ZUR ANALYSE VON ZEITINTERVALLEN UNTER BERÜCKSICHTIGUNG UNBEOBACHTETER HETEROGENITÄT

Anwendungen auf Rückfallintervalle nach der Haftentlassung und die Dauer der Arbeitslosigkeit

Andreas DIEKMANN

Forschungsbericht/
Research Memorandum No. 197
Februar 1984

*) Für kritische Hinweise und Anmerkungen bedanke ich mich bei Peter Mitter, Iain Paterson, Victoria Poupko und Camillo Signor.

Die in diesem Forschungsbericht getroffenen Aussagen liegen im Verantwortungsbereich des Autors und sollen daher nicht als Aussagen des Instituts für Höhere Studien wiedergegeben werden.

INHALTSVERZEICHNIS

Abstract

Transition rate models are very appropriate for the analysis of social processes as demonstrated by the work of Coleman, Hannan, Tuma and other authors. In recent years these models became increasingly popular in empirical social research concerned with time related data. In particular, there is a great need for techniques which are able to deal with the problem of observed and unobserved heterogeneity. In this paper some characteristics of a model with unobserved heterogeneity are analyzed. It is assumed that the hazard rate depends on a gamma distributed error term. The model is applied to recidivism and unemployment data and compared with alternative hazard rate models. Estimation of parameters is performed with Tuma's program RATE. Important implications of the model are derived which allow clear and unambigious interpretations of estimated parameters.

Zusammenfassung

Wie die Arbeiten von Coleman, Hannan, Tuma und anderen Autoren demonstrieren, erweisen sich Übergangsraten-Modelle zur Analyse zeitlicher Abläufe als besonders zweckmäßig. Diese Modelle wurden in jüngster Zeit in wachsendem Maße zur Untersuchung von Lebensverläufen und sozialen Prozessen herangezogen. Dabei benötigt man insbesondere Verfahren, die der Heterogenität der untersuchten Population in geeigneter Weise Rechnung tragen. In dem vorliegenden Aufsatz werden die Eigenschaften eines Modells mit sogenannter unbeobachteter Heterogenität untersucht, wobei angenommen wird, daß die Übergangsrate u.a. von einem gamma-verteilten Fehlerterm abhängig ist. Anhand von Strafvollzugsdaten und Arbeitslosigkeitsdaten werden Anwendungen des Modells sowie alternativer Modelle demonstriert. Die Schätzung der Parameter erfolgt mit Tumas Programm RATE. Weiterhin werden verschiedene Modellkonsequenzen aufgezeigt, die eine eindeutige und anschauliche Interpretation der geschätzten Parameter erlauben.

1. Beobachtete und unbeobachtete Heterogenität

Zur Analyse von Zeitintervallen mit stochastischen Modellen in heterogenen Populationen wird häufig die folgende Strategie eingeschlagen: Man spezifiziert eine Gleichung für die Hazardrate in Abhängigkeit von bestimmten Kovariaten und eventuell auch von der Verweildauer. Sodann wird mit einer geeigneten Methode (z.B. Maximum-Likelihood) anhand der beobachteten Zeitintervalle die Stärke der Effekte der Kovariate auf die Hazardrate geschätzt. Das Resultat ist ein "gemischter" Markov- oder Semi-Markov-Prozeß. Jeder Kovariaten-Konstellation entspricht eine Hazardrate, die den Prozeß (d.h. die erwartete Verteilung der Ankunftszeiten) eindeutig bestimmt. Bei diesem Modell wird der beobachteten Heterogenität durch die Einführung von Kovariaten Rechnung zu tragen versucht. Beispiele hierfür sind die Arbeiten von Tuma (1976), Tuma, Hannan und Groeneveld (1979), Hannan und Carroll (1981) oder Sørensen (1984).

Durch die Berücksichtigung von Kovariaten wird jedoch in der Regel die Heterogenität der untersuchten Population nicht vollständig erfaßt. Neben der beobachteten kann auch unbeobachtete Heterogenität auftreten. Dies ist immer dann der Fall, wenn die Rate nicht eindeutig durch die Kovariate determiniert wird. Als unangenehme Folge vernachlässigter Heterogenität ergeben sich aber scheinbare Verweildauerabhängigkeiten. Unter den Ökonomen haben insbesondere Flinn und Heckman (1982) sowie Heckman und Singer (1982) auf diesen Umstand aufmerksam gemacht und realistischere Modelle unter Einfluß unbeobachteter Heterogenität vorgeschlagen.

In der Soziologie ist die Anwendung derartiger Modelle noch sehr selten, obwohl Tumas (198o) Programm RATE Schätzverfahren zur Berücksichtigung eines bestimmten Typs unbeobachteter Heterogenität zur Verfügung stellt. Eine Anwendung stellt Tumas (1982) Arbeit über Berufsmobilität dar.

Die Annahme unbeobachteter Heterogenität impliziert, daß die Rate nicht mehr als Konstante, sondern selbst als Zufallsvariable aufgefaßt wird. Diese Annahme finden wir auch schon in einem der ältesten stochastischen Prozesse, dem Yule-Greenwood-Prozeß, realisiert (siehe z.B.Chiang 1968). Greenwood und Yule sind davon ausgegangen, daß die Hazardrate in der Population gamma-verteilt ist, wobei jede Ausprägung der Hazardrate einen Poisson-Prozeß unterschiedlicher Intensität definiert. Das mit RATE schätzbare Modell ist im Prinzip eine Generalisierung des Yule-Greenwood-Prozesses, wobei sich die Generalisierung auf die Einbeziehung von Kovariaten bezieht.

Unbeobachtete Heterogenität kann - in Analogie zur Regressionsanalyse - durch Einführung eines Fehlerterms ε in der Ratengleichung berücksichtigt werden. Das allgemeine Modell mit beobachteter und unbeobachteter Heterogenität sowie Verweildauerabhängigkeit hat dann die folgende Form:

(1) $r = f$ (Kovariate, Verweildauer, ε)

Dabei steht "r" für die Hazardrate oder kurz "Rate".
In diesem Aufsatz diskutieren wir zwei wichtige Spezialfälle. In Teil 2 wird ein Modell mit ausschließlich unbeobachteter Heterogenität analysiert. Aus dem Yule-Greenwood-Prozeß wird die Ankunftszeitenverteilung abgeleitet und gezeigt, wie die Parameter mit RATE geschätzt werden können. Mit dem Modell werden Rückfalldaten untersucht, die nur in tabellarischer Form und ohne Kovariateninformation vorliegen. Sofern von der Annahme der Heterogenität ausgegangen wird - und dafür sprechen theoretische Gründe - zwingt die spezielle Datenlage dazu, Heterogenität nur in unbeobachteter

Form (d.h. als Fehlerterm) zuzulassen. Nach einem Vergleich von vier verschiedenen Hazardraten-Modellen in Teil 3 wird in Teil 4 der Arbeit die Erweiterung des Ansatzes unter Einschluß von Kovariaten anhand der Analyse von Daten der Arbeitslosigkeitsdauer diskutiert.

Das Hauptziel der vorliegenden Arbeit ist darin zu sehen, Konsequenzen des Modells mit gamma-verteiltem Fehlerterm einer genauen Analyse zu unterziehen. Insbesondere fragt es sich, wie die geschätzen Parameter und Koeffizienten zu interpretieren sind und welche Schlüsse aus dem Modell gezogen werden können. Darüber hinaus ist es auch Zweck der Arbeit, darauf aufmerksam zu machen, daß ein Modell mit gamma-verteilter Hazardrate einen geeigneten Kandidaten zur Analyse von Zeitintervallen bis zum Eintreffen des Ereignisses "abweichendes Verhalten" darstellt (siehe auch Diekmann, 1981). Rückfalldaten werden im deutschsprachigen Raum selten methodisch adäquat ausgewertet. Obwohl die Probleme der Datenanalyse formal ähnlich der Problemlage von Biometrikern und Medizinstatistikern sind, haben die in diesen Gebieten hoch entwickelten Methoden der Survival-Analyse m.W. noch nie in einer deutschsprachigen Rückfallstudie Verwendung gefunden. Auch wird die Debatte über die Anwendung geeigneter Survival-Modelle in der amerikanischen Kriminologie (z.B. Barton und Turnbull 1979, Harris und Moitra 1978, Lloyd und Joe 1979) von empirisch arbeitenden Kriminologen im deutschsprachigen Raum vollständig ignoriert. Dies ist um so unverständlicher, als die Survival-Analyse bei geeigneter Datenlage für zahlreiche Probleme bei der Auswertung von Rückfalldaten eine Lösung anbieten kann. Es sei hier nur an das Problem zensierter Daten, an die Schwierigkeit der Konstruktion geeigneter Rückfallmaße, an das Problem der Bewertung des Kausaleinflusses unabhängiger Variablen und an das Problem der Aufstellung von Prognosen über den Beobachtungszeitraum hinaus erinnert.

2. Die Analyse von Rückfalldaten unter der Annahme gamma-verteilter Hazardraten

2.1. Das Modell

Wir gehen von einem Modell des Typs $r = f(\varepsilon) = A\varepsilon$ aus, d.h. wir untersuchen ausschließlich unbeobachtete Heterogenität, wobei der Fehler ε und damit auch die Rate gamma-verteilt ist und A eine Konstante darstellt. Bezeichnen wir mit f(t) die Dichteverteilung der Ankunftszeit, d.h. des Zeitintervalls bis zu einem eventuellen Rückfall nach Haftentlassung, und mit G(t) die Überlebensfunktion, so gilt für das momentane Risiko eines Rückfalls: [1)]

$$(2) \qquad r = \frac{f(t)}{G(t)}$$

Grob gesprochen drückt die Rate den Anteil derjenigen Personen, die im "nächsten Moment" den Zustand wechseln (f(t)), an den "Überlebenden" G(t) bis zum Zeitpunkt t und damit das momentane Risiko eines Zustandswechsels (= Auftreten eines Ereignisses wie Rückfall) aus.

Wir nehmen ferner an, daß die Rate im Zeitablauf konstant ist, jedoch bei verschiedenen Personen unterschiedliche Werte aufweisen kann. Wie oben erwähnt, folgt die Rate einer Gamma-Verteilung:

$$(3) \qquad f(r) = \frac{\lambda^c \, r^{c-1} \, e^{-\lambda r}}{\Gamma(c)}$$

mit den Parametern λ und c. Die Gamma-Verteilung erfüllt die Bedingung, daß sie für positive Zufallsvariablen definiert ist. Zweitens handelt es sich um eine relativ allgemeine Verteilung, [2)] die sowohl eine J-förmige ($c < 1$) als auch eine unimodale Form annehmen kann. Drittens ist die Gamma-Verteilung mathematisch relativ gut handhabbar.

Bei normiertem Verhalten erscheint die Gamma-Verteilung aus inhaltlichen Gründen besonders angemessen zu sein. So wird bei kriminellen Delikten eine J-Verteilung vermutet, d.h.

extreme abweichende Reaktionen werden seltener erwartet (Allport 1934, Kaiser 1971, S. 11f.). Nach dieser Theorie sollte sich ein Paramterwert von $c < 1$ ergeben. Entsprechend Allports (1934) Bild einer Sanddüne, die je nach Stärke des sozialen Drucks mehr oder minder verformt wird, sinkt die Intensität einer Norm, wenn sich die Masse der Verteilung nach rechts verschiebt, d.h. wenn der Wert von c anwächst. Im Kontext normierten Verhaltens kann c als ein Maß der Normierung angesehen werden (dazu auch Hofstätter 1972, S. 41ff. und Diekmann 1981).

Bei festem Wert von r ist die Ankunftszeit exponential verteilt mit der Dichteverteilung:

$$(4)\quad g(t|r) = r.e^{-rt} .$$

Zur Ermittlung der unbedingten Dichteverteilung der Ankunftszeiten ist es erforderlich, $g(t|r)$ mit $f(r)$ zu multiplizieren und über alle Werte von r zu integrieren:

$$(5)\quad f(t) = \int_0^\infty g(t|r).f(r)dr = \frac{\lambda^c}{\Gamma(c)} \int_0^\infty r^c e^{-r(\lambda+t)} dr .$$

Die Lösung des Integrals liefert uns die gesuchte Dichte des "zusammengesetzten Prozesses" $f(t)$. Durch Integration von $f(t)$ erhält man die kumulierte Verteilung der Ankunftszeit $F(t)$ und hieraus aufgrund der Beziehung $G(t) = 1-F(t)$ die Überlebensfunktion sowie durch Division von (6) und (8) die Hazardfunktion des zusammengesetzten Prozesses (9): [3)]

$$(6)\quad f(t) = \frac{c\lambda^c}{(\lambda+t)^{c+1}}$$

$$(7)\quad F(t) = 1 - \left(\frac{\lambda}{\lambda+t}\right)^c$$

$$(8)\quad G(t) = \left(\frac{\lambda}{\lambda+t}\right)^c$$

$$(9)\quad r(t) = \frac{f(t)}{G(t)} = \frac{c}{\lambda+t} .$$

Gleichung (9) macht auf einen wichtigen Punkt aufmerksam. Obwohl die in der Population verteilten Hazardraten zeitunabhängig sind, ergibt sich für den beobachtbaren zusammengesetzten Prozeß eine im Zeitablauf fallende Hazardfunktion (Abbildung 1).

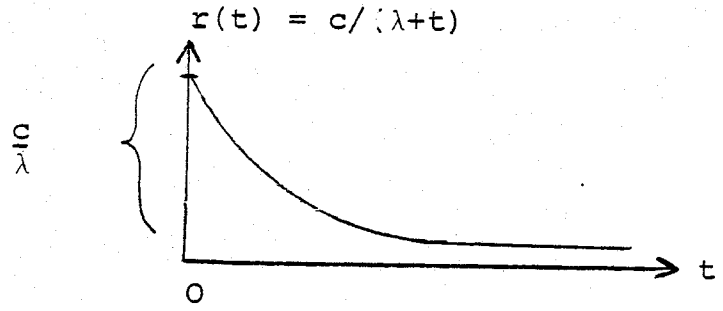

Abbildung 1: Hazardfunktion des zusammengesetzten Prozesses

Nur für t = 0 entspricht r(0) genau dem Erwartungswert c/λ der gamma-verteilten individuellen Raten. Intuitiv läßt sich das Ergebnis leicht deuten. Bei Personen mit hohen r-Werten tritt in einem Zeitintervall ein Ereignis mit größerer Wahrscheinlichkeit auf als bei Personen mit geringer Hazardrate. Die ersteren Fälle scheiden früher aus dem Prozeß aus, so daß die zusammengesetzte Hazardrate r(t) für die Gruppe der verbleibenden Personen absinkt.

Verfügt man nur über "Makro-Daten" auf der Ebene des zusammengesetzten Prozesses, d.h. kennt man nur die Ankunftszeitenverteilung für die untersuchte Population, so könnte man fälschlicherweise auf eine Verweildauerabhängigkeit der Hazardrate bei den einzelnen Mitgliedern der Population schließen. Anders formuliert sind auf der Ebene der globalen Verteilung Verweildaueреffekte (Zeitabhängigkeit der Rate) nicht von dem Effekt der Heterogenität zu trennen.

2.2. Die Anwendung des Modells auf Rückfalldaten

In einer Untersuchung von Berckhauer und Hasenpusch (1982) werden die Rückfallintervalle für eine 1974er Kohorte von 520 Strafentlassenen über einen Beobachtungszeitraum von 60 Monaten berichtet, wobei als Rückfallkriterium eine erneute Verurteilung herangezogen wurde.

Sterbetafel-Analyse der Rückfalldaten

Die nach den prozentualen Angaben von Berckhauer und Hasenpusch (1982, S. 301) umgerechneten Häufigkeiten gehen aus der zweiten Spalte von Tabelle 1 hervor.[4] (Siehe Tabelle 1, Seite 8.)

Die nicht-parametrische "Sterbetafel"-Analyse der gruppierten Zeitintervalle ist vor allem auch für explorative Zwecke sinnvoll. Man erkennt, daß entsprechend der Modell-Implikation (9) die nicht-parametrisch geschätzte Hazardrate $\hat{r}_i$ (Spalte 6 der Tabelle) im Zeitablauf abnimmt.

Die einfache Analyse in Form einer "Sterbetafel" liefert auch dem Kriminologen wertvolle Informationen, zumal zensierte Daten in statistisch angemessener Weise verarbeitet werden.[5] Ferner kann der Median (= 18,82 Monate) als robustes und vergleichbares Rückfallmaß dienen, und schließlich stellt die Survival-Analyse inferenzstatistische Tests zum Vergleich von Rückfallverteilungen - z.B. Sozialtherapie versus bedingte Entlassung versus Entlassung nach voller Strafhaft - zur Verfügung.

Parameterschätzung mit RATE

Zur Schätzung der Parameter λ und c des Modells mit unbeobachtbarer Heterogenität benutzen wir das Programm RATE von Tuma (1980). Dazu ist es zunächst erforderlich, eine Verbindung zwischen den mit der Maximum-Lixelihood-Methode geschätzten Größen des Programms und den Parametern λ und c herzuleiten.

Zeitintervall Monate ( ]	n_i'	d_i	$\hat{q}_i = \frac{d_i}{n_i'}$	$\hat{p}_i = 1-\hat{q}_i$	$\hat{r}_i = \frac{2\,\hat{q}_i}{6(1+\hat{q}_i)}$	$\hat{G}_i$
- 6	504	145	0,29	0,71	0,057	0,71
6 - 12	359	67	0,19	0,81	0,035	0,58
12 - 18	292	37	0,13	0,87	0,023	0,51
18 - 24	255	22	0,09	0,91	0,016	0,46
24 - 30	233	30	0,13	0,87	0,023	0,40
30 - 36	203	11	0,05	0,95	0,009	0,38
36 - 42	192	11	0,06	0,94	0,010	0,36
42 - 48	181	15	0,08	0,92	0,014	0,33
48 - 54	166	11	0,07	0,93	0,012	0,31
54 - 60	155	7	0,05	0,95	0,009	0,29
> 60	148					

Median = 18,82 Monate

n_i' = Risikomenge (Personen, die zu Beginn des Intervalls i keinen Rückfall aufweisen)

d_i = Rückfälle im Intervall i

$\hat{q}_i$ = Bedingte Wahrscheinlichkeit eines Rückfalls im Intervall i

$\hat{p}_i$ = Bedingte "Überlebenswahrscheinlichkeit" im Intervall i

$\hat{r}_i$ = Hazardrate im Intervall i

$\hat{G}_i$ = Überlebenswahrscheinlichkeit bis zum Ende des Intervalls i

Tabelle 1: "Sterbetafel" der Häufigkeiten von Rückfällen

Mit Modell Typ 2 des RATE-Programms kann die folgende Ratengleichung formuliert werden:

$$(10) \quad r = A\,\varepsilon \ .$$

Hierbei ist ε ein gamma-verteilter Fehlerterm mit dem Erwartungswert 1 und der Varianz:

$$(11) \quad \mathrm{Var}(\varepsilon) = B \ .$$

A und B spezifieren wahlweise lineare oder log-lineare Effekte der Kovariate. In unserem Zusammenhang interessieren nur die Konstanten der Kovariatenvektoren, so daß bei log-linearer Modellwahl:

$$(12) \quad A = e^{\alpha} \ ,$$
$$(13) \quad B = e^{\beta}$$

mit den zu schätzenden Paramtern α und β gilt. Für den Mittelwert und die Varianz der gamma-verteilten Variablen r folgt:

$$(14) \quad E(r) = A\,E(\varepsilon) = e^{\alpha}$$
$$(15) \quad \mathrm{Var}(r) = A^2\mathrm{Var}(\varepsilon) = e^{2\alpha+\beta}$$

Unter Berücksichtigung des Erwartungswertes:

$$(16) \quad E(r) = \frac{c}{\lambda}$$

und der Varianz:

$$(17) \quad \mathrm{Var}(r) = \frac{c}{\lambda^2}$$

der Gamma-Verteilung (3) können die Paramter λ und c in Abhängigkeit von α und β ausgedrückt werden:

$$(18) \qquad \lambda = e^{-(\alpha+\beta)}$$

$$(19) \qquad c = e^{-\beta} = \frac{1}{Var(\varepsilon)}$$

Die Daten in Tabelle 1 sind gruppiert nach Sechs-Monats-Intervallen. Zur Maximum-Likelihood-Schätzung von α und ß mit RATE sind jedoch die "exakten" Ankunftszeiten erforderlich. Da die Rohdaten nicht vorliegen, wird jeweils die "exakte" Ankunftszeit durch die Intervallmitte geschätzt (3, 9, 12, ..., 57 Monate). [6] Die Ergebnisse der Berechnung enthält Tabelle 2.

	Schätzwerte der Parameter
$\hat{\alpha}$	- 2,771 (o,1o46)*
$\hat{\beta}$	o,433 (o,1157)*
$\hat{\lambda}$	1o,358
$\hat{c}$	o,648
$\hat{E}(r)$	o,o63
$\hat{Var}(\varepsilon)$	1,543
N	5o4

* Standardfehler

Tabelle 2: Geschätzte Parameter

Es zeigt sich, daß der Wert von $\hat{c}$ kleiner als eins ist; die Gamma-Verteilung der individuellen Hazardraten hat demnach - die Angemessenheit des Modells unterstellt - die Form einer J-Verteilung.

Mit Hilfe der Gleichungen (6) - (8) und den geschätzten Werten für λ und c kann ferner die erwartete Verteilung der Ankunftszeit sowie die Überlebensfunktion $\hat{G}(t)$ berechnet werden. Die erwarteten Häufigkeiten ($N \cdot \hat{G}(t)$) sind Tabelle 4 in Abschnitt 3 zu entnehmen.

Die erwartete Hazardfunktion kann mittels Gleichung (9) prognostiziert werden:

$$(2o)\quad \hat{r}(t) = \frac{\hat{c}}{\hat{\lambda}+t} = \frac{o,648}{1o,358+t}$$

Abbildung 2 stellt den Verlauf der Funktion im Vergleich mit den "beobachteten" nicht-parametrisch geschätzten Hazardraten dar.

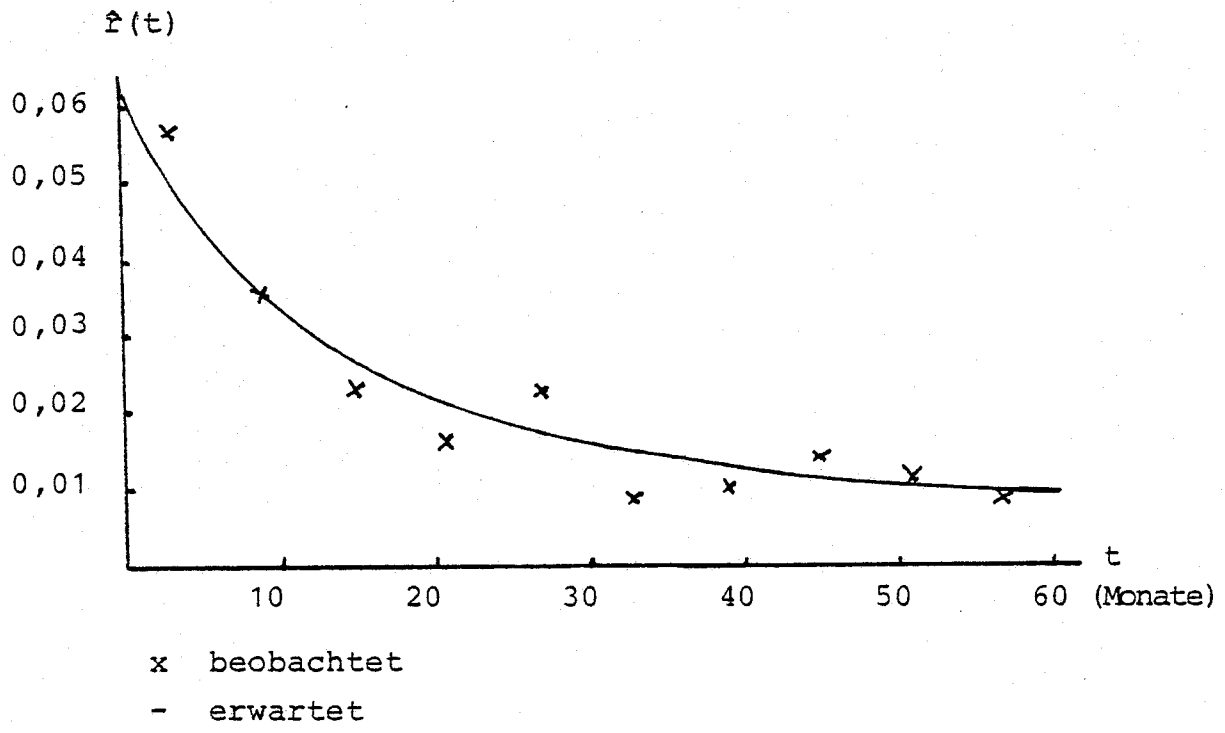

Abbildung 2: "Beobachtete" und erwartete Hazardraten des zusammengesetzten Prozesses.

Wie man sieht, ist die Übereinstimmung ziemlich gut. Freilich ist zu beachten, daß - wie oben erwähnt - ein ähnlicher Verlauf der Ratenfunktion auch aus Modellen mit negativer Verweildauerabhängigkeit der Hazardrate abgeleitet werden kann.

2.3. Einige Folgerungen aus dem Modell

Ein Vorteil parametrischer Modellbildung ist darin zu sehen, daß das Modell die Ableitung verschiedener Charakteristika des untersuchten Prozesses gestattet. In diesem Abschnitt wollen wir kurz einige Folgerungen bezüglich des Erwartungswerts der Ankunftszeiten, der Varianz von T und des Medians erörtern. Diese Beziehungen sind auch von Bedeutung für die Interpretation der Koeffizienten des Modells mit Kovariaten in Abschnitt 4.

Für den Erwartungswert $\overline{T}$ erhält man nach Einsetzen von (6) für f(t) als Lösung des Integrals:

$$(21)\ \overline{T} = \int_0^\infty t \cdot f(t)\,dt = \frac{\lambda}{c-1}\ ,\ c > 1.$$

$\overline{T}$ existiert nur für $c > 1$, ist also bei den Rückfalldaten nicht berechenbar. Ähnliches gilt für die Varianz:

$$(22)\ \mathrm{Var}(T) = \frac{c}{c-2}\ \overline{T}^{\,2}\ ,\quad c > 2.$$

Durch Umformung von (21) und (22) unter Benutzung von (16) und (19) erkennt man den Einfluß der Fehlervarianz Var(ε):

$$(23)\ \overline{T} = \frac{1}{E(r)\left[1 - \mathrm{Var}(\varepsilon)\right]}\ ,\ 0 \leq \mathrm{Var}(\varepsilon) < 1$$

$$(24)\ \mathrm{Var}(T) = \frac{1}{1 - 2\,\mathrm{Var}(\varepsilon)}\ \overline{T}^{\,2},\ 0 \leq \mathrm{Var}(\varepsilon) < \frac{1}{2}\ .$$

Vorausgesetzt, Mittelwert und Varianz der Verteilung existieren, so beeinflußt wachsende Fehlervarianz beide Größen in positiver Weise. Für den Extremfall einer Fehlervarianz von null liefern die Formeln (23) und (24) den Mittelwert und die Varianz der Exponentialverteilung. Liegt keine Heterogenität vor ($\mathrm{Var}(\varepsilon) = 0$), so erhalten wir als Spezialfall den einfachen Poisson-Prozeß.

Zur Berechnung des Medians T^* greifen wir auf Gleichung (8) zurück:

$$(25)\quad G(T^*) = \left(\frac{\lambda}{\lambda + T^*}\right)^c = \frac{1}{2} \quad .$$

Hieraus folgt:

$$(26)\quad T^* = \lambda\left[2^{\frac{1}{c}} - 1\right] = \frac{2^{Var(\varepsilon)} - 1}{E(r)Var(\varepsilon)} \quad .$$

Auch der Median verschiebt sich nach "rechts" bei wachsender Fehlervarianz. Zunehmende Heterogenität erhöht also den Mittelwert, die Varianz und den Median der erwarteten Verteilung.

Bei den Rückfalldaten stimmt der erwartete Median, berechnet nach Formel (26), mit einem Wert von 19,81 Monaten recht gut mit den Beobachtungen überein.

Betrachten wir noch unter Berücksichtigung von (21) und (26) das Verhältnis des Medians zum Mittelwert:

$$(27)\quad \frac{T^*}{\bar{T}} = \left(2^{\frac{1}{c}} - 1\right)(c-1) < 1.$$

Aus (27) folgt $T^* < \bar{T}$. Wie bei rechtsschiefen Verteilungen nicht anders zu erwarten, liegt der Median vor dem Mittelwert. Im Extremfall der Fehlervarianz von null ($Var(\varepsilon) \to 0$ bzw. $c \to \infty$) erhalten wir aus (27):

$$(28)\quad \lim_{c \to \infty} \frac{T^*}{\bar{T}} = \ln 2 = 0,693,$$

was genau dem Quotienten aus Median und Mittelwert im Fall des einfachen Poisson-Prozesses (Exponentialverteilung) entspricht.

3. Ein Vergleich von vier Hazardraten-Modellen

Neben dem Poisson-Modell (Exponentialverteilung der Ankunftszeiten) als Basis-Zufallshypothese berücksichtigen wir bei dem Vergleich drei weitere Modelle, die folgende Gemeinsamkeiten aufweisen: Es handelt sich jeweils um zwei-parametrige Modelle, mit denen ein monoton fallender Verlauf der Hazardfunktion nachgebildet werden kann. Während jedoch der Weibull- und Gampertz-Prozeß von der Hypothese "echter" negativer Verweildauerabhängigkeit ausgeht, ist beim Heterogenitätsmodell die im Zeitablauf abnehmende Hazardrate Resultat einer heterogenen Population.

Tabelle 3 informiert über die alternativen Ratenfunktionen, die Überlebenskurven der Modelle und die geschätzten Parameterwerte.

Der graphische Log-minus-log-Test liefert auch im Falle des Weibull-Modells eine gute Übereinstimmung mit den Daten [7] (Abbildung 3).

	Exponential (Poisson)	Weibull	Gompertz	Heterogenität (Gamma)
G(t)	$\exp(-rt)$	$\exp\{-(\lambda t)^c\}$	$\exp\{-\frac{\lambda}{c}(e^{ct}-1)\}$	$\left(\frac{\lambda}{\lambda+t}\right)^c$
r(t)	r	$\lambda c(\lambda t)^{c-1}$	λe^{ct}	$\left(\frac{c}{\lambda+t}\right)$
Parameter*	$\hat{r}$= o,o251	$\hat{\lambda}$ = o,o265 $\hat{c}$ = o,55	$\hat{\lambda}$ = o,o52 $\hat{c}$ =-o,o377	$\hat{\lambda}$=10,36 $\hat{c}$=0,648
erwarteter Median	27,6	19,4	18,6	19,8
"Beobachteter" Median	18,82 *			

* Die Parameter wurden wie folgt geschätzt: Im Poisson-Fall ist der ML-Schätzer $\hat{r} = n/\Sigma t_i$, d.h. die Anzahl der Rückfälle wird durch die Summe der beobachteten und der zensierten Rückfallzeiten dividiert. Die Weibull-Parameter wurden grob mit einem graphischen Verfahren abgeschätzt (siehe Diekmann und Mitter 1984, S. 154ff.). Zur Schätzung der Parameter des Gompertz- und des Gamma-Modells wurde das Programm RATE (Tuma 198o) benutzt. Als "exakte" Ankunftszeit wurde jeweils die Intervallmitte genommen.

Tabelle 3: Geschätzte Parameter von vier Hazardraten-Modellen

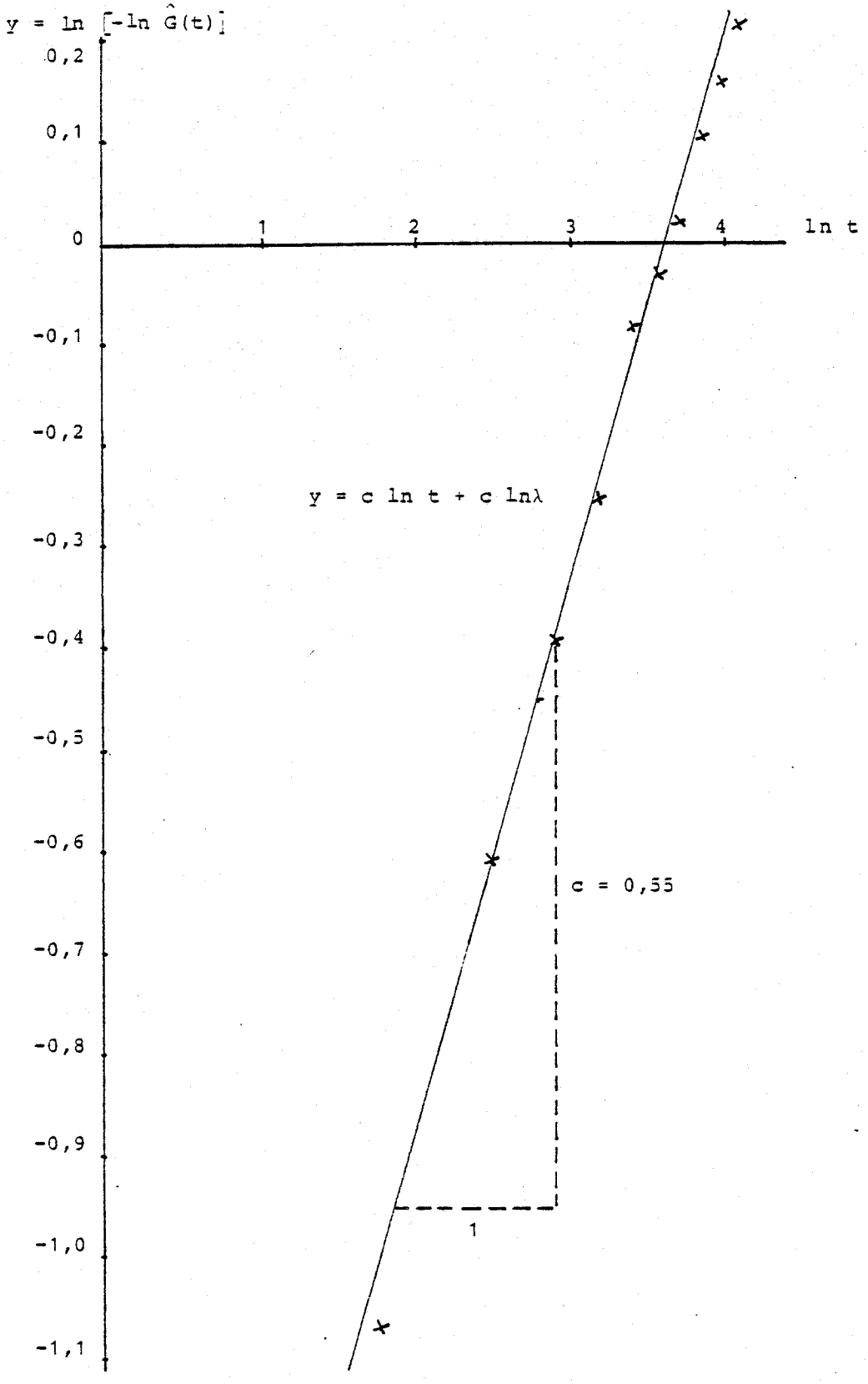

Abbildung 3: Log-minus-log-Test des Weibull-Modells

Die prognostizierten Überlebenshäufigkeiten $N.\hat{G}(t)$ für die vier Modelle sind Tabelle 4 zu entnehmen.

Intervall-ende	beob-achtet	exponential (Poisson)	Weibull	Gompertz	Hetero-genität (Gamma)
6	145	70	154	123	129
12	67	61	54	76	69
18	37	52	37	49	44
24	22	45	28	34	30
30	30	39	22	24	23
36	11	33	19	17	18
42	11	28	15	12	15
48	15	25	14	10	12
54	11	21	12	7	10
60	7	18	10	5	8
> 60	148	112	139	147	146
N	504	504	504	504	504
$\Sigma \lvert \text{beob.} - \text{erwart.} \rvert$		234	62	80	58
$\text{Chi}^2 = \Sigma \frac{(\text{beob.}-\text{erwart.})^2}{\text{erwart.}}$ mit df = 10		151	13,92	21,47	12,21

Tabelle 4: Beobachtete und erwartete Häufigkeiten auf der Basis von vier alternativen Survival-Modellen.

Im Vergleich mit der beobachteten Überlebenshäufigkeit zeigt sich eine extrem schlechte Anpassung für das Poisson-Modell. Die beobachteten Häufigkeiten werden zu Beginn des Prozesses kraß unter- und später überschätzt. Auch der Median (Tab. 3) ist wesentlich größer als der beobachtete Median. Natürlich ist dieses Ergebnis nicht verwunderlich in Anbetracht der sinkenden nicht-parametrisch geschätzten Hazardrate in Abbildung 2, setzt doch die Nullhypothese des Poisson-Modells eine konstante Hazardrate voraus.

Für die drei übrigen Modelle ergibt sich eine wesentlich bessere Übereinstimmung. Nimmt man als Kriterium für die Übereinstimmung die Summe der absoluten Abweichungen oder die Summe der durch die erwarteten Häufigkeiten dividierten quadratischen Abweichungen (Chi^2 mit df = 10), so schneidet das Heterogenitätsmodell am besten ab, wobei der Unterschied zum Weibull-Modell jedoch sehr gering ist. In beiden Fällen ist der Unterschied erwartete versus beobachtete Verteilung für p = 0,05 nicht signifikant. Zu berücksichtigen ist außerdem, daß die Weibull-Parameter graphisch geschätzt wurden und daher einen Bias aufweisen können. Aufgrund der vorliegenden Daten kann keine Entscheidung über die Gültigkeit der Heterogenitäts-Hypothese versus der Verweildauer-Hypothese des Weibull-Typs getroffen werden. Beide Hypothesen stimmen gleichermaßen mit den Daten überein. Dies bedeutet aber nicht, daß eine Diskriminierung zwischen den Hypothesen prinzipiell unmöglich wäre. Bleibt bei Berücksichtigung relevanter Kovariate, d.h. bei "Kontrolle" der Heterogenität, eine deutliche Zeitabhängigkeit der Hazardrate erhalten, so würde dieser Befund für "echte" Verweildaueffekte sprechen. Eine Sekundäranalyse der Rohdaten der Berckhauer-Hasenpusch-Studie könnte hierüber Aufschluß geben.

4. Kovariate und unbeobachtete Heterogenität am Beispiel von Arbeitslosendaten

Die bisherigen Überlegungen behalten auch bei der Generalisierung des Modells durch die Einführung von Kovariaten ($r = f$ (Kovariate, ε)) ihre Gültigkeit, wenn e^{α} und e^{β} - bei log-linearer Kovariatenabhängigkeit - durch die folgenden Ausdrücke ersetzt werden:

$$(29) \quad E(r) = A = e^{\alpha_o + \alpha_1 x_1 + \ldots + \alpha_m x_m}$$

$$(30) \quad \mathrm{Var}(\varepsilon) = B = e^{\beta_o + \beta_1 x_1 + \ldots \beta_m x_m}$$

Dabei bezeichnen $x_1, \ldots, x_m$ die Kovariate und α_i bzw. β_i die Koeffizienten. Als wichtiger Spezialfall ist ein Modell anzusehen, bei dem die Kovariate zwar den Erwartungswert der Rate beeinflussen, nicht jedoch die Fehlervarianz. Es gilt dann $\beta_1 = \beta_2 = \ldots = \beta_m = 0$ und somit $B = e^{\beta_o}$. Wie lassen sich in diesem Fall die Koeffizienten interpretieren?

In "Multiplikatoren-Schreibweise" lautet (29):

$$(31) \quad E(r) = a_o\, a_1^{x_1}\, a_2^{x_2} \ldots a_m^{x_m} .$$

$(a_i - 1) \cdot 100$ ist somit der %-Effekt einer Veränderung der unabhängigen Variable um eine Einheit auf den Erwartungswert der Rate.

Ein anschauliches Maß ist der Effekt auf die erwartete Verweildauer $\bar{T}$ und den Median T^*. Rufen wir uns hierzu die Formeln (23) und (26) ins Gedächtnis:

$$(32) \quad \bar{T} = \frac{1}{E(r)} \cdot \left[\frac{1}{1 - \mathrm{Var}(\varepsilon)} \right]$$

$$(33) \quad T^* = \frac{1}{E(r)} \cdot \left[\frac{2^{\mathrm{Var}(\varepsilon)} - 1}{\mathrm{Var}(\varepsilon)} \right] .$$

Die Ausdrücke in eckigen Klammern sind konstante Faktoren, falls Var(ε) von den Kovariaten unabhängig ist. Wie beim Modell mit konstanter Rate (einfacher Poisson-Prozeß)[8] informiert somit:

$$(34) \qquad \left(\frac{1}{a_i} - 1\right) \cdot 100$$

über den %-Effekt bei Veränderung der unabhängigen Variable um eine Einheit auf die mittlere Ankunftszeit (sofern $\bar{T}$ existiert) und auf den Median.

Unter Berücksichtigung von (29) und (30) sowie (16) und (17) ist es ferner möglich, mittels der Gleichungen (6) - (9) und (21) - (26) für beliebige Kombinationen von Kovariaten Voraussagen bezüglich der Verteilung der Ankunftszeiten, der Überlebensfunktion, der mittleren Ankunftszeit, dem Median und der Varianz zu formulieren.

Als Anwendungsbeispiel analysieren wir Daten über die Dauer der Arbeitslosigkeit einer österreichischen 1979er-Kohorte von 1055 Arbeitslosen, wobei der Zensierungszeitpunkt 140 Tage beträgt. Kovariate sind das Lebensalter (x_1), der Sozialstatus (x_2), das Geschlecht (x_3) und die Höhe der Arbeitslosenunterstützung (x_4). Es wird das folgende Modell mit beobachteter und unbeobachteter Heterogenität formuliert:

$$(35) \quad E(r) = e^{\alpha_o + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \alpha_4 x_4}$$

$$(36) \quad Var(\varepsilon) = e^{\beta_o}$$

Die mit RATE geschätzten Parameterwerte gehen aus Tabelle 5 hervor:

Tabelle 5: Parameter-Schätzwerte für die Dauer der Arbeitslosigkeit

	Koeffizienten $\hat{\alpha}_i, \hat{\beta}_o$	Standardfehler	$\hat{a}_i = e^{\hat{\alpha}_i}$, $\hat{Var}(\varepsilon) = e^{\hat{\beta}}$	%-Effekt auf E(r)	%-Effekt auf $\bar{T}$
Konstante	-3,744	0,144	0,0237	-	-
Alter (x_1)	-0,0313	0,0034	0,9692	- 3 %	+ 3 %
Sozialstatus* (x_2)	-0,740	0,086	0,4771	-52 %	+110 %
Geschlecht** (x_3)	0,147	0,077	1,1584	+16 %	- 14 %
Arbeitslosengeld (x_4)	$1,56 \cdot 10^{-4}$	$2,59 \cdot 10^{-5}$	1,00016	$+16 \cdot 10^{-3}$ %	$- 16 \cdot 10^{-3}$ %
	$\hat{\beta}_o$ = -2,673	0,475	Var(ε)=0,0691		

* Angestellter = 1, Arbeiter = 0

** Männer = 1, Frauen = 0

Alter in Jahren und Arbeitslosengeld in österr. Schillingen

Unterstrichene Werte sind signifikant (p = 0,05)

$E(r) = e^{(\alpha_o + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \alpha_4 x_4)}$

$Var(\varepsilon) = e^{\beta_o}$

Wie man der Tabelle entnehmen kann, ist die erwartete Dauer der Arbeitslosigkeit um so höher, je älter ein Arbeitsloser ist, wenn er Angestellter ist, weiblich und wenig Arbeitslosengeld erhält, wobei der Geschlechtseffekt allerdings einen knapp nicht-signifikanten Wert hat.

Vergleicht man die Ergebnisse mit den Schätzwerten für das Modell mit konstanter Rate (ohne unbeobachtete Heterogenität), so zeigen sich nur äußerst geringe Unterschiede. Gleiches gilt für die Schätzung mit dem semi-parametrischen Cox-Modell (siehe Diekmann und Mitter 1984, S. 116 und 137 ff.). Beim einfachen Poisson-Modell ergeben sich die folgenden Multiplikator-Werte $a_i : a_1 = 0,968$, $a_2 = 0,478$, $a_3 = 1,117$ und $a_4 = 1,00015$.

Bei unterschiedlicher Modellspezifikation (Cox-Regression, Poisson-Prozeß, unbeobachtete Heterogenität) sind die Schätzwerte ausgesprochen robust.

Es ist jedoch davor zu warnen, diesen Befund vorschnell zu generalisieren. Die nur geringen Unterschiede beim Heterogenitätsmodell gegenüber den Modellen ohne unbeobachtete Heterogenität rühren daher, daß die Fehlervarianz mit 0,0691 ($c = 14,47$) einen sehr kleinen Wert aufweist. Betrachten wir den Ausdruck in eckigen Klammern bei Formel (32) als Korrekturfaktor für die mittlere Ankunftszeit gegenüber dem Modell mit konstanter Rate (Poisson-Prozeß mit Kovariaten), so erhalten wir einen nur wenig von eins verschiedenen Wert von $1/(1-0,0691) = 1,07$. Im Gegensatz zu dem Modell mit ausschließlich unbeobachteter Heterogenität wird die vorhandene Heterogenität bei den Arbeitslosendaten offenbar weitgehend durch die Kovariate ausgeschöpft.[9)]

Anmerkungen

1) Zu den grundlegenden Definitionen und Ableitungen sei auf die Literatur der Survival-Analyse verwiesen. Siehe hierzu z.B. Diekmann und Mitter 1984, Kap. 2.

2) Z.B. enthält sie die Exponential- und die Chi^2- Verteilung als Spezialfälle.

3) Bei dem Yule-Greenwood-Prozeß ergibt sich bei gamma-verteiltem Parameter der Poisson-Verteilung für den zusammengesetzten Prozeß eine negative Binomialverteilung für die Zahl der Ereignisse k (Chiang 1968):

$$p_k(t) = \binom{k+c-1}{k} \left(\frac{t}{\lambda+t}\right)^k \left(\frac{\lambda}{\lambda+t}\right)^c$$

Für k = 0 erhält man auf einfache Weise die zugehörige Überlebensverteilung bezüglich des ersten Ereignisses: $p_o(t) = G(t)$. Wie man erkennt, stimmt $p_o(t) = (\lambda/\lambda+t)^c$ mit (8) überein.

4) Für 15 der 520 Personen ist nur bekannt, daß eine erneute Verurteilung vorliegt, nicht jedoch der Zeitpunkt des Rückfalls. Diese Fälle bleiben bei der nachfolgenden Analyse unberücksichtigt. Der Bias durch Weglassung von 15 Fällen, für die einzig die Information vorliegt, daß ein Rückfall innerhalb von 60 Monaten aufgetreten ist, dürfte sehr gering sein. Bei der Schätzung des Parameters der Exponentialverteilung (dazu Abschnitt 3) wurde zur Kontrolle jeweils die Annahme gemacht, daß alle 15 Fälle a) zum Zeitpunkt t = 0 und b) zum Zeitpunkt t = 60 ein Ereignis aufweisen. Der unverzerrte Schätzwert sollte dann zwischen den beiden extremen Schätzungen liegen. Tatsächlich ist der Unterschied zwischen der Behandlung der 15 Fälle nach Methode a) und b) relativ gering. Theoretisch bestünde

auch die Möglichkeit, die 15 Fälle gemäß einer bestimmten Verteilungsannahme auf die Zeitraum-Kategorien der Tabelle 1 aufzuteilen; praktisch dürfte jedoch der Mehraufwand kaum lohnenswert sein. Bleiben diese Fälle unberücksichtigt, so erhält man - bei zusätzlicher Einkalkulierung von Rundungsfehlern bei der Umrechnung der von Berckhauer und Hasenpusch angegebenen Prozentwerte - die 504 Fälle der Tabelle 1.

5) In der Tabelle 1 treten alle Zensierungen am Ende der Beobachtungszeit, also nach 60 Monaten auf. In diesem unproblematischeren Fall stimmt die jeweilige Risikomenge n_i' mit den Rohwerten n_i (Spalte 2) direkt überein. Bei zwischenzeitlichen Zensierungen ermöglicht die Survival-Analyse entsprechende Korrekturen. Siehe auch das Anwendungsbeispiel in Andreß 1982 sowie Diekmann und Mitter 1984, Kap. 3.

6) Dabei wird implizit die (dem Modell widersprechende) Annahme der Gleichverteilung innerhalb eines Sechs-Monats-Intervalls unterstellt. Der Fehler ist bei gruppierten Daten um so geringer, je weniger grob die Intervalleinteilung vorgenommen wurde.

   Zur Eingabe gruppierter Daten kann bei RATE eine WEIGHT-Karte geschrieben werden, wobei die Ankunftszeit mit der jeweiligen Häufigkeit gewichtet wird.

   Einem Hinweis von Peter Mitter folgend ist auch eine einfache graphische Test- und näherungsweise Schätzmethode auf der Basis der Hazardfunktion (9) konstruierbar. Demnach gilt: $1/r(t)=\lambda/c+(1/c).t$, d.h. als Graph der reziproken Hazardrate ist eine Gerade mit dem Anstieg $1/c$ zu erwarten. Nimmt man als Schätzung für $r(t)$ den "Life-Table-Schätzer" in Tabelle 1, wobei sich t jeweils auf die Mitte

des Intervalls i bezieht, so können die Parameter λ und c graphisch oder mit der Methode der kleinsten Quadrate grob geschätzt werden. Es zeigte sich, daß im vorliegenden Fall die Abweichung von der Maximum-Likelihood-Schätzung mit RATE nur geringfügig ist.

7) Die Überlebensfunktion des Weibull-Modells ist $G(t) = \exp\{-(\lambda t)^c\}$. Doppeltes Logarithmieren liefert: $\ln[-\ln G(t)] = c \ln t + c \ln\lambda$. Die beobachteten $\hat{G}(t)$-Werte in Tabelle 1 sollten also nach der $\ln[-\ln\hat{G}(t)]$- Transformation gegen ln t aufgetragen eine Gerade ergeben mit der Steigung c.

8) Zur Interpretation der Koeffizienten beim Modell mit konstanter Rate siehe Diekmann und Mitter 1984, S. 122.

9) Schätzt man das Modell in Abschnitt 2 mit ausschließlich unbeobachteter Heterogenität an den Arbeitslosendaten, d.h. werden die Kovariate nicht miteinbezogen, so erhält man als Fehlervarianz $\hat{Var}(\varepsilon) = \exp(\hat{\beta}) = 0{,}357$. Für den Erwartungswert der Rate ergibt sich ferner: $E(r) = \exp(\hat{\alpha}) = 0{,}0179$. Ohne Kovariate ist die Fehlervarianz also mehr als fünfmal so hoch. Aus den angegebenen Werten folgt mittels der Formeln in Abschnitt 2: $c = 2{,}80$; $\lambda = 156{,}49$. Die erwartete Verweildauer $\overline{T}$ beträgt 87 Tage und der prognostizierte Median T* liegt bei 44 Tagen. Durch die Einführung von Kovariaten wird eine proportionale Fehlerreduktion von $[\,Var(\varepsilon_o) - Var(\varepsilon_1)\,]/Var(\varepsilon_o) = (0{,}357 - 0{,}0691)/0{,}357 = 0{,}81$, also von 81% erzielt.

Literatur

ALLPORT, F.H.: The J-curve Hypothesis of Conforming Behavior, Journal of Social Psychology, Bd. 5, 1934.

ANDRESS, H.J.: Tätigkeitswechsel und Berufserfahrung. Analyse zeitbezogener Daten mit Hilfe von Sterbetafeln anhand eines Beispiels aus der Mobilitätsforschung, Zeitschrift für Soziologie, Jg. 11, 4/1982, S. 380 - 400.

BARTON, R.R. und TURNBULL, B.W.: Evaluation of Recidivism Data. Use of Failure Rate Regression Models, Evaluation Quarterly, Bd. 3, 1979, S. 629 - 642.

BERCKHAUER, F. und HASENPUSCH, B.: Legalbewährung nach Strafvollzug. Zur Rückfälligkeit der 1974 aus dem niedersächsischen Strafvollzug Entlassenen. In: H.D. Schwind und G. Steinhilper, Modelle zur Kriminalitätsvorbeugung und Resozialisierung, Heidelberg 1982, S. 281 - 319.

CHIANG, C.L.: Introduction to Stochastic Processes in Biostatistics, New York 1968.

DIEKMANN, A.: Ein einfaches stochastisches Modell zur Analyse von Häufigkeitsverteilungen abweichenden Verhaltens, Zeitschrift für Soziologie, Jg. 10, 1981, S. 319 - 325.

DIEKMANN, A. und MITTER, P.: Methoden zur Analyse von Zeitverläufen, Stuttgart 1984.

FLINN, Chr.J. und HECKMAN, J.J.: New Methods for Analyzing Individual Event Histories. In: S. Leinhardt, Hrsg., Sociological Methodology 1982, San Francisco 1982.

HANNAN, M.T. und CAROLL, G.R.: Dynamics of Formal Political Structure, American Sociological Review, Jg. 46, 1981, S. 19 - 35.

HARRIS, C.M. und MOITRA, S.O.: Improved Statistical Techniques for the Measurement of Recidivism, Journal of Research in Crime and Delinquency, 1978, S. 194 - 213.

HECKMAN, J.J. und SINGER, B.: Heterogeneity in Demographic Models. In: K. Land und A. Rogers, Hrsg., Multidimensional Mathematical Demography, S. 567 - 599, New York 1982.

HOFSTÄTTER, P.R.: Individuum und Gesellschaft, Frankfurt - Berlin - Wien 1972.

KAISER, G.: Kriminologie. Eine Einführung in die Grundlagen, Karlsruhe 1971.

LLOYD, M.R. und JOE, G.W.: Recidivism Comparisons across Groups. Methods of Estimation and Tests of Significance for Recidivism Rates and Asymptotes, Evaluation Quarterly, Bd. 3, 1979, S. 105 - 117.

SØRENSEN, A.B.: Interpreting Time Dependencies in Career Processes. In: A. Diekmann und P. Mitter, Hrsg. Stochastic Modelling of Social Processes, New York 1984.

TUMA, N.B.: Rewards, Resources, and the Rate of Mobility, American Sociological Review, Jg. 41, 1976, S. 338 - 360.

TUMA, N.B.: Invoking Rate, Arbeitspapier des Zentrums für Umfrageforschung in Mannheim (ZUMA), 1980.

TUMA, N.B.: Effects of Labor Market Structure on Job-Shift Patterns, unveröffentlichtes Manuskript, Stanford University, California, Dezember 1982.

TUMA, N.B.; HANNAN, M.T. und GROENEVELD, L.P.: Dynamic Analysis of Event Histories, American Journal of Sociology, Jg. 84, 1979, S. 820 - 854.