Hans-Peter Blossfeld, Alfred Hamerle,
Karl Ulrich Mayer

Ereignisanalyse

Statistische Theorie und Anwendung in den Wirtschafts- und Sozialwissenschaften

Campus Verlag
Frankfurt/New York

CIP-Kurztitelaufnahme der Deutschen Bibliothek

Blossfeld, Hans-Peter:
Ereignisanalyse : statist. Theorie u. Anwendung in d. Wirtschafts- u. Sozialwiss. / Hans-Peter Bloss= feld ; Alfred Hamerle ; Karl Ulrich Mayer. - Frankfurt/Main ; New York : Campus Verlag, 1986.
(Campus : Studium ; Bd. 569)
ISBN 3-593-32569-1

NE: Hamerle, Alfred:; Mayer, Hans Ulrich:; Campus / Studium

Umschlaggestaltung: Atelier Warminski, Büdingen
Druck und Bindung: Beltz Offsetdruck, Hemsbach
Printed in Germany

Vorwort

In den empirisch forschenden Disziplinen, vor allem in den Wirtschafts- und Sozialwissenschaften, ist das Interesse an der Analyse von Prozessen und Verläufen neu erwacht. Dabei treten im Vergleich zu den herkömmlichen Panel- oder Zeitreihenstudien zunehmend ereignisorientierte Datenstrukturen in den Vordergrund, die dem Wandel und der Dynamik empirischer Phänomene besser gerecht werden können. Für jede Untersuchungseinheit informieren sie über die genauen Zeitdauern bis zu einem Zustandswechsel beziehungsweise bis zum Eintreffen bestimmter Ereignisse und über deren Abfolge. Beispiele hierfür sind die Überlebenszeiten von Patienten in medizinischen Studien; die Arbeitslosigkeitsphasen in ökonomischen Untersuchungen; die „Lebensdauer" politischer Systeme im Bereich der Politologie; die Zeitspanne, während deren ein technisches Gerät störungsfrei arbeitet, auf dem Gebiet der Qualitäts- und Materialprüfung; die Dauer von Lernprozessen in der psychologischen Forschung; die Zeitspanne bis zum Umzug in eine andere Region bei Wanderungs-und Mobilitätsanalysen; der Zeitraum bis zur Rückfälligkeit von Straftätern in kriminologischen Untersuchungen; die Verweildauer der Kinder im elterlichen Haushalt bis zum Auszug in der Jugend- und Familiensoziologie usw.

Die Darstellung der statistischen Theorie und der konkreten Anwendung der Ereignisanalyse in diesem Buch ist deshalb für einen breiten Leserkreis von Bedeutung. Sie kommt aber in den ausgewählten Beispielen vor allem den Bedürfnissen der modernen Wirtschafts- und Sozialforschung entgegen.

Der Text ist als Lehrbuch für die Vermittlung der Ereignisanalyse an Studenten oder zum Selbststudium sowie als Handbuch und Nachschlagewerk für den Anwender in der Forschung geschrieben worden. Neben der Behandlung der statistischen Grundlagen der Ereignisanalyse haben wir ganz besonderen Wert darauf gelegt, den gesamten Forschungsweg von den Problemen der Erhebung ereignisorientierter Daten über die spezifischen Fragen der Datenorganisation bis hin zur konkreten Anwendung von EDV-Programmen und der Interpretation der Analysebefunde zu vermitteln.

Verglichen mit anderen, meist englischsprachigen Lehrbüchern zu diesem Gebiet der angewandten Statistik ist es die besondere Absicht des vorliegenden Buches, die Möglichkeiten des Einbezugs von Kovariablen in semiparametri-

schen und parametrischen Regressionsmodellen anhand vieler konkreter Beispiele didaktisch nachvollziehbar zu machen, dies aber in enger Verbindung mit einer unverkürzten Darstellung der statistischen Theorie. Daneben werden parameterfreie Verfahren zur Analyse von Ereignisdaten und die Möglichkeit ihrer graphischen Präsentation ausführlich mit Anwendungsbeispielen behandelt. Großer Raum wird den spezifischen Problemen von Mehr-Zustands- und Mehr-Episoden-Modellen, der Aufnahme zeitveränderlicher Kovariablen und den Fragen der unbeobachteten Populationsheterogenität gegeben. Detaillierte Beispiele demonstrieren die Überprüfung von Modellannahmen, die Durchführung von Hypothesentests und die Auswahl geeigneter Modelle.

Als Datengrundlage für die anwendungsorientierten Beispiele diente die Lebensverlaufsstudie aus dem Sonderforschungsbereich 3 „Mikroanalytische Grundlagen der Gesellschaftspolitik" der Deutschen Forschungsgemeinschaft, die gegenwärtig vom Max-Planck-Institut für Bildungsforschung in Berlin durchgeführt wird.

Das vorliegende Lehrbuch ist im Rahmen des Projekts „Lebensverläufe und gesellschaftlicher Wandel" am Max-Planck-Institut für Bildungsforschung in Berlin in Zusammenarbeit mit der Abteilung für Statistik an der Universität Konstanz entstanden. Wir danken dem Max-Planck-Institut für Bildungsforschung für die materielle Unterstützung bei der Erstellung der Druckvorlage.

Unser besonderer Dank gilt Frau Doris Gampig, die sehr professionell und mit großer Akribie den Satz des Buches erstellt hat. Danken möchten wir auch Herrn Axel Rieke und Herrn Dieter Schmidt für die Aufbereitung der Abbildungen, Tabellen und Programmbeispiele. Herrn Gottfried Pfeffer danken wir für die redaktionelle Bearbeitung des Manuskripts und die Betreuung der Verlagsbeziehungen. Herrn Gerhard Tutz von der Universität Regensburg danken wir für die Durchsicht von Teilen des Manuskripts. Frau Bettina Althainz und Herr Peter Baumann haben als wissenschaftliche Hilfskräfte mit großem Engagement bei der Vorbereitung der Anwendungsbeispiele mitgeholfen und viele der notwendigen Rechenläufe durchgeführt. Schließlich danken wir Herrn Trond Petersen von der Harvard Universität in Cambridge für die Erlaubnis, sein BMDP-Unterprogramm P3RFUN dokumentieren zu dürfen.

Berlin und Konstanz, im April 1986

Hans-Peter Blossfeld
Alfred Hamerle
Karl Ulrich Mayer

Inhaltsverzeichnis

Kapitel 1:
Ziel und Aufbau des Buches

Dieses Buch soll eine zusammenfassende Darstellung der wichtigsten Methoden der Ereignisanalyse bieten. Mit dem Begriff Ereignisanalyse bezeichnen wir statistische Verfahren zur Untersuchung von Zeitintervallen zwischen aufeinanderfolgenden Zustandswechseln beziehungsweise Ereignissen. Die von den Untersuchungseinheiten eingenommenen Zustände sind dabei abzählbar, und die Ereignisse können zu beliebigen Zeitpunkten eintreten. Es handelt sich also um statistische Verfahren zur Analyse stochastischer Prozesse mit diskreten Zuständen und stetiger Zeit.

Die Statistik bietet heute eine Fülle von Modellen, Ansätzen und Methoden zur Analyse von Ereignisdaten an, die allerdings in den gängigen Statistiklehrbüchern kaum enthalten sind. Der Grund dafür liegt zum einen darin, daß in der Ereignisanalyse mit stochastischen Modellen gearbeitet wird, die bisher in der normalen Anwendung keine große Rolle gespielt haben, und zum anderen darin, daß unvollständige Stichproben (mit zensierten Daten) nur in bestimmten Problemzusammenhängen vorkommen. Auch aufgrund der Entwicklung und Anwendung dieser Verfahren in verschiedenen Disziplinen wie Medizin, Demographie, Technik oder Wirtschafts- und Sozialwissenschaften ist die Terminologie sehr uneinheitlich und deshalb dem Anwender nicht leicht zugänglich.

Im Vordergrund des vorliegenden Textes stehen deswegen sowohl die systematische Aufarbeitung der statistischen Grundlagen der Ereignisanalyse als auch ihre konkrete Umsetzung in die Forschungspraxis. In enger Verbindung mit einer vereinheitlichenden Darstellung der statistischen Theorie wird anhand konkreter Beispiele aus der Forschungspraxis die Anwendung der Ereignisanalyse aufgezeigt.

Nach diesem Überblick (Kapitel 1) soll im 2. Kapitel zunächst auf dreierlei Weise die spezifische Art der Problemkonzeption und der Problemlösungsfähigkeit der Ereignisanalyse veranschaulicht werden. Wir zeigen zuerst an einer fachlich breiten Palette das Anwendungsspektrum der Ereignisanalyse (Abschnitt 2.1) und erläutern anschließend die methodischen Besonderheiten der Lebensverlaufsstudie aus dem Sonderforschungsbereich 3 der Deutschen Forschungsgemeinschaft, der die empirischen Anwendungsbeispiele in den Kapiteln 4 bis 6

entnommen sind (Abschnitt 2.2). Danach werden in allgemeiner Form die Vorzüge der ereignisorientierten Datenstruktur im Vergleich zu Querschnitts- und traditionellen Paneldaten aufgezeigt (Abschnitt 2.3).
Das Kapitel 3 hat dann die statistischen Grundlagen zum Gegenstand. Neben der Einordnung der Ereignisanalyse in den Rahmen der stochastischen Prozesse werden die Grundkonzepte der Ereignisanalyse wie zum Beispiel Hazardrate, Survivorfunktion, kumulative Hazardrate usw. definiert (Abschnitt 3.1) und die nichtparametrischen Schätzverfahren wie Sterbetafel-Methode und Kaplan-Meier-Schätzung behandelt (Abschnitt 3.2). Von besonderer Bedeutung für das gesamte Lehrbuch ist der Abschnitt 3.3, in dem die Einbeziehung erklärender Variablen in semiparametrische Cox-Modelle und parametrische Modelle wie zum Beispiel das Exponential-, das Weibull-, das Gompertz-(Makeham-) und das log-logistische Modell dargestellt werden. Die allgemeine Theorie der Mehr-Zustands- und Mehr-Episoden-Fälle steht im Mittelpunkt der Abschnitte 3.4 und 3.5. Danach werden die Maximum-Likelihood-Schätzung der unbekannten Modellparameter (Abschnitt 3.6), Verfahren zur Konstruktion von Hypothesentests sowie die Probleme der Modellauswahl (Abschnitt 3.7), die Einbeziehung zeitabhängiger Kovariablen (Abschnitt 3.8) und Modelle mit expliziter Berücksichtigung unbeobachteter Populationsheterogenität (Abschnitt 3.9) behandelt. Den Abschluß des theorieorientierten 3. Kapitels bildet schließlich ein kurzer Abriß von Hazardraten-Modellen in diskreter Zeit.
Die Kapitel 4 bis 6 wenden sich vor allem an die Anwender der Ereignisanalyse in der Forschung. Man kann sie aber auch als Werkbuch benutzen, um in die empirische Analyse von Berufs- und Erwerbsverläufen im Rahmen der Arbeitsmarktforschung einzuführen. Auf der Grundlage der Lebensverlaufsstudie des Sonderforschungsbereichs 3 der Deutschen Forschungsgemeinschaft werden schrittweise die Strategien der Aufbereitung und Auswertung von Ereignisdaten dargestellt. Anhand konkreter Beispiele wird gezeigt, wie inhaltliche Fragestellungen methodisch-statistisch umgesetzt werden, welche EDV-Programmpakete (SPSS, BMDP, GLIM, RATE) für welche Analysezwecke zur Verfügung stehen, wie die Steuerkarten jeweils aufgebaut sein müssen und wie die Ergebnisse der Auswertungsläufe zu interpretieren und zu bewerten sind.
Zunächst werden in Kapitel 4 nach einem Blick auf die programmtechnische Aufbereitung ereignisorientierter Datenstrukturen (Abschnitt 4.1) verschiedene Möglichkeiten ihrer graphischen Präsentation gezeigt (Abschnitt 4.2). Anschließend wird die Anwendung der Sterbetafel-Methode und des Kaplan-Meier-Schätzers vorgeführt (Abschnitt 4.3).
Im Zentrum des 5. Kapitels steht die Anwendung des Cox-Modells und der Partial-Likelihood-Schätzung. Nach den Überprüfungsmöglichkeiten der Proportionalitätsannahme (Abschnitt 5.1), wird ausführlich die Interpretation eines Cox-Modells aufgezeigt (Abschnitt 5.2). Die Modellauswahl mit Hilfe der schrittweisen Regression wird in Abschnitt 5.3 demonstriert. Besonders wichtig für die Anwendung der Ereignisanalyse in den Wirtschafts- und Sozialwissen-

schaften sind die Beispiele dafür, wie zeitveränderliche unabhängige Variablen in das Cox-Modell aufgenommen (Abschnitt 5.4) und wie Mehr-Zustands-Fälle praktisch gehandhabt werden (Abschnitt 5.5).
Das Kapitel 6 schließlich ist der Anwendung parametrischer Modelle gewidmet. Nach der graphischen Überprüfung der Verteilungsannahmen (Abschnitt 6.1) wird zuerst ausführlich auf das Exponential-Modell, auf seine Interpretation und auf die Überprüfung der Residuen eingegangen (Abschnitt 6.2). Danach folgen Beispiele zur Aufnahme zeitveränderlicher unabhängiger Variablen mit Hilfe des Episodensplittings (Abschnitt 6.3) und Beispiele zu Modellen mit periodisierter Verweildauer (Abschnitt 6.4). Spezielle Verweildauer-Modelle werden im Abschnitt 6.5 dargestellt. Dabei werden ausführliche Interpretationsbeispiele und Residuentests zur Gompertz-(Makeham-) (Abschnitt 6.5.1), zur Weibull- (Abschnitt 6.5.2) und zur log-logistischen Verteilung (Abschnitt 6.5.3) gegeben. Anwendungsbeispiele zur unbeobachteten Populationsheterogenität beschließen die Ausführungen zu den parametrischen Modellen (Abschnitt 6.6).
Mit dem 7. Kapitel schließt das Buch. In diesem Kapitel werden einige noch nicht gelöste Probleme bei der Anwendung der Ereignisanalyse und ihrer statistischen Grundlagen skizziert.

Kapitel 2:
Zum Anwendungsfeld der Ereignisanalyse

In den Wirtschafts- und Sozialwissenschaften gibt es gute Gründe dafür, Prozesse und Verläufe zu untersuchen. Zunächst erfordert eine angemessene Abbildung der Wirklichkeit die systematische Beschreibung von Veränderungsprozessen und Wandlungstendenzen. Dies ist freilich nicht neu, aber das Interesse daran wächst noch in einer Zeit, die als Wendepunkt mittel- und langfristiger ökonomischer und gesellschaftlicher Entwicklungen betrachtet wird. Neueren Datums ist hingegen die Erkenntnis, daß angemessene Erklärungen auf der Grundlage von Querschnittsdaten nur in den relativ seltenen Fällen gewonnen werden können, in denen sich Prozesse in einem Gleichgewichtszustand befinden. In allen anderen Situationen hingegen können Prozesse nur mit Längsschnittdaten zuverlässig erfaßt werden; und nur der Realität angemessene Prozeßmodelle können Grundlage für rationale politische Interventionen sein.
In der Vergangenheit waren im Bereich der Wirtschafts- und Sozialwissenschaften die Möglichkeiten, Prozesse zu messen und durch mathematische Modelle abzubilden, sowohl von den verfügbaren Daten her als auch aufgrund des vorhandenen statistischen Instrumentariums eng begrenzt. Das Instrument der Differentialgleichungssysteme erfordert kontinuierlich gemessene Variablen in stetiger Zeit, die in den Wirtschafts- und Sozialwissenschaften nur selten verfügbar sind. Zwei- und Mehr-Panelwellen-Studien erfassen – wie wir in Abschnitt 2.3 zeigen werden – die Prozesse im Zeitablauf nur unvollständig und in der Regel durch die extern vorgegebenen Erhebungszeitpunkte verzerrt. Zeitreihenanalysen und die Mehrzahl der ökonometrischen Modelle erfordern dagegen viele Meßpunkte.
In zunehmendem Maße werden jedoch heute ereignisorientierte Daten neu erhoben oder zugänglich gemacht, bei denen die Zustandswechsel von Untersuchungseinheiten mit ihren genauen Zeitpunkten registriert sind. Solche Datenstrukturen informieren über die genauen Zeitdauern bis zum Eintreffen von Ereignissen und über deren Abfolge. Zusätzlich zu diesen Verweildauern beziehungsweise Wartezeiten interessieren häufig Variablen, die einzeln oder in Kombination die Zeiten bis zum Auftreten eines Ereignisses beeinflussen. Das können zeitstabile Merkmale sein oder Merkmale, die sich im Zeitablauf verändern.

2.1 Anwendungsbeispiele

Im folgenden soll an einer Reihe von Beispielen die spezifische Art der Problemkonzeption und der Problemlösungsfähigkeit der Ereignisanalyse aufgezeigt werden. Dabei soll deutlich werden, daß sich die Ereignisanalyse für ein breites Anwendungsspektrum eignet.

Beispiel 1: Arbeitslosigkeitsstudien

In den Wirtschaftswissenschaften wurden Ereignisanalysen bereits häufig zur Untersuchung der Arbeitslosigkeit herangezogen (Heckman/Borjas 1980; Flinn/Heckman 1983; Heckman/Singer 1982, 1984a; Hamerle 1985c). Hintergrund dieser Untersuchungen ist die Erkenntnis, daß zur Analyse der Arbeitslosigkeit Querschnitte nur bedingt aufschlußreich sind, da sie es unter anderem nicht erlauben, zwischen Kurzzeit- und Langzeitarbeitslosen zu unterscheiden und zeitveränderliche Variablen einzubeziehen.
In den Arbeitslosigkeitsstudien bilden die aufeinanderfolgenden Arbeitslosigkeitsphasen eines Arbeitnehmers die Verweildauern. Arbeitslosigkeitsphasen können ferner aus verschiedenen Gründen beendet werden, so zum Beispiel durch die Wiederaufnahme einer neuen Erwerbstätigkeit, durch Arbeitsbeschaffungsmaßnahmen, Umschulung, Verrentung oder die Zuerkennung einer Erwerbsunfähigkeit. Solche unterschiedlichen Zielzustände lassen sich als „competing-risks-" oder Mehr-Zustands-Modelle untersuchen.

Beispiel 2: Studien zum Konsumentenverhalten

Breite Anwendungsmöglichkeiten der Ereignisanalyse liegen auf dem Gebiet der Konsumentenforschung. Auf einem Markt werden verschiedene Produktmarken angeboten. Konsumenten wählen und erwerben eine dieser Marken, kaufen dann zu einem späteren Zeitpunkt diese wieder oder wechseln zu einer anderen Marke über. In diesem Beispiel entspricht eine Episode oder Verweildauer der Nutzungsdauer eines Produkts. Die Zustände werden durch die verschiedenen Marken indiziert.
Bei den in diesem Buch vorgestellten Methoden kann die Nutzungsdauer auch in Verbindung mit exogenen Einflußgrößen untersucht werden, von denen sich einige im Laufe der Zeit ändern können. Bei derartigen Faktoren kann es sich um demographische Merkmale (z. B. Alter, Geschlecht, Familienstand, Haushaltsgröße), sozio-ökonomische Merkmale (z. B. Einkommen, Schulbildung, Berufstätigkeit, soziale Schicht), geographische Merkmale (Großstadt, Land) oder psychologische Bedingungen (z. B. Einstellungen, Präferenzen, Preisbewußtsein, Qualitätsbewußtsein, Kaufgewohnheiten) handeln. Darüber hinaus können für die Dauer der Nutzung eines Gutes frühere Erfahrungen mit dem Produkt von Bedeutung sein. Die Einbeziehung von Daten der Vorgeschichte des Prozesses ist mit den in diesem Buch vorgestellten Verfahren ebenfalls möglich.

Beispiel 3: Studien über Krankheitsverläufe

In den letzten Jahren wurden wichtige Anwendungen der in diesem Buch zu besprechenden Verfahren zur Analyse von Heilungs- und Überlebenszeiten in medizinischen und epidemiologischen Studien eingesetzt (vgl. zum Beispiel Kalbfleisch/Prentice 1980 und die dort angeführten Beispiele). Dabei handelt es sich in der Regel um Untersuchungen mit einem oder mehreren absorbierenden Zielzuständen. Absorbierend bedeutet, daß der betreffende Zielzustand nicht mehr verlassen werden kann, wie das beispielsweise beim Tod eines Patienten der Fall ist.
Empirische Mehr-Episoden-Modelle sind im medizinischen Bereich bislang noch selten, obwohl sie häufig angemessen wären. So besteht der Verlauf einer Krankheit meist aus einer Abfolge verschiedener Krankheitsstadien, die gekennzeichnet sind durch Ereignisse wie Remission, Rezidiv oder Tod. Hamerle (1985c) hat zum Beispiel für Patientinnen nach einer Brustkrebsoperation den krankheitsfreien Verlauf bis zum Auftreten der konkurrierenden Risiken „Auftreten von Metastasen" und „Tod" beziehungsweise nach dem Auftreten von Metastasen die Überlebenszeit bis zum Tod untersucht. Interessant an diesem Beispiel ist unter anderem, daß die Zeit bis zum Wiederauftreten von Metastasen nach der Operation ein besonders guter Prädiktor für die Überlebenszeit danach zu sein scheint. Es zeigte sich außerdem in diesem Beispiel, daß getrennte Untersuchungen der einzelnen Phasen mit Ein-Episoden-Modellen nicht problemadäquat wären, da sie die gegenseitige

Abhängigkeitsstruktur der Ereignisse sowie ihre zeitliche Abfolge nicht berücksichtigen würden. In diesem Buch werden Hinweise gegeben, wie solche Spezialfälle behandelt werden können.

Beispiel 4: Lernexperimente in der Psychologie und der Unterrichtsforschung

In der Lernpsychologie können Ereignisanalysen Aufschluß geben über den zeitlichen Verlauf des Lernverhaltens. Die Verweildauern sind hier die Zeitspannen, die zum Lernen eines bestimmten Sachverhalts benötigt werden. Auch hier kann die Lerngeschwindigkeit wiederum in Abhängigkeit von personen- oder umweltbezogenen Faktoren betrachtet werden.
In der Unterrichtsforschung sind Ereignisanalysen zum Beispiel auch für die Auswertung von Videoaufzeichnungen über Aufmerksamkeitsspannen bei Schülern verwandt worden. Dabei sollte geklärt werden, ob Gruppenbildung nach Leistungsniveau innerhalb von Klassen die Aufmerksamkeit beeinflußt (Felmlee/Eder 1983).

Beispiel 5: Untersuchungen in der Unfallforschung

In der Unfallforschung können die Verfahren der Ereignisanalyse unter anderem zur Untersuchung des bedingten Unfallrisikos in Abhängigkeit von der Zeit und von möglichen Risikofaktoren herangezogen werden. Die Verweildauer kann hier beispielsweise durch die Zeit des unfallfreien Fahrens eines Kraftfahrzeugs gegeben sein. Mögliche erklärende Variablen wären Alter, Anzahl der gefahrenen Kilometer, Verkehrskontext, Fahrzeugtyp usw.

Beispiel 6: Studien über regionale Wanderungen

Auch zur Untersuchung von Wohnungswechseln und Wanderungen zwischen Regionen eignet sich die Ereignisanalyse besonders gut. Die Zustände werden hier durch Wohnungsgrößen oder Regionen und die Episoden durch die jeweiligen Aufenthaltsdauern repräsentiert. Die Wanderungsraten können in bezug auf verschiedene Faktoren und Motive wie zum Beispiel Verdienstmöglichkeiten, Wohnungsversorgung, Siedlungs- und Infrastruktur, Eigentums- und Besitzverhältnisse, Freizeitwert usw. untersucht werden. Wichtig ist dabei die Unterscheidung, ob Wanderungen durch veränderliche Ressourcen von Personen beeinflußt oder durch andere Lebensereignisse wie Arbeitsplatzwechsel, Heirat oder Geburt eines Kindes ausgelöst werden (Courgeau 1984; Mayer/Wagner 1986; Wagner 1986).

Beispiel 7: Analysen der Familienbildung und der Geburtenentwicklung

Als Untersuchungsgegenstand der Ereignisanalyse eignen sich die Heirat, die Geburt von Kindern sowie das Scheidungsverhalten besonders gut, obwohl die bevölkerungswissenschaftliche Forschung bislang nur die einfache Sterbetafel-Methode anwendet und nur zögernd von ihren Aggregatdaten auf individuelle Verläufe übergeht. Bei diesen Prozessen ist von wesentlicher Bedeutung, wie sie vom Lebensalter beziehungsweise von der Ehedauer abhängen und wie sich die „Risiken" über die Zeit verteilen. So ist das Scheidungsrisiko unmittelbar nach der Eheschließung gering, steigt nach wenigen Ehejahren an und fällt danach monoton ab. Es stellt sich in diesem Falle zunächst das Problem, den richtigen Funktionsverlauf für diesen Zusammenhang zu finden. Wie solche parametrischen Modelle ausgewählt werden und ob es besser ist, auf Verfahren zurückzugreifen, die die zeitliche Entwicklung des Risikos unspezifiziert lassen, wird später noch ausführlich erörtert.
Am Scheidungsbeispiel läßt sich auch das Problem „heterogener Populationen" illustrieren. So können zu einer Bevölkerung auch Gruppen gehören, welche einem Scheidungsrisiko – zum Beispiel aufgrund religiöser Überzeugungen – überhaupt nicht ausgesetzt sind (Diekmann/Mitter 1984).

Beispiel 8: Studien aus der Kriminologie und Rechtstatsachenforschung

In der Kriminologie wird die Ereignisanalyse beispielsweise zur Untersuchung der Neigung zur Rückfälligkeit bei Straftätern, die aus der Haft entlassen wurden, angewandt. Die Verweildauer wird durch die Zeitspanne zwischen Haftentlassung und neuerlicher Straffälligkeit definiert und kann zu den Strafvollzugsbedingungen, zu bestimmten Resozialisierungs- und Betreuungsmaßnahmen oder zur Einkommens- und Wohnsituation in Bezug gesetzt werden. Auch die Anzahl und Dauer vorheriger Inhaftierungen können zur Erklärung herangezogen werden (Diekmann 1980). Ein anderes Gebiet stellt die Rechstatsachenforschung dar. Dort kann zum Beispiel die Dauer bis zur Abwicklung von Zivil- oder Strafgerichtsverfahren in Abhängigkeit von Verfahrens- und Richtermerkmalen oder in bezug auf gesetzliche Veränderungen analysiert werden.

Beispiel 9: Organisations- und Unternehmensforschung

Die Überlebenszeit politischer Regime, von Firmen, Arbeitsgruppen und ähnlichem bietet ein weiteres wichtiges Anwendungsgebiet für die Ereignisanalyse, insbesondere im Bereich des neuen Theorie- und Forschungszweigs der „organization ecology" (Carroll/Huo 1985, 1986; Hannan/Freeman 1977; Freeman/Carroll/Hannan 1983; Carroll 1984; Carroll/Delacroix 1982). Dabei wurden bislang insbesondere die Entwicklungsdynamik von Zeitungen, Restaurants und lokalen Gewerkschaftsorganisationen im 19. Jahrhundert untersucht.

Diese Beispiele, die nur einige Anregungen zum Einsatz von Ereignisanalysen in empirisch forschenden Disziplinen geben sollten, sind natürlich bei weitem nicht vollständig. Viele weitere Anwendungsmöglichkeiten sind denkbar, vor allem in der Technik. Bei Zuverlässigkeitsuntersuchungen erscheint die Ermittlung des Einflusses zeitveränderlicher Kovariablen auf die Lebensdauer, insbesondere bei Lebensdauertests unter besonderer Materialbelastung oder allgemein „accelerated life tests", erfolgversprechend. Für den wirtschafts- und sozialwissenschaftlichen Bereich wird in den Kapiteln 4 bis 6 anhand weiterer konkreter Beispiele die praktische Umsetzung der Methoden der Ereignisanalyse vorgeführt.

2.2 Zur Lebensverlaufsstudie

Die rasch steigende Nachfrage nach den Methoden der Ereignisanalyse im Bereich der Wirtschafts- und Sozialwissenschaften ist eng mit dem wachsenden Interesse an der Erforschung von Lebensverläufen verbunden. Diese Entwicklung wird durch eine Vielzahl von Forschungsprojekten dokumentiert, in deren Verlauf mehrere umfangreiche, ereignisorientierte Datenbestände entstanden sind und entstehen.

In der Bundesrepublik Deutschland sind hier beispielsweise, neben der Lebensverlaufsstudie des Max-Planck-Instituts für Bildungsforschung und des DFG-Sonderforschungsbereichs 3 (Mayer 1984a, 1984b), folgende Vorhaben zu nennen: das Sozio-ökonomische Panel, das ebenfalls im Rahmen des DFG-Sonderforschungsbereichs 3 durchgeführt wird und am Deutschen Institut für Wirtschaftsforschung in Berlin angesiedelt ist (Hanefeld 1984; Krupp 1985); die Gymnasiasten-Wiederholungsbefragung von Meulemann und Wiese am Zentralarchiv in Köln (Meulemann u. a. 1984) und schließlich die Projekte „Generatives Verhalten in Nordrhein-Westfalen" (Strohmeier/Schultz/Kaufmann 1985) und „Arbeitsmarktdynamik, Familienentwicklung und generatives Verhalten" (Birg u. a. 1985), beide am Institut für Bevölkerungsforschung und Sozialpolitik der Universität Bielefeld.

Während die meisten dieser Vorhaben sich noch in einem relativ frühen Stadium befinden, liegen aus internationalen Vergleichsprojekten bereits eine Reihe von Auswertungsbeispielen vor. Dies gilt insbesondere für die Analysen der amerikanischen, norwegischen, französischen und israelischen Lebensverlaufsunter-

suchungen (Featherman/Sørensen 1983; Matras 1983; Courgeau 1984; Michael/Tuma 1985).

Allen diesen Projekten ist gemeinsam, daß sie sich in *dynamischer Perspektive* dem *Lebensverlauf* zuwenden, der in bestimmte *historische Zeitperioden eingebettet* ist; die beobachteten Veränderungen beschränken sich nicht nur auf einen Lebensbereich, sondern beziehen sich auf eine Vielzahl von Dimensionen in *mehreren Lebensbereichen*; und die Lebensverläufe werden retrospektiv, prospektiv oder auf der Grundlage prozeßproduzierter Daten beobachtet, wobei die *Veränderung von diskreten Zuständen in kontinuierlicher Zeit erhoben* wird.

Als zunehmend eigenständige Forschungsrichtung ist die Lebensverlaufsforschung aus der Fortentwicklung und Konvergenz einer Reihe von Theorieansätzen und empirisch-methodischen Arbeitsgebieten entstanden. Dazu sind zu zählen die Forschungen zur sozialen und beruflichen Mobilität, zur Fertilitätsgeschichte und zum Familienzyklus, zu Ausbildungs- und Berufsverläufen, zur Qualifikations- und Absorptionsforschung, zur Humankapitalbildung und Einkommensentstehung, zum Arbeitsmarktverhalten und zur Frauenerwerbstätigkeit, zur Binnenwanderung, zur Soziologie des Alterns, zum Generationenwandel und zur Kohortendifferenzierung. Wichtige Anstöße kamen ferner aus der Sozialindikatorenbewegung, der Demographie und der historischen Familienforschung (Mayer 1986).

In dieser Tradition ist es das allgemeine Ziel der in diesem Buch zu Demonstrationszwecken herangezogenen Studie „Lebensverlauf und gesellschaftlicher Wandel" des Max-Planck-Instituts für Bildungsforschung und des DFG-Sonderforschungsbereichs 3, die im folgenden immer als *Lebensverlaufsstudie* bezeichnet wird (Mayer 1984a, 1984b, 1986), die neuere Sozialgeschichte seit dem Ende des Zweiten Weltkrieges zu rekonstruieren und die Auswirkungen gesellschaftlicher Institutionen, insbesondere des Bildungssystems, des Beschäftigungssystems und der Familie, auf individuelle Lebensverläufe zu untersuchen. Konkret sollen mit der Studie unter anderem folgende Fragen beantwortet werden: Wie sehen die Prozesse der Familienbildung aus und in welchem Maße haben sich die Lebensverläufe von Frauen verändert? Gibt es in unserer Gesellschaft altersnormierte Prozesse und wie strukturieren diese die individuellen Lebensverläufe? Wie hat sich das Verhältnis von Bildungs- und Beschäftigungssystem gewandelt und wie schlägt sich dieser Wandel in den Karriereverläufen unterschiedlicher Geburtskohorten nieder? Welche quantitative Bedeutung kommt im Lebenslauf bestimmten Wanderungspfaden zu und welches Ausmaß an räumlicher Mobilität gibt es im Verlauf des Lebens?

Da es allerdings das Hauptziel dieses Buches ist, zu zeigen, wie man schrittweise bei der Analyse eines konkreten Ereignisdatensatzes vorgeht, werden sich die Analysebeispiele hauptsächlich auf ein Gebiet der Lebensverlaufsstudie konzentrieren, und zwar auf die Untersuchungen über die „Strukturen und Bedingungen von Berufsverläufen".

Dazu steht uns mit der Lebensverlaufsstudie ein Fundus von detaillierten Verlaufsinformationen über 2171 Frauen und Männer aus den Geburtsjahrgängen

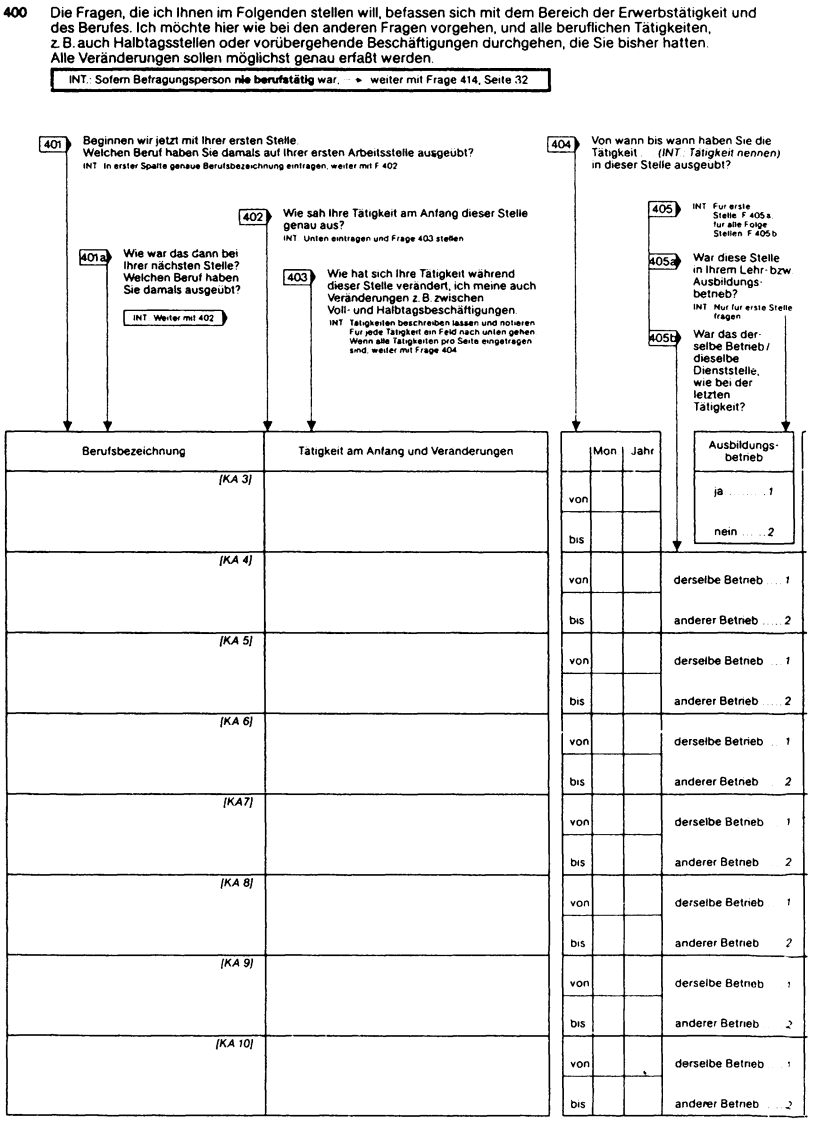
1929–31, 1939–41 und 1949–51 zur Verfügung, die in den Jahren 1981 bis 1983 erhoben worden sind. Die Geburtskohorten wurden dabei strategisch so gewählt, daß die Phase des Übergangs vom Bildungs- in das Beschäftigungssystem jeweils in besondere historische Perioden fällt. So liegt bei den um 1930 Geborenen diese Übergangsphase in der unmittelbaren Nachkriegszeit, bei den um 1940 Geborenen in der Periode des extensiven Wirtschaftswachstums und bei den um 1950 Geborenen in der Phase des Ausbaus des Wohlfahrtsstaates. Dahinter steht die Hypothese, daß diese besonderen historischen Bedingungen in der Berufseinstiegsphase einen prägenden Einfluß auf den späteren Erwerbsverlauf haben.

Die Lebensverläufe der Lebensverlaufsstudie wurden *retrospektiv im ereignisorientierten Erhebungsdesign* erfaßt. Dieses Schema verdeutlicht ein Ausschnitt aus dem Fragebogen, in dem die Erwerbstätigkeitsepisoden abgefragt worden sind (Abbildung 2.1). Kennzeichnend ist, daß zu jeder einzelnen Berufstätigkeit, neben einer Reihe theoretisch interessanter Informationen wie Branche, Anzahl der geleisteten Arbeitsstunden, Einkommen usw., die Anfangs- und Endzeitpunkte auf den Monat genau festgehalten wurden. In Verbindung mit den ebenfalls im ereignisorientierten Erhebungsdesign erfaßten Ausbildungs- und Unterbrechungszeiten ist so der *Bildungs- und Berufsverlauf eines Individuums lückenlos rekonstruierbar*. Für jedes Individuum und für jeden Zeitpunkt des Beobachtungszeitraums steht dann eine exakte Zustandsinformation über den Karriereprozeß zur Verfügung.

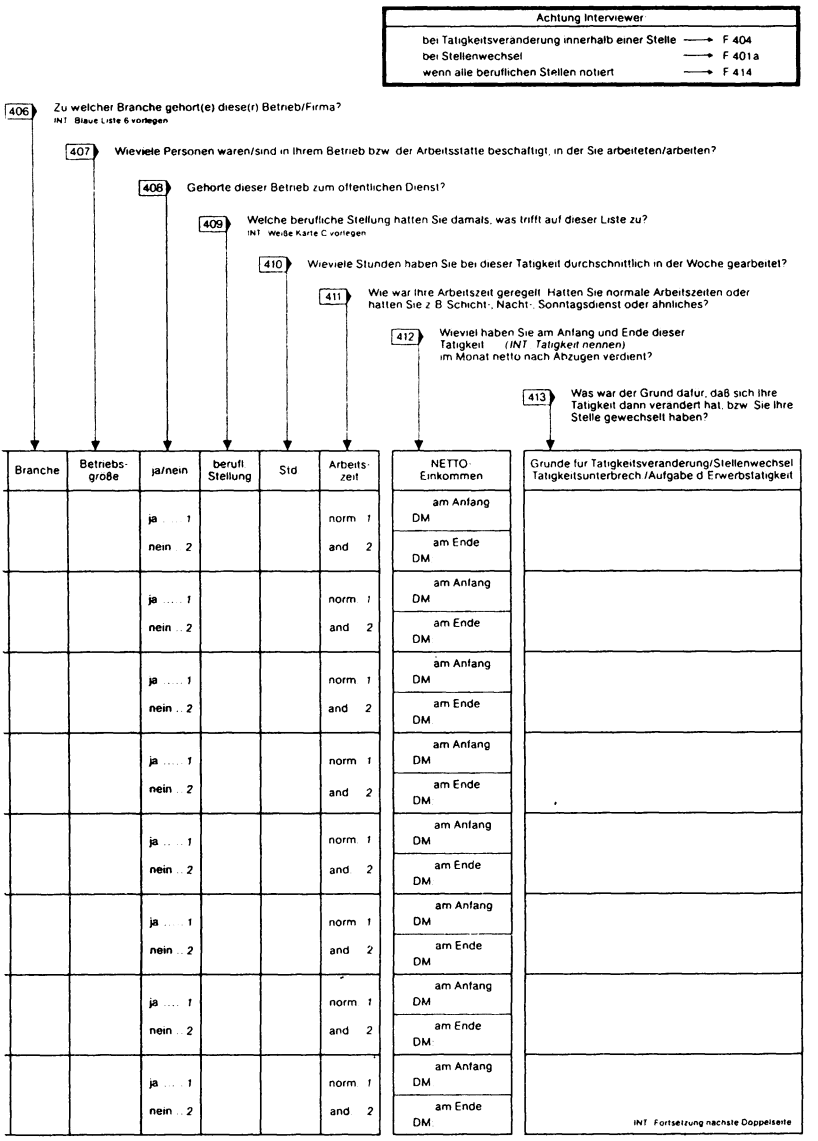
Obgleich solchen Rückerinnerungsdaten häufig eine höhere Fehlerhaftigkeit zugeschrieben wird, ließ eine vor der eigentlichen Stichprobenziehung durchgeführte Vorstudie deutlich erkennen, daß die Zuverlässigkeit retrospektiv erhobener Daten zur objektiven Lebensgeschichte nicht einschneidend durch eine mangelnde Antwortbereitschaft und Erinnerungsfähigkeit beeinträchtigt wird (Papastefanou 1980; Tölke 1980). Außerordentlich wichtig für die Qualität der Antworten ist allerdings die Art und Präzision des Erhebungsinstruments. Als besonders angemessen hat sich das jeweils getrennte Abfragen der Lebensgeschichten in den verschiedenen Bereichen (Bildung, Beruf, Wohnen usw.) erwiesen. Langwierige und aufwendige Dateneditionen, Datenrecherchen und Quervergleiche bürgen darüber hinaus für die Güte der erhobenen Informationen (Brückner u. a. 1984). Schließlich zeigte eine Überprüfung der repräsentativen Qualität der Lebensverlaufsdaten mit Hilfe von Zensus- und Mikrozensuserhebungen, daß die sozialstrukturellen Querschnitte aus der Vergangenheit außerordentlich gut von der Lebensverlaufsstudie wiedergegeben werden können (Blossfeld 1985a).

Da die Daten nicht nur den Bildungs- und Berufsverlauf umfassen, sondern darüber hinaus die gesamte *Breite der verschiedenen Lebensbereiche* für Analysezwecke verfügbar machen (d. h. Informationen über die soziale Herkunft, die Familiengeschichte, die Geschichte des Ehepartners, die Wohngeschichte usw.), können die Effekte von Ereignissen aus anderen *parallelen Prozessen* (z. B. bei der Familiengeschichte das Ereignis Heirat) auf die Erwerbskarriere (z. B. die

Abbildung 2.1: Beispiel eines ereignisorientierten Erhebungsdesigns



Stabilität von Berufsverläufen) analysiert werden. Gleiches gilt auch für die *Vorgeschichte*, die im Hinblick darauf untersucht werden kann, inwieweit sie den späteren Berufsverlauf präformiert und in bestimmte Bahnen lenkt. Insge-



samt bietet diese Datenbasis sehr gute Voraussetzungen, um an ihrem Beispiel den Gang des Analyseprozesses bei Ereignisdaten mit allen methodisch interessanten Varianten darzustellen.

2.3 Vorzüge der ereignisorientierten Datenstruktur

Worin bestehen eigentlich die Vorzüge, die eine ereignisorientierte Datenstruktur für die moderne Wirtschafts- und Sozialforschung so außerordentlich attraktiv machen? Wenden wir uns zur Beantwortung dieser Frage einem einfachen Beispiel zu, in dem die Erhebung des Bildungs- und Berufsverlaufs einer Person mit Hilfe eines Querschnitts, eines Panels und eines ereignisorientierten Designs dargestellt wird (Abbildung 2.2). Zur Charakterisierung des Karriereverlaufs werden sieben Zustände (Ausbildung, Beruf 1, Beruf 2, Beruf 3, Beruf 4, Arbeitslosigkeit und Krankheit) unterschieden, in denen sich eine Person befinden kann.

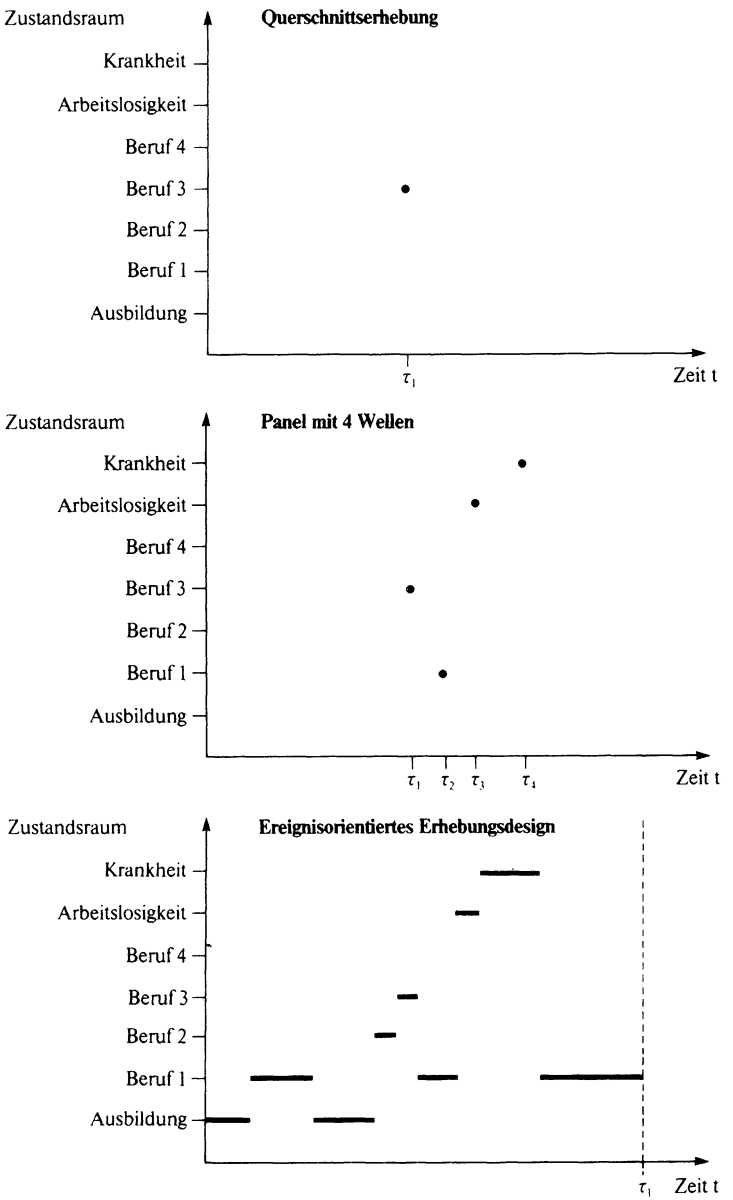
Aus Abbildung 2.2 ist zunächst zu erkennen, daß in der *Querschnittserhebung* der Bildungs- und Berufsverlauf der Person nur durch einen einzigen Punkt, den Zustand zum Zeitpunkt der Stichprobenziehung, repräsentiert wird. Etwas mehr Informationen liefert dagegen das vierwellige *Panel*, in dem die Zustände der Person schon zu vier verschiedenen Zeitpunkten beobachtet werden können. Allerdings ist unklar, wie sich der Prozeß zwischen diesen vier Wellen des Panels entwickelt hat. Erst ein *ereignisorientiertes Erhebungsdesign*, bei dem die Zustandsänderungen und ihre genauen Zeitpunkte (z.B. wie in der Lebensverlaufsstudie retrospektiv) erfragt worden sind, erlaubt es, den Bildungs- und Berufsverlauf in seinen einzelnen Phasen und für jeden beliebigen Zeitpunkt detailliert zu rekonstruieren.

An diesem Beispiel wird folgendes deutlich:

— Der sinnvolle Einsatz von Querschnittsanalysen impliziert in der Regel eine *Gleichgewichtsannahme*. Das heißt, die zu einem bestimmten Zeitpunkt sich ergebende Verteilung ist nur dann aussagekräftig, wenn der zugrundeliegende Prozeß in der Zeit einigermaßen stabil bleibt. Bei größeren Schwankungen und Wandlungsprozessen ist die Momentaufnahme eines Querschnitts nicht angemessen, weil die Analyseergebnisse dann davon abhängen, wann die Erhebung durchgeführt wurde. Panel- und ereignisorientierte Daten tragen dagegen dem *Wandel und der Dynamik* explizit Rechnung. Jede an wirtschafts- und gesellschaftspolitischen Problemen orientierte Untersuchung müßte deswegen auf Informationen zurückgreifen, die auf Längsschnittsdaten für die einzelnen Beobachtungseinheiten beruhen.

— Aber auch wenn in der empirischen Realität beträchtliche Stabilität vorherrschen sollte, besitzen Panel- und ereignisorientierte Daten im Vergleich zu Querschnitten den Vorteil eines *höheren Informationsgehalts*. So können Querschnittsdaten zunächst einmal als Spezialfall von Panel- und ereignisorientierten Daten angesehen werden, weil sich aus diesen Querschnitte ohne weiteres rekonstruieren lassen. Im empirischen Anwendungsfall kann überdies eigentlich nur die Erhebung von Panel- oder ereignisorientierten Daten Aufschluß darüber geben, ob über die Zeit tatsächlich Stabilität vorliegt. Schließlich dürften die im Vergleich zu Querschnitten bei Panel- oder ereignisorientierten Daten vorliegenden Informationen über die Vorgeschichte dazu

Abbildung 2.2: Erfassung des Bildungs- und Berufsverlaufs einer Person mit Hilfe einer Querschnittserhebung, eines Panels und eines ereignisorientierten Erhebungsdesigns



beitragen, die Erklärungs- und Prognosekraft statistischer Modelle zu verbessern.

— Bleibt im Panel der Verlauf zwischen den einzelnen Erhebungszeitpunkten offen, so ermöglicht das ereignisorientierte Erhebungsdesign dagegen die *Rekonstruktion des kontinuierlichen Prozesses*. Zwar ist auch das Panel zur Erfassung des zeitlichen Verlaufs geeignet, wenn die Zustandsänderungen zu fest vorgegebenen Zeitpunkten stattfinden, die mit den Erhebungsintervallen übereinstimmen (z. B. die monatliche Erfassung des Monatseinkommens), oder wenn eine kontinuierliche Variable (z. B. das Körpergewicht eines Menschen) sinnvoll nur auf der Basis zeitdiskreter Erhebungen gemessen werden kann; aber alle anderen Veränderungen nicht-metrischer Variablen, die zu beliebigen Zeitpunkten eintreten können, erfordern zur vollständigen Rekonstruktion eine genaue Registrierung von Art und Zeitpunkt der Zustandsänderungen. Das ereignisorientierte Erhebungsdesign erweist sich damit in vielen konkreten Anwendungsgebieten als eine notwendige Voraussetzung, um Wandlungsprozesse adäquat abbilden zu können.

— Denkt man schließlich an die dynamische Analyse *komplexer Kopplungs- und Rückkopplungsprozesse* im wirtschafts- und sozialwissenschaftlichen Bereich, dann scheint die kontinuierliche Erhebung nicht-metrischer Variablen mit Hilfe ereignisorientierter Designs die einzige adäquate Möglichkeit zu sein, empirische Wandlungs- und Veränderungstendenzen zu erfassen. Dies gilt insbesondere dann, wenn die Ereignisse dieser parallelen Prozesse nicht nur zu beliebigen Zeitpunkten eintreten, sondern darüber hinaus auch zeitverzögert aufeinander einwirken.

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

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

Teile der Lebensverläufe der zwischen 1929–31, 1939–41 und 1949–51 Geborenen nur noch retrospektiv zugänglich. Im allgemeinen werden solche Daten aber mit dem Einwand unzureichender Zuverlässigkeit konfrontiert; insbesondere dann, wenn die zu erinnernden Ereignisse weit in der Vergangenheit zurückliegen. Die retrospektive Erhebung von Ereignisdaten erfordert deswegen ein vergleichsweise sehr hohes Maß an Sorgfalt und Kontrolle, wie es in der Regel nur durch aufwendige Datenrecherchen und zeitraubende Dateneditionen zu erreichen ist. Werden die Daten darüber hinaus nur ein einziges Mal retrospektiv erfragt, so ist die Gefahr groß, daß die Datenbasis relativ schnell veraltet.

Deswegen werden beispielsweise beim Sozio-ökonomischen Panel (Hanefeld 1984) die Vorteile des *traditionellen Panels mit der retrospektiven Erhebung von Ereignisdaten verbunden.* Mit jeder neuen Panel-Welle stehen dann nicht nur jeweils aktuelle Informationen bereit, sondern durch die retrospektiven Fragen werden auch die wichtigsten Veränderungen und ihre genauen Zeitpunkte zwischen den Wellen erfaßt (zum Vergleich von Panel- und Retrospektivstudien vgl. auch Featherman 1979-80).

Welches der beschriebenen Verfahren zur Erhebung von Ereignisdaten auch immer herangezogen wird, es handelt sich stets um außerordentlich *aufwendige und kostenintensive Prozeduren.* Doch dürfte dies nicht der einzige Grund dafür sein, daß ereignisorientierte Erhebungsdesigns bisher nur im geringen Umfang in den Wirtschafts- und Sozialwissenschaften eingesetzt worden sind. Ein weiterer Grund ist sicherlich auch, daß viele Wirtschafts- und Sozialwissenschaftler *zuwenig mit den Methoden der dynamischen Analyse vertraut* sind. Die Datenstruktur wird meist als zu komplex betrachtet, die zur Ereignisanalyse erforderlichen Wahrscheinlichkeitsmodelle sind zuwenig bekannt und die bei unvollständigen Stichproben notwendigen Schätzer und Statistikprogramme sind zuwenig verbreitet. Indessen besteht inzwischen eine starke, meist inhaltlich motivierte Nachfrage nach dynamischen Analysen von Prozessen und Verläufen im Bereich der Wirtschafts- und Sozialwissenschaften. Sie dürfte zunehmend dazu führen, daß ereignisorientierte Datenstrukturen auch dort bereitgestellt und adäquat analysiert werden.

Kapitel 3:
Statistische Theorie der Ereignisanalyse

Dieses Kapitel hat die statistischen Grundlagen der Ereignisanalyse zum Gegenstand. Nach der Einordnung der Ereignisanalyse in den Rahmen der stochastischen Prozesse in Abschnitt 3.1 werden in Abschnitt 3.2 die statistischen Grundkonzepte der Ereignisanalyse wie Hazardrate, Survivorfunktion, kumulative Hazardrate und einige wichtige Verteilungsklassen für die Episodendauer (Verweildauer, Lebenszeit) ausführlich behandelt. Die Begriffe werden für den Ein-Episoden-Fall definiert, viele der statistischen Konzepte lassen sich aber auch auf komplexere Situationen wie mehrere aufeinanderfolgende Episoden oder mehrere Endzustände (competing risks) übertragen. Den Schluß dieses Abschnitts bildet die Darstellung nichtparametrischer Schätzverfahren wie der Sterbetafel-Methode und der Kaplan-Meier-Schätzung der Survivorfunktion sowie von Tests zum Vergleich von Survivorfunktionen.
Von zentraler Bedeutung für dieses Kapitel ist Abschnitt 3.3, in dem die Einbeziehung von Kovariablen in Regressionsansätzen behandelt wird. Es werden parametrische Modelle, wie zum Beispiel das Exponential-, das Weibull-, das Gompertz- oder das log-logistische Regressionsmodell, und das semiparametrische Cox-Modell ausführlich dargestellt. Alle weiteren Abschnitte dieses Kapitels setzen die Einbeziehung von Kovariablen oder prognostischen Faktoren voraus.
In den Abschnitten 3.4 und 3.5 werden allgemeine Mehr-Zustands- und Mehr-Episoden-Regressionsmodelle erörtert. Es wird eine allgemeine Theorie zur Darstellung der Modelle entwickelt, die als zentrales Konzept die episoden- und zustandsspezifische Hazardrate enthält.
Abschnitt 3.6 hat die Maximum-Likelihood-Schätzung der unbekannten Modellparameter zum Gegenstand. Nach einer kurzen Einführung in das allgemeine Prinzip der Maximum-Likelihood-Schätzung wird auf die Zensierungsproblematik bei Ereignisdaten eingegangen, die in die Schätzprozedur einzubeziehen ist. Anschließend wird für die in den vorangegangenen Abschnitten vorgestellten Modelle die Durchführung der Maximum-Likelihood-Schätzung beschrieben, wobei das Cox-Modell eine Sonderstellung einnimmt.
In Abschnitt 3.7 werden Verfahren zur Konstruktion von Hypothesentests und zur Modellwahl behandelt, während in Abschnitt 3.8 die Möglichkeit der Einbe-

ziehung zeitabhängiger und stochastischer Kovariablen dargestellt wird. In Abschnitt 3.9 werden schließlich einige Methoden zur Berücksichtigung unbeobachteter Populationsheterogenität mit Hilfe einer individuenspezifischen „Fehlervariablen" sowie dabei auftretende Probleme und einige Lösungsmöglichkeiten beschrieben.
Der letzte Abschnitt dieses Kapitels enthält eine kurze Darstellung von Regressionsansätzen für diskrete Hazardraten bei „gruppierten" Verweildauern.

3.1 Ereignisanalyse als spezieller stochastischer Prozeß

Das statistische Grundmodell der Ereignisanalyse untersucht die Länge der Zeitintervalle zwischen aufeinanderfolgenden *Zustandswechseln beziehungsweise Ereignissen*. Für jede Untersuchungseinheit sind dabei die Zeitpunkte der Zustandswechsel beziehungsweise des Eintreffens bestimmter Ereignisse und ihre Abfolge gegeben. Ist die Länge der Zeitintervalle beziehungsweise die Dauer der Episoden exakt angebbar, handelt es sich um *stochastische Prozesse mit stetiger Zeit*. Die Zeit ist eine stetige Variable, weil Ereignisse beziehungsweise Zustandswechsel zu jedem beliebigen Zeitpunkt eintreten können. Die *Zustandsvariable besitzt* hingegen nur *endlich viele Ausprägungen*.
Im statistischen Modell werden die Zeitpunkte, zu denen Zustandswechsel beziehungsweise Übergänge auftreten, repräsentiert durch eine Folge von nichtnegativen Zufallsvariablen $0 = T_0 \leq T_1 \leq T_2 \leq \ldots$ und die Zustandsvariable wird festgelegt durch $\{Y_k : k = 0, 1, 2, \ldots\}$, eine Folge von Zufallsvariablen mit endlichem Zustandsraum. Der zum stochastischen Prozeß $(Y, T) = \{(Y_k, T_k) : k = 1, 2, \ldots\}$ gehörende Prozeß

$$Z = \{Z(t) : t \geq 0\}$$

mit $$Z(t) = Y_k \text{ für } T_{k-1} \leq t < T_k,\ k = 1, 2, \ldots$$

ist ein stochastischer Prozeß mit endlichem Zustandsraum und stetiger Zeit.
Obwohl unter theoretischen Gesichtspunkten der Zustandsraum auch abzählbar unendlich viele Zustände enthalten kann, werden bei praktischen Anwendungen in der Regel nur wenige Zustände zu berücksichtigen sein. Beispielsweise kann bei einer Untersuchung der Arbeitslosigkeit eine Einteilung des Zustandsraums in „erwerbstätig" – „nicht erwerbstätig" – „arbeitslos" sinnvoll sein (vgl. Abbildung 3.1). Gelegentlich handelt es sich bei den Zeitverläufen um Zeitspannen bis zum erneuten Auftreten ein und desselben Ereignisses. Beispiele hierfür sind die Zeitspannen zwischen aufeinanderfolgenden Geburten bei demographischen Studien oder die störungsfreien Laufzeiten eines technischen Geräts bis zum Auftreten eines bestimmten Defekts. In einem solchen Fall ist der Prozeß Y_k *degeneriert*, das heißt, der Zustandsraum enthält nur ein Element, und die Akzente liegen auf der Analyse der Zeiten T_k, $k = 1, 2, \ldots$ bis zur Wiederkehr des in Frage stehenden Ereignisses. In jedem Fall sollte die Festle-

gung des Zustandsraums durch die inhaltliche Problemstellung bestimmt und das Ergebnis theoretischer Vorüberlegungen sein, denn durch die Wahl des Zustandsraums werden sowohl für das statistische Modell als auch für die Interpretation der Ergebnisse wesentliche Vorentscheidungen getroffen.
Abgesehen von dem eben beschriebenen Spezialfall korrespondiert der Terminus „Ereignis" stets mit einem Wechsel in Z(t), also mit dem Übergang von einem bestimmten Zustand in einen anderen Zustand.
Der Terminus *„Episode"* kennzeichnet die Zeitdauer zwischen aufeinanderfolgenden Ereignissen. Von besonderem Interesse sind die *Zeitdauern*

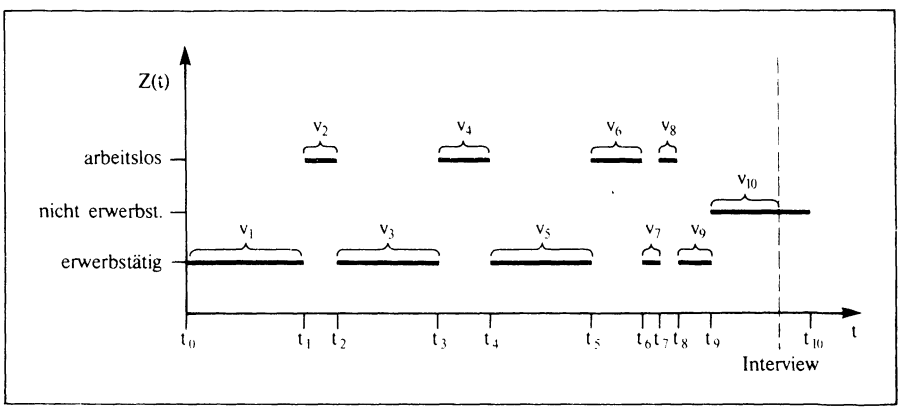
$$V_k = T_k - T_{k-1} \quad , k = 1, 2, \ldots$$

die *Verweildauern* oder *Wartezeiten* genannt werden.

Beispiel:

In einer Untersuchung der Arbeitslosigkeitsdauer in mehreren aufeinanderfolgenden Episoden werden die Zustände „erwerbstätig", „nicht erwerbstätig" und „arbeitslos" unterschieden. Der zeitliche Verlauf des Prozesses für eine Person ist in Abbildung 3.1 dargestellt. Die horizontale Achse in Abbildung 3.1 repräsentiert die Zeit und auf der vertikalen Achse sind drei Zustände aufgetragen, in denen sich die Person zum Zeitpunkt t befinden kann.

Abbildung 3.1: Zeitlicher Verlauf der Erwerbstätigkeit



Ein besonders wichtiger Spezialfall ergibt sich, wenn Prozesse mit nur *einer* Episode, *einem* Anfangszustand und *einem* (absorbierenden) Endzustand analysiert werden. Methoden zur Behandlung dieses Spezialfalles wurden vor allem in der „Survival-Analyse" bei der Untersuchung von Lebens- beziehungsweise Überlebenszeiten entwickelt. Existieren mehrere (absorbierende) Zielzustände, handelt es sich um *Mehr-Zustands-Modelle*, die in der Biostatistik meist als Competing-Risks-Modelle bezeichnet werden. *Mehr-Episoden-Modelle* liegen

vor, wenn im Laufe der Zeit mehrfache Übergänge möglich sind oder wenn ein bestimmtes Ereignis wiederholt auftreten kann.
Da das Ende des gesamten Beobachtungszeitraums in der Regel vorgegeben ist (z.B. durch den Zeitpunkt der Stichprobenziehung bei einer retrospektiven Erhebung von Ereignisdaten), ist die letzte Episode eines Individuums beziehungsweise Objekts unter Umständen nicht abgeschlossen. Man spricht in einem solchen Fall von *rechtszensierten Daten*. Beispielsweise treten die Untersuchungsobjekte zu bestimmten Zeitpunkten in die Untersuchung ein, etwa am Tag ihrer ersten Arbeitslosmeldung, und danach wird ihre Ereignisgeschichte über einen Zeitraum hinweg bis zu einem Stichtag verfolgt. Gemessen werden die Zeitdauern der aufeinanderfolgenden Episoden, in denen die Personen arbeitslos sind. In einem solchen Fall ist es möglich, daß die letzte Episode der Arbeitslosigkeit am Stichtag noch andauert. Für die Beschreibung verschiedener Zensierungsmechanismen vergleiche man zum Beispiel Kalbfleisch/Prentice (1980, Kap. 5), oder Lawless (1982). Die statistische Modellierung der wichtigsten Zensierungsmechanismen wird in Abschnitt 3.6.2 behandelt.
Daneben existiert für viele Problemstellungen auch die Möglichkeit der *Zensierung von links*, das heißt, die Zeitdauer beziehungsweise die Zeitspanne, die ein Individuum beziehungsweise Objekt bereits im Zustand y_0 verbracht hat, ist unbekannt. Dieser Fall ist wesentlich schwieriger zu behandeln als Zensierung von rechts, da es im allgemeinen nicht möglich ist, die Auswirkungen der nicht bekannten Vorgeschichte auf zukünftige Ereignisse einzuschätzen. Im folgenden setzen wir stets voraus, daß entweder der Startzeitpunkt und der Startzustand fest vorgegeben sind (ohne Beschränkung der Allgemeinheit dann $t_0 = 0$) oder daß die Vorgeschichte des Prozesses vor dem Beobachtungszeitraum den weiteren Verlauf des Prozesses nicht beeinflußt. Da in der Lebensverlaufsstudie von den Kohorten 1929–31, 1939–41 und 1949–51 jeweils der gesamte Lebensverlauf erhoben worden ist, tritt das Problem linkszensierter Daten bei den Anwendungsbeispielen in den Kapiteln 4 bis 6 nicht auf.
In vielen Fällen werden neben den Verweildauern bei jeder Untersuchungseinheit eine Reihe weiterer Kovariablen oder prognostischer Faktoren erhoben, die einzeln und in Kombination die Verweildauer beeinflussen können. Ein wichtiges Ziel der statistischen Analyse von Zeitverläufen ist deshalb die quantitative Ermittlung des Einflusses dieser exogenen oder endogenen Variablen mit Hilfe geeigneter Regressionsmodelle. Die Kovariablen können dabei quantitativ oder qualitativ sein. Kategoriale Kovariablen sind in den Regressionsansätzen durch geeignete Dummy-Variablen zu kodieren. Man vergleiche dazu die Ausführungen in Abschnitt 3.3.1. Einige der erhobenen Kovariablen können auch zeitabhängig und stochastisch sein.
Im einfachsten Fall ist eine zeitabhängige Kovariable eine fest vorgegebene Funktion der Zeit, etwa das Alter. Einige Kovariablen können aber auch stochastische Prozesse sein, die parallel zu dem in Frage stehenden Prozeß verlaufen und zusammen mit ihm beobachtet werden. Dabei kann es sich um einen externen Prozeß handeln, dessen Pfad nicht von der Verweildauer des

Individuums beeinflußt wird, es kann aber auch ein Prozeß sein, der von der Verweildauer des Individuums im untersuchten Zustand abhängig ist.
Befindet sich ein Individuum gerade in der k-ten Episode, so wird die Information über die Vorgeschichte des Prozesses in H_{k-1} zusammengefaßt. H_{k-1} umfaßt die *Vorgeschichte des Prozeßverlaufs* bis zum Zeitpunkt t_{k-1}, also

$$H_{k-1} = \{t_0, y_0, t_1, y_1, \{x_1(u) : u < t_1\}, \ldots, t_{k-1}, y_{k-1}, \{x_{k-1}(u) : t_{k-2} \leq u < t_{k-1}\}\}.$$

Dabei wurde der Index für das Individuum der Einfachheit halber weggelassen. Die Kovariablen werden in den nächsten Abschnitten als zeitunabhängig angenommen. In Abschnitt 3.8 werden dann die Probleme, die bei der Einbeziehung zeitabhängiger Kovariablen auftreten können, ausführlich erörtert.
Insgesamt erfordert die vollständige *Ereignisgeschichte* eines Individuums i im Beobachtungszeitraum die Angabe folgender Daten:

1. y_{i0}	Startzustand
2. n_i	Anzahl der Episoden im Beobachtungszeitraum
3. $t_{i1} \leq t_{i2} \leq \ldots \leq t_{in_i}$	Zeitpunkte, zu denen Zustandswechsel stattfinden beziehungsweise bestimmte Ereignisse eintreten
4. $y_{i1}, y_{i2}, \ldots, y_{in_i}$	Zustände, die zu den obigen Zeitpunkten angenommen werden
5. δ_i	Indikator, ob die n_i-te Episode zensiert ist oder nicht
6. $x_{i1}, x_{i2}, \ldots, x_{in_i}$	Vektoren von Kovariablen, die zu Beginn jeder Episode gemessen werden. Sie werden zunächst als zeitunabhängig angenommen.

3.2 Statistische Grundkonzepte (Ein-Episoden-Fall)

Der einfachste Fall der Ereignisanalyse liegt dann vor, wenn lediglich die Zeitdauer vom Eintritt in einen Anfangszustand bis zum Erreichen eines bestimmten Endzustandes gemessen wird. Anwendungen findet man vor allem bei der Untersuchung von Lebens- beziehungsweise Überlebenszeiten in medizinischen Studien, aber auch zum Beispiel bei der Analyse der Lebensdauer politischer oder gesellschaftlicher Organisationen. Viele der für den Ein-Episoden-Fall entwickelten statistischen Konzepte können auf komplexere Situationen, wie mehrere aufeinanderfolgende Episoden oder mehrere Endzustände (competing risks), übertragen werden.
Die Dauer der Episode wird im statistischen Modell repräsentiert durch eine nicht-negative Zufallsvariable T. Kann der Endzeitpunkt exakt angegeben werden, ist T eine *stetige Zufallsvariable*. Lassen sich lediglich Zeitintervalle angeben, in denen der Endzustand erreicht werden kann, ist T eine *diskrete Zufalls-*

variable. T = t bedeutet dann, daß im t-ten Zeitintervall ein Übergang stattgefunden hat. In den Abschnitten 3.2 bis 3.9 wird T als stetig vorausgesetzt. Diskrete Modelle werden in Abschnitt 3.10 kurz erörtert.

3.2.1 Dichte- und Verteilungsfunktion, Survivorfunktion, Hazardrate

Im folgenden werden wichtige statistische Kenngrößen der Ereignisanalyse im Ein-Episoden-Fall mit einem Anfangs- und einem Endzustand eingeführt. Dabei wird zunächst von einer homogenen Population ausgegangen, das heißt, interindividuelle Heterogenität in bezug auf verschiedene Merkmale bleibt unberücksichtigt. Die Einbeziehung von Kovariablen und prognostischen Faktoren wird dann in Abschnitt 3.3 ausführlich behandelt.
Die *Dichte-* und die *Verteilungsfunktion* der Episodendauer T ($T \geq 0$) seien mit f(t) beziehungsweise mit F(t) bezeichnet. Dabei gilt wie üblich der Zusammenhang

$$F(t) = P(T \leq t) = \int_0^t f(u)\, du, \tag{3.2.1}$$

und an allen Stellen, an denen F(t) differenzierbar ist, gilt

$$f(t) = F'(t). \tag{3.2.2}$$

Die *Survivorfunktion*

$$S(t) = P(T \geq t) \tag{3.2.3}$$

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

$$S(t) = 1 - F(t). \tag{3.2.4}$$

Die Survivorfunktion ist in Abhängigkeit von der Zeit monoton fallend (vgl. Abbildung 3.2).

Die *Hazardrate* ist

$$\lambda(t) = \lim_{\substack{\Delta t \to 0 \\ \Delta t > 0}} \frac{1}{\Delta t} P(t \leq T < t + \Delta t | T \geq t). \tag{3.2.5}$$

Die *Hazardrate* kann aufgefaßt werden als der Grenzwert der bedingten Wahrscheinlichkeit, daß die Episode im Intervall $[t, t + \Delta t)$ zu Ende geht unter der Voraussetzung, daß die Episode bis zum Beginn dieses Intervalls andauert. Andere Bezeichnungen für die Hazardrate sind *Intensitäts-* oder *Risikofunktion*, *Übergangsrate* oder *Mortalitätsrate*.

Abbildung 3.2: Typischer Verlauf einer Survivorfunktion

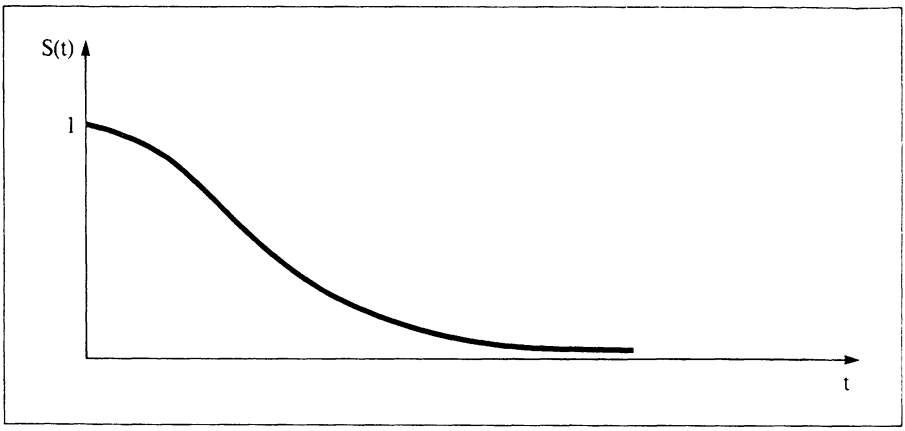
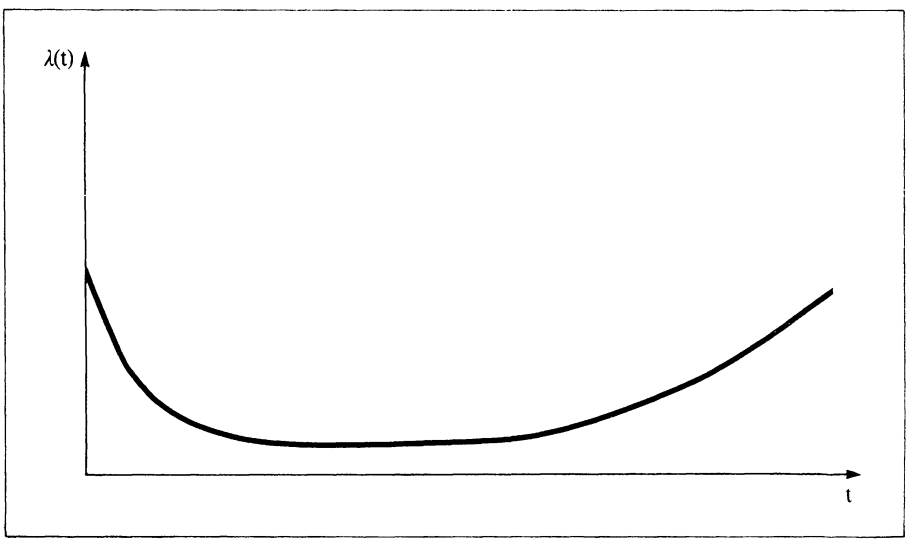


Abbildung 3.3: Hazardrate mit „badewannenförmigem" Verlauf



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

$$\Lambda(t) = \int_0^t \lambda(u)\, du \tag{3.2.6}$$

wird als *kumulative Hazardrate* bezeichnet.

Die Hazardrate stellt ein zentrales Konzept bei der Analyse von Verlaufsdaten dar. „Überlebt" ein Individuum den Zeitpunkt t, so informiert die Hazardrate über „den weiteren Verlauf". Häufig besitzt man bei praktischen Anwendungen zumindest qualitative Vorinformationen über die Hazardrate. Dies soll an dem Beispiel des Sterberisikos einer Population verdeutlicht werden. Die Hazardrate hat hier typischerweise einen „badewannenförmigen" Verlauf (vgl. Abbildung 3.3).
Zu Beginn des Prozesses ist das Sterberisiko wegen der Kindersterblichkeit relativ hoch, es fällt dann und bleibt über einen bestimmten Zeitraum konstant auf niedrigem Niveau, bis es mit zunehmendem Alter wieder anwächst.
Ähnlich verhält sich die Hazardrate bei vielen technischen Geräten. Aufgrund von „Kinderkrankheiten" und „Defekten beim ersten Einschalten" ist das Ausfallrisiko zunächst relativ hoch, fällt dann ab und wächst wieder, wenn Alterungsprozesse und Materialermüdungserscheinungen auftreten.
Daneben sind natürlich auch andere Formen der Hazardrate denkbar, zum Beispiel ständig zunehmende oder abnehmende Hazardraten.
Aus Definition (3.2.5) folgt unmittelbar die Beziehung zwischen Hazardrate und Survivorfunktion

$$\lambda(t) = \frac{f(t)}{S(t)}, \tag{3.2.7}$$

und da T als stetig vorausgesetzt wurde, gilt auch

$$\lambda(t) = \frac{f(t)}{1 - F(t)}. \tag{3.2.8}$$

Umgekehrt ergibt sich die Survivorfunktion in Abhängigkeit von der Hazardrate, wenn man $\lambda(t)$ integriert und die Beziehungen (3.2.7) und (3.2.8) verwendet.

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

$$= -\ln(1 - F(t)) = -\ln S(t). \tag{3.2.9}$$

Dies führt zur wichtigen Beziehung

$$S(t) = \exp\left(-\int_0^t \lambda(u)\,du\right). \tag{3.2.10}$$

Die Dichtefunktion f(t) ergibt sich aus (3.2.7) und (3.2.10) in Abhängigkeit von der Hazardrate

$$f(t) = \lambda(t) \cdot S(t) = \lambda(t) \cdot \exp\left(-\int_0^t \lambda(u)\,du\right). \tag{3.2.11}$$

Aus den Beziehungen (3.2.1) bis (3.2.11) wird ersichtlich, daß jede der drei Größen f(t), S(t) und $\lambda(t)$ zur Beschreibung der Dauer der Episode herangezogen werden kann. Ist eine der Größen festgelegt, so sind die beiden anderen

eindeutig daraus ableitbar. Kennt man insbesondere die Hazardrate, so ist dadurch der Prozeßverlauf vollständig beschrieben.
Gelegentlich ist noch ein weiterer Zusammenhang nützlich. Aus (3.2.3) und (3.2.11) ergibt sich für die Survivorfunktion

$$S(t) = \int_t^\infty f(u)\,du = \int_t^\infty \lambda(u)\,S(u)\,du. \tag{3.2.12}$$

Während in (3.2.10) die Vergangenheit des Prozesses (bis zum Zeitpunkt t) eine Rolle spielt, geht in (3.2.12) auch die „Zukunft" ein.
Schließlich erhält man für die (bedingte) Wahrscheinlichkeit, daß im Intervall $[t_1, t_2]$, $t_1 < t_2$ ein Ereignis eintritt unter der Voraussetzung, daß der Zeitpunkt t_1 erreicht wurde

$$P(t_1 \le T \le t_2 \,|\, T \ge t_1) = \frac{S(t_1) - S(t_2)}{S(t_1)}, \tag{3.2.13}$$

und für eine Menge von aufeinanderfolgenden Zeitpunkten $t_0 = 0 < t_1 < t_2 < \ldots < t_k$ resultiert

$$\begin{aligned} S(t_k) &= \exp\left(-\int_0^{t_k} \lambda(u)\,du\right) = \exp\left(-\sum_{i=0}^{k-1} \int_{t_i}^{t_{i+1}} \lambda(u)\,du\right) \\ &= \prod_{i=0}^{k-1} \exp\left(-\int_{t_i}^{t_{i+1}} \lambda(u)\,du\right) = \prod_{i=0}^{k-1} \frac{S(t_{i+1})}{S(t_i)}. \end{aligned} \tag{3.2.14}$$

Dies ergibt schließlich

$$S(t_k) = \prod_{i=0}^{k-1} P(T \ge t_{i+1} \,|\, T \ge t_i). \tag{3.2.15}$$

Beziehung (3.2.15) ist insbesondere für die Konstruktion von Sterbetafeln von Bedeutung, man vergleiche Abschnitt 3.2.3.

3.2.2 Spezielle Wahrscheinlichkeitsverteilungen für die Dauer der Episode

Exponentialverteilung – zeitunabhängige Hazard- beziehungsweise Übergangsrate

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

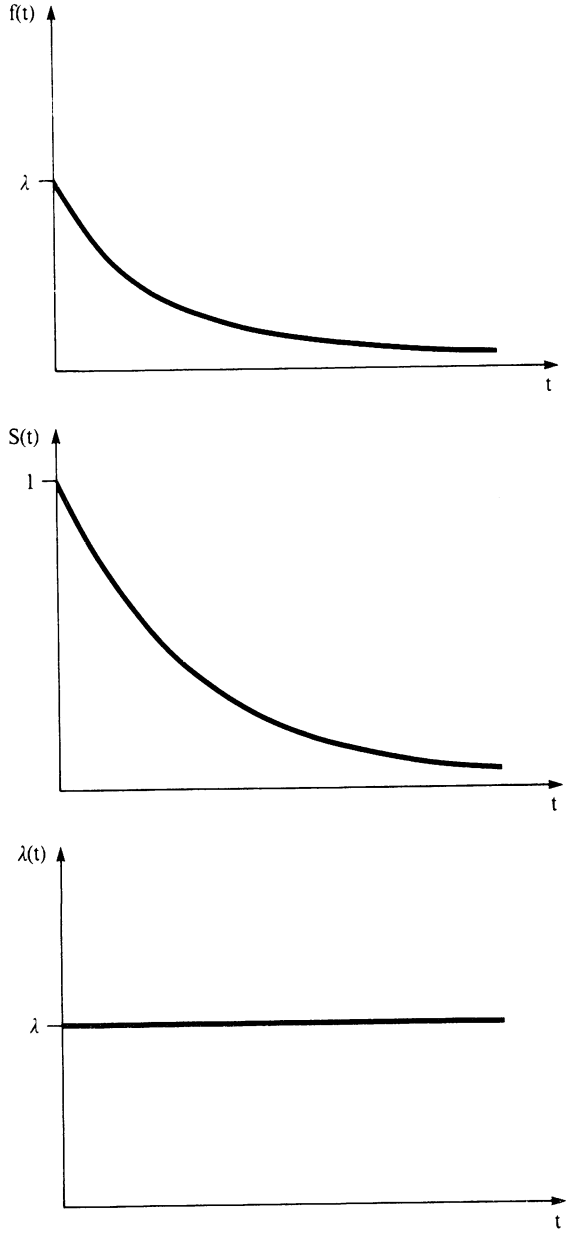
$$\lambda(t) = \lambda, \qquad t \ge 0, \quad \lambda > 0. \tag{3.2.16}$$

Für Dichte- und Survivorfunktion folgen (vgl. Abbildung 3.4).

$$S(t) = \exp(-\lambda t), \tag{3.2.17}$$

$$f(t) = \lambda \exp(-\lambda t) \tag{3.2.18}$$

Abbildung 3.4: Dichtefunktion, Survivorfunktion und Hazardrate der Exponentialverteilung



Für die „mittlere Verweildauer" erhält man

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

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

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

Weibull-Verteilung

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

$$\lambda(t) = \lambda\alpha\,(\lambda t)^{\alpha-1} \qquad (t > 0) \tag{3.2.19}$$

mit den Parametern $\lambda > 0$ und $\alpha > 0$. Für den Spezialfall $\alpha = 1$ erhält man wieder die Exponentialverteilung.
Die Survivorfunktion ist

$$S(t) = \exp(-(\lambda t)^{\alpha}) \tag{3.2.20}$$

und die Dichtefunktion von T ist

$$f(t) = \lambda\alpha\,(\lambda t)^{\alpha-1}\exp(-(\lambda t)^{\alpha}). \tag{3.2.21}$$

Die Hazard- beziehungsweise Übergangsrate der Weibull-Verteilung ist monoton steigend für $\alpha > 1$, abnehmend für $\alpha < 1$ und konstant für $\alpha = 1$. Das Weibull-Modell ist sehr flexibel und daher für eine Vielzahl von Modellen für Verweildauern und Lebenszeiten angemessen (vgl. Abbildung 3.5).
Für den Erwartungswert E(T) der Verweildauer ergibt sich

$$E(T) = \Gamma\left(\frac{1+\alpha}{\alpha}\right) / \lambda, \tag{3.2.22}$$

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

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

Extremwert-Verteilung

Eine mit der Weibull-Verteilung in enger Beziehung stehende Verteilung ist die Extremwert-Verteilung. Survivor- und Dichtefunktion sind gegeben durch

$$S(y) = \exp\left[-\exp\left(\frac{y-\mu}{\sigma}\right)\right] \qquad -\infty < y < \infty \tag{3.2.23}$$

$$f(y) = \frac{1}{\sigma}\exp\left[\frac{y-\mu}{\sigma} - \exp\left(\frac{y-\mu}{\sigma}\right)\right] \qquad -\infty < y < \infty \tag{3.2.24}$$

Abbildung 3.5: Dichtefunktion, Survivorfunktion und Hazardrate der Weibull-Verteilung (jeweils für $\alpha = 0{,}5$, $\alpha = 1$ und $\alpha = 3$)

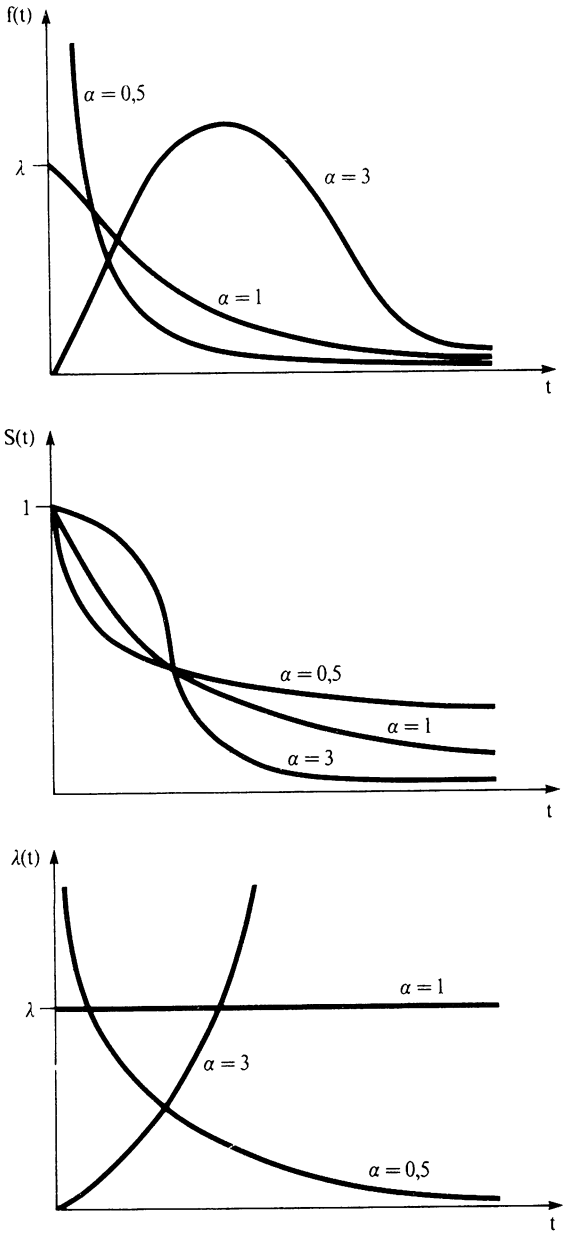
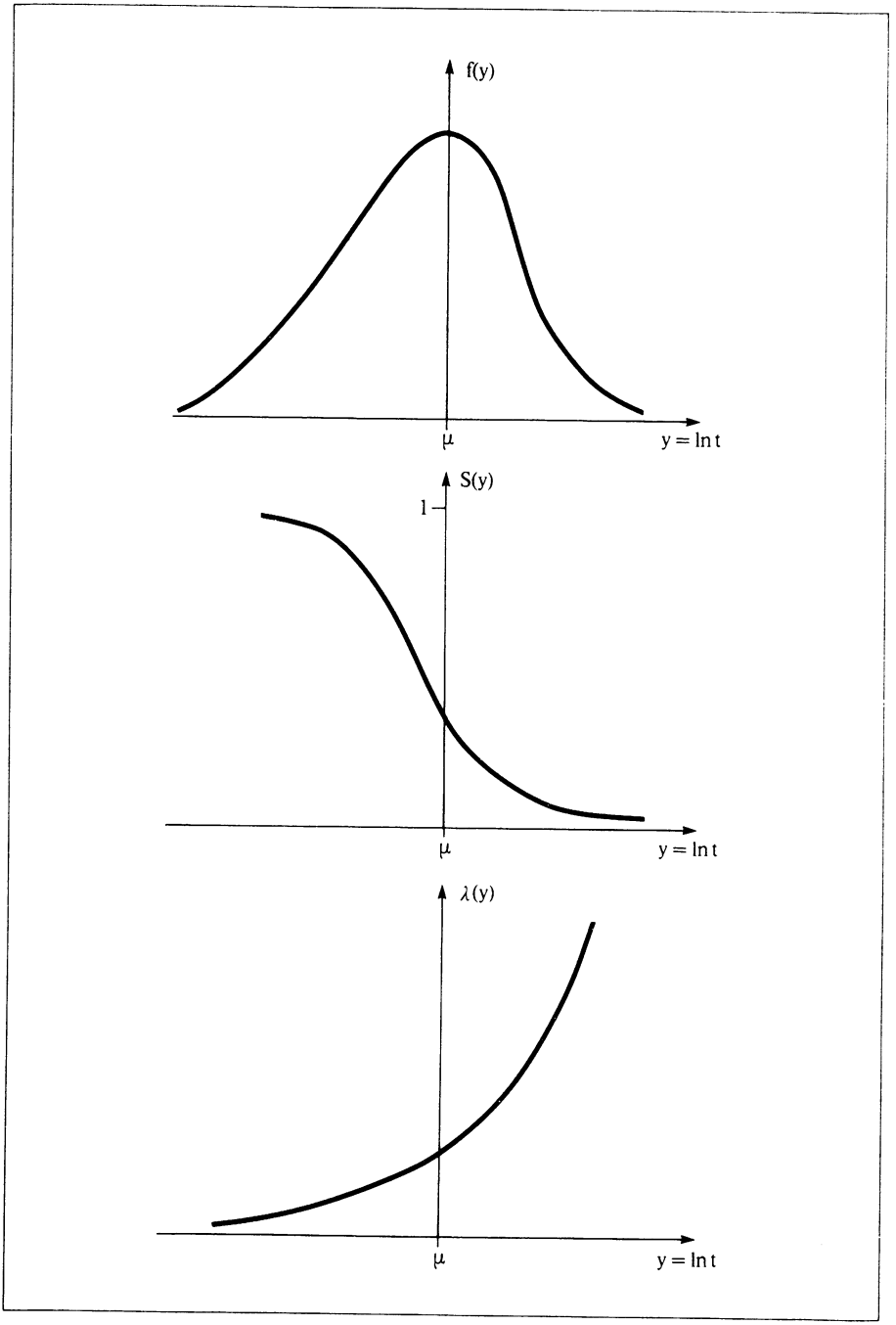


Abbildung 3.6: Dichtefunktion, Survivorfunktion und Hazardrate der Standard-Extremwert-Verteilung



und die Hazardrate ist

$$\lambda(y) = \frac{1}{\sigma} \exp\left(\frac{y-\mu}{\sigma}\right) \qquad -\infty < y < \infty. \qquad (3.2.25)$$

Mit der Weibull-Verteilung besteht folgender Zusammenhang:
Besitzt T eine Weibull-Verteilung mit der Dichtefunktion (3.2.21), so hat $Y = \ln T$ eine Extremwert-Verteilung mit $\sigma = \alpha^{-1}$ und $\mu = -\ln \lambda$.
Da insbesondere bei Lebenszeitanalysen häufig mit logarithmierten Daten gearbeitet wird, ist die Extremwert-Verteilung für die logarithmierten Werte adäquat, wenn die ursprünglichen Zeitdauern weibullverteilt sind.
Der Spezialfall $\mu = 0$ und $\sigma = 1$ wird als „Standard-Extremwert-Verteilung" bezeichnet (vgl. Abbildung 3.6).

Für Erwartungswert und Varianz erhält man

$$E(Y) = \mu + \sigma \cdot \Gamma'(1) \quad \text{und}$$
$$\text{Var}(Y) = \frac{\sigma^2 \pi^2}{6}. \qquad (3.2.26)$$

Dabei ist $\Gamma'(1)$ die erste Ableitung der Gamma-Funktion $\Gamma(n)$ an der Stelle $n = 1$.

Log-logistische Verteilung

Setzt man

$$\ln T = \mu + \sigma\, \omega$$

und nimmt für ω eine logistische Verteilung mit der Dichtefunktion

$$f(\omega) = \frac{\exp(\omega)}{[1 + \exp(\omega)]^2} \qquad (3.2.27)$$

an, so ergibt sich für T selbst die log-logistische Verteilung mit der Dichtefunktion

$$f(t) = \lambda\, \alpha\, (\lambda t)^{\alpha-1}\, [1 + (\lambda t)^{\alpha}]^{-2}, \qquad (3.2.28)$$

wobei $\lambda = e^{-\mu}$ und $\alpha = \sigma^{-1}$ ist.
Survivorfunktion und Hazardrate sind gegeben durch

$$S(t) = \frac{1}{1 + (\lambda t)^{\alpha}} \qquad (3.2.29)$$

und

$$\lambda(t) = \frac{\lambda \alpha\, (\lambda t)^{\alpha-1}}{1 + (\lambda t)^{\alpha}} \qquad (3.2.30)$$

(vgl. Abbildung 3.7).

Abbildung 3.7: Dichtefunktion, Survivorfunktion und Hazardrate der log-logistischen Verteilung

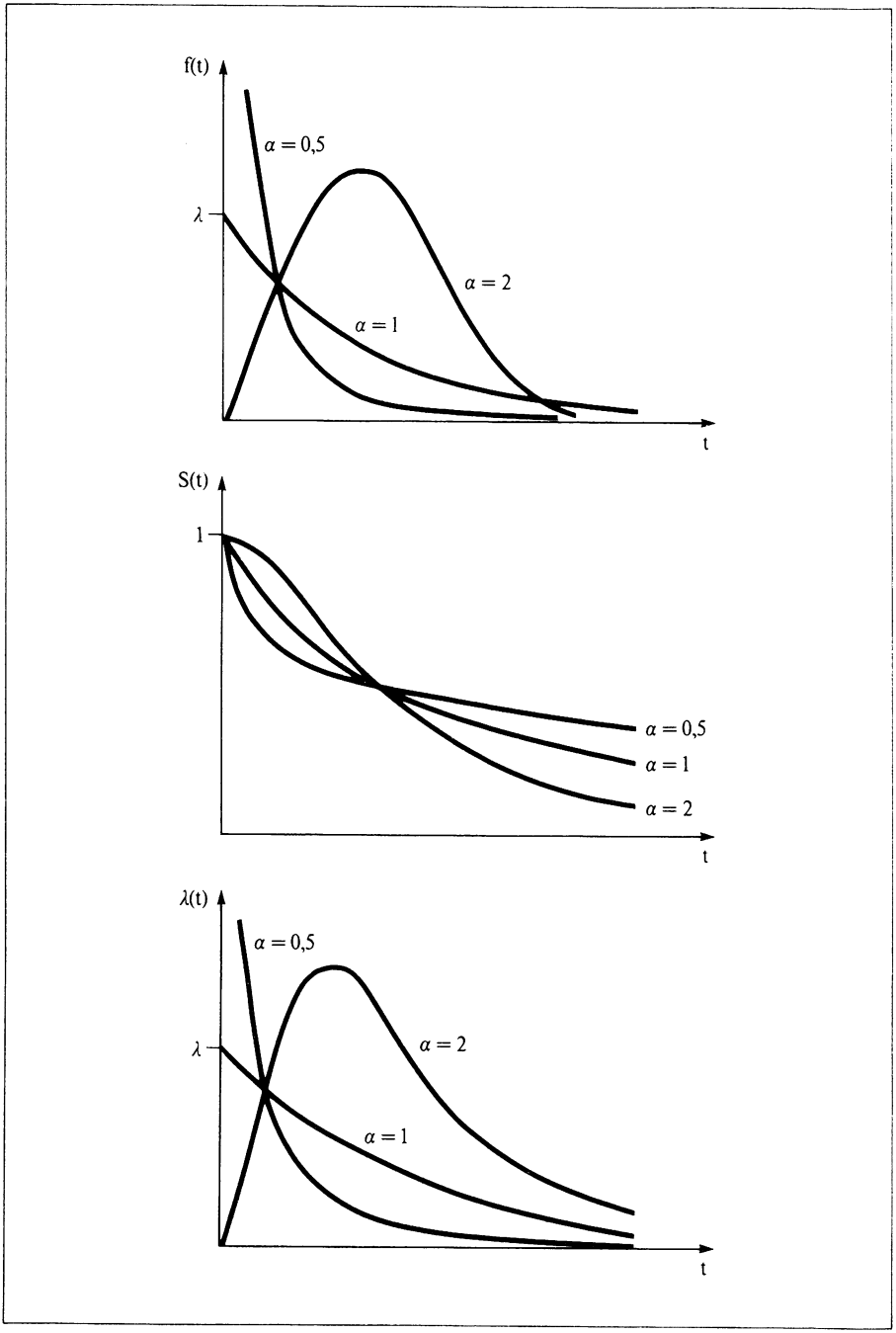
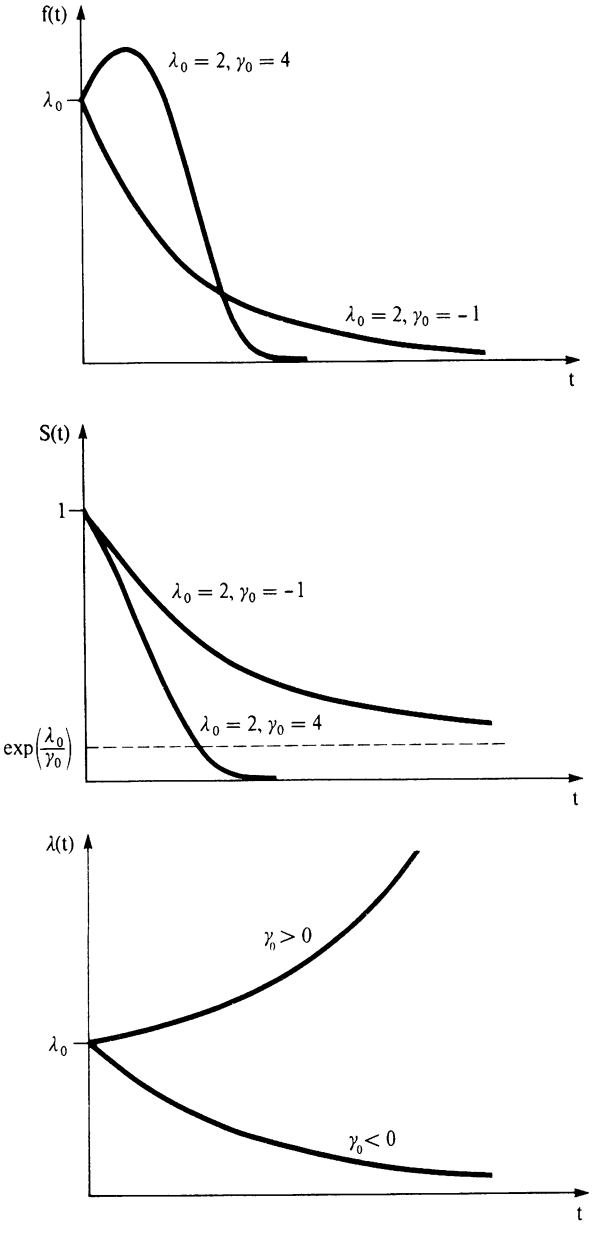


Abbildung 3.8: Dichtefunktion, Survivorfunktion und Hazardrate der Gompertz-Verteilung



Neben den hier beschriebenen Verteilungen werden bei der Analyse von Ereignisdaten gelegentlich auch die Gamma-Verteilung, die Log-Normalverteilung, die Pareto-Verteilung oder die verallgemeinerte F-Verteilung angewendet. Man vergleiche hierzu beispielsweise Kalbfleisch/Prentice (1980, Kap. 2), Miller (1981, Kap 1) oder Lawless (1982, Kap. 1.3).
Ein in der Demographie bewährtes Modell ist die Gompertz-Verteilung, die von Gompertz bereits 1825 eingeführt wurde und die im folgenden kurz beschrieben wird.

Die Gompertz-(Makeham-)Verteilung

Die Gompertz-Verteilung geht aus von der Hazardrate

$$\lambda(t) = \lambda_0 \exp(\gamma_0 t) \quad (t \geq 0) \tag{3.2.31}$$

mit den Parametern $\lambda_0 > 0$ und $-\infty < \gamma_0 < \infty$.
Die Gompertz-Verteilung fand außer in der Demographie auch im Versicherungswesen weite Verbreitung, wobei in der Regel die Substitution $\exp(\gamma_0) = c$ verwendet wird. Die Hazardrate hat dann die Form

$$\lambda(t) = \lambda_0 c^t. \tag{3.2.32}$$

Die Gompertz-Verteilung geht aus der Extremwert-Verteilung hervor, wenn diese beim Wert 0 abgeschnitten wird, so daß keine negativen Ausprägungen möglich sind. Sie ergibt sich auch als Extremwert-Verteilung bei sehr vielen voneinander unabhängig wirkenden Todesursachen, wobei die Lebenszeit durch die am ehesten eintretende Todesursache beendet wird. Addiert man zur Gompertz-Hazardrate noch eine Konstante α_0 ($\alpha_0 > 0$), die Todesfälle durch Unfälle (zusätzlich zu den Todesfällen durch natürliche Todesursachen) erfassen soll, resultiert die Gompertz-Makeham-Hazardrate

$$\lambda(t) = \alpha_0 + \lambda_0 \exp(\gamma_0 t). \tag{3.2.33}$$

Mit Hilfe von (3.2.10) und (3.2.11) erhält man für die Survivorfunktion

$$S(t) = \exp(-\alpha_0 t - \frac{\lambda_0}{\gamma_0}(\exp(\gamma_0 t) - 1))$$

und für die Dichtefunktion

$$f(t) = (\alpha_0 + \lambda_0 \exp(\gamma_0 t)) \exp(-\alpha_0 t - \frac{\lambda_0}{\gamma_0}(\exp(\gamma_0 t) - 1))$$

(vgl. Abbildung 3.8).

3.2.3 Die Sterbetafel-Methode

Die Sterbetafel-Methode ist eine der gebräuchlichsten Methoden zur Analyse von Verweildauern und Lebenszeiten. Sie ist im Prinzip nonparametrisch und

wurde vorwiegend in Demographie und Versicherung in der Form der Populationssterbetafel angewendet.
Die Zeitachse wird zerlegt in $q + 1$ Intervalle $[a_{k-1}, a_k)$, $k = 1, \ldots, q + 1$, wobei $a_0 = 0$ und $a_{q+1} = \infty$ ist. Für a_q wird in der Regel der letztmögliche Beobachtungszeitpunkt gewählt. Die Intervalle müssen nicht äquidistant sein. Die Sterbetafel-Methode ist ein Verfahren für „gruppierte" Zeiten ohne explizite Berücksichtigung von Kovariablen. Modelle mit Kovariablen für gruppierte Zeiten werden in Abschnitt 3.10 behandelt.
Die Hazardrate des k-ten Intervalls

$$\lambda_k = P(T \epsilon [a_{k-1}, a_k) \mid T \geq a_{k-1}) \tag{3.2.34}$$

ist die bedingte Wahrscheinlichkeit, daß im k-ten Zeitintervall ein Ereignis eintritt, unter der Voraussetzung, daß das Zeitintervall erreicht wurde.

Seien

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

und

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

Man erhält

$$\begin{aligned} P_k &= P(T \geq a_k \mid T \geq a_{k-1}) \cdot \ldots \cdot P(T \geq a_1 \mid T \geq a_0)\, P(T \geq a_0) \\ &= p_k \cdot \ldots \cdot p_1. \end{aligned} \tag{3.2.35}$$

Die erhobenen Daten sind:

n Gesamtanzahl der Individuen bzw. Objekte zu Beginn der Studie,
d_k Anzahl der Fälle, für die im k-ten Intervall ein Ereigniss eintritt,
w_k Anzahl der Zensierungen im k-ten Intervall, $k = 1, \ldots, q + 1$.

Für die „Risikomenge" R_k, das heißt die Anzahl derjeniger Individuen beziehungsweise Objekte, die zu Beginn des k-ten Intervalls noch kein Ereignis hatten und auch nicht zensiert sind, erhält man

$$R_1 = n \quad \text{und}$$
$$R_k = R_{k-1} - d_{k-1} - w_{k-1} \quad \text{für } k = 2, \ldots, q + 1.$$

Sind im k-ten Intervall keine Zensierungen beobachtet worden, kann die Hazardrate λ_k direkt durch die relative Häufigkeit d_k/R_k geschätzt werden. Im Falle $w_k > 0$ wird jedoch diese relative Häufigkeit die tatsächliche Hazardrate in der Regel unterschätzen. Bei der Sterbetafel-Methode wird die Risikomenge R_k korrigiert und um $w_k/2$ vermindert. Als Schätzung ergibt sich

$$\hat{\lambda}_k = \frac{d_k}{R_k - w_k/2}. \tag{3.2.36}$$

Mit $\hat{p}_k = 1 - \hat{\lambda}_k$ erhält man aus (3.2.35) Schätzungen

$$\hat{P}_k = \hat{p}_k \cdot \ldots \cdot \hat{p}_1 \tag{3.2.37}$$

für die Survivorfunktion $S(a_k)$. Daraus ergibt sich unmittelbar die geschätzte Ereignis-Wahrscheinlichkeit im k-ten Intervall

$$\hat{P}(T \in [a_{k-1}, a_k)) = \hat{P}_{k-1} - \hat{P}_k \tag{3.2.38}$$

und für die Ereignis-Wahrscheinlichkeit im k-ten Intervall bezogen auf eine Zeiteinheit erhält man die „Dichte"

$$\hat{f}_k = \frac{\hat{P}_{k-1} - \hat{P}_k}{h_k} = \frac{\hat{P}_{k-1}\,\hat{\lambda}_k}{h_k}, \tag{3.2.39}$$

wobei $h_k = a_k - a_{k-1}$ die Länge des k-ten Intervalls bezeichnet. Gewöhnlich wird für h_k ein fester Zeitabstand, zum Beispiel ein Tag, ein Monat oder ein Jahr, gewählt.
$\hat{f}_k$ aus (3.2.39) kann verwendet werden, um eine Schätzung der Hazardrate der zugrundeliegenden stetigen Verweildauer zu ermitteln. Eine „mittlere Hazardrate" in der Mitte m_k des k-ten Intervalls läßt sich schätzen durch

$$\hat{\lambda}(m_k) = \frac{\hat{f}_k}{\hat{P}(T \geq m_k)} = \frac{\hat{f}_k}{(\hat{P}_{k-1} + \hat{P}_k)/2} = \frac{2\,\hat{\lambda}_k}{h_k(\hat{P}_{k-1} + \hat{P}_k)/\hat{P}_{k-1}} = \frac{2\,\hat{\lambda}_k}{h_k\,(1 + \hat{p}_k)}. \tag{3.2.40}$$

3.2.4 Der Produkt-Limit-Schätzer (Kaplan-Meier-Schätzer) der Survivorfunktion

Eine Möglichkeit zur Schätzung der Survivorfunktion wurde bereits im letzten Abschnitt mit (3.2.37) dargestellt. Dazu ist eine Zerlegung der Zeitachse erforderlich, wobei die Wahl der Grenzen willkürlich ist. Die Vorgehensweise beim sogenannten Produkt-Limit-Schätzer, der von Kaplan/Meier (1958) abgeleitet wurde, ist ähnlich. Der Unterschied liegt darin, daß jetzt die beobachteten Ereigniszeitpunkte als Intervallgrenzen gewählt werden.
Seien $t_{(1)} < t_{(2)} < \ldots < t_{(m)}$ die geordneten Ereigniszeitpunkte ($m \leq n$), wobei zunächst angenommen wird, daß keine Ties oder Bindungen vorliegen. Dann werden die Intervalle

$$[0,\ t_{(1)}),\ [t_{(1)},\ t_{(2)}),\ \ldots,\ [t_{(m)},\ \infty)$$

gebildet. Es gilt

$$\hat{p}_k = 1 - \frac{1}{R_k},$$

wobei R_k die Risikomenge zum Zeitpunkt $t_{(k-1)}$ ist. Als Schätzung der Survivorfunktion ergibt sich

$$\hat{S}(t) = \begin{cases} 1 & \text{für } t \leq t_{(1)} \\ \prod\limits_{k|t_{(k)}<t} (1 - \frac{1}{R_k}) & \text{für } t > t_{(1)} \end{cases} \tag{3.2.41}$$

Falls Ties, das heißt mehrere Ereignisse zum gleichen Zeitpunkt auftreten, ist $1 - \frac{1}{R_k}$ in (3.2.41) durch $1 - \frac{d_k}{R_k}$ zu ersetzen, wobei d_k die Anzahl der Ereigniszeitpunkte an der Stelle $t_{(k)}$ ist.
Treten zensierte Beobachtungen zum gleichen Zeitpunkt wie Ereignisse auf, so wird die Annahme getroffen, daß die Ereigniszeitpunkte etwas vor den Zensierungszeitpunkten liegen.
Ist die letzte Beobachtung zensiert, so ist $\hat{S}(t) > 0$ für $t \to \infty$. Man wird in diesem Fall $\hat{S}(t)$ nur für die Zeitspanne bis zum größten Ereigniszeitpunkt als definiert betrachten.
Der Produkt-Limit-Schätzer kann auch als Maximum-Likelihood-Schätzer abgeleitet werden. Man vergleiche dazu Kalbfleisch/Prentice (1980, S. 10 ff.), Lawless (1982, S. 74 ff.) oder Johansen (1978). Darüber hinaus läßt sich zeigen, daß der Produkt-Limit-Schätzer aus der mit Hilfe der Sterbetafel gewonnenen Schätzung für die Survivorfunktion hervorgeht, wenn $q \to \infty$ und gleichzeitig $\max|a_k - a_{k-1}| \to 0$ gilt.
Eine Schätzung der (asymptotischen) Varianz von $\hat{S}(t)$ ist

$$\hat{Var}(\hat{S}(t)) = \hat{S}(t)^2 \sum_{t_{(k)} \leq t} \frac{d_k}{R_k (R_k - d_k)}. \tag{3.2.42}$$

Eine Schätzung für die erwartete mittlere Verweildauer beziehungsweise Lebenszeit ist

$$\hat{\mu} = \sum_{k=1}^{n} \hat{S}(t_{(k-1)}) (t_{(k)} - t_{(k-1)}), \tag{3.2.43}$$

wobei in (3.2.43) über sämtliche der Größe nach geordneten Beobachtungen (zensiert oder nicht-zensiert) zu summieren ist.
Ist die Beobachtung $t_{(n)}$ zensiert, so wird $\hat{\mu}$ im allgemeinen den wahren Erwartungswert unterschätzen. Eine Formel für die geschätzte Varianz von $\hat{\mu}$ kann ebenfalls angegeben werden (vgl. Gross/Clark 1975).
Es besteht auch die Möglichkeit, Schätzungen für beliebige Quantile der Verteilung der Verweildauer beziehungsweise Lebenszeit zu ermitteln. Auf eine ausführliche Darstellung wird hier verzichtet. Man vergleiche dazu wieder Gross/Clark (1975).

3.2.5 Vergleich von Survivorfunktionen

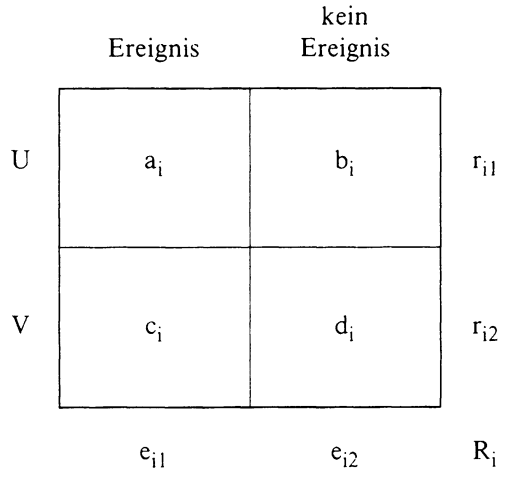
Wurde die Population in s Teilgesamtheiten gespalten, so läßt sich für jede Subpopulation eine Survivorfunktion – etwa nach der Produkt-Limit-Methode – schätzen, und von besonderem Interesse ist dann die Überprüfung, ob die Survivorfunktionen der einzelnen Gruppen übereinstimmen oder sich signifikant unterscheiden. Wegen $S(t) = 1 - F(t)$ ist dies gleichbedeutend mit der Prüfung, ob die Verteilungen der Verweildauer beziehungsweise Lebenszeiten übereinstimmen. Liegen keine Zensierungen vor, können dafür bekannte nonparametrische Verfahren im Mehr-Stichproben-Fall eingesetzt werden. Man vergleiche dazu zum Beispiel Schaich/Hamerle (1984, Kap. 5).
In der Regel werden jedoch bei der Ereignisanalyse zensierte Daten vorliegen. In diesem Fall müssen die Verfahren modifiziert werden. In der Literatur wurde bereits eine Reihe von Methoden entwickelt, so daß eine ausführliche Darstellung all dieser Verfahren den Rahmen dieser Einführung überschreiten würde. Wir werden uns deshalb darauf beschränken, die prinzipielle Vorgehensweise bei der Anwendung der Tests zu erläutern, und zwar zunächst für den einfachen Fall zweier Gruppen ($s = 2$). Wir folgen dabei der Darstellung von Tarone/Ware (1977).
Seien $u_1, \ldots, u_{n_1}$ und $v_1, \ldots, v_{n_2}$ die Meßwerte in den beiden Stichproben, $n = n_1 + n_2$, und δ_i beziehungsweise ϵ_j, $i = 1, \ldots, n_1$, $j = 1, \ldots, n_2$ seien die dazugehörigen Zensierungsindikatoren. Die Survivorfunktionen seien $S_1(t)$ und $S_2(t)$, und zu prüfen ist

$$H_0 : S_1(t) = S_2(t) \quad \text{für alle } t.$$

Die beiden Stichproben werden gepoolt und die Meßwerte der Größe nach geordnet. Seien

$$T_{(1)} \leq T_{(2)} \leq \ldots \leq T_{(m)}$$

die der Größe nach geordneten, nicht-zensierten Meßwerte der gepoolten Stichprobe ($m \leq n$). Die Risikomenge R_i ist wieder die Anzahl der Personen, die unmittelbar vor dem Zeitpunkt $T_{(i)}$ noch kein Ereignis hatten und nicht zensiert sind.
Für jeden nicht-zensierten Zeitpunkt $T_{(i)}$, $i = 1, \ldots, m$ wird eine Vierfelder-Kontingenztabelle in der folgenden Form erstellt.



a_i ist die Zahl der Personen aus der ersten Gruppe, die zum Zeitpunkt $t_{(i)}$ ein Ereignis haben. Liegen keine Ties vor, so ist a_i entweder 0 oder 1. e_{i1} ist dann stets 1. Sind Ties zum Zeitpunkt $t_{(i)}$ vorhanden, so ist e_{i1} die Anzahl der Ties und a_i und c_i geben an, wie sich die Ereignisse zum Zeitpunkt $t_{(i)}$ auf die beiden Gruppen verteilen. r_{i1} ist die Risikomenge bezogen auf die Mitglieder der ersten Gruppe, r_{i2} besitzt eine analoge Interpretation für die zweite Gruppe. Wir wollen die Vorgehensweise an einem kleinen Beispiel verdeutlichen.

Beispiel:

Die $n_1 = 4$ Werte (U_i, δ_i) der ersten Gruppe seien

(5, 1), (4.5, 1), (12, 1), (6, 0)

und die $n_2 = 3$ Werte (V_j, ϵ_j) der zweiten Gruppe seien

(7, 1), (5, 1), (8, 0).

Nimmt man beide Stichproben zusammen, ordnet die nicht-zensierten Werte der Größe nach, erhält man (m = 5)

4.5 5 5 7 12.

In der folgenden Tabelle sind die Werte für R_i, e_{i1}, r_{i1} und a_i unterhalb der fünf Beobachtungen angegeben

$z_{(i)}$	4.5	5	5	7	12
R_i	7	6		3	1
e_{i1}	1	2		1	1
r_{i1}	4	3		1	1
a_i	1	1		0	1

Im Prinzip genügt für jede Vierfeldertafel der Wert von a_i, da sich die anderen Werte dann aus den gegebenen Randwerten ermitteln lassen.
Bezeichnen $E_0(a_i)$ und $Var_0(a_i)$ den Erwartungswert und die Varianz der Häufigkeit a_i bei Gültigkeit der Nullhypothese, die sich berechnen lassen, bildet man die Teststatistik

$$S = \frac{\sum_{i=1}^{m} \omega_i (a_i - E_0(a_i))}{[\sum_{i=1}^{m} \omega_i^2 \, Var_0(a_i)]^{1/2}}, \tag{3.2.44}$$

wobei die ω_i's Gewichte darstellen.
Wählt man $\omega_i = 1$ für alle i, so ergibt sich die Mantel-Haenszel-Prüfgröße. Dieser Test wird auch als Log-rank-Test oder Mantel-Cox-Test bezeichnet.
Wählt man $\omega_i = R_i$, ergibt sich die Prüfgröße von Gehan, die man auch aus dem Wilcoxon-Test in der Mann-Whitney-Variante (vgl. z. B. Schaich/Hamerle 1984, S. 116 ff.) ableiten kann. Im Gehan-Test werden kleine Beobachtungen höher gewichtet.
Tarone/Ware schlagen $\omega_i = \sqrt{R_i}$ vor und geben an, daß bei dieser Wahl der Gewichte der Test in bezug auf eine große Klasse von Alternativhypothesen eine besonders hohe Effizienz aufweist.
Bei $s > 2$ Stichproben ist die Vorgehensweise entsprechend zu verallgemeinern. Am grundlegenden Prinzip ändert sich nichts. Als Erweiterung des Gehan-Tests ergibt sich der Breslow-Test, und auch der Mantel/Haenszel-Test läßt sich auf $s > 2$ Stichproben verallgemeinern. Für die exakten Formeln für die Prüfgrößen vergleiche man zum Beispiel Breslow (1970), Gehan (1965), Lee/Desu (1972) oder Tarone/Ware (1977).
Die Prüfgrößen besitzen bei Gültigkeit von H_0 stets asymptotisch eine χ^2-Verteilung mit s - 1 Freiheitsgraden. Es handelt sich bei den angegebenen Tests um Omnibus-Tests, die lediglich die globale Nullhypothese $H_0 : S_1(t) = \ldots = S_s(t)$ für alle t überprüfen. Man kann aber auch Tests auf Trend gegen die Alternativhypothese $S_1(t) > S_2(t) > \ldots > S_s(t)$ konstruieren (siehe dazu z. B. Miller 1981, S. 110 ff.).
Insbesondere bei kleinen Stichprobenumfängen ist darauf zu achten, daß sich die Zensierungsmuster in den Subgruppen nicht allzusehr unterscheiden, denn unterschiedliche Zensierungsmechanismen können die Verteilung der Teststatistiken beeinflussen.

3.3 Einbeziehung von Kovariablen: Regressionsmodelle

3.3.1 Quantitative und qualitative Kovariablen

Neben der Verweildauer beziehungsweise Lebenszeit werden in der Regel für jedes Individuum oder Objekt eine Reihe von weiteren Kovariablen oder pro-

gnostischen Faktoren erhoben, und ein wichtiges Ziel der statistischen Analyse besteht in der quantitativen Ermittlung des Einflusses dieser exogenen oder endogenen Variablen.
Bei den Kovariablen kann es sich um *quantitative* oder um *qualitative Merkmale* handeln. Ein quantitatives Merkmal x_j wird wie in der herkömmlichen multiplen Regression mit einem Parameter β_j gewichtet und mit $x_j\,\beta_j$ in das Modell aufgenommen. Bei kategorialen Merkmalen geht man in Analogie zur Varianzanalyse über zu einer Kodierung der einzelnen Kategorien durch Dummy-Variablen.
Eine Möglichkeit für die Kodierung der Kategorien qualitativer Merkmale besteht in der *(0,1)-Kodierung* („cornered effects"). Besitzt ein Merkmal A I Kategorien (Ausprägungen, Klassen, Faktorstufen), so lassen sich diese durch I - 1 Dummy-Variablen erfassen in der Form

$$x_i^A = \begin{cases} 1 & \text{falls Kategorie i der Variablen A vorliegt} \\ 0 & \text{sonst} \end{cases}$$

$$i = 1, \ldots, I - 1. \qquad (3.3.1)$$

Die i-te Dummy-Variable x_i^A (i = 1, ..., I - 1) kodiert dabei nur das Vorliegen beziehungsweise Nicht-Vorliegen der i-ten Ausprägung. Das Vorliegen der I-ten (Referenz-)Kategorie ist implizit erfaßt durch die Kodierungen $x_i^A = 0$ für i = 1, ..., I - 1. Die Wahl der I-ten Kategorie als Referenzkategorie ist prinzipiell beliebig. Im Hinblick auf die Interpretation der Ergebnisse sollte jedoch eine Kategorie gewählt werden, auf die sich alle anderen Ausprägungen leicht beziehen lassen, da die Parameter β_j jeweils die „Abstände" der j-ten Ausprägung zur Referenzkategorie darstellen.
Besonders einfach ist der Spezialfall eines dichotomen unabhängigen Merkmals. Dann ist I = 2 und man erhält nur eine Dummy-Variable

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

Im allgemeinen Fall lassen sich mit $x_1^A, x_2^A, \ldots, x_{I-1}^A$ sämtliche Kategorien der qualitativen Variablen A kodieren. Die zugehörigen Regressionskoeffizienten β_j werden gewöhnlich wie in der Varianzanalyse *Haupteffekte* genannt.
Die (0,1)-Kodierung ist insbesondere bei Ansätzen mit gemischt quantitativ/qualitativen Kovariablen zweckmäßig. Bei ausschließlich qualitativen Kovariablen wird häufig auch die *Effekt-Kodierung* („centered effects") verwendet, die unmittelbar an die herkömmliche Varianzanalyse angelehnt ist. Man vergleiche dazu beispielsweise Hamerle/Kemény/Tutz (1984 S. 214).
Im Rahmen von Regressionsmodellen für Verweildauern und Lebenszeiten, insbesondere bei kategorialen unabhängigen Merkmalen, kommen auch *Interaktionswirkungen* als Einflußgrößen in Frage. Sie messen den gemeinsamen Einfluß einer bestimmten Kombination von Kategorien von zwei oder mehreren unabhängigen Merkmalen. Formal können sie in einfacher Weise durch die

Bildung entsprechender Produkte der Dummy-Variablen in den Regressionsansatz einbezogen werden. Für die Zwei-Faktor-Interaktionswirkungen der Faktoren A und B ergeben sich die Produkte $x_i^A x_j^B$, $i = 1, \ldots, I-1, j = 1, \ldots, J-1$, für die Drei-Faktor-Interaktionen die Produkte $x_i^A x_j^B x_k^C$, usw.
Die Werte der quantitativen Kovariablen einer Person beziehungsweise eines Objekts i sowie die Kodierungen für sämtliche Haupteffekte und im Modell enthaltene Interaktionswirkungen der qualitativen Kovariablen werden in einem Daten- oder Designvektor $\mathbf{x}_i$ zusammengefaßt. Die Dimension von $\mathbf{x}_i$ sei p.
Von Interesse ist die Art des Einwirkens der Kovariablen auf die Verweildauern beziehungsweise Lebenszeiten. Im allgemeinen wird – wie bei herkömmlichen Regressionsansätzen – davon ausgegangen, daß der Einfluß der Kovariablen oder prognostischen Faktoren linear in den Parametern erfolgt, also über eine Linearkombination

$$\eta_i = \mathbf{x}_i'\beta$$

mit einem unbekannten p-dimensionalen Parametervektor β. Die Parameter $\beta_1, \ldots, \beta_p$ repräsentieren die Einflußgewichte der Kovariablen. Im Gegensatz zur klassischen multiplen Regression geht man aber nicht davon aus, daß die Linearkombination $\eta_i = \mathbf{x}_i'\beta$ die Verweildauern beziehungsweise Lebenszeiten T_i direkt beeinflußt, sondern in der Regel eine Funktion von T_i, etwa $\ln T_i$.
Ein weiterer wichtiger Unterschied zur herkömmlichen Regression liegt darin, daß in den hier behandelten Verweildauermodellen einige Kovariablen selbst zeitabhängig sein können. Dies ist beispielsweise dann der Fall, wenn eine bestimmte medizinische Therapie nur während eines bestimmten Zeitraums angewendet wird. Das Untersuchungsziel könnte dann darin bestehen, den Einfluß dieser Therapie während der eigentlichen Anwendung oder in ihren Nachwirkungen zu überprüfen. Dafür definiert man zwei Dummy-Variablen, etwa $x_1(t)$ und $x_2(t)$ mit

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

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

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

Eine Möglichkeit der Analyse des Einflusses von Kovariablen auf die Verweildauern beziehungsweise Lebenszeiten besteht darin, ein Regressionsmodell zu formulieren, bei dem die Verteilung der Verweildauer beziehungsweise Lebenszeit von den Kovariablen abhängt. Bezeichnet $\mathbf{x}$ den Vektor der Kovariablen, so ist ein Modell für die Verweildauer beziehungsweise Lebenszeit T bei gegebenem Kovariablenvektor $\mathbf{x}$ zu spezifizieren.
Es ist naheliegend, eine Vorgehensweise analog zur herkömmlichen Regression zu wählen und die zu Beginn dieses Abschnitts eingeführten Verteilungen, wie zum Beispiel Exponential- oder Weibull-Verteilung, so zu verallgemeinern, daß ein oder mehrere Parameter in Abhängigkeit von den Kovariablen $\mathbf{x}$ modelliert werden. In der Regel wird dann die Verweildauer-Verteilung durch die zu den Kovariablen $\mathbf{x}$ gehörenden Regressionskoeffizienten und einen weiteren Parametervektor θ determiniert. Ein anderer Ansatz wurde von Cox (1972) vorgeschlagen. Dabei handelt es sich um einen semi-parametrischen Ansatz, der mit weniger Annahmen über die zugrundeliegenden Verteilungen der Verweildauern beziehungsweise Lebenszeiten auskommt. Das Cox-Modell wird im Anschluß an die vollparametrisierten Modelle behandelt, die im folgenden dargestellt werden.

3.3.2 Parametrische Regressionsmodelle

Das Exponential-Modell

Anhand des einfachen Modells einer exponentialverteilten Verweildauer beziehungsweise Lebenszeit soll die Vorgehensweise zur Konstruktion von Regressionsmodellen ausführlich demonstriert werden.
Nach (3.2.18) ist die Dichte von T ohne Berücksichtigung von Kovariablen

$$f(t) = \lambda \exp(-\lambda t) \quad , t \geq 0, \lambda > 0.$$

Die Exponentialverteilung wird determiniert durch einen Parameter λ. Die durchschnittliche Verweildauer ist $1/\lambda$. Es liegt nahe, den Einfluß der Kovariablen $\mathbf{x}$ auf die Verweildauer über diesen Parameter, etwa in der Form $1/\lambda = g(\mathbf{x};\beta)$ mit einem unbekannten Parametervektor β, zu modellieren. Dies entspricht der Vorgehensweise der herkömmlichen multiplen Regression, bei der für eine nach $N(\mu, \sigma^2)$ verteilten abhängigen Variablen y der Erwartungswert in der Form $\mu = \mathbf{x}'\beta$ parametrisiert wird. Geht man hier davon aus, daß die Kovariablen über eine Linearkombination $\mathbf{x}'\beta$ auf die Verweildauer beziehungsweise Lebenszeit T einwirken, erhält man die Parametrisierung

$$1/\lambda = g(\mathbf{x}'\beta).$$

Nun ist nur noch die Funktion g zu spezifizieren. Die Wahl der identischen Funktion $g(z) = z$, also

$$1/\lambda = \mathbf{x}'\beta,$$

ist hier nicht zweckmäßig, da der Parameter λ der Restriktion $\lambda > 0$ unterliegt. Diese Nebenbedingung hätte beim Ansatz $1/\lambda = \mathbf{x}'\boldsymbol{\beta}$ unerwünschte und möglicherweise nicht kontrollierbare Restriktionen für die Parameter $\boldsymbol{\beta}$ zur Folge (vgl. z. B. Mantel/Myers 1971). Es ist besser, von vornherein eine Funktion g zu wählen, die nur positive Werte annehmen kann. Eine einfache Möglichkeit ist

$$g(\mathbf{x}'\boldsymbol{\beta}) = \exp(\mathbf{x}'\boldsymbol{\beta})$$

oder

$$g(\mathbf{x}'\boldsymbol{\beta}) = \lambda_0 \exp(\mathbf{x}'\boldsymbol{\beta}) \qquad (\lambda_0 > 0). \tag{3.3.2}$$

Setzt man $\beta_0 = \ln \lambda_0$, so resultiert für (3.3.2)

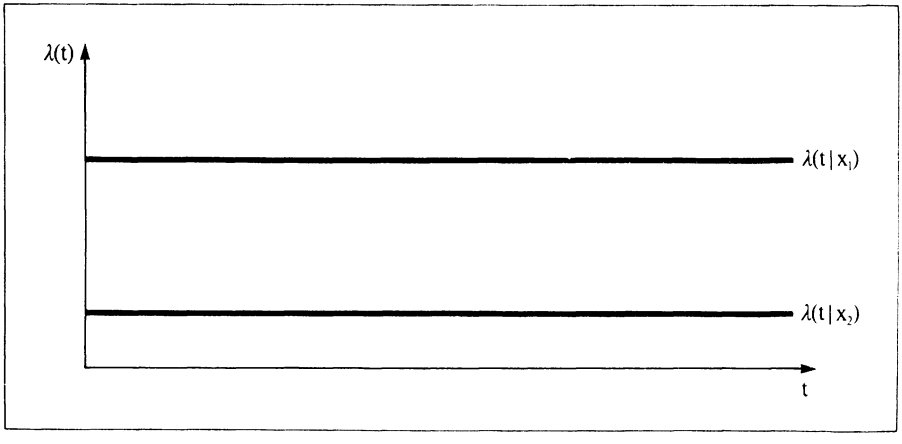
$$g(\mathbf{x}'\boldsymbol{\beta}) = \exp(\beta_0 + \mathbf{x}'\boldsymbol{\beta}). \tag{3.3.3}$$

λ_0 beziehungsweise β_0 ist das konstante Glied des Regressionsansatzes. Mit (3.3.3) besitzt nun T eine Exponentialverteilung mit dem Parameter $\lambda = \exp(-\beta_0 - \mathbf{x}'\boldsymbol{\beta})$, und die Hazardrate ist

$$\lambda(t|\mathbf{x}) = \exp(-\beta_0 - \mathbf{x}'\boldsymbol{\beta}). \tag{3.3.4}$$

Die Heterogenität innerhalb der Population ist in den Kovariablen enthalten und über die Zeit konstant. Unterscheiden sich zwei Individuen in den Kovariablen, so unterscheiden sich auch die Hazardraten der beiden Individuen. Dies wird in Abbildung 3.9 veranschaulicht.

Abbildung 3.9: Hazardraten zweier Individuen bei exponentialverteilten Verweildauern



Die Hazardraten sind zeitunabhängig, und die Rate des einen Individuums ist ein Vielfaches der Rate des anderen Individuums. Die letzte Eigenschaft der „proportionalen Hazards", die beim Exponentialmodell trivialerweise erfüllt ist, gilt auch für einige wichtige Modellklassen mit zeitabhängiger Hazardrate.
Die Ähnlichkeit zu den herkömmlichen Modellen der multiplen Regression wird

noch deutlicher, wenn man zu $y = \ln T$, den logarithmierten Werten der Verweildauer, übergeht. Die Variable $y = \ln T$ besitzt die Dichtefunktion

$$g(y) = \exp(y + \ln \lambda - \exp(y + \ln \lambda)) \qquad -\infty < y < \infty. \tag{3.3.5}$$

Schreibt man den Regressionsansatz in der gewohnten Form ($\ln \lambda = -\beta_0 - \mathbf{x}'\beta$)

$$y = \beta_0 + \mathbf{x}'\beta + \omega, \tag{3.3.6}$$

so besitzt die Fehlervariable ω eine Standardextremwertverteilung mit der Dichte

$$f(\omega) = \exp(\omega - \exp(\omega)) \qquad -\infty < \omega < \infty. \tag{3.3.7}$$

Eine graphische Darstellung der Dichtefunktion der Standardextremwertverteilung wurde in Abbildung 3.6 gegeben.
Man beachte, daß die Fehlervariable ω im Regressionsmodell (3.3.6) eine Verteilung besitzt, in der kein Parameter mehr frei wählbar ist. Dies bedeutet eine starke Einschränkung und ist im herkömmlichen Ansatz der multiplen Regression nicht der Fall. Hier wird im einfachsten Modell angenommen, daß die Fehlervariable $N(0;\sigma^2)$ verteilt ist. Dabei ist σ^2 ein weiterer unbekannter Parameter, der aus den Daten zu schätzen ist. Aus diesem Grunde liegt es nahe, den Ansatz (3.3.6) zu erweitern und wie in der multiplen Regression

$$y = \beta_0 + \mathbf{x}'\beta + \sigma\ \omega$$

zu modellieren, wobei ω die Verteilung (3.3.7) besitzt. Dies geschieht im nächsten Abschnitt und ist eng mit der Weibull-Verteilung für die Verweildauern beziehungsweise Lebenszeiten verknüpft.

Das Weibull-Regressionsmodell

Geht man aus vom Regressionsansatz ($y = \ln T$)

$$y = \beta_0 + \mathbf{x}'\beta + \sigma\ \omega, \tag{3.3.8}$$

wobei ω wieder eine Standardextremwertverteilung mit der Dichte (3.3.7) besitzt, so läßt sich dieses Modell leicht auf die Verweildauer beziehungsweise Lebenszeit T selbst übertragen. T besitzt bei gegebenem Kovariablenvektor $\mathbf{x}$ die Dichtefunktion

$$f(t|\mathbf{x}) = \frac{\delta}{\exp(\beta_0 + \mathbf{x}'\beta)} \cdot \left(\frac{t}{\exp(\beta_0 + \mathbf{x}'\beta)}\right)^{\delta-1} \exp\left[-\left(\frac{t}{\exp(\beta_0 + \mathbf{x}'\beta)}\right)^{\delta}\right]$$

$$t \geq 0,\ \delta = 1/\sigma. \tag{3.3.9}$$

(3.3.9) entspricht aber genau der Dichte einer Weibull-Verteilung, wobei für den Parameter λ die Parametrisierung $\lambda = \exp(-\beta_0 - \mathbf{x}'\beta)$ beziehungsweise $1/\lambda = \exp(\beta_0 + \mathbf{x}'\beta)$ gewählt wurde. Der andere Parameter $\delta = 1/\sigma$ hängt nicht von den Kovariablen ab und ist für alle Individuen gleich.

Die Hazardrate des Weibull-Regressionsmodells ist gegeben durch

$$\lambda(t|x) = \frac{\delta}{\exp(\beta_0 + x'\beta)} \left(\frac{t}{\exp(\beta_0 + x'\beta)} \right)^{\delta-1}. \qquad (3.3.10)$$

Das Weibull-Regressionsmodell gehört zur sogenannten Klasse der Proportional-Hazards-Modelle, wie man leicht nachrechnet. Besitzen zwei Individuen die Kovariablenvektoren x_1 und x_2, so erhält man für den Quotienten der beiden Hazardraten

$$\frac{\lambda(t|x_1)}{\lambda(t|x_2)} = \exp((x_2 - x_1)'\beta)$$

unabhängig von der Zeit t.
Die Annahme, daß der Parameter $\delta = 1/\sigma$ für alle Individuen beziehungsweise Objekte derselbe ist, entspricht der Voraussetzung der Varianzhomogenität in der herkömmlichen multiplen Regression.
Für $\delta = \sigma = 1$ ergibt sich wieder das Exponential-Regressionsmodell.
Die Analogie von (3.3.8) zum Modell der multiplen Regression ist unmittelbar ersichtlich. Im herkömmlichen multiplen Regressionsmodell wird für ω eine Standardnormalverteilung angenommen. Die Fehlervariable ω hat dann den Erwartungswert 0. Dies ist bei der Standardextremwertverteilung nicht der Fall. Der Erwartungswert ist hier -0.5772 ..., wobei 0.5772 ... $= \Gamma'(1)$ die Eulersche Konstante ist. Die Varianz der Standardextremwertverteilung ist $\frac{\pi^2}{6} = 1.64493 \ldots$ Die Varianz von y ist $\frac{\pi^2}{6}\sigma^2$ und der Erwartungswert von y ist $\beta_0 + x'\beta - 0.5772\,\sigma$, in dem auch der Parameter σ enthalten ist.

Log-Normalverteilungs-Regressionsmodelle

Eine naheliegende Möglichkeit besteht darin, für ω in (3.3.8) eine Standardnormalverteilung anzunehmen. Die Verweildauer beziehungsweise Lebenszeit T selbst besitzt dann eine sogenannte Log-Normalverteilung mit der Dichtefunktion

$$f(t|x) = \frac{1}{\sqrt{2\pi}\,\sigma\,t} \exp\left[-\frac{1}{2}\left(\frac{\ln t - \mu(x)}{\sigma}\right)^2\right] \qquad t > 0, \qquad (3.3.11)$$

wobei $\mu(x) = \beta_0 + x'\beta$ bedeutet.
Survivorfunktion und Hazardrate berechnen sich zu

$$S(t|x) = 1 - \Phi\left(\frac{\ln t - \mu(x)}{\sigma}\right)$$

und

$$\lambda(t|x) = f(t|x) / S(t|x), \qquad (3.3.12)$$

wobei $\Phi(z)$ die Verteilungsfunktion der Standardnormalverteilung bezeichnet. Die Hazardrate steigt zunächst an bis zu einem maximalen Wert und fällt dann wieder bis 0 für $t \to \infty$. Aufgrund des umfangreichen numerischen Aufwands bei zensierten Daten wird die Log-Normalverteilung bei der Analyse von Verweildauern oder Lebenszeiten nur wenig eingesetzt. Sie kann gut durch die log-logistische Verteilung (vgl. (3.3.13)) approximiert werden.
Man beachte, daß Log-Normalverteilungs-Regressionsmodelle nicht zur Klasse der Proportional-Hazards-Modelle gehören.

Log-logistische Regressionsmodelle

Gemäß (3.2.30) ist die log-logistische Hazardrate gegeben durch

$$\lambda(t) = \frac{\lambda\,\alpha(\lambda t)^{\alpha-1}}{1+(\lambda t)^{\alpha}}.$$

Ein Regressionsmodell erhält man dadurch, daß man den Parameter λ in Abhängigkeit von den Kovariablen parametrisiert, etwa in der Form $\lambda(x) = \exp(x'\beta)$. Damit ergibt sich

$$\lambda(t|x) = \frac{\exp(x'\beta)\,\alpha(\exp(x'\beta)\,t)^{\alpha-1}}{1+(\exp(x'\beta)t)^{\alpha}}. \tag{3.3.13}$$

Allgemeine log-lineare Regressionsmodelle

Die Regressionsansätze (3.3.6), (3.3.8) und (3.3.11) gehören zu den *log-linearen Regressionsmodellen*, bei denen ein linearer Zusammenhang zwischen den Kovariablen und der logarithmierten Verweildauer beziehungsweise Lebenszeit $y = \ln T$ unterstellt wird.
Allgemein ist ein log-lineares Regressionsmodell gegeben durch den Ansatz

$$y = \beta_0 + x'\beta + \sigma\,\omega \tag{3.3.14}$$

mit einer von x unabhängigen Verteilung von ω. Die Verteilung von y kann geschrieben werden als

$$\frac{1}{\sigma} f(u)$$

mit $u = (y - \beta_0 - x'\beta)/\sigma$. Der Parameter σ ist ein Skalenparameter und β_0 kennzeichnet die allgemeine Lokalisation von $\ln T$.

Das Gompertz-(Makeham-)Regressionsmodell

Die Gompertz-Makeham-Hazardrate ohne Berücksichtigung von Kovariablen (für eine homogene Population) ist in (3.2.33) gegeben durch

$$\lambda(t) = \alpha_0 + \lambda_0 \exp(\gamma_0 t). \tag{3.3.15}$$

Es besteht nun die Möglichkeit, in Analogie zu den bisher betrachteten Ansätzen einen oder mehrere der Parameter in (3.3.15) in Abhängigkeit vom Kovaria-

blenvektor $\mathbf{x}$ zu modellieren, etwa in der Form $\lambda_0(\mathbf{x}) = \exp(\mathbf{x}'\beta)$ und $\gamma_0(\mathbf{x}) = \mathbf{x}'\gamma$. Allerdings ist darauf zu achten, daß das Modell identifizierbar bleibt, das heißt, daß nicht zwei verschiedene Parameterkonstellationen beobachtungsäquivalent sind und zur selben Wahrscheinlichkeitsverteilung für die beobachteten Daten führen.

Periodische Veränderung von Kovariablen oder Parametern

Eine von Tuma/Hannan (1984, Kap. 7.2) dargestellte weitere Möglichkeit der Einführung von zeitabhängigen Hazardraten besteht darin, die Zeitachse in Intervalle zu unterteilen und innerhalb dieser Intervalle die Hazardrate mit einem speziellen Ansatz zu modellieren, allerdings mit möglicherweise von Intervall zu Intervall variierenden Kovariablen und/oder Parametern.

Sei die Zeitachse zerlegt in $q + 1$ Intervalle

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

wobei in der Regel $a_0 = 0$ gesetzt und für a_q das Ende des Beobachtungszeitraums genommen wird. Für die Hazardrate wird im einfachsten Fall

$$\lambda_p(t|\mathbf{x}_p) = \exp(\mathbf{x}'_p\beta_p) \qquad \text{für } t \in [a_{p-1}, a_p) \tag{3.3.16}$$

angesetzt, das heißt, innerhalb jedes Intervalls wird eine exponentialverteilte Verweildauer postuliert.
In (3.3.16) wird angenommen, daß die Intervallgrenzen für alle Individuen dieselben sind. Die Intervallgrenzen müssen exogen vorgegeben sein. Darüber hinaus ist zu beachten, daß das Modell selbst bei einer mittleren Zahl von Zeitintervallen bereits eine große Zahl zu schätzender Parameter enthält und entsprechend große Datenmengen benötigt. Es sind auch Spezialfälle von (3.3.16) möglich, so können zum Beispiel nur die Kovariablen von Intervall zu Intervall variieren und die Parameter β über die Zeit hinweg konstant bleiben.
Zur Schätzung der unbekannten Parameter ist die Survivorfunktion von Bedeutung (vgl. Abschnitt 3.6.3). Analog zu (3.2.10) erhält man

$$S(t|\mathbf{x}) = \exp(-\int_0^t \lambda(u|\mathbf{x})\, du).$$

Aus der Linearität des Integrals folgt

$$S(t|\mathbf{x}) = \exp(-\sum_{i=1}^{p-1} \exp(\mathbf{x}'_i\beta_i)\,(a_i - a_{i-1}) - \exp(\mathbf{x}'_p\beta_p)\,(t - a_{p-1}))$$

$$\text{für } a_{p-1} \leq t < a_p, \quad k = 1, \ldots, q+1 \quad (a_{q+1} = \infty). \tag{3.3.17}$$

Dabei enthält der Kovariablenvektor $\mathbf{x}$ in (3.3.16) und (3.3.17) sämtliche bis zum Zeitpunkt t erhobenen Kovariablen eines Individuums.

Abschließend sei noch bemerkt, daß in einem konkreten Anwendungsfall statt (3.3.16) auch die in Abschnitt 3.10 behandelten diskreten Regressionsmodelle in Betracht gezogen werden sollten.

3.3.3 Das Proportional-Hazards-Regressionsmodell von Cox

Das Proportional-Hazards-Modell (PH-Modell) wurde von Cox (1972) vorgeschlagen. Während die bisher betrachteten Ansätze davon ausgingen, daß die Hazardrate und damit die Verteilung der Verweildauer beziehungsweise Lebenszeit bis auf einige Parameter bekannt sind, handelt es sich beim Cox-Modell um einen semiparametrischen Ansatz mit einer unspezifizierten „Baseline"-Hazardrate.
Die Kovariablen werden wieder zusammengefaßt zu einem p-dimensionalen Vektor $\mathbf{x}$, β sei der zugehörige p-dimensionale Parametervektor und T sei die Verweildauer beziehungsweise Lebenszeit. Die Hazardrate des PH-Modells ist

$$\lambda(t|\mathbf{x}) = \lambda_0(t)\ \exp(\mathbf{x}'\beta). \tag{3.3.18}$$

$\lambda_0(t)$ ist die beliebige, nicht spezifizierte Grundhazardrate. Dadurch wird mehr Flexibilität in der Modellierung erreicht, allerdings sind bei der Parameterschätzung andere Methoden zu verwenden als bei den bisher betrachteten Modellen.
Die Proportionalität der Hazardraten ergibt sich aus der Betrachtung des Quotienten

$$\frac{\lambda(t|\mathbf{x}_1)}{\lambda(t|\mathbf{x}_2)}$$

für zwei Individuen beziehungsweise Objekte mit den Kovariablen $\mathbf{x}_1$ und $\mathbf{x}_2$. Aus (3.3.18) erhält man

$$\frac{\lambda(t|\mathbf{x}_1)}{\lambda(t|\mathbf{x}_2)} = \exp((\mathbf{x}_1 - \mathbf{x}_2)'\beta).$$

Der Quotient hängt nicht von der Zeit t ab (vgl. Abbildung 3.10).
Die Proportionalität der Hazardraten basiert im wesentlichen auf der multiplikativen Einwirkung der Kovariablen auf die Baseline-Hazardrate und auf der Faktorisierung der Hazardrate in einen Term, der nur von der Zeit, und in einen Term, der nur von den Kovariablen abhängt, im allgemeinen

$$\lambda(t|\mathbf{x}) = \lambda_0(t)\ g(\mathbf{x};\beta) \qquad , g(\cdot) > 0. \tag{3.3.19}$$

Im Cox-Modell wird $g(\mathbf{x};\beta) = \exp(\mathbf{x}'\beta)$ gesetzt.
Die Annahme proportionaler Hazardraten bedeutet natürlich auch eine Einschränkung der Anwendungsmöglichkeiten des Modells. So darf beispielsweise bei Einbeziehung der Kovariablen „Geschlecht" das Verhältnis der Hazardraten von Männern und Frauen nicht mit der Zeit variieren. Die Voraussetzung

proportionaler Hazardraten kann etwas gelockert werden, indem man schichtspezifische Hazardraten einführt. Besitzen eine oder mehrere (kategoriale oder kategorisierte) Kovariablen keinen multiplikativen Effekt auf die Hazardrate, so können die Kategorien dieser Kovariablen zur Bildung von Schichten beziehungsweise Teilpopulationen herangezogen werden. Erhält man J Schichten, so wird für jede Schicht der Ansatz

$$\lambda_j(t|\mathbf{x}) = \lambda_{0j}(t)\ \exp(\mathbf{x}'\beta) \qquad j = 1, \ldots, J, \tag{3.3.20}$$

mit einer schichtspezifischen Baseline-Hazardrate formuliert. In dem Kovariablenvektor $\mathbf{x}$ sind dann nur noch die verbleibenden Kovariablen enthalten.
Es besteht die Möglichkeit, mit Hilfe von speziell konstruierten zeitabhängigen Kovariablen einen statistischen Test zur Überprüfung der Proportionalität durchzuführen. Man vergleiche dazu die Ausführungen in Abschnitt 3.7.

Für die Survivorfunktion des Cox-Modells resultiert nach Anwendung von (3.2.10)

$$\begin{aligned} S(t|\mathbf{x}) &= \exp(-\int_0^t \lambda(u|\mathbf{x})\,du) \\ &= \exp(-\int_0^t \lambda_0(u)\exp(\mathbf{x}'\beta)\,du) \\ &= \exp(-\int_0^t \lambda_0(u)\,du)^{\exp(\mathbf{x}'\beta)} \\ &= S_0(t)^{\exp(\mathbf{x}'\beta)} \end{aligned} \tag{3.3.21}$$

und für die Dichtefunktion der Verweildauer beziehungsweise Lebenszeit ergibt sich

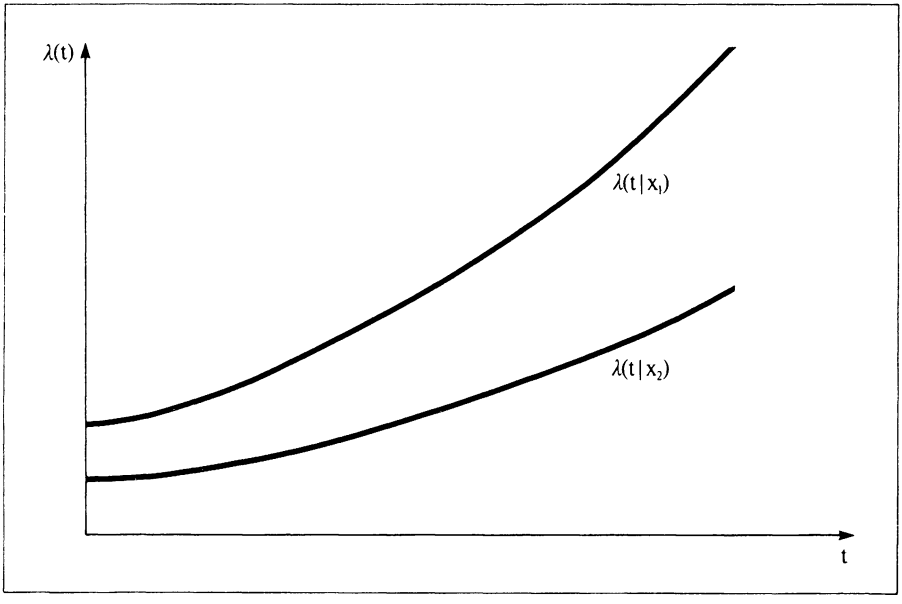
$$f(t|\mathbf{x}) = \lambda(t|\mathbf{x}) \cdot S(t|\mathbf{x}) = \lambda_0(t)\exp(\mathbf{x}'\beta)\,S_0(t)^{\exp(\mathbf{x}'\beta)}. \tag{3.3.22}$$

Das Proportional-Hazards-Modell (3.3.18) wurde von Cox (1972) vorgeschlagen, der auch eine Schätzmethode zur Schätzung von β und $\lambda_0(t)$ angegeben hat, ohne daß über die Baseline-Hazardrate besondere Annahmen (außer natürlich $\lambda_0(t) \geq 0$) getroffen werden müssen. Mittlerweile sind eine Vielzahl von Literaturbeiträgen zum Cox-Modell erschienen, und es ist das am meisten angewendete Modell in der Praxis. Eine ausführliche Darstellung des PH-Modells findet man bei Kalbfleisch/Prentice (1980, Kap. 4 und 5) oder Lawless (1982, Kap. 7). Wird für die Baseline-Hazardrate eine spezielle parametrische Form angenommen, erhält man ein voll parametrisiertes Proportional-Hazards-Modell. Spezialfälle davon sind die bereits behandelten Weibull- und Exponential-Regressionsmodelle. Im Weibull-Regressionsmodell ist

$$\lambda_0(t) = \delta\,\lambda_0\,(\lambda_0 t)^{\delta-1}, \tag{3.3.23}$$

das Exponentialmodell erhält man aus (3.3.23) mit $\delta = 1$.

Abbildung 3.10: Verlauf von zwei proportionalen Hazardraten



3.4 Mehr-Zustands-Modelle — Competing Risks

Im vorangegangenen Abschnitt wurde bei den Ein-Episoden-Modellen stets davon ausgegangen, daß lediglich ein Übergang in einen absorbierenden Endzustand erfolgen kann. Eine naheliegende Erweiterung besteht darin, Übergänge in mehrere Endzustände zu betrachten. In der Analyse von Lebens- oder Überlebenszeiten werden diese „competing risks" gewöhnlich durch verschiedene Todesursachen (oder Ereignisarten) repräsentiert, in der Technik durch verschiedenartige Defekte, die jeweils den Ausfall eines Geräts bewirken.
Über Competing-Risks-Modelle ohne Einbeziehung von Kovariablen existieren bereits viele Literaturbeiträge. Für einen Überblick zur vorhandenen Literatur vergleiche man Gail (1975). Der Ein-Episoden-Fall mit Kovariablen wird unter anderem von David/Moeschberger (1978), Seal (1977) und in der „Survival"-Literatur von Holt (1978), Prentice/Breslow (1978) sowie Prentice u.a. (1978) behandelt.
Für jedes Individuum beziehungsweise Objekt wird neben der Verweildauer beziehungsweise Lebenszeit T nun eine Zustandsvariable Y beobachtet, die Werte aus der Menge der möglichen Endzustände (oder Zielzustände) {1, ..., m} annehmen kann. Darüber hinaus wird für jedes Individuum beziehungsweise Objekt eine Reihe von Kovariablen oder prognostischen Faktoren erhoben, die zusammen mit zusätzlich aufgenommenen Interaktionswirkungen

zum p-dimensionalen Kovariablenvektor x zusammengefaßt werden. Der Einfachheit halber werden die Kovariablen oder prognostischen Faktoren zunächst wieder als zeitunabhängig vorausgesetzt.
Ein geeigneter Ausgangspunkt zur Analyse der Relationen zwischen Kovariablen und den Übergängen in die verschiedenen Endzustände ist wieder die Hazardrate. Eine übergangsspezifische (ursachenspezifische; ereignisspezifische) Hazardrate läßt sich wie folgt definieren:

$$\lambda_j(t|\mathbf{x}) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} P(t \leq T < t + \Delta t,\ Y = j | T \geq t, \mathbf{x}). \tag{3.4.1}$$

(3.4.1) ist der Grenzwert der (bedingten) Wahrscheinlichkeit, daß ein Individuum im Zeitintervall $[t, t + \Delta t)$ zum Zustand j wechselt, gegeben die Kovariablen und die Voraussetzung, daß bis zum Zeitpunkt t noch kein Übergang in irgendeinen der konkurrierenden Zielzustände stattgefunden hat.
Die Gesamthazardrate zum Zeitpunkt t ergibt sich dann aus der Summe aller zielspezifischen Hazardraten

$$\lambda(t|\mathbf{x}) = \sum_{j=1}^{m} \lambda_j(t|\mathbf{x}). \tag{3.4.2}$$

Die Gesamthazardrate ist der Grenzwert der (bedingten) Wahrscheinlichkeit, daß im Zeitintervall $[t, t + \Delta t)$ ein Übergang stattfindet, gegeben die Kovariablen und die Voraussetzung, daß bis zum Zeitpunkt t noch kein Übergang stattgefunden hat.
Die Survivorfunktion ist

$$S(t|\mathbf{x}) = \exp\left(-\int_0^t \lambda(u|\mathbf{x})\, du\right). \tag{3.4.3}$$

Setzt man in (3.4.3) die übergangsspezifischen Hazardraten ein, erhält man

$$\begin{aligned} S(t|\mathbf{x}) &= \exp\left(-\int_0^t \sum_{j=1}^{m} \lambda_j(u|\mathbf{x})\, du\right) \\ &= \prod_{j=1}^{m} \exp\left(-\int_0^t \lambda_j(u|\mathbf{x})\, du\right). \end{aligned} \tag{3.4.4}$$

Gelegentlich ist es zweckmäßig, auch „übergangsspezifische Dichten" $f_j(t|\mathbf{x})$ einzuführen durch

$$\begin{aligned} f_j(t|\mathbf{x}) &= \lim_{\Delta t \to 0} \frac{1}{\Delta t} P(t \leq T < t + \Delta t,\ Y = j|\mathbf{x}) \\ &= \lambda_j(t|\mathbf{x}) \cdot S(t|\mathbf{x}), \qquad j = 1, \ldots, m. \end{aligned} \tag{3.4.5}$$

Man beachte, daß $f_j(t|\mathbf{x})$ nicht die Dichtefunktion der Verweildauer beziehungsweise Lebenszeit ist. Insbesondere gilt

$$\int_0^\infty f_j(t|\mathbf{x})\, dt = P(Y = j|\mathbf{x}) = \pi_j(\mathbf{x}). \tag{3.4.6}$$

$\pi_j(\mathbf{x})$ ist die Wahrscheinlichkeit eines Übergangs in den j-ten Endzustand (j = 1, ..., m) bei gegebenem Kovariablenvektor $\mathbf{x}$. Man hat die Beziehung

$$\sum_{j=1}^{m} \pi_j(\mathbf{x}) = 1.$$

Spezifikationen der Hazardrate

Im Prinzip können alle bisher behandelten Regressionsmodelle auf den Mehr-Zustands-Fall übertragen werden. Dabei können sowohl die zu den Kovariablen gehörenden Parameter als auch die sonstigen in der Hazardrate enthaltenen Parameter von Endzustand zu Endzustand variieren. Allerdings ist darauf zu achten, daß das Gesamtmodell nicht zu viele Parameter enthält, denn dies könnte die Schätzgenauigkeit beeinträchtigen und sehr große Stichprobenumfänge erfordern. Ein Ziel der Modellbildung sollte auch darin bestehen, den empirischen Sachverhalt möglichst einfach zu beschreiben, das heißt ein Modell mit möglichst wenigen Parametern zu finden. Allerdings sollten die einfachen Modelle dem empirischen Befund auch angemessen sein.
Geht man vom Cox-Modell aus, ergeben sich die übergangsspezifischen Hazardraten

$$\lambda_j(t|\mathbf{x}) = \lambda_{0j}(t) \exp(\mathbf{x}'\beta_j) \qquad j = 1, \ldots, m. \tag{3.4.7}$$

Dabei sind sowohl die Baseline-Hazardfunktionen $\lambda_{0j}(t)$ als auch die Regressionskoeffizienten β_j vom Endzustand abhängig. Einen wichtigen Spezialfall erhält man, wenn die Baseline-Hazardraten als zueinander proportional angenommen werden mit einem Proportionalitätsfaktor $\exp(\beta_{0j})$, also

$$\lambda_{0j}(t) = \lambda_0(t) \exp(\beta_{0j}). \tag{3.4.8}$$

Damit die Beziehung (3.4.8) eindeutig wird, kann man beispielsweise $\beta_{01} = 0$ setzen. Setzt man (3.4.8) in (3.4.7) ein, resultiert

$$\lambda_j(t|\mathbf{x}) = \lambda_0(t) \exp(\beta_{0j} + \mathbf{x}'\beta_j). \tag{3.4.9}$$

Eine Konsequenz des Modells (3.4.9) ist die relativ einfache Berechenbarkeit von $P(Y = j|\mathbf{x}) = \int_0^\infty f_j(t|\mathbf{x})dt$, der Wahrscheinlichkeit, daß ein Individuum mit gegebenem Kovariablenvektor $\mathbf{x}$ im j-ten Endzustand endet. Es läßt sich zeigen (Kalbfleisch/Prentice, 1980, S. 171), daß

$$P(Y = j|\mathbf{x}) = \frac{\exp(\beta_{0j} + \mathbf{x}'\beta_j)}{\sum_{k=1}^{m} \exp(\beta_{0k} + \mathbf{x}'\beta_j)} \qquad j = 1, \ldots, m \tag{3.4.10}$$

gilt. (3.4.10) ist ein logistisches Modell für die Übergangswahrscheinlichkeiten. Eine weitere wichtige Spezifikation ist der Weibull-Regressionsansatz für Competing Risks. Läßt man sowohl die Regressionskoeffizienten als auch die Para-

meter der Baseline-Hazardrate von Endzustand zu Endzustand variieren, erhält man aus (3.3.23) und aus (3.3.10) nach einer Umparametrisierung

$$\lambda_j(t|x) = \delta_j \lambda_{0j} (\lambda_{0j} t)^{\delta_j - 1} \exp(x'\beta_j) \qquad j = 1, \ldots, m \tag{3.4.11}$$

für die übergangsspezifischen Hazardraten. Die unbekannten Parameter δ_j, λ_{0j} der Baseline-Hazardrate sowie die zu den Kovariablen gehörenden Parametervektoren $\beta_1, \ldots, \beta_m$ sind aus dem Datenmaterial zu schätzen.

3.5 Regressionsmodelle für den Mehr-Episoden-Fall

Die in den vorangegangenen Abschnitten behandelten Modelle für eine Episode sind insbesondere bei der Analyse von Lebens- beziehungsweise Überlebenszeiten einsetzbar, bei denen der Endzustand beziehungsweise die Endzustände in der Regel absorbierend sind. Handelt es sich bei den Verweildauern nicht um Zeitdauern, bei denen sämtliche Zielzustände absorbierend sind, so können für jedes Individuum beziehungsweise Objekt mehrere Übergänge auftreten. Beispiele hierfür sind die Verweildauern in verschiedenen Berufen bei der Untersuchung von Berufskarrieren, die Dauern der Arbeitslosigkeit in möglicherweise mehreren aufeinanderfolgenden Episoden, die Dauern bis zum Umzug in eine andere Region bei Wanderungs- und Mobilitätsanalysen, die Nutzungsdauern von langlebigen Konsumgütern usw.
Im folgenden betrachten wir zunächst nur den Spezialfall, daß ein bestimmtes Ereignis wiederholt auftritt. Im nächsten Abschnitt wird dann der allgemeine Fall eines Mehr-Episoden- und Mehr-Zustands-Modells behandelt.
Im statistischen Modell werden die Zeitpunkte, zu denen Ereignisse stattfinden, repräsentiert durch nicht-negative Zufallsvariablen $0 = T_0 < T_1 < T_2 < \ldots$. Die Verweildauern, das heißt die Länge der Episoden, sind gegeben durch

$$V_k = T_k - T_{k-1} \qquad , k = 1, 2, \ldots \tag{3.5.1}$$

Für jedes Individuum beziehungsweise Objekt wird in jeder Episode ein Vektor von Kovariablen x_k gemessen, von denen einige auch zeitabhängig sein können. Wir gehen aus Gründen der Einfachheit zunächst wieder davon aus, daß alle Kovariablen zeitunabhängig sind. Die Anzahl der erhobenen Kovariablen kann jedoch von Episode zu Episode variieren.
Durchläuft ein Individuum i den Prozeß bis zur n_i-ten Episode, so wird der Pfad des Individuums folgendermaßen generiert:
1. Der Pfad des Individuums beginnt zum Zeitpunkt $T_0 = 0$. Ist der Anfangszeitpunkt der ersten Episode nicht bekannt, spricht man von einer *Zensierung von links*. Dieses Problem ist wesentlich schwieriger zu behandeln als die Zensierung von rechts, da es im allgemeinen nicht möglich ist, die Auswirkungen der nicht bekannten Vorgeschichte auf zukünftige Ereignisse abzuschätzen. Man benötigt dann restriktive Annahmen, wie zum Beispiel eine vorgegebene Verteilung des Anfangszeitpunkts, die gewöhnlich empirisch nicht überprüfbar sind.

Im folgenden setzen wir stets voraus, daß entweder der Startzeitpunkt vorgegeben ist (ohne Beschränkung der Allgemeinheit dann $t_0 = 0$) oder daß die Vorgeschichte des Prozesses vor dem Beobachtungszeitraum den weiteren Verlauf des Prozesses nicht beeinflußt. Dies ist nur dann der Fall, wenn die Verweildauern exponentialverteilt sind.
Die Dauer der ersten Episode wird gesteuert durch die Hazardrate beziehungsweise Survivorfunktion

$$\lambda^1(t|x_1) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} P(t \leq T_1 < t + \Delta t \mid T_1 \geq t, x_1) \tag{3.5.2}$$

beziehungsweise

$$S^1(t|x_1) = \exp(-\int_0^t \lambda^1(u|x_1)\, du). \tag{3.5.3}$$

Für die Dichtefunktion von T_1 erhält man

$$f^1(t|x_1) = \lambda^1(t|x_1)\ S^1(t|x_1) \tag{3.5.4}$$

wie im Ein-Episoden-Fall.
2. Zum Zeitpunkt $T_1 = t_1$ tritt für das Individuum das erste Ereignis ein und die zweite Episode beginnt. Die Hazardrate der zweiten Episode ist gegeben durch

$$\lambda^2(t|x_2, H_1) \qquad t \geq t_1. \tag{3.5.5}$$

Man beachte, daß $\lambda^2(t|x_2, H_1)$ identisch gleich Null ist für $t < t_1$. In H_1 wird die Vorgeschichte des Prozesses (mit den Kovariablen) zusammengefaßt, also $H_1 = \{t_1, x_1\}$. Die Hazardrate $\lambda^2(t|x_2, H_1)$ kann auch von Daten aus H_1 abhängen. So wird beispielsweise die Dauer der zweiten Arbeitslosigkeit mit großer Sicherheit von der Dauer der vorangegangenen Arbeitslosigkeitsepisode (also hier t_1) beeinflußt (vgl. auch Hamerle, 1985c). Das Konzept der Survivorfunktion kann ohne Schwierigkeiten übertragen werden. Man erhält

$$S^2(t|x_2, H_1) = P(T_2 \geq t|x_2, H_1) = \exp(-\int_{t_1}^t \lambda^2(u|x_2, H_1)\, du), \qquad t \geq t_1. \tag{3.5.6}$$

Die Dauer der zweiten Episode ist

$$V_2 = T_2 - T_1. \tag{3.5.7}$$

Die Survivorfunktion kann auch in Abhängigkeit von der Verweildauer V_2 ausgedrückt werden.
3. Auf diese Weise wird fortgefahren, und man erhält sukzessive den zeitlichen Verlauf der aufeinanderfolgenden Episoden für das Individuum. Die Zeitpunkte, zu denen Ereignisse stattfinden, sind $t_1 < t_2 < t_3 < \ldots$ In jeder Episode wird möglicherweise ein neuer Vektor von Kovariablen erhoben. Für die k-te Episode wird dieser Vektor mit x_k bezeichnet. Die Vorgeschichte ist

$$H_{k-1} = \{t_1, x_1, \ldots, t_{k-1}, x_{k-1}\}.$$

Gelegentlich wird der für die k-te Episode relevante Teil der Vorgeschichte, also zum Beispiel die Dauer der k-1 vorangegangenen Episoden oder bestimmte früher erhobene Kovariablen, in den aktuellen Kovariablenvektor $\mathbf{x}_k$ aufgenommen. In diesem Fall kann die Abhängigkeit der Hazardrate oder Survivorfunktion von H_{k-1} weggelassen werden.
Die Hazardrate der k-ten Episode (k = 1,2, ...) ist

$$\lambda^k(t|\mathbf{x}_k, H_{k-1}) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} P(t \leq T_k < t + \Delta t \mid T_k \geq t, \mathbf{x}_k, H_{k-1}), \qquad t \geq t_{k-1}. \tag{3.5.8}$$

Für $t < t_{k-1}$ gilt $\lambda^k(t|\mathbf{x}_k, H_{k-1}) = 0$.

Für die Survivorfunktion erhält man

$$S^k(t|\mathbf{x}_k, H_{k-1}) = P(T_k > t|\mathbf{x}_k, H_{k-1}), \qquad t \geq t_{k-1}, \tag{3.5.9}$$

und in Abhängigkeit von der Hazardrate

$$S^k(t|\mathbf{x}_k, H_{k-1}) = \exp(-\int_{t_{k-1}}^{t} \lambda^k(u|\mathbf{x}_k, H_{k-1})\, du). \tag{3.5.10}$$

Schließlich resultiert für die Dichte von T_k, gegeben $\mathbf{x}_k$ und die Vorgeschichte H_{k-1}

$$f^k(t|\mathbf{x}_k, H_{k-1}) = \lambda^k(t|\mathbf{x}_k, H_{k-1}) \exp(-\int_{t_{k-1}}^{t} \lambda^k(u|\mathbf{x}_k, H_{k-1})\, du), \qquad t \geq t_{k-1}. \tag{3.5.11}$$

Jede der Funktionen (3.5.8), (3.5.10) und (3.5.11) kann zur Analyse von Zeitverläufen bei mehreren aufeinanderfolgenden Episoden verwendet werden. Kennt man eine dieser Funktionen, so lassen sich die beiden anderen Funktionen daraus ermitteln. Für Regressionsansätze ist es am zweckmäßigsten, die Übergangs- beziehungsweise Hazardrate zu modellieren.
Im Prinzip können alle in Abschnitt 3.3 vorgestellten Regressionsansätze auf den Mehr-Episoden-Fall übertragen werden. Geht man beispielsweise von einem Weibull-Regressionsmodell aus, so ist die Hazardrate der k-ten Episode

$$\lambda^k(t|\mathbf{x}_k) = \delta_k \lambda_{0k} (\lambda_{0k} t)^{\delta_{k-1}} \exp(\mathbf{x}_k' \beta_k), \tag{3.5.12}$$

wobei der relevante Teil von H_{k-1} bereits in $\mathbf{x}_k$ aufgenommen wurde. Für das allgemeine Proportional-Hazards-Modell ergibt sich

$$\lambda^k(t|\mathbf{x}_k) = \lambda_{0k}(t) \exp(\mathbf{x}_k' \beta_k) \tag{3.5.13}$$

mit der nicht spezifizierten Baseline-Hazardrate $\lambda_{0k}(t)$. Es besteht auch die Möglichkeit, für die aufeinanderfolgenden Episoden unterschiedliche Modellklassen zu wählen, also zum Beispiel für die erste Episode ein Weibull-Regressionsmodell, die zweite Episode ein PH-Modell usw. Allerdings ist die Wahl stets im Kontext des Anwendungszusammenhangs zu begründen.
Für die Modellierung der Abhängigkeit der Hazardrate der k-ten Episode von der Dauer früherer Episoden sind eine Reihe von Varianten denkbar. Eine Möglichkeit besteht darin, etwa bei der Untersuchung der Dauer der Arbeitslo-

sigkeit, in die Hazardrate der k-ten Episode von der Vorgeschichte nur die durchschnittliche Arbeitslosigkeitsdauer der vorangegangenen Perioden aufzunehmen, zum Beispiel in der Form

$$\lambda^k(t|x_k, H_{k-1}) = \lambda^k(t|x_k, \frac{1}{k-1} \sum_{j=1}^{k-1} (t_j - t_{j-1})).$$

Eine derartige Spezifikation ohne Berücksichtigung von Kovariablen wählten Braun/Hoem (1978).
Heckman/Borjas (1980) untersuchten die Dauer der Arbeitslosigkeit mit einer Hazardrate, die außer von einem zeitunabhängigen Kovariablenvektor nur von der Anzahl der früheren Arbeitslosigkeitsepisoden einer Person abhängt.

Regressionsmodelle für den Mehr-Episoden- und Mehr-Zustands-Fall

Die zufälligen Übergangszeiten werden wieder repräsentiert durch nicht-negative Zufallsvariablen $0 = T_0 < T_1 < T_2 < \ldots$, und die Verweildauern, das heißt die Länge der Episoden, sind

$$V_k = T_k - T_{k-1} \quad , k = 1,2, \ldots .$$

Zusätzlich wird nun eine Zustandsvariable $\{Y_k : k = 0,1,2, \ldots\}$ festgelegt, eine Folge von Zufallsvariablen mit endlichem Zustandsraum.
Wir betrachten die k-te Episode, $k = 1,2, \ldots$. Die (k-1)-te Episode endete im Zustand y_{k-1}. Die Menge der vom Zustand y_{k-1} aus erreichbaren Zustände sei $M(y_{k-1})$. Diese Mengen können im allgemeinen von Zustand zu Zustand und von Episode zu Episode variieren. Wir nehmen hier der Einfachheit halber stets den gesamten Zustandsraum mit allen m Zuständen. Ist ein bestimmter Übergang ausgeschlossen, so wird dies dadurch zum Ausdruck gebracht, daß die entsprechende übergangsspezifische Hazardrate beziehungsweise die entsprechende Übergangswahrscheinlichkeit gleich Null gesetzt wird.
Sei $Y_k = j$, dann wird die übergangsspezifische Hazardrate der k-ten Episode folgendermaßen definiert:

$$\lambda_j^k(t|x_k, H_{k-1}) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} P(t \leq T_k < t + \Delta t, Y_k = j | T_k \geq t, H_{k-1}, x_k), \tag{3.5.14}$$

wobei in H_{k-1} die Vorgeschichte des Prozeßverlaufs bis zum Zeitpunkt t_{k-1} zusammengefaßt wird, also

$$H_{k-1} = \{t_0, y_0, t_1, y_1, x_1, \ldots, t_{k-1}, y_{k-1}, x_{k-1}\}.$$

Man beachte, daß $\lambda_j^k(t|x_k, H_{k-1})$ wieder identisch gleich Null ist für $t < t_{k-1}$.

Anmerkung:

Gelegentlich werden in der Literatur (vgl. z.B. Tuma/Hannan, 1984) die Hazardrate sowie die zu den Kovariablen gehörenden Parameter explizit in Abhängigkeit vom Zustand y_{k-1} formuliert, also $\lambda_{ij}^k(t|x_k, H_{k-1})$ beziehungs-

weise β_{ij}^k. Dabei muß allerdings eine beträchtliche Zunahme der Zahl der unbekannten Parameter in Kauf genommen werden. Im Ansatz (3.5.14) ist der Ausgangszustand y_{k-1} in der Vorgeschichte H_{k-1} enthalten, und sein Einfluß kann durch einen Parameter repräsentiert werden. Der Übergang auf Hazardraten $\lambda_{ij}^k(t|x_k, H_{k-1})$ beziehungsweise eine Parametrisierung mit Parametern β_{ij}^k ist ohne Schwierigkeiten möglich, falls es notwendig erscheint.
Die Gesamthazardrate $\lambda^k(t|x_k, H_{k-1})$, in der k-ten Episode den Zustand y_{k-1} zu verlassen, ist

$$\lambda^k(t|x_k, H_{k-1}) = \sum_{j=1}^{m} \lambda_j^k(t|x_k, H_{k-1}). \tag{3.5.15}$$

Die Survivorfunktion kann ebenfalls auf den Mehr-Episoden- und Mehr-Zustands-Fall erweitert werden. Sie gibt die Wahrscheinlichkeit an, ausgehend vom Zustand y_{k-1} den Zeitpunkt t zu „überleben", das heißt, daß bis zu diesem Zeitpunkt der k-te Übergang noch nicht stattgefunden hat,

$$S^k(t|x_k, H_{k-1}) = P(T_k > t|x_k, H_{k-1}) \quad \text{für } t \geq t_{k-1}. \tag{3.5.16}$$

Definiert man die Verweildauer

$$V_k = T_k - T_{k-1},$$

erhält man die Survivorfunktion in Abhängigkeit von der Verweildauer

$$S^k(v|x_k, H_{k-1}) = P(V_k > v|x_k, H_{k-1}), \quad v \geq 0. \tag{3.5.17}$$

Für den Zusammenhang zwischen Survivorfunktion und Hazardrate ergibt sich

$$S^k(t|x_k, H_{k-1}) = \exp[-\int_{t_{k-1}}^{t} \lambda^k(u|x_k, H_{k-1})\, du], \quad t \geq t_{k-1} \tag{3.5.18}$$

und mit (3.5.15)

$$\begin{aligned} S^k(t|x_k, H_{k-1}) &= \exp[-\int_{t_{k-1}}^{t} \sum_{j=1}^{m} \lambda_j^k(u|x_k, H_{k-1})\, du] = \\ &= \prod_{j=1}^{m} \exp[-\int_{t_{k-1}}^{t} \lambda_j^k(u|x_k, H_{k-1})\, du]. \end{aligned} \tag{3.5.19}$$

Schließlich erhält man für die „Dichtefunktion"

$$f_j^k(t|x_k, H_{k-1}) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} P(t \leq T_k < t + \Delta t,\ Y_k = j|x_k, H_{k-1})$$

in Abhängigkeit von Hazardrate und Survivorfunktion

$$f_j^k(t|x_k, H_{k-1}) = \lambda_j^k(t|x_k, H_{k-1})\, S^k(t|x_k, H_{k-1}), \quad t \geq t_{k-1}. \tag{3.5.20}$$

Zur Modellierung der Hazardrate (3.5.14) für Regressionsansätze können im Prinzip alle in Abschnitt 3.3 vorgestellten Spezifikationen verwendet werden. In den Kapiteln 5 und 6 wird die Anwendung einiger Modelle für einen speziellen Mehr-Episoden- und Mehr-Zustands-Fall anhand der Lebensverlaufsstudie mit vielen Beispielen ausführlich behandelt.

Ein wichtiger Spezialfall von Mehr-Episoden-Modellen, der in den Kapiteln 5 und 6 stets zugrundegelegt wird, ergibt sich dadurch, daß die episodenspezifischen Hazardraten $\lambda^k(t|x_k, H_{k-1})$ nur von der Verweildauer $v = t - t_{k-1}$ abhängen, also

$$\lambda^k(t|x_k, H_{k-1}) = \lambda^{*k}(t - t_{k-1}|x_k, H_{k-1}). \tag{3.5.21}$$

Legt man das Cox-Modell zugrunde und nimmt man darüber hinaus an, daß sich die Grundhazardrate und die Einflußgewichte der Kovariablen von Episode zu Episode nicht verändern, erhält man

$$\lambda^k(t|x_k) = \lambda_0(v) \exp(x_k'\beta) \quad , v = t - t_{k-1}. \tag{3.5.22}$$

Für das Modell (3.5.22) vereinfacht sich die Parameterschätzung beträchtlich. Eine Beschreibung findet man am Ende von Abschnitt 3.6.6.

3.6 Maximum-Likelihood-Schätzung

Nach der Konstruktion eines statistischen Modells für die vorliegende Ereignisgeschichte sind die unbekannten Parameter aus den erhobenen Daten zu schätzen. In diesem Abschnitt wird ausschließlich die Maximum-Likelihood-Methode behandelt, die den Erfordernissen der Ereignisanalyse unter Einbeziehung rechtszensierter Daten am besten gerecht wird. Zunächst wird ein kurzer Abriß der allgemeinen Theorie der Maximum-Likelihood-Schätzung (ML-Schätzung) gegeben. Dann wird auf die Zensierungsproblematik näher eingegangen und die notwendigen Modifikationen werden erläutert. In den folgenden Teilabschnitten werden die ML-Schätzprozedur für einige wichtige parametrische Regressionsmodelle (vgl. Abschnitt 3.3) sowie die speziell für das Proportional-Hazards-Modell von Cox entwickelte Partial-Likelihood-Schätzprozedur erörtert. Schließlich wird noch die ML-Schätzung für Competing-Risks-Modelle und für allgemeine Mehr-Episoden- und Mehr-Zustands-Regressionsmodelle behandelt.

3.6.1 Allgemeine Theorie der Maximum-Likelihood-Schätzung

Im folgenden werden die grundlegenden Prinzipien der Maximum-Likelihood-Schätzung nur skizziert. Für ausführliche Darstellungen, die allerdings gelegentlich maßtheoretische Kenntnisse voraussetzen, vergleiche man beispielsweise Cox/Hinkley (1974), Rao (1973) oder Witting/Nölle (1970).
Den Ausgangspunkt der Maximum-Likelihood-Schätzung bilden die beobachteten Daten. Im einfachsten Fall handelt es sich um unabhängige Realisierungen $x_1, \ldots, x_n$ eines zufälligen Merkmals X. Die Dichtefunktion (bei stetigen Merkmalen) beziehungsweise die Wahrscheinlichkeitsfunktion (bei diskreten Merkmalen) sei bis auf einen unbekannten Parametervektor θ vollständig spezifiziert.

Wir schreiben dafür $f(x; \theta)$. Der Parametervektor $\theta = (\theta_1, \ldots, \theta_p)$ ist zu schätzen. Die *Likelihood-Funktion* der Stichprobe ist

$$L(\theta; x_1, \ldots, x_n) = \prod_{i=1}^{n} f_i(x_i|\theta). \tag{3.6.1}$$

Für das wahre, aber unbekannte θ ist die rechte Seite von (3.6.1) die gemeinsame Dichte beziehungsweise Wahrscheinlichkeit von $x_1, \ldots, x_n$. Die Likelihood-Funktion $L(\theta; x_1, \ldots, x_n)$ wird dagegen für die feste Stichprobe $x_1, \ldots, x_n$ als Funktion von θ, das in einem zulässigen Parameterbereich Θ variieren darf, aufgefaßt.

Das *Maximum-Likelihood-Prinzip* besteht darin, bei Vorliegen der Daten $x_1, \ldots, x_n$ einen Parameterschätzwert $\hat{\theta} = \hat{\theta}(x_1, \ldots, x_n) \in \Theta$ so zu wählen, daß für diesen Schätzwert der Beobachtung eine maximale Wahrscheinlichkeitsdichte (im diskreten Fall maximale Wahrscheinlichkeit) zukommt.

$\hat{\theta} = \hat{\theta}(x_1, \ldots, x_n)$ heißt also *Maximum-Likelihood-(ML-)Schätzwert für* θ, wenn

$$L(\hat{\theta}; x_1, \ldots, x_n) \geq L(\theta; x_1, \ldots, x_n) \quad \text{für alle } \theta \in \Theta \tag{3.6.2}$$

gilt.

Gewöhnlich wird aus rechentechnischen Gründen nicht die Likelihood-Funktion selbst maximiert, sondern die logarithmierte Likelihood-Funktion, da der Logarithmus als streng monotone Transformation die Stelle eines Maximums unverändert läßt.

Die Log-Likelihood-Funktion ist

$$l(\theta; x_1, \ldots, x_n) = \ln L(\theta; x_1, \ldots, x_n) = \sum_{i=1}^{n} \ln f_i(x_i; \theta). \tag{3.6.3}$$

Für die konkrete Berechnung der ML-Schätzwerte sind die Ableitungen

$$\frac{\partial}{\partial \theta_j} l(\theta; x_1, \ldots, x_n)$$

von Bedeutung. Der Spaltenvektor dieser Ableitungen

$$s(\theta; x_1, \ldots, x_n) = \left(\frac{\partial}{\partial \theta_1} l(\theta; x_1, \ldots, x_n), \ldots, \frac{\partial}{\partial \theta_p} l(\theta; x_1, \ldots, x_n) \right)'$$

wird als *Score-Funktion* bezeichnet. p ist die Dimension des unbekannten θ.

Zur Maximierung der Log-Likelihood-Funktion sind die Nullstellen der ML-Gleichungen

$$\frac{\partial}{\partial \theta_j} l(\theta; x_1, \ldots, x_n) = 0 \qquad j = 1, \ldots, p \tag{3.6.4}$$

zu ermitteln. Nur in wenigen Fällen, etwa bei der Normal- oder der Exponentialverteilung, ist ein explizites Auflösen der ML-Gleichungen möglich. Meistens müssen die ML-Schätzwerte durch numerische, in der Regel iterative Verfahren berechnet werden.

Newton-Raphson-Technik

Ein wichtiges iteratives Verfahren zur Lösung von Maximierungsproblemen beziehungsweise zur Lösung der ML-Gleichungen ist das Newton-Verfahren. Obwohl heute meist sogenannte modifizierte oder Quasi-Newton-Verfahren angewendet werden, ist die Technik des Newton-Verfahrens immer noch von grundlegender Bedeutung. Im folgenden wird lediglich der Grundgedanke des Verfahrens kurz skizziert, für ausführliche Einführungen siehe zum Beispiel Luenberger (1973) oder Stoer (1976, Kap. 5). Man vergleiche auch Fahrmeir/ Hamerle (1984, Kap. 3.2.4).

Die Newton-Raphson-Technik zur Maximierung der logarithmierten Likelihood-Funktion $\ln L(\theta; x_1, \ldots, x_n)$ basiert auf der Taylor-Entwicklung der Scorefunktion $s(\theta) = \partial \ln L / \partial \theta$. Das Verfahren ist iterativ aufgebaut. Ausgehend von einem Startwert θ_0 konstruiert man sukzessive Näherungen $\theta_1, \theta_2, \ldots, \theta_k, \ldots$. Entwickelt man $s(\theta)$ um θ_k, so resultiert in Vektor- beziehungsweise Matrizenschreibweise

$$s(\theta) \approx s(\theta_k) + \frac{\partial^2 \ln L}{\partial \theta \, \partial \theta'} (\theta_k)(\theta - \theta_k). \tag{3.6.5}$$

Wäre diese Näherung korrekt, würde man eine Nullstelle von $s(\theta)$ für

$$\theta_{k+1} = \theta_k - \left(\frac{\partial^2 \ln L}{\partial \theta \, \partial \theta'} (\theta_k) \right)^{-1} s(\theta_k) \tag{3.6.6}$$

erhalten. Da die Näherung (3.6.5) im allgemeinen aber nicht exakt stimmt, wendet man (3.6.6) iterativ an und fährt mit der Iteration solange fort, bis θ_{k+1} mit ausreichender Genauigkeit mit θ_k übereinstimmt.

Das Newton-Raphson-Verfahren (3.6.6) soll für den einfachen Fall eines eindimensionalen Parameters θ demonstriert werden. Für (3.6.6) ergibt sich dann

$$\theta_{k+1} = \theta_k - \frac{s(\theta_k)}{\frac{d^2}{d\theta^2} \ln L(\theta_k)} = \theta_k - \frac{s(\theta_k)}{s'(\theta_k)}. \tag{3.6.7}$$

In der Abbildung 3.11 ist die Scorefunktion $s(\theta)$, deren Nullstelle $\hat{\theta}$ gesucht wird, eingezeichnet.

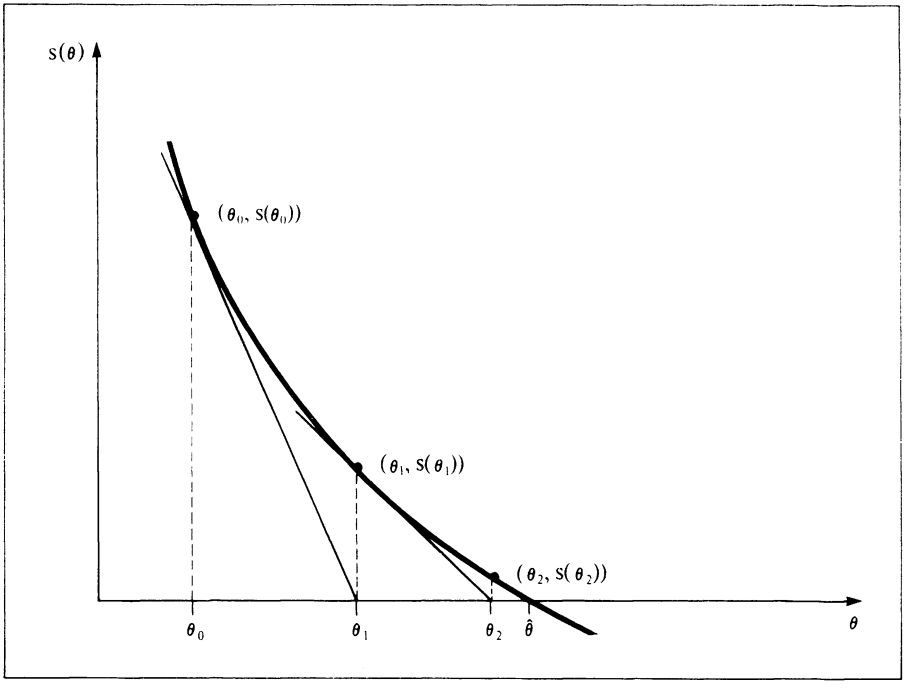
Der Startwert sei θ_0. Legt man an den Punkt $(\theta_0, s(\theta_0))$ die Tangente, so ist die Gleichung dieser Tangente gegeben durch

$$y = s'(\theta_0)(\theta - \theta_0) + s(\theta_0).$$

Berechnet man den Schnittpunkt der Tangente mit der x-Achse und bezeichnet ihn mit θ_1, erhält man

$$\theta_1 = \theta_0 - \frac{s(\theta_0)}{s'(\theta_0)}.$$

Abbildung 3.11: Graphische Veranschaulichung des Newton-Raphson-Verfahrens



Dies ist gerade die Iterationsvorschrift (3.6.7) für k = 0. Dieser Prozeß wird nun laufend wiederholt bis $\theta_{k+1} \approx \theta_k$ ist.
Da ein Maximum der Log-Likelihood-Funktion gesucht wird, muß die Matrix der zweiten Ableitungen von $\ln L(\theta; x_1, \ldots, x_n)$ negativ definit sein. Ist

$$\mathbf{H} = \frac{\partial^2 l}{\partial \theta \, \partial \theta'} = \left(\frac{\partial^2 l}{\partial \theta_i \, \partial \theta_j} \right)$$

die (p × p)-Matrix der zweiten Ableitungen der Log-Likelihood-Funktion, so heißt

$$\mathbf{I}(\theta) = \mathrm{E}\left(- \frac{\partial^2 l(\theta; x_1, \ldots, x_n)}{\partial \theta \, \partial \theta'} \right)$$

(Fishersche) Informationsmatrix der Stichprobe $x_1, \ldots, x_n$.

Asymptotische Eigenschaften der ML-Schätzer

Zur Konstruktion von Konfidenzintervallen für $\hat{\theta}_j$ sowie zur Prüfung von Hypothesen über bestimmte θ_j oder Teilvektoren von θ benötigt man die Wahrscheinlichkeitsverteilung der ML-Schätzungen $\hat{\theta}$. Außer in Ausnahmefällen kann die

Verteilung von $\hat{\theta}$ für endliche Stichprobenumfänge n nicht angegeben werden. Für gewisse Standardsituationen lassen sich jedoch asymptotische Eigenschaften von $\hat{\theta}$, das heißt für $n \to \infty$, ableiten. Eine derartige Standardsituation ist zum Beispiel der Fall unabhängiger und identisch verteilter Beobachtungen $x_1, \ldots, x_n$. Die Maximum-Likelihood-Schätzungen sind dann konsistent und asymptotisch normalverteilt mit Erwartungwert θ und Varianz-Kovarianzmatrix $(\mathbf{I}(\theta))^{-1}$, der Inversen der Informationsmatrix, wenn die Log-Likelihood-Funktion gewisse Regularitätseigenschaften besitzt.
Dies bedeutet für die praktische Anwendung, daß für große Stichprobenumfänge $\hat{\theta}$ als approximativ normalverteilt angenommen werden kann mit Erwartungswert θ und Varianz-Kovarianzmatrix $[\mathbf{I}(\theta)]^{-1}/n$. Allerdings enthält die Informationsmatrix den wahren Parameter θ. Es ist daher notwendig, diese Informationsmatrix zu schätzen. Eine Möglichkeit besteht darin, die Matrix der zweiten Ableitungen der Log-Likelihood-Funktion an der Stelle $\hat{\theta}$ auszuwerten und

$$-\hat{\mathbf{H}} = \left(-\frac{\partial^2 l}{\partial\theta\,\partial\theta'}(\hat{\theta})\right) \tag{3.6.8}$$

als Schätzung der Informationsmatrix zu verwenden. Bei Verwendung des Newton-Verfahrens bereitet dies keine Probleme, da diese Matrix ohnehin beim letzten Iterationsschritt berechnet wird.
Die Hauptdiagonalelemente der Matrix $(-\hat{\mathbf{H}})^{-1}/n$ enthalten die geschätzten (asymptotischen) Varianzen der ML-Schätzungen $\hat{\theta}_j$, $j = 1, \ldots, p$, die insbesondere für die Konstruktion von Konfidenzintervallen oder für Tests von Bedeutung sind.
Es ist zu beachten, daß die in diesem Buch behandelten Modelle im allgemeinen *nicht* zu den oben erwähnten Standardsituationen gehören, insbesondere wenn zensierte Daten vorliegen. Die asymptotischen Eigenschaften der Maximum-Likelihood-Schätzungen sind dann erst zu beweisen. Dies ist in einigen Fällen noch nicht vollständig gelöst. Wir werden in Abschnitt 3.6.4 näher darauf eingehen.
Gelten die asymptotischen Aussagen oder unterstellt man ihre Gültigkeit, so können einzelne Parameter oder bestimmte Modellteile auf Signifikanz geprüft werden. Derartige Hypothesen lassen sich stets zusammengefaßt darstellen als

$$H_0 : C\,\theta = 0 \qquad , m = rg\,(C) \tag{3.6.9}$$

mit einer geeigneten Matrix C. Der Test einer solchen linearen Hypothese kann zum Beispiel mit Hilfe des Likelihood-Quotienten-Tests erfolgen. Die Teststatistik

$$2(\ln L(\hat{\theta}) - \ln L(\tilde{\theta})), \tag{3.6.10}$$

wobei $\tilde{\theta}$ die ML-Schätzungen unter der Nebenbedingung $C\theta = 0$ und $\hat{\theta}$ die ML-Schätzungen ohne Restriktionen sind, besitzt unter H_0 asymptotisch eine (zentrale) χ^2-Verteilung mit m Freiheitsgraden.

Zur Prüfung der linearen Hypothese $C\theta = 0$ können auch die Score-Statistik oder die Wald-Statistik herangezogen werden. Die asymptotische Verteilung ist dieselbe wie die der Likelihood-Quotienten-Teststatistik. Man vergleiche dazu etwa Rao (1973, S. 351 ff.).

3.6.2 Zensierung

Bei der Anwendung der Maximum-Likelihood-Schätzung – wie auch bei anderen Schätzverfahren – muß für jedes Stichprobenelement eine Realisation des in Frage stehenden zufälligen Merkmals vorliegen. Da in der Ereignisanalyse das Ende des gesamten Beobachtungszeitraums in der Regel vorgegeben ist, ist die Dauer der Episode unter Umständen nicht abgeschlossen. Man spricht in einem solchen Fall von rechtszensierten Daten. Die Stichprobenrealisation t_i eines Individuums besagt dann lediglich, daß die Dauer der Episode *mindestens* t_i Zeiteinheiten beträgt. Die exakte Zeitdauer läßt sich nicht angeben. In der Regel liegt eine Stichprobe vor, bei der einige Werte t_i exakte Zeitdauern sind, während es sich beim Rest um zensierte Daten handelt. Man bringt dies mit Hilfe eines *Zensierungsindikators* δ_i zum Ausdruck mit

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

Die Möglichkeit, die zensierten Daten einfach zu ignorieren und den Stichprobenumfang zu reduzieren, ist nicht zu empfehlen, da dies verzerrte Resultate zur Folge haben kann.
Die Maximum-Likelihood-Methode bietet die Möglichkeit, rechtszensierte Daten explizit im Schätzvorgang zu berücksichtigen. Zu diesem Zweck ist der Zensierungsmechanismus, der den Daten zugrundeliegt, genau zu analysieren und in ein statistisches Modell zu fassen. Für das Zustandekommen von zensierten Daten sind je nach Anwendungsbereich mehrere statistische Konzepte denkbar. Hier sollen zwei Modelle behandelt werden, die für die Anwendungen in diesem Buch von besonderem Interesse sind.

Zensierungs-Modell I

In Modell I ist für jedes Individuum i, $i = 1, \ldots, n$, ein fester Beobachtungszeitraum L_i vorgegeben. Bei der Lebensverlaufsstudie handelt es sich dabei um die Zeit von der Geburt bis zum Zeitpunkt der Befragung. Die Zeiten L_i können von Individuum zu Individuum variieren. Die Zeitdauern T_i, deren Dichtefunktion mit $f_i(t)$ und deren Survivorfunktion mit $S_i(t)$ bezeichnet werden, können nur exakt beobachtet werden, falls $T_i \leq L_i$ gilt. Der Zensierungsindikator ist

$$\delta_i = \begin{cases} 1 & \text{falls } T_i \leq L_i \\ 0 & \text{falls } T_i > L_i \end{cases}$$

und die in der Stichprobe vorliegenden Zeiten t_i sind

$$t_i = \min(T_i, L_i) \quad , i = 1, \ldots, n.$$

Die gemeinsame Wahrscheinlichkeitsverteilung von (T_i, δ_i) berechnet sich wie folgt:
Für $\delta_i = 0$ gilt stets $T_i > L_i$ und damit $t_i = L_i$ und die Wahrscheinlichkeit dafür ist $S_i(L_i)$. Die Dichte von $(T_i, \delta_i = 1)$ ist

$$\begin{aligned} f_i(t_i, \delta_i = 1) &= f_i(t_i \mid \delta_i = 1) \cdot P(\delta_i = 1) \\ &= f_i(t_i \mid T_i \leq L_i) \cdot P(T_i \leq L_i) \\ &= \frac{f_i(t_i)}{1 - S_i(L_i)} (1 - S_i(L_i)) \\ &= f_i(t_i) \end{aligned}$$

und insgesamt ergibt sich für den Beitrag (t_i, δ_i) des Individuums i zur Likelihood-Funktion

$$L_i = f_i(t_i)^{\delta_i} S_i(L_i)^{1-\delta_i}. \tag{3.6.11}$$

Die Gesamt-Likelihood-Funktion ist dann

$$L = \prod_{i=1}^{n} f_i(t_i)^{\delta_i} S_i(L_i)^{1-\delta_i}. \tag{3.6.12}$$

Zensierungs-Modell II (random censoring)

Eine andere Variante besteht darin, die Zensierungszeiten ebenfalls als Realisierungen von Zufallsvariablen anzusehen und nicht wie bei Modell I als fest vorgegeben. Die Zeitdauern T_i und die Zensierungszeiten C_i werden als unabhängige Zufallsvariablen vorausgesetzt mit den Dichten $f_i(t)$ beziehungsweise $g_i(t)$ und den Survivorfunktionen $S_i(t)$ beziehungsweise $G_i(t)$.
Für eine unzensierte Beobachtung $(t_i, \delta_i = 1)$, das heißt $T_i = t_i$ und $T_i \leq C_i$, ist der Wert der gemeinsamen Verteilung wegen der Unabhängigkeit von T_i und C_i (ohne Berücksichtigung der Kovariablen)

$$f_i(t_i)\ G_i(t_i),$$

für eine zensierte Beobachtung $(t_i, \delta_i = 0)$, das heißt $T_i > C_i$, $C_i = t_i$ ist der Wert der gemeinsamen Verteilung

$$g_i(t_i)\ S_i(t_i).$$

Dies läßt sich zusammenfassen, und für den Beitrag (t_i, δ_i) von Individuum i zur Likelihood-Funktion ergibt sich

$$L_i = [f_i(t_i)\ G_i(t_i)]^{\delta_i} [g_i(t_i)\ S_i(t_i)]^{1-\delta_i} = f_i(t_i)^{\delta_i} S_i(t_i)^{1-\delta_i} g_i(t_i)^{1-\delta_i} G_i(t_i)^{\delta_i}. \tag{3.6.13}$$

Unterstellt man, daß die Verteilung der Zensierungszeiten nicht von den für f_i und S_i relevanten Parametern abhängt, insbesondere nicht von eventuellen

Regressionskoeffizienten, so können die beiden letzten Faktoren in (3.6.13) zu einem (in Abhängigkeit von den relevanten Parametern) konstanten Term c zusammengefaßt werden, der die Maximierung der Likelihood-Funktion beziehungsweise der Log-Likelihood-Funktion nicht beeinflußt. Die Gesamt-Likelihood-Funktion ist dann

$$L = c \cdot \prod_{i=1}^{n} f_i(t_i)^{\delta_i} S_i(t_i)^{1-\delta_i}. \tag{3.6.14}$$

Die Likelihood-Funktion (3.6.14) ist wiederum von der Form (3.6.12). Diese Form der Likelihood-Funktion ist typisch für die ML-Schätzung bei rechtszensierten Daten. Sie gilt auch bei noch allgemeineren Zensierungsmechanismen. Man vergleiche dazu zum Beispiel Kalbfleisch/Prentice (1980, Kap. 5) oder Lagakos (1979). Im wesentlichen muß sichergestellt werden, daß die Verteilung, die den Zensierungsmechanismus steuert, nicht von den zu den Kovariablen gehörenden Regressionskoeffizienten sowie weiteren die Episodendauerverteilung determinierenden Parametern abhängt (nicht informativer Zensierungsmechanismus).

3.6.3 Maximum-Likelihood-Schätzung für parametrische Regressionsmodelle

In diesem Abschnitt wird die Vorgehensweise bei der ML-Schätzung für Regressionsmodelle skizziert, bei denen die Hazardrate und damit auch die Verteilung der Dauer der Episode vollständig spezifiziert ist. Wir werden dabei nicht auf jedes Modell, das in Abschnitt 3.3.2 vorgestellt wurde, im Detail eingehen, da die Berechnungen stets völlig analog erfolgen und im konkreten Anwendungsfall immer mit Hilfe des Computers durchgeführt werden.
Für jedes Individuum beziehungsweise Objekt i, i = 1, ..., n, liegen folgende Daten vor:

t_i : Dauer der Episode (eventuell zensiert)
δ_i : Zensierungsindikator
$\mathbf{x}_i$: Vektor von Kovariablen.

Nach (3.6.12) beziehungsweise (3.6.14) ist die Likelihood-Funktion (eventuell bis auf einen Proportionalitätsfaktor) gegeben durch

$$L = \prod_{i=1}^{n} f_i(t|\mathbf{x}_i)^{\delta_i} S_i(t_i|\mathbf{x}_i)^{1-\delta_i}.$$

Berücksichtigt man den Zusammenhang zwischen Hazardrate und Survivorfunktion (vgl. (3.2.10)), so ergibt sich

$$\begin{aligned} L &= \prod_{i=1}^{n} \lambda_i(t_i|\mathbf{x}_i)^{\delta_i} S_i(t_i|\mathbf{x}_i) \\ &= \prod_{i=1}^{n} \lambda_i(t_i|\mathbf{x}_i)^{\delta_i} \exp\left(-\int_0^t \lambda_i(u|\mathbf{x}_i)\, du\right) \end{aligned} \tag{3.6.15}$$

In (3.6.15) steht die Likelihood-Funktion nur in Abhängigkeit von der Hazardrate. Für die ML-Schätzung von besonderem Interesse, da meist rechentechnisch einfacher, ist die Log-Likelihood-Funktion. Wird die Hazardrate von einem unbekannten Parametervektor θ, der insbesondere auch die Regressionskoeffizienten β_j enthält, vollständig determiniert, erhält man für die Log-Likelihood-Funktion

$$l(\theta; x_1, \ldots, x_n) = \ln L(\theta; x_1, \ldots, x_n) = \sum_{i=1}^{n} (\delta_i \ln \lambda_i(t_i|x_i, \theta) - \int_0^t \lambda_i (u|x_i, \theta)\, du). \tag{3.6.16}$$

Die Log-Likelihood-Funktion (3.6.16) ist in Abhängigikeit von θ zu maximieren. Dies geschieht in der Regel mit Hilfe eines iterativen Verfahrens, etwa des Newton-Verfahrens oder eines modifizierten Newton-Verfahrens. Wie in Abschnitt 3.6.1 ausgeführt, wird die Inverse der an der Stelle $\hat{\theta}$ ausgewerteten Informationsmatrix als Schätzung der asymptotischen Kovarianzmatrix $\mathbf{Cov}(\hat{\theta})$ herangezogen.

Die Vorgehensweise der ML-Schätzung bei einem vollständig parametrisierten Regressionsmodell wird anhand des Exponential-Regressionsmodells verdeutlicht.

Die Hazardrate hängt gemäß (3.6.2) beziehungsweise (3.6.3) von einem Vektor $x = (x_1, \ldots, x_p)'$ von Kovariablen ab, wobei die erste Komponente gleich 1 ist, das heißt, die Hazardrate ist gegeben durch

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

Damit erhält man die Likelihood-Funktion ($\theta = \beta$)

$$L(\beta; x_1, \ldots, x_n) = \prod_{i=1}^{n} \exp(\delta_i\, x_i'\beta) \exp(-\exp(x_i'\beta)\, t_i) \tag{3.6.17}$$

und für die Log-Likelihood-Funktion

$$l(\beta; x_1, \ldots, x_n) = \sum_{i=1}^{n} [\delta_i\, x_i'\beta - \exp(x_i'\beta)\, t_i]. \tag{3.6.18}$$

Die Scorefunktion besitzt die Komponenten

$$s_j(\beta; x_1, \ldots, x_n) = \frac{\partial l(\beta; x_1, \ldots, x_n)}{\partial \beta_j} = \sum_{i=1}^{n} x_{ij}\, (\delta_i - \exp(x_i'\beta)\, t_i), \qquad j = 1, \ldots, p,$$

und die ML-Gleichungen sind

$$\sum_{i=1}^{n} x_{ij}\, (\delta_i - \exp(x_i'\beta)\, t_i) = 0, \qquad j = 1, \ldots, p. \tag{3.6.19}$$

Die Matrix der zweiten Ableitungen der Log-Likelihood-Funktion besitzt in der j-ten Zeile und der k-ten Spalte das Element

$$h_{jk}(\beta) = \frac{\partial^2 l\, (\beta; x_1, \ldots, x_n)}{\partial \beta_j\, \partial \beta_k} = \sum_{i=1}^{n} x_{ij}\, x_{ik} \exp(x_i'\beta)\, t_i.$$

Die Lösung der ML-Gleichungen ergibt die ML-Schätzungen $\hat{\beta}$, falls die Matrix $\mathbf{H}(\hat{\beta}) = (h_{jk}(\hat{\beta}))$ an dieser Stelle negativ definit ist. Als Schätzung der asymptotischen Kovarianzmatrix dient $\mathbf{H}(\hat{\beta})^{-1}/n$.
Wählt man für die Hazardrate einen anderen Ansatz, etwa das Weibull-Modell oder die Gompertz-Makeham-Rate, so ist die Vorgehensweise völlig analog, lediglich die numerischen Berechnungen sind aufwendiger, da zu den Regressionskoeffizienten weitere unbekannte Parameter hinzukommen.

3.6.4 Das Proportional-Hazards-Modell von Cox: Partial-Likelihood

Im Proportional-Hazards-Regressionsmodell (PH-Modell), das erstmals von Cox (1972) vorgeschlagen wurde, ist die Hazardrate

$$\lambda(t|\mathbf{x}) = \lambda_0(t)\, \exp(\mathbf{x}'\beta)$$

mit einer beliebigen, nicht spezifizierten Baseline-Hazardrate $\lambda_0(t)$. Der zweite Faktor $\exp(\mathbf{x}'\beta)$ kann auch durch eine andere positive Funktion $g(\mathbf{x}; \beta)$ ersetzt werden.
Die Likelihood-Funktion für das PH-Modell (Cox-Modell) ist

$$L(\beta, \lambda_0(t), \mathbf{x}_1, \ldots \mathbf{x}_n) = \prod_{i=1}^{n} [\lambda_0(t) \exp(\mathbf{x}_i'\beta)]^{\delta_i} \exp[-\int_0^t \lambda_0(u) \exp(\mathbf{x}_i'\beta)\, du]. \quad (3.6.20)$$

(3.6.20) enthält die unbekannte Baseline-Hazardrate $\lambda_0(t)$, das heißt nicht nur die unbekannten Parameter β, sondern darüber hinaus noch die sogenannte „Nuisance-Funktion" $\lambda_0(t)$. Deshalb kann (3.6.20) zur Schätzung von β nicht herangezogen werden.
Cox (1972, 1975) schlug vor, die Likelihood (3.6.20) zu faktorisieren. Seien $t_{(1)} < \ldots < t_{(k)}$ die Zeitdauern der Individuen, die nicht zensiert sind ($k \leq n$), und sei $R(t)$ die *„Risikomenge"*, das heißt die Menge der Individuen, deren Episode unmittelbar vor dem Zeitpunkt t noch nicht beendet ist und die nicht zensiert sind. Aus (3.6.20) erhält man dann durch Erweiterung

$$L(\beta, \lambda_0(t); \mathbf{x}_1, \ldots, \mathbf{x}_n) = \prod_{i=1}^{k} \frac{\exp(\mathbf{x}_i'\beta)}{\sum_{l \in R(t_{(i)})} \exp(\mathbf{x}_l'\beta)} \sum_{l \in R(t_{(i)})} \lambda_0(t_{(i)}) \exp(\mathbf{x}_l'\beta) \prod_{i=1}^{n} S_0(t_i)^{\exp(\mathbf{x}_i'\beta)} \quad (3.6.21)$$

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

Den ersten Faktor

$$PL(\beta; \mathbf{x}_1, \ldots, \mathbf{x}_n) = \prod_{i=1}^{k} \frac{\exp(\mathbf{x}_i'\beta)}{\sum_{l \in R(t_{(i)})} \exp(\mathbf{x}_l'\beta)}, \quad (3.6.22)$$

der nur von β abhängt, bezeichnete Cox (1972, 1975) als „partial likelihood", und schlug vor, (3.6.22) wie eine gewöhnliche Likelihood-Funktion zu behandeln und in Abhängigkeit von β zu maximieren.
Da der zweite Faktor auf der rechten Seite von (3.6.21) ebenfalls den Parameter β enthält, geht beim Übergang auf die Partial-Likelihood-Funktion etwas Information verloren. Dies kann sich insbesondere bei kleinen Stichprobenumfängen auf die Güte der Schätzung auswirken. Asymptotisch wurden aber auch für die Partial-ML-Schätzungen im Cox-Modell die üblichen wünschenswerten Eigenschaften wie Konsistenz und asymptotische Normalität unter bestimmten Voraussetzungen für die Kovariablen nachgewiesen. Man vergleiche dazu Tsiatis (1981), Bailey (1983), Naes (1982), Prentice/Self (1983) und insbesondere Andersen/Gill (1982).
Gelegentlich ist eine andere Interpretation der Partial Likelihood von Nutzen.

Der Term

$$\frac{\lambda\,(t_{(i)}|x_i)}{\sum\limits_{l\in R(t_{(i)})} \lambda\,(t_{(i)}|x_l)} = \frac{\exp(x_i'\beta)}{\sum\limits_{l\in R(t_{(i)})} \exp(x_l'\beta)}$$

auf der rechten Seite von (3.6.22) kann interpretiert werden als (bedingte) Wahrscheinlichkeit, daß zum Zeitpunkt $t_{(i)}$ gerade für Individuum i ein Ereignis stattfindet, unter der Voraussetzung, daß für die Individuen der Risikomenge $R(t_{(i)})$ unmittelbar vor dem Zeitpunkt $t_{(i)}$ die in Frage stehende Episode noch nicht beendet ist und daß zum Zeitpunkt $t_{(i)}$ genau ein Ereignis stattfindet. Das Produkt über alle k Zeitpunkte, an denen Ereignisse stattfinden, ergibt die Partial Likelihood (3.6.22).
Die Anwendung der Partial Likelihood (3.6.22) setzt voraus, daß die Zeitdauern t_i ausreichend genau gemessen werden können, so daß keine gleichen Meßwerte (Verbundwerte; Ties) auftreten. Bei praktischen Anwendungen treten jedoch häufig gleiche Meßwerte auf, da entweder nur ungenau gemessen wird oder nur Zeitintervalle angegeben werden können, in denen Ereignisse stattfinden. In solchen Fällen muß die Partial Likelihood korrigiert werden. Breslow (1974) schlägt vor, (3.6.22) durch

$$PL(\beta; x_1, \ldots, x_n) = \prod_{i=1}^{k} \frac{\exp(s_i'\beta)}{\left[\sum\limits_{l\in R(t_{(i)})} \exp(x_l'\beta)\right]^{d_i}} \qquad (3.6.23)$$

zu approximieren. Dabei ist d_i die Anzahl der gleichen Verweildauerzeiten zum Zeitpunkt $t_{(i)}$, und s_i ist die Summe der Kovariablenvektoren dieser d_i Individuen.
Ist die Anzahl der Ties groß, empfiehlt es sich, ein diskretes Modell zu verwenden. Man vergleiche dazu Abschnitt 3.10.

Schätzung der Baseline-Hazardrate und der Survivorfunktion

Unter dem PH-Modell ist

$$S(t|x) = \exp(-\exp(x'\beta)\int_0^t \lambda_0(u)\,du). \tag{3.6.24}$$

Zur Schätzung von $S(t|x)$ benötigt man neben den Schätzungen $\hat{\beta}$ auch die Schätzung der Baseline-Hazardrate $\lambda_0(t)$. Hierzu existieren mehrere Vorschläge, von denen hier nur zwei behandelt werden sollen, da sie in einigen Programmpaketen (z.B. BMDP) implementiert sind.
Breslow (1974) schlägt vor, $\lambda_0(t)$ als konstant zwischen den beobachteten Verweildauerzeiten $t_{(1)} < t_{(2)} < \ldots < t_{(k)}$ anzunehmen:

$$\lambda_0(t) = \lambda_i \qquad \text{für } t_{(i-1)} < t \leq t_{(i)},\ i = 1, \ldots, k$$

mit $t_{(0)} = 0$. Daraus ergibt sich der ML-Schätzer

$$\hat{\lambda}_i = \frac{1}{t_{(i)} - t_{(i-1)}} \frac{d_i}{\sum\limits_{l \in R(t_{(i)})} \exp(x_l'\beta)} \quad \text{für } i = 1, \ldots, k. \tag{3.6.25}$$

Dabei ist d_i die Anzahl der zum Zeitpunkt $t_{(i)}$ gerade zuendegehenden Episoden. Für genügend genaue Messungen ist $d_i = 1$.
Link (1984) verwendet das Integral des Breslow-Schätzers (dies entspricht einer linearen Interpolation) zur Schätzung der kumulierten Hazardrate

$$\Lambda_0(t) = \int_0^t \lambda_0(u)\,du.$$

Man erhält

$$\hat{\Lambda}_0(t) = \int_0^t \hat{\lambda}_0(u)\,du = \sum_{i=1}^{s} (t_{(i)} - t_{(i-1)})\,\hat{\lambda}_i + (t - t_{(s)})\,\hat{\lambda}_{s+1}, \tag{3.6.26}$$

wobei zu gegebenem t der Wert s so zu wählen ist, daß $t_{(s)} < t$ und $t_{(s+1)} \geq t$.
Als Schätzung der Survivorfunktion erhält man

$$\hat{S}(t|x) = \exp(-\hat{\Lambda}(t|x)) = \exp(-\exp(x'\hat{\beta})\,\hat{\Lambda}_0(t)). \tag{3.6.27}$$

3.6.5 Maximum-Likelihood-Schätzung für Competing-Risks-Modelle

Die statistische Darstellung von Mehr-Zustands- oder Competing-Risks-Modellen der Ereignisanalyse erfolgte in Abschnitt 3.4. In diesem Kapitel wird die Maximum-Likelihood-Schätzung der Modellparameter für diese Modelle behandelt.

Es seien m Ereignisarten oder Zielzustände möglich, die mit y, $y \epsilon \{1, \ldots, m\}$, bezeichnet werden. Der Beitrag eines Individuums i zur Likelihood-Funktion ist bei gegebenem Kovariablenvektor $\mathbf{x}_i$

$$L_i = f(t_i, y_i|\mathbf{x}_i)^{\delta_i} S(t_i|\mathbf{x}_i)^{1-\delta_i} = \lambda_{y_i}(t_i|\mathbf{x}_i)^{\delta_i} S(t_i|\mathbf{x}_i). \tag{3.6.28}$$

Dies ergibt in Abhängigkeit von den übergangsspezifischen Hazardraten

$$\begin{aligned} L_i &= \lambda_{y_i}(t|\mathbf{x}_i)^{\delta_i} \exp(-\int_0^{t_i} \lambda(u|\mathbf{x}_i)\, du) \\ &= \lambda_{y_i}(t|\mathbf{x}_i)^{\delta_i} \exp(-\int_0^{t_i} \sum_{j=1}^{m} \lambda_j(u|\mathbf{x}_i)\, du) \\ &= \lambda_{y_i}(t|\mathbf{x}_i)^{\delta_i} \prod_{j=1}^{m} \exp(-\int_0^{t_i} \lambda_j(u|\mathbf{x}_i)\, du). \end{aligned} \tag{3.6.29}$$

Für die gesamte Likelihood-Funktion erhält man

$$L = \prod_{i=1}^{n} \lambda_{y_i}(t|\mathbf{x}_i)^{\delta_i} \prod_{j=1}^{m} \exp(-\int_0^{t_i} \lambda_j(u|\mathbf{x}_i)\, du). \tag{3.6.30}$$

Dabei ist δ_i wieder der Zensierungsindikator. Im folgenden wird (3.6.30) etwas umgeformt. Seien $t_{j1} < t_{j2} < \ldots < t_{jn_j}$ die n_j beobachteten nicht zensierten Zeiten bis zum Übergang in den Zustand j beziehungsweise bei denen Ereignisart j aufgetreten ist ($j \epsilon \{1, \ldots, m\}$). Dann läßt sich die Likelihood-Funktion umschreiben in

$$L = \prod_{j=1}^{m} \prod_{k=1}^{n_j} \lambda_j(t_{jk}|\mathbf{x}_{jk}) \prod_{i=1}^{n} S_j(t_i|\mathbf{x}_i). \tag{3.6.31}$$

Dabei ist $\mathbf{x}_{jk}$ der Kovariablenvektor des Individuums mit der beobachteten nicht zensierten Zeitdauer t_{jk} und

$$S_j(t_i|\mathbf{x}_i) = \exp(-\int_0^{t_i} \lambda_j(u|\mathbf{x}_i)\, du).$$

Aus (3.6.31) wird ersichtlich, daß sich die Likelihood-Funktion aufspalten läßt in das Produkt

$$L = \prod_{j=1}^{m} L_j \quad \text{mit } L_j = \prod_{k=1}^{n_j} \lambda_j(t_{jk}|\mathbf{x}_{jk}) \prod_{i=1}^{n} S_j(t_i|\mathbf{x}_i).$$

Die Faktoren L_j lassen sich noch umformen zu

$$L_j = \prod_{i=1}^{n} [\lambda_j(t_i|\mathbf{x}_i)]^{\delta_{ij}} S_j(t_i|\mathbf{x}_i) \tag{3.6.32}$$

$$\text{mit } \delta_{ij} = \begin{cases} 1 & \text{Individuum i geht zum Zeitpunkt } t_i \text{ in den Zustand j über} \\ & \text{(bzw. Ereignisart j tritt ein)} \\ 0 & \text{sonst.} \end{cases}$$

Hängen die übergangsspezifischen Hazardraten $\lambda_j(t|x)$ von Parametervektoren θ_j ab, j = 1, ..., m, die keine Komponenten gemeinsam haben, so kann die Log-Likelihood-Funktion $\ln L = \sum_{j=1}^{m} \ln L_j$ getrennt für jedes j maximiert werden.
Man kann insbesondere die bisher entwickelten Programme verwenden, wobei alle Zeitdauern mit einem von j verschiedenen Zielzustand als zensierte Beobachtungen betrachtet werden. Dies wird aus (3.6.32) deutlich.
Zur Modellierung der übergangsspezifischen Hazardrate können im Prinzip alle in Abschnitt 3.3 behandelten Ansätze verwendet werden. Es ist allerdings zu beachten, daß mit zunehmender Zahl der Zielzustände beziehungsweise Ereignisarten die Anzahl der Parameter im Gesamtmodell stark ansteigt, so daß eine Verschlechterung der Schätzgenauigkeit zu erwarten ist. Im konkreten Anwendungsfall ist stets ein Kompromiß zwischen den inhaltlichen Erfordernissen und den statistischen Konsequenzen anzustreben, insbesondere bei kleinen und mittleren Stichprobenumfängen.

3.6.6 Maximum-Likelihood-Schätzung im Mehr-Episoden-Fall

Schließlich wird in diesem Abschnitt noch die Erweiterung der Schätzverfahren auf den Mehr-Episoden-Fall erörtert. Für jedes Individuum i (i = 1, ..., n) der Stichprobe muß die vollständige Ereignisgeschichte im Beobachtungszeitraum bekannt sein. Dies erfordert die Angabe folgender Daten:

y_{i0}	Startzustand,
n_i	Anzahl der Episoden im Beobachtungszeitraum,
$t_{i1} < t_{i2} < \ldots < t_{in_i}$	Zeitpunkte, zu denen Zustandswechsel stattfinden beziehungsweise Ereignisse eintreten,
$y_{i1}, y_{i2}, \ldots, y_{in_i}$	Zustände, die zu den obigen Zeitpunkten angenommen werden,
δ_i	Indikator, ob die n_i-te Episode zensiert ist oder nicht,
$x_{i1}, x_{i2}, \ldots, x_{in_i}$	Vektoren von Kovariablen, die zu Beginn jeder Episode gemessen werden.

Im folgenden wird bei der Bildung des Beitrages des Individuums i zur Likelihood-Funktion aus Gründen der einfacheren Schreibweise der Index i weggelassen. Die gesamte Likelihood-Funktion ergibt sich wieder als Produkt der Beiträge aller Individuen der Stichprobe.
Wir gehen zunächst davon aus, daß keine zensierten Beobachtungen der letzten Episode vorliegen, daß also $\delta_i = 1$ ist für alle i.
Der Beitrag eines Individuums zur Likelihood-Funktion ist, gegeben der Anfangszustand und die Kovariablenvektoren

$$L_i = f(t_{n_i}, y_{n_i}, \ldots, t_1, y_1 | y_0, x_1, \ldots, x_{n_i}). \tag{3.6.33}$$

Es wird die Annahme getroffen, daß sich (3.6.33) faktorisieren läßt in

$$L_i = \prod_{k=1}^{n_i} f(t_k, y_k | H_{k-1}, x_k) \cdot c, \tag{3.6.34}$$

wobei in $H_{k-1} = \{t_{k-1}, y_k, x_{k-1}, \ldots, t_1, y_1, x_1, y_0\}$, $H_0 = \{y_0\}$, die Vorgeschichte des Prozesses zusammengefaßt und c ein Term ist, der nicht von den relevanten Parametern abhängt.
Die einzelnen Faktoren in (3.6.34) lassen sich umformen zu

$$f(t_k, y_k | H_{k-1}, x_k) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} P(t_k \leq T_k < t_k + \Delta t, y_k | H_{k-1}, x_k)$$

$$= \lim_{\Delta t \to 0} \frac{1}{\Delta t} P(t_k \leq T_k < t_k + \Delta t, y_k | T_k \geq t_k, H_{k-1}, x_k)\, P(T_k \geq t_k | H_{k-1}, x_k). \tag{3.6.35}$$

Der erste Term auf der rechten Seite von (3.6.35) ist gerade die Übergangsrate in den Zustand y_k. Der zweite Term gibt die Wahrscheinlichkeit dafür an, während der k-ten Episode den Zustand y_{k-1} bis zum Zeitpunkt t_k nicht zu verlassen. Diese Wahrscheinlichkeit ist gegeben durch die Survivorfunktion (3.5.16). Es ergibt sich mit (3.5.18)

$$L_i = \prod_{k=1}^{n_i} \lambda_{y_k}^k (t_k | H_{k-1}, x_k) \exp\left(-\int_{t_{k-1}}^{t_k} \lambda^k (u | H_{k-1}, x_k)\, du\right).$$

Nun ist noch zu berücksichtigen, daß die letzte Episode eines Individuums zensiert sein kann. Man erhält

$$L_i = \prod_{k=1}^{n_i} [\lambda_{y_k}^k (t_k | H_{k-1}, x_k)]^{\delta_k} \exp\left(-\int_{t_{k-1}}^{t_k} \lambda^k (u | H_{k-1}, x_k)\, du\right) \tag{3.6.36}$$

mit $\delta_k = 1$ für $k = 1, \ldots, n_i - 1$ und $\delta_{n_i} = 0$, falls die letzte Episode des Individuums zensiert ist beziehungsweise $\delta_{n_i} = 1$ sonst.
Die gesamte Likelihood-Funktion ist dann

$$L = \prod_{i=1}^{n} \prod_{k=1}^{n_i} [\lambda_{y_{ik}}^k (t_{ik} | H_{i,\,k-1}, x_{ik})]^{\delta_{ik}} \exp\left(-\int_{t_{i,\,k-1}}^{t_{ik}} \lambda^k (u | H_{i,\,k-1}, x_{ik})\, du\right). \tag{3.6.37}$$

In Erweiterung von (3.6.32) läßt sich die Likelihood-Funktion (3.6.37) umformen in

$$L = \prod_{k} \prod_{j=1}^{m} \prod_{i=1}^{n} [\lambda_j^k (t_{ik} | H_{i,\,k-1}, x_{ik})]^{\delta_{ikj}} [S_j^k (t_{ik} | H_{i,\,k-1}, x_{ik})]^{\epsilon_{ik}} \tag{3.6.38}$$

mit

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

$$\epsilon_{ik} = \begin{cases} 1 & \text{falls Individuum i die k-te Episode erlebt} \\ 0 & \text{sonst} \end{cases}$$

und

$$S_j^k(t_{ik}|H_{i,k-1}, x_{ik}) = \exp(-\int_{t_{i,k-1}}^{t_{ik}} \lambda_j^k(u|H_{i,k-1}, x_{ik})\, du).$$

Aus (3.6.38) wird ersichtlich, daß die Log-Likelihood-Funktion getrennt für jedes k und jedes j maximiert werden kann, wenn die Hazardraten $\lambda_j^k(\cdot)$ von Parametervektoren θ_{jk} abhängen, die keine Komponenten gemeinsam haben. Für die k-te Episode sind nur die Individuen der Stichprobe heranzuziehen, die auch mindestens k Episoden erlebt haben. In bezug auf die einzelnen Zielzustände (Ereignisarten) ist so zu verfahren wie am Ende des letzten Abschnitts beschrieben wurde. Wird gerade der Zielzustand j untersucht, so sind alle Zeitdauern der laufenden Episode mit einem von j verschiedenen Zielzustand als zensierte Beobachtungen zu betrachten.
Zur Modellierung der Hazardraten $\lambda_j^k(t|H_{k-1}, x_k)$ kommen im Prinzip wieder sämtliche Ansätze aus Abschnitt 3.3 in Frage. Für das Cox-Modell beispielsweise ist die Hazardrate

$$\lambda_j^k(t|H_{k-1}, x_k) = \lambda_{0j}^k(t) \exp(x_k'\beta_j^k), \tag{3.6.39}$$

wobei der relevante Teil der Vorgeschichte H_{k-1} in den aktuellen Kovariablenvektor x_k aufgenommen wird. Auf die Konstruktion der Partial Likelihood wird hier nicht eingegangen. Sie ist in Hamerle (1984) ausführlich beschrieben.
An dieser Stelle wird noch kurz auf das in den Kapiteln 5 und 6 vorwiegend verwendete spezielle Modell (3.5.22) eingegangen. Seine Hazardrate ist

$$\lambda^k(t|x_k) = \lambda_0(v) \exp(x_k'\beta) \tag{3.6.40}$$

mit $v = t - t_{k-1}$. Eine Partial-Likelihood für dieses Modell ist

$$\prod_k \prod_{i=1}^{d_k} \frac{\exp(x_{ik}'\beta)}{\sum_k \sum_{l \in R_k(v_{ik})} \exp(x_{lk}'\beta)}. \tag{3.6.41}$$

Dabei sind $v_{ik}, \ldots, v_{d_k k}$ die d_k (nicht zensierten) Verweildauern der k-ten Episode und $R_k(v)$ ist die Risikomenge der k-ten Episode. Dieselbe Partial-Likelihood erhält man aber auch, wenn man sämtliche Episoden zusammen nimmt und ein Ein-Episoden-Modell verwendet. Hat eine Person n_i Episoden durchlaufen, so gehen diese Episoden als n_i unabhängige Episoden in das Ein-Episoden-Modell ein, wobei sich die Kovariablenvektoren unterscheiden können. Damit besteht für das Modell (3.6.40) die Möglichkeit, für den Ein-Episoden-Fall konzipierte Programmsysteme auch im Mehr-Episoden-Fall zu verwenden. Man beachte jedoch, daß dies für allgemeine Mehr-Episoden-Modelle wie (3.6.39) im allgemeinen nicht gilt.

3.7 Hypothesentests und Modellwahl

Im Rahmen einer konkreten Datenanalyse ist die Festlegung der allgemeinen Struktur der Beziehung zwischen den Kovariablen oder prognostischen Faktoren und den Verweildauern oder Lebenszeiten von besonderem Interesse. Neben der Auswahl geeigneter Kovariablen kommt der Spezifikation des Modells, das in der Regel durch die Hazardrate determiniert wird, besondere Bedeutung zu. Bei einer schrittweisen Modellevaluation besteht die Möglichkeit, einige der Modellannahmen zu überprüfen. In einfachen Situationen mit nur einer Kovariablen x bringt meist schon eine graphische Darstellung von x gegenüber t oder lnt Aufschluß über die Form des Zusammenhangs. Darüber hinaus kann die Berechnung der mittleren Verweildauern beziehungsweise Lebenszeiten bei verschiedenen Ausprägungen von x nützlich sein. Diese Methoden reichen jedoch im allgemeinen nicht aus, wenn viele Kovariablen in das Modell aufgenommen werden. Außerdem sind viele Modelle an weitere Zusatzannahmen gebunden. So geht beispielsweise in das Cox-Modell oder das Weibull-Modell bei zeitunabhängigen Kovariablen die Annahme proportionaler Risiken ein. Diese Voraussetzung muß ebenfalls überprüft werden.
Im folgenden Abschnitt werden einige Methoden der Residualanalyse sowie graphische Modelltests vorgestellt. Dann werden speziell Tests für das Cox-Modell behandelt, und im letzten Abschnitt werden Möglichkeiten zur Signifikanzprüfung einzelner Regressionskoeffizienten oder Modellteile sowie der Variablenselektion erörtert.

3.7.1 Residuenanalyse und Modelltests

Betrachtet man den Zusammenhang zwischen der Survivorfunktion und der kumulativen Hazardrate (bei gegebenem Kovariablenvektor)

$$S(t|\mathbf{x}) = \exp(-\int_0^t \lambda(u|\mathbf{x})\,du) = \exp(-\Lambda(t|\mathbf{x})), \tag{3.7.1}$$

erkennt man, daß die durch

$$r = \Lambda(T|\mathbf{x})$$

definierte Zufallsvariable eine Exponentialverteilung mit Parameter $\lambda = 1$ besitzt.
Liegt eine Stichprobe mit nicht-zensierten Daten $(t_1, \mathbf{x}_1), \ldots, (t_n, \mathbf{x}_n)$ vor, so sind die $\Lambda(t_i|\mathbf{x}_i)$'s unabhängige Realisierungen einer standard-exponentialverteilten Zufallsvariablen. Es ist daher naheliegend (Cox/Snell 1968; Kay 1977), durch

$$\hat{r}_i = \hat{\Lambda}(t_i|\mathbf{x}_i) = -\ln\hat{S}(t_i|\mathbf{x}_i) \tag{3.7.2}$$

„Residuen" zu definieren, wobei $\hat{\Lambda}(t_i|\mathbf{x}_i)$ die Schätzungen der unbekannten Parameter enthalten (vgl. z. B. (3.6.24) bis (3.6.27) für das Cox-Modell). In einer

ersten Näherung werden $\hat{r}_1, \ldots, \hat{r}_n$ ebenfalls als Zufallsstichprobe aus einer standard-exponentialverteilten Grundgesamtheit aufgefaßt. Man beachte, daß die in (3.7.2) definierten Residuen im allgemeinen weder unabhängig sind noch eine identische Verteilung besitzen, so daß die geschilderte Vorgehensweise lediglich als Approximation aufzufassen ist.

Beispiel:

Es ergibt sich für die Survivorfunktion des Weibull-Regressionsmodells

$$S(t|\mathbf{x}) = \exp(-(t\exp(-\mathbf{x}'\beta))^{\delta}), \tag{3.7.3}$$

wobei die erste Komponente von $\mathbf{x}$ stets gleich 1 ist, und daraus die kumulativen Hazardraten

$$r_i = \Lambda(t_i|\mathbf{x}_i) = (t_i \exp(-\mathbf{x}_i'\beta))^{\delta} \qquad , i = 1, \ldots, n.$$

Die r_i's sind unabhängig und standard-exponentialverteilt. Setzt man in r_i die zugehörigen Parameterschätzungen ein, erhält man die Residuen

$$\hat{r}_i = (t_i \exp(-\mathbf{x}_i'\hat{\beta}))^{\hat{\delta}} \qquad , i = 1, \ldots, n. \tag{3.7.4}$$

Man kann die Residuen auch auf anderem Wege definieren, indem man vom Regressionsansatz

$$y = \mathbf{x}'\beta + \sigma\omega$$

ausgeht, wobei $y = \ln T$ sowie $\sigma = 1/\delta$ ist und ω eine Standard-Extremwertverteilung besitzt. In Analogie zur herkömmlichen multiplen Regression definiert man die Residuen

$$\hat{\omega}_i = \frac{y_i - \mathbf{x}_i'\hat{\beta}}{\hat{\sigma}} \qquad i = 1, \ldots, n. \tag{3.7.5}$$

Die in (3.7.5) definierten Residuen hängen mit (3.7.4) durch $\hat{\omega}_i = \ln\hat{r}_i$ zusammen, und im Prinzip kann je nach Untersuchungsziel sowohl mit (3.7.5) als auch mit (3.7.4) gearbeitet werden.
Die Residuen können zu verschiedenen Zwecken eingesetzt werden. So können mit Hilfe von Residuen-Plots Verteilungsannahmen über die r_i graphisch überprüft werden, oder man kann die Residuen zu bestimmten Regressionsvariablen graphisch in Beziehung setzen, um auf diese Weise über die Angemessenheit des Modells Aufschluß zu erhalten.
Betrachtet man die Survivorfunktion des Weibull-Regressionsmodells in (3.7.3), so erhält man durch zweimaliges Logarithmieren

$$\ln(-\ln S(t|\mathbf{x})) = \delta \ln t - \mathbf{x}'\beta.$$

Aus dieser Beziehung läßt sich ein Modelltest zur Überprüfung der Gültigkeit des Weibull-Modells konstruieren. Man kategorisiert die stetigen Kovariablen, so daß für jeden Kovariablenvektor $\mathbf{x}_j$ mehrere Beobachtungen vorliegen. Bildet

man in jeder Gruppe $\mathbf{x}_j$ die Schätzungen $\hat{S}(t|\mathbf{x}_j)$ und trägt diese gegen lnt auf, so müssen diese Plots näherungsweise linear und parallel zueinander sein. Ein anderer graphischer Modelltest wird im nächsten Abschnitt im Zusammenhang mit dem Cox-Modell behandelt.

Bisher sind wir davon ausgegangen, daß keine zensierten Beobachtungen vorliegen. Bei zensierten Daten ist die Vorgehensweise zu modifizieren. Eine Möglichkeit besteht darin, zur Schätzung von $\hat{S}(t|\mathbf{x}_j)$ die Produkt-Limit-Methode (Kaplan-Meier-Schätzer) zu verwenden. Analoges gilt für die Residuen.

Ist t_i^* eine zensierte Verweildauer, so ist das zugehörige Residuum ebenfalls zensiert, und man kann die Residuen $\hat{r}_1, \ldots, \hat{r}_n$ als zensierte Stichprobe behandeln. Dann wird die Produkt-Limit-Schätzung oder die empirische Hazardrate aus den Residuen ermittelt und zur Schätzung der zugrundeliegenden Survivorfunktion herangezogen. Graphische Tests können dann Informationen über die Verteilung der Residuen liefern.

Eine andere Möglichkeit (vgl. Lawless 1982, S. 281) besteht darin, bei zensierten Daten Schätzungen der mittleren Restverweildauer zu berechnen. Die Schätzung wird dann zur zensierten Beobachtung hinzuaddiert und der resultierende Wert als unzensierte Beobachtung in die Stichprobe aufgenommen. Besonders einfach ist dies bei der Exponentialverteilung. Ist die Verweildauer T_i, gegeben $\mathbf{x}_i$, exponentialverteilt mit dem Mittelwert $1/\lambda_i = \exp(\mathbf{x}_i\beta)$, und liegt die zensierte Beobachtung t_i^* vor, so prüft man leicht nach, daß

$$E(T_i|T_i \geq t_i^*) = t_i^* + 1/\lambda_i$$

gilt. Definiert man die Residuen

$$\hat{r}_i = \hat{\lambda}_i\, t_i = t_i \exp(-\mathbf{x}_i'\hat{\beta}) \tag{3.7.6}$$

für nicht-zensierte Beobachtungen, erhält man für eine zensierte Beobachtung t_i^* eine Adjustierung durch

$$\hat{r}_i = t_i^* \exp(-\mathbf{x}_i'\hat{\beta}) + 1. \tag{3.7.7}$$

Da die $r_i = \Lambda(t_i|\mathbf{x}_i)$ eine Standard-Exponentialverteilung besitzen, kann man die gemäß (3.7.2) definierten Residuen $\hat{r}_i = \hat{\Lambda}(t_i|\mathbf{x}_i)$ $(i = 1, \ldots, n)$ in erster Näherung als Stichprobe aus einer standard-exponentialverteilten Grundgesamtheit auffassen. Für eine zensierte Beobachtung t_i^* wird entsprechend (3.7.7) ein adjustiertes Residuum

$$\hat{r}_i = \hat{\Lambda}(t_i^*|\mathbf{x}_i) + 1 \tag{3.7.8}$$

berechnet. Damit läßt sich wieder ein graphischer Modelltest konstruieren. Trägt man in einem Plot $-\ln\hat{S}(r)$ gegen r auf, so sollte sich bei Gültigkeit des postulierten Modells approximativ eine Gerade mit Steigung 1 ergeben. $\hat{S}(r)$ ist dabei die Produkt-Limit-Schätzung der Survivorfunktion der r_i. Für eine weitere Möglichkeit der Festlegung von Residuen und der Modellkontrolle für das Cox-Modell vergleiche man Schoenfeld (1982).

3.7.2 Modelltests für das Proportional-Hazards-Modell

Einige der im letzten Abschnitt behandelten Verfahren können auch zur Prüfung der Gültigkeit des Cox-Modells, insbesondere der Annahme proportionaler Risiken, verwendet werden. Das Cox-Modell besitzt die Hazardrate

$$\lambda(t|\mathbf{x}) = \lambda_0(t)\ \exp(\mathbf{x}'\boldsymbol{\beta})$$

mit der unspezifizierten Baseline-Hazardrate $\lambda_0(t)$ und der Survivorfunktion (vgl. (3.3.21))

$$S(t|\mathbf{x}) = S_0(t)^{\exp(\mathbf{x}'\boldsymbol{\beta})}. \tag{3.7.9}$$

Da das am Ende des letzten Abschnitts dargestellte Verfahren für beliebige Modelle durchführbar ist, kann es auch im Falle des Cox-Modells angewendet werden. Man definiert die Residuen gemäß (3.7.2)

$$\hat{r}_i = \hat{\Lambda}(t_i|\mathbf{x}_i) = -\ln\hat{S}(t_i|\mathbf{x}_i).$$

Im Falle des Cox-Modells ergibt sich aus (3.7.9)

$$\hat{r}_i = -\ln\hat{S}_0(t_i)\ \exp(\mathbf{x}_i'\hat{\boldsymbol{\beta}}), \tag{3.7.10}$$

wobei $\hat{S}_0(t)$ eine mit den Methoden aus Abschnitt 3.6.4 ermittelte Schätzung der „Baseline"-Survivorfunktion ist (vgl. (3.6.25)). $\hat{\boldsymbol{\beta}}$ ist die Maximum-Partial-Likelihood-Schätzung. Ermittelt man mit den Werten aus (3.7.10) die Produkt-Limit-Schätzung $\hat{S}(r)$ und plottet $\ln\hat{S}(r)$ gegen r, so sollte sich näherungsweise eine Gerade mit Steigung −1 ergeben.
Allerdings dürfte die eben beschriebene Methode nur funktionieren, wenn für die Baseline-Hazardrate ein parametrisches Modell formuliert wird. Werden jedoch $\boldsymbol{\beta}$ mit Hilfe der Partial-Likelihood-Methode und $\lambda_0(t)$ mit einem nichtparametrischen Schätzer der Form (3.6.25) oder (3.6.26) geschätzt, dann kann die Verteilung der gebildeten Residuen beträchtlich von der Exponentialverteilung abweichen. Lagakos (1981) demonstriert dies anhand von Beispielen. Deshalb ist nach Crowley/Storer (1983) der eben beschriebene graphische Residuentest bei unspezifizierter Grundhazardrate nicht zu empfehlen. In Kapitel 5 wird deshalb auf die Anwendung des graphischen Residuentests verzichtet. Weitere Anpassungstests für das Cox-Modell sind in Schoenfeld (1980), Andersen (1982) und Kemény/Rothmeier/Hamerle (1985) beschrieben.
Auf der Schätzung $\hat{S}_0(t)$ der Baseline-Survivorfunktion beruht eine weitere Möglichkeit der graphischen Überprüfung des Weibull-Modells, eines speziellen Proportional-Hazards-Modells. Dazu geht man aus vom Cox-Modell und führt mit den Methoden von Abschnitt 3.6.4 eine Schätzung von $S_0(t)$ durch. Das Weibull-Modell ist ein Spezialfall des Cox-Modells mit der Baseline-Hazardrate

$$\lambda_0(t) = \lambda\delta(\lambda t)^{\delta-1}$$

und der Baseline-Survivorfunktion

$$S_0(t) = \exp(-(\lambda t)^{\delta}).$$

Zweimaliges Logarithmieren von $S_0(t)$ ergibt

$$\ln(-\ln S_0(t)) = \delta \ln t + \delta \ln\lambda,$$

und eine graphische Darstellung von $\ln(-\ln S_0(t))$ gegen $\ln t$ müßte bei Gültigkeit des Modells näherungsweise eine Gerade ergeben. Für weitere Details vergleiche man Kay (1977) oder Kemény/Rothmeier/Hamerle (1985).

Interessiert man sich speziell für die Überprüfung der Proportionalitätsannahme des Cox-Modells, so kann dies durch Schichtenbildung und Einführung von geeigneten zeitabhängigen Kovariablen erfolgen. Der Einfachheit halber nehmen wir an, daß die Grundgesamtheit nur in zwei Teilgesamtheiten aufgeteilt wird (z. B. nach Geschlecht), bezüglich deren die Proportionalitätsannahme in Frage steht. Liegen proportionale Risiken vor, sind die Hazardraten in den beiden Teilgesamtheiten

$$h_1(t|\mathbf{x}) = \lambda_0(t) \exp(\mathbf{x}'\beta) \quad \text{bzw.} \quad h_2(t|\mathbf{x}) = \lambda_0(t) \exp(\alpha_1 + \mathbf{x}'\beta), \qquad (3.7.11)$$

wobei $\mathbf{x}$ der Vektor der anderen im Modell enthaltenen Kovariablen ist. Mit der Dummy-Variablen z_1, die für Individuen aus der zweiten Teilgesamtheit den Wert 1 annimmt und sonst 0 ist, läßt sich (3.7.11) zusammenfassen zu

$$h(t|\mathbf{x}, z_1) = \lambda_0(t) \exp(z_1\alpha_1 + \mathbf{x}'\beta).$$

Definiert man nun eine zusätzliche Kovariable z_2 durch

$$z_2 = z_1 \ln t$$

mit dem dazugehörigen Parameter α_2, ergibt sich

$$\begin{aligned} h(t|\mathbf{x}, z_1, z_2) &= \lambda_0(t) \exp(z_1\alpha_1 + z_2\alpha_2 + \mathbf{x}'\beta) \\ &= \lambda_0(t)\, t^{z_1\alpha_2} \exp(z_1\alpha_1 + \mathbf{x}'\beta), \end{aligned} \qquad (3.7.12)$$

und die Überprüfung von $H_0 : \alpha_2 = 0$ liefert einen Test der Proportionalitätsannahme. Dafür können dann die im nächsten Abschnitt behandelten Methoden auf der Basis der Partial-Likelihood verwendet werden.

Gelegentlich wird vorgeschlagen (Kalbfleisch/McIntosh 1977), die zeitabhängige Kovariable durch

$$z_2 = z_1(\ln t - c) \qquad (3.7.13)$$

festzulegen, wobei c das arithmetische Mittel der $\ln t_i$ ist. Dadurch kann nach Auffassung der Autoren eine zu hohe asymptotische Korrelation der Parameterschätzungen $\hat{\alpha}_1$ und $\hat{\alpha}_2$ vermieden werden.

Ist die Voraussetzung proportionaler Risiken in bezug auf ein oder mehrere Merkmale verletzt, so besteht die Möglichkeit, nach den Ausprägungen dieser Merkmale zu schichten (eventuell nach vorheriger Kategorisierung) und ein stratifiziertes Cox-Modell zu verwenden. Man vergleiche Abschnitt 3.3.4. Hat man s Schichten (Teilpopulationen) gebildet, so sind die schichtspezifischen Hazardraten

$$\lambda_j(t|\mathbf{x}) = \lambda_{0j}(t) \exp(\mathbf{x}'\beta).$$

Die Grundhazardfunktionen $\lambda_{01}(t), \ldots, \lambda_{0s}(t)$ müssen nicht zueinander proportional sein. Die Regressionskoeffizienten β bleiben für alle Schichten gleich.

3.7.3 Tests für Regressionskoeffizienten oder Modellteile

Wurde die allgemeine Modellstruktur, etwa die Form der Zeitabhängigkeit der Hazardrate und die Beziehung zwischen Kovariablen und Hazardrate, geprüft und hat man sich für einen Modellansatz entschieden, so können auch einzelne Regressionskoeffizienten oder Modellteile auf Signifikanz getestet werden. Die Tests beruhen auf den asymptotischen Eigenschaften, insbesondere der asymptotischen Normalität der Maximum-Likelihood-Schätzungen beziehungsweise der Maximum-Partial-Likelihood-Schätzungen beim Cox-Modell. Dabei ist zu beachten, daß die asymptotischen Eigenschaften noch nicht für alle in diesem Buch behandelten Situationen vollständig nachgewiesen sind.
Bezeichnet θ den Parametervektor des Modells, der den Vektor β der Regressionskoeffizienten sowie weitere ins Modell aufgenommene Parameter enthält (z. B. den Parameter δ im Weibull-Modell), lassen sich Hypothesen über einzelne Regressionskoeffizienten oder Modellteile zusammenfassen in einer allgemeinen linearen Hypothese

$$\mathbf{C}\,\theta = \mathbf{0}, \qquad \mathrm{rg}(\mathbf{C}) = m \tag{3.7.14}$$

mit einer geeigneten Matrix $\mathbf{C}$.
Prüft man $H_0 : \beta_i = 0$, so besteht $\mathbf{C}$ nur aus einer Zeile mit der 1 an der entsprechenden Stelle und sonst lauter Nullen. Für $H_0 : \beta_i = \beta_j = 0$ besteht $\mathbf{C}$ aus zwei Zeilen, die analog aufgebaut sind. In der Regel ist der Rang m von $\mathbf{C}$ gleich der Anzahl der Zeilen von $\mathbf{C}$.
Im Prinzip kann auch die Hypothese

$$\mathbf{C}\,\theta = \xi, \tag{3.7.15}$$

wobei ξ ein fester Wert ist, überprüft werden, bei praktischen Anwendungen ist aber meist die Hypothese (3.7.14) mit $\xi = \mathbf{0}$ von Interesse. Wir werden uns deshalb auf diesen Fall beschränken.

Prüfung einzelner Regressionskoeffizienten

Will man speziell die Hypothese $H_0 : \beta_i = 0$ überprüfen, so kann dies mit Hilfe der Schätzung $\hat{\beta}_i$ und der geschätzten (asymptotischen) Varianz $\widehat{\mathrm{Var}}(\hat{\beta}_i)$ erfolgen. $\widehat{\mathrm{Var}}(\hat{\beta}_i)$ ist das entsprechende Diagonalelement in der geschätzten Kovarianzmatrix von $\hat{\beta}$, der Inversen der beobachteten Informationsmatrix (vgl. Abschnitt 3.6.1, insbesondere Beziehung (3.6.8) und nachfolgende Bemerkungen). Die Teststatistik

$$\frac{\hat{\beta}_i}{\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_i)}} \tag{3.7.16}$$

ist bei Gültigkeit von H_0 asymptotisch standardnormalverteilt. Möchte man allgemeiner $H_0 : \beta_i = \xi$ überprüfen, so gilt diese Aussage für die Teststatistik

$$\frac{\hat{\beta}_i - \xi}{\sqrt{\hat{\text{Var}}\,(\hat{\beta}_i)}}. \qquad (3.7.17).$$

Simultane Prüfung mehrerer Regressionskoeffizienten beziehungsweise Parameter

Zur Prüfung der Hypothese, daß mehrere Parameter gleich 0 sind, wird die Hypothese in die Form (3.7.14) gebracht. Zur expliziten Durchführung des Tests können verschiedene Prüfgrößen gewählt werden.

a) Die Likelihood-Quotienten-Teststatistik ist

$$Lq = 2(\ln L(\hat{\theta}) - \ln L(\tilde{\theta})), \qquad (3.7.18)$$

wobei $\tilde{\theta}$ die Maximum-Likelihood-Schätzungen unter der Nebenbedingung $\mathbf{C}\theta = \mathbf{0}$ und $\hat{\theta}$ die Maximum-Likelihood-Schätzungen ohne Restriktionen sind.

b) Die Wald-Teststatistik ist

$$W = (\mathbf{C}\hat{\theta})' \, [\mathbf{C}\, \mathbf{Cov}\, (\hat{\theta})\, \mathbf{C}']^{-1} \, (\mathbf{C}\hat{\theta}). \qquad (3.7.19)$$

c) Neben der Likelihood-Quotienten- und der Wald-Teststatistik kann auch die Score-Statistik zur Prüfung der allgemeinen linearen Hypothese verwendet werden. Die Score-Statistik geht aus von der Score-Funktion $\mathbf{s}(\theta)$, dem Vektor der Ableitungen der Log-Likelihood-Funktion (vgl. Abschnitt 3.6.1). Man bildet die Prüfgröße

$$S = \mathbf{s}(\tilde{\theta})' \, \mathbf{I}(\tilde{\theta})^{-1} \, \mathbf{s}(\tilde{\theta}). \qquad (3.7.20)$$

$\mathbf{I}(\tilde{\theta})^{-1}$ ist die Inverse der beobachteten Informationsmatrix, ausgewertet an der Stelle $\tilde{\theta}$.

Die asymptotische Verteilung unter H_0 ist für alle drei Teststatistiken dieselbe. Sie besitzen bei Gültigkeit von H_0 asymptotisch eine zentrale χ^2-Verteilung mit $m = rg\,(\mathbf{C})$ Freiheitsgraden.

Man beachte, daß sämtliche Aussagen in diesem Abschnitt über Verteilungen von Prüfgrößen nur asymptotisch gelten. Die Anwendung der Tests setzt deshalb in der Praxis große Stichprobenumfänge voraus.

Variablenselektion

Bei der schrittweisen Regression zur Variablenselektion wird auf der Basis der berechneten Signifikanzwahrscheinlichkeiten bei jedem Schritt eine Variable in das Modell aufgenommen oder aus dem Modell eliminiert. Bei der Likelihood- beziehungsweise Partial-Likelihood-Quotientenmethode werden zur Aufnahme beziehungsweise Elimination von Kovariablen Signifikanzwahrscheinlichkeiten auf der Basis eines Likelihood- beziehungsweise Partial-Likelihood-Quotiententests ermittelt. Die Teststatistik lautet in Analogie zu (3.7.18):

$$2(\ln L(\hat{\beta}_\nu) - \ln L(\hat{\beta}_{\nu+1})) \qquad , \nu = 1, 2, \ldots,$$

wobei $\hat{\beta}_\nu$ der beim ν-ten Schritt berechnete ML-Schätzer ist. Diese Testgröße ist asymptotisch $\chi^2(1)$-verteilt. Für eine ausführliche Diskussion der Variablenselektion bei Regressionsmodellen, insbesondere auch im Hinblick auf numerisch effiziente Verfahren, vergleiche man Fahrmeir/Hamerle (1984, Kap. 4.).

3.8 Einbeziehung von zeitabhängigen Kovariablen

Bisher wurde davon ausgegangen, daß die Kovariablen zu Beginn einer Episode gemessen werden und sich ihre Werte im Verlauf der Episode nicht ändern. In einer Reihe von Anwendungen können jedoch auch die Kovariablen von der Zeit beziehungsweise Verweildauer abhängen. Beispiele hierfür sind Alter, Einkommen, Familienstand oder eine Therapie, die nur während eines bestimmten Zeitraums angewendet wird. Für die Art der Zeitabhängigkeit der Kovariablen gibt es verschiedene Möglichkeiten. Eine einfache Form der Zeitabhängigkeit besteht beispielsweise bei den Variablen Alter und Berufserfahrung. Hier läßt sich die Abhängigkeit von der Zeit in eine vorher festgelegte funktionale Form fassen. Kalbfleisch/Prentice (1980, Kap. 5) sprechen in einem solchen Fall von „*definierten*" Kovariablen. Daneben existieren Kovariablen, die ihrerseits Realisierungen eines stochastischen Prozesses sind. Kalbfleisch/Prentice (1980) unterscheiden hier zwischen *externen* und *internen* zeitabhängigen Kovariablen. Bei externen zeitabhängigen Kovariablen wird der Pfad des Kovariablenvektors $\mathbf{x}(t)$ nicht von der Verweildauer beeinflußt. Umgekehrt kann er aber auf die Verweildauer sehr wohl einwirken. Umweltfaktoren, berufliche oder wirtschaftliche Rahmenbedingungen zählen häufig zu diesem Typ von zeitabhängigen Kovariablen. Im Gegensatz dazu wird bei internen zeitabhängigen Kovariablen der Kovariablenprozeß von den Individuen, deren Verweildauer analysiert wird, selbst generiert, das heißt, die Beobachtung einer konkreten Ausprägung von $\mathbf{x}(t)$ hängt von den Ergebnissen des Prozesses selbst ab. Beispielsweise hängt die Entscheidung für Weiterbildungsmaßnahmen vom konkreten Berufsverlauf ab (Andreß 1985).

Im folgenden gehen wir zunächst von externen zeitabhängigen Kovariablen aus. Die bei internen zeitabhängigen Kovariablen auftretenden Probleme werden am Schluß dieses Abschnitts behandelt. Wir legen den Fall einer Episode zugrunde, die Methoden können ohne Schwierigkeiten auf den Mehr-Episoden-Fall oder auf Competing-Risks-Ansätze übertragen werden.

Die Hazardrate wird in der Regel wie bei zeitunabhängigen Kovariablen definiert. Bezeichnet $\mathbf{x}(t)$ den Kovariablenvektor zum Zeitpunkt t, so ist

$$\lambda(t|\mathbf{x}(t)) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} P(t \leq T \leq t + \Delta t | T \geq t, \mathbf{x}(t)) \tag{3.8.1}$$

die Hazardrate. Survivor- und Dichtefunktion können bei externen zeitabhän-

gigen Kovariablen ebenfalls in Analogie zur Definition bei fixen Kovariablen festgelegt werden.

$$S(t|\mathbf{x}(t)) = P(T \geq t|\mathbf{x}(t)) = \exp(-\int_0^t \lambda(u|\mathbf{x}(u))\, du) \tag{3.8.2}$$

$$f(t|\mathbf{x}(t)) = \lambda(t|\mathbf{x}(t)) \cdot \exp(-\int_0^t \lambda(u|\mathbf{x}(u))\, du) \tag{3.8.3}$$

Aus (3.8.2) wird ersichtlich, daß die numerische Berechnung der Survivorfunktion Schwierigkeiten bereiten kann, da auch über die Kovariablen zu integrieren ist.

Liegen ausschließlich „definierte" zeitabhängige Kovariablen vor und ist die funktionale Abhängigkeit der Kovariablen von der Zeit nicht zu kompliziert, schafft die Ermittlung von (3.8.2) im allgemeinen keine großen Probleme. Das gleiche gilt für Kovariablen, deren Pfade die Gestalt einer Treppenfunktion besitzen. Sie sind stückweise konstant und die kumulative Hazardrate kann in eine Summe von Integralen zerlegt werden. Bezeichnen $t_0 < t_1 \ldots < t_s$ die Änderungszeitpunkte des Kovariablenvektors im Intervall $[0, t)$ und sei $t_{s+1} = t$, so ergibt sich für die Survivorfunktion

$$\begin{aligned} S(t|\mathbf{x}(t)) &= \exp(-\sum_{r=1}^{s+1} \int_{t_{r-1}}^{t_r} \lambda(u|\mathbf{x}(t_{r-1}))\, du) \\ &= \prod_{r=1}^{s+1} S(t_r|t_{r-1}, \mathbf{x}(t_{r-1})) \end{aligned} \tag{3.8.4}$$

mit

$$S(t_r|t_{r-1}, \mathbf{x}(t_{r-1})) = P(T \geq t_r|T \geq t_{r-1}, \mathbf{x}(t_{r-1})).$$

Mit Hilfe von (3.8.1) und (3.8.4) läßt sich dann die Likelihood-Funktion konstruieren.

Ändern sich die Kovariablen kontinuierlich im Zeitablauf, so kann die kumulative Hazardrate eventuell durch numerische Integration ermittelt werden, falls ausreichend viele Meßwerte des Kovariablenvektors vorliegen. Ist dies nicht der Fall, kann die Integration nicht durchgeführt werden. Man kann dann versuchen, die Kovariablen durch Treppenfunktionen oder von Meßpunkt zu Meßpunkt stückweise lineare Funktionen zu approximieren und wie in (3.8.4) zu verfahren. Eine andere Möglichkeit sind die von Tuma u. a. (1979) beziehungsweise Tuma/Hannan (1984) eingeführten periodenspezifischen Modelle, die in Abschnitt 3.3.2 behandelt wurden. Dabei werden für jede Periode die aktualisierten Werte der zeitabhängigen Kovariablen eingesetzt. Schließlich stehen auch die in Abschnitt 3.10 behandelten diskreten Modelle als Approximation zur Verfügung.

Eine weitere Möglichkeit zur Einbeziehung von Kovariablen, die sich im Zeitablauf kontinuierlich verändern, bietet die Anwendung des Cox-Modells mit der Hazardrate

$$\lambda(t|\mathbf{x}(t)) = \lambda_0(t) \exp(\mathbf{x}'(t)\boldsymbol{\beta}). \tag{3.8.5}$$

Die Ausführungen von Abschnitt 3.6.4 zur Ableitung der Partial-Likelihood-Funktion lassen sich übertragen, und man erhält

$$PL(\beta) = \prod_{i=1}^{k} \frac{\exp(x'_i(t_{(i)})\,\beta)}{\sum_{l \in R(t_{(i)})} \exp(x'_l(t_{(i)})\,\beta)}. \tag{3.8.6}$$

Dabei sind $t_{(1)} < t_{(2)} < \ldots < t_{(k)}$ die Ereigniszeitpunkte ($k \leq n$), und $R(t_{(i)})$ ist die Risikomenge zum Zeitpunkt $t_{(i)}$. Für Ties ist (3.8.6) entsprechend (3.6.23) zu modifizieren. Auch die Anwendung des Cox-Modells und die Schätzung auf der Basis von (3.8.6) erfordert eine relativ genaue Aufzeichnung der Kovariablenpfade. Zu jedem Ereigniszeitpunkt $t_{(i)}$ müssen für alle Individuen der Risikomenge $R(t_{(i)})$, die bis zu diesem Zeitpunkt noch kein Ereignis hatten und nicht zensiert sind, die Werte der Kovariablen vorliegen. Dies wird im allgemeinen nur bei regelmäßig erhobenen Kovariablen der Fall sein. So werden bei medizinischen Studien bestimmte Merkmale wie z.B. Blutdruck mehrmals am Tag gemessen. In anderen Situationen, in denen nicht so viele Meßwerte der Kovariablen vorliegen, behilft man sich gelegentlich damit, daß man statt $x(t_{(i)})$ den Wert der Kovariablen an einem tatsächlich gemessenen Zeitpunkt nimmt, der $t_{(i)}$ am nächsten liegt. Ist der Zustandsraum der Kovariablen diskret und kennt man die Zustandswechsel der Kovariablen im Beobachtungszeitraum, so kann (3.8.6) in der Regel ohne Probleme angewendet werden. Eine ausführliche Diskussion der Inklusion zeitabhängiger Kovariablen findet man bei Petersen (1985).

Handelt es sich bei einigen der Kovariablen um interne zeitabhängige Kovariablen, können besondere Probleme auftreten, zum Beispiel bei medizinischen Lebenszeitstudien. Hier enthalten diese Kovariablen stets unmittelbare Informationen über die Lebenszeit. Hat man eine interne Kovariable bis zum Zeitpunkt t beobachtet, so muß das Individuum bis zu diesem Zeitpunkt gelebt haben und es gilt

$$P(T \geq t | x(t)) = 1.$$

Damit besitzt dieser Ausdruck nicht die Interpretation einer Survivorfunktion wie in (3.8.2) und die Likelihood-Funktion läßt sich nicht auf dem in Abschnitt 3.6.2 geschilderten Weg konstruieren. Man kann die Likelihood-Funktion jedoch auf anderem Weg ableiten (vgl. Kalbfleisch/Prentice 1980, Kap. 5.3) und zeigen, daß sich insbesondere für das PH-Modell wieder die Partial-Likelihood (3.8.6) ergibt.

Die bei der Einbeziehung von internen zeitabhängigen Kovariablen auftretenden Probleme sind noch nicht vollständig erforscht. So ist bei der Interpretation eine gewisse Vorsicht geboten, vor allem, wenn es sich um den Vergleich der Wirkungsweise verschiedener Treatments, Programme oder Kampagnen handelt und der Kovariablenprozeß seine Werte erst nach der Treatment-Zuordnung annimmt. Für ein illustratives Beispiel vergleiche man Kalbfleisch/Prentice (1980, S. 126).

Interne zeitabhängige Kovariablen können auch zum Problem der wechselseitigen Abhängigkeit von Ereignissen führen, denn bei den Kovariablen und der Dauer bis zum Eintreffen eines wiederholbaren Ereignisses handelt es sich oft um Prozesse, die sich gegenseitig beeinflussen, und das Eintreffen des einen Ereignisses verändert das Risiko für das andere Ereignis und umgekehrt. Beispiele sind Familienstand und Dauer der Beschäftigung in einem Job oder Schulbesuch (ja/nein) und „Leben bei den Eltern" (ja/nein) (vgl. Sørensen/Sørensen 1983). Auch diese Problematik von interdependenten Prozessen wurde noch nicht ausreichend untersucht. Man vergleiche dazu Tuma/Hannan (1984, Kap. 9 und 16), Coleman (1984b), Clayton/Cuzick (1985) und Petersen (1985).

3.9 Einbeziehung unbeobachteter Populationsheterogenität

3.9.1 Beispiele zur unbeobachteten Heterogenität

In der Regel werden neben den in das Modell aufgenommenen Kovariablen weitere personen- und umweltspezifische Merkmale, die nicht erhoben werden oder nicht bekannt sind, die Übergangs- beziehungsweise Hazardrate beeinflussen. Bisher wurde davon ausgegangen, daß die erhobenen Kovariablen die Hazardrate vollständig determinieren, und eventuell unbeobachtete Merkmale wurden nicht berücksichtigt. Wie sich eine derartige Aggregation auf die Hazardrate auswirken kann, wird an zwei einfachen Beispielen demonstriert.

Beispiel 1

Angenommen, die Population sei in zwei Teilgesamtheiten unterteilt, die mit x_1 und x_2 bezeichnet werden. In jeder der Teilgesamtheiten sei die Hazardrate konstant mit

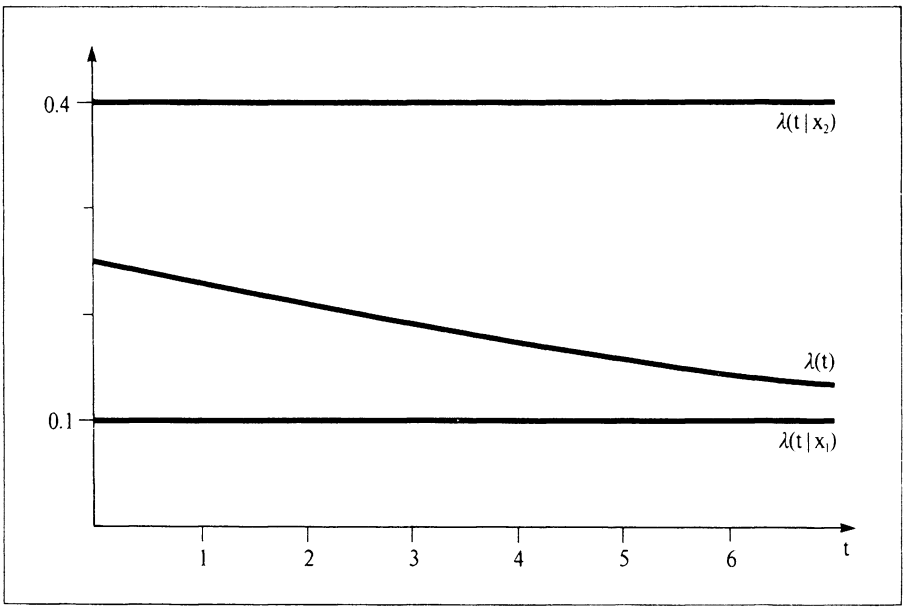
$$\lambda(t|x_1) = 0.1, \qquad \lambda(t|x_2) = 0.4,$$

und die Wahrscheinlichkeiten, daß ein Individuum der ersten beziehungsweise zweiten Teilgesamtheit entstammt, seien p_1 und p_2. Wird die Unterteilung in die beiden Teilgesamtheiten nicht berücksichtigt, erhält man für die Übergangsrate in der Gesamtpopulation

$$\lambda(t) = \frac{\sum_i f(t|x_i)\, p_i}{\sum_i S(t|x_i)\, p_i} \tag{3.9.1}$$

wobei $S(t|x_i)$ die Survivorfunktion in der Teilgesamtheit x_i bezeichnet. Für die oben angegebenen Werte von $\lambda(t|x_1)$ und $\lambda(t|x_2)$ sowie $p_1 = p_2 = 0.5$ ist der Verlauf der Hazardraten in der Abbildung 3.12 wiedergegeben.
Die Nichtberücksichtigung der Populationsheterogenität führt zu einer abnehmenden Hazardrate. Dies läßt sich auf heuristischem Weg plausibel machen.

Abbildung 3.12: Verlauf der Hazardraten in den Teilpopulationen und in der Gesamtpopulation bei $\lambda(t|x_1) = 0{,}1$ und $\lambda(t|x_2) = 0{,}4$



Zuerst werden im Durchschnitt diejenigen Individuen ein Ereignis haben, deren Übergangsrate hoch ist, und in der Risikomenge verbleiben tendenziell diejenigen mit der niedrigen Übergangsrate. Dadurch kommt insgesamt eine in Abhängigkeit von der Zeit abnehmende Hazardrate zustande.

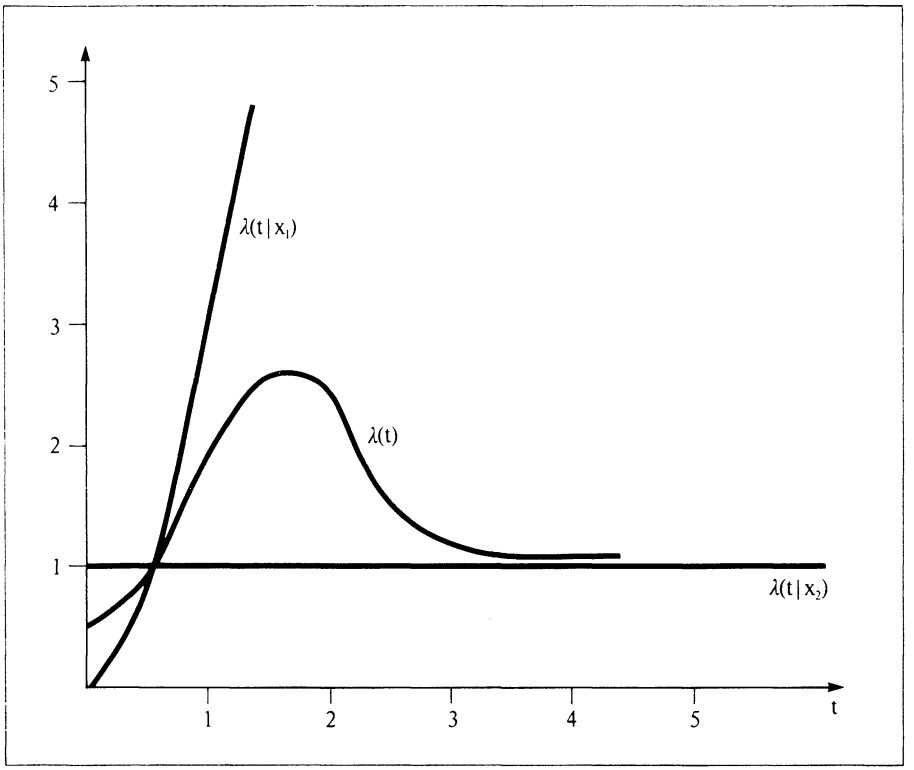
Beispiel 2

Die Ausgangssituation sei wie in Beispiel 1. Für die Hazardraten in den beiden Teilgesamtheiten gelte

$$\lambda(t|x_1) = 3t^2 \qquad \text{und } \lambda(t|x_2) = 1.$$

$\lambda(t|x_1)$ ist eine Weibull-Hazardrate mit $\alpha = 3$ und $\lambda = 1$. Außerdem seien $p_1 = p_2 = 0.5$. Die Hazardraten sind in der Abbildung 3.13 wiedergegeben.
Während die Hazardrate $\lambda(t|x_1)$ in der ersten Teilgesamtheit rasch ansteigt, ist die Hazardrate $\lambda(t|x_2)$ in der zweiten Teilgesamtheit konstant gleich 1. Für die Hazardrate $\lambda(t)$ der Gesamtpopulation ergibt sich ein völlig anderes Bild. Sie steigt zwar zunächst an, nimmt aber ab einem relativ frühen Zeitpunkt kontinuierlich ab und nähert sich dem Wert 1.
Die in den beiden Beispielen aufgezeigte Tendenz gilt allgemein. Wird unbeobachtete Populationsheterogenität nicht berücksichtigt, das heißt, wird bei der Bildung der Hazardrate über die nicht beobachteten Merkmale aggregiert, so

Abbildung 3.13: Verlauf der Hazardrate in den Teilpopulationen und in der Gesamtpopulation bei $\lambda(t|x_1) = 3t^2$ und $\lambda(t|x_2) = 1$



bewirkt dies eine tendenzielle Änderung der Hazardrate in Richtung negativer Zeitabhängigkeit, das heißt abnehmender Hazardrate. Einen Beweis findet man zum Beispiel bei Heckman/Singer (1984a). Man beachte, daß es sich hierbei nur um eine Tendenzaussage handelt. Dies bedeutet nicht, daß die aggregierte Hazardrate stets monoton fallend sein muß. Man betrachte dazu folgendes Beispiel:

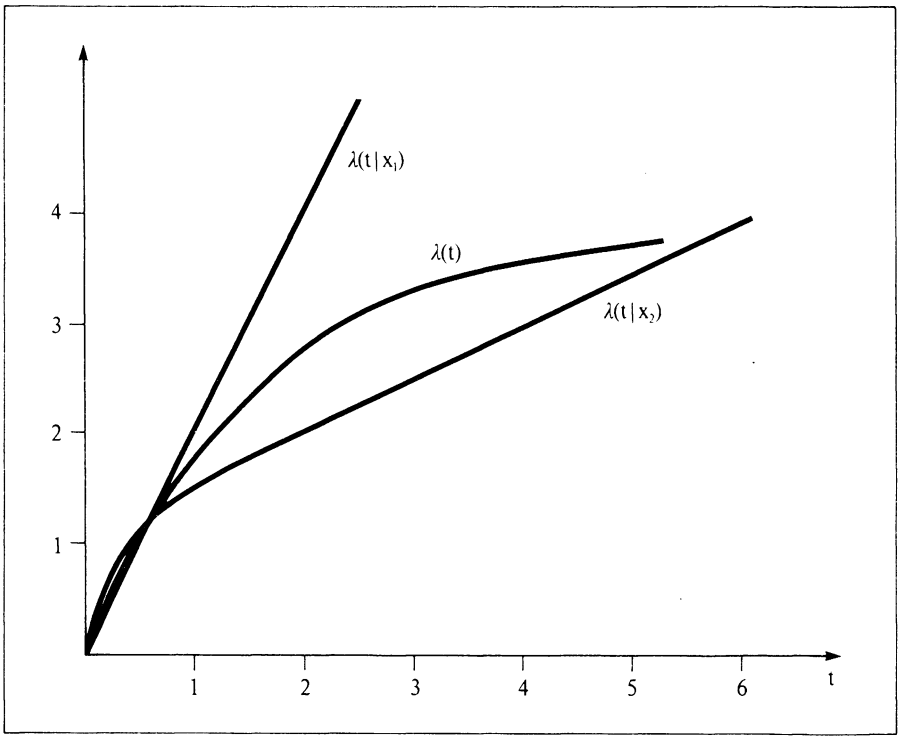
Beispiel 3

Die Ausgangssituation sei wie in den Beispielen 1 und 2. Für die Hazardraten in den beiden Teilgesamtheiten gelte

$$\lambda(t|x_1) = 2t \qquad \text{und } \lambda(t|x_2) = 3t^2.$$

Es handelt sich dabei um Weibull-Hazardraten mit $\lambda_1 = 1$, $\alpha_1 = 2$ beziehungsweise $\lambda_2 = 1$, $\alpha_2 = 3$. Die Hazardraten sind in der Abbildung 3.14 dargestellt ($p_1 = p_2 = 0.5$).

Abbildung 3.14: Verlauf der Hazardraten in den Teilpopulationen und in der Gesamtpopulation bei $\lambda(t|x_1) = 2t$ und $\lambda(t|x_2) = 3t^2$



Die oben getroffene Tendenzaussage bedeutet, daß sich mit zunehmender Zeit t die Steigerungsraten der Hazardrate ohne Berücksichtigung der unbeobachteten Merkmale verringert.

Modelle mit unbeobachteter Populationsheterogenität wurden bisher fast ausschließlich für den Ein-Episoden-Fall behandelt. Ausnahmen bilden Flinn/Heckman (1982) und Newman/McCulloch (1984). Insbesondere vom theoretischen Standpunkt aus sind die Modelle noch nicht zureichend untersucht. Im folgenden werden Möglichkeiten vorgestellt, unbeobachtete Populationsheterogenität explizit in den Modellansatz aufzunehmen.

Die unbeobachtete Populationsheterogenität wird durch eine (reellwertige) Zufallsvariable ϵ repräsentiert, deren Realisation die Hazardrate als „Abweichung" beeinflußt. Es liegt nahe, diese Abweichung als multiplikativen Term in die Hazardrate einzuführen, etwa in der Form

$$\lambda(t|x, \epsilon) = \lambda(t|x) \cdot \epsilon. \qquad (3.9.2)$$

Da die Hazardrate nicht negativ ist, muß ϵ auf positive Werte beschränkt sein. In der Regel nimmt man an, daß

$$E(\epsilon) = 1 \tag{3.9.3}$$

gilt, das heißt, im Durchschnitt ergibt sich $\lambda(t|\mathbf{x})$.
Die „Abweichung" ϵ variiert von Individuum zu Individuum. Sie ist nicht direkt beobachtbar. Die Vorgehensweise entspricht der Einbeziehung eines Personenparameters im Random-Effect-Modell der Varianzanalyse.
Die Verteilung von ϵ in der Population wird mit $G(\epsilon)$ bezeichnet. Die Modelle (3.9.2) werden im Bereich der Demographie und Biostatistik auch als „Frailty"-Modelle bezeichnet (vgl. Vaupel u. a. 1979). In der Event-History-Analyse wurden sie von Tuma (1978) eingeführt für den Fall einer auf der individuellen Ebene zeitunabhängigen Hazardrate $\lambda(t|\mathbf{x})$ in (3.9.2). Zur Einbeziehung unbeobachteter Populationsheterogenität siehe auch Heckman/Singer (1982, 1984a, 1984b), Flinn/Heckman (1982), Elbers/Ridder (1982).

3.9.2 Modelle und Parameterschätzung bei gegebener Verteilung der Heterogenitätskomponente

Ist die Hazardrate bei gegebenem Kovariablenvektor und gegebener Heterogenitätskomponente $\lambda(t|\mathbf{x}, \epsilon)$, so gilt für die Rand-Dichte von T bei gegebenem Kovariablenvektor

$$f(t|\mathbf{x}) = \int_0^\infty f(t|\mathbf{x}, \epsilon)\, dG(\epsilon) = \int_0^\infty \lambda(t|\mathbf{x}, \epsilon)\, S(t|\mathbf{x}, \epsilon)\, dG(\epsilon). \tag{3.9.4}$$

Verteilungen wie $f(t|\mathbf{x})$ werden als *Mischverteilungen* bezeichnet. $G(\epsilon)$ repräsentiert dabei die *mischende* Verteilung.
(3.9.4) ist die Verteilung der beobachtbaren Größen des Modells und erfordert das „Hinausintegrieren" der Abweichung ϵ. Dazu muß die Verteilung $G(\epsilon)$ voll spezifiziert werden. Wird für $G(\epsilon)$ eine parametrische Verteilung vorgegeben und wird $\lambda(t|\mathbf{x})$ spezifiziert, so lassen sich die unbekannten Parameter von $G(\epsilon)$ zusammen mit den Parametern von $\lambda(t|\mathbf{x})$ mit Hilfe der Maximum-Likelihood-Methode auf der Basis von (3.9.4) schätzen. Bezeichnet δ_i den Zensierungsindikator, so erhält man eine „marginale" Log-Likelihood-Funktion durch (vgl. Abschnitte 3.6.2 und 3.6.3)

$$\ln L_M = \sum_{i=1}^{n} \ln \int_0^\infty \lambda(t_i|\mathbf{x}_i, \epsilon)^{\delta_i}\, S(t_i|\mathbf{x}_i, \epsilon)\, dG(\epsilon). \tag{3.9.5}$$

Die Maximierung von (3.9.5) kann erhebliche numerische Probleme verursachen, da in jedem Schritt eine Integration bezüglich der Verteilung von ϵ notwendig ist.
Wird die Heterogenitätskomponente als stetig vorausgesetzt mit der Dichtefunktion $g(\epsilon)$, ergibt sich

$$\ln L_M = \sum_{i=1}^{n} \ln \int_0^\infty \lambda(t_i|\mathbf{x}_i, \epsilon)^{\delta_i}\, S(t_i|\mathbf{x}_i, \epsilon)\, g(\epsilon)\, d\epsilon. \tag{3.9.6}$$

Im folgenden wird die Vorgehensweise anhand einer einfachen Spezifikation von $\lambda(t|x)$ verdeutlicht, die von Tuma (1978) und Tuma/Hannan (1984, Kap. 6.3), gewählt wurde. Die Hazardrate wird als zeitunabhängig vorausgesetzt mit

$$\lambda(t|x) = \phi(x'\beta)$$

mit einer (deterministischen) nicht-negativen Funktion ϕ. Als Verteilung der Heterogenitätskomponente wird die Gamma-Verteilung angenommen mit der Dichte

$$g(\epsilon) = \frac{\theta}{\Gamma(\alpha)} (\theta\,\epsilon)^{\alpha-1} \exp(-\theta\,\epsilon), \quad \epsilon \geq 0,\ \alpha, \theta > 0, \tag{3.9.7}$$

wobei $\Gamma(\alpha)$ die Gamma-Funktion

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x}\, dx$$

bezeichnet. Erwartungswert und Varianz einer gammaverteilten Zufallsvariablen sind

$$E(\epsilon) = \frac{\alpha}{\theta} \text{ und } \mathrm{Var}(\epsilon) = \frac{\alpha}{\theta^2}.$$

Damit entsprechend der Annahme (3.9.3) $E(\epsilon) = 1$ gilt, muß $\alpha = \theta$ sein. Daraus folgt für die Dichte der Heterogenitätskomponente

$$g(\epsilon) = \frac{\alpha}{\Gamma(\alpha)} (\alpha\,\epsilon)^{\alpha-1} \exp(-\alpha\,\epsilon). \tag{3.9.8}$$

Gemäß der Annahme $\lambda(t|x) = \phi(x'\beta)$ ist die Verteildauer T exponentialverteilt, und für die Dichte von T bei gegebener Heterogenitätskomponente und gegebenem Kovariablenvektor erhält man

$$f(t|x, \epsilon) = \phi(x'\beta)\,\epsilon \exp(-\phi(x'\beta)\,\epsilon\, t),$$

und aus (3.9.4) folgt

$$f(t|x) = \int_0^\infty \phi(x'\beta)\,\epsilon \exp(-\phi(x'\beta)\,\epsilon\, t) \quad \frac{\alpha}{\Gamma(\alpha)} (\alpha\,\epsilon)^{\alpha-1} \exp(-\alpha\,\epsilon)\, d\epsilon. \tag{3.9.9}$$

Mit geeigneter Substitution und Verwendung der Beziehung $\Gamma(\alpha + 1) = \alpha\,\Gamma(\alpha)$ läßt sich (3.9.9) integrieren. Es resultiert

$$f(t|x) = \frac{\phi(x'\beta)\,\alpha^{\alpha+1}}{(\phi(x'\beta)\,t + \alpha)^{\alpha+1}}. \tag{3.9.10}$$

Führt man in

$$S(t|x) = 1 - \int_0^t f(u|x)\, du$$

die Integration auf der rechten Seite durch, ergibt sich für die Survivorfunktion

$$S(t|x) = \left(\frac{\alpha}{\phi(x'\beta)\, t + \alpha)}\right)^{\alpha} \tag{3.9.11}$$

und wegen $\lambda(t|x) = f(t|x) / S(t|x)$ erhält man für die Hazardrate

$$\lambda(t|x) = \frac{\phi(x'\beta)\,\alpha}{\phi(x'\beta)\, t + \alpha}. \tag{3.9.12}$$

Man beachte, daß die Hazardrate in (3.9.12) sehr wohl von der Zeit t abhängt, obwohl die individuellen Hazardraten $\lambda(t|x, \epsilon) = \phi(x'\beta) \cdot \epsilon$ als zeitunabhängig vorausgesetzt wurden. Dies ist wieder eine Folge des eingangs beschriebenen Resultats, daß unbeobachtete Heterogenität die Hazardrate in der Gesamtpopulation in Richtung auf negative Zeitabhängigkeit verschiebt.
Aus (3.9.5) beziehungsweise (3.9.6) wird ersichtlich, daß die Wahl von $G(\epsilon)$ die Form der marginalen Likelihood-Funktion und damit auch die Schätzung der strukturellen Modellparameter β beeinflussen kann. Heckman/Singer (1982) gelangen bei der Analyse eines empirischen Datensatzes für eine angenommene Weibull-Hazardrate bei verschiedenen $G(\epsilon)$ zu recht unterschiedlichen Resultaten für die Schätzung der β. Sie schlagen deshalb eine alternative Strategie vor, auf die wir im nächsten Abschnitt eingehen werden. Zu einem anderen Ergebnis kommen Newman/McCulloch (1984) bei einem Datensatz über die Zeitabschnitte zwischen aufeinanderfolgenden Geburten. Sie wählen als „mischende" Verteilung $G(\epsilon)$ verschiedene diskrete Approximationen der Gammaverteilung sowie der Lognormalverteilung und kamen zu dem Ergebnis, daß sich die Schätzungen für die β's nur wenig unterscheiden. Ein Grund für die unterschiedlichen Resultate des von Heckman/Singer analysierten Datensatzes könnte auch darin liegen, daß das Weibull-Modell den Daten nicht angepaßt ist.
Für mehrere aufeinanderfolgende Zeitperioden ist der Sachverhalt komplizierter. Wir wollen hier nur den Repeated-Event-Fall kurz skizzieren. Unterstellt man für jede Episode eine andere Heterogenitätskomponente ϵ_k, so ist eine Annahme über die gemeinsame Verteilung der ϵ_k notwendig, da die einzelnen Komponenten im allgemeinen nicht unabhängig sind. Die Rechtfertigung einer bestimmten Wahl dieser gemeinsamen Verteilung ist – wie schon die Wahl von $G(\epsilon)$ – aus dem Sachzusammenhang heraus in der Regel schwierig. Gelegentlich wird die vereinfachende Annahme

$$\epsilon_k = \gamma_k\, \epsilon$$

getroffen, wobei γ_k ein episodenspezifischer Parameter ist.
Werden die Hazardraten $\lambda^k(t|x_k, \gamma_k, \epsilon)$ sowie die Verteilung $G(\epsilon)$ spezifiziert, so können aus den in Abschnitt 3.6.6 beschriebenen Likelihood-Funktionen durch Integration über ϵ wieder „marginale" Likelihood-Funktionen konstruiert werden, die man zur Schätzung der unbekannten Modellparameter heranziehen kann.
An dieser Stelle scheint eine kritische Anmerkung angebracht. Gewöhnlich wird die Annahme getroffen, daß die Heterogenitätskomponente unabhängig ist von

den beobachteten Kovariablen. Dies ist insbesondere in sämtlichen Beiträgen, die im folgenden Abschnitt zitiert werden, der Fall. Auf der anderen Seite wird die Einbeziehung unbeobachtbarer Populationsheterogenität in der Regel durch den Einwand motiviert, man könne in empirischen Anwendungen niemals alle relevanten Einflußgrößen erheben und man habe bei Nichtberücksichtigung unbeobachteter Merkmale mit einer Verzerrung der Resultate zu rechnen (omitted variables bias). Mit Sicherheit werden jedoch die unbeobachteten Merkmale bei einem Individuum nicht unabhängig sein von den erhobenen Merkmalen. Unterstellt man Unabhängigkeit, so wird das Problem nicht beobachteter Einflußgrößen hinausdefiniert, und der omitted variables bias ist nicht beseitigt. Man vergleiche dazu auch Chamberlain (1980).
Soll die mögliche Abhängigkeit der Heterogenitätskomponente und der Kovariablen berücksichtigt werden, müßte man nicht eine Verteilung $G(\epsilon)$, sondern eine Verteilung $G(\epsilon|\mathbf{x})$ spezifizieren. Dies ist im praktischen Anwendungsfall schwierig. Eine Möglichkeit besteht darin, einen Regressionsansatz

$$\epsilon = \mathbf{x}'\pi + u$$

zu formulieren und in das Modell aufzunehmen. Der Parametervektor π ist mitzuschätzen. Allerdings wird dadurch die Zahl der unbekannten Parameter stark erhöht, und darüber hinaus können Identifikationsprobleme auftreten.

3.9.3 Simultane Schätzung der strukturellen Modellparameter und der Verteilung der Heterogenitätskomponente

Heckman/Singer (1982, 1984a, 1984b) schlagen aufgrund der möglichen Sensitivität der Schätzungen der strukturellen Modellparameter gegenüber der Wahl der Verteilung der unbeobachteten Heterogenität eine simultane Schätzung der Modellparameter und der Verteilung der Heterogenitätskomponente vor, ähnlich der empirischen Bayes-Schätzung (vgl. z. B. Maritz 1971).
Im folgenden betreffen alle Aussagen den Ein-Episoden-Fall, und der Kovariablenvektor wird als zeitunabhängig vorausgesetzt. Sei die bedingte Verteilung der Verweildauer, gegeben die Kovariablen und die Heterogenitätskomponente, mit $f(t|\mathbf{x}, \epsilon; \theta)$ bezeichnet, wobei die Abhängigkeit vom Vektor θ der strukturellen Parameter deutlich gemacht wird. θ enthält die Regressionskoeffizienten sowie weitere die Hazardrate und damit auch $f(\cdot)$ determinierende Parameter. Die Verteilung der Heterogenität ϵ, die mischende Verteilung, sei $G(\epsilon)$, und daraus ergibt sich für die Dichtefunktion von T gegeben $\mathbf{x}$

$$h(t|\mathbf{x}; \theta) = \int f(t|\mathbf{x}, \epsilon; \theta)\, dG(\epsilon). \tag{3.9.13}$$

Das Ziel besteht darin, neben den Parameterschätzungen $\hat{\theta}$ auch eine Schätzung $\hat{G}(\epsilon)$ zu finden, die wünschenswerte Eigenschaften besitzt, also mit zunehmendem Stichprobenumfang zumindest nach Wahrscheinlichkeit gegen $G(\epsilon)$ konvergiert.

Bevor dieses Ziel näher untersucht werden kann, sind zuerst verschiedene Identifikationsprobleme zu lösen. Läßt man das Schätzproblem gänzlich außer acht, so stellt sich die Frage, ob die Kenntnis der Mischverteilung $h(t|x;\theta)$ ausreicht, damit die Integralgleichung (3.9.13) mit eindeutig bestimmten Funktionen $f(t|x,\epsilon;\theta)$ und $G(\epsilon)$ erfüllt ist. Ohne weitere Zusatzannahmen ist dies sicher nicht gewährleistet. Auch bei Spezifikationen der bedingten Verteilung $f(t|x,\epsilon;\theta)$ ist nicht in jedem Fall gesichert, daß nicht zwei verschiedene mischende Verteilungen $G_1(\epsilon)$ und $G_2(\epsilon)$ dieselbe Mischverteilung ergeben. Für Beispiele vergleiche man Hamerle/Pape (1977) oder Heckman/Singer (1984b).
Für stetige Verweildauermodelle (bei einer Verweildauer und einem Endzustand) haben Elbers/Ridder (1982) für die Klasse der Proportional-Hazards-Modelle mit

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

die Identifizierbarkeit gezeigt. Eine wichtige Bedingung ist dabei, daß der Kovariablenvektor mindestens eine stetige Komponente enthält. Man vergleiche dazu auch Hougaard (1984) und Heckman/Singer (1984a, 1984b). Eine weitere zentrale Forderung, die in allen Beiträgen enthalten ist, betrifft die Unabhängigkeit der Heterogenitätskomponente von den Kovariablen. Wie im letzten Abschnitt ausgeführt wurde, ist damit das „omitted-variables"-Problem im allgemeinen nicht gelöst.
Kiefer/Wolfowitz (1956) geben allgemeine Bedingungen für die Existenz eines konsistenten Schätzers der mischenden Verteilung und der strukturellen Modellparameter an. Ihr Aufsatz enthält aber keinen Hinweis auf eine konstruktive Vorgehensweise bei der numerischen Ermittlung der ML-Schätzungen. Heckman/Singer (1984b) verifizieren die Kiefer/Wolfowitz-Bedingungen für Proportional-Hazards-Modelle mit zeitabhängigen Kovariablen und möglicherweise zensierten Daten. Darüber hinaus schlagen sie einen nicht parametrischen Maximum-Likelihood-Schätzalgorithmus vor, der auf der theoretischen Charakterisierung der ML-Schätzung bei Mischverteilungen von Lindsay (1983a, 1983b) beruht. Sie propagieren dazu die Verwendung des EM-Algorithmus (vgl. Dempster u.a. 1977). Zur Anwendung des EM-Algorithmus in diesem Zusammenhang siehe auch Arminger (1984a).

3.10 Diskrete Hazardraten-Regressionsmodelle

In den vorangegangenen Abschnitten wurde davon ausgegangen, daß die Zeitpunkte, zu denen Zustandswechsel beziehungsweise Ereignisse stattfinden, exakt angegeben werden können. In vielen Fällen ist jedoch die exakte Angabe der Zeitpunkte von Zustandswechseln nicht möglich, sondern es können lediglich Zeitintervalle angegeben werden, in denen Zustandswechsel aufgetreten oder bestimmte Ereignisse eingetreten sind. Legt man dennoch ein zeitstetiges Modell zugrunde, ist die Zahl gleicher Beobachtungswerte (Ties) bei den gemessenen

Verweildauern sehr hoch. Man erhält dann bei vielen Modellen (z. B. im Cox-Modell) unbrauchbare Parameterschätzungen. Darüber hinaus ist in allen theoretischen Ableitungen, etwa zum Beweis der asymptotischen Eigenschaften der Parameterschätzungen, die Annahme enthalten, daß gleiche Meßwerte der Verweildauern in der Stichprobe die Wahrscheinlichkeit 0 besitzen. Diese Annahme ist in der eben geschilderten Situation verletzt, so daß dem zugrundegelegten stetigen Modell die theoretische Fundierung fehlt. Deshalb ist es zweckmäßig, in derartigen Situationen von vorneherein ein zeitdiskretes Modell zu verwenden, das der Datenerhebung besser angepaßt ist. Damit teilen wir nicht die Auffassung, daß zeitstetige Modelle stets zeitdiskreten Modellen vorzuziehen sind, auch wenn sie der Datenerhebung nicht angemessen sind (vgl. auch Allison 1982).

Wir werden in diesem Abschnitt nur eine kurze Einführung in die Modelle für diskrete Hazardraten geben. Es werden auch in den Kapiteln 4 bis 6 keine Anwendungsbeispiele dargestellt, da bei der Lebensverlaufsstudie die Zeitpunkte der Zustandswechsel mit ausreichender Genauigkeit gemessen wurden. Wir beschränken uns auf die wichtigsten statistischen Konzepte bei einer Episode und ohne Competing Risks. Für diskrete Competing-Risks-Regressionsansätze siehe Hamerle (1985c). Eine ausführliche Behandlung diskreter Hazardraten-Modelle im Ein- und Mehr-Episoden-Fall findet man in Hamerle/Tutz (1986), an deren Darstellung wir uns hier anlehnen.

Die Zeitachse wird zerlegt in $q + 1$ Intervalle

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

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

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

Neben der Verweildauer beziehungsweise Lebenszeit wird für jedes Individuum beziehungsweise Objekt ein p-dimensionaler Vektor **x** von Kovariablen beziehungsweise prognostischen Faktoren erhoben. Die Kovariablen werden hier als zeitunabhängig vorausgesetzt. Die Einbeziehung von externen und internen zeitabhängigen Kovariablen ist möglich. Man vergleiche Hamerle/Tutz (1986).

In Analogie zu den Abschnitten 3.2 und 3.3 können auch im diskreten Fall Hazardrate und Survivorfunktion definiert werden. Die Hazardrate ist gegeben durch

$$\lambda(t|x) = P(T = t|T \geq t, x). \tag{3.10.1}$$

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

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

$$P(T > t|T \geq t, x) = 1 - \lambda(t|x). \tag{3.10.2}$$

Die Survivorfunktion ist

$$S(t|x) = P(T \geq t|x), \tag{3.10.3}$$

die (unbedingte) Wahrscheinlichkeit, das Zeitintervall t zu „erleben", das heißt, daß bis zum Beginn dieses Intervalls noch kein Ereignis stattgefunden hat. Den Zusammenhang zwischen Survivorfunktion und Hazardrate erhält man durch sukzessive Anwendung von

$$P(T \geq s|x) = P(T \geq s|T \geq s-1, x) \cdot P(T \geq s-1|x)$$

und mit (3.10.2). Es ergibt sich

$$S(t|x) = \prod_{s=1}^{t-1} (1 - \lambda(s|x)). \tag{3.10.4}$$

Schließlich erhält man für die unbedingte Ereigniswahrscheinlichkeit beziehungsweise die Wahrscheinlichkeit, den Endzustand im Zeitintervall t zu erreichen, gegeben die Kovariablen

$$P(T = t|x) = P(T = t|T \geq t, x) \cdot P(T \geq t|x) = \lambda(t|x) \prod_{s=1}^{t-1} (1 - \lambda(s|x)). \tag{3.10.5}$$

Mit (3.10.5) läßt sich auch $P(T = t|x)$ durch die Hazardrate ausdrücken, und es ist wie bei den Modellen mit stetig gemessener Zeit zweckmäßig, die Hazardrate in Abhängigkeit von den Kovariablen zu parametrisieren. Dafür gibt es verschiedene Möglichkeiten. Die wichtigsten Modelle sind in Hamerle/Tutz (1986) ausführlich beschrieben. Hier werden nur zwei Spezifikationen skizziert, das logistische Modell und das gruppierte Cox-Modell.
Die Hazardrate des logistischen Modells ist

$$\lambda(t|x) = P(T = t|T \geq t, x) = \frac{\exp(\beta_{0t} + x'\beta)}{1 + \exp(\beta_{0t} + x'\beta)} \quad t = 1, \ldots, q. \tag{3.10.6}$$

Eine äquivalente Formulierung des Modells erhält man durch

$$\ln \frac{P(T = t|x)}{P(T > t|x)} = \beta_{0t} + x'\beta. \tag{3.10.7}$$

Die Parameter $\beta_{01}, \ldots, \beta_{0q}$ repräsentieren wie im Cox-Modell bei stetig gemessener Zeit eine „Grundhazardrate" ohne Berücksichtigung der Kovariablen, die allen Individuen gemeinsam ist.
Die Hazardrate des gruppierten Cox-Modells ist

$$\lambda(t|x) = 1 - \exp(-\exp(\beta_{0t} + x'\beta)). \tag{3.10.8}$$

Modell (3.10.8) läßt sich als Cox-Modell bei diskreten Beobachtungen auffassen. Gilt für die zugrundeliegende stetige Verweildauer das Cox-Modell und wird die Dauer aber nur diskret beobachtet, so gilt für die diskreten Beobach-

tungen Modell (3.10.8). Der Parametervektor β, der den Einfluß der Kovariablen steuert, bleibt bei der Diskretisierung unverändert. Der Vektor β in (3.10.8) ist identisch mit dem entsprechenden Gewichtsvektor des stetigen Cox-Modells.

Maximum-Likelihood-Schätzung

Auch für die diskreten Modelle läßt sich in Analogie zu den Ausführungen in Abschnitt 3.6.2 der Beitrag des Individuums i zur Likelihood-Funktion ableiten. Es ergibt sich

$$L_i = P(T_i = t_i|\mathbf{x}_i)^{\delta_i} P(T_i \geq t_i|\mathbf{x}_i)^{1-\delta_i} = \lambda(t_i|\mathbf{x}_i)^{\delta_i} P(T_i \geq t_i|\mathbf{x}_i), \quad (3.10.9)$$

wobei δ_i wieder den Zensierungsindikator bezeichnet. Mit (3.10.4) resultiert

$$L_i = \lambda(t_i|\mathbf{x}_i)^{\delta_i} \prod_{s=1}^{t_i-1} (1-\lambda(s|\mathbf{x}_i)). \quad (3.10.10)$$

Die meisten der diskreten Verweildauer-Modelle lassen sich auf der Basis von (3.10.9) im Rahmen der verallgemeinerten linearen Modelle schätzen. Verallgemeinerte lineare Modelle sind in Fahrmeir/Hamerle (1984, Kap. 7) oder McCullagh/Nelder (1983) ausführlich dargestellt. Man betrachte t_i unabhängige dichotome Zufallsvariablen $Y_{i1}, \ldots, Y_{it_i}$. Dann gilt für die zugehörige Likelihood-Funktion

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

Falls die Y_{ir} bedingte Zufallsvariablen in Abhängigkeit von Einflußvariablen $\mathbf{x}_i$ sind, gilt

$$L_i = \prod_{r=1}^{t_i} P(Y_{ir} = 1|\mathbf{x}_i)^{y_{ir}} (1 - P(Y_{ir} = 1|\mathbf{x}_i))^{1-y_{ir}}. \quad (3.10.11)$$

Dies entspricht der Likelihood von t_i Beobachtungen eines verallgemeinerten linearen Modells, wobei $P(Y_{ir} = 1|\mathbf{x}_i)$ mit Hilfe einer Responsefunktion, nämlich $\lambda(r|\mathbf{x}_i)$, in Abhängigkeit vom Parameter β spezifiziert wird. Setzt man für den beobachteten Vektor $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{it_i}) = (0, \ldots, 0, 1)$, erhält man aus (3.10.11) den Beitrag der Likelihood-Funktion der i-ten Person im Verweildauer-Modell für den Fall $\delta_i = 1$. Betrachtet man nur $t_i - 1$ Zufallsvariablen, erhält man mit

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

den Beitrag (3.10.10) zur Likelihood-Funktion für $\delta_i = 0$, wobei wiederum $P(Y_{ir} = 1|\mathbf{x}_i) = \lambda(r|\mathbf{x}_i)$ gesetzt wird und der Beobachtungsvektor $\mathbf{Y}_i = (y_{i1}, \ldots, y_{i,t_i-1}) = (0, \ldots, 0)$ vorliegt.

Die Möglichkeit der Schätzung der Verweildauer-Modelle im Rahmen der verallgemeinerten linearen Modelle bietet den Vorteil, daß die numerische Auswertung mit bereits vorhandenen Programmpaketen, wie etwa BMDP (für das

logistische Modell), GLIM oder GLAMOUR (nähere Informationen über GLAMOUR sind vom Lehrstuhl für Statistik der Universität Regensburg erhältlich), durchgeführt werden kann.
Die diskreten Modelle können dahingehend erweitert werden, daß sowohl externe als auch interne zeitabhängige Kovariablen einbezogen werden. Bei internen zeitabhängigen Kovariablen muß allerdings die Likelihood-Funktion auf anderem Wege abgeleitet werden. Man vergleiche dazu Hamerle/Tutz (1986). Zur Berücksichtigung von Linkszensierungen bei diskreten Modellen siehe Hamerle (1986a). Verallgemeinerungen der Modelle unter Einbeziehung unbeobachteter Populationsheterogenität sind ebenfalls möglich. Man vergleiche dazu Hamerle (1986b).

Kapitel 4:
Datenorganisation und beschreibende Verfahren

Das Hauptziel der Datenanalyse wird es in der Regel sein, ein sparsames statistisches Modell zu finden, das der Realität möglichst nahe kommt und sich inhaltlich gut interpretieren läßt. Allerdings kennt man meist zu Beginn des Analyseprozesses das zu untersuchende Datenmaterial nicht genau, so daß man in einem ersten Schritt versuchen wird, mit Hilfe von deskriptiven und graphischen Methoden einen Überblick über die Verteilungen der erhobenen Variablen zu gewinnen. Weitere Gruppenvergleiche und Tests können dann in einem zweiten Schritt zusätzliche Hinweise auf Zusammenhänge zwischen den Variablen geben, die schließlich in einem dritten Schritt bei Kontrolle anderer Einflußfaktoren in einem angemessenen Modell zur Erklärung herangezogen werden können.
In diesem Kapitel werden zunächst die programmtechnische Aufbereitung ereignisorientierter Datenstrukturen (Abschnitt 4.1) sowie verschiedene Möglichkeiten ihrer graphischen Präsentation dargestellt (Abschnitt 4.2). Anschließend wird auf der Grundlage der Lebensverlaufsstudie exemplarisch die Anwendung der Sterbetafel-Methode und des Kaplan-Meier-Schätzers aufgezeigt (Abschnitt 4.3). Großes Gewicht wird dabei auf die Herausarbeitung der spezifischen Probleme und Vorzüge der Ereignisanalyse gelegt, die anhand anschaulicher Programmbeispiele, Tests und Ergebnisinterpretationen dokumentiert werden sollen.

4.1 Die Handhabung ereignisorientierter Datenstrukturen

Erste, nicht zu unterschätzende Probleme im Analyseprozeß treten bereits bei der Frage nach der geeigneten Datenspeicherung und Aufbereitung von Ereignisdaten für statistische Auswertungen auf. *Ereignisorientierte Datenstrukturen sind* zunächst *weit komplexer* als Querschnittsdaten, weil mit den jeweiligen *Zustandsinformationen* gleichzeitig auch die genauen *Anfangs- und Endzeitpunkte* zu berücksichtigen sind. Häufig treten in den Wirtschafts- und Sozialwissenschaften darüber hinaus *wiederholbare Ereignisse* auf, *deren Zahl über die*

Untersuchungseinheiten erheblich schwankt und zur Frage nach geeigneten Speicherverfahren führt. Werden beispielsweise wie in der Lebensverlaufsstudie die Erwerbsgeschichten von 3 Geburtskohorten kontinuierlich erhoben, so gibt es zwischen den Personen große Unterschiede in der Anzahl der Berufsepisoden, die von 0 (nie erwerbstätig gewesen) bis zu 19 reicht. Speichert man diese Daten für jedes Individuum zeilenweise in Rechtecksform ab, so entsteht eine Matrix mit vielen leeren Speicherzellen, und der verfügbare Speicherplatz wird unökonomisch genutzt. Ein *geringfügiger Wechsel der Fragestellung*, der zu einer *Neudefinition des Zustandsraums* führt, erfordert schließlich meist eine grundlegende Reorganisation der Datendatei, weil bei den Programmpaketen wie SPSS, BMDP, RATE, SAS oder GLIM immer nur bestimmte Episoden die Analyseeinheit darstellen.

In der Praxis werden diese Probleme des Datenmanagements bei Ereignisdaten am besten mit Hilfe eines *Datenbanksystems* gelöst, welches die Daten in hierarchischer und ökonomischer Weise abspeichert und ein flexibles Retrieval erlaubt. Ein solches Datenbanksystem, mit dem beispielsweise die Lebensverlaufsdaten *personenbezogen abgespeichert* wurden, stellt das Programmpaket SIR (Robinson u.a. 1980) dar. Aus einer solchen Datenbank können dann je nach aktueller Fragestellung beliebige *ereignisorientierte Datensätze* erzeugt werden, die mit den Programmpaketen SPSS, BMDP, RATE, SAS oder GLIM analysierbar sind.

Von einem ereignisorientierten Datensatz spricht man dann, wenn sich *jeder Satz einer Datei genau auf ein Ereignis oder eine Episode bezieht*. Wurde bei jedem Untersuchungsobjekt jeweils nur eine Episode beobachtet, dann stimmt die Anzahl der Sätze in der Datei genau mit der Anzahl der Untersuchungseinheiten überein. Handelt es sich aber um wiederholbare Ereignisse (z. B. Berufstätigkeiten), deren Zahl bei jeder Person unterschiedlich hoch sein kann, so ergibt sich die Zahl der Sätze der Datei aus der Summe über diese personenspezifischen Episoden.

Jede Episode ist vollständig durch einen *Anfangs-* (TA) und einen *Endzeitpunkt* (TE) sowie durch einen *Anfangs-* (ZA) und einen *Endzustand* (ZE) charakterisiert. Häufig wird anstatt der Anfangs- und Endzeitpunkte auch nur die dadurch bestimmte *Zeitdauer* (TD) angegeben, welche die eigentliche abhängige Variable bei Ereignisanalysen darstellt. Haben Anfangs- und Endzustand denselben Wert, dann handelt es sich um eine *rechtszensierte Beobachtung*, weil die Episode dann nicht durch einen Zustandswechsel beendet, sondern, beispielsweise zum Zeitpunkt des retrospektiven Interviews, abgeschnitten wurde. Häufig wird zur Indizierung rechtszensierter Beobachtungen auch eine eigene *Zensierungs-(oder Indikator-)variable* (ZEN) eingeführt, die den Wert 1 annimmt, wenn die Episode durch einen Zustandswechsel regulär beendet wurde, und ansonsten den Wert 0 hat. Zu jeder der so beschriebenen Episoden können aus der Datenbank dann *Kovariablen* dazugespielt werden, die *zeitkonstant* (x) oder *zeitveränderlich* (x(t)) sein können und sich auf die *Vorgeschichte* oder auf einen *anderen parallelen Prozeß* beziehen können.

Tabelle 4.1 zeigt ein Beispiel für einen ereignisorientierten Datensatz zur Untersuchung von Erwerbsverläufen. Jeder Satz dieser Datei bezieht sich auf eine Erwerbstätigkeitsepisode, und die Erwerbstätigkeiten einer Person sind in sukzessiven Sätzen abgespeichert. Zu jeder Episode sind der genaue Beginn (TA) und das genaue Ende (TE) der Erwerbstätigkeit sowie die sich damit ergebende Erwerbstätigkeitsdauer (TD) gegeben. Um die Zeitangaben der drei Kohorten in der Lebensverlaufsstudie einfach vergleichen und verrechnen zu können, wurden sie nach der Anzahl der Monate vom Beginn dieses Jahrhunderts an vercodet. Die Zahl 590 entspricht beispielsweise dem 2. Monat (Februar) im Jahre 1949 ($49 \cdot 12 + 2 = 590$). Der Zustandsraum konstituiert sich in diesem Beispiel aus zwölf Berufen. Person 1 hatte danach zuerst im Beruf 1 ($ZA_1 = 1$) gearbeitet, bis sie in den Beruf 2 ($ZE_1 = 2$ und $ZA_2 = 2$) und schließlich in den Beruf 3 ($ZE_2 = 3$ und $ZA_3 = 3$) überwechselte. Dort war sie dann bis zum Zeitpunkt des Interviews tätig ($ZE_3 = 3$). Während die ersten beiden Episoden dieser ersten Person durch reguläre Ereignisse beendet wurden (ZEN = 1), wurde die 3. Episode durch das Interview abgeschnitten, also rechtszensiert (ZEN = 0). In den restlichen Spalten werden schließlich für jede der Episoden neben der Nummer der jeweiligen Erwerbstätigkeitsepisode eine Reihe von Kovariablen gespeichert. Beispiele für Variablen, die die Erwerbsgeschichte beeinflussen und über die Episode hinweg konstant bleiben, wären etwa Geschlecht, Herkunftsschicht und Kohortenzugehörigkeit. Beispiele für zeitveränderliche Variablen, die sich auf die Vorgeschichte des zu untersuchenden Prozesses selbst beziehen, sind die Zahl der Berufserfahrungsmonate zu Beginn jeder Episode oder die Anzahl der zuvor ausgeübten verschiedenen Berufstätigkeiten. Ein Beispiel für eine zeitveränderliche Variable, die sich auf einen parallelen Prozeß bezieht und die sich während einer Episode verändern kann, wäre etwa der Zeitpunkt der Heirat. Man könnte hier danach fragen, ob der Erwerbsverlauf nach der Heirat stabiler geworden ist oder nicht.

Die Berufswechsel können in dieser ereignisorientierten Form in Abhängigkeit von den Kovariablen mit den Programmpaketen SPSS, BMDP, RATE oder GLIM untersucht werden. Die Episoden (oder Untersuchungseinheiten) sind hier die einzelnen Berufstätigkeiten, egal, ob es sich beim Zustandswechsel um einen Aufstieg, einen Abstieg oder um eine horizontale Mobilität handelt. Wollte man aber beispielsweise untersuchen, wie lange es dauert, bis jemand in einen Beruf aufsteigt, der über dem höchsten bis dahin erreichten Niveau liegt (Sørensen 1984), dann müßte man den Zustandsraum und die Episodenlänge neu konstruieren. Alle Berufstätigkeitszeiten bis zu einem Aufstieg über das höchste bisher erreichte Niveau müßten in diesem Fall zu einer Episode zusammengefaßt werden. Ob man dabei etwaige Erwerbsunterbrechungszeiten mitberücksichtigt oder nicht, hängt von der Definition des Prozesses und vom theoretischen Interesse ab. Bei Personen ohne Aufstieg müßte die Episode zum Zeitpunkt des Interviews als zensiert betrachtet werden, da bisher noch kein Ereignis vorlag, aber die Person vielleicht nach dem Interview in der Lage ist, über das höchste bisher erreichte Niveau aufzusteigen. Für diese etwas modifizierte

Tabelle 4.1: Ausschnitt aus einem ereignisorientierten Datensatz zur Analyse der Erwerbsgeschichte

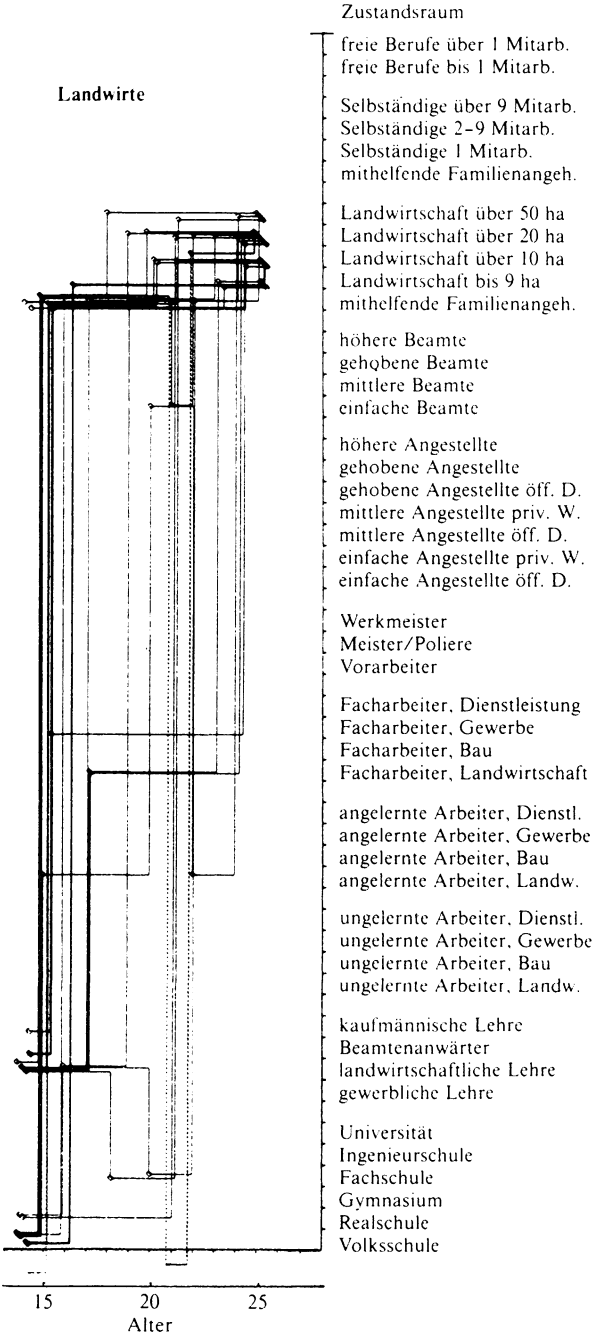
Nr. des Falls	Beginn der Episode (TA)	Ende der Episode (TE)	Zeitdauer der Episode (TD)	Zustand zu Beginn der Episode (ZA)	Zustand am Ende der Episode (ZE)	Zensierungsvariable (ZEN)	Nummer der Episode	Kovariablen, die über die Episode konstant bleiben							Kovariablen, die sich auf einen anderen parallelen Prozeß beziehen		
								Geschlecht	Herkunft	...	Kohorte	Berufserfahrung zu Beginn jeder Episode	...	Anzahl der vorher ausgeübten Berufe	Zeitpunkt der ersten Heirat	...	Zeitpunkt des Auszugs aus dem Elternhaus
								x_1	x_2	...	.	... .	...	x_m	$x_{m+1}(t)$	...	$x_p(t)$
1	590	626	36	1	2	1	1	1	2	...	1	0	...	0	660	...	665
1	626	698	72	2	3	1	2	1	2	...	1	36	...	1	660	...	665
1	698	981	313	3	3	0	3	1	2	...	1	108	...	2	660	...	665
2	610	680	70	10	11	1	1	2	1	...	1	0	...	0	705	...	700
.	.	.	.	.	.	.	.	.	.	...	.	.	...	.	.	...	...
.	.	.	.	.	.	.	.	.	.	...	.	.	...	.	.	...	...
.	.	.	.	.	.	.	.	.	.	...	.	.	...	.	.	...	...

Fragestellung wäre die ereignisorientierte Datei der Tabelle 4.1 allerdings nicht ohne weiteres verwendbar, da hier jeder Berufswechsel als Episode behandelt wurde. Deswegen müßte man aus der Datenbank (oder durch eigene Programmierung) einen neuen ereignisorientierten Datensatz mit anderer Struktur erzeugen.
In der Praxis hat sich bei der EDV-technischen Handhabung ereignisorientierter Datenstrukturen die Speicherung in Datenbanksystemen (z. B. SIR) oder Programmpaketen mit Datenbankeigenschaften (z. B. SAS) durchgesetzt, aus denen heraus für spezifische Fragestellungen selbst außerordentlich kompliziert aufgebaute ereignisorientierte Datensätze komfortabel erstellt werden können.

4.2 Die graphische Präsentation von Verläufen

Aufgrund der hohen Informationsdichte von Ereignisdaten müssen nicht nur andere Wege der Datenspeicherung und -aufbereitung beschritten werden als bei Querschnitten, sondern es stellt sich auch das Problem, wie man dieses zeitbezogene Datenmaterial möglichst übersichtlich und verständlich darstellt. Dabei sind insbesondere Methoden erforderlich, die dem in den Wirtschafts- und Sozialwissenschaften häufig auftretenden Mehr-Zustands- und Mehr-Episoden-Fall gerecht werden. In diesem Abschnitt sollen drei Möglichkeiten der graphischen Präsentation von Ereignisdaten vorgestellt werden: die Darstellung von individuellen Verläufen, von Ereignissequenzen und von Zustandsverteilungen.
Am umfassendsten ist die hohe Informationsdichte von Ereignisdaten in der *graphischen Darstellung individueller Verläufe repräsentiert.* Für jede Untersuchungseinheit wird der gesamte Verlauf in das Schaubild gezeichnet (Abbildung 4.1). Die x-Achse stellt dabei die historische Zeit oder das Lebensalter dar, und auf der y-Achse werden die diskreten Zustände, die die Untersuchungseinheit einnehmen kann, abgetragen. Das Ziel ist es, durch die Darstellung vieler individueller Verläufe in einem Schaubild zu Typologien zu kommen.
Ein Anwendungsbeispiel für diese Art der Darstellung von Ereignisdaten finden wir bei Müller (1978) (Abbildung 4.1). Die durchgezogenen waagerechten Linien im Rahmen dieses Schaubilds symbolisieren die Verweildauern in verschiedenen Zuständen des Ausbildungs- beziehungsweise des Beschäftigungssystems. Die vertikalen Linien stellen Zustandswechsel dar. Mit diesem doch sehr komplexen Zustandsraum lassen sich die Berufsverläufe von 30 zufällig ausgewählten Männern des Geburtsjahrgangs 1946, die im Jahre 1971 als Landwirt gearbeitet hatten, detailliert verfolgen. Man sieht, daß sich diese Erwerbstätigengruppe vornehmlich aus Volksschulabsolventen, die danach eine landwirtschaftliche oder gewerbliche Lehre absolviert und schließlich einige Zeit als mithelfende Familienangehörige gearbeitet haben, rekrutiert.
Obwohl durch die Entwicklung von Plot-Programmen (vgl. Carr-Hill/Macdonald 1973; Müller 1978) der Einsatz dieses Verfahrens beträchtlich er-

Abbildung 4.1: Berufsverläufe von 30 zufällig ausgewählten Männern des Geburtsjahrgangs 1946, die im Jahre 1971 als Landwirt tätig waren



Quelle: Müller, W.: Klassenlage und Lebenslauf. Habilitationsschrift. Mannheim 1978.

leichtert worden ist, besteht der prinzipielle Nachteil bei der Erstellung von Mehrfach-Plots darin, daß mit zunehmender Zahl der Zustände und der zu zeichnenden Verläufe die Komplexität der Plots sehr schnell steigt und die Übersichtlichkeit spürbar sinkt. Es läßt sich dann nicht mehr identifizieren, zu welchem individuellen Verlauf eine bestimmte Linie gehört und was die Vorgeschichte eines individuellen Zustandswechsels war. Zwar kann durch die Auswahl bestimmter Subgruppen (z. B. die Auswahl von Personen mit gemeinsamen Anfangs- und Endzuständen) die Übersichtlichkeit der Plots gesteigert werden, aber diesem Verfahren sind doch enge Grenzen gesetzt.

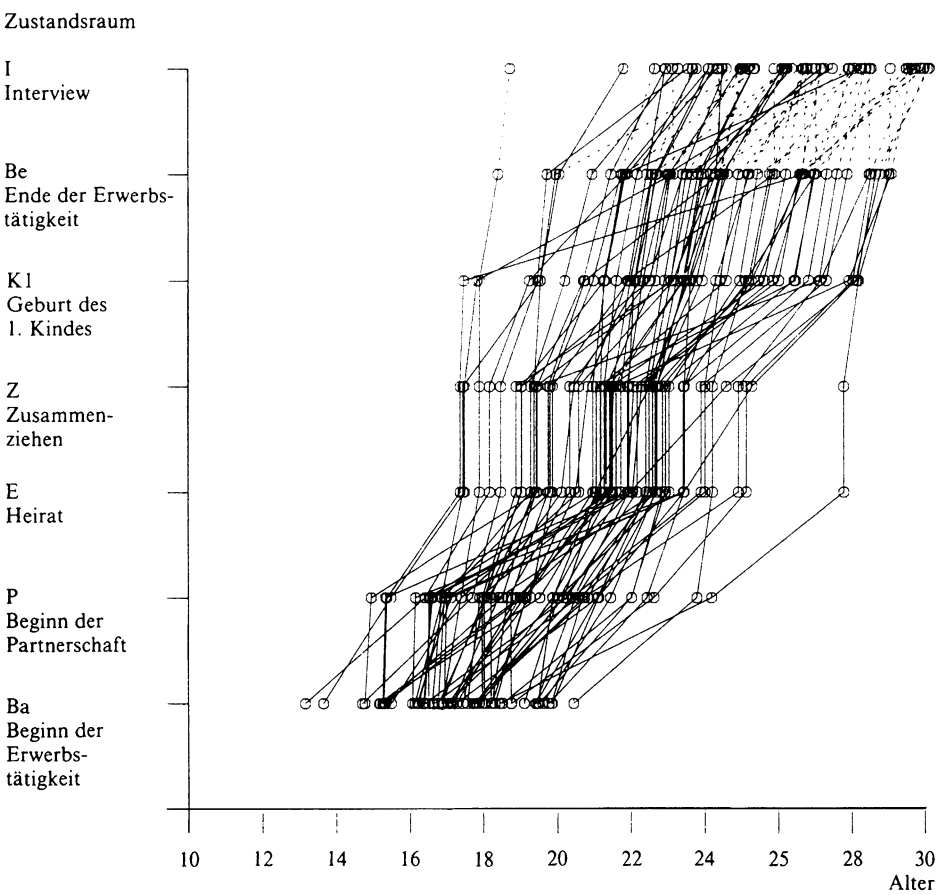
An diesem Punkt setzt ein anderes Verfahren zur Präsentation von Ereignisdaten an, die *Darstellung von Ereignissequenzen*. Die Übersichtlichkeit der individuellen Verlaufsplots wird dort gesteigert, indem nur Subpopulationen mit einer bestimmten, festgelegten Folge von Ereignissen gezeichnet werden. Auf der Abszisse wird wiederum die Zeit oder das Alter und auf der Ordinate ein bestimmtes Sequenzmuster abgetragen. In das Diagramm werden nur diejenigen Ereignissequenzen gezeichnet, die sich diesem Sequenzmuster zuordnen lassen. Ein Anwendungsbeispiel dieses Verfahrens wurde von Schulz und Strohmeier (1985) vorgelegt, in dem die Ereignisse verschiedener biographischer Karrieren (Familie, Partnerschaft, Beruf) miteinander verbunden wurden (Abbildung 4.2). An der Steilheit der Kurvenzüge kann man die Geschwindigkeit ablesen, mit der ein bestimmtes Sequenzmuster durchlaufen wurde. Die Breite der Kurvenschar gibt Informationen über den Grad der Altershomogenität. Heterogenität der Verweildauern ist im Diagramm an einer hohen Anzahl überkreuzter Linien sichtbar.

Der Vorteil dieses Verfahrens besteht darin, daß die hohe Informationsdichte der individuellen Verläufe relativ überschaubar bleibt. Es ist erkennbar, wie bestimmte Ereignisfolgen über das Lebensalter streuen. Allerdings wird bei diesem Verfahren immer eine Auswahl getroffen, und man sieht nicht, wie typisch beziehungsweise atypisch ein bestimmtes Sequenzmuster überhaupt ist. Auch in diesem Bild sinkt die Übersichtlichkeit mit der Anzahl der gezeichneten individuellen Verläufe und der sich überlagernden Kurven.

Ganz anders ist dies, wenn die Ereignisdaten nicht zur Abbildung individueller Verläufe, sondern zur *Beschreibung von Aggregaten* in der Form *kumulierter Verteilungen* benutzt werden. Hier steigt mit der Anzahl der Untersuchungseinheiten die Stabilität der Verteilungen und damit die Überschaubarkeit der Zeichnungen an. Für jeden Zeitpunkt wird die Verteilung der Untersuchungseinheiten auf eine vorgegebene Zahl von Zuständen berechnet und kumuliert. Werden diese Punkte verbunden, dann entsteht ein Bild über die Strukturveränderungen in der Zeit.

Ein Beispiel für die Anwendung kumulierter Verteilungen zur Darstellung des Bildungs- und Berufsverlaufs der Geburtskohorten von 1929–31, 1939–41 und 1949–51 liefert Blossfeld (1985b) (Abbildung 4.3). Auf der x-Achse ist dort das jeweilige Lebensalter (die kleinste Einheit ist ein Monat) der Kohorten aufgetragen, und in der y-Richtung wird deren kumulative Verteilung auf die Zustände

Abbildung 4.2: Ereignissequenz bei 61 Fällen mit demselben Verlaufsmuster

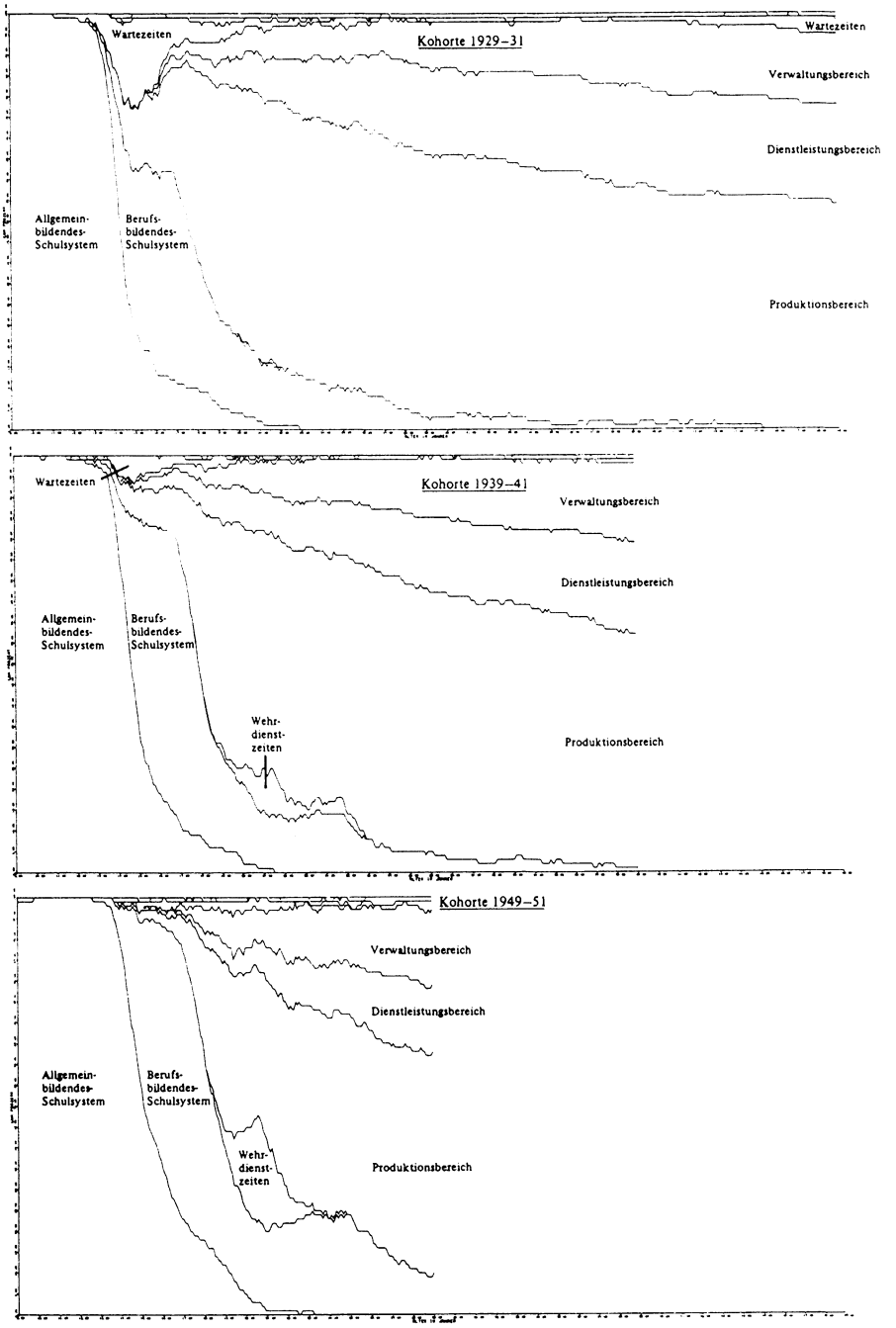


Quelle: SCHULZ, M., und STROHMEIER, K.: „Familienkarriere und Berufskarriere" In: FRANZ, H.-W.(Hrsg): 22. Deutscher Soziologentag 1984. Sektions- und Ad-hoc-Gruppen. Opladen 1985.

im Bildungs- und Beschäftigungssystem sowie auf wichtige Unterbrechungsarten (in Prozent) wiedergegeben. Entlang der Altersachse lassen sich so die mit zunehmendem Alter verbundenen Übergänge zwischen verschiedenen Bereichen des Bildungs- und Beschäftigungssystems im Saldo verfolgen. Die unterschiedlichen Chancen, denen die drei Kohorten in den 40 Jahren des (maximalen) Beobachtungszeitraums begegnet sind, können so sehr einfach verglichen werden.

Obwohl dieses Verfahren bei Mehr-Zustands- und Mehr-Episoden-Modellen wichtige Hinweise über den Prozeßverlauf geben kann, besteht sein großer

Abbildung 4.3: Bildungs- und Berufsverlauf der Geburtskohorten 1929–31, 1939–41 und 1949–51



Nachteil darin, daß die individuellen Verläufe verlorengehen. So hängt die Aussagekraft des Bildes sehr stark von den individuellen Bewegungen ab, die sich hinter der Struktur verbergen, weil zu jedem Zeitpunkt nur die Salden der individuellen Bewegungen geplottet werden. Die Stabilität auf der Ebene des Aggregats muß deswegen nicht unbedingt mit einer Stabilität der Verläufe auf der Ebene von Individuen einhergehen. Trotz allem scheint sich in der Forschungspraxis wegen der erwähnten Unübersichtlichkeit individueller Verlaufsplots die Darstellung von Aggregaten zunehmend durchzusetzen.
Insgesamt sind mit den drei vorgestellten Verfahren zur graphischen Präsentation von Ereignisdaten nützliche Instrumente verfügbar, mit denen ein erster Einblick in den zu analysierenden Prozeß auch bei außerordentlich komplex strukturierten Mehr-Zustands- und Mehr-Episoden-Fällen gewonnen werden kann. Diese Plots können Hinweise auf wichtige Zusammenhänge zwischen Variablen geben. Eine wichtige Funktion ist aber auch in der Überprüfung der Plausibilität von Ergebnissen aus komplexeren statistischen Modellen während der späteren Phasen des Analyseprozesses zu sehen. Nicht selten kann sie vor Fehlschlüssen und Methodenartefakten schützen.

4.3 Sterbetafel-Methode und Kaplan-Meier-Schätzung

Haben sich die drei bisher genannten Verfahren zur Präsentation von Ereignisdaten von vornherein auf die Darstellung des gesamten Prozeßverlaufs bezogen, so werden in diesem Abschnitt *parameterfreie Schätzmethoden* besprochen, die sich zunächst auf die einzelnen Episoden eines solchen Verlaufs konzentrieren. Zwar können auch diese Methoden zur Untersuchung von Mehr-Zustands- und Mehr-Episoden-Fällen herangezogen werden, doch werden wir darauf erst später zu sprechen kommen.
Wenden wir uns zuerst dem *Ein-Episoden-Fall* zu und beschränken uns zugleich auf nur *eine bestimmte Ereignisart*. Typische Anwendungsbeispiele sind die Untersuchung der Zeitdauer, bis ein an einer bestimmten Krankheit leidender Mensch stirbt (Medizin), oder die Untersuchung der Zeitdauer, in der eine Maschine nach Neuinstallation störungsfrei arbeitet (Technik). Aber auch in den Wirtschafts- und Sozialwissenschaften gibt es viele Anwendungsmöglichkeiten dieses Spezialfalls, wenn man beispielsweise an die Zeit denkt, die von der Markteinführung eines Produkts bis zur Kaufentscheidung der Kunden vergeht (Marketing), oder wenn man untersuchen möchte, wie lange es dauert, bis die Kinder ihr Elternhaus verlassen (Familiensoziologie). Liegen wiederholbare Ereignisse vor, wie sie beispielsweise durch die Erwerbstätigkeitsepisoden einer Berufsgeschichte gegeben sind, und interessiert man sich zum Beispiel nur für den Wechsel der ersten Erwerbstätigkeit, dann liegt wiederum der Ein-Episoden-Fall mit einer Ereignisart vor.
Zur Schätzung der Survivorfunktionen bei solchen Modellen stehen die Sterbetafel-Methode (z. B. implementiert in den Programmpaketen SPSS und BMDP)

und der Kaplan-Meier-Schätzer (z. B. implementiert im Programmpaket BMDP) zur Verfügung. Bei beiden Methoden handelt es sich um parameterfreie Schätzverfahren, die keine Verteilungsannahme über den zu untersuchenden Prozeß machen und sich deswegen außerordentlich gut für erste explorative Untersuchungen des Datenmaterials eignen. Besonders hilfreich sind dabei die graphischen Darstellungen der geschätzten Survivorfunktion und ihres Logarithmus, der geschätzten Hazardfunktion und der geschätzten Dichtefunktion der Verweildauern, die einen detaillierten Einblick in den Verlauf des Prozesses geben können.

Die Sterbetafel-Methode

Bei der *Sterbetafel-Methode*, die bereits in Abschnitt 3.2.3 beschrieben wurde, wird die Wartezeit (bis zum Eintritt des Ereignisses oder bis zur Zensierung) in feste Intervalle eingeteilt, die beliebig lang sein können. Für jedes dieser Intervalle werden die Anzahl der Untersuchungseinheiten, die am Beginn des Intervalls noch dem Ereignisrisiko ausgesetzt sind, die Anzahl der Untersuchungseinheiten, die im Intervall ein Ereignis erfahren, und die Anzahl der Untersuchungseinheiten, die im Intervall zensiert werden, gezählt. Auf dieser Basis werden dann die Dichtefunktion der Verweildauern, die Survivorfunktion und die Hazardfunktion geschätzt. Die Schätzung stützt sich auf die Annahme, daß die Zensierungen in jedem Intervall gleichverteilt sind (vgl. Abschnitt 3.2.3).
Als ein Beispiel für die konkrete Anwendung der Methode der Sterbetafelanalyse mit dem Programmpaket SPSS soll uns die Untersuchung der Zeitdauer bis zum ersten Wechsel der Erwerbstätigkeit dienen. Unser Interesse gilt dabei insbesondere der Frage, ob sich Unterschiede zwischen Männern und Frauen sowie zwischen den drei Kohorten der Lebensverlaufsstudie (den 1929–31, 1939–41 und 1949–51 Geborenen) zeigen.
Bei den Eingabedaten handelt es sich um einen ereignisorientierten Datensatz, wie er bereits in Tabelle 4.1 dargestellt ist, der neben einer Reihe von interessanten Variablen für jede Erwerbstätigkeitsepisode die Anfangs- und Endzeitpunkte enthält, vercodet in Anzahl von Monaten vom Beginn dieses Jahrhunderts an. Mit dem folgenden SPSS-Programmlauf sollen für die Männer und die Frauen beziehungsweise für die drei Kohorten jeweils getrennt die Sterbetafeln berechnet und die Verläufe der verschiedenen geschätzten Funktionen gezeichnet werden:

Programmbeispiel 4.1:

```
GET FILE        DATA
COMPUTE         DUR = M51 - M50 + 1
COMPUTE         ZEN = 1
IF              (M51 EQ M47) ZEN = 0
IF              (M48 GE 348 AND LE 384) KOHO = 1
IF              (M48 GE 468 AND LE 504) KOHO = 2
IF              (M48 GE 588 AND LE 624) KOHO = 3
*SELECT IF      (M5 EQ 1)
```

```
SURVIVAL        TABLES = DUR BY M3(1,2),KOHO(1,3)/
                INTERVALS=THRU 24 BY 3, THRU 300 BY 12/
                STATUS = ZEN(1)/
                PLOTS (ALL)/
                COMPARE
OPTIONS         5,8
FINISH
```

Nach dem Einlesen des SPSS-Systemfiles wird mit der zweiten Karte zunächst die Erwerbstätigkeitsdauer DUR berechnet, indem vom Zeitpunkt des Endes jeder Erwerbstätigkeitsepisode (M51) der Zeitpunkt des Beginns jeder Erwerbstätigkeitsepisode (M50) subtrahiert wird (vgl. Anhang 1). Die Erwerbstätigkeitsdauer jeder Episode steht dann in Monatseinheiten zur Verfügung. Mit den nächsten zwei Anweisungen wird eine Zensierungsvariable erzeugt, die die Ausprägung 1 erhält, wenn es sich um ein reguläres Erwerbstätigkeitsende handelt, und sonst (bei Zensierung) den Wert 0 annimmt, wenn das Ende der jeweiligen Erwerbstätigkeitsepisode (M51) mit dem Zeitpunkt des Interviews (M47) identisch ist. Die nächsten drei Anweisungen bilden aus dem Geburtszeitpunkt (M48), der auch durch die Anzahl von Monaten vom Beginn des Jahrhunderts an vercodet ist, eine neue Variable (KOHO), welche die drei Kohorten durch drei Ausprägungen unterscheidet. Da uns nicht alle Episoden der Erwerbsgeschichte interessieren, die in dem ereignisorientierten Datensatz enthalten sind, sondern nur die erste Erwerbstätigkeit, wird mit Hilfe der Sequenznummer der Episoden (M5) nur jeweils die erste herausgefiltert. Die eigentliche Sterbetafelanalyse wird mit der Kennung SURVIVAL aufgerufen. In der TABLES-Angabe werden neben der Überlebenszeit DUR die Gruppierungsvariablen Geschlecht (M3) und Kohorte (KOHO) genannt, nach deren Werten (in Klammern gegeben) die Gesamtstichprobe in Teilstichproben gruppiert werden soll.
Der in den Sterbetafeln zu berücksichtigende Zeitraum und die Einteilung in Intervalle werden nach dem Schlüsselwort INTERVALS bezeichnet. Die Wartezeit wurde in diesem Beispiel während der ersten 24 Monate in 3-Monats-Intervalle und danach bis zum 300. Monat (dem 25. Berufsjahr) in 12-Monats-Intervalle eingeteilt. Dies vor allem deswegen, weil Wechsel in den ersten Jahren einer Erwerbstätigkeit besonders häufig sein dürften (Probezeit, Enttäuschung der Erwartungen von Arbeitnehmern oder Arbeitgebern) und in späteren Phasen aufgrund von Humankapitalinvestitionen langsam zurückgehen müßten. In der STATUS-Angabe wird dem Programm die Zensierungsvariable (ZEN) zugeordnet. Die Ausprägung in Klammern bedeutet, daß reguläre Ereignisse die Ausprägung 1 haben, während alle anderen Werte als zensiert behandelt werden. Durch die PLOTS-Angabe mit dem Schlüsselwort ALL werden die Survivorfunktion und ihr Logarithmus, die Hazardfunktion und die Dichtefunktion der Verweildauern gezeichnet. Dabei wird mit COMPARE veranlaßt, daß die Überlebensfunktionen der jeweiligen Subgruppen verglichen und statistisch dahingehend geprüft werden, ob sie sich signifikant voneinander unterscheiden. Die OPTIONS-Karte bewirkt schließlich, daß bei Speicherplatzmangel die Si-

gnifikanzprüfung nur approximativ durchgeführt wird (Ausprägung 5) und daß die erstellten Sterbetafeln zusätzlich in Rohdatenform in eine Datei ausgegeben werden, die dann mit einem Plotprogramm ausgewertet werden kann (Ausprägung 8).

Mit diesem SPSS-Programm werden fünf Sterbetafeln (für die Männer und Frauen sowie für jede der drei Kohorten) erzeugt, von denen als Beispiel nur die der Männer besprochen werden soll (Tabelle 4.2). Die 1. Spalte dieser Sterbetafel enthält dabei die Untergrenzen der gewählten Intervalle (a_{k-1}). In der Spalte 2 wird jeweils zu Beginn jedes Intervalls gezählt, wie viele Männer noch dem Risiko eines Erwerbstätigkeitswechsels ausgesetzt waren (Risikomenge 1: R_k). Im ersten Intervall sind dies beispielsweise $R_1 = 1077$ Männer und im zweiten Intervall nur noch $R_2 = 1063$ Männer.

In der Spalte 3 wird die Anzahl der Männer ausgegeben, deren Berufsepisode im jeweiligen Intervall zensiert worden ist (w_k). Gab es beispielsweise im ersten Intervall keine Zensierungen ($w_1 = 0$), so wurde im zweiten Intervall nur eine Episode zensiert ($w_2 = 1$). Die 4. Spalte beinhaltet die um die zensierten Beobachtungen korrigierte Zahl der Männer, die dem Risiko eines Berufswechsels ausgesetzt waren (Risikomenge 2: $R_k - \frac{w_k}{2}$). Für das zweite Intervall ergibt sich damit beispielsweise ein Wert der Risikomenge 2 von $1063 - \frac{1}{2} = 1062{,}5$.

Tabelle 4.2: Beispiel einer Sterbetafel

LIFE TABLE
SURVIVAL VARIABLE DUR FOR M3

INTVL START TIME	NUMBER ENTRNG THIS INTVL	NUMBER WDRAWN DURING INTVL	NUMBER EXPOSD TO RISK	NUMBER OF TERMNL EVENTS	PROPN TERMI-NATING	PROPN SURVI-VING	CUMUL PROPN SURV AT END	PROBA-BILITY DENSTY	HAZARD RATE	SE OF CUMUL SURV-IVING	SE OF PROB-ABILTY DENS	SE OF HAZRD RATE
0.0	1077.0	0.0	1077.0	14.0	0.0130	0.9870	0.9870	0.0043	0.0044	0.003	0.001	0.001
3.0	1063.0	1.0	1062.5	44.0	0.0414	0.9586	0.9461	0.0136	0.0141	0.007	0.002	0.002
6.0	1018.0	1.0	1017.5	64.0	0.0629	0.9371	0.8866	0.0198	0.0216	0.010	0.002	0.003
9.0	953.0	0.0	953.0	46.0	0.0483	0.9517	0.8438	0.0143	0.0165	0.011	0.002	0.002
12.0	907.0	1.0	906.5	85.0	0.0938	0.9062	0.7647	0.0264	0.0328	0.013	0.003	0.004
15.0	821.0	0.0	821.0	36.0	0.0438	0.9562	0.7312	0.0112	0.0149	0.014	0.002	0.002
18.0	785.0	2.0	784.0	45.0	0.0574	0.9426	0.6892	0.0140	0.0197	0.014	0.002	0.003
21.0	738.0	0.0	738.0	42.0	0.0569	0.9431	0.6500	0.0131	0.0195	0.015	0.002	0.003
24.0	696.0	5.0	693.5	183.0	0.2639	0.7361	0.4785	0.0143	0.0253	0.015	0.001	0.002
36.0	508.0	0.0	508.0	92.0	0.1811	0.8189	0.3918	0.0072	0.0166	0.015	0.001	0.002
48.0	416.0	3.0	414.5	75.0	0.1809	0.8191	0.3209	0.0059	0.0166	0.014	0.001	0.002
60.0	338.0	3.0	336.5	47.0	0.1397	0.8603	0.2761	0.0037	0.0125	0.014	0.001	0.002
72.0	288.0	1.0	287.5	40.0	0.1391	0.8609	0.2377	0.0032	0.0125	0.013	0.000	0.002
84.0	247.0	2.0	246.0	27.0	0.1098	0.8902	0.2116	0.0022	0.0097	0.013	0.000	0.002
96.0	218.0	3.0	216.5	21.0	0.0970	0.9030	0.1911	0.0017	0.0085	0.012	0.000	0.002
108.0	194.0	3.0	192.5	20.0	0.1039	0.8961	0.1712	0.0017	0.0091	0.012	0.000	0.002
120.0	171.0	2.0	170.0	21.0	0.1235	0.8765	0.1501	0.0018	0.0110	0.011	0.000	0.002
132.0	148.0	4.0	146.0	13.0	0.0890	0.9110	0.1367	0.0011	0.0078	0.011	0.000	0.002
144.0	131.0	7.0	127.5	9.0	0.0706	0.9294	0.1271	0.0008	0.0061	0.010	0.000	0.002
156.0	115.0	5.0	112.5	8.0	0.0711	0.9289	0.1180	0.0008	0.0061	0.010	0.000	0.002
168.0	102.0	7.0	98.5	5.0	0.0508	0.9492	0.1120	0.0005	0.0043	0.010	0.000	0.002
180.0	90.0	2.0	89.0	11.0	0.1236	0.8764	0.0982	0.0012	0.0110	0.010	0.000	0.003
192.0	77.0	2.0	76.0	4.0	0.0526	0.9474	0.0930	0.0004	0.0045	0.009	0.000	0.002
204.0	71.0	3.0	69.5	3.0	0.0432	0.9568	0.0890	0.0003	0.0037	0.009	0.000	0.002
216.0	65.0	2.0	64.0	1.0	0.0156	0.9844	0.0876	0.0001	0.0013	0.009	0.000	0.001
228.0	62.0	1.0	61.5	6.0	0.0976	0.9024	0.0791	0.0007	0.0085	0.009	0.000	0.003
240.0	55.0	2.0	54.0	2.0	0.0370	0.9630	0.0761	0.0002	0.0031	0.009	0.000	0.002
252.0	51.0	2.0	50.0	3.0	0.0600	0.9400	0.0716	0.0004	0.0052	0.009	0.000	0.003
264.0	46.0	5.0	43.5	3.0	0.0690	0.9310	0.0666	0.0004	0.0060	0.009	0.000	0.003
276.0	38.0	3.0	36.5	1.0	0.0274	0.9726	0.0648	0.0002	0.0023	0.009	0.000	0.002
288.0	34.0	4.0	32.0	0.0	0.0000	1.0000	0.0648	0.0000	0.0000	0.009	0.000	0.000
300.0+	30.0	24.0	18.0	6.0	0.3333	0.6667	0.0432	**	**	0.009	**	**

** THESE CALCULATIONS FOR THE LAST INTERVAL ARE MEANINGLESS.

THE MEDIAN SURVIVAL TIME FOR THESE DATA IS 34.49

Die regulären Ereignisse für jedes Intervall werden schließlich in der 5. Spalte ausgegeben (d_k). Danach gab es beispielsweise im zweiten Intervall $d_2 = 44$ Ereignisse.
Auf der Grundlage dieser Daten werden dann die bedingte Wahrscheinlichkeit zum Jobwechsel in jedem Intervall (Spalte 6)

$$\hat{\lambda}_k = \frac{d_k}{R_k - \frac{w_k}{2}},$$

die bedingte Wahrscheinlichkeit zum Verbleib im Job für jedes Intervall (Spalte 7)

$$\hat{p}_k = 1 - \hat{\lambda}_k,$$

die Survivorfunktion (Spalte 8)

$$\hat{P}_k = \hat{p}_k \cdot \ldots \cdot \hat{p}_1,$$

die Dichtefunktion der Verweildauern (Spalte 9)

$$\hat{f}_k = \frac{\hat{P}_{k-1} - \hat{P}_k}{h_k},$$

die Hazardfunktion (Spalte 10)

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

sowie die dazugehörigen Standardfehler (Spalten 11 bis 13) geschätzt (vgl. Abschnitt 3.2.3).
Für das zweite Intervall ergibt sich danach beispielsweise eine Schätzung der bedingten Wahrscheinlichkeit zum Jobwechsel von

$$\hat{\lambda}_2 = \frac{44}{1063 - \frac{1}{2}} = 0{,}0414,$$

eine Schätzung der bedingten Wahrscheinlichkeit zum Verbleib von

$$\hat{p}_2 = 1 - 0{,}0414 = 0{,}9586,$$

eine Schätzung für die Survivorfunktion von

$$\hat{P}_2 = 0{,}9870 \cdot 0{,}9586 = 0{,}9461,$$

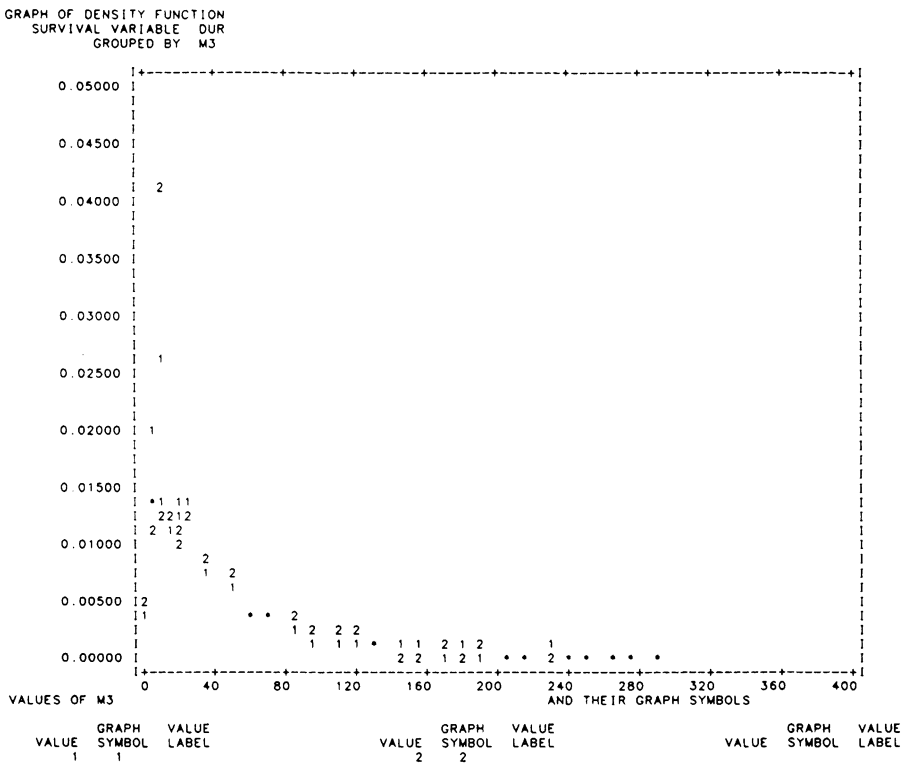
eine Schätzung für die Dichtefunktion der Verweildauer von

$$\hat{f}_2 = \frac{0{,}9870 - 0{,}9461}{3} = 0{,}0136$$

und eine Schätzung für die Hazardfunktion von

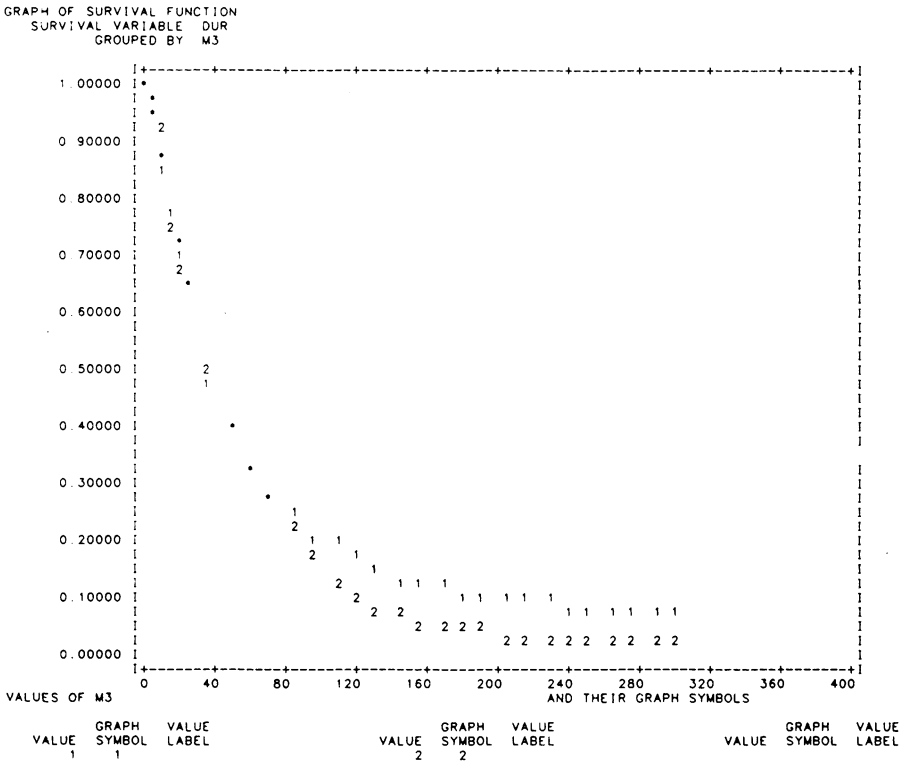
$$\hat{\lambda}(m_2) = \frac{2 \cdot 0{,}0414}{3(1+0{,}9586)} = 0{,}0141\,.$$

Abbildung 4.4: Beispiel für einen Plot der Dichtefunktion der Verweildauern, geschätzt nach der Sterbetafel-Methode



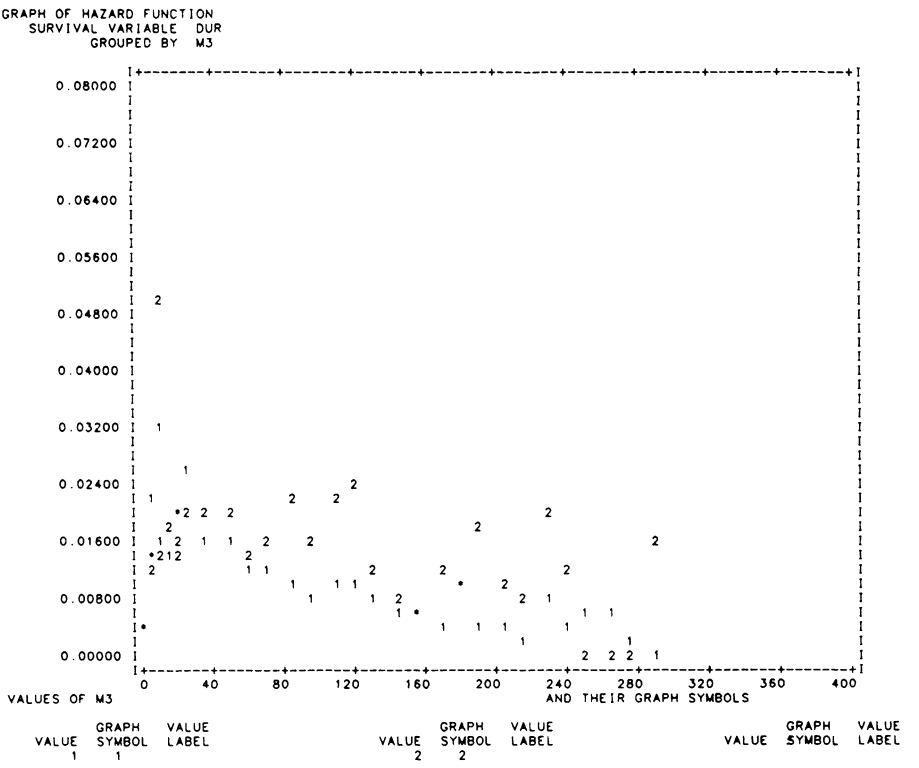
Der dazugehörige Standardfehler für die Survivorfunktion beträgt 0,007, für die Dichtefunktion der Verweildauer 0,002 und für die Hazardfunktion 0,002.
Unterhalb der Sterbetafel wird schließlich noch eine Schätzung des Medians der Verweildauer in der ersten Erwerbstätigkeit ausgedruckt (ein Wert von 34,49 bedeutet hier etwa zwei Jahre und zehn Monate).
Die Sterbetafel, die den Prozeß der ersten Erwerbstätigkeit sehr detailliert beschreibt, ist natürlich sehr komplex und eignet sich nur wenig zum Vergleich von Subgruppen. Für den Vergleich von Subpopulationen sind dagegen die von SPSS zur Verfügung gestellten Plots von großem Nutzen.
Betrachten wir zuerst die *Dichtefunktion der Verweildauern* (vgl. Formel 3.2.2) in der ersten Erwerbstätigkeit von Männern (1) und Frauen (2) (Abbildung 4.4). Dabei wird deutlich, daß bei beiden Geschlechtern die Dichte am Beginn des Prozesses rasch ansteigt und sich dann asymptotisch mit zunehmender Verweildauer der x-Achse annähert. Die Dichtefunktion zeigt also für beide Geschlechter denselben rechtsschiefen Verlauf, und die Unterschiede zwischen den beiden Prozessen sind, rein optisch betrachtet, minimal.

Abbildung 4.5: Beispiel für einen Plot der Survivorfunktion, geschätzt nach der Sterbetafel-Methode



Eine andere Darstellung des Prozesses zeigt ein Plot der *Survivorfunktion* (Abbildung 4.5). Die Survivorfunktion gibt in diesem Falle an, wie groß der Anteil der Männer (1) beziehungsweise der Frauen (2) ist, die bis zum jeweiligen Zeitpunkt ihren ersten Beruf noch nicht gewechselt haben. Aus dem Plot ist ersichtlich, daß bereits nach 48 Monaten (also 4 Jahren) etwa 70 Prozent der Männer und Frauen aus dem ersten Beruf ausgeschieden sind. Der Prozeß verläuft bis zu diesem Zeitpunkt für Männer und Frauen sehr ähnlich. Doch dann werden die Unterschiede zwischen beiden Geschlechtern zunehmend größer. Die Frauen haben bei zunehmender Verweildauer im ersten Beruf die Tendenz, vergleichsweise schneller die erste Erwerbstätigkeit zu verlassen. Ihre Survivorfunktion fällt deswegen unter die der Männer.
Auf der Basis der Schätzung der *Hazardfunktion* (vgl. Formel 3.2.5) werden Hinweise darauf gegeben, ob es sich bei dem Prozeß des Ausscheidens aus der ersten Erwerbstätigkeit um einen zeitabhängigen Prozeß handelt oder nicht. So ist in Abbildung 4.6 zu erkennen, daß die Wahrscheinlichkeit eines Berufswechsels in den ersten 9 Monaten bei beiden Geschlechtern ansteigt, um dann mit

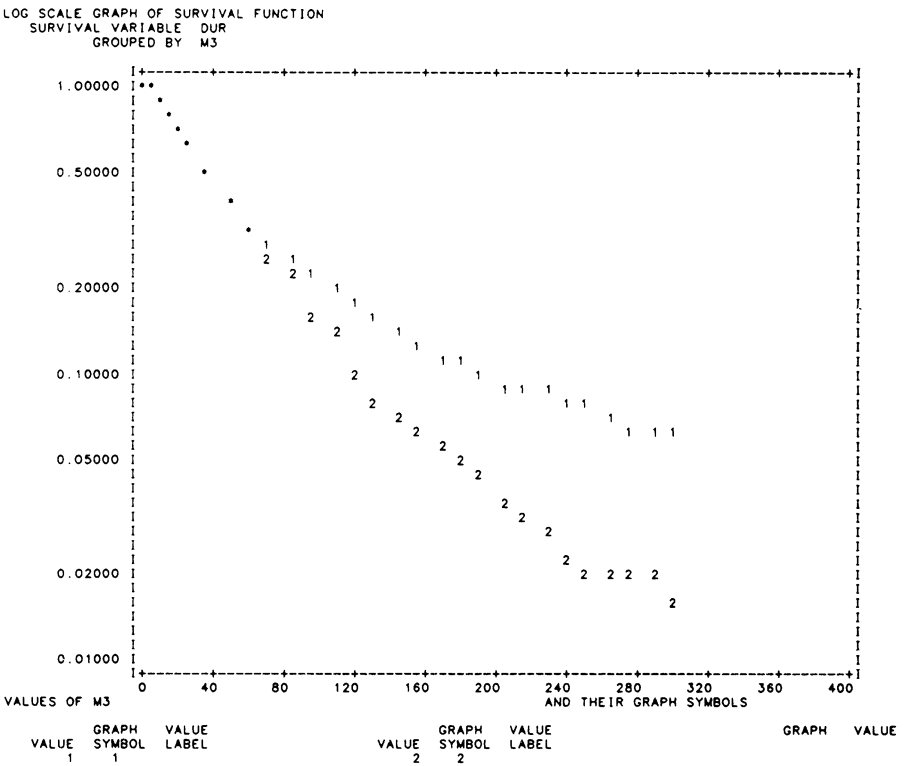
Abbildung 4.6: Beispiel für einen Plot der Hazardfunktion, geschätzt nach der Sterbetafel-Methode



zunehmender Dauer zu sinken. Bei den Frauen zeigt sich allerdings zwischen dem 60. Monat und dem 110. Monat der Erwerbsdauer wieder ein deutlicher Anstieg des Berufswechselrisikos. Nach etwa dem 180. Monat (also dem 15. Jahr der Berufstätigkeit im ersten Beruf) sind die Plots der Hazardfunktion allerdings nur noch schwer zu interpretieren, da aufgrund der geringen Zahl der Personen, die danach noch dem Risiko ausgesetzt sind, die Schätzungen des Verlaufs der Hazardfunktion sehr instabil werden und größeren Schwankungen unterliegen.

Mit der *logarithmierten Survivorfunktion* beziehungsweise der *kumulierten Hazardfunktion* (vgl. Formel 3.2.9) in Abbildung 4.7 ist schließlich noch einmal eine andere Darstellung möglich, wie sich das Risiko, den ersten Beruf zu wechseln, mit zunehmender Verweildauer entwickelt. Würde der Prozeß in der Zeit konstant verlaufen, dann müßte die logarithmierte Survivorfunktion linear mit einer konstanten Steigung fallen. Vermindert sich das Risiko, wie dies beispielsweise bei den Männern (1) ab dem 12. Monat der Fall ist, dann biegt sich die Kurve nach oben, steigt dagegen das Risiko, wie das zum Beispiel bei den

Abbildung 4.7: Beispiel für einen Plot der logarithmierten Survivorfunktion (bzw. kumulierten Hazardfunktion), geschätzt nach der Sterbetafel-Methode



Frauen (2) ab dem 60. Monat gegeben ist, dann wird die Kurve nach unten gebogen.

Zusätzlich zu den obigen Plots kann in SPSS anhand der *Lee-Desu-Teststatistik* (Lee/Desu 1972) überprüft werden, ob sich die Survivorfunktionen von Männern und Frauen signifikant unterscheiden. Es handelt sich dabei um einen modifizierten Wilcoxon-Test (vgl. Abschnitt 3.2.5). Voraussetzung dieses Rangtests ist, daß sich die Survivorfunktionen der Subpopulationen nicht überschneiden (d. h. die Ränge dadurch nicht umgedreht werden) und innerhalb der Subpopulationen identische Zensierungsmuster vorliegen. Unter der Nullhypothese, daß sich die Survivorfunktionen der Subgruppen nicht unterscheiden, ist diese Teststatistik asymptotisch χ^2-verteilt mit k–1 Freiheitsgraden (k = Anzahl der Subgruppen). Der Vergleich der Überlebensfunktionen von Männern und Frauen im obigen Beispiel ergab einen Wert von 0,048 mit 1 d. f. Die Überlebensfunktionen von Männern und Frauen unterscheiden sich also bei einem Signifikanzniveau von 0,05 nicht signifikant voneinander. Es ist allerdings zu berück-

sichtigen, daß dieser Test eher sensitiv auf Unterschiede der Survivorfunktionen zu Beginn des Prozesses reagiert. Da sich die Unterschiede zwischen Männern und Frauen im ersten Beruf aber erst nach einer Verweildauer von 48 Monaten herauskristallisieren, sollte man besser auf die Cox-Mantel-Teststatistik, die im Programmpaket BMDP angeboten wird, zurückgreifen. Diese Teststatistik ist in bezug auf Unterschiede gegen Ende des Prozesses besonders sensitiv. Diese Teststatistik wird später im Zusammenhang mit der Kaplan-Meier-Schätzung noch genauer dargestellt.

Der Kaplan-Meier-Schätzer

Obwohl die Sterbetafelmethode, insbesondere bei großen Stichproben, ein nützliches Verfahren zur ersten Analyse von Ereignisdaten darstellt, hängt die Genauigkeit der Schätzungen doch stark von den gewählten Intervallbreiten ab. Je größer diese sind, desto schlechter und ungenauer dürften in der Regel die Schätzungen der Funktionen werden. Störend ist ferner die Tatsache, daß bei unterschiedlicher Wahl der Intervalleinteilung normalerweise jeweils unterschiedliche Schätzergebnisse zu erwarten sind. Bei nicht zu großen Stichproben sollte man deswegen auf den *Kaplan-Meier-Schätzer (oder den Produkt-Limit-Schätzer)* zurückgreifen, der im Programmpaket BMDP verfügbar ist. Diese Schätzmethode, die bereits in Abschnitt 3.2.4 ausführlich beschrieben wurde, arbeitet nicht mit nach Intervallen gruppierten Daten, sondern mit den tatsächlich gemessenen Ereignis- und Zensierungszeiten. Die Grundidee dieses Schätzers ist, daß durch die Einteilung der Verweildauer in immer kleinere Intervalle schließlich ein Punkt erreicht wird, wo jede Ereignis- oder Zensierungszeit nur in ein bestimmtes Intervall fällt. Tatsächlich sind Sterbetafel- und Kaplan-Meier-Schätzer dann identisch, wenn man bei der Sterbetafelmethode die Intervalle entsprechend klein wählt. Die gemessenen Ereignis-und Zensierungszeiten werden dann der Größe nach geordnet. Dabei werden zensierte Beobachtungen, die zum gleichen Zeitpunkt wie Ereignisse auftreten, als etwas verzögert betrachtet. Auf der Basis einer solchen eindeutigen Rangreihe von Ereignis- und Zensierungszeiten werden Schätzungen nur für die Ereigniszeitpunkte vorgenommen, während die zensierten Zeiten nur jeweils die Risikomenge der später eintretenden Ereignisse verringern. Bei der nach der Methode von Kaplan und Meier geschätzten Survivorfunktion handelt es sich deswegen um eine Stufenfunktion mit diskreten Sprungstellen an den Ereigniszeitpunkten. Allerdings tritt bei dieser Methode ein Problem auf, wenn es zensierte Zeiten in dieser Rangreihe gibt, die größer sind als die größte Ereigniszeit. Die geschätzte Survivorfunktion kann in diesen Fällen nicht mehr gegen Null streben, und der wahre Mittelwert wird unterschätzt. In der Praxis wird man in solchen Fällen nur mehr die Zeitspanne bis zum größten Ereigniszeitpunkt zur Interpretation heranziehen.

Zur Illustration der konkreten Anwendung des Kaplan-Meier-Schätzers soll uns wieder das Beispiel der Zeitdauern bis zum Wechsel der ersten Erwerbstätigkeit

dienen. Es handelt sich um denselben ereignisorientierten Datensatz, den wir diesmal jedoch in das Programm PL1 von BMDP einlesen.

Programmbeispiel 4.2:

```
/INPUT UNIT IS 30.
       CODE IS DATA.
/VARIABLE NAMES ARE (63) DUR,(64)ZEN,(65)KOHO.
           ADD = 3.
/TRANSFORM DUR = M51 - M50 + 1.
           ZEN = 1.
           IF (M51 EQ M47) THEN ZEN = 0.
           IF (M48 GE 348 AND M48 LE 384) THEN KOHO = 1.
           IF (M48 GE 468 AND M48 LE 504) THEN KOHO = 2.
           IF (M48 GE 588 AND M48 LE 624) THEN KOHO = 3.
           USE = M5 EQ 1.
/GROUP CODES (65) ARE 1,2,3.
       NAMES (65) ARE KOHO1,KOHO2,KOHO3.
       CODES (3) ARE 1,2.
       NAMES (3) ARE MAENNER,FRAUEN.
/FORM TIME IS DUR.
      STATUS IS ZEN.
      RESPONSE IS 1.
/ESTIMATE METHOD IS PROD.
          GROUP IS KOHO.
          PLOTS ARE SURV,LOG.
          STATISTICS ARE BRESLOW,MANTEL.
/ESTIMATE METHOD IS PROD.
          GROUP IS M3.
          PLOTS ARE SURV,LOG.
          STATISTICS ARE BRESLOW,MANTEL.
/END
```

Im obigen BMDP-Programm werden nach dem Einlesen der BMDP-Systemdatei (vgl. Anhang 1), die 62 Variablen enthält, im Paragraph VARIABLE drei zusätzliche Variablen definiert, die in diesem Lauf noch benötigt werden. Im TRANSFORM-Paragraph wird, wie im vorhergehenden SPSS-Beispiel, zuerst die Verweildauer DUR berechnet, dann eine Zensierungsvariable (ZEN) gebildet, eine Kohortenvariable (KOHO) erzeugt und schließlich mit USE nur die jeweils erste Berufstätigkeit herausgefiltert.
Im FORM-Paragraph wird dem BMDP-Programm die Verweildauervariable (TIME IS DUR) und die Zensierungsvariable (STATUS IS ZEN) zugeordnet. Letztere hat bei einem regulären Ereignis den Wert 1 (RESPONSE IS 1). Die Schätzung soll jeweils getrennt nach Kohorte (KOHO) und Geschlecht (M3) erfolgen. Dabei ist der Produkt-Limit-Schätzer oder Kaplan-Meier-Schätzer (METHOD IS PROD) zu verwenden. Geplottet werden sollen jeweils die Survivor- (SURV) und die logarithmierte Survivorfunktion (LOG). Darüber hinaus sollen die Breslow- (BRESLOW) und die Cox-Mantel-Statistik (MANTEL) verwendet werden.
Tabelle 4.3 zeigt zunächst am Beispiel der Männer, wie mit der Methode von Kaplan und Meier die Survivorfunktion für den ersten Berufswechsel geschätzt

Tabelle 4.3 Beispiel für eine Kaplan-Meier-Schätzung

TIME VARIABLE IS DUR

CASE LABEL	CASE NUMBER	TIME	STATUS	CUMULATIVE SURVIVAL	STANDARD ERROR	CUM DEATHS	CUM LOST	REMAIN AT RISK
	1788	1.00	DEAD			1	0	1077
	1944	1.00	DEAD			2	0	1076
	2396	1.00	DEAD			3	0	1075
	3632	1.00	DEAD			4	0	1074
	4644	1.00	DEAD			5	0	1073
	4770	1.00	DEAD	0.9944	0.0023	6	0	1072
	1843	2.00	DEAD			7	0	1071
	2645	2.00	DEAD			8	0	1070
	2660	2.00	DEAD			9	0	1069
	2936	2.00	DEAD			10	0	1068
	2951	2.00	DEAD			11	0	1067
	3339	2.00	DEAD			12	0	1066
	4875	2.00	DEAD			13	0	1065
	6023	2.00	DEAD			14	0	1064
	6653	2.00	DEAD	0.9861	0.0036	15	0	1063
	69	3.00	DEAD			16	0	1062
	758	3.00	DEAD			17	0	1061
	871	3.00	DEAD			18	0	1060
	1413	3.00	DEAD			19	0	1059
	1468	3.00	DEAD			20	0	1058
	2484	3.00	DEAD			21	0	1057
	3155	3.00	DEAD			22	0	1056
	3465	3.00	DEAD			23	0	1055
	4448	3.00	DEAD			24	0	1054
	4527	3.00	DEAD			25	0	1053
	4662	3.00	DEAD			26	0	1052
	5113	3.00	DEAD			27	0	1051
	5298	3.00	DEAD			28	0	1050
	5509	3.00	DEAD			29	0	1049
	6370	3.00	DEAD			30	0	1048
	6391	3.00	DEAD	0.9712	0.0051	31	0	1047
	6432	3.00	CENSORED			31	0	1046
	136	4.00	DEAD			32	0	1045
	313	4.00	DEAD			33	0	1044
	855	4.00	DEAD			34	0	1043
	897	4.00	DEAD			35	0	1042
	1270	4.00	DEAD			36	0	1041
	1389	4.00	DEAD			37	0	1040
	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.
	4198	331.00	DEAD	0.0551	0.0086	976	0	21
	2866	333.00	CENSORED			976	0	20
	4319	334.00	CENSORED			976	0	19
	1612	363.00	DEAD	0.0522	0.0086	977	0	18
	180	367.00	CENSORED			977	0	17
	4689	368.00	CENSORED			977	0	16
	3119	377.00	CENSORED			977	0	15
	244	388.00	CENSORED			977	0	14
	4346	392.00	CENSORED			977	0	13
	372	397.00	CENSORED			977	0	12
	5194	398.00	CENSORED			977	0	11
	553	404.00	CENSORED			977	0	10
	6675	410.00	DEAD	0.0470	0.0092	978	0	9
	3565	413.00	CENSORED			978	0	8
	6590	413.00	CENSORED			978	0	7
	375	414.00	CENSORED			978	0	6
	1986	416.00	CENSORED			978	0	5
	6161	416.00	CENSORED			978	0	4
	5217	422.00	CENSORED			978	0	3
	6186	426.00	CENSORED			978	0	2
	1	428.00	CENSORED			978	0	1
	3458	429.00	CENSORED			978	0	0

MEAN SURVIVAL TIME = 73.83 LIMITED TO 429.00 S.E. = 3.291

wird. In der Spalte CASE NUMBER werden die Fallnummern der Männer nach ihrer Verweildauer in der TIME-Spalte geordnet. Ob es sich um ein reguläres Ereignis (DEAD) oder um eine zensierte Beobachtung (CENSORED) handelt, wird in der STATUS-Spalte ausgewiesen. Die Regel, daß bei gleicher Dauer von Ereigniszeit und Zensierungszeit die Zensierungszeit als etwas verzögert zu behandeln ist, wird am Fall mit der Nummer 6432 deutlich, der nach den Ereigniszeiten mit ebenfalls 3 Monaten rangiert. Da in der Lebensverlaufsstudie die Erwerbsgeschichte auf Monatsebene erhoben wurde, ist es durchaus möglich, daß mehrere Fälle die gleiche Verweildauer haben. In Tabelle 4.3 gibt es beispielsweise sechs Fälle mit einer einmonatigen Verweildauer. Gemäß der Formel 3.2.41 mit einer Korrektur für Ties,

$$\hat{S}(t) = \begin{cases} 1 & \text{für } t \leq t_{(1)} \\ \prod\limits_{k|t_{(k)}<t} \left(1 - \frac{1}{R_k}\right) & \text{für } t > t_{(1)} \end{cases}$$

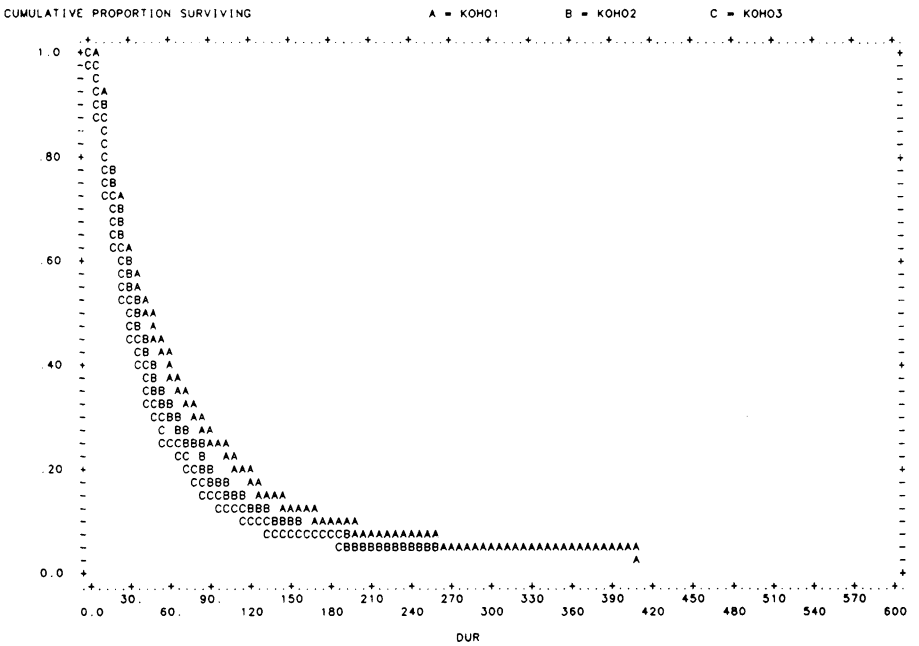
springt die Survivorfunktion in der Spalte CUMULATIVE SURVIVAL nach dem ersten Monat um den Wert $\frac{6}{1077} = 0{,}0056$ von 1 auf 0,9944 und in der Spalte STANDARD ERROR wird an der Sprungstelle der Standardfehler ausgedruckt (vgl. Formel 3.2.42). Ist der Anteil solcher Bindungen aufgrund ungenauer Erhebung der Verweildauern recht groß, kann die Sterbetafelmethode das effizientere und robustere Verfahren sein. Schließlich wird in der Spalte CUM DEATHS die Anzahl der regulären Ereignisse gezählt und in der Spalte REMAIN AT RISK die Zahl der Männer ausgewiesen, die jeweils noch dem Risiko eines Berufswechsels ausgesetzt sind.

Die größte Ereigniszeit besitzt der Fall mit der Nummer 6675 mit 410 Monaten. Danach folgen aber noch weitere 9 zensierte Fälle, die bei der Schätzung nicht berücksichtigt werden. Man muß deswegen davon ausgehen, daß die durchschnittliche Verweildauer (MEAN SURVIVAL TIME) in der ersten Erwerbstätigkeit mit 73,83 Monaten etwas unterschätzt wird (vgl. Formel 3.2.43).

In der Regel ist der Ausdruck des detaillierten Berechnungsschemas der Überlebensfunktion nach der Methode von Kaplan und Meier (Tabelle 4.3) wenig hilfreich. Der Ausdruck ist meist sehr umfangreich (im obigen Beispiel nur für die Männer bereits 1077 Zeilen) und kann gegebenenfalls mit der Anweisung NO PRINT unterdrückt werden. Er vermittelt aber einen Eindruck davon, wie groß der Sortier- und Speicherplatzaufwand zur Berechnung eines Kaplan-Meier-Schätzers ist. Bei großen Stichproben ist die Sterbetafelmethode dem Kaplan-Meier-Schätzer deswegen vorzuziehen.

Von großem Nutzen für den Vergleich von Subpopulationen sind die auch von BMDP zur Verfügung gestellten Plots. Abbildung 4.8 zeigt den *Verlauf der Survivorfunktion* für die Kohorten 1929–31 (A), 1939–41 (B) und 1949–51 (C). Dabei wird deutlich, daß die Survivorfunktionen bis zu einer Verweildauer von etwa 30 Monaten etwa gleich verlaufen und sich danach Unterschiede zwischen

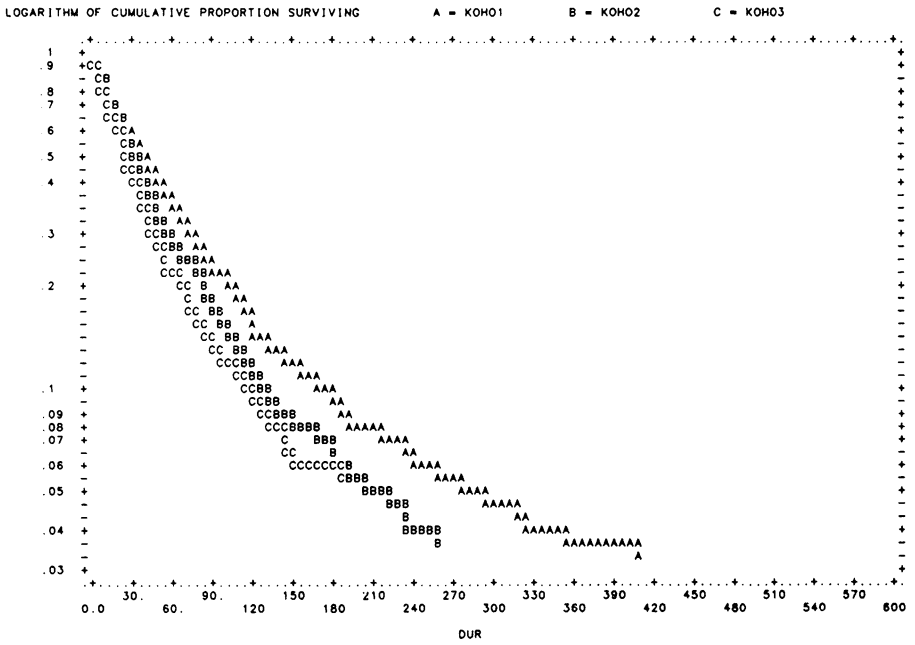
Abbildung 4.8: Beispiel für einen Plot der Survivorfunktion auf der Basis einer Kaplan-Meier-Schätzung



den Kohorten herauskristallisieren. Je jünger die Kohorte ist, desto schneller wird die erste Erwerbstätigkeit verlassen. Dies ist auch an den *logarithmierten Survivorfunktionen* (bzw. den *kumulierten Hazardfunktionen*) (Abbildung 4.9) sichtbar (vgl. Formel 3.2.9), die von der ältesten zur jüngsten Kohorte zunehmend nach unten gekrümmt sind. Dies besagt, daß das Risiko, den ersten Job zu wechseln, von der ältesten bis zur jüngsten Kohorte ansteigt.

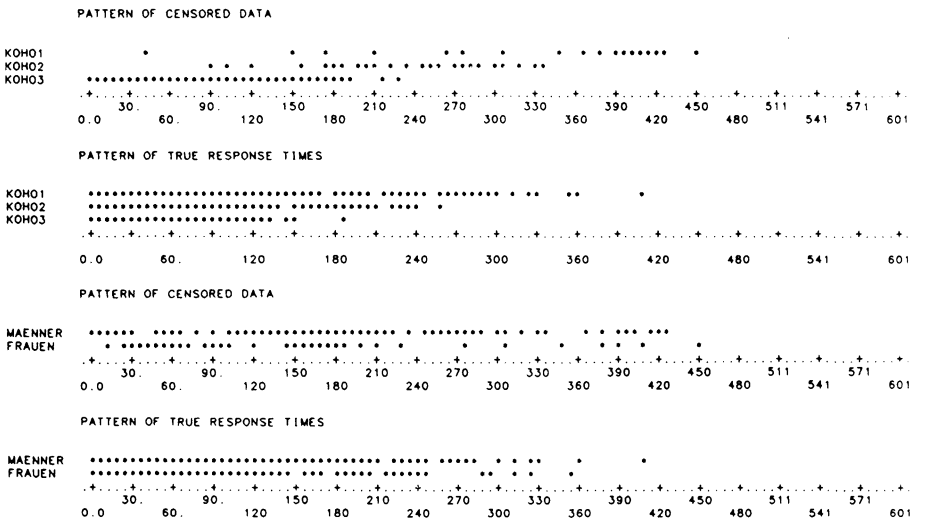
Die rein visuelle Überprüfung von Unterschieden zwischen den zu untersuchenden Subgruppen auf der Grundlage von Plots kann auch im BMDP durch die Durchführung statistischer Tests ergänzt werden (vgl. Abschnitt 3.2.5). Dazu berechnet BMDP die *Wilcoxon-Statistik in der Version von Breslow* (1970) und eine *Log-Rang-Statistik in der Version von Cox und Mantel* (1966). Beide Teststatistiken setzen voraus, daß sich die Survivorfunktionen nicht überschneiden und daß in den Subgruppen gleiche Zensierungsmuster vorherrschen. Die zweite dieser Annahmen kann bequem auf der Basis der von BMDP ausgedruckten *Zensierungs- und Ereignismuster* (Abbildung 4.10) beurteilt werden. Für unser Kohortenbeispiel zeigt sich dort ein von Kohorte zu Kohorte sich verschiebendes Zensierungs- und Ereignismuster. Die Erklärung ist in der Tatsache zu suchen, daß die jeweils älteren Kohorten bis zum Zeitpunkt des Interviews jeweils länger die Möglichkeit hatten, erwerbstätig zu sein, und somit

Abbildung 4.9: Beispiel für einen Plot der logarithmierten Survivorfunktion (bzw. kumulierten Hazardfunktion) auf der Basis einer Kaplan-Meier-Schätzung



auch längere Zensierungs- und Ereigniszeiten haben. Damit korreliert aber das Auftreten der Zensierungen mit der Kohortenzugehörigkeit, und die Voraussetzung der Tests zur Überprüfung von Kohortenunterschieden ist nicht gegeben. Beschränken wir uns bei der Interpretation der Testergebnisse deswegen auf den Vergleich von Männern und Frauen, wo sich die Zensierungszeiten über den gesamten Bereich ziemlich ähnlich verteilen und die Annahme konstanter Zensierungsmuster eher als erfüllt gelten kann (Abbildung 4.10). Beide Tests sind unter der Annahme, daß sich die Survivorfunktionen der zu untersuchenden Subpopulationen nicht unterscheiden, asymptotisch χ^2-verteilt, mit k–1 Freiheitsgraden (k = Anzahl der Subgruppen). Die Breslow-Statistik, die ebenso wie die Lee-Desu-Teststatistik in SPSS die Unterschiede der Survivorfunktionen zu Beginn des Prozesses betont, hat einen Wert von 0,079 (mit 1 d.f.) und ist bei einem Signifikanzniveau 0,05 nicht signifikant. Die Cox-Mantel-Statistik dagegen, die auf mit zunehmender Verweildauer sich verstärkende Unterschiede anspricht, ist bei einem Freiheitsgrad und einem χ^2-Wert von 8,745 signifikant (Signifikanzniveau = 0,05). Verschiedene Teststatistiken können also bei ein und denselben Survivorfunktionen (vgl. Abbildung 4.5) zu unterschiedlichen Resultaten kommen. Bei der Beurteilung der Testresultate ist deswegen nicht nur

Abbildung 4.10: Beispiel für den Ausdruck von Ereignis- und Zensierungsmustern



darauf zu achten, daß die Zensierungsmuster sich in etwa gleichen und sich die Survivorfunktionen nicht schneiden, sondern es sollte auf der Basis visueller Inspektion der Survivorplots auch vorher entschieden werden, welche der Teststatistiken für die jeweilige Situation das sensitivere und damit angemessenere Instrument ist.

Der Mehr-Episoden-Fall

Die Darstellung der verteilungsfreien Verfahren hat sich bisher auf den Fall beschränkt, daß nur eine Episode (z. B. die Verweildauer im ersten Beruf) und eine bestimmte Ereignisart (z. B. das Verlassen des ersten Berufs) vorliegt. Dieser Fall wird in den Wirtschafts- und Sozialwissenschaften allerdings nur selten auftreten. Gewöhnlich liegen dort pro Untersuchungseinheit mehrere Episoden vor, wie sie beispielsweise durch die Berufstätigkeitsepisoden in einer Erwerbskarriere oder die Wohnungsepisoden in einer Wanderungskarriere gegeben sind. Damit stellt sich die Frage, ob sich diese Verfahren auch zur Untersuchung des *Mehr-Episoden-Falls* heranziehen lassen und wie man dabei praktisch am besten vorgeht.

Zunächst ist es naheliegend, die verschiedenen Episoden eines Untersuchungsobjekts als eigenständige Einheiten zu betrachten, sie in einem ereignisorientierten Datensatz nacheinander abzuspeichern (Tabelle 4.1) und die gerade besprochenen Methoden zur Untersuchung aller Episoden heranzuziehen. Die entscheidende Frage bei diesem Vorgehen ist dann aber, ob die verschiedenen Episoden einer Untersuchungseinheit in dieser Weise verarbeitet werden dürfen

oder nicht. Personen mit unterschiedlich vielen Berufstätigkeiten wären dadurch beispielsweise in einem Datensatz unterschiedlich oft durch ihre Episoden repräsentiert und würden unterschiedlich oft in die Analyse eingehen. Dies wäre solange kein Problem, soweit es sich dabei um eine homogene Population, also um Personen mit gleichen Eigenschaften in bezug auf den zu untersuchenden Prozeß, handelt. Werden durch dieses Vorgehen aber heterogene Subpopulationen miteinander vermischt, so kann dies zu scheinbarer Zeitabhängigkeit und zu falschen Schlußfolgerungen führen (vgl. Abschnitt 3.9.1). Im Grunde geht es hier also um dieselbe Homogenitätsannahme, die implizit bereits bei der Darstellung der bisherigen Beispiele getroffen wurde. Denn auch hinter dem Verlauf der Survivorfunktionen von Männern und Frauen, beziehungsweise dem der drei Kohorten, können sich unterschiedliche Mischungen von Subpopulationen verbergen, die diese Verläufe erzeugen. Generell ist im Bereich der Wirtschafts- und Sozialwissenschaften, wo hohe Interdependenzbeziehungen zwischen den Merkmalen eher die Regel als die Ausnahme sind, die Verletzung der Homogenitätsannahme sehr wahrscheinlich. Man kann sich manchmal dadurch behelfen, daß man die Stichprobe nach diesen wichtigen Merkmalen disaggregiert. Leider stößt dieses Verfahren aber häufig nicht nur an die Grenze der mangelnden *Verfügbarkeit solcher Variablen im Datensatz* (Problem der unbeobachteten Heterogenität), sondern auch an das *Limit des Stichprobenumfangs*, der notwendig wäre, um für sehr kleine Subpopulationen noch aussagekräftige Survivorfunktionen schätzen zu können. Im Prozeß der Datenanalyse kann es sich bei den oben dargestellten Methoden deswegen nur um erste heuristische Verfahren handeln, die Aufschlüsse über die Struktur des Datenmaterials geben können.

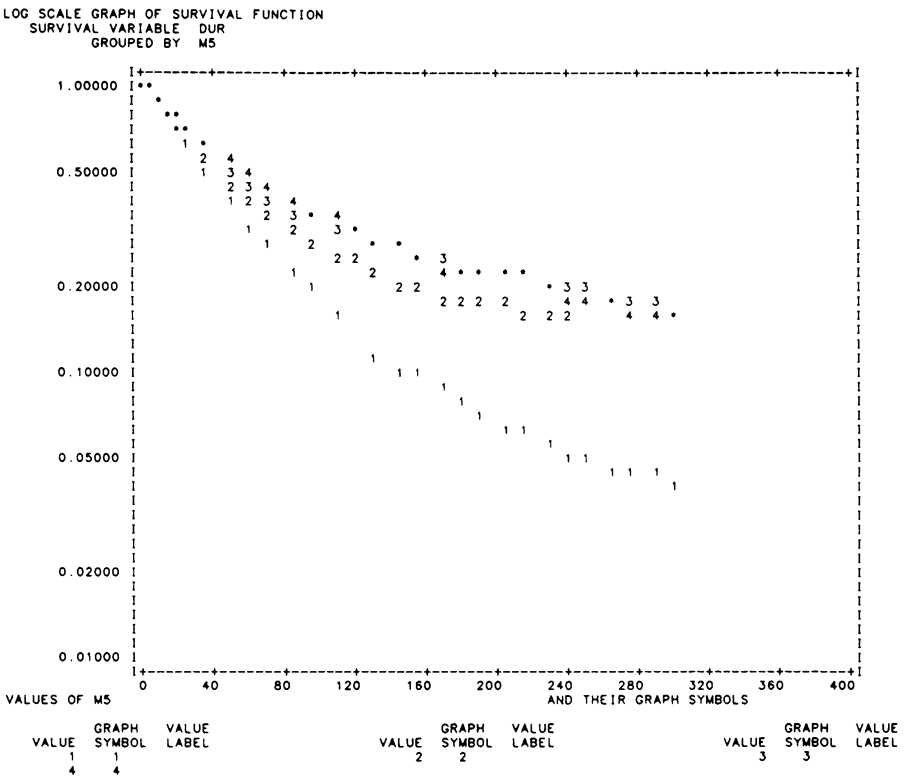
Im Mehr-Episoden-Fall der Erwerbskarriere kann die Survivoranalyse beispielsweise Hinweise darüber geben, ob die Verweildauer in einem Beruf von der Vorgeschichte abhängt; dazu muß man überprüfen, ob sich die Survivorfunktionen beispielsweise bei den ersten vier Erwerbstätigkeiten (mit denen man im Lebensverlaufs-Datensatz etwa 85 Prozent aller Erwerbstätigkeitsepisoden erfaßt) unterscheiden. Dazu soll uns das folgende SPSS-Programm dienen, welches verglichen mit dem Programmbeispiel 4.1 statt des Geschlechts (M3) und der Kohortenvariable (KOHO), die Sequenznummer der Erwerbstätigkeitsepisode (M5) als Gruppierungsmerkmal verwendet (vgl. Anhang 1):

Programmbeispiel 4.3:

```
GET FILE        DATA
COMPUTE         DUR = M51 - M50 + 1
COMPUTE         ZEN = 1
IF              (M51 EQ M47) ZEN = 0
SURVIVAL        TABLES = DUR BY M5 (1,4)/
                INTERVALS = THRU 24 BY 3,THRU 300 BY 12/
                STATUS = ZEN (1)/
                PLOTS (ALL)/
                COMPARE
OPTIONS         5,8
FINISH
```

Die mit diesem Programm geschätzten logarithmierten Survivorfunktionen zeigen, daß insbesondere zwischen dem ersten Beruf und den späteren Berufen größere Unterschiede bestehen (Abbildung 4.11). Der Kurvenverlauf des ersten Berufs ist deutlich nach unten gebogen und weist darauf hin, daß das Risiko, den ersten Beruf mit zunehmender Verweildauer zu verlassen, weit größer ist als bei den nachfolgenden Erwerbstätigkeitsepisoden. Je höher allerdings die Sequenznummer der Erwerbstätigkeit wird, desto geringer werden die Unterschiede zur jeweils vorhergehenden Erwerbstätigkeit. Insgesamt wird man aufgrund der Verlaufsmuster in Abbildung 4.11 davon ausgehen müssen, daß der Prozeß für verschiedene Berufsepisoden unterschiedlich ist, was bei der Schätzung von Regressionsmodellen (Cox- oder parametrischen Ratenmodellen) zu berücksichtigen ist.

Abbildung 4.11: Beispiel für einen Plot der logarithmierten Survivorfunktion (bzw. kumulierten Hazardfunktion) zur Überprüfung der Verteilung im Mehrepisodenfall



Der Mehr-Zustands-Fall

Eine letzte Erweiterung der Anwendungsmöglichkeiten der oben dargestellten Methoden besteht in der Einbeziehung des *Mehr-Zustands-Falls*. Hier liegt nicht nur eine Ereignisart (z. B. der Berufswechsel als solcher) vor, sondern es werden verschiedene Ereignisse (z. B. der Wechsel vom Arbeiter zum Angestellten und der Wechsel vom Arbeiter zum Selbständigen), die als miteinander konkurrierend betrachtet werden („competing risks"), unterschieden. Gerade im Bereich der Wirtschafts- und Sozialwissenschaften, wo meist Zustandsräume mit vielen Ausprägungen untersucht werden, ist dieser Fall von besonderer praktischer Bedeutung.
Die konkrete Realisierung der Mehr-Zustands-Modelle erfolgt, indem bei Betrachtung einer bestimmten Ereignisart (oder eines bestimmten Zustandswechsels) die jeweils konkurrierenden Ereignisse als zensiert behandelt werden.
Ein Anwendungsbeispiel soll die programmtechnische Vorgehensweise beim Mehr-Zustands-Fall demonstrieren. Dazu greifen wir wieder auf den bereits bekannten ereignisorientierten Erwerbstätigkeits-Datensatz zurück. Die Berufstätigkeiten der Erwerbsgeschichten wurden dazu in Berufsgruppen mit 12 Ausprägungen (siehe Tabelle 4.4) klassifiziert. Für jede Erwerbstätigkeitsepisode wurde nicht nur der jeweilige Berufsgruppenschlüssel (Ausgangszustand), sondern auch der Berufsgruppenschlüssel der jeweils nächsten Erwerbstätigkeitsepisode (Endzustand) abgespeichert. Handelt es sich um die letzte Berufstätigkeitsepisode, die keinen Nachfolger hat, dann wurde als Endzustand die Ausprägung 0 (zensiert) codiert. Der folgende SPSS-Lauf (vgl. Anhang 1) soll jetzt für den spezifischen Übergang von einem unqualifizierten manuellen Beruf (Ausprägung 2) zu einem qualifizierten manuellen Beruf (Ausprägung 3) die Sterbetafel jeweils getrennt für Männer und Frauen schätzen:

Programmbeispiel 4.4:

```
GET FILE        DATA
COMPUTE         DUR = M51 - M50 + 1
COMPUTE         ZEN = 1
IF              (M51 EQ M47) ZEN = 0
IF              (M62 NE 3) ZEN = 0
*SELECT IF      (M61 EQ 2)
SURVIVAL        TABLES = DUR BY M3(1,2)/
                INTERVALS = THRU 36 BY 3,THRU 180 BY 12/
                STATUS = ZEN(1) FOR DUR/
                PLOTS(LOGSURV,SURVIVAL)/
                COMPARE
FINISH
```

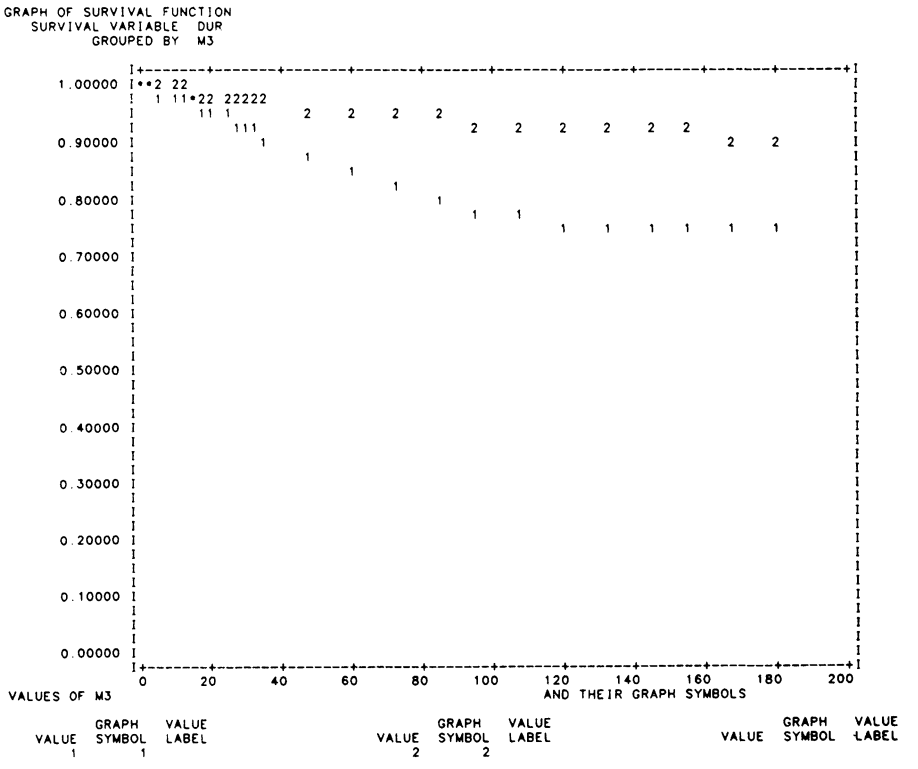
Nach der Berechnung der Verweildauer DUR wird die Zensierungsvariable ZEN bei allen Episoden auf 1 gesetzt. Fällt dann das Ende der Erwerbstätigkeitsepisode (M51) mit dem Zeitpunkt des Interviews (M47) zusammen, dann handelt es sich um eine zensierte Beobachtung, und ZEN wird mit einer 0 überschrieben. Die programmtechnische Umsetzung konkurrierender Ereig-

Tabelle 4.4: Klassifikation der Berufe

Bezeichnung der Berufsgruppe	Beschreibung der Berufsgruppe	Zusammensetzung der Berufsgruppen nach der deutschen Klassifizierung der Berufe (1970)	Zusammensetzung der Berufsgruppen nach der internationalen Klassifizierung der Berufe (ISCO)	Beispiele
Produktion				
Agrarberufe (AGR)	Berufe mit dominant landwirtschaftlicher Orientierung	011-022,041-051,053-062	6-11 bis 6-49	Landwirte, Tierzüchter, familieneigene Landarbeitskräfte, Gärtner, Waldarbeiter usw.
Einfache manuelle Berufe (EMB)	alle manuellen Berufe, die 1970 mindestens einen 60prozentigen Anteil von Ungelernten aufwiesen	071-133,135-141,143,151-162,164, 176-193,203-213,222-244,252,263, 301,313,321-323,332-346,352-371, 373,375-377,402-403,412,423-433, 442,452-463,465-472,482,486,504, 512-531,543-549	3-92,7-11 bis 7-23,7-26 bis 7-34,7-51 bis 7-61,7-71,7-72,7-74,7-75,7-77,7-79 bis 7-89,7-93 bis 7-95,7-99,8-02,8-12,8-20,8-35,8-39,8-72,8-91 bis 9-01, 9-10,9-25,9-39,9-42 bis 9-49,9-52,9-53,9-9-56,9-59,9-69,9-72 bis 9-74,9-82,9-99	Förderleute, Schrämhauer, Steinbrecher, Papier- und Zellstoffhersteller, Holzaufbereiter, Drukkerhelfer, Schweißer, Nieter, Löter, Hilfsarbeiter, Bauhelfer, Gleisbauer, Straßenbauer usw.
Qualifizierte manuelle Berufe (QMB)	alle manuellen Berufe, die 1970 höchstens einen 40prozentigen Anteil von Ungelernten aufwiesen	134,142,144,163,171-175,201-202, 221,251,261-262,270-291,302,305-312,314-315,331,351,372,374,378-401,411,421-422,441,451,464,481, 483-485,491-503,511,541-542	7-41 bis 7-49,9-02,9-26 bis 9-31,7-24, 7-25,8-31 bis 8-34,8-71,8-73 bis 8-80, 8-41 bis 8-59,9-41,7-91,8-01,8-03,7-62,7-92,7-76,7-73,5-31,7-78,9-51,9-54,9-55,9-57,7-96,8-11,8-19,9-61	Glasbläser, Buchbinder, Schriftsetzer, Schlosser, Feinmechaniker, Elektriker, Funk- und Fernsehgerätebauer, Weinküfer, Brauer, Zimmerer usw.
Techniker (TEC)	alle technischen Fachkräfte	303,304,621-635,721-722,733,857	0-75,0-32 bis 0-39,0-14,7-00,0-54,0-84,9-27,0-42,0-43,0-62,0-66,0-69	Maschinenbautechniker, Techniker des Elektrofaches, Bau- und Vermessungstechniker, Berg- und Hüttenbautechniker usw.
Ingenieure (ING)	hochqualifizierte Fachkräfte zur Lösung naturwissenschaftlicher und technischer Probleme	032,052,601-612,726,883	0-21 bis 0-31,0-11 bis 0-13,0-82,0-83 0-41,0-51 bis 0-53	Architekten, Bauingenieure, Elektroingenieure, Fertigungsingenieure, Chemiker, Physiker, Mathematiker usw.
Dienstleistung				
Einfache Dienste (EDI)	alle einfachen persönlichen Dienste	685-686,688,706,713-716,723-725, 741-744,791-794,805,838,911-913, 923-937	4-52,4-90,3-59,9-81,9-85 bis 9-89,3-91,9-71,9-79,5-89,5-51,5-92,1-75,1-80,3-94,5-00,5-10,5-32,5-40,5-52,5-60	Wascher, Raum- und Gebäudereiniger, Gastwirte, Kellner usw.
Qualifizierte Dienste (QDI)	im wesentlichen Ordnungs- und Sicherheitsberufe sowie qualifizierte Dienstleistungsberufe	684,704-705,711-712,801-804,812, 814,831,837,851-852,854-857,892-902,921-922	4-43,9-83,9-84,3-51,3-60,5-82,1-71, 1-63,0-79,0-76,0-72,5-99,1-49,5-70, 5-20,4-31,5-81	Polizisten, Feuerwehrleute, Makler, Schienenfahrzeugführer, Rechtspfleger, Photographen, Friseure, Hauswirtschaftsberater usw.
Semiprofessionen (SEMI)	Dienstleistungsberufe, die sich durch eine Verwissenschaftlichung der Berufspositionen auszeichnen	821-823,853,861-864,873-877	1-51,1-59,1-79,1-91,1-93,1-95,0-71, 0-73,0-74,0-77,1-94,1-33 bis 1-39	Krankenschwestern, Sozialarbeiter, Sozialpädagogen, Real- und Volksschullehrer usw.
Professionen (PROF)	Freie Berufe und hochqualifizierte Dienstleistungsberufe	811,813,841-844,871-872,881-882, 891	1-21 bis 1-29,0-61,0-63 bis 0-65,0-67, 0-68,1-31,1-32,0-81,0-90,1-92,1-99, 1-41	Zahnärzte, Ärzte, Apotheker, Richter, Gymnasiallehrer, Sozial- und Geisteswissenschaftler usw.
Verwaltung				
Einfache kaufmännische und Verwaltungsberufe (EVB)	relativ unqualifizierte Büro- und Handelsberufe	682,687,731-732,734,782-784,773	4-51,4-32,3-52,3-70,3-80,3-21,3-22, 3-99	Posthalter, Telefonisten, Verkäufer- und Verkaufshilfen, Kassierer, Maschinenschreiber, Bürohilfskräfte usw.
Qualifizierte kaufmännische und Verwaltungsberufe (QVB)	Berufe mit mittleren und höheren verwaltenden und distributiven Funktionen	031,681,683,691-703,771-772,774-781	6-00,4-00 bis 4-22,3-39,4-41,4-42,5-91,3-31,3-41,3-42,3-00,3-10,3-93,3-95	Bankfachleute, Speditionsfachleute, Großhandelskaufleute, Datenverarbeitungsfachleute, Bürofachkräfte usw.
Manager (MAN)	Berufe, die die Kontrolle und Entscheidungsgewalt über den Einsatz von Produktionsfaktoren besitzen, sowie Funktionäre in Organisationen	751-763	2-01 bis 2-19,1-10	Unternehmer, Geschäftsführer, Organisatoren, Geschäftsbereichsleiter, Abgeordnete, Minister, Verbandsleiter, Funktionare.

nisse besteht nun darin, in einem zweiten Schritt auch alle Episoden, die nicht durch den Übergang in den interessierenden Endzustand (M62) „qualifizierter manueller Beruf" (mit der Ausprägung 3) beendet werden, als zensiert zu betrachten. Dahinter verbirgt sich die Überlegung, daß diese Personen so lange dem Risiko eines Wechsels zu den qualifizierten manuellen Berufen ausgesetzt sind, bis eines der konkurrierenden Ereignisse (Wechsel in eine der anderen Berufsgruppen) eingetreten ist. Es ist deshalb notwendig, diese Verweildauern bei der Schätzung der Sterbetafel auch zu berücksichtigen. Mit der SELECT IF-Karte werden schließlich nur diejenigen Episoden ausgewählt, in denen sich eine Person tatsächlich im Ausgangszustand „unqualifizierter manueller Beruf" (mit der Ausprägung 2) befindet. Zu Beginn des Prozesses ist damit die eigentliche Risikomenge definiert. In Kombination mit der Zensierungsvariable wird dann für Männer und Frauen jeweils getrennt eine Sterbetafel für den spezifischen Übergang vom unqualifizierten manuellen Beruf zum qualifizierten manuellen Beruf geschätzt.

Abbildung 4.12: Beispiel eines Plots der Suvivorfunktion für den Übergang von einem unqualifizierten manuellen Beruf zu einem qualifizierten manuellen Beruf



Anhand der Survivorfunktionen für die Männer (1) und Frauen (2) (Abbildung 4.12) zeigt sich zunächst, daß der Übergang vom unqualifizierten manuellen Beruf zum qualifizierten manuellen Beruf außerordentlich träge verläuft. Nach etwa 120 Monaten (oder 10 Berufsjahren) Verweildauer in einem unqualifizierten manuellen Beruf haben nur etwa 25 Prozent der Männer und nur etwa 8 Prozent der Frauen diesen Übergang zum qualifizierten manuellen Beruf vollzogen. Danach bleiben, insbesondere für die Männer, die Survivorfunktionen weitgehend unverändert, das heißt, es findet kein Übergang mehr statt. Da die Survivorfunktion für die Männer bis zum 120. Monat etwas steiler verläuft als für die Frauen, ist die Chance dieses spezifischen Übergangs für die Männer auch etwas größer.
Ähnlich wie in diesem Beispiel kann man nun alle anderen Übergänge schätzen und so zu multiplen Sterbetafeln kommen, die eine adäquate Behandlung des Mehr-Zustands-Falls erlauben.

Zusammenfassung

Wenn wir die Ausführungen über die Sterbetafel-Methode und die Kaplan-Meier-Schätzung zusammenfassen, so kann man sagen, daß ihre Hauptfunktion im Prozeß der Datenanalyse darin besteht, einen ersten, eher heuristischen Einblick in den Prozeßverlauf zu geben. Im Mittelpunkt werden meist Subgruppenvergleiche stehen, die erste Aufschlüsse über die Wichtigkeit bestimmter Variablen geben können.
Die Annahme, daß in solche Vergleiche nur homogene Subpopulationen eingehen, ist allerdings im Bereich der Wirtschafts- und Sozialwissenschaften wenig plausibel, da dort in der Regel hohe Interkorrelationen zwischen den Merkmalen auftreten. Der Ausweg, durch Disaggregierung nur sehr spezifische Subgruppen miteinander zu vergleichen, stößt bei vielen Datensätzen sehr schnell an die Grenze des Stichprobenumfangs. Vergleiche zwischen den Subgruppen und die Beurteilung von Zeitabhängigkeiten auf der Basis von Survivorfunktionen oder kumulierten Survivorfunktionen dürften deswegen eher heuristischen Charakter haben. Besondere Vorsicht ist im Mehr-Episoden-Fall dann angebracht, wenn sich die Verteilungen der Episoden stark unterscheiden und zu vermuten ist, daß dadurch Personen mit bestimmten Eigenschaften überrepräsentiert in die Auswertungen eingehen. In den meisten Fällen läßt sich das Problem homogener Subpopulationen angemessener in den später zu besprechenden Regressionsmodellen lösen, wo durch die Einführung von Kovariablen nicht nur Subgruppenunterschiede, sondern auch die Vorgeschichte des Prozesses berücksichtigt werden können. Daß die Sterbetafel-Methode und die Kaplan-Meier-Schätzung auch zur Untersuchung von Mehr-Zustands-Modellen herangezogen werden können, ist ebenfalls an einem Beispiel demonstriert worden. Obgleich bei der Interpretation der Ergebnisse der dargestellten Verfahren sicherlich große Vorsicht angebracht ist, geben diese Methoden doch sehr hilfreiche Hinweise darauf, welche Variablen im weiteren Analyseprozeß von Bedeutung sind.

Kapitel 5:
Semiparametrische Regressionsmodelle: Das Proportional-Hazards-Modell von Cox

Bei der Sterbetafel-Methode und bei der Kaplan-Meier-Schätzung tritt nicht nur die Schwierigkeit auf, daß mit wachsender Zahl der zu kontrollierenden Subgruppen oft sehr schnell ein Punkt erreicht wird, von dem ab die *Schätzung von Survivorfunktionen wegen zu geringer Fallzahlen nicht mehr sinnvoll* scheint. Man wird auch mit dem Problem konfrontiert, daß die *Subgruppenvergleiche mit steigender Zahl unübersichtlich* werden und nur noch schwer zu interpretieren sind. In den letzten Jahren haben sich in der praktischen Anwendung deswegen zunehmend Regressionsansätze zur Analyse von Ereignisdaten durchgesetzt, bei denen die zur Beschreibung des Prozesses zentralen Hazardraten in Abhängigkeit von der Verweildauer und von Kovariablen modelliert werden. Damit sind nicht nur *zeitkonstante Subgruppenunterschiede* (wie etwa die Differenzen nach Geschlecht, Geburtskohorte usw.) und die *Einflüsse der Vorgeschichte* (z. B. die bereits gemachten beruflichen Erfahrungen) leicht zu berücksichtigen, sondern es lassen sich auch die *Wirkungen eines oder mehrerer paralleler Prozesse* über die Einführung zeitabhängiger Variablen analysieren.

Im Vergleich zur Sterbetafel-Methode oder zum Kaplan-Meier-Schätzer, bei denen die kontinuierlichen *Merkmale unter Informationsverlust klassifiziert* werden müssen, können in die Regressionsmodelle neben den qualitativen Kovariablen auch quantitative eingehen. Wo es unter meßtheoretischen Gesichtspunkten angemessen scheint, ergibt sich damit auch die Möglichkeit, eine Reihe von Effekten zu kontrollieren, indem man sie in metrisierter Form als Proxies in die Analysen aufnimmt. Bekannte Beispiele dafür sind die Einbeziehung der sozialen Schicht in Form von Statusscores (Treiman 1977; Handl/Mayer/Müller 1977; Wegener 1985) oder die Einbeziehung des formalen Qualifikationsniveaus durch die Verwendung der für die Ausbildung jeweils erforderlichen Zahl von Jahren (Helberger 1980; Blossfeld 1985c). Aber auch die Vorgeschichte kann beispielsweise bei Karriereanalysen flexibel durch die Anzahl der vorher ausgeübten Berufstätigkeiten oder der bereits im Beschäftigungssystem verbrachten Monate eingehen. Mit den so gebildeten Variablen können die entsprechenden Einflüsse in den Regressionsmodellen berücksichtigt werden, ohne daß dadurch die Zahl der zu schätzenden Parameter deutlich steigt.

In diesem Kapitel stehen zunächst die Anwendung des Cox-Modells und der

Partial-Likelihood-Schätzung im Mittelpunkt. Nach der Überprüfung der Proportionalitätsannahme (Abschnitt 5.1), wird ausführlich die Interpretation eines Cox-Modells aufgezeigt (Abschnitt 5.2). Die Auswahl von Modellen mit Hilfe der schrittweisen Regression wird in Abschnitt 5.3 demonstriert. Besonders wichtig für die Anwendung der Ereignisanalyse in den Wirtschafts- und Sozialwissenschaften sind schließlich die Beispiele, wie zeitveränderliche unabhängige Variablen in das Cox-Modell aufgenommen (Abschnitt 5.4) und wie Mehr-Zustands-Fälle praktisch gehandhabt werden (Abschnitt 5.5).

Der Einsatz von Cox-Modellen ist in der Regel dann angezeigt, wenn der Einfluß von Kovariablen ohne zusätzliche Annahmen über die Zeitabhängigkeit zu bestimmen ist (vgl. Abschnitt 3.3.3). Sei es, daß man manchmal keinerlei Vorinformationen über den zeitlichen Verlauf der Hazardrate hat, daß der bekannte Verlauf nicht adäquat durch ein parametrisches Modell erfaßt werden kann oder, daß bei Kontrolle der Veränderungen im Zeitablauf nur die Größe und die Richtung der Wirkung von Kovariablen interessieren. Das Cox-Modell ist damit außerordentlich flexibel und kann in vielen Situationen eingesetzt werden.

Die nicht näher spezifizierte Verweildauerabhängigkeit geht in das Cox-Modell als sogenannte Baseline-Hazardrate $\lambda_0(t)$ ein. Die Kovariablen werden in log-linearer Form $\exp(\mathbf{x}'\beta)$ in das Modell aufgenommen. Baseline-Hazardrate und log-linearer Kovariablenvektor werden schließlich multiplikativ miteinander verbunden:

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

Da die Baseline-Hazardrate nicht spezifiziert und unbekannt ist, hat Cox zur Schätzung der β-Koeffizienten des Modells die Partial-Likelihood-Methode vorgeschlagen (vgl. Abschnitt 3.6.4). Dabei werden nur für die tatsächlichen Ereigniszeitpunkte die Wahrscheinlichkeiten berücksichtigt, daß bei gegebener Risikomenge ein bestimmtes Individuum ein Ereignis erfährt. Die zensierten Beobachtungen verringern dagegen nur die Risikomenge der jeweils später eintretenden Ereignisse. Analog zum Kaplan-Meier-Schätzer müssen die Verweildauern deswegen bei der Schätzung in aufsteigender Reihenfolge sortiert werden. Die Bestimmung der β-Schätzwerte erfolgt schließlich in den Programmpaketen wie etwa BMDP, SAS, RATE und GLIM auf iterativem Wege. Zur iterativen Ermittlung der Schätzungen vergleiche man die Ausführungen in Abschnitt 3.6.1.

Stellt man die gegenwärtig vorhandenen Programmpakete zur Schätzung des Cox-Modells in bezug auf ihre Benutzerfreundlichkeit und ihre Möglichkeiten zur Modellauswahl und -evaluation einander gegenüber, so empfiehlt sich vor allem das Programmpaket BMDP. In GLIM und RATE müssen beispielsweise die bei der Schätzung notwendigen Sortierläufe der Verweildauern vom Benutzer selbst vorgenommen werden. Die folgenden Anwendungsbeispiele zur Partial-Likelihood-Schätzung beziehen sich deswegen nur auf das Unterprogramm P2L von BMDP.

5.1 Die Überprüfung der Proportionalitätsannahme

Trotz großer Einsatzbreite ist das Cox-Modell an die Annahme proportionaler Risiken gebunden. Damit ist gemeint, daß das Verhältnis der Hazardraten zweier beliebiger Individuen über die gesamte Verweildauer hinweg konstant ist. Haben zwei Individuen die zeitunabhängigen Kovariablenvektoren $\mathbf{x}_i$ und $\mathbf{x}_j$, so ist das Verhältnis der Hazardraten

$$\frac{\lambda(t|\mathbf{x}_i)}{\lambda(t|\mathbf{x}_j)} = \frac{\lambda_0(t)\exp(\mathbf{x}_i'\beta)}{\lambda_0(t)\exp(\mathbf{x}_j'\beta)} = \exp(\beta'(\mathbf{x}_i - \mathbf{x}_j))$$

unabhängig von t. Ein Beispiel für den Verlauf zweier proportionaler Hazardraten findet man in Abbildung 3.10.

Die graphische Überprüfung der Proportionalitätsannahme

Erste Hinweise darauf, ob die Proportionalitätsannahme tatsächlich erfüllt ist, erhält man, wenn die Stichprobe nach den Ausprägungen der in das Modell aufzunehmenden Variablen geschichtet wird und die subgruppenspezifischen Survivorfunktionen unter der Annahme geschätzt werden, daß der Einfluß der bereits in das Modell einbezogenen Kovariablen in allen Schichten identisch ist (vgl. dazu die Ausführungen in Abschnitt 3.7.2). Nach doppelt logarithmierter Transformation der Survivorfunktionen sollten sich die geplotteten Kurvenverläufe über die gesamte Verweildauer hinweg dann nur durch einen konstanten Faktor unterscheiden. Die Logik dieses Verfahrens soll anhand der Variablen Geschlecht demonstriert werden. Bezeichnet man die im Cox-Modell nicht näher spezifizierte Baseline-Survivorfunktion mit $S_0(t)$, dann lassen sich die Survivorfunktionen von Männern und Frauen wie folgt schreiben:

Männer $S_M(t|\mathbf{x}) = S_0(t)^{\exp(\mathbf{x}'\beta)\exp(\beta_g 1)}$

Frauen $S_F(t|\mathbf{x}) = S_0(t)^{\exp(\mathbf{x}'\beta)}$

Transformiert man diese Gleichungen durch doppelte Logarithmierung, dann erhält man:

Männer $\ln(-\ln S_M(t|\mathbf{x})) = \ln(-\ln S_0(t)) + \mathbf{x}'\beta + \beta_g$

Frauen $\ln(-\ln S_F(t|\mathbf{x})) = \ln(-\ln S_0(t)) + \mathbf{x}'\beta$

und damit:

$$\ln(-\ln S_M(t|\mathbf{x})) = \ln(-\ln S_F(t|\mathbf{x})) + \beta_g.$$

Plottet man die so transformierten geschätzten Survivorfunktionen von Männern und Frauen, dann dürften sich bei Vorliegen von Proportionalität beide

Kurven über die gesamte Verweildauer hinweg nur durch β_g unterscheiden. Wir gehen im folgenden stets aus von der Hazardrate

$$\lambda^k(t|x_k) = \lambda_0(t - t_{k-1}) \exp(x_k'\beta) \qquad k = 1, 2, \ldots,$$

das heißt, die Hazardrate hängt nur ab von der Verweildauer $v = t - t_{k-1}$ und die Grundhazardrate sowie die Einflußgewichte β sind für alle Episoden gleich. Nach den Ausführungen am Ende von Abschnitt 3.6.6 kann das Modell auf den Ein-Episoden-Fall zurückgeführt werden, wobei die aufeinanderfolgenden Episoden einer Person als unabhängig betrachtet werden. Sie werden jeweils als neue Episoden in das Modell aufgenommen. Daß es sich um die k-te Episode eines Individuums handelt, kommt lediglich in den Kovariablen zum Ausdruck, die zum Teil, etwa bei der Berufserfahrung (BERF), jeweils am Beginn der Episode gemessen werden. Im folgenden schreiben wir statt $\lambda^k(t|x_k)$ stets $\lambda^k(v|x_k)$.
Ein Beispiel für die Schätzung eines Cox-Modells mit dem Programmpaket BMDP liefert der folgende Programmlauf. Wir greifen dabei wieder auf den bereits bekannten ereignisorientierten Erwerbstätigkeits-Datensatz zurück (vgl. dazu auch Anhang 1). Nach Einführung der Variablen Bildung (BILDG), Berufserfahrung (BERF), Anzahl der vorher ausgeübten Berufe (BANZ) und Kohorte (KOHO2, KOHO3) in das Cox-Modell soll für alle Erwerbstätigkeitsepisoden (Mehr-Episoden-Fall) überprüft werden, ob die Risiken von Männern und Frauen über die gesamte Verweildauer hinweg zueinander proportional sind.

Programmbeispiel 5.1:

```
/INPUT UNIT IS 30.
       CODE IS DATA.
/VARIABLE NAMES ARE (63)DUR,(64)ZEN,(65)BILDG,(66)BERF,
                    (67)BANZ,(68)KOHO2,(69)KOHO3.
          ADD IS 7.
/TRANSFORM DUR = M51 - M50 + 1.
           ZEN = 1.
           IF (M51 EQ M47) THEN ZEN = 0.
           IF (M41 EQ 1 AND M42 EQ 1) THEN BILDG = 9.
           IF (M41 EQ 1 AND (M42 EQ 2 OR M42 EQ 3)) THEN BILDG = 11.
           IF (M41 EQ 2 AND M42 EQ 1) THEN BILDG = 10.
           IF (M41 EQ 2 AND (M42 EQ 2 OR M42 EQ 3)) THEN BILDG = 12.
           IF (M41 EQ 3 AND (M42 EQ 1 OR M42 EQ 2 OR M42 EQ 3))
           THEN BILDG = 13.
           IF (M42 EQ 4) THEN BILDG = 17.
           IF (M42 EQ 5) THEN BILDG = 19.
           KOHO2 = 0.
           KOHO3 = 0.
           IF(M48 GE 468 AND M48 LE 504) THEN KOHO2 = 1.
           IF(M48 GE 588 AND M48 LE 624) THEN KOHO3 = 1.
```

```
          BERF = M50 - M43.
          BANZ = M5 - 1.
/FORM TIME IS DUR.
      STATUS IS ZEN.
      RESPONSE IS 1.
/REGRESSION COVARIATES ARE BILDG,M59,BANZ,BERF,
                           KOHO2,KOHO3.
            STRATA IS M3.
/GROUP CODES (3) ARE 1,2.
       NAMES (3) ARE 'MAENNER','FRAUEN'.
/TEST ELIMINATE = BILDG,M59,BANZ,BERF,KOHO2,KOHO3.
      STATISTICS = WALD,LRATIO,SCORE.
/PLOT TYPE = LOG.
/PRINT CASES ARE 0.
/END
```

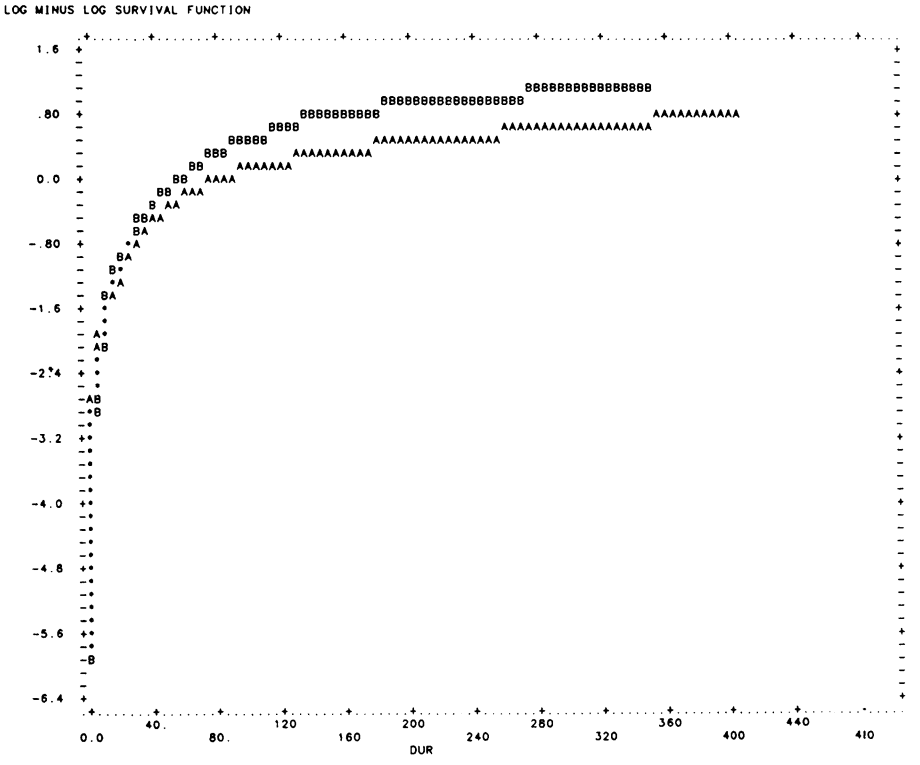
Im TRANSFORM-Paragraph wird im obigen Lauf zunächst für jede Berufsepisode die Verweildauer (DUR) berechnet und die Zensierungsinformation (ZEN) erzeugt. Da das Ausbildungsniveau der Individuen zu Beginn jeder Erwerbstätigkeitsepisode in metrisierter Form in das Cox-Modell aufgenommen werden soll, wird der Variablen BILDG die durchschnittlich für einen Ausbildungsabschluß benötigte Zahl von Schuljahren zugeordnet. Bei der Zuordnung von Schuljahren zu Schulabschlüssen wird von folgenden Werten ausgegangen: Hauptschulabschluß ohne Berufsausbildung entspricht 9 Jahren, Hauptschulabschluß mit Berufsausbildung entspricht 11 Jahren, Mittlere Reife ohne Berufsausbildung entspricht 10 Jahren, Mittlere Reife mit Berufsausbildung entspricht 12 Jahren, Abitur entspricht 13 Jahren, Fachhochschulabschluß entspricht 17 Jahren und Hochschulabschluß entspricht 19 Jahren. Die Unterscheidung der drei Geburtskohorten soll im Cox-Modell durch die Einführung von Dummy-Variablen erfolgen. Dabei wird die älteste Kohorte (die 1929–31 Geborenen) als Referenzkategorie gewählt, während die Kohorten von 1939–41 und 1949–51 durch die Variablen KOHO2 und KOHO3 eingehen:

	KOHO2	KOHO3
Kohorte 1929–31	0	0
Kohorte 1939–41	1	0
Kohorte 1949–51	0	1

Jeweils zu Beginn jeder Erwerbstätigkeitsepisode wird die Berufserfahrung seit Eintritt in das Beschäftigungssystem (BERF) in Monaten gemessen. Die Anzahl der vor der jeweiligen Erwerbstätigkeitsepisode bereits ausgeübten Berufe wird schließlich in der Variablen BANZ abgespeichert.
Im FORM-Paragraph wird dem BMDP-Programm die Verweildauervariable (TIME IS DUR) und die Zensierungsvariable (STATUS IS ZEN) übergeben. Letztere hat bei einem Ereignis den Wert 1 (RESPONSE IS 1).

Mit dem REGRESSION-Paragraph wird ein Cox-Modell spezifiziert, in das die Variablen BILDG, M59 (Prestige) (vgl. Wegener 1985), BANZ, BERF, KOHO2 und KOHO3 eingehen. Da die Variable Geschlecht (M3) zunächst auf ihre Proportionalität hin überprüft werden soll, wird sie als Schichtungsvariable in das Modell einbezogen (STRATA IS M3). Mit dem GROUP-Paragraph werden die entsprechenden Labels zugeordnet. Der Auftrag, die Log-minus-log-Survivorfunktion jeweils für die Männer und Frauen zu plotten, erfolgt im PLOT-Paragraph durch die Anweisung TYPE=LOG. Schließlich werden mit dem TEST-Paragraph ein Partial-Likelihood-Quotiententest (LRATIO), ein Wald-Test (WALD) und ein Score-Test (SCORE) zur Überprüfung der Hypothese angefordert, daß keine der eingeführten Kovariablen einen signifikanten Einfluß auf die Rate hat.

Abbildung 5.1 Beispiel für einen Plot der Log-minus-log-Survivorfunktion zur Überprüfung der Proportionalitätsannahme



Interpretieren wir zuerst Abbildung 5.1, in der die Kurvenverläufe der Log-minus-log-Survivorfunktion von Männern (A) und Frauen (B) gegen die Ver-

weildauer (DUR) geplottet werden. Bei Erfüllung der Proportionalitätsannahme müßte für alle Zeitpunkte der Verweildauer der senkrechte (d. h. parallel zur y-Achse verlaufende) Abstand zwischen beiden Kurven unverändert bleiben. In Abbildung 5.1 ist zunächst für die ersten 40 Monate wegen der starken Steigung beider Survivorfunktionen aus dem Plot der Abstand zwischen den beiden Kurven überhaupt nicht zu erkennen. Danach zeigt sich, daß mit zunehmender Verweildauer der Abstand zwischen beiden Kurven größer wird. Da die Zunahme des Abstands aber nicht erheblich ist, stellt sich die Frage, ob dadurch die Proportionalitätsannahme bereits verletzt wird oder ob die Abweichung in tolerierbaren Grenzen liegt. Die bloße visuelle Inspektion der geplotteten Kurven kann darüber allerdings keine Auskunft geben. Wünschenswert ist deswegen ein statistischer Test.

Ein Test zur Überprüfung der Proportionalitätsannahme

Im Programmpaket BMDP läßt sich ein solcher Test auf elegante Weise über die Einführung einer zeitabhängigen Variablen realisieren. Wie in Abschnitt 3.7.2 bereits ausgeführt, darf es zwischen der zu überprüfenden Variablen und der Verweildauer keinen signifikanten Interaktionseffekt geben, wenn das Verhältnis der Hazardraten zweier beliebiger Individuen mit den Kovariablen-Vektoren $\mathbf{x}_i$ und $\mathbf{x}_j$ unabhängig von der Verweildauer v ist. Im Falle der Variablen Geschlecht, die als Dummy-Variable in die Regressionsrechnung aufzunehmen ist, definieren wir folgende Interaktionsvariable[1]:

$$z = x_g\,(\ln v - \ln c) \qquad \text{mit } x_g = \begin{cases} 1 \text{ für Männer} \\ 0 \text{ für Frauen} \end{cases}$$

Die Regressionsgleichung im Cox-Modell lautet dann wie folgt:

$$\lambda^k(v|\mathbf{x}) = \lambda_0(v)\exp(\mathbf{x}_k'\beta + \beta_g x_g + \beta_z z), \qquad , k = 1, 2, \ldots,$$

oder:

$$\lambda^k(v|\mathbf{x}) = \lambda_0(v)\exp(\mathbf{x}_k'\beta + \beta_g x_g)\,(\frac{v}{c})^{\beta_z x_g} \qquad , k = 1, 2, \ldots.$$

Wenn die Proportionalitätsannahme erfüllt ist, dann darf der Koeffizient β_z nicht signifikant von Null verschieden sein, so daß sich die Hazardraten von Männern und Frauen nur durch den konstanten Faktor $\exp(\beta_g)$ voneinander unterscheiden.
Die Realisierung dieses Tests kann in BMDP durch geringfügige Modifikation des Programmbeispiels 5.1 erfolgen:

---

[1] Zur besseren Schätzbarkeit der Parameter β_g und β_z wird neben der logarithmierten Verweildauer noch der logarithmierte Mittelwert der Verweildauer berücksichtigt. Eine Schätzung des Mittelwerts der Verweildauer (c) erhält man durch einen Lauf mit dem BMDP-Unterprogramm P1L (siehe Tabelle 4.3).

Programmbeispiel 5.2:

```
/ INPUT UNIT IS 30.
        CODE IS DATA.
/VARIABLE NAMES ARE (63)DUR,(64)ZEN,(65)BILDG,
                    (66)BERF,(67)BANZ,(68)KOHO2,
                    (69)KOHO3,(70)GESCHL.
          ADD IS 8.
/TRANSFORM DUR = M51 - M50 + 1.
           ZEN = 1.
           IF (M51 EQ M47) THEN ZEN = 0.
           IF (M41 EQ 1 AND M42 EQ 1) THEN BILDG = 9.
           IF (M41 EQ 1 AND (M42 EQ 2 OR M42 EQ 3)) THEN BILDG = 11.
           IF (M41 EQ 2 AND M42 EQ 1) THEN BILDG = 10.
           IF (M41 EQ 2 AND (M42 EQ 2 OR M42 EQ 3)) THEN BILDG = 12.
           IF (M41 EQ 3 AND (M42 EQ 1 OR M42 EQ 2 OR M42 EQ 3))
           THEN BILDG = 13.
           IF (M42 EQ 4) THEN BILDG = 17.
           IF (M42 EQ 5) THEN BILDG = 19.
           KOHO2 = 0.
           KOHO3 = 0.
           IF (M48 GE 468 AND M48 LE 504) THEN KOHO2 = 1.
           IF (M48 GE 588 AND M48 LE 624) THEN KOHO3 = 1.
           BERF = M50 - M43.
           BANZ = M5 - 1.
           GESCHL = 0.
           IF (M3 EQ 1) THEN GESCHL = 1.
/FORM TIME IS DUR.
      STATUS IS ZEN.
      RESPONSE IS 1.
/REGRESSION COVARIATES ARE BILDG,M59,BANZ,BERF,KOHO2,
                           KOHO3,GESCHL.
           ADD IS Z2.
/FUNCTION Z2 = GESCHL * (LN(TIME)-LN(87.37)).
/PRINT CASES ARE 0.
/END.
```

Nach der Definition der neuen Variablen GESCHL im Paragraph VARIABLES wird im TRANSFORM-Paragraph des obigen Programmbeispiels aus der Variablen M3, welche die Ausprägungen 1 für Männer und 2 für Frauen enthält, die Dummy-Variable GESCHL konstruiert. Diese nimmt den Wert 1 an, wenn es sich um einen Mann handelt, und sonst den Wert 0. Die Frauen sind in diesem Beispiel also die Referenzgruppe.
Im REGRESSION-Paragraph wird ein Cox-Modell spezifiziert, welches im Vergleich zu Programmbeispiel 5.1 als zusätzliche Kovariable die Dummy-Variable GESCHL einschließt. Danach wird mit dem ADD-Statement die Interaktionsvariable Z2 als zeitveränderliche Kovariable einbezogen. Z2 wird im FUNCTION-Paragraph konstruiert und stellt die Interaktion zwischen der Variablen GESCHL und der Verweildauer TIME dar, wobei letztere noch durch

den Logarithmus der durchschnittlichen Verweildauer von 87,37 Monaten gewichtet wird.
Als Ergebnis dieses Laufs[2] erhält man für die Interaktionsvariable Z2 einen signifikanten Effekt ($\hat{\beta}_z/SE(\hat{\beta}_z) = -0{,}2071/0{,}0282 = -7{,}3548$). Das Verhältnis der Hazardraten von Männern und Frauen, das zu Beginn des Prozesses den Wert $\exp(\hat{\beta}_g) = 0{,}5803$ hat, ist also über die Verweildauer hinweg nicht konstant, sondern verändert sich. Die Proportionalitätsannahme ist damit nicht erfüllt, und die Variable Geschlecht kann in diesem Fall nicht als Kovariable, sondern muß als Schichtungsvariable in das Cox-Modell aufgenommen werden.

5.2 Zur Interpretation der Schätzergebnisse

Interpretieren wir deswegen nur die Ergebnisse des Programmbeispiels 5.1 (Tabelle 5.1), in dem das Cox-Modell unter den Annahmen schichtspezifischer Baseline-Funktionen $\lambda_{0j}(v)$ für die Subgruppen der Männer (j = 1) und Frauen (j = 2) sowie identischer Kovariableneinflüsse über diese beiden Schichten im Mehrepisodenfall geschätzt wurde:

$$\lambda_j^k(v|x_k) = \lambda_{0j}(v)\exp(x_k'\beta) \qquad j = 1,2 \qquad k = 1, 2, \ldots.$$

Zunächst liefert die GLOBAL CHI-SQUARE-Statistik einen Anhaltspunkt dafür, wieviel das geschätzte Modell, verglichen mit einem Modell ohne Kovariablen, zusätzlich an Heterogenität erklärt. Unter der Nullhypothese, daß keine der Kovariablen einen Einfluß auf den Jobwechsel hat, ist diese Prüfgröße asymptotisch χ^2-verteilt mit p (= Anzahl der geschätzten Parameter) Freiheitsgraden:

$$\text{GLOBAL CHI-SQUARE} = s(\hat{\beta}_0)' \, I(\hat{\beta}_0)^{-1} \, s(\hat{\beta}_0),$$

wobei β_0 gleich Null gesetzt wird. Die Prüfgröße entspricht der Teststatistik (3.7.20) des Score-Tests aus Abschnitt 3.7.3. Ein χ^2-Wert von 667,39 mit sechs Freiheitsgraden (vgl. Tabelle 5.1) signalisiert, daß die Nullhypothese, keine der

---

[2] Ergebnis des Cox-Modells aus Programmbeispiel 5.2

LOG LIKELIHOOD = -37109.8619
GLOBAL CHI-SQUARE = 845.12 D.F.= 8 P-VALUE =0.0000

	VARIABLE	COEFFICIENT	STANDARD ERROR	COEFF./S.E.	EXP(COEFF.)
65	BILDG	0.0261	0.0100	2.6030	1.0265
59	M59	-0.0078	0.0010	-7.5614	0.9922
67	BANZ	0.1470	0.0100	14.7182	1.1584
66	BERF	-0.0070	0.0003	-22.2626	0.9930
68	KOHO2	0.1031	0.0350	2.9477	1.1086
69	KOHO3	0.1780	0.0391	4.5561	1.1948
70	GESCHL	-0.5442	0.0442	-12.3025	0.5803
71	Z2	-0.2071	0.0282	-7.3548	0.8129

Tabelle 5.1: Ergebnis des Cox-Modells aus Programmbeispiel 5.1

```
   LOG LIKELIHOOD = -33877.3618
GLOBAL CHI-SQUARE =    667.39  D.F.=   6  P-VALUE =0.0000

                                    STANDARD
    VARIABLE       COEFFICIENT        ERROR    COEFF./S.E.    EXP(COEFF.)
    --------       -----------      --------   -----------    -----------
65 BILDG              0.0258          0.0100       2.5628         1.0261
59 M59               -0.0077          0.0010      -7.5078         0.9923
67 BANZ               0.1470          0.0100      14.7181         1.1584
66 BERF              -0.0070          0.0003     -22.2621         0.9930
68 KOHO2              0.1038          0.0350       2.9690         1.1094
69 KOHO3              0.1766          0.0391       4.5218         1.1932


  STATISTIC       CHI-SQUARE    D.F.    P-VALUE
    LRATIO           821.35       6     0.0000
    SCORE            667.39       6     0.0000
    WALD             653.40       6     0.0000
```

eingeführten Variablen erkläre etwas, zu verwerfen ist (P-VALUE = 0,0000). Zu dem gleichen Ergebnis kommen auch die anderen Teststatistiken, die explizit im TEST-Paragraph angefordert wurden und mit denen in diesem Beispiel ebenfalls das gesamte Modell getestet wird. Der Partial-Likelihood-Quotiententest (vgl. Abschnitt 3.7.3)

$$\text{LRATIO} = -2\ln\left(\frac{\text{L (Modell ohne Kovariablen)}}{\text{L (aktuelles Modell)}}\right)$$

hat dabei einen χ^2-Wert von LRATIO = 821,35.
Die Wald-Statistik

$$\text{WALD} = \hat{\beta}' \, \text{Cov}(\hat{\beta})^{-1} \, \hat{\beta},$$

die sich aus der allgemeinen Wald-Teststatistik (3.7.19) ergibt, wenn für $\hat{\theta}$ der Wert $\hat{\beta}$ und für C die Einheitsmatrix gesetzt wird, hat einen Wert von WALD = 653,40 bei sechs Freiheitsgraden.
Welche der Variablen nun tatsächlich einen signifikanten Einfluß hat, kann im einzelnen in der Spalte der standardisierten Koeffizienten (COEFF./S.E.) abgelesen werden. Nach den Ausführungen in Abschnitt 3.7.3 handelt es sich dabei um asymptotisch normalverteilte Schätzwerte, die sich bei einer Irrtumswahrscheinlichkeit von $\alpha = 0{,}05$ signifikant von 0 unterscheiden, wenn ihr Betrag den Wert von 1,96 übersteigt. Anhand dieses Kriteriums beurteilt, haben alle in das obige Modell einbezogenen Variablen einen signifikanten Effekt (vgl. Tabelle 5.1).
Die Wirkung einer Kovariablen x_i läßt sich besonders anschaulich interpretieren, wenn man bei Konstanthaltung der jeweils anderen Variablen $x'\beta$ zeigt, um

wieviel Prozent sich die Rate bei Erhöhung der Kovariablen x_i um einen bestimmten Wert Δx_i verändert:

$$\zeta_{\Delta x_i} = \frac{\lambda_0(v)\exp(x'\beta + \beta_i(x_i + \Delta x_i)) - \lambda_0(v)\exp(x'\beta + \beta_i x_i)}{\lambda_0(v)\exp(x'\beta + \beta_i x_i)} \cdot 100\%$$

$$\zeta_{\Delta x_i} = \frac{\exp(\beta_i(x_i + \Delta x_i)) - \exp(\beta_i x_i)}{\exp(\beta_i x_i)} \cdot 100\%$$

$$\zeta_{\Delta x_i} = (\exp(\beta_i \Delta x_i) - 1) \cdot 100\%$$

oder:

$$\zeta_{\Delta x_i} = (\exp(\beta_i)^{\Delta x_i} - 1) \cdot 100\%.$$

Die Antilogarithmen $\exp(\beta_i)$ der β_i-Koeffizienten sind in der Spalte EXP (COEFF.) von Tabelle 5.1 zu finden und werden in der Literatur als α_i-Effekte bezeichnet. Sie nehmen den Wert 1 an, wenn die Variable keinen Effekt auf die Rate hat ($\beta_i = 0$), und einen Wert kleiner beziehungsweise größer als 1, wenn die Variable einen vermindernden ($\beta_i < 0$) beziehungsweise erhöhenden ($\beta_i > 0$) Einfluß ausübt. Erhöht man den Wert der Variablen x_i um genau eine Einheit, dann verändert sich die Rate also um

$$\zeta_1 = (\exp(\beta_i) - 1) \cdot 100\%.$$

Der $\hat{\beta}$-Koeffizient der Variablen Bildung (BILDG) hat ein positives Vorzeichen (vgl. Tabelle 5.1). Jedes zusätzliche Schuljahr steigert damit die Neigung zum Berufswechsel um 2,61 Prozent [$(\exp(\hat{\beta}_{BILDG}) - 1) \cdot 100\% = (1{,}0261 - 1) \cdot 100\% = 2{,}61\%$]. Erhöhend auf die Rate wirkt sich auch die Anzahl der vorher ausgeübten Berufstätigkeiten (BANZ) aus. Jede dieser Tätigkeiten steigert die Neigung, den gegenwärtigen Beruf zu verlasssen, um 15,84 Prozent. Schließlich haben auch die $\hat{\beta}$-Koeffizienten der Geburtskohorten-Dummys einen positiven Wert. Im Vergleich zur ältesten Kohorte, den zwischen 1929 und 1931 Geborenen, ist die Neigung zum Jobwechsel bei den zwischen 1939 und 1941 Geborenen (KOHO2) um 10,94 und bei den zwischen 1949 und 1951 Geborenen (KOHO3) um 19,32 Prozent höher.
Der $\hat{\beta}$-Koeffizient der Variablen Prestige (M59) ist dagegen negativ. Je höher ein Beruf in der Hierarchie der Berufspyramide angesiedelt ist, desto geringer ist die Tendenz, den Beruf zu verlassen. Die gleiche Wirkungsrichtung hat auch die Variable Berufserfahrung (BERF), welche die Rate um 0,70 Prozent pro Monat senkt.
Wollte man die Wirkung der Berufserfahrung in Jahren statt in Monaten ausdrücken, dann wäre aufgrund der log-linearen Formulierung im Cox-Modell besondere Vorsicht geboten. Die Verminderung der Rate bei einem Jahr Berufserfahrung berechnet sich nämlich nicht wie bei den normalen Regressionsmodel-

len durch $12 \cdot (-0{,}7\%) = -8{,}4\%$, sondern ergibt sich nach obiger Formel über die Beziehung $((\exp(\hat{\beta}_{BERF}))^{12} - 1) \cdot 100\% = (0{,}9923^{12} - 1) \cdot 100\% = -8{,}86\%$.
Es ist bei log-linearen Kovariableneinflüssen außerdem darauf hinzuweisen, daß sich gleichzeitige Veränderungen von zwei und mehr Kovariablen nicht additiv, sondern multiplikativ auswirken:

$$\gamma_{\Delta x} = \frac{\lambda_0(v) \exp(\beta_1 (x_1 + \Delta x_1) + \ldots + \beta_p (x_p + \Delta x_p)) - \lambda_0(v) \exp(\beta_1 x_1 + \ldots + \beta_p x_p)}{\lambda_0(v) \exp(\beta_1 x_1 + \ldots + \beta_p x_p)} \cdot 100\%$$

$$\gamma_{\Delta x} = \frac{\exp(\beta_1 (x_1 + \Delta x_1) + \ldots + \beta_p (x_p + \Delta x_p)) - \exp(\beta_1 x_1 + \ldots + \beta_p x_p)}{\exp(\beta_1 x_1 + \ldots + \beta_p x_p)} \cdot 100\%$$

$$\gamma_{\Delta x} = (\exp(\beta_1)^{\Delta x_1} \cdot \ldots \cdot \exp(\beta_p)^{\Delta x_p} - 1) \cdot 100\%$$

Erhöht man dementsprechend die Anzahl der vorher ausgeübten Berufe (BANZ) um eine Einheit und das Prestige (M59) um 20 Einheiten, was einem beruflichen Aufstieg entsprechen würde, dann verändert sich das Risiko, den Job zu verlassen, um $(1{,}1584^1 \cdot 0{,}9923^{20} - 1) \cdot 100\% = -0{,}75\%$.

5.3 Die schrittweise Regression im Cox-Modell

Obwohl man die Auswahl von Variablen und die Interpretation von Ergebnissen in der Regel auf der Grundlage theoretischer Vorüberlegungen und inhaltlicher Hypothesen vornehmen wird, können manchmal Situationen auftreten, in denen man nur vage Vermutungen über die Wirkungsrichtung der Variablen und die Zusammenhänge im Datenmaterial hat. Eine wichtige Hilfe der Modellbildung am empirischen Material kann dann die Methode der schrittweisen Regression (stepwise regression) sein, die bereits in Abschnitt 3.7.3 beschrieben und auch im Unterprogramm P2L von BMDP implementiert ist. Dabei wird aus einem vorher spezifizierten Pool von Variablen Schritt für Schritt auf der Grundlage von Signifikanzüberprüfungen entschieden, ob eine dieser Variablen in das Cox-Modell einbezogen oder ausgeschlossen werden soll. Das Ergebnis ist ein für den gegebenen Variablen-Pool und nach den jeweils gewählten statistischen Kriterien optimales Modell.
In BMDP kann man sich zwischen zwei Methoden entscheiden, die den stufenweisen Auswahlprozeß steuern: einem Partial-Likelihood-Quotiententest (**m**aximum **p**artial **l**ikelihood **r**atio **t**est, MPLR) und der Teststatistik von **P**eduzzy, **H**ardy und **H**olford (PHH). Zwar ist MPLR das statistisch zufriedenstellendere und genauere Verfahren, doch benötigt es weit mehr Rechenzeit als PHH. Beide Verfahren können auch bei ein und demselben Variablen-Pool zu unterschiedlichen Ergebnissen kommen. Bei einer großen Zahl von Variablen ist deswegen zu empfehlen, zuerst mit PHH alle Variablen ohne Wirkung oder mit geringen Einflüssen auf die Hazardrate zu eliminieren und dann in einem zweiten Schritt

mit MPLR ein geeignetes Modell zu suchen. Dabei ist zu berücksichtigen, daß ein multiples Testproblem vorliegt, was zu einer Erhöhung des α-Fehlers führen kann.
Für jede Variable kann in BMDP spezifiziert werden, ob sie bereits im ersten Schritt in das Modell einbezogen werden soll, und man kann angeben, wie oft jede der Variablen während der schrittweisen Regression herausgenommen beziehungsweise aufgenommen werden darf. Der Auswahlprozeß läßt sich darüber hinaus noch über die Vergabe von Anfangswerten für die zu schätzenden Parameter und die Festlegung einer Reihe von Abbruchkriterien für den dabei verwendeten Newton-Raphson-Algorithmus (vgl. dazu Abschnitt 3.6) steuern.
Ein einfaches Beispiel für die Anwendung der schrittweisen Regression in BMDP liefert der folgende Lauf, mit dem das Programmbeispiel 5.1 etwas modifiziert wurde.

Programmbeispiel 5.3:

```
/INPUT UNIT IS 30.
       CODE IS DATA.
/VARIABLE NAMES ARE (63)DUR,(64)ZEN,(65)BILDG,(66)BERF,
                    (67)BANZ,(68)KOHO2,(69)KOHO3.
          ADD IS 7.
/TRANSFORM DUR = M51 - M50 + 1.
           ZEN = 1.
           IF (M51 EQ M47) THEN ZEN = 0.
           IF (M41 EQ 1 AND M42 EQ 1) THEN BILDG = 9.
           IF (M41 EQ 1 AND (M42 EQ 2 OR M42 EQ 3)) THEN BILDG = 11.
           IF (M41 EQ 2 AND M42 EQ 1) THEN BILDG = 10.
           IF (M41 EQ 2 AND (M42 EQ 2 OR M42 EQ 3)) THEN BILDG = 12.
           IF (M41 EQ 3 AND (M42 EQ 1 OR M42 EQ 2 OR M42 EQ 3))
           THEN BILDG = 13.
           IF (M42 EQ 4) THEN BILDG = 17.
           IF (M42 EQ 5) THEN BILDG = 19.
           KOHO2 = 0.
           KOHO3 = 0.
           IF(M48 GE 468 AND M48 LE 504) THEN KOHO2 = 1.
           IF(M48 GE 588 AND M48 LE 624) THEN KOHO3 = 1.
           BERF = M50 - M43.
           BANZ = M5 - 1.
/FORM TIME IS DUR.
      STATUS IS ZEN.
      RESPONSE IS 1.
/REGRESSION COVARIATES ARE BILDG,M59,BANZ,BERF,
                           KOHO2,KOHO3.
            STRATA IS M3.
            STEPWISE = MPLR.
/GROUP CODES (3) ARE 1,2.
       NAMES (3) ARE 'MAENNER','FRAUEN'.
/PRINT CASES ARE 0.
/END
```

Im Vergleich zu Programmbeispiel 5.1 wird im REGRESSION-Paragraph des obigen Laufs kein festes Cox-Modell spezifiziert, sondern es wird damit zunächst nur ein Variablen-Pool vorgegeben, aus dem schrittweise nach statistischen Kriterien ein Cox-Modell aufgebaut werden soll. Die Aufnahme beziehungsweise der Ausschluß von Variablen erfolgt dabei auf der Grundlage des Partial-Likelihood-Quotiententests (STEPWISE = MPLR) und lautet

$$\chi^2_{MPLR} = 2(\ln L(\hat{\beta}_\nu) - \ln L(\hat{\beta}_{\nu+1})) \qquad , \nu = 1, 2, \ldots,$$

wobei $\hat{\beta}_\nu$ der beim ν-ten Schritt berechnete MPL-Schätzer ist. Implizit werden in diesem Programmlauf durch die Voreinstellung von Werten noch eine Reihe von weiteren Entscheidungen über die Vorgehensweise im Auswahlprozeß getroffen (vgl. Dixon u.a. 1983, S. 589). So zum Beispiel, daß keine der Variablen von vornherein in den ersten Regressionsschritt aufgenommen werden soll, oder, daß jede Variable nur zweimal aufgenommen beziehungsweise ausgeschlossen werden darf.
Das Ergebnis dieses Programms ist in Tabelle 5.2 ausführlich dargestellt. In STEP NUMBER 0 werden zunächst für alle Variablen aus dem vorgegebenen Pool auf der Basis des Partial-Likelihood-Quotiententests die χ^2-Werte und die dazugehörigen Signifikanzwahrscheinlichkeiten (P-VALUE) unter der Annahme ermittelt, daß jede dieser Variablen einzeln als erste in das Cox-Modell einbezogen wird. Auf dieser Grundlage ergibt sich, daß die Variable Berufserfahrung (BERF) mit einem χ^2-Wert von 541,11 von allen Variablen am meisten zur Erklärung der Rate des Berufswechsels beitragen kann. Diese Variable wird deswegen in STEP NUMBER 1 als Kovariable in das Cox-Modell aufgenommen. Ihr $\hat{\beta}$-Koeffizient beträgt -0,0046 und ist hochsignifikant (COEFFICIENT/STANDARD ERROR -0,0046/0,0002 = -21,2017).

Tabelle 5.2: Beispiel für eine Variablenselektion mit Hilfe der schrittweisen Cox-Regression

```
STEP NUMBER   0          THERE ARE NO TERMS IN THE MODEL AT THIS STEP.
---------------

STATISTICS TO ENTER OR REMOVE VARIABLES
---------------------------------------

                    APPROX.        APPROX.
    VARIABLE        CHI-SQ.        CHI-SQ.                      LOG
NO. N A M E         ENTER          REMOVE        P-VALUE     LIKELIHOOD

 65 BILDG             5.86                       0.0155    -34285.1049
 59 M59              68.79                       0.0000    -34253.6425
 67 BANZ             34.78                       0.0000    -34270.6435
 66 BERF            541.11                       0.0000    -34017.4808
 68 KOHO2             1.99                       0.1579    -34287.0382
 69 KOHO3            69.79                       0.0000    -34253.1410


STEP NUMBER   1              BERF     IS ENTERED
---------------

                  LOG LIKELIHOOD = -34017.4808
IMPROVEMENT CHI-SQ  ( 2*(LN(MPLR) ) =    541.11  D.F.=   1  P = 0.0000
                GLOBAL CHI-SQUARE =    462.24  D.F.=   1  P = 0.0000
```

Fortsetzung von Tabelle 5.2

	VARIABLE	COEFFICIENT	STANDARD ERROR	COEFF./S.E.	EXP(COEFF.)
66	BERF	-0.0046	0.0002	-21.2017	0.9954

STATISTICS TO ENTER OR REMOVE VARIABLES

NO.	VARIABLE N A M E	APPROX. CHI-SQ. ENTER	APPROX. CHI-SQ. REMOVE	P-VALUE	LOG LIKELIHOOD
65	BILDG	10.88		0.0010	-34012.0421
59	M59	56.60		0.0000	-33989.1796
67	BANZ	194.76		0.0000	-33920.0995
66	BERF		541.11	0.0000	-34288.0354
68	KOHO2	2.01		0.1566	-34016.4775
69	KOHO3	9.49		0.0021	-34012.7363

STEP NUMBER 2 BANZ IS ENTERED

LOG LIKELIHOOD = -33920.0995

IMPROVEMENT CHI-SQ (2*(LN(MPLR)) = 194.76 D.F.= 1 P = 0.0000

GLOBAL CHI-SQUARE = 583.41 D.F.= 2 P = 0.0000

	VARIABLE	COEFFICIENT	STANDARD ERROR	COEFF./S.E.	EXP(COEFF.)
67	BANZ	0.1520	0.0100	15.2560	1.1642
66	BERF	-0.0074	0.0003	-23.8515	0.9926

STATISTICS TO ENTER OR REMOVE VARIABLES

NO.	VARIABLE N A M E	APPROX. CHI-SQ. ENTER	APPROX. CHI-SQ. REMOVE	P-VALUE	LOG LIKELIHOOD
65	BILDG	10.81		0.0010	-33914.6965
59	M59	55.89		0.0000	-33892.1523
67	BANZ		194.76	0.0000	-34017.4808
66	BERF		701.09	0.0000	-34270.6435
68	KOHO2	0.51		0.4737	-33919.8428
69	KOHO3	6.54		0.0105	-33916.8291

STEP NUMBER 3 M59 IS ENTERED

LOG LIKELIHOOD = -33892.1523

IMPROVEMENT CHI-SQ (2*(LN(MPLR)) = 55.89 D.F.= 1 P = 0.0000

GLOBAL CHI-SQUARE = 643.02 D.F.= 3 P = 0.0000

	VARIABLE	COEFFICIENT	STANDARD ERROR	COEFF./S.E.	EXP(COEFF.)
59	M59	-0.0053	0.0007	-7.2320	0.9947
67	BANZ	0.1505	0.0099	15.2358	1.1624
66	BERF	-0.0073	0.0003	-23.6724	0.9927

Fortsetzung von Tabelle 5.2

STATISTICS TO ENTER OR REMOVE VARIABLES

NO.	VARIABLE N A M E	APPROX. CHI-SQ. ENTER	APPROX. CHI-SQ. REMOVE	P-VALUE	LOG LIKELIHOOD
65	BILDG	8.32		0.0039	-33887.9899
59	M59		55.89	0.0000	-33920.0995
67	BANZ		194.05	0.0000	-33989.1796
66	BERF		691.27	0.0000	-34237.7891
68	KOHO2	0.97		0.3249	-33891.6676
69	KOHO3	13.64		0.0002	-33885.3330

STEP NUMBER 4 KOHO3 IS ENTERED

LOG LIKELIHOOD = -33885.3330
IMPROVEMENT CHI-SQ (2*(LN(MPLR)) = 13.64 D.F.= 1 P = 0.0002
GLOBAL CHI-SQUARE = 657.14 D.F.= 4 P = 0.0000

	VARIABLE	COEFFICIENT	STANDARD ERROR	COEFF./S.E.	EXP(COEFF.)
59	M59	-0.0057	0.0007	-7.6556	0.9943
67	BANZ	0.1493	0.0099	15.0449	1.1611
66	BERF	-0.0071	0.0003	-22.8015	0.9929
69	KOHO3	0.1264	0.0340	3.7215	1.1348

STATISTICS TO ENTER OR REMOVE VARIABLES

NO.	VARIABLE N A M E	APPROX. CHI-SQ. ENTER	APPROX. CHI-SQ. REMOVE	P-VALUE	LOG LIKELIHOOD
65	BILDG	7.13		0.0076	-33881.7704
59	M59		62.99	0.0000	-33916.8291
67	BANZ		189.62	0.0000	-33980.1443
66	BERF		628.57	0.0000	-34199.6192
68	KOHO2	9.46		0.0021	-33880.6030
69	KOHO3		13.64	0.0002	-33892.1523

STEP NUMBER 5 KOHO2 IS ENTERED

LOG LIKELIHOOD = -33880.6030
IMPROVEMENT CHI-SQ (2*(LN(MPLR)) = 9.46 D.F.= 1 P = 0.0021
GLOBAL CHI-SQUARE = 663.40 D.F.= 5 P = 0.0000

	VARIABLE	COEFFICIENT	STANDARD ERROR	COEFF./S.E.	EXP(COEFF.)
59	M59	-0.0060	0.0008	-7.9313	0.9941
67	BANZ	0.1465	0.0100	14.7194	1.1578
66	BERF	-0.0070	0.0003	-22.3572	0.9930
68	KOHO2	0.1075	0.0349	3.0753	1.1135
69	KOHO3	0.1838	0.0390	4.7171	1.2017

Fortsetzung von Tabelle 5.2

STATISTICS TO ENTER OR REMOVE VARIABLES

NO.	VARIABLE NAME	APPROX. CHI-SQ. ENTER	APPROX. CHI-SQ. REMOVE	P-VALUE	LOG LIKELIHOOD
65	BILDG	6.48		0.0109	-33877.3618
59	M59		67.82	0.0000	-33914.5117
67	BANZ		182.24	0.0000	-33971.7225
66	BERF		597.73	0.0000	-34179.4681
68	KOHO2		9.46	0.0021	-33885.3330
69	KOHO3		22.13	0.0000	-33891.6676

STEP NUMBER 6 BILDG IS ENTERED

LOG LIKELIHOOD = -33877.3618
IMPROVEMENT CHI-SQ (2*(LN(MPLR)) = 6.48 D.F.= 1 P = 0.0109
GLOBAL CHI-SQUARE = 667.39 D.F.= 6 P = 0.0000

	VARIABLE	COEFFICIENT	STANDARD ERROR	COEFF./S.E.	EXP(COEFF.)
65	BILDG	0.0258	0.0100	2.5628	1.0261
59	M59	-0.0077	0.0010	-7.5078	0.9923
67	BANZ	0.1470	0.0100	14.7181	1.1584
66	BERF	-0.0070	0.0003	-22.2621	0.9930
68	KOHO2	0.1038	0.0350	2.9690	1.1094
69	KOHO3	0.1766	0.0391	4.5218	1.1932

STATISTICS TO ENTER OR REMOVE VARIABLES

NO.	VARIABLE NAME	APPROX. CHI-SQ. ENTER	APPROX. CHI-SQ. REMOVE	P-VALUE	LOG LIKELIHOOD
65	BILDG		6.48	0.0109	-33880.6030
59	M59		58.80	0.0000	-33906.7628
67	BANZ		182.48	0.0000	-33968.6034
66	BERF		591.83	0.0000	-34173.2762
68	KOHO2		8.82	0.0030	-33881.7704
69	KOHO3		20.34	0.0000	-33887.5322

NO TERM PASSES THE REMOVE AND ENTER LIMITS (0.1500 0.1000) .

SUMMARY OF STEPWISE RESULTS

STEP NO	VARIABLE ENTERED	DF	VARIABLE REMOVED	LOG LIKELIHOOD	IMPROVEMENT CHI-SQUARE	IMPROVEMENT P-VALUE	GLOBAL CHI-SQUARE	GLOBAL P-VALUE
0				-34288.035				
1	66 BERF	1		-34017.481	541.109	0.000	462.240	0.000
2	67 BANZ	2		-33920.100	194.762	0.000	583.409	0.000
3	59 M59	3		-33892.152	55.894	0.000	643.018	0.000
4	69 KOHO3	4		-33885.333	13.639	0.000	657.144	0.000
5	68 KOHO2	5		-33880.603	9.460	0.002	663.397	0.000
6	65 BILDG	6		-33877.362	6.482	0.011	667.389	0.000

Nach Hereinnahme der Variablen Berufserfahrung in das Modell werden für die restlichen Variablen wiederum die χ^2-Werte sowie ihre Signifikanzwahrscheinlichkeiten berechnet, und zwar unter der Annahme, daß im nächsten Schritt jeweils nur eine von ihnen zusätzlich in das Modell aufgenommen wird. Dabei zeigt sich, daß die Variable Anzahl der vorher ausgeübten Berufe (BANZ) mit einem χ^2-Wert von 194,76 von allen Variablen am meisten zusätzlich erklären kann. Dementsprechend wird diese Variable in STEP NUMBER 2 in das Cox-Modell aufgenommen. Durch die Hereinnahme dieser Variablen verändert sich allerdings der $\hat{\beta}$-Koeffizient der bereits im Modell befindlichen Variablen Berufserfahrung von -0,0046 auf -0,0074. Das heißt, bei Kontrolle der Variablen Anzahl der vorher ausgeübten Berufe, die sich signifikant erhöhend auf die Jobwechsel-Rate auswirkt ($\hat{\beta}_{BANZ} = 0{,}1520$), tritt der reduzierende Effekt der Berufserfahrung auf die Neigung zum Jobwechsel noch deutlicher hervor.
In STEP NUMBER 3 wird dann die Variable Prestige (M59) mit einem zusätzlichen Erklärungsbeitrag von $\chi^2 = 55{,}89$ hereingenommen. Bemerkenswert dabei ist, daß sich die β-Schätzungen und die χ^2-Werte der bereits im Modell befindlichen Variablen dadurch kaum ändern. Man kann deswegen davon ausgehen, daß durch diese Variable eine weitgehend neue Erklärungsdimension in die Regressionsgleichung kommt, die unabhängig von $\hat{\beta}_{BANZ}$ und $\hat{\beta}_{BERF}$ ist.
Nach diesem dritten Schritt wird anhand der χ^2-Werte deutlich, daß sich die jüngste Kohorte (KOHO3) weit mehr von den beiden restlichen Kohorten unterscheidet (χ^2-Wert = 13,64), als dies bei der mittleren Kohorte (KOHO2) der Fall ist (χ^2-Wert = 0,97). Der χ^2-Wert von KOHO3 ist außerdem größer als der χ^2-Wert für die Bildungsvariable (8,32). In STEP NUMBER 4 wird deswegen die Variable KOHO3 in das Cox-Modell eingeschlossen.
Bei Kontrolle der Unterschiede zur jüngsten Kohorte kann allerdings die Dummy-Variable KOHO2, die jetzt nur noch den Unterschied zur ältesten Kohorte erfaßt, auch einen signifikanten Erklärungsbeitrag leisten (χ^2-Wert = 9,46). In STEP NUMBER 5 wird diese Variable in das Modell aufgenommen.
Als letzte Kovariable wird in STEP NUMBER 6 schließlich die Variable BILDG einbezogen, die ebenfalls noch einen signifikanten Erklärungsbeitrag leisten kann. Dabei ist zu beobachten, daß sich insbesondere die $\hat{\beta}$-Koeffizienten der Kohorten-Dummies etwas verringern. Ein Teil der Kohortenunterschiede läßt sich deswegen durch Bildungsdifferenzen zwischen den Kohorten erklären.
Nach diesem letzten Schritt zeigt sich, daß alle Variablen einen signifikanten Effekt haben und die Herausnahme einer dieser Variablen das Cox-Modell spürbar verschlechtern würde (siehe dazu die χ^2-Werte in der Spalte APPROX. CHI-SQ. REMOVE).
Obwohl die Methode der schrittweisen Regression zum gleichen Modell wie in Programmbeispiel 5.1 (Tabelle 5.1) führt, verbergen sich hinter beiden Vorgehensweisen unterschiedliche Prinzipien. Während die Variablen in Programmbeispiel 5.1 von vornherein auf der Grundlage inhaltlicher Überlegungen einbezogen wurden, geschah dies bei der schrittweisen Regression nach statistischen

Kriterien. Das Resultat einer schrittweisen Regression muß deswegen nicht immer inhaltlich sinnvoll sein und hängt auch von den Vorgaben ab, die den Auswahlmechanismus steuern. Insbesondere kann es dazu kommen, daß eine inhaltlich gut zu interpretierende Variable nur deswegen nicht in die Regressionsgleichung aufgenommen wird, weil sie mit weniger gut zu interpretierenden Variablen hoch korreliert, die bereits in früheren Schritten in das Modell aufgenommen worden sind. Umgekehrt kann es aber auch sein, daß der Erklärungsbeitrag einer Variablen erst dann zum Vorschein kommt, wenn andere Einflußfaktoren schon kontrolliert sind, wie dies bei der Dummy-Variablen KOHO2 der Fall war. Insgesamt sollte man bei der Anwendung der schrittweisen Regression deswegen große Vorsicht walten lassen. Vor allem sollte man prüfen, ob man auf verschiedenen Wegen (z. B. durch Vorwärts- und Rückwärtsselektion) sowie durch die Variation der den Newton-Raphson-Algorithmus steuernden Vorgaben jeweils zu ein und demselben Modell kommt oder nicht.

5.4 Die Einbeziehung zeitveränderlicher unabhängiger Variablen

Sind mit dem soeben besprochenen Modell nur zeitkonstante unabhängige Variablen modelliert worden, so läßt sich mit dem BMDP-Unterprogramm P2L aber auch der Einfluß zeitveränderlicher unabhängiger Variablen schätzen. Unter Anwendungsgesichtspunkten besonders interessant ist dies, weil man damit den Einfluß von Kovariablen auf die Hazardrate realitätsgerechter formulieren und zwei oder mehrere parallele Prozesse direkt miteinander koppeln kann. Nicht selten resultiert darüber hinaus auch aus der bloßen Tatsache, daß zeitveränderliche Kovariablen im Modell als zeitkonstant behandelt werden, scheinbare Verweildauerabhängigkeit, da die beobachteten Zustände dann Aggregationen von unbeobachteten heterogenen Zuständen sind (vgl. Abschnitt 3.9.1).
Datentechnisch betrachtet spricht man von zeitkonstanten Kovariablen dann, wenn diese zu Beginn der Episode k gemessen (oder aktualisiert) werden und ihre Werte über die Verweildauer $v_k = t - t_{k-1}$ hinweg unverändert bleiben (vgl. Abbildung 5.2(c)). Zeitveränderliche Kovariablen können hingegen ihren Wert innerhalb der Episode k verändern. Bei diskreten zeitveränderlichen Kovariablen bleiben die Werte dabei über gewisse Subintervalle v_{k_i}, $v_k = \sum_{i=1}^{s} v_{k_i}$ konstant (vgl. Abbildung 5.2(b)), während sich stetige zeitveränderliche Variablen kontinuierlich verändern.

5.4.1 Diskrete zeitveränderliche unabhängige Variablen

In den meisten Fällen wird man es in der Wirtschafts- und Sozialwissenschaft mit Variablen zu tun haben, die sich nicht stetig in der Zeit verändern. Diese

folgen in der Zeit einer Treppenfunktion und wirken direkt auf die Verweildauer ein, indem sie die Rate innerhalb der Episoden verändern. Geht man beispielsweise davon aus, daß sich bei Männern das Ereignis „Heirat" im Familiensystem stabilisierend auf den Erwerbsprozeß im Beschäftigungssystem auswirkt (vgl. Abbildung 5.2), dann läßt sich dieser Zusammenhang über die Einführung einer zeitveränderlichen Kovariablen testen.
Bezeichnet man für das Individuum i mit $t_{i,k-1}$ den Beginn der Berufsepisode k und mit t_{ik} deren Endzeitpunkt, sowie mit v die Verweildauer der Episode k, und ist t_i^H der Heiratszeitpunkt des Individuums i, dann ergibt sich der Wert der zeitveränderlichen Dummy-Variablen Heirat $x_{ik}^H(v)$ wie folgt:

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

Der Kovariablen-Vektor $\mathbf{x}_k$ im Cox-Modell hängt dann von der Zeit ab, und das Modell lautet:

$$\lambda^k(v|\mathbf{x}_k(v)) = \lambda_0(v) \exp(\mathbf{x}_k'(v)\,\beta) \qquad , k = 1, 2, \ldots.$$

Das folgende Programm zeigt die Realisierung eines solchen Modells in BMDP, welches sich wiederum durch geringfügige Veränderung des Programmlaufs 5.1 ergibt. Im Unterschied zu Programmlauf 5.1, in dem nach Geschlecht geschichtet wurde, konzentriert sich das Beispiel hier nur auf die Männer (USE = (M3 EQ 1)). Das Ereignis Heirat dürfte sich nämlich bei den Frauen wegen der sich meist anschließenden familienbedingten Erwerbsunterbrechungen ganz anders auf den Erwerbsprozeß auswirken:

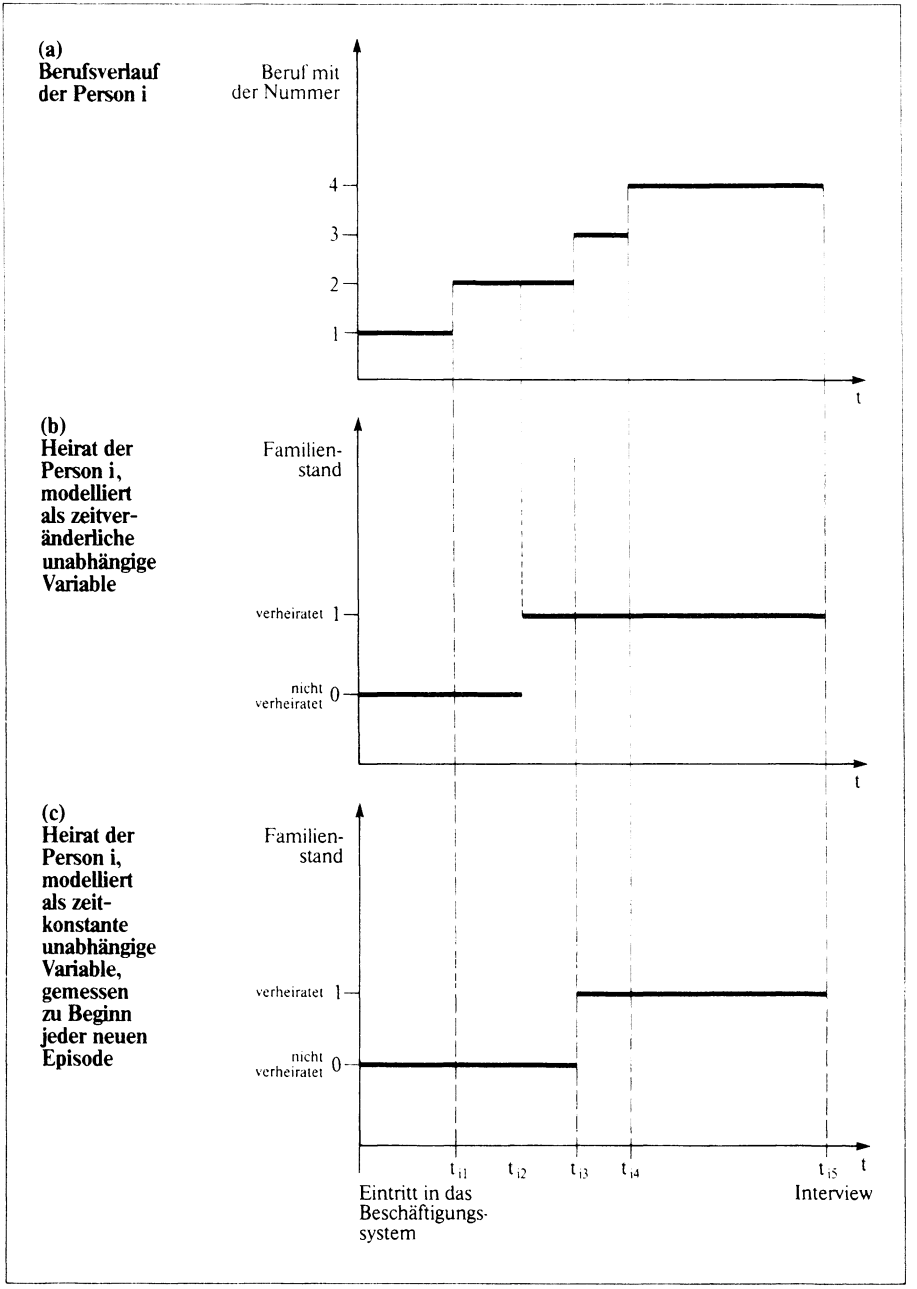
Programmbeispiel 5.4:

```
/INPUT UNIT IS 30.
       CODE IS DATA.
/VARIABLE NAMES ARE (63)DUR,(64)ZEN,(65)BILDG,(66)BERF,
                    (67)BANZ,(68)KOHO2,(69)KOHO3.
          ADD IS 7.
/TRANSFORM USE = (M3 EQ 1).
           DUR = M51 - M50 + 1.
           ZEN = 1.
           IF (M51 EQ M47) THEN ZEN = 0.
           IF (M41 EQ 1 AND M42 EQ 1) THEN BILDG = 9.
           IF (M41 EQ 1 AND (M42 EQ 2 OR M42 EQ 3)) THEN BILDG = 11.
           IF (M41 EQ 2 AND M42 EQ 1) THEN BILDG = 10.
           IF (M41 EQ 2 AND (M42 EQ 2 OR M42 EQ 3)) THEN BILDG = 12.
           IF (M41 EQ 3 AND (M42 EQ 1 OR M42 EQ 2 OR M42 EQ 3))
           THEN BILDG = 13.
           IF (M42 EQ 4) THEN BILDG = 17.
           IF (M42 EQ 5) THEN BILDG = 19.
           KOHO2 = 0.
           KOHO3 = 0.
```

```
            IF(M48 GE 468 AND M48 LE 504) THEN KOHO2 = 1.
            IF(M48 GE 588 AND M48 LE 624) THEN KOHO3 = 1.
            IF (M49 EQ 0) THEN M49 = 10000.
            M49 = M49 - M50.
            BERF = M50 - M43.
            BANZ = M5 - 1.
/FORM TIME IS DUR.
      STATUS IS ZEN.
      RESPONSE IS 1.
/REGRESSION COVARIATES ARE BILDG,M59,BANZ,BERF,
                           KOHO2,KOHO3.
             ADD IS HEIRAT.
             AUXILIARY IS M49.
/FUNCTION HEIRAT = 0.0.
          IF (TIME GT M49) THEN HEIRAT = 1.0.
/PRINT CASES ARE 0.
/END
```

Im REGRESSION-Paragraph werden wiederum die bereits bekannten Kovariablen eingeführt (vgl. Anhang 1). Als zeitveränderliche Kovariable dient uns die Dummy-Variable mit dem Namen HEIRAT, die zusätzlich mit dem ADD-Statement in das Modell aufgenommen wird. Die Hilfsvariable M49, die den Zeitpunkt der Heirat in Anzahl der Monate vom Beginn des Jahrhunderts an enthält und im FUNCTION-Paragraph verwendet werden soll, wird mit dem AUXILIARY-Statement definiert.
Die Erzeugung der zeitveränderlichen Variablen HEIRAT erfolgt im FUNCTION-Paragraph. Dabei wird die Variable HEIRAT zuerst auf die Ausprägung 0 gesetzt. Falls die Verweildauer im Beruf (TIME) über den Zeitpunkt der Heirat (M49) hinausgeht, das heißt, sobald der Mann geheiratet hat, soll die zeitveränderliche Variable HEIRAT den Wert 1 erhalten.
Voraussetzung dieses Vergleichs ist allerdings, daß der Zeitpunkt der Heirat vorher auf die Verweildauer (TIME) hin, die bei jeder neuen Berufsepisode neu bei 0 beginnt, standardisiert worden ist. Im TRANSFORM-Paragraph wird deswegen vom Zeitpunkt der Heirat (M49) der Zeitpunkt des Eintritts in die k-te Berufsepisode M50 (beide vercodet in Anzahl von Monaten seit Beginn des Jahrhunderts) subtrahiert.
Sollte es sich dabei um einen bereits vor dem Eintritt in einen Beruf verheirateten Mann handeln, dann wird M49 negativ, und beim Vergleich mit der Verweildauer (TIME) im FUNCTION-Paragraph ist die Abfrage immer erfüllt. Die Variable HEIRAT erhält dann über die gesamte Verweildauer hinweg immer den Wert 1 zugewiesen. Umgekehrt wird allen Männern, die während des gesamten Beobachtungszeitraums noch nicht geheiratet hatten und die in der Variablen Zeitpunkt der Heirat (M49) die Ausprägung 0 haben, im TRANSFORM-Paragraph ein so großer Wert (10.000 Monate) zugewiesen, daß die Abfrage im FUNCTION-Paragraph nie erfüllt sein kann und die Variable HEIRAT über die gesamte Prozeßzeit immer den Wert 0 annimmt.

Abbildung 5.2: Modellierung des Einflusses der diskreten Variablen Heirat (a) auf den Berufsverlauf, (b) als zeitveränderliche unabhängige Variable und (c) als zeitkonstante unabhängige Variable, gemessen zu Beginn jeder neuen Episode



Die Partial-Likelihood-Schätzung des Einflusses der zeitveränderlichen Variablen HEIRAT ist natürlich außerordentlich aufwendig und rechenzeitintensiv, da für jeden Ereigniszeitpunkt und für alle dabei noch dem Risiko ausgesetzten Individuen der Vergleich im FUNCTION-Paragraph durchgeführt und die Risikomenge jeweils neu aktualisiert werden muß. Das Ergebnis der Schätzung von Programmbeispiel 5.4 ist in Tabelle 5.3 dargestellt.

Tabelle 5.3: Ergebnis des Cox-Modells aus Programmbeispiel 5.4

```
   LOG LIKELIHOOD = -19012.0115
GLOBAL CHI-SQUARE =    511.69   D.F.=   7   P-VALUE =0.0000
```

	VARIABLE	COEFFICIENT	STANDARD ERROR	COEFF./S.E.	EXP(COEFF.)
65	BILDG	0.0113	0.0145	0.7756	1.0113
59	M59	-0.0045	0.0014	-3.2215	0.9955
67	BANZ	0.1609	0.0121	13.2832	1.1745
66	BERF	-0.0076	0.0005	-15.6909	0.9925
68	KOHO2	0.0581	0.0469	1.2379	1.0598
69	KOHO3	0.1888	0.0533	3.5452	1.2078
70	HEIRAT	-0.3208	0.0511	-6.2807	0.7256

An dem GLOBAL CHI-SQUARE-Wert von 511,69 (bei sieben Freiheitsgraden) wird zunächst wieder deutlich, daß die Hypothese, keine der eingeführten Kovariablen erkläre etwas an der Rate der Berufswechsel von Männern, zu verwerfen ist (P-VALUE = 0,0000).
Eine genauere Überprüfung der standardisierten Koeffizienten in der Spalte COEFF./S.E. zeigt allerdings, daß nicht alle Variablen einen signifikanten Einfluß haben. So sind die Absolutbeträge der standardisierten $\hat{\beta}$-Koeffizienten der Variablen Bildung ($|\hat{\beta}_{BILDG}/S.E.(\hat{\beta}_{BILDG})| = 0{,}7756$) und der Dummy-Variablen KOHO2 ($|\hat{\beta}_{KOHO2}/S.E.(\hat{\beta}_{KOHO2})| = 1{,}2379$) kleiner als der Wert von 1,96 und damit bei einer Irrtumswahrscheinlichkeit von 0,05 nicht signifikant von 0 verschieden. Unterschiede im Bildungsniveau wirken sich bei Männern also nicht auf die Jobwechselrate aus, und zwischen den Männern der ältesten und der mittleren Kohorte gibt es keine prinzipiellen Unterschiede im Berufswechselverhalten. Die Variablen Prestige (M59), Anzahl der vorher ausgeübten Berufe (BANZ), Berufserfahrung (BERF) und KOHO3 haben dagegen die gleiche Wirkungsrichtung wie in Tabelle 5.1 und lassen sich auf dieselbe Weise inhaltlich interpretieren.
Die eigentlich interessante Kovariable aus dem obigen Programmbeispiel ist natürlich die zeitveränderliche Variable HEIRAT. Ihr $\hat{\beta}$-Koeffizient ist signifikant und hat erwartungsgemäß ein negatives Vorzeichen. Es besagt, daß sich die Neigung zum Berufswechsel nach einer Heirat deutlich vermindert. Im Vergleich zu den unverheirateten Männern vermindert sich die Mobilitätsrate bei den Ehemännern um 27,44 Prozent [$(0{,}7256 - 1) \cdot 100\% = -27{,}44\%$].

Ein anderes Beispiel für die Kopplung paralleler Prozesse durch die Einbeziehung diskreter zeitabhängiger Kovariablen in einem Cox-Modell liefern Mayer und Wagner (1986). Ebenfalls auf der Basis der Lebensverlaufsstudie untersuchten sie, wie sich zeitsynchrone Ereignisse aus dem Ausbildungs- und Berufsbereich sowie dem Fertilitätsverhalten auf den Auszug aus dem elterlichen Haushalt auswirken.

Die Verweildauer ist in diesem Fall über die Zeitspanne zwischen dem 15. Lebensjahr und dem Auszugsalter aus dem Haushalt der Eltern definiert worden (Ein-Episoden-Fall). In der Untersuchung wurden zwei verschiedene Ereignisarten modelliert (vgl. dazu den nächsten Abschnitt über die Konstruktion von Mehr-Zustands-Modellen): Der Auszug aus dem Elternhaus ohne gleichzeitige Heirat (NS = nicht synchronisiert) und der Auszug aus dem Elternhaus mit gleichzeitiger Heirat (S = synchronisiert). Auszugszeitpunkt und Heiratstermin wurden dann als synchronisiert betrachtet, wenn die Heirat höchstens zwei Monate vor oder nach dem Auszug erfolgte. Für jede dieser Ereignisarten wurde getrennt nach den drei Geburtskohorten (den 1929–31, 1939–41 und den 1949–51 Geborenen) eine Regressionsgleichung geschätzt (vgl. Tabelle 5.4). In diese Schätzungen wurden, neben einer Reihe zeitkonstanter Kovariablen, auf die hier nicht näher eingegangen werden soll, vier zeitabhängige Dummy-Variablen eingeführt: die Variable Berufserfahrung (JOBERF), die zeigt, ob der Befragte vor dem Auszug bereits irgendwann einmal erwerbstätig war oder nicht; die Variable Ausbildung (AUSB), die indiziert, ob gleichzeitig mit dem Auszug (d.h. in einem Zeitraum von plus/minus zwei Monaten) eine Ausbildung begonnen wurde oder nicht; die Variable Berufstätigkeit (JOB), die aufzeigt, ob gleichzeitig mit dem Auszug (d.h. in einem Zeitraum von plus/minus zwei Monaten) eine Erwerbstätigkeit neu aufgenommen wurde oder nicht, und die Variable Geburt (GEB), die angibt, ob in einem Zeitraum von zwei Monaten vor bis fünf Monaten nach dem Auszugstermin ein Kind geboren wurde oder nicht. Die Ergebnisse der Partial-Likelihood-Schätzungen sind in Tabelle 5.4 zusammengestellt. Dabei zeigt sich, daß sich die Neigung, aus dem elterlichen Haushalt auszuziehen, mit dem Eintritt in eine Berufstätigkeit (Variable JOB) für alle Kohorten signifikant erhöht. Wenn der Auszug nicht mit einer Heirat verbunden ist (vgl. die Spalten NS), ist die Wirkung der mit einer Erwerbstätigkeit einhergehenden größeren Autarkie und finanziellen Unabhängigkeit auf das Auszugsverhalten sogar größer.

In dieselbe Richtung wirken auch die bereits vor dem Auszug gemachten beruflichen Erfahrungen (Variable JOBERF). Sie erweisen sich allerdings nur dann als signifikant, wenn der Auszug aus dem Elternhaus mit der Gründung einer eigenen Familie verbunden ist (vgl. die Spalten S).

Beschleunigend auf den Auszug aus dem Elternhaus wirkt sich auch der Einstieg in eine neue Ausbildungsphase aus (Variable AUSB). Die Wirkung ist dabei für diejenigen Personen größer, die beim Auszug nicht heiraten (vgl. die Spalten NS).

Auch das Fertilitätsverhalten (Variable GEB) beeinflußt den Zeitpunkt, zu dem

Tabelle 5.4: Partial-Likelihood-Schätzungen der Effekte von zeitveränderlichen und zeitkonstanten Kovariablen auf die Auszugsrate aus dem elterlichen Haushalt[1]

	Geburtsjahr					
	1929–31		1939–41		1949–51	
	S[2]	NS[2]	S[2]	NS[2]	S[2]	NS[2]
SEX	0.54*	-0.01	0.55*	0.56*	1.14*	0.48*
MS	-0.15	0.01	0.05	0.09	0.13	0.11
ABI	0.49	0.22	-0.89*	-0.40	-0.75	0.45*
AUSB	2.61*	-[3]	1.92*	4.48*	1.91*	2.82*
JOBERF	0.67*	-0.26	0.74*	0.14	1.07*	-0.05
JOB	3.80*	4.19*	2.73*	3.98*	1.94*	3.01*
GEB	1.61*	1.45*	2.12*	1.18*	2.43*	0.72*
BILDV	-0.02	0.09*	0.03	-0.01	0.00	0.00
BILDM	0.12*	-0.01	-0.06	-0.04	-0.07	0.02
ERWERBM	-0.38*	0.07	-0.19	-0.03	0.10	0.10
UNFAM	0.32	-0.12	0.14	0.16	0.15	0.04
LANDWVA	-0.57*	-0.10	0.14	0.07	-0.40	0.45*
SELSTVA	0.22	-0.14	-0.36	-0.06	0.10	0.16
BEAMTVA	-0.26	0.14	-0.11	-0.30	0.11	-0.06
ANGESTVA	-0.53*	0.03	-0.45*	0.24	-0.13	0.00
BILDDIF	0.06*	0.00	0.02	0.05*	-0.03	0.00
ALTDIF	-0.05*	0.00	0.00	0.00	0.01	-0.01
GEZIFF	0.04	0.00	0.05	0.04	0.09*	0.03
GEBLAND	-0.49*	1.04*	-0.12	0.39*	-0.24	-0.18
ANZMIG15	-0.03	-0.04	0.04	0.04	-0.16*	0.04
ORT15	-0.05	0.07	-0.32*	0.24	-0.16	0.13
WOART15	-0.19	-0.25	-0.29*	-0.22	-0.36*	-0.40*
BELDI15	-0.19	0.16	-0.38*	-0.20	0.04	-0.06
χ^2	800.34	2902,96	733,32	4068.64	911,96	2824,90
d.f.	23	23	23	23	23	23
Anzahl der Ereignisse	234	283	283	280	270	334
Anzahl der Personen	607	607	649	649	701	701

[1] Werte mit einen * sind auf einem 95%-Niveau signifikant

[2] S: Auszug mit erster Heirat synchronisiert. NS: Auszug mit erster Heirat nicht synchronisiert.

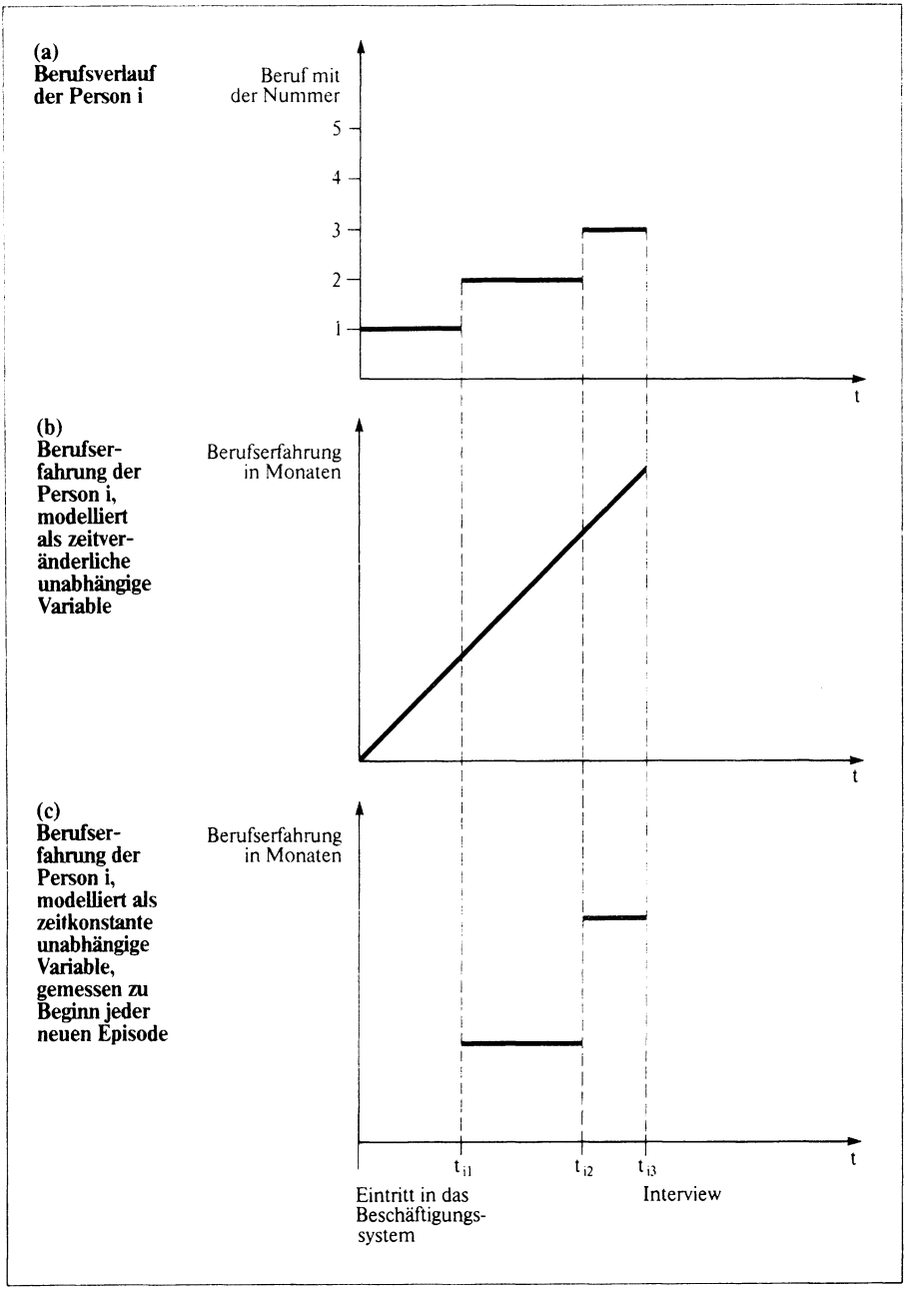
[3] Das Modell NS bei Kohorte 1929–31 konnte nur geschätzt werden, wenn die Variable AUSB aus den Berechnungen ausgeschlossen wurde.

der elterliche Haushalt verlassen wird. Wenn ein Kind geboren wird, dann steigt die Rate des Auszugs aus dem Elternhaus signifikant an. Dieser Effekt ist besonders groß, wenn der Auszug mit einer Heirat zusammenfällt (vgl. die Spalten S).

5.4.2 Stetige zeitveränderliche unabhängige Variablen

Die beiden soeben dargestellten Beispiele zur Einbeziehung diskreter zeitveränderlicher unabhängiger Variablen machen deutlich, daß sich die Erklärungsmodelle in inhaltlicher Hinsicht viel realitätsgerechter formulieren lassen und, daß damit natürlich die Attraktivität der Erhebung und Analyse von Ereignisdaten in den Wirtschafts- und Sozialwissenschaften steigt.
Dies gilt entsprechend auch für die Modellierung stetiger zeitveränderlicher unabhängiger Variablen. Im Vergleich zu den später bei den parametrischen Standard-Modellen (vgl. dazu Abschnitt 6.3) zu besprechenden Möglichkeiten der Einbeziehung stetiger zeitveränderlicher Kovariablen bietet das Cox-Modell hier sogar den Vorzug, daß man direkt auf die Verweildauer (im BMDP-Programm P2L die reservierte Variable mit dem Namen TIME) zugreifen kann. Denn bei der Partial-Likelihood-Schätzung werden zum einen nur für die Ereigniszeitpunkte die Wahrscheinlichkeiten berücksichtigt, daß bei gegebener Risikomenge ein Individuum ein Ereignis erfährt, und es lassen sich zum anderen für jeden dieser Ereigniszeitpunkte und für alle Individuen, die dann jeweils noch dem Risiko ausgesetzt und nicht zensiert sind, die Kovariablenvektoren im FUNCTION-Paragraph aktualisieren.
Kennt man demnach die funktionale Form der stetigen unabhängigen Variablen in der Zeit, oder läßt sich ihr Verlauf durch ein Polynom oder eine Treppenfunktion approximieren, dann kann man sie ohne Schwierigkeiten über die Variable TIME mit dem zu untersuchenden Prozeß koppeln. Für das Beispiel der Untersuchung von Berufsverläufen heißt dies zunächst, daß auch die Veränderungen sozialstruktureller und ökonomischer Rahmenbedingungen, in welche die Karriereprozesse eingebunden sind, in die Modelle aufgenommen werden können. Beispiele dafür sind die Veränderung von Arbeitslosenquoten, das Wachstum des Volkseinkommens, der Wandel des Anteils des tertiären Sektors usw. Mit diesen Makrodaten, die von den statistischen Ämtern meist als lange Reihen angeboten werden, ließen sich etwa Hypothesen prüfen wie die, daß mit steigender Arbeitslosigkeit die Neigung zum Jobwechsel sinkt oder daß mit zunehmendem Pro-Kopf-Einkommen und der damit einhergehenden wachsenden ökonomischen Unabhängigkeit die Neigung zum Berufswechsel steigt. Da die Mikrodaten (die an Individuen gemessenen Berufsverläufe) und die Makrodaten (die im Aggregat vorliegenden sozialstrukturellen und ökonomischen Kennziffern) über die Zeitdimension synchronisierbar oder durch lag-Funktionen verbindbar sind, ist mit der Ereignisanalyse auch die Möglichkeit gegeben, Makrodaten kausal auf Individualdaten zu beziehen.
Aber nicht nur Periodeneffekte, zyklische Schwankungen und langfristige Trendentwicklungen auf der Makroebene, sondern auch stetige Entwicklungen auf der Mikroebene wie etwa beim Verdienst können über die Verweildauer hinweg zeitveränderlich eingebunden werden. Hat man etwa, wie in der Lebensverlaufsstudie, jeweils das Einkommen zu Beginn und am Ende jeder Berufsepisode gemessen und geht man von der Annahme einer linearen Einkommensent-

Abbildung 5.3: Modellierung des Einflusses der stetigen Variablen Berufserfahrung (a) auf den Berufsverlauf, (b) als zeitveränderliche unabhängige Variable und (c) als zeitkonstante unabhängige Variable, gemessen zu Beginn jeder neuen Episode



wicklung in der Berufsepisode aus, dann kann durch lineare Interpolation für jeden Zeitpunkt der Verweildauer das Einkommen approximiert werden. Statt dessen könnte man aber auch versuchen, eine nicht-lineare Funktion der Lebenseinkommensentwicklung über mehrere Berufsepisoden hinweg auf der Basis von Stützpunkten zu schätzen und diese dann für den Verlauf in jeder einzelnen Berufsepisode in Abhängigkeit vom Lebensalter zu verwenden. Für beide Approximationen gilt allerdings, daß die Annäherung um so besser (schlechter) wird, je kürzer (länger) die Berufsepisoden sind.
Insgesamt ist es für die Modellierung einer stetigen zeitveränderlichen Variablen günstig, deren Verlauf in einem ersten Schritt durch eine parametersparsame, aber realitätsgerechte lineare, einfache nicht-lineare oder Treppenfunktion zu approximieren und diese dann in einem zweiten Schritt im FUNCTION-Paragraph mit der Verweildauer TIME zu verbinden.
Mit der Partial-Likelihood-Methode nicht schätzbar sind allerdings zeitveränderliche Kovariablen, mit denen die Verweildauerabhängigkeit additiv in der Form $x(v) = x^* + g(v)$ in das Cox-Modell aufgenommen wird. Versucht man beispielsweise die Variable Beruferfahrung (BERF), die bisher als zeitkonstante Kovariable in das Cox-Modell einbezogen wurde (vgl. Abbildung 5.3(c)), als über die Verweildauer hinweg zeitveränderliche Kovariable mit einer linearen Funktion zu approximieren (vgl. Abbildung 5.3(b); $x_{BERF}(t) = x_{BERF}(t_{i,k-1}) + v$, mit $v = t - t_{i,k-1}$), dann kürzt sich, wie anhand von Formel (3.6.22) unmittelbar sichtbar ist, bei jedem Ereigniszeitpunkt die Verweildauerabhängigkeit v heraus und die Berufserfahrung wird als zeitkonstante Kovariable geschätzt. Diese Art von Effekten lassen sich nur mit parametrischen Verfahren schätzen (vgl. dazu die Abschnitte 6.3.2 und 6.5).

5.5 Zur Modellierung von Mehr-Zustands-Modellen

Unter Anwendungsgesichtspunkten in den Wirtschafts- und Sozialwissenschaften ist man aber meist nicht nur an einer bestimmten Ereignisart (z. B. dem Ereignis Berufswechsel an sich oder dem Ereignis Auszug aus dem Elternhaus an sich) interessiert, sondern man will darüber hinaus häufig auch wissen, wie sich bestimmte Kovariablen (in Art, Richtung und Effektgröße) auf die Übergänge zu verschiedenen, miteinander konkurrierenden Zielzuständen (wie etwa auf den Übergang vom Arbeiter zum Angestellten und vom Arbeiter zum Selbständigen oder auf den Auszug aus dem Elternhaus mit oder ohne gleichzeitige Heirat) auswirken. In diesem Abschnitt soll deswegen gezeigt werden, wie man mit der Partial-Likelihood-Methode von Cox Mehr-Zustands- oder Competing-Risks-Modelle bearbeitet.
Wie in den Abschnitten 3.4 und 4.3 bereits ausgeführt, erfolgt die Realisierung eines Mehr-Zustands-Modells dadurch, daß man bei Betrachtung einer bestimmten Ereignisart (oder eines bestimmten End- oder Zielzustandes) die jeweils konkurrierenden Ereignisse als zensiert behandelt.

Als Grundlage für die Demonstration der programmtechnischen Vorgehensweise zur Modellierung eines Mehr-Zustands-Modells in BMDP soll uns wieder der bereits bekannte ereignisorientierte Erwerbstätigkeits-Datensatz dienen (vgl. Anhang 1).
Um mehrere Ausgangs- und Endzustände zu erhalten, wurden die Berufstätigkeiten der Erwerbsgeschichte wieder in 12 Berufsgruppen (vgl. Tabelle 4.4) klassifiziert. Für jede Erwerbstätigkeitsepisode wurde dann der entsprechende Berufsgruppenschlüssel (Variable M61) (Ausgangszustand) und der Berufsgruppenschlüssel der jeweils nächsten Erwerbstätigkeitsepisode (Variable M62) (Zielzustand) abgespeichert. Wenn es sich um die letzte Berufstätigkeitsepisode in einer Erwerbsgeschichte handelt, die keinen Nachfolger hat, dann wurde als Endzustand die Ausprägung 0 (zensiert) codiert.
Mit dem folgenden BMDP-Lauf, mit dem das Programmbeispiel 5.1 wiederum nur geringfügig modifiziert wird, soll für den spezifischen Übergang von einem unqualifizierten manuellen Beruf (Ausprägung 2) zu einem qualifizierten manuellen Beruf (Ausprägung 3) ein Cox-Modell geschätzt werden:

$$\lambda^{k}_{23}(v|\mathbf{x}) = \lambda_{023}(v)\ \exp(\mathbf{x}_k\beta) \qquad , k = 1, 2, \ldots.$$

Programmbeispiel 5.5

```
/INPUT UNIT IS 30.
       CODE IS DATA.
/VARIABLE NAMES ARE (63)DUR,(64)ZEN,(65)BILDG,(66)BERF,
                    (67)BANZ,(68)KOHO2,(69)KOHO3.
          ADD IS 7.
/TRANSFORM IF (M3 NE 1 OR M61 NE 2) THEN USE = -1.
           DUR = M51 - M50 + 1.
           ZEN = 1.
           IF (M51 EQ M47) THEN ZEN = 0.
           IF (M62 NE 3) THEN ZEN = 0.
           IF (M41 EQ 1 AND M42 EQ 1) THEN BILDG = 9.
           IF (M41 EQ 1 AND (M42 EQ 2 OR M42 EQ 3)) THEN BILDG = 11.
           IF (M41 EQ 2 AND M42 EQ 1) THEN BILDG = 10.
           IF (M41 EQ 2 AND (M42 EQ 2 OR M42 EQ 3)) THEN BILDG = 12.
           IF (M41 EQ 3 AND (M42 EQ 1 OR M42 EQ 2 OR M42 EQ 3))
           THEN BILDG = 13.
           IF (M42 EQ 4) THEN BILDG = 17.
           IF (M42 EQ 5) THEN BILDG = 19.
           KOHO2 = 0.
           KOHO3 = 0.
           IF(M48 GE 468 AND M48 LE 504) THEN KOHO2 = 1.
           IF(M48 GE 588 AND M48 LE 624) THEN KOHO3 = 1.
           BERF = M50 - M43.
           BANZ = M5 - 1.
/FORM TIME IS DUR.
      STATUS IS ZEN.
      RESPONSE IS 1.
```

```
/REGRESSION COVARIATES ARE BILDG,M59,BANZ,BERF,
                           KOHO2,KOHO3.
/PLOT TYPE = LOG.
/PRINT CASES ARE 0.
/END
```

Im Unterschied zu Programmbeispiel 5.1 betrachten wir wiederum nur die Männer (M3) und filtern alle diejenigen Episoden heraus, in denen sich einer der Befragten in einem einfachen manuellen Beruf befindet (M61). Die Risikomenge ist damit durch die Gruppe der Männer gebildet, die in einem unqualifizierten manuellen Beruf tätig sind.
Die Verarbeitung konkurrierender Ereignisse erfolgt nun dadurch, daß alle Episoden, die nicht mit dem interessierenden Endzustand (M62) „qualifizierter manueller Beruf" (mit der Ausprägung 3) übereinstimmen, als zensiert behandelt werden. Im TRANSFORM-Paragraph wird deswegen bei allen anderen Ereignissen (M62 NE 3) die Zensierungsvariable ZEN mit einer 0 überschrieben. Dahinter steht die Überlegung, daß die Männer in den unqualifizierten manuellen Berufen solange dem Risiko eines Wechsels zu den qualifizierten manuellen Berufen ausgesetzt sind, bis eines der konkurrierenden Ereignisse (Wechsel in eine der anderen Berufsgruppen) eingetreten ist. Wie in Abschnitt 4.3 bereits ausgeführt, gehen diese zensierten Episoden in die Partial-Likelihood-Schätzung insofern ein, als sie die Risikomenge der jeweils später eintretenden Ereignisse verringern.
Mit den restlichen Anweisungen im FORM- und REGRESSION-Paragraph wird dann, wie bereits mehrfach beschrieben, ein Cox-Modell mit den bekannten Kovariablen Bildung (BILDG), Prestige (M59), Anzahl der vorher ausgeübten Berufe (BANZ), Berufserfahrung (BERF) und den Kohorten-Dummies KOHO2 und KOHO3 geschätzt.
Das Ergebnis der Schätzung dieses spezifischen Übergangs ist in Tabelle 5.5 zu finden. Nach dem GLOBAL CHI-SQUARE-Wert von 26,11, der bei sechs Freiheitsgraden einer Signifikanzwahrscheinlichkeit von 0,0002 entspricht, muß die Hypothese abgelehnt werden, keine der eingeführten Kovariablen erkläre etwas in bezug auf die Rate, von einem unqualifizierten manuellen Beruf zu einem qualifizierten manuellen Beruf aufzusteigen ($\alpha = 0{,}05$). ·
In der Spalte der standardisierten $\hat{\beta}$-Koeffizienten (COEFF./S.E.) sieht man allerdings, daß diesmal nur die Variable Bildung (BILDG) tatsächlich etwas erklärt. Alle anderen standardisierten $\hat{\beta}$-Koeffizienten sind absolut kleiner als der Wert 1,96 und damit bei einer Irrtumswahrscheinlichkeit von 0,05 nicht signifikant. Die Neigung, von einem unqualifizierten manuellen Beruf zu einem qualifizierten manuellen Beruf zu wechseln, hängt also nur von der Bildung ab und steigt mit jedem zusätzlichen Ausbildungsjahr um 61,91 Prozent [(1,6191 − 1) · 100% = 61,91%].
Obwohl dieses Ergebnis unter inhaltlichen Gesichtspunkten sehr plausibel ist, sollte man aber darauf hinweisen, daß die Ergebnisse der Signifikanztests von

Tabelle 5.5: Ergebnis des Cox-Modells aus Programmbeispiel 5.5

```
   LOG LIKELIHOOD =    -404.0954
GLOBAL CHI-SQUARE =      26.11   D.F.=    6   P-VALUE =0.0002
```

	VARIABLE	COEFFICIENT	STANDARD ERROR	COEFF./S.E.	EXP(COEFF.)
65	BILDG	0.4819	0.1080	4.4617	1.6191
59	M59	-0.0273	0.0144	-1.8920	0.9731
67	BANZ	0.0863	0.0761	1.1340	1.0902
66	BERF	-0.0046	0.0025	-1.8230	0.9955
68	KOH02	-0.1413	0.2825	-0.5002	0.8682
69	KOH03	0.2323	0.3264	0.7116	1.2615

der Anzahl der tatsächlichen Ereignisse und damit indirekt vom Stichprobenumfang abhängen. So reduziert sich die Zahl der Episoden durch die Beschränkung auf die Männer und den Ausgangszustand der unqualifizierten manuellen Berufe zunächst von 6.732 im Gesamtfile auf 705. Da nur wenige der in unqualifizierten manuellen Berufen tätigen Männer den Aufstieg zu einem qualifizierten manuellen Beruf schaffen, vermindert sich die Anzahl der Ereignisse schließlich auf 72, was einem Anteil von 89,79 Prozent an Zensierungen entspricht. Es könnte deswegen durchaus sein, daß schwächere Effekte sich erst bei einem größeren Stichprobenumfang und einer damit verbundenen höheren Zahl von Ereignissen als signifikant erweisen. Einer Disaggregierung der Stichprobe nach einer Vielzahl von Ausgangs- und Endzuständen sind somit bei Mehr-Zustands-Modellen enge Grenzen durch den Stichprobenumfang gesetzt.

Ähnlich wie im obigen Beispiel, das nur die programmtechnische Realisierung von Mehr-Zustands-Modellen zeigen sollte, könnte man auch alle anderen Übergänge schätzen und die Koeffizienten der Cox-Regressionsrechnungen miteinander vergleichen.

Wir wollen statt dessen jedoch ein zweites Beispiel für die Anwendung von Mehr-Zustands-Modellen geben, das von Carroll und Mayer (1986) vorgelegt wurde. Auch sie untersuchten Karriereprozesse auf der Grundlage der Lebensverlaufsstudie. Ziel war es, zu zeigen, inwieweit Aufwärts-, Abwärts- und horizontale Mobilität sowie die Mobilität innerhalb und zwischen Unternehmen von der Organisationsgröße abhängen, wenn man verschiedene individuelle und strukturelle Merkmale als Hintergrundsvariablen kontrolliert. Zu prüfen war die Hypothese, die Organisationsgröße wirke sich negativ auf die Berufswechselrate aus, weil Beschäftigte in großen Organisationen in der Erwartung einer Beförderung weniger geneigt sind, ihre gegenwärtige Position zu verlassen. Umgekehrt sehen Beschäftigte in kleinen Organisationen nur durch einen Jobwechsel eine Möglichkeit, ihre gegenwärtige Lage zu verbessern.

Im Vergleich dazu sollten sich die erwarteten Differenzen im Mobilitätsmuster großer und kleiner Firmen allerdings als paradox erweisen, wenn man davon ausgeht, daß bürokratische Karriereleitern „rationalisiert" sind. „Rationalisie-

rung" meint in diesem Zusammenhang, daß universalistische Standards zur Bestimmung der Rate und des Niveaus von Beförderungen herangezogen werden. Zu viele Beförderungen, zu schnell ausgesprochen, oder Beförderungen, die einen zu großen Zuwachs an Belohnungen bedeuteten, bedrohen nämlich die Legitimität etablierter Hierarchiesysteme. Statt dessen bieten große Organisationen deswegen bürokratisierte Karriereleitern mit vielen Stufen und geringfügigen relativen Zuwächsen an. Es ist damit zu vermuten, daß sich bei der tatsächlich realisierten Mobilität zwar die Organisationsgröße positiv auf den Berufswechsel innerhalb eines Unternehmens, aber negativ auf die Berufsbewegungen zwischen den Unternehmen auswirkt. Diese Jobbewegungen im Unternehmen müßten allerdings eher horizontalen als nach oben gerichteten Charakter haben.

Von einer Aufwärtsbewegung sprechen beide Autoren dann, wenn sich der Anfangsverdienst eines nachfolgenden Berufes um mehr als 15 Prozentpunkte des Endverdienstes des vorhergehenden Berufes gesteigert hat. Ein beruflicher Abstieg liegt dann vor, wenn sich der Verdienst vermindert, und von einer horizontalen beruflichen Mobilität wird dann gesprochen, wenn die Veränderung des Verdienstes zwischen 0 und 15 Prozentpunkten liegt.

Wie im vorhergehenden Beispiel bereits gezeigt, werden bei der Schätzung von Mehr-Zustands-Modellen die jeweils nicht interessierenden Ereignisse als zensiert behandelt. Wird beispielsweise das Ereignis „Aufwärtsmobilität" untersucht, dann werden neben der normalen Zensierung alle anderen Berufswechsel, die nicht mit einem Zuwachs im Verdienst um mehr als 15 Prozentpunkte verbunden sind, bei der Schätzung ebenfalls zensiert (vgl. die Spalte UPWARD MOVES). Entsprechend werden alle anderen Modelle in Tabelle 5.6 geschätzt.

Das erste Modell in Tabelle 5.6 betrifft zunächst alle Jobwechsel (ANY MOVE). Aus dieser Gleichung ist ersichtlich, daß für die jeweils jüngeren Kohorten (C2 ≙ Kohorte 1939–41 und C3 ≙ Kohorte 1949–51), für Frauen (SEX) und für die verschiedenen Niveaus der Allgemeinbildung (GED) jeweils signifikant höhere Berufswechselraten prognostiziert werden. Signifikant niedrigere Raten des Jobwechsels ergeben sich dagegen für den ersten Beruf (JN1), die Statusvariable (STATUS), den Verdienst (LNWAGE) und für die jeweils höheren Stufen der beruflichen Bildung (OED). Die Variable Organisationsgröße (LNSIZE), die eigentlich im Zentrum der Untersuchung steht, wirkt sich wie erwartet ebenfalls negativ auf die Neigung zum Berufswechsel aus. Dies spricht für die Hypothese, daß die Beschäftigten in großen Organisationen in der Erwartung von Beförderungen weniger geneigt sind, ihren gegenwärtigen Beruf zu verlassen, als Personen, die in kleinen Organisationen arbeiten.

Die anderen Schätzungen in Tabelle 5.6 sprechen für die „Rationalisierungs-Hypothese". Die Organisationsgröße hat ihren größten negativen Effekt in bezug auf die Aufwärtsmobilität (UPWARD MOVES) und weniger starke Wirkungen auf die horizontalen (LATERAL MOVES) sowie die nach unten gerichteten (DOWNWARD MOVES) Jobbewegungen. Wichtiger noch, die

Tabelle 5.6: Partial-Likelihood-Schätzungen der Effekte der Organisationsgröße auf die Raten zum Berufswechsel (Standardfehler der β-Schätzungen in Klammern)

	Any Move	Upward Moves	Laternal Moves	Downward Moves	Within Firm	Across Firms	Upward Within Firm	Laternal Within Firm	Upward Across Firms	Laternal Across Firms
C2	.140 (.077)	.188 (.046)	.071 (.088)	.241 (.114)	.120 (.114)	.192 (.054)	.138 (.211)	.049 (.202)	.147 (.084)	.060 (.099)
C3	.178 (.092)	.309 (.053)	.226 (.101)	.331 (.130)	.467 (.132)	.203 (.064)	.488 (.253)	.565 (.224)	.146 (.100)	.124 (.116)
SEX	.128 (.072)	-.111 (.040)	-.063 (.083)	.853 (.100)	-.563 (.109)	-.089 (.049)	-.527 (.210)	-.609 (.205)	-.074 (.077)	.080 (.093)
LFX	-.006 (.001)	-.005 (.000)	-.005 (.001)	-.005 (.001)	-.001 (.001)	-.006 (.000)	-.001 (.002)	-.002 (.001)	-.007 (.001)	-.006 (.001)
JN1	-.126 (.083)	-.285 (.048)	-.334 (.094)	-.701 (.126)	.149 (.120)	-.420 (.057)	.142 (.228)	.041 (.209)	-.157 (.089)	-.483 (.109)
STATUS	-.008 (.002)	-.005 (.001)	-.004 (.002)	-.007 (.003)	-.007 (.003)	-.004 (.002)	-.012 (.006)	-.005 (.005)	-.007 (.003)	-.004 (.003)
GED	.293 (.069)	.169 (.039)	-.115 (.075)	-.009 (.095)	.323 (.095)	.136 (.048)	.569 (.177)	.409 (.153)	.247 (.076)	-.009 (.089)
OED	-.048 (.054)	-.026 (.029)	-.022 (.056)	-.223 (.074)	.180 (.069)	-.108 (.036)	.089 (.133)	.134 (.111)	-.076 (.059)	.010 (.066)
LNSIZE	-.088 (.015)	-.045 (.008)	-.025 (.016)	-.028 (.020)	.074 (.020)	-.082 (.010)	.063 (.039)	.146 (.034)	-.116 (.017)	-.078 (.018)
LNWAGE	-.486 (.039)	-.189 (.025)	-.084 (.052)	.367 (.073)	-.382 (.061)	-.178 (.030)	-.665 (.107)	-.301 (.121)	-.452 (.042)	-.040 (.059)
χ^2	644.1	570.9	111.9	206.3	142.4	446.6	76.6	72.8	597.3	113.7
d.f.	10	10	10	10	10	10	10	10	10	10

$\hat{\beta}$-Koeffizienten der Variablen Organisationsgröße unterscheiden sich bei Berufswechseln innerhalb eines Unternehmens (WITHIN FIRM) und zwischen den Unternehmen (ACROSS FIRMS) im Vorzeichen. Die Chancen beruflicher Veränderungen sind innerhalb des gleichen Unternehmens um so höher, je größer dieses ist, während Wechsel zwischen den Unternehmen mit abnehmender Organisationsgröße zunehmen.
Die restlichen vier Spalten in Tabelle 5.6 geben einen detaillierteren Einblick in den Zusammenhang von Berufsmobilität und Unternehmensgröße. Obgleich die Variable Organisationsgröße (LNSIZE) auf beide Arten des Berufswechsels innerhalb eines Unternehmens (UPWARD WITHIN FIRM und LATERAL WITHIN FIRM) eine positive Wirkung hat, ist sie nur für die horizontalen beruflichen Bewegungen (LATERAL WITHIN FIRM) signifikant. Innerhalb der großen Unternehmen gibt es also tatsächlich nur moderate Aufstiege, was der „Rationalisierungs-Hypothese" entspricht. Umgekehrt wirkt sich die Organisationsgröße bei beruflichen Veränderungen mit gleichzeitigem Unternehmenswechsel sowohl auf die Aufstiege (UPWARD ACROSS FIRMS) als auch auf die horizontalen Bewegungen (LATERAL ACROSS FIRMS) negativ aus. Das heißt, je kleiner das Unternehmen ist, desto wahrscheinlicher sind Aufstiege durch Firmenwechsel.
Bei der Gegenüberstellung der verschiedenen Modelle in Tabelle 5.6 ist allerdings wiederum auf das Problem hinzuweisen, daß diese jeweils auf einer unterschiedlichen Zahl von Ereignissen beruhen. Wenn sich wie in diesem Beispiel die Variable Organisationsgröße (LNSIZE) im Modell UPWARD WITHIN FIRM nicht als signifikant erweist, dann könnte das auch damit in Zusammenhang stehen, daß bei dem gegebenen Stichprobenumfang vergleichsweise nur sehr wenige Ereignisse zur Schätzung dieses Übergangs zur Verfügung stehen und sich nur die besonders starken Effekte als signifikant herausstellen können.
Obwohl bei den Mehr-Zustands-Modellen die Zahl der zu schätzenden Parameter mit der Zahl der Ausgangs- und Endzustände spürbar steigt und zur Schätzung entsprechend große Datenmengen benötigt werden, hat doch das soeben vorgestellte Beispiel deutlich gemacht, daß man mit Hilfe der Ereignisanalyse durch die geschickte Wahl multipler Zielzustände außerordentlich differenzierte inhaltliche Argumentationen überprüfen kann. Da sich bei den parametrischen Verfahren an der Modellierung von Mehr-Zustands-Modellen nichts ändert und die dort ausgeführten Beispiele ohne weiteres auch auf den Mehr-Zustands-Fall übertragen werden können, soll im folgenden bei den Anwendungsbeispielen nicht mehr näher auf den Mehr-Zustands-Fall eingegangen werden.

Kapitel 6:
Parametrische Regressionsmodelle

Der Vorteil eines Cox-Modells, den Einfluß von Kovariablen ohne weitere Annahmen über die Form des Hazardratenverlaufs bestimmen zu können, wird natürlich mit dem Nachteil erkauft, daß ein Teil des Regressionsmodells unbekannt und unspezifiziert bleibt. Dies ist nicht problematisch, solange keine spezifischen Hypothesen über den Hazardratenverlauf vorliegen, die Veränderung des Ereignisrisikos über die Verweildauer hinweg tatsächlich unbekannt ist oder so unsystematisch ausfällt, daß sie sich nicht durch ein parametrisches Modell approximieren läßt. Immer dann allerdings, wenn man begründete Hypothesen oder klare Vorstellungen über die Verweildauerabhängigkeit besitzt, sollte man – soweit dies möglich ist – auf parametrische Modelle zurückgreifen, da mit der Partial-Likelihood-Schätzung ein Effizienzverlust in Kauf genommen wird.

In diesem Kapitel soll für die in der Forschungspraxis der Wirtschafts- und Sozialwissenschaften gebräuchlichsten parametrischen Modelle, die bereits in den Abschnitten 3.2.2, 3.3.2, 3.6.3 und 3.9.2 ausführlich dargestellt wurden, eine Reihe von Anwendungs- und Interpretationsbeispielen gegeben werden. Nach der graphischen Überprüfung der Verteilungsannahmen (Abschnitt 6.1) wird zuerst ausführlich das Exponential-Modell, seine Interpretation und auf die Überprüfung der Residuen eingegangen (Abschnitt 6.2). Die Aufnahme zeitveränderlicher unabhängiger Variablen bei parametrischen Modellen wird dann in Abschnitt 6.3 gezeigt. Danach folgen Beispiele zu Modellen mit periodisierter Verweildauer (Abschnitt 6.4). Spezielle Verweildauermodelle werden im Abschnitt 6.5 dargestellt. Dabei werden ausführliche Interpretationsbeispiele und Residuentests zur Gompertz-(Makeham-) (Abschnitt 6.5.1), zur Weibull- (Abschnitt 6.5.2) und zur log-logistischen Verteilung (Abschnitt 6.5.3) gegeben. Anwendungsbeispiele zur unbeobachteten Populationsheterogenität beschließen die Ausführungen zu den parametrischen Modellen (Abschnitt 6.6).

Sieht man einmal von jenen Programmsystemen ab, mit denen durch benutzereigene Makros und Unterprogramme (wie etwa in GLIM oder P3R von BMDP) Maximum-Likelihood-Schätzungen berechnet werden können und die in der Regel höhere mathematisch-statistische Kenntnisse sowie Programmiererfahrung voraussetzen, so existiert heute noch kein allgemein zugängliches und

benutzerfreundliches Programmpaket zur Schätzung der oben angesprochenen parametrischen Modelle[1].

Es ist deshalb auch ein Ziel dieses Kapitels, anhand von Beispielen den Leser mit mehreren Programmsystemen zur Schätzung parametrischer Modelle etwas vertraut zu machen. Dementsprechend werden wir das Exponential-Modell und das Gompertz-Modell mit dem von Trond Petersen (1985) geschriebenen FORTRAN-Unterprogramm P3RFUN, das jeder Benutzer leicht in das Unterprogramm P3R von BMDP einbinden kann (siehe dazu das Quellenprogramm im Anhang 2), und mit dem Programmsystem RATE (Version 2E) von Nancy Tuma (1979) schätzen. Zur Schätzung des Weibull- und des log-logistischen Modells wird ebenfalls auf das Programm P3RFUN sowie auf das Programmsystem GLIM (siehe dazu das Listing der dazu verwendeten Makros im Anhang 3) (Baker/Nelder 1978; Roger/Peacock 1983) zurückgegriffen. Beim Heterogenitäts-Modell (mit einer Gamma-Verteilung) und den parametrischen Modellen mit periodisierter Verweildauer wird schließlich das Programmsystem RATE herangezogen.

6.1 Graphische Überprüfung der Verteilungsannahmen

Da bei den parametrischen Modellen der Verweildauerabhängigkeit das Regressionsmodell vollständig spezifiziert wird, kommt der Frage der Angemessenheit des jeweils gewählten Verteilungsmodells natürlich besondere Wichtigkeit zu. Bei der Darstellung der einzelnen parametrischen Modelle werden wir später noch auf die Möglichkeit zurückkommen, die Residuen nach der Schätzung eines Modells zu analysieren sowie die Möglichkeit aufzeigen, die Schätzung eines Modells durch die Schätzung eines allgemeineren Verteilungsmodells zu überprüfen, das dieses als Sonderfall enthält (z. B. enthält das Weibull-Modell das Exponential-Modell als Spezialfall). Zunächst befassen wir uns aber damit, daß man bereits aus dem Verlauf der durch parameterfreie Methoden geschätzten Survivorfunktion erste Hinweise auf das Vorliegen eines bestimmten Verteilungstyps erhalten kann (vgl. Abschnitt 3.7.1).

Die Survivorfunktion muß dabei so transformiert werden, daß mit dem bloßen Auge tatsächlich eine Beurteilung des Verteilungstyps vorgenommen werden kann. Dies ist dann besonders einfach, wenn die Beziehung in die Form einer Geraden ($y = a + bx$) gebracht wird und sich Abweichungen von der Verteilungsannahme als Abweichungen von der Geraden niederschlagen.

---

[1] Mit der Prozedur LIFEREG von SAS können beispielsweise zwar das Exponential-, das Weibull-, das log-normale und das log-logistische Modell, aber nicht das Gompertz-Modell geschätzt werden.

Für das *Exponentialmodell* lautet die Survivorfunktion nach (3.2.17)

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

Durch Logarithmieren ergibt sich die Gerade

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

mit: $y = \ln S(t)$, $a = 0$, $b = -\lambda$, $x = t$.
Trägt man somit den geschätzten Verlauf $\hat{S}(t)$ gegen t auf, dann muß sich bei Erfüllung der Verteilungsannahme approximativ eine Gerade durch den Ursprung ergeben. Unter der Annahme, daß das Modell gültig ist, kann man auf der Basis des Verlaufs der so geschätzten und transformierten Survivorfunktion Kleinst-Quadrate-Schätzungen für die Parameter a und b angeben. Dann müßte $\hat{a} \sim 0$ und $-\hat{b} \sim \hat{\lambda}$ sein.
Für das *Weibull-Modell* lautet die Survivorfunktion nach (3.2.20)

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

Durch Logarithmierung ergibt sich

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

und durch nochmaliges Logarithmieren folgt die Gerade

$$\ln(-\ln S(t)) = \alpha \ln\lambda + \alpha \ln t,$$

mit: $y = \ln(-\ln S(t))$, $a = \alpha \ln\lambda$, $b = \alpha$, $x = \ln t$.
Trägt man die doppelt logarithmierte Survivorfunktion $\ln(-\ln\hat{S}(t))$ gegen den Logarithmus der Zeit lnt auf, dann muß sich bei Erfüllung der Verteilungsannahme approximativ eine Gerade ergeben. Unter der Annahme, daß das Modell gültig ist, kann man auf der Basis des Verlaufs der so geschätzten und transformierten Survivorfunktion Kleinst-Quadrate-Schätzungen für die Parameter a und b angeben. Dann müßte $\hat{a} \sim \hat{\alpha}\ln\hat{\lambda}$ und $\hat{b} \sim \hat{\alpha}$ sein.
Für das *Gompertz-Modell* lautet die Survivorfunktion

$$S(t) = \exp\left(-\frac{\lambda_0}{\gamma_0}(\exp(\gamma_0 t)-1)\right).$$

Durch Logarithmierung erhält man

$$\ln S(t) = -\frac{\lambda_0}{\gamma_0}(\exp(\gamma_0 t)-1).$$

Für den Zeitpunkt t+1 ergibt sich entsprechend

$$\ln S(t+1) = -\frac{\lambda_0}{\gamma_0}(\exp(\gamma_0(t+1))-1).$$

Bildet man die Differenz, dann ergibt sich

$$\ln\frac{S(t)}{S(t+1)} = -\frac{\lambda_0}{\gamma_0}\exp(\gamma_0 t)(1-\exp(\gamma_0))$$

und durch nochmaliges Logarithmieren folgt die Gerade

$$\ln(\ln\frac{S(t)}{S(t+1)}) = \ln(-\frac{\lambda_0}{\gamma_0}(1-\exp(\gamma_0))) + \gamma_0 t,$$

mit: $y = \ln(\ln\frac{S(t)}{S(t+1)})$, $a = \ln(-\frac{\lambda_0}{\gamma_0}(1-\exp(\gamma_0)))$, $b = \gamma_0$, $x = t$.

$\ln(\ln\frac{\hat{S}(t)}{\hat{S}(t+1)})$, gegen t aufgetragen, muß somit beim Vorliegen einer Gompertz-Verteilung ungefähr eine Gerade ergeben. Unter der Annahme, daß das Modell gültig ist, kann man auf der Grundlage des Verlaufs der geschätzten und transformierten Survivorfunktion wieder Kleinst-Quadrate-Schätzungen für die Parameter a und b dieser Geraden angeben. Dann müßte $\hat{a} \sim \ln(-\frac{\hat{\lambda}_0}{\hat{\gamma}_0}(1-\exp(\hat{\gamma}_0)))$ und $\hat{b} \sim \hat{\gamma}_0$ sein.

Schließlich lautet für das *log-logistische Modell* die Survivorfunktion nach (3.2.29)

$$S(t) = \frac{1}{1+(\lambda t)^{\alpha}}.$$

Daraus folgt

$$1-S(t) = \frac{(\lambda t)^{\alpha}}{1+(\lambda t)^{\alpha}}$$

und damit

$$\frac{1-S(t)}{S(t)} = (\lambda t)^{\alpha}.$$

Durch Logarithmierung ergibt sich die Gerade

$$\ln\frac{1-S(t)}{S(t)} = \alpha\ln\lambda + \alpha\ln t,$$

mit: $y = \ln\frac{1-S(t)}{S(t)}$, $a = \alpha\ln\lambda$, $b = \alpha$, $x = \ln t$.

Wenn die Verteilungsannahme richtig ist, dann müßte die so transformierte Survivorfunktion wiederum approximativ eine Gerade ergeben. Unter der Annahme, daß das Modell gültig ist, kann man auf der Grundlage des Verlaufs der geschätzten und transformierten Survivorfunktion wieder Kleinst-Quadrate-Schätzungen für die Parameter a und b dieser Geraden angeben. Dann müßte $\hat{a} \sim \hat{\alpha}\ln\hat{\lambda}$ und $\hat{b} \sim \hat{\alpha}$ sein.

Als ein Anwendungsbeispiel für den Einsatz der graphischen Verfahren zur Überprüfung der Verteilungsannahmen soll im folgenden untersucht werden, ob das Berufswechselrisiko der Männer über die Verweildauer hinweg konstant ist oder eher einer Weibull-, einer Gompertz- oder einer log-logistischen Verteilung folgt. Dazu schätzt man beispielsweise zuerst mit dem Programmpaket

SPSS die Survivorfunktion $\hat{S}(t)$ und schreibt diese als Rohdatenfile heraus (OPTIONS 8)[2]. Die Intervallbreite wird dabei möglichst klein gewählt (in diesem Fall ein Monat), um eine möglichst genaue Schätzung zu erhalten.
Die Rohdaten können dann in einem zweiten Schritt mit SPSS wieder eingelesen, nach den oben beschriebenen Transformationen umgeformt und mit SCATTERGRAM geplottet werden:

Programmbeispiel 6.1:

```
VARIABLE LIST    TYPE,TABNR,BOINT,NEIT,WITHDR,RISK,EVENTS,
                 TERMI,PROBSURV,SURV,DENS,HAZ,SE1,SE2,SE3
INPUT FORMAT     FIXED (F2.0,F5.0,F6.2,4F8.2/7X,5F8.6,3F7.4)
INPUT MEDIUM     DISK
N OF CASES       300
COMPUTE          TIME = BOINT
SELECT IF        (TIME LE 160)
COMPUTE          LOGTIME = LN(BOINT)
COMPUTE          EXPOSURV = LN(SURV)
COMPUTE          WEIBSURV = LN(-LN(SURV))
LAG              GOSURV = SURV
COMPUTE          GOMPSURV = LN(LN(GOSURV/SURV))
COMPUTE          LOGLOG =LN((1-SURV)/SURV)
TASK NAME        **** GRAPHISCHER TEST - EXPONENTIALVERTEILUNG ****
SCATTERGRAM      EXPOSURV(-4,0) WITH BOINT (0,160)
OPTIONS          4
TASK NAME        ***** GRAPHISCHER TEST - WEIBULLVERTEILUNG *******
SCATTERGRAM      WEIBSURV(-3,2) WITH LOGTIME (0,6)
OPTIONS          4
TASK NAME        ***** GRAPHISCHER TEST - GOMPERTZVERTEILUNG ******
SCATTERGRAM      GOMPSURV(-20,20) WITH TIME (0,160)
OPTIONS          4
TASK NAME        ***** GRAPHISCHER TEST - LOG-LOG-VERTEILUNG *******
SCATTERGRAM      LOGLOG(-3,5) WITH LOGTIME(0,6)
OPTIONS          4
FINISH
```

---

[2] SPSS-Programm zur Erstellung einer Sterbetafel in Rohdatenform:

```
SPACE            100000
GET FILE         DATA
SELECT IF        (M3 EQ 1)
COMPUTE          DUR = M51 - M50 + 1
COMPUTE          ZEN = 1
IF               (M51 EQ M47) ZEN = 0
RAW OUTPUT UNIT  15
SURVIVAL         TABLES = DUR/
                 INTERVALS = THRU 300 BY 1/
                 STATUS = ZEN (1)/
                 PLOTS (LOGSURV)
OPTIONS          8
FINISH
```

Eine genaue Beschreibung des Formats und der Variablen findet man bei Hull/Nie 1981.

Das Ergebnis dieses Vorgehens ist in den Abbildungen 6.1 bis 6.4 zu sehen. Die visuelle Inspektion dieser Kurven läßt sich noch dadurch erleichtern, daß man für die einzelnen Schaubilder mit Hilfe einer Kleinst-Quadrate-Schätzung die Geraden ermittelt und anhand der R^2-Werte die Anpassung der Datenpunkte an diese Geraden beurteilt:

Programmbeispiel 6.2:

```
VARIABLE LIST    TYPE,TABNR,BOINT,NEIT,WITHDR,RISK,EVENTS,
                 TERMI,PROBSURV,SURV,DENS,HAZ,SE1,SE2,SE3
INPUT FORMAT     FIXED (F2.0,F5.0,F6.2,4F8.2/7X,5F8.6,3F7.4)
INPUT MEDIUM     DISK
N OF CASES       300
COMPUTE          TIME = BOINT
SELECT IF        (TIME LE 160)
COMPUTE          LOGTIME = LN(BOINT)
COMPUTE          EXPOSURV = LN(SURV)
COMPUTE          WEIBSURV = LN(-LN(SURV))
LAG              GOSURV = SURV
COMPUTE          GOMPSURV = LN(LN(GOSURV/SURV))
COMPUTE          LOGLOG =LN((1-SURV)/SURV)
TASK NAME        REGRESSION TEST VERTEILUNGEN
REGRESSION       VARIABLES = EXPOSURV,WEIBSURV,GOMPSURV,
                             GOMPSURV,LOGLOG,
                             TIME,LOGTIME/
                 REGRESSION = EXPOSURV WITH TIME (2)/
                 REGRESSION = WEIBSURV WITH LOGTIME (2)/
                 REGRESSION = GOMPSURV WITH TIME (2)/
                 REGRESSION = LOGLOG WITH LOGTIME (2)/
STATISTICS       ALL
FINISH
```

Dabei ergeben sich für die verschiedenen Verteilungstypen die in Tabelle 6.1 angeführten Geraden-Schätzungen.

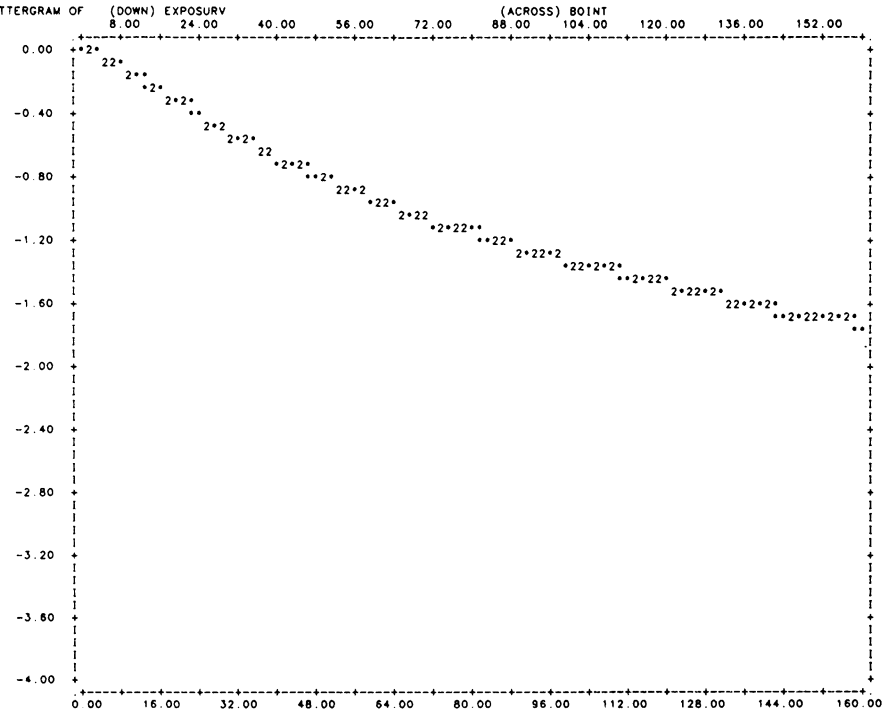
Tabelle 6.1: Regressionsschätzung der Geraden zur Beurteilung der Verteilungsannahmen

Modell	$\hat{a}$	$\hat{b}$	R^2
Exponential-Verteilung	-0,2029	-0,0106	0,9642
Weibull-Verteilung	-3,6950	0,8621	0,8534
Gompertz-Verteilung	-3,9092	-0,0095	0,4114
Log-logistische Verteilung	-4,2100	1,1313	0,9068

Für die Exponential-Verteilung erhält man aus Tabelle 6.1 als Näherung für den Koeffizienten λ den Wert 0,0106 [$\hat{b} = -\hat{\lambda} = -0{,}0106$]. Der Kurvenverlauf in Abbildung 6.1 ist allerdings leicht nach unten gekrümmt, und auf der Grundlage

der Regressionsschätzung ergibt sich, daß die Gerade nicht durch den Ursprung verläuft, sondern die y-Achse bei â = -0,2029 schneidet. Trotz dieser Abweichungen hat die Exponential-Verteilung von allen überprüften Verteilungstypen noch den „besten Datenfit" ($R^2 = 0,9642$).

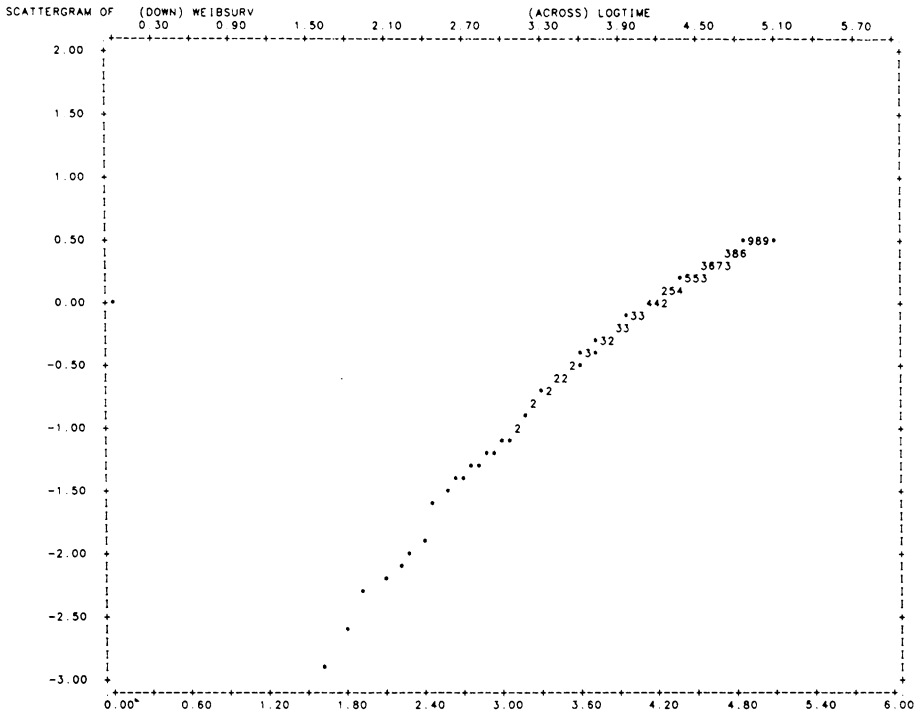
Abbildung 6.1: Graphischer Test der Exponential-Verteilung



Bei der Weibull-Verteilung schätzen wir $\hat{\alpha} = 0,8621$ [$\hat{b} = \hat{\alpha} = 0,8621$] und $\hat{\lambda} = 0,0138$ [$-3,6950 = 0,8621 \ln\hat{\lambda}$]. Da der Koeffizient $\hat{\alpha} < 1$ ist, erwartet man einen monoton fallenden Verlauf der Hazardrate über die Verweildauer hinweg (vgl. Abbildung 3.5). Allerdings ist die „Anpassung" an eine Gerade schon weniger gut als im Exponential-Modell ($R^2 = 0,8534$), was auch in Abbildung 6.2 deutlich sichtbar ist.

Für das Gompertz-Modell ergibt sich als Schätzung für den Koeffizienten γ_0 der Wert -0,0095 [$\hat{b} = \hat{\gamma}_0 = -0,0095$] und für den Koeffizienten λ_0 der Wert 0,00202 [$-3,9092 = \ln(-\frac{\hat{\lambda}_0}{-0,0095}(1-\exp(-0,0095)))$]. Da $\hat{\gamma}_0 < 0$ ist, ist wiederum ein mo-

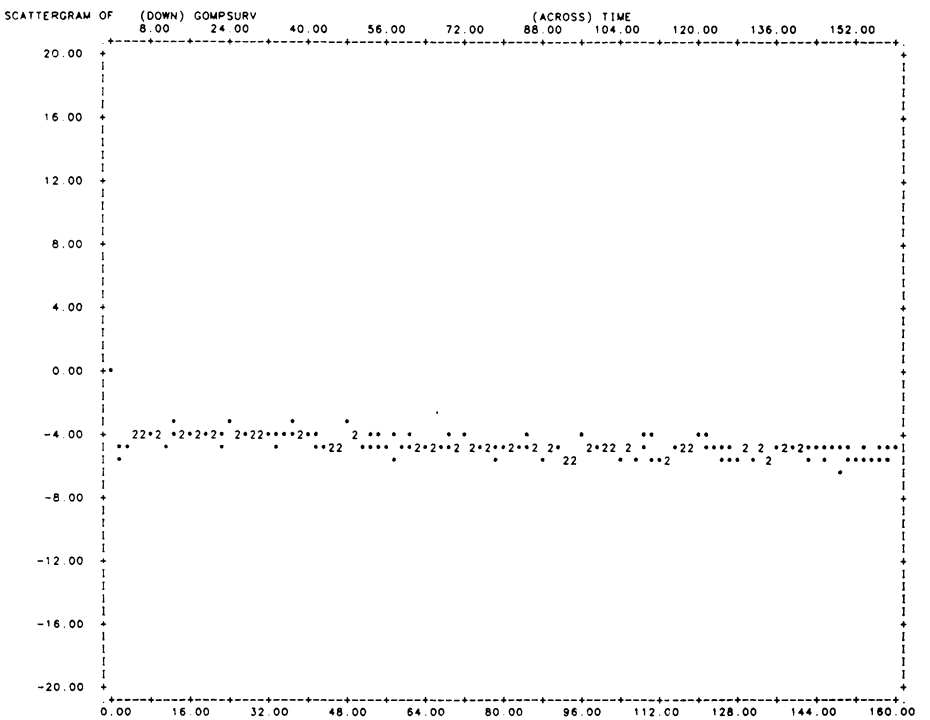
Abbildung 6.2: Graphischer Test der Weibull-Verteilung



noton fallender Verlauf der Hazardrate über die Verweildauer hinweg zu erwarten (vgl. Abbildung 3.8). Die „Anpassung" der Datenpunkte an die Gerade in Abbildung 6.3, mit einem R^2 von 0,4114, ist jedoch noch einmal schlechter als für die Weibull-Verteilung.

Schließlich erhält man als Schätzung für den α-Koeffizienten der log-logistischen Verteilung den Wert 1,1313 [$\hat{b} = \hat{\alpha} = 1{,}1313$] und für den λ-Koeffizienten den Wert 0,024 [$-4{,}2100 = 1{,}1313 \ln\hat{\lambda}$]. Nach dieser Schätzung ergäbe sich über die Verweildauer hinweg ein nicht-monotoner Verlauf (vgl. Abbildung 3.7): Das Risiko, den Beruf zu wechseln, würde bei den Männern zu Beginn jedes neuen Berufes steigen, einen Höhepunkt nach etwa sieben Monaten [$(\hat{\alpha} - 1)^{1/\hat{\alpha}}/\hat{\lambda} = (1{,}1313 - 1)^{1/1{,}1313}/0{,}024 = 6{,}92$] erreichen und mit zunehmender Verweildauer wieder fallen. Vergleicht man die „Anpassung" der Datenpunkte an die geschätzte Gerade in Abbildung 6.4 mit der „Anpassung" der Datenpunkte im Weibull- oder Gompertz-Modell, dann ist diese wieder relativ gut (R^2 = 0,9068).

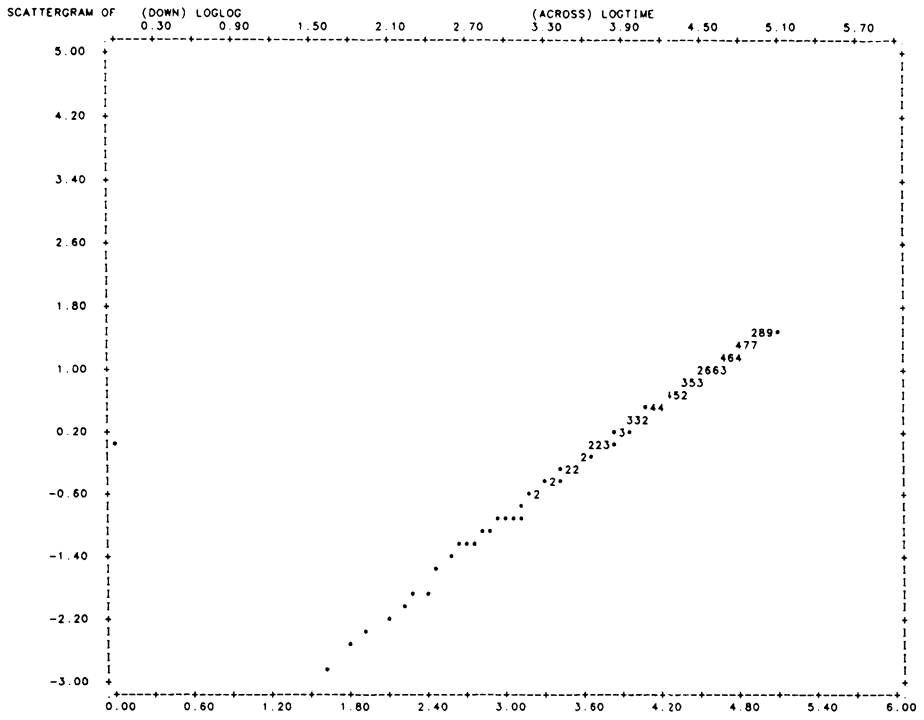
Abbildung 6.3: Graphischer Test der Gompertz-Verteilung



Insgesamt erhält man also aufgrund der graphischen Überprüfung der Verteilungsannahmen ein doch recht widersprüchliches Bild über den Verlauf des Berufswechsel-Risikos bei Männern: Die Überprüfung des Exponential-Modells, das zwar von allen untersuchten Modellen den „besten Fit" hat, jedoch die Annahme $a \sim 0$ verletzt und einen nach unten gebogenen Verlauf aufweist, führt zu einem Modell mit konstanter Rate; die Überprüfungen der Weibull- und der Gompertz-Verteilung legen, allerdings bei „schlechterem Datenfit", gemeinsam einen monoton fallenden Verlauf der Rate nahe, und nach der Überprüfung der log-logistischen Verteilung ist man, bei „relativ gutem Datenfit", zur Annahme eines nicht-monotonen Risikos geneigt.

Für jeden dieser Verläufe lassen sich gute inhaltliche Gründe vorbringen, und es wird deutlich, daß man die Methode der graphischen Überprüfung der Verteilungsannahmen nicht als abschließenden Test im Prozeß der Modellfindung mißverstehen sollte. Es lassen sich damit nur erste Hinweise über eventuell vorliegende Verweildauerabhängigkeiten gewinnen.

Abbildung 6.4: Graphischer Test der log-logistischen Verteilung



Auch sollte man bei diesen graphischen Tests die Entscheidung zwischen den verschiedenen Verteilungsmodellen nicht alleine auf der Grundlage der Größe des R^2-Wertes treffen, da die bei der Kleinst-Quadrate-Schätzung gemachte Annahme der Homoskedastizität nicht erfüllt ist und die Störgrößen korreliert sind. Darüber hinaus ist zu bedenken, daß es sich hierbei um eine Überprüfung handelt, die ohne Kontrolle weiterer Kovariablen vorgenommen wurde, und daß die damit einhergehende Aggregation über die Raten heterogener Subgruppen, auch wenn diese zeitkonstant sind, an sich schon zu scheinbarer Verweildauerabhängigkeit führen wird (vgl. Abschnitt 3.9.1).
Zur Überprüfung der Verteilungsannahmen mit einem graphischen Test wäre es deswegen angemessener, wenn man für verschiedene Ausprägungen der unabhängigen Merkmale jeweils die Survivorfunktion in einem ungeschichteten Cox-Modell schätzen und diese Schätzungen wie oben gezeigt transformieren und plotten würde. Zwar kann mit dem Unterprogramm P2L von BMDP für jede Konstellation der Kovariablen die geschätzte Survivorfunktion $S(t|\mathbf{x})$ ge-

plottet werden, indem man mit der PATTERN-Anweisung des PLOT-Paragraphen die entsprechenden Werte der Kovariablen angibt, es ist aber leider nicht möglich, auf diese Schätzungen zuzugreifen und sie den für die graphische Überprüfung notwendigen Transformationen zu unterwerfen.
Aber auch dann, wenn man beispielsweise mit dem Programmsystem GLIM die Survivorfunktionen für verschiedene Subgruppen schätzt und transformiert, was relativ aufwendig ist und auch gute GLIM-Kenntnisse voraussetzt, erhält man eine Vielzahl von Kurven, die meist ebenfalls die Entscheidung für oder gegen ein bestimmtes parametrisches Modell nicht unbedingt erleichtern. Im Prozeß der Modellfindung sollte man sich deswegen nicht zu sehr auf diese graphischen Verteilungstests stützen, sondern zusätzlich auch Residuentests, Vergleiche von Mittelwert- und Medianschätzungen der Verweildauer sowie Vergleiche von verschiedenen Modellen (z. B. zwischen parametrischen Modellen und dem Cox-Modell oder im Falle der Exponential-Verteilung mit einem Weibull-Modell oder einem Modell mit unbeobachteter Heterogenität) heranziehen.

6.2 Modelle ohne Verweildauerabhängigkeit der Hazardrate: Das Exponential-Modell

Das Exponential-Modell geht von der Annahme eines konstanten Ereignisrisikos und damit implizit auch von der Annahme proportionaler Risiken über die gesamte Verweildauer hinweg aus. Es läßt sich einfach interpretieren und wird in der Forschungspraxis häufig als Basis- oder Referenz-Modell benutzt, mit dem dann die Schätzungen komplexerer Verteilungs-Modelle verglichen werden.

6.2.1 Das Exponential-Modell ohne Kovariablen

Zur Illustration der Interpretation des Exponential-Modells soll zunächst auf der Basis des Mehr-Episoden-Falls für das durchschnittliche Berufswechselrisiko der Männer ein Modell ohne Kovariablen, daß heißt nur mit einer Regressionskonstanten β_0, geschätzt werden:

$$\lambda^k(v) = \exp(\beta_0), \qquad k = 1, 2, \ldots .$$

Die Maximum-Likelihood-Schätzung wird dabei mit Hilfe des Unterprogramms P3RFUN von Trond Petersen (1985) berechnet, dessen Quellenprogramm im Anhang 2 zu finden und von jedem Benutzer leicht in das Unterprogramm P3R von BMDP einzubinden ist. Dieses Programm bietet insbesondere den Vorteil, daß man eine Reihe von parametrischen Modellen (das Exponential-, das Weibull-, das Gompertz- und das log-logistische Modell) mit dem weitverbreiteten Programmpaket BMDP schätzen kann und sich nicht in eine neue Syntax spezieller Programmsysteme einarbeiten muß.

Programmbeispiel 6.3:

```
/INPUT UNIT IS 30.
       CODE IS DATA.
/VARIABLE NAMES ARE (63)DUR,(64)ZEN,(65)BILDG,(66)BERF,
                    (67)BANZ,(68)KOHO2,(69)KOHO3,(70)X1,(71)DP.
          ADD IS 9.
/TRANSFORM USE = (M3 EQ 1).
           DUR = M51 - M50 + 1.
           ZEN = 1.
           IF (M51 EQ M47) THEN ZEN = 0.
           IF (M41 EQ 1 AND M42 EQ 1) THEN BILDG = 9.
           IF (M41 EQ 1 AND (M42 EQ 2 OR M42 EQ 3)) THEN BILDG = 11.
           IF (M41 EQ 2 AND M42 EQ 1) THEN BILDG = 10.
           IF (M41 EQ 2 AND (M42 EQ 2 OR M42 EQ 3)) THEN BILDG = 12.
           IF (M41 EQ 3 AND (M42 EQ 1 OR M42 EQ 2 OR M42 EQ 3))
           THEN BILDG = 13.
           IF (M42 EQ 4) THEN BILDG = 17.
           IF (M42 EQ 5) THEN BILDG = 19.
           KOHO2 = 0.
           KOHO3 = 0.
           IF(M48 GE 468 AND M48 LE 504) THEN KOHO2 = 1.
           IF(M48 GE 588 AND M48 LE 624) THEN KOHO3 = 1.
           BERF = M50 - M43.
           BANZ = M5 - 1.
           DP = 0.
           X1 = 0.
/REGRESS DEPENDENT IS DP.
         PARAMETERS ARE 1.
         PRINT IS 0.
         MEANSQUARE IS 1.0.
         ITERATIONS ARE 100.
         LOSS.
/PARAMETER INITIAL ARE -3.8.
           NAMES ARE KONST.
/END.
/COMMENT'
M 1
T 70 63
D 71
C 64
I  0
```

Zunächst wird im Programmbeispiel 6.3, wie das bereits bei den Beispielen zum Cox-Modell der Fall war, der BMDP-Systemfile eingelesen. Zusätzlich zu den bereits bekannten Variablen benötigt man aus programmtechnischen Gründen die Variable X1 (Anfangswert der Verweildauer in Episode k) und DP (abhängige Variable, dependent variable), die beide im TRANSFORM-Paragraph auf den Wert 0 gesetzt werden.
Im REGRESSION-Paragraph ist bei DEPENDENT IS der Name dieser abhängigen Variablen, also DP, anzugeben. Es wird mit diesem Programm genau

ein Parameter (nämlich β_0) geschätzt (PARAMETERS ARE 1). Die Anweisung PRINT IS 0 bewirkt, daß weder die Residuen noch die beobachteten und geschätzten Werte der abhängigen Variablen ausgegeben werden. Mit MEAN-SQUARE IS 1.0 wird der Wert von „residual mean square" bei der Schätzung der asymptotischen Standardabweichung durch den Wert 1 ersetzt. Maximal sollen 100 Iterationen durchgeführt werden (ITERATIONS ARE 100). Tritt nach 100 Iterationen noch keine Konvergenz ein, wird der Algorithmus beendet. Die Anweisung LOSS bewirkt schließlich, daß bei der Maximum-Likelihood-Schätzung (vgl. Abschnitt 3.6.3) zur Konvergenzberechnung nicht die Quadratsumme der Residuen, sondern die LOSS-Funktion (≙ -(Log-Likelihood)) herangezogen wird. Der Gauß-Newton-Algorithmus des Programms P3R versucht somit denjenigen Set von Parametern zu finden, bei dem die Log-Likelihood-Funktion maximiert beziehungsweise die LOSS-Funktion minimiert wird. Die restlichen Entscheidungen über den Iterationsalgorithmus werden über die Voreinstellung von Werten getroffen (vgl. Dixon u.a. 1983).
Nach dem END-Statement der BMDP-Steuerkarten muß noch das Unterprogramm P3RFUN mit Parametern versorgt werden. Diese werden als Kommentarkarten in der Form /COMMENT '....' nach den BMDP-Steuerkarten eingelesen. Die erste Karte spezifiziert dabei zunächst ein Exponential-Modell, welches im Programm P3RFUN die Nummer 1 trägt (M 1). Mit der T-Karte (Time) werden die BMDP-internen Nummern der Variablen angegeben, in denen die Anfangs- und Endwerte der Verweildauer stehen. Dies sind die Nummer 70 für die Variable X1 (Beginn der Verweildauer) und die Nummer 63 für die Variable DUR (Ende der Verweildauer). Danach muß auch für P3RFUN noch auf der D-Karte (dependent) die BMDP-interne Nummer der abhängigen Variablen DP, also 71, angeführt werden. Die Zensierungsinformation, die in der Variablen ZEN steht, muß auf der C-Karte (censored) mit der BMDP-internen Nummer 64 eingegeben werden. Da nur das Modell mit einer Regressionskonstanten β_0 geschätzt werden soll, wird auf der I-Karte (independent) die Zahl 0, keine Kovariable, vermerkt.

Tabelle 6.2: Ergebnis des Exponential-Modells aus Programmbeispiel 6.3

PARAMETER	ESTIMATE	ASYMPTOTIC STANDARD DEVIATION
KONST	-4.581673	0.015316

Das Ergebnis der Schätzung ist in Tabelle 6.2 abgebildet. Danach ergibt sich bei einem Wert der Log-Likelihood-Funktion (≙ -LOSS) von -14434,35 die Schätzung $\hat{\beta}_0 = -4{,}582$. Die geschätzte durchschnittliche Rate der Männer, den Beruf zu wechseln, beträgt damit $\hat{\lambda} = \exp(-4{,}582) = 0{,}0102$. Vergleicht man diese

Maximum-Likelihood-Schätzung mit der Schätzung bei den graphischen Verfahren aus Abschnitt 6.1 ($\hat{\lambda} = 0{,}0106$), dann sieht man, daß beide Schätzungen sehr eng beieinander liegen.
Gemäß Abschnitt 3.2.2 erhält man über die Beziehung

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

eine Schätzung der mittleren Verweildauer der Männer im Beruf, also

$$\frac{1}{\hat{\lambda}} = \frac{1}{0{,}0102} = 98{,}04 \text{ Monate.}$$

Es dauert somit im Durchschnitt ein wenig mehr als acht Jahre, bis ein Mann seinen Beruf wechselt.
Über die Beziehung (3.2.17)

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

läßt sich darüber hinaus auch der Median der Verweildauer M*, mit

$$S(M^*) = 0{,}5,$$

abschätzen:

$$S(\hat{M}^*) = \exp(-0{,}0102\ \hat{M}^*)$$

$$\hat{M}^* = 67{,}96 \text{ Monate.}$$

Der Median der Verweildauer im Beruf ist kleiner als der Mittelwert der Verweildauer im Beruf, weil die Wartezeiten bei der Exponential-Verteilung eine rechtsschiefe Form haben. Er beträgt genau 69,34 Prozent des Mittelwerts $[0{,}5 = \exp(\frac{1}{E(T)} M^*)]$[3].

Die Wahrscheinlichkeit, daß ein Mann nach einem Zeitraum von acht Jahren noch in demselben Beruf arbeitet, beträgt mit $\hat{S}(96) = \exp(-0{,}0102 \cdot 96) = 0{,}3756$, also 37,56 Prozent, und umgekehrt, daß er diesen Beruf bis dahin bereits verlassen hat, 62,44 Prozent. Die soeben gegebenen inhaltlichen Interpretationen sind natürlich nur insoweit gültig, als man bei den Männern tatsächlich von einem „durchschnittlichen Berufswechselrisiko" sprechen kann. Die Cox-Modelle in Abschnitt 5.2 haben aber bereits den signifikanten Einfluß einer

---

[3] Da bei exponentialverteilten Verweildauern die Anzahl N der während einer bestimmten Zeitspanne eintretenden Ereignisse poissonverteilt mit dem Erwartungswert

$$E(N) = \lambda t$$

ist, prognostizieren wir in einem Jahr durchschnittlich $\hat{\lambda} \cdot v = 0{,}0102 \cdot 12 = 0{,}1224$ Berufswechsel bei den Männern. Oder anders ausgedrückt: In einem Zeitraum von acht Jahren erwarten wir für jeden Mann durchschnittlich einen Berufswechsel.

Reihe von Kovariablen bestätigt, so daß diese auch in das Exponential-Modell aufgenommen werden sollten.

6.2.2 Das Exponential-Modell mit zeitkonstanten Kovariablen

Werden bei der Schätzung zusätzlich Kovariablen berücksichtigt, dann verbessern sich nicht nur die Interpretationsmöglichkeiten, sondern die Modelle werden auch realitätsgerechter. Im folgenden soll deswegen das Berufswechselrisiko der Männer in Abhängigkeit von den bereits bekannten Variablen Bildung (BILDG), Prestige (M59), Anzahl der vorher ausgeübten Berufe (BANZ), Berufserfahrung zu Beginn jedes Jobs (BERF) sowie den Kohorten-Dummies KOHO2 und KOHO3 geschätzt werden (vgl. Anhang 1):

$$\lambda^k(v|x_k) = \exp(x_k'\beta), \qquad k = 1, 2, \ldots, .$$

Die Berechnung der Maximum-Likelihood-Schätzwerte erfolgt wieder mit dem BMDP-Programm P3R und dem darin eingebundenen Unterprogramm P3RFUN von Trond Petersen:

Programmbeispiel 6.4:

```
/INPUT UNIT IS 30.
       CODE IS DATA.
/VARIABLE NAMES ARE (63)DUR,(64)ZEN,(65)BILDG,(66)BERF,
                    (67)BANZ,(68)KOHO2,(69)KOHO3,(70)X1,(71)DP.
          ADD IS 9.
/TRANSFORM USE = (M3 EQ 1).
           DUR = M51 - M50 + 1.
           ZEN = 1.
           IF (M51 EQ M47) THEN ZEN = 0.
           IF (M41 EQ 1 AND M42 EQ 1) THEN BILDG = 9.
           IF (M41 EQ 1 AND (M42 EQ 2 OR M42 EQ 3)) THEN BILDG = 11.
           IF (M41 EQ 2 AND M42 EQ 1) THEN BILDG = 10.
           IF (M41 EQ 2 AND (M42 EQ 2 OR M42 EQ 3)) THEN BILDG = 12.
           IF (M41 EQ 3 AND (M42 EQ 1 OR M42 EQ 2 OR M42 EQ 3))
           THEN BILDG = 13.
           IF (M42 EQ 4) THEN BILDG = 17.
           IF (M42 EQ 5) THEN BILDG = 19.
           KOHO2 = 0.
           KOHO3 = 0.
           IF(M48 GE 468 AND M48 LE 504) THEN KOHO2 = 1.
           IF(M48 GE 588 AND M48 LE 624) THEN KOHO3 = 1.
           BERF = M50 - M43.
           BANZ = M5 - 1.
           DP = 0.
           X1 = 0.
/REGRESS DEPENDENT IS DP.
         PARAMETERS ARE 7.
         PRINT IS 0.
         MEANSQUARE IS 1.0.
         ITERATIONS ARE 100.
         LOSS.
```

```
/PARAMETER INITIAL ARE -3.8,0.04,-0.01,0.15,-0.01,0.1,0.17.
          NAMES ARE KONST,BILDG,M59,BANZ,BERF,KOHO2,KOHO3.
/END.
/COMMENT'
M 1
T 70 63
D 71
C 64
I  6 65 59 67 66 68 69
```

Die Aufbereitung der Variablen im TRANSFORM-Paragraph geschieht genauso wie in den Programmbeispielen 5.1 und 6.3 und ist dort bereits ausführlich beschrieben. Das obige Steuerkartenbeispiel soll nur noch einmal dokumentieren, daß sich die Programmläufe zur Schätzung des Cox-Modells mit geringfügigen Veränderungen auch bei P3R verwenden lassen.
Im Vergleich zu Programmbeispiel 6.3 werden jetzt allerdings 7 β-Koeffizienten (PARAMETERS ARE 7) geschätzt, die Regressionskonstante β_0 und die 6 β-Koeffizienten für die unabhängigen Variablen. Für jeden dieser β-Koeffizienten wird im Paragraph PARAMETER ein Startwert vorgegeben (INITIAL ARE). Diese lehnen sich an die bereits in Kapitel 5 durchgeführten Cox-Schätzungen an.
Auch die Parameter für das Unterprogramm P3RFUN bleiben, bis auf die Karte, auf der die unabhängigen Variablen angegeben werden (I-Karte), im Vergleich zu Programmbeispiel 6.3 unverändert. Dort wird zuerst die Anzahl der Kovariablen eingesetzt, in diesem Beispiel die Zahl 6, und dann folgen die Kovariablen mit ihren BMDP-internen Nummern: 65 (≙ BILDG) 59 (≙ M59, Prestige), 67 (≙ BANZ), 66 (≙ BERF), 68 (≙ KOHO2) und 69 (≙ KOHO3). Das Ergebnis dieses Programms ist in Tabelle 6.3 dargestellt.

Tabelle 6.3: Ergebnis des Exponential-Modells aus Programmbeispiel 6.4

PARAMETER	ESTIMATE	ASYMPTOTIC STANDARD DEVIATION
KONST	-4.337501	0.103741
BILDG	0.012903	0.011682
M59	-0.005213	0.001116
BANZ	0.171426	0.008598
BERF	-0.008856	0.000369
KOHO2	0.179615	0.035365
KOHO3	0.486224	0.041957

Zunächst bekommt man im Beispiel der Tabelle 6.3 einen Wert der Log-Likelihood-Funktion (≙ -LOSS) von -14081,40. Wenn man dieses Modell mit dem Modell ohne Kovariablen aus Abschnitt 6.2.1 vergleicht, dann ergibt sich auf

der Basis des Likelihood-Quotienten-Tests (vgl. Abschnitt 3.7.3) ein χ^2-Wert von

$$Lq = 2(-14081{,}40 - (-14434{,}35)) = 705{,}90,$$

mit sechs Freiheitsgraden. Die eingeführten Kovariablen können also zusätzlich etwas zur Erklärung des Berufswechselrisikos bei Männern beitragen, und die Nullhypothese, keiner der zusätzlich aufgenommenen β-Koeffizienten ist von Null verschieden, muß verworfen werden.
Eine Signifikanzprüfung der einzelnen Regressionsparameter wird durchgeführt, indem man die Koeffizienten $\hat{\beta}_i$ durch ihre geschätzten asymptotischen Standardabweichungen $s(\hat{\beta}_i)$ dividiert. Unter der Hypothese $H_0 : \beta_i = 0$ sind diese Prüfgrößen näherungsweise standardnormalverteilt (vgl. Abschnitt 3.7.3). Geht man wieder von einer Signifikanzwahrscheinlichkeit von 0,05 und beidseitiger Fragestellung aus, dann haben die Kovariablen einen signifikanten Effekt, wenn der Betrag ihrer standardisierten Koeffizienten

$$\left| \frac{\hat{\beta}_i}{s(\hat{\beta}_i)} \right|$$

größer als der Wert 1,96 ist.

Dies ist außer bei der Konstanten β_0 (KONST) bei den Variablen M59 (Prestige), BANZ, BERF, KOHO2 und KOHO3 der Fall. Nur die Variable Bildung (BILDG) hat wieder keinen signifikanten Einfluß auf die Rate des Berufswechsels bei Männern.
Wie beim Cox-Modell (vgl. Abschnitt 5.2), so kann man auch hier wieder die Wirkung einer Kovariablen x_i anschaulich interpretieren, indem man bei Konstanthaltung der jeweils anderen Variablen $\mathbf{x}'\boldsymbol{\beta}$ zeigt, um wieviel Prozent sich die Rate bei Erhöhung der Kovariablen x_i um einen bestimmten Wert Δx_i verändert. So ergibt sich beispielsweise bei einer Erhöhung der Anzahl der vorher ausgeübten Berufe (BANZ) um eine Einheit eine Erhöhung der Rate um 18,70 Prozent [$(\exp(0{,}171426) - 1) \cdot 100\% = 18{,}70\%$]. Eine Erhöhung des Prestiges (M59) um 20 Einheiten führt dagegen zu einer Verminderung der Neigung zum Berufswechsel um 9,9 Prozent [$(\exp(-0{,}00521)^{20} - 1) \cdot 100\% = -9{,}9\%$]. Die gleichzeitige Veränderung von BANZ um eine Einheit und von Prestige (M59) um 20 Einheiten, was einem beruflichen Aufstieg entspräche, erhöht die Rate allerdings nur um 6,95 Prozent [$(\exp(0{,}171426)^1 \cdot \exp(-0{,}00521)^{20} - 1) \cdot 100\% = 6{,}95\%$] und nicht um 8,80 Prozent [$18{,}70\% - 9{,}9\% = 8{,}80\%$].
Über die Beziehung

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

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

$$\delta_{\Delta x_i} = \frac{\frac{1}{\exp(x'\beta + \beta_i (x_i + \Delta x_i))} - \frac{1}{\exp(x'\beta + \beta_i x_i)}}{\frac{1}{\exp(x'\beta + \beta_i x_i)}} \cdot 100\%$$

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

Danach folgt bei Erhöhung der Anzahl der vorher bereits ausgeübten Berufe (BANZ) um eine Einheit eine Verminderung der durchschnittlichen Verweildauer im Beruf um 15,75 Prozent [(1/exp(0,171426) - 1)·100% = -15,75%]. Eine Erhöhung des Prestiges (M59) um 20 Einheiten führt dagegen zu einer Erhöhung der Verweildauer im Beruf um 10,98 Prozent [(1/exp(-0,00521)^{20} - 1) · 100% = 10,98%]. Die gleichzeitige Veränderung von BANZ um eine Einheit und von Prestige (M59) um 20 Einheiten, was wieder einem beruflichen Aufstieg entsprechen würde, vermindert die durchschnittliche Verweildauer aber um 6,5 Prozent [(1/(exp(0,171426)^1 exp(-0,00521)^{20}) - 1) · 100% = -6,5%] und nicht nur um 4,77 Prozent [10,98% - 15,75% = -4,77%].

Für beliebige Subgruppen lassen sich natürlich auch Prognosen über die durchschnittliche Verweildauer, den Median der Verweildauer, die in einer bestimmten Zeitspanne durchschnittlich eintretenden Ereignisse und die Wahrscheinlichkeit geben, bis zu einem bestimmten Zeitpunkt noch im selben Zustand zu sein.

Betrachtet man beispielsweise einen Mann der Kohorte von 1939–41 (KOHO2 = 1 und KOHO3 = 0), der in einem mit 50 Prestige-Punkten (M59 = 50) bewerteten Beruf arbeitet und der vorher bereits 10 Berufe (BANZ = 10) ausgeübt sowie dabei eine Berufserfahrung von 100 Monaten (BERF = 100) gesammelt hat, so kommt man zu folgender Prognose-Gleichung für die Rate des Berufswechsels[4]:

$$\hat{\lambda} = \exp(-4{,}338 - 0{,}005 \cdot 50 + 0{,}171 \cdot 10 - 0{,}009 \cdot 100 + 0{,}180 \cdot 1) = 0{,}0274.$$

Daraus folgt, daß man bei dieser Person eine mittlere Verweildauer von 36,5 Monaten [$\frac{1}{\hat{\lambda}} = 1/0{,}0274 = 36{,}496$] erwartet, die weit unter der des Durchschnitts von 98,04 Monaten liegt. Darüber hinaus kann man einen Median der Verweildauer im Beruf von $\hat{M}^* = 0{,}6934 \cdot 36{,}5 = 25{,}31$ Monaten prognostizieren und in einem Jahr durchschnittlich $\hat{\lambda}v = 0{,}0274 \cdot 12 = 0{,}3288$ Berufswechsel erwarten. Die Wahrscheinlichkeit schließlich, daß diese Person noch nach acht Jahren in ihrem Beruf arbeitet, beträgt 7,2 Prozent [$\hat{S}(96) = \exp(-0{,}0274 \cdot 96) = 0{,}072$], während sie sich beim Durchschnitt aller Männer auf 37,56 Prozent beläuft.

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

Durch entsprechende Prognosen für weitere Subgruppen kann man insgesamt ein sehr differenziertes Bild von dem Berufswechselverhalten der Männer und der Bedeutung unterschiedlicher Einflußfaktoren geben.

6.2.3 Die Überprüfung der Residuen im Exponential-Modell

Die inhaltlichen Interpretationen der Exponential-Modelle in den Abschnitten 6.2.1 und 6.2.2 sind natürlich nur unter der Voraussetzung gültig, daß tatsächlich ein konstantes Berufswechselrisiko über die Verweildauer hinweg vorliegt. Daß dieser Unterstellung eine gewisse Berechtigung zukommt, hat bereits die graphische Überprüfung der Verteilungsannahmen in Abschnitt 6.2.1 gezeigt. Dort ist allerdings auch darauf aufmerksam gemacht worden, daß es notwendig ist, zusätzliche Kriterien zur Modellevaluation heranzuziehen. Eine solche weitere Überprüfungsmöglichkeit besteht nun darin, ähnlich wie bei der herkömmlichen Regression, die Residuen der Modellschätzungen zu beurteilen.
Wie in Abschnitt 3.7.1 schon ausführlich diskutiert, kann man allgemein bei Gültigkeit der jeweiligen Verteilungsannahme die kumulativen Hazardraten $\Lambda(t|x_i)$ als Residuen r_i betrachten, die einer Standard-Exponential-Verteilung (mit $\lambda = 1$) folgen:

$$r_i = \Lambda(t|x_i) = -\ln S(t|x_i).$$

Die Survivorfunktion dieser exponentialverteilten Zufallsvariablen R lautet dann

$$S(r|x) = \exp(-r)$$

und ergibt bei Erfüllung der jeweiligen Verteilungsannahme nach logarithmischer Transformation (vgl. Abschnitt 4.4.2.1) eine Gerade mit

$$y = \ln S(r|x),\ a = 0,\ b = -1,\ x = r.$$

Für den vorliegenden Fall bei mehreren Episoden erhält man eine Schätzung der Residuen r_{ik} wie folgt:

$$\hat{r}_{ik} = \hat{\Lambda}\,(v_{ik}|x_{ik}) = \exp(x'_{ik}\hat{\beta}) \cdot v_{ik}.$$

Aus diesen so berechneten Residuen läßt sich bei zensierten Daten die Survivorfunktion mit Hilfe der Produkt-Limit-Methode schätzen und nach Logarithmierung gegen r auftragen. Bei Erfüllung der Annahme konstanter Risiken über die Verweildauer hinweg müßte sich dann eine Gerade mit der Steigung –1 ergeben.
Die folgenden zwei Programmläufe mit dem Unterprogramm P1L von BMDP zeigen die Realisierung dieses Residuen-Tests für das Exponential-Modell ohne Kovariablen und das Exponential-Modell mit Kovariablen:

Programmbeispiel 6.5:

```
/INPUT UNIT IS 30.
       CODE IS DATA.
/VARIABLE NAMES ARE (63)DUR,(64)ZEN,(65)BILDG,(66)BERF,
                    (67)BANZ,(68)KOHO2,(69)KOHO3,(70)RESID1.
          ADD IS 8.
/TRANSFORM USE = (M3 EQ 1).
           DUR = M51 - M50 + 1.
           ZEN = 1.
           IF (M51 EQ M47) THEN ZEN = 0.
           IF (M41 EQ 1 AND M42 EQ 1) THEN BILDG = 9.
           IF (M41 EQ 1 AND (M42 EQ 2 OR M42 EQ 3)) THEN BILDG = 11.
           IF (M41 EQ 2 AND M42 EQ 1) THEN BILDG = 10.
           IF (M41 EQ 2 AND (M42 EQ 2 OR M42 EQ 3)) THEN BILDG = 12.
           IF (M41 EQ 3 AND (M42 EQ 1 OR M42 EQ 2 OR M42 EQ 3))
           THEN BILDG = 13.
           IF (M42 EQ 4) THEN BILDG = 17.
           IF (M42 EQ 5) THEN BILDG = 19.
           KOHO2 = 0.
           KOHO3 = 0.
           IF(M48 GE 468 AND M48 LE 504) THEN KOHO2 = 1.
           IF(M48 GE 588 AND M48 LE 624) THEN KOHO3 = 1.
           BERF = M50 - M43.
           BANZ = M5 - 1.
           RESID1 = DUR * EXP(-4.581673).
/FORM TIME IS RESID1.
      STATUS IS ZEN.
      RESPONSE IS 1.
/ESTIMATE METHOD IS PROD.
          PLOTS ARE LOG.
/PRINT CASES ARE 0.
/END
```

Programmbeispiel 6.6:

```
/INPUT UNIT IS 30.
       CODE IS DATA.
/VARIABLE NAMES ARE (63)DUR,(64)ZEN,(65)BILDG,(66)BERF,
                    (67)BANZ,(68)KOHO2,(69)KOHO3,(70)RESID2.
          ADD IS 8.
/TRANSFORM USE = (M3 EQ 1).
           DUR = M51 - M50 + 1.
           ZEN = 1.
           IF (M51 EQ M47) THEN ZEN = 0.
           IF (M41 EQ 1 AND M42 EQ 1) THEN BILDG = 9.
           IF (M41 EQ 1 AND (M42 EQ 2 OR M42 EQ 3)) THEN BILDG = 11.
           IF (M41 EQ 2 AND M42 EQ 1) THEN BILDG = 10.
           IF (M41 EQ 2 AND (M42 EQ 2 OR M42 EQ 3)) THEN BILDG = 12.
           IF (M41 EQ 3 AND (M42 EQ 1 OR M42 EQ 2 OR M42 EQ 3))
           THEN BILDG = 13.
           IF (M42 EQ 4) THEN BILDG = 17.
           IF (M42 EQ 5) THEN BILDG = 19.
           KOHO2 = 0.
           KOHO3 = 0.
```

```
          IF(M48 GE 468 AND M48 LE 504) THEN KOHO2 = 1.
          IF(M48 GE 588 AND M48 LE 624) THEN KOHO3 = 1.
          BERF = M50 - M43.
          BANZ = M5 - 1.
          RESID2 = DUR * EXP(-4.337501+0.012903*BILDG-0.00521*M59
                    +0.171426*BANZ-0.008856*BERF+0.179615*KOHO2
                    +0.486224*KOHO3).
/FORM TIME IS RESID2.
      STATUS IS ZEN.
      RESPONSE IS 1.
/ESTIMATE METHOD IS PROD.
          PLOTS ARE LOG.
/PRINT CASES ARE 0.
/END
```

Im TRANSFORM-Paragraph der obigen Programmbeispiele werden zuerst aufgrund der β-Schätzungen für die Exponential-Modelle aus den Tabellen 6.2 und 6.3 die Residuen (RESID1 bzw. RESID2) geschätzt. Im FORM-Paragraph werden diese neben der Zensierungsvariablen (STATUS IS ZEN) als Verweildauer eingegeben (TIME IS RESID1 bzw. TIME IS RESID2). Zur Schätzung der Survivorfunktion wird jeweils die Produkt-Limit-Methode herangezogen (METHOD IS PROD), und geplottet werden sollen schließlich die logarithmierten Survivorfunktionen (PLOTS ARE LOG).
Die Ergebnisse dieser Läufe sind in den Abbildungen 6.5 (für das Modell ohne Kovariablen aus Abschnitt 6.2.1) und 6.6 (für das Modell mit Kovariablen aus Abschnitt 6.2.2) dargestellt.
Aus Abbildung 6.5 ist zunächst ersichtlich, daß der Verlauf der transformierten Survivorfunktion $\ln\hat{S}(r)$ deutlich von einer Geraden abweicht. Insbesondere bei kleinen Residuen ist die Kurve stark nach unten gekrümmt, was gegen die Annahme exponentialverteilter Verweildauern mit der Rate $\exp(\beta_0)$ spricht.
Auch beim Exponential-Modell mit Kovariablen ist die Anpassung an eine Gerade mit der Steigung -1 nicht viel besser (Abbildung 6.6). Der Verlauf der transformierten Survivorfunktion ist ebenfalls deutlich nach unten gekrümmt und spricht deswegen auch gegen die Annahme einer konstanten Rate über die Verweildauer hinweg.
Obwohl man somit aufgrund dieser Befunde geneigt ist, die Annahme exponentialverteilter Verweildauern im Beruf zurückzuweisen, sollte man berücksichtigen, daß die oben definierten Residuen als Zufallsvariablen weder unabhängig sind noch eine identische Verteilung besitzen (vgl. Abschnitt 3.7.1), so daß das eben beschriebene Verfahren nur als Näherung an einen strengen Test aufgefaßt werden kann. Weitere Hinweise darauf, ob die Annahme konstanter Risiken beim Berufswechsel der Männer tatsächlich verworfen werden muß, erhält man bei der Schätzung des Weibull-Modells, welches die Exponential-Verteilung als Spezialfall enthält. Bevor allerdings weitere Verfahren mit spezifischen Verweildauerabhängigkeiten ausführlicher dargestellt werden, wird im folgenden Abschnitt aufgezeigt, wie man bei den parametrischen Modellen zeitveränderliche Kovariablen aufnimmt.

Abbildung 6.5: Residuen-Plot für das Exponential-Modell ohne Kovariablen (nur mit der Regressions-Konstanten β_0)

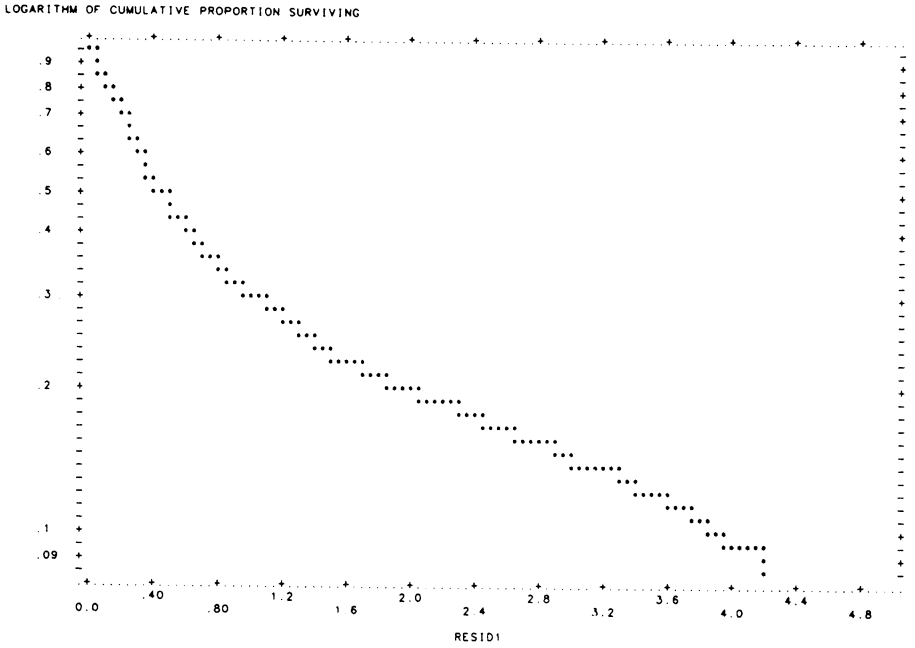
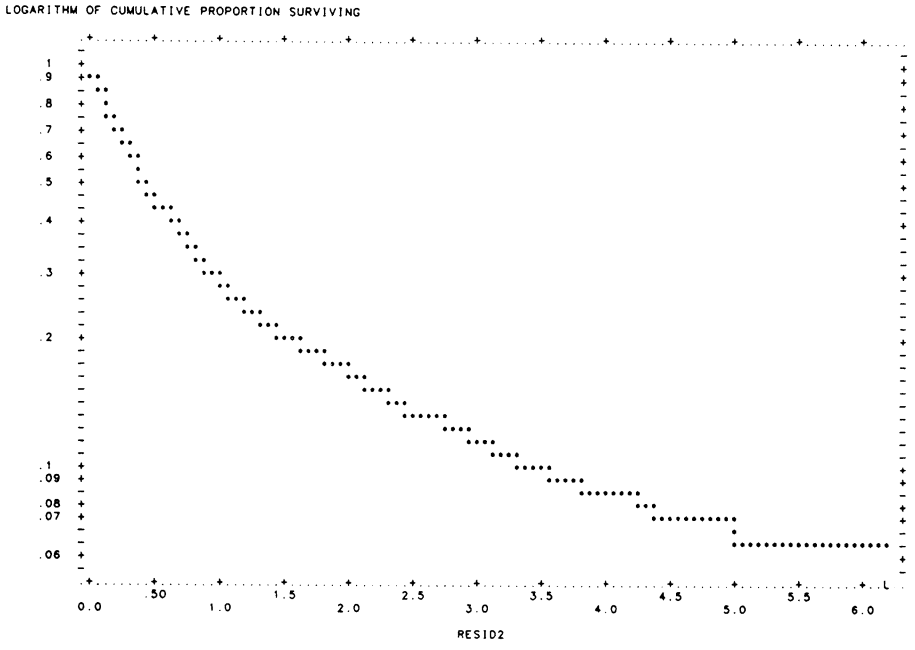


Abbildung 6.6: Residuen-Plot für das Exponential-Modell mit Kovariablen



6.3 Die Aufnahme zeitveränderlicher unabhängiger Variablen bei parametrischen Modellen

Die große Relevanz zeitveränderlicher unabhängiger Variablen für die Anwendung der Ereignisanalyse im Bereich der Wirtschafts- und Sozialwissenschaften ist bereits im Abschnitt 5.4 ausführlich dargelegt worden. So kann in vielen Fällen mit zeitveränderlichen unabhängigen Variablen nicht nur der Einfluß von Kovariablen auf die Hazardrate realitätsgerechter formuliert werden, sondern es lassen sich damit auch zwei oder mehrere parallele Prozesse miteinander verbinden.
Im Vergleich zur Maximum-Partial-Likelihood-Schätzung, bei der die jeweils beim Ereigniszeitpunkt vorliegenden Kovariablen-Vektoren der noch zur Risikomenge gehörenden Individuen berücksichtigt werden und die sich beispielsweise mit dem Programm P2L von BMDP über die Variable TIME bequem aktualisieren lassen (vgl. die Abschnitte 5.4.1 und 5.4.2), besteht bei den parametrischen Verfahren diese Möglichkeit der Einbeziehung zeitveränderlicher unabhängiger Variablen nicht. Man muß hier auf andere Lösungen zurückgreifen.

6.3.1 Die Methode des Episodensplittings bei diskreten zeitveränderlichen unabhängigen Variablen

Relativ einfach läßt sich bei der Maximum-Likelihood-Schätzung noch die Einbeziehung diskreter zeitveränderlicher unabhängiger Variablen bewerkstelligen. Diese folgen in der Zeit der Gestalt einer Treppenfunktion und sind stückweise konstant (vgl. Abbildung 5.2(b)). Bezeichnen $t_0 < t_1 < \ldots < t_s$ die Änderungszeitpunkte des Kovariablen-Vektors im Verweildauer-Intervall $[0,t)$ und sei $t_{s+1} = t$, dann kann nach den Ausführungen in Abschnitt 3.8 die kumulative Hazardrate in eine Summe von Integralen zerlegt werden, und die Wahrscheinlichkeit, daß bis zum Zeitpunkt t kein Ereignis auftritt, ergibt sich aus dem Produkt der Survivorfunktionen der Subepisoden, in denen der Kovariablen-Vektor unverändert bleibt:

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

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

(1) Die Ausprägungen der Kovariablen zu Beginn der Subepisode;

(2) die Verweildauer zu Beginn und am Ende der Subepisode (die Verweildauer als solche ist nur beim Exponential-Modell ausreichend);
(3) eine Zensierungsinformation, ob die Subepisode mit einem Ereignis (ZEN = 1) endete oder nicht (ZEN = 0).

Um die konkrete Anwendung der Methode des Episodensplittings bei diskreten zeitveränderlichen Kovariablen zu demonstrieren, greifen wir wieder auf das Heiratsbeispiel aus Abschnitt 5.4.1 zurück. Diesmal soll mit Hilfe eines Exponential-Modells untersucht werden, ob sich bei Männern das Ereignis „Heirat" im Familiensystem stabilisierend auf den Erwerbsprozeß im Beschäftigungssystem auswirkt oder nicht. Die Schätzung soll dabei mit dem Programm RATE vorgenommen werden. Da im Programm RATE die Datentransformationen und -recodierungen nur umständlich vorgenommen werden können[5], empfiehlt es sich, die Daten zuvor, beispielsweise mit dem Programmpaket SPSS, aufzubereiten und als Rohdatenfile auszugeben:

Programmbeispiel 6.7:

```
GET FILE
SELECT IF          (M3 EQ 1)
COMPUTE            DUR = M51 - M50 + 1
COMPUTE            ZEN = 1
IF                 (M51 EQ M47) ZEN = 0
IF                 (M41 EQ 1 AND M42 EQ 1) BILDG = 9
IF                 (M41 EQ 1 AND (M42 EQ 2 OR M42 EQ 3)) BILDG = 11
IF                 (M41 EQ 2 AND M42 EQ 1) BILDG = 10
IF                 (M41 EQ 2 AND (M42 EQ 2 OR M42 EQ 3)) BILDG = 12
IF                 (M41 EQ 3 AND (M42 EQ 1 OR M42 EQ 2 OR M42 EQ 3))
                   BILDG = 13
IF                 (M42 EQ 4) BILDG = 17
IF                 (M42 EQ 5) BILDG = 19
COMPUTE            KOHO2 = 0
COMPUTE            KOHO3 = 0
IF                 (M48 GE 468 AND M48 LE 504) KOHO2 = 1
IF                 (M48 GE 588 AND M48 LE 624) KOHO3 = 1
COMPUTE            BERF = M50 - M43
COMPUTE            BANZ = M5 - 1
COMPUTE            TANF = M50
COMPUTE            TEND = M51 + 1
COMPUTE            PRES = M59
COMPUTE            JOBN = M61
COMPUTE            JOBN1 = M62
COMPUTE            THEIRAT = M49
WRITE CASES        (13F5.0) TANF,TEND,ZEN,BILDG,PRES,BANZ,BERF,
                   KOHO2,KOHO3,DUR,JOBN,JOBN1,THEIRAT
FINISH
```

[5] Dies trifft bei der neuen RATE-Version nicht mehr zu.

Das Ergebnis des obigen SPSS-Programms ist ein ereignisorientierter Datenfile (Tabelle 6.4) in den die Variablen TANF (Anfangszeitpunkt der Episode, gemessen in Monaten vom Beginn des Jahrhunderts an), TEND (Endzeitpunkt der Episode, gemessen in Monaten vom Beginn des Jahrhunderts an), ZEN (Zensierungsvariable), BILDG (Bildungsniveau, gemessen in Anzahl von Schuljahren), PRES (Prestigescore von Wegener (1985)), BANZ (Anzahl der vorher ausgeübten Berufe), BERF (Berufserfahrung beim Eintritt in die Episode, gemessen in Monaten), KOH2 (Dummy-Variable für die Kohorte 1939–41), KOH3 (Dummy-Variable für die Kohorte 1949–51), DUR (Verweildauer in Monaten), JOBN (Berufsgruppenschlüssel des Berufs in der Episode), JOBN1 (Berufsgruppenschlüssel des nächstfolgenden Berufs) und THEIRAT (Zeitpunkt der Heirat, gemessen in Monaten vom Beginn des Jahrhunderts an) eingehen (vgl. Anhang 1).

Tabelle 6.4: Beispiel für einen ereignisorientierten Datensatz

TANF	TEND	ZEN	BILDG	PRES	BANZ	BERF	KOH2	KOH3	DUR	JOBN	JOBN1	THEIRAT
555	983	0	11	66	0	0	0	0	428	11	0	679
583	651	1	11	50	0	0	0	0	68	3	3	701
651	788	1	11	50	1	68	0	0	137	3	3	701
788	983	0	11	50	2	205	0	0	195	3	0	701
691	717	1	9	39	0	0	1	0	26	6	3	781
728	754	1	11	56	1	37	1	0	26	3	3	781
771	847	1	11	56	2	80	1	0	76	3	0	781
.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.

Der erste Satz der Datei in Tabelle 6.4 läßt sich beispielsweise wie folgt interpretieren: Es handelt sich um die erste Berufsepisode (BANZ = 0, BERF = 0) eines Mannes aus der Kohorte von 1929–31 (KOH2 = 0, KOH3 = 0), der nach einem Hauptschulabschluß mit Berufsausbildung (BILDG = 11) in einem qualifizierten kaufmännischen oder Verwaltungsberuf (JOBN = 11) tätig war (wobei sich dieser Beruf mit einem Prestige-Score von 66 Punkten, also PRES = 66, bewerten läßt). Der Eintrittszeitpunkt in diesen Beruf war der März 1946 (TANF = 555 Monate), und der Mann hat bis zum Zeitpunkt des Interviews (ZEN = 0, JOBN1 = 0), im November 1981 (TEND = 983 Monate), ununterbrochen dort gearbeitet (DUR = 428). Schließlich ist bekannt, daß dieser Mann im Juli 1956 geheiratet hat (THEIRAT = 679).
Will man nun, wie im vorliegenden Beispiel, den Familienstand (verheiratet -nicht verheiratet) als zeitveränderliche unabhängige Variable bei der Schätzung des Berufswechselrisikos von Männern in einem Exponential-Modell berücksichtigen, dann bricht man die Berufsepisoden der Datei aus Tabelle 6.4 nach dem Zeitpunkt der Heirat auf. Die neue ereignisorientierte Datei (vgl. Tabelle 6.5) wird dabei so aufbereitet, daß für jedes Zeitintervall innerhalb einer gegebenen Berufsepisode, in der die Kovariable Familienstand unverändert

bleibt, ein eigener Datensatz erzeugt wird. In diesen gehen dann, neben den Ausprägungen der Kovariablen zu Beginn jeder Subepisode (einschließlich der neuen Dummy-Variablen HEIRAT), der aktualisierte Anfangs- (TANF) und Endzeitpunkt (TEND), die sich daraus ergebende Verweildauer (DUR)[6] sowie die Zensierungsinformation ein, ob die Subepisode mit dem Ereignis „Berufswechsel" (ZEN = 1) beendet wurde oder nicht (ZEN = 0).
Leider kann man das Episodensplitting nicht mit den satzorientierten Programmpaketen wie SPSS oder BMDP durchführen, so daß man generell zur Lösung dieses Problems auf ein Datenbanksystem (wie etwa SIR oder SAS), auf Programmsysteme wie GLIM oder auf eigene Anwenderprogrammierung angewiesen ist. Für das vorliegende Beispiel wurde die neue ereignisorientierte Datei (vgl. Tabelle 6.5) mit Hilfe eines FORTRAN-Programms[7] erzeugt.

Tabelle 6.5: Beispiel für einen ereignisorientierten Datensatz, dessen Episoden nach dem Zeitpunkt der Heirat in Subepisoden aufgesplittet wurden.

TANF	TEND	ZEN	BILDG	PRES	BANZ	BERF	KOH2	KOH3	DUR	JOBN	JOBN1	THEI	HEIRAT
555	679	0	11	66	0	0	0	0	124	11	0	679	0
679	983	0	11	66	0	0	0	0	304	11	0	679	1
583	651	1	11	50	0	0	0	0	68	3	3	701	0
651	701	0	11	50	1	68	0	0	50	3	3	701	0
701	788	1	11	50	1	68	0	0	87	3	3	701	1
788	983	0	11	50	2	205	0	0	195	3	0	701	1
691	717	1	9	39	0	0	1	0	26	6	3	781	0
728	754	1	11	56	1	37	1	0	26	3	3	781	0
771	781	0	11	56	2	80	1	0	10	3	0	781	0
781	847	1	11	56	2	80	1	0	66	3	0	781	1
.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.

In der neuen ereignisorientierten Datei (Tabelle 6.5) liegt nun beispielsweise die erste Episode aus der Datei in Tabelle 6.4 aufgesplittet in zwei Subepisoden vor. Die erste Subepisode beginnt mit dem Eintritt in das Beschäftigungssystem im März 1946 (TANF = 555) und dauert bis zum Zeitpunkt der Heirat im Juli 1956 (TEND = 679, THEI = 679, HEIRAT = 0 und ZEN = 0). Die zweite Subepisode beginnt mit dem Heiratszeitpunkt (TANF = 679, THEI = 679, HEIRAT = 1)

---

[6] Bei Exponential-Modellen mit zeitkonstanter Rate genügt die Angabe der Verweildauer DUR, während für das Weibull-, das Gompertz- und das log-logistische Modell sowohl der Wert der Verweildauer zu Beginn und der Wert der Verweildauer am Ende der Subepisoden benötigt werden. Im Programm P3RFUN kann man beispielsweise den Beginn der Verweildauer über die Variable X1 und das Ende der Verweildauer über die Variable DUR (vgl. Programmbeispiel 6.4) zur Schätzung eingeben. Auch im Programm RATE kann man die Anfangs- und Endzeitpunkte der Subepisoden auf der T-AND-S-Karte spezifizieren.

[7] Das Quellenprogramm ist im Anhang 3 zu finden.

und endet im November 1981 (TEND = 983), dem Zeitpunkt des Interviews (ZEN = 0). Entsprechend sind alle anderen Berufsepisoden behandelt worden. An dieser Stelle sollte man darauf hinweisen, daß durch das eben beschriebene Episodensplitting sich an der Verweildauer im Zustand selbst nichts ändert, obgleich die Zahl der Datensätze von 3516 in der Datei von Tabelle 6.4 auf 4268 in der Datei von Tabelle 6.5 ansteigt.
Der neue ereignisorientierte Datensatz mit den nach dem Heiratszeitpunkt aufgebrochenen Episoden (vgl. Tabelle 6.5), kann nun wie im Falle zeitkonstanter Kovariablen behandelt und in das Programm RATE zur Schätzung eines Exponential-Modells mit zeitveränderlicher Heiratsvariablen

$$\lambda^k(v|x_k(v)) = \exp(x_k'(v)\beta), \qquad k = 1, 2, \ldots$$

eingelesen werden:

Programmbeispiel 6.8:

```
RUN NAME        EXPONENTIAL-MODELL
N OF CASES      UNKNOWN
VARIABLES       14
TANF            1
TEND            2
IZEN            3
BILDG           4
PRES            5
BANZ            6
IBERF           7
KOHO2           8
KOHO3           9
IDUR            10
JOBN            11
JOBN1           12
THEIRAT         13
HEIRAT          14
READ DATA
(14F5.0)
T AND S         10 3
MODEL           (1) A=1
VECTOR          (1) 4 5 6 7 8 9 14
SOLVE
FINISH
```

Im obigen RATE-Programmbeispiel (vgl. Anhang 1) wird nach der Angabe der Variablennamen, die der Reihe nach durchnumeriert werden, und der Angabe des Formats nach der READ DATA-Karte auf der T-AND-S-Karte die RATE-interne Nummer der Verweildauervariablen IDUR (≙ 10) und der Zensierungsvariablen IZEN (≙ 3) eingesetzt. Beim Exponential-Modell handelt es sich um das RATE-Modell mit der Nummer (1), in dem der Buchstabe A log-linear mit dem Kovariablen-Vektor x modelliert wird (A = 1) : A = $\exp(x'\beta)$. In den

Kovariablen-Vektor **x** werden die auf der VECTOR-Karte mit ihren RATE-internen Nummern spezifizierten Variablen aufgenommen: 4 ≙ Bildung (BILDG), 5 ≙ Prestige (PRES), 6 ≙ Anzahl der vorher ausgeübten Berufe (BANZ), 7 ≙ Berufserfahrung zu Beginn der ursprünglichen Episode (IBERF), 8 ≙ Kohorten-Dummy für die 1939–41 Geborenen (KOHO2), 9 ≙ Kohorten-Dummy für die 1949–51 Geborenen (KOHO3) und 14 ≙ Familienstand (HEIRAT). Mit der SOLVE-Karte wird schließlich die Berechnung einer Maximum-Likelihood-Schätzung der β-Koeffizienten (einschließlich der Regressionskonstanten β_0) veranlaßt. Das Ergebnis dieses Laufs ist in Tabelle 6.6 dargestellt.

Tabelle 6.6: Ergebnis des Exponential-Modells aus Programmbeispiel 6.8

DESTINATION	UNWEIGHTED FREQUENCY	WEIGHTED FREQUENCY	PROPORTION OF ALL	PROPORTION OF ORIGIN	ESTIMATED RATE	LOG OF RATE	MAX(LOG OF L)
1	2586	2586.0	0.60449	0.60449	1.02370-02	-4.5817D+00	-1.443430+04
NO CHANGE	1692	1692.0	0.39551	0.39551			

DESTINATION	UNWEIGHTED FREQUENCY	WEIGHTED FREQUENCY	MAX(LOG OF L) NULL HYPOTHESIS	MAX(LOG OF L) ALTERNATIVE HYPOTHESIS	PSEUDO R-SQUARED	CHI-SQUARED	OF	PROBABILITY LEVEL
1	2586	2586.0	-1.4434320+04	-1.3949390+04	0.0336	969.87	7	0.000+00

DESTINATION 1 LETTER A LOG-LINEAR TIME-INDEPENDENT VECTOR

INTERNAL NUMBER	VARIABLE NUMBER	VARIABLE NAME	VECTOR 1 PARAMETER	PARAMETER STANDARD ERROR	PARAMETER F RATIO	ANTILOG OF THE PARAMETER	ANTILOG STANDARD ERROR	ANTILOG F RATIO
1		(CONSTANT)	-4.2830+00	1.2450-01	1183.109	1.3800-02		
2	4	BILDG	2.5240-02	1.4410-02	3.069	1.0260+00	1.4780-02	2.993
3	5	PRES	-4.0880-03	1.4080-03	8.426	9.9590-01	1.4030-03	8.461
4	6	BANZ	1.7310-01	1.2110-02	204.442	1.1890+00	1.4400-02	172.374
5	7	IBERF	-6.5930-03	4.7660-04	191.393	9.9340-01	4.7340-04	192.659
6	8	KOHO2	1.5580-01	4.6630-02	11.163	1.1690+00	5.4490-02	9.572
7	9	KOHO3	4.1500-01	5.2310-02	62.937	1.5140+00	7.9210-02	42.161
8	14	HEIRAT	-7.1410-01	4.4590-02	256.438	4.8970-01	2.1830-02	546.344

Im ersten Abschnitt des RATE-Outputs in Tabelle 6.6 ist zunächst ersichtlich, daß 2586 der eingelesenen Subepisoden aus der Datei in Tabelle 6.5 mit einem Berufswechsel beendet und 1692 Subepisoden zensiert wurden. Das entspricht einem Anteil an Zensierungen von 39,55 Prozent.
Für das RATE-Modell ohne Kovariablen wird die geschätzte Rate $\hat{\lambda}$ (ESTIMATED RATE = 0,010237) und deren logarithmierter Wert, also die Schätzung für β_0 (LOG OF RATE = -4,5817), ausgedruckt. Diese Schätzung stimmt ebenso wie der Log-Likelihood-Wert von -14434,3 mit dem Ergebnis aus Tabelle 6.2 überein, und man sieht, daß sich durch das Episodensplitting die Schätzung dieses Modells nicht verändert hat.
Im zweiten Abschnitt des obigen RATE-Outputs wird dann für das Modell mit Kovariablen der Log-Likelihood-Wert in der Spalte MAX (LOG OF L) ALTERNATIVE HYPOTHESIS ausgegeben. Er beträgt -13949,39. Für den Vergleich des Modells ohne Kovariablen mit dem Modell mit Kovariablen, auf der Basis eines Likelihood-Quotienten-Tests, ergibt sich bei sieben Freiheitsgraden ein χ^2-Wert von 969,87 [Lq = 2(-13949,39 - (-14434,32)) = 969,87]. Mindestens eine der einbezogenen Kovariablen leistet demnach einen signifikanten Erklärungsbeitrag.

RATE gibt zusätzlich den Wert eines PSEUDO-R^2 aus, der sich wie folgt errechnet:

$$\text{PSEUDO R-SQUARED} = 1 - \frac{\ln L(\text{aktuelles Modell})}{\ln L(\text{Modell ohne Kovariablen})}$$

Im obigen Beispiel ist der Wert von PSEUDO R-SQUARED = 1 - (-13943,39)/(-14434,32) = 0,0336. Leider läßt sich dieses PSEUDO-R^2 nicht wie bei der normalen Regression als Anteil erklärter Varianz interpretieren, sondern drückt nur die relative Verminderung der Log-Likelihood-Funktion des gegenwärtigen Modells gegenüber dem Modell ohne Kovariablen aus.
Im dritten Teil des RATE-Outputs werden schließlich für die einzelnen Kovariablen die Schätzungen der β-Koeffizienten $\hat{\beta}_i$, deren asymptotische Standardabweichungen $s(\hat{\beta}_i)$ und zur Signifikanzprüfung die sich ergebenden F-Werte (mit $d_1 = 1$ und $d_2 = 1$ Freiheitsgraden) ausgegeben. Da die Wurzel einer F-verteilten Zufallsvariablen (mit $d_1 = 1$ und $d_2 = 1$) eine t-verteilte Zufallsvariable (mit einem Freiheitsgrad) ergibt und diese wiederum bei großem Stichprobenumfang annähernd normalverteilt ist, sind im obigen Beispiel die $\hat{\beta}$-Koeffizienten bei einer Irrtumswahrscheinlichkeit von 0,05 dann signifikant, wenn der Wert der Prüfgröße größer als der Wert $1{,}96^2 = 3{,}84$ ist. Man sieht, daß alle Kovariablen bis auf die Variable Bildung (BILDG) signifikant von Null verschieden sind.
Da der Einfluß dieser Kovariablen auf das Berufswechselrisiko von Männern bereits mit einem Cox-Modell geschätzt wurde (vgl. Tabelle 5.3), können die Schätzungen des obigen Exponential-Modells damit verglichen werden.
Beide Modelle kommen zunächst bei den einzelnen $\hat{\beta}$-Koeffizienten zu denselben Vorzeichen und zu denselben Signifikanzentscheidungen. Auch die absolute Größe der $\hat{\beta}$-Koeffizienten stimmt in etwa überein. Größere Unterschiede gibt es lediglich bei der Variablen Bildung (BILDG), der Dummy-Variablen KOHO3 und der Variablen HEIRAT, deren $\hat{\beta}$-Koeffizienten im obigen Exponential-Modell ungefähr doppelt so groß sind wie im Cox-Modell (Tabelle 5.3). Für die zeitveränderliche unabhängige Variable HEIRAT heißt dies beispielsweise, daß sich nach dem Exponential-Modell die Mobilitätsrate bei den Ehemännern um 51,03 Prozent [$(0{,}4897 - 1) \cdot 100\% = -51{,}03\%$] vermindern würde, während sich nach dem Cox-Modell nur eine Verminderung um 27,44 Prozent ergeben würde. Der Einfluß der Variablen HEIRAT auf die Neigung zum Berufswechsel ist im Exponential-Modell also deutlich höher. Da beide Modelle von proportionalen Risiken ausgehen und das Cox-Modell das umfassendere Modell ist (die Baseline-Hazardrate des Cox-Modells wäre bei Gültigkeit des Exponential-Modells konstant), ist zu vermuten, daß tatsächlich keine konstante Rate vorliegt, sondern daß das Berufswechselrisiko mit dem obigen Exponential-Modell fehlspezifiziert ist. In einem nächsten Schritt soll deswegen zunächst eine stetige zeitveränderliche unabhängige Variable in das Exponential-Modell aufgenommen werden.

6.3.2 Die Methode des Episodensplittings bei stetigen zeitveränderlichen unabhängigen Variablen

Während sich diskrete zeitveränderliche unabhängige Variablen einfach in die parametrischen Raten-Modelle aufnehmen lassen, indem die ursprünglichen Verweildauern in Subepisoden aufgesplittet werden, innerhalb deren diese dann konstant sind, besteht bei stetigen zeitveränderlichen unabhängigen Variablen diese einfache Möglichkeit nicht.
Eine unmittelbare Lösung ist nur dann gegeben, wenn die stetigen zeitveränderlichen unabhängigen Variablen eine bestimmte vorgegebene Funktion der Verweildauer sind und direkt ein Weibull-, ein Gompertz-(Makeham-) oder ein log-logistisches Modell zur Schätzung herangezogen werden kann. Auf diese Verfahren werden wir in den Abschnitten 6.5.1 bis 6.5.3 noch genauer eingehen. Wenn dies allerdings nicht der Fall ist, dann besteht eine Lösung darin, den Verlauf der kumulativen Hazardrate zu approximieren (vgl. Abschnitt 3.8), indem man die stetige unabhängige Variable stückweise über Subepisoden als konstant betrachtet und wie im Falle diskreter zeitveränderlicher Variablen vorgeht. Die Werte der stetigen zeitveränderlichen unabhängigen Variablen werden dann zu bestimmten fest vorgegebenen Zeitpunkten ermittelt und die Episoden an diesen Stellen künstlich aufgesplittet. Wie Abbildung 6.7 zeigt, ist eine solche Approximation durch eine Treppenfunktion natürlich um so besser, je kleiner die Intervalle gewählt und je öfter die stetigen zeitveränderlichen unabhängigen Variablen neu gemessen werden.
Als ein Beispiel für die Approximation des Verlaufs einer stetigen zeitveränderlichen unabhängigen Variablen durch eine solche über Subepisoden jeweils als konstant betrachtete Treppenfunktion soll uns wieder die Berufserfahrung dienen. Diese wurde bisher bei den Exponential-Modellen als zeitkonstant behandelt und nur jeweils zu Beginn jeder neuen Berufsepisode neu gemessen. Geht man wie bisher wieder davon aus, daß sich die Berufserfahrung linear mit der Zeit erhöht, die eine Person im Beschäftigungssystem verbringt, dann kann man die Berufserfahrung durch die beim Eintritt in jede Subepisode bereits im Beschäftigungssystem verbrachten Monate approximieren (vgl. Abbildung 6.7(c)).
Die Tabelle 6.7 zeigt einen ereignisorientierten Datensatz, in dem die ursprünglichen Verweildauern im Beruf aus der Datei in Tabelle 6.4 künstlich in maximal 60 Monate lange Subepisoden aufgesplittet worden sind, bei denen die Berufserfahrung jeweils zu Beginn neu gemessen wurde[8].

[8] Das FORTRAN-Programm, mit dem diese Datei aufbereitet worden ist, ist im Anhang 4 zu finden.

Abbildung 6.7: Modellierung der stetigen Variablen Berufserfahrung (a) in ihrem Einfluß auf den Berufsverlauf, (b) als zeitveränderliche unabhängige Variable und (c) als Approximation mit Hilfe einer Treppenfunktion

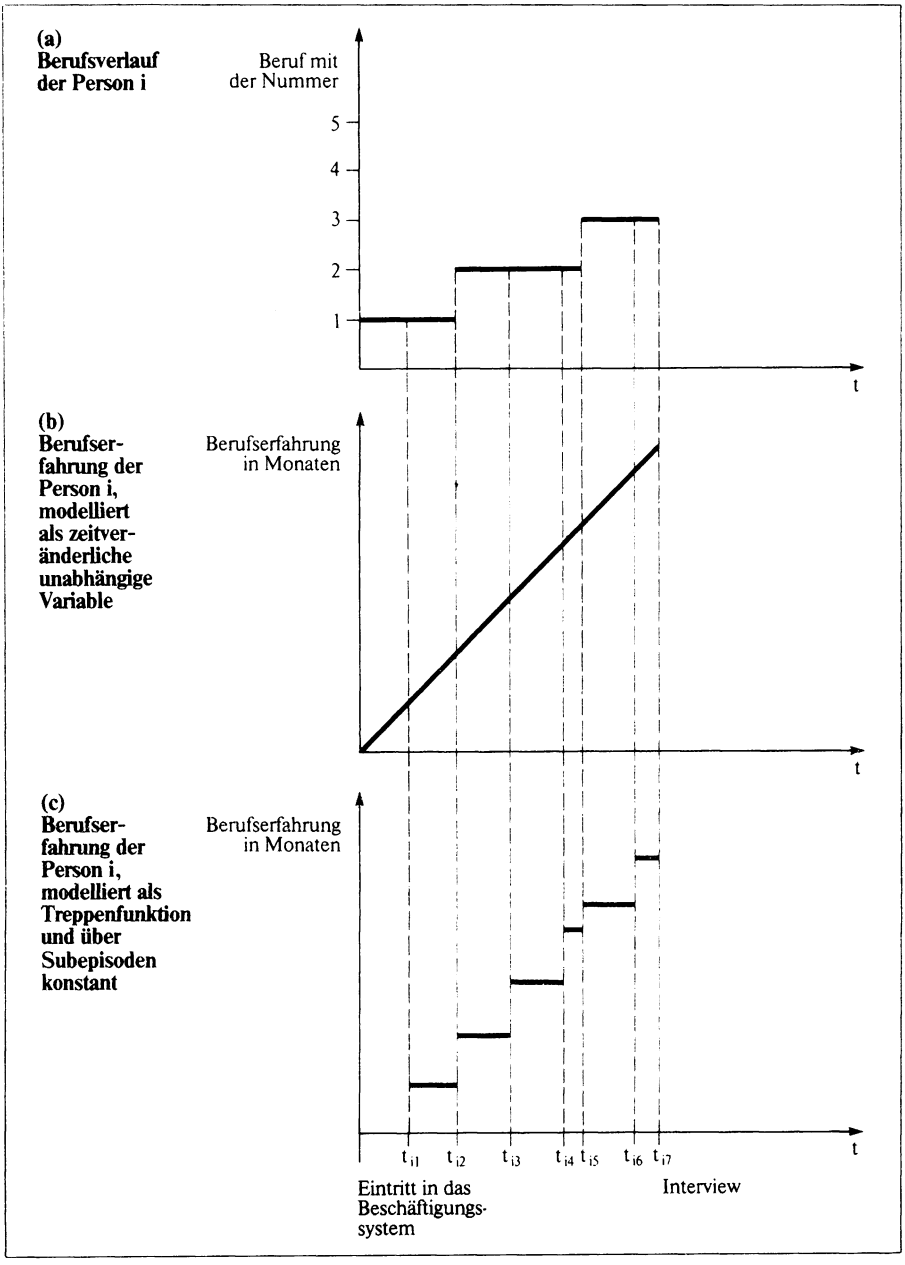


Tabelle 6.7: Beispiel für einen ereignisorientierten Datensatz, dessen Episoden in maximal 60 Monate lange Subepisoden aufgesplittet wurden

TANF	TEND	ZEN	BILDG	PRES	BANZ	BERF	KOH2	KOH3	DUR	JOBN	JOBN1	THEIRAT
555	615	0	11	66	0	0	0	0	60	11	0	679
615	675	0	11	66	0	60	0	0	60	11	0	679
675	735	0	11	66	0	120	0	0	60	11	0	679
735	795	0	11	66	0	180	0	0	60	11	0	679
795	855	0	11	66	0	240	0	0	60	11	0	679
855	915	0	11	66	0	300	0	0	60	11	0	679
915	975	0	11	66	0	360	0	0	60	11	0	679
975	983	0	11	66	0	420	0	0	8	11	0	679
583	643	0	11	50	0	0	0	0	60	3	3	701
643	651	1	11	50	0	60	0	0	8	3	3	701
651	711	0	11	50	1	68	0	0	60	3	3	701
711	771	0	11	50	1	128	0	0	60	3	3	701
771	788	1	11	50	1	188	0	0	17	3	3	701
788	848	0	11	50	2	205	0	0	60	3	0	701
848	908	0	11	50	2	265	0	0	60	3	0	701
908	968	0	11	50	2	325	0	0	60	3	0	701
968	983	0	11	50	2	385	0	0	15	3	0	701
691	717	1	9	39	0	0	1	0	26	6	3	781
728	754	1	11	56	1	37	1	0	26	3	3	781
771	831	0	11	56	2	80	1	0	60	3	0	781
.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.

In der ereignisorientierten Datei der Tabelle 6.7 liegt beispielsweise die erste Episode aus der Datei in Tabelle 6.4 aufgesplittet in acht Subepisoden vor, bei denen jeweils zu Beginn die Variable Berufserfahrung (BERF) aktualisiert wurde.

Auch hier ist wieder darauf hinzuweisen, daß sich durch dieses Episodensplitting an der Verweildauer im Zustand selbst nichts ändert.

Um zu zeigen, wie sich die Schätzungen verhalten, wenn die maximale Intervallbreite der Subepisoden verkleinert und damit die Approximation an die zeitveränderliche unabhängige Variable Berufserfahrung verbessert wird, wurden noch zwei weitere ereignisorientierte Datensätze erzeugt, in denen die Episoden nach der maximalen Intervallbreite von 24 Monaten und 12 Monaten aufgebrochen wurden. Mit dem folgenden RATE-Programm (Programmbeispiel 6.9) wurde für diese drei ereignisorientierten Eingabedateien jeweils ein Exponential-Modell mit den Variablen Bildung (BILDG), Prestige (PRES), Anzahl der vorher ausgeübten Berufe (BANZ), Berufserfahrung zu Beginn jeder Subespisode (IBERF) sowie mit den Kohorten-Dummies KOHO2 und KOHO3 geschätzt:

$$\lambda^k(v|x_k(v)) = \exp(x_k'(v)\beta), \qquad k = 1, 2, \ldots .$$

Programmbeispiel 6.9:

```
RUN NAME        EXPONENTIAL-MODELL
N OF CASES      UNKNOWN
VARIABLES       13
TANF            1
TEND            2
IZEN            3
BILDG           4
PRES            5
BANZ            6
IBERF           7
KOHO2           8
KOHO3           9
IDUR            10
JOBN            11
JOBN1           12
THEIRAT         13
READ DATA
(13F5.0)
T AND S         10 3
MODEL           (1) A=1
VECTOR          (1) 4 5 6 7 8 9
SOLVE
FINISH
```

Das obige RATE-Programm ist mit dem Programmbeispiel 6.8 bis auf die Variable HEIRAT identisch, die im Programmbeispiel 6.9 weder eingelesen noch in den Kovariablen-Vektor aufgenommen wird. Die Erklärung der RATE-Steuerkarten für das Programmbeispiel 6.8 kann deswegen auf das Programmbeispiel 6.9 übertragen werden. Die Ergebnisse der drei RATE-Schätzungen für die maximalen Subepisodenlängen 60 Monate, 24 Monate und 12 Monate sind in der Tabelle 6.8 zusammengestellt.

Zunächst kann man das Exponential-Modell, in dem die stetige zeitveränderliche Variable Berufserfahrung auf der Basis von Subepisoden mit der maximalen Intervallbreite von 60 Monaten approximiert wird (vgl. Tabelle 6.8(a)), mit dem Exponential-Modell vergleichen, in dem die Berufserfahrung über die Verweildauer hinweg als zeitkonstant betrachtet wird (vgl. Tabelle 6.3). Dabei wird zunächst sichtbar, daß sich gegenüber dem Modell ohne Kovariablen eine deutliche Verbesserung des χ^2-Werts von 705,9, bei zeitkonstanter Berufserfahrung, auf 1164,16, bei zeitveränderlicher Berufserfahrung, ergibt. Durch die Dynamisierung der Berufserfahrung kann die Erklärungskraft des Modells also deutlich erhöht werden.

Aber nicht nur das Modell als Ganzes verbessert sich, sondern es ändern sich auch die inhaltlichen Schlußfolgerungen, die man daraus ziehen muß. Insbesondere verschwindet der signifikante Effekt der Kohorten-Variable KOHO2 in Tabelle 6.3. Das heißt, der Unterschied zwischen der Kohorte 1929–31 und der Kohorte 1939–41 in Tabelle 6.3 kann durch unterschiedlich lange Berufserfah-

Tabelle 6.8: Ergebnisse der Exponential-Modelle aus Programmbeispiel 6.9 (Berufsepisoden aufgebrochen nach maximaler Intervallbreite von 60, 24 und 12 Monaten)

(a) Maximale Intervallbreite 60 Monate

DESTINATION	UNWEIGHTED FREQUENCY	WEIGHTED FREQUENCY	MAX(LOG OF L) NULL HYPOTHESIS	MAX(LOG OF L) ALTERNATIVE HYPOTHESIS	PSEUDO R-SQUARED	CHI-SQUARED	DF	PROBABILITY LEVEL
1	2586	2586.0	-1.4434320D+04	-1.3852250D+04	0.0403	1184.16	6	0.000D+00

DESTINATION 1 LETTER A LOG-LINEAR TIME-INDEPENDENT VECTOR

INTERNAL NUMBER	VARIABLE NUMBER	VARIABLE NAME	VECTOR 1 PARAMETER	PARAMETER STANDARD ERROR	PARAMETER F RATIO	ANTILOG OF THE PARAMETER	ANTILOG STANDARD ERROR	ANTILOG F RATIO
1		(CONSTANT)	-3.8130+00	1.280D-01	887.078	2.209D-02		
2	4	BILDG	4.505D-03	1.452D-02	0.096	1.005D+00	1.458D-02	0.096
3	5	PRES	-5.172D-03	1.376D-03	14.118	9.948D-01	1.369D-03	14.191
4	6	BANZ	1.503D-01	1.052D-02	204.019	1.162D+00	1.223D-02	175.883
5	7	IBERF	-8.367D-03	3.101D-04	727.855	9.917D-01	3.076D-04	733.975
6	8	KOHO2	4.051D-02	4.661D-02	0.755	1.041D+00	4.854D-02	0.725
7	9	KOHO3	1.763D-01	5.280D-02	11.146	1.193D+00	6.298D-02	9.369

(b) Maximale Intervallbreite 24 Monate

DESTINATION	UNWEIGHTED FREQUENCY	WEIGHTED FREQUENCY	MAX(LOG OF L) NULL HYPOTHESIS	MAX(LOG OF L) ALTERNATIVE HYPOTHESIS	PSEUDO R-SQUARED	CHI-SQUARED	DF	PROBABILITY LEVEL
1	2586	2586.0	-1.4434320D+04	-1.3834430D+04	0.0416	1199.79	6	0.000+00

DESTINATION 1 LETTER A LOG-LINEAR TIME-INDEPENDENT VECTOR

INTERNAL NUMBER	VARIABLE NUMBER	VARIABLE NAME	VECTOR 1 PARAMETER	PARAMETER STANDARD ERROR	PARAMETER F RATIO	ANTILOG OF THE PARAMETER	ANTILOG STANDARD ERROR	ANTILOG F RATIO
1		(CONSTANT)	-3.698D+00	1.284D-01	828.997	2.477D-02		
2	4	BILDG	3.255D-03	1.450D-02	0.050	1.003D+00	1.455D-02	0.050
3	5	PRES	-5.073D-03	1.374D-03	13.632	9.949D-01	1.367D-03	13.701
4	6	BANZ	1.472D-01	1.040D-02	200.415	1.159D+00	1.204D-02	173.300
5	7	IBERF	-8.227D-03	2.963D-04	770.781	9.918D-01	2.939D-04	777.133
6	8	KOHO2	3.607D-02	4.662D-02	0.599	1.037D+00	4.833D-02	0.578
7	9	KOHO3	1.583D-01	5.289D-02	8.957	1.172D+00	6.198D-02	7.662

(c) Maximale Intervallbreite 12 Monate

DESTINATION	UNWEIGHTED FREQUENCY	WEIGHTED FREQUENCY	MAX(LOG OF L) NULL HYPOTHESIS	MAX(LOG OF L) ALTERNATIVE HYPOTHESIS	PSEUDO R-SQUARED	CHI-SQUARED	DF	PROBABILITY LEVEL
1	2586	2586.0	-1.4434320D+04	-1.3834250D+04	0.0416	1200.16	6	0.000+00

DESTINATION 1 LETTER A LOG-LINEAR TIME-INDEPENDENT VECTOR

INTERNAL NUMBER	VARIABLE NUMBER	VARIABLE NAME	VECTOR 1 PARAMETER	PARAMETER STANDARD ERROR	PARAMETER F RATIO	ANTILOG OF THE PARAMETER	ANTILOG STANDARD ERROR	ANTILOG F RATIO
1		(CONSTANT)	-3.656D+00	1.287D-01	807.335	2.583D-02		
2	4	BILDG	3.057D-03	1.449D-02	0.045	1.003D+00	1.454D-02	0.044
3	5	PRES	-5.050D-03	1.374D-03	13.517	9.950D-01	1.367D-03	13.585
4	6	BANZ	1.461D-01	1.038D-02	198.144	1.157D+00	1.201D-02	171.512
5	7	IBERF	-8.179D-03	2.940D-04	773.943	9.919D-01	2.918D-04	780.303
6	8	KOHO2	3.587D-02	4.662D-02	0.592	1.037D+00	4.832D-02	0.571
7	9	KOHO3	1.573D-01	5.290D-02	8.844	1.170D+00	6.192D-02	7.572

rung innerhalb von Berufen erklärt werden. Das gleiche gilt auch für den Unterschied zwischen der Kohorte 1929–31 und 1949–51 (vgl. die $\hat{\beta}$-Koeffizienten für KOHO3), der sich in der Tabelle 6.8(a) deutlich verringert, aber noch immer signifikant bleibt. Die restlichen $\hat{\beta}$-Koeffizienten sind, wenn man von der

nicht signifikanten Variablen Bildung absieht, weitgehend unverändert geblieben.
In einem zweiten Schritt kann man nun untersuchen, wie sich das Modell verhält, wenn man die maximale Intervallbreite von 60 Monaten auf 24 Monate und schließlich auf 12 Monate herabsetzt und dabei die Approximation an die stetige zeitveränderliche Variable Berufserfahrung zunehmend verbessert.
Es zeigt sich, daß sich gegenüber dem Modell ohne Kovariablen der χ^2-Wert von 1164,16, bei einer maximalen Intervallbreite von 60 Monaten, auf 1199,79, bei einer maximalen Intervallbreite von 24 Monaten, nur leicht verbessert. Ein Vergleich der einzelnen $\hat{\beta}$-Koeffizienten und F-Werte macht darüber hinaus deutlich, daß sich die Schätzungen nur geringfügig verändern. Zu nahezu demselben Resultat kommt man schließlich, wenn man das Modell mit der maximalen Intervallbreite von 24 Monaten mit dem Modell mit der maximalen Intervallbreite von 12 Monaten vergleicht. Insgesamt folgt daraus, daß man im vorliegenden Beispiel bereits bei einer maximalen Intervallbreite von 24 Monaten unter inhaltlichen Gesichtspunkten zu einer relativ befriedigenden Approximation an die stetige zeitveränderliche Variable Berufserfahrung kommt.

6.4 Modelle mit periodisierter Verweildauer

Bei den Beispielen in Abschnitt 6.3 sind wir davon ausgegangen, daß zwar die Ausprägungen der Kovariablen über die Verweildauer hinweg variieren können, daß aber die Kovariablen an sich und ihre β-Gewichte dabei unverändert bleiben. Eine solche Formulierung des Berufswechselrisikos muß aber nicht unbedingt sinnvoll sein, wie das folgende Beispiel zeigen wird. Wie bereits im Abschnitt 3.3 ausgeführt, besteht nämlich eine weitere Möglichkeit der Modellierung zeitabhängiger Hazardraten darin, die Verweildauer in Perioden zu unterteilen und die Hazardrate mit von Periode zu Periode variierenden Kovariablen und/oder β-Koeffizienten zu schätzen.
Im Falle der Untersuchung des Berufswechselverhaltens bei Männern könnte man beispielsweise mit den Vertretern der Filter- oder Signaltheorie (vgl. Arrow 1973; Spence 1973, 1974) argumentieren, daß bestimmte Merkmale wie Bildungsabschluß, Anzahl der vorher bereits ausgeübten Berufe und bisherige Berufserfahrung für die Arbeitgeber eine Signalfunktion besitzen und besonders häufig zu Beginn jeder neuen Berufstätigkeit zur Produktivitätsbeurteilung mit herangezogen werden. In dem Maße aber, in dem die Arbeitgeber die Leistung der Arbeitnehmer aufgrund eigener Erfahrungen besser berurteilen können, dürfte die Bedeutung dieser Signalmechanismen für die Personalentscheidungen zurückgehen. Umgekehrt gilt natürlich auch, daß das Verhalten des Arbeitnehmers insbesondere zu Beginn jeder neuen Berufstätigkeit stark vom Image

(oder Prestige) des jeweiligen Berufes bestimmt wird und es sich erst nach und nach an die tatsächlichen Gegebenheiten anpaßt. Beide Argumente führen zu der Vermutung, daß die Wirkung bestimmter fixer Variablen in Form ihrer β-Koeffizienten über die Verweildauer hinweg nicht konstant ist.
Im vorliegenden Berufswechselbeispiel ist es deswegen sinnvoll, die Verweildauer im Beruf in Perioden aufzuteilen und die β-Koeffizienten der Kovariablen über diese Perioden variieren zu lassen. Das besondere Problem besteht dabei allerdings darin, wie viele Perioden mit welcher Länge modelliert werden sollen. Einerseits ist zu beachten, wie bereits in Abschnitt 3.3 diskutiert, daß das Perioden-Modell selbst bei einer mittleren Zahl von Perioden schon zu einer großen Zahl zu schätzender Parameter führt und entsprechend große Datenmengen erfordert. Auch im Sinne einer parametersparsamen Modellbildung ist es sinnvoll, nicht zu viele Perioden zu wählen. Andererseits sollte die Schätzung eines konstanten Risikos in jeder der Perioden tatsächlich angemessen sein und nicht zu einer Fehlspezifikation führen. Für das Berufswechselrisiko wählen wir deswegen drei Perioden: eine Periode vom Eintritt in jeden neuen Beruf bis zur Verweildauer von zwei Jahren (24 Monate), eine zweite Periode von der Verweildauer von zwei Jahren bis zur Verweildauer von fünf Jahren (60 Monate) und eine dritte Periode, die nach einer Verweildauer von fünf Jahren beginnt. Es ist anzunehmen, daß das Berufswechselrisiko in der ersten Periode besonders stark durch die wechselseitige Anpassung der Erwartungen von Arbeitnehmern und Arbeitgebern bestimmt wird und daß den „Filtervariablen" hier große Bedeutung zukommt. In der zweiten Periode müßte die Wirkung dieser Variablen zurückgehen, weil sich beide Seiten stärker an den tatsächlich gegebenen Verhältnissen orientieren können. In der dritten Periode schließlich dürften diese Variablen fast keine Rolle mehr für das Berufswechselrisiko spielen.
Die Schätzung dieses periodenspezifischen Exponential-Modells

$$\lambda_p^k(v|x_k) = \exp(x_k'\beta_p) \qquad \text{mit } p = 1, 2, 3, k = 1, 2, \ldots$$

erfolgt mit dem folgenden RATE-Programmlauf:

Programmbeispiel 6.10:

```
RUN NAME        MODELL MIT PERIODISIERUNG
N OF CASES      3516
VARIABLES       12
TANF            1
TEND            2
ZEN             3
BILDG           4
PRES            5
BANZ            6
BERF            7
KOHO2           8
KOHO3           9
DUR             10
JOBN            11
JOBN1           12
```

```
READ DATA
(12F5.0)
TIME INTERVALS  0.0 24.0 60.0
T AND S         10 3
MODEL           (4) A=2 B=0 C=0
VECTOR          (1) 4 5 6 7 8 9
SOLVE           (2)=30
FINISH
```

Wie in den Programmbeispielen 6.8 und 6.9 werden im obigen RATE-Lauf zuerst die Variablen mit ihren RATE-internen Nummern aufgelistet und mit dem angegebenen Format eingelesen. Mit der TIME-INTERVALS-Karte gibt man die Intervallgrenzen für die Perioden an, in diesem Fall also 0, 24 und 60 Monate. Auf der MODEL-Karte wird das Modell mit der Nummer (4) spezifiziert. Die Buchstaben B und C werden dabei auf Null gesetzt und der Buchstabe A wird periodenspezifisch und log-linear mit den Kovariablen-Vektoren verbunden (A = 2) : $A = \exp(x_k'\beta_p)$[9]. Da auf der Vektorkarte nur der Vektor (1) angegeben wurde, werden in allen drei Perioden automatisch dieselben Variablen in die periodenspezifischen Kovariablen-Vektoren aufgenommen. Es handelt sich dabei um die Variablen, die bereits in den Programmbeispielen 6.8 und 6.9 besprochen wurden. Auf der SOLVE-Karte, mit der die Maximum-Likelihood-Schätzung der periodenspezifischen Koeffizienten angefordert wird, wird schließlich noch mit der Anweisung (2) = 30 die maximale Anzahl der durchzuführenden Iterationen auf 30 hochgesetzt. Das Ergebnis dieses RATE-Programms ist in Tabelle 6.9 zu finden.

Zunächst erhält man auf der Grundlage des Likelihood-Quotienten-Tests für das periodenspezifische Exponential-Modell in Tabelle 6.9, verglichen mit dem Exponential-Modell ohne Kovariablen und nur einer Regressionskonstanten β_0, einen χ^2-Wert von 1163,79 – jetzt allerdings bei 20 Freiheitsgraden, da für jede dieser drei Perioden sieben β-Parameter geschätzt wurden. Das Modell ist signifikant (PROBABILITY LEVEL = 0,00), und die Nullhypothese, keiner der 20 zusätzlich eingeführten Regressionsparameter erkläre etwas am Risiko des Berufswechsels bei Männern, muß abgelehnt werden.

Die Ergebnisse der ersten Periode (0 bis 24 Monate) zeigen (Tabelle 6.9(a)), daß die Variablen Bildung (BILDG) und KOHO2 keinen signifikanten Einfluß haben (beide F-Teststatistiken sind kleiner als der Wert 3,84 (≙ Signifikanzniveau 0,05)). Entgegen unserer ursprünglichen Hypothese besitzt der Bildungsabschluß also auch in der ersten Phase jeder neuen Erwerbstätigkeit keine große Relevanz für die Erklärung des Berufswechselrisikos. Ganz anders ist dies bei den Variablen Prestige (PRES), Anzahl der vorher ausgeübten Berufe (BANZ) und der Berufserfahrung (BERF). Ein Vergleich ihrer $\hat{\beta}$-Koeffizienten mit den entsprechenden $\hat{\beta}$-Koeffizienten des einfachen Exponential-Modells in Tabelle

[9] Mit zusätzlicher linearer und periodenspezifischer Spezifikation des Buchstaben C (C = -2) : $C = x_k'\beta_p$ könnte man aber auch ein periodenspezifisches Gompertz-Modell in RATE modellieren.

Tabelle 6.9: Ergebnis des Exponential-Modells mit periodisierter Verweildauer aus Programmbeispiel 6.10

DESTINATION	UNWEIGHTED FREQUENCY	WEIGHTED FREQUENCY	MAX(LOG OF L) NULL HYPOTHESIS	MAX(LOG OF L) ALTERNATIVE HYPOTHESIS	PSEUDO R-SQUARED	CHI-SQUARED	DF	PROBABILITY LEVEL
1	2586	2586.0	-1.4434320+04	-1.3852430+04	0.0403	1163.79	20	0.000+00

(a) Periode 0 – 24 Monate

INTERNAL NUMBER	VARIABLE NUMBER	VARIABLE NAME	DESTINATION 1 LETTER A VECTOR 1 PARAMETER	LOG-LINEAR TIME-DEPENDENT VECTOR, PERIOD 1 PARAMETER STANDARD ERROR	PARAMETER F RATIO	ANTILOG OF THE PARAMETER	ANTILOG STANDARD ERROR	ANTILOG F RATIO
1		(CONSTANT)	-3.8450+00	1.8770-01	419.460	2.1400-02		
2	4	BILDG	1.7420-02	2.1330-02	0.667	1.0180+00	2.1700-02	0.656
3	5	PRES	-6.7790-03	2.0900-03	10.524	9.9320-01	2.0760-03	10.596
4	6	BANZ	1.9500-01	1.5820-02	152.043	1.2150+00	1.9220-02	125.501
5	7	BERF	-9.0600-03	6.9720-04	168.866	9.9100-01	6.9090-04	170.404
6	8	KOHO2	6.6010-02	7.3840-02	0.799	1.0680+00	7.8880-02	0.748
7	9	KOHO3	2.1600-01	7.6750-02	7.919	1.2410+00	9.5250-02	6.406

(b) Periode 24 – 60 Monate

INTERNAL NUMBER	VARIABLE NUMBER	VARIABLE NAME	DESTINATION 1 LETTER A VECTOR 2 PARAMETER	LOG-LINEAR TIME-DEPENDENT VECTOR, PERIOD 2 PARAMETER STANDARD ERROR	PARAMETER F RATIO	ANTILOG OF THE PARAMETER	ANTILOG STANDARD ERROR	ANTILOG F RATIO
8		(CONSTANT)	-4.0610+00	2.2290-01	331.862	1.7230-02		
9	4	BILDG	1.1300-02	2.5690-02	0.194	1.0110+00	2.5980-02	0.191
10	5	PRES	-3.7930-03	2.4280-03	2.440	9.9620-01	2.4190-03	2.450
11	6	BANZ	1.1050-01	2.5260-02	19.139	1.1170+00	2.8210-02	17.154
12	7	BERF	-8.3320-03	8.3400-04	99.814	9.9170-01	8.2710-04	100.650
13	8	KOHO2	1.5910-01	8.5350-02	3.473	1.1720+00	1.0010-01	2.968
14	9	KOHO3	2.9310-01	9.3050-02	9.918	1.3410+00	1.2470-01	7.452

(c) Periode über 60 Monate

INTERNAL NUMBER	VARIABLE NUMBER	VARIABLE NAME	DESTINATION 1 LETTER A VECTOR 3 PARAMETER	LOG-LINEAR TIME-DEPENDENT VECTOR, PERIOD 3 PARAMETER STANDARD ERROR	PARAMETER F RATIO	ANTILOG OF THE PARAMETER	ANTILOG STANDARD ERROR	ANTILOG F RATIO
15		(CONSTANT)	-4.4940+00	2.7570-01	265.719	1.1170-02		
16	4	BILDG	-2.4470-02	3.1230-02	0.614	9.7580-01	3.0470-02	0.629
17	5	PRES	-2.9720-03	2.8310-03	1.102	9.9700-01	2.8220-03	1.106
18	6	BANZ	8.7620-02	3.1190-02	7.927	1.0920+00	3.4050-02	7.265
19	7	BERF	-8.8760-03	9.0820-04	57.308	9.9310-01	9.0200-04	57.705
20	8	KOHO2	4.8010-02	8.6710-02	0.307	1.0490+00	9.0970-02	0.292
21	9	KOHO3	1.6870-01	1.3260-01	1.620	1.1840+00	1.5690-01	1.372

6.3 zeigt deutlich, daß zu Beginn jeder neuen Erwerbsphase der Einfluß dieser Variablen stärker als im Durchschnitt ist: Prestige (–0,0068 im Vergleich zu –0,0052), Anzahl der vorher ausgeübten Berufe (0,1950 im Vergleich zu 0,1714) und Berufserfahrung (–0,0091 im Vergleich zu –0,0084). Dies spricht für die Hypothese, daß diesen Variablen insbesondere zu Beginn jeder neuen Berufsepisode eine große Bedeutung zukommt, und stützt somit die „Filter"- oder „Signal"-Hypothese.

Noch klarer wird dieses Resultat, wenn man die absoluten und standardisierten $\hat{\beta}$-Koeffizienten dieser Variablen über die drei Perioden vergleicht. So hat beispielsweise die Variable Prestige nach der ersten Periode keinen signifikanten Effekt mehr, und die signifikanten Variablen BANZ und BERF verlieren von

Periode zu Periode an Wirkung: $\hat{\beta}^{P1}_{BANZ} = 0{,}1950$, $\hat{\beta}^{P2}_{BANZ} = 0{,}1105$ und $\hat{\beta}^{P3}_{BANZ} = 0{,}0878$ sowie $\hat{\beta}^{P1}_{BERF} = -0{,}0091$, $\hat{\beta}^{P2}_{BERF} = -0{,}0083$ und $\hat{\beta}^{P3}_{BERF} = -0{,}0069$. Dies spricht ebenfalls für die Hypothese, daß sich das Verhalten mit zunehmender Verweildauer im Beruf mehr und mehr an den tatsächlichen Verhältnissen orientiert und deswegen die Erklärungskraft der „Filter"-Variablen für das Berufswechselverhalten zurückgeht.

Natürlich sollte man bei dem Ergebnis in Tabelle 6.9 in einem zweiten Schritt alle diejenigen Kovariablen, die in den verschiedenen Perioden keinen Erklärungsbeitrag leisten können, jeweils aus dem Modell herausnehmen. Man würde damit zu einem Modell kommen, in dem nicht nur die β-Koeffizienten, sondern auch die Kovariablen von Periode zu Periode variieren. Dies läßt sich mit RATE einfach realisieren[10], soll aber an dieser Stelle nicht weiter ausgeführt werden.

Insgesamt bietet sich mit der Möglichkeit, die Hazardraten periodenspezifisch variieren zu lassen, ein zusätzliches Instrument an, mit dem man stochastische Prozesse im Bereich der Wirtschafts- und Sozialwissenschaften realitätsgerechter modellieren kann. Die Hazardrate läßt sich insbesondere bei unregelmäßigem Verlauf, bei dem die herkömmlichen Verfahren der Verweildauerabhängigkeit meist versagen, noch mit einem periodenspezifischen Ratenmodell analysieren. Der Nachteil dieser Modelle besteht aber darin, daß sich mit steigender Zahl der Perioden und Kovariablen die Anzahl der zu schätzenden Parameter rasch erhöht. Dies erfordert bereits bei mittlerer Periodenzahl nicht nur große Datenmengen, sondern steigert auch die Komplexität der zu interpretierenden Ergebnisse. Wenn möglich sollte man deswegen besser gleich auf ein parametrisches Verweildauermodell zurückgreifen und versuchen, den zu erklärenden Prozeß parametersparsam in den Griff zu bekommen.

6.5 Modelle mit Verweildauerabhängigkeit der Hazardrate: Das Gompertz-(Makeham-), das Weibull- und das log-logistische Modell

Die Modellierung parametrischer Modelle der Verweildauerabhängigkeit wirft natürlich unter inhaltlichen Gesichtspunkten zunächst die Frage auf, wie im Bereich der Wirtschafts- und Sozialwissenschaften die Zeitabhängigkeit zu interpretieren ist.

Dabei sollte man sich am Beispiel der Berufserfahrung als erstes bewußt machen, daß mit dem Exponential-Modell in Tabelle 6.8 nicht nur die stetige zeitveränderliche Variable Berufserfahrung, sondern gleichzeitig bereits auch ein spezielles Gompertz-Modell über eine Treppenfunktion approximiert wurde. Geht man nämlich wie in diesem Beispiel davon aus, daß sich die

[10] Man gibt für das obige Beispiel auf der VECTOR-Karte nur die jeweils signifikanten Variablen an: VECTOR (1) 5 6 7 9 (2) 6 7 9 (3) 6 7.

Berufserfahrung linear mit der Zeit erhöht, die eine Person i im Beschäftigungssystem verbringt,

$$x_{BERF}(t) = x_{BERF}(t_{i,k-1}) + v \qquad \text{, mit } v = t - t_{i,k-1} \; (v \geq 0),$$

und nimmt man diese dynamisierte Berufserfahrungsvariable $x_{BERF}(t)$ in ein Exponential-Modell auf, dann läßt sich dieses Modell wie folgt umformen:

$$\begin{aligned}\lambda^k(v|x_k(t)) &= \exp(x_k'(t)\beta) \\ &= \exp(x_k'\beta + \beta_{BERF}\, x_{BERF}(t)) \\ &= \exp(x_k'\beta + \beta_{BERF}\, (x_{BERF}(t_{k-1}) + v)) \\ &= \exp(x_k'\beta + \beta_{BERF}\, x_{BERF}(t_{k-1}))\, \exp(\beta_{BERF}\, v),\end{aligned}$$

wobei $x_{BERF}(t_{k-1})$ die Berufserfahrung zu Beginn jeder neuen Berufsepisode k und v die Verweildauer in der Episode k bezeichnet. Dies aber ist gemäß (3.3.15) ein Gompertz-Modell mit $\lambda_0(x_k) = \exp(x_k'\beta + \beta_{BERF}\, x_{BERF}(t_{k-1}))$ und $\gamma_0 = \beta_{BERF}$ (vgl. dazu auch Abschnitt 6.5.1):

$$\lambda^k(v) = \lambda_0(x_k)\, \exp(\gamma_0 \cdot v).$$

Entsprechend hätte man ein Weibull-Modell erhalten, wenn man gute inhaltliche Gründe dafür gehabt hätte, daß sich die Berufserfahrung bei einer Person i nicht linear, sondern logarithmisch mit der Zeit t erhöhen würde, die diese im Beschäftigungssystem verbringt:

$$x_{BERF}(t) = x_{BERF}(t_{i,k-1}) + \ln v \qquad \text{, mit } v = t - t_{i,k-1} \; (v \geq 0).$$

Das Exponential-Modell ließe sich dann wie folgt schreiben:

$$\lambda^k(v|x_k(t)) = \exp(x_k'\beta + \beta_{BERF}\, x_{BERF}(t_{k-1}))\, v^{\beta_{BERF}},$$

und gemäß (3.2.19) mit $\lambda^*(x_k) = \exp(x_k'\beta + \beta_{BERF}\, x_{BERF}(t_{k-1}))$ sowie mit $\gamma^* = \beta_{BERF}$ erhielte man das folgende spezielle Weibull-Modell (vgl. dazu auch Abschnitt 6.5.2):

$$\lambda^k(v) = \lambda^*(x_k)\, v^{\gamma^*}.$$

In beiden Fällen wäre man also aufgrund inhaltlicher Hypothesen und über die Modellierung einer stetigen zeitveränderlichen unabhängigen Variablen direkt zu einer substantiellen Interpretation der Zeitabhängigkeit gekommen.
Leider ist man im Bereich der Wirtschafts- und Sozialwissenschaften aber nicht immer in der glücklichen Lage, den Wandel stetiger zeitveränderlicher unabhängiger Variablen genau messen zu können. Vielmehr ist man häufig auf Proxy-Variablen angewiesen, mit deren Hilfe man den Effekt dieser Variablen stellvertretend modelliert[11].

[11] In der Tat ist in den obigen Beispielen die Berufserfahrung ja auch über die Proxy-Variable „Anzahl von Jahren im Beschäftigungssystem" in das Modell eingegangen (vgl. auch Tuma 1985, S. 340).

Zu diesen Proxy-Variablen zählt natürlich auch die Verweildauer selbst, die man als stetige zeitveränderliche unabhängige Variable über bestimmte Verteilungsmodelle auf die Hazardrate wirken lassen kann. Es hängt dann jeweils von den inhaltlichen Überlegungen und der sich daraus ergebenden angenommenen Wirkungsweise der stellvertretend zu modellierenden unabhängigen Variablen ab, auf welches konkrete Verteilungsmodell der Verweildauerabhängigkeit man dabei zurückgreift. Prinzipiell kann dazu jede beliebige Wahrscheinlichkeitsverteilung herangezogen werden, soweit sie sich überhaupt zur Beschreibung von Verweildauern eignet. Auf der Grundlage von Forschungserfahrungen aus den Bereichen Medizin, Biologie, Demographie, Technik, Psychologie sowie aus dem Bereich der Wirtschafts- und Sozialwissenschaften hat sich allerdings gezeigt, daß man sich auf einige wenige Verteilungsmodelle als die wichtigsten beschränken kann. Zur Modellierung von Kovariablen, die bei der Hazardrate zu einem monoton fallenden beziehungsweise monoton steigenden Verlauf führen, wie es in den Wirtschafts- und Sozialwissenschaften häufig zu beobachten ist, wird in der Regel auf die Gompertz-(Makeham-) (vgl. Abbildung 3.8) oder auf die Weibull-Verteilung (vgl. Abbildung 3.5) zurückgegriffen und bei den Einflußfaktoren, die zuerst zu einem monoton steigenden und später nach einem bestimmten Punkt zu einem monoton fallenden Verlauf der Hazardrate führen, hat sich in der Praxis meist das log-logistische Modell durchgesetzt (vgl. Abbildung 3.7). Die folgenden Anwendungsbeispiele und Interpretationen beschränken sich deswegen auf diese drei Verteilungstypen, die für die wichtigsten Fälle genügen.

6.5.1 Das Gompertz-(Makeham-)Modell

Wie in Abschnitt 3.2 bereits ausgeführt, fand die Gompertz-(Makeham-)Verteilung vor allem in der Demographie und im Versicherungswesen eine weite Verbreitung. Sie ergibt sich dann, wenn es sehr viele voneinander unabhängige Todesursachen gibt, von denen die am ehesten eintretende die Lebenszeit beendet.
Das sich daraus ergebende Gompertz-Makeham-Gesetz, nach dem die Hazardrate mit zunehmender Verweildauer monoton fällt, hat man sich insbesondere beim Studium der Lebenszeit von Organisationen (vgl. Carroll/Delacroix 1982; Freeman/Carroll/Hannan 1983) und bei der Untersuchung beruflicher Mobilitätsprozesse (Sørensen/Tuma 1981; Sørensen 1979) zunutze gemacht.

Das Gompertz-Modell ohne Kovariablen

Zur Veranschaulichung der Interpretation des Gompertz-Modells soll zunächst wieder auf der Grundlage des Mehr-Episoden-Falls für das durchschnittliche Berufswechselrisiko der Männer ein Modell ohne Kovariablen geschätzt werden. Dabei wird der λ_0-Koeffizient der Gompertz-Verteilung log-linear mit

einer Regressionskonstanten β_0 verknüpft und γ_0 direkt geschätzt:

$$\lambda^k(v) = \exp(\beta_0) \cdot \exp(\gamma_0 v) \quad , k = 1, 2, \ldots .$$

Die Verweildauer v soll uns in diesem Beispiel als Proxy-Variable für die in jedem neuen Beruf jeweils neu zu erwerbenden berufsspezifischen Kenntnisse und Fähigkeiten dienen. Das heißt, die berufsspezifische Erfahrung beginnt wie die Verweildauer selbst mit jedem neuen Beruf jeweils bei Null und wächst linear mit der im Beruf verbrachten Zeit an. Da der Erwerb dieser berufsspezifischen Kenntnisse und Fähigkeiten als Investition verstanden werden kann, die Kosten verursacht und die zum Großteil verloren wäre, wenn man den Beruf wechselt, gehen wir von der Hypothese aus, daß mit zunehmender Berufserfahrung die Hazardrate monoton sinkt. Dies wiederum heißt bei einer Gompertz-Verteilung, daß bei Richtigkeit der Theorie ein signifikantes $\hat{\gamma}_0$ mit einem negativen Vorzeichen zu erwarten ist (vgl. Abbildung 3.8).
Die Schätzung dieses Modells soll wieder mit Hilfe des BMDP-Programms P3R und dem darin eingebundenen Unterprogramm P3RFUN von Trond Petersen berechnet werden:

Programmbeispiel 6.11:

```
/INPUT UNIT IS 30.
       CODE IS DATA.
/VARIABLE NAMES ARE (63)DUR,(64)ZEN,(65)BILDG,(66)BERF,
                    (67)BANZ,(68)KOHO2,(69)KOHO3,(70)X1,(71)DP.
          ADD IS 9.
/TRANSFORM USE = (M3 EQ 1).
           DUR = M51 - M50 + 1.
           ZEN = 1.
           IF (M51 EQ M47) THEN ZEN = 0.
           IF (M41 EQ 1 AND M42 EQ 1) THEN BILDG = 9.
           IF (M41 EQ 1 AND (M42 EQ 2 OR M42 EQ 3)) THEN BILDG = 11.
           IF (M41 EQ 2 AND M42 EQ 1) THEN BILDG = 10.
           IF (M41 EQ 2 AND (M42 EQ 2 OR M42 EQ 3)) THEN BILDG = 12.
           IF (M41 EQ 3 AND (M42 EQ 1 OR M42 EQ 2 OR M42 EQ 3))
           THEN BILDG = 13.
           IF (M42 EQ 4) THEN BILDG = 17.
           IF (M42 EQ 5) THEN BILDG = 19.
           KOHO2 = 0.
           KOHO3 = 0.
           IF(M48 GE 468 AND M48 LE 504) THEN KOHO2 = 1.
           IF(M48 GE 588 AND M48 LE 624) THEN KOHO3 = 1.
           BERF = M50 - M43.
           BANZ = M5 - 1.
           DP = 0.
           X1 = 0.
/REGRESS DEPENDENT IS DP.
         PARAMETERS ARE 2.
         PRINT IS 0.
         MEANSQUARE IS 1.0.
         ITERATIONS ARE 100.
         LOSS.
```

```
/PARAMETER INITIAL ARE -3.8,-0.01.
          NAMES ARE KONST,VDABH.
/END.
/COMMENT'
M 2
T 70 63
D 71
C 64
I 0
```

Das obige BMDP-Programm unterscheidet sich von dem Programmbeispiel 6.3 nur insofern, als jetzt bei PARAMETERS ARE die Zahl 2 für die zwei zu schätzenden Parameter und bei INITIAL ARE zusätzlich für den zu schätzenden Parameter γ_0 ein Startwert vorgegeben wird (VDABH) (vgl. Anhang 1). Auch bei den Steuerkarten für das Unterprogramm P3RFUN, die nach dem END-Paragraphen folgen, ändert sich nur die Modell-Karte M, in die jetzt die Nummer 2 für das Gompertz-Modell eingesetzt werden muß. Das Ergebnis der Schätzung ist in Tabelle 6.10 zu finden.

Tabelle 6.10: Ergebnis des Gompertz-Modells aus Programmbeispiel 6.11

PARAMETER	ESTIMATE	ASYMPTOTIC STANDARD DEVIATION
KONST	-4.080915	0.026859
VDABH	-0.008132	0.000393

Nach Tabelle 6.10 ergeben sich bei einem Wert der Log-Likelihood-Funktion ($\triangleq$ - LOSS) von -14147,10 die Schätzungen $\hat{\beta}_0 = -4{,}0809$ und $\hat{\gamma}_0 = -0{,}0081$. Ein Vergleich dieses Gompertz-Modells mit dem Exponential-Modell ohne Kovariablen in Tabelle 6.2 führt damit auf der Grundlage eines Likelihood-Quotienten-Tests zu einem χ^2-Wert von

$$Lq = 2(-14147{,}10 - (-14434{,}35)) = 574{,}5,$$

mit einem Freiheitsgrad. Die Aufnahme der Proxy-Variablen Verweildauer, die in diesem Modell für den Erwerb berufsspezifischer Kenntnisse steht, verbessert die Schätzung also beträchtlich.
Ein Vergleich der obigen Maximum-Likelihood-Schätzungen $\hat{\lambda}_0 = \exp(-4{,}0809) = 0{,}0169$ und $\hat{\gamma}_0 = -0{,}0081$ mit den Schätzungen aus Abschnitt 6.1 für das Gompertz-Modell ($\hat{\lambda}_0 = 0{,}0202$ und $\hat{\gamma}_0 = -0{,}0095$) zeigt wieder, daß die Schätzungen auf der Basis der graphischen Verfahren bereits relativ treffsicher waren und die Abweichungen nur geringfügig sind.
Unter inhaltlichen Gesichtspunkten besonders wichtig ist, daß der geschätzte γ_0-Koeffizient tatsächlich das erwartete Vorzeichen besitzt, das heißt also, die Neigung zum Berufswechsel mit zunehmenden berufsspezifischen Kentnissen monoton sinkt. Vergleicht man beispielsweise einen Mann, der gerade in einen neuen Beruf eingetreten ist [$\lambda^k(0) = \exp(-4{,}0809) = 0{,}0169$], mit einem Mann,

der bereits zehn Jahre in ein und demselben Beruf arbeitet [$\lambda^k(120) = \exp(-4{,}0809) \cdot \exp(-0{,}0081 \cdot 120) = 0{,}0064$], dann hat sich bei dem zweiten Mann die Neigung zum Berufswechsel durch die Akkumulation der berufsspezifischen Kenntnisse um 62,13 Prozent [$((0{,}0064 - 0{,}0169)/0{,}0169) \cdot 100\% = -62{,}13\%$] verringert.
Über die Beziehung

$$S(t) = \exp(-\frac{\lambda_0}{\gamma_0}(\exp(\gamma_0 t)-1))$$

läßt sich wieder der Median der Verweildauer von Männern im Beruf M*, mit

$$S(M^*) = 0{,}5,$$

abschätzen:

$$S(\hat{M}^*) = \exp(-\frac{0{,}0169}{-0{,}008132}(\exp(-0{,}008132 \cdot \hat{M}^*) - 1))$$

$$\hat{M}^* = 49{,}90 \text{ Monate.}$$

Der Median nach dem Gompertz-Modell beträgt also etwas mehr als vier Jahre und ist damit weit kleiner als der Median bei der Exponential-Verteilung in Abschnitt 6.2.1.
Darüber hinaus kann man wieder die Wahrscheinlichkeit angeben, daß beispielsweise ein Mann nach einem Zeitraum von acht Jahren noch in demselben Beruf arbeitet. Sie beträgt bei der Gompertz-Verteilung mit $\hat{S}(96) = \exp(-\frac{0{,}0169}{-0{,}008132}(\exp(-0{,}008132 \cdot 96)-1)) = 0{,}32426$, also 32,43 Prozent. Im Vergleich dazu hatte sich beim Exponential-Modell in Abschnitt 6.2.1 eine Wahrscheinlichkeit von 37,56 Prozent ergeben.
Die soeben gegebenen Interpretationen gelten natürlich wieder nur unter der Voraussetzung, daß man bei den Männern tatsächlich von einer homogenen Population sprechen kann. Daß dies nicht der Fall ist, wurde bereits mit vielen Beispielen gezeigt. Dabei ist das Argument heterogener Subpopulationen im vorliegenden Beispiel besonders wichtig, da sich bei konstanten Raten in den Subgruppen eine scheinbar monoton fallende Verweildauerabhängigkeit ergeben und in einem signifikanten γ_0-Koeffizienten ausdrücken könnte. Die inhaltliche Erklärung, daß die Hazardrate wegen der Anhäufung berufsspezifischer Kenntnisse und Fähigkeiten falle, wäre dann eine Fehlinterpretation. Es soll deswegen in einem nächsten Schritt das Gompertz-Modell mit Berücksichtigung beobachteter Heterogenität geschätzt werden.

Das Gompertz-Modell mit Berücksichtigung von Kovariablen im λ_0-Term

Die Aufnahme von Kovariablen in das Gompertz-Modell kann zunächst in der Weise geschehen, daß der Parameter λ_0 log-linear mit dem Kovariablen-Vektor verbunden wird (vgl. Abschnitt 3.3.2): $\lambda_0(\mathbf{x}) = \exp(\mathbf{x}'\boldsymbol{\beta})$.
Im Berufswechselbeispiel heißt dies, daß beim Eintritt in jeden neuen Beruf,

wenn die berufsspezifischen Kenntnisse (gemessen über die Proxy-Variable Verweildauer) gleich Null sind, eine Reihe zeitunabhängiger Kovariablen wie beim Exponential-Modell auf die Neigung zum Berufswechsel wirkt. Mit zunehmender Verweildauer im Beruf und den damit wachsenden berufsspezifischen Kenntnissen sollte dann allerdings die Hazardrate nach unserer Hypothese monoton fallen. Dies ist dann der Fall, wenn die Schätzung von γ_0 signifikant von Null verschieden ist und ein negatives Vorzeichen besitzt. Das Gompertz-Modell lautet dann wie folgt:

$$\lambda^k(v|x_k) = \exp(x_k'\beta) \cdot \exp(\gamma_0 v) \qquad , k = 1, 2, \ldots .$$

In den Kovariablen-Vektor x sollen wieder die bereits bekannten Variablen Bildung (BILDG), Prestige (M59), Anzahl der bereits vorher ausgeübten Berufe (BANZ), Berufserfahrung beim Eintritt in den Beruf (BERF) sowie die Kohorten-Dummies KOHO2 und KOHO3 eingehen (vgl. Anhang 1). Die tatsächliche Berufserfahrung zum Zeitpunkt t ($x_{BERF}(t)$) wird in diesem Modell also in zwei Komponenten aufgeteilt: einerseits in die allgemeine Berufserfahrung ($x_{BERF}(t_{k-1})$), die jemand mitbringt, wenn er in einen Beruf eintritt, und andererseits in die berufsspezifische Berufserfahrung ($x_{BERF}(v)$), die jemand innerhalb des jeweiligen Berufes akkumuliert[12]:

$$x_{BERF}(t) = \beta_{BERF} x_{BERF}(t_{k-1}) + \gamma_0 x_{BERF}(v)$$

Die Maximum-Likelihood-Schätzung der Parameter soll wieder mit dem BMDP-Programm P3R und dem darin eingebundenen Unterprogramm P3RFUN von Trond Petersen berechnet werden.

Programmbeispiel 6.12:

```
/INPUT UNIT IS 30.
       CODE IS DATA.
/VARIABLE NAMES ARE (63)DUR,(64)ZEN,(65)BILDG,(66)BERF,
                    (67)BANZ,(68)KOHO2,(69)KOHO3,(70)X1,(71)DP.
          ADD IS 9.
/TRANSFORM USE = (M3 EQ 1).
           DUR = M51 - M50 + 1.
           ZEN = 1.
           IF (M51 EQ M47) THEN ZEN = 0.
           IF (M41 EQ 1 AND M42 EQ 1) THEN BILDG = 9.
           IF (M41 EQ 1 AND (M42 EQ 2 OR M42 EQ 3)) THEN BILDG = 11.
           IF (M41 EQ 2 AND M42 EQ 1) THEN BILDG = 10.
           IF (M41 EQ 2 AND (M42 EQ 2 OR M42 EQ 3)) THEN BILDG = 12.
           IF (M41 EQ 3 AND (M42 EQ 1 OR M42 EQ 2 OR M42 EQ 3))
           THEN BILDG = 13.
           IF (M42 EQ 4) THEN BILDG = 17.
           IF (M42 EQ 5) THEN BILDG = 19.
           KOHO2 = 0.
           KOHO3 = 0.
```

[12] Beide Arten von Berufserfahrung lassen sich natürlich nur im Mehr-Episoden-Fall unterscheiden, bei dem die Variable „Berufserfahrung zu Beginn jedes Berufs" tatsächlich variiert.

```
            IF(M48 GE 468 AND M48 LE 504) THEN KOHO2 = 1.
            IF(M48 GE 588 AND M48 LE 624) THEN KOHO3 = 1.
            BERF = M50 - M43.
            BANZ = M5 - 1.
            DP = 0.
            X1 = 0.
/REGRESS DEPENDENT IS DP.
         PARAMETERS ARE 8.
         PRINT IS 0.
         MEANSQUARE IS 1.0.
         ITERATIONS ARE 100.
         LOSS.
/PARAMETER INITIAL ARE -3.8,-0.01,0.04,-0.01,0.15,-0.01,0.1,0.17.
           NAMES ARE KONST,VDABH,BILDG,M59,BANZ,BERF,KOHO2,KOHO3.
/END.
/COMMENT'
M 2
T 70 63
D 71
C 64
I  6 65 59 67 66 68 69
```

Die Aufbereitung der Variablen im TRANSFORM-Paragraph geschieht wieder genauso wie in den Programmbeispielen 5.1 und 6.3 und ist dort bereits ausführlich beschrieben worden. Im Vergleich zu Programmbeispiel 6.11 werden jetzt allerdings 8 Koeffizienten (PARAMETERS ARE 8) geschätzt, die Regressionskonstante β_0, die 6 β-Koeffizienten für die unabhängigen Variablen und der γ_0-Koeffizient für die Verweil**dauer**-**abh**ängigkeit (VDABH). Für jeden dieser Koeffizienten wird im Paragraph PARAMETER ein Startwert vorgegeben. Diese Werte lehnen sich wieder an die bereits in Kapitel 5 durchgeführten Cox-Schätzungen an.

Auch die Parameter für das Unterprogramm P3RFUN bleiben im Vergleich zum Programmbeispiel 6.11 wieder bis auf die I-Karte, auf der die BMDP-internen Nummern der unabhängigen Variablen angeführt werden müssen, unverändert. Dort wird zuerst die Anzahl der Kovariablen, in diesem Beispiel die Zahl 6, eingesetzt, und dann folgen die Kovariablen mit ihren BMDP-internen Nummern: 65 (≙ BILDG), 59 (≙ M59, Prestige), 67 (≙ BANZ), 66 (≙ BERF), 68 (≙ KOHO2) und 69 (≙ KOHO3). Das Resultat der Schätzung ist in Tabelle 6.11 dargestellt.

Bei dem Modell in Tabelle 6.11 erhält man einen Wert der Log-Likelihood-Funktion (≙ -LOSS) von -13854,60. Vergleicht man dieses Modell mit dem Gompertz-Modell ohne Kovariablen, dann ergibt sich auf der Basis des Likelihood-Quotienten-Tests (vgl. Abschnitt 3.7) ein χ^2-Wert von

$$Lq = 2(-13854{,}60 - (-14147{,}10)) = 585{,}00$$

bei sechs Freiheitsgraden. Die Nullhypothese, keine der zusätzlich eingeführten Kovariablen erkläre etwas am Berufswechselrisiko der Männer, muß damit abgelehnt werden.

Tabelle 6.11: Ergebnis des Gompertz-Modells aus Programmbeispiel 6.12

PARAMETER	ESTIMATE	ASYMPTOTIC STANDARD DEVIATION
KONST	-3.651509	0.127242
VDABH	-0.007335	0.000397
BILDG	0.003205	0.013920
M59	-0.004916	0.001314
BANZ	0.155978	0.009915
BERF	-0.008679	0.000416
KOHO2	0.037889	0.046693
KOHO3	0.167833	0.052601

Die Signifikanzprüfung der einzelnen Koeffizienten erbringt wieder das bereits bekannte Bild: Die Variablen BILDG und KOHO2 sind nicht signifikant, und die Variablen Prestige (M59), BANZ, BERF und KOHO3 haben einen signifikanten Effekt mit den üblichen Vorzeichen (Signifikanzniveau: 0,05).
Wichtig ist, daß auch der $\hat{\gamma}_0$-Koeffizient (VDABH) signifikant von Null verschieden ist und ein negatives Vorzeichen trägt. Also auch bei Kontrolle von relevanten Kovariablen bleibt der monoton fallende Hazardratenverlauf erhalten. Dies spricht für die Hypothese der mobilitätshemmenden Wirkung der berufsspezifischen Qualifikation. Vergleicht man den unstandardisierten Koeffizienten der Berufserfahrung zu Beginn jeder neuen Erwerbstätigkeit ($\hat{\beta}_{BERF}$) mit dem der Berufserfahrung innerhalb des Berufes ($\hat{\gamma}_0 \triangleq$ VDABH), dann wird deutlich, daß beide in etwa die gleiche Größe haben (-0,0087 und -0,0073). Es spielt also für das Berufswechselverhalten keine große Rolle, ob man die Berufserfahrung in ein und demselben Beruf oder über mehrere Berufe erwirbt.

Die Überprüfung der Residuen im Gompertz-Modell

Die inhaltlichen Interpretationen der gerade besprochenen Gompertz-Modelle sind wiederum nur unter der Bedingung gültig, daß es sich dabei jeweils um homogene Subpopulationen handelt. Wie in den Abschnitten 3.7.1 und 6.2.3 schon ausführlich diskutiert, kann man bei Gültigkeit der Verteilungsannahme die kumulativen Hazardraten $\Lambda(t_i|\mathbf{x}_i)$ als Residuen r_i betrachten und diese zur Modellevaluation heranziehen.
Für den Fall der Gompertz-Verteilung bei mehreren Episoden erhält man eine Schätzung der Residuen r_{ik} wie folgt:

$$\hat{r}_{ik} = \hat{\Lambda}(v_{ik}|\mathbf{x}_{ik}) = \frac{\exp(\mathbf{x}'_{ik}\hat{\beta})}{\hat{\gamma}_0}(\exp(\hat{\gamma}_0 v_{ik}) - 1).$$

Aus diesen so berechneten Residuen läßt sich bei zensierten Daten die Survivorfunktion wieder mit Hilfe der Produkt-Limit-Methode schätzen, entsprechend transformieren und gegen r auftragen. Bei Richtigkeit der Annahme einer gompertzverteilten Hazardrate müßte sich dann eine Gerade mit der Steigung -1 ergeben.

Die folgenden zwei Programmbeispiele mit dem Unterprogramm P1L von BMDP zeigen die Realisierung dieses Residuen-Tests für das Gompertz-Modell ohne Kovariablen und das Gompertz-Modell, in dem der λ_0-Koeffizient mit dem Kovariablen-Vektor x log-linear verbunden wurde:

Programmbeispiel 6.13:

```
/INPUT UNIT IS 30.
       CODE IS DATA.
/VARIABLE NAMES ARE (63)DUR,(64)ZEN,(65)BILDG,(66)BERF,
                    (67)BANZ,(68)KOHO2,(69)KOHO3,(70)RESID1.
          ADD IS 8.
/TRANSFORM USE = (M3 EQ 1).
           DUR = M51 - M50 + 1.
           ZEN = 1.
           IF (M51 EQ M47) THEN ZEN = 0.
           IF (M41 EQ 1 AND M42 EQ 1) THEN BILDG = 9.
           IF (M41 EQ 1 AND (M42 EQ 2 OR M42 EQ 3)) THEN BILDG = 11.
           IF (M41 EQ 2 AND M42 EQ 1) THEN BILDG = 10.
           IF (M41 EQ 2 AND (M42 EQ 2 OR M42 EQ 3)) THEN BILDG = 12.
           IF (M41 EQ 3 AND (M42 EQ 1 OR M42 EQ 2 OR M42 EQ 3))
           THEN BILDG = 13.
           IF (M42 EQ 4) THEN BILDG = 17.
           IF (M42 EQ 5) THEN BILDG = 19.
           KOHO2 = 0.
           KOHO3 = 0.
           IF(M48 GE 468 AND M48 LE 504) THEN KOHO2 = 1.
           IF(M48 GE 588 AND M48 LE 624) THEN KOHO3 = 1.
           BERF = M50 - M43.
           BANZ = M5 - 1.
           RESID1= (EXP(-4.080915)/(-0.008132))*(EXP((-0.008132)*DUR)-1)
/FORM TIME IS RESID1.
      STATUS IS ZEN.
      RESPONSE IS 1.
/ESTIMATE METHOD IS PROD.
          PLOTS ARE LOG.
/PRINT CASES ARE 0.
/END
```

Programmbeispiel 6.14:

```
/INPUT UNIT IS 30.
       CODE IS DATA.
/VARIABLE NAMES ARE (63)DUR,(64)ZEN,(65)BILDG,(66)BERF,
                    (67)BANZ,(68)KOHO2,(69)KOHO3,(70)RESID2.
          ADD IS 8.
/TRANSFORM USE = (M3 EQ 1).
           DUR = M51 - M50 + 1.
           ZEN = 1.
           IF (M51 EQ M47) THEN ZEN = 0.
           IF (M41 EQ 1 AND M42 EQ 1) THEN BILDG = 9.
```

```
          IF (M41 EQ 1 AND (M42 EQ 2 OR M42 EQ 3)) THEN BILDG = 11.
          IF (M41 EQ 2 AND M42 EQ 1) THEN BILDG = 10.
          IF (M41 EQ 2 AND (M42 EQ 2 OR M42 EQ 3)) THEN BILDG = 12.
          IF (M41 EQ 3 AND (M42 EQ 1 OR M42 EQ 2 OR M42 EQ 3))
          THEN BILDG = 13.
          IF (M42 EQ 4) THEN BILDG = 17.
          IF (M42 EQ 5) THEN BILDG = 19.
          KOHO2 = 0.
          KOHO3 = 0.
          IF(M48 GE 468 AND M48 LE 504) THEN KOHO2 = 1.
          IF(M48 GE 588 AND M48 LE 624) THEN KOHO3 = 1.
          BERF = M50 - M43.
          BANZ = M5 - 1.
          RESID2= (EXP(-3.651509+0.003205*BILDG-0.004916*M59
                +0.155978*BANZ-0.008679*BERF+0.037889*KOHO2
                +0.167833*KOHO3)/(-0.007335))*(EXP((-0.007335)*DUR)-1)
/FORM TIME IS RESID2.
      STATUS IS ZEN.
      RESPONSE IS 1.
/ESTIMATE METHOD IS PROD.
         PLOTS ARE LOG.
/PRINT CASES ARE 0.
/END
```

Im TRANSFORM-Paragraph der obigen Programmbeispiele werden wieder zuerst aufgrund der Koeffizienten-Schätzungen die Residuen (RESID1 bzw. RESID2) berechnet. Im FORM-Paragraph werden diese neben der Zensierungsvariablen (STATUS IS ZEN) als Verweildauer eingegeben (TIME IS RESID1 bzw. TIME IS RESID2). Zur Schätzung der Survivorfunktion wird jeweils die Produkt-Limit-Methode herangezogen (METHOD IS PROD), und geplottet werden sollen schließlich die logarithmierten Survivorfunktionen (PLOTS ARE LOG).
Die Resultate dieser Läufe sind in den Abbildungen 6.8 (für das Gompertz-Modell ohne Kovariablen) und 6.9 (für das Gompertz-Modell, in dem der λ_0-Koeffizient log-linear mit dem Kovariablen-Vektor verbunden wurde) dargestellt. Für das Gompertz-Modell ohne Kovariablen zeigt sich danach bei kleinen Residuen eine relativ gute Anpassung an die Gerade (vgl. Abbildung 6.8). Nur bei den großen Residuen weicht die Schätzung stark davon ab. Durch die Aufnahme der Kovariablen kann die Annäherung an eine Gerade zwar noch einmal verbessert werden (vgl. Abbildung 6.9), aber auch hier ist insbesondere bei den größeren Residuen die Abweichung von der Geraden nicht zu übersehen. Vergleicht man dieses Resultat mit den Residuen-Plots bei den Exponential-Modellen (vgl. Abbildungen 6.5 und 6.6), dann kann man wohl davon ausgehen, daß es sich bei dem Gompertz-Modell, in dem der λ_0-Koeffizient log-linear mit den Kovariablen verbunden wurde, um ein weitaus angemesseneres Modell handelt. Allerdings ist auch an dieser Stelle wieder darauf hinzuweisen, daß es sich bei diesem Residuen-Test nicht um einen Test im strengen Sinne handelt, da die Residuen weder unabhängig noch identisch verteilt sind.

Abbildung 6.8: Residuen-Plot für das Gompertz-Modell ohne Kovariablen

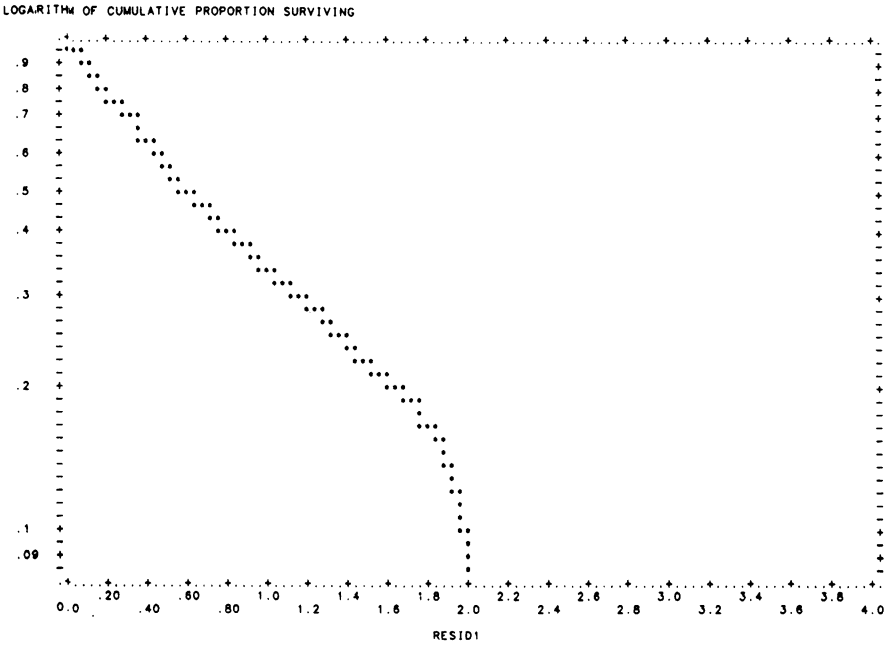
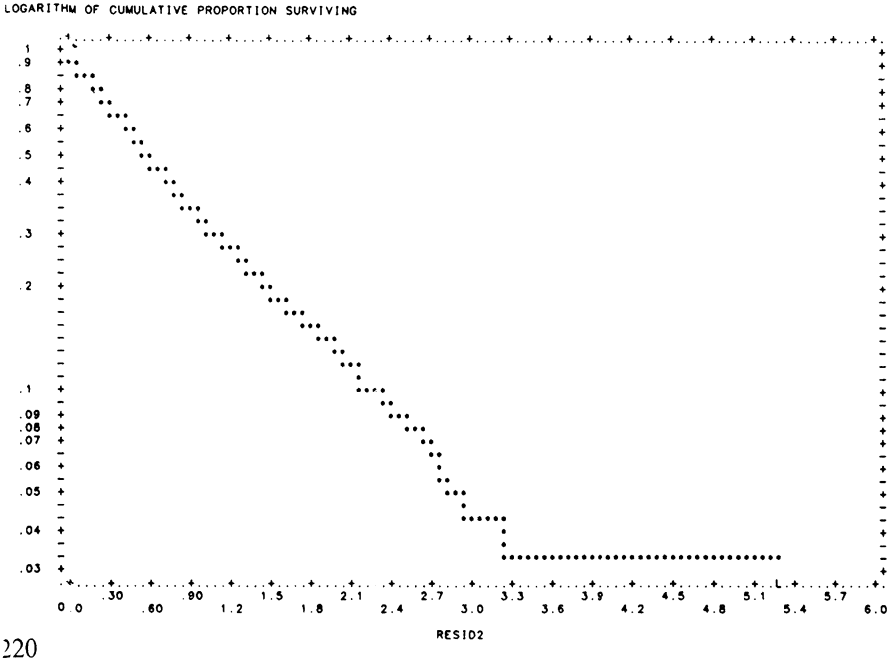


Abbildung 6.9: Residuen-Plot für das Gompertz-Modell mit Kovariablen im λ_0-Term



Das Gompertz-Modell mit Berücksichtigung von Kovariablen im λ_0- und γ_0-Term

Bereits bei den Modellen mit periodisierter Verweildauer ist diskutiert worden, daß die Wirkung zeitkonstanter Kovariablen über die Verweildauer hinweg nicht unverändert bleiben muß. Vor dem Hintergrund der Signal- oder Filtertheorie ist im Berufswechselbeispiel sogar zu erwarten, daß die Wirkung der „Signal"-Variablen mit zunehmender Verweildauer abnimmt. Die Verminderung des Einflusses solcher Variablen kann bei der Gompertz-Verteilung in das Modell aufgenommen werden, indem man zusätzlich zu $\lambda_0(x) = \exp(x'\beta)$ auch noch den Koeffizienten γ_0 in linearer Abhängigkeit vom Kovariablen-Vektor x modelliert: $\gamma_0(x) = x'\gamma$. Es ergibt sich dann das folgende Gompertz-Modell:

$$\lambda^k(v|x_k) = \exp(x_k'\beta) \cdot \exp((x_k'\gamma)v) \quad , k = 1, 2, \ldots .$$

Falls die Signaltheorie richtig ist, müßten die Koeffizienten der „Filter"-Variablen im γ_0-Term genau das entgegengesetzte Vorzeichen wie im λ_0-Term haben und ihre Wirkung mit zunehmender Verweildauer selbst mehr und mehr aufheben. Die Schätzung dieses speziellen Gompertz-Modells soll mit dem Programm RATE vorgenommen werden. Als Eingabedatensatz dient uns die ereignisorientierte Datei aus Tabelle 6.4 (ohne die Variable THEIRAT).

Programmbeispiel 6.15:

```
RUN NAME          GOMPERTZ-MODELL
N OF CASES        3516
VARIABLES         12.
TANF              1
TEND              2
ZEN               3
BILDG             4
PRES              5
BANZ              6
BERF              7
KOHO2             8
KOHO3             9
DUR               10
JOBN              11
JOBN1             12
READ DATA
(12F5.0)
T AND S           10 3
MODEL             (4) A=0 B=1 C=-1
VECTOR            (1) 4 5 6 7 8 9 (2) 4 5 6 7 8 9
SOLVE             (2)=100
FINISH
```

Im obigen RATE-Programm wird auf der MODEL-Karte ein Gompertz-Makeham-Modell spezifiziert, welches die Nummer 4 trägt: A + B exp(Ct). Der Buchstabe A wird dabei auf 0 gesetzt. Der Buchstabe B wird log-linear (B = 1)

und der Buchstabe C linear (C = -1) mit dem Kovariablen-Vektor **x** verbunden, so daß sich folgendes Modell ergibt: $\exp(\mathbf{x}'\beta)\exp((\mathbf{x}'\gamma)v)$. In die beiden Kovariablen-Vektoren werden mit der VECTOR-Karte jeweils dieselben Variablen über ihre RATE-internen Nummern aufgenommen: 4 ≙ Bildung (BILDG), 5 ≙ Prestige (PRES), 6 ≙ Anzahl der vorher bereits ausgeübten Berufe (BANZ), 7 ≙ Berufserfahrung zu Beginn der Episode (BERF), 8 ≙ Kohorten-Dummy für die 1939-41 Geborenen (KOHO2) und 9 ≙ Kohorten-Dummy für die 1949-51 Geborenen (KOHO3). Das Resultat der Schätzung ist in Tabelle 6.12 zu finden.

Tabelle 6.12: Ergebnis des Gompertz-Modells aus Programmbeispiel 6.15

DESTINATION	UNWEIGHTED FREQUENCY	WEIGHTED FREQUENCY	MAX(LOG OF L) NULL HYPOTHESIS	MAX(LOG OF L) ALTERNATIVE HYPOTHESIS	PSEUDO R-SQUARED	CHI-SQUARED	DF	PROBABILITY LEVEL
1	2586	2586.0	-1.4434320+04	-1.3843990+04	0.0409	1180.67	13	0.000+00

DESTINATION 1 LETTER B LOG-LINEAR TIME-INDEPENDENT VECTOR

INTERNAL NUMBER	VARIABLE NUMBER	VARIABLE NAME	VECTOR 1 PARAMETER	PARAMETER STANDARD ERROR	PARAMETER F RATIO	ANTILOG OF THE PARAMETER	ANTILOG STANDARD ERROR	ANTILOG F RATIO
1		(CONSTANT)	-3.8560+00	1.7040-01	512.321	2.1150-02		
2	4	BILDG	1.1840-02	1.9490-02	0.369	1.0120+00	1.9720-02	0.365
3	5	PRES	-4.3600-03	1.8790-03	5.385	9.9560-01	1.8710-03	5.409
4	6	BANZ	1.8260-01	1.6030-02	129.856	1.2000+00	1.9240-02	108.482
5	7	BERF	-9.0140-03	6.3080-04	204.174	9.9100-01	6.2520-04	206.024
6	8	KOHO2	1.4750-01	6.4590-02	5.217	1.1590+00	7.4850-02	4.510
7	9	KOHO3	2.8740-01	7.3120-02	15.452	1.3330+00	9.7470-02	11.672

DESTINATION 1 LETTER C LINEAR TIME-INDEPENDENT VECTOR

INTERNAL NUMBER	VARIABLE NUMBER	VARIABLE NAME	VECTOR 2 PARAMETER	PARAMETER STANDARD ERROR	PARAMETER F RATIO
8		(CONSTANT)	-2.8830-03	2.6560-03	1.178
9	4	BILDG	-2.0550-04	3.0360-04	0.458
10	5	PRES	-1.2660-05	2.8680-05	0.195
11	6	BANZ	-7.6440-04	2.9780-04	6.590
12	7	BERF	9.3230-06	9.2190-06	1.023
13	8	KOHO2	-2.0690-03	8.5850-04	5.811
14	9	KOHO3	-3.0810-03	1.4400-03	4.580

Anhand des Likelihood-Quotienten-Tests wird deutlich, daß das vorliegende Modell im Vergleich zum Gompertz-Modell mit Kovariablen nur im λ_0-Term signifikant besser ist. Man erhält einen signifikanten χ^2-Wert von

$$Lq = 2(-13843{,}99 - (-13854{,}60)) = 21{,}22,$$

bei sechs Freiheitsgraden ($\chi^2_{0,95;6} = 12{,}59$). Damit beschreibt das obige Modell die Verhältnisse etwas angemessener.

Eine Gegenüberstellung der $\hat{\beta}$-Koeffizienten im λ_0-Term (LETTER B) des obigen Modells mit den Ergebnissen der ersten Periode (0 bis 24 Monate) im Perioden-Modell (vgl. Tabelle 6.9) zeigt zunächst eine hohe Übereinstimmung in den unstandardisierten β-Koeffizienten. Die erste Periode im Perioden-Modell scheint deswegen im großen und ganzen gut gewählt worden zu sein. Größere Unterschiede ergeben sich außer bei der nicht-signifikanten Variablen Bildung (BILDG) nur bei der Dummy-Variablen KOHO2, die im vorliegenden Gompertz-Modell nun ebenfalls signifikant ist.

Von den $\hat{\gamma}$-Koeffizienten im γ_0-Term (LETTER C) sind nur die Variablen BANZ, KOHO2 und KOHO3 signifikant von Null verschieden. Diese Koeffizienten haben auch das erwartete entgegengesetzte Vorzeichen im Vergleich zu ihren komplementären $\hat{\beta}$-Koeffizienten im λ_0-Term (LETTER B). Dies besagt, daß mit zunehmender Verweildauer der absolute Einfluß dieser Variablen nach und nach zurückgeht.
Die Variablen PRES und BERF sind zwar im λ_0-Term (LETTER B), nicht aber im γ_0-Term (LETTER C) signifikant von Null verschieden. Damit verändert sich die Wirkung dieser Variablen über die Verweildauer hinweg nicht. Bei der Variablen Berufserfahrung war das auch zu erwarten. Die Stabilität der Wirkung der Variablen Prestige (PRES) widerspricht allerdings der Signal-Hypothese und auch den Ergebnissen aus dem Perioden-Modell (vgl. Abschnitt 6.4). Dort hatte sich ergeben, daß die Variable Prestige nur in der ersten Periode einen Effekt hat und in den darauffolgenden Perioden nicht mehr signifikant von Null verschieden ist.

Das Gompertz-Makeham-Modell mit Berücksichtigung von Kovariablen im α_0-, λ_0- und γ_0-Term

Bisher ist völlig von der Eigenschaft des Gompertz-Modells abgesehen worden, daß sich bei negativem γ_0-Koeffizienten und zunehmender Verweildauer ($v \to \infty$) die Hazardrate asymptotisch an die Zeitachse annähert (vgl. Abbildung 3.8). Bei großen Verweildauern wäre die Hazardrate dann annähernd Null, und das Eintreten eines Ereignisses wäre in diesem Modell sehr unwahrscheinlich.
Im Falle des Berufswechsels bei Männern hieße dies beispielsweise, daß für große Verweildauern in ein und demselben Beruf die Neigung zum Berufswechsel gegen Null geht und daß damit ab einem bestimmten Punkt fast kein Berufswechsel mehr eintreten kann. Es ist damit ein statistisches Modell gewählt worden, das den realen Verhältnissen nicht unbedingt entspricht. So wird man sicherlich auch noch nach sehr langen Verweildauern im Beruf Mobilitätsprozesse erwarten können. Daß das Modell mit der Gompertz-Verteilung besonders bei großen Verweildauern fehlspezifiziert ist, wurde ja auch bereits bei den Residuen-Plots in den Abbildungen 6.8 und 6.9 deutlich.
Wie in Abschnitt 3.2.2 bereits ausgeführt, läßt sich dieses Problem bei der Gompertz-Verteilung dadurch entschärfen, daß man eine Konstante α_0 ($\alpha_0 > 0$) zum Gompertz-Modell addiert und damit das sogenannte Gompertz-Makeham-Modell erhält. Bei großen Verweildauern strebt das Risiko dann nicht asymptotisch gegen Null, sondern gegen den Wert dieser Konstanten α_0. Natürlich kann auch diese Konstante neben den anderen Parametern der Gompertz-Verteilung ($\lambda_0(\mathbf{x}) = \exp(\mathbf{x}'\beta)$; $\gamma_0(\mathbf{x}) = \mathbf{x}'\gamma$) wieder in Abhängigkeit von einem Kovariablen-Vektor $\mathbf{x}$ modelliert werden. Da die Bedingung $\alpha_0 > 0$ erfüllt sein soll, wird der Kovariablen-Vektor $\mathbf{x}$ dabei mit dem Parameter α_0 log-linear verbunden: $\alpha_0(\mathbf{x}) = \exp(\mathbf{x}'\alpha)$.
Würde man im Falle des Berufswechsels von Männern ein solches Gompertz-

Makeham-Modell zur Schätzung benutzen, dann sagen die α-Koeffizienten des α_0-Terms etwas darüber aus, wie langfristig die Kovariablen auf das Berufswechselverhalten wirken. Geht man wieder von der Signaltheorie aus, dann dürften die „Filter"-Variablen wie das Bildungsniveau, das Prestige oder die bereits vor dem Eintritt in einen Beruf erworbene Berufserfahrung im α_0-Term nicht signifikant von Null verschieden sein. Einen langfristigen Einfluß auf das Berufswechselverhalten würde man vielmehr von relativ stabilen Persönlichkeitsvariablen erwarten. Leider liegen in der Lebensverlaufsstudie keine psychologischen Indikatoren vor, mit denen sich eine Person als prinzipiell „eher stabil" oder „eher instabil" kennzeichnen ließe. Man muß sich deswegen wieder mit Proxy-Variablen behelfen. Als eine Proxy-Variable für die generelle Tendenz einer Person zu eher stabilen beziehungsweise eher instabilen Arbeitsverhältnissen könnte in diesem Fall die Zahl der bereits vor dem entsprechenden Beruf ausgeübten Tätigkeiten dienen (BANZ). Je größer diese Zahl ist, desto eher wird man bei einer Person von einer Neigung zu eher instabilen Arbeitsverhältnissen sprechen können. Um diese Hypothese und die Signal-Hypothese noch einmal zu überprüfen, schätzen wir mit dem Programm RATE das folgende Gompertz-Modell:

$$\lambda^k(v|x_k) = \exp(x_k'\alpha) + \exp(x_k'\beta) \cdot \exp((x_k'\gamma)\, v) \quad \text{mit } k = 1, 2, \ldots .$$

Programmbeispiel 6.16:

```
RUN NAME        GOMPERTZ - MAKEHAM - MODELL
N OF CASES      3516
VARIABLES       12
TANF            1
TEND            2
ZEN             3
BILDG           4
PRES            5
BANZ            6
BERF            7
KOHO2           8
KOHO3           9
DUR             10
JOBN            11
JOBN1           12
READ DATA
(12F5.0)
T AND S         10 3
MODEL           (4) A=1 B=1 C=-1
VECTOR          (1) 4 5 6 7 8 9 (2) 4 5 6 7 8 9 (3) 4 5 6 7 8 9
SOLVE           (2)=100
FINISH
```

Der Unterschied zum Programmbeispiel 6.15 besteht im obigen RATE-Beispiel darin, daß auf der MODEL-Karte zusätzlich auch der Buchstabe A log-linear mit dem Kovariablen-Vektor verbunden wird (A = 1) und daß auf der VEC-

TOR-Karte für diesen zusätzlichen Kovariablen-Vektor die RATE-internen Nummern der entsprechenden Kovariablen eingetragen werden. Das Ergebnis dieses Programms ist in Tabelle 6.13 zu finden.

Wenn man dieses Gompertz-Makeham-Modell zunächst mit dem Gompertz-Modell mit Kovariablen im λ_0-Term und γ_0-Term vergleicht, dann erhält man auf der Basis des Likelihood-Quotienten-Tests einen χ^2-Wert von

$$Lq = 2(-13829{,}68) - (-13843{,}99) = 28{,}62$$

bei sechs Freiheitsgraden. Da dieser Wert größer ist als der Wert von $\chi^2_{0,95;6} = 12{,}59$, muß demnach mindestens einer der α-Koeffizienten im α_0-Term eine signifikante Wirkung haben, und es bestätigt sich die Vermutung, daß mit dem vorliegenden Gompertz-Makeham-Modell das Berufswechselrisiko bei großen Verweildauern etwas adäquater modelliert werden kann. Dies wird auch im Residuenplot in Abbildung 6.10 deutlich sichtbar[13].

---

[13] PIL-Programmlauf zum Test der Residuen des Gompertz-Makeham-Modells, in dem die Kovariablen im α_0-, λ_0- und γ_0-Term berücksichtigt wurden:

```
/INPUT UNIT IS 30.
      CODE IS DATA.
/VARIABLE NAMES ARE (63)DUR,(64)ZEN,(65)BILDG,(66)BERF,
                    (67)BANZ,(68)KOHO2,(69)KOHO3,(70)RESID,
                    (71)A,(72)B,(73)C.
          ADD IS 11.
/TRANSFORM USE = (M3 EQ 1).
           DUR = M51 - M50 + 1.
           ZEN = 1.
           IF (M51 EQ M47) THEN ZEN = 0.
           IF (M41 EQ 1 AND M42 EQ 1) THEN BILDG = 9.
           IF (M41 EQ 1 AND (M42 EQ 2 OR M42 EQ 3)) THEN BILDG = 11.
           IF (M41 EQ 2 AND M42 EQ 1) THEN BILDG = 10.
           IF (M41 EQ 2 AND (M42 EQ 2 OR M42 EQ 3)) THEN BILDG = 12.
           IF (M41 EQ 3 AND (M42 EQ 1 OR M42 EQ 2 OR M42 EQ 3))
           THEN BILDG = 13.
           IF (M42 EQ 4) THEN BILDG = 17.
           IF (M42 EQ 5) THEN BILDG = 19.
           KOHO2 = 0.
           KOHO3 = 0.
           IF(M48 GE 468 AND M48 LE 504) THEN KOHO2 = 1.
           IF(M48 GE 588 AND M48 LE 624) THEN KOHO3 = 1.
           BERF = M50 - M43.
           BANZ = M5 -1.
           A = EXP(-7.285 + 0.05935*BILDG - 0.0004507*M59 +
                  0.1055*BANZ + 0.0008332*BERF - 0.02829 * KOHO2 -
                  12.31*KOHO3).
           B = EXP(-3.836 + 0.01899*BILDG - 0.006956*M59 +
                  0.2321*BANZ - 0.01212*BERF + 0.1538*KOHO2 +
                  0.3841*KOHO3).
           C =    (-0.001277 - 0.0006124*BILDG + 0.00002117*M59 -
                  0.001545*BANZ + 0.000004597*BERF - 0.001397*KOHO2 -
                  0.001316*KOHO3).
           RESID = A*DUR + B/C * (EXP(C*DUR)- 1).
/FORM TIME IS RESID.
     STATUS IS ZEN.
     RESPONSE IS 1.
/ESTIMATE METHOD IS PROD.
          PLOTS ARE LOG.
/PRINT CASES ARE 0.
/END
```

Tabelle 6.13: Ergebnis des Gompertz-Makeham-Modells aus Programmbeispiel 6.16

DESTINATION	UNWEIGHTED FREQUENCY	WEIGHTED FREQUENCY	MAX(LOG OF L) NULL HYPOTHESIS	MAX(LOG OF L) ALTERNATIVE HYPOTHESIS	PSEUDO R-SQUARED	CHI-SQUARED	DF	PROBABILITY LEVEL
1	2586	2586.0	-1.4434320D+04	-1.3829680D+04	0.0419	1209.29	20	0.000D+00

DESTINATION 1 LETTER A LOG-LINEAR TIME-INDEPENDENT VECTOR

INTERNAL NUMBER	VARIABLE NUMBER	VARIABLE NAME	VECTOR 1 PARAMETER	PARAMETER STANDARD ERROR	PARAMETER F RATIO	ANTILOG OF THE PARAMETER	ANTILOG STANDARD ERROR	ANTILOG F RATIO
1		(CONSTANT)	-7.285D+00	8.060D-01	81.691	6.855D-04		
2	4	BILOG	5.935D-02	6.177D-02	0.923	1.061D+00	6.555D-02	0.870
3	5	PRES	-4.507D-04	7.181D-03	0.004	9.995D-01	7.178D-03	0.004
4	6	BANZ	1.055D-01	3.880D-02	7.397	1.111D+00	4.312D-02	8.643
5	7	BERF	8.332D-04	1.768D-03	0.222	1.001D+00	1.770D-03	0.222
6	8	KOHO2	-2.829D-02	2.938D-01	0.009	9.721D-01	2.856D-01	0.010
7	9	KOHO3	-1.231D+01	3.672D+02	0.001	4.489D-06	1.648D-03	368073.525

DESTINATION 1 LETTER B LOG-LINEAR TIME-INDEPENDENT VECTOR

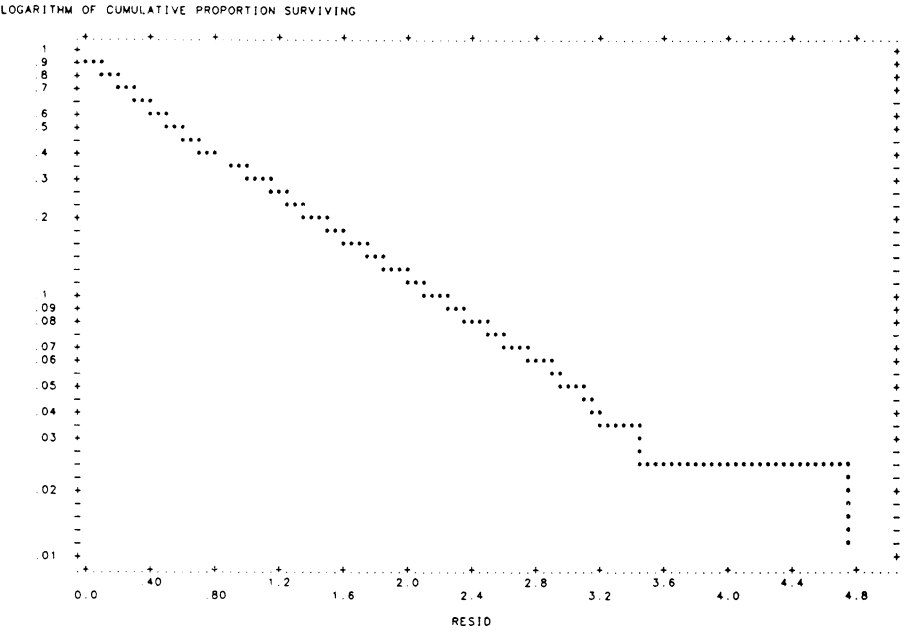
INTERNAL NUMBER	VARIABLE NUMBER	VARIABLE NAME	VECTOR 2 PARAMETER	PARAMETER STANDARD ERROR	PARAMETER F RATIO	ANTILOG OF THE PARAMETER	ANTILOG STANDARD ERROR	ANTILOG F RATIO
8		(CONSTANT)	-3.836D+00	2.060D-01	346.700	2.158D-02		
9	4	BILOG	1.899D-02	2.383D-02	0.635	1.019D+00	2.429D-02	0.623
10	5	PRES	-6.956D-03	2.429D-03	8.201	9.931D-01	2.412D-03	8.258
11	6	BANZ	2.321D-01	2.173D-02	114.044	1.261D+00	2.741D-02	90.828
12	7	BERF	-1.212D-02	1.101D-03	121.153	9.880D-01	1.088D-03	122.631
13	8	KOHO2	1.538D-01	8.277D-02	3.455	1.166D+00	9.653D-02	2.968
14	9	KOHO3	3.841D-01	8.428D-02	20.774	1.468D+00	1.237D-01	14.323

DESTINATION 1 LETTER C LINEAR TIME-INDEPENDENT VECTOR

INTERNAL NUMBER	VARIABLE NUMBER	VARIABLE NAME	VECTOR 3 PARAMETER	PARAMETER STANDARD ERROR	PARAMETER F RATIO
15		(CONSTANT)	-1.277D-03	4.084D-03	0.098
16	4	BILOG	-6.124D-04	4.543D-04	1.817
17	5	PRES	2.117D-05	4.491D-05	0.222
18	6	BANZ	-1.545D-03	5.900D-04	6.858
19	7	BERF	4.597D-06	1.724D-05	0.071
20	8	KOHO2	-1.397D-03	1.313D-03	1.133
21	9	KOHO3	-1.316D-03	1.751D-03	0.564

Die Überprüfung der Koeffizienten im α_0-Term (LETTER A) zeigt, daß tatsächlich nur die Variable Anzahl der vorher bereits ausgeübten Berufe eine Wirkung auf das Berufswechselrisiko hat. Alle anderen Kovariablen wirken sich bei großen Verweildauern nicht mehr signifikant aus. Dies spricht für die Signaltheorie, die davon ausgeht, daß die Wirkung der „Filter"-Variablen zu Beginn jedes neuen Berufs groß ist und mit zunehmender Verweildauer langsam schwindet, und für die Vermutung, daß langfristig nur relativ stabile Persönlichkeitsvariablen noch ausschlaggebend sind.
Im γ_0-Term (LETTER C) hat nur die Variable Anzahl der vorher bereits ausgeübten Berufe (BANZ) eine signifikante Wirkung. Der $\hat{\gamma}$-Koeffizient hat dabei das entgegengesetzte Vorzeichen wie im λ_0-Term (LETTER B). Das heißt, bei kurzen und mittleren Verweildauern im Beruf nimmt der Einfluß der Variablen Anzahl der vorher bereits ausgeübten Berufe nach und nach ab.
Signifikante Effekte nur im λ_0-Term (LETTER B) haben schließlich die Kovariablen Prestige (PRES), Berufserfahrung zu Beginn jeder neuen Episode (BERF) und KOHO3. Bei kurzen und mittleren Verweildauern bleibt ihr Effekt damit auf das Berufswechselrisiko weitgehend erhalten. Dies spricht allerdings gegen die Filtertheorie, nach der auch die Wirkung der Variablen Prestige und Berufserfahrung vor Eintritt in den jeweiligen Beruf nachlassen sollte.

Abbildung 6.10: Residuen-Plot für das Gompertz-Makeham-Modell, in dem der α_0-, der λ_0- und der γ_0-Term mit dem Kovariablen-Vektor x verbunden wurden



Insgesamt ist anhand des Berufswechselbeispiels deutlich geworden, daß durch die sukzessive Aufnahme von Kovariablen in die Parameter der Gompertz-(Makeham-)Verteilung zwar ein zunehmend komplexeres, aber auch realitätsgerechteres Modell für die Determinanten der Berufswechselprozesse gewonnen werden kann.

Sicherlich müßte man im Anschluß an dieses Modell in einem weiteren Schritt alle nicht-signifikanten Kovariablen aus den entsprechenden Kovariablen-Vektoren im RATE-Programm herausnehmen, das Modell neu schätzen und damit zu einem genauso guten, aber parametersparsameren Modell kommen. An dieser Stelle soll aber auf diesen Schritt verzichtet werden. Wir wenden uns statt dessen einem speziellen Gompertz-Modell zu, mit dem der Versuch gemacht werden soll, den Einfluß der stetigen zeitveränderlichen Kovariablen Berufserfahrung von der reinen Verweildauerabhängigkeit zu trennen.

Das Gompertz-Modell mit einer stetigen zeitveränderlichen Kovariablen

Mit dem Gompertz-Modell in Programmbeispiel 6.12 ist versucht worden, die kumulierte Berufserfahrung bis zum Zeitpunkt t in zwei Komponenten aufzuteilen: einerseits in die allgemeine Berufserfahrung ($x_{BERF}(t_{k-1})$), die jemand mit-

bringt, wenn er in einen neuen Beruf eintritt, und andererseits in die berufsspezifische Erfahrung $x_{BERF}(v)$, die jemand innerhalb des jeweiligen Berufs akkumuliert:

$$\dot{x}_{BERF}(t) = \beta_{BERF} x_{BERF}(t_{k-1}) + \gamma_0 x_{BERF}(v)$$

Alle Effekte, die sich mit zunehmender Verweildauer im Beruf ($x_{DUR}(v) = t - t_{k-1}$) ergeben, sind damit automatisch mit der Variablen $x_{BERF}(v)$ gleichgesetzt worden, und man kann die Gleichung wie folgt umschreiben:

$$\dot{x}_{BERF}(t) = \beta_{BERF} x_{BERF}(t_{k-1}) + \gamma_o x_{DUR}(v)$$

Ein solches Modell ist natürlich nur dann angemessen, wenn über die Verweildauer hinweg keine anderen Effekte als die Berufserfahrung innerhalb eines Berufes wirksam werden und die Berufserfahrung innerhalb eines Berufes $x_{BERF}(v)$ tatsächlich mit der Verweildauerabhängigkeit $x_{DUR}(v)$ gleichgesetzt werden kann. Dies ist allerdings nicht plausibel, da neben den berufsspezifischen Humankapitalinvestitionen noch andere Einflüsse wirksam werden können. Zum Beispiel kann man sich vorstellen, daß durch Gewöhnungsprozesse, den Aufbau sozialer Kontaktnetze in einer bestimmten Arbeitsumgebung usw. Effekte entstehen, die das Berufswechselverhalten von Arbeitnehmern zusätzlich hemmen; oder daß mit zunehmender Verweildauer innerhalb eines Berufes und der damit verbundenen wachsenden Erkenntnis, daß ein weiteres berufliches Fortkommen unter den gegebenen Bedingungen nicht mehr möglich ist, die Neigung zum Berufswechsel erhöht wird.
In dem folgenden Modell soll deswegen der „reine" Effekt der Berufserfahrung von dem „reinen" Verweildauereffekt im Beruf getrennt werden. Dabei gehen wir von einer Berufserfahrung aus, bei der es nicht darauf ankommt, ob die Kenntnisse innerhalb eines Berufes oder über mehrere Berufe hinweg erworben wurden:

$$x_{BERF}(t) = x_{BERF}(t_{k-1}) + x_{BERF}(v),$$

und kommen zu folgendem Gompertz-Modell:

$$\lambda^k(v|\mathbf{x}_k) = \exp(\mathbf{x}_k'\beta + \beta_{BERF} x_{BERF}(t)) \exp(\gamma_0 x_{DUR}(v)) \qquad \text{mit } k = 1, 2, \ldots .$$

Unterstellt man ferner, daß sich die Berufserfahrung und die anderen Einflüsse, die sich mit zunehmender Verweildauerabhängigkeit ergeben, entsprechend der Verweildauer v erhöhen:

$$x_{BERF}(t) = x_{BERF}(t_{k-1}) + v$$

und

$$x_{DUR}(v) = v,$$

dann läßt sich das obige Gompertz-Modell wie folgt umschreiben

$$\lambda^k(v|\mathbf{x}_k) = \exp(\mathbf{x}_k'\beta + \beta_{BERF}(x_{BERF}(t_{k-1}) + v)) \exp(\gamma_0 v) \qquad \text{mit } k = 1, 2, \ldots$$

und mit dem Sondermodell M 3 des Programms P3RFUN von Trond Petersen schätzen[14]:

Programmbeispiel 6.17:

```
/INPUT UNIT IS 30.
       CODE IS DATA.
/VARIABLE NAMES ARE (63)DUR,(64)ZEN,(65)BILDG,(66)BERFA,(67)BANZ,
                    (68)KOHO2,(69)KOHO3,(70)X1,(71)DP,(72)BERFE.
          ADD IS 10.
/TRANSFORM USE = (M3 EQ 1).
           DUR = M51 - M50 + 1.
           ZEN = 1.
           IF (M51 EQ M47) THEN ZEN = 0.
           IF (M41 EQ 1 AND M42 EQ 1) THEN BILDG = 9.
           IF (M41 EQ 1 AND (M42 EQ 2 OR M42 EQ 3)) THEN BILDG = 11.
           IF (M41 EQ 2 AND M42 EQ 1) THEN BILDG = 10.
           IF (M41 EQ 2 AND (M42 EQ 2 OR M42 EQ 3)) THEN BILDG = 12.
           IF (M41 EQ 3 AND (M42 EQ 1 OR M42 EQ 2 OR M42 EQ 3))
           THEN BILDG = 13.
           IF (M42 EQ 4) THEN BILDG = 17.
           IF (M42 EQ 5) THEN BILDG = 19.
           KOHO2 = 0.
           KOHO3 = 0.
           IF(M48 GE 468 AND M48 LE 504) THEN KOHO2 = 1.
           IF(M48 GE 588 AND M48 LE 624) THEN KOHO3 = 1.
           BERFA = M50 - M43.
           BERFE = BERFA + DUR.
           BANZ = M5 - 1.
           DP = 0.
           X1 = 0.
/REGRESS DEPENDENT IS DP.
         PARAMETERS ARE 8.
         PRINT IS 0.
         MEANSQUARE IS 1.0.
         ITERATIONS ARE 100.
         LOSS.
/PARAMETER INITIAL ARE -3.8,0.001,-0.01,0.04,-0.01,0.15,0.1,0.17.
           NAMES ARE KONST,VDABH,BERF,BILDG,M59,BANZ,KOHO2,KOHO3.
/END.
/COMMENT'
M 3
T 70 63
D 71
C 64
I  5 65 59 67 68 69
L 66 72
```

Wie im Programmbeispiel 6.12 werden im obigen Programmlauf zuerst im TRANSFORM-Paragraph die Variablen aufbereitet (vgl. Anhang 1). Da im Unterschied zu diesem Programm neben der Verweildauerabhängigkeit (VDABH) die Berufserfahrung nicht als zeitkonstante, sondern als zeitverän-

[14] β_{BERF} und γ_0 lassen sich natürlich nur im Mehrepisodenfall identifizieren, da sonst $x_{BERF}(t) = x_{DUR}(v) = t-t_{k-1} = v$ ist.

Tabelle 6.14: Ergebnis des Gompertz-Modells mit stetiger zeitveränderlicher Berufserfahrungsvariablen aus Programmbeispiel 6.17

PARAMETER	ESTIMATE	ASYMPTOTIC STANDARD DEVIATION
KONST	-3.651521	0.127242
VDABH	0.001344	0.000579
BERF	-0.008679	0.000416
BILDG	0.003206	0.013920
M59	-0.004916	0.001314
BANZ	0.155966	0.009915
KOHO2	0.037890	0.046693
KOHO3	0.167832	0.052601

derliche Variable aufgenommen werden soll und man dazu für jede Berufsepisode die Berufserfahrung zu Beginn (BERFA) und am Ende (BERFE) benötigt, werden diese Variablen zusätzlich im TRANSFORM-Paragraph erzeugt.
Wie im Programmbeispiel 6.12 werden wieder 8 Parameter (PARAMETERS ARE 8) geschätzt. Bei PARAMETER INITIAL ARE muß allerdings eine Umordnung der Startwerte erfolgen, da der Startwert für den β-Koeffizienten der stetigen zeitveränderlichen Berufserfahrungsvariablen BERF jetzt nach der Regressionskonstanten (KONST) und der Verweildauerabhängigkeit (VDABH) an der dritten Stelle einzugeben ist.
Nach der END-Karte werden wieder die Parameter für das Unterprogramm P3RFUN eingegeben. Es handelt sich diesmal um das Modell mit der Nummer 3. Auf der I-Karte (independent) werden im Vergleich zu Programmbeispiel 6.12 aber nur die zeitkonstanten fünf Kovariablen aufgeführt. Die zeitveränderliche Variable Berufserfahrung, repräsentiert durch BERFA und BERFE, muß dagegen auf der L-Karte (labor force) mit den BMDP-internen Nummern spezifiziert werden: 66 = BERFA und 72 = BERFE. Das Ergebnis der Schätzung ist in Tabelle 6.14 zu finden.
Obwohl man mit dem Gompertz-Modell in Tabelle 6.14 den Effekt der „reinen" Verweildauerabhängigkeit vom „reinen" Effekt der Berufserfahrung trennen kann, ist es nicht möglich, dieses Modell gegen das Gompertz-Modell in Tabelle 6.11 zu testen, in das die Berufserfahrung zu Beginn jeder Berufsepisode als zeitkonstante Kovariable einging. Denn das gerade geschätzte Gompertz-Modell läßt sich auch wie folgt schreiben:

$$\lambda^k(v|x_k) = \exp(x_k'\beta + \beta_{BERF} x_{BERF}(t_{k-1})) \exp((\beta_{BERF} + \gamma_0)\, v), \qquad k = 1, 2, \ldots,$$

wobei die Berufserfahrung zu Beginn jeder Berufsepisode den Effekt β_{BERF} und die Verweildauer den Effekt $\beta_{BERF} + \gamma_0$ hat. Dementsprechend liefert das Modell in Tabelle 6.14 einen Wert der Log-Likelihood-Funktion ($\hat{=}$ -LOSS) von -13854,60, der genau mit dem des Gompertz-Modells aus der Tabelle 6.11 übereinstimmt. Auch die unstandardisierten $\hat{\beta}$-Koeffizienten der zeitkonstanten

Kovariablen BILDG, M59 (Prestige), BANZ, KOHO2 und KOHO3 ändern sich gegenüber diesem Gompertz-Modell nicht.
Der Vorteil des gerade geschätzten Gompertz-Modells ist deswegen nur darin zu sehen, die Wirkung der „reinen" Berufserfahrung (BERF) von der „reinen" Verweildauerabhängigkeit (VDABH) interpretativ unterscheiden zu können. Da sowohl der Koeffizient der Berufserfahrung $\hat{\beta}_{BERF} = -0{,}008679$ als auch der Koeffizient der Verweildauerabhängigkeit $\hat{\gamma}_0 = 0{,}001344$ signifikant von Null verschieden sind, folgt, daß der mobilitätsmindernde Einfluß der Berufserfahrung innerhalb eines Berufes durch andere Mechanismen ($\hat{\gamma}_0 > 0$) teilweise wieder kompensiert wird. Wie oben bereits ausgeführt, könnte eine Erklärung dieser Tendenz darin liegen, daß man mit zunehmender Verweildauer in ein und demselben Beruf nach und nach erkennt, daß ein berufliches Fortkommen nur mehr durch einen Arbeitsplatzswechsel zu erzielen ist.
Mit den dargestellten Beispielen zum Exponential- und Gompertz-Modell konnten natürlich nur erste Hinweise auf die Modellierungs- und Interpretationsmöglichkeiten gegeben werden, die sich bei der Anwendung von parametrischen Modellen ergeben. Die folgenden Ausführungen werden sich nun auf die Weibull- und die log-logistische Verteilung konzentrieren, die beide ebenfalls aufgrund der graphischen Tests beim Berufswechselbeispiel mit einer gewissen Berechtigung herangezogen werden können.

6.5.2 Das Weibull-Modell

Das Weibull-Modell ist sehr flexibel und bei einer Vielzahl von Situationen angemessen. Wie die Gompertz-Verteilung kann auch die Weibull-Verteilung zur Modellierung eines monoton fallenden ($0 < \alpha < 1$) oder monoton steigenden ($\alpha > 1$) Risikos herangezogen werden (vgl. Abschnitt 3.2.2). Für den Spezialfall $\alpha = 1$ erhält man die Exponential-Verteilung, und man kann insbesondere die Nullhypothese eines konstanten Ereignisrisikos gegen die Alternative $\alpha \neq 1$ testen.

Das Weibull-Modell ohne Kovariablen

Zur Verdeutlichung der Interpretation des Weibull-Modells soll zunächst wieder auf der Grundlage des Mehr-Episoden-Falls für das durchschnittliche Berufswechselrisiko der Männer ein Weibull-Modell ohne Kovariablen geschätzt werden. Dabei wird der λ-Koeffizient der Weibull-Verteilung log-linear mit einer Regressionskonstanten β_0^* verbunden und α direkt geschätzt:

$$\lambda^k(v) = \exp(\beta_0^*)^\alpha \; \alpha \; v^{\alpha-1}$$
$$= \exp(\beta_0) \; \alpha \; v^{\alpha-1} \qquad k = 1, 2, \ldots .$$
$$\text{mit } \beta_0 = \alpha\beta_0^*.$$

Die Verweildauer v soll uns in diesem Beispiel wieder als Proxy-Variable für die in jedem neuen Beruf jeweils neu zu erwerbenden berufsspezifischen Kenntnisse

und Fertigkeiten dienen. Bei Richtigkeit der Hypothese, daß die Neigung zum Berufswechsel mit zunehmender Akkumulation der berufsspezifischen Kenntnisse abnimmt, erwarten wir bei der Weibull-Verteilung ein signifikantes $\hat{\alpha}$, das zwischen 0 und 1 liegt.
Die Schätzung dieses Modells soll diesmal mit Hilfe des Programmsystems GLIM vorgenommen werden. Da man dabei auf bereits vorhandene GLIM-Makros (vgl. Roger/Peacock 1983) zurückgreifen kann, die im Anhang 5 zu finden und von jedem Benutzer schnell einzugeben sind, werden dazu nur wenige Programmzeilen benötigt.

Programmbeispiel 6.18:

```
$INPUT 40 MAKROS
UNITS 3517$
$DATA TANF TEND ZEN BILDG PRES BANZ BERF K2 K3 DUR JN JN1 THEIRAT$
$DINPUT 20$
$CALC U = %LOG(DUR)$
$CALC C = ZEN$
$USE SETW $
$END$
$STOP$
```

Im obigen GLIM-Programmbeispiel werden mit INPUT zunächst über den Eingabekanal 40 die GLIM-Makros (vgl. Anhang 5), die in einer externen Datei abgelegt und zur Schätzung des Weibull-Modells benötigt werden, geladen.
Als Eingabedatensatz dient uns wieder die bereits mehrfach verwendete ereignisorientierte Datei aus Tabelle 6.4 (vgl. Anhang 1). Die Variablen sind dort so formatiert, daß zwischen jeder Datenspalte mindestens ein Leerzeichen steht und damit formatfreies Einlesen möglich ist. Mit der DATA-Anweisung wird den Variablen-Vektoren entsprechend ihrer Stellung in der Eingabedatei ein Name zugewiesen, und die Daten werden über den Eingabekanal 20 eingelesen (DINPUT 20). Aus programmtechnischen Gründen muß die Länge der GLIM-Vektoren mit der Anweisung UNITS auf die Zahl der einzulesenden Episoden + 1, im obigen Beispiel also auf die Zahl 3516 + 1 = 3517 gesetzt werden.
Das zur Schätzung des Weibull-Modells zu verwendende Makro SETW setzt voraus, daß in den Vektor C die Zensierungsinformation ZEN und in den Vektor U die logarithmierten Verweildauern DUR übergeben wurden[15] und wird mit der Anweisung USE SETW gestartet. Das Ergebnis der Schätzung ist in Tabelle 6.15 zu finden.
Aus Tabelle 6.15 erhalten wir einen Devianz-Wert (SCALED DEVIANCE) von 6237. Die Verminderung der Devianz durch das Schätzen zusätzlicher Parame-

[15] Wenn man mit CALC U = DUR setzen würde, dann würde das Makro SETW automatisch eine Extremwert-Verteilung schätzen.

Tabelle 6.15: Ergebnis des Weibull-Modells aus Programmbeispiel 6.18

CYCLE	SCALED DEVIANCE	DF
4	6237.	3515

	ESTIMATE	S.E.	PARAMETER
1	-3.581	0.5980E-01	GM
2	0.7866	0.1231E-01	U

ter ist asymptotisch χ^2-verteilt mit k Freiheitsgraden (k $\triangleq$ Anzahl der zusätzlich gefitteten Parameter).
Ein Vergleich der Maximum-Likelihood-Schätzungen in Tabelle 6.15 $\hat{\lambda} = \exp(-3{,}581/0{,}7866) = 0{,}0105$ und $\hat{\alpha} = 0{,}7866$ mit den Schätzungen bei den graphischen Verfahren aus Abschnitt 6.1 ($\hat{\lambda} = 0{,}0138$ und $\hat{\alpha} = 0{,}8621$) zeigt wieder, daß diese Schätzungen bereits relativ gut waren.
Prüft man die Nullhypothese eines konstanten oder monoton steigenden Ereignisrisikos $H_0 : \alpha \geq 1$ gegen die Alternativhypothese eines monoton fallenden Verlaufs der Hazardrate $H_1 : \alpha < 1$ mit der folgenden standardnormalverteilten Prüfgröße (vgl. Abschnitt 3.7.3):

$$z = \frac{\hat{\alpha} - 1}{s(\hat{\alpha})} = \frac{0{,}7866 - 1}{0{,}01231} = -17{,}34,$$

dann muß die Nullhypothese verworfen werden (Signifikanzniveau: 0,05). Es liegt also ein mit zunehmender Verweildauer monoton fallendes Berufswechselrisiko vor. Dies spricht für die Hypothese, daß mit zunehmender Verweildauer wegen der Akkumulation berufsspezifischer Kenntnisse die Neigung zum Berufswechsel nach und nach sinkt.
Vergleicht man beispielsweise einen Mann, der gerade einen Monat in einem Beruf tätig ist [$\lambda^k(1) = 0{,}0105 \cdot 0{,}7866 \cdot (0{,}0105 \cdot 1)^{-0{,}2134} = 0{,}02183$], mit einem Mann, der bereits zehn Jahre in ein und demselben Beruf arbeitet [$\lambda^k(120) = 0{,}0105 \cdot 0{,}7866 \cdot (0{,}0105 \cdot 120)^{-0{,}2134} = 0{,}00786$], dann hat sich bei dem zweiten Mann durch die Akkumulation der berufsspezifischen Kenntnisse die Neigung zum Berufswechsel um etwa 64 Prozent [$(0{,}00786 - 0{,}02183)/0{,}02183 \cdot 100\% = -63{,}99\%$] verringert. Im Vergleich dazu hatte sich bei der Gompertz-Verteilung eine Verminderung um etwa 62 Prozent ergeben (vgl. Abschnitt 6.5.1).
Über die Beziehung (3.2.20)

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

kann man wieder den Median der Verweildauer von Männern im Beruf M*, mit $S(M^*) = 0{,}5$, abschätzen:

$$S(\hat{M}^*) = \exp((-0{,}0105 \cdot \hat{M}^*)^{0{,}7866})$$
$$\hat{M}^* = 59{,}77 \text{ Monate.}$$

Der Median der Weibull-Verteilung liegt damit zwischen dem Median der Gompertz-Verteilung (mit etwa vier Jahren, vgl. Abschnitt 6.5.1) und dem

Median der Exponential-Verteilung (mit etwa fünfeinhalb Jahren, vgl. Abschnitt 6.2.1).
Darüber hinaus kann man wieder die Wahrscheinlichkeit berechnen, daß ein Mann nach einem Zeitraum von acht Jahren noch in demselben Beruf arbeitet. Sie beträgt mit $\hat{S}(96) = \exp(-(0{,}0105 \cdot 96)^{0{,}7866}) = 0{,}3656$, also 36,56 Prozent. Im Vergleich dazu hatte sich beim Gompertz-Modell in Abschnitt 6.5.1 eine Wahrscheinlichkeit von 32,43 Prozent und beim Exponential-Modell in Abschnitt 6.2.1 eine Wahrscheinlichkeit von 37,56 Prozent ergeben.
Natürlich sind diese Interpretationen wiederum nur unter der Voraussetzung richtig, daß man bei den Männern von einer homogenen Population sprechen kann. Ist dies nicht der Fall, dann kann die monoton fallende Verweildauerabhängigkeit auch das Resultat einer Aggregation über heterogene Subgruppen mit jeweils konstantem Ereignisrisiko sein (vgl. Abschnitt 3.9.1) und die Erklärung, daß das Berufswechselrisiko aufgrund wachsender berufsspezifischer Kenntnisse falle, wäre eine Fehlinterpretation. In einem weiteren Schritt sollen deswegen zusätzlich Kovariablen im Weibull-Modell berücksichtigt werden.

Das Weibull-Modell mit Kovariablen im λ-Term

Die Aufnahme von Kovariablen in das Weibull-Modell soll in der Weise geschehen, daß der Parameter λ log-linear mit dem Kovariablen-Vektor x verbunden wird (vgl. Abschnitt 3.3.2): $\lambda(\mathbf{x}) = \exp(\mathbf{x}'\beta^*)$. Das Weibull-Modell lautet dann wie folgt:

$$\lambda^k(v|\mathbf{x}_k) = \exp(\mathbf{x}_k'\beta^*)^\alpha \alpha\, v^{\alpha-1}$$
$$= \exp(\mathbf{x}_k'\beta)\alpha\, v^{\alpha-1} \qquad , k = 1, 2, \ldots$$
$$\text{mit } \beta = \alpha\beta^*.$$

In den Kovariablen-Vektor x sollen wieder die bereits bekannten Variablen Bildung (BILDG), Prestige (PRES), Anzahl der bereits vorher ausgeübten Berufe (BANZ), Berufserfahrung beim Eintritt in den Beruf (BERF) sowie die Kohorten-Dummies K2 und K3 eingehen (vgl. Anhang 1). Die berufsspezifische Erfahrung, die mit zunehmender Verweildauer in einem Beruf ansteigt, sollte sich wieder in einer monoton fallenden Neigung zum Berufswechsel ausdrücken ($0 < \alpha < 1$).
Die Schätzung der Parameter soll mit dem Programmsystem GLIM vorgenommen werden[16].

Programmbeispiel 6.19:

```
$INPUT 40 MAKROS
$UNITS 3517$
$DATA TANF TEND ZEN BILDG PRES BANZ BERF K2 K3 DUR JN JN1 THEIRAT$
$DINPUT 20$
$CALC BILDG(%NU)=0 $
$CALC PRES(%NU)=0 $
```

```
$CALC BANZ(%NU)=0 $
$CALC BERF(%NU)=0 $
$CALC K2(%NU)=0 $
$CALC K3(%NU)=0 $
$CALC U = %LOG(DUR)$
$CALC C = ZEN$
$USE SETW $
$FIT +BILDG+PRES+BANZ+BERF+K2+K3
$DIS E
$END$
$STOP$
```

---

16) Würde man dieses Weibull-Modell mit dem BMDP-Programm P3R und dem darin eingebundenen Unterprogramm P3RFUN von Trond Petersen berechnen, das von der Ratenfunktion $\lambda^k(v|x_k) = \exp(x_k'\beta)v^{\alpha^*}$ ausgeht, dann muß man die geschätzten Parameter $\hat{\alpha}^*$ und das konstante Glied $\hat{\beta}_0^1$ wie folgt transformieren, um auf die Parameter $\hat{\alpha}$ und $\hat{\beta}_0$des obigen Modells zu kommen:

$$\hat{\alpha} = \hat{\alpha}^* + 1$$

$$\hat{\beta}_0 = \hat{\beta}_0^1 - \ln\hat{\alpha}.$$

Alle anderen Regressionskoeffizienten sind in beiden Ansätzen identisch. Das dazu notwendige BMDP-Programm stimmt mit dem Programmbeispiel 6.12 bis auf die Modell-Karte überein, auf der jetzt das Modell mit der Nummer 4 (M 4) spezifiziert werden muß:

```
/INPUT UNIT IS 30.
       CODE IS DATA.
/VARIABLE NAMES ARE (63)DUR,(64)ZEN,(65)BILDG,(66)BERF,
                    (67)BANZ,(68)KOHO2,(69)KOHO3,(70)X1,(71)DP.
          ADD IS 9.
/TRANSFORM USE = (M3 EQ 1).
           DUR = M51 - M50 + 1.
           ZEN = 1.
           IF (M51 EQ M47) THEN ZEN = 0.
           IF (M41 EQ 1 AND M42 EQ 1) THEN BILDG = 9.
           IF (M41 EQ 1 AND (M42 EQ 2 OR M42 EQ 3)) THEN BILDG = 11.
           IF (M41 EQ 2 AND M42 EQ 1) THEN BILDG = 10.
           IF (M41 EQ 2 AND (M42 EQ 2 OR M42 EQ 3)) THEN BILDG = 12.
           IF (M41 EQ 3 AND (M42 EQ 1 OR M42 EQ 2 OR M42 EQ 3))
           THEN BILDG = 13.
           IF (M42 EQ 4) THEN BILDG = 17.
           IF (M42 EQ 5) THEN BILDG = 19.
           KOHO2 = 0.
           KOHO3 = 0.
           IF(M48 GE 468 AND M48 LE 504) THEN KOHO2 = 1.
           IF(M48 GE 588 AND M48 LE 624) THEN KOHO3 = 1.
           BERF = M50 - M43.
           BANZ = M5 - 1.
           DP = 0.
           X1 = 0.
/REGRESS DEPENDENT IS DP.
         PARAMETERS ARE 8.
         PRINT IS 0.
         MEANSQUARE IS 1.0.
         ITERATIONS ARE 100.
         LOSS.
/PARAMETER INITIAL ARE -3.8,-0.14,0.04,-0.01,0.15,-0.01,0.1,0.17.
           NAMES ARE KONST,VDABH,BILDG,M59,BANZ,BERF,KOHO2,KOHO3.
/END.
/COMMENT'
M 4
T 70 63
D 71
C 64
I  6 65 59 67 66 68 69
'
```

Im obigen Programmbeispiel müssen im Vergleich zu Programmbeispiel 6.18 zuerst aus programmtechnischen Gründen die (N+1)-ten Elemente (%NU) aller Variablen-Vektoren, die als Kovariablen aufgenommen werden sollen, auf Null gesetzt werden. Danach folgen nach der USE SETW-Anweisung zusätzlich die FIT- und die DIS-Anweisung. Mit der FIT-Anweisung werden die Namen der aufzunehmenden Kovariablen, jeweils versehen mit einem Plus-Zeichen, spezifiziert, und mit der DIS-Anweisung wird die Ausgabe der Schätzwerte sowie ihrer asymptotischen Standardabweichungen angefordert. Das Ergebnis der Schätzung ist in Tabelle 6.16 zu finden[17].

Aus Tabelle 6.16 ist zunächst ersichtlich, daß sich der Devianz-Wert im Vergleich zum Weibull-Modell ohne Kovariablen von 6237 auf 5632 durch die Schätzung der zusätzlichen Parameter vermindert hat. Die Verminderung von -2logL (vgl. Aitkin/Clayton 1980, S. 161) entspricht einem χ^2-Wert von 605 (bei sechs Freiheitsgraden) und besagt, daß sich der Erklärungswert des Modells durch die Einführung der Kovariablen signifikant verbessert hat.

Tabelle 6.16: Ergebnis des Weibull-Modells aus Programmbeispiel 6.19

CYCLE	SCALED DEVIANCE	DF
3	5632.	3509

	ESTIMATE	S.E.	PARAMETER
1	-3.424	0.1433	GM
2	0.8266	0.1293E-01	U
3	0.7150E-02	0.1446E-01	BILD
4	-0.4906E-02	0.1376E-02	PRES
5	0.1592	0.1222E-01	BANZ
6	-0.8552E-02	0.4506E-03	BERF
7	0.1218	0.4648E-01	K2
8	0.3325	0.5300E-01	K3

[17] Bei dem Modell M 4 von Trond Petersen kommt man bei einem Wert der Log-Likelihood-Funktion (≙ -LOSS) von -14000,8 zu folgendem Ergebnis:

PARAMETER	ESTIMATE	ASYMPTOTIC STANDARD DEVIATION
KONST	-3.667665	0.136723
VDABH	-0.161433	0.016871
BILDG	0.007218	0.012912
M59	-0.004928	0.001224
BANZ	0.160115	0.009639
BERF	-0.008542	0.000395
KOHO2	0.124740	0.040516
KOHO3	0.340973	0.049549

Sieht man einmal von den Rundungsfehlern ab, dann stimmen die $\hat{\beta}$-Koeffizienten der Kovariablen mit denen aus Tabelle 6.16 überein, und die Schätzungen aus dem Programm P3RFUN $\hat{\alpha}^*$ (VDABH) und $\hat{\beta}_0^1$ (KONST) lassen sich in die GLIM-Schätzungen $\hat{\alpha}$ (U) und $\hat{\beta}_0$ (GM) wie folgt umrechnen:

$$\hat{\alpha} = \hat{\alpha}^* + 1 = -0{,}161433 + 1 = 0{,}83856$$

$$\hat{\beta}_0 = \hat{\beta}_0^1 - \ln\hat{\alpha} = -3{,}6676 - \ln 0{,}83856 = -3{,}492.$$

Die geschätzten β-Koeffizienten stimmen in Einflußrichtung und -größe bis auf die Kohorten-Dummies und die wiederum nicht-signifikante Variable Bildung (BILD) relativ gut mit denen im Gompertz-Modell in Tabelle 6.11 überein. Unabhängig von der gewählten Verteilung (Gompertz- oder Weibull-Verteilung) würde man deswegen zu annähernd denselben inhaltlichen Schlußfolgerungen in bezug auf die Wirkungsweise der Kovariablen kommen.
Die Prüfung der Nullhypothese $H_0: \alpha \geq 1$ gegen die Alternativhypothese $H_1: \alpha < 1$ zeigt wiederum, daß auch bei Kontrolle der Kovariablen die monoton fallende Neigung zum Berufswechsel erhalten bleibt:

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

Auch in dieser Hinsicht kommt man also zur gleichen inhaltlichen Schlußfolgerung wie beim Gompertz-Modell in Tabelle 6.11.

Die Überprüfung der Residuen im Weibull-Modell

Wie in den Abschnitten 3.7.1 und 6.2.3 schon ausführlich diskutiert, kann man generell die kumulativen Hazardraten $\Lambda(t_i|\mathbf{x}_i)$ als Residuen r_i betrachten und diese zur Modellevaluation heranziehen. Im Falle der Weibull-Verteilung erhält man eine Schätzung der Residuen r_{ik} wie folgt:

$$\hat{r}_{ik} = \hat{\Lambda}(v_{ik}|\mathbf{x}_{ik}) = \exp(\mathbf{x}'_{ik}\hat{\beta})v_{ik}^{\hat{\alpha}}.$$

Aus diesen so berechneten Residuen kann man bei zensierten Daten die Survivorfunktion wieder mit Hilfe der Produkt-Limit-Methode schätzen, entsprechend transformieren und gegen r auftragen. Bei Richtigkeit der Annahme einer weibullverteilten Hazardrate müßte sich dann eine Gerade mit der Steigung -1 ergeben.
Die folgenden zwei Programmläufe mit dem Unterprogramm P1L von BMDP zeigen die Realisierung dieses Residuen-Tests für das Weibull-Modell ohne Kovariablen und das Weibull-Modell, in dem der λ-Koeffizient mit dem Kovariablen-Vektor log-linear verbunden wurde:

Programmbeispiel 6.20:

```
/INPUT UNIT IS 30.
       CODE IS DATA.
/VARIABLE NAMES ARE (63)DUR,(64)ZEN,(65)BILDG,(66)BERF,
                    (67)BANZ,(68)KOHO2,(69)KOHO3,
                    (70)RESID1,(71)ALPHA,(72)LAMBDA.
          ADD IS 10.
/TRANSFORM USE = (M3 EQ 1).
           DUR = M51 - M50 + 1.
           ZEN = 1.
           IF (M51 EQ M47) THEN ZEN = 0.
           IF (M41 EQ 1 AND M42 EQ 1) THEN BILDG = 9.
           IF (M41 EQ 1 AND (M42 EQ 2 OR M42 EQ 3)) THEN BILDG = 11.
```

```
            IF (M41 EQ 2 AND M42 EQ 1) THEN BILDG = 10.
            IF (M41 EQ 2 AND (M42 EQ 2 OR M42 EQ 3)) THEN BILDG = 12.
            IF (M41 EQ 3 AND (M42 EQ 1 OR M42 EQ 2 OR M42 EQ 3))
            THEN BILDG = 13.
            IF (M42 EQ 4) THEN BILDG = 17.
            IF (M42 EQ 5) THEN BILDG = 19.
            KOHO2 = 0.
            KOHO3 = 0.
            IF(M48 GE 468 AND M48 LE 504) THEN KOHO2 = 1.
            IF(M48 GE 588 AND M48 LE 624) THEN KOHO3 = 1.
            BERF = M50 - M43.
            BANZ = M5 -1.
            ALPHA = 0.7866.
            LAMBDA = EXP(-3.581).
            RESID1 = LAMBDA * DUR ** ALPHA.
/FORM TIME IS RESID1.
      STATUS IS ZEN.
      RESPONSE IS 1.
/ESTIMATE METHOD IS PROD.
         PLOTS ARE LOG.
/PRINT CASES ARE 0.
/END
```

Programmbeispiel 6.21:

```
/INPUT UNIT IS 30.
       CODE IS DATA.
/VARIABLE NAMES ARE (63)DUR,(64)ZEN,(65)BILDG,(66)BERF,
                    (67)BANZ,(68)KOHO2,(69)KOHO3,
                    (70)RESID2,(71)ALPHA,(72)LAMBDA.
          ADD IS 10.
/TRANSFORM USE = (M3 EQ 1).
           DUR = M51 - M50 + 1.
           ZEN = 1.
           IF (M51 EQ M47) THEN ZEN = 0.
           IF (M41 EQ 1 AND M42 EQ 1) THEN BILDG = 9.
           IF (M41 EQ 1 AND (M42 EQ 2 OR M42 EQ 3)) THEN BILDG = 11.
           IF (M41 EQ 2 AND M42 EQ 1) THEN BILDG = 10.
           IF (M41 EQ 2 AND (M42 EQ 2 OR M42 EQ 3)) THEN BILDG = 12.
           IF (M41 EQ 3 AND (M42 EQ 1 OR M42 EQ 2 OR M42 EQ 3))
           THEN BILDG = 13.
           IF (M42 EQ 4) THEN BILDG = 17.
           IF (M42 EQ 5) THEN BILDG = 19.
           KOHO2 = 0.
           KOHO3 = 0.
           IF(M48 GE 468 AND M48 LE 504) THEN KOHO2 = 1.
           IF(M48 GE 588 AND M48 LE 624) THEN KOHO3 = 1.
           BERF = M50 - M43.
           BANZ = M5 -1.
           ALPHA = 0.8266.
           LAMBDA = EXP(-3.424 + 0.007150*BILDG - 0.004906*M59 +
                    0.1592*BANZ - 0.008552*BERF + 0.1218*KOHO2 +
                    0.3325*KOHO3).
           RESID2 = LAMBDA * DUR ** ALPHA.
```

```
/FORM TIME IS RESID2.
      STATUS IS ZEN.
      RESPONSE IS 1.
/ESTIMATE METHOD IS PROD.
         PLOTS ARE LOG.
/PRINT CASES ARE 0.
/END
```

Im TRANSFORM-Paragraph der obigen Programmläufe werden wieder zuerst aufgrund der Koeffizienten-Schätzungen aus den Tabellen 6.15 und 6.16 jeweils die Residuen (RESID1 bzw. RESID2) berechnet. Im FORM-Paragraph werden diese neben der Zensierungsvariablen (STATUS IS ZEN) als Verweildauer eingegeben (TIME IS RESID1 bzw. TIME IS RESID2). Zur Schätzung der Survivorfunktionen wird jeweils die Produkt-Limit-Methode herangezogen (METHOD IS PROD), und geplottet werden sollen schließlich die logarithmierten Survivorfunktionen (PLOTS ARE LOG).
Die Ergebnisse dieser Läufe sind in den Abbildungen 6.11 (für das Weibull Modell ohne Kovariablen) und 6.12 (für das Weibull-Modell mit Kovariablen) dargestellt. Beide Plots weichen deutlich von einer Geraden mit der Steigung -1 ab, so daß die Weibull-Verteilung in diesem Fall den Verlauf des Berufswechselrisikos weit schlechter beschreibt als die Gompertz-Verteilung.

Abbildung 6.11: Residuen-Plot für das Weibull-Modell ohne Kovariablen

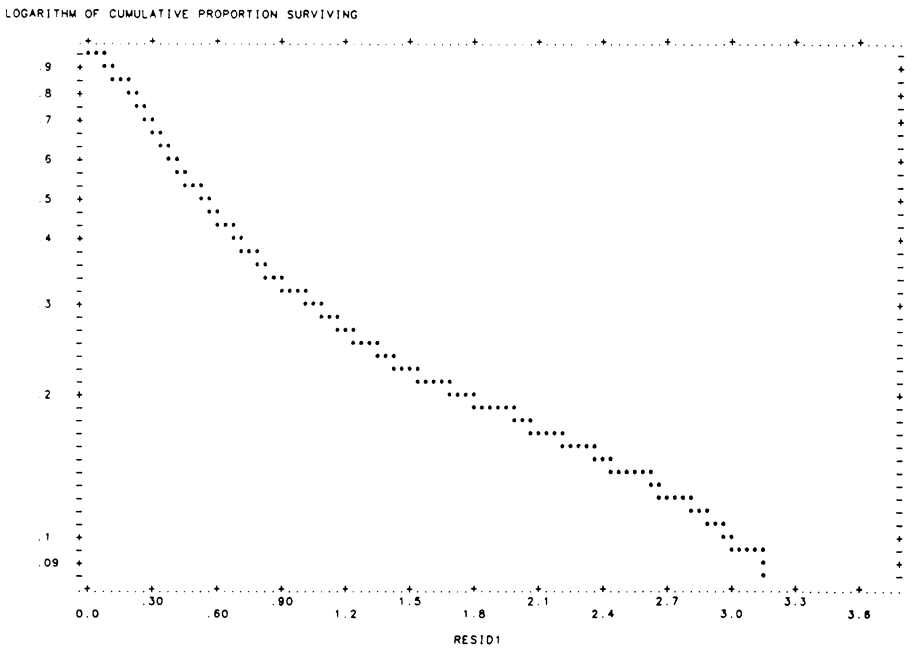
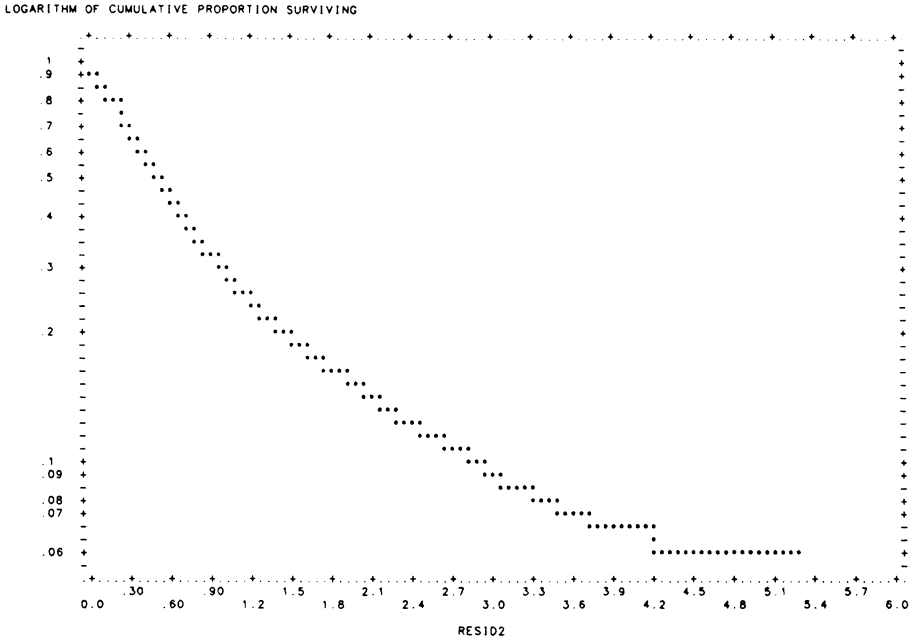


Abbildung 6.12: Residuen-Plot für das Weibull-Modell, in dem der λ-Koeffizient log-linear mit dem Kovariablen-Vektor verbunden wurde



6.5.3 Das log-logistische Modell

Die log-logistische Verteilung ist ebenso wie die Weibull-Verteilung außerordentlich flexibel und bei verschiedenen Situationen einsetzbar. Für $\alpha \leq 1$ erhält man einen monoton fallenden Verlauf der Hazardrate, und für $\alpha > 1$ steigt das Risiko zunächst bis zu einem bestimmten Zeitpunkt der Verweildauer monoton auf ein Maximum an ($v_{max} = (\alpha - 1)^{1/\alpha}/\lambda$), um danach wieder monoton zu fallen (vgl. Abbildung 3.7). In der Literatur ist das log-logistische Modell deswegen neben dem Log-Normalverteilungs-Modell (welches allerdings mathematisch nicht so einfach zu handhaben ist) die am häufigsten empfohlene Verteilung, wenn zuerst ein steigendes und dann ein fallendes Risiko vermutet werden kann[18].

[18] Zur Modellierung solcher nicht-monotonen Verläufe im Bereich der Wirtschafts- und Sozialwissenschaften ist von Diekmann und Mitter (1983, 1984) auch noch die Sichelverteilung vorgeschlagen worden.

Das log-logistische Modell ohne Kovariablen

Zur Verdeutlichung der Interpretation des log-logistischen Modells soll zunächst wieder auf der Grundlage des Mehr-Episoden-Falls für das durchschnittliche Berufswechselrisiko der Männer ein log-logistisches Modell ohne Kovariablen geschätzt werden. Dabei wird der λ-Koeffizient des log-logistischen Modells log-linear mit einer Regressionskonstanten β_0^* verbunden und der Parameter α direkt geschätzt:

$$\lambda^k(v) = \frac{\exp(\beta_0^*)^\alpha \, \alpha \, v^{\alpha-1}}{1 + \exp(\beta_0^*)^\alpha \, v^\alpha}$$

$$= \frac{\exp(\beta_0)\alpha \, v^{\alpha-1}}{1 + \exp(\beta_0)v^\alpha} \qquad , k = 1, 2, \ldots .$$

mit $\beta_0 = \alpha\beta_0^*$.

Unter inhaltlichen Gesichtspunkten kann mit diesem Modell die Hypothese eines von Beginn an mit zunehmenden berufsspezifischen Kenntnissen monoton fallenden Berufswechselrisikos ($\alpha \leq 1$) gegen die Alternativhypothese eines zuerst steigenden und dann fallenden Berufswechselrisikos getestet werden ($\alpha > 1$). Der nicht-monotone Verlauf des Berufswechselrisikos läßt sich dabei plausibel über die bei jeder Neueinstellung ablaufenden Anpassungsprozesse begründen, die insbesondere in der ersten Zeit das Berufswechselrisiko erhöhen dürften. Nachdem sich dann aber die Erwartungen der Arbeitgeber und Arbeitnehmer aneinander angepaßt haben, müßte die Neigung zum Berufswechsel mit zunehmender Akkumulation berufsspezifischen Wissens wieder fallen.
Die Schätzung dieses Modells soll wieder mit dem Programmsystem GLIM vorgenommen werden.

Programmbeispiel 6.22:

```
$INPUT 40 MAKROS
$UNITS 3517$
$DATA TANF TEND ZEN BILDG PRES BANZ BERF K2 K3 DUR JN JN1 THEIRAT$
$DINPUT 20$
$CALC U = %LOG(DUR)$
$CALC C = ZEN$
$USE SETL $
$END$
$STOP$
```

Im Unterschied zu Programmbeispiel 6.18 wird im obigen Programmbeispiel die Anweisung USE SETW nur durch die Anweisung USE SETL ersetzt[19]. Das Ergebnis der Schätzung ist in Tabelle 6.17 zu sehen.

[19] Wenn man mit CALC U = DUR setzen würde, dann würde das Makro SETL automatisch eine logistische Verteilung schätzen.

Tabelle 6.17: Ergebnis des log-logistischen Modells aus Programmbeispiel 6.22

```
        SCALED
CYCLE   DEVIANCE        DF
   4     3881.        3515

       ESTIMATE        S.E.        PARAMETER
   1   -4.464       0.7530E-01   GM
   2    1.144       0.1856E-01   U
```

Zunächst erhalten wir bei einem Wert der Devianz-Funktion von 3881 ein $\hat{\lambda} = \exp(-4{,}464/1{,}144) = 0{,}0202$ und eine Schätzung für α von 1,144. Beide Schätzungen liegen damit wieder sehr nahe bei den Schätzungen der graphischen Verfahren aus Abschnitt 6.1 ($\hat{\lambda} = 0{,}024$ und $\hat{\alpha} = 1{,}1313$).
Die Überprüfung der Nullhypothese eines monoton fallenden Berufswechselrisikos ($H_0 : \alpha \leq 1$) gegen die Alternative eines zuerst steigenden und dann fallenden Berufswechselrisikos ($H_1 : \alpha > 1$) zeigt, daß die Nullhyothese abgelehnt werden muß (Signifikanzniveau: 0,05):

$$z = \frac{1{,}144 - 1}{0{,}01856} = 7{,}759.$$

Das Berufswechselrisiko steigt also nach der Rekrutierung eines neuen Arbeitnehmers wegen der dabei ablaufenden Anpassungsprozesse an, erreicht nach etwa neun Monaten einen Höhepunkt [$v_{max} = (1{,}144 - 1)^{1/1{,}144}/0{,}0202) = 9{,}10$] und fällt danach monoton mit zunehmenden berufsspezifischen Kenntnissen ab.

Über die Beziehung (3.2.29)

$$S(t) = \frac{1}{1 + (\lambda t)^{\alpha}}$$

kann wieder der Median der Verweildauer von Männern im Beruf M*, mit $S(M^*) = 0{,}5$, abgeschätzt werden:

$$S(\hat{M}^*) = \frac{1}{1 + (0{,}0202\ \hat{M}^*)^{1{,}144}}$$

$$\hat{M}^* = 49{,}50 \text{ Monate.}$$

Der Median beträgt nach dem log-logistischem Modell also etwas mehr als vier Jahre und liegt damit in etwa auf dem Niveau der Gompertz-Verteilung (mit 49,90 Monaten, vgl. Abschnitt 6.5.1), aber unter dem Niveau der Weibull-Verteilung (etwa fünf Jahre, vgl. Abschnitt 6.5.2) und dem der Exponential-Verteilung (etwa fünfeinhalb Jahre, vgl. Abschnitt 6.2.1).
Darüber hinaus kann man wieder die Wahrscheinlichkeit berechnen, daß ein Mann nach einem Zeitraum von acht Jahren noch in ein und demselben Beruf arbeitet. Sie beträgt $\hat{S}(96) = (1 + (0{,}0202 \cdot 96)^{1{,}144})^{-1} = 0{,}3192$, also 31,92 Prozent.

Im Vergleich dazu hatte sich beim Gompertz-Modell (vgl. Abschnitt 6.5.1) eine Wahrscheinlichkeit von 32,56 Prozent, beim Weibull-Modell (vgl. Abschnitt 6.5.2) eine Wahrscheinlichkeit von 36,56 Prozent und beim Exponential-Modell (vgl. Abschnitt 6.2.1) eine Wahrscheinlichkeit von 37,56 Prozent ergeben.
Diese Interpretationen sind natürlich wieder nur unter der Voraussetzung gültig, daß es sich bei den Männern um eine homogene Population handelt und man tatsächlich von einem durchschnittlichen Berufswechselrisiko sprechen kann. In einem weiteren Schritt sollen nun wieder die bereits bekannten Kovariablen in das Modell aufgenommen werden (vgl. Anhang 1), und man kann prüfen, ob sich dadurch die Form der Verweildauerabhängigkeit ändert.

Das log-logistische Modell mit Kovariablen im λ-Term

Die Aufnahme der Kovariablen in das log-logistische Modell soll wieder in der Weise geschehen, daß der Parameter λ log-linear mit dem Kovariablen-Vektor **x** verbunden wird (vgl. Abschnitt 3.3.2): $\lambda(\mathbf{x}) = \exp(\mathbf{x}'\beta^*)$. Das log-logistische Modell lautet dann wie folgt:

$$\lambda^k(v|\mathbf{x}_k) = \frac{\exp(\mathbf{x}_k'\beta^*)^\alpha \; \alpha \; v^{\alpha-1}}{1 + \exp(\mathbf{x}_k'\beta^*)^\alpha \; v^\alpha}$$

$$= \frac{\exp(\mathbf{x}_k'\beta) \; \alpha \; v^{\alpha-1}}{1 + \exp(\mathbf{x}_k'\beta) \; v^\alpha} \qquad k = 1, 2, \ldots$$

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

In den Kovariablen-Vektor sollen wieder die Variablen Bildung (BILDG), Prestige (PRES), Anzahl der bereits vorher ausgeübten Berufe (BANZ), Berufserfahrung beim Eintritt in den Beruf (BERF) sowie die Kohorten-Dummies K2 und K3 eingehen (vgl. Anhang 1). Wenn bei jeder Neueinstellung tatsächlich Anpassungsprozesse ablaufen, die das Berufswechselrisiko zu Beginn jeder Erwerbstätigkeit erhöhen, dann müßte der nicht-monotone Verlauf des Berufswechselrisikos ($\alpha > 1$) auch noch nach Kontrolle dieser Kovariablen zu finden sein.
Die Schätzung der Parameter soll wieder mit dem Programmsystem GLIM vorgenommen werden[20].

Programmbeispiel 6.23:

```
$INPUT 40 MAKROS
$UNITS 3517$
$DATA TANF TEND ZEN BILDG PRES BANZ BERF K2 K3 DUR JN JN1 THEIRAT$
$DINPUT 20$
$CALC BILDG(%NU)=0 $
$CALC PRES(%NU)=0 $
$CALC BANZ(%NU)=0 $
$CALC BERF(%NU)=0 $
$CALC K2(%NU)=0 $
```

```
$CALC K3(%NU)=0 $
$CALC U = %LOG(DUR)$
$CALC C = ZEN$
$USE SETL $
$FIT +BILDG+PRES+BANZ+BERF+K2+K3
$DIS E
$END$
$STOP$
```

---

[20] Würde man dieses log-logistische Modell mit dem BMDP-Programm P3R und dem darin eingebundenen Unterprogramm P3RFUN von Trond Petersen berechnen, das von folgender Ratenfunktion ausgeht

$$\lambda^k(v|x_k) = \frac{\exp(x_k'\beta)\,(\alpha^* + 1)\,v^{\alpha^*}}{1 + \exp(x_k'\beta)\,v^{\alpha^*+1}},$$

dann ist der geschätzte Koeffizient $\hat{\alpha}^*$ wie folgt zu transformieren, um auf den Parameter $\hat{\alpha}$ des obigen Modells zu kommen:

$$\hat{\alpha} = \hat{\alpha}^* + 1.$$

Die $\hat{\beta}$-Koeffizienten (einschließlich des konstanten Glieds $\hat{\beta}_0$) sind in beiden Ansätzen identisch. Das dazu notwendige BMDP-Programm stimmt mit dem Programmbeispiel 6.12 bis auf die Modell-Karte überein, auf der jetzt das Modell mit der Nummer 5 (M 5) spezifiziert werden muß:

```
/INPUT UNIT IS 30.
       CODE IS DATA.
/VARIABLE NAMES ARE (63)DUR,(64)ZEN,(65)BILDG,(66)BERF,
                    (67)BANZ,(68)KOHO2,(69)KOHO3,(70)X1,(71)DP.
          ADD IS 9.
/TRANSFORM USE = (M3 EQ 1).
           DUR = M51 - M50 + 1.
           ZEN = 1.
           IF (M51 EQ M47) THEN ZEN = 0.
           IF (M41 EQ 1 AND M42 EQ 1) THEN BILDG = 9.
           IF (M41 EQ 1 AND (M42 EQ 2 OR M42 EQ 3)) THEN BILDG = 11.
           IF (M41 EQ 2 AND M42 EQ 1) THEN BILDG = 10.
           IF (M41 EQ 2 AND (M42 EQ 2 OR M42 EQ 3)) THEN BILDG = 12.
           IF (M41 EQ 3 AND (M42 EQ 1 OR M42 EQ 2 OR M42 EQ 3))
           THEN BILDG = 13.
           IF (M42 EQ 4) THEN BILDG = 17.
           IF (M42 EQ 5) THEN BILDG = 19.
           KOHO2 = 0.
           KOHO3 = 0.
           IF(M48 GE 468 AND M48 LE 504) THEN KOHO2 = 1.
           IF(M48 GE 588 AND M48 LE 624) THEN KOHO3 = 1.
           BERF = M50 - M43.
           BANZ = M5 - 1.
           DP = 0.
           X1 = 0.
/REGRESS DEPENDENT IS DP.
         PARAMETERS ARE 8.
         PRINT IS 0.
         MEANSQUARE IS 1.0.
         ITERATIONS ARE 100.
         LOSS.
/PARAMETER INITIAL ARE -3.8,1.14,0.04,-0.01,0.15,-0.01,0.1,0.17.
           NAMES ARE KONST,VDABH,BILDG,M59,BANZ,BERF,KOHO2,KOHO3.
/END.
/COMMENT'
M 5
T 70 63
D 71
C 64
I  6 65 59 67 66 68 69
```

Im obigen Programmbeispiel müssen im Vergleich zu Programmbeispiel 6.22 zuerst aus programmtechnischen Gründen die (N+1)-ten Elemente (%NU) aller Variablen-Vektoren, die als Kovariablen aufgenommen werden sollen, auf Null gesetzt werden. Danach folgen nach der USE SETL-Anweisung zusätzlich die FIT- und die DIS-Anweisung. Mit der FIT-Anweisung werden die Namen der aufzunehmenden Kovariablen, jeweils versehen mit einem Plus-Zeichen, spezifiziert, und mit der DIS-Anweisung wird die Ausgabe der Schätzwerte sowie ihrer asymptotischen Standardabweichungen veranlaßt. Das Ergebnis der Schätzung ist in Tabelle 6.18 zu finden[21].

Tabelle 6.18: Ergebnis des log-logistischen Modells aus Programmbeispiel 6.23

```
        SCALED
CYCLE   DEVIANCE        DF
   2     3294.         3509

      ESTIMATE        S.E.        PARAMETER
   1   -4.333        0.1992        GM
   2    1.234        0.1954E-01    U
   3  0.1038E-01     0.2088E-01    BILD
   4 -0.7245E-02     0.1963E-02    PRES
   5  0.2661         0.1990E-01    BANZ
   6 -0.1264E-01     0.6258E-03    BERF
   7  0.1244         0.7263E-01    K2
   8  0.3244         0.8027E-01    K3
```

Aus Tabelle 6.18 ist zunächst ersichtlich, daß sich der Devianz-Wert im Vergleich zum log-logistischen Modell ohne Kovariablen von 3881 auf 3294 durch die Schätzung der zusätzlichen Parameter vermindert hat. Die Verminderung von -2logL entspricht einem χ^2-Wert von 587 (bei sechs Freiheitsgraden) und besagt, daß sich der Erklärungswert des Modells durch die Einführung der Kovariablen signifikant verbessert hat.

---

[21] Bei dem Modell M 5 von Trond Petersen würde man bei einem Wert der Log-Likelihood-Funktion (≙ -LOSS) von 13830,3 zu folgendem Ergebnis kommen:

```
PARAMETER       ESTIMATE        ASYMPTOTIC
                              STANDARD DEVIATION

KONST          -4.338965          0.200337
VDABH           0.236721          0.021759
BILDG           0.010408          0.020212
M59            -0.007254          0.001841
BANZ            0.268493          0.017634
BERF           -0.012778          0.000553
KOHO2           0.124088          0.071432
KOHO3           0.324118          0.082421
```

Sieht man wieder von Rundungsfehlern ab, dann stimmen die $\hat{\beta}$-Koeffizienten der Kovariablen (einschließlich der Regressionskonstanten $\hat{\beta}_0$) mit denen aus Tabelle 6.18 überein, und die Schätzungen des Koeffizienten α lassen sich wie folgt ineinander umrechnen:

$$\hat{\alpha} = \hat{\alpha}^* + 1 = 0{,}236721 + 1 = 1{,}236721.$$

Obwohl sich die unstandardisierten β-Koeffizienten des log-logistischen Modells in Tabelle 6.18 beispielsweise deutlich von den β-Koeffizienten des Gompertz- (vgl. Abschnitt 6.5.1) und des Weibull-Modells (vgl. Abschnitt 6.5.2) unterscheiden, bleibt die Wirkungsrichtung der Kovariablen und deren Signifikanz unverändert.
Die Prüfung der Nullhypothese $H_0 : \alpha \leq 1$ gegen die Alternativhypothese $H_1 : \alpha > 1$ zeigt wiederum, daß auch bei Kontrolle der Kovariablen der nicht-monotone Verlauf des Berufswechselrisikos erhalten bleibt.

$$z = \frac{\hat{\alpha} - 1}{s(\hat{\alpha})} = \frac{1{,}234 - 1}{0{,}01954} = 11{,}98.$$

Dies spricht damit wieder für die Hypothese eines zunächst steigenden Berufswechselrisikos, das erst nach einer ersten Anpassungsphase in jedem neuen Beruf mit zunehmenden berufsspezifischen Kenntnissen fällt.

Die Überprüfung der Residuen im log-logistischen Modell

Auch beim log-logistischen Modell kann man, wie in den Abschnitten 3.7.1 und 6.2.3 schon ausführlich dargestellt, die kumulativen Hazardraten $\Lambda(t_i|x_i)$ als Residuen r_i betrachten und diese zur Modellevaluation heranziehen. Im Falle der log-logistischen Verteilung erhält man eine Schätzung der Residuen r_{ik} wie folgt:

$$\hat{r}_{ik} = \hat{\Lambda}(v_{ik}|x_{ik}) = \ln(1 + \exp(x'_{ik}\hat{\beta})v_{ik}^{\hat{\alpha}}).$$

Aus diesen so berechneten Residuen kann man bei zensierten Daten die Survivorfunktion wieder mit Hilfe der Produkt-Limit-Methode schätzen, transformieren und gegen r auftragen. Bei Richtigkeit der Annahme einer log-logistisch-verteilten Hazardrate müßte sich dann eine Gerade mit der Steigung -1 ergeben. Die folgenden zwei Programmläufe mit dem Unterprogramm P1L von BMDP zeigen die Realisierung dieses Residuen-Tests für das log-logistische Modell ohne Kovariablen und das log-logistische Modell, in dem der λ-Koeffizient mit dem Kovariablen-Vektor log-linear verbunden wurde.

Programmbeispiel 6.24:

```
/INPUT UNIT IS 30.
       CODE IS DATA.
/VARIABLE NAMES ARE (63)DUR,(64)ZEN,(65)BILDG,(66)BERF,
                    (67)BANZ,(68)KOHO2,(69)KOHO3,(70)RESID1.
          ADD IS 8.
/TRANSFORM USE = (M3 EQ 1).
           DUR = M51 - M50 + 1.
           ZEN = 1.
           IF (M51 EQ M47) THEN ZEN = 0.
           IF (M41 EQ 1 AND M42 EQ 1) THEN BILDG = 9.
           IF (M41 EQ 1 AND (M42 EQ 2 OR M42 EQ 3)) THEN BILDG = 11.
           IF (M41 EQ 2 AND M42 EQ 1) THEN BILDG = 10.
           IF (M41 EQ 2 AND (M42 EQ 2 OR M42 EQ 3)) THEN BILDG = 12.
           IF (M41 EQ 3 AND (M42 EQ 1 OR M42 EQ 2 OR M42 EQ 3))
```

```
            THEN BILDG = 13.
            IF (M42 EQ 4) THEN BILDG = 17.
            IF (M42 EQ 5) THEN BILDG = 19.
            KOHO2 = 0.
            KOHO3 = 0.
            IF(M48 GE 468 AND M48 LE 504) THEN KOHO2 = 1.
            IF(M48 GE 588 AND M48 LE 624) THEN KOHO3 = 1.
            BERF = M50 - M43.
            BANZ = M5 -1.
            RESID1 = LN(1+EXP(-4.464)*DUR**1.144).
/FORM TIME IS RESID1.
      STATUS IS ZEN.
      RESPONSE IS 1.
/ESTIMATE METHOD IS PROD.
          PLOTS ARE LOG.
/PRINT CASES ARE 0.
/END
```

Programmbeispiel 6.25:

```
/INPUT UNIT IS 30.
       CODE IS DATA.
/VARIABLE NAMES ARE (63)DUR,(64)ZEN,(65)BILDG,(66)BERF,
                    (67)BANZ,(68)KOHO2,(69)KOHO3,
                    (70)RESID2,(71)ALPHA,(72)LAMBDA.
          ADD IS 10.
/TRANSFORM USE = (M3 EQ 1).
            DUR = M51 - M50 + 1.
            ZEN = 1.
            IF (M51 EQ M47) THEN ZEN = 0.
            IF (M41 EQ 1 AND M42 EQ 1) THEN BILDG = 9.
            IF (M41 EQ 1 AND (M42 EQ 2 OR M42 EQ 3)) THEN BILDG = 11.
            IF (M41 EQ 2 AND M42 EQ 1) THEN BILDG = 10.
            IF (M41 EQ 2 AND (M42 EQ 2 OR M42 EQ 3)) THEN BILDG = 12.
            IF (M41 EQ 3 AND (M42 EQ 1 OR M42 EQ 2 OR M42 EQ 3))
            THEN BILDG = 13.
            IF (M42 EQ 4) THEN BILDG = 17.
            IF (M42 EQ 5) THEN BILDG = 19.
            KOHO2 = 0.
            KOHO3 = 0.
            IF(M48 GE 468 AND M48 LE 504) THEN KOHO2 = 1.
            IF(M48 GE 588 AND M48 LE 624) THEN KOHO3 = 1.
            BERF = M50 - M43.
            BANZ = M5 -1.
            ALPHA = 1.234.
            LAMBDA = EXP(-4.333 + 0.01038*BILDG -0.007245*M59 +
                     0.2661*BANZ - 0.01264*BERF +0.1244*KOHO2 +
                     0.3244*KOHO3).
            RESID2 = LN(1 + LAMBDA * DUR ** ALPHA).
/FORM TIME IS RESID2.
      STATUS IS ZEN.
      RESPONSE IS 1.
/ESTIMATE METHOD IS PROD.
          PLOTS ARE LOG.
          NO PRINT.
/PRINT CASES ARE 0.
/END
```

Im TRANSFORM-Paragraph der obigen Programmläufe werden wieder zuerst aufgrund der Koeffizienten-Schätzungen aus den Tabellen 6.17 und 6.18 jeweils die Residuen (RESID1 bzw. RESID2) berechnet. Im FORM-Paragraph werden diese neben der Zensierungsvariablen (STATUS IS ZEN) als Verweildauer eingegeben (TIME IS RESID1 bzw. TIME IS RESID2). Zur Schätzung der Survivorfunktionen wird jeweils die Produkt-Limit-Methode herangezogen (METHOD IS PROD), und geplottet werden sollen schließlich die logarithmierten Survivorfunktionen (PLOTS ARE LOG).
Die Ergebnisse dieser Tests sind in den Abbildungen 6.13 (für das log-logistische Modell ohne Kovariablen) und 6.14 (für das log-logistische Modell mit Kovariablen) dargestellt. Beide Plots zeigen Verläufe, die relativ gut mit einer Geraden mit der Steigung -1 übereinstimmen, und durch die Aufnahme von Kovariablen kann diese Anpassung sogar noch etwas verbessert werden. Vergleicht man diese Residuen-Plots mit den Residuen-Plots bei der Gompertz-Verteilung, so sieht man, daß beide Verteilungen den Prozeß des Berufswechsels zwar bei kleinen und mittleren Residuen relativ gut beschreiben, daß aber im Bereich großer Residuen noch deutliche Abweichungen bestehen. Das heißt, insbesondere bei langen Verweildauern sind beide Modelle weniger angemessen.

Abbildung 6.13: Residuen-Plot für das log-logistische Modell ohne Kovariablen

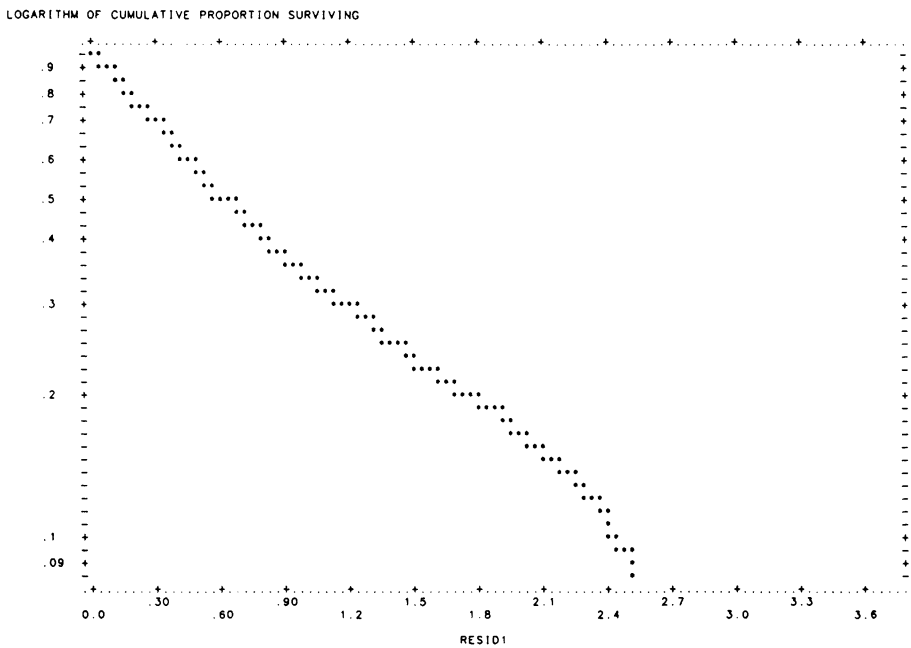
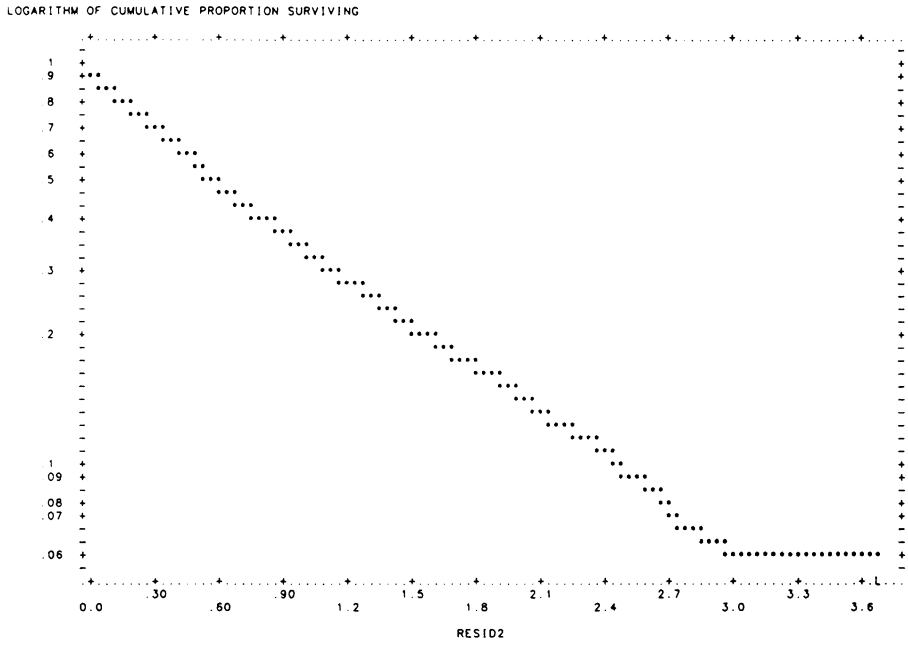


Abbildung 6.14: Residuen-Plot für das log-logistische Modell, in dem der λ-Koeffizient log-linear mit dem Kovariablen-Vektor verbunden wurde

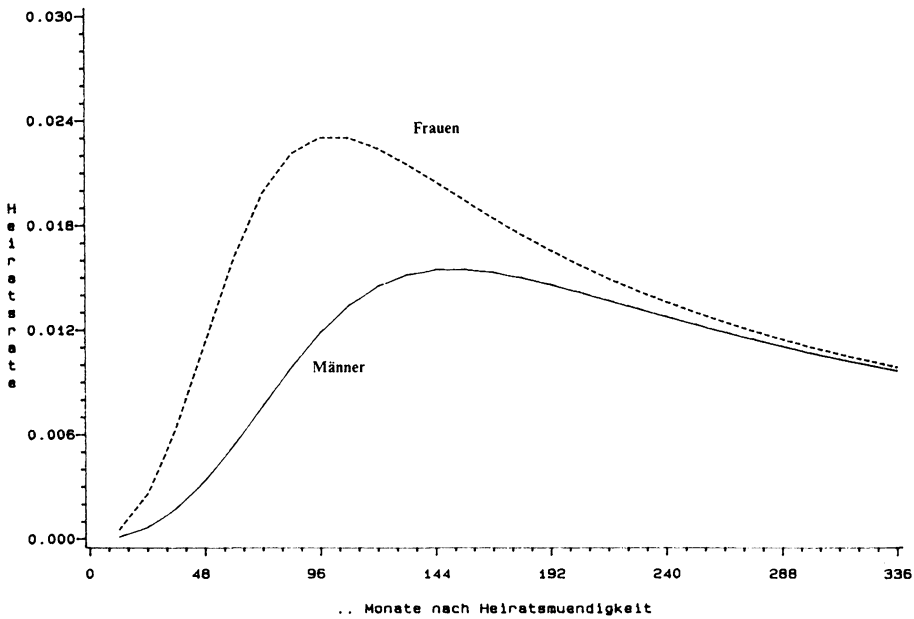


Die Analyse des Heiratsprozesses mit Hilfe eines log-logistischen Modells

Ein anderes instruktives Beispiel für die Anwendung der log-logistischen Verteilung in der Forschungspraxis liefert Papastefanou (1986). Auf der Grundlage der Lebensverlaufsstudie untersucht er, wie der Heiratsprozeß für Männer und Frauen nach dem gesetzlichen Heiratsalter verläuft. Da die Heiratsrate einen nicht-monotonen Verlauf hat, der zunächst ansteigt und dann ab einem bestimmten Punkt fällt, bietet sich die log-logistische Verteilung zur Modellierung dieses Verlaufs an. Dieser nicht-monotone Verlauf der Heiratsrate ist dabei nicht nur historisch (vgl. Papastefanou 1986) und interkulturell (vgl. Coale 1971) stabil, sondern läßt sich darüber hinaus auch suchtheoretisch begründen (vgl. Keely 1979; Hernes 1972; Sørensen/Sørensen 1984).
Das Ziel der Untersuchung von Papastefanou ist es, zu zeigen, inwieweit der Heiratsprozeß bereits durch sozialstrukturelle Merkmale (wie Bildungsniveau, berufliche Stellung, Berufserfahrung usw.) bestimmt ist, bei denen Frauen und Männer große Unterschiede aufweisen, und wie groß die dann noch verbleibende geschlechtsspezifische Differenz ist, wenn diese Hintergrundsvariablen kontrolliert sind. Nach Coleman (1984a) ließe sich dieser verbleibende Unterschied zwischen Männern und Frauen auf psychobiologische Reifungs- und

Entwicklungsprozesse zurückführen, die auf Frauen einen größeren Druck ausüben als auf Männer. Deswegen sind die Heiratschancen von Frauen aufgrund von Attraktivität und Fertilität, unabhängig von sozialstrukturellen Unterschieden wie Bildungsniveau oder berufliche Stellung, deutlicher als bei Männern auf ein spezifisches Lebenslaufsegment eingeengt.
Der Beginn der Verweildauer oder Wartezeit ist in diesem Beispiel durch den Eintritt in das gesetzlich vorgeschriebene Heiratsalter definiert. Die Verweildauer ist damit durch die Anzahl von Monaten bestimmt, die zwischen der Heiratsmündigkeit und dem Zeitpunkt der ersten Eheschließung verstreichen. Es handelt sich also um einen Ein-Episoden-Fall mit zwei Zuständen (nicht-verheiratet, verheiratet).

Abbildung 6.15: Durchschnittliche Heiratsraten von Männern und Frauen bei Kontrolle relevanter sozialstruktureller Merkmale (geschätzt mit einem log-logistischen Modell)



Die Schätzung des log-logistischen Modells erbrachte zunächst, daß der Unterschied zwischen der Heiratsrate von Männern und Frauen hoch-signifikant ist und auch noch nach Kontrolle relevanter Hintergrundsvariablen bestehen bleibt (vgl. Papastefanou 1986). Der partielle Effekt des Geschlechts auf die Heiratsrate läßt sich allerdings im log-logistischen Modell nicht unmittelbar interpretieren, da er aufgrund der log-logistischen Ratenfunktion von der Verweildauer abhängt und je nach dem Zeitpunkt der Verweildauer unterschiedlich groß ist. Dies wird in Abbildung 6.15 deutlich. Dort ist der nach dem log-logisti-

schen Modell geschätzte „durchschnittliche" Verlauf der Heiratsrate von Frauen und Männern dargestellt, wobei die relevanten Hintergrundsvariablen durch Einsetzen der jeweiligen Mittelwerte kontrolliert wurden. Es wird deutlich, daß der Unterschied zwischen Männern und Frauen zu Beginn des Prozesses ansteigt, etwa acht Jahre nach der Heiratsmündigkeit seinen Höhepunkt erreicht und mit fortschreitender Verweildauer zunehmend geringer wird. Damit ist die Neigung zur Heirat bei Frauen insbesondere in der ersten Phase des Prozesses weit größer als bei Männern, was als Hinweis auf die Hypothese von Coleman (1984a) gewertet werden kann, daß die Heiratsneigung von Frauen aufgrund psychobiologischer Reifungs- und Entwicklungsprozesse deutlicher als bei Männern auf ein spezifisches Lebenslaufsegment eingeengt ist.

6.6 Modelle mit unbeobachteter Heterogenität

Bei den bisher dargestellten Beispielen zur Anwendung parametrischer Modelle sind wir davon ausgegangen, daß alle relevanten Einflußgrößen gemessen und in die Modelle aufgenommen wurden. Die Hazardrate war vollständig durch die unabhängigen Variablen, einschließlich der als Proxy-Variable fungierenden Verweildauerabhängigkeit, determiniert, und Unterschiede in der Hazardrate konnten damit nur das Ergebnis unterschiedlicher Ausprägungen in diesen unabhängigen Merkmalen sein.
Eine solche Modellannahme ist natürlich restriktiv und insbesondere im Bereich der Wirtschafts- und Sozialwissenschaften nur selten erfüllt. In der Regel werden neben den in das Modell aufgenommenen Kovariablen weitere Merkmale, die nicht erhoben wurden oder nicht bekannt sind, die Hazardrate beeinflussen. Wird aber bei der Bildung der Hazardrate über diese unberücksichtigten Unterschiede aggregiert, so ergibt sich, wie in Abschnitt 3.9.1 bereits ausführlich dargestellt, als unangenehme Nebenfolge scheinbare Verweildauerabhängigkeit, und auf der Ebene der zu untersuchenden Hazardrate kann nicht mehr unterschieden werden, ob die Hazardrate mit zunehmender Verweildauer bei jedem Individuum in derselben Weise fällt oder ob dies bloß ein Artefakt aufgrund vernachlässigter Unterschiede zwischen Individuen ist. In den Beispielen zum Berufswechselverhalten der Männer könnte man deswegen auch argumentieren, daß die negative Verweildauerabhängigkeit nicht das Ergebnis zunehmender berufsspezifischer Humankapitalinvestitionen, sondern nur der methodenbedingte Niederschlag von nicht beobachteter Heterogenität ist.
In diesem Abschnitt wollen wir deswegen im Berufswechselbeispiel die negative Verweildauerabhängigkeit unter diesem methodischen Gesichtspunkt betrachten und unbeobachtete Heterogenität in Form eines Fehlerterms in den Regressionsmodellen zulassen. Dazu greifen wir auf das relativ einfache Modell von Tuma (1978) zurück, in dem die individuellen Hazardraten per definitionem als zeitkonstant vorausgesetzt werden. Damit wird auf der individuellen Ebene keine Verweildauerabhängigkeit zugelassen. Die Hazardraten der Individuen

können allerdings, je nach Zugehörigkeit zu relevanten Subgruppen, unterschiedliche Werte aufweisen, wobei die Verteilung dieser Werte als gammaverteilt betrachtet wird. Wie in Abschnitt 3.9.2 bereits ausgeführt, schlägt sich vernachlässigte Heterogenität in diesem Modell dann als negative Verweildauerabhängigkeit bei der erwarteten Hazardrate nieder.

Das Modell ohne beobachtete Heterogenität

Zur Verdeutlichung der Interpretation des Modells soll zunächst auf der Grundlage des Mehr-Episoden-Falls für das Berufswechselrisiko der Männer ein Modell berechnet werden, in dem nur unbeobachtete Heterogenität auftritt. Wir betrachten wie in Abschnitt 3.9.2 – allerdings hier noch ohne Kovariablen – ein Modell der Form

$$\lambda^k(v|\epsilon) = \lambda\epsilon \quad , k = 1, 2, \ldots .$$

Damit wird die Hazardrate auf der individuellen Ebene als zeitunabhängig vorausgesetzt, und für die Heterogenitätskomponente wird eine Gamma-Verteilung angenommen. In Abschnitt 3.9.2 wurde bereits abgeleitet, daß wegen $E(\epsilon) = 1$ die Gamma-Verteilung nur einen frei variierenden Parameter α besitzt (vgl. Beziehung 3.9.8), der in RATE gleich $\alpha = \exp(-\gamma_0)$ gesetzt wird. Ferner wird in RATE $\lambda = \exp(\beta_0)$ gesetzt.

Programmbeispiel 6.26:

```
RUN NAME        GAMMA - MODELL
N OF CASES      3516
VARIABLES       12
TANF            1
TEND            2
ZEN             3
BILDG           4
PRES            5
BANZ            6
BERF            7
KOHO2           8
KOHO3           9
DUR             10
JOBN            11
JOBN1           12
READ DATA
(12F5.0)
T AND S         10 3
MODEL           (2) A=1 B=1
VECTOR          (1) (2)
SOLVE
FINISH
```

Im obigen RATE-Programm wird auf der MODEL-Karte das Modell mit der Nummer (2) spezifiziert, in dem die Buchstaben A und B jeweils log-linear mit dem Kovariablen-Vektor x modelliert werden (A = 1, B = 1) A = $\exp(x'\beta)$, B =

$\exp(\mathbf{x}'\gamma)$. Da auf der VECTOR-Karte allerdings weder in den ersten (1) noch in den zweiten (2) Vektor Kovariablen aufgenommen werden, werden jeweils nur die Konstanten β_0 und γ_0 geschätzt. Das Ergebnis der Schätzung ist in Tabelle 6.19 zu finden.

Tabelle 6.19: Ergebnis des Modells mit unbeobachteter Heterogenität aus Programmbeispiel 6.26

DESTINATION	UNWEIGHTED FREQUENCY	WEIGHTED FREQUENCY	MAX(LOG OF L) NULL HYPOTHESIS	MAX(LOG OF L) ALTERNATIVE HYPOTHESIS	PSEUDO R-SQUARED	CHI-SQUARED	DF	PROBABILITY LEVEL
1	2586	2586.0	-1.4434320+04	-1.4153160+04	0.0195	562.34	1	0.000+00

DESTINATION 1 LETTER A LOG-LINEAR TIME-INDEPENDENT VECTOR

INTERNAL NUMBER	VARIABLE NUMBER	VARIABLE NAME	VECTOR 1 PARAMETER	PARAMETER STANDARD ERROR	PARAMETER F RATIO	ANTILOG OF THE PARAMETER	ANTILOG STANDARD ERROR	ANTILOG F RATIO
1		(CONSTANT)	-3.974D+00	3.481D-02	13029.359	1.880D-02		

DESTINATION 1 LETTER B LOG-LINEAR TIME-INDEPENDENT VECTOR

INTERNAL NUMBER	VARIABLE NUMBER	VARIABLE NAME	VECTOR 2 PARAMETER	PARAMETER STANDARD ERROR	PARAMETER F RATIO	ANTILOG OF THE PARAMETER	ANTILOG STANDARD ERROR	ANTILOG F RATIO
2		(CONSTANT)	-1.244D-01	5.622D-02	4.893	8.831D-01		

Aus Tabelle 6.19 erhalten wir eine Schätzung für β_0 von -3,974 und für γ_0 von -0,1244. Damit ergibt sich eine geschätzte Fehlervarianz von

$$\hat{\mathrm{Var}}(\epsilon) = \frac{1}{\hat{\alpha}} = \exp(\hat{\gamma}_0) = 0{,}833.$$

Setzt man in (3.9.12) $\phi(\mathbf{x}'\beta) = \exp(\beta_0)$ und $\alpha = \exp(-\gamma_0)$, so erhält man für die Hazardrate $\lambda(v)$ ohne Berücksichtigung der unbeobachteten Heterogenität

$$\lambda(v) = \frac{\exp(\beta_0 - \gamma_0)}{\exp(\beta_0)\, v + \exp(-\gamma_0)} = \frac{\exp(-\gamma_0)}{\exp(-\beta_0 - \gamma_0) + v}$$

$$= \frac{\exp(0.1244)}{\exp(3.974 + 0.1244) + v} = \frac{1.1325}{60.2438 + v}.$$

Obwohl die individuellen Hazardraten bei gegebener Heterogenitätskomponente als zeitunabhängig vorausgesetzt wurden, ergibt sich für die Hazardrate ohne Berücksichtigung der unbeobachteten Heterogenität mit zunehmender Verweildauer also ein monoton fallender Verlauf.

Das Regressionsmodell mit beobachteter und unbeobachteter Heterogenität

Nachdem wir mit der Fehlervarianz $\mathrm{Var}(\epsilon)$ ein Maß für die insgesamt vorhandene Heterogenität berechnet haben, kann in einem zweiten Schritt untersucht werden, wieviel Prozent der insgesamt vorhandenen Variation durch die in das Modell aufgenommenen Kovariablen (beobachtete Heterogenität) erklärt wird und welcher Anteil von unbeobachteter Heterogenität dann noch verbleibt. Dazu wird der Kovariablen-Vektor $\mathbf{x}$ wie beim Exponential-Modell log-linear mit der Konstanten λ verknüpft ($\lambda(\mathbf{x}) = \exp(\mathbf{x}'\beta)$) und die Varianz des Fehler-

terms ϵ log-linear mit dem Parameter γ_0 verbunden ($\mathrm{Var}(\epsilon) = \exp(\gamma_0)$), so daß folgende Ratengleichung formuliert werden kann

$$\lambda^k(v|x,\epsilon) = \exp(x'\beta)\epsilon \quad \text{mit } k = 1, 2, \ldots .$$

Programmbeispiel 6.27:

```
RUN NAME        GAMMA - MODELL
N OF CASES      3516
VARIABLES       12
TANF            1
TEND            2
ZEN             3
BILDG           4
PRES            5
BANZ            6
BERF            7
KOHO2           8
KOHO3           9
DUR             10
JOBN            11
JOBN1           12
READ DATA
(12F5.0)
T AND S         10 3
MODEL           (2) A=1 B=1
VECTOR          (1) 4 5 6 7 8 9 (2)
SOLVE
FINISH
```

Im Vergleich zu Programmbeispiel 6.26 werden im obigen RATE-Programmlauf auf der VECTOR-Karte im ersten Vektor die bereits bekannten Kovariablen mit ihren RATE-internen Nummern spezifiziert (vgl. Anhang 1). Das Ergebnis der Schätzung ist in Tabelle 6.20 zu finden.

Nach der Aufnahme der Kovariablen verbleibt eine Fehlervarianz von $\mathrm{Var}(\epsilon) = \exp(-0{,}4523) = 0{,}6362$. Mit Hilfe eines PRE-Maßes, mit dem die Fehlervarianz des Modells ohne Einbeziehung der Kovariablen $\mathrm{Var}(\epsilon_0)$ mit der Fehlervarianz des obigen Modells $\mathrm{Var}(\epsilon)$ in Beziehung gesetzt werden kann, ergibt sich, daß durch die Aufnahme der Kovariablen die insgesamt gemessene Fehlervarianz nur um 23,65 Prozent verringert werden kann und ein zu erklärender Rest von 76,35 Prozent übrig bleibt:

$$\mathrm{PRE} = \frac{\mathrm{Var}(\epsilon_0) - \mathrm{Var}(\epsilon)}{\mathrm{Var}(\epsilon_0)} = \frac{0{,}833 - 0{,}636}{0{,}833} = 23{,}65.$$

Dieses Ergebnis ist natürlich nicht verwunderlich, hatten wir doch bei der Gompertz-, der Weibull- und der log-logistischen Verteilung jeweils auch einen signifikanten Effekt der Verweildauerabhängigkeit bekommen.

Tabelle 6.20: Ergebnis des Modells mit unbeobachteter und beobachteter Heterogenität und aus Programmbeispiel 6.27

DESTINATION	UNWEIGHTED FREQUENCY	WEIGHTED FREQUENCY	MAX(LOG OF L) NULL HYPOTHESIS	MAX(LOG OF L) ALTERNATIVE HYPOTHESIS	PSEUDO R-SQUARED	CHI-SQUARED	DF	PROBABILITY LEVEL
1	2586	2586.0	-1.4434320+04	-1.3877530+04	0.0386	1113.59	7	0.000+00

DESTINATION 1 LETTER A LOG-LINEAR TIME-INDEPENDENT VECTOR

INTERNAL NUMBER	VARIABLE NUMBER	VARIABLE NAME	VECTOR 1 PARAMETER	PARAMETER STANDARD ERROR	PARAMETER F RATIO	ANTILOG OF THE PARAMETER	ANTILOG STANDARD ERROR	ANTILOG F RATIO
1		(CONSTANT)	-3.617D+00	1.677D-01	465.263	2.686D-02		
2	4	BILDG	6.965D-03	1.882D-02	0.137	1.007D+00	1.895D-02	0.136
3	5	PRES	-6.173D-03	1.779D-03	12.043	9.938D-01	1.768D-03	12.118
4	6	BANZ	2.193D-01	1.751D-02	156.814	1.245D+00	2.180D-02	126.442
5	7	BERF	-1.103D-02	5.797D-04	361.990	9.890D-01	5.733D-04	366.008
6	8	KOHO2	9.830D-02	6.409D-02	2.352	1.103D+00	7.071D-02	2.134
7	9	KOHO3	2.650D-01	7.074D-02	14.031	1.303D+00	9.221D-02	10.828

DESTINATION 1 LETTER B LOG-LINEAR TIME-INDEPENDENT VECTOR

INTERNAL NUMBER	VARIABLE NUMBER	VARIABLE NAME	VECTOR 2 PARAMETER	PARAMETER STANDARD ERROR	PARAMETER F RATIO	ANTILOG OF THE PARAMETER	ANTILOG STANDARD ERROR	ANTILOG F RATIO
8		(CONSTANT)	-4.523D-01	6.511D-02	48.265	6.361D-01		

Die unter inhaltlichen Gesichtspunkten entscheidende Frage allerdings, ob es sich im Berufswechselbeispiel um tatsächliche oder nur um scheinbare Verweildauerabhängigkeit handelt, kann mit dem obigen Modell mit unbeobachteter Heterogenität aber nicht geklärt werden, da dieses Modell per definitionem von konstanten Hazardraten auf der individuellen Ebene ausgeht. Hierzu müßte man auch auf der individuellen Ebene zusätzlich Verweildauerabhängigkeit zulassen, wie dies beispielsweise in der Arbeit von Heckman/Singer (1982) der Fall war. Leider existieren dazu allerdings noch keine allgemein verfügbaren Programme.

Kapitel 7:
Schlußbemerkungen

Im folgenden werden noch einige ausgewählte Problemkreise angesprochen, die in den vorangegangenen Kapiteln nur kurz oder überhaupt nicht behandelt wurden.

Ein noch nicht vollständig gelöstes Problem betrifft die *Linkszensierung*. Eine Zensierung von links liegt vor, wenn sich ein Individuum zu Beginn des Untersuchungszeitraums bereits im in Frage stehenden Zustand befindet und die Zeitspanne, die es in diesem Zustand verbracht hat, nicht bekannt ist. Linkszensierungen sind unproblematisch, wenn die untersuchten Episoden exponentialverteilt sind oder wenn aus anderen Gründen vorausgesetzt werden kann, daß die Vorgeschichte des Prozesses vor dem Beobachtungszeitraum den weiteren Verlauf des Prozesses nicht beeinflußt. In allen anderen Fällen sind zum Teil restriktive Annahmen zu treffen, um die Parameterschätzung durchführen zu können. Man vergleiche in diesem Zusammenhang auch Amemyia (1985, Kap. 11) und Flinn/Heckman (1982). Gelegentlich wird versucht, linkszensierte Daten im Rahmen von Modellen mit unbeobachteter Populationsheterogenität zu behandeln (Arminger 1984b), allerdings ist noch nicht geklärt, wie in diesem Fall die Parameterschätzung konkret durchzuführen ist. Erfolgversprechender scheint hingegen die Verwendung zeitdiskreter Modelle (vgl. Hamerle 1986a, 1986c), bei denen linkszensierte Beobachtungen und unter bestimmten Voraussetzungen auch kurzfristige Unterbrechungen während der Beobachtung eines Individuums im Modell berücksichtigt werden können.

In der vorliegenden Monographie wurden fast ausschließlich Verweildauer-Modelle mit stetig gemessener Zeit behandelt. Lediglich in Abschnitt 3.10 wurden *diskrete Hazardraten-Modelle* kurz skizziert. Ob Modelle mit stetig oder mit diskret gemessener Zeit zu verwenden sind, ist im wesentlichen eine Frage der Datenerhebung. Bei der Lebensverlaufsstudie wurden die Zeitverläufe exakt erhoben, so daß in den Kapiteln 4 bis 6 ausschließlich stetige Modelle angewendet wurden. Können aber nur Zeitintervalle angegeben werden, in denen Zustandswechsel aufgetreten oder bestimmte Ereignisse eingetreten sind, ist die Zahl gleicher Beobachtungswerte („ties") bei den gemessenen Verweildauern im allgemeinen hoch. Dies trifft insbesondere für neuere Panel-Erhebungen zu, bei denen die Vorteile des traditionellen Panels mit der retrospektiven Erhebung

von Ereignisdaten verbunden werden. Ein Beispiel einer solchen Studie ist das Sozio-ökonomische Panel des Sonderforschungsbereichs 3 (vgl. Hanefeld 1984). Hier wird versucht, wichtige Veränderungen zum Beispiel in der Erwerbstätigkeit auf Monatsebene zu registrieren. Da im Zeitablauf häufig Zustandswechsel stattfinden, ist die Anzahl der „ties" in der Regel groß. Für die Daten der ersten Welle des Sozio-ökonomischen Panels wird dies aus der Übersicht 3 in Hujer/Schneider (1986) deutlich. Man beachte, daß die Anzahl der „ties" nicht nur eine Frage der gewählten Intervallbreite ist, sondern auch durch die Häufigkeit der Übergänge im Zeitablauf beeinflußt wird. In Fällen mit einer großen Anzahl von „ties" ist es nicht gerechtfertigt, ein dynamisches Übergangsraten-Modell mit stetig gemessener Zeit zugrundezulegen, bei dem vom theoretisch-statistischen Standpunkt aus implizit die Annahme enthalten ist, daß gleiche Beobachtungswerte bei den Verweildauern die Wahrscheinlichkeit Null besitzen. In diesen Situationen ist es zweckmäßig, von vornherein ein zeitdiskretes Modell zu verwenden, das der Datenerhebung besser angepaßt ist. Eine ausführliche Beschreibung zeitdiskreter Hazardraten-Modelle findet man in Hamerle/Tutz (1986) und Hamerle (1985c, 1986a, 1986b).

In den Abschnitten 3.9 und 6.6 wurden Theorie und Anwendung von *Modellen mit unbeobachteter Populationsheterogenität* kurz dargestellt. Bei diesen Modellen existieren insbesondere auf der theoretischen Ebene noch eine Reihe offener Probleme. Einige wurden bereits in Abschnitt 3.9 angesprochen, zum Beispiel die kritische, jedoch vom theoretischen Standpunkt aus erforderliche Annahme der Unabhängigkeit zwischen Heterogenitätskomponente und beobachteten Kovariablen. In der Regel werden die nicht beobachteten Merkmale bei einem Individuum nicht unabhängig sein von den erhobenen Merkmalen und der sogenannte „omitted variables bias" kann auf diese Weise nicht beseitigt werden. Außerdem existieren *ungelöste Probleme im Mehr-Episoden-Fall* und bei der simultanen Schätzung der strukturellen Modellparameter und der Heterogenitätskomponente, zum Beispiel in bezug auf die Schätzalgorithmen und die asymptotischen Eigenschaften der Schätzungen. Hamerle (1986b) behandelt zeitdiskrete Modelle mit unbeobachteter Populationsheterogenität, bei denen diese Probleme zum Teil umgangen beziehungsweise gelöst werden können. In neuerer Zeit wurde versucht, Tests zu entwickeln, die eine Fehlspezifikation des Modells und insbesondere auch eine mögliche Vernachlässigung unbeobachteter Heterogenität überprüfen. Man vergleiche dazu Kiefer (1984), Lancaster (1985) und Arminger (1986).

Ein weiteres Problem betrifft die *statistische Analyse bivariater Prozesse*. Werden zwei parallel verlaufende Prozesse simultan untersucht, interessiert man sich gewöhnlich für die gegenseitige Beeinflussung der Prozesse, etwa für die Frage, ob die Dauer bis zum Eintreffen eines bestimmten Ereignisses in bezug auf den ersten Prozeß (z. B. Heirat) die Hazardrate des zweiten Prozesses (z. B. Jobdauer) verändert und umgekehrt. Derartige asymmetrische Fragestellungen können mit den in dieser Monographie dargestellten Verfahren behandelt werden, indem man den einen Prozeß als zeitabhängige Kovariable in den Ansatz

aufnimmt und die Hazardrate für den anderen Prozeß in Abhängigkeit davon modelliert. Man vergleiche dazu Abschnitt 3.8. Liegen hingegen die Akzente auf der simultanen Analyse und der Korrelationsstruktur zwischen den beiden Prozessen, sind andere Verfahren anzuwenden. Zu dieser Problematik liegen bisher nur wenige Resultate vor. Man vergleiche zum Beispiel Tuma/Hannan (1984, Kap. 9 und Kap. 16), Coleman (1984b), Petersen (1985) und Clayton/ Cuzick (1985).

Anhänge

Anhang 1: Übersicht über die in den Beispielen verwendeten Variablennamen

Variablenname	Bedeutung
BANZ	Anzahl der vorher ausgeübten Berufe
BERF, IBERF[2)]	Berufserfahrung in Anzahl von Monaten
BERFA	Berufserfahrung in Monaten zu Beginn der Berufsepisode
BERFE	Berufserfahrung in Monaten am Ende der Berufsepisode
BILDG	Ausbildungsniveau in Anzahl von durchschnittlichen Schuljahren zu Beginn der Berufsepisode: 9 Jahre ≙ Volksschul- oder Hauptschulabschluß ohne Berufsausbildung 10 Jahre ≙ Mittlere Reife ohne Berufsausbildung 11 Jahre ≙ Volksschul- oder Hauptschulabschluß mit Berufsausbildung 12 Jahre ≙ Mittlere Reife mit Berufsausbildung 13 Jahre ≙ Abitur 17 Jahre ≙ Fachhochschulabschluß 19 Jahre ≙ Hochschulabschluß
DP	Hilfsvariable zur Bezeichnung der abhängigen Variablen (dependent variable) für das Programm P3RFUN von Trond Petersen
DUR, IDUR[2)]	Verweildauer in Monaten in einem Beruf
GESCHL	Geschlecht: 1 ≙ männlich 1 ≙ weiblich
HEIRAT	Familienstand: 0 ≙ unverheiratet 1 ≙ verheiratet
KOHO	Kohortenvariable: 1 ≙ Kohorte 1929–31 2 ≙ Kohorte 1939–41 3 ≙ Kohorte 1949–51

Fortsetzung von Anhang 1

Variablenname	Bedeutung
KOHO2, K2[1], KOH2[4]	Dummy-Variable für die Kohorte 1939–41: 1 ≙ Kohorte 1939–41 0 ≙ sonst.
KOHO3, K3[1], KOH3[4]	Dummy-Variable für die Kohorte 1949–51: 1 ≙ Kohorte 1949–51 0 ≙ sonst.
KONST	Bezeichung der Regressionskonstanten bei P3RFUN-Programmläufen
M3	Dummy-Variable für die Männer: 1 ≙ Männer 0 ≙ sonst.
M5	Sequenznummer der Erwerbstätigkeitsepisode
M41	Höchster allgemeinbildender Schulabschluß zu Beginn der Berufsepisode: 1 ≙ Hauptschulabschluß 2 ≙ Mittlere Reife 3 ≙ Abitur
M42	Höchster berufsbildender Schulabschluß zu Beginn der Berufsepisode: 1 ≙ kein beruflicher Ausbildungsabschluß 2 ≙ Lehrausbildung oder adäquate Berufsausbildung 3 ≙ Meister- oder Technikerabschluß 4 ≙ Fachhochschulabschluß 5 ≙ Universitätsabschluß
M43	Zeitpunkt des Eintritts in das Beschäftigungssystem in Anzahl von Monaten seit Beginn des Jahrhunderts
M47	Zeitpunkt des Interviews in Anzahl von Monaten seit Beginn des Jahrhunderts
M48	Zeitpunkt der Geburt in Anzahl von Monaten seit Beginn des Jahrhunderts
M50, TANF[3]	Beginn der Berufsepisode in Anzahl von Monaten seit Beginn des Jahrhunderts
M51, TEND[3]	Ende der Berufsepisode in Anzahl von Monaten seit Beginn des Jahrhunderts
M59, PRES[3]	Prestige, gemessen nach der Prestigeskala von Wegener (1985)
M61, JOBN[3], JN[1]	Berufsgruppe, in der sich der Befragte während der Berufsepisode befindet (12 Ausprägungen, siehe Tabelle 4.4)
M62, JOBN1[3], JN1[1]	Berufsgruppe, in der sich der Befrage in der nächstfolgenden Berufsepisode befindet (12 Ausprägungen, siehe Tabelle 4.4)
RESID, RESID1[5], RESID2[5]	Residuenvariable zur Überprüfung der Modellgüte
M59, THEIRAT[3], THEI[4]	Zeitpunkt der Heirat in Anzahl von Monaten seit Beginn des Jahrhunderts

Fortsetzung von Anhang 1

Variablenname	Bedeutung
U	Variable, die die logarithmierten Verweildauern für die Schätzung mit GLIM enthält
VDABH	Variable, die die Verweildauerabhängigkeit bei parametrischen Modellen bezeichnet
X1	Zeitpunkt des Beginns der Episoden bzw. Subepisoden im Programm P3RFUN von Trond Petersen
ZEN, IZEN[2], C[1]	Indikatorvariable, die die Zensierungsinformation enhält: 1 ≙ Episode wird durch ein Ereignis beendet 0 ≙ Episode wird zensiert
Z2	Interaktionsvariable des Geschlechts mit der Zeit

[1] Bezeichnung bei Analysen mit dem Programmsystem GLIM. Dort werden nun die ersten drei Zeichen unterschieden.

[2] Bezeichnung beim Episodensplitting.

[3] Bezeichnung nach der SPSS-Aufbereitung für RATE-Analysen

[4] Abkürzende schreibweise beim Episodensplitting

[5] Zur Unterscheidung von Residuen-Werten

```
C       WRITTEN BY TROND PETERSEN, DAVID DICKENS AND NANCY WILLIAMSON.
      SUBROUTINE P3RFUN( F,    DF,   P,    X,     N,
     *                   KASE, NVAR, NPAR, IPASS, XLOSS,
     *                   IDEP )
C
C     P3RFUN - FUNCTION SUBROUTINE
C                                                                      **
C     F      = FUNCTION VALUE (OUTPUT)
C     DF     = ARRAY OF DERIVATIVES WITH RESPECT TO P (OUTPUT)
C     P      = CURRENT VALUE OF PARAMETERS (INPUT)
C     X      = CURRENT CASE (INPUT)
C     N      = CODE NUMBER FOR FUNCTION (INPUT)
C     KASE   = CURRENT CASE NUMBER (INPUT)
C     NVAR   = NUMBER OF VARIABLES (INPUT)
C     NPAR   = NUMBER OF PARAMETERS (INPUT)
C     IPASS  = INDEX OF PASS (INPUT, SEE MANUAL)
C     XLOSS = LOSS VALUE (OUTPUT, SEE MANUAL) NOT USED HERE.
C          INCLUDED ONLY SO USER SUPPLIED FUN CAN USE IT.
C     IDEP   = INDEX OF THE DEPENDENT VARIABLE (INPUT)
C ********************************************************************
      IMPLICIT REAL*8 (A-H,O-Z)
      DIMENSION INDEP(100),A(6),H(9),DF(NPAR),P(NPAR),X(NVAR)
C
C
      LOGICAL ERR,ALPHA(26),FIRST,FDPROG,FDMODL
        DATA ERR,ALPHA,FIRST/28*.FALSE./
        DATA FDPROG,FDMODL/2*.FALSE./
      CHARACTER * 256 CONTRL

C     VALMAX IS THE MAXIMUM VALUE THAT SHOULD BE USED AS AN
C     ARGUMENT TO THE EXP FUNCTION TO AVOID OVER AND UNDERFLOWS
C     IN SUBROUTINE LSTSQ. ( 2*VALMAX ) SHOULD NOT OVERFLOW.

      DATA VALMAX / 40. /
      IF (.NOT. FIRST) THEN
                FIRST = .TRUE.

997         DO 9971 I = 1,256
9971         CONTRL(I:I) = ' '
                    READ (1,1000,END=980,ERR=8888)CONTRL
1000         FORMAT (A256)
8888                WRITE (40,2222) CONTRL
2222     FORMAT (A256)
         REWIND 40
1818    DO 999 I = 1,256
              IF (CONTRL(I:I) .NE. ' ') GO TO 998
999         CONTINUE

998         CONTINUE


        IF (CONTRL(I:I)  EQ. '''') GO TO 990
```

```
      IF (.NOT. FDPROG) THEN
            FDPROG = .TRUE.
            IF (.NOT.(CONTRL(I:I+1) .EQ. '/C')) THEN
                  WRITE(35,1001) CONTRL(1:60)
1001              FORMAT (' FIRST BMDP CONTROL CARD READ SHOULD
     * BE /C OR /C AND',
     *            INSTEAD HAVE READ'/5X,A60)
                  ERR = .TRUE.
                  GO TO 980
            END IF
            GO TO 997
      END IF

      IF (.NOT.FDMODL) THEN
            FDMODL = .TRUE.
            IF (CONTRL(I:I) .EQ.'M' .OR. CONTRL(I:I) .EQ. 'M')THEN
                  ALPHA(13) = .TRUE.
                  READ (40,2223) IMOD
2223     FORMAT (1X,I3)
         REWIND 40
                  WRITE (35,10021) IMOD
10021              FORMAT (/' MODEL CARD SAYS RUNNING MODEL',I5)
                  GO TO 997
            ELSE
                  WRITE(35,1002) CONTRL(1:60)
1002               FORMAT (' FIRST P3RFUN CARD SHOULD BE MODEL CARD,
     *        BUT INSTEAD IS '/5X,A60)
                  ERR = .TRUE.
                  GO TO 980
            END IF
      END IF

      IF (CONTRL(I:I) .EQ. 'T' .OR. CONTRL(I:I) .EQ. 'T')THEN
            ALPHA(20) = .TRUE.
            READ (40,2224) ITIME1,ITIME2
2224        FORMAT(1X,2I3)
            REWIND 40
            WRITE (35,1003) ITIME1,ITIME2
1003       FORMAT (/' TIME CARD SHOWS BEGIN AND END TIME VARIABLES
     *        AS',2I5)

      ELSE IF (CONTRL(I:I) .EQ. 'C' .OR. CONTRL(I:I) .EQ. 'C')THEN
            ALPHA(3) = .TRUE.
            READ (40,2225) ICNVAR
2225        FORMAT(1X,I3)
            REWIND 40
            WRITE (35,1004)  ICNVAR
1004       FORMAT (/' CENSOR CARD SHOWS CENSOR VARIABLE AS',I5)

      ELSE IF (CONTRL(I:I) .EQ. 'D' .OR. CONTRL(I:I) .EQ. 'D')THEN
            ALPHA(4) = .TRUE.
            READ(40,2226) IDEPVR
```

```
2226          FORMAT(1X,I3)
              REWIND 40
              WRITE (35,1005) IDEPVR
1005        FORMAT (/' DEPENDENT VARIABLE CARD SHOWS DEPVAR AS',I5)
          ELSE IF (CONTRL(I:I) .EQ. 'I' .OR.CONTRL(I:I) .EQ. 'I')THEN
              READ (40,2227)MDNPAR
2227          FORMAT (1X,I3)
              REWIND 40

              IF (MDNPAR .EQ. 0) THEN
                    ALPHA(9) = .TRUE.
                    WRITE(35,1009)
1009                    FORMAT(/' INDEPENDENT VARIABLE CARD SHOWS NO',
     *                    ' COVARIATES')
                    GO TO 997
              END IF
                    ALPHA(9) = .TRUE.
           READ (40,2228)(INDEP(J),J=1,MDNPAR)
2228      FORMAT (4X,40I3)
          REWIND 40
                    WRITE (35,1006) MDNPAR,(INDEP(J),J=1,MDNPAR)
1006               FORMAT (/' INDEPENDENT VARIABLE CARD SHOWS',I3,
     *     ' COVARIATES TO BE',
     *               /2X,20I3/2X,20I3)
        REWIND 40
          ELSE IF (CONTRL(I:I) .EQ. 'L' .OR. CONTRL(I:I) .EQ. 'L') THEN
              READ (40,2229)LFX1,LFX2
              ALPHA(12) = .TRUE.
              WRITE (35,1007) LFX1,LFX2
2229     FORMAT(1X,2I3)
1007               FORMAT (/' LABOR FORCE EXPERIENCE (OR SIMILAR`
     *      VARIALE) AT',
     *               ' BEGINNING AND END'/5X,'OF RECORD SHOWN AS
     *              VARIABLES',2I5)
                   REWIND 40
          ELSE
              WRITE (35,1008) CONTRL(1:60)
1008               FORMAT(' READ A CONTROL CARD TYPE NOT
     *     EXPECTING'/11X,A60)
                   ERR = .TRUE.
          END IF

          GO TO 997


980        WRITE (35,981)
981          FORMAT (//'*** RUN STOPPED BECAUSE OF FATAL ERRORS ***')
               STOP
990         CONTINUE
          IF (.NOT.( ALPHA(13) .AND. ALPHA(20) .AND. ALPHA(3) .AND.
     *        ALPHA(9) .AND. ALPHA(4))) THEN
              WRITE (35,991)
```

Fortsetzung von Anhang 2

```
991          FORMAT (//' DID NOT FIND ALL REQUIRED (M,T,C,I,AND D)',
     *         ' CONTROLS CARDS')
             ERR = .TRUE.
        END IF
        IF (IMOD .EQ. 3 .AND. .NOT. ALPHA(12)) THEN
             WRITE (35,992)
992             FORMAT (//' SPECIFIED MODEL 3 BUT NO L CONTROL CARD')
             ERR = .TRUE.
        END IF

        IF (ERR) GO TO 980

        WRITE (35,995)
995          FORMAT(//' END OF READING P3RFUN COMMAND INFORMATION',
     *     ' FROM TITLE CARD OF BMDP COMMAND FILE.')
       END IF


C       THERE ARE ALTOGETHER FIVE MODELS.
C       THESE ARE:
C       MODEL 1: THE EXPONENTIAL DISTRIBUTION
C       MODEL 2: THE GOMPERTZ DISTRIBUTION
C       MODEL 3: THE GOMPERTZ DISTRIBUTION WITH AGE OR
C            LABOR FORCE EXPERIENCE IN ADDITION TO AGE
C            AS A CONTINUOUSLY VARYING VARIABLE WITHIN
C            THE STATE.
C       MODEL 4: THE WEIBULL DISTRIBUTION
C       MODEL 5: THE LOG-LOGISTIC DISTRIBUTION.

C       ALL MODELS ALLOW FOR THE INCLUSION OF TIME-DEPENDENT
C       COVARIATES.

C       THE PARAMETERS TO BE ESTIMATED ALWAYS COME IN THIS ORDER.

C       COMMON TO ALL MODELS:
C       P(1): THE CONSTANT TERM

C       IN MODEL 1:
C       P(2),P(3) ETC.: THE EFFECTS OF VARIABLES WHICH
C       FOLLOW STEP-FUNCTIONS OVER TIME.

C       IN MODEL 2-5:
C       P(2): THE DURATION DEPENDENCE

C       IN MODEL 2,4,5:
C       P(3),P(4) ETC.: THE EFFECTS OF VARIABLES WHICH FOLLOW
C       STEP FUNCTIONS OVER TIME.

C       IN MODEL 3:
C       P(3): THE EFFECT OF AGE OR LABOUR FORCE EXPERIENCE
C       OR SOME SIMILAR VARIABLE AS A CONTINUOUSLY VARYING
C       VARIABLE WITHIN A STATE IN ADDITION TO DURATION.
C       THIS VARIABLE MUST BE A DETERMINISTIC AND LINEAR
```

```
C       FUNCTION OF SOME INITIAL VALUE AT THE BEGINNING OF
C       A STATE AND THE TIME SPENT IN THE STATE.

C       P(4),P(5) ETC.: THE EFFECTS OF VARIABLES WHICH
C       FOLLOW STEP-FUNCTIONS OVER TIME.

C  ESTABLISH PARAMETERS FOR SPECIFIC MODELS, FOLLOWED BY
C  CALCULATION OF CV BEFORE BRANCHING BASED ON MODELS.

      IF (IMOD .EQ. 2 .OR. IMOD .EQ. 4 .OR. IMOD .EQ. 5)      THEN
        IBEGIN = 3
      ELSE IF (IMOD .EQ. 1)THEN
        IBEGIN = 2
      ELSE IF (IMOD .EQ. 3) THEN
        IBEGIN = 4
      END IF


C       WE START WITH DEFINING THE CRUCIAL VARIABLES
C       ON DURATION, CENSORING AND THE COVARIATES.

      CV=0.0
      IF (NPAR .GE. IBEGIN) THEN
        DO 11 I=IBEGIN,NPAR
11        CV=CV+P(I)*X(INDEP(I-(IBEGIN-1)))
      END IF

      IF (IMOD .EQ. 1) THEN

C       THEN COMES A BUNCH OF FUNCTIONS WHICH WE NEED IN
C       WRITING DOWN THE LOG-LIKELIHOOD AS WELL AS THE GRADIENT
C       VECTOR USED IN THE MODIFIED SCORING ALGORITHM.

        H(1)=P(1)+CV
        DUR=X(ITIME2)-X(ITIME1)
        A(1) = DUR*DEXP(H(1))

C       HERE COMES THE LOGLIKELIHOOD.

        F= X(ICNVAR)*H(1) - A(1)


        DF(1) = X(ICNVAR) - A(1)


      ELSE IF (IMOD .EQ. 2) THEN

C       NOW COMES THE MODEL FOR THE GOMPERTZ DISTRIBUTION. THIS
C       IS THE MODEL WHERE WE DO NOT TAKE ACCOUNT OR AGE OR
C       LABOR FORCE EXPERIENCE AS A CONTINUOUSLY VARYING VARIABLE
C       WITHIN A SPELL.
```

```
        H(1)=P(1)+P(2)*X(ITIME2)+CV
        H(2)=DEXP(H(1))
        H(3)=DEXP(P(1)+P(2)*X(ITIME1)+CV)
        A(1)=(1.0/P(2))*(H(2)-H(3))
        A(2)=(1.0/(P(2)*P(2)))*(H(2)-H(3))
        A(3)=(1.0/P(2))*(X(ITIME2)*H(2)-X(ITIME1)*H(3))

C       THEN COMES THE LOGLIKELIHOOD.

        F=X(ICNVAR)*H(1)-A(1)

C       THEN COMES THE GRADIENT VECTOR.

        DF(1)=X(ICNVAR) - A(1)
        DF(2)=X(ICNVAR)*X(ITIME2)+A(2)-A(3)



      ELSE IF (IMOD .EQ. 3) THEN

C       THIS IS THE SET-UP FOR DOING THE GUMPERTZ DISTRIBUTION
C       WITH TIME-DEPENDENT COVARIATES, AND IN PARTICULAR WITH
C       AGE OR LABOR FORCE EXPERIENCE IN ADDITION TO DURATION AS
C       A CONTINUOUSLY VARYING VARIABLE WITHIN THE STATE.


C       NOW, COMES SOME AUXILLIARY FUNCTIONS.
C       THESE ARE USED IN WRITING DOWN THE LOG-LIKELIHOOD
C       AND THE GRADIENT VECTOR.

        H(1)=P(1)+P(2)*X(ITIME2)+P(3)*X(LFX2)+CV
        H(2) = DEXP(H(1))
        H(3) = DEXP(P(1) + P(2)*X(ITIME1) + P(3)*X(LFX1) + CV)
        A(1) = (1.0/(P(2)+P(3)))*(H(2)-H(3))
        A(2) = (1.0/((P(2)+P(3))*(P(2)+P(3))))*(H(2)-H(3))
        A(3) = (1.0/(P(2)+P(3)))*(X(ITIME2)*H(2)-X(ITIME1)*H(3))
        A(4) = (1.0/(P(2)+P(3)))*(X(LFX2)*H(2)-X(LFX1)*H(3))

C       THEN COMES THE LOGLIKELIHOOD FUNCTION.

        F = X(ICNVAR)*H(1) - A(1)

C       THEN COMES THE GRADIENT VECTOR.

        DF(1) = X(ICNVAR) - A(1)
        DF(2) = X(ICNVAR)*X(ITIME2) + A(2) - A(3)
        DF(3) = X(ICNVAR)*X(LFX2) + A(2) - A(4)


      ELSE IF (IMOD .EQ. 4) THEN

C       NOW COMES THE MODEL FOR THE WEIBULL DISTRIBUTION. THIS
C       IS THE MODEL WHERE WE DO NOT TAKE ACCOUNT OR AGE OR
C       LABOR FORCE EXPERIENCE AS A CONTINUOUSLY VARYING VARIABLE
```

```
C       WITHIN A SPELL.
C       THE HAZARD IS OF THE FORM:
C       H(T) = EXP(BX + A*LN(T))
C       WHEN A IS LESS THAT ZERO THERE IS NEGATIVE DURATION
C       DEPENDENCE, OTHERWISE IT IS POSITIVE. THE A COEFFICIENT
C       SHOULD BE GREATER THAN -1.0.

C       THIS PARAMETRIZATION IS SOMEWHAT DIFFERENT FROM THE
C       ONE WE USUALLY FIND IN THE STATISTICAL LITERATURE. IN THE
C       USUAL PARAMETRIZATION THE DURATION DEPENDENCE PARAMETER
C       IS ALWAYS GREATER THAN 0.0, AND WE HAVE NEGATIVE DURATION
C       DEPENDENCE WHEN THE PARAMETER IS LESS THAN 1.0.
C       THE USUAL PARAMETER CAN BE OBTAINED FROM THE MODEL
C       ESTIMATED HERE BY JUST ADDING 1.0 TO THE PRESENT ESTIMATE
C       OF THE DURATION DEPENDENCE PARAMETER.


        H(1)=P(1)+CV
        H(2)=DEXP(H(1))
        H(3) = DLOG(X(ITIME2))
        H(4) = P(2) + 1.0
        H(5) = (1/H(4))
        H(6) = X(ITIME2)**H(4)

        IF (X(ITIME1) .EQ. 0.0) THEN
              H(7) = 0.0
        ELSE
              H(7) = X(ITIME1)**H(4)
        END IF
        H(8) = H(6) - H(7)

        IF (X(ITIME1) .LE. 1.0) THEN
              H(9) = H(6)*H(3)
        ELSE
               H(9) = H(6)*H(3) - H(7)*DLOG(X(ITIME1))
        END IF

        A(1) = H(5)*H(2)*H(8)
        A(2) = X(ICNVAR) - A(1)
        A(3) = H(5)*H(2)*(H(5)*H(8) - H(9))

C       THEN COMES THE LOG-LIKELIHOOD.

        F = X(ICNVAR)*(H(1) + P(2)*H(3)) - A(1)

C       THEN COMES THE GRADIENT VECTOR.

        DF(1) = A(2)
        DF(2) = X(ICNVAR)*H(3) + A(3)


      ELSE IF (IMOD .EQ. 5) THEN
```

```
C       NOW COMES THE MODEL FOR THE LOG-LOGISTIC DISTRIBUTION. THIS
C       IS THE MODEL WHERE WE DO NOT TAKE ACCOUNT OR AGE OR
C       LABOR FORCE EXPERIENCE AS A CONTINUOUSLY VARYING VARIABLE
C       WITHIN A SPELL.

C       THE HAZARD IS:
C       R(T)= (P+1)*(EXP(BX + P*LN(T)))/(1 + EXP(BX + P*LN(T)))
C       HERE, P SHOULD BE GREATER THAN -1.0.
C       THERE IS MONOTONE NEGATIVE DURATION DEPENDENCE WHEN
C       P IS LESS THAN ZERO. FOR P LARGER THAN ZERO THE DURATION
C       DEPENDENCE IS FIRST POSITIVE AND AFTER SOME TIME IT
C       BECOMES NEGATIVE. THE SWITCH POINT DEPENDS BOTH ON P
C       AND ON BX. IT SHOULD BE CALCULATED FROM THE ESTIMATES.

C       NOTE THAT THIS PARAMETRIZATION IS SOMEWHAT DIFFERENT
C       FROM THE ONE WE USUALLY FIND IN THE LITERATURE.
C       IN PARTICULAR: THE TRADITIONAL DURATION DEPENDENCE PARAMETER
C       SHOULD ALWAYS BE LARGER THAN 0.0, WHICH GIVES MONOTONE NEGATIVE
C       DURATION DEPENDENCE FOR ALL VALUES OF THE PARAMETER EQUAL
C       TO OR LESS THAN 1.0 . THE TRADIONAL DURATION DEPENDENCE
C       CAN BE OBTAINED FROM THE PRESENT BY ADDING 1.0 TO THE
C       DURATION DEPENDENCE PARAMETER ESTIMATED HERE.
C       I FIND THE PRESENT PARAMETRIZATION PREFERABLE BECAUSE IT
C       MAXIMIZES COMPARABILITY OF THE DURATION DEPENDENCE ACROSS THE
C       FIVE MODELS.

      H(1)=P(1)+CV
      H(2)=DEXP(H(1))
      H(3) = P(2) + 1.0
      H(4) = X(ITIME2)**H(3)

      IF (X(ITIME1) .EQ. 0.0) THEN
            H(5) = 0.0
      ELSE
            H(5) = X(ITIME1)**H(3)
      END IF

      H(6) = DLOG(X(ITIME2))

      IF (X(ITIME1) .LE. 1.0) THEN
            H(7) = 0.0
      ELSE
            H(7) = DLOG(X(ITIME1))
      END IF

      H(8) = H(2)*H(5)
      H(9) = H(2)*H(4)
      A(1) = DLOG(1 + H(8))
      A(2) = DLOG(1 + H(9))
      A(3) = (1./(1+H(8)))* H(8)
      A(4) = (1./(1+H(9)))* H(9)
      A(5) = A(3)*H(7)
      A(6) = A(4)*H(6)
```

```
C       THEN COMES THE LOG-LIKELIHOOD.

              IF (X(ICNVAR) .EQ. 1.0) THEN
                    F = H(1) +DLOG(H(3)) + P(2)*H(6) + A(1) - 2.*A(2)
              ELSE
                    F = A(1) - A(2)
              END IF
C       THEN COMES THE GRADIENT VECTOR.

              IF (X(ICNVAR) .EQ. 1.0) THEN
                    DF(1) = 1. + A(3) - 2.*A(4)
                    DF(2) = (1./H(3)) + H(6) + A(5) - 2.*A(6)
              ELSE
                    DF(1) = A(3) - A(4)
                    DF(2) = A(5) - A(6)
              END IF
      END IF

C  NOW THE CALCULATIONS THAT ARE DONE IN COMMON FOR ALL MODELS.


      DO 21 I=IBEGIN,NPAR
21      DF(I)=X(INDEP(I-(IBEGIN-1)))*DF(1)

C       THIS IS THE TRICK TO REDEFINE THE DEPENDENT VARIABLE
C       IN THE SCORING ALGORITHM USED FOR NON-LINEAR LEAST SQUARES
C       PROBLEMS OR FOR MAXIMUM LIKELIHOOD PROBLEMS IN MODELS
C       FALLING WITHIN AN EXPONENTIAL FAMILY.

      X(IDEPVR)= F+1.0

C       FINALLY COMES THE LOG-LIKELIHOOD CONVERGENCE CRITERION.

      XLOSS=-F
          RETURN
       END
```

Anhang 3: Listing des FORTRAN-Programms zum Episodensplitting bei diskreten zeitveränderlichen unabhängigen Variablen

```
      PROGRAM TRANSF
      INTEGER TANF,TEND,ZEN,BILDG,PRES,BANZ,BERF
      INTEGER KOHO2,KOHO3,DUR,JOBN,JOBN1,THEIRAT
      INTEGER HEIRAT
      N = 0
      M = 0
      MDUR = 0
      NDUR = 0
1     READ(20,1001,END=999) TANF,TEND,ZEN,BILDG,PRES,BANZ,BERF,
     *                      KOHO2,KOHO3,DUR,JOBN,JOBN1,THEIRAT
      M = M + 1
      MDUR = MDUR + DUR
      IF(THEIRAT .EQ. 0) THEIRAT = 10000
      IF(THEIRAT .GE. TEND) THEN
        HEIRAT = 0
        IDUR = DUR
        ITANF = TANF
        ITEND = TEND
        IZEN = ZEN
        WRITE(30,1002) ITANF,ITEND,IZEN,BILDG,PRES,BANZ,BERF,
     *                 KOHO2,KOHO3,IDUR,JOBN,JOBN1,THEIRAT,HEIRAT
        N = N + 1
        NDUR = NDUR + IDUR
      ELSE IF(THEIRAT .LE. TANF) THEN
        HEIRAT = 1
        IDUR = DUR
        ITANF = TANF
        ITEND = TEND
        IZEN = ZEN
        WRITE(30,1002) ITANF,ITEND,IZEN,BILDG,PRES,BANZ,BERF,
     *                 KOHO2,KOHO3,IDUR,JOBN,JOBN1,THEIRAT,HEIRAT
        N = N + 1
        NDUR = NDUR + IDUR
      ELSE
        HEIRAT = 0
        IDUR = THEIRAT - TANF
        ITANF = TANF
        ITEND = THEIRAT
        IZEN = 0
        WRITE(30,1002) ITANF,ITEND,IZEN,BILDG,PRES,BANZ,BERF,
     *                 KOHO2,KOHO3,IDUR,JOBN,JOBN1,THEIRAT,HEIRAT
        N = N + 1
        NDUR = NDUR + IDUR
        HEIRAT = 1
        IDUR = TEND - THEIRAT
        ITANF = TEND - IDUR
        ITEND = TEND
        IZEN = ZEN
        WRITE(30,1002) ITANF,ITEND,IZEN,BILDG,PRES,BANZ,BERF,
     *                 KOHO2,KOHO3,IDUR,JOBN,JOBN1,THEIRAT,HEIRAT
```

```
        N = N + 1
        NDUR = NDUR + IDUR
      END IF
      GOTO 1
999   WRITE (2,2000) M,N,MDUR,NDUR
2000  FORMAT(//' ','ES WURDEN ',I5,' SAETZE GELESEN UND ',I6,'SAETZE',
     *' AUSGEGEBEN'/' ','DABEI WURDEN ',I10,' ZEITEINHEITEN GELESEN'
     *' UND ',I15,' ZEITEINHEITEN AUSGEGEBEN')
1001   FORMAT(13I5)
1002   FORMAT(14I5)
      END
```

Anhang 4: Listing des FORTRAN-Programms zum Episodensplitting bei stetigen zeitveränderlichen unabhängigen Variablen

```
      PROGRAM TRANSF
      INTEGER TANF,TEND,ZEN,BILDG,PRES,BANZ,BERF
      INTEGER KOHO2,KOHO3,DUR,JOBN,JOBN1,THEIRAT
      INTEGER INTERVALL/60/,OBERGRENZE
      N = 0
      M = 0
      MDUR = 0
      NDUR = 0
1     READ(20,1000,END=999) TANF,TEND,ZEN,BILDG,PRES,BANZ,BERF,
     *                      KOHO2,KOHO3,DUR,JOBN,JOBN1,THEIRAT
      M = M + 1
      MDUR = MDUR + DUR
      OBERGRENZE = INTERVALL
      IBERFA = BERF
      ITANF = TANF
2     IF (DUR .LE. INTERVALL) THEN
         IBERF = IBERFA
         IZEN = ZEN
         IDUR = DUR
         ITANF = ITANF
         ITEND = ITANF + INTERVALL
         IF (ITEND .GT. TEND) ITEND = TEND
      ELSE
         IBERF = IBERFA
         IZEN = 0
         IDUR = INTERVALL
         ITANF = ITANF
         ITEND = ITANF + INTERVALL
         IF (ITEND .GT. TEND) ITEND = TEND
      END IF
      WRITE(30,1000) ITANF,ITEND,IZEN,BILDG,PRES,BANZ,IBERF,
     *               KOHO2,KOHO3,IDUR,JOBN,JOBN1,THEIRAT
      N = N + 1
      NDUR = NDUR + IDUR
      OBERGRENZE = OBERGRENZE + INTERVALL
      IBERFA = IBERFA + INTERVALL
      DUR = DUR - INTERVALL
      ITANF = ITANF + INTERVALL
      IF (DUR .LE. 0) GOTO 1
      GOTO 2
999   WRITE (2,2000) M,N,MDUR,NDUR
2000  FORMAT(//' ','ES WURDEN ',I5,' SAETZE GELESEN UND ',I6,'SAETZE',
     *' AUSGEGEBEN'/' ','DABEI WURDEN ',I10,' ZEITEINHEITEN GELESEN'
     *' UND ',I15,' ZEITEINHEITEN AUSGEGEBEN')
1000  FORMAT(1315)
      END
```

Anhang 5: Listing der GLIM-Makros zum Schätzen von Weibull- und log-logistischen Modellen von Roger und Peacock

```
$SUBFILE MAKROS
$M ML1
$CA %LP=%IF(%LT(%LP,78),%LP,78)
$CA %LP=%IF(%GT(%LP,-78),%LP,-78)
$CA %FV=N/(1+%EXP(-%LP))
$CA %FV(%NU)=%W/2
$END
$C
$C
$C
$M ML2
$CA %DR=N/(%FV*(N-%FV))
$CA %DR(%NU)=%LP(%NU)/%FV(%NU)
$END
$C
$C
$C
$M ML3
$CA %VA=%FV*(N-%FV)/N
$CA %VA(%NU)=%W/4
$END
$C
$C
$C
$M ML4
$CA %DI=2*((R-N)*%LOG(N-%FV)-R*%LOG(%FV))
$CA %DI(%NU) = -2*%W*%LOG(%LP(%NU))
$END
$C
$C
$C
$M SETL
$CA GM=1
$CA GM(%NU)=0
$CA U(%NU)=1
$CA C(%NU)=0
$CA N=%EQ(C,1)
$CA %W=%CU(N)
$CA N=N+1
$CA R=%NE(C,0)
$CA R(%NU)=N(%NU)=%W
$YVAR R
$OWN ML1 ML2 ML3 ML4
$SCALE 1$
$CA %LP=%LOG(R+0.1)/(N-R+0.1)
$CA %LP(%NU)=0.5
$FIT GM - %GM + U
$DIS E $
$END
$C
$C
```

```
$C
$MAC MW1 $
$CA %LP=%IF(%LT(%LP,78),%LP,78)
$CA %LP=%IF(%GT(%LP,-78),%LP,-78)
$CA %FV=%EXP(%LP)
$CA %FV(%NU)=%W/2
$END
$C
$C
$C
$MAC MW2 $
$CA %DR=1/%FV
$CA %DR(%NU)=%LP(%NU)/%FV(%NU)
$END
$C
$C
$C
$MAC MW3 $
$CA %VA=%FV
$CA %VA(%NU)=%W/4
$END
$C
$C
$C
$MAC MW4 $
$CA %DI=2*(%FV-C*(%LP+1))
$CA %DI(%NU)=-2*%W*%LOG(%LP(%NU))
$END
$C
$C
$C
$MAC SETW $
$CA GM=1
$CA GM(%NU)=0
$CA U(%NU)=1
$CA C(%NU)=0
$CA %W=%CU(C)
$CA C(%NU)=%W
$YVAR C
$OWN MW1 MW2 MW3 MW4
$SCALE 1 $
$CA %LP=%LOG(C*0.8+0.1)
$CA %LP(%NU)=0.5
$FIT GM-%GM+U
$DIS E$
$END
$C
$C
$C
$RETURN
$FINISH
```

Literaturverzeichnis

ALLISON, P.: „Discrete time methods for the analysis of event histories". In: LEINHARDT, S. (Hrsg.): Sociological methodology. San Francisco 1982, S. 61–98.

AMEMIYA, T.: Advanced econometrics. Oxford 1985.

ANDERSEN, P. K.: „Testing goodness of fit of Cox's regression and life model". In: Biometrics, 1982, 38, S. 67–77.

ANDERSEN, P. K., und GILL, R. D.: „Cox's regression model for counting processes: A large sample study". In: Annals of Statistics, 1982, 10, S. 1100–1120.

ANDRESS, H. J.: Multivariate Analyse von Verlaufsdaten. ZUMA-Methodentexte, Bd. 1. Mannheim 1985.

ARMINGER, G.: „Modelltheoretische und methodische Probleme bei der Analyse von Paneldaten mit qualitativen Variablen". In: Vierteljahreshefte des Deutschen Instituts für Wirtschaftsforschung, 1984a, 4, S. 470–480.

DERS.: EM estimation for compound generalized linear models. Manuskript. Wuppertal 1984b.

DERS.: Testing against misspecification in parametric rate models. International Conference on Applications of Life History Analysis in Life Course Research. Max-Planck-Institut für Bildungsforschung, Berlin, Juni 1986.

ARROW, K.: „Higher education as a filter". In: Journal of Public Economics, 1973, 2, S. 193–216.

BAILEY, K. R.: „The asymptotic joint distribution of regression and survival parameter estimates in the Cox regression model". In: Annals of Statistics, 1983, 11, S. 39–48.

BAJENESCU, T. I.: Zuverlässigkeit elektronischer Komponenten. Berlin und Offenbach 1985.

BAKER, R. J., und NELDER, J. A.: The GLIM system. Oxford 1978.

BIRG, H.: „Arbeitsmarktdynamik, Familienentwicklung und generatives Verhalten – Eine biographietheoretische Konzeption für Untersuchungen demographisch relevanter Verhaltensweisen". In: SCHMID, J., und SCHWARZ, I. (Hrsg.): Politische und prognostische Tragweite von Forschungen zum generativen Verhalten. Berlin 1985, S. 209–223.

BLOSSFELD, H.-P.: Zur Repräsentativität der SfB-3-Lebensverlaufsstudie – Ein Vergleich mit Daten aus der amtlichen Statistik. Arbeitspapier Nr. 163 des SfB 3 „Mikroanalytische Grundlagen der Gesellschaftspolitik". Frankfurt a. M. und Mannheim 1985a (erscheint in: Allgemeines Statistisches Archiv).

DERS.: „Berufseintritt und Berufsverlauf. Eine Kohortenanalyse über die Bedeutung des ersten Berufs in der Erwerbsbiographie". In: Mitteilungen aus der Arbeitsmarkt- und Berufsforschung, 1985b, 2, S. 177–197.

DERS.: Bildungsexpansion und Berufschancen. Frankfurt a. M. und New York 1985c.

BRAUN, H., und HOEM, J.: Modelling cohabitational birth intervals in the current Danish population. A progress report. Working Paper, Nr. 24, Kopenhagen 1978.

BRESLOW, N. E.: „A generalized Kruskal-Wallis test for comparing K samples subject to unequal patterns of censorship". In: Biometrika, 1970, 57, S. 579–594.

DERS.: „Covariance analysis of censored survival data". In: Biometrics, 1974, 30, S. 89–100.

BRÜCKNER, E., u. a.: Methodenbericht „Lebensverläufe". ZUMA-Arbeitsbericht, Nr. T 84/08. Mannheim 1984 .

CARR-HILL, R. A., und MACDONALD, K. I.: „Problems in the analysis of life histories". In: Sociological Review Monograph, 1973, 19, S. 57–95.

CARROLL, G. R.: „Organizational ecology". In: Annual Review of Sociology, 1984, 10, S. 71–93.

CARROLL, G. R., und DELACROIX, J.: „Organizational mortality in the newspaper industries of Argentina and Ireland: An ecological approach". In: Administrative Science Quarterly, 1982, 27, S. 169–198.

CARROLL, G. R., und HUO, Y. P.: Organizational task and institutional environments in ecological perspective: Findings from the local newspaper industry. American Sociological Association Meetings, Washington, D.C. 1985.

DIES.: „Losing by winning: The paradox of electoral success by organized labor parties in the Knights of Labor era". In: DIEKMANN, A., und MITTER, P. (Hrsg.): Paradox consequences of social behavior. 1986 (im Druck).

CARROLL, G. R., und MAYER, K. U.: „Job-shift patterns in the Federal Republic of Germany: The effects of social class, industrial sector and organizational size". (erscheint in: American Sociological Review, 1986)

CHAMBERLAIN, G.: „Analysis of covariance with qualitative data". In: Review of Economic Studies, 1980, 47, S. 225–238.

CLAYTON, D. G., und CUZICK, J.: „Multivariate generalizations of the proportional hazards model". In: Journal of the Royal Statistical Society A, 1985, 148, S. 82–117.

COALE, A. J.: „Age patterns of marriage". In: Population Studies, 1971, 25, S. 193–214.

COLEMAN, J. S.: „Stochastic models of market structures". In: DIEKMANN, A., und MITTER, P. (Hrsg.): Stochastic modelling of social processes. New York 1984a, S. 189–213.
DERS.: „Interdependence among qualitative attributes". In: Journal of Mathematical Sociology, 1984b, 10, S. 29–50.
COURGEAU, D.: „Relations entre cycle de vie et migrations". In: Population, 1984, 3, S. 483–514.
CROWLEY, J., und STORER, B. E.: „Comment". In: Journal of the American Statistical Association, 1983, 78, S. 277–281.
COX, D. R.: „Regression models and life-tables (with discussion)". In: Journal of the Royal Statistical Society B, 1972, 34, S. 187–220.
DERS.: „Partial likelihood". In: Biometrika, 1975, 62, S. 269–276.
COX, D. R., und HINKLEY, D. V.: Theoretical statistics. London 1974.
COX, D. R., und OAKES, D.: Analysis of survival data. London 1984.
COX, D. R., und SNELL, E. J.: „A general definition of residuals (with discussion)". In: Journal of the Royal Statistical Society B, 1968, 30, S. 248–275.
DAVID, H. A., und MOESCHBERGER, M. L.: The theory of competing risks. London 1978.
DEMPSTER, A. P., LAIRD, N. M., und RUBIN, D. B.: „Maximum likelihood from incomplete data via the EM algorithm (with discussion)". In: Journal of the Royal Statistical Society B, 1977, 39, S. 1–38.
DIEKMANN, A.: Dynamische Modelle sozialer Prozesse. München 1980.
DIEKMANN, A., und MITTER, P.: „The ‚Sickle Hypothesis'". In: Journal of Mathematical Sociology, 1983, 9, S. 85–101.
DIES.: „A comparison of the „Sickle Function" with alternative stochastic models of divorce rates for Austrian and U.S. marriage cohorts". In: DIEKMANN, A., und MITTER, P. (Hrsg.): Stochastic modelling of social processes. New York 1984.
DIXON, W. J., u.a.: BMDP statistical software. Berkeley, Los Angeles und London 1983.
ELBERS, C., und RIDDER, G.: „True and spurious duration dependence: The identifiability of the proportional hazards model". In: Review of Economic Studies, 1982, 49, S. 403–410.
FAHRMEIR, L., und HAMERLE, A.: Multivariate statistische Verfahren. Berlin 1984.
FEATHERMAN, D. I.: „Retrospective longitudinal research: Methodological considerations". In: Journal of Economics and Business, 1979–1980, 32, S. 152–169.
FEATHERMAN, D. I., und SØRENSEN, A. B.: „Societal transformation in Norway and change in the life course transition into adulthood". In: Acta Sociologica, 1983, 26, S. 105–126.

FELMLEE, D., und EDER, D.: „Contextual effects in the classroom: The impact of ability groups on student attention". In: Sociology of Education, 1983, S. 77–87.

FLINN, Ch. J., und HECKMAN, J. J.: „Models for the analysis of labor force dynamics". In: BASMANN, R., und RHODES, G. (Hrsg.): Advances in econometrics. Bd. I, Greenwich, Conn. 1982, S. 35–95.

DIES.: „Are unemployment and out of the labor force behaviorally distinct labor force states?" In: Journal of Labor Economics, 1983, 1, S. 28–42.

FREEMAN, J., CARROLL, G. R., und HANNAN, M. T.: „The liability of newness: Age dependence in organizational death rates". In: American Sociological Review, 1983, 48, S. 692–710.

GAIL, M. H.: „A review and critique of some models used in competing risks analysis". In: Biometrics, 1975, 31, S. 209–222.

GEHAN, E. A.: „A generalized Wilcoxon test for comparing arbitrarily single-censored samples". In: Biometrika, 1965, 52, 203–223.

GROSS, A. J., und CLARK, V. A.: Survival distributions: Reliability applications in the Biomedical Sciences. New York 1975.

HAMERLE, A.: Zur statistischen Analyse von Zeitverläufen. Diskussionsbeitrag Nr. 180. Universität Regensburg 1984.

DERS.: Zählprozeß-Modelle zur statistischen Analyse von Ereignisdaten mit Kovariablen bei konkurrierenden Risiken und mehreren Episoden. Diskussionsbeitrag Nr. 90/s. Universität Konstanz 1985a.

DERS.: Diskrete Modelle zur statistischen Analyse von Verweildauern und Lebenszeiten. Diskussionspapier Nr. 79. Universität Konstanz 1985b.

DERS.: „Regressionsmodelle für gruppierte Verweildauern und Lebenszeiten". In: Zeitschrift für Operations Research, Serie B: Praxis, 1985c, 29, S. 243–260.

DERS.: Ein dynamisches Logitmodell für diskrete Hazardraten. Diskussionsbeitrag Nr. 92/s. Universität Konstanz 1986a.

DERS.: Diskrete Hazardraten-Modelle mit unbeobachteter Populationsheterogenität zur Analyse von Panel-Daten. Manuskript. Universität Konstanz 1986b.

HAMERLE, A., und PAPE, H.: „Über einen stochastischen Ansatz zur Lösung von Klassifikationsproblemen". In: Statistische Hefte, 1977, 18, S. 142 bis 146.

HAMERLE, A., KEMÉNY, P., und TUTZ, G.: „Kategoriale Regression". In: FAHRMEIR, L., und HAMERLE, A. (Hrsg.): Multivariate statistische Verfahren. Berlin 1984, Kap. 6.

HAMERLE, A., und TUTZ, G.: Diskrete Modelle zur Analyse von Verweildauern. Manuskript. Konstanz und Regensburg 1986.

HANEFELD, U.: „Das Sozio-ökonomische Panel – Eine Längsschnittstudie für die Bundesrepublik Deutschland". In: Vierteljahreshefte zur Wirtschaftsforschung, 1984, 4, S. 391–406.

HANDL, J., MAYER, K. U., und MÜLLER, W.: Klassenlagen und Sozialstruktur. Frankfurt a.M. und New York 1977.
HANNAN, M. T., und FREEMAN, J.: „The population ecology of organizations". In: American Journal of Sociology, 1977, 82, S. 929–964.
HECKMAN, J. J., und BORJAS, G.: „Does unemployment cause future unemployment? Definitions, questions and answers from a continuous time model of heterogeneity and state dependence". In: Econometrica, 1980, 47, S. 247–283.
HECKMAN, J. J., und SINGER, B.: „The identification problem in econometric models for duration data". In: HILDENBRAND, W. (Hrsg.): Advances in econometrics: Proceedings of world meetings of the Econometric Society. Cambridge 1980, S. 39–77.
DIES.: „Econometric duration analysis". In: Journal of Econometrics, 1984a, 24, S. 63–132.
DIES.: „A method for minimizing the impact of distributional assumptions in econometric models for duration data". In: Econometrica, 1984b, 52, S. 271–320.
HELBERGER, C.: Veränderungen der bildungsspezifischen Einkommensunterschiede zwischen 1969/71 und 1978. SfB-3-Arbeitspapier, Nr. 51. Frankfurt a.M. und Mannheim 1980.
HERNES, G.: „The process of entry into first marriage". In: American Sociological Review, 1972, 37, S. 173–182.
HOLT, J. D.: Competing risks analysis with special reference to matched pair experiments". In: Biometrika, 1978, 65, S. 159–166.
HOUGAARD, Ph.: „Life table methods for heterogeneous populations: Distributions describing the heterogeneity". In: Biometrika, 1984, 71, S. 75–83.
HUJER, R., und SCHNEIDER, H.: Ökonomische Ansätze zur Analyse von Paneldaten: Schätzung und Vergleich von Übergangsratenmodellen. Manuskript. Frankfurt a.M. 1986.
HULL, C. H., und NIE, N. H. (Hrsg.): SPSS-Update 7–9. New York 1981.
JOHANSEN, S.: „The product limit estimator as maximum likelihood estimator". In: Scandinavian Journal of Statistics, 1978, 5, S. 195–199.
KALBFLEISCH, J. D., und McINTOSH, A. A.: „Efficiency in survival distributions with time-dependent covariables". In: Biometrika, 1977, 64, S. 47–50.
KALBFLEISCH, J. D., und PRENTICE, R. L.: „The statistical analysis of failure time data. New York 1980.
KAPLAN, E. L., und MEIER, P.: „Nonparametric estimation from incomplete observations". In: Journal of the American Statistical Association, 1958, 53, S. 457–481.
KAY, R.: „Proportional hazard regression models and the analysis of censored survival data". In: Applied Statistics, 1977, 26, S. 227–237.
KEELEY, M. C.: „An analysis of the age pattern of first marriage". In: International Economic Review, 1979, 20, S. 527–544.

KEMÉNY, P., ROTHMEIER, F., und HAMERLE, A.: „Explorative Variablenselektion und Anpassungstests bei Regressionsmodellen zur Analyse der stationären Aufenthaltsdauer nach Unfallverletzungen im Schulsport". In: EDV in Medizin und Biologie, 1985, 16 (im Druck).
KIEFER, J., und WOLFOWITZ, J.: „Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters". In: Annals of Mathematical Statistics, 1956, 27, S. 887–906.
KIEFER, N.: „A simple test for heterogeneity in exponential models of duration". In: Journal of Labour Economics, 1984, 2, S. 539–549.
KRUPP, H.-J.: Das Sozio-ökonomische Panel. Bericht über die Forschungstätigkeit 1983–1985. Antrag auf Förderung der Forschungsphase 1986–1988. Frankfurt a. M. und Berlin 1985.
LAGAKOS, S. W.: „General right censoring and its impact on the analysis of survival data". In: Biometrics, 1979, 35, S. 139–156.
DERS.: „The graphical evaluation of explanatory variables in proportional hazard regression models". In: Biometrika, 1981, 68, S. 93–98.
LANCASTER, T.: „Generalized residuals and heterogeneous duration models with applications to the Weibull model". In: Journal of Econometrics, 1985, 28, S. 155–169.
LAWLESS, J. F.: Statistical models and methods for life-time data. New York 1982.
LEE, E., und DESU, M.: „A computer program for comparing K samples with right-censored data". In: Computer Programs in Biomedicine, 1972, 2, S. 315–321.
LEE, E., DESU, M., und GEHAN, E. H.: „A Monte Carlo study of the power of some two-sample tests". In: Biometrika, 1975, 62, S. 425–432.
LINDSAY, B. G.: „The geometry of mixture likelihoods: A general theory". In: Annals of Statistics, 1983a, 11, S. 86–94.
DERS.: „The geometry of mixture likelihoods, part II: The exponential family". In: Annals of Statistics, 1983b, 11, S. 783–792.
LUENBERGER, D. G.: Introduction to linear and nonlinear programming. Reading, Penn. 1973.
MANTEL, N.: „Evaluation of survival data and two new rank order statistics arising in its consideration". In: Cancer Chemotherapy Reports, 1966, 50, S. 163–170.
MANTEL, N., und MYERS, M.: „Problems of convergence of maximum likelihood iterative procedures in multiparameter situations". In: Journal of the American Statistical Association, 1971, 66, S. 484–491.
MARITZ, J. S.: Empirical Bayes methods. London 1971.
MATRAS, J.: On schooling and employment in the transition of Israeli males to adulthood. Manuskript. Jerusalem: Brookdale Institute 1983.
MAYER, K. U.: Lebensverläufe und Wohlfahrtsentwicklung. Bericht über die Forschungstätigkeit in der zweiten Forschungsphase 1982–1984. Frankfurt a. M. und Mannheim 1984a, S. 119–142.

DERS.: Lebensverläufe und Wohlfahrtsentwicklung. Antrag auf Förderung für die dritte Forschungsphase 1985–1987. Frankfurt a. M. und Mannheim 1984b, S. 131–171.
DERS: „Lebensverlaufsforschung". In: VOGES, W. (Hrsg.): Soziologie der Lebensalter. Methoden der Biographie und Lebenslaufforschung. Opladen 1986 (im Druck).
MAYER, K. U., und WAGNER, M.: Wann verlassen die Kinder ihr Elternhaus? Untersuchungen zu den Geburtsjahrgängen 1929–31, 1939–41, 1949–51. IBS-Materialien. Universität Bielefeld, Institut für Bevölkerungsforschung und Sozialpolitik 1986 (im Druck).
McCULLAGH, P., und NELDER, J. A.: Generalized linear models. London 1983.
MEULEMANN, H., u. a.: Bildung und Lebenslauf. Antrag auf Gewährung einer Sachbeihilfe (Neuantrag). Manuskript. o.O. 1984.
MICHAEL, R. T., und TUMA, N. B.: „Entry into marriage and parenthood by young men and women: The influence of family background". In: Demography, 1985, 22, S. 515–544.
MILLER, R. G.: Survival analysis. New York 1981.
MÜLLER, W.: Klassenlage und Lebenslauf. Habilitationsschrift. Mannheim 1978.
NAES, T.: „The asymptotic distribution of the estimator for the regression parameter in Cox's regression model". In: Scandinavian Journal of Statistics, 1982, 9, S. 107–115.
NEWMAN, J. L., und McCULLOCH, C. E.: „A hazard rate approach to the timing of births". In: Econometrica, 1984, 52, S. 939–961.
PAPASTEFANOU, G.: Zur Güte von retrospektiven Daten – Eine Anwendung gedächtnispsychologischer Theorie und Ergebnisse einer Nachbefragung. Arbeitspapier Nr. 29 des SfB 3 „Mikroanalytische Grundlagen der Gesellschaftspolitik". Frankfurt a. M. und Mannheim 1980.
DERS.: Veränderungen der Familienbildung in der Bundesrepublik seit dem zweiten Weltkrieg. Manuskript. Berlin 1986.
PETERSEN, T.: Incorporating time-dependent covariates in models for analysis of duration data. CDE Working Paper, 1985.
PRENTICE, R. L., und BRESLOW, N. E.: „Retrospective studies and failure time models". In: Biometrika, 1978, 65, S. 153–158.
PRENTICE, R. L., und SELF, S. G.: „Asymptotic distribution theory for Cox-type regression models with general relative risk form". In: Annals of Statistics, 1983, 11, S. 804–813.
PRENTICE, R. L., u. a.: „The analysis of failure time in the presence of competing risks". In: Biometrics, 1978, 34, S. 541–554.
RAO, C. R.: Lineare statistische Methoden und ihre Anwendungen. Berlin 1973.
ROBINSON, B. N., u. a.: SIR, scientific information retrieval, user's manual, version 2. Evanston 1980.

ROGER, J. H., und PEACOCK, S. D.: „Fitting the scale as a GLIM parameter for Weibull, extreme value, logistic and log-logistic regression models with censored data". In: GLIM-Newsletter, 1983, 6, S. 30–37.
SCHAICH, E., und HAMERLE, A.: Verteilungsfreie statistische Prüfverfahren. Heidelberg und Berlin 1984.
SCHOENFELD, D.: „Goodness-of-fit tests for the proportional hazards regression model". In: Biometrika, 1980, 67, S. 145–154.
DERS.: „Partial residuals for the proportional hazards regression model". In: Biometrika, 1982, 69, S. 239–241.
SCHULZ, M., und STROHMEIER, K. P.: „Familienkarriere und Berufskarriere". In: FRANZ, H.-W. (Hrsg.): 22. Deutscher Soziologentag 1984, Beiträge der Sektions- und Ad-hoc-Gruppen. Opladen 1985, S. 167–171.
SEAL, H. L.: „Studies in the history of probability and statistics. Multiple decrements or competing risks". In: Biometrika, 1977, 64, S. 429–439.
SØRENSEN, A., und SØRENSEN, A. B.: „An event history analysis of the process of entry into first marriage". In: KERTZER, D. I.: Family relations in life course perspective. New York 1984 (im Druck).
SØRENSEN, A. B.: „A model and a metric for the analysis of the intragenerational status attainment process". In: American Journal of Sociology, 1979, 85, S. 361–384.
DERS.: „Interpreting time dependency in career processes". In: DIEKMANN, A., und MITTER, P. (Hrsg).: Stochastic modelling of social processes. New York 1984, S. 89–122.
SØRENSEN, A. B., und SØRENSEN, A.: Modeling interdependence of life course events with event history data. Paper prepared for the meeting of the American Sociological Association. Detroit, 2. September 1983.
SØRENSEN, A. B., und TUMA, N. B.: „Labor market structures and job mobility". In: Research in Social Stratification and Mobility, 1981, 1, S. 67–94.
SPENCE, A. M.: „Job market signaling". In: Quartely Journal of Economics, 1973, 87, S. 355–374.
DERS.: Market signaling. Cambridge, Mass. 1974.
STOER, J.: Einführung in die numerische Mathematik. 2. Aufl., Berlin 1976.
STROHMEIER, K. P., SCHULTZ, M., und KAUFMANN, F.-X.: „Modellierung und Mikrosimulation von Prozessen der Familienentwicklung. Bericht aus dem Projekt „Generatives Verhalten in Nordrhein-Westfalen". In: SCHMID, J., und SCHWARZ, K. (Hrsg.): Politische und prognostische Tragweite von Forschungen zum generativen Verhalten. Berlin 1985.
TARONE, R. E., und WARE, J.: „On distribution-free tests for equality of survival distributions". In: Biometrika, 1977, 64, S. 156–160.
TÖLKE, A.: Zuverlässigkeit retrospektiver Verlaufsdaten – Qualitative Ergebnisse einer Nachbefragung. Arbeitspapier Nr. 30 des SfB 3 „Mikroanalytische Grundlagen der Gesellschaftspolitik". Frankfurt a. M. und Mannheim 1980.

TREIMAN, D. J.: Occupational prestige in comparative perspective. New York 1977.
TSIATIS, A. A.: „A large sample study of Cox's regression model". In: Annals of Statistics, 1981, 9, S. 93–108.
TUMA, N. B.: „Effects of labor market structure on job-shift patterns". In: HECKMAN, J. J., und SINGER, B. (Hrsg.): Longitudinal analysis of labor market data. Cambridge, Mass. 1985.
TUMA, N. B., und HANNAN, M. T.: Social dynamics: Models and methods. New York 1984.
TUMA, N. B., HANNAN, M. T., und GROENEVELD, L. P.: „Dynamic analysis of event histories". In: American Journal of Sociology, 1979, 84, S. 820–854.
VAUPEL, J. W., MANTON, K. G., und STALLARD, E.: „The impact of heterogeneity in individual frailty on the dynamics of morality". In: Demography, 1979, 16, S. 439–454.
WAGNER, M.: Bildung und Migration. Manuskript. Berlin: Max-Planck-Institut für Bildungsforschung 1986.
WEGENER, B.: „Gibt es Sozialprestige?" In: Zeitschrift für Soziologie, 1985, 14, S. 209–235.
WITTING, H., und NÖLLE, G.: Angewandte mathematische Statistik. Stuttgart 1970.

Einige wichtige Programmpakete

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

GLIM Generalized Linear Interactive Modeling
Numerical Algorithms Group
7 Banbury Road
Oxford OX2 6NN
England

RATE Invoking RATE
Tuma, N.B.
Sociology Department
Stanford University
Stanford, CA 94305
USA

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

SIR Scientific Information Retrieval
SIR Inc.
P.O. Box 1404
Evanston, IL 60204
USA

SPSS Statistical Package for the Social Science
SPSS Inc.
Suite 3300
444 North Michigan Avenue
Chicago, IL 60611
USA

Register