CARL VON OSSIETZKY universität OLDENBURG

Diplomstudiengang Mathematik

DIPLOMARBEIT

Statistische Modelle und Methoden in der Analyse von Lebenszeitdaten

vorgelegt von: Patricia Glomb

Betreuende Gutachterin: Prof. Dr. Christine Müller
Zweitgutachter: Prof. Dr. Dietmar Pfeifer

Oldenburg, 15. Juni 2007

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Symbol- und Abkürzungsverzeichnis

$:=$	definierende Gleichung
$\{\ldots\}$	Menge
$\emptyset$	leere Menge
$\mathbb{N}$	Menge der natürlichen Zahlen $\{1, 2, 3, \ldots\}$
$\mathbb{R}$	Körper der reellen Zahlen
$\mathbb{R}^p$	reeller Standardraum der Dimension $p \in \mathbb{N}$, $\mathbb{R}^p = \{\mathbf{x} = (x_1, \ldots, x_p) \mid x_i \in \mathbb{R},\ i \in \{1, \ldots, p\}\}$
$(a,\, b)$	für $a < b$ offenes Intervall $I = \{x \in \mathbb{R} \mid a < x < b\}$, sonst $I := \emptyset$
$[a,\, b]$	für $a < b$ abgeschlossenes Intervall $I = \{x \in \mathbb{R} \mid a \leq x \leq b\}$, sonst $I := \emptyset$
$(a,\, b]$	für $a < b$ links offenes, rechts abgeschlossenes Intervall $I = \{x \in \mathbb{R} \mid a < x \leq b\}$, sonst $I := \emptyset$
$[a,\, b)$	für $a < b$ links abgeschlossenes, rechts offenes Intervall $I = \{x \in \mathbb{R} \mid a \leq x < b\}$, sonst $I := \emptyset$
$\mathbf{x}$	Vektor (Spaltenform)
A	Matrix
$\mathbf{x}',\ A'$	Transponierte des Vektors $\mathbf{x}$ bzw. der Matrix A
$\forall$	für alle
$\exists$	es existiert
∂	partieller Ableitungsoperator
$\exp(\cdot)$	Exponentialfunktion
$\ln(\cdot)$	natürliche Logarithmusfunktion
$\log(\cdot)$	dekadische Logarithmusfunktion

$\mathbb{1}_A(\cdot)$	Indikatorfunktion einer Menge A
	($\mathbb{1}_A(x) = 1$, falls $x \in A$ und $\mathbb{1}_A(x) = 0$, falls $x \notin A$)
$\mathbb{1}(A)$	Indikatorfunktion einer Aussage A
	($\mathbb{1}(A) = 1$, falls A wahr und $\mathbb{1}(A) = 0$, falls A falsch)
$\text{sign}(\cdot)$	Vorzeichenfunktion
	($\text{sign}(x) = 1$, falls $x > 0$ und $\text{sign}(x) = -1$, falls $x < 0$)
$\lvert \cdot \rvert$	Betragsfunktion
	($\lvert x \rvert = x$, falls $x \geq 0$ und $\lvert x \rvert = -x$, falls $x < 0$)
$o(\cdot)$	Landau-Symbol, bezeichnet eine Funktion $g : \mathbb{R} \to \mathbb{R}$
	mit $\lim_{x \to 0}[g(x)/x] = 0$.
$\int_a^b g(u)\, du$	Riemann-Integral
$\int_{(a,\,b]} h(u)\, d\,G(u)$	Riemann-Stieltjes-Integral (S. 11)
$\mathcal{P}_a^b\,[\,1 + d\,G(u)\,]$	Produkt-Integral (S. 12)
T	Zufallsgröße
t	Realisierung der Zufallsgröße T
$\mathbf{T}$	n-dimensionaler Zufallsvektor $\mathbf{T} = (T_1, \ldots, T_n)'$
$\mathbf{t}$	Realisierung des Zufallsvektors $\mathbf{T}$
F_T	Verteilungsfunktion von T
f_T, p_T	stetige bzw. diskrete Dichte von T
$S_T(\cdot)$	Survival-Funktion zu $T \geq 0$
$\lambda_T(\cdot)$	Hazard-Funktion zu $T \geq 0$
$\Lambda_T(\cdot)$	kumulierte Hazard-Funktion zu $T \geq 0$
$mrl(\cdot)$	Erwartete Restlebensdauer zu $T \geq 0$
$\text{supp}\,(T)$	Träger von T
Δ	Zensurindikator (S. 36)
δ	Realisierung des Zensurindikators Δ
$T_n \xrightarrow{\mathcal{P}} T$	$(T_n)_{n \in \mathbb{N}}$ konvergiert für $n \to \infty$ stochastisch gegen T
$T_n \xrightarrow{\mathcal{D}} T$	$(T_n)_{n \in \mathbb{N}}$ konvergiert für $n \to \infty$ in Verteilung gegen T
X	erklärende Variable (Kovariable)
x	Realisierung der erklärenden Variable X

$\mathbf{X}$	p-dimensionaler Vektor von erklärenden Variablen, $\mathbf{X} = (X_1, \ldots, X_p)'$
$\mathbf{x}$	Realisierung des Kovariablenvektors $\mathbf{X}$
$\mathrm{P}(\cdot)$	Wahrscheinlichkeitsmaß
$\mathrm{E}(\cdot)$	Erwartungswert
$\mathrm{Var}(\cdot)$	Varianz
$\mathrm{Cov}(\cdot)$	Kovarianz
Θ	Parameterraum, $\Theta \subset \mathbb{R}^p$
$\boldsymbol{\theta}$	p-dimensionaler Parameter, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)' \in \Theta$
$L(\boldsymbol{\theta})$	Likelihood-Funktion für $\boldsymbol{\theta}$
$l(\boldsymbol{\theta})$	Log-Likelihood-Funktion für $\boldsymbol{\theta}$
$\hat{\boldsymbol{\theta}}$	Schätzer für $\boldsymbol{\theta}$
$I(\boldsymbol{\theta})$	Informationsmatrix (S. 42)
$\mathcal{I}(\boldsymbol{\theta})$	Fisher-Matrix, erwartete Informationsmatrix (S. 42)
$I(\boldsymbol{\theta})$	beobachtete Informationsmatrix (S. 42)
$L^P(\boldsymbol{\beta})$	partielle Likelihood-Funktion für $\boldsymbol{\beta} \in \Theta$ (S. 87)
$l^P(\boldsymbol{\beta})$	partielle Log-Likelihood-Funktion für $\boldsymbol{\beta} \in \Theta$
$f_{\mathcal{N}(0,1)}(\cdot)$	Dichtefunktion bei Standard-Normal-Verteilung
$\Phi(\cdot)$	Verteilungsfunktion bei Standard-Normal-Verteilung
ML	Maximum-Likelihood (Schätzer)
KM	Kaplan-Meier (Schätzer)
AFT	Accelerated failure time
$\mathrm{Exp}(\lambda)$	Exponential-Verteilung (S. 59)
$\mathrm{Weib}(\lambda, \beta)$	Weibull-Verteilung (S. 59)
$\mathrm{EV}(u, b)$	Extremwert-Verteilung (S. 60)
$\mathcal{N}(\mu, \sigma^2)$	Normal-Verteilung (S. 60)
$\mathcal{N}_p(\boldsymbol{\mu}, \Sigma)$	p-dimensionale Normal-Verteilung
$Log\mathcal{N}(\mu, \sigma^2)$	Log-Normal-Verteilung (S. 61)
$Logist(u, b)$	Logistik-Verteilung (S. 61)

$LogLogist(\alpha, \beta)$	Log-Logistik-Verteilung (S. 61)
χ^2_q	Chiquadrat-Verteilung mit q Freiheitsgraden
$\chi^2_{q;\alpha}$	α-Quantil der χ^2_q -Verteilung

Kapitel 1

Einleitung

Die statistische Analyse von „Lebensdauern“ spielt in medizinischen und biologischen Studien eine wichtige Rolle. Sie dient beispielsweise der Beurteilung von Therapien bei bestimmten Krankheiten oder kann Auskunft darüber geben, wie stark die toxische Wirkung von Schadstoffen auf Organismen ist. Zu weiteren Anwendungsgebieten zählen unter anderem das Ingenieurwesen, die Ökonomie oder die Sozialwissenschaften. Beispiele von „Lebensdaten“ in diesen Fachgebieten sind die Funktionsdauer von elektronischen oder mechanischen Komponenten und Systemen, die Dauer von Streiks, die Periode der Arbeitslosigkeit bis zur Neueinstellung, die Länge eines Zeitschriftenabonnements oder die Dauer einer Ehe. Da es sich nicht immer um eine Lebensdauer im biologischen oder funktionstechnischen Sinn handelt, ist der Ausdruck „Lebensdauer“ lediglich als Fachbegriff zu verstehen, mit dem die Zeitspanne zwischen einem wohldefinierten Startzeitpunkt t_0 und dem Eintritt des interessierenden Ereignisses, dem Endzeitpunkt $t_0 + t$, gemeint ist. In der medizinischen Forschung entspricht der Startpunkt oft der Rekrutierung eines Patienten in eine klinische Studie und geht mit dem Beginn einer bestimmten Therapie einher. Das interessierende Ereignis könnte dann – je nach Krankheitsstadium und Behandlungsmethode – die Genesung des Patienten, sein Tod oder eine Linderung der Symptome sein.

Lebenszeitdaten zeichnen sich dadurch aus, dass sie häufig einer Zensur unterliegen. Die Lebensdauer eines Individuums nennt man zensiert, wenn der interessierende Endzeitpunkt $t_0 + t$ für dieses Individuum nicht beobachtet werden kann. Ursache hierfür kann beispielsweise ein Beenden der Studie sein, obwohl das interessierende Ereignis noch nicht bei allen Individuen eingetreten ist. Ist in einem solchen Fall

$t_0 + c$ der Zeitpunkt, in dem ein Individuum zum letzten Mal beobachtet worden ist, so ist c die zensierte Lebenszeit. Man weiß nur, dass die eigentliche Lebensdauer größer ist als c und spricht in diesem Zusammenhang von Rechtszensierung. Zu weiteren Zensurmechanismen zählen die Linkszensur, bei der lediglich bekannt ist, dass ein Individuum das interessierende Ereignis vor dem Beginn der Studie erlebt hat, und die Intervallzensur, bei der die Information darin besteht, dass das Ereignis innerhalb eines bestimmten Zeitintervalls stattgefunden hat. Die Charakteristika dieser Zensursysteme werden in Kapitel 3, Abschnitt 3.1 sorgfältiger definiert.

Aufgrund der unvollständigen Beobachtungen können Lebenszeitdaten nicht mit statistischen Standardverfahren analysiert werden, sondern erfordern besondere Auswertungsmethoden. Das Ziel dieser Arbeit besteht daher in der Darstellung von statistischen Modellen und Methoden der Lebenszeitanalyse, welche in der Literatur auch als „Survival Analysis“ bekannt ist. Fundamentale Fragestellungen sind dabei die nicht-parametrische Schätzung der Verteilungsfunktion, die Modellierung des Einflusses von erklärenden Variablen auf die Lebensdauern beobachteter Individuen, der Vergleich von Überlebenswahrscheinlichkeiten für verschiedene Gruppen von Probanden und schließlich die Bewertung der Güte einer Modellanpassung.

Kapitel 2 beginnt mit grundlegenden Definitionen der „Survival Analysis“ und beinhaltet die Charakterisierung einer Lebensdauer T. Die Verteilung der nicht-negativen Zufallsvariable T kann neben der Dichtefunktion durch die Survival-Funktion, die Hazard-Funktion und die erwartete Restlebensdauer eindeutig beschrieben werden. Nach Einführung dieser Funktionen wird in Abschnitt 2.4 die Beziehung zwischen ihnen zusammenfassend dargestellt.

Im ersten Abschnitt des Kapitels 3 erfolgt eine ausführliche Beschreibung der bereits erwähnten Zensierungsmechanismen. Da die Rechtszensur für die Praxis die größte Relevanz hat, wird der Schwerpunkt dieser Arbeit auf die Beschreibung und Analyse von Datensätzen unter dieser Zensurform gelegt. Weil zensierte Stichproben mit Hilfe klassischer Maximum-Likelihood-Methoden ausgewertet werden können, wird in Abschnitt 3.2 die Likelihood-Funktion für rechtszensierte Daten konstruiert. Abschnitt 3.3 fasst Aspekte der asymptotischen Statistik zusammen, die für die Thematik nachfolgender Kapitel bedeutsam sind.

Als grundlegendes Werkzeug der Lebenszeitanalyse wird in Kapitel 4 der von Kaplan und Meier [Kap-58] eingeführte nicht-parametrische Schätzer für die Survival-

Funktion diskutiert. Nach seiner Herleitung aus der in Abschnitt 3.2 konstruierten Likelihood-Funktion, wird im zweiten Teil des Kapitels die Varianz dieses Schätzers approximiert und in Abschnitt 4.3 lokale Konfidenzintervalle für die Survival-Funktion angegeben.

Obwohl theoretisch jede nicht-negative Zufallsvariable die Lebensdauern von Individuen einer Population repräsentieren kann, nehmen in der parametrischen Lebensdauerstatistik bestimmte Verteilungen eine zentrale Position ein. Sie werden in Kapitel 5 vorgestellt. Ferner wird in diesem Kapitel eine Methode eingeführt, mit der graphisch beurteilt werden kann, inwieweit sich ein parametrisches Modell für die Beschreibung einer zensierten Stichprobe eignet.

In bestimmten Situationen besteht zwar das Interesse an der Lebenszeitverteilung einer einzelnen Gruppe, doch kommt es weitaus häufiger vor, dass die Ausfallzeiten von Individuen mehrerer Gruppen miteinander verglichen werden sollen, so zum Beispiel in einer klinischen Studie zur Bewertung der Wirksamkeit verschiedener Therapien. Auch können in einem Experiment Kovariablen beobachtet werden, um zu überprüfen welche von ihnen Auswirkungen auf die Lebenszeit haben. Beides führt zu Regressionsmodellen. Sowohl parametrische als auch nicht-parametrische Methoden sind dabei gebräuchlich. In den Kapiteln 6 und 7 werden mit dem parametrischen Accelerated-Failure-Time-Modell (AFT-Modell) und dem semi-parametrischen Cox-Hazard-Modell die beiden wichtigsten unter ihnen vorgestellt.

Likelihood-Methoden für AFT-Modelle basieren auf der in Kapitel 3 hergeleiteten Likelihood-Funktion. Sie werden in Abschnitt 6.1 allgemein behandelt. Abschnitt 6.2 enthält dann die Anpassung eines Weibull-, Log-Logistik und Log-Normal-AFT-Modells an einen konkreten Datensatz.

Das von Cox [Cox-72] vorgeschlagene Cox-Hazard-Modell ist bezüglich der Lebenszeit T an keine spezielle Verteilung gebunden. Parametrisch modelliert werden nur die Auswirkungen der erklärenden Variablen. Für die Schätzung der Regressionsparameter hat Cox [Cox-72] [Cox-75] eine partielle Likelihood-Funktion vorgeschlagen. Diese wird in den Abschnitten 7.1 und 7.2 sowohl für bindungsfreie Datensätze als auch für solche mit mehrfachen Messwerten hergeleitet. Die Identifikation relevanter Kovariablen wird in Abschnitt 7.3 nur für das Cox-Hazard-Modell erörtert. Die dort beschriebene Vorgehensweise kann aber ganz entsprechend auf das AFT-Modell

übertragen werden. Die Interpretation geschätzter Regressionsparameter im Cox-Hazard-Modell wird in Abschnitt 7.4 beschrieben und an einem konkreten Beispiel durchgeführt. Sowohl das Kapitel zum AFT-Modell als auch das zum Cox-Hazard-Modell schließen mit einem graphischen Verfahren zur Überprüfung des jeweiligen Modells.

Kapitel 2

Grundbegriffe der Lebenszeitanalyse

Die grundlegende Aufgabe bei der Modellierung von Lebensdaten besteht darin, Aussagen über die Zeit T zu machen, zu der das interessierende Ereignis eintritt. T sei demnach stets eine nicht-negative Zufallsvariable mit Verteilungsfunktion F, die die Ausfallzeit von Individuen einer zunächst homogenen Population beschreibt.

Im Folgenden werden Funktionen vorgestellt, mit deren Hilfe die Verteilung von T charakterisiert werden kann. Neben der Dichtefunktion handelt es sich dabei um die Survival-Funktion, die Hazard-Funktion und die erwartete Restlebensdauer. Eingeführt werden diese Funktionen für stetig, diskret und gemischt verteilte Zufallsvariablen. Diskutiert wird außerdem der Zusammenhang zwischen ihnen, denn ist eine der Funktionen bekannt, so können die anderen drei eindeutig bestimmt werden. Die Ausführungen des Kapitels folgen im Wesentlichen denen von Klein und Moeschberger [Kle-97, Kapitel 2]. Weitere Quellen werden an gegebener Stelle benannt.

2.1 Die Survival-Funktion

Die Survival-Funktion ist eine wesentliche Größe bei der Beschreibung von Lebenszeiten. Sie gibt die Wahrscheinlichkeit dafür an, dass ein aus der Population beliebig ausgewähltes Individuum den Zeitpunkt t überlebt. Synonym verwendete Bezeichnungen sind Survivor-Funktion, Überlebensfunktion und – besonders im technischen Kontext – Zuverlässigkeitsfunktion (reliability function).

Definition 2.1 (**Survival-Funktion**). [1] *Die Survival Funktion S_T zur Ausfallzeit T ist definiert als*

$$S_T(t) = P(T > t), \quad t \geq 0. \tag{2.1}$$

Im Folgenden wird der Index T weggelassen, wenn klar ist, welche Zufallsvariable der Survival-Funktion zugrunde liegt.

Satz 2.1 (**Eigenschaften der Survival-Funktion**). *Für die Survival-Funktion gilt:*

1. $S(t) = 1 - F(t), \quad t \geq 0.$
2. *S ist eine monoton fallende Funktion mit $S(0) = 1$ und $\lim_{t\to\infty} S(t) = 0$.*
3. *Ist T stetig verteilt mit positiver Dichte f auf $(0, \infty)$, so ist S streng monoton fallend und es gilt*

$$S(t) = \int_t^\infty f(u)\, du. \tag{2.2}$$

In diesem Fall lässt sich f schreiben als

$$f(t) = -\frac{\partial S(t)}{\partial t} \tag{2.3}$$

$$= \lim_{\Delta\to 0^+} \frac{P(t \leq T < t + \Delta)}{\Delta}. \tag{2.4}$$

4. *Ist T diskret verteilt mit Werten in $0 < t_1 < t_2 < \dots$, d.h. diskreter Dichte $p(t_j) = P(T = t_j)$, $j = 1, 2, \dots$, so ist S eine monoton fallende rechts-stetige Treppenfunktion:*

$$S(t) = \sum_{t_j > t} p(t_j). \tag{2.5}$$

Beweis: Alle Gleichungen folgen direkt aus der Definition bzw. den Eigenschaften der Verteilungsfunktion: Es gilt

$$S(t) = P(T > t) = 1 - P(T \leq t) = 1 - F(t),$$

[1] Andere Autoren, so zum Beispiel Cox & Oakes [Cox-84, S. 13] oder Lawless [Law-03, S. 9], definieren $S_T(t)$ als die Wahrscheinlichkeit dafür, mindestens bis zum Zeitpunkt t am Leben zu bleiben, d.h. $S_T(t) = P(T \geq t)$, $t \geq 0$.

daraus folgt im stetigen Fall $\frac{\partial S(t)}{\partial t} = -f(t)$ und damit die Gleichung (2.3). Weiter gilt $S(0) = 1 - F(0) = 1$ und $\lim_{t\to\infty} S(t) = \lim_{t\to\infty} 1 - F(t) = 0$. Ebenso schnell sieht man (2.2): Da $\operatorname{supp}(T) = (0, \infty)$, ist

$$S(t) = 1 - F(t) = 1 - \int_0^t f(u)\, du = \int_t^\infty f(u)\, du.$$

Die Gleichheit in (2.4) ergibt sich mit der Stetigkeit von T:

$$\begin{aligned}
\frac{P(t \le T < t + \Delta)}{\Delta} &= \frac{P(T < t + \Delta) - P(T < t)}{\Delta} \\
&= \frac{1 - P(T \ge t + \Delta) - (1 - P(T \ge t))}{\Delta} \\
&= \frac{1 - P(T > t + \Delta) - (1 - P(T > t))}{\Delta} \\
&= \frac{-S(t + \Delta) + S(t)}{\Delta} \longrightarrow -\frac{\partial S(t)}{\partial t}, \text{ für } \Delta \to 0^+.
\end{aligned}$$

Und (2.5) gilt aufgrund der Definition von S. □

Nach Satz 2.1 ist $f(t)\Delta$ die approximative Wahrscheinlichkeit dafür, dass das interessierende Ereignis zum Zeitpunkt t eintritt.

2.2 Die Hazard-Funktion

Die Hazard-Funktion[2] ist für die Charakterisierung von Lebensverteilungen von grundlegender Bedeutung. Sie gibt an, wie sich das Ausfallrisiko in Abhängigkeit vom Alter im Verlauf der Zeit verändert. Diese Beziehung ist in vielen Anwendungen von großem Interesse. Informationen über den Verlauf der Hazard-Funktion helfen überdies bei der Auswahl geeigneter parametrischer Modelle. Im ersten Teil dieses Abschnitts wird die Hazard-Funktion und ihre Beziehung zur Survival-Funktion sowohl für stetige als auch diskrete Verteilungen vorgestellt. Im zweiten Teil wird ein allgemeiner Rahmen entwickelt, innerhalb dessen stetige, diskrete und gemischte Lebensverteilungen gemeinsam behandelt werden können. Der letzte Teil liefert

[2]Abhängig vom wissenschaftlichen Anwendungsgebiet ist die Hazard-Funktion auch unter anderen Namen bekannt. Während man in der Zuverlässigkeits-Theorie zum Beispiel den Begriff bedingte Ausfallrate (conditional failure rate) benutzt, ist in der Demographie der Begriff Sterblichkeitskraft (force of mortality) gebräuchlicher. In der Epidemiologie ist die Hazard-Funktion als altersspezifische Ausfallrate (age-specific failure rate) bekannt und im Zusammenhang mit stochastischen Prozessen wird der Begriff Intensitätsfunktion (intensity function) verwendet.

schließlich allgemeine Informationen zur Hazard-Funktion im Zusammenhang mit parametrischer Modellbildung.

2.2.1 Definition und Eigenschaften

Definition 2.2 (**Hazard-Funktion**). *Für stetige Überlebenszeiten T definiert man die Hazard-Funktion λ als*

$$\lambda(t) := \lim_{\Delta \to 0^+} \frac{P(t \leq T < t + \Delta \mid T \geq t)}{\Delta}, \quad t \geq 0. \tag{2.6}$$

Ist T diskret verteilt mit Träger $\{t_1, t_2, \dots\}$, $0 < t_1 < t_2 < \dots$, so ist die Hazard-Funktion gegeben durch

$$\lambda(t_j) := P(T = t_j \mid T \geq t_j), \quad j = 1, 2, \dots. \tag{2.7}$$

Die Hazard-Funktion (auch Hazard-Rate) ist nicht-negativ und gibt im stetigen Fall Auskunft darüber, wie schnell Individuen eines bestimmten Alters das interessierende Ereignis erleben. Handelt es sich dabei um Ausfall oder Tod, so ist die Größe $\lambda(t)\Delta$ ein approximatives Maß für das Risiko im infinitesimalen Intervall $(t, t + \Delta)$ zu sterben, falls man bis zur Zeit t überlebt hat. Im diskreten Fall ist $\lambda(t_j)$ die bedingte Wahrscheinlichkeit dafür, dass ein Individuum, welches das Alter t_j erreicht hat, zum Zeitpunkt t_j stirbt.

Satz 2.2. *Ist die Ausfallzeit T stetig verteilt, so gilt für die Hazard-Funktion*

$$\lambda(t) = \frac{f(t)}{S(t)} = -\frac{S'(t)}{S(t)} = -\frac{\partial}{\partial t} \ln S(t). \tag{2.8}$$

Beweis: Es gilt

$$\begin{aligned}
\lambda(t) &\overset{Def}{=} \lim_{\Delta \to 0^+} \frac{P(t \leq T < t + \Delta \mid T \geq t)}{\Delta} = \lim_{\Delta \to 0^+} \frac{P(t \leq T < t + \Delta)}{P(T \geq t)\Delta} \\
&= \lim_{\Delta \to 0^+} \frac{P(T < t + \Delta) - P(T < t)}{P(T \geq t)\Delta} \overset{\substack{T \\ stetig}}{=} \lim_{\Delta \to 0^+} \frac{1}{P(T > t)} \frac{F(t + \Delta) - F(t)}{\Delta} \\
&= \frac{f(t)}{S(t)} \overset{(2.3)}{=} -\frac{S'(t)}{S(t)} = -\frac{\partial}{\partial t} \ln S(t).
\end{aligned}$$

□

Eine verwandte Größe ist die kumulierte Hazard-Funktion.

Definition 2.3 (**Kumulierte Hazard-Funktion I**). *Für eine stetig verteilte Lebenszeit T wird die kumulierte Hazard-Funktion Λ definiert als*

$$\Lambda(t) := \int_0^t \lambda(u)\, du. \tag{2.9}$$

Ist die Zufallsvariable T diskret, so definiert man die kumulierte Hazard-Funktion Λ entsprechend als

$$\Lambda(t) := \sum_{t_j \le t} \lambda(t_j). \tag{2.10}$$

Satz 2.3 (**Darstellung von S durch λ I**). *1. Ist T eine stetig verteilte Zufallsvariable, so lassen sich Survival- und Hazard-Funktion in folgende Beziehung bringen:*

$$S(t) = \exp\Big(- \int_0^t \lambda(u)\, du\Big) = \exp\big(- \Lambda(t)\big). \tag{2.11}$$

2. Ist die Zufallsvariable T hingegen diskret mit Wahrscheinlichkeiten $p(t_j)$, $0 < t_1 < t_2 < \dots$, so ist der Zusammenhang gegeben durch:

$$S(t) = \prod_{t_j \le t} [\, 1 - \lambda(t_j)\,]. \tag{2.12}$$

Beweis: Mit (2.8) hat man $\int_0^t \lambda(u)\, du = -\ln S(t)$ und somit den stetigen Fall (2.11). Ist T diskret verteilt auf dem Zeitraster $0 < t_1 < t_2 < \dots$, so gilt für die Hazard-Funktion

$$\begin{aligned} \lambda(t_j) &\overset{Def}{=} P(T = t_j \mid T \ge t_j) \\ &= \frac{P(T = t_j)}{P(T \ge t_j)} = \frac{p(t_j)}{P(T > t_{j-1})} \\ &= \frac{p(t_j)}{S(t_{j-1})}, \quad j = 1,\ 2, \dots, \end{aligned} \tag{2.13}$$

wobei $S(t_0) := 1$. Da $p(t_j) = S(t_{j-1}) - S(t_j)$, gilt weiter

$$\lambda(t_j) = 1 - \frac{S(t_j)}{S(t_{j-1})}, \quad j = 1, 2, \dots.$$

Die Gleichheit (2.12) erhält man schließlich durch

$$\begin{aligned}\prod_{t_j \le t}[\,1-\lambda(t_j)\,] &= \prod_{t_j \le t}\left[\,1-\left(1-\frac{S(t_j)}{S(t_{j-1})}\right)\right] = \prod_{t_j \le t}\frac{S(t_j)}{S(t_{j-1})} \\ &= \frac{S(t_1)\,S(t_2)\,\ldots\,S(t_{k-1})\,S(t_k)}{S(t_0)\,S(t_1)\,S(t_2)\,\ldots\,S(t_{k-1})}, \quad t_k := \max\{t_j \mid t_j \le t\} \\ &= S(t_k) \\ &\overset{Def\ t_k}{=} S(t).\end{aligned}$$

□

Bemerkung 2.1. Zu beachten ist, dass die Beziehung $S(t) = \exp\big(-\Lambda(t)\big)$ für die diskrete Definition von Λ nicht gilt. Ein Analogon zu (2.11) wäre nach (2.12) durch

$$\Lambda(t) := -\ln\left(\prod_{t_j \le t}[\,1-\lambda(t_j)\,]\right) = -\sum_{t_j \le t}\ln[\,1-\lambda(t_j)\,] \tag{2.14}$$

gegeben. Im nächsten Abschnitt werden zwei neue Integral-Begriffe eingeführt, das Riemann-Stieltjes-Integral und das Produkt-Integral. Durch sie kann die Unterscheidung zwischen stetigem und diskretem Fall formal umgangen werden und man wird sehen, dass dabei gerade die in (2.9) gegebe Definition von Λ der in (2.14) vorzuziehen ist.

2.2.2 Eine allgemeine Formulierung

Möchte man den stetigen und diskreten Fall nicht getrennt voneinander untersuchen oder betrachtet man eine Zufallsvariable T, die sowohl diskrete als auch stetige Komponenten besitzt, so bietet es sich an, die kumulierte Hazard-Funktion mittels des Riemann-Stieltjes-Integrals zu definieren. Die diskrete und stetige Definition von Λ können dadurch zusammengefasst und folglich auch gemeinsam behandelt werden. Ein zweites nützliches Werkzeug ist das sogenannte Produkt-Integral. Mit seiner Hilfe kann die Darstellung der Survival-Funktion durch die Hazard-Rate ebenfalls für beliebige Überlebensverteilungen in einem Ausdruck vereint werden. Der Begriff des Produkt-Integrals geht auf den italienischen Mathematiker Vito Volterra [Vol-1887] zurück und die Idee des Kalküls besteht darin, die wohlbekannte Verwandtschaft von Summation und Integration auf Produkte zu übertragen. Da die Definition des Produkt-Integrals in Analogie zu der des Riemann-Stieltjes-Integrals erfolgt, soll

dieses im Folgenden zuerst eingeführt werden. Der Inhalt des Abschnitts basiert dabei auf den Ausführungen von Lawless [Law-03, S. 11–13].

Definition 2.4 (**Riemann-Stieltjes-Integral**). [3] *Sei G eine schwach monoton wachsende Funktion auf $\mathbb{R}$, die rechtsstetig ist, linksseitige Grenzwerte besitzt und in jedem endlichen Intervall höchstens endlich viele Diskontinuitäten hat. Sei h eine Funktion, die in jedem Intervall $(a, b]$ ebenfalls höchstens endlich viele Unstetigkeitsstellen hat. Diese seien jedoch verschieden von denen von G. Unter der Annahme, dass $g(u) = G'(u)$ für alle u existiert, in denen G stetig ist, und dass in den Unstetigkeitsstellen a_j von G gilt $G(a_j) - \lim_{\Delta \to 0} G(a_j - \Delta) = g_j$, ist das Riemann-Stieltjes-Integral von h bezüglich G über dem Intervall $(a, b]$ durch*

$$\int_{(a,\,b]} h(u)\, d\, G(u) \;=\; \int_a^b h(u)\, g(u)\, du + \sum_{j:\, a < a_j \leq b} h(a_j)\, g_j \tag{2.15}$$

gegeben. Bei dem ersten Integral auf der rechten Seite der Gleichung (2.15) handelt es sich um das wohlbekannte Riemann-Integral.

Bemerkung 2.2. Die Verteilungsfunktion $F(t) = P(T \leq t)$ einer Zufallsvariable T ist eine rechtsstetige und monoton wachsende Funktion mit Sprüngen in Punkten a_j, für die $P(T = a_j) = p_j > 0$ gilt, und mit Wahrscheinlichkeitsdichte $f(u) = F'(u)$ für alle u, in denen F stetig ist. Nach (2.15) ist $P(a < T \leq b)$ demnach mit dem Integranden $h \equiv 1$ gegeben durch

$$\begin{aligned} F(b) - F(a) &= \int_{(a,\,b]} d\, F(u) && (2.16) \\ &= \int_a^b f(u)\, du + \sum_{j:\, a < a_j \leq b} p_j. && (2.17) \end{aligned}$$

Ist die Zufallsvariable T stetig, so hat die Verteilungsfunktion F keine Sprungstellen und die Summe im letzten Teil der Gleichungskette fällt weg. Ist T hingegen diskret verteilt, so ist F eine Treppenfunktion mit $f(u) = 0$ für alle u, in denen F stetig ist, so dass das Riemann-Integral im letzten Ausdruck dann gleich 0 ist. Das Riemann-Stieltjes-Integral liefert folglich ein Werkzeug, mit dem der diskrete Fall

[3] Diese Definition liefert lediglich eine Spezialform des Riemann-Stieltjes-Integrals, die für Verteilungsfunktionen geeignet ist. Allgemein wird das Riemann-Stieltjes-Integral für schwach monoton wachsende Integratoren G erklärt und über untere und obere Darboux-Summen eingeführt. Eine sinnvolle Definition ist auch für Integratoren von beschränkter Variation möglich. Siehe dazu [Pfe-05, S. 79ff].

und der stetige Fall gemeinsam abgehandelt werden können. Mit seiner Hilfe kann eine allgemeine Definition der kumulierten Hazard-Funktion angegeben werden.

Definition 2.5 (Kumulierte Hazard-Funktion II). *Sei T eine nicht-negative Zufallsvariable und λ die Hazard-Funktion von T. Nach den Überlegungen aus Abschnitt 2.2.1 sei die Hazard-Funktion für alle u, in denen F stetig ist, durch $\lambda(u) = f(u)/S(u)$ gegeben und $\lambda_j = P(T = a_j \mid T \geq a_j)$ seien die diskreten Hazard-Werte in den Sprungstellen a_j von F. Die kumulierte Hazard-Funktion kann dann mit Hilfe des Riemann-Stieltjes-Integrals definiert werden als*

$$\Lambda(t) \;=\; \int_0^t d\,\Lambda(u) \;=\; \int_0^t \lambda(u)\,du + \sum_{j:a_j \leq t} \lambda_j\,. \tag{2.18}$$

In Satz 2.3 wurde der wichtige Zusammenhang zwischen Survival- und Hazard-Funktion für stetige und diskrete Verteilungen getrennt angegeben. Um den Sachverhalt allgemeingültig formulieren zu können, benötigt man das Produkt-Integral.

Definition 2.6 (Produkt-Integral). *Sei $Z^{(m)} := (\{u_0^m, u_1^m, \ldots, u_m^m\})_m$ mit $a = u_0^m < u_1^m < \ldots < u_m^m = b$ eine Folge von Zerlegungen des Intervalls $(a, b]$, für die mit $\Delta u_i^m = u_i^m - u_{i-1}^m$ gilt:*

$$\max(\Delta u_i^m) \longrightarrow 0 \quad \textit{für} \quad m \to \infty.$$

Sei G wie in Definition 2.4 gegeben. Das Produkt-Integral von G über $(a, b]$ ist dann für jede solcher Zerlegungsfolgen $Z^{(m)}$ definiert als[4]

$$\mathcal{P}_a^b\,[\,1 + d\,G(u)\,] \;=\; \lim_{m\to\infty} \prod_{i=1}^{m} \{1 + G(u_i^m) - G(u_{i-1}^m)\}. \tag{2.19}$$

Die Eigenschaften von G, wie sie in Definition 2.4 gegeben sind, garantieren dabei die Existenz des Grenzwertes.

Zu Gunsten der Übersichtlichkeit wird im Folgenden bei den Folgengliedern der Zerlegungsfolge $Z^{(m)} := (\{u_0^m, u_1^m, \ldots, u_m^m\})_m$ in den meisten Fällen auf den Index m verzichtet.

[4] Entsprechend der allgemeinen Definition des Riemann-Stieltjes-Integrals für Integratoren von beschränkter Variation, ist die obige Definition des Produkt-Integrals auch für Funktionen G sinnvoll, die von beschränkter Variation sind, vgl. z.B. [Gil-01].

Satz 2.4 (**Multiplikativität des Produkt-Integrals**). *Existiert das Produkt-Integral von G über dem Intervall $(a, b]$ und ist c beliebig mit $a < c < b$, so gilt*

$$\mathcal{P}_a^b [\, 1 + d\, G(u) \,] \;=\; \mathcal{P}_a^c [\, 1 + d\, G(u) \,] \, \mathcal{P}_c^b [\, 1 + d\, G(u) \,]. \tag{2.20}$$

Beweis: Nach Definition 2.6 ist $\mathcal{P}_a^b [1 + d\, G(u)] = \lim_{m\to\infty} \prod_{i=1}^m \{1 + G(u_i) - G(u_{i-1})\}$ für eine beliebige Zerlegungsfolge $Z^{(m)} = \{a = u_0 < u_1 < \ldots < u_m = b\}$ mit $\max(u_i - u_{i-1}) \to 0$ für $m \to \infty$. Es sei nun $a = \bar{u}_0 < \bar{u}_1 < \ldots < \bar{u}_m = b$ eine gemeinsame Verfeinerung von Z^m und $a < c < b$. Dann gilt mit $c = \bar{u}_k$

$$\begin{aligned} &\prod_{i=1}^m \{1 + G(\bar{u}_i) - G(\bar{u}_{i-1})\} \\ &\qquad = \; \prod_{i=1}^k \{1 + G(\bar{u}_i) - G(\bar{u}_{i-1})\} \prod_{i=k+1}^m \{1 + G(\bar{u}_i) - G(\bar{u}_{i-1})\}. \end{aligned}$$

Da G auf den Teilintervallen $(a, c]$ und $(c, b]$ die gleichen Eigenschaften besitzt wie auf $(a, b]$, existieren für $m \to \infty$ die Limiten der beiden Produkte auf der rechten Seite der Gleichung und entsprechen gerade den Produkt-Integralen von G über $(a, c]$ und $(c, b]$. Es gilt also

$$\mathcal{P}_a^b [\, 1 + d\, G(u) \,] \;=\; \mathcal{P}_a^c [\, 1 + d\, G(u) \,] \, \mathcal{P}_c^b [\, 1 + d\, G(u) \,].$$

□

Lemma 2.1. *1. Für $x \in \mathbb{R}$ mit x nahe 0 (insbesondere $|x| < 1$) gilt:*

$$\ln(1 + x) \;=\; x + r(x), \tag{2.21}$$

wobei für den Fehlerterm $r(x) = o(|x|)$ für $x \to 0$ gilt. In der Lagrange-Form ist er gegeben durch

$$r(x) \;=\; -\frac{x^2}{2(1+\xi)^2} \;\leq\; 0$$

mit $\xi \in (0, x)$, falls $x > 0$ bzw. $\xi \in (x, 0)$, falls $x < 0$. Insbesondere erhält man die Abschätzung

$$\ln(1 + x) \;\leq\; x. \tag{2.22}$$

2. *Für $x > 0$, x nahe 0, gilt*

$$\ln(1-x) \;\geq\; -2x. \tag{2.23}$$

Beweis: 1. Nach dem Satz von Taylor, siehe zum Beispiel [Kab-96, S. 263], kann $f(x) = \ln(x+1)$ in einer Umgebung von 0 durch ein Polynom 1. Ordnung approximiert werden:

$$f(x) = f(0) + f'(0)(x) + r(x).$$

Der Fehler $r(x)$ erfüllt dabei $\lim_{x\to 0} r(x)/\,|x| = 0$, bzw. $r(x) = o(|x|)$ für $x \to 0$. In der Lagrange'schen Form kann $r(x)$ angegeben werden als $r(x) = \frac{f''(\xi)}{2!}x^2$ mit $\xi \in (0,\, x)$ bzw. $\xi \in (x,\, 0)$, je nachdem ob x positiv oder negativ ist. Mit $f(x) = \ln(1+x)$, $f'(x) = (1+x)^{-1}$ und $f''(x) = -(1+x)^{-2}$ erhält man demnach die Behauptung.
2. Die Abschätzung (2.23) ist klar, da $\ln(1-x)$ in 0 die Steigung -1 hat. □

Satz 2.5 (**Weitere Eigenschaften des Produkt-Integrals**). *Sei G wie in Definition 2.4 festgelegt. Darüberhinaus sei im stetigen Fall $G' = g$ beschränkt auf $(a, b]$.*

1. *Ist G in allen $u \in (a, b]$ stetig, so gilt:*

$$\mathcal{P}_a^b\,[\,1 + d\,G(u)\,] \;=\; \lim_{m\to\infty} \prod_{i=1}^{m} [1 + g(u_i)\,\Delta u_i] \tag{2.24}$$

$$\;=\; \exp\left(\int_a^b g(u)\,du\right). \tag{2.25}$$

2. *Ist G eine Treppenfunktion mit Sprüngen der Höhe g_j in den Punkten a_j, $j = 1, 2, \ldots$, dann ist das Produkt-Integral gegeben durch*

$$\mathcal{P}_a^b\,[\,1 + d\,G(u)\,] \;=\; \prod_{j:a<a_j\leq b} (1 + g_j). \tag{2.26}$$

3. *Allgemeiner gilt: Hat G Diskontinuitäten in a_j mit Sprunghöhen g_j, $j = 1, 2, \ldots$, so ist*

$$\mathcal{P}_a^b\,[\,1 + d\,G(u)\,] \;=\; \exp\left(\int_a^b g(u)\,du\right) \prod_{j:a<a_j\leq b} (1 + g_j). \tag{2.27}$$

Beweis: 1. Ist die Funktion G auf dem gegebenen Intervall stetig, so ist sie nach der Festlegung in Definition 2.4 auch differenzierbar mit $G'(u) = g(u) \geq 0$. Für jede Partition $a = u_0 < u_1 < \ldots < u_m = b$ des Intervalls $(a, b]$ gilt mit $\Delta u_i = u_i - u_{i-1},\ i = 1, \ldots, m$, also die folgende Approximation

$$\begin{aligned} G(u_i) - G(u_{i-1}) &= G(u_i) - G(u_i - \Delta u_i) \\ &\overset{Taylor}{=} G(u_i) - \big[G(u_i) - G'(u_i)\Delta u_i + r(\Delta u_i)\big] \\ &= g(u_i)\Delta u_i + r(\Delta u_i). \end{aligned} \tag{2.28}$$

Für den Fehlerterm gilt dabei: $r(\Delta u_i) = r(\Delta u_i^m) = o(|\Delta u_i^m|)$ für $\Delta u_i^m \to 0$. Das bedeutet

$$\forall\, \varepsilon > 0\ \ \exists\, \delta > 0 : \quad |\Delta u_i^m| \leq \delta \ \Rightarrow\ |r(\Delta u_i^m)| \leq \varepsilon\, |\Delta u_i^m|. \tag{2.29}$$

Da hier $\Delta u_i^m \geq 0$ und $\Delta u_i^m \to 0$ für $m \to \infty$, gibt es nach (2.29) insbesondere zu jedem $\varepsilon > 0$ ein m_0, so dass für alle $m \geq m_0$

$$|r(\Delta u_i^m)| \leq \varepsilon\, \Delta u_i^m. \tag{2.30}$$

Für das Produkt-Integral erhält man damit zunächst (2.24):

$$\begin{aligned} \mathcal{P}_a^b\,[\,1 + d\,G(u)\,] &\overset{Def}{=} \lim_{m\to\infty} \prod_{i=1}^{m} \{1 + G(u_i) - G(u_{i-1})\} \\ &\overset{(2.28)}{=} \lim_{m\to\infty} \prod_{i=1}^{m} \{1 + g(u_i)\Delta u_i + r(\Delta u_i)\} \\ &\overset{!}{=} \lim_{m\to\infty} \prod_{i=1}^{m} \{1 + g(u_i)\Delta u_i\} \end{aligned}$$

Die Gültigkeit der letzten Gleichheit soll mit Hilfe der Abschätzung (2.30) explizit nachgerechnet werden:

(i) Im Fall $r(\Delta u_i) \geq 0$ ist „$\geq$“ klar. Es bleibt zu zeigen

$$\lim_{m\to\infty} \prod_{i=1}^{m} \{1 + g(u_i)\Delta u_i + r(\Delta u_i)\} \ \leq\ \lim_{m\to\infty} \prod_{i=1}^{m} \{1 + g(u_i)\Delta u_i\}. \tag{2.31}$$

Sei dazu $\varepsilon > 0$ beliebig und m_0 so gewählt, dass (2.30) gilt. Sei $m \geq m_0$. Dann folgt

$$\prod_{i=1}^{m}\{1 + g(u_i)\Delta u_i + r(\Delta u_i)\} \;\leq\; \prod_{i=1}^{m}\{1 + g(u_i)\Delta u_i + \varepsilon\Delta u_i\}.$$

Setze $a_i := 1 + g(u_i)\Delta u_i$ und zeige

$$\exists\, C_\varepsilon : \qquad \prod_{i=1}^{m}\{a_i + \varepsilon\Delta u_i\} \;\leq\; C_\varepsilon \prod_{i=1}^{m} a_i \quad \text{und} \quad \lim_{\varepsilon\to 0} C_\varepsilon = 1. \tag{2.32}$$

Mit einem solchen C_ε gilt dann

$$\lim_{m\to\infty}\prod_{i=1}^{m}\{1 + g(u_i)\Delta u_i + r(\Delta u_i)\}$$
$$\leq \lim_{m\to\infty} C_\varepsilon \prod_{i=1}^{m}\{1 + g(u_i)\Delta u_i\} \overset{\varepsilon\to 0}{\longrightarrow} \lim_{m\to\infty}\prod_{i=1}^{m}\{1 + g(u_i)\Delta u_i\}$$

und damit (2.31). Zum Beweis von (2.32) wähle $C_\varepsilon := \exp[\varepsilon(b-a)]$. Wegen der Stetigkeit von $\exp(\cdot)$ gilt $\lim_{\varepsilon\to 0} C_\varepsilon = 1$. Weiter ist

$$\ln C_\varepsilon \;=\; \varepsilon(b-a) \;=\; \sum_{i=1}^{m}\varepsilon\Delta u_i$$
$$\overset{a_i\geq 1}{\geq} \sum_{i=1}^{m}\frac{\varepsilon\Delta u_i}{a_i} \overset{(2.22)}{\geq} \sum_{i=1}^{m}\ln\left(1 + \frac{\varepsilon\Delta u_i}{a_i}\right) \;=\; \ln\left[\prod_{i=1}^{m}\left(1 + \frac{\varepsilon\Delta u_i}{a_i}\right)\right].$$

Es gilt also

$$C_\varepsilon \;\geq\; \prod_{i=1}^{m}\left(1 + \frac{\varepsilon\Delta u_i}{a_i}\right) \;=\; \frac{\prod_{i=1}^{m}\{a_i + \varepsilon\Delta u_i\}}{\prod_{i=1}^{m} a_i}$$

und damit (2.32).

(ii) Ist $r(\Delta u_i) \leq 0$, so ist die Abschätzung „$\leq$“ klar und lediglich

$$\lim_{m\to\infty}\prod_{i=1}^{m}\{1 + g(u_i)\Delta u_i + r(\Delta u_i)\} \;\geq\; \lim_{m\to\infty}\prod_{i=1}^{m}\{1 + g(u_i)\Delta u_i\} \tag{2.33}$$

zu zeigen. Sei wieder $\varepsilon > 0$ beliebig und m_0 so, dass (2.30) gilt. Für $m \geq m_0$ hat

man zunächst $r(\Delta u_i) \geq -\varepsilon \Delta u_i$ und damit

$$\prod_{i=1}^{m}\{1 + g(u_i)\Delta u_i + r(\Delta u_i)\} \;\geq\; \prod_{i=1}^{m}\{1 + g(u_i)\Delta u_i - \varepsilon \Delta u_i\}.$$

Ist $a_i := 1 + g(u_i)\Delta u_i$, so ist zum Beweis von (2.33) zu zeigen:

$$\exists\, C_\varepsilon : \qquad \prod_{i=1}^{m}\{a_i - \varepsilon \Delta u_i\} \;\geq\; C_\varepsilon \prod_{i=1}^{m} a_i \quad \text{und} \quad \lim_{\varepsilon \to 0} C_\varepsilon = 1. \tag{2.34}$$

Wähle dazu $C_\varepsilon := \exp\{-2\varepsilon(b-a)\}$. Es ist $\lim_{\varepsilon\to 0} C_\varepsilon = 1$ und

$$\begin{aligned} \ln C_\varepsilon \;&=\; -2\varepsilon(b-a) \;=\; \sum_{i=1}^{m} -2\varepsilon \Delta u_i \\ &\overset{a_i \geq 1}{\leq} \sum_{i=1}^{m} -\frac{2\varepsilon \Delta u_i}{a_i} \overset{(2.23)}{\leq} \sum_{i=1}^{m} \ln\Big(1 - \frac{\varepsilon \Delta u_i}{a_i}\Big) \;=\; \ln\left[\prod_{i=1}^{m}\Big(1 - \frac{\varepsilon \Delta u_i}{a_i}\Big)\right]. \end{aligned}$$

Damit ist

$$C_\varepsilon \;\leq\; \prod_{i=1}^{m}\Big(1 - \frac{\varepsilon \Delta u_i}{a_i}\Big) \;=\; \frac{\prod_{i=1}^{m}\{a_i - \varepsilon \Delta u_i\}}{\prod_{i=1}^{m} a_i}.$$

Es gilt also (2.34). Daraus folgt

$$\begin{aligned} &\lim_{m\to\infty} \prod_{i=1}^{m}\{1 + g(u_i)\Delta u_i + r(\Delta u_i)\} \\ &\qquad\qquad \geq \lim_{m\to\infty} C_\varepsilon \prod_{i=1}^{m}\{1 + g(u_i)\Delta u_i\} \overset{\varepsilon\to 0}{\longrightarrow} \lim_{m\to\infty} \prod_{i=1}^{m}\{1 + g(u_i)\Delta u_i\} \end{aligned}$$

und damit (2.33). Die Gültigkeit von (2.25) zeigt die folgende Gleichungskette.

$$\begin{aligned} \lim_{m\to\infty} \prod_{i=1}^{m}[1 + g(u_i)\Delta u_i] \;&=\; \exp\left\{\ln\left(\lim_{m\to\infty} \prod_{i=1}^{m}[1 + g(u_i)\Delta u_i]\right)\right\} \\ &\overset{\ln \textit{ stetig}}{=} \exp\left\{\lim_{m\to\infty} \sum_{i=1}^{m} \ln\big[1 + g(u_i)\Delta u_i\big]\right\} \\ &\overset{(*)}{=} \exp\left\{\lim_{m\to\infty} \sum_{i=1}^{m} \Big[g(u_i)\Delta u_i + r\big(g(u_i)\Delta u_i\big)\Big]\right\} \\ &\overset{(**)}{=} \exp\left\{\lim_{m\to\infty} \sum_{i=1}^{m} g(u_i)\Delta u_i\right\} \;=\; \exp\left\{\int_a^b g(u)\, du\right\}. \end{aligned}$$

Für großes m ist Δu_i klein, ebenso $g(u_i)\Delta u_i$. $(*)$ gilt also nach Lemma 2.1. Bei dem Beweis von $(**)$ ist zu beachten, dass der Fehlerterm $r\big(g(u_i)\Delta u_i\big)$ nach Lemma 2.1 durch

$$r\big(g(u_i)\Delta u_i\big) \;=\; -\frac{[g(u_i)\Delta u_i]^2}{2(1+\xi)^2} \;\le\; 0, \quad \xi \in (0, g(u_i)\Delta u_i),$$

gegeben ist und damit lediglich

$$\lim_{m\to\infty} \sum_{i=1}^{m} r(g(u_i)\Delta u_i) \;\ge\; 0. \tag{2.35}$$

zu zeigen bleibt. Es gilt $r\big(g(u_i)\Delta u_i\big) = r\big(g(u_i^m)\Delta u_i^m\big) = o\big(g(u_i^m)\Delta u_i^m\big)$ für $g(u_i^m)\Delta u_i^m \to 0$, mit anderen Worten

$$\forall\, \varepsilon > 0 \;\; \exists\, m_0 \;\; \forall\, m \ge m_0 : \quad \big|r\big(g(u_i^m)\Delta u_i^m\big)\big| \le \varepsilon\, g(u_i^m)\Delta u_i^m. \tag{2.36}$$

Sei $\varepsilon > 0$ beliebig und m_0 so gewählt, dass (2.36) gilt. Wegen der Beschränktheit von g folgt dann mit $m \ge m_0$

$$\begin{aligned}
\sum_{i=1}^{m} r\big(g(u_i)\Delta u_i\big) \;&\ge\; \sum_{i=1}^{m} -\varepsilon\, g(u_i)\Delta u_i \\
&\ge\; -\varepsilon\, \sup\{\, g(u) \mid u \in (a,b]\,\} \sum_{i=1}^{m} \Delta u_i \\
&=\; -\varepsilon\, \sup\{\, g(u) \mid u \in (a,b]\,\}\, (b-a).
\end{aligned}$$

Da $\varepsilon > 0$ in (2.36) beliebig klein gewählt werden kann, gilt also

$$\lim_{m\to\infty} \sum_{i=1}^{m} r\big(g(u_i)\Delta u_i\big) \;\ge\; -\varepsilon\, \sup\{\, g(u) \mid u \in (a,b]\,\}\, (b-a) \stackrel{\varepsilon\to 0}{\longrightarrow} 0$$

und damit (2.35) bzw. $(**)$.

2. Als Treppenfunktion ist G zwischen ihren Sprungstellen konstant. Es tragen folglich nur die Intervalle $(u_{i-1}, u_i]$ zum Produkt-Integral bei, die mindestens eine der Unstetigkeitsstellen enthalten. Für hinreichend großes m ist maximal eine Sprungstelle a_j Element des Intervalls $(u_{i-1},\, u_i]$. Ist dies der Fall, so ist

$1 + G(u_i) - G(u_{i-1}) = 1 + g_j$. Es gilt also

$$\mathcal{P}_a^b\,[\,1 + d\,G(u)\,] \;=\; \lim_{m\to\infty} \prod_{i=1}^{m} \{1 + G(u_i) - G(u_{i-1})\} \;=\; \prod_{j:a<a_j\le b} (1 + g_j).$$

3. Sind $a_j,\ j = 1, 2, \ldots$, die Sprungstellen von G, so setze man o.B.d.A. $a_1 = \min\{a_j \mid a_j > a\}$ und $a_n = \max\{a_j \mid a_j \le b\}$. Ist $b = a_n$ und setzt man $a = a_0$, so kann nach Satz 2.4 das Produkt-Integral $\mathcal{P}_a^b\,[1 + d\,G(u)]$ zerlegt werden in

$$\begin{aligned}\mathcal{P}_{(a,\,b]}\,[1 + d\,G(u)] \;=\;& \mathcal{P}_{(a_0,\,a_1-\Delta]}\,[1 + d\,G(u)]\;\mathcal{P}_{(a_1-\Delta,\,a_1]}\,[1 + d\,G(u)] \\ & \mathcal{P}_{(a_1,\,a_2-\Delta]}\,[1 + d\,G(u)]\;\mathcal{P}_{(a_2-\Delta,\,a_2]}\,[1 + d\,G(u)]\ldots \\ & \ldots\mathcal{P}_{(a_{n-1},\,a_n-\Delta]}\,[1 + d\,G(u)]\;\mathcal{P}_{(a_n-\Delta,\,a_n]}\,[1 + d\,G(u)].\end{aligned}$$

Da G auf den Intervallen $(a_{j-1}, a_j - \Delta],\ j = 1, \ldots, n$, stetig ist, ist mit Teil 1 des Satzes zunächst die folgende Darstellung möglich

$$\begin{aligned}\mathcal{P}_{(a,\,b]}\,[1 + d\,G(u)] \;=\;& \exp\left(\int_{a_0}^{a_1-\Delta} g(u)\,du\right)\mathcal{P}_{(a_1-\Delta,\,a_1]}\,[1 + d\,G(u)] \\ & \exp\left(\int_{a_1}^{a_2-\Delta} g(u)\,du\right)\mathcal{P}_{(a_2-\Delta,\,a_2]}\,[1 + d\,G(u)]\ldots \\ & \ldots\exp\left(\int_{a_{n-1}}^{a_n-\Delta} g(u)\,du\right)\mathcal{P}_{(a_n-\Delta,\,a_n]}\,[1 + d\,G(u)].\end{aligned}$$

Für $\Delta \to 0$ ist $\mathcal{P}_{(a_j-\Delta,\,a_j]}\,[1 + d\,G(u)] = 1 + g_j,\ j = 1, \ldots, n$. Wegen der Kommutativität der Multipikation vereinfacht sich der obige Ausdruck des Produkt-Integrals damit schließlich zu

$$\mathcal{P}_{(a,\,b]}\,[1 + d\,G(u)] \;=\; \exp\left(\int_a^b g(u)\,du\right)\prod_{j=1}^{n}(1 + g_j),$$

was mit der Definition von a_1 und a_n gerade der Behauptung entspricht. Für $b > a_n$ zerlegt man das Intervall $(a, b]$ in $a = a_0 < a_1 - \Delta < a_1 < a_2 - \Delta < a_2 < \ldots < a_n - \Delta < a_n < b$ und erhält mit entsprechenden Überlegungen das gleiche Ergebnis. □

Satz 2.6 (Darstellung von S durch λ II). *Für eine beliebige Lebenszeit T kann die Survival-Funktion S mittels der kumulierten Hazard-Funktion Λ und des Produkt-Integrals $\mathcal{P}$ wie folgt dargestellt werden*

$$S(t) \;=\; \mathcal{P}_0^t\,[\,1 - d\,\Lambda(u)\,]. \tag{2.37}$$

Beweis: Sei die kumulierte Hazard-Funktion Λ wie in Definition 2.5 gegeben. Die allgemeine Darstellung (2.37) erhält man aufgrund der Tatsache, dass für jede beliebige Zerlegung $0 = u_0 < u_1 < \ldots < u_m = t$ von $(0, t]$ gilt

$$S(t) \;=\; P(T > t) \;=\; \prod_{i=1}^{m} P(T > u_i \mid T > u_{i-1}),$$

insbesondere also auch

$$S(t) \;=\; \lim_{m\to\infty} \prod_{i=1}^{m} P(T > u_i \mid T > u_{i-1}). \tag{2.38}$$

Der folgende Beweis zeigt, dass der Limes im rechten Teil der Gleichung (2.38) gerade dem Produkt-Integral in (2.37) entspricht. Man hat zunächst

$$\begin{aligned} P(T > u_i \mid T > u_{i-1}) &= 1 - P(T \le u_i \mid T > u_{i-1}) \\ &= 1 - \frac{P(u_{i-1} < T \le u_i)}{P(T > u_{i-1})} \overset{(2.16)}{=} 1 - \frac{1}{S(u_{i-1})} \int_{u_{i-1}}^{u_i} d\,F(u) \\ &\overset{(2.17)}{=} 1 - \left[\frac{1}{S(u_{i-1})} \int_{u_{i-1}}^{u_i} f(u)\,du + \frac{1}{S(u_{i-1})} \sum_{u_{i-1} < a_j \le u_i} p(a_j)\right]. \end{aligned}$$

Für den diskreten Teil des letzten Ausdrucks ist im Beweis zu Satz 2.3 gezeigt worden, dass $\lambda(a_j) = p(a_j)/S(a_{j-1})$. Es ist also zunächst

$$\frac{1}{S(u_{i-1})} \sum_{u_{i-1} < a_j \le u_i} p(a_j) \;=\; \frac{1}{S(u_{i-1})} \sum_{u_{i-1} < a_j \le u_i} \lambda(a_j)\,S(a_{j-1}). \tag{2.39}$$

Da bei einem hinreichend kleinen $\Delta u_i = u_i - u_{i-1}$ das Intervall $(u_{i-1},\, u_i]$ höchstens nur eine Sprungstelle a_j von F enthält, ist

$$S(a_{j-1}) \overset{(2.5)}{=} \sum_{a_k > a_{j-1}} p(a_k) = \sum_{a_k > u_{i-1}} p(a_k) \overset{(2.5)}{=} S(u_{i-1})$$

und (2.39) kann weiter umgeformt werden zu

$$\frac{1}{S(u_{i-1})} \sum_{u_{i-1}<a_j\le u_i} p(a_j) \;=\; \sum_{u_{i-1}<a_j\le u_i} \lambda(a_j). \tag{2.40}$$

Für den stetigen Teil gilt mit $\Delta u_i = u_i - u_{i-1}$

$$\begin{aligned}
\frac{1}{S(u_{i-1})} \int_{u_{i-1}}^{u_i} f(u)\,du &= \frac{1}{S(u_{i-1})}\,[F(u_i) - F(u_{i-1})] \\
&= \frac{1}{S(u_{i-1})}\,[F(u_{i-1}+\Delta u_i) - F(u_{i-1})] \\
&\overset{Taylor}{=} \frac{1}{S(u_{i-1})}\,[F(u_{i-1}) + f(u_{i-1})\Delta u_i + r(\Delta u_i) - F(u_{i-1})] \\
&= \frac{f(u_{i-1})}{S(u_{i-1})}\,\Delta u_i + \frac{r(\Delta u_i)}{S(u_{i-1})},
\end{aligned}$$

mit $r(\Delta u_i) = o(\Delta u_i)$ für $\Delta u_i \to 0$. Da S beschränkt ist, gilt folglich auch $r_1(\Delta u_i) := r(\Delta u_i)/S(u_{i-1}) = o(\Delta u_i)$ für $\Delta u_i \to 0$. Wegen $\lambda(t) = f(t)/S(t)$ erhält man ferner

$$\begin{aligned}
\frac{1}{S(u_{i-1})} \int_{u_{i-1}}^{u_i} f(u)\,du &= \lambda(u_{i-1})\Delta u_i + r_1(\Delta u_i) \\
&= \int_{u_{i-1}}^{u_i} \lambda(u)\,du + r_1(\Delta u_i) + r_2(\Delta u_i),
\end{aligned}$$

wobei für den Fehler $r_2(\Delta u_i)$ gilt

$$\begin{aligned}
& r_2(\Delta u_i) \;\le\; \Delta u_i \sup\big\{\, |\lambda(x) - \lambda(y)| \;\big|\; x, y \in (u_{i-1}, u_i] \,\big\} \\
\iff\quad & \frac{r_2(\Delta u_i)}{\Delta u_i} \;\le\; \sup\big\{\, |\lambda(x) - \lambda(y)| \;\big|\; x, y \in (u_{i-1}, u_i] \,\big\} \longrightarrow 0 \quad \text{für} \quad \Delta u_i \to 0.
\end{aligned}$$

Setzt man $r_3(\Delta u_i) = r_1(\Delta u_i) + r_2(\Delta u_i)$, so ergibt sich

$$\begin{aligned}
P(T > u_i \mid T > u_{i-1}) &= 1 - \left[\int_{u_{i-1}}^{u_i} \lambda(u)\,du + \sum_{u_{i-1}<a_j\le u_i} \lambda(a_j) + r_3(\Delta u_i) \right] \\
&= 1 - \big[\Lambda(u_i) - \Lambda(u_{i-1}) + r_3(\Delta u_i)\big]
\end{aligned}$$

bzw.

$$\lim_{m\to\infty} \prod_{i=1}^{m} P(T > u_i \mid T > u_{i-1}) \;=\; \lim_{m\to\infty} \prod_{i=1}^{m} \big\{1 - \big[\Lambda(u_i) - \Lambda(u_{i-1}) + r_3(\Delta u_i)\big]\big\}.$$

Da für für den Fehler $r_3(\Delta u_i) = r_1(\Delta u_i) + r_2(\Delta u_i)$ ebenfalls $r_3(\Delta u_i) = o(\Delta u_i)$ für $\Delta u_i \to 0$ gilt, folgt mit der Definition des Produkt-Integrals insgesamt

$$\begin{aligned} S(t) &= \lim_{m\to\infty} \prod_{i=1}^{m} \{1 - [\Lambda(u_i) - \Lambda(u_{i-1}) + r(\Delta u_i)]\} \\ &\overset{(*)}{=} \lim_{m\to\infty} \prod_{i=1}^{m} \{1 - [\Lambda(u_i) - \Lambda(u_{i-1})]\} \\ &= \mathcal{P}_0^t [\, 1 - d\,\Lambda(u)\,]. \end{aligned}$$

Die Gleichheit bei $(*)$ zeigt man dabei mit $a_i := 1 - [\Lambda(u_i) - \Lambda(u_{i-1})]$ ganz analog zu dem Beweis von Gleichung (2.24) in Satz 2.5, siehe dazu S. 15ff. □

Bemerkung 2.3. Die Beziehung (2.37) gilt für alle Typen von Verteilungen. Für stetige Zufallsvariablen erhält man die aus Abschnitt 2.2.1 bekannte Darstellung $S(t) = \exp\big(-\int_0^t \lambda(u)\,du\big)$ durch Anwenden des ersten Teils von Satz 2.5. Im diskreten Fall folgt der Ausdruck $S(t) = \prod_{t_j \le t}[\, 1 - \lambda(t_j)\,]$ direkt aus Gleichung (2.26).

Bemerkung 2.4. Alternativ kann $\mathcal{P}$ analog zum Riemann-Integral definiert werden. In diesem Fall ergibt sich die folgende Darstellung für die Survival-Funktion:

$$\mathcal{P}_0^t[\, 1 - \lambda(u)\,du\,] = \lim_{m\to\infty} \prod_{i=1}^{m} [1 - \lambda(\xi_i)(u_i - u_{i-1})]\}. \tag{2.41}$$

Auch hier wird eine immer feinere Zerlegung des Intervalls $(0, t]$ betrachtet, d.h. $0 = u_0 < u_1 < u_1 < \ldots < u_m = t$ mit $\Delta u_i = u_i - u_{i-1}$ und $\max(\Delta u_i) \to 0$ für $m \to \infty$, sowie $\xi_i \in (u_{i-1}, u_i]$. Zu finden ist der Ausdruck (2.41) zum Beispiel in [Cox-84, S. 15].

2.2.3 Bemerkungen

Die Bedeutung der Hazard-Funktion liegt in erster Linie darin, das unmittelbare Sterberisiko eines Individuums zu bestimmen, von welchem bekannt ist, dass es zu einem bestimmten Zeitpunkt t noch am Leben ist. Der Verlauf der Hazard-Funktion beschreibt dabei, wie sich das augenblickliche Ausfallrisiko in Abhängigkeit von der bereits verlebten Zeit verändert.

In der parametrischen Überlebenszeitanalyse verwendet man die Hazard-Funktion zum Bestimmen der Ausfallverteilung. Solche auf der Hazard-Rate basierenden Modelle verwenden qualitative Informationen über den Ausfallmechanismus. Es gibt

mehrere allgemeine Formen der Hazard-Funktion. So zeigen einige einen (streng) monoton wachsenden Verlauf, andere sind (streng) monoton fallend, Badewannen- oder Hügel-fömig. Modelle mit wachsender Hazard-Rate ergeben sich im Zusammenhang mit natürlicher Alterung oder Abnutzung. Fallende Hazard-Funktionen sind weitaus weniger üblich, finden jedoch gelegentlich Gebrauch bei sehr frühen Sterbewahrscheinlichkeiten, wie sie sich zum Beispiel nach Organtransplantationen ergeben. Badewannen-förmige Hazard-Funktionen eignen sich für von Geburt an beobachtete Populationen. So folgen die meisten Sterblichkeitstafeln dieser Hazard-Form: Die vielen Todesfälle der ersten Lebensphase erklären sich primär durch typische Kinderkrankheiten. Es folgt eine Phase mit niedriger Sterbewahrscheinlichkeit und anschließend – basierend auf natürlichen Alterungsprozessen – eine wachsende Hazard-Rate. Hügel-förmige Hazard-Funktionen werden verwendet, um das Überleben nach chirurgischen Behandlungen zu modellieren. Dem operativen Eingriff folgt aufgrund von Infektionen, Blutungen oder anderen Komplikationen ein anfänglicher Zuwachs des Sterberisikos. Erholt sich der Patient, so sinkt dieses Risiko wieder ab.

Das Erwägen der Hazard-Funktion ist schließlich auch dann nützlich, wenn Individuen mehrerer Gruppen miteinander verglichen werden sollen (siehe dazu Kapitel 7, insbesondere Abschnitt 7.3.2 und 7.4.4).

2.3 Die erwartete Restlebensdauer

Die erwartete Restlebensdauer (mean residual life function) ist der vierte grundlegende Parameter in der Analyse von Lebenszeiten. Für Individuen des Alters t gibt sie an, welche restliche Lebensdauer diesen im Mittel noch verbleibt.

Definition 2.7. *Die erwartete Restlebensdauer ist definiert als*

$$mrl(t) = E(T - t \mid T > t), \quad t \geq 0. \tag{2.42}$$

Satz 2.7. *Es gelten folgende Beziehungen:*

1. *Ist die Zufallsvariable T stetig verteilt, so gilt:*

$$mrl(t) = \frac{1}{S(t)} \int_t^\infty S(u)\, du, \quad t \geq 0. \tag{2.43}$$

Die Größe $mrl(t)$ entspricht also der Fläche unterhalb des Graphen der Survival-Funktion rechts von t geteilt durch $S(t)$.
Für diskrete Zufallsvariablen auf dem Zeitraster $0 < t_1 < t_2 < \ldots$ gilt daher entsprechend:

$$mrl(t) = \frac{(t_{i+1} - t)\, S(t_i) + \sum_{j \geq i+1} (t_{j+1} - t_j)\, S(t_j)}{S(t)}, \tag{2.44}$$

für $t_i \leq t < t_{i+1}$.

2. *Ist T stetig verteilt, so hat man:*

$$E(T) = mrl(0), \tag{2.45}$$

$$Var(T) = 2 \int_0^\infty t\, S(t)\, dt - \left(\int_0^\infty S(t)\, dt \right)^2. \tag{2.46}$$

Beweis: 1. Hat T eine stetige Dichte f, so gilt

$$\begin{aligned} \mathrm{E}(T - t \mid T > t) &= \frac{\int_t^\infty (u-t) f(u)\, du}{S(t)} \\ &\overset{part.Int,\ f(t) = -S'(t)}{=} \frac{1}{S(t)} \left(-(u-t)S(u)\Big|_t^\infty + \int_t^\infty S(u)\, du \right) \\ &\overset{lim_{t\to\infty} S(t) = 0}{=} \frac{1}{S(t)} \int_t^\infty S(u)\, du. \end{aligned}$$

Für eine diskrete Zufallsvariable erhält man die Formel (2.44) mittels partieller Summation: Es gilt

$$\mathrm{E}(T - t \mid T > t) = \frac{1}{S(t)} \sum_{t_j > t} (t_j - t)\, p(t_j).$$

Dabei ist für $t_i \leq t < t_{i+1}$

$$\begin{aligned} &\sum_{t_j > t} (t_j - t)\, p(t_j) = \\ &= (t_{i+1} - t) \sum_{j \geq i+1} p(t_j) + (t_{i+2} - t_{i+1}) \sum_{j \geq i+2} p(t_j) + (t_{i+3} - t_{i+2}) \sum_{j \geq i+3} p(t_j) + \ldots \\ &= (t_{i+1} - t)\, S(t_i) + (t_{i+2} - t_{i+1})\, S(t_{i+1}) + (t_{i+3} - t_{i+2})\, S(t_{i+3}) + \ldots \\ &= (t_{i+1} - t)\, S(t_i) + \sum_{j \geq i+1} (t_{j+1} - t_j) S(t_j). \end{aligned}$$

Die folgende Abbildung 2.1 veranschaulicht diese Summation für eine diskrete Zufallsvariable T mit $p(t_1) = \frac{1}{4},\ p(t_2) = \frac{1}{8},\ p(t_3) = \frac{3}{8},\ p(t_4) = \frac{1}{4}$ und $t_1 < t < t_2$.

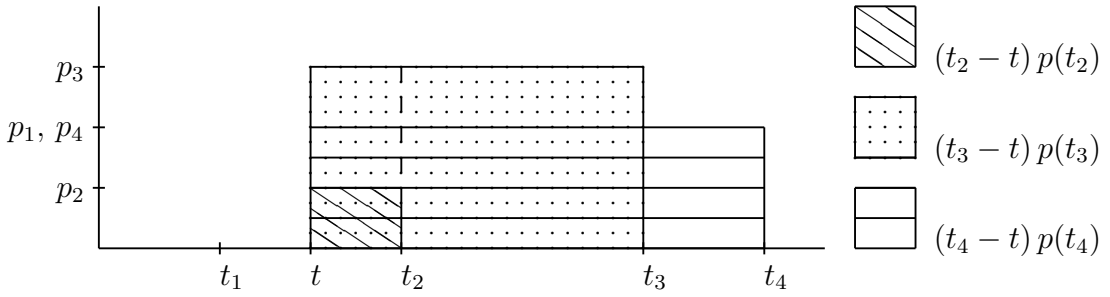

Abbildung 2.1: Hilfsskizze zum Beweis von Satz 2.7.

2. Der Ausdruck (2.45) folgt direkt aus der Definition des Erwartungswertes und (2.46) ergibt sich dann wegen $\mathrm{Var}(T) = \mathrm{E}(T^2) - [\mathrm{E}(T)]^2$: Es ist

$$\begin{aligned} \mathrm{E}(T^2) &= \int_0^\infty t^2\, f(t)\, dt = -t^2\, S(t)\Big|_0^\infty + \int_0^\infty 2\, t\, S(t)\, dt \\ &= \int_0^\infty 2\, t\, S(t)\, dt, \quad \text{da } S(t) \stackrel{t\to\infty}{\longrightarrow} 0 \end{aligned}$$

und damit

$$\mathrm{Var}(T) = 2 \int_0^\infty t\, S(t)\, dt - \left(\int_0^\infty S(t)\, dt \right)^2 .$$

□

2.4 Zusammenfassung

Wie zu Beginn des Kapitels erwähnt, lassen sich – falls eine der eingeführten Größen bekannt ist – die anderen drei eindeutig bestimmen. Der Zusammenhang zwischen erwarteter Restlebensdauer, Survival-, Hazard- und Dichte-Funktion lässt sich folgendermaßen zusammenfassen:

Satz 2.8. *Ist die Lebenszeit T stetig verteilt, so gilt*

1. *für die Survival-Funktion:*

$$\begin{aligned} S(t) &\overset{Def}{=} P(T > t) \\ &\overset{(2.2)}{=} \int_t^\infty f(u)\,du \\ &\overset{(2.11)}{=} \exp\left(-\int_0^t \lambda(u)\,du\right) \\ &\overset{Def\,\Lambda}{=} \exp\left(-\Lambda(t)\right) \\ &\overset{\text{a)}}{=} \frac{mrl(0)}{mrl(t)} \exp\left(-\int_0^t \frac{1}{mrl(u)}\,du\right), \end{aligned}$$

2. *für die Dichtefunktion:*

$$\begin{aligned} f(t) &\overset{(2.3)}{=} -\frac{\partial}{\partial t} S(t) \\ &\overset{(2.8)}{=} \lambda(t)\,S(t) \\ &\overset{\text{b)}}{=} \left[\frac{d}{dt} mrl(t) + 1\right] \left[\frac{mrl(0)}{mrl^2(t)}\right] \exp\left(-\int_0^t \frac{1}{mrl(u)}\,du\right), \end{aligned}$$

3. *für die Hazard-Funktion:*

$$\begin{aligned} \lambda(t) &\overset{Def}{=} \lim_{\Delta \to 0^+} \frac{P(t \le T < t + \Delta \mid T \ge t)}{\Delta} \\ &\overset{(2.8)}{=} -\frac{\partial}{\partial t} \ln S(t) \\ &\overset{(2.8)}{=} \frac{f(t)}{S(t)} \\ &\overset{\text{c)}}{=} \left[\frac{d}{dt} mrl(t) + 1\right] \frac{1}{mrl(t)}, \end{aligned}$$

4. für die erwartete Restlebensdauer:

$$\begin{aligned} mrl(t) &\overset{Def}{=} E\left(T-t \mid T>t\right) \\ &\overset{(2.43)}{=} \frac{1}{S(t)}\int_t^\infty S(u)\,du \\ &\overset{Bew\,(2.43)}{=} \frac{1}{S(t)}\int_t^\infty (u-t)f(u)\,du. \end{aligned}$$

Ist die Lebenszeit T diskret verteilt, so gilt

1. für die Survival-Funktion:

$$\begin{aligned} S(t) &\overset{(2.5)}{=} \sum_{t_j>t} p(t_j) \\ &\overset{(2.12)}{=} \prod_{t_j\le t}[\,1-\lambda(t_j)\,], \end{aligned}$$

2. für die diskrete Dichte:

$$p(t_j) = S(t_{j-1}) - S(t_j) \overset{(2.13)}{=} \lambda(t_j)S(t_{j-1}), \quad j=1,2,\ldots,$$

3. für die Hazard-Funktion:

$$\lambda(t_j) \overset{(2.13)}{=} \frac{p(t_j)}{S(t_{j-1})}, \quad j=1,2,\ldots,$$

4. für die erwartete Restlebensdauer:

$$mrl(t) = \frac{(t_{i+1}-t)\,S(t_i) + \sum_{j\ge i+1}(t_{j+1}-t_j)\,S(t_j)}{S(t)}, \quad t_i \le t < t_{i+1}.$$

Beweis: Zu zeigen bleibt lediglich die Gleichheit bei a), b) und c): Mit $mrl(t) = \frac{1}{S(t)}\int_t^\infty S(u)\,du$ folgt c):

$$\begin{aligned} \frac{1}{mrl(t)}\left[\frac{\partial}{\partial t}mrl(t)+1\right] &= \frac{1}{mrl(t)}\left[\frac{f(t)}{S^2(t)}\int_t^\infty S(u)\,du - \frac{S(t)}{S(t)} + 1\right] \\ &= \frac{1}{mrl(t)}\left[\frac{f(t)}{S^2(t)}\int_t^\infty S(u)\,du\right] \\ &= \frac{1}{mrl(t)}\left[\frac{f(t)}{S(t)}mrl(t)\right] = \frac{f(t)}{S(t)}. \end{aligned}$$

Damit lässt sich a) nachrechnen:

$$\begin{aligned}
&\exp\left(-\int_0^t \lambda(u)\,du\right)\\
&\quad\overset{c)}{=}\ \exp\left[-\int_0^t \left(\frac{\partial}{\partial u}\,mrl(u)+1\right)\frac{1}{mrl(u)}\,du\right]\\
&\quad=\ \exp\left[-\int_0^t \frac{\partial}{\partial u}\Big(mrl(u)\Big)\frac{1}{mrl(u)}\,du\right]\exp\left[-\int_0^t \frac{1}{mrl(u)}\,du\right]\\
&\quad=\ \exp\left[-\int_0^t \frac{\partial}{\partial u}\Big(\ln\big[mrl(u)\big]\Big)\,du\right]\exp\left[-\int_0^t \frac{1}{mrl(u)}\,du\right]\\
&\quad=\ \exp\Big[\ln\big(mrl(0)\big)-\ln\big(mrl(t)\big)\Big]\exp\left[-\int_0^t \frac{1}{mrl(u)}\,du\right]\\
&\quad=\ \frac{mrl(0)}{mrl(t)}\exp\left(-\int_0^t \frac{1}{mrl(u)}\,du\right)
\end{aligned}$$

und schließlich b):

$$\begin{aligned}
&\left[\frac{d}{dt}mrl(t)+1\right]\left[\frac{mrl(0)}{mrl^2(t)}\right]\exp\left(-\int_0^t \frac{1}{mrl(u)}\,du\right)\\
&\quad\overset{Bew\,c)}{=}\ \left[\frac{f(t)}{S(t)}mrl(t)\right]\left[\frac{mrl(0)}{mrl^2(t)}\right]\exp\left(-\int_0^t \frac{1}{mrl(u)}\,du\right)\\
&\quad\overset{(2.8)}{=}\ \lambda(t)\,\frac{mrl(0)}{mrl(t)}\exp\left(-\int_0^t \frac{1}{mrl(u)}\,du\right)\\
&\quad\overset{a)}{=}\ \lambda(t)\,S(t).
\end{aligned}$$

□

Kapitel 3

Zensierte Daten und Konstruktion der Likelihood-Funktion

Es kann vorkommen, dass das interessierende Ereignis – wie zum Beispiel der Tod eines Krebspatienten oder der Ausfall eines bestimmten Maschinenbauteils – nicht beobachtet werden kann. Bei einer Krebsstudie kann ein solcher Fall zum Beispiel dann eintreten, wenn ein Patient wegen eines Wohnortwechsels aus der Studie ausscheidet oder aufgrund anderer Ursachen als der untersuchten Krebsart verstirbt. Auch das vorzeitige Beenden der Studie selbst kann dazu führen, dass Lebensdauern nicht vollständig beobachtet werden. Man spricht von zensierten Daten.

Um zensierte Daten geeignet auswerten zu können, ist es wichtig zu wissen, durch welchen Mechanismus sie hervorgerufen werden. Bei den verschiedenen Arten der Zensur unterscheidet man zwischen rechts-, links- und intervallweise zensierten Ausfallzeiten. Im ersten Abschnitt dieses Kapitels werden diese drei Typen der Zensur vorgestellt. Da die Rechtszensierung jedoch den für die Praxis wichtigsten Fall darstellt, soll der Schwerpunkt in der Analyse rechtszensierter Lebensdaten liegen. Die Auswertung der unvollständigen Datensätze wird dabei hauptsächlich auf der Maximum-Likelihood-Theorie basieren. In Abschnitt 3.2 wird die Likelihood-Funktion für rechtszensierte Daten konstruiert und in Abschnitt 3.3 die für die folgenden Kapitel relevanten Aspekte der Large-Sample-Theorie zusammengefasst.

3.1 Zensierungsformen

In diesem Abschnitt werden die drei Grundtypen der Zensur vorgestellt: die Rechts-, Links- und Intervall-Zensierung. Rechtszensierten Datensätzen soll dabei die größte Aufmerksamkeit gewidmet werden. Die Beschreibung der verschiedenen Zensierungsformen folgt dabei im Wesentlichen den Darstellungen von Klein und Moeschberger [Kle-97, S. 55–65], Collet [Col-03, S. 1–4] und Lawless [Law-03, S. 52–57]. Bezüglich der Charakterisierung weiterer Zensurarten, wie der progressiven Typ-II-Zensur oder der Stutzung (truncation) wird an dieser Stelle auf die Ausführungen in [Law-03] und [Kle-97] verwiesen.

3.1.1 Typ-I-Rechtszensur

Wird aus Zeit- oder Kostengründen eine Studie beendet, bevor das interessierende Ereignis bei allen Teilnehmern oder Einheiten eingetreten ist, so spricht man von der Typ-I-Rechtszensierung. In derartigen Experimenten beobachtet man ein Kollektiv von n Individuen vom Zeitpunkt $t = 0$ des Studienbeginns bis zu einem vorher festgelegten Zeitpunkt $t = c$, zu dem das Experiment abgebrochen wird. Ein Ausfall wird demnach nur dann beobachtet, wenn er sich vor dem festgesetzten Zeitpunkt c ereignet. Andernfalls hat man lediglich die Information, dass die Lebensdauer des Individuums größer ist als c. Die Typ-I-Rechtszensierung ergibt sich typischerweise bei ingenieurwissenschaftlichen Untersuchungen zur Funktionsdauer von elektronischen oder mechanischen Maschinenbauteilen. Abbildung 3.1 veranschaulicht den möglichen Verlauf einer solchen Studie mit fünf Einheiten. Bei der Auswertung von Daten aus medizinischen Untersuchungen muss beachtet werden, dass Patienten in der Regel nicht gleichzeitig in die Studie eintreten, sondern während einer Anfangsphase zu verschiedenen Zeitpunkten rekrutiert werden. In diesem Zusammenhang unterscheidet man die Studien-Zeit (auch Kalender-Zeit oder study time) von der Patienten-Zeit (patient time). Beide Begriffe geben die Zeitperiode an, in der ein Individuum unter Beobachtung steht. Während die Studien-Zeit dabei die relative Beobachtungsdauer angibt, handelt es sich bei der Patienten-Zeit um die zeitliche Differenz zwischen Beobachtungsbeginn und Todeszeitpunkt. Als absolute Beobachtungsdauer ist sie damit die eigentlich interessierende Lebenszeit. Abbildung 3.2 stellt die Studien-Zeit für sechs Patienten bei Typ-1-Rechtszensur dar. Die entsprechende Patienten-Zeit wird in Abbildung 3.3 veranschaulicht. Beide Graphiken sind, wie die restlichen Abbildungen des Kapitels, an [Lee-03, Figure 1.1 – Figure 1.3] angelehnt.

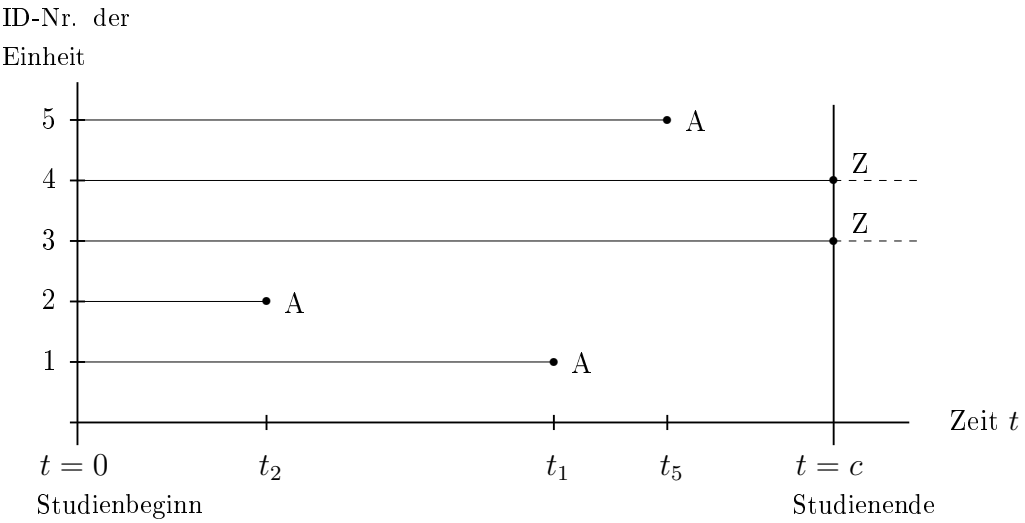

Abbildung 3.1: Typ-I-Rechtszensur unter Laborbedingungen. A: Einheit ausgefallen; Z: Lebensdauer der Einheit zensiert; $t_1,\ t_2, \ldots$ beobachtete Ausfallzeiten der Versuchseinheiten.

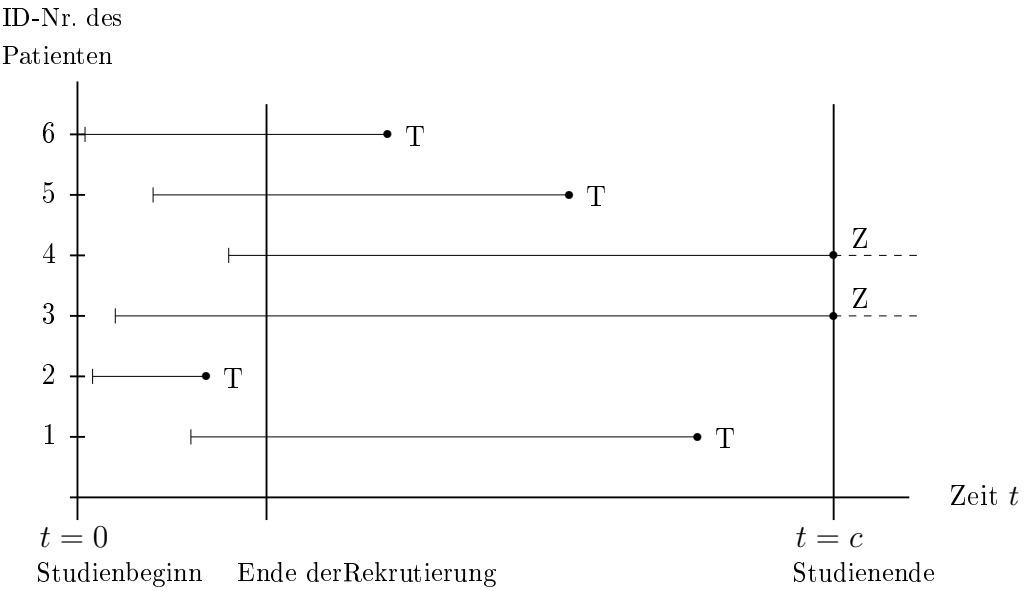

Abbildung 3.2: Studien-Zeit bei Typ-I-Rechtszensur. Die Individuen treten zu unterschiedlichen Zeitpunkten in die Studie ein. | : Rekrutierung des Patienten; T: Interessierendes Ereignis (hier T für Tod) ist beobachtet worden; Z: Zensierung.

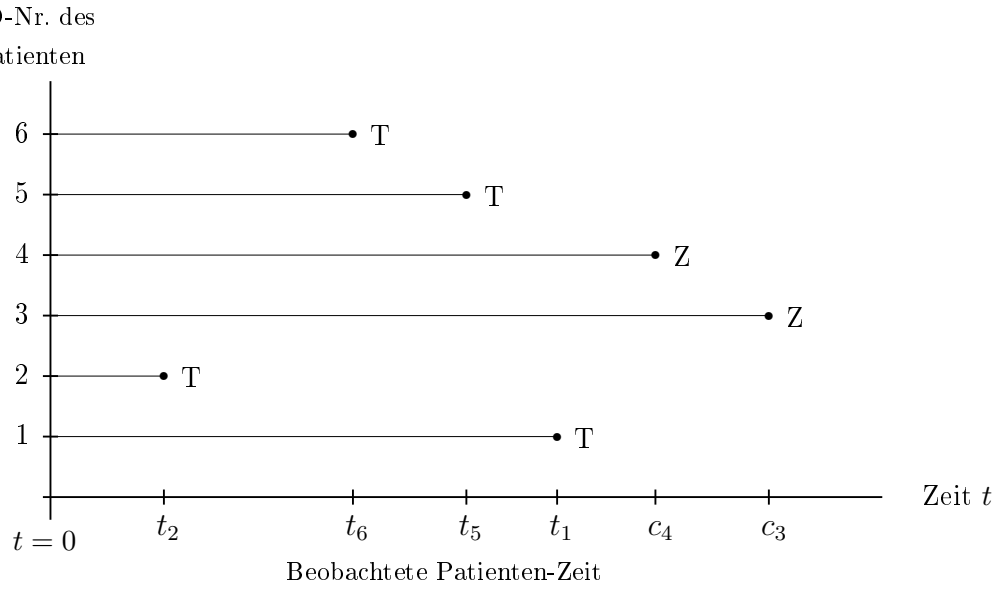

Abbildung 3.3: Patienten-Zeit bei Typ-I-Rechtszensur. Der individuelle Beobachtungsbeginn ist bei allen Patienten auf Null gesetzt worden. Bei T bzw. t_i, $i = 1, 2, 5, 6$, ist der Tod des Patienten beobachtet worden. Bei Z bzw. c_3 und c_4 konnte die tatsächliche Lebensdauer aufgrund des Studienabbruchs nicht gemessen werden.

3.1.2 Typ-II-Rechtszensur

Bei der Typ-II-Rechtszensur werden n Einheiten so lange beobachtet, bis die ersten k von ihnen ausfallen. Die Zahl $k < n$ wird dabei vor Beginn der Untersuchung festgesetzt. Experimente zur Lebensdauer von Industrieprodukten beinhalten oft diese Art der Zensur. Da das Prüfen der Einheiten abgebrochen wird, sobald eine bestimmte Anzahl von ihnen versagt, erfordern solche Verfahren – ebenso wie die mit Typ-I-Rechtszensur – weitaus weniger Geld und Zeit als diejenigen, die den Ausfall aller Prüfgegenstände abwarten. Obwohl die Typ-II-Zensur bei einigen Lebenszeit-Experimenten Anwendung findet, hat sie den Nachteil, dass die gesamte Dauer des Versuchs von der Ausfallzeit $t_{(k)}$ abhängt und diese zu Beginn des Tests unbekannt ist. In geplanten Experimenten ist die Typ-I-Rechtszensur daher viel weiter verbreitet. Abbildung 3.4 stellt die Typ-II-Zensur am Beispiel von sechs Einheiten dar. Dabei ist $k = 4$ gesetzt, d.h. der Abbruchzeitpunkt des Experiments ist derjenige, zu dem die vierte Versuchseinheit ausfällt.

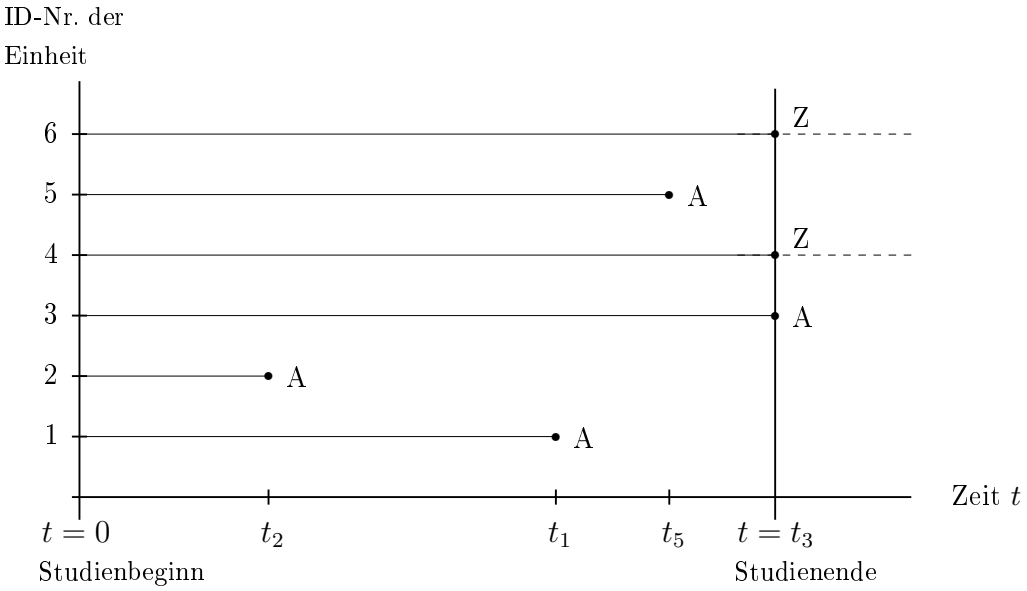

Abbildung 3.4: Typ-II-Rechtszensur unter Laborbedingungen. A: Einheit ausgefallen; Z: Lebensdauer der Einheit zensiert; $t_1,\ t_2, \ldots$ beobachtete Ausfallzeiten der Versuchseinheiten.

3.1.3 Zufällige Rechtszensur

Versuchseinheiten oder Patienten können aus zahlreichen Gründen noch vor dem Beenden der Studie der Beobachtung entzogen werden. Studienteilnehmer können den Wohnort wechseln, an anderen als der untersuchten Krankheit sterben oder sich aus Desinteresse nicht mehr an der Studie beteiligen. In ingenieurwissenschaftlichen Untersuchungen ist es möglich, dass Maschinen und Geräte aufgrund des Versagens anderer als der zu prüfenden Bauteile ausfallen. Der Eintritt des interessierenden Ereignisses ist dann nicht beobachtbar. Man spricht von zufälliger Rechtszensierung. Bei klinischen Tests ist die Art der Zensierung häufig eine Mischung aus Zufalls- und Typ-I-Zensur. Diese Kombination wird im Folgenden als zufällige Typ-I-Rechtszensur bezeichnet. Die beiden Abbildungen 3.5 und 3.6 stellen sie graphisch dar.

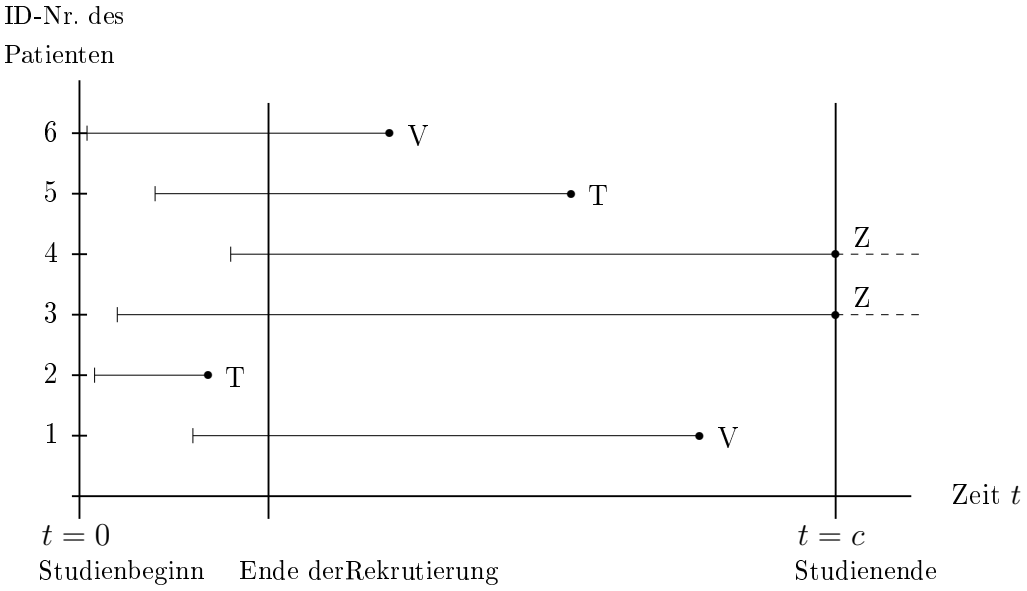

Abbildung 3.5: Studien-Zeit bei zufälliger Typ-I-Rechtszensur. Die Individuen treten zu unterschiedlichen Zeitpunkten in die Studie ein. Z: Zensur – Patient überlebt das Ende der Studie; T: Tod – Patient verstirbt aufgrund der untersuchten Krankheit; V: Verlust – Patient scheidet aufgrund anderer Ursachen aus der Studie aus.

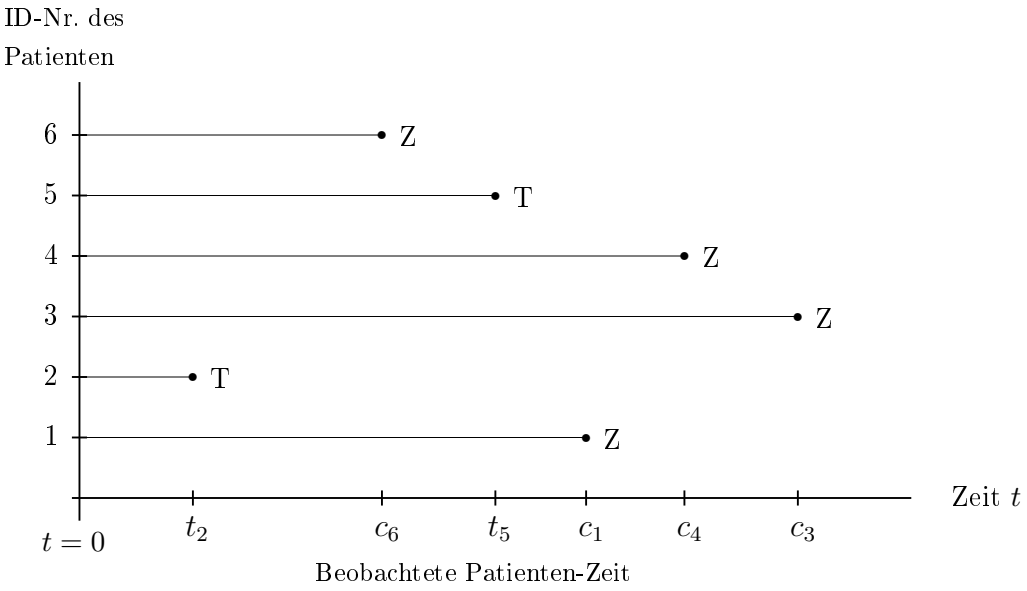

Abbildung 3.6: Patienten-Zeit bei zufälliger Typ-I-Rechtszensur. Der individuelle Beobachtungsbeginn ist bei allen Patienten auf Null gesetzt worden. Bei T bzw. t_2 und t_5 ist der Tod des Patienten tatsächlich beobachtet worden. Bei Z bzw. c_i, $i = 1, 3, 4, 6$, konnte die tatsächliche Lebensdauer aufgrund der zufälligen Rechtszensur bzw. des Studienabbruchs nicht gemessen werden.

3.1.4 Links- und Intervallzensur

Von linksseitig zensierten Ausfalldaten spricht man, wenn das interessierende Ereignis bereits vor Beginn der Studie eingetreten ist. Der genaue Zeitpunkt ist dann unbekannt, man weiß nur, dass das Ereignis irgendwann vor dem Startpunkt der Beobachtung stattgefunden hat. Dieser Typ zensierter Daten kommt gelegentlich in empirischen Untersuchungen vor. Soll mit Hilfe einer Umfrage ermittelt werden, wie die Verteilung der Zeit ist, zu der Jugendliche das erste Mal nach einer Zigarette greifen, so kann sich als mögliche Antwort ergeben, dass die befragte Person zwar Raucher ist, aber nicht mehr wisse seit wann. Die Information ist folglich, dass der Jugendliche zum interessierenden Zeitpunkt jünger gewesen ist als beim Interview.

Ein allgemeinerer Typ der Zensur, die sogenannte Intervallzensur, ergibt sich, wenn Ereignisse in bestimmte Zeitintervalle $(L, R]$ fallen. Wird bei einer ärztlichen Untersuchung eine Krankheit diagnostiziert, so ist lediglich bekannt, dass der Zeitpunkt ihres Ausbruchs zwischen den beiden letzten Arztbesuchen liegt. Bei industriellen Experimenten kommt es zur Intervallzensierung, wenn Maschinenbauteile nur periodisch auf ordnungsgemäßes Funktionieren überprüft werden.

3.2 Konstruktion der Likelihood-Funktion für rechtszensierte Daten

Lebenszeiten unterliegen häufig den in Abschnitt 3.1 beschriebenen Mechanismen der Rechtszensur. Ein Ziel ist demnach die Entwicklung von Methoden, mit denen solche unvollständigen Datensätze adäquat ausgewertet werden können. Die statistische Inferenz kann dabei auf der wohlbekannten Maximum-Likelihood-Theorie beruhen. Als Basis für die Thematik nachstehender Kapitel wird im Folgenden die Likelihood-Funktion entwickelt. Dabei wird sich überraschenderweise Weise herausstellen, dass sie bei den verschiedenen Typen der Rechtszensur dieselbe Form annimmt. Die Ausführungen dieses Abschnitts beruhen auf den Darstellungen von Klein und Moeschberger [Kle-97, S. 65–70], sowie denen von Lawless [Law-03, S. 52–57]. Entsprechend der dortigen Vorgehensweise werden zunächst einige Bezeichnungen eingeführt.

3.2.1 Notation für die mathematische Modellierung der Rechtszensur

Für das mathematische Modell der Rechtszensur betrachte man eine endliche Folge $T_1, \ldots, T_n$ von unabhängigen identisch verteilten nicht-negativen Zufallsvariablen, die die wahren Lebensdauern einer n-elementigen Population beschreiben und durch T repräsentiert werden. Die Zeiten $T_i,\ i = 1, \ldots, n$, seien dem Risiko ausgesetzt von rechts zensiert zu werden. Beobachtungen sind folglich Werte y_i, die entweder den wahren Lebenszeiten entsprechen ($y_i = t_i$), oder aber zensiert sind. Die explizite Definition der, einer Beobachtung y_i zugrundeliegenden Zufallsvariable Y_i soll bei der Betrachtung der einzelnen Zensurmechanismen in den folgenden Abschnitten formuliert werden. Zur Präzision des Sachverhalts definiere man an dieser Stelle jedoch den Zensurindikator:

$$\Delta_i \ = \ \mathbb{1}(T_i = y_i) \ \ = \ \ \begin{cases} 1\,, & \text{falls} \quad T_i = y_i \quad \text{(wahre Lebensdauer beobachtet)} \\ 0\,, & \text{falls} \quad T_i > y_i \quad \text{(Zensur-Zeit beobachtet)} \end{cases}$$

Zu beachten ist, dass der Zensurindikator eine Zufallsvariable ist. Um im Folgenden allgemein zwischen der Zufallsvariable Δ_i und ihrer Realisierung zu unterscheiden, wird für die letztgenannte die Bezeichnung δ_i verwendet.Das Ergebnis eines Experiments mit n Individuen wäre also der Datensatz

$$\big((y_1, \delta_1), \ldots, (y_n, \delta_n)\big).$$

Im Fall $\delta_i = 0$ kann die tatsächliche Lebensdauer zwar nicht gemessen werden, aber es ist bekannt, dass Patient i mindestens bis zum Zeitpunkt $T_i = y_i$ am Leben gewesen ist. Das gänzliche Ignorieren dieser Beobachtung würde damit zu einem Informationsverlust führen.

3.2.2 Typ-I-Rechtszensur

Unterliegen die Lebenszeiten der betrachteten n-elementigen Population einer Rechtszensur vom Typ I, so hat jedes Individuum eine feste potentielle Zensur-Zeit c_i. Treten die Individuen zu unterschiedlichen Zeitpunkten in die Studie ein, so gilt im Allgemeinen $c_i \neq c_j$ für $i \neq j$. Bei Experimenten unter Laborbedingungen ist dagegen $c_i = c$ für alle $i = 1, \ldots, n$. Man vergleiche hierzu die Abbildungen 3.1, 3.2 und 3.3. Die Lebenszeit T_i wird in beiden Fällen nur beobachtet, wenn $T_i \leq c_i$

gilt. Andernfalls lautet die Information $T_i > c_i$. Die Stichprobe besteht aus Realisierungen der Zufallsvariablen $(Y_i\,,\,\Delta_i)\,,\ i = 1, \ldots, n$, die mit den oben festgelegten Bezeichnungen und den soeben gemachten Überlegungen wie folgt dargestellt werden können:

$$Y_i = \min(T_i\,,\,c_i) \qquad \text{und} \qquad \Delta_i = \begin{cases} 1, & \text{falls} \quad T_i \leq c_i, \\ 0, & \text{falls} \quad T_i > c_i. \end{cases}$$

Die Likelihood-Funktion für eine Typ-I-zensierte Stichprobe basiert auf der gemeinsamen Verteilung von $(Y_i\,,\,\Delta_i)\,,\ i = 1, \ldots, n$. Diese wird im folgenen Satz angegeben.

Satz 3.1 (Dichte von (Y_i, Δ_i) bei Typ-I-Zensur). *Im Modell der Typ-I-Rechtszensur ist die Dichte der gemeinsamen Verteilung von $(Y_i, \Delta_i)\,,\ i = 1, \ldots, n$, gegeben durch*

$$f_{(Y_i,\Delta_i)}(y_i\,,\,\delta_i) \;=\; f_{T_i}(y_i)^{\delta_i}\, S_{T_i}(c_i)^{1-\delta_i}. \tag{3.1}$$

Dabei gibt c_i die potentielle Zensurzeit des Individuums i an. S_{T_i} ist die Survival-Funktion zu T_i und mit f_{T_i} wird hier sowohl eine Dichtefunktion als auch eine Zähldichte bezeichnet, je nachdem ob die Verteilung von T_i stetig oder diskret ist.

Beweis: Die Zensurzeiten c_i sind feste Konstanten und für die Beobachtungen gilt $y_i \leq c_i$. Ist die Lebenszeit T_i diskret mit Zähldichte $p_{T_i}(t_j) = f_{T_i}(t_j)$, so ist

$$\begin{aligned} P\big[(Y_i, \Delta_i) = (y_i, 1)\big] &= P\big[T_i = y_i\big] = f_{T_i}(y_i), \\ P\big[(Y_i, \Delta_i) = (c_i, 0)\big] &= P\big[T_i > c_i\big] = S_{T_i}(c_i). \end{aligned}$$

Ist T_i eine absolut stetige Zufallsvariable, so betrachte man mit $\varepsilon > 0$ die Approximation $P\big[Y_i \in [y_i, y_i + \varepsilon),\ \Delta_i = \delta_i\big]$. Es ist

$$\begin{aligned} \lim_{\varepsilon \to 0} P\big[Y_i \in [y_i, y_i + \varepsilon),\ \Delta_i = 1\big] &= \lim_{\varepsilon \to 0} P\big[T_i \in [y_i, y_i + \varepsilon)\big] \\ &= \lim_{\varepsilon \to 0} \int_{y_i}^{y_i + \varepsilon} f_{T_i}(u)\, du \;=\; f_{T_i}(y_i), \\ \lim_{\varepsilon \to 0} P\big[Y_i \in [y_i, y_i + \varepsilon),\ \Delta_i = 0\big] &= P\big[T_i > c_i\big] = S_{T_i}(c_i). \end{aligned}$$

Für die gemeinsame Verteilung ergibt sich in beiden Fällen also wie behauptet $f_{(Y_i,\Delta_i)}(y_i\,,\,\delta_i) = f_{T_i}(y_i)^{\delta_i} S_{T_i}(c_i)^{1-\delta_i}$. □

Die Likelihood-Funktion kann nun im Fall der Typ-I-Rechtszensierung formuliert werden.

Satz 3.2 (**Likelihood-Funktion für $(\boldsymbol{y}, \boldsymbol{\delta})$ bei Typ-I-Zensur**). *Die Likelihood-Funktion für die Stichprobe* $(\mathbf{y}, \boldsymbol{\delta}) = \big((y_1, \delta_1), \ldots, (y_n, \delta_n)\big)$ *ist bei Typ-I-Rechtszensur gegeben durch*

$$L\big(P_T\,;\,(\mathbf{y}, \boldsymbol{\delta})\big) \;=\; \prod_{i=1}^{n} f_{T_i}(y_i)^{\delta_i} S_{T_i}(y_i)^{1-\delta_i}. \tag{3.2}$$

Beweis: Wegen der Unabhängigkeit von $(Y_i\,,\,\Delta_i)$, $i = 1, \ldots, n$, gilt für die Likelihood-Funktion

$$\begin{aligned} L\big(P_T\,;\,(y,\delta)\big) &= f_{\big((Y_1,\Delta_1),\ldots,(Y_n,\Delta_n)\big)}\big((y_1,\delta_1),\ldots,(y_n,\delta_n)\big) \\ &= \prod_{i=1}^{n} f_{(Y_i,\Delta_i)}(y_i,\delta_i) \\ &\overset{(3.1)}{=} \prod_{i=1}^{n} f_{T_i}(y_i)^{\delta_i}\, S_{T_i}(c_i)^{1-\delta_i}. \end{aligned}$$

Die Darstellung (3.2) ist also eine direkte Folgerung aus Satz 3.1. □

3.2.3 Zufällige Rechtszensur

Im Modell der zufälligen Rechtszensur wird angenommen, dass jedes Individuum i eine Lebenszeit T_i und eine Zensurzeit C_i besitzt. Neben den nicht-negativen Zufallsvariablen $T_1, \ldots, T_n$, die nach Voraussetzung unabhängig identisch verteilt sind, betrachtet man nun eine endliche Folge von identisch verteilten Zensurzeiten, $C_1, \ldots, C_n$. Man nehme an, dass sowohl die Zenssurzeiten $C_1, \ldots, C_n$ untereinander als auch T_i und C_i stochastisch unabhängig sind für $i = 1, \ldots, n$. Als weitere Voraussetzung lege man fest, dass die Verteilungsfunktion F_{C_i} in keinem Parameter mit F_{T_i} übereinstimmt. In diesem Zusammenhang spricht man gelegentlich auch von unabhängiger rechtsseitiger Zufallszensur. Wie im Fall der Typ-I-Rechtszensur erhält man im Experiment die Realisierungen der Zufallsvariablen

$$Y_i = \min(T_i\,,\,C_i) \qquad \text{und} \qquad \Delta_i = \begin{cases} 1, & \text{falls} \quad Y_i = T_i, \quad \text{d.h.} \quad T_i \le C_i, \\ 0, & \text{falls} \quad Y_i = C_i, \quad \text{d.h.} \quad T_i > C_i. \end{cases}$$

Die Zufallsvariablen $(Y_1, \Delta_1), \ldots, (Y_n, \Delta_n)$ sind identisch und unabhängig verteilt. Die gemeinsame Verteilung von (Y_i, Δ_i) erhält man analog zur Typ-I-Zensur.

Satz 3.3 (Dichte von (Y_i, Δ_i) bei Zufallszensur). *Im Modell der zufälligen Rechtszensur ist die Dichte der Verteilung von* (Y_i, Δ_i), $i = 1, \ldots, n$, *gegeben durch*

$$f_{(Y_i,\Delta_i)}(y_i\,,\,\delta_i) \;=\; \left[f_{T_i}^{\delta_i}(y_i) S_{T_i}^{1-\delta_i}(y_i)\right]\left[f_{C_i}^{1-\delta_i}(y_i) P^{\delta_i}(C_i \geq y_i)\right]. \tag{3.3}$$

S_{T_i}, f_{T_i} und f_{C_i} bezeichenen dabei in üblicher Weise die Survival-Funktion und – abhängig davon, ob T_i und C_i stetig oder diskret sind – die Dichtefunktionen oder Zähldichten der entsprechenden Randverteilungen. Sind die Zensurzeiten stetig verteilt, so gilt natürlich $P(C_i \geq y_i) = S_{C_i}(y_i)$.

Beweis: Da T_i und C_i stochastisch unabhängig sind, gilt im diskreten Zeitmodell

$$\begin{aligned} P\big[(Y_i, \Delta_i) = (y_i, 1)\big] &= P\big[T_i = y_i,\, C_i \geq y_i\big] \;=\; f_{T_i}(y_i)\, P(C_i \geq y_i), \\ P\big[(Y_i, \Delta_i) = (y_i, 0)\big] &= P\big[T_i > y_i,\, C_i = y_i\big] \;=\; S_{T_i}(y_i)\, f_{C_i}(y_i), \end{aligned}$$

mit diskreten Zähldichten $f_{T_i}(t_j)$ und $f_{C_i}(c_j)$. Bei stetigen Wahrscheinlichkeitsverteilungen betrachte man für $\varepsilon > 0$ erneut die Approximation $P\big[Y_i \in [y_i, y_i + \varepsilon),\ \Delta_i = \delta_i\big]$:

$$\begin{aligned} \lim_{\varepsilon\to 0} P\big[Y_i \in [y_i, y_i+\varepsilon),\ \Delta_i = 1\big] &= \lim_{\varepsilon\to 0} P\big[T_i \in [y_i, y_i+\varepsilon),\ C_i \geq T_i\big] \\ &= \lim_{\varepsilon\to 0} \int_{y_i}^{y_i+\varepsilon} \int_{u}^{\infty} f_{(T_i,C_i)}(u,v)\, dv\, du \\ &= \lim_{\varepsilon\to 0} \int_{y_i}^{y_i+\varepsilon} \int_{u}^{\infty} f_{T_i}(u)\, f_{C_i}(v)\, dv\, du \\ &= \lim_{\varepsilon\to 0} \int_{y_i}^{y_i+\varepsilon} f_{T_i}(u)\, S_{C_i}(u)\, du \\ &= f_{T_i}(y_i)\, S_{C_i}(y_i) \end{aligned}$$

und analog

$$\begin{aligned} \lim_{\varepsilon\to 0} P\big[Y_i \in [y_i, y_i+\varepsilon),\ \Delta_i = 0\big] &= \lim_{\varepsilon\to 0} P\big[T_i > C_i,\ C_i \in [y_i, y_i+\varepsilon)\big] \\ &= \lim_{\varepsilon\to 0} \int_{y_i}^{y_i+\varepsilon} \int_{v}^{\infty} f_{T_i}(u)\, f_{C_i}(v)\, du\, dv \\ &= \lim_{\varepsilon\to 0} \int_{y_i}^{y_i+\varepsilon} S_{T_i}(v)\, f_{C_i}(v)\, dv \\ &= S_{T_i}(y_i)\, f_{C_i}(y_i) \end{aligned}$$

mit stetigen Dichten $f_{T_i}(t)$ und $f_{C_i}(c)$. Damit hat man also wie behauptet:

$$\begin{aligned} f_{(Y_i,\Delta_i)}(y_i,\delta_i) &= \big[f_{T_i}(y_i)P\big(C_i \geq y_i\big)\big]^{\delta_i}\big[S_{T_i}(y_i)f_{C_i}(y_i)\big]^{1-\delta_i} \\ &= \big[f_{T_i}^{\delta_i}(y_i)S_{T_i}^{1-\delta_i}(y_i)\big]\,\big[f_{C_i}^{1-\delta_i}(y_i)P^{\delta_i}(C_i \geq y_i)\big]\,. \end{aligned}$$

□

Die Darstellung der Likelihood-Funktion ergibt sich wieder direkt aus der Verteilung von (Y_i, Δ_i):

Satz 3.4 (**Likelihood-Funktion für $(\boldsymbol{y}, \boldsymbol{\delta})$ bei Zufallszensur**). *Die Likelihood-Funktion für die Stichprobe* $(\mathbf{y}, \boldsymbol{\delta}) = \big((y_1,\delta_1),\ldots,(y_n,\delta_n)\big)$ *ist bei zufälliger Rechtszensur gegeben durch*

$$L\big(P_T, P_C\,;\,(\mathbf{y},\boldsymbol{\delta})\big) = \prod_{i=1}^{n} f_{T_i}^{\delta_i}(y_i)S_{T_i}^{1-\delta_i}(y_i)\prod_{i=1}^{n} f_{C_i}^{1-\delta_i}(y_i)P^{\delta_i}(C_i \geq y_i). \tag{3.4}$$

Korollar 3.1. *Da die Verteilungsfunktionen F_{T_i} und F_{C_i} keine gemeinsamen Parameter besitzen, können die beiden Faktoren der Likelihood-Funktion in (3.4) einzeln maximiert werden, P_{T_i} also insbesondere unabhängig von P_{C_i} geschätzt werden. Als Likelihood-Funktion für die interessierende Lebensverteilung erhält man demnach wie im Fall der Typ-I-Zensur*

$$L\big(P_T\,;\,(\mathbf{y},\boldsymbol{\delta})\big) = \prod_{i=1}^{n} f_{T_i}(y_i)^{\delta_i}S_{T_i}(y_i)^{1-\delta_i}. \tag{3.5}$$

Bemerkung 3.1. Fällt die Masse der Verteilung von C_i jeweils auf einen einzigen festen Punkt c_i, so erhält man als Spezialfall der Zufallszensur die Typ I-Zensur aus Abschnitt 3.2.2.

3.2.4 Typ-II-Rechtszensur

Eine Rechtszensur vom Typ II ergibt sich, wenn n Einheiten mit Lebensdauern $T_1,\ldots,T_n$ gleichzeitig in die Studie aufgenommen werden und die Studie abgebrochen wird, sobald eine bestimmte Anzahl von ihnen ausfällt. Die Zahl k der Ausfälle wird dabei vor Beginn des Experiments festgelegt und der Datensatz besteht aus den k kleinsten Lebenszeiten, das heißt aus den Ordnungsstatistiken $T_{(1)} \leq \ldots \leq T_{(k)}$. Ist die Verteilung von T stetig, so folgt die Gestalt der Likelihood-Funktion direkt aus einem Ergebnis für Ordnungsstatistiken.

Satz 3.5 (Ordnungsstatistiken). *Seien $X_1, \ldots, X_n$ unabhängig identisch verteilte Zufallsvariablen mit stetiger Dichte f und Verteilungsfunktion F. Werden diese n Zufallsvariablen in aufsteigender Reihenfolge angeordnet und notiert als*

$$X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)},$$

so heißt $X_{(k)}$, $1 \leq k \leq n$, die k-te Ordnungsstatistik. Für die gemeinsame Dichte von $X_{(1)}, \ldots, X_{(k)}$ gilt

$$f_{(X_{(1)},\ldots,X_{(k)})}(x_1, \ldots, x_k) \;=\; \frac{n!}{(n-k)!} \left(\prod_{i=1}^{k} f(x_{(i)}) \right) \left[1 - F(x_{(k)})\right]^{n-k}. \tag{3.6}$$

Beweis: Siehe Arnold, Balakrishnan und Nagaraja [Arn-92, Kapitel 2], insbesondere Beweis zu Theorem 2.4.3. □

Satz 3.6. *Im Modell der Typ-II-Rechtszensur sei f_T die Dichtefunktion und S_T die Survival-Funktion der Lebenszeiten $T_1, \ldots, T_n$. Die gemeinsame Dichte der k kleinsten Lebenszeiten $T_{(1)} < \ldots < T_{(k)}$ ist dann gegeben durch*

$$f_{(T_{(1)},\ldots,T_{(k)})}(t_1, \ldots, t_k) \;=\; \frac{n!}{(n-k)!} \left(\prod_{i=1}^{k} f_{T_i}(t_{(i)}) \right) S_{T_i}(t_{(k)})^{n-k} \tag{3.7}$$

und die Likelihood-Funktion lässt sich wie bei den anderen Typen der Rechtszensur darstellen als

$$L\big(P_T\,;\,(\mathbf{y}, \boldsymbol{\delta})\big) \;=\; \prod_{i=1}^{n} f_{T_i}(y_i)^{\delta_i} S_{T_i}(y_i)^{1-\delta_i}\,. \tag{3.8}$$

Beweis: Die Darstellung (3.7) entspricht der Dichte aus Satz 3.5. Die Likelihood-Funktion erhält man daraus wie folgt. Ist $(y_{(1)}, \delta_1) \leq \ldots \leq (y_{(n)}, \delta_n)$ der geordnete Datensatz, so gilt $(y_{(i)}, \delta_i) = (y_{(i)}, 1) = (t_{(i)}, 1)$ für $i = 1, \ldots, k$ und $(y_{(i)}, \delta_i) = (y_{(i)}, 0) = (t_{(k)}, 0)$ für $i = k+1, \ldots, n$. Die Likelihood-Funktion basiert also auf der gemeinsamen Verteilung der k kleinsten Ordnungsstatistiken $T_{(1)} \leq \ldots \leq T_{(k)}$

$$\begin{aligned} L\big(P_T\,;\,(\mathbf{y}, \boldsymbol{\delta})\big) \;&=\; f_{(T_{(1)},\ldots,T_{(k)})}(y_{(1)}, \ldots y_{(k)}) \;=\; \frac{n!}{(n-k)!} \left(\prod_{i=1}^{k} f_{T_i}(y_{(i)}) \right) S_{T_i}(y_{(k)})^{n-k} \\ &=\; \frac{n!}{(n-k)!} \prod_{i=1}^{n} f_{T_i}(y_i)^{\delta_i} S_{T_i}(y_i)^{1-\delta_i} \end{aligned}$$

und der Faktor $\frac{n!}{(n-k)!}$ kann als Konstante aus der Likelihood-Funktion entfernt werden. □

3.3 Aspekte der Large-Sample-Theorie

In diesem Abschnitt soll ein Überblick über diejenigen Resultate der asymptotischen Statistik gegeben werden, die für die Thematik der nachfolgenden Kapitel von Bedeutung sind. Einleitend soll mit Definitionen und Sätzen der Likelihood-Theorie begonnen werden. Dazu betrachte man n identisch und unabhängig verteilte Zufallsvariablen $Y_1, \ldots, Y_n \sim Y$ mit Dichtefunktion $f_Y(y; \boldsymbol{\theta})$, wobei $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)' \in \Theta$ ein p-dimensionaler Parameter-Vektor ist. Sei $y = (y_1, \ldots, y_n)$ der Datensatz eines Experiments, dann ist die Likelihood-Funktion zu dieser Beobachtung bekanntermaßen gegeben durch

$$L(\boldsymbol{\theta}) \;=\; \prod_{i=1}^{n} f_Y(y_i; \boldsymbol{\theta}). \tag{3.9}$$

Sei $l(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta})$ die Log-Likelihood-Funktion und $\hat{\boldsymbol{\theta}} \in \Theta$ der Maximum-Likelihood-Schätzer für $\boldsymbol{\theta}$.

Definition 3.1 (**Informationsmatrix**). *Die $p \times p$-Matrix $I(\boldsymbol{\theta})$ mit den Einträgen*

$$I_{ij}(\boldsymbol{\theta}) \;=\; \frac{-\partial^2 l(\boldsymbol{\theta})}{\partial\theta_i \partial\theta_j}, \quad i,\ j = 1, \ldots p, \tag{3.10}$$

heißt Informationsmatrix.

Definition 3.2 (**Erwartete Informationsmatrix (Fisher-Matrix)**). [Paw-01, S. 216], [Lin-96, S. 98], [Wit-95, S. 154f] *Die $p \times p$-Matrix $\mathcal{I}(\boldsymbol{\theta})$ mit den Einträgen*

$$\mathcal{I}_{ij}(\boldsymbol{\theta}) \;=\; E_{\boldsymbol{\theta}}\left(\frac{-\partial^2 l(\boldsymbol{\theta})}{\partial\theta_i \partial\theta_j}\right), \quad i,\ j = 1, \ldots p, \tag{3.11}$$

heißt erwartete Informationsmatrix oder Fisher-Matrix.

Definition 3.3 (**Beobachtete Informationsmatrix**). [Paw-01, S. 31f], [Lin-96, S. 96] *Ist $I(\boldsymbol{\theta})$ die Informationsmatrix und $\hat{\boldsymbol{\theta}}$ der Maximum-Likelihood-Schätzer für $\boldsymbol{\theta}$, so heißt $I(\hat{\boldsymbol{\theta}})$ die beobachtete Informationsmatrix.*

Satz 3.7. *Unter schwachen Regularitätsbedingungen, die den angegeben Beweisen zu entnehmen sind, gilt:*

1. *Der ML-Schätzer $\hat{\boldsymbol{\theta}}$ ist ein konsistenter Schätzer für $\boldsymbol{\theta}$.*

2. $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{\mathcal{D}} \mathcal{N}_p\big(\mathbf{0}, \mathcal{I}^{-1}(\boldsymbol{\theta})\big)$

3. *$n^{-1}I(\hat{\boldsymbol{\theta}})$ ist ein konsistenter Schätzer für $n^{-1}\mathcal{I}(\boldsymbol{\theta})$.*

Beweis: 1. [Sch-95, Theorem 7.49, S. 415f] oder [Wit-95, Satz 6.34, S. 201f]
2. [Sch-95, Theorem 7.63, S. 421ff] oder [Wit-95, Satz 6.35, S. 202ff]
3. [Sch-95, im Beweis zu Theorem 7.63, S. 421ff] □

Satz 3.8 (Delta-Methode). [Rao-73, S. 385f] *Sei $(T_n)_{n\in\mathbb{N}}$ eine Folge von Zufallsvariablen, die asymptotisch normal-verteilt ist mit*

$$\sqrt{n}(T_n - \mu) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2). \tag{3.12}$$

Ist f eine differenzierbare Funktion mit $f'(\mu) \neq 0$, so ist auch $f(T_n)$ asymptotisch normal-verteilt und es gilt:

$$\sqrt{n}\,\big[f(T_n) - f(\mu)\big] \xrightarrow{\mathcal{D}} \mathcal{N}\big(0, \sigma^2[f'(\mu)]^2\big). \tag{3.13}$$

Beweis: Nach dem Satz von Taylor kann f in einer Umgebung von μ linear approximiert werden. Für $f(T_n)$ bedeutet das

$$f(T_n) \;=\; f(\mu) + f'(\mu)(T_n - \mu) + r(T_n - \mu), \tag{3.14}$$

wobei $\varepsilon_n := r(T_n - \mu)/(T_n - \mu)$ für $T_n \to \mu$ gegen 0 konvergiert. Es gilt also

$$\forall\, \varepsilon > 0 \;\; \exists\, \delta > 0 : \quad |T_n - \mu| < \delta \;\Rightarrow\; |\varepsilon_n| < \varepsilon$$

und nach Voraussetzung folglich

$$P(\,|\varepsilon_n| < \varepsilon\,) \geq P(\,|T_n - \mu| < \delta\,) \longrightarrow 1 \quad \text{für} \quad n \to \infty.$$

Weil $\varepsilon > 0$ beliebig klein gewählt werden kann, ist insbesondere $\varepsilon_n \xrightarrow{\mathcal{P}} 0$. Da weiter

$\sqrt{n}(T_n - \mu)$ nach Voraussetzung in Verteilung konvergiert, folgt zunächst[1]

$$\sqrt{n}\left[f(T_n) - f(\mu)\right] - \sqrt{n}\left[f'(\mu)(T_n - \mu)\right] \;=\; \sqrt{n}\,(T_n - \mu)\,\varepsilon_n \xrightarrow{\mathcal{P}} 0$$

und schließlich mit dem Lemma von Slutzky[2], dass die asymptotische Verteilung von $\sqrt{n}\left[f(T_n) - f(\mu)\right]$ mit der von $\sqrt{n}\left[f'(\mu)(T_n - \mu)\right]$ übereinstimmt, also durch $\mathcal{N}(0,\, \sigma^2\,[f'(\mu)]^2)$ gegeben ist. □

[1] Für zwei Folgen $(X_n)_{n\in\mathbb{N}}$ und $(Y_n)_{n\in\mathbb{N}}$ von Zufallsvariablen mit $X_n \xrightarrow{\mathcal{P}} 0$ und $Y_n \xrightarrow{\mathcal{D}} Y$ gilt: $X_n\,Y_n \xrightarrow{\mathcal{P}} 0$, siehe z.B. [Rao-73, S. 122f].

[2] Das Lemma von Slutzky besagt: Sind $(X_n)_{n\in\mathbb{N}}$ und $(Y_n)_{n\in\mathbb{N}}$ zwei Folgen von Zufallsvariablen mit $X_n \xrightarrow{\mathcal{D}} X$ und $|X_n - Y_n| \xrightarrow{\mathcal{P}} 0$, so gilt $Y_n \xrightarrow{\mathcal{D}} X$, siehe z.B. [Mue-05, Satz 10.4, S. 133].

Kapitel 4

Der Kaplan-Meier-Schätzer

In diesem Kapitel soll auf der Grundlage von zufallszensierten Daten der Maximum-Likelihood-Schätzer (ML-Schätzer) für die Survival-Funktion S bestimmt werden. Da in vielen Fällen nicht bekannt ist, zu welcher Klasse von Wahrscheinlichkeitsverteilungen die Ausfallzeit T gehört, sollte dieser dann nach Möglichkeit nicht-parametrisch sein. Werden die Zeiten $T_i,\ i = 1, \ldots, n$, vollständig beobachtet, so liefert die empirische Überlebensfunktion in natürlicher Weise einen geeigneten Schätzer:

$$\hat{S}(t) \quad = \quad 1 - \hat{F}_n(t) = 1 - \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{(-\infty, t]}(t_i). \tag{4.1}$$

Unterliegt der Datensatz jedoch einer Zensur, so kann $\hat{F}_n(t)$ gar nicht ermittelt werden und (4.1) ist in einer solchen Situation offenbar nicht geeignet. Eine für zensierte Lebensdaten modifizierte Form der empirischen Überlebensfunktion ist der Kaplan-Meier-Schätzer (KM-Schätzer). Erstmals angegeben und diskutiert von Kaplan und Meier [Kap-58], ist der KM-Schätzer ein fundamentales Instrument der Lebenszeit-Analyse. Im ersten Abschnitt dieses Kapitels wird er als nicht-parametrischer ML-Schätzer hergeleitet. Die nicht-parametrische Schätzung der Survival-Funktion bildet ferner eine Basis für graphische Verfahren, mit deren Hilfe parametrische Modelle überprüft werden können. Sie werden in Kapitel 5 vorgestellt.

4.1 Der KM-Schätzer als ML-Schätzer

Die nachstehende Herleitung des Schätzers für S basiert auf den Darlegungen in [Cox-84, S. 48–50] und [Haf-01, S. 103–107]. Es wird dabei eine nicht-negative

Zufallsvariable T betrachtet, die die Lebenszeiten einer n-elementigen Population repräsentiert. Weiter werden den Überlegungen die mathematischen Modelle der Rechtszensur aus Abschnitt 3.2 zugrundegelegt. Unterliegt der Datensatz $(\mathbf{y}, \boldsymbol{\delta}) = \big((y_1, \delta_1), \ldots, (y_n, \delta_n)\big)$ einem der dort beschriebenen Zensur-Mechanismen, so nimmt die auf $(\mathbf{y}, \boldsymbol{\delta})$ basierende Likelihood-Funktion die folgende Form an:

$$L\big(P_T\,;\,(\mathbf{y}, \boldsymbol{\delta})\big) \;=\; \prod_{i=1}^{n} f_{T_i}(y_i)^{\delta_i} S_{T_i}(y_i)^{1-\delta_i}. \tag{4.2}$$

Will man (4.2) auf der Menge aller Verteilungen maximieren, so stellt man fest, dass es innerhalb der stetigen Verteilungen keinen ML-Schätzer geben kann, weil $L\big(P_T\,;\,(\mathbf{y}, \boldsymbol{\delta})\big) = \prod_{i=1}^{n} f_{T_i}(y_i)^{\delta_i} S_{T_i}(y_i)^{1-\delta_i}$ – interpretiert als Wahrscheinlichkeitsdichte – dann beliebig groß werden kann. Da aber $L\big(P_T\,;\,(\mathbf{y}, \boldsymbol{\delta})\big)$ für stetiges P_T als Wahrscheinlichkeit null ist, für diskretes P_T, sofern die Zeitpunkte $y_1, \ldots, y_n$ positive Wahrscheinlichkeit tragen, eine ebenfalls positive Wahrscheinlichkeit darstellt, reicht es zur Maximierung der Likelihood-Funktion (4.2) aus, P_T über die Menge der diskreten Verteilungen laufen zu lassen. (Man bemerke, dass auch die empirische Verteilungsfunktion ein diskreter ML-Schätzer für eine Verteilungsfunktion F ist, die sowohl stetig als auch diskret sein kann.)

Weil die Survival-Funktion $S = S_{T_i}$ nach Satz 2.3 durch die Hazard-Rate dargestellt werden kann, besteht die Idee nun darin, die Likelihood-Funktion $L\big(P_T\,;\,(\mathbf{y}, \boldsymbol{\delta})\big)$ als Funktion in $\lambda = \lambda_{T_i}$ darzustellen. Dazu werden einige Bezeichnungen vereinbart.

Annahmen und Bezeichnungen 4.1. Es wird o.B.d.A. angenommen, dass der Datensatz $\big((y_1, \delta_1), \ldots, (y_n, \delta_n)\big)$ nach steigenden Werten von y_i geordnet ist:

$$y_1 \leq y_2 \leq \cdots \leq y_n.$$

$\big\{z_i \in \{y_1, \ldots, y_n\} \mid z_i \neq z_j \text{ für } i \neq j;\ i, j = 1, \ldots, k\big\}$ mit $k \leq n$ sei die Menge der paarweise verschiedenen Beobachtungszeiten, die ebenfalls steigend geordnet sind:

$$z_1 < z_2 < \cdots < z_k.$$

Weiter wird vereinbart, dass:

- $\mathbf{d_j} = \sharp\{\,(y_j, \delta_j) \mid y_j = z_j,\ \delta_j = 1\,\}$ die Anzahl der Individuen ist, die zum Zeitpunkt z_j ausfallen,

- $\mathbf{c_j} = \sharp\{(y_j, \delta_j) \mid y_j = z_j,\ \delta_j = 0\}$ die Anzahl der Individuen ist, die zum Zeitpunkt z_j zensiert werden und

- $\mathbf{n_j} = \sharp\{(y_j, \delta_j) \mid y_j \geq z_j\}$ die Anzahl der Einheiten, die unmittelbar vor dem Zeitpunkt z_j noch unter Beobachtung stehen. Dabei gilt: $n_1 = n$ und $n_{j+1} = n_j - d_j - c_j$.

Satz 4.1 (Darstellung von $L\big(P_T\,;\,(y,\delta)\big)$ durch die Hazard-Funktion). *Mit Hilfe der Hazard-Rate λ und der Bezeichnungen 4.1 kann die Likelihood-Funktion (4.2) geschrieben werden als*

$$L\big(P_T\,;\,(\mathbf{y},\boldsymbol{\delta})\big) \;=\; \prod_{j=1}^{k} \lambda(z_j)^{d_j}[1-\lambda(z_j)]^{n_j-d_j}. \tag{4.3}$$

Beweis: Da die Zufallsvariablen $T_1, \ldots, T_n$ identisch verteilt sind, setze $p = f = f_{T_i}$ und $S = S_{T_i}$. Mit den Bezeichnungen 4.1 kann die Likelihood-Funktion (4.2) dann zunächst wie folgt dargestellt werden:

$$L\big(P_T\,;\,(\mathbf{y},\boldsymbol{\delta})\big)\big) \;=\; \prod_{j=1}^{k} p(z_j)^{d_j}\, S(z_j)^{c_j}.$$

Da nach Satz 2.3 für die Survival- und Hazard-Funktion einer diskreten Zufallsvariable auf $0 < z_1 < \ldots < z_k$ gilt, dass

$$\begin{aligned} S(z) &= \prod_{z_i \leq z} [1-\lambda(z_i)] \quad \text{und} \\ \lambda(z_i) &= \frac{p(z_i)}{S(z_{i-1})} \qquad \text{für} \quad i = 1, \ldots, k \quad \text{mit} \quad S(z_0) = 1, \end{aligned}$$

können die Wahrscheinlichkeiten $p_i := p(z_i)$ mit Hilfe der diskreten Hazards $\lambda_i := \lambda(z_i)$ folgendermaßen dargestellt werden

$$\begin{aligned} p_1 &= \lambda_1 S(z_0) &&= \lambda_1 \\ p_2 &= \lambda_2 S(z_1) &&= \lambda_2[1-\lambda_1] \\ &\vdots \\ p_k &= \lambda_k S(z_{k-1}) &&= \lambda_k \prod_{i=1}^{k-1}[1-\lambda_i]\,. \end{aligned}$$

Die Likelihood-Funktion ist also gegeben durch

$$L\big(P_T\,;\,(\mathbf{y},\boldsymbol{\delta})\big) \;=\; \prod_{j=1}^{k}\Big(\lambda_j\prod_{i=1}^{j-1}[1-\lambda_i]\Big)^{d_j}\Big(\prod_{i=1}^{j}[1-\lambda_i]\Big)^{c_j}.$$

Um $L\big(P_T\,;\,(\mathbf{y},\boldsymbol{\delta})\big)$ wie in (4.3) darzustellen, betrachte man die einzelnen Faktoren des Produkts $\prod_{j=1}^{k}\Big(\lambda_j\prod_{i=1}^{j-1}[1-\lambda_i]\Big)^{d_j}\Big(\prod_{i=1}^{j}[1-\lambda_i]\Big)^{c_j}$:

$$\begin{aligned}
j=1:\quad & \lambda_1^{d_1}(1-\lambda_1)^{c_1}\\
j=2:\quad & \lambda_2^{d_2}(1-\lambda_1)^{d_2}(1-\lambda_1)^{c_2}(1-\lambda_2)^{c_2}\\
&\;\vdots\\
j=k:\quad & \lambda_k^{d_k}(1-\lambda_1)^{d_k}(1-\lambda_2)^{d_k}\ldots(1-\lambda_{k-1})^{d_k}\\
& \quad (1-\lambda_1)^{c_k}(1-\lambda_2)^{c_k}\ldots(1-\lambda_{k-1})^{c_k}(1-\lambda_k)^{c_k}.
\end{aligned}$$

Es ergibt sich

$$\begin{aligned}
L\big(P_T\,;\,(\mathbf{y},\boldsymbol{\delta})\big) \;=\;& \lambda_1^{d_1}\,(1-\lambda_1)^{\overbrace{c_1+d_2+c_2+\cdots+d_k+c_k}^{=n_1-d_1}}\\
& \lambda_2^{d_2}\,(1-\lambda_2)^{\overbrace{c_2+d_3+c_3+\cdots+d_k+c_k}^{=n_2-d_2}}\\
& \qquad\cdots\\
& \lambda_k^{d_k}\,(1-\lambda_k)^{\overbrace{c_k}^{=n_k-d_k}}\\
=\;& \prod_{j=1}^{k}\lambda_j^{d_j}(1-\lambda_j)^{n_j-d_j},
\end{aligned}$$

was mit der Notation $\lambda_j=\lambda(z_j)$ gerade dem Ausdruck (4.3) entspricht. □

Logarithmieren und Differenzieren des Ausdrucks (4.3) führt zum Maximum-Likelihood-Schätzer für $\boldsymbol{\lambda}$: Als Log-Likelihood-Funktion erhält man mit $\lambda_j=\lambda(z_j)$

$$\begin{aligned}
l(\boldsymbol{\lambda}) \;:=\; \log L\big(P_T\,;\,(\mathbf{y},\boldsymbol{\delta})\big) \;&=\; \sum_{j=1}^{k}\log\Big(\lambda_j^{d_j}(1-\lambda_j)^{n_j-d_j}\Big)\\
&=\; \sum_{j=1}^{k}\Big[d_j\,\log\lambda_j+(n_j-d_j)\,\log(1-\lambda_j)\Big]. \quad (4.4)
\end{aligned}$$

Differenzieren nach $\boldsymbol{\lambda}$ liefert

$$\operatorname{grad} l(\boldsymbol{\lambda}) \;=\; \left(\frac{\partial l}{\partial \lambda_1}, \quad \dots \quad , \frac{\partial l}{\partial \lambda_k}\right) \;=\; \left(\frac{d_1}{\lambda_1} - \frac{n_1 - d_1}{1-\lambda_1}, \quad \dots \quad , \frac{d_k}{\lambda_k} - \frac{n_k - d_k}{1-\lambda_k}\right).$$

$$\operatorname{grad} l(\boldsymbol{\lambda}) \;\stackrel{!}{=}\; \mathbf{0} \qquad \Longleftrightarrow \qquad (\lambda_1\,, \quad \dots \quad , \lambda_k) = \left(\frac{d_1}{n_1}, \quad \dots \quad , \frac{d_k}{n_k}\right).$$

Die Hesse-Matrix ist folglich

$$\begin{pmatrix} \frac{\partial^2}{\partial\lambda_1\partial\lambda_1} l(\lambda) & \dots & \frac{\partial^2}{\partial\lambda_1\partial\lambda_k} l(\lambda) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial\lambda_k\partial\lambda_1} l(\lambda) & \dots & \frac{\partial^2}{\partial\lambda_k\lambda_k} l(\lambda) \end{pmatrix} = \begin{pmatrix} \frac{d_1-n_1}{(1-\lambda_1)^2} - \frac{d_1}{\lambda_1^2} & & 0 \\ & \ddots & \\ 0 & & \frac{d_k-n_k}{(1-\lambda_k)^2} - \frac{d_k}{\lambda_k^2} \end{pmatrix}.$$

Da für alle $j = 1, \dots, k$ gilt: $d_j,\, n_j,\, \lambda_j > 0$ und $d_j \le n_j$, ist diese Matrix negativ definit und der gefundene stationäre Punkt von L

$$\hat{\lambda} = \left(\hat{\lambda}_1\,, \quad \dots \quad , \hat{\lambda}_k\right) = \left(\frac{d_1}{n_1}, \quad \dots \quad , \frac{d_k}{n_k}\right) \tag{4.5}$$

ein Maximum-Likelihood-Schätzer für λ. Damit $\log(1-\hat{\lambda}_j)$ in (4.4) für alle $j = 1, \dots, k$ definiert ist, muss vorausgesetzt werden, dass zum letzten beobachteten Zeitpunkt z_k nicht sämtliche Versuchseinheiten ausfallen, d.h. dass $d_k < n_k$ gilt. Weil klinische Studien jedoch in der Regel beendet werden, bevor das interessierende Ereignis bei allen Patienten eingetreten ist, stellt diese Annahme für die Praxis keine Einschränkung dar. Das Ergebnis ist der folgende Satz:

Satz 4.2 (Der Kaplan-Meier-Schätzer für die Survival-Funktion). *Der Maximum-Likelihood-Schätzer für die Survival-Funktion S – basierend auf dem Datensatz $(\mathbf{y}, \boldsymbol{\delta}) = \big((y_1, \delta_1)\,, \dots\,, (y_n, \delta_n)\big)$ – ist für $t \le z_k$ gegeben durch:*

$$\hat{S}(t) \;=\; \prod_{z_j \le t} \Big(1 - \hat{\lambda}(z_j)\Big) = \prod_{j:\, z_j \le t} \left(1 - \frac{d_j}{n_j}\right). \tag{4.6}$$

Dabei werden mit $z_1 < \dots < z_k$ die verschiedenen Zeitpunkte unter den $y_1, \dots, y_n$ bezeichnet. n_j kennzeichnet die Anzahl der Einheiten, die unmittelbar vor z_j noch unter Beobachtung stehen und d_j die Anzahl der Einheiten, die zum Zeitpunkt z_j ausfallen. $\hat{S}(\cdot)$ heißt Kaplan-Meier-Schätzer.

Bemerkung 4.1. Gilt mit den Bezeichnungen 4.1 $c_k = n_k - d_k > 0$, d.h. werden zum letzten Beobachtungszeitpunkt Individuen zensiert, so ist der KM-Schätzer (4.6) für $t > z_k$ nicht definiert. Dann ist nämlich $1 - d_k/n_k > 0$ und somit auch $\hat{S}(z_k) > 0$. Da jedoch der Verlauf von F rechts von z_k keine Auswirkungen auf die Likelihood-Funktion hat, können die Werte der Survival-Funktion S für $t > z_k$ nicht geschätzt werden.

Korollar 4.1. *Für unzensierte Datensätze stimmt der KM-Schätzer (4.6) mit der empirischen Überlebensfunktion (4.1) überein.*

Beweis: Enthält der gegebene Datensatz $\big((y_1, \delta_1), \ldots, (y_n, \delta_n)\big)$ nur unzensierte Beobachtungen, d.h. gilt $(y_j, \delta_j) = (t_j, 1)$ für alle $j = 1, \ldots, n$, so gilt mit den Bezeichnungen 4.1: $n_{j+1} = n_j - d_j - c_j = n_j - d_j$ und damit

$$\begin{aligned} \hat{S}(t) \quad &= \quad \prod_{j:\, z_j \le t} \left(1 - \frac{d_j}{n_j}\right) \;=\; \prod_{j:\, z_j \le t} \frac{n_{j+1}}{n_j} \;\overset{n_1 = n}{=}\; \frac{n_{j+1}}{n} \\ &\overset{y_j = t_j}{=} \quad \frac{1}{n} \sum_{j=1}^{n} \mathbb{1}_{(t,\infty)}(t_j) \;=\; 1 - \frac{1}{n} \sum_{j=1}^{n} \mathbb{1}_{(-\infty,\, t]}(t_j). \end{aligned}$$

□

Bemerkung 4.2. Der Kaplan-Meier-Schätzer $\hat{S}(\cdot)$ hat einige wünschenswerte Eigenschaften, darunter Konsistenz und asymptotische Normal-Verteilung. Mit Hilfe von Zähl-Prozessen und Martingalen kann dies allgemein für alle Typen von Lebensverteilungen gezeigt werden. Eine sehr ausführliche Darstellung der Theorie geben Fleming und Harrington [Fle-91] und Andersen et al. [And-93]. Die Konsistenz von $\hat{S}(\cdot)$ wird in [Fle-91, Theorem 3.4.2] bewiesen. Asymptotische Eigenschaften dieses Schätzers werden dort in Kapitel 6 behandelt. Eine zusammenfassende Abhandlung der wichtigsten Ergebnisse ist auch in [Law-03, Abschnitt 3.2.4] zu finden.

4.2 Die Varianz des Kaplan-Meier-Schätzers

Eine Schätzung für die Varianz von $\hat{S}(\cdot)$ erhält man mittels der Large-Sample-Theorie für Maximum-Likelihood-Verfahren. Sie wird im Folgenden gemäß der Darstellung in [Cox-84, Abschnitt 4.3] hergeleitet. Nimmt man an, dass die möglichen Ausfallzeiten z_1, z_2, ... , z_k fest sind und dass der Ausfallmechanismus dazu führt, dass die Anzahl der Ausfälle in jedem z_j mit derselben Rate steigt wie der Stich-

probenumfang n, so können die in Abschnitt 3.3 vorgestellten Resultate verwendet werden.

Ist $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_k) = (\lambda(z_1), \ldots, \lambda(z_k))$ der interessierende Parameter, so konvergiert $\sqrt{n}\,(\hat{\boldsymbol{\lambda}}_n - \boldsymbol{\lambda})$ in Verteilung gegen eine Zufallsvariable, die multivariat normalverteilt ist mit Erwartungswert $\mathbf{0} \in \mathbb{R}^k$ und Kovarianz-Matrix $\mathcal{I}^{-1}(\boldsymbol{\lambda})$:

$$\sqrt{n}\,(\hat{\boldsymbol{\lambda}}_n - \boldsymbol{\lambda}) \xrightarrow{\mathcal{D}} \mathcal{N}_k\big(\mathbf{0},\, \mathcal{I}^{-1}(\boldsymbol{\lambda})\big). \tag{4.7}$$

Der Index n in $\hat{\boldsymbol{\lambda}}_n$ bringt dabei zum Ausdruck, dass die Schätzung für $\boldsymbol{\lambda}$ auf einer n-elementigen Stichprobe basiert. $\mathcal{I}(\boldsymbol{\lambda})$ bezeichnet die Fisher-Matrix und kann durch die beobachtete Informationsmatrix

$$I(\hat{\boldsymbol{\lambda}}) \;=\; \left(-\frac{\partial^2 l(\boldsymbol{\lambda})}{\partial\lambda_i \partial\lambda_j}\bigg|_{\boldsymbol{\lambda}=\hat{\boldsymbol{\lambda}}} \right)_{i,\, j=1,\ldots,p}$$

geschätzt werden. Es ist

$$\begin{aligned}
\frac{\partial^2 l(\boldsymbol{\lambda})}{\partial\lambda_i \partial\lambda_i}\bigg|_{\boldsymbol{\lambda}=\hat{\boldsymbol{\lambda}}} &= \frac{\partial^2}{\partial\lambda_i \partial\lambda_i} \left(\sum_{j=1}^{k} \Big[d_j \,\log\lambda_j + (n_j - d_j)\,\log(1-\lambda_j) \Big] \right)\bigg|_{\boldsymbol{\lambda}=\hat{\boldsymbol{\lambda}}} \\
&= \frac{\partial}{\partial\lambda_i} \left(\frac{d_i}{\lambda_i} - \frac{n_i - d_i}{1-\lambda_i} \right)\bigg|_{\boldsymbol{\lambda}=\hat{\boldsymbol{\lambda}}} \\
&= \frac{\partial}{\partial\lambda_i} \left(\frac{d_i - \lambda_i n_i}{\lambda_i(1-\lambda_i)} \right)\bigg|_{\boldsymbol{\lambda}=\hat{\boldsymbol{\lambda}}} \\
&= \frac{-n_i\lambda_i(1-\lambda_i) - (d_i - \lambda_i n_i)(1-2\lambda_i)}{[\lambda_i(1-\lambda_i)]^2}\bigg|_{\boldsymbol{\lambda}=\hat{\boldsymbol{\lambda}}} \\
&= -\frac{n_i}{\lambda_i(1-\lambda_i)}, \\
&\quad \text{da} \quad (d_i - \lambda_i n_i) = 0 \quad \text{für} \quad \lambda_i = \hat{\lambda}_i = \frac{d_i}{n_i}\,.
\end{aligned}$$

Weil $\partial^2 l / \partial\lambda_i \partial\lambda_j = 0$ für $i \neq j$, sind die Einträge der geschätzten Varianz-Kovarianz-Matrix gegeben durch

$$-\frac{\partial^2 l(\boldsymbol{\lambda})}{\partial\lambda_i \partial\lambda_j}\bigg|_{\boldsymbol{\lambda}=\hat{\boldsymbol{\lambda}}} \;=\; \begin{cases} \frac{\hat{\lambda}_i(1-\hat{\lambda}_i)}{n_i}, & i = j, \\ 0, & i \neq j. \end{cases} \tag{4.8}$$

Es folgt $\mathrm{Cov}\,(\hat{\lambda}_i,\, \hat{\lambda}_j) = 0$ asymptotisch für $i \neq j$, was gleichbedeutend damit ist, dass $\hat{\lambda}_i$ und $\hat{\lambda}_j$ asymptotisch unabhängig sind für alle $i \neq j$.

Satz 4.3 (Varianz des KM-Schätzers). *Die Varianz des Kaplan-Meier-Schätzers kann wie folgt approximiert werden:*

$$Var\big(\hat{S}(t)\big) \;\approx\; \big[\hat{S}(t)\big]^2 \sum_{j:\, z_j \le t} \frac{d_j}{n_j\,(n_j - d_j)}. \tag{4.9}$$

Beweis: (i) Es ist $\log \hat{S}(t) = \log\Big(\prod_{j:\, z_j \le t}(1-\hat{\lambda}_j)\Big) = \sum_{j:\, z_j \le t} \log(1-\hat{\lambda}_j)$.
(ii) Mit der Delta-Methode aus Satz 3.8 und $f(x) = \log(1-x)$ erhält man

$$\mathrm{Var}\big(\log(1-\hat{\lambda}_j)\big) \;\approx\; \left(\frac{1}{1-\lambda_j}\right)^2 \mathrm{Var}\big(\hat{\lambda}_j\big).$$

(iii) Nach (4.8) hat man außerdem

$$\mathrm{Var}\big(\hat{\lambda}_j\big) \;=\; \mathrm{Cov}\big(\hat{\lambda}_j,\, \hat{\lambda}_j\big) \;\approx\; \frac{\hat{\lambda}_j(1-\hat{\lambda}_j)}{n_j}.$$

Insgesamt folgt mit (i) - (iii):

$$\begin{aligned}
\mathrm{Var}\big(\log \hat{S}(t)\big) &\overset{(i)}{=} \mathrm{Var}\left(\sum_{j:\, z_j \le t} \log\left(1-\hat{\lambda}_j\right)\right) \\
&\overset{\substack{\hat{\lambda}_j \text{ asympt.} \\ \text{unabh.}}}{\approx} \sum_{j:\, z_j \le t} \mathrm{Var}\left(\log\left(1-\hat{\lambda}_j\right)\right) \\
&\overset{(ii)}{\approx} \sum_{j:\, z_j \le t} \left(\frac{1}{1-\lambda_j}\right)^2 \mathrm{Var}\big(\hat{\lambda}_j\big) \\
&\overset{(iii)}{\approx} \sum_{j:\, z_j \le t} \left(\frac{1}{1-\lambda_j}\right)^2 \frac{\hat{\lambda}_j(1-\hat{\lambda}_j)}{n_j} \\
&\overset{\lambda_j \approx \hat{\lambda}_j}{\approx} \sum_{j:\, z_j \le t} \left(\frac{1}{1-\hat{\lambda}_j}\right)^2 \frac{\hat{\lambda}_j(1-\hat{\lambda}_j)}{n_j} \\
&= \sum_{j:\, z_j \le t} \frac{\hat{\lambda}_j}{(1-\hat{\lambda}_j)\,n_j} \\
&\overset{\hat{\lambda}_j = \frac{d_j}{n_j}}{=} \sum_{t_j \le t} \frac{d_j}{n_j\,(n_j - d_j)}.
\end{aligned} \tag{4.10}$$

Ein entsprechendes Anwenden der Delta-Methode auf $\mathrm{Var}(\log \hat{S}(t))$ liefert mit

$f(x) = \log(x)$: $\mathrm{Var}(\log \hat{S}(t)) = (\hat{S}(t))^{-2} \mathrm{Var}(\hat{S}(t))$. Man erhält schließlich

$$\mathrm{Var}\big(\hat{S}(t)\big) \approx \big[\hat{S}(t)\big]^2 \sum_{t_j \le t} \frac{d_j}{n_j\,(n_j - d_j)}.$$

□

Bemerkung 4.3. Die Approximation (4.9) für die Varianz des Kaplan-Meier-Schätzers ist als Greenwood-Formel bekannt. Greenwood [Gre-26] hat sie im Rahmen seiner Untersuchungen zur natürlichen Dauer von Krebs aufgestellt.

4.3 Lokale Konfidenzintervalle für $S(t)$

Konfidenzintervalle für die Survival-Funktion zu einem festen Zeitpunkt t können auf verschiedene Weise konstruiert werden. Zwei verwandte, auf der Normal-Approximation basierende Verfahren werden in diesem Abschnitt vorgestellt. Ist $\hat{S}(t)$ der ML-Schätzer, so ist er unter schwachen Regularitätsbedingungen asymptotisch normalverteilt:

$$\frac{\hat{S}(t) - S(t)}{\sqrt{\mathrm{Var}\big(\hat{S}(t)\big)}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,\,1). \tag{4.11}$$

$\mathrm{Var}\big(\hat{S}(t)\big)$ kann mit Hilfe der Formel (4.9) berechnet werden. Lokale Konfidenzintervalle werden dann konstruiert über den Annahmebereich des zweiseitigen Tests zum Niveau α für:

$$H_0: \quad S(t_0) = S_0 \qquad \text{vs.} \qquad H_1: \quad S(t_0) \neq S_0,$$

wobei t_0 ein beliebiger aber fest gewählter Zeitpunkt ist. Sei $(\mathbf{y}, \boldsymbol{\delta}) = (\,(y_1,\,\delta_1), \ldots, (y_n,\,\delta_n)\,)$ der Beobachtungsvektor. Dann ist der Test durch die folgende Entscheidungsfunktion gegeben:

$$\varphi(\mathbf{y}, \boldsymbol{\delta}) = \mathbb{1}_{\left\{\frac{|\hat{S}(t_0) - S_0|}{\sqrt{Var(\hat{S}(t_0))}} > \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right\}}(\mathbf{y}, \boldsymbol{\delta}).$$

Der Annahmebereich ist

$$\begin{array}{rrcccl} & \frac{\left|\hat{S}(t_0)-S_0\right|}{\sqrt{Var(\hat{S}(t_0))}} & \leq & \Phi^{-1}\left(1-\frac{\alpha}{2}\right) & & \\ \iff & -\Phi^{-1}\left(1-\frac{\alpha}{2}\right) & \leq & \frac{\hat{S}(t_0)-S_0}{\sqrt{Var(\hat{S}(t_0))}} & \leq & \Phi^{-1}\left(1-\frac{\alpha}{2}\right) \\ \iff & \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\sqrt{\mathrm{Var}(\hat{S}(t_0))} & \geq & \hat{S}(t_0)-S_0 & \geq & -\Phi^{-1}\left(1-\frac{\alpha}{2}\right)\sqrt{\mathrm{Var}(\hat{S}(t_0))}. \end{array}$$

Als Konfidenzintervall zum Niveau $\beta = 1-\alpha$ erhält man mit $\hat{\sigma}_S(t_0) := \left[\mathrm{Var}(\hat{S}(t_0))\right]^{1/2}$ also

$$\left[\hat{S}(t_0) - \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\hat{\sigma}_S(t_0)\,,\;\; \hat{S}(t_0) + \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\hat{\sigma}_S(t_0)\right]. \tag{4.12}$$

Diese Bereichsschätzung darf allerdings nicht als Konfidenzstreifen für den gesamten Verlauf der Survival-Funktion verstanden werden. Sie stellt lediglich ein Konfidenzintervall an einer fest gewählten Stelle t_0 dar.

Ist innerhalb der Stichprobe die Anzahl der unzensierten Lebenszeiten klein, so ist die Normal-Approximation von $(\hat{S}(t) - S(t))\,(\mathrm{Var}(\hat{S}(t)))^{-1/2}$ in der Regel nicht besonders gut und das Konfidenzintervall (4.12) kann Werte außerhalb von (0, 1) enthalten [Kal-80, S. 14]. Bessere Konfidenzintervalle ergeben sich mittels der log-log-Transformation. Statt der Survival-Funktion betrachtet man

$$\psi:\ (0,\infty) \to \mathbb{R}, \qquad \psi(t) \;=\; \ln\big(-\ln S(t)\big)$$

mit dem ML-Schätzer $\hat{\psi}(t) = \ln\big(-\ln \hat{S}(t)\big)$. Die Delta-Methode liefert eine Approximation für die Varianz von $\hat{\psi}(t)$:

$$\begin{array}{rcl} \mathrm{Var}\big(\hat{\psi}(t)\big) & \approx & \left(\dfrac{1}{\ln \hat{S}(t)}\right)^2 \mathrm{Var}\big(\ln \hat{S}(t)\big) \\ & \overset{(4.10)}{\approx} & \left(\dfrac{1}{\ln \hat{S}(t)}\right)^2 \displaystyle\sum_{t_j \leq t} \frac{d_j}{n_j\,(n_j - d_j)}. \end{array} \tag{4.13}$$

Der Ausdruck

$$\frac{\hat{\psi}(t) - \psi(t)}{\sqrt{\mathrm{Var}(\hat{\psi}(t))}} \tag{4.14}$$

wird durch die Standard-Normal-Verteilung besser approximiert als die entsprechende Größe in (4.11) und liefert demnach auch bessere Bereichsschätzungen.

Ein lokales Konfidenzintervall für $\psi(t_0)$ erhält man entsprechend der Konstruktion von (4.12). Durch Rücktransformation ergibt sich dann ein Intervall für $S(t_0)$: Ist $\hat{\sigma}_\psi(t_0) := \left[\mathrm{Var}\big(\hat{\psi}(t_0)\big)\right]^{1/2}$ und

$$\left[\hat{\psi}(t_0) - \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\hat{\sigma}_\psi(t_0)\,,\;\; \hat{\psi}(t_0) + \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\hat{\sigma}_\psi(t_0)\right] \tag{4.15}$$

das Konfidenzintervall zum Niveau $\beta = 1-\alpha$ für $\psi(t_0)$, so ist das für $S(t_0)$ zum gleichen Niveau gegeben durch

$$\left[\exp\left\{-\mathrm{e}^{\hat{\psi}(t_0)+\Phi^{-1}\left(1-\frac{\alpha}{2}\right)\hat{\sigma}_\psi(t_0)}\right\},\;\; \exp\left\{-\mathrm{e}^{\hat{\psi}(t_0)-\Phi^{-1}\left(1-\frac{\alpha}{2}\right)\hat{\sigma}_\psi(t_0)}\right\}\right]. \tag{4.16}$$

Bemerkung 4.4. Borgan & Leistøl [Bor-90] haben gezeigt, dass das Konfidenzintervall (4.16) selbst für 25-elementige Stichproben mit bis zu 50% zensierten Lebenszeiten noch gute Ergebnisse liefert.

Im folgenden Beispiel wird in Anlehnung an die Beipiele 1.1, 2.3 und 2.5 in [Col-03] der Kaplan-Meier-Schätzer für einen konkreten Datensatz berechnet. Für die beobachteten Ausfallzeiten werden außerdem die lokalen Konfidenzintervalle nach (4.12) und (4.16) angegeben.

Beispiel 4.1 (Zeit bis zum Abbruch einer IUP-Anwendung). Die Weltgesundheitsorganisation (World Health Organisation, WHO) hat in [WHO-87] die Daten klinischer Studien zu verschiedenen Methoden der Kontrazeption veröffentlicht. Ein Teil dieses Datensatzes enthält die Zeitspannen zwischen dem Anwendungsbeginn bestimmter Verhütungsmethoden und ihren Abbrüchen. Die Daten in Anhang A, Tabelle A.1 beziehen sich auf die Anzahl der Wochen vom Beginn einer Intrauterinpessar (IUP)-Anwendung bis zur Einstellung dieser Kontrazeptionsform aufgrund von Blutungsproblemen. Die Survival-Funktion beschreibt für diesen Datensatz die Wahrscheinlichkeit dafür, dass eine Frau die IUP-Anwendung nach einem Zeitpunkt t unterbricht. Den Kaplan-Meier-Schätzer für die Survival-Funktion erhält man nach Satz 4.2 mit den in Tabelle 4.1 dargelegten Berechnungen. Tabelle 4.2 enthält die Werte des geschätzten Standard-Fehlers, sowie die nach (4.12) und (4.16) berechneten lokalen Konfidenzintervalle für S. Abbildung 4.1 stellt den Graphen des KM-Schätzers sowie die Grenzen der gemäß (4.16) berechneten lokalen Konfidenzintervalle dar. Zu beachten ist, dass der letzte Abbruchzeitpunkt zensiert ist und der KM-Schätzer somit für $t > 107$ nicht definiert ist, vergleiche hierzu Bemerkung 4.1. Der zur Berechnung des KM-Schätzers, seiner Standard-Fehler und

der Konfidenzintervalle benötigte „R“-Quellcode ist in Anhang B, Abschnitt B.1 zu finden.

Tabelle 4.1: Berechnungen für den KM-Schätzer zum Datensatz A.1.

Ausfallzeitpunkt t_j	n_j	d_j	$1 - n_j/d_j$	$\hat{S}(t_j)$
10	18	1	0.944	0.944
19	15	1	0.933	0.881
30	13	1	0.923	0.814
36	12	1	0.917	0.746
59	8	1	0.875	0.653
75	7	1	0.857	0.559
93	6	1	0.833	0.466
97	5	1	0.800	0.373
107	3	1	0.667	0.249

Tabelle 4.2: Standard-Fehler (standard error, se) des KM-Schätzers zum Datensatz A.1 und lokale Konfidenzintervalle (KI) zum Niveau 0.95.

Zeitintervall	$\hat{S}(t_j)$	se$[\hat{S}(t_j)]$	0.95KI nach (4.12)	0.95KI nach (4.16)
[0, 10)	1.000	0.0000	–	–
[10, 19)	0.944	0.0540	[0.8386, 1.000]	[0.844, 1.000]
[19, 30)	0.881	0.0790	[0.7267, 1.000]	[0.739, 1.000]
[30, 36)	0.814	0.0978	[0.6220, 1.000]	[0.643, 1.000]
[36, 59)	0.746	0.1107	[0.5290, 0.963]	[0.558, 0.998]
[59, 75)	0.653	0.1303	[0.3972, 0.908]	[0.441, 0.965]
[75, 93)	0.559	0.1412	[0.2827, 0.836]	[0.341, 0.917]
[93, 97)	0.466	0.1452	[0.1816, 0.751]	[0.253, 0.858]
[97, 107)	0.373	0.1430	[0.0927, 0.653]	[0.176, 0.791]
107	0.249	0.1392	[0.0000, 0.522]	[0.083, 0.745]

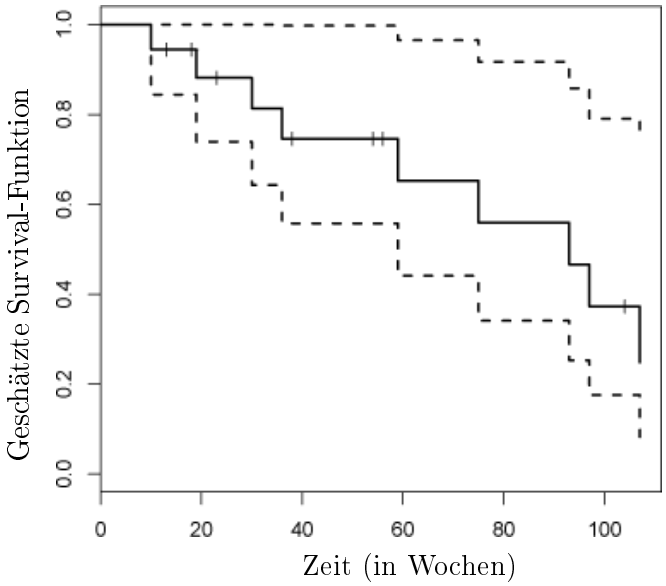

Abbildung 4.1: Geschätzte Survival-Funktion und Grenzen der lokalen 0.95-Konfidenzintervalle zum Datensatz A.1. Die Zensurzeitpunkte sind mit einem senkrechten Strich kenntlich gemacht.

Kapitel 5

Parametrische Modelle

In diesem Kapitel werden Wahrscheinlichkeitsverteilungen vorgestellt, die für die Analyse von Lebensdaten besonders nützlich sind. Obwohl jede nicht-negative Zufallsvariable einen möglichen Kandidaten darstellt, nehmen innerhalb der univariaten Modelle einige Verteilungen eine zentrale Position ein. Dazu gehören die Exponential-Verteilung, die Weibull-Verteilung, die Log-Normal-Verteilung und die Log-Logistik-Verteilung. Informationen über den Ausfall-Mechanismus suggerieren – insbesondere über die Hazard-Rate (vgl. Abschnitt 2.2.3) – in vielen Situationen ein bestimmtes Modell. Ob eine bestimmte Verteilung sich zur Beschreibung eines gegebenen Datensatzes eignet, kann mit Hilfe des KM-Schätzers durch graphische Verfahren überprüft werden. Diese werden in Abschnitt 5.2 vorgestellt.

5.1 Spezielle Verteilungen

5.1.1 Die Exponential-Verteilung

Die Exponential-Verteilung zählt zu den ersten umfassend diskutierten Lebenszeit-Modellen. Große Verwendung fand sie insbesondere in Studien zur Funktionsdauer industriell hergestellter Güter, siehe z.B. Davis [Dav-52], Epstein und Sobel [Eps-54], sowie Epstein [Eps-58]. Aufgrund ihrer Gedächtnislosigkeit und der daraus resultierenden konstanten Hazard-Rate ist die Exponential-Verteilung in der modernen Lebenszeitanalyse allerdings nur sehr begrenzt anwendbar. Die Dichte-, Survival- und Hazard-Funktionen zu $T \sim \text{Exp}(\lambda)$ werden nachstehend angegeben. Die einzelnen Zuordnungsvorschriften resultieren aus den in Kapitel 2 eingeführten Definitionen

und ihren Zusammenhängen. Für $t \geq 0$ ist

$$f(t) = \lambda \exp(-\lambda t), \tag{5.1}$$
$$S(t) = \exp(-\lambda t), \tag{5.2}$$
$$\lambda(t) = \lambda. \tag{5.3}$$

5.1.2 Die Weibull-Verteilung

Die Weibull-Verteiung liefert ein Lebenszeit-Modell, das in vielen verschiedenen Bereichen eingesetzt wird.[1] Obwohl erstmals von Rosen und Rammler [Ros-33] verwendet, ist die Wahrscheinlichkeitsverteilung nach Weibull [Wei-51] benannt. Ist T Weibull-verteilt mit Skalenparameter α und Formparameter β, d.h. $T \sim \text{Weib}(\lambda, \beta)$, so sind Dichte-, Survival- und Hazard-Funktion für $t > 0$ wie folgt gegeben

$$f(t) = \frac{\beta}{\alpha}\Big(\frac{t}{\alpha}\Big)^{\beta-1} \exp\Big[-\Big(\frac{t}{\alpha}\Big)^{\beta}\Big], \tag{5.4}$$
$$S(t) = \exp\Big[-\Big(\frac{t}{\alpha}\Big)^{\beta}\Big], \tag{5.5}$$
$$\lambda(t) = \frac{\beta}{\alpha}\Big(\frac{t}{\alpha}\Big)^{\beta-1}. \tag{5.6}$$

Über den Formparameter $\beta > 0$ liefert die Weibull-Verteilung je nach Bedarf wachsende ($\beta > 1$), fallende ($\beta < 1$) oder konstante ($\beta = 1$) Hazard-Funktionen. Wegen dieser Flexibilität und der relativ einfachen Form von Survival-, Hazard- und Dichtefunktion ist die Weibull-Verteilung ein weit verbreitetes parametrisches Modell. Mit $\beta = 1$ und $\alpha = \lambda^{-1}$ erhält man als Spezialfall die Exponential-Verteilung.

Satz 5.1 (**Transformationssatz für Dichten**). *Sei X eine Zufallsvariable mit stetiger Dichte f_X und $h : \mathbb{R} \to \mathbb{R}$ eine stetig differenzierbare, streng monotone Funktion mit $h'(x) \neq 0$ für alle $x \in \mathbb{R}$. Ist $Y = h(X)$, so hat auch Y eine Dichte f_Y, für die mit $g(y) = h^{-1}(y)$ gilt*

$$f_Y(y) = f_X(g(y))\,|g'(y)|\,, \quad y \in \mathbb{R}. \tag{5.7}$$

[1] Industrielle Anwendbarkeit wird u.A. von Kao [Kao-59] und Berretoni [Ber-64] diskutiert. Doll [Dol-71] betrachtet die Weibull-Verteilung im Zusammenhang mit Zeitdauern bis zur Tumorbildung in menschlichen und tierischen Organismen.

Beweis: Ist h streng monoton wachsend, so hat man

$$F_Y(y) = P(Y \le y) = P(h(X) \le y) = P(X \le g(y)) = \int_{-\infty}^{g(y)} f_X(x)\,dx.$$

Damit ist F_Y differenzierbar und es gilt $f_Y(y) := F_Y'(y) = f_X(g(y))\,g'(y)$. Ist h streng monoton fallend, so folgt entsprechend

$$F_Y(y) = P(Y \le y) = P(h(X) \le y) = P(X \ge g(y)) = \int_{g(y)}^{\infty} f_X(x)\,dx$$

und $f_Y(y) = -\,f_X(g(y))\,g'(y) = f_X(g(y))\,|g'(y)|$. □

Die Extremwert-Verteilung

Die Extremwert-Verteilung, $\mathrm{EV}(u,\,b)$ (EV von extreme value), ist auch als Gumbel-Verteilung bekannt und hat für $y \in (-\infty,\,\infty)$ die folgende Dichte- und Survival-Funktion

$$f(y) \;=\; \frac{1}{b}\exp\Big[\frac{y-u}{b} - \exp\Big(\frac{y-u}{b}\Big)\Big], \tag{5.8}$$

$$S(y) \;=\; \exp\Big[-\exp\Big(\frac{y-u}{b}\Big)\Big]. \tag{5.9}$$

Die Extremwert-Verteilung ist mit der Weibull-Verteilung eng verwandt. Ist nämlich $T \sim \mathrm{Weib}(\alpha,\beta)$, so ist mit Hilfe des Transformationssatzes 5.1 leicht nachzurechnen, dass $Y = \ln T$ eine Extremwert-Verteilung mit den Parametern $b = \beta^{-1}$ und $u = \ln\alpha$ besitzt. Nach diesem ist nämlich

$$\begin{aligned} f_Y(y) \quad &= \quad f_T(\exp(y))\,|\exp(y)| \;=\; \frac{\beta}{\alpha}\Big(\frac{\exp(y)}{\alpha}\Big)^{\beta-1}\exp\Big[-\Big(\frac{\exp(y)}{\alpha}\Big)^{\beta}\Big]\exp(y) \\ &\overset{\substack{\alpha=\exp u\\ \beta=1/b}}{=} \quad \frac{1}{b}\exp(-u)\exp\Big(\frac{y-u}{b}\Big)\exp(u-y)\exp\Big[-\exp\Big(\frac{y-u}{b}\Big)\Big]\exp(y) \\ &= \quad \frac{1}{b}\exp\Big[\frac{y-u}{b} - \exp\Big(\frac{y-u}{b}\Big)\Big]. \end{aligned}$$

5.1.3 Die Log-Normal-Verteilung

Eine Zufallsvariable T heißt log-normal-verteilt, falls $Y = \ln T$ der Normal-Verteilung folgt und damit durch die Dichte

$$f(y) \;=\; \frac{1}{\sqrt{2\pi\sigma^2}}\exp\Big[-\frac{1}{2}\Big(\frac{y-\mu}{\sigma}\Big)^2\Big] \tag{5.10}$$

gegeben ist. Auch in diesem Fall ist schnell nachzurechnen, dass die Dichte- und Survival-Funktion zu $T = \exp Y \sim \mathit{Log}\mathcal{N}(\mu, \sigma^2)$ für $t > 0$ wie folgt gegeben sind

$$f(t) = \frac{1}{(2\pi)^{1/2}\sigma t} \exp\Big\{ -\frac{1}{2}\Big(\frac{\log t - \mu}{\sigma}\Big)^2\Big\}, \tag{5.11}$$

$$S(t) = 1 - \Phi\Big(\frac{\log t - \mu}{\sigma}\Big). \tag{5.12}$$

Die Hazard-Funktion lässt sich gemäß Kapitel 2 berechnen als $\lambda(t) = f(t)/S(t)$. Es lässt sich zeigen, dass die Hazard-Funktion in 0 den Wert 0 annimmt, dann zu einem Maximum anwächst und für $t \to \infty$ gegen 0 konvergiert [Law-03, S. 22]. Durch diese hügelförmige Hazard-Rate eignet sich die Log-Normal-Verteilung besonders gut zur Modellierung des Überlebens von Patienten nach bestimmten medizinischen Behandlungen oder operativen Eingriffen: Ist die Therapie erfolgreich gewesen und hat sich der Patient erholt, so sinkt sein Sterberisiko.

5.1.4 Die Log-Logistik-Verteilung

Die Wahrscheinlichkeitsdichte der Log-Logistik-Verteilung ist für $t > 0$ definiert als

$$f(t) = \frac{\big(\frac{\beta}{\alpha}\big)\big(\frac{t}{\alpha}\big)^{\beta-1}}{\big[1 + \big(\frac{t}{\alpha}\big)^{\beta}\big]^2}. \tag{5.13}$$

Dabei sind $\alpha > 0$ und $\beta > 0$ die Parameter der Verteilung. Die Survival- und Hazard-Funktion einer Lebenszeit $T \sim \mathit{LogLogist}(\alpha, \beta)$ lassen sich nach Kapitel 2 daraus berechnen als

$$S(t) = \Big[1 + \Big(\frac{t}{\alpha}\Big)\Big]^{-1}, \tag{5.14}$$

$$\lambda(t) = \frac{\big(\frac{\beta}{\alpha}\big)\big(\frac{t}{\alpha}\big)^{\beta-1}}{1 + \big(\frac{t}{\alpha}\big)^{\beta}}. \tag{5.15}$$

Die logarithmierte Zufallsvariable $Y = \ln T$ hat dann eine Logistik-Verteilung $\mathit{Logist}(u, b)$ mit Dichtefunktion

$$f(y) = \frac{\frac{1}{b}\exp\big(\frac{y-u}{b}\big)}{\big[1 + \exp\big(\frac{y-u}{b}\big)\big]^2}, \quad y \in (-\infty, \infty), \tag{5.16}$$

und den Parametern $u = \log\alpha$ und $b = \beta^{-1}$ (zu berechnen nach Satz 5.1).

5.1.5 Log-Lokation-Skalen-Modelle

Die in den Abschnitten 5.1.2, 5.1.3 und 5.1.4 vorgestellten Modelle gehören sämtlich zu den Log-Lokation-Skalen-Verteilungen. Diese werden in der Lebenszeitanalyse am häufigsten verwendet und sollen daher nun in einem allgemeinen Rahmen behandelt werden. Die Ausführungen folgen dabei im Wesentlichen denen von Lawless in [Law-03, S. 27ff]. Ein parametrisches Lokation-Skalen-Modell für eine Variable Y ist durch folgende Dichtefunktion gegeben:

$$f_Y(y) = \frac{1}{b} f_Z\left(\frac{y-u}{b}\right), \qquad -\infty < y < \infty. \tag{5.17}$$

Dabei ist u , $-\infty < u < \infty$, der Lokations-Parameter, $b > 0$ der Skalen-Parameter und f_Z die Dichte der standardisierten Zufallsvariable $Z = (Y - u)/b$. Man nennt (5.17) mit $u = 0$ und $b = 1$ den Standard-Typen der Lokation-Skalen-Familie. Die Zufallsvariable Y lässt sich darstellen als

$$Y = bZ + u. \tag{5.18}$$

Bezeichnet F_Z die Verteilungs- und S_Z die Survival-Funktion zu Z, so gilt für die entsprechenden Funktionen zu Y (gemäß Satz 5.1)

$$F_Y(y) = F_Z\Big(\frac{y-u}{b}\Big), \qquad -\infty < y < \infty, \tag{5.19}$$

$$S_Y(y) = S_Z\Big(\frac{y-u}{b}\Big), \qquad -\infty < y < \infty. \tag{5.20}$$

Ist T eine Lebenszeit und gilt $\ln T = Y$ mit f_Y wie in (5.17) gegeben, so sagt man, dass T eine Log-Lokation-Skalen-Verteilung besitzt. Die Dichte- und Survival-Funktion von $T = \exp Y$ lassen sich dann nach Satz 5.1 mit $t > 0$ darstellen als

$$f_T(t) = \frac{1}{b\,t} f_Z\left(\frac{\ln t - u}{b}\right), \tag{5.21}$$

$$S_T(t) = S_Z\left(\frac{\ln t - u}{b}\right). \tag{5.22}$$

Weibull-, Log-Normal- und Log-Logistik-Verteilungen sind von dieser Form. Die zugehörigen Lokation-Skalen-Verteilungen für Y sind durch Extremwert-, Normal-

und Logistik-Verteilungen gegeben. Die Survival-Funktionen der entsprechenden Standard-Typen sind

$$\begin{aligned} S_Z(z) &= \exp[-\exp(z)] \quad \text{(Extremwert-Verteilung)}, \\ S_Z(z) &= 1 - \Phi(z) \quad \text{(Normal-Verteilung)}, \\ S_Z(z) &= [1 + \exp(z)]^{-1} \quad \text{(Logistik-Verteilung)}, \end{aligned}$$

wobei Φ die Verteilungsfunktion der Standard-Normal-Verteilung bezeichnet.

5.2 Graphische Beurteilung eines parametrischen Modells

Plots von KM-Schätzern liefern nicht nur graphische Darstellungen univariater Datensätze, sondern lassen auch beurteilen, wie gut sich ein parametrisches Modell zur Beschreibung bestimmter Lebenszeitdaten eignet. Eine graphische Beurteilung ist in der Regel zwar sehr subjektiv, in vielen Situationen aber nützlich. Es folgt daher ein kurzer Überblick über ein Verfahren, das sich besonders für die in der Lebenszeitanalyse wichtigen Log-Lokation-Skalen-Verteilungen eignet.

Sei $S(t;\boldsymbol{\theta})$ die Survival-Funktion in einem parametrischen Modell. Existieren Funktionen g und h derart, dass die Transformation $g[S(t;\boldsymbol{\theta})]$ linear in $h(t)$ ist, so ist – falls ein parametrisches Modell geeignet ist – der Plot von $g[\hat{S}(t;\boldsymbol{\theta})]$ gegen $h(t)$ nahezu linear.

Unterstellt man einem Datensatz beispielsweise die Exponential-Verteilung, so erfüllt die zugehörige Survival-Funktion die folgende Gleichung

$$\ln S(t) \quad = \quad -\lambda t. \tag{5.23}$$

Sind $t_1 < \ldots < t_k$ die verschiedenen Beobachtungszeiten, zu denen sich Ausfälle ereignen und $\hat{S}(\cdot)$ der KM-Schätzer für die Survival-Funktion, so ist das Exponential-Modell nach obiger Überlegung geeignet, falls die Punkte $\big(t_j,\ \ln \hat{S}(t_j)\big)$ nahe einer Gerade liegen, die durch den Ursprung geht und negative Steigung hat. Eine graphische Schätzung für λ ergibt sich dabei aus der Steigung einer Regressionsgerade. Die Exponential-Verteilung kann mit $\beta = 1$ und $\alpha = \lambda^{-1}$ auch als Spezialfall der Weibull-Verteilung behandelt werden. Die Survival-Funktion einer Weibull-verteilten Zufalls-

variable genügt

$$\ln(-\ln S(t)) \;=\; \beta \ln t - \beta \ln \alpha. \tag{5.24}$$

Eine Weibull-Verteilung erscheint also angemessen, wenn die Punkte $\big(\ln(t_j),\ \ln\big(-\ln \hat{S}(t_j)\big)\big)$ nahezu auf einer Gerade liegen. Ist dies der Fall, so liefert die Regressionsgerade Schätzer für α und β: Die Steigung der Geraden ist ein Schätzer für β und ihr Schnittpunkt mit der $\ln t$-Achse der für $\ln \alpha$.

Dieses Linearisierungsverfahren ist im Allgemeinen für alle Lebenszeit-Verteilungen T geeignet, zu denen eine Transformation $Y = g(T)$ existiert, die eine Lokation-Skalen-Verteilung besitzt. Das gilt nach Abschnitt 5.1.5 mit $g(t) = \ln t$ insbesondere für Log-Lokation-Skalen-Verteilungen. Unter der Annahme, dass $Y = g(T) = \ln T$ ist, gilt nach (5.20) und (5.22)

$$S_Y(y) \;=\; S_Z\Big(\frac{y-u}{b}\Big) \;=\; S_T(t) \tag{5.25}$$

mit $t = \exp(y)$, $b > 0$ und $u \in (-\infty, \infty)$. Damit ist

$$S_Z^{-1}\big[S_T(t)\big] \;=\; \frac{1}{b}\, y - \frac{u}{b} \tag{5.26}$$

eine lineare Funktion in $y = g(t) = \ln t$. Sind $t_1, \ldots, t_n$ die beobachteten Ausfall- und Zensurzeiten und ist $\hat{S}_T(\cdot)$ der auf ihnen basierende KM-Schätzer, so werden die Daten durch das betrachtete Modell gut beschrieben, wenn der Plot von $S_Z^{-1}[\hat{S}_T(t)]$ gegen $g(t)$ annähernd linear ist. In der Regel werden lediglich die Punkte $\big(\mathrm{g}(t_j),\ S_Z^{-1}[\hat{S}_T(t_j)]\big)$ für die verschiedenen Ausfallzeiten unter den $t_1, \ldots, t_n$, graphisch dargestellt werden.

Für eine Weibull-verteilte Lebenszeit T hat $Y = \ln T$ Extremwert-Verteilung mit Standard-Survival-Funktion $S_Z(z) = \exp[-\exp(z)]$, was mit (5.26) zu der anfangs vorgeschlagenen Linearisierung (5.24) führt. Die graphische Überprüfung von Log-Normal- und Log-Logistik-Modellen für T erfolgt ebenfalls mit (5.26) und $g(t) = \ln(t)$, die Standard-Survival-Funktionen der zugehörigen Lokation-Skalen-Verteilungen sind durch $S_Z(z) = 1 - \Phi(z)$ bzw. $S_Z(z) = [1 + \exp(z)]^{-1}$ gegeben.

Man beachte, dass $S_Z^{-1}(p-1)$ das p-te Quantil der Variable $Z = (Y-u)/b$ ist. Die beschriebenen Graphiken sind also Q-Q-Plots.

Beispiel 5.1 (Zeit bis zum Abbruch einer IUP-Anwendung – Fortsetzung). In Beispiel 4.1 ist der KM-Schätzer für die Survival-Funktion zum Datensatz A.1, den Zeiten bis zum Abbruch einer IUP-Anwendung, berechnet worden. Inwiefern angenommen werden kann, dass diesem Datensatz eine Weibull-verteilte Zufallsvariable zugrundeliegt, kann nach den obigen Ausführungen mit Hilfe des KM-Schätzers über die graphische Darstellung der Punkte $\big(\ln(t_j),\ \ln[-\ln \hat{S}(t_j)]\big)$ beurteilt werden. Abbildung 5.1 stellt einen solchen Plot dar und bestärkt die Weibull-Verteilung als geeignetes Modell. Mit Hilfe des „R"-Codes in Anhang B.2 erhält man die Regressionsgerade $g_W(x) = 1.322\,x - 6.186,\ \ x = \ln(t)$. Die Steigung liefert einen Schätzer für den Formparameter β, nämlich $\hat{\beta}^* = 1.322$. Da der Schnittpunkt der Regressionsgeraden mit der $\ln t$-Achse eine Schätzung für $\ln \alpha$ ist, ergibt sich ferner $\hat{\alpha}^* = \exp(4.679) = 107.662$. Der geschätzte Wert für β liegt ziemlich nahe bei 1 und lässt somit darauf schließen, dass die vorliegende Stichprobe der Abbruchzeiten auch durch eine Exponential-Verteilung angemessen modelliert werden kann. Der Plot von t_j gegen $\ln \hat{S}(t_j)$ wird zur Bestätigung ebenfalls in Abbildung 5.1 dargestellt. Die zugehörige Regressionsgerade, $g_E(x) = -0.0118\,x + 0.14783$, liefert den graphischen Schätzer $\hat{\lambda}^* = 0.0118$.

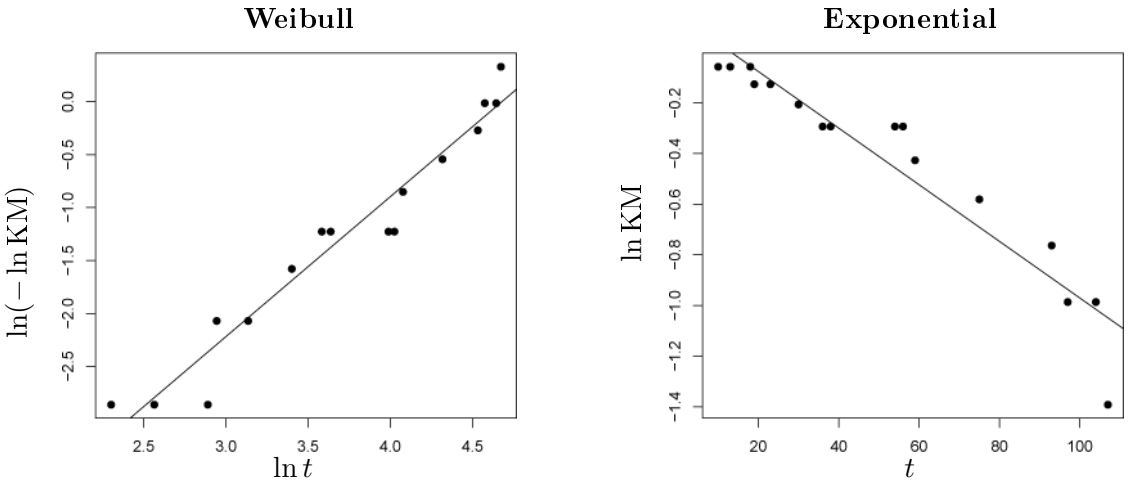

Abbildung 5.1: Plots zur Bewertung der Weibull- und Exponential-Verteilung als Modell zum Datensatz A.1.

Nimmt man an, dass der gegebene Datensatz durch eine Exponential-Verteilung adäquat beschrieben wird, so kann mit Hilfe der Likelihood-Funktion aus Kapitel 3,

$$L(\lambda) \quad = \quad \prod_{i=1}^{n} f_{T_i}(y_i)^{\delta_i} S_{T_i}(y_i)^{1-\delta_i} \tag{5.27}$$

ein ML-Schätzer für den Parameter λ bestimmt werden. Da für eine Exp(λ)-verteilte Zufallsvariable T gilt $f_T(t) = \lambda \exp(-\lambda t)$ und $S_T(t) = \exp(-\lambda t)$, ergibt sich

$$L(\lambda) \;=\; \prod_{i=1}^{n} \Big[\lambda \exp(-\lambda y_i)\Big]^{\delta_i} \Big[\exp(-\lambda y_i)\Big]^{1-\delta_i} \;=\; \prod_{i=1}^{n} \lambda^{\delta_i} \exp(-\lambda y_i).$$

Mit $r = \sum_{i=1}^{n} \delta_i$ erhält man damit die logarithmierte Likelihood-Funktion

$$l(\lambda) \;=\; \sum_{i=1}^{n} \delta_i \ln \lambda - \sum_{i=1}^{n} \lambda\, y_i \;=\; r \ln \lambda - \lambda \sum_{i=1}^{n} y_i.$$

Weiter ist

$$\frac{\partial l(\lambda)}{\partial \lambda} \;=\; \frac{r}{\lambda} - \sum_{i=1}^{n} y_i \;\stackrel{!}{=}\; 0 \;\Longleftrightarrow\; \lambda \;=\; \frac{r}{\sum_{i=1}^{n} y_i},$$

so dass man wegen $\partial^2 l(\lambda)/\partial \lambda^2 = -r/\lambda^2 < 0$ für den gegebenen Datensatz mit $n = 18$, $r = 9$ und $\sum_{i=1}^{18} y_i = 1046$ den ML-Schätzer $\hat{\lambda} = 0.0086$ erhält.

Auf entsprechende Weise können die Parameter der Weibull-Verteilung und die der anderen in diesem Kapitel vorgestellten Modelle geschätzt werden.

Kapitel 6

Das Accelerated-Failure-Time-Modell

Bislang wurden ausschließlich Überlebenswahrscheinlichkeiten homogener Populationen betrachtet. In der Lebenszeitanalyse trifft man jedoch häufig auf das Problem, dass die beobachteten Daten Werte von Kovariablen enthalten, die Auswirkungen auf die Zufallsvariable T haben. In einer Krebsstudie können der allgemeine Zustand des Patienten, sein Alter oder die Behandlungsart für die Lebensdauer entscheidend sein. Sind relevante Zusatzinformationen vorhanden, so betrachtet man neben der Lebenszeit T einen Vektor $\mathbf{X} = (X_1, \ldots, X_p)'$ von erklärenden Variablen, die quantitativ (Blutdruck, Alter, Gewicht), qualitativ (Geschlecht, Behandlungsart) und/oder zeitabhängig sein können. Zu zeitabhängigen Kovariablen gehören zum Beispiel wiederholende Messungen von Variaben oder die Information, ob ein bestimmtes Zwischenereignis bis zum Zeitpunkt t eingetreten ist oder nicht. Erklärende Variablen, die sich im Verlauf der Zeit verändern, notiert man als $\mathbf{X} = (X_1(t), \ldots, X_p(t))', \ t \geq 0$. Sie werden innerhalb dieser Arbeit jedoch nicht diskutiert.

In diesem Kapitel wird das Accelerated-Failure-Time-Modell (AFT-Modell) für zeitunabhängige Kovariablen vorgestellt. Es handelt sich dabei um ein parametrisches Regressionsmodell, in dem entsprechend seines Namens angenommen wird, dass sich erklärende Variablen auf das Vergehen der Lebenszeit eines Individuums auswirken, sie entweder beschleunigen oder verlangsamen. Die folgenden Ausführungen erfolgen in Anlehnung an [Law-03, S. 292–295] und [Kle-97, S. 373–375].

Definition 6.1 (AFT-Modell). *Sei $T \geq 0$ eine Lebensdauer und $\mathbf{X} = (X_1, \ldots, X_p)'$ ein p-dimensionaler Vektor von erklärenden Variablen. Dann ist das*

AFT-Modell durch folgenden Zusammenhang definiert:

$$S\big(t \mid \mathbf{X}\big) \;=\; S_0\big[\exp(\boldsymbol{\gamma}'\mathbf{X})\, t\big], \quad t \geq 0. \tag{6.1}$$

Dabei ist $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)' \in \mathbb{R}^p$ *ein Vektor von Regressionskoeffizienten und* S_0 *die Survival-Funktion eines Individuums mit Kovariablenvektor* $\mathbf{X} = \mathbf{0}$*. Man nennt* S_0 *Basis-Survival-Funktion (baseline survival function) und* $\exp(\boldsymbol{\gamma}'\mathbf{X})$ *Beschleunigungsfaktor (acceleration factor). Entsprechend heißen die Hazard-Funktion und das p-Quantil der Lebenszeit eines Individuums mit* $\mathbf{X} = \mathbf{0}$ *Basis-Hazard-Funktion (baseline hazard function) und Basis-p-Quantil. Ihre Bezeichnungen sind* λ_0 *bzw.* $t_{0\,p}$*.*

Die erklärenden Variablen $\mathbf{X} \neq \mathbf{0}$ eines Individuums verändern seine Zeitskala gegenüber der eines Individuums, dessen Kovariablen sämtlich 0 sind, also um den Faktor $\exp(\boldsymbol{\gamma}'\boldsymbol{X})$. In Abhängigkeit von $\operatorname{sign}(\boldsymbol{\gamma}'\mathbf{X})$ wird die Zeit entweder beschleunigt (accelerated) oder verlangsamt.

Satz 6.1. *Mit Hilfe der Basis-Hazard-Funktion* λ_0 *und des Basis-p-Quantils* $t_{0\,p}$ *können im AFT-Modell (6.1) die Hazard-Funktion und das p-Quantil zu einem beliebigen Kovariablen-Vektor* $\mathbf{X}$ *wie folgt dargestellt werden:*

$$\lambda(t \mid \mathbf{X}) \;=\; \exp(\boldsymbol{\gamma}'\mathbf{X})\, \lambda_0\big[\exp(\boldsymbol{\gamma}'\mathbf{X})\, t\big], \quad t \geq 0, \tag{6.2}$$

$$t_p(\mathbf{X}) \;=\; \exp(-\boldsymbol{\gamma}'\mathbf{X})\, t_{0\,p}. \tag{6.3}$$

Beweis: Nach Satz 2.2 gilt $\lambda(t) = -\partial/\partial t\, (\ln S(t))$ und damit

$$\begin{aligned}
\lambda(t \mid \mathbf{X}) &= -\frac{\partial}{\partial t} \ln S(t \mid \mathbf{X}) \\
&= -\frac{\partial}{\partial t} \ln S_0\big[\exp(\boldsymbol{\gamma}'\mathbf{X})\, t\big] \\
&= -\frac{\partial}{\partial z} \ln S_0(z)\, \exp(\boldsymbol{\gamma}'\mathbf{X})\,, \qquad z = \exp(\boldsymbol{\gamma}'\mathbf{X})\, t \\
&= \lambda_0\big[\exp(\boldsymbol{\gamma}'\mathbf{X})\, t\big]\, \exp(\boldsymbol{\gamma}'\mathbf{X}).
\end{aligned}$$

Beachtet man, dass für das p-Quantil $t_p(\mathbf{X}) = S^{-1}(1 - p \mid \mathbf{X})$ gilt, so ergibt sich auch die zweite Aussage direkt aus (6.1): Aus $S(t_p(\mathbf{X}) \mid \mathbf{X}) = 1 - p$ folgt $S_0\big[\exp(\boldsymbol{\gamma}'\mathbf{X})\, t_p(\mathbf{X})\big] = 1 - p$ und schließlich $t_p(\mathbf{X}) = S_0^{-1}(1 - p)\, \exp(\text{-}\,\boldsymbol{\gamma}'\mathbf{X}) = t_{0p}\, \exp(\text{-}\,\boldsymbol{\gamma}'\mathbf{X})$. □

Für $T > 0$ kann das AFT-Modell alternativ als lineares Modell für die logarithmierte Lebenszeit $Y = \ln T$ dargestellt werden.

Satz 6.2. *Ist $T > 0$ eine Lebensdauer, $Y = \ln T$ und $\mathbf{X} = (X_1, \ldots, X_p)'$ ein p-dimensionaler Vektor von erklärenden Variablen, so wird im AFT-Modell der Einfluss der Kovariablen auf die Lebenszeit durch folgenden Zusammenhang beschrieben:*

$$Y = \ln(T) = \beta_0 + \boldsymbol{\beta}'\mathbf{X} + bZ, \tag{6.4}$$

mit einem Regressionsvektor $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)' \in \mathbb{R}^p$ und einer Fehlerverteilung Z.

Beweis: Wird mit S_0 die Survival-Funktion von $T = \exp(Y)$ mit $\mathbf{X} = \mathbf{0}$ bezeichnet, so gilt für die Überlebenswahrscheinlichkeit eines Individuums mit beliebigem Vektor $\mathbf{X}$:

$$\begin{aligned} S(t \mid \mathbf{X}) &= P\big[\exp(\beta_0 + \boldsymbol{\beta'}\boldsymbol{X} + bZ) > t\big] \\ &= P\big[\exp(\beta_0 + bZ) > t\exp(-\boldsymbol{\beta'}\boldsymbol{X})\big] \\ &= S_0\big[t\exp(-\boldsymbol{\beta'}\boldsymbol{X})\big]. \end{aligned}$$

Das lineare Modell (6.4) entspricht also der Darstellung (6.1) mit $\boldsymbol{\gamma} = -\boldsymbol{\beta}$. □

Das AFT-Modell eignet sich insbesondere für die in Kapitel 5 vorgestellten Log-Lokation-Skalen-Verteilungen. In der Lebenszeitanalyse nimmt man daher in der Regel an, dass der Fehler Z in (6.4) eine standardverteilte Zufallsvariable aus der Familie der Normal-, Extremwert- oder Logistik-Verteilungen ist. In den folgenden Abschnitten sollen Maximum-Likelihood-Schätzungen für b, β_0 und $\boldsymbol{\beta}$ bestimmt werden. Dazu werden zunächst allgemeine Likelihood-Methoden für AFT-Modelle behandelt, mit deren Hilfe dann die einzelnen Log-Lokation-Skalen-Modelle aus Kapitel 5 untersucht werden können.

6.1 Likelihood-Methoden

Dieser Abschnitt behandelt ML-Schätzungen für das lineare Regressionsmodell (6.4), das mit $u = u(\mathbf{X}) = \beta_0 + \boldsymbol{\beta}'\mathbf{X}$ dem Log-Lokation-Skalen-Modell (5.18) entspricht. Über die in Abschnitt 5.1.5 dargestellten Eigenschaften von Lokation-Skalen-Verteilungen können dann Aussagen über die Lebensdauer T gemacht werden.

Gegeben seien die Realisierungen der identisch und unabhängig verteilten Zufallsvariablen $(Y_i, \Delta_i, \mathbf{X}_i)$, $i = 1, \ldots, n$. Mit $Y_i = \ln T_i$ wird hier entweder eine logarith-

mierte Lebenszeit oder eine logarithmierte Zensur-Zeit bezeichnet, abhängig davon, ob für den beobachteten Wert des Zensurindikators $\delta_i = 1$ oder $\delta_i = 0$ gilt.[1]

Satz 6.3. *Die Likelihood-Funktion für die Stichprobe* $\big((y_1, \delta_1, \mathbf{x}_1), \ldots, (y_n, \delta_n, \mathbf{x}_n)\big)$ *ist im log-linearen Modell (6.4) mit* $u(\mathbf{x}) = \beta_0 + \boldsymbol{\beta}'\mathbf{x}$ *durch*

$$L\big(u(\mathbf{x}),\, b\big) \;=\; \prod_{i=1}^{n} \left[\frac{1}{b} f_{Z_i}\left(\frac{y_i - u(\mathbf{x}_i)}{b}\right)\right]^{\delta_i} S_{Z_i}\left(\frac{y_i - u(\mathbf{x}_i)}{b}\right)^{1-\delta_i} \tag{6.5}$$

gegeben. Dabei bezeichnet f_{Z_i} *die Wahrscheinlichkeitsdichte und* S_{Z_i} *die Survival-Funktion der Fehlerverteilung* Z_i.

Beweis: Ist die logarithmierte Lebenszeit des i-ten Individuums als $\ln T_i = Y_i = u(\mathbf{X}_i) + bZ_i$ gegeben, so ist nach (5.17) bzw. (5.20)

$$f_{Y_i}(y_i) = \frac{1}{b} f_{Z_i}\left(\frac{y_i - u(\mathbf{x}_i)}{b}\right) \quad \text{und} \quad S_{Y_i}(y_i) = S_{Z_i}\left(\frac{y_i - u(\mathbf{x}_i)}{b}\right).$$

Die Likelihood-Funktion (6.5) entspricht also der Likelihood-Funktion aus Abschnitt 3.2. □

Setzt man

$$z_i = \frac{y_i - u_i}{b} \qquad \text{mit} \qquad u_i = u(\mathbf{x}_i) = \beta_0 + \boldsymbol{\beta}'\mathbf{x}_i, \;\; \mathbf{x}_i = (x_{i1}, \ldots, x_{ip})',$$

so erhält man aus (6.5) die Log-Likelihood-Funktion

$$\begin{aligned} l(\beta_0,\, \boldsymbol{\beta},\, b) \;&=\; \sum_{i=1}^{n} \big[-\delta_i \ln b + \delta_i f_{Z_i}(z_i) + (1-\delta_i) \ln S_{Z_i}(z_i)\big] \\ &=\; -r \ln b + \sum_{i=1}^{n} \big[\delta_i \ln f_{Z_i}(z_i) + (1-\delta_i) \ln S_{Z_i}(z_i)\big] \end{aligned} \tag{6.6}$$

mit $r = \sum_{i=1}^{n} \delta_i$. Für die ersten partiellen Ableitungen von $z_i(\beta_0, \boldsymbol{\beta}, b)$ gilt mit $\beta_j \neq \beta_0$, $j = 1, \ldots, p$,

$$\frac{\partial z_i}{\partial \beta_0} = -\frac{1}{b}, \qquad \frac{\partial z_i}{\partial \beta_j} = -\frac{1}{b} x_{ij} \qquad \text{und} \qquad \frac{\partial z_i}{\partial b} = -\frac{1}{b} z_i. \tag{6.7}$$

[1] In den Kapiteln 3 und 4 ist die Bezeichnung $Y_i = \min(T_i, C_i)$ für die Beobachtungszeiten benutzt worden. Dabei war T_i die Lebenszeit und C_i die Zensurzeit des i-ten Individuums. Um Verwechslungen mit den logarithmierten Lebenszeiten zu vermeiden, wird hier auf diese Schreibweise verzichtet.

Damit können die partiellen Ableitungen von (6.6) für $\beta_j \neq \beta_0$, $\beta_k \neq \beta_0$, $j, k = 1, \ldots, p$ und $f_Z = f_{Z_i}$, sowie $S_Z = S_{Z_i}$ angegeben werden als

$$\begin{aligned}
\frac{\partial l}{\partial \beta_0} &= \sum_{i=1}^{n} \left[\delta_i \frac{\partial \ln f_Z(z_i)}{\partial z_i} \frac{\partial z_i}{\partial \beta_0} + (1-\delta_i) \frac{\partial \ln S_Z(z_i)}{\partial z_i} \frac{\partial z_i}{\partial \beta_0} \right] \\
&= -\frac{1}{b} \sum_{i=1}^{n} \left[\delta_i \frac{\partial \ln f_Z(z_i)}{\partial z_i} + (1-\delta_i) \frac{\partial \ln S_Z(z_i)}{\partial z_i} \right], \qquad (6.8)
\end{aligned}$$

$$\begin{aligned}
\frac{\partial l}{\partial \beta_j} &= \sum_{i=1}^{n} \left[\delta_i \frac{\partial \ln f_Z(z_i)}{\partial z_i} \frac{\partial z_i}{\partial \beta_j} + (1-\delta_i) \frac{\partial \ln S_Z(z_i)}{\partial z_i} \frac{\partial z_i}{\partial \beta_j} \right] \\
&= -\frac{1}{b} \sum_{i=1}^{n} \left[\delta_i \frac{\partial \ln f_Z(z_i)}{\partial z_i} + (1-\delta_i) \frac{\partial \ln S_Z(z_i)}{\partial z_i} \right] x_{ij}, \qquad (6.9)
\end{aligned}$$

$$\begin{aligned}
\frac{\partial l}{\partial b} &= -\frac{r}{b} + \sum_{i=1}^{n} \left[\delta_i \frac{\partial \ln f_Z(z_i)}{\partial z_i} \frac{\partial z_i}{\partial b} + (1-\delta_i) \frac{\partial \ln S_Z(z_i)}{\partial z_i} \frac{\partial z_i}{\partial b} \right] \\
&= -\frac{r}{b} - \frac{1}{b} \sum_{i=1}^{n} \left[\delta_i \frac{\partial \ln f_Z(z_i)}{\partial z_i} + (1-\delta_i) \frac{\partial \ln S_Z(z_i)}{\partial z_i} \right] z_i, \qquad (6.10)
\end{aligned}$$

$$\begin{aligned}
\frac{\partial^2 l}{\partial \beta_0^2} &= -\frac{1}{b} \sum_{i=1}^{n} \left[\delta_i \frac{\partial^2 \ln f_Z(z_i)}{\partial z_i^2} \frac{\partial z_i}{\partial \beta_0} + (1-\delta_i) \frac{\partial^2 \ln S_Z(z_i)}{\partial z_i^2} \frac{\partial z_i}{\partial \beta_0} \right] \\
&= \frac{1}{b^2} \sum_{i=1}^{n} \left[\delta_i \frac{\partial^2 \ln f_Z(z_i)}{\partial z_i^2} + (1-\delta_i) \frac{\partial^2 \ln S_Z(z_i)}{\partial z_i^2} \right], \qquad (6.11)
\end{aligned}$$

$$\begin{aligned}
\frac{\partial^2 l}{\partial \beta_0 \, \partial b} &= -\frac{1}{b} \sum_{i=1}^{n} \left[\delta_i \frac{\partial^2 \ln f_Z(z_i)}{\partial z_i^2} \frac{\partial z_i}{\partial \beta_0} + (1-\delta_i) \frac{\partial^2 \ln S_Z(z_i)}{\partial z_i^2} \frac{\partial z_i}{\partial \beta_0} \right] z_i \\
&\quad - \frac{1}{b} \sum_{i=1}^{n} \left[\delta_i \frac{\partial \ln f_Z(z_i)}{\partial z_i} + (1-\delta_i) \frac{\partial \ln S_Z(z_i)}{\partial z_i} \right] \frac{\partial z_i}{\partial \beta_0} \\
&= \frac{1}{b^2} \sum_{i=1}^{n} \left[\delta_i \frac{\partial^2 \ln f_Z(z_i)}{\partial z_i^2} + (1-\delta_i) \frac{\partial^2 \ln S_Z(z_i)}{\partial z_i^2} \right] z_i \\
&\quad + \frac{1}{b^2} \sum_{i=1}^{n} \left[\delta_i \frac{\partial \ln f_Z(z_i)}{\partial z_i} + (1-\delta_i) \frac{\partial \ln S_Z(z_i)}{\partial z_i} \right], \qquad (6.12)
\end{aligned}$$

$$\begin{aligned}
\frac{\partial^2 l}{\partial\beta_0\,\partial\beta_j} &= -\frac{1}{b}\sum_{i=1}^{n}\left[\delta_i\,\frac{\partial^2 \ln f_Z(z_i)}{\partial z_i^2}\,\frac{\partial z_i}{\partial\beta_0} + (1-\delta_i)\,\frac{\partial^2 \ln S_Z(z_i)}{\partial z_i^2}\,\frac{\partial z_i}{\partial\beta_0}\right] x_{ij} \\
&= \frac{1}{b^2}\sum_{i=1}^{n}\left[\delta_i\,\frac{\partial^2 \ln f_Z(z_i)}{\partial z_i^2} + (1-\delta_i)\,\frac{\partial^2 \ln S_Z(z_i)}{\partial z_i^2}\right] x_{ij}, \qquad (6.13)
\end{aligned}$$

$$\begin{aligned}
\frac{\partial^2 l}{\partial\beta_j\,\partial b} &= -\frac{1}{b}\sum_{i=1}^{n}\left[\delta_i\,\frac{\partial^2 \ln f_Z(z_i)}{\partial z_i^2}\,\frac{\partial z_i}{\partial\beta_j} + (1-\delta_i)\,\frac{\partial^2 \ln S_Z(z_i)}{\partial z_i^2}\,\frac{\partial z_i}{\partial\beta_j}\right] z_i \\
&\quad - \frac{1}{b}\sum_{i=1}^{n}\left[\delta_i\,\frac{\partial \ln f_Z(z_i)}{\partial z_i} + (1-\delta_i)\,\frac{\partial \ln S_Z(z_i)}{\partial z_i}\right]\frac{\partial z_i}{\partial\beta_j} \\
&= \frac{1}{b^2}\sum_{i=1}^{n}\left[\delta_i\,\frac{\partial^2 \ln f_Z(z_i)}{\partial z_i^2} + (1-\delta_i)\,\frac{\partial^2 \ln S_Z(z_i)}{\partial z_i^2}\right] x_{ij}\,z_i \\
&\quad + \frac{1}{b^2}\sum_{i=1}^{n}\left[\delta_i\,\frac{\partial \ln f_Z(z_i)}{\partial z_i} + (1-\delta_i)\,\frac{\partial \ln S_Z(z_i)}{\partial z_i}\right] x_{ij}, \qquad (6.14)
\end{aligned}$$

$$\begin{aligned}
\frac{\partial^2 l}{\partial\beta_k\,\partial\beta_j} &= -\frac{1}{b}\sum_{i=1}^{n}\left[\delta_i\,\frac{\partial^2 \ln f_Z(z_i)}{\partial z_i^2}\,\frac{\partial z_i}{\partial\beta_k} + (1-\delta_i)\,\frac{\partial^2 \ln S_Z(z_i)}{\partial z_i^2}\,\frac{\partial z_i}{\partial\beta_k}\right] x_{ij} \\
&= \frac{1}{b^2}\sum_{i=1}^{n}\left[\delta_i\,\frac{\partial^2 \ln f_Z(z_i)}{\partial z_i^2} + (1-\delta_i)\,\frac{\partial^2 \ln S_Z(z_i)}{\partial z_i^2}\right] x_{ij}\,x_{ik}, \qquad (6.15)
\end{aligned}$$

$$\begin{aligned}
\frac{\partial^2 l}{\partial b^2} &= \frac{r}{b^2} + \frac{1}{b^2}\sum_{i=1}^{n}\left[\delta_i\,\frac{\partial \ln f_Z(z_i)}{\partial z_i} + (1-\delta_i)\,\frac{\partial \ln S_Z(z_i)}{\partial z_i}\right] z_i \\
&\quad - \frac{1}{b}\sum_{i=1}^{n}\left[\left(\delta_i\,\frac{\partial^2 \ln f_Z(z_i)}{\partial z_i^2}\,\frac{\partial z_i}{\partial b} + (1-\delta_i)\,\frac{\partial^2 \ln S_Z(z_i)}{\partial z_i^2}\,\frac{\partial z_i}{\partial b}\right) z_i\right. \\
&\qquad \left. + \left(\delta_i\,\frac{\partial \ln f_Z(z_i)}{\partial z_i} + (1-\delta_i)\,\frac{\partial \ln S_Z(z_i)}{\partial z_i}\right)\frac{\partial z_i}{\partial b}\right] \\
&= \frac{r}{b^2} + \frac{1}{b^2}\sum_{i=1}^{n}\left[\delta_i\,\frac{\partial \ln f_Z(z_i)}{\partial z_i} + (1-\delta_i)\,\frac{\partial \ln S_Z(z_i)}{\partial z_i}\right] z_i \\
&\quad + \frac{1}{b^2}\sum_{i=1}^{n}\left[\delta_i\,\frac{\partial^2 \ln f_Z(z_i)}{\partial z_i^2} + (1-\delta_i)\,\frac{\partial^2 \ln S_Z(z_i)}{\partial z_i^2}\right] z_i^2 \\
&\quad + \frac{1}{b^2}\sum_{i=1}^{n}\left[\delta_i\,\frac{\partial \ln f_Z(z_i)}{\partial z_i} + (1-\delta_i)\,\frac{\partial \ln S_Z(z_i)}{\partial z_i}\right] z_i
\end{aligned}$$

$$
\begin{aligned}
&= \frac{r}{b^2} + \frac{2}{b^2}\sum_{i=1}^{n}\left[\delta_i \frac{\partial \ln f_Z(z_i)}{\partial z_i} + (1-\delta_i)\frac{\partial \ln S_Z(z_i)}{\partial z_i}\right] z_i \\
&\quad + \frac{1}{b^2}\sum_{i=1}^{n}\left[\delta_i \frac{\partial^2 \ln f_Z(z_i)}{\partial z_i^2} + (1-\delta_i)\frac{\partial^2 \ln S_Z(z_i)}{\partial z_i^2}\right] z_i^2. \qquad (6.16)
\end{aligned}
$$

Als Informationsmatrix ergibt sich schließlich

$$
\mathrm{I}(\beta_0, \boldsymbol{\beta}, b) = \begin{pmatrix}
-\frac{\partial^2 l}{\partial \beta_0^2} & -\frac{\partial^2 l}{\partial \beta_1 \partial \beta_0} & \cdots & -\frac{\partial^2 l}{\partial \beta_p \partial \beta_0} & -\frac{\partial^2 l}{\partial b \partial \beta_0} \\
-\frac{\partial^2 l}{\partial \beta_0 \partial \beta_1} & -\frac{\partial^2 l}{\partial \beta_1^2} & \cdots & -\frac{\partial^2 l}{\partial \beta_p \partial \beta_1} & -\frac{\partial^2 l}{\partial b \partial \beta_1} \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
-\frac{\partial^2 l}{\partial \beta_0 \partial \beta_p} & -\frac{\partial^2 l}{\partial \beta_1 \partial \beta_p} & \cdots & -\frac{\partial^2 l}{\partial \beta_p^2} & -\frac{\partial^2 l}{\partial b \partial \beta_p} \\
-\frac{\partial^2 l}{\partial \beta_0 \partial b} & -\frac{\partial^2 l}{\partial \beta_1 \partial b} & \cdots & -\frac{\partial^2 l}{\partial \beta_p \partial b} & -\frac{\partial^2 l}{\partial b^2}
\end{pmatrix}.
$$

Bei großen Stichproben kann die gemeinsame Verteilung von $(\hat{\beta}_0, \hat{\boldsymbol{\beta}}, \hat{b})$ durch die $(p+2)$-dimensionale Normalverteilung mit Erwartungsvektor $(\hat{\beta}_0, \hat{\boldsymbol{\beta}}, \hat{b})'$ und Varianz-Kovarianz-Matrix $\mathrm{I}^{-1}(\hat{\beta}_0, \hat{\boldsymbol{\beta}}, \hat{b})$ approximiert werden (vgl. Abschnitt 3.3).

Bemerkung 6.1. Da die Ableitungen (6.8), (6.9) und (6.10) im ML-Schätzer $(\hat{\beta}_0, \hat{\boldsymbol{\beta}}, \hat{b})$ den Wert 0 annehmen, vereinfachen sich die partiellen Ableitungen (6.12), (6.14) und (6.16) in der geschätzten Varianz-Kovarianz-Matrix $\mathrm{I}(\hat{\beta}_0, \hat{\boldsymbol{\beta}}, \hat{b})$: Für $(\beta_0, \boldsymbol{\beta}\ , b) = (\hat{\beta}_0, \hat{\boldsymbol{\beta}}, \hat{b})$ ist der zweite Term in (6.12) gleich 0, ebenso der zweite in (6.14). Der zweite Term in (6.16) vereinfacht sich zu $-\frac{2r}{\hat{b}^2}$. Mit $\hat{z}_i = (y_i - \hat{\beta}_0 - \hat{\boldsymbol{\beta}})/\hat{b}$ bedeutet das

$$
\begin{aligned}
\frac{\partial^2 l(\hat{\beta}_0, \hat{\boldsymbol{\beta}}, \hat{b})}{\partial \beta_0 \partial b} &= \frac{1}{\hat{b}^2}\sum_{i=1}^{n}\left[\delta_i \frac{\partial^2 \ln f_Z(\hat{z}_i)}{\partial z_i^2} + (1-\delta_i)\frac{\partial^2 \ln S_Z(\hat{z}_i)}{\partial z_i^2}\right]\hat{z}_i, \\
\frac{\partial^2 l(\hat{\beta}_0, \hat{\boldsymbol{\beta}}, \hat{b})}{\partial \beta_j \partial b} &= \frac{1}{\hat{b}^2}\sum_{i=1}^{n}\left[\delta_i \frac{\partial^2 \ln f_Z(\hat{z}_i)}{\partial z_i^2} + (1-\delta_i)\frac{\partial^2 \ln S_Z(\hat{z}_i)}{\partial z_i^2}\right]\hat{z}_i\, x_{ij}, \\
\frac{\partial^2 l(\hat{\beta}_0, \hat{\boldsymbol{\beta}}, \hat{b})}{\partial b^2} &= -\frac{r}{\hat{b}^2} + \frac{1}{\hat{b}^2}\sum_{i=1}^{n}\left[\delta_i \frac{\partial^2 \ln f_Z(\hat{z}_i)}{\partial z_i^2} + (1-\delta_i)\frac{\partial^2 \ln S_Z(\hat{z}_i)}{\partial z_i^2}\right]\hat{z}_i^2.
\end{aligned}
$$

Die gängigen Software-Pakete berechnen die ML-Schätzer für die unbekannten Parameter $\beta_0, \beta_1, \ldots, \beta_p$ und b mittels des sogenannten Newton-Raphson-Verfahrens, welches unter anderem in [Lee-03, S. 428–432] beschrieben wird. In den beiden folgenden Abschnitten werden das Weibull-, das Log-Logistik- und das Log-Normal-AFT-Modell an einen konkreten Datensatz angepasst und miteinander verglichen.

6.2 Anpassung von AFT-Modellen – Prognose für Brustkrebs-Patientinnen

Brustkrebs, auch Mammakarzinom genannt, ist die bei Frauen am häufigsten vorkommende Krebsform. Während gutartige Brusttumore lediglich aus einer Ansammlung von inaktiven Zellen bestehen, breiten sich bösartige Tumorzellen im Körper aus und bilden sogenannte Metastasen.[2] Wird Brustkrebs vor Beginn der Metastisierung diagnostiziert, so ist er in der Regel vollständig heilbar. Das Ziel einer von Leathem und Brooks [Lea-87] dokumentierten retrospektiven Studie bestand daher in der Bewertung eines histochemischen Markers, mit dessen Hilfe zwischen metastisierenden Tumorzellen und solchen ohne Metastasen differenziert werden kann. Der Marker, das Lecithin *Helix Pomatia* agglutinin (HPA), bindet an Brustkrebszellen mit Metastasen in den regionären Lymphknoten, wodurch diese dann mikroskopisch identifiziert werden können. Tabelle A.2 in Anhang A enthält die Lebensdauern von 45 Patientinnen, bei denen zur Behandlung eines Tumors im Stadium II, III oder IV – d.h. insbesondere bei nachgewiesenen Lymphknotenmetastasen – zwischen Januar 1969 und Dezember 1971 eine Mastektomie durchgeführt worden ist. Teile der entfernten Tumore sind mit HPA behandelt und hinsichtlich der HPA-Markierung als positiv oder negativ klassifiziert worden. In Anlehnung an das Beispiel 6.3 in [Col-03, S. 217–220] werden im Folgenden AFT-Modelle an die Lebenszeitdaten dieser Bruskrebs-Patientinnen angepasst. Die Modellkonstruktion erfolgt dabei über die log-lineare Darstellung (6.4) und die Likelihood-Funktion (6.5). Die hierfür notwendigen Berechnungen werden mit der Statistik-Software „R“ durchgeführt, die zugehörigen Quellcodes sind in Anhang B, Abschnitt B.3 zu finden.

6.2.1 Das Weibull-AFT-Modell

Nimmt man an, dass die Lebensdauern der Brustkrebs-Patientinnen einer Weibull-Verteilung folgen und dass X als dichotomer Faktor das Resultat der HPA-Markierung beschreibt, so ist das log-lineare Modell gemäß Satz 6.2 durch

$$\ln T \;=\; \beta_0 + \beta_1 X + bZ \tag{6.17}$$

[2] Eine ausführlichere Beschreibung des Krankheitsverlaufs ist in [Psc-04, S. 1109f] zu finden.

gegeben, wobei der Fehler Z eine Standard-Extremwert-Verteilung besitzt und die Parameter $\beta_0,\ \beta_1$ und b unbekannt sind. Die Kovariable X nehme bei positiver HPA-Markierung den Wert 1 und bei negativer HPA-Markierung den Wert 0 an. Über die Likelihood-Funktion (6.5) erhält man mit Hilfe von „R“ die Schätzer

$$\hat{\beta}_0 \;=\; 5.8544, \qquad \hat{\beta}_1 \;=\; -0.9967, \qquad \hat{b} \;=\; 1.0668. \tag{6.18}$$

Der Acceleration-Faktor für die i-te Patientin ist damit $\exp(-\hat{\beta}_1 X_i)$. Die Lebenszeit einer Frau mit positiver HPA-Markierung ist also gegenüber der einer Patientin mit negativer Markierung um den Faktor $\exp(-\hat{\beta}_1) \approx 2.7093$ beschleunigt. Nach Abschnitt 5.1.5 kann die Survival-Funktion zur Weibull-verteilten Lebensdauer T mit Hilfe der Survival-Funktion zu Z, $S_Z(z) = \exp[-\exp(z)]$, wie folgt dargestellt werden:

$$S_T(t \mid X) \;=\; \exp\left[-\exp\left(\frac{\ln t - \beta_0 - \beta_1 X}{b} \right) \right], \quad t > 0. \tag{6.19}$$

Bei gegebenem Datensatz A.3 ist die geschätzte Survival-Funktion für die i-te Patientin damit durch

$$\begin{aligned} \hat{S}_{T_i}(t \mid X_i) \;&=\; \exp\left[-\exp\left(\frac{\ln t - \hat{\beta}_0 - \hat{\beta}_1 X_i}{\hat{b}} \right) \right] \\ &=\; \exp\left[-\exp\left(\frac{\ln t - 5.8544 + 0.9967\, X_i}{1.0668} \right) \right] \\ &=\; \exp\left[-\, t^{0.9374} \exp\left(-5.4878 + 0.9343\, X_i \right) \right], \quad t > 0 \end{aligned} \tag{6.20}$$

gegeben. Die Graphen der geschätzten Survival-Funktionen für beide Gruppen von Patientinnen werden in Abbildung 6.1 dargestellt. Da für das p-Quantil t_p einer Lebenszeit T gilt $t_p = S_T^{-1}(1-p)$, kann der Median einer Weibull-verteilten Überlebenszeit T im AFT-Modell mit $u(X) = \beta_0 + \beta_1 X$ berechnet werden als

$$\begin{aligned} t_{0.5}(X) \;&=\; S_T^{-1}(0.5 \mid X) \\ &\overset{(6.19)}{=}\; \exp\left[\ln(-\ln 0.5)\, b + u(X) \right] \\ &=\; \exp\left[\ln(\ln 2)\, b + u(X) \right]. \end{aligned}$$

Für die i-te Patientin erhält man also den Schätzer

$$\hat{t}_{0.5}(X_i) \;=\; \exp\left[\ln(\ln 2)\, 1.0668 + 5.8544 - 0.9967 X_i \right].$$

Der geschätzte Median für die Lebenszeit einer Patientin mit negativer HPA-Markierung ist mit $\hat{t}_{0.5}(0) \approx 236$ Monaten also etwa 2.7-mal so hoch wie der für die Lebenszeit einer Patientin mit positiver HPA-Markierung, $\hat{t}_{0.5}(1) \approx 87$. Mit dem oben geschätzten Beschleunigungsfaktor $\exp(0.9967) \approx 2.7093$ entspricht dieses Ergebnis der Aussage (6.3) in Satz 6.1. Für die Schätzung der Hazard-Rate $\lambda_T(\cdot \mid X)$ betrachte man zunächst die folgende Gleichungskette

$$\begin{aligned} \lambda_T(t \mid X) &\overset{(2.8)}{=} -\frac{\partial}{\partial t} \ln S_T(t \mid X) \\ &\overset{(6.19)}{=} \frac{1}{t\,b} \exp\left(\frac{\ln t - \beta_0 - \beta_1 X}{b}\right) \\ &= \frac{1}{t^{1-b^{-1}}\,b} \exp\left(\frac{-\beta_0 - \beta_1 X}{b}\right), \quad t > 0. \end{aligned}$$

Bei gegebenem Datensatz ergibt sich daraus der Hazard-Schätzer

$$\begin{aligned} \hat{\lambda}_{T_i}(t \mid X_i) &= \frac{1}{t^{1-\hat{b}^{-1}}\,\hat{b}} \exp\left(\frac{-\hat{\beta}_0 - \hat{\beta}_1 X_i}{\hat{b}}\right) \\ &\overset{(6.18)}{=} \frac{1}{t^{1-1.0668^{-1}}\,1.0668} \exp\left(\frac{-5.8544 + 0.9967\,X_i}{1.0668}\right) \\ &= 0.9374\,t^{-0.0626} \exp(-5.4878 + 0.9343\,X_i), \quad t > 0. \end{aligned}$$

Abbildung 6.1 enthält den Plot des Schätzers $\hat{\lambda}_{T_i}(\cdot \mid X_i)$ für $X_i = 1$ und $X_i = 0$.

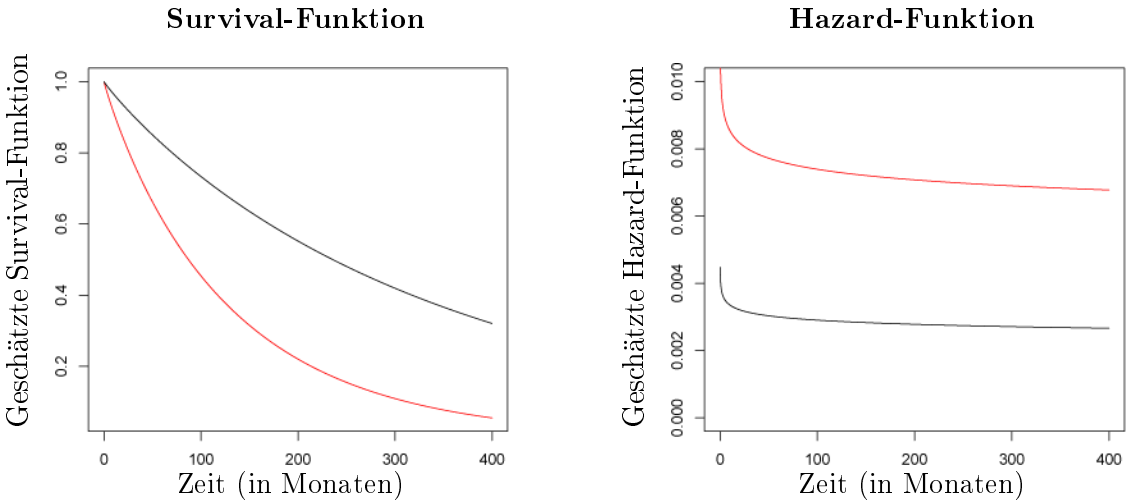

Abbildung 6.1: Geschätzte Survival- und Hazard-Funktionen im Weibull-AFT-Modell für Brustkrebs-Patientinnen mit positiver (rot) und negativer (schwarz) HPA-Markierung.

6.2.2 Das Log-Logistik-AFT-Modell

Berechnet man zum selben Datensatz die Parameter-Schätzer für die log-lineare Form eines Log-Logistik-AFT-Modells,

$$\ln T \;=\; \beta_0 + \beta_1 X + bZ, \tag{6.21}$$

mit Standard-Logistik-verteiltem Fehler Z, so erhält man

$$\hat{\beta}_0 \;=\; 5.4611, \qquad \hat{\beta}_1 \;=\; -1.1491, \qquad \hat{b} \;=\; 0.8047. \tag{6.22}$$

Der geschätzte Beschleunigungsfaktor $\exp(-\hat{\beta}_1 X_i)$ ist für $X_i = 1$ gleich 3.1554 und damit etwas größer als der unter dem Weibull-AFT-Modell in Abschnitt 6.2.1. Nach Abschnitt 5.1.5 kann die Schätzung der Survival-Funktion zur Log-Logistik-verteilten Lebensdauer T und Kovariable X_i mit Hilfe der Survival-Funktion von Z, $S_Z(z) = [1 + \exp(z)]^{-1}$, berechnet werden:

$$\begin{aligned}
\hat{S}_{T_i}(t \mid X_i) \quad &= \quad S_{Z_i}\left(\frac{\ln t - \hat{\beta}_0 - \hat{\beta}_1 X_i}{\hat{b}}\right) \\
&= \quad \left[1 + \exp\left(\frac{\ln t - \hat{\beta}_0 - \hat{\beta}_1 X_i}{\hat{b}}\right)\right]^{-1} \\
&\overset{(6.22)}{=} \quad \left[1 + \exp\left(\frac{\ln t - 5.4611 + 1.1491\, X_i}{0.8047}\right)\right]^{-1} \\
&= \quad \left[1 + t^{1.2427}\, \exp\big(-6.7865 + 1.4280\, X_i\big)\right]^{-1}, \quad t > 0.
\end{aligned} \tag{6.23}$$

Abbildung 6.2 enthält die Graphen der geschätzten Survival-Funktionen für $X_i = 1$ und $X_i = 0$. Der geschätzte Median von T_i ist

$$\begin{aligned}
\hat{t}_{0.5}(X_i) \quad &= \quad \hat{S}_{T_i}^{-1}(0.5 \mid X_i) \\
&\overset{(6.23)}{=} \quad \exp\big(\ln(1)\, \hat{b} + \hat{\beta}_0 + \hat{\beta}_1 X_i\big) \\
&\overset{(6.22)}{=} \quad \exp\big(5.4611 - 1.1491\, X_i\big).
\end{aligned} \tag{6.24}$$

Für die Lebenszeit einer Patientin mit positiv markiertem Tumor beträgt der geschätzte Median damit $\hat{t}_{0.5}(1) = \exp(5.4611 - 1.1491) \approx 75$ Monate und für die Lebenszeit einer Patientin mit negativ markiertem Tumor $\hat{t}_{0.5}(0) = \exp(5.4611) \approx 235$ Monate. Beide Werte liegen relativ nahe bei den entsprechenden Schätzungen für eine Weibull-verteilte Lebensdauer T. Als Schätzer für die Hazard-Funktion der i-ten

Patientin erhält man

$$
\begin{aligned}
\hat{\lambda}_{T_i}(t \mid X_i) &\overset{(2.8)}{=} -\frac{\partial}{\partial t} \ln \hat{S}_{T_i}(t \mid X_i) \\
&= -\frac{\partial}{\partial t} \ln S_{Z_i}\left(\frac{\ln t - \hat{\beta}_0 - \hat{\beta}_1 X_i}{\hat{b}}\right), \qquad S_{Z_i}(z) = [1+\exp(z)]^{-1} \\
&= -\frac{\partial}{\partial t} \ln \left[1 + \exp\left(\frac{\ln t - \hat{\beta}_0 - \hat{\beta}_1 X_i}{\hat{b}}\right)\right]^{-1} \\
&= \frac{\partial}{\partial t} \ln \left[1 + \exp\left(\frac{\ln t - \hat{\beta}_0 - \hat{\beta}_1 X_i}{\hat{b}}\right)\right] \\
&= \left[1 + \exp\left(\frac{\ln t - \hat{\beta}_0 - \hat{\beta}_1 X_i}{\hat{b}}\right)\right]^{-1} \exp\left(\frac{\ln t - \hat{\beta}_0 - \hat{\beta}_1 X_i}{\hat{b}}\right) \frac{1}{\hat{b}\, t} \\
&= \frac{1}{\hat{b}\, t} \left[1 + \exp\left(-\frac{\ln t - \hat{\beta}_0 - \hat{\beta}_1 X_i}{\hat{b}}\right)\right]^{-1} \\
&= \hat{b}^{-1}\, t^{-1} \left[1 + t^{-\hat{b}^{-1}} + \exp\left(\frac{\hat{\beta}_0 + \hat{\beta}_1 X_i}{\hat{b}}\right)\right]^{-1} \\
&= 1.2427\, t^{-1} \left[1 + t^{-1.2427} \exp\left(6.7865 - 1.4280\, X_i\right)\right]^{-1}, \; t > 0 \quad (6.25)
\end{aligned}
$$

Der Graph dieser Funktion wird in Abbildung 6.2 für beide Gruppen von Patientinnen dargestellt.

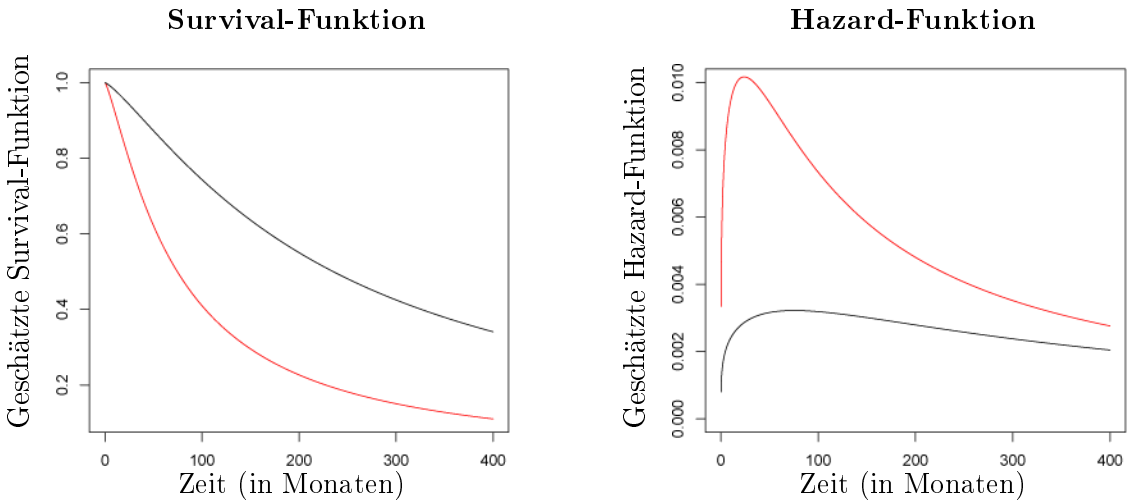

Abbildung 6.2: Geschätzte Survival- und Hazard-Funktionen im Log-Logistik-AFT-Modell für Brustkrebs-Patientinnen mit positiver (rot) und negativer (schwarz) HPA-Markierung.

6.2.3 Das Log-Normal-AFT-Modell

Bei einer log-normal-verteilten Zufallsvariable T erhält man das log-lineare Modell

$$\ln T \;=\; \beta_0 + \beta_1 X + bZ, \tag{6.26}$$

in dem der Fehler Z eine Standard-Normal-Verteilung besitzt. Die geschätzten Parameter hierfür sind

$$\hat{\beta}_0 \;=\; 5.4917, \qquad \hat{\beta}_1 \;=\; -1.1512, \qquad \hat{b} \;=\; 1.3595. \tag{6.27}$$

Die Schätzung des Beschleunigungsfaktors für $X_i = 1$ ist $\exp(-\hat{\beta}_1) = 3.1619$, was ziemlich genau mit der entsprechenden Schätzung im Log-Logistik-AFT-Modell übereinstimmt. Da Z_i eine Standard-Normal-Verteilung besitzt, ist die geschätzte Survival-Funktion zu T_i nach Abschnitt 5.1.5 durch

$$\begin{aligned}
\hat{S}_{T_i}(t \mid X_i) \;&=\; S_{Z_i}\!\left(\frac{\ln t - \hat{\beta}_0 - \hat{\beta}_1 X_i}{\hat{b}}\right) \\
&=\; 1 - \Phi\!\left(\frac{\ln t - \hat{\beta}_0 - \hat{\beta}_1 X_i}{\hat{b}}\right) \\
&\overset{(6.27)}{=}\; 1 - \Phi\big(0.7356\,\ln t - 4.0395 + 0.8468\,X_i\big), \quad t > 0
\end{aligned} \tag{6.28}$$

gegeben. Die Graphen dieser Funktion für $X_i = 0$ und $X_i = 1$ werden in Abbildung 6.3 dargestellt. Als Schätzung für den Median ergibt sich

$$\begin{aligned}
\hat{t}_{0.5}(X_i) \;&=\; \hat{S}^{-1}_{T_i}(0.5 \mid X_i) \\
&\overset{(6.28)}{=}\; \exp\left(\Phi^{-1}(0.5)\,\hat{b} + \hat{\beta}_0 + \hat{\beta}_1 X_i\right) \\
&\overset{(6.27)}{=}\; \exp\big(5.4917 - 1.1512\,X_i\big),
\end{aligned}$$

woraus sich bei positiver HPA-Markierung $\hat{t}_{0.5}(1) \approx 77$ Monate und bei negativer HPA-Markierung $\hat{t}_{0.5}(0) \approx 243$ Monate ergeben. Die Hazard-Rate für die i-te Patientin kann schließlich durch

$$\begin{aligned}
\hat{\lambda}_{T_i}(t \mid X_i) \;&\overset{(2.8)}{=}\; -\frac{\partial}{\partial t} \ln \hat{S}_{T_i}(t \mid X_i) \\
&=\; -\frac{\partial}{\partial t} \ln \left[1 - \Phi\!\left(\frac{\ln t - \hat{\beta}_0 - \hat{\beta}_1 X_i}{\hat{b}}\right)\right]
\end{aligned}$$

$$
\begin{aligned}
&= \left[1 - \Phi\left(\frac{\ln t - \hat{\beta}_0 - \hat{\beta}_1 X_i}{\hat{b}}\right)\right]^{-1} \Phi'\left(\frac{\ln t - \hat{\beta}_0 - \hat{\beta}_1 X_i}{\hat{b}}\right) t^{-1} b^{-1} \\
&\overset{(6.27)}{=} 0.7356\, t^{-1} \left[1 - \Phi\big(0.7356\, \ln t - 4.0395 + 0.8468\, X_i\big)\right]^{-1} \\
&\qquad f_{\mathcal{N}(0,1)}\big(0.7356\, \ln t - 4.0395 + 0.8468\, X_i\big), \quad t > 0
\end{aligned}
$$

geschätzt werden. Die Graphen dieser Funktion werden für beide Gruppen von Patientinnen in Abbildung 6.3 dargestellt.

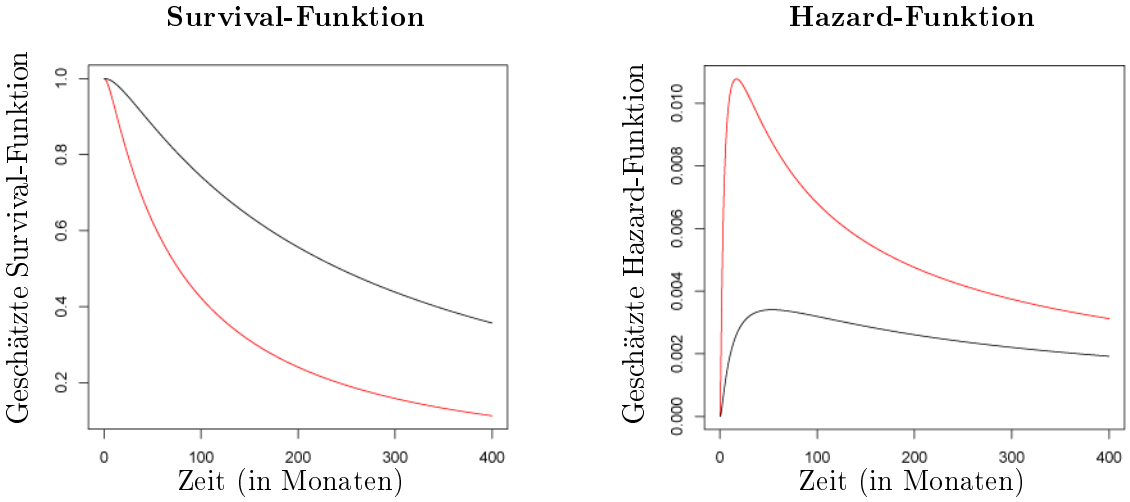

Abbildung 6.3: Geschätzte Survival- und Hazard-Funktionen im Log-Normal-AFT-Modell für Brustkrebs-Patientinnen mit positiver (rot) und negativer (schwarz) HPA-Markierung.

Sowohl bei positiver als auch negativer HPA-Markierung stimmen die geschätzten Werte für den Lebensdauer-Median unter allen drei Modellen weitgehend überein. Während sich für Patientinnen mit negativ markierten Tumoren auch die Graphen der Hazard-Funktionen ähneln, besteht bei positiver HPA-Markierung zwischen der Hazard-Funktion des Weibull-Modells und den Hazard-Funktionen der beiden anderen Modelle ein großer Unterschied. Mit Hilfe der im Folgenden diskutierten Cox-Snell-Residuen wird in Beispiel 6.1 versucht, zwischen den Modellen auszuwählen.

6.3 Überprüfung des AFT-Modells

Im vorangegangenen Abschnitt sind verschiedene AFT-Modelle an die Lebenszeitdaten von Brustkrebs-Patientinnen angepasst worden. Unterschiede zwischen den einzelnen Modellen bestanden dabei vor allem im Verlauf der Hazard-Rate bei positiv markierten Tumoren. Um zwischen den einzelnen Modellen auswählen zu können,

soll im Folgenden ein Verfahren eingeführt werden, mit dessen Hilfe die Güte der Modellanpassungen graphisch überprüft werden kann. Grundlegend sind dabei die sogenannten Cox-Snell-Residuen, die in einer allgemeinen Form von Cox und Snell [Cox-68] definiert worden sind. Die Ausführungen hier erfolgen in Anlehnung an [Col-03, S. 122 und S. 232] und [Kle-97, S. 329].

Gegeben seien die Realisierungen der identisch und unabhängig verteilten Zufallsvariablen $(T_1, \Delta_1, \mathbf{X}_1), \ldots, (T_n, \Delta_n, \mathbf{X}_n) \sim (T, \Delta, \mathbf{X})$, wobei mit T_i entweder eine Lebenszeit oder eine Zensur-Zeit bezeichnet wird, abhängig davon, ob für den beobachteten Wert des Zensurindikators $\delta_i = 1$ oder $\delta_i = 0$ gilt. Für das i-te Individuum seien $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})'$ die beobachteten Werte der p erklärenden Variablen $\mathbf{X} = (X_1, \ldots, X_p)'$. Wird für T ein AFT-Modell angenommen, so gilt nach Satz 6.2

$$\ln T \quad = \quad \beta_0 + \boldsymbol{\beta}'\mathbf{X} + bZ \tag{6.29}$$

mit unbekannten Parametern β_0, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ und b. Die Wahrscheinlichkeitsverteilung des Fehlers Z hängt von der für T angenommenen Verteilung ab. Für die in Abschnitt 6.2 betrachteten Modelle ist Z eine standardverteilte Zufallsvariable aus der Familie der Extremwert-, Logistik- oder Normal-Verteilungen. Nach Abschnitt 5.1.5 ist die geschätzte Survival-Funktion für das i-te Individuum dann durch

$$\hat{S}_T(t \mid \mathbf{x}_i) \quad = \quad S_Z\left(\frac{\ln t - \hat{\beta}_0 - \hat{\boldsymbol{\beta}}\mathbf{x}_i}{\hat{b}}\right), \quad t > 0 \tag{6.30}$$

gegeben, wobei S_Z die Survival-Funktion von Z ist und $\hat{\beta}_0$, $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)'$, $\hat{b}$ die ML-Schätzer der unbekannten Parameter.

Definition 6.2 (Cox-Snell-Residuen für das AFT-Modell). *Für das AFT-Modell sind Cox-Snell-Residuen definiert als*

$$r_i \quad = \quad \hat{\Lambda}_T(t_i \mid \mathbf{x}_i) \overset{(2.11)}{=} -\ln \hat{S}_T(t_i \mid \mathbf{x}_i), \quad i = 1, \ldots, n. \tag{6.31}$$

Dabei ist $\hat{S}_T(t_i \mid \mathbf{x}_i)$ die geschätzte, im Beobachtungszeitpunkt t_i ausgewertete Survival-Funktion (6.30) und $\hat{\Lambda}_T(t_i \mid \mathbf{x}_i)$ der entsprechende Schätzer für die kumulierte Hazard-Funktion.

Inwiefern Cox-Snell-Residuen für die Überprüfung eines Modells geeignet sind, beantwortet das folgende Lemma.

Lemma 6.1. *Sei T eine beliebige Lebensdauer mit Dichte f_T. Ist Λ_T die kumulierte Hazard-Funktion zu T, so ist die Zufallsvariable $Z := \Lambda_T(T)$ Exp(1)-verteilt.*

Beweis: Mit den Bezeichnungen aus dem Transformationssatz 5.1 gilt hier $h(t) = \Lambda_T(t) = -\ln S_T(t)$ und $g(z) = h^{-1}(z) = S_T^{-1}(\exp(-z))$. Die entsprechenden Ableitungen sind

$$h'(t) \;=\; -\frac{S_T'(t)}{S_T(t)} \;=\; \frac{f_T(t)}{S_T(t)}, \qquad g'(z) \;=\; \frac{1}{h'(g(z))} \;=\; \frac{S_T\big[S_T^{-1}(\exp(-z))\big]}{f_T\big[S_T^{-1}(\exp(-z))\big]}.$$

Es ist also

$$\begin{aligned} f_Z(z) &= f_T(g(z))\,|g'(z)| \\ &= f_T\big[S_T^{-1}(\exp(-z))\big]\,\frac{S_T\big[S_T^{-1}(\exp(-z))\big]}{f_T\big[S_T^{-1}(\exp(-z))\big]} \\ &= S_T\big[S_T^{-1}(\exp(-z))\big] \\ &= \exp(-z), \end{aligned}$$

und damit $Z \sim \text{Exp}(1)$. □

Ist T_i nun die Lebensdauer, $\mathbf{X}_i$ der Kovariablenvektor und $\Lambda(\cdot \mid \mathbf{X}_i)$ die kumulierte Hazard-Funktion des i-ten Individuums einer Beobachtungsgruppe, so folgt nach Lemma 6.1 $\Lambda(T_i \mid \mathbf{X}_i)$ einer Exponential-Verteilung mit Parameter $\lambda = 1$. Bei einem korrekt angepassten Modell sollten die Residuen $r_i = \hat{\Lambda}_T(t_i \mid \mathbf{x}_i)$ folglich einer zensierten Stichprobe mit zugrundeliegender Exp(1)-Verteilung entsprechen.

Betrachtet man die nach (6.31) berechneten Residuen $r_i,\ i = 1, \ldots, n$, als einen zensierten Datensatz, in dem ein Residuum r_i genau dann als zensiert angesehen wird, wenn die ihm zugrundeliegende Beobachtung t_i zensiert ist, und berechnet man für sie die geschätzte kumulierte Hazard-Funktion mit Hilfe des Kaplan-Meier-Schätzers $\hat{S}_{KM}(\cdot)$ aus Kapitel 4 nach

$$\hat{\Lambda}(\cdot) \;=\; -\ln \hat{S}_{KM}(\cdot), \tag{6.32}$$

so kann die graphische Bewertung über einen Plot von $\hat{\Lambda}(r_i)$ versus r_i erfolgen. Die Modellanpassung ist zufriedenstellend, falls der auf den Residuen r_i basierende Schätzer für Λ der kumulierten Hazard-Funktion $\Lambda_{\text{Exp}}(t) = t$ einer Exp(1)-Verteilung gleicht und die Punkte $(r_i, \hat{\Lambda}(r_i))$ damit nahe der ersten Winkelhalbierenden liegen.

Bemerkung 6.2. 1. Cox-Snell-Residuen eignen sich am besten für die Prüfung der Gesamtanpassung eines Modells. Eine Exp(1)-Verteilung der Residuen liegt vor, wenn für ihre Berechnung gemäß (6.31) in (6.30) die tatsächlichen Parameterwerte benutzt werden. Gebraucht man lediglich die Schätzer für β_0, $\boldsymbol{\beta}$ und b, so kann die Ursache einer Abweichung von der Exponential-Verteilung in den Ungenauigkeiten dieser Schätzer liegen. Umgekehrt besteht die Möglichkeit, dass der Plot der kumulierten Hazards für die Residuen auch dann einer Geraden mit Steigung 1 ähnelt, wenn die Anpassung des Modells nicht ausreichend ist.

2. Für die Beobachtungszeitpunkte $t_1, \ldots, t_n$ und die ML-Schätzer $\hat{\beta}_0$, $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)'$, $\hat{b}$ heißen

$$r_{S_i} = \frac{\ln t_i - \hat{\beta}_0 - \hat{\boldsymbol{\beta}}\mathbf{x}_i}{\hat{b}}, \quad i = 1, \ldots, n, \tag{6.33}$$

standardisierte Residuen. Bei einem zufriedenstellend angepassten AFT-Modell folgen sie gemäß (6.29) derselben Verteilung wie Z und können in ganz entprechender Weise zur Beurteilung der Güte einer Modellanpassung dienen. Man vergleiche hierzu [Col-03, S.231].

Beispiel 6.1 (Überprüfung der AFT-Modelle aus Abschnitt 6.2). In Abschnitt 6.2 sind drei AFT-Modelle an die Lebenszeitdaten von Krebspatientinnen angepasst worden. Für die logarithmierte Lebensdauer $\ln T$ ergaben sich dabei die folgenden Modelle:

$$\begin{array}{lrcll}
\text{Weibull:} & \ln T &=& 5.8544 - 0.9967\,X + 1.0668\,Z, & Z \sim \mathrm{EV}(0,1), \\
\text{Log-Logistik:} & \ln T &=& 5.4611 - 1.1491\,X + 0.8047\,Z, & Z \sim \mathit{Logist}(0,1), \\
\text{Log-Normal:} & \ln T &=& 5.4917 - 1.1512\,X + 1.3595\,Z, & Z \sim \mathcal{N}(0,1).
\end{array}$$

Berechnet man zunächst gemäß (6.33) die standardisierten Residuen r_{S_i}, so erhält man die Cox-Snell-Residuen r_i nach (6.30) und (6.31) wie folgt:

$$\begin{array}{lrcl}
\text{Weibull:} & r_i &=& \exp(r_{S_i}), \\
\text{Log-Logistik:} & r_i &=& \ln\big[1 + \exp(r_{S_i})\big], \\
\text{Log-Normal:} & r_i &=& -\ln\big[1 - \Phi(r_{S_i})\big].
\end{array}$$

Für die Residuen r_i, $i = 1, \ldots, 45$, wird der Schätzer für die kumulierte Hazard-Funktion über den Kaplan-Meier-Schätzer nach (6.32) berechnet und in r_i ausge-

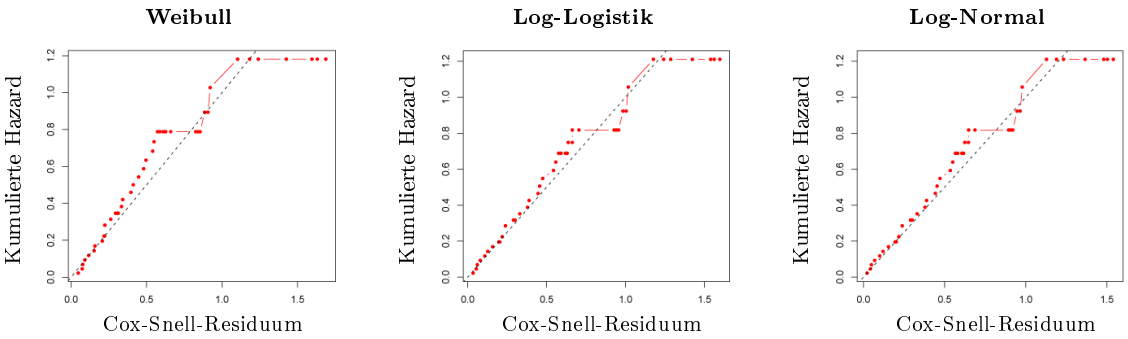

Abbildung 6.4: Überprüfung der AFT-Modelle aus Abschnitt 6.2 mit Hilfe von Cox-Snell-Residuen. Die erste Winkelhalbierende wird gestrichelt dargestellt.

wertet. Die Plots von $\hat{\Lambda}(r_i)$ versus r_i werden für die drei Modelle in Abbildung 6.4 dargestellt, vergleiche dazu den „R“-Quellcode in Anhang B.4.

Vernachlässigt man die Residuen zu den letzten Beobachtungen, so liegen die Punkte $(r_i, \hat{\Lambda}(r_i))$ bei allen betrachteten Modellen nahe einer Geraden, die durch den Ursprung geht und die Steigung eins hat. Weil die Schätzer zu späteren Zeitpunkten ohnehin auf weniger Individuen basieren und dadurch ungenauer sind, spricht nichts gegen die in Abschnitt 6.2 angepassten Modelle. Da zwischen den drei Plots jedoch eine große Ähnlichkeit besteht, kann keins der Modelle den anderen beiden klar vorgezogen werden.

Kapitel 7

Das Cox-Hazard-Modell

Das Cox-Hazard-Modell ist ein semiparametrisches Regressionsmodell, das bezüglich der Lebenszeit T an keinen speziellen Verteilungstyp gebunden ist. Parametrisch modelliert wird lediglich die Auswirkung der erklärenden Variablen. Das Modell ist von Cox [Cox-72] eingeführt worden und geht davon aus, dass sich Kovariablen direkt auf die Hazard-Rate eines Individuums auswirken.

Definition 7.1 (**Cox-Hazard-Modell**). *Sei $T \geq 0$ eine Lebensdauer und $\mathbf{X} = (X_1, \ldots, X_p)'$ ein p-dimensionaler Vektor von erklärenden Variablen. Das Cox-Hazard-Modell postuliert, dass die Form der Hazard-Funktion in Abhängigkeit von $\mathbf{X}$ durch*

$$\lambda(t \mid \mathbf{X}) \quad = \quad \lambda_0(t) \exp(\boldsymbol{\beta}'\mathbf{X}), \quad t \geq 0 \tag{7.1}$$

gegeben ist, wobei $\lambda_0(t) > 0$ für $t \geq 0$. Die nicht näher spezifizierte Funktion λ_0 heißt – entsprechend den Bezeichnungen im AFT-Modell – Basis-Hazard-Funktion (baseline hazard function) und ist die Hazard-Rate eines Individuums mit Kovariablenvektor $\mathbf{X} = \mathbf{0}$. $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)' \in \mathbb{R}^p$ ist ein Vektor von Regressionskoeffizienten.

Bemerkung 7.1.

1. Aufgrund der Beliebigkeit von λ_0 enthält die lineare Komponente des Modells keinen konstanten Term β_0. Wäre die Hazard-Funktion durch $\lambda(t \mid \mathbf{X}) = \lambda_0(t) \exp(\beta_0 + \boldsymbol{\beta}'\mathbf{X}),\ t \geq 0$ gegeben, so würde ein Teilen der Basis-Hazard-Funktion durch $\exp(\beta_0)$ zur Form (7.1) führen.

2. Das Cox-Hazard-Modell ist auch unter dem Namen Proportionale-Hazards-Modell (PH-Modell) bekannt. Diese Bezeichnung rührt daher, dass bei zeitunabhängigen Kovariablen die Hazard-Raten zweier Indidividuen stets proportio-

nal zueinander sind. Betrachtet man im Cox-Hazard-Modell zwei Individuen mit den erklärenden Variablen $\mathbf{X}$ und $\bar{\mathbf{X}}$, so ist das Verhältnis ihrer Hazard-Raten konstant: Für $t \geq 0$ ist

$$\frac{\lambda(t \mid \mathbf{X})}{\lambda(t \mid \bar{\mathbf{X}})} \;=\; \frac{\lambda_0(t)\,\exp(\boldsymbol{\beta}'\mathbf{X})}{\lambda_0(t)\,\exp(\boldsymbol{\beta}'\bar{\mathbf{X}})} \;=\; \exp(\boldsymbol{\beta}'(\mathbf{X}-\bar{\mathbf{X}})). \tag{7.2}$$

Das Risiko-Verhältnis (hazard ratio) (7.2) ist im Cox-Hazard-Modell also unabhängig von der Zeit.

3. Der Faktor $\exp(\boldsymbol{\beta}'\mathbf{X})$ im Modell (7.1) kann durch jede andere positive Funktion $r(\mathbf{X}) > 0$ ersetzt werden. Möglichkeiten für alternative Parametrisierungen von $r(\mathbf{X})$ wären $r(\mathbf{X}) = 1 + \boldsymbol{\beta}'\mathbf{X}$ und $r(X) = \log\,[1 + \exp(\boldsymbol{\beta}'\mathbf{X})]$, vergleiche [Cox-84, S. 91]. Cox [Cox-72] hat die log-lineare-Form in (7.1) benutzt. Sie ist bis heute am weitesten verbreitet und auch die folgenden Ausführungen sollen sich auf diese Spezifizierung beschränken.

Satz 7.1. *Im Cox-Hazard-Modell lässt sich die Survival-Funktion bei stetiger Lebenszeit darstellen als*

$$S(t \mid \mathbf{X}) \;=\; S_0(t)^{\exp(\boldsymbol{\beta}'\mathbf{X})}, \quad t \geq 0. \tag{7.3}$$

Beweis: Für stetige Lebensdauern gilt die Beziehung $S(t) = \exp[-\Lambda(t)]$. Daraus folgt

$$\begin{aligned} S(t \mid \mathbf{X}) \;&=\; \exp\Big(-\int_0^t \lambda_0(u)\,\exp(\boldsymbol{\beta}'\mathbf{X})\,du\Big) \;=\; \Big[\exp\Big(-\int_0^t \lambda_0(u)\,du\Big)\Big]^{\exp(\boldsymbol{\beta}'\mathbf{X})} \\ &=\; S_0(t)^{\exp(\boldsymbol{\beta}'\mathbf{X})}. \end{aligned}$$

□

7.1 Die partielle Likelihood-Funktion für bindungsfreie Datensätze

Die unbekannten Regressionsparameter $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)' \in \mathbb{R}^p$ im semiparametrischen PH-Modell (7.1) können mit der üblichen Maximum-Likelihood-Methode geschätzt werden. Cox [Cox-72] hat hierfür eine partielle Likelihood-Funktion vorgeschlagen, die wie eine klassische Likelihood-Funktion genutzt werden kann. Ba-

sierend auf seinen Ausführungen in [Cox-72] und [Cox-75] soll diese im Folgenden zunächst definiert und anschließend auf das PH-Modell (7.1) angewendet werden.

Definition 7.2 (Partielle Likelihood-Funktion). [Cox-75] [Kal-02, S. 99] *Sei* $\mathbf{y}$ *ein Beobachtungsvektor, der durch die Zufallsvariable* $\mathbf{Y}$ *mit Dichte* $f(\mathbf{y};\boldsymbol{\theta},\boldsymbol{\beta})$ *repräsentiert wird. Sei* $\boldsymbol{\beta}$ *der Vektor von interessierenden Parametern und* $\boldsymbol{\theta}$ *ein „Nuisance"-Parameter. Kann* $\mathbf{Y}$ *in eine Folge von Zufallsvariablen* $A_1, B_1, \ldots, A_m, B_m$ *transformiert werden, so dass die gemeinsame Dichte von* $A^{(m)} = (A_1, A_2, \ldots, A_m)$ *und* $B^{(m)} = (B_1, B_2, \ldots, B_m)$ *durch*

$$\prod_{j=1}^{m} f_{B_j|B^{(j-1)},A^{(j-1)}}(b_j \mid b^{(j-1)}, a^{(j-1)}; \boldsymbol{\beta}, \boldsymbol{\theta}) \prod_{j=1}^{m} f_{A_j|B^{(j)},A^{(j-1)}}(a_j \mid b^{(j)}, a^{(j-1)}; \boldsymbol{\beta}) \quad (7.4)$$

gegeben ist, dann heißt das zweite Produkt in (7.4) die partielle Likelihood-Funktion (partial likelihood) von $\boldsymbol{\beta}$*, basierend auf* $\{A_j\}$ *in der Folge* $\{A_j, B_j\}$*. Die Anzahl* m *der Terme kann dabei fest oder zufällig sein und die Zufallsvariablen* A_i*,* B_i*,* $i = 1, \ldots, m$*, können sowohl Zufallsgrößen als auch Zufallsvektoren sein.*

Bemerkung 7.2. Die auf der partiellen Likelihood-Funktion

$$L^P(\boldsymbol{\beta}) \;=\; \prod_{j=1}^{m} f_{A_j|B^{(j)},A^{(j-1)}}(a_j \mid b^{(j)}, a^{(j-1)}; \boldsymbol{\beta}) \quad (7.5)$$

basierende ML-Schätzung für $\boldsymbol{\beta}$ ist unter bestimmten Voraussetzungen an die erklärenden Variablen konsistent und asymptotisch normal-verteilt. Man vergleiche dazu [Tsi-81], [And-82] und [Kal-02, S. 99–101].

Zur Herleitung der partiellen Likelihood-Funktion für das Modell (7.1) betrachte man einen n-elementigen Datensatz von Überlebenszeitdaten als Realisierungen von $(Y_i, \Delta_i, \mathbf{X}_i)$, $i = 1, \ldots, n$. Dabei beschreibe $Y_i = \min(T_i, C_i)$ die Beobachtungszeit, T_i die wahre Lebensdauer und C_i die Zensur-Zeit des i-ten Individuums. Die erklärenden Variablen

$$\mathbf{X}_i \;=\; (X_{i1}, \ldots, X_{ip})', \quad i = 1, \ldots, n,$$

seien zeitunabhängig und ihre Werte $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})'$ zu Beginn der Studie aufgezeichnet. Insbesondere erfolge z.B. im Falle einer Studie zur Wirksamkeit verschiedener Therapien die Zuordnung eines Patienten zu einer Behandlungsgruppe zufällig. Weiter werde vorausgesetzt, dass bei gegebenem Kovariablenvektor die Ausfall- und

Zensur-Zeiten für das i-te Individuum unabhängig voneinander agieren. Ebenso seien die Ausfälle und Zensuren untereinander jeweils stochastisch unabhängig. Überdies werden weitere Bezeichnungen vereinbart:

Annahmen und Bezeichnungen 7.1.

- Die Hazard-Funktion des i-ten Individuums sei

$$\lambda(t \mid \mathbf{X}_i) = \lambda_0(t) \exp(\boldsymbol{\beta}' \mathbf{X}_i), \quad t \geq 0.$$

- $\mathbf{r}$ sei die Anzahl der beobachteten Ausfälle $(y_i, 1, \mathbf{x}_i) = (t_i, 1, \mathbf{x}_i)$, von denen in diesem Abschnitt angenommen wird, dass sie keine gleichen Messwerte enthalten, also frei von Bindungen (ties) sind. Liegt den Beobachtungen eine stetige Verteilung zugrunde und sind sie zudem exakt gemessen worden, so sind Bindungen theoretisch nicht möglich, insbesondere finden Ausfall und Zensur nicht gleichzeitig statt. Die Ausfallzeiten seien aufsteigend angeordnet:

$$t_{(1)} < \ldots < t_{(r)}.$$

- $\mathbf{n - r}$ ist dann die Anzahl der rechtszensierten Lebensdauern und

- $\mathbf{R(t_{(i)})}$, $i = 1, \ldots, r$, bezeichne die Risiko-Menge zum Zeitpunkt $t_{(i)}$, d.h. die Menge aller Individuen, die unmittelbar vor $t_{(i)}$ noch unter Beobachtung stehen.

Das grundlegende Argument bei der Konstruktion der partiellen Likelihood-Funktion für das Cox-Hazard-Modell (7.1) ist, dass die Zeitintervalle zwischen den aufeinanderfolgenden Ausfallzeiten $t_{(1)} < \ldots < t_{(r)}$ nur wenig über den Effekt aussagen, den die erklärenden Variablen auf die Hazard-Rate haben. Da die Basis-Hazard-Funktion λ_0 unbestimmt ist, ist es denkbar, dass sie auf den Intervallen, in denen keine Ausfälle stattfinden, nahe bei Null ist und dort folglich auch $\lambda(\cdot \mid \mathbf{X}) \approx 0$ gilt. Das widerum bedeutet, dass diese Intervalle nur wenig Information über die Regressionparameter $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ enthalten. Bei bindungsfreien Daten reicht es approximativ also aus, diejenigen Zeitpunkte $t_{(i)}$ zu betrachten, in denen Ausfälle tatsächlich beobachtet werden. Ist gemäß Definition 7.2

- $\mathbf{B_i}$ die Zufallsvariable, welche sowohl die Zensierung in dem Intervall $[t_{(i-1)}, t_{(i)})$ spezifiziert als auch die Information, dass *ein* Individuum in $t_{(i)}$ ausfällt und

- $\mathbf{A_i}$ die Zufallsvariable, die bestimmt, *welches* Individuum in $t_{(i)}$ ausfällt,

so entspricht der i-te Term der partiellen Likelihood-Funktion,

$$f_{A_i|B^{(i)},A^{(i-1)}}(a_i \mid b^{(i)}, a^{(i-1)}; \boldsymbol{\beta}), \tag{7.6}$$

der bedingten Wahrscheinlichkeit dafür, dass bei gegebener Risikomenge $R(t_{(i)})$ der Ausfall zum Zeitpunkt $t_{(i)}$ gerade bei dem beobachteten Individuum mit Kovariablenvektor $\mathbf{X}_{(i)}$ erfolgt.

Satz 7.2. *Die bedingte Wahrscheinlichkeit dafür, dass der Ausfall zum Zeitpunkt $t_{(i)}$ gerade bei dem beobachteten Individuum – d.h. dem mit Kovariablenvektor $\mathbf{X}_{(i)}$ – stattfindet, ist bei gegebener Risikomenge $R(t_{(i)})$ gegeben durch*

$$\frac{\exp\left(\boldsymbol{\beta}'\mathbf{X}_{(i)}\right)}{\sum_{j\in R(t_{(i)})}\exp\left(\boldsymbol{\beta}'\mathbf{X}_j\right)}. \tag{7.7}$$

Beweis: Sei $\Delta > 0$, dann gilt zunächst

$$\begin{aligned} &P\big[i \in R(t_{(i)}) \text{ mit } \mathbf{X}_{(i)} \text{ stirbt in } [t_{(i)}, t_{(i)}+\Delta) \mid \text{ein } j \in R(t_{(i)}) \text{ stirbt in } [t_{(i)}, t_{(i)}+\Delta)\big] \\ &\qquad = \frac{P\big[i \in R(t_{(i)}) \text{ mit } \mathbf{X}_{(i)} \text{ stirbt in } [t_{(i)}, t_{(i)}+\Delta)\big]}{P\big[\text{ein } j \in R(t_{(i)}) \text{ stirbt in } [t_{(i)}, t_{(i)}+\Delta)\big]}. \end{aligned}$$

Da angenommen wird, dass die Ausfallzeiten stetig verteilt sind, ist der Nenner des obigen Ausdrucks bei einem hinreichend klein gewählten Δ gerade die Summe über die Ausfallwahrscheinlichkeiten in $[t_{(i)}, t_{(i)}+\Delta)$ aller Individuen, die dann noch der Risikogruppe $R(t_{(i)})$ angehören. Das führt zu

$$\begin{aligned} &\frac{P\big[\, i \in R(t_{(i)}) \text{ mit } \mathbf{X}_{(i)} \text{ stirbt in } [t_{(i)}, t_{(i)}+\Delta)\big]}{P\big[\text{ein } j \in R(t_{(i)}) \text{ stirbt in } [t_{(i)}, t_{(i)}+\Delta)\,\big]} \\ &\qquad = \frac{P\big[\, i \in R(t_{(i)}) \text{ mit } \mathbf{X}_{(i)} \text{ stirbt in } [t_{(i)}, t_{(i)}+\Delta)\big]}{\sum_{j\in R(t_{(i)})} P\big[\, j \text{ stirbt in } [t_{(i)}, t_{(i)}+\Delta)\,\big]} \\ &\qquad = \frac{P\big[\, i \in R(t_{(i)}) \text{ mit } \mathbf{X}_{(i)} \text{ stirbt in } [t_{(i)}, t_{(i)}+\Delta)\big] \cap P\big[\, T_{\mathbf{X}_{(i)}} \geq t_{(i)}\,\big]}{\sum_{j\in R(t_{(i)})} \big\{P\big[\, j \text{ stirbt in } [t_{(i)}, t_{(i)}+\Delta)\,\big] \cap P\big[\, T_{\mathbf{X}_j} \geq t_{(i)}\,\big]\big\}}, \end{aligned}$$

wobei hier mit $T_{\mathbf{X}_{(i)}}$ die Lebensdauer des Individuums bezeichnet wird, das der Risikogruppe $R(t_{(i)})$ angehört und die erklärenden Variablen $\mathbf{X}_{(i)}$ besitzt. Entsprechend ist die Lebenszeit $T_{\mathbf{X}_j}$ definiert. Da $P(T_{\mathbf{X}_{(i)}} \geq t_{(i)}) = P(T_{\mathbf{X}_j} \geq t_{(i)})$ gilt, ist der letzte

Quotient identisch mit

$$\frac{\big(P\big[\, i \in R(t_{(i)}) \text{ mit } \mathbf{X}_{(i)} \text{ stirbt in } [t_{(i)}, t_{(i)}+\Delta)\big] \cap P\big[\, T_{\mathbf{X}_{(i)}} \geq t_{(i)} \,\big]\big) / P\big[\, T_{\mathbf{X}_{(i)}} \geq t_{(i)} \,\big]}{\sum_{j \in R(t_{(i)})} \big\{\big(P\big[\, j \text{ stirbt in } [t_{(i)}, t_{(i)}+\Delta)\,\big] \cap P\big[\, T_{\mathbf{X}_j} \geq t_{(i)} \,\big]\big) / P\big[\, T_{\mathbf{X}_j} \geq t_{(i)} \,\big]\big\}}$$
$$= \frac{P\big[\, i \in R(t_{(i)}) \text{ mit } \mathbf{X}_{(i)} \text{ stirbt in } [t_{(i)}, t_{(i)}+\Delta) \mid T_{\mathbf{X}_{(i)}} \geq t_{(i)} \,\big]}{\sum_{j \in R(t_{(i)})} P\big[\, j \text{ stirbt in } [t_{(i)}, t_{(i)}+\Delta) \mid T_{\mathbf{X}_j} \geq t_{(i)} \,\big]}.$$

Erweitern mit $1/\Delta$ liefert

$$\frac{P[\, i \in R(t_{(i)}) \text{ mit } \mathbf{X}_{(i)} \text{ stirbt in } [t_{(i)},\, t_{(i)}+\Delta) \mid T_{\mathbf{X}_{(i)}} \geq t_{(i)} \,]/\Delta}{\sum_{j \in R(t_{(i)})} P\big[\, j \text{ stirbt in } [t_{(i)}, t_{(i)}+\Delta) \mid T_{\mathbf{X}_j} \geq t_{(i)} \,\big]/\Delta}$$

und beim Grenzwertübergang $\Delta \to 0$ erhält man

$$\frac{\lim_{\Delta\to 0} P[\, i \in R(t_{(i)}) \text{ mit } \mathbf{X}_{(i)} \text{ stirbt in } [t_{(i)},\, t_{(i)}+\Delta) \mid T_{\mathbf{X}_{(i)}} \geq t_{(i)} \,]/\Delta}{\sum_{j \in R(t_{(i)})} \lim_{\Delta\to 0} P[\, j \text{ stirbt in } [t_{(i)},\, t_{(i)}+\Delta) \mid T_{\mathbf{X}_j} \geq t_{(i)} \,]/\Delta},$$

was nach Definition der Hazard-Rate gerade

$$\frac{\lambda(t_{(i)} \mid \mathbf{X}_{(i)})}{\sum_{j \in R(t_{(i)})} \lambda(t_{(i)} \mid \mathbf{X}_j)} = \frac{\lambda_0(t_{(i)}) \exp(\boldsymbol{\beta}'\mathbf{X}_{(i)})}{\sum_{j \in R(t_{(i)})} \lambda_0(t_{(i)}) \exp(\boldsymbol{\beta}'\mathbf{X}_j)} = \frac{\exp(\boldsymbol{\beta}'\mathbf{X}_{(i)})}{\sum_{j \in R(t_{(i)})} \exp(\boldsymbol{\beta}'\mathbf{X}_j)}$$

ist. □

Weil jeder Ausfall einen Faktor der Form (7.7) liefert, erhält man die partielle Likelihood-Funktion als Produkt über alle r Ausfallzeiten $t_{(1)} < \ldots < t_{(r)}$:

Satz 7.3 (**Partielle Likelihood-Funktion**). *Die Likelihood-Funktion für das Cox-Hazard-Modell (7.1) ist bei einem bindungsfreien Datensatz gegeben durch*

$$L^P(\boldsymbol{\beta}) = \prod_{i=1}^{r} \frac{\exp(\boldsymbol{\beta}'\mathbf{x}_{(i)})}{\sum_{j \in R(t_{(i)})} \exp(\boldsymbol{\beta}'\mathbf{x}_j)}. \tag{7.8}$$

Dabei ist r die Anzahl der beobachteten Ausfälle, $R(t_{(i)})$ die in 7.1 definierte Risikogruppe, $\mathbf{x}_{(i)}$ der Kovariablenvektor des Individuums, das zum Zeitpunkt $t_{(i)}$ ausfällt und $\mathbf{x}_j$ entsprechend der des j-ten Individuums in $R(t_{(i)})$.

Bei der Konstruktion der partiellen Likelihood-Funktion (7.8) werden die eigentlichen Zensur- und Ausfallzeiten nicht direkt benutzt. Von den Ausfallzeiten wird nur die Information über die Reihenfolge des Ausfalls verwendet und die zensierten Lebensdauern sind lediglich in der Summation über die Risikomengen zu den

Ausfallzeiten enthalten, die sich vor dem Zeitpunkt ihrer Zensierung ereignen. Die auf (7.8) basierende, partielle ML-Schätzung für $\boldsymbol{\beta}$ hat nach Bemerkung 7.2 die asymptotischen Eigenschaften einer klassischen ML-Schätzung. Man erhält sie in der gewohnten Weise durch Logarithmieren und Differenzieren von $L^P(\boldsymbol{\beta})$. Die logarithmierte partielle Likelihood-Funktion ist

$$\begin{aligned} l^P(\boldsymbol{\beta}) &= \ln L^P(\boldsymbol{\beta}) = \sum_{i=1}^{r} \Big[\boldsymbol{\beta}' \mathbf{x}_{(i)} - \ln \Big(\sum_{j \in R(t_{(i)})} [\exp(\boldsymbol{\beta}' \mathbf{x}_j)] \Big) \Big] \\ &= \sum_{i=1}^{r} \boldsymbol{\beta}' \mathbf{x}_{(i)} - \sum_{i=1}^{r} \ln \Big(\sum_{j \in R(t_{(i)})} [\exp(\boldsymbol{\beta}' \mathbf{x}_j)] \Big) \Big] \end{aligned} \tag{7.9}$$

und als erste partielle Ableitungen erhält man

$$\frac{\partial l^P(\boldsymbol{\beta})}{\partial \beta_h} = \sum_{i=1}^{r} x_{(i)h} - \sum_{i=1}^{r} \frac{\sum_{j \in R(t_{(i)})} \exp(\boldsymbol{\beta}' \mathbf{x}_j)\, x_{jh}}{\sum_{j \in R(t_{(i)})} \exp(\boldsymbol{\beta}' \mathbf{x}_j)}, \quad h = 1, \ldots, p. \tag{7.10}$$

Daraus ergibt sich die Informationsmatrix $I(\boldsymbol{\beta}) = \left(\frac{-\partial^2 l^P(\boldsymbol{\beta})}{\partial \beta_g \partial \beta_h} \right)_{\substack{g=1,\ldots,p \\ h=1,\ldots,p}}$ mit

$$\begin{aligned} \frac{-\partial^2 l^P(\boldsymbol{\beta})}{\partial \beta_g \partial \beta_h} &= \sum_{i=1}^{r} \frac{\sum_{j \in R(t_{(i)})} \exp(\boldsymbol{\beta}' \mathbf{x}_j)\, x_{jh}\, x_{jg}}{\sum_{j \in R(t_{(i)})} \exp(\boldsymbol{\beta}' \mathbf{x}_j)} \\ &\quad - \sum_{i=1}^{r} \frac{\Big(\sum_{j \in R(t_{(i)})} \exp(\boldsymbol{\beta}' \mathbf{x}_j)\, x_{jh} \Big) \Big(\sum_{j \in R(t_{(i)})} \exp(\boldsymbol{\beta}' \mathbf{x}_j)\, x_{jg} \Big)}{\Big(\sum_{j \in R(t_{(i)})} \exp(\boldsymbol{\beta}' \mathbf{x}_j) \Big)^2} \end{aligned} \tag{7.11}$$

Die partiellen ML-Schätzungen für die Regressionskoeffizienten findet man durch Lösen des nicht-linearen Gleichungssytems $\partial l^P(\boldsymbol{\beta}) / \partial \beta_h = 0,\ h = 1, \ldots, p$. Die gängigen Software-Pakete benutzen dazu das bereits in Abschnitt 6.1 erwähnte Newton-Raphson-Verfahren.

7.2 Die partielle Likelihood-Funktion für Datensätze mit mehrfachen Messwerten

Bei der Konstruktion der partiellen Likelihood-Funktion (7.8) ist angenommen worden, dass die gegebenen Daten auf einer stetigen Verteilung basieren und hinreichend genau gemessen worden sind, somit also keine Bindungen enthalten. In der Praxis können für die Ereignisse jedoch häufig nur Zeitintervalle angegeben werden. Die Be-

obachtungszeiten werden auf den Tag, den Monat oder das Jahr genau aufgezeichnet und der Rundungsprozess führt zu mehrfachen Messwerten.

Finden zu einem Zeitpunkt nicht nur mehrere Ausfälle statt, sondern gleichzeitig auch eine oder mehr Zensierungen, so wird angenommen, dass sich letztere erst nach sämtlichen Ausfällen ereignen. Eventuelle Unklarheiten darüber, welche Individuen zu einem Ausfallzeitpunkt in die Risikomenge aufgenommen werden sollen, werden dadurch beseitigt und es bleibt lediglich die Aufgabe, die partielle Likelihood-Funktion (7.8) an mehrfache Ausfallzeiten anzupassen. Cox [Cox-72] selbst konstruierte sie unter der Annahme einer echt diskreten Zeitskala. In der Diskussion zu seiner Veröffentlichung fand diese Vorgehensweise jedoch leichte Kritik, da das Vorkommen von Bindungen bei stetig verteilten Ereignissen lediglich eine schwache Gruppierung darstellt. Eine exakte Likelihood-Funktion für den Fall mehrfach auftretender Messwerte ist in [Kal-02, S. 105] zu finden. Diese hat allerdings eine sehr komplizierte Form und führt zudem zu einer sehr zeitaufwendigen Berechnung von $\hat{\boldsymbol{\beta}}$. In der Literatur finden sich mehrere Vorschläge für Approximationen, die gegenüber der exakten Methode rechnerische Vorteile haben. Die Ausführungen hier beschränken sich auf die Konstruktion von Peto [Pet-72], die bei nicht all zu vielen Bindungen eine recht gute Approximation liefert.

Wie im vorhergehenden Abschnitt sei $\mathbf{X}_i = (X_1, \ldots, X_p)'$ der Kovariablenvektor des i-ten Individuums und $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)' \in \mathbb{R}^p$ der interessierende Vektor von Regressionskoeffizienten. Entsprechend den Bezeichnungen in 7.1 wird weiter vereinbart:

Annahmen und Bezeichnungen 7.2.

- Die Hazard-Funktion des i-ten Individuums sei

$$\lambda(t \mid \mathbf{X}_i) = \lambda_0(t) \exp(\boldsymbol{\beta}'\mathbf{X}_i), \quad t \geq 0.$$

- $\mathbf{r}$ sei die Anzahl der verschiedenen Ausfallzeiten $t_{(1)} < \ldots < t_{(r)}$ und
- $\mathbf{d_i}$ die Anzahl der Ausfälle in $t_{(i)},\ i = 1, \ldots, r$.
- Man nehme an, dass dem Datensatz das zeitstetige Modell zugrunde liegt und Bindungen durch das Gruppieren von Messwerten entstehen. Die d_i Ausfälle zum Zeitpunkt $t_{(i)}$ sind folglich als verschieden und aufeinanderfolgend anzusehen. Wie zuvor sei weiter
- $\mathbf{R(t_{(i)})},\ i = 1, \ldots, r$, die Risiko-Menge zum Zeitpunkt $t_{(i)}$, d.h. die Menge aller Individuen, die unmittelbar vor $t_{(i)}$ noch unter Beobachtung stehen.

Satz 7.4. *Die bedingte Wahrscheinichkeit dafür, dass es sich bei den d_i Ausfällen zum Zeitpunkt $t_{(i)}$ gerade um die beobachteten handelt, ist bei gegebener Risikomenge $R(t_{(i)})$ approximativ gegeben durch*

$$\frac{d_i!\,\exp\left(\boldsymbol{\beta}'\mathbf{S}_i\right)}{\left[\sum_{j\in R(t_{(i)})}\exp\left(\boldsymbol{\beta}'\mathbf{X}_j\right)\right]^{d_i}}. \tag{7.12}$$

Dabei ist $\mathbf{S}_i$ die Summe der Kovariablenvektoren der d_i Individuen, die zum Zeitpunkt $t_{(i)}$ ausfallen.

Beweis: In Satz 7.2 ist gezeigt worden, dass im Fall eines einzigen Ausfalls, die Wahrscheinlichkeit dafür, dass es gerade der Ausfall des beobachteten Individuums mit Kovariablenvektor $\mathbf{X}_{(i)}$ ist, durch

$$\frac{\exp\left(\boldsymbol{\beta}'\mathbf{X}_{(i)}\right)}{\sum_{j\in R(t_{(i)})}\exp\left(\boldsymbol{\beta}'\mathbf{X}_j\right)}$$

gegeben ist. Um (7.12) entsprechend herzuleiten, soll zunächst der Fall zwei gleichzeitiger Ausfälle betrachtet werden. Seien also Ind_1 und Ind_2 mit den erklärenden Variablen $\mathbf{X}^1_{(i)}$ und $\mathbf{X}^2_{(i)}$ die beiden Individuen, deren Ausfall zum Zeitpunkt $t_{(i)}$ beobachtet wird. Im zeitstetigen Modell ist bei gegebener Risikomenge die bedingte Wahrscheinlichkeit dafür, dass zum Zeitpunkt $t_{(i)}$ gerade die Individuen Ind_1 und Ind_2 ausfallen, die Summe aus der bedingten Wahrscheinlichkeit dafür, dass zuerst Ind_1 ausfällt und dann Ind_2 und der bedingten Wahrscheinlichkeit dafür, dass Ind_2 als erstes ausfällt und Ind_1 als zweites:

$$\begin{aligned}\frac{\exp\left(\boldsymbol{\beta}'\mathbf{X}^1_{(i)}\right)}{\sum_{j\in R(t_{(i)})}\exp\left(\boldsymbol{\beta}'\mathbf{X}_j\right)}\,\frac{\exp\left(\boldsymbol{\beta}'\mathbf{X}^2_{(i)}\right)}{\sum_{j\in R(t_{(i)})}\exp\left(\boldsymbol{\beta}'\mathbf{X}_j\right)-\exp\left(\boldsymbol{\beta}'\mathbf{X}^1_{(i)}\right)}\\ +\ \frac{\exp\left(\boldsymbol{\beta}'\mathbf{X}^2_{(i)}\right)}{\sum_{j\in R(t_{(i)})}\exp\left(\boldsymbol{\beta}'\mathbf{X}_j\right)}\,\frac{\exp\left(\boldsymbol{\beta}'\mathbf{X}^1_{(i)}\right)}{\sum_{j\in R(t_{(i)})}\exp\left(\boldsymbol{\beta}'\mathbf{X}_j\right)-\exp\left(\boldsymbol{\beta}'\mathbf{X}^2_{(i)}\right)}\end{aligned} \tag{7.13}$$

Die beiden Summanden des Ausdrucks sind unter den gegebenen Voraussetzungen eine direkte Folgerung aus Satz 7.2. Bei gegebener Risikomenge $R(t_{(i)})$ ist die bedingte Wahrscheinlichkeit für einen Ausfall von Ind_1 durch

$$\frac{\exp\left(\boldsymbol{\beta}'\mathbf{X}^1_{(i)}\right)}{\sum_{j\in R(t_{(i)})}\exp\left(\boldsymbol{\beta}'\mathbf{X}_j\right)}$$

gegeben und

$$\frac{\exp\left(\boldsymbol{\beta}'\mathbf{X}_{(i)}^2\right)}{\sum_{j\in R(t_{(i)})}\exp\left(\boldsymbol{\beta}'\mathbf{X}_j\right)-\exp\left(\boldsymbol{\beta}'\mathbf{X}_{(i)}^1\right)}$$

ist dann die bedingte Wahrscheinlichkeit dafür, dass bei gegebener Risikomenge $R(t_{(i)})$ der Ausfall von Ind_2 nach dem von Ind_1 erfolgt, das zuerst ausgefallene Individuum Ind_1 ist aus der Menge $R(t_{(i)})$ zu entfernen. Der zweite Summand in (7.13) erklärt sich entsprechend. Der Ausdruck (7.13) kann unglücklicherweise nicht weiter vereinfacht werden und die partielle Log-Likelihood-Funktion wäre als Summe aus Logarithmen von Termen der Form (7.13) rechnerisch nur sehr schwer zu handhaben. Peto [Pet-72] schlägt daher eine Approximation vor, die im Wesentlichen darin besteht, die Summen in den Nennern der Brüche über alle Individuen der entsprechenden Risikogruppe zu bilden. Das führt im Fall von zwei gleichzeitigen Ausfällen zu

$$\frac{2!\,\exp\left(\boldsymbol{\beta}'\mathbf{X}_{(i)}^1\right)\,\exp\left(\boldsymbol{\beta}'\mathbf{X}_{(i)}^2\right)}{\left[\sum_{j\in R(t_{(i)})}\exp\left(\boldsymbol{\beta}'\mathbf{X}_j\right)\right]^2}.$$

Als Verallgemeinerung für d_i Ausfälle in $t_{(i)}$ ergibt sich

$$\frac{d_i!\,\exp\left(\boldsymbol{\beta}'\mathbf{X}_{(i)}^1\right)\ldots\exp\left(\boldsymbol{\beta}'\mathbf{X}_{(i)}^{d_i}\right)}{\left[\sum_{j\in R(t_{(i)})}\exp\left(\boldsymbol{\beta}'\mathbf{X}_j\right)\right]^{d_i}} \;=\; \frac{d_i!\,\exp\left(\boldsymbol{\beta}'\mathbf{S}_i\right)}{\left[\sum_{j\in R(t_{(i)})}\exp\left(\boldsymbol{\beta}'\mathbf{X}_j\right)\right]^{d_i}}$$

mit $\mathbf{S}_i$ als Summe der Kovariablenvektoren der d_i Individuen, die zum Zeitpunkt $t_{(i)}$ ausfallen. □

Die partielle Likelihood-Funktion erhält man erneut als Produkt der Faktoren (7.12) für die verschiedenen Ausfallzeitpunkte $t_{(1)}<\ldots<t_{(r)}$.

Satz 7.5 (Partielle Likelihood-Funktion). *Bei mehrfachen Messwerten ist die partielle Likelihood-Funktion gegeben durch*

$$L^P(\boldsymbol{\beta}) \;=\; \prod_{i=1}^{r}\frac{d_i!\,\exp(\boldsymbol{\beta}'\mathbf{s}_i)}{\left[\sum_{j\in R(t_{(i)})}\exp(\boldsymbol{\beta}'\mathbf{x}_j)\right]^{d_i}}. \tag{7.14}$$

Dabei ist r die Anzahl der verschiedenen Ausfallzeitpunkte und d_i die Anzahl der

Individuen die in $t_{(i)}$, $t_{(1)} < \ldots < t_{(r)}$, ausfallen. $R(t_{(i)})$ bezeichnet die in 7.2 definierte Risikogruppe, $\mathbf{x}_j$ den Kovariablenvektor des j-ten Individuums in $R(t_{(i)})$ und $\mathbf{s}_i$ ist die Summe der Kovariablenvektoren derjenigen Individuen, die zum Zeitpunkt $t_{(i)}$ ausfallen.

Bemerkung 7.3. Bis auf die Konstante $\prod_{i=1}^{r} d_i!$ entspricht (7.14) der von Breslow [Bre-74] vorgeschlagenen Approximation

$$L^P(\boldsymbol{\beta}) \;=\; \prod_{i=1}^{r} \frac{\exp(\boldsymbol{\beta}'\mathbf{s}_i)}{\left[\sum_{j\in R(t_{(i)})} \exp(\boldsymbol{\beta}'\mathbf{x}_j)\right]^{d_i}}. \tag{7.15}$$

Diese ist in den meisten statistischen Softwarepaketen implementiert. Sowohl (7.14) als auch (7.15) stimmen im Fall eines bindungsfreien Datensatzes mit der partiellen Likelihood-Funktion (7.8) überein.

7.3 Modellbildung

Die Strategie bei der Auswahl eines geeigneten PH-Modells ist wie bei gewöhnlicher Regression in gewisser Hinsicht vom Ziel der Studie abhängig. In Anwendungen kann es zum einen vorkommen, dass man über eine Vielzahl erklärender Variablen verfügt und die Aufgabe zunächst darin besteht herauszufinden, welche von ihnen Einfluss auf die Hazard-Rate und damit auf die Lebenszeit haben. Eine andere Situation ergibt sich, wenn das Interesse auf die Auswirkungen ganz bestimmter Kovariablen fokussiert ist, so zum Beispiel auf den Effekt einer neuen Therapie. Da das Ausmaß eines Therapieeffekts in der Regel von anderen Faktoren und Kovariaten beeinflusst wird, werden diese bei der Modellierung berücksichtigt und bei signifikantem Einfluss ins Modell integriert. Diese beiden grundlegenden Probleme der Modellbildung werden im Folgenden an zwei einfachen Beispielen exemplarisch erörtert.

Soll die Relevanz gegebener Kovariablen für das gesuchte Modell statistisch bewertet werden, so bietet sich das Verwenden des Likelihood-Quotienten-Tests an. Mit seiner Hilfe können Modelle, die einzelne Kovariablen oder Kombinationen von erklärenden Variablen enthalten, untereinander und mit dem Nullmodell, d.h. dem Modell ohne erklärende Variablen, verglichen werden.

Satz 7.6 (Likelihood-Quotienten-Test zum Prüfen einer einfachen Nullhypothese). *Seien $Y_1, \ldots, Y_n$ identisch und unabhängig verteilte Zufallsgrößen mit*

unbekannter Dichte $f_Y(y, \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \mathbb{R}^p$ *und sei* $\mathbf{y} = (y_1, \ldots, y_n)$ *der aus den Realisierungen dieser Zufallsvariablen bestehende Datensatz. Ist* $L_{(n)}(\cdot,\ \mathbf{y})$ *die auf diesen Daten basierende (partielle) Likelihood-Funktion und* $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_p)'$ *der (partielle) Maximum-Likelihood-Schätzer für* $\boldsymbol{\theta}$*, so wird der Likelihood-Quotienten-Test (LQ-Test) für*

$$H_0: \quad \boldsymbol{\theta} = \boldsymbol{\theta}_0 \qquad \textit{vs.} \qquad H_1: \quad \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$$

durch die folgende Prüfgröße definiert

$$T_n(\mathbf{y}) \quad = \quad \frac{L_{(n)}(\hat{\boldsymbol{\theta}},\ \mathbf{y})}{L_{(n)}(\boldsymbol{\theta}_0,\ \mathbf{y})}. \tag{7.16}$$

Es gilt (unter schwachen Regularitätsvoraussetzungen):

1. $2 \log T_n(\mathbf{y}) = -2\,[\log L_{(n)}(\boldsymbol{\theta_0},\ \mathbf{y}) - \log L_{(n)}(\hat{\boldsymbol{\theta}},\ \mathbf{y})\,] \xrightarrow{\mathcal{D}} \chi^2_p$
2. $\varphi_n(\mathbf{y}) = \mathbb{1}(2 \log T_n(\mathbf{y}) > \chi^2_{p;\alpha})$ *ist ein asymptotischer Test zum Niveau* α*.*

Beweis: Siehe zum Beispiel [Wit-95, Satz 6.47, S. 217f]. □

Satz 7.7 (Likelihood-Quotienten-Test zum Prüfen einer zusammengesetzten Nullhypothese). *Seien* $Y_1, \ldots, Y_n$ *identisch und unabhängig verteilte Zufallsgrößen mit unbekannter Dichte* $f_Y(y, \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \mathbb{R}^p$ *und sei* $\mathbf{y} = (y_1, \ldots, y_n)$ *der aus den Realisierungen dieser Zufallsvariablen bestehende Datensatz. Ist* $\boldsymbol{\theta} = (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2')'$ *mit* $\boldsymbol{\theta}_1' \in \mathbb{R}^q$ *und* $\boldsymbol{\theta}_2' \in \mathbb{R}^{p-q}$*, so ist der Likelihood-Quotienten-Test für*

$$H_0: \quad \boldsymbol{\theta}_1 = \boldsymbol{\theta}_{10} \qquad \textit{vs.} \qquad H_1: \quad \boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_{10}$$

durch die folgende Prüfgröße definiert

$$T_n(\mathbf{y}) \quad = \quad \frac{L_{(n)}(\hat{\boldsymbol{\theta}},\ \mathbf{y})}{L_{(n)}(\hat{\boldsymbol{\theta}}_2(\boldsymbol{\theta}_{10}),\ \mathbf{y})}. \tag{7.17}$$

Dabei bezeichnet $L_{(n)}(\cdot,\ \mathbf{y})$ *die auf* $\mathbf{y}$ *basierende (partielle) Likelihood-Funktion,* $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_p)'$ *den globalen (partiellen) Maximum-Likelihood-Schätzer für* $\boldsymbol{\theta}$ *und* $\hat{\boldsymbol{\theta}}_2(\boldsymbol{\theta}_{10})$ *den (partiellen) Maximum-Likelihood-Schätzer für* $\boldsymbol{\theta}$ *unter der Restriktion* $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_{10}$*. Es gilt (unter schwachen Regularitätsvoraussetzungen):*

1. $2 \log T_n(\mathbf{y}) = -2\,[\log L_{(n)}(\hat{\boldsymbol{\theta}}_2(\boldsymbol{\theta}_{10}),\ \mathbf{y}) - \log L_{(n)}(\hat{\boldsymbol{\theta}},\ \mathbf{y})\,] \xrightarrow{\mathcal{D}} \chi^2_q$

2. $\varphi_n(\mathbf{y}) = \mathbb{1}(2\log T_n(\mathbf{y}) > \chi^2_{q;\alpha})$ *ist ein asymptotischer Test zum Niveau* α*.*

Beweis: Siehe zum Beispiel [Wit-95, Satz 6.50, S. 220ff]. □

7.3.1 Identifikation von erklärenden Variablen mit Einfluss auf die Lebenszeit von Patienten mit Plasmozytom

Das Plasmozytom, auch als Kahler-Krankheit bekannt, ist eine bösartige Erkrankung, bei der es zu einer geschwulstartigen Wucherung des Knochenmarks kommt. Die Geschwulstzellen bilden abnorme Eiweißkörper, sogenannte Paraproteine, die im Blut auftauchen (Paraproteinanämie) und im Urin als Bence-Jones-Proteine ausgeschieden werden. Die fortschreitende Ausbreitung abnormer Plasmazellen führt zu Knochendefekten. Durch Verdrängung des blutbildenden Gewebes kommt es im Krankheitsverlauf häufig zur Anämie.[1]

Das im Folgenden behandelte Beispiel beruht auf den Daten in Anhang A, Tabelle A.3. Sie entstammen einer vom medizinischen Zentrum der West Virginia Universität durchgeführten Studie zur Lebensdauer von Patienten mit Plasmozytom. Das primäre Ziel dieser Studie war die Untersuchung des Einflusses bestimmter Kovariablen auf die Lebensdauer betroffener Patienten. Der vollständige Datensatz enthält Messungen von insgesamt 16 Kovariablen und ist samt statistischer Auswertung in [Kra-75] zu finden. Auf einem Teil des vollständigen Datensatzes basieren auch zahlreiche Beispiele in [Col-03]. Während die dort betrachteten Kovariablen mit denen in Tabelle A.3 übereinstimmen, werden in der Auswertung lediglich die Messungen von nur 48 der insgesamt 65 Patienten berücksichtigt. Da die Reduktion des Datensatzes weder begründet noch in ihrer Art präzisiert wird, sind die Ausführungen dieses Abschnitts zwar an die Beispiele 1.3, 3.2 und 3.5 in [Col-03] angelehnt, konkreten Berechnungen wird aber der Datensatz in Tabelle A.3 zugrundegelegt. Neben den beobachteten Lebenszeiten der 65 Patienten enthält dieser den Patienten-Status, d.h. den zugehörigen Wert des Zensurindikators, und die bei der Diagnose aufgezeichneten Messungen folgender Kovariablen:

[1] Eine genauere Beschreibung des Krankheitsverlaufs ist in [Psc-04, S. 1431f] zu finden.

X_1	(age)	Alter des Patienten in Jahren,
X_2	(sex)	Geschlecht des Patienten (0 = männlich, 1 = weiblich),
X_3	(bun)	Blut-Urea-Stickstoff (blood urea nitrogen),
X_4	(ca)	Serum-Calcium,
X_5	(hb)	Hämoglobin,
X_6	(pcells)	prozentualer Anteil von Plasmazellen im Knochenmark,
X_7	(protein)	Bence-Jones-Proteine im Urin (0 = nein, 1 = ja).

Aus den Kovariablen $X_1, \ldots, X_7$ werden nun diejenigen ausgewählt, die für die Lebensdauer von Plasmozytom-Patienten signifikant sind. Mögliche Interaktionen zwischen den Kovariablen werden dabei nicht berücksichtigt. Es wird außerdem angenommen, dass kein medizinischer Grund dafür besteht, bestimmte Kovariablen auf jeden Fall ins Modell aufzunehmen.

Ist die Anzahl der in Betracht zu ziehenden erklärenden Variablen nicht allzu groß, so können Modelle mit sämtlichen Variablenkombinationen betrachtet werden und mittels des Likelihood-Quotienten-Tests für einfache bzw. zusammengesetzte Hypothesen ein geeignetes unter ihnen ausgewählt werden. Da hier jedoch die Messwerte von sieben Kovariablen vorliegen, erhält man, selbst unter der Annahme, dass keine Interaktionen stattfinden, $2^7 = 128$ Kombinationsmöglichkeiten. Es wird daher ein stufenweises Verfahren benutzt, das in [Col-03, S. 83] beschrieben ist. Die dazu notwendigen Berechnungen werden mit der Statistik-Software „R“ durchgeführt. Der Quellcode ist an den entsprechenden Stellen Anhang B, Abschnitt B.5 zu entnehmen.

1. Der erste Schritt bei der Auswahl eines geeigneten Modells besteht in der Bestimmung von Kovariablen, die sich einzeln signifikant auf die Lebenszeit auswirken. Man betrachtet dazu die Modelle

$$\lambda(t \mid X_i) \quad = \quad \lambda_0(t)\, \exp(\beta_i X_i), \quad t \geq 0, \quad i = 1, \ldots, 7 \tag{7.18}$$

und vergleicht sie mit dem Nullmodell. Wird die Null-Hypothese des LQ-Tests für

$$H_0 : \quad \beta_i = 0 \qquad \text{vs.} \qquad H_1 : \quad \beta_i \neq 0 \tag{7.19}$$

signifikant verworfen, so kann davon ausgegangen werden, dass X_i Auswirkungen auf die Lebensdauer der Patienten hat. Tabelle 7.1 enthält die Werte von

$-2\log L(\hat{\beta}_i)$ $(i = 0, \ldots, 7)$ mit $\hat{\beta}_0 = 0$. Die geschätzten Regressionskoeffizienten in den Modellen mit X_3, X_4, X_5 und X_7 führen zu den niedrigsten Werten von $-2\log L^P(\hat{\beta}_i)$. Die Ergebnisse von Tests gemäß (7.19) mit $i = 3, 4, 5, 7$ werden in Tabelle 7.2 dargestellt. Signifikant abgelehnt werden die Nullhypothesen $H_0: \beta_3 = 0$ (zum Niveau 0.1%) und $H_0: \beta_5 = 0$ (zum Niveau 5%). Obwohl der ermittelte P-Wert zum Test von $H_0: \beta_7 = 0$ mit 0.1522 relativ groß ist, soll die Kovariable X_7 in Schritt zwei weiterhin betrachtet werden. Nicht verworfen werden kann die Nullhypothese $H_0: \beta_4 = 0$, folglich also auch nicht die Hypothesen $H_0: \beta_i = 0$, $i = 1, 2, 6$. Bei ihrer einzelnen Betrachtung zeigen also die Kovariablen X_3 (bun), X_5 (ca) und X_7 (protein) Auswirkungen auf die Lebenszeit der Patienten.

Tabelle 7.1: Werte von $-2\log L^P(\hat{\beta}_i)$ für Cox-Hazard-Modelle mit jeweils einer Kovariable X_i, basierend auf den Daten in Tabelle A.3.

Kovariable X_i im Modell	$-2\log L^P(\hat{\beta}_i)$
keine (Nullmodell)	309.9035
X_1 (age)	309.8883
X_2 (sex)	309.5511
X_3 (bun)	298.0386
X_4 (ca)	308.9772
X_5 (hb)	305.0127
X_6 (pcells)	309.0668
X_7 (protein)	307.8539

Tabelle 7.2: Ergebnisse der LQ-Tests zu den Modellen $\lambda(t) = \lambda_0(t)\exp(\beta_i X_i)$ für $H_0: \beta_i = 0$, $i = 3, 4, 5, 7$, basierend auf den Daten in Tabelle A.3.

Nullhypothese	$-2[\log L^P(\hat{\beta}_i) - \log L^P(0)]$	P-Wert
$H_0: \beta_3 = 0$	11.8649	0.0006
$H_0: \beta_4 = 0$	0.92634	0.3358
$H_0: \beta_5 = 0$	4.89084	0.0270
$H_0: \beta_7 = 0$	2.04957	0.1522

2. Im zweiten Schritt ist zu überprüfen, ob die Variablen, die gemäß vorangegangener Tests einzeln Auswirkungen auf die Lebenszeit gezeigt haben, diese in Kombination mit den anderen immer noch haben. Nach obigen Ergebnissen ist also das

Modell

$$\lambda(t \mid (X_3, X_5, X_7)) \;=\; \lambda_0(t)\,\exp(\beta_3 X_3 + \beta_5 X_5 + \beta_7 X_7), \quad t \geq 0 \tag{7.20}$$

zu betrachten und für $\boldsymbol{\beta} = (\beta_3, \beta_5, \beta_7)'$ die folgenden Hypothesen zu testen:

$$H_0: \quad \boldsymbol{\beta} = (\beta_3, \beta_5, 0)', \tag{7.21}$$

$$H_0: \quad \boldsymbol{\beta} = (\beta_3, 0, \beta_7)', \tag{7.22}$$

$$H_0: \quad \boldsymbol{\beta} = (0, \beta_5, \beta_7)'. \tag{7.23}$$

Im Modell einbehalten bleiben nur die Kovariablen X_i, für die $\beta_i = 0$ signifikant abgelehnt wird. Tabelle 7.3 zeigt die Ergebnisse der entsprechenden LQ-Tests. Während die Nullhypothesen (7.22) und (7.23) abgelehnt werden können, kann $H_0 : \boldsymbol{\beta} = (\beta_3, \beta_5, 0)'$ nicht verworfen werden. Folglich wird die Kovariable X_7 (protein) aus dem Modell entfernt.

Tabelle 7.3: Ergebnisse von LQ-Tests zum Modell $\lambda(t) = \lambda_0(t)\,\exp(\beta_3 X_3 + \beta_5 X_5 + \beta_7 X_7)$, basierend auf den Daten in Tabelle A.3. Es ist dabei $\boldsymbol{\beta} = (\beta_3, \beta_5, \beta_7)'$ und $\hat{\boldsymbol{\beta}}_{\mathbf{0}}$ der Schätzer für $\boldsymbol{\beta}$ unter H_0.

Nullhypothese		$-2[\log L^P(\hat{\boldsymbol{\beta}}_{\mathbf{0}}) - \log L^P(\hat{\boldsymbol{\beta}})]$	P-Wert
$H_0:$	$\boldsymbol{\beta} = (\beta_3, \beta_5, 0)'$	1.78470	0.1816
$H_0:$	$\boldsymbol{\beta} = (\beta_3, 0, \beta_7)'$	3.41584	0.0646
$H_0:$	$\boldsymbol{\beta} = (0, \beta_5, \beta_7)'$	12.1422	0.0005

3. Zuletzt bleibt zu untersuchen, ob die Variablen X_1, X_2, X_4 und X_6 in das Modell mit X_3 und X_5 aufzunehmen sind, denn obwohl sie bei der separaten Betrachtung in Schritt 1 als unwichtig eingestuft worden sind, kann ihr Effekt auf die Lebenszeit der Patienten in Gegenwart von X_3 und X_5 wesentlich sein. Für $t \geq 0$ sind zu den Modellen

$$\lambda(t \mid X_3, X_5, X_i) \;=\; \lambda_0(t)\,\exp(\beta_3 X_3 + \beta_5 X_5 + \beta_i X_i), \quad i = 1, 2, 4, 6 \tag{7.24}$$

die Hypothesen

$$H_0: \quad (\beta_3, \beta_5, \beta_i) = (\beta_3, \beta_5, 0), \quad i = 1, 2, 4, 6 \tag{7.25}$$

zu testen. Die Ergebnisse der entprechenden LQ-Tests werden Tabelle 7.4 dargestellt. Da keine der Nullhypothesen signifikant verworfen werden kann, ist zum Datensatz A.3 ein geeignetes Modell durch

$$\lambda(t) \quad = \quad \lambda_0(t)\,\exp(\beta_3 X_3 + \beta_5 X_5), \quad t \geq 0 \tag{7.26}$$

gegeben. Die Schätzer für die Regressionskoeffizienten sind dabei $\hat{\beta}_3 = 0.0192$ und $\hat{\beta}_5 = -0.1273$.

Tabelle 7.4: Ergebnisse der LQ-Tests zum Modell $\lambda(t) = \lambda_0(t)\,\exp(\beta_3 X_3 + \beta_5 X_5 + \beta_i X_i)$, $i = 1, 2, 4, 6$, basierend auf den Daten in Tabelle A.3. Es ist dabei $\hat{\boldsymbol{\beta}}_{\mathbf{0}}$ der Schätzer für $\boldsymbol{\beta}$ unter H_0.

Nullhypothese		$-2[\log L^P(\hat{\boldsymbol{\beta}}_{\mathbf{0}}) - \log L^P(\hat{\boldsymbol{\beta}})]$	P-Wert
$H_0:$	$(\beta_3, \beta_5, \beta_1) = (\beta_3, \beta_5, 0)$	0.69085	0.40587
$H_0:$	$(\beta_3, \beta_5, \beta_2) = (\beta_3, \beta_5, 0)$	1.05281	0.30486
$H_0:$	$(\beta_3, \beta_5, \beta_4) = (\beta_3, \beta_5, 0)$	2.16098	0.14155
$H_0:$	$(\beta_3, \beta_5, \beta_6) = (\beta_3, \beta_5, 0)$	1.24193	0.26510

Bemerkung 7.4. Die Basis-Hazard-Funktion λ_0 ist per Definition die Hazard-Funktion eines Individuums, dessen Kovariablen sämtlich den Wert Null haben. In diesem Beispiel ist sie also die Hazard-Rate eines 0-jährigen Mannes, bei dem im Urin keine Bence-Jones-Proteine gefunden worden sind und dessen Messwerte für bun, ca, hb und pcells alle null sind. Eine derart unrealistische Interpretation der Basis-Hazard-Rate kann umgangen werden, wenn von den Variablen X_1 (age), X_3 (bun), X_4 (ca), X_5 (hb) und X_6 (pcells) jeweils die Werte eines Durchschnittspatienten subtrahiert werden. Die Basis-Hazard-Funktion entspräche dann einem durchschnittlichen männlichen Patienten. Weil die Interpretation von λ_0 für das statistische Auswerten des Kovariablen-Effekts jedoch nicht relevant ist, ist auf eine derartige Transformation der erklärenden Variablen verzichtet worden.

7.3.2 Vergleich zweier Therapien bei Prostata-Krebs

Ein weiteres Modellierungsproblem besteht in der Auswahl geeigneter Kovariablen, wenn einige von ihnen von besonderem Interesse sind, man also bereits vor der Konstruktion eines geeigneten Modells das Überprüfen einer bestimmten Hypothese im Auge hat. Die Vorgehensweise in einer solchen Situation wird nun exemplarisch

für den Effekt einer Therapie bei Prostata Krebs dargestellt. Erforderliche Berechnungen werden auch hier mit „R“ durchgeführt, der Quellcode ist in Anhang B, Abschnitt B.6 zu finden.

Eine von der VACURG[2] durchgeführte und in [VAC-67] veröffentlichte Studie diente dem Vergleich von vier verschiedenen Therapien bei Prostata-Krebs. Die Daten wurden randomisiert erfasst und sind vollständig in [AnH-85, S. 261–274] zu finden. Dieses Beispiel soll sich lediglich auf den Effekt einer Behandlung mit 1,0mg Diethylstilbestrol (DES) beschränken. Die DES-Therapie wird dabei mit einer Placebo-Behandlung verglichen. Der betrachtete Datensatz wird weiter auf Patienten eingeschränkt, deren Tumor sich im sogenannten Stadium III befindet. Zusätzliche Auswahlkriterien sind ein normales EKG zu Studienbeginn, keine aktuellen oder vergangenen Herz-Kreislauf-Erkrankungen und keine verordnete Bettlägerigkeit. Neben der Therapie-Art werden hier die folgenden Kovariablen betrachtet:

X_1	(age)	Alter,
X_2	(shb)	Serum-Hämoglobin in gm/100 ml,
X_3	(size)	Größe des Tumors in cm^2,
X_4	(index)	Gleason-Index[3].

Die Messwerte dieser vier erklärenden Variablen sind bei Beobachtungsbeginn erfasst worden. Ihre Auflistung erfolgt neben den Lebensdauern der Patienten in Anhang A, Tabelle A.4.

Bevor der eigentliche Therapie-Effekt beurteilt werden kann, muss entschieden werden, welche der betrachteten Kovariablen Auswirkungen auf die Lebenszeit der Patienten haben. Dazu bleibt die Therapie-Art zunächst unberücksichtigt und für die Auswahl signifikanter Kovariablen verfährt man analog zum vorangegangenen Abschnitt. Da nun lediglich vier Variablen vorliegen, erhält man – unter der Annahme, dass zwischen ihnen keine Interaktionen bestehen – $2^4 = 16$ Kombinationsmöglichkeiten. Für jede dieser Möglichkeiten werden mit Hilfe der partiellen Likelihood-Funktion die Regressionskoeffizienten geschätzt und die Werte von $-2 \log L^P(\hat{\boldsymbol{\beta}})$ berechnet, vergleiche dazu Tabelle 7.5. Die Ergebnisse der im Folgenden erläuterten Tests sind Tabelle 7.6 zu entnehmen.

[2] Veteran's Administration Cooperative Urological Research Group

[3] Der Gleason-Index ist nach dem amerikanischen Arzt Dr. Donald Gleason benannt. Er basiert auf bestimmten Eigenschaften der Prostatakrebszellen und gibt Aufschluss über die Aggressivität des Tumors, je höher sein Wert, desto schneller die Vermehrung der Krebszellen.

Tabelle 7.5: Werte von $-2 \log L^P(\hat{\boldsymbol{\beta}})$, basierend auf den Daten in Tabelle A.4.

Kovariablen im Modell	$-2 \log L^P(\hat{\boldsymbol{\beta}})$
keine (Nullmodell)	36.34891
X_1 (age)	36.26942
X_2 (shb)	36.19577
X_3 (size)	29.04156
X_4 (index)	29.12706
X_1 (age) + X_2 (shb)	36.15064
X_1 (age) + X_3 (size)	28.85423
X_1 (age) + X_4 (index)	28.75962
X_2 (shb) + X_3 (size)	29.01906
X_2 (shb) + X_4 (index)	27.98141
X_3 (size) + X_4 (index)	23.53326
X_1 (age) + X_2 (shb) + X_3 (size)	28.85219
X_1 (age) + X_2 (shb) + X_4 (index)	27.89329
X_1 (age) + X_3 (size) + X_4 (index)	23.26935
X_2 (shb) + X_3 (size) + X_4 (index)	23.50802
X_1 (age) + X_2 (shb) + X_3 (size) + X_4 (index)	23.23089

Tabelle 7.6: Ergebnisse von LQ-Tests für Cox-Hazard-Modelle zum Datensatz aus Tabelle A.4. Die Cox-Hazard-Modelle werden durch ihre lineare Komponente angegeben. Je nach Nullhypothese und Kovariablenanzahl entspricht die LQ-Teststatistik T entweder $-2[\log L^P(\hat{\beta}_i) - \log L^P(0)]$ oder $-2[\log L^P(\hat{\boldsymbol{\beta}}_{\mathbf{0}}) - \log L^P(\hat{\boldsymbol{\beta}})]$. Im letzteren Fall ist $\hat{\boldsymbol{\beta}}_{\mathbf{0}}$ die Schätzung für $\boldsymbol{\beta}$ unter H_0.

Lineare Komponente	Nullhypothese H_0	T	P-Wert
$\beta_2 X_2$	$\beta_2 = 0$	0.15314	0.6956
$\beta_3 X_3$	$\beta_3 = 0$	7.30735	0.0069
$\beta_4 X_4$	$\beta_4 = 0$	7.22185	0.0072
$\beta_3 X_3 + \beta_4 X_4$	$\beta_4 = 0$	5.50830	0.0189
$\beta_3 X_3 + \beta_4 X_4$	$\beta_3 = 0$	5.59380	0.0180
$\beta_1 X_1 + \beta_3 X_3 + \beta_4 X_4$	$\beta_1 = 0$	0.26391	0.6074
$\beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$	$\beta_2 = 0$	0.02525	0.8738
$\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$	$(\beta_1, \beta_2) = (0, 0)$	0.30238	0.8597

Betrachtet man wie in Abschnitt 7.3.1 zunächst die Modelle mit jeweils einer erklärenden Variable, so zeigen LQ-Tests für die Nullhypothesen $H_0 : \beta_i = 0,\ i = 2, 3, 4$, dass bei getrennter Betrachtung nur die Kovariablen X_3 (size) und X_4 (index) signifikante Auswirkungen auf die Lebensdauer der Patienten haben. Da für das Modell

$$\lambda(t \mid X_3, X_4) \quad = \quad \lambda_0(t) \exp(\beta_3 X_3 + \beta_4 X_4), \quad t \geq 0 \tag{7.27}$$

jede der Hypothesen

$$H_0 : \quad (\beta_3, \beta_4) = (\beta_3, 0) \tag{7.28}$$

$$H_0 : \quad (\beta_3, \beta_4) = (0, \beta_4) \tag{7.29}$$

zum Niveau 5% signifikant verworfen werden kann, wirken sich die Kovariablen X_3 (size) und X_4 (index) auch in Kombination bedeutsam auf die Lebensdauer aus. Tabelle 7.5 ist zu entnehmen, dass ein Hinzufügen der Variablen X_1 (age) und X_2 (shb) zum Modell (7.27) nur eine geringe Reduktion von $-2 \log L^P(\hat{\boldsymbol{\beta}})$ bewirkt. Mit entprechenden LQ-Tests werden in den Modellen

$$\lambda(t \mid X_i, X_3, X_4) \quad = \quad \lambda_0(t) \exp(\beta_i X_i + \beta_3 X_3 + \beta_4 X_4),\ t \geq 0,\ i = 1, 2 \tag{7.30}$$

und

$$\lambda(t \mid X_1, X_2, X_3, X_4) \quad = \quad \lambda_0(t) \exp(\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4),\ t \geq 0 \tag{7.31}$$

die Nullhypothesen

$$H_0 : \quad (\beta_i, \beta_3, \beta_4) = (0, \beta_3, \beta_4), \quad i = 1, 2 \quad \text{bzw.} \tag{7.32}$$

$$H_0 : \quad (\beta_1, \beta_2, \beta_3, \beta_4) = (0, 0, \beta_3, \beta_4) \tag{7.33}$$

nicht verworfen (vergleiche Tabelle 7.6). Zum gegebenen Datensatz ist also bei Nichtberücksichtigung der Therapieart mit

$$\lambda(t \mid X_3, X_4) \quad = \quad \lambda_0(t) \exp(\beta_3 X_3 + \beta_4 X_4), \quad t \geq 0 \tag{7.34}$$

ein geeignetes Modell gefunden.

Für die nun durchzuführende Beurteilung des Therapieeffekts sollen die Kovariablen

X_3 und X_4 umbenannt werden:

$$\begin{aligned} X_3 &=: X_1^* \text{ (size)} \\ X_4 &=: X_2^* \text{ (index).} \end{aligned}$$

Weiter sei X_3^* die Indikator-Variable, die die Therapieart beschreibt und den Wert 0 annimmt, falls dem Patienten ein Placebo verabreicht wird, und 1, falls dieser DES erhält. Erweitert man das obige Modell mit Größe und Index um die Variable X_3^*, so erhält man für $-2\log L^P((\hat{\beta}_1^*, \hat{\beta}_2^*, \hat{\beta}_3^*))$ den Wert 22.5723. Das führt zu $-2\,[\log L^P((\hat{\beta}_1^*, \hat{\beta}_2^*, 0)) - \log L^P(\hat{\beta}_1^*, \hat{\beta}_2^*, \hat{\beta}_3^*)\,] = 0.96094$ mit P-Wert 0.3270, womit die Nullhypothese

$$H_0: \quad (\beta_1^*, \beta_2^*, \beta_3^*) = (\beta_1^*, \beta_2^*, 0)$$

nicht verworfen werden kann. Um die Vermutung, dass eine Behandlung mit DES ohne Effekt ist, zu untermauern, soll weiter überprüft werden, ob Interaktionen zwischen X_1^* und X_3^* bzw. zwischen X_2^* und X_3^* bestehen. Dazu werden die Produkte $X_{13}^* := X_1^* \times X_3^*$ und $X_{23}^* := X_2^* \times X_3^*$ dem linearen Teil des Cox-Hazard-Modells hinzugefügt

$$\begin{aligned} \lambda(t) &= \lambda_0(t)\,\exp(\beta_1^* X_1^* + \beta_2^* X_2^* + \beta_3^* X_3^* + \beta_{13}^* X_{13}^*), \quad t \geq 0 \\ \lambda(t) &= \lambda_0(t)\,\exp(\beta_1^* X_1^* + \beta_2^* X_2^* + \beta_3^* X_3^* + \beta_{23}^* X_{23}^*), \quad t \geq 0 \\ \lambda(t) &= \lambda_0(t)\,\exp(\beta_1^* X_1^* + \beta_2^* X_2^* + \beta_3^* X_3^* + \beta_{13}^* X_{13}^* + \beta_{23}^* X_{23}^*), \quad t \geq 0 \end{aligned}$$

und die Hypothesen

$$H_0: \quad (\beta_1^*, \beta_2^*, \beta_3^*, \beta_{13}^*) = (\beta_1^*, \beta_2^*, \beta_3^*, 0) \tag{7.35}$$

$$H_0: \quad (\beta_1^*, \beta_2^*, \beta_3^*, \beta_{23}^*) = (\beta_1^*, \beta_2^*, \beta_3^*, 0) \tag{7.36}$$

$$H_0: \quad (\beta_1^*, \beta_2^*, \beta_3^*, \beta_{13}^*, \beta_{23}^*) = (\beta_1^*, \beta_2^*, \beta_3^*, 0, 0) \tag{7.37}$$

getestet. Die Ergebnisse der Tests sind in Tabelle 7.7 aufgelistet und zeigen, dass keine der Nullhypothesen abgelehnt werden kann. Es besteht also kein Beweis dafür, dass der Effekt einer DES-Behandlung von der Tumorgröße oder dem Gleason-Index abhängt. Basierend auf dem gegebenen Datensatz stellt sich insgesamt heraus, dass die Verabreichung von 0,1mg DES keine signifikanten Auswirkungen auf die Lebens-

dauer der Patienten hat. Eine quantitative Betimmung des Therapie-Effekts erfolgt in Abschnitt 7.4.4.

Tabelle 7.7: Ergebnisse von LQ-Tests für Cox-Hazard-Modelle zum Datensatz aus Tabelle A.4. Die Cox-Hazard-Modelle werden durch ihre lineare Komponente angegeben. Es ist $T = -2[\log L^P(\hat{\boldsymbol{\beta}}_{\mathbf{0}}^{*}) - \log L^P(\hat{\boldsymbol{\beta}}^{*})]$, $\hat{\boldsymbol{\beta}}_{\mathbf{0}}^{*}$ die Schätzung für $\boldsymbol{\beta}^{*}$ unter H_0 und $X_{13}^* = X_1^* \times X_3^*$, sowie $X_{23}^* = X_2^* \times X_3^*$.

Lineare Komponente	Nullhypothese H_0	T	P-Wert
$\beta_1^* X_1^* + \beta_2^* X_2^* + \beta_3^* X_3^*$	$\beta_3^* = 0$	0.96094	0.3270
$\beta_1^* X_1^* + \beta_2^* X_2^* + \beta_3^* X_3^* + \beta_{13}^* X_{13}^*$	$\beta_{13}^* = 0$	1.74359	0.1867
$\beta_1^* X_1^* + \beta_2^* X_2^* + \beta_3^* X_3^* + \beta_{23}^* X_{23}^*$	$\beta_{23}^* = 0$	1.78059	0.1821
$\beta_1^* X_1^* + \beta_2^* X_2^* + \beta_3^* X_3^* + \beta_{13}^* X_{13}^* + \beta_{23}^* X_{23}^*$	$(\beta_{13}^*, \beta_{23}^*) = (0,0)$	2.86756	0.2384

7.4 Interpretation geschätzter Parameter

In der Einführung zu diesem Kapitel ist in Bemerkung 7.1 die Zeitunabhängigkeit des Risikoverhältnisses (hazard ratio)

$$\mathrm{HR}(t, \mathbf{x}, \bar{\mathbf{x}}) \;=\; \frac{\lambda(t \mid \mathbf{x})}{\lambda(t \mid \bar{\mathbf{x}})} \;=\; \exp(\boldsymbol{\beta}'(\mathbf{x} - \bar{\mathbf{x}})), \quad t \geq 0 \tag{7.38}$$

bei zwei (unterschiedlichen) Beobachtungen $\mathbf{x}$ und $\bar{\mathbf{x}}$ bereits als charakteristische Eigenschaft des Cox-Hazard-Modells angegeben worden. Die geschätzten Regressionsparameter können direkt über dieses Risikoverhältnis interpretiert werden. Im Folgenden wird zunächst auf den generellen Effekt von Kovariablen eingegangen, um anschließend auf das Beispiel aus Abschnitt 7.3.2 zurückzugreifen und das Ausmaß des dort beschriebenen Therapie-Effekts genauer zu analysieren. Hinweise zur Interpretation von Regressionskoeffizienten im Cox-Hazard-Modell sind in [Sac-06, S. 625–627] und [Col-03, Abschnitt 3.7] gegeben.

7.4.1 Stetige Kovariablen

Sind im Cox-Hazard-Modell stetige Kovariablen enthalten, so erfolgt die Interpretation der entsprechenden Regressionsparameter über konstante Intervalle. Handelt es sich bei der stetigen Variable X um die einzige erklärende Variable des Modells

und wird für ein Individuum der Wert $X = x$ beobachtet, so ist seine Hazard-Rate durch

$$\lambda(t \mid x) \quad = \quad \lambda_0(t) \exp(\beta x), \quad t \geq 0 \tag{7.39}$$

gegeben. Für das Hazard-Verhältnis $\mathrm{HR}(t, x + c, x)$ ergibt sich nach (7.38) folglich

$$\mathrm{HR}(t, x + c, x) \quad = \quad \exp(c\beta), \quad t \geq 0. \tag{7.40}$$

Verändert sich also der Wert der Kovariable X um c Einheiten, so führt das zu einer Veränderung des Risikoverhältnisses um $\exp(c\beta)$. Die geschätzte Änderung der Hazard-Ratio ist $\exp(c\hat{\beta})$ und die der logarithmierten Hazard-Ratio entsprechend $c\hat{\beta}$. Mit Hilfe des Standard-Fehlers $(\mathrm{Var}(c\hat{\beta}))^{1/2}$ von $c\hat{\beta}$ können über Konfidenzintervalle für $c\beta$ die für das Hazard-Verhältnis $\mathrm{HR}(t, x + c, x) = \exp(c\beta)$ konstruiert werden.

7.4.2 Faktoren

Faktoren sind Kovariablen, die nur eine endliche Menge an Werten annehmen können. Die Funktionswerte heißen dann Stufen, Stadien, Kategorien oder Levels des Faktors. Die Berücksichtigung von Faktoren ist über Linearkombinationen von Indikatorvariablen möglich. Betrachtet man beispielsweise einen Faktor A mit a Stufen, so können zur Modellierung $a - 1$ Indikatorvariablen in folgender Weise definiert werden:

Stufe von A	X_2	X_3	X_4	$\ldots$	X_a
1	0	0	0	$\ldots$	0
2	1	0	0	$\ldots$	0
3	0	1	0	$\ldots$	0
4	0	0	1	$\ldots$	0
$\vdots$	$\vdots$	$\vdots$		$\ddots$	
a	0	0	0	$\ldots$	1

Die Hazard-Funktion hat dann mit $\mathbf{X} = (X_2, \ldots, X_a)'$, $X_i = \mathbb{1}(A = i)$, die Form

$$\lambda(t \mid \mathbf{X}) \quad = \quad \lambda_0(t) \exp(\beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_a X_a), \quad t \geq 0 \tag{7.41}$$

und die Basis-Hazard-Funktion λ_0 ist die Hazard-Rate von Individuen, für die der Faktor A im ersten Level ist. Entsprechen die Stufen des Faktors bestimmten The-

rapien, so ist das relative Verhältnis der Hazard-Rate eines Patienten in Therapie i, $i \geq 2$, und der eines Individuums in Therapie 1, für $t \geq 0$ gegeben durch

$$\mathrm{HR}(t, \mathbf{e}_i, \mathbf{0}) \;=\; \frac{\lambda_0(t)\,\exp(\beta_i)}{\lambda_0(t)} \;=\; \exp(\beta_i), \tag{7.42}$$

wobei $\mathbf{e}_i \in \mathbb{R}^{a-1}$ hier den i-ten Einheitsvektor bezeichnet. Der geschätzte Parameter $\hat{\beta}_i$ entspricht damit der Schätzung für den Logarithmus des relativen Hazard-Verhältnisses zwischen den beiden Therapie-Gruppen.

7.4.3 Kombinationen von Kovariablen

Enthält das an einen Datensatz angepasste Cox-Hazard-Modell mehrere erklärende Variablen, so können die Regressionsparameter ebenfalls als Logarithmen von Hazard-Verhältnissen interpretiert werden. Bei der Betrachtung des geschätzten Effekts $\hat{\beta}_i$ von Variable X_i, müssen die übrigen Kovariablen des Modells jedoch berücksichtigt werden. Man spricht in diesem Zusammenhang von einem adjustierten, d.h. an die anderen Variablen des Modells angepassten Effekt (adjusted effect). Die Kovariablen können sich gegenseitig beeinflussen. Eine Interaktion besteht zum Beispiel dann, wenn das relative Risiko für zwei Stufen eines Faktors davon abhängig ist, in welchem Level sich ein anderer Faktor befindet.

Das Beipiel aus Abschnitt 7.3.2 (Vergleich zweier Therapien bei Prostata-Krebs) enthält sowohl Faktoren als auch stetige Kovariablen. In einer Fortsetzung soll die Interpretation geschätzter Parameter exemplarisch dargestellt werden. Neben dem Ausmaß des Therapie-Effekts werden dabei die Auswirkungen von Gleason-Index und Tumorgröße auf die Lebensdauer der Patienten bewertet.

7.4.4 Vergleich zweier Therapien bei Prostata-Krebs – Fortsetzung

In Abschnitt 7.3 wurde – basierend auf den Daten in Tabelle A.4 – gezeigt, dass die Größe des Tumors und der Wert des Gleason-Index' Auswirkungen auf die Lebensdauer von Patienten mit Prostata-Krebs haben. Der Effekt einer Behandlung mit 0,1mg DES schien dagegen unwesentlich zu sein. Da das Ziel der ursprünglichen Studie der Vergleich verschiedener Therapien war, soll die Variable, welche eine DES-Behandlung indiziert, weiter im Modell bleiben und der Therapie-Effekt

quantifiziert werden. Mit den erklärenden Variablen

X_1	(size)	Größe des Tumors in cm^2,
X_2	(index)	Gleason-Index,
X_3	(treatment)	Therapie (1 = DEF, 0 = Placebo)

wird das Modell

$$\lambda(t) \quad = \quad \lambda_0(t)\, \exp(\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3), \quad t \geq 0 \tag{7.43}$$

betrachtet und die Regressionparameter β_1, β_2 und β_3 geschätzt. Die Schätzungen und ihre Standard-Fehler sind Tabelle 7.8 zu entnehmen.

Tabelle 7.8: Geschätzte Regressionsparameter im Cox-Hazard-Modell $\lambda(t) = \lambda_0(t)\, \exp(\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3)$ zum Datensatz A.4.

Kovariable		$\hat{\beta}_i$	$\exp(\hat{\beta}_i)$	$\mathrm{se}(\hat{\beta})$
X_1	(size)	0.0826	1.0861	0.0475
X_2	(index)	0.7102	2.0345	0.3379
X_3	(treatment)	-1.1127	0.3287	1.2031

Bei gleicher Größe des Tumors und identischem Gleason-Index beträgt das geschätzte relative Verhältnis der Hazard-Rate eines Individuums der DES-Gruppe zu der von Patienten mit Placebo-Behandlung $\exp(\hat{\beta}_3) = 0.329$. Dieser an die Kovariablen X_1 (size) und X_2 (index) adjustierte Wert besagt, dass die Hazard-Rate eines Placebo-Patienten zu einem Zeitpunkt t etwa dreimal so hoch ist wie die eines DES-Patienten mit identischen Werten von Index und Tumor-Größe. Dass der Behandlungseffekt damit aber überschätzt wird, sieht man an einer entprechenden Schätzung für das Hazard-Verhältnis in einem Modell, das nur die Kovariable X_3 (treatment) enthält. In einem solchen Modell ist $\hat{\beta}_3 = -1.9780$ und die unadjustierte Schätzung für das Hazard-Verhältnis 0.1383.

Vergleicht man die Hazard-Raten von Individuen derselben Behandlungsgruppe und mit dem gleichen Index-Wert bezüglich ihrer Tumor-Größen miteinander, so ist $\exp(\hat{\beta}_1) = 1.0861$ das geschätzte Verhältnis der Hazard-Rate eines dieser Patienten relativ zu der Hazard-Rate eines anderen, dessen Tumor um eine Einheit kleiner ist. Nach Abschnitt 7.4.1 steigt die Hazard-Rate bei gleicher Therapie und gleichem Gleason-Index also mit der Größe des Tumors.

Entsprechend ist $\exp(\hat{\beta}_2) = 2.0344$ das geschätzte relative Verhältnis der Hazard-Rate eines Patienten zu der eines anderen, der identisch behandelt wird und die gleiche Tumor-Größe hat, dessen Index-Wert aber um eine Einheit kleiner ist. In der Gruppe von Patienten mit gleicher Therapie und gleicher Tumor-Größe, führt die Erhöhung des Indizes um eine Einheit zur Verdoppelung der Hazard-Rate.

7.5 Schätzen der Survival-Funktion

Bislang sind in diesem Kapitel lediglich die Schätzungen für die Regressionsparameter des Cox-Hazard-Modells behandelt worden. Für die Beurteilung des Effekts von erklärenden Variablen sind diese zwar völlig ausreichend, es besteht jedoch häufig auch das Interesse an der Überlebenswahrscheinlichkeit eines neuen Individuums mit Kovariablenvektor $\mathbf{X} = \mathbf{x}_0$. Sind die Regressionskoeffizienten geschätzt worden, so besteht die Konstruktion des Schätzers für

$$S(t \mid \mathbf{x}_0) \overset{Satz\ 7.1}{=} S_0(t)^{\exp(\boldsymbol{\beta}'\mathbf{x}_0)}, \quad t \geq 0 \tag{7.44}$$

lediglich in einer Schätzung der Basis-Survival-Funktion S_0. Aufgrund der für stetige Verteilungen gültigen Beziehung

$$S(t \mid \mathbf{X}) \;=\; \exp\big(-\Lambda(t \mid \mathbf{X})\big) \;=\; \exp\big(-\Lambda_0(t)\exp(\boldsymbol{\beta}'\mathbf{x})\big), \quad t \geq 0 \tag{7.45}$$

soll diese über einen Schätzer für die kumulierte Basis-Hazard-Funktion Λ_0 entwickelt werden, der darüberhinaus in Abschnitt 7.6 für die graphische Überprüfung eines Cox-Hazard-Modells benutzt werden wird. Die Herleitung von $\hat{\Lambda}_0$ basiert im Folgenden auf den Ausführungen von Klein und Moeschberger [Kle-97, S.236f].

Satz 7.8 (**Schätzer für die kumulierte Basis-Hazard-Funktion**). *Seien $X_1, \ldots, X_p$ die erklärenden Variablen in der linearen Komponente eines Cox-Hazard-Modells, $t_{(1)} < \ldots < t_{(r)}$ die verschiedenen Ausfallzeiten eines beobachteten Datensatzes und d_i die Anzahl der Ausfälle in $t_{(i)}$. Sind $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)'$ die partiellen ML-Schätzer für die Regressionskoeffizenten des Modells, so kann die kumulierte Basis-Hazard-Funktion durch*

$$\hat{\Lambda}_0(t) \;=\; \sum_{t_{(i)} \leq t} \frac{d_i}{\sum_{j \in R(t_{(i)})} \exp(\hat{\boldsymbol{\beta}}'\mathbf{X}_j)}, \quad t \geq 0 \tag{7.46}$$

geschätzt werden, wobei $\mathbf{X}_j$ *den Kovariablenvektor des* j*-ten Individuums bezeichnet.*

Beweis: Ist $\big(\,(y_1,\delta_1,\mathbf{x}_1),\ldots,(y_n,\delta_n,\mathbf{x}_n)\,\big)$ der auf $(Y_i,\Delta_i,\mathbf{X}_i)$ mit $Y_i=\min(T_i,C_i)$ und $\mathbf{X}_i=(X_1,\ldots,X_p)'$, $i=1,\ldots,n$, basierende Datensatz, T_i die Lebensdauer und C_i die Zensurzeit des i-ten Individuums, so ist nach Kapitel 3 die Likelihood-Funktion durch

$$\begin{aligned} L(\boldsymbol{\beta},\lambda_0(\cdot)) &= \prod_{i=1}^{n} f_T(y_i\mid\mathbf{x}_i)^{\delta_i} S_T(y_i\mid\mathbf{x}_i)^{1-\delta_i} \\ &\overset{\lambda(y)=f(y)/S(y)}{=} \prod_{i=1}^{n} \lambda_T(y_i\mid\mathbf{x}_i)^{\delta_i} S_T(y_i\mid\mathbf{x}_i) \end{aligned}$$

gegeben. Da für $y\geq 0$ im Cox-Hazard-Modell $\lambda(y\mid\mathbf{x})=\lambda_0(y)\exp(\boldsymbol{\beta}'\mathbf{x})$ (Definition 7.1) und $S(y\mid\mathbf{x})=S_0(y)^{\exp(\boldsymbol{\beta}'\mathbf{x})}$ (Satz 7.1) gilt, kann die Likelihood-Funktion umformuliert werden zu

$$L(\boldsymbol{\beta},\lambda_0(\cdot)) = \prod_{i=1}^{n}\big[\lambda_0(y_i)\exp(\boldsymbol{\beta}'\mathbf{x}_i)\big]^{\delta_i} S_0(y_i)^{\exp(\boldsymbol{\beta}'\mathbf{x}_i)}. \tag{7.47}$$

Man betrachte $\boldsymbol{\beta}$ von nun an als konstant und die Likelihood-Funktion (7.47) lediglich als Funktion in $\lambda_0(\cdot)$. Wegen der Beziehung (7.45) ist diese dann identisch mit

$$L(\lambda_0(\cdot)) = \prod_{i=1}^{n}\big[\lambda_0(y_i)\exp(\boldsymbol{\beta}'\mathbf{x}_i)\big]^{\delta_i}\exp\big[-\Lambda_0(y_i)\exp(\boldsymbol{\beta}'\mathbf{x}_i)\big]. \tag{7.48}$$

Will man (7.48) auf der Menge aller Hazard-Funktionen maximieren, so stellt man fest, dass es innerhalb der stetigen Hazard-Funktionen keinen ML-Schätzer geben kann, weil $L(\lambda_0(\cdot))$ dann beliebig groß werden kann. (Man vergleiche dazu das entsprechende Argument bei der Herleitung des Kaplan-Meier-Schätzers in Kapitel 4.) Es reicht also aus, (7.48) auf der Menge der diskreten Hazard-Funktionen zu maximieren, für die $\lambda_0(t)=0$ gilt, außer in Zeitpunkten zu denen Ausfälle $(y_i,\delta_i,\mathbf{x}_i)=(t_i,1,\mathbf{x}_i)$ beobachtet werden.

Seien $t_{(1)},\ldots,t_{(r)}$ die verschiedenen unter den beobachteten Ausfallzeiten und d_i die Anzahl der Ausfälle in $t_{(i)}$. Dann ist

$$\Lambda_0(t) = \sum_{t_{(i)}\leq t}\lambda_0(t_{(i)}),\quad t\geq 0 \tag{7.49}$$

und (7.48) entspricht

$$L(\lambda_0(\cdot)) \;=\; \prod_{i=1}^{r} \big[\lambda_0(t_{(i)}) \exp(\boldsymbol{\beta}'\mathbf{x}_i)\big]^{d_i} \prod_{j=1}^{n} \exp\big[-\Lambda_0(y_j)\exp(\boldsymbol{\beta}'\mathbf{x}_j)\big].$$

Der konstante Faktor $\prod_{i=1}^{r} \exp(\boldsymbol{\beta}'\mathbf{x}_i)^{d_i}$ kann aus der Likelihood-Funktion entfernt werden und mit (7.49) erhält man

$$\begin{aligned} L(\lambda_0(\cdot)) &= \left[\prod_{i=1}^{r} \lambda_0(t_{(i)})^{d_i}\right] \exp\left[-\sum_{j=1}^{n} \Lambda_0(y_j)\exp(\boldsymbol{\beta}'\mathbf{x}_j)\right] \\ &= \left[\prod_{i=1}^{r} \lambda_0(t_{(i)})^{d_i}\right] \exp\left[-\sum_{j=1}^{n} \left(\sum_{t_{(i)}\le y_j} \lambda_0(t_{(i)})\right) \exp(\boldsymbol{\beta}'\mathbf{x}_j)\right]. \end{aligned}$$

Wegen

$$\begin{aligned} &\sum_{j=1}^{n} \left(\sum_{t_{(i)}\le y_j} \lambda_0(t_{(i)})\right) \exp(\boldsymbol{\beta}'\mathbf{x}_j) \\ &\qquad = \sum_{t_{(i)}\le y_1} \lambda_0(t_{(i)}) \exp(\boldsymbol{\beta}'\mathbf{x}_1) + \ldots + \sum_{t_{(i)}\le y_n} \lambda_0(t_{(i)}) \exp(\boldsymbol{\beta}'\mathbf{x}_n) \\ &\qquad = \sum_{i=1}^{r} \left(\lambda_0(t_{(i)}) \sum_{j\in R(t_{(i)})} \exp(\boldsymbol{\beta}'\mathbf{x}_j)\right) \end{aligned}$$

ergibt sich

$$\begin{aligned} L(\lambda_0(\cdot)) &= \left[\prod_{i=1}^{r} \lambda_0(t_{(i)})^{d_i}\right] \exp\left[-\sum_{i=1}^{r} \left(\lambda_0(t_{(i)}) \sum_{j\in R(t_{(i)})} \exp(\boldsymbol{\beta}'\mathbf{x}_j)\right)\right] \\ &= \prod_{i=1}^{r} \left[\lambda_0(t_{(i)})^{d_i} \exp\left(-\lambda_0(t_{(i)}) \sum_{j\in R(t_{(i)})} \exp(\boldsymbol{\beta}'\mathbf{x}_j)\right)\right]. \end{aligned}$$

Sei $\lambda_{0i} := \lambda_0(t_{(i)}),\ i = 1, \ldots, r$. Dann ist die logarithmierte Likelihood-Funktion durch

$$l(\lambda_{01}, \ldots, \lambda_{0r}) \;=\; \sum_{i=1}^{r} d_i \ln \lambda_{0i} - \sum_{i=1}^{r} \left(\lambda_{0i} \sum_{j\in R(t_{(i)})} \exp(\boldsymbol{\beta}'\mathbf{x}_j)\right)$$

gegeben.

Als partielle Ableitungen erhält man für $i,\ h = 1, \ldots, r$:

$$\frac{\partial l}{\partial \lambda_{0i}} \;=\; \frac{d_i}{\lambda_{0i}} - \sum_{j \in R(t_{(i)})} \exp(\boldsymbol{\beta}' \mathbf{x}_j) \qquad \text{und} \qquad \frac{\partial^2 l}{\partial \lambda_{0h} \partial \lambda_{0i}} \;=\; \begin{cases} 0, & h \neq i, \\ -\frac{d_i}{\lambda_{0i}^2}, & h = i. \end{cases}$$

Das Lösen der Gleichungen $\frac{\partial l}{\partial \lambda_{0i}} = 0,\ i = 1, \ldots, r$, liefert

$$\left(\hat{\lambda}_{01}, \ldots, \hat{\lambda}_{0r}\right) \;=\; \left(\frac{d_1}{\sum_{j \in R(t_{(1)})} \exp(\boldsymbol{\beta}' \mathbf{x}_j)}, \ldots, \frac{d_r}{\sum_{j \in R(t_{(r)})} \exp(\boldsymbol{\beta}' \mathbf{x}_j)} \right). \quad (7.50)$$

Und da die Hesse-Matrix wegen $d_i \geq 1$ negativ definit ist, ist dieser stationäre Punkt der Likelihood-Funktion ein ML-Schätzer für $(\lambda_{01}, \ldots, \lambda_{0r})$. Mit (7.49) ergibt sich schließlich

$$\hat{\Lambda}_0(t) \;=\; \sum_{t_{(i)} \leq t} \frac{d_i}{\sum_{j \in R(t_{(i)})} \exp(\boldsymbol{\beta} \mathbf{x}_j)}, \quad t \geq 0,$$

was mit den nach Abschnitt 7.2 berechneten partiellen ML-Schätzern $\hat{\beta}_1, \ldots, \hat{\beta}_p$ der Behauptung (7.46) entspricht. □

Satz 7.9 (Schätzer der Survival-Funktion). *Seien* $\mathbf{X} = (X_1, \ldots, X_p)'$ *die erklärenden Variablen in der linearen Komponente eines Cox-Hazard-Modells,* $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)'$ *die partiellen ML-Schätzer für die Koeffizienten dieser Variablen und* $\hat{\Lambda}_0(\cdot)$ *der Schätzer (7.46). Dann ist eine Schätzung der Survival-Funktion zum Kovariablenvektor* $\mathbf{X} = \mathbf{x}_0$ *gegeben durch*

$$\hat{S}(t \mid \mathbf{x}_0) \;=\; \exp\left[- \hat{\Lambda}_0(t) \exp(\hat{\boldsymbol{\beta}}' \mathbf{x}_0) \right], \quad t \geq 0. \quad (7.51)$$

Beweis: Für $t \geq 0$ gilt $S(t \mid \mathbf{x}) = S_0(t)^{\exp(\boldsymbol{\beta}' \mathbf{x})}$ und $S_0(t) = \exp(-\Lambda_0(t))$, damit also

$$\hat{S}(t \mid \mathbf{x}_0) \;=\; \hat{S}_0(t)^{\exp(\hat{\boldsymbol{\beta}}' \mathbf{x}_0)} \;=\; \left(\exp\left[- \hat{\Lambda}_0(t) \right] \right)^{\exp(\hat{\boldsymbol{\beta}}' \mathbf{x}_0)} \;=\; \exp\left[- \hat{\Lambda}_0(t) \exp(\hat{\boldsymbol{\beta}}' \mathbf{x}_0) \right].$$

□

Beispiel 7.1 (Schätzung der Survival-Funktion für Patienten mit Plasmozytom). Zu den Lebenszeitdaten aus Tabelle A.3 ist in Abschnitt 7.3.1 ein geeignetes Cox-Hazard-Modell konstruiert worden. Einfluss auf die Hazard-Rate der Patienten haben dabei die Kovariablen X_3 (bun) und X_5 (hb) gezeigt, so dass die Hazard-Rate für ein Individuum mit Kovariablenvektor $\mathbf{x} = (x_3, x_5)'$ wie folgt ange-

geben werden konnte

$$\lambda(t \mid \mathbf{x}) \quad = \quad \lambda_0(t) \exp(0.0192\, x_3 - 0.1273\, x_5), \quad t \geq 0. \tag{7.52}$$

Für festgelegte Werte von (x_3, x_5) kann die Überlebenszeit mit der Statistik-Software „R“ nach Satz 7.9 geschätzt und graphisch dargestellt werden. Abbildung 7.1 stellt die Graphen der geschätzten Survival-Funktionen zu den Kovariablen-Werten $(15, 14)$, $(8, 14)$, $(15, 5)$ und $(160, 8)$ dar. Die geschätzte Basis-Survival-Funktion $\hat{S}_0(\cdot)$ wird zum Vergleich ebenfalls angegeben. Der verwendete „R“-Quellcode ist Anhang B, Abschnitt B.8 zu entnehmen.

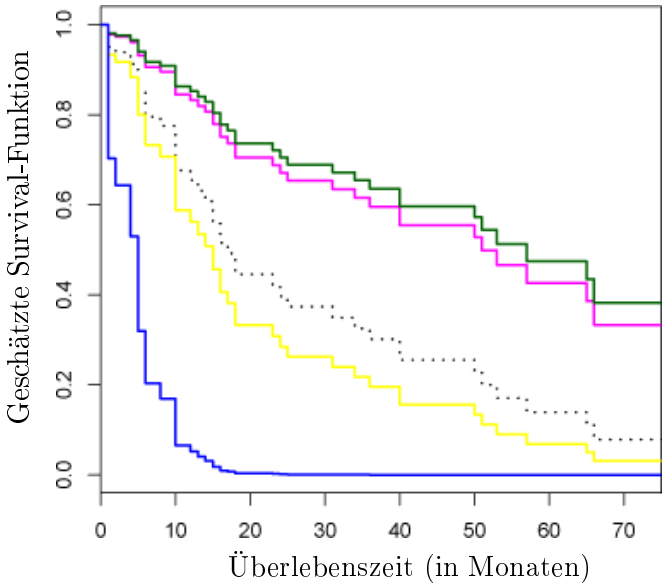

Abbildung 7.1: Geschätzte Survival-Funktionen von Patienten mit Plasmozytom und den folgenden Werten der Kovariablen: (bun, hb) = $(15, 14)$ (magenta), $(8, 14)$ (grün), $(15, 5)$ (gelb), $(160, 14)$ (blau). Die geschätzte Basis-Survival-Funktion ist gestrichelt dargestellt.

7.6 Überprüfung des Cox-Hazard-Modells

7.6.1 Güte der Modellanpassung – Cox-Snell-Residuen

In den Abschnitten 7.3 und 7.4 sind an zwei konkreten Beispielen Methoden zur Konstruktion und Auswertung eines geeigneten Cox-Hazard-Modells vorgestellt worden. Unter der Annahme von proportionalen Hazards sollen im Folgenden die bereits in Abschnitt 6.3 eingeführten Cox-Snell-Residuen dazu benutzt werden, die Güte der Modellanpassung zu bewerten. Auch für das Cox-Hazard-Modell sind diese definiert als die geschätzten Werte der kumulierten Hazard-Funktion in den Beobachtungszeiten $y_i,\ i = 1, \ldots, n$. Der Unterschied zum AFT-Modell besteht darin, dass für die Lebenszeitverteilung hier keine Annahmen gemacht werden und die kumulierte Hazard-Funktion daher durch den nicht-parametrischen Schätzer (7.46) approximiert werden muss. Im Zusammenhang mit dem PH-Modell werden Cox-Snell-Residuen in [Kle-97, S. 329] dargestellt.

Definition 7.3 (**Cox-Snell-Residuen für das Cox-Hazard-Modell**). *Sei eine n-elementige Stichprobe als Realisierung der Zufallsvariablen* $(Y_i, \Delta_i, \mathbf{X}_i),\ i = 1, \ldots, n$, *gegeben. Sei dabei in gewohnter Weise* $\mathbf{X}_i$ *der Kovariablenvektor und* $Y_i = \min(T_i, C_i)$ *die Beobachtungszeit des i-ten Individuums. Wird ein Cox-Hazard-Modell postuliert, so ist das Cox-Snell-Residuum für das i-te Individuum wie folgt definiert*

$$r_i \;=\; \exp\left(\hat{\boldsymbol{\beta}}'\mathbf{x}_i\right)\hat{\Lambda}_0(y_i), \quad i = 1, \ldots, n. \tag{7.53}$$

Dabei sind $\hat{\boldsymbol{\beta}} = \left(\hat{\beta}_1, \ldots, \hat{\beta}_p\right)'$ *die partiellen ML-Schätzer für die Regressionskoeffizienten des Modells und* $\hat{\Lambda}_0(\cdot)$ *der Schätzer aus Satz 7.8.*

Bemerkung 7.5. Für die Cox-Snell-Residuen (7.53) gilt

$$r_i \;=\; \hat{\Lambda}(y_i \mid \mathbf{x}_i) \;=\; -\ln \hat{S}(y_i \mid \mathbf{x}_i), \quad i = 1, \ldots, n.$$

Ist die Gesamtanpassung des Modells zufriedenstellend, so sollten die Residuen (7.53) nach Lemma 6.1 einer zensierten Stichprobe entsprechen, der eine Exp(1)-Verteilung zugrunde liegt. Wie im Zusammenhang mit AFT-Modellen, wird zur Überprüfung des Cox-Hazard-Modells die geschätzte kumulierte Hazard-Funktion basierend auf der Residuen-Menge $\{r_i = \exp\left(\hat{\boldsymbol{\beta}}'\mathbf{x}_i\right)\hat{\Lambda}_0(y_i) \mid i = 1, \ldots, n\}$ über den

Kaplan-Meier-Schätzer $\hat{S}_{KM}(\cdot)$ nach

$$\hat{\Lambda}(\cdot) \quad = \quad -\ln \hat{S}_{KM}(\cdot) \tag{7.54}$$

berechnet. Erhält man bei einem Plot von $\hat{\Lambda}(r_i)$ versus r_i Punkte, die nahe der ersten Winkelhalbierenden liegen, so kann die Anpassung des Modells als gut bewertet werden. Man vergleiche hierzu die begründenden Erläuterungen auf Seite 82.

Beispiel 7.2 (Überprüfung des Modells aus Abschnitt 7.3.1). In Abschnitt 7.3.1 ist festgestellt worden, dass für den Datensatz A.3, die Überlebenszeiten von Patienten mit Plasmozytom, ein geeignetes Cox-Hazard-Modell durch

$$\lambda(t \mid \mathbf{X}) \quad = \quad \lambda_0(t) \exp(0.0192\, X_3 - 0.1273\, X_5), \quad t \geq 0 \tag{7.55}$$

gegeben ist. Die erklärenden Variablen X_3 und X_5 beschreiben dabei den Gehalt an Blut-Urea-Stickstoff (bun) bzw. Hämoglobin (hb). Ob diese Modellanpassung angemessen ist, soll nun mit Hilfe von Cox-Snell-Residuen beurteilt werden. Abbildung 7.2 enthält den Plot der geschätzten kumulierten Hazards in den nach (7.53) berechneten Cox-Snell-Residuen. Die Schätzung für die kumulierte Hazard-Funktion ist gemäß (7.54) über den Kaplan-Meier-Schätzer erfolgt. Der für die Berechnungen verwendete „R"-Quellcode ist in Anhang B.9 zu finden. Die durch die Punkte $(r_i, \hat{\Lambda}(r_i))$ angedeutete Line in Abbildung 7.2 geht durch den Ursprung und hat ziemlich genau die Steigung eins. Von diesem Plot ausgehend spricht also nichts gegen das Modell (7.55).

Bemerkung 7.6. Wie bereits im Zusammenhang mit AFT-Modellen erwähnt, eignen sich Cox-Snell-Residuen am besten für die Prüfung der Gesamtanpassung eines Modells. Die in Bemerkung 6.2 angesprochene Gefahr der Fehlinterpretation von Plots der kumulierten Hazard-Rate in den Residuen ist hier durch das Benutzen eines nicht-parametrischen Schätzers für Λ_0 sogar noch größer.

7.6.2 Prüfen der Proportional-Hazards-Annahme

Die grundlegende Voraussetzung für die Gültigkeit des Cox'schen Regressionsmodells ist, dass die Hazard-Raten zweier Individuen mit verschiedenen Kovariablenwerten proportional zueinander sind. Abschließend soll daher eine Methode vorgestellt werden, mit der die Annahme der proportionalen Hazards graphisch überprüft

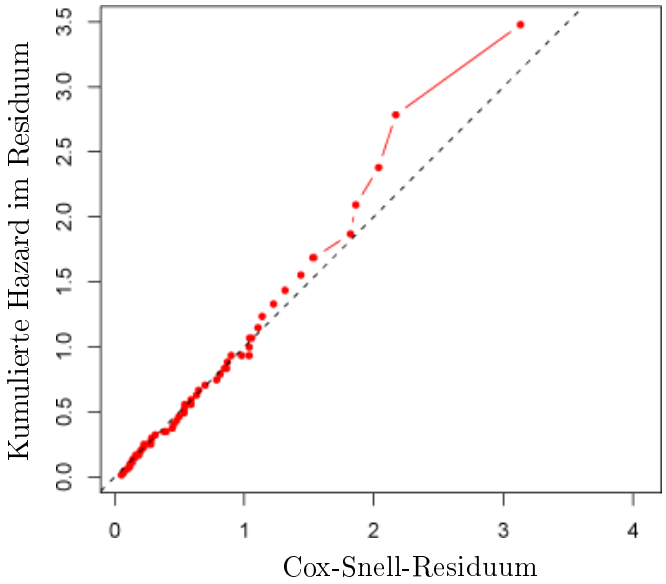

Abbildung 7.2: Überprüfung des Modells aus Abschnitt 7.3.1 mit Hilfe von Cox-Snell-Residuen. Die erste Winkelhalbierende wird gestrichelt dargestellt.

werden kann. Im Cox-Hazard-Modell ist die Hazard-Funktion eines Individuums mit Kovariablen-Vektor $\mathbf{X} = \mathbf{x}$ durch

$$\lambda(t \mid \mathbf{x}) \quad = \quad \lambda_0(t)\, \exp(\boldsymbol{\beta}'\mathbf{x}), \quad t \geq 0 \tag{7.56}$$

gegeben. Integrieren und anschließendes Logarithmieren der beiden Seiten dieser Gleichung führt zu

$$\begin{aligned} & \int_0^t \lambda(u \mid \mathbf{x})\, du \;=\; \exp(\boldsymbol{\beta}'\mathbf{x}) \int_0^t \lambda_0(t), \quad t \geq 0 \\ \iff \quad & \Lambda(t \mid \mathbf{x}) \;=\; \exp(\boldsymbol{\beta}'\mathbf{x})\, \Lambda_0(t), \quad t \geq 0 \\ \iff \quad & \ln \Lambda(t \mid \mathbf{x}) \;=\; \boldsymbol{\beta}'\mathbf{x} + \ln \Lambda_0(t), \quad t \geq 0. \end{aligned}$$

Bei einem Plot der log-kumulierten Hazard-Funktionen $\ln \Lambda(\cdot \mid \mathbf{x})$ für Individuen mit verschiedenen Werten ihrer erklärenden Variablen $\mathbf{x}$ gegen die Zeit, sollte – falls die Annahme der proportionalen Hazards korrekt ist – der Abstand zwischen den resultierenden Graphen folglich konstant sein.

Für die Prüfung der Proportional-Hazards-Annahme bei einem konkreten Datensatz müssen die beobachteten Werte stetiger Kovariablen zunächst klassifiziert werden. Die Ausfall- und Zensurzeiten $(y_i, \delta_i, \mathbf{x}_i)$ der Individuen werden dann bezüglich dieser Klassen und der Level von Faktoren gruppiert. Innerhalb jeder Gruppe $g = 1, \ldots, G$ wird über den Kaplan-Meier-Schätzer der Schätzer für die kumulierte

Hazard-Funktion berechnet und in den Beobachtungszeiten y_i ausgewertet:

$$\hat{\Lambda}_g(y_i) \quad = \quad -\ln \hat{S}_g(y_i), \quad g = 1, \ldots, G.$$

Über die Plots von $\hat{\Lambda}_g(y_i)$ versus y_i kann dann nach obiger Argumentation bewertet werden, inwiefern proportionale Hazard-Raten angenommen werden können. Insbesondere besteht die Möglichkeit, die Proportional-Hazards-Annahme für einzelne Kovariablen zu überprüfen.

Beispiel 7.3 (Überprüfung der PH-Annahme für den Datensatz A.3). Es soll erneut auf den Datensatz A.3 zugegriffen und geprüft werden, ob die Hazard-Raten von Patienten mit Plasmozytom bezüglich der Kovariable X_5 (hb, Hämoglobin) proportional zueinander sind. Weil es sich hierbei um eine stetige Kovariable handelt, sind die Patienten je nach beobachtetem Hämoglobin-Wert in vier Gruppen zu unterteilen: (i) hb ≤ 7, (ii) $7 <$ hb ≤ 10, (iii) $10 <$ hb ≤ 13 und (iv) hb ≥ 13. Mit Hilfe des „R"-Quellcodes aus Anhang B.10 kann für jede dieser Gruppen der Kaplan-Meier-Schätzer $\hat{S}(\cdot)$ berechnet werden. In Abbildung 7.3 wird der daraus resultierende Schätzer für die log-kumulierte Hazard-Funktion graphisch dargestellt. Neben dem Plot von $\ln \hat{\Lambda}(y_i)$ gegen die tatsächliche Beobachtungszeit y_i (links), wird zur besseren Beurteilung der Proportionalität $\ln \hat{\Lambda}(y_i)$ auch gegen $\ln y_i$ aufgetragen (rechts).

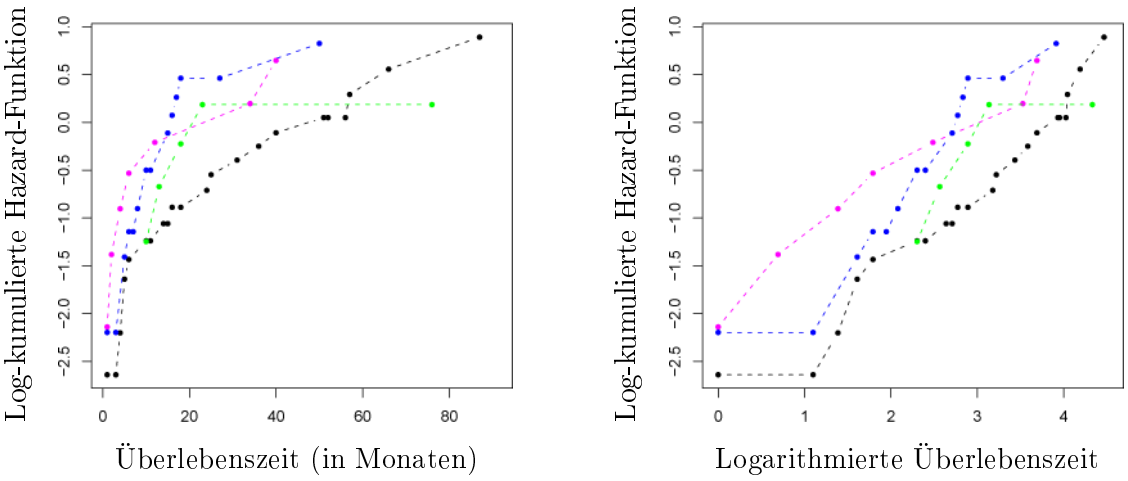

Abbildung 7.3: Überprüfung der Proportional-Hazards-Annahme für die Kovariable X_5 (hb) aus Datensatz A.3. Die Darstellung der log-kumulierten Hazard-Funktion erfolgt für die Gruppen hb ≤ 7 (magenta), $7 <$ hb ≤ 10 (blau), $10 <$ hb ≤ 13 (schwarz) und hb ≥ 13 (grün).

Während die log-kumulierten Hazard-Raten bei niedrigen Ausfall- und Zensur-Zeiten noch einen annähernd konstanten Abstand haben, schneiden sich die Graphen bei höheren Beobachtungszeiten sogar. Vor allem die log-kumulierte Hazard-Rate zu $10 < hb \leq 13$ ist für hohe Zeiten auffallend steil. Da die dem Plot zugrundeliegenden Schätzer bei späteren Zeitpunkten nur noch auf wenigen Beobachtungen basieren, darf diesbezüglich jedoch nicht überinterpretiert werden. Ferner werden die Werte der erklärenden Variable X_3 (bun, Blut-Urea-Stickstoff) nicht in Betracht gezogen. Es ist aber denkbar, dass gerade die Lebenszeiten der Individuen in der dritten Hämoglobin-Gruppe von ihren bun-Werten beeinflusst worden sind. Insgesamt sollte an der Proportionalität der Hazard-Raten dennoch gezweifelt werden.

Kapitel 8

Abschließende Bemerkungen

Das Ziel der vorliegenden Arbeit ist es gewesen, grundlegende Fragestellungen und Methoden der Lebenszeitanalyse darzustellen. Nach Einführung der fundamentalen Größen und Funktionen dieses Gebietes, sowie der ausführlichen Abhandlung der verschiedenen Zensurmechanismen ist mit der Konstruktion der Likelihood-Funktion für rechtszensierte Daten in Abschnitt 3.2 eine Basis für die statistische Auswertung von Lebenszeitdaten geschaffen worden. Im Anschluss daran sind sowohl parametrische als auch nicht-parametrische Verfahren zur Beschreibung der Lebensdauer von Individuen homogener Populationen vorgestellt worden. Um den Zusammenhang zwischen Überlebenszeiten und möglicherweise prognostisch relevanten Faktoren analysieren zu können, sind mit dem AFT-Modell in Kapitel 6 und dem Cox-Hazard-Modell in Kapitel 7 zwei Regressionsansätze für die Auswertung zensierter Daten diskutiert worden. Mit Hilfe von Cox-Snell-Residuen konnte für beide Methoden ein graphisches Verfahren zur Beurteilung der Gesamtanpassung eines Modells beschrieben werden. Im Zusammenhang mit dem Cox-Hazard-Modell sind anhand von konkreten Beispielen Aspekte der Konstruktion und Auswertung eines Modells erörtert worden.

Das Cox-Hazard-Modell hat gegenüber dem AFT-Modell den Vorteil, dass es bezüglich der Lebenszeit T an keine spezielle Verteilung gebunden ist und nur die Auswirkungen erklärender Variablen parametrisch modelliert. Andererseits liefern parametrische Modelle, falls sie korrekt ausgewählt werden, in der Regel bessere Schätzer für die interessierenden Größen, wie zum Beispiel den Median einer Überlebenszeit.

Aufgrund des sehr umfangreichen Themenbereichs konnte im Rahmen dieser Arbeit

nur auf einzelne Aspekte der Lebenszeitanalyse eingegangen werden. So sind beispielsweise im Zusammenhang mit Regressionsmodellen ausschließlich Auswirkungen zeitunabhängiger Kovariablen behandelt worden. In Lebensdauerstudien werden die Werte erklärender Variablen, wie zum Beispiel die Größe eines Tumors, jedoch häufig nicht nur zu Studienbeginn aufgezeichnet, sondern in regelmäßigen Abständen mehrfach gemessen. Die Modellierung zeitabhängiger Kovariablen, insbesondere unter AFT- und Cox-Hazard-Modellen, müsste demzufolge noch erörtert werden.

Des Weiteren sind als Mittel zur graphischen Beurteilung eines Regressionsmodells allein Cox-Snell-Residuen vorgestellt worden. Diese eignen sich jedoch lediglich zur Bewertung der Gesamtanpassung eines Modells. Von großem Interesse wären folglich weitere Verfahren, mit deren Hilfe zum Beispiel Ausreißer in einer zensierten Stichprobe identifiziert werden könnten. Ferner wäre es wichtig zu überprüfen, ob für die Kovariablen eines Modells die richtige funktionelle Form angenommen worden ist. In den konkreten Beispielen dieser Arbeit sind Einflussgrößen zwar stets linear modelliert worden, doch kann die Transformation von Kovariablen die Güte der Modellanpassung unter Umständen erheblich verbessern. Diesbezüglich wäre die Betrachtung weiterer diagnostischer Methoden sinnvoll.

Abschließend soll erwähnt werden, dass für die theoretische Untersuchung von statistischen Verfahren der „Survival Analysis“ heute vor allem Punktprozesse und Martingalmethoden eingesetzt werden. Da die vorliegende Arbeit jedoch die Ambition hatte, eine anwendungsorientierte Einführung in die Thematik der Lebenszeitanalyse zu vermitteln, lag ihr Schwerpunkt nicht in der Entwicklung dieser Techniken, sondern in der Herleitung und Anwendung klassischer Verfahren, die nicht auf der Theorie der Zählprozesse basieren.

Anhang

Anhang A

Datensätze

A.1 Zeit bis zum Abbruch einer IUP-Anwendung

patient	time	status	patient	time	status
1	10	1	10	56	0
2	13	0	11	59	1
3	18	0	12	75	1
4	19	1	13	93	1
5	23	0	14	97	1
6	30	1	15	104	0
7	36	1	16	107	1
8	38	0	17	107	0
9	54	0	18	107	0

Variablen des Datensatzes:

patient: Nummer der Patientin (1–18)
time: beobachtete Zeit bis zum Abbruch der IUP-Anwendung in Wochen
status: Status der Patientin (0 = zensiert, 1 = beobachtet)

Quelle: [WHO-87] bzw. [Col-03, S. 5]

A.2 Überlebenszeiten von Brustkrebs-Patientinnen

patient	HPA	time	status	patient	HPA	time	status
1	0	23	1	24	1	40	1
2	0	47	1	25	1	41	1
3	0	69	1	26	1	48	1
4	0	70	0	27	1	50	1
5	0	71	0	28	1	59	1
6	0	100	0	29	1	61	1
7	0	101	0	30	1	68	1
8	0	148	1	31	1	71	1
9	0	181	1	32	1	76	0
10	0	198	0	33	1	105	0
11	0	208	0	34	1	107	0
12	0	212	0	35	1	109	0
13	0	224	0	36	1	113	1
14	1	5	1	37	1	116	0
15	1	8	1	38	1	118	1
16	1	10	1	39	1	143	1
17	1	13	1	40	1	154	0
18	1	18	1	41	1	162	0
19	1	24	1	42	1	188	0
20	1	26	1	43	1	212	0
21	1	26	1	44	1	217	0
22	1	31	1	45	1	225	0
23	1	35	1				

Variablen des Datensatzes:

patient: Nummer der Patientin (1–45)
HPA: HPA-Markierung (0 = negativ, 1 = positiv)
time: beobachtete Lebensdauer in Monaten
status: Status der Patientin (0 = zensiert, 1 = tot)

Quelle: [Lea-87] bzw. [Col-03, S. 7]

A.3 Überlebenszeiten von Patienten mit Plasmozytom

patient	time	status	age	sex	bun	ca	hb	pcells	protein
1	1	1	67	1	165	10	9.4	90	0
2	1	1	57	1	20	9	5.1	100	1
3	1	1	75	1	56	12	11.3	18	0
4	1	1	38	1	87	18	12.0	90	1
5	1	1	81	1	33	15	9.8	100	1
6	2	1	46	2	35	10	6.7	86	0
7	4	1	50	2	172	9	10.1	46	1
8	4	1	74	1	48	9	6.5	54	0
9	5	1	60	1	13	10	9.7	25	0
10	5	1	67	2	26	8	10.4	49	0
11	5	1	70	2	130	8	10.2	23	0
12	5	1	77	1	23	8	9.0	29	0
13	6	1	53	2	15	13	11.4	33	1
14	6	1	61	2	11	10	5.1	100	0
15	6	1	48	1	95	10	9.5	37	0
16	8	1	55	1	53	12	8.2	55	0
17	10	1	65	1	20	10	13.2	66	0
18	10	1	70	1	37	12	7.5	47	0
19	10	1	51	2	12	9	9.6	80	0
20	10	1	61	1	13	10	14.0	19	0
21	10	1	43	1	17	9	12.0	15	1
22	12	1	60	2	6	10	5.5	25	0
23	13	1	66	1	25	10	14.6	18	1
24	14	1	70	1	40	11	10.6	27	0
25	15	1	62	2	21	10	8.8	5	0
26	15	1	48	1	22	10	9.0	100	0
27	16	1	68	1	39	10	11.2	41	0
28	16	1	53	1	17	9	10.0	28	0
29	17	1	65	2	28	8	7.5	8	0
30	18	1	51	1	12	15	14.4	100	0

patient	time	status	age	sex	bun	ca	hb	pcells	protein
31	18	1	60	2	18	9	7.5	85	1
32	23	1	56	2	20	9	14.6	3	0
33	24	1	67	1	10	10	12.4	44	0
34	25	1	49	2	17	11	11.2	100	1
35	31	1	46	1	21	9	10.6	43	0
36	34	1	48	1	13	10	7.0	15	1
37	36	1	63	1	40	9	11.0	16	1
38	40	1	69	1	10	10	10.2	30	1
39	40	1	70	2	14	9	5.0	22	0
40	50	1	74	1	37	13	7.7	11	1
41	51	1	60	2	10	10	10.1	45	1
42	53	1	49	1	18	10	9.0	50	1
43	57	1	42	2	16	10	12.1	38	1
44	65	1	59	1	28	9	6.6	66	0
45	66	1	52	1	21	10	12.8	11	1
46	87	1	47	2	15	9	10.6	57	1
47	88	1	63	1	21	9	14.0	42	1
48	91	1	58	2	27	11	11.0	26	1
49	3	0	59	1	90	10	10.2	6	1
50	3	0	49	2	84	13	10.0	42	0
51	6	0	48	2	13	10	12.4	72	0
52	6	0	81	1	34	11	10.2	76	0
53	7	0	57	2	12	8	9.9	45	0
54	10	0	60	1	41	9	14.0	70	1
55	11	0	66	2	25	9	8.8	23	0
56	11	0	46	2	14	7	11.6	14	0
57	12	0	71	2	46	9	4.9	62	0
58	15	0	55	1	14	9	13.0	8	0
59	18	0	69	2	21	10	10.8	33	0
60	18	0	59	2	21	10	13.0	100	0
61	27	0	82	2	17	9	7.3	47	0
62	40	0	72	1	57	9	12.8	28	1
63	52	0	66	1	13	11	12.0	100	0

patient	time	status	age	sex	bun	ca	hb	pcells	protein
64	56	0	66	1	18	11	12.5	90	0
65	76	0	60	1	12	12	14.0	9	0

Variablen des Datensatzes:

patient: Nummer des Patienten (1–65)
time: beobachtete Lebensdauer ab Zeitpunkt der Diagnose in Monaten
status: Status des Patienten (0 = lebend, 1 = tot)
age: Alter des Patienten in Jahren
sex: Geschlecht des Patienten (1 = männlich, 2 = weiblich)
bun: Blut-Urea-Stickstoff-Wert (blood urea nitrogen)
ca: Serum-Calcium-Wert
hb: Hämoglobin-Wert
pcells: prozentualer Anteil von Plasma-Zellen im Knochenmark
protein: Bence-Jones-Proteine im Urin bei Diagnose (0 = nein, 1 = ja)

Die Werte der Variablen age, bun, ca, hb, pcells und protein entsprechen den Messungen zum Zeitpunkt der Diagnose. Die Daten entstammen einem umfangreicheren Datensatz aus [Kra-75]. Dieser enthält Werte von insgesamt 16 erklärenden Variablen, die in dem Beispiel des Abschnitts 7.3.1 jedoch nicht alle berücksichtigt worden sind. Die Messeinheiten von bun, ca und hb werden in [Kra-75] nicht angegeben.

A.4 Vergleich zweier Therapien bei Prostata-Krebs

patient	treatment	time	status	age	shb	size	index
1	0	65	0	67	13.4	34	8
2	1	61	0	60	14.6	4	10
3	1	60	0	77	15.6	3	8
4	0	58	0	64	16.2	6	9
5	1	51	0	65	14.1	21	9
6	0	51	0	61	13.5	8	8
7	0	14	1	73	12.4	18	11
8	0	43	0	60	13.6	7	9
9	1	16	0	73	13.8	8	9
10	0	52	0	73	11.7	5	9
11	0	59	0	77	12.0	7	10
12	1	55	0	74	14.3	7	10
13	1	68	0	71	14.5	19	9
14	1	51	0	65	14.4	10	9
15	0	2	0	76	10.7	8	9
16	0	67	0	70	14.7	7	9
17	1	66	0	70	16.0	8	9
18	1	66	0	70	14.5	15	11
19	1	28	0	75	13.7	19	10
20	1	50	1	68	12.0	20	11
21	0	69	1	60	16.1	26	9
22	0	67	0	71	15.6	8	8
23	1	65	0	51	11.8	2	6
24	0	24	0	71	13.7	10	9
25	1	45	0	72	11.0	4	8
26	1	64	0	74	14.2	4	6
27	0	61	0	75	13.7	10	12
28	0	26	1	72	15.3	37	11
29	0	42	1	57	13.9	24	12
30	1	57	0	72	14.6	8	10
31	1	70	0	72	13.8	3	9

patient	treatment	time	status	age	shb	size	index
32	1	5	0	74	15.1	3	9
33	1	54	0	51	15.8	7	8
34	0	36	1	72	16.4	4	9
35	1	70	0	71	13.6	2	10
36	1	67	0	73	13.8	7	8
37	0	23	0	68	12.5	2	8
38	0	62	0	63	13.2	3	8

Variablen des Datensatzes:

patient: Nummer des Patienten (1–65)
treatment: Therapie (0 = Placebo, 1 = DES)
time: beobachtete Lebensdauer in Monaten
status: Status des Patienten (0 = lebend, 1 = tot)
age: Alter des Patienten in Jahren
shb: Serum-Hämoglobin in gm/100ml
size: Größe des Tumors in cm^2
index: Gleason-Index

Messungen der Kovariablen age, shb, size und index erfolgten zu Studienbeginn. Die Daten der Tabelle sind Teil eines Datensatzes in [AnH-85, S. 261–274]. In dem hier dargestellten Umfang sind sie auch in [Col-03, S. 10] zu finden.

Anhang B

R-Quellcodes

B.1 Zeit bis zum Abbruch einer IUP-Anwendung

```
library(splines)
library(survival)
iud<- read.table("C:\\ ... \\iud.txt",header=T)

fit.1<- survfit(Surv(time, status), conf.type="plain", iud)
summary(fit.1)

# Konfidenzintervalle nach log-log-Transformation:
fit.2<- survfit(Surv(time, status), iud)
summary(fit.2)

# Plot des KM-Schätzers mit 0.95-Konfidenzintervallen:
plot(fit.2, xlab="Zeit (in Wochen)", ylab="Geschätzte Survival-Funktion")
```

B.2 Zeit bis zum Abbruch einer IUP-Anwendung – Fortsetzung

```
library(splines)
library(survival)
iud<- read.table("C:\\ ... \\iud.txt",header=T)

## KM-Schätzer:
```

```
fit<- survfit(Surv(time, status), iud)

## Plot zum Prüfen auf Weibull-Verteilung:
plot(log(fit$time), log(-log(fit$surv)), pch=19,
+ xlab= "ln t", ylab="ln(-ln KM)", main="Weibull")

# Regressionsgerade:
rg_w<- lm(log(-log(fit$surv))~log(fit$time))
rg_w
abline(rg_w)

## Plot zum Prüfen auf Exponential-Verteilung:
plot(fit$time, log(fit$surv), pch=19,
+ xlab="t", ylab="ln KM", main="Exponential")

# Regressionsgerade:
rg_e<- lm(log(fit$surv)~fit$time)
rg_e
abline(rg_e)
```

B.3 Anpassung von AFT-Modellen – Prognose für Brustkrebs-Patientinnen

```
library(splines)
library(survival)
breast.cancer<- read.table("C:\\ ... \\breast.cancer.txt",header=T)


## 1. WEIBULL-VERTEILUNG:

# Schätzen der Parameter im Lokation-Skalen-Modell:
fit.w<- survreg(Surv(time, status) ~ HPA, breast.cancer, dist="weibul")
fit.w$coef
fit.w$scale

# Plot Survival-Funktion:
```

```
surv.func.w.0<- function(x){exp(-exp((log(x)-5.8544)/1.0668))}
surv.func.w.1<- function(x){exp(-exp((log(x)-5.8544+0.9967)/1.0668))}

x<- seq(0,400,0.1)
plot(x, surv.func.w.1(x), type="l", col="red", xlab="Zeit (in Monaten)",
+ ylab="Geschätzte Survival-Funktion", main="Survival-Funktion")
lines(x, surv.func.w.0(x))

# Plot Hazard-Rate:
hazard.rate.w.0<- function(x)
+ {1.066777^{-1}*x^{1.066777^{-1}-1}*exp(-5.8543638/1.066777)}
hazard.rate.w.1<- function(x)
+ {1.066777^{-1}*x^{1.067^{-1}-1}*exp((-5.8543638+0.9966647)/1.066777)}

x<- seq(0,400,0.1)
plot(x,hazard.rate.w.0(x), ylim=range(0,0.01), type="l",
+ xlab="Zeit (in Monaten)", ylab="Geschätzte Hazard-Funktion",
+ main="Hazard-Funktion")
lines(x,hazard.rate.w.1(x), col="red", type="l")


## 2. LOG-LOGISTIK-VERTEILUNG:

# Schätzen der Parameter im Lokation-Skalen-Modell:
fit.l<- survreg(Surv(time, status) ~ HPA,
+ breast.cancer, dist="loglogistic")
fit.l$coef
fit.l$scale

# Plot Survival-Funktion:
surv.func.l.0<- function(x){(1+exp((log(x)-5.4611)/0.8047))^{-1}}
surv.func.l.1<- function(x){(1+exp((log(x)-5.4611+1.1491)/0.8047))^{-1}}

x<- seq(0,400,0.1)
plot(x, surv.func.l.1(x), col="red", type="l",
+ xlab="Zeit (in Monaten)", ylab="Geschätzte Survival-Funktion",
+ main="Survial-Funktion")
```

```
lines(x, surv.func.l.0(x), type="l")

# Plot Hazard-Rate:
hazard.rate.l.0<- function(x)
+ {1.2427*x^{-1}*(1+x^{-1.2427}*exp(6.7865))^{-1}}
hazard.rate.l.1<- function(x)
+ {1.2427*x^{-1}*(1+x^{-1.2427}*exp(6.7865-1.4280))^{-1}}

x<- seq(0,400,0.1)
plot(x,hazard.rate.l.0(x), ylim=range(0,0.01), type="l",
+ xlab="Zeit (in Monaten)", ylab="Geschätzte Hazard-Funktion",
+ main="Hazard-Funktion")
lines(x,hazard.rate.l.1(x), col="red", type="l")


## 3. LOG-NORMAL-VERTEILUNG:

# Schätzen der Parameter im Lokation-Skalen-Modell:
fit.n<- survreg(Surv(time, status) ~ HPA, breast.cancer, dist="lognormal")
fit.n$coef
fit.n$scale

# Plot Survival-Funktion:
surv.func.n.0<- function(x)
+ {1-pnorm((log(x)- 5.491726)/1.359451, mean=0, sd=1)}
surv.func.n.1<- function(x)
+ {1-pnorm((log(x)- 5.491726 + 1.151172)/1.359451, mean=0, sd=1)}

x<- seq(0,400,0.1)
plot(x, surv.func.n.1(x), col="red", type="l",
+ xlab="Zeit (in Monaten)", ylab="Geschätzte Survival-Funktion"
+ main="Survival-Funktion")
lines(x, surv.func.n.0(x), type="l")

# Plot Hazard-Rate:
hazard.rate.n.0<- function(x)
+ {0.7356*x^{-1}*(1-pnorm(0.7356*log(x)-4.0395, mean=0, sd=1))^{-1}*
```

```
+ dnorm(0.7356*log(x)-4.0395)}
hazard.rate.n.1<- function(x)
+ {0.7356*x^{-1}*(1-pnorm(0.7356*log(x)-4.0395+0.8468, mean=0, sd=1))^{-1}*
+ dnorm(0.7356*log(x)-4.0395+0.8468)}

x<- seq(0,400,0.1)
plot(x,hazard.rate.n.1(x), col="red", type="l",
+ xlab="Zeit (in Monaten)", ylab="Geschätzte Hazard-Funktion"
+ main="Hazard-Funktion")
lines(x,hazard.rate.n.0(x), type="l")
```

B.4 Überprüfung der Modelle aus Abschnitt 6.2

```
library(splines)
library(survival)
breast.cancer<- read.table("C:\\ ... \\breast.cancer.txt",header=T)


## 1. WEIBULL-VERTEILUNG:

# Standardisierte Residuen:
s.resid_w<- (log(breast.cancer$time)-5.8544
+ +0.9967*breast.cancer$HPA)/1.0668

# Cox-Snell-Residuen:
cs.resid_w<- exp(s.resid_w)

# Berechnung des Kaplan-Meier-Schätzers:
km.cs_w<- survfit(Surv(cs.resid_w, breast.cancer$status))

# Berechnung der Werte des Kaplan-Meier-Schätzers
# und des Schätzers für die kumulierte Hazard-Rate in den Residuen:
cs.S_w<- km.cs_w$surv
cs.S_w
cs.H_w<- -log(cs.S_w)
cs.H_w
```

```
# Plot zur graphischen Überprüfung des Weibull-AFT-Modells:
cs.times_w<- km.cs_w$time
plot(cs.times_w, cs.H_w, type="b",lty=1, pch=20, col="red",
+ xlab="Cox-Snell-Residuum", ylab="Kumulierte Hazard", main="Weibull")
abline(0,1,lty=2)


## 2. LOG-LOGISTIK-VERTEILUNG:

# Standardisierte Residuen:
s.resid_l<- (log(breast.cancer$time)-5.4611
+ +1.1491*breast.cancer$HPA)/0.8047

# Cox-Snell-Residuen:
cs.resid_l<- log(1+exp(s.resid_l))

# Berechnung des Kaplan-Meier-Schätzers:
km.cs_l<- survfit(Surv(cs.resid_l, breast.cancer$status))

# Berechnung der Werte des Kaplan-Meier-Schätzers
# und des Schätzers für die kumulierte Hazard-Rate in den Residuen:
cs.S_l<- km.cs_l$surv
cs.S_l
cs.H_l<- -log(cs.S_l)
cs.H_l

# Plot zur graphischen Überprüfung des Log-Logistik-AFT-Modells:
cs.times_l<- km.cs_l$time
plot(cs.times_l, cs.H_l, type="b",lty=1, pch=20, col="red",
+ xlab="Cox-Snell-Residuum", ylab="Kumulierte Hazard", main="Log-Logistik")
abline(0,1,lty=2)


## 3. LOG-NORMAL-VERTEILUNG:

# Standardisierte Residuen:
s.resid_n<- (log(breast.cancer$time)-5.4917
```

```
+ +1.1512*breast.cancer$HPA)/1.3595

# Cox-Snell-Residuen:
cs.resid_n<- -log(1-pnorm(s.resid_n, mean=0, sd=1))

# Berechnung des Kaplan-Meier-Schätzers:
km.cs_n<- survfit(Surv(cs.resid_n, breast.cancer$status))

# Berechnung der Werte des Kaplan-Meier-Schätzers
+ und des Schätzers für die kumulierte Hazard-Rate in den Residuen:
cs.S_n<- km.cs_n$surv
cs.S_n
cs.H_n<- -log(cs.S_n)
cs.H_n

# Plot zur graphischen Überprüfung des Log-Normal-AFT-Modells:
cs.times_n<- km.cs_n$time
plot(cs.times_n, cs.H_n, type="b",lty=1, pch=20, col="red",
+ xlab="Cox-Snell-Residuum", ylab="Kumulierte Hazard", main="Log-Normal")
abline(0,1,lty=2)
```

B.5 Identifikation von erklärenden Variablen mit Einfluss auf die Lebenszeit

```
library(splines)
library(survival)
myeloma<- read.table("C:\\ ... \\sort.daten.myeloma.txt",header=T)

# Anpassung des vollen Modells mittels des Schätzers von Breslow:

fitm<-coxph(Surv(time,status)~ age + sex + bun + ca +
+ hb + pcells + protein, myeloma, method="breslow")

# SCHRITT 1: Null-Modell und Modelle mit jeweils einer Kovariable,
# Bestimmung von Kovariablen,
# die einzeln Auswirkungen auf die Lebenszeit haben:
```

```
Age<- update(fitm, . ~ . - sex - bun - ca - hb - pcells - protein)
Sex<- update(fitm, . ~ . - age - bun - ca - hb - pcells - protein)
Bun<- update(fitm, . ~ . - age - sex - ca - hb - pcells - protein)
Ca<- update(fitm, . ~ . -age - sex - bun - hb - pcells - protein)
Hb<- update(fitm, . ~ . -age - sex - bun - ca - pcells - protein)
Pcells<- update(fitm, . ~ . -age - sex - bun - ca - hb - protein)
Protein<- update(fitm, . ~ . -age - sex - bun - ca - hb - pcells)

L.none<- -2*fitm$loglik[1]
L.Age<- -2*Age$loglik[2]
L.Sex<- -2*Sex$loglik[2]
L.Bun<- -2*Bun$loglik[2]
L.Ca<- -2*Ca$loglik[2]
L.Hb<- -2*Hb$loglik[2]
L.Pcells<- -2*Pcells$loglik[2]
L.Protein<- -2*Protein$loglik[2]

L.none
L.Age
L.Sex
L.Bun
L.Ca
L.Hb
L.Pcells
L.Protein

lq.Bun<- L.none - L.Bun
lq.Bun
pchisq(lq.Bun, 1, lower.tail=F)

lq.Ca<- L.none - L.Ca
lq.Ca<- L.none - L.Ca
pchisq(lq.Ca, 1, lower.tail=F)

lq.Hb<- L.none - L.Hb
lq.Hb
```

```
pchisq(lq.Hb, 1, lower.tail=F)

lq.Protein<- L.none - L.Protein
lq.Protein
pchisq(lq.Protein, 1, lower.tail=F)

# Signifikant abgelehnt werden die Nullhypothesen
# beta(bun)=0 [zum Niveau 0.1%] und
# beta(hb)=0 [zum Niveau 5%].
# Nicht verworfen werden kann die Hypothese beta(ca)=0.
# Obwohl P-Wert zu protein relativ hoch (P=0.1522497),
# soll Protein zunächst im Modell bleiben.

# SCHRITT 2: Modell-Konstruktion mit Bun, Hb und Protein,
# testen, ob in Kombination einige von ihnen unwichtig werden.

Bun.Hb.Protein<- update(fitm, . ~ . - age - sex -  ca - pcells)
Bun.Hb<- update(fitm, . ~ . - age - sex -  ca - pcells -protein)
Bun.Protein<- update(fitm, . ~ . - age - sex -  ca - pcells - hb)
Hb.Protein<- update(fitm, . ~ . - age - sex -  ca - pcells - bun)

L.Bun.Hb.Protein<- -2*Bun.Hb.Protein$loglik[2]
L.Bun.Hb<- -2*Bun.Hb$loglik[2]
L.Bun.Protein<- -2*Bun.Protein$loglik[2]
L.Hb.Protein<- -2*Hb.Protein$loglik[2]

L.Bun.Hb.Protein
L.Bun.Hb
L.Bun.Protein
L.Hb.Protein

lq.Bun.Hb<- L.Bun.Hb - L.Bun.Hb.Protein
lq.Bun.Hb
pchisq(lq.Bun.Hb, 1, lower.tail=F)

lq.Bun.Protein<- L.Bun.Protein - L.Bun.Hb.Protein
lq.Bun.Protein
```

```
pchisq(lq.Bun.Protein, 1, lower.tail=F)

lq.Hb.Protein<- L.Hb.Protein - L.Bun.Hb.Protein
lq.Hb.Protein
pchisq(lq.Hb.Protein, 1, lower.tail=F)

# Signifikant abgeleht werden die beiden Nullhypothesen
# [beta(bun), beta(hb), beta(protein)] = [0, beta(hb), beta(protein)]
# [zum Niveau 0.1%] und
# [beta(bun), beta(hb), beta(protein)] = [beta(bun), 0, beta(protein)]
# [zum Niveau 7%].
# Nicht verworfen werden kann
# [beta(bun), beta(hb), beta(protein)] = [beta(bun), beta(hb), 0],
# protein wird also aus dem Modell entfernt.

# SCHRITT 3: Soll das bun-hb-Modell durch die Kovariablen
# age, sex, ca, und Pcells ergänzt werden?
# Einzeln werden diese Kovariablen dem bestehenden Modell
# hinzugefügt und Tests auf Null-Effekt durchgeführt.

Bun.Hb.Age<- update(fitm, . ~ . - sex - ca - pcells - protein)
Bun.Hb.Sex<- update(fitm, . ~ . - age - ca - pcells - protein)
Bun.Hb.Ca<- update(fitm, . ~ . - age - sex - pcells - protein)
Bun.Hb.Pcells<- update(fitm, . ~ . - age - sex - ca - protein)

L.Bun.Hb.Age<- -2*Bun.Hb.Age$loglik[2]
L.Bun.Hb.Sex<- -2*Bun.Hb.Sex$loglik[2]
L.Bun.Hb.Ca<- -2*Bun.Hb.Ca$loglik[2]
L.Bun.Hb.Pcells<- -2*Bun.Hb.Pcells$loglik[2]

L.Bun.Hb.Age
L.Bun.Hb.Sex
L.Bun.Hb.Ca
L.Bun.Hb.Pcells

lq.Bun.Hb.Age<- L.Bun.Hb - L.Bun.Hb.Age
lq.Bun.Hb.Sex<- L.Bun.Hb - L.Bun.Hb.Sex
```

```
lq.Bun.Hb.Ca<- L.Bun.Hb - L.Bun.Hb.Ca
lq.Bun.Hb.Pcells<- L.Bun.Hb - L.Bun.Hb.Pcells

lq.Bun.Hb.Age
lq.Bun.Hb.Sex
lq.Bun.Hb.Ca
lq.Bun.Hb.Pcells

pchisq(lq.Bun.Hb.Age, 1, lower.tail=F)
pchisq(lq.Bun.Hb.Sex, 1, lower.tail=F)
pchisq(lq.Bun.Hb.Ca, 1, lower.tail=F)
pchisq(lq.Bun.Hb.Pcells, 1, lower.tail=F)

# Keine der Nullhypothesen kann signifikant verworfen werden.
# Also: Das am besten geeignete Modell enthält die Kovariablen bun und hb.
# Schätzer der zugehörigen Koeffizienten:

fit<-coxph(Surv(time,status)~ bun + hb, myeloma, method="breslow")
fit
```

B.6 Vergleich zweier Therapien bei Prostata-Krebs

```
library(splines)
library(survival)
prostatic.cancer<-read.table("C:\\ ... \\prostatic.cancer.txt",header=T)

# TEIL 1: Auswahl relevanter Kovariablen.
# Welche der Kovariablen age, shb, size und index
# beeinflussen die Lebensdauer der Patienten?
# Hier: 2^4=16 Kombinationsmöglichkeiten. Alle Modelle werden konstruiert.

fitm<-coxph(Surv(time, status) ~ age + shb + size + index,
+ prostatic.cancer, method="breslow")

Age<- update(fitm, . ~ . - shb - size - index)
Shb<- update(fitm, . ~ . - age - size - index)
Size<- update(fitm, . ~ . - age - shb - index)
```

```
Index<- update(fitm, . ~ . - age - shb - size)

Age.Shb<- update(fitm, . ~ . - size - index)
Age.Size<- update(fitm, . ~ . - shb - index)
Age.Index<- update(fitm, . ~ . - shb - size)
Shb.Size<- update(fitm, . ~ . - age - index)
Shb.Index<- update(fitm, . ~ . - age - size)
Size.Index<- update(fitm, . ~ . - age - shb)

Age.Shb.Size<- update(fitm, . ~ . - index)
Age.Shb.Index<- update(fitm, . ~ . - size)
Age.Size.Index<- update(fitm, . ~ . - shb)
Shb.Size.Index<- update(fitm, . ~ . - age)

# Werte von -2*log L(beta) für jedes dieser Modelle:

L.none<- -2*fitm$loglik[1]

L.Age<- -2*Age$loglik[2]
L.Shb<- -2*Shb$loglik[2]
L.Size<- -2*Size$loglik[2]
L.Index<- -2*Index$loglik[2]

L.Age.Shb<- -2*Age.Shb$loglik[2]
L.Age.Size<- -2*Age.Size$loglik[2]
L.Age.Index<- -2*Age.Index$loglik[2]
L.Shb.Size<- -2*Shb.Size$loglik[2]
L.Shb.Index<- -2*Shb.Index$loglik[2]
L.Size.Index<- -2*Size.Index$loglik[2]

L.Age.Shb.Size<- -2*Age.Shb.Size$loglik[2]
L.Age.Shb.Index<- -2*Age.Shb.Index$loglik[2]
L.Age.Size.Index<- -2*Age.Size.Index$loglik[2]
L.Shb.Size.Index<- -2*Shb.Size.Index$loglik[2]

L.full<- -2*fitm$loglik[2]
```

```
L.none
L.Age
L.Shb
L.Size
L.Index

L.Age.Shb
L.Age.Size
L.Age.Index
L.Shb.Size
L.Shb.Index
L.Size.Index

L.Age.Shb.Size
L.Age.Shb.Index
L.Age.Size.Index
L.Shb.Size.Index

L.full

# Betrachte Modelle mit jeweils einer Kovariable.
# Teste: H_0: beta(i)=0, i=age, shb, size, index.
# Die Kovariablen index, size und shb führen einzeln
# zu den drei kleinsten Werten von -2*logL(beta).
# -> Tests: H_0: beta(i)=0, i=index, size, shb.

lq.Index<- L.none - L.Index
lq.Index
pchisq(lq.Index, 1, lower.tail=F)

lq.Size<- L.none - L.Size
lq.Size
pchisq(lq.Size, 1, lower.tail=F)

lq.Shb<- L.none - L.Shb
lq.Shb
pchisq(lq.Shb, 1, lower.tail=F)
```

```
# Bei einzelner Betrachtung: index und size
# haben signifikanten Einfluss auf Lebenszeit.
# -> Im Modell mit index und size ist zu testen,
# ob sie in Kombination immernoch wichtig sind:
# H_0: [beta(index), beta(size)] = [0, beta(size)]
# H_0: [beta(index), beta(size)] = [beta(index), 0]

lq.I<- L.Size - L.Size.Index
lq.I
pchisq(lq.I, 1, lower.tail=F)

lq.S<- L.Index - L.Size.Index
lq.S
pchisq(lq.S, 1, lower.tail=F)

# Beide Hypothesen können abgelehnt werden.
# Zu prüfen ist, ob dem Modell mit index und size die Variablen
# age und shb hinzugefügt werden sollen.
# Dazu: Betrachte das Modell mit (index, size, age), (index, size, shb)
# bzw. (index, size, age, shb) und teste
# H_0: [beta(age), beta(index), beta(size)] = [0, beta(index), beta(size)]
# H_0: [beta(shb), beta(index), beta(size)] = [0, beta(index), beta(size)]
# H_0: [beta(age), beta(shb), beta(index), beta(size)]
# = [0, 0, beta(index), beta(size)]

lq.A<- L.Size.Index - L.Age.Size.Index
lq.A
pchisq(lq.A, 1, lower.tail=F)

lq.Sh<- L.Size.Index - L.Shb.Size.Index
lq.Sh
pchisq(lq.Sh, 1, lower.tail=F)

lq.A.Sh<- L.Size.Index - L.full
lq.A.Sh
pchisq(lq.A.Sh, 2, lower.tail=F)
```

```
# Keine der Hypothesen kann abgelehnt werden.
# Also: Das geeignete Modell enthält size und index.


# TEIL 2: Hat eine Behandlung mit DES signifikante Auswirkungen?

fitm0<- coxph(Surv(time, status) ~ treatment + size + index,
+ prostatic.cancer, method="breslow")

L.Treat.Size.Index<- -2*fitm0$loglik[2]
L.Treat.Size.Index

# Teste:
# H_0: [beta(treatment), beta(size), beta(index)]
# = [0, beta(size), beta(index)]

lq.T<- L.Size.Index - L.Treat.Size.Index
lq.T
pchisq(lq.T, 1, lower.tail=F)

# Nullhypothese kann nicht abgelehnt werden.
# Teste, ob Interaktionen treatment/size und treatment/index bestehen:
# H_0: [beta(size), beta(index), beta(treat), beta(treat:size)]
# = [beta(size), beta(index), beta(treat), 0]
# H_0: [beta(size), beta(index), beta(treat), beta(treat:index)]
# = [beta(size), beta(index), beta(treat), 0]
# H_0: [beta(size), beta(index), beta(treat), beta(treat:size),
#  beta(treat:index)]= [beta(size), beta(index), beta(treat), 0, 0]

fitm1<- coxph(Surv(time, status) ~ treatment + size +
+ index + treatment:size, prostatic.cancer, method="breslow")
fitm2<- coxph(Surv(time, status) ~ treatment + size +
+ index + treatment:index, prostatic.cancer, method="breslow")
fitm3<- coxph(Surv(time, status) ~ treatment + size + index +
+ treatment:size + treatment:index, prostatic.cancer, method="breslow")
```

```
L.TrSize<- -2*fitm1$loglik[2]
L.TrIndex<- -2*fitm2$loglik[2]
L.TrSize.TrIndex<- -2*fitm3$loglik[2]
L.no.int<- -2*fitm0$loglik[2]

L.TrSize
L.TrIndex
L.TrSize.TrIndex
L.no.int

lq.TrSize<- L.no.int - L.TrSize
lq.TrSize
pchisq(lq.TrSize, 1, lower.tail=F)

lq.TrIndex<- L.no.int - L.TrIndex
lq.TrIndex
pchisq(lq.TrIndex, 1, lower.tail=F)

lq.TrSize.TrIndex<- L.no.int - L.TrSize.TrIndex
lq.TrSize.TrIndex
pchisq(lq.TrSize.TrIndex, 2, lower.tail=F)

# Keine der Nullhypothesen kann signifikant verworfen werden!
# Keine Interaktionen der Form treatment/size, treatment/index!
# Behandlung mit DES ohne Auswirkung auf Lebenszeit!
```

B.7 Vergleich zweier Therapien bei Prostata-Krebs – Fortsetzung

```
library(splines)
library(survival)
prostatic.cancer<-read.table("C:\\ ... \\prostatic.cancer.txt",header=T)
prostatic.cancer

# Modell:
# lambda(t) =
```

```
# lambda_0(t) exp[beta(size)size + beta(index)index + beta(treat)treat]

fitm<-coxph(Surv(time, status) ~ size + index + treatment,
+ prostatic.cancer, method="breslow")
fitm$coef
exp(fitm$coef)

# Vergleich:
# Modell, das nur die Kovariable treatment enthält:

fitt<- coxph(Surv(time, status) ~ size + index + treatment,
+ prostatic.cancer, method="breslow")
fitt$coeff
exp(fitt$coef)
```

B.8 Schätzung der Survival-Funktion für Patienten mit Plasmozytom

```
library(splines)
library(survival)
myeloma<- read.table("C:\\ ... \\sort.daten.myeloma.txt",header=T)

fitm<-coxph(Surv(time,status)~ bun + hb, myeloma, method="breslow")

plot( survfit(fitm), conf.int=FALSE, lty=3, lwd=1, xlim=c(0,75), xlab =
+ "Überlebenszeit (in Monaten)", ylab = "Geschätzte Survival-Funktion")
lines( survfit(fitm, newdata=data.frame(bun=15, hb=14)), col="magenta",
+ lty=1, lwd=2)
lines( survfit(fitm, newdata=data.frame(bun=8, hb=14)), col="green",
+ lty=1, lwd=2)
lines( survfit(fitm, newdata=data.frame(bun=15, hb=5)), col="yellow",
+ lty=1, lwd=2)
lines( survfit(fitm, newdata=data.frame(bun=160, hb=14)), col="blue",
+ lty=1, lwd=2)
```

B.9 Überprüfung des Modells aus Abschnitt 7.3.1

```
library(splines)
library(survival)
myeloma<- read.table("C:\\ ... \\sort.daten.myeloma.txt",header=T)
fitm<-coxph(Surv(time,status)~ bun + hb, myeloma, method="breslow")

# Berechnung der Cox-Snell-Residuen:
m.resid<- resid(fitm)
cs.resid<- myeloma$status - m.resid
cs.resid

# Berechnung des Kaplan-Meier-Schätzers:
km.cs<- survfit(Surv(cs.resid, myeloma$status))

# Berechnung der Werte des Kaplan-Meier-Schätzers
# und des Schätzers für die kumulierte Hazard-Rate in den Residuen:
cs.S<- km.cs$surv
cs.S
cs.H<- -log(cs.S)
cs.H

# Plot zur graphischen Überprüfung des Cox-Hazard-Modells:
cs.times<- km.cs$time
plot(cs.times,cs.H, type="b",lty=1, pch=20, col="red" ,
+ xlab = "Cox-Snell-Residuum", ylab = "Kumulierte Hazard im Residuum")
abline(0,1,lty=2)
```

Bemerkung B.1. Die Berechnung der Cox-Snell-Residuen (`cs.resid`) erfolgt hier gemäß [Sac-06, S. 629] über Martingal-Residuen (`m.resid`). Ist r_i das Cox-Snell-Residuum für die i-te Beobachtung und δ_i der entsprechende Zensurindikator, so gilt bei einem rechtszensierten Datensatz und zeitunabhängigen Kovariablen: $m_i = \delta_i - r_i$. Siehe dazu [Kle-97, Abschnitt 11.3].

B.10 Überprüfung der PH-Annahme

```
library(splines)
```

```
library(survival)
myeloma<- read.table("C:\\ ... \\sort.daten.myeloma.txt",header=T)

# Kategorisierung der Daten:

hb.leq7<- subset(myeloma, hb<=7)
hb.g7<- subset(myeloma, hb>7)
hb.g7.leq10<- subset(hb.g7, hb<=10)
hb.g10<- subset(myeloma, hb>10)
hb.g10.leq13<- subset(hb.g10, hb<=13)
hb.g13<- subset(myeloma, hb>13)

# Berechnung des KM-Schätzers für jede Gruppe:

km.hb.leq7<- survfit(Surv(hb.leq7$time, hb.leq7$status))
km.hb.g7.leq10<- survfit(Surv(hb.g7.leq10$time,
+ hb.g7.leq10$status))
km.hb.g10.leq13<- survfit(Surv(hb.g10.leq13$time,
+ hb.g10.leq13$status))
km.hb.g13<- survfit(Surv(hb.g13$time, hb.g13$status))

# Berechnung der Werte des KM-Schätzers und des Schätzers
# für die log-kumulierte Haz-Fkt in den Beobachtungszeiten:

S.hb.leq7<- km.hb.leq7$surv
H.hb.leq7<- -log(S.hb.leq7)
log.H.hb.leq7<- log(H.hb.leq7)

S.hb.g7.leq10<- km.hb.g7.leq10$surv
H.hb.g7.leq10<- -log(S.hb.g7.leq10)
log.H.hb.g7.leq10<- log(H.hb.g7.leq10)

S.hb.g10.leq13<- km.hb.g10.leq13$surv
H.hb.g10.leq13<- -log(S.hb.g10.leq13)
log.H.hb.g10.leq13<- log(H.hb.g10.leq13)

S.hb.g13<- km.hb.g13$surv
```

```
H.hb.g13<- -log(S.hb.g13)
log.H.hb.g13<- log(H.hb.g13)

# Plot zur graphischen Überprüfung der PH-Annahme:

# kumulierte Hazard-Rate vs. Zeit:
times.hb.leq7<- km.hb.leq7$time
times.hb.g7.leq10<- km.hb.g7.leq10$time
times.hb.g10.leq13<- km.hb.g10.leq13$time
times.hb.g13<- km.hb.g13$time

plot(times.hb.g10.leq13, log.H.hb.g10.leq13, type="b", lty=2,
+ pch=20, col="black", xlab="Überlebenszeit (in Monaten)",
+ ylab="Log-kumulierte Hazard-Funktion")
lines(times.hb.leq7,log.H.hb.leq7, type="b", lty=2, pch=20,
+ col="magenta")
lines(times.hb.g13,log.H.hb.g13, type="b",lty=2, pch=20,
+ col="green")
lines(times.hb.g7.leq10,log.H.hb.g7.leq10, type="b",lty=2, pch=20,
+ col="blue")

# Kumulierte Hazard-Rate vs. log-Zeit:
log.times.hb.leq7<- log(km.hb.leq7$time)
log.times.hb.g7.leq10<- log(km.hb.g7.leq10$time)
log.times.hb.g10.leq13<- log(km.hb.g10.leq13$time)
log.times.hb.g13<- log(km.hb.g13$time)

plot(log.times.hb.g10.leq13, log.H.hb.g10.leq13, type="b", lty=2,
+ pch=20, col="black",
+ xlab="Logarithmierte Überlebenszeit",
+ ylab="Log-kumulierte Hazard-Funktion")
lines(log.times.hb.leq7,log.H.hb.leq7, type="b", lty=2,
+ pch=20, col="magenta")
lines(log.times.hb.g7.leq10,log.H.hb.g7.leq10, type="b", lty=2,
+ pch=20,col="blue")
lines(log.times.hb.g13,log.H.hb.g13, type="b", lty=2,
+ pch=20, col="green")
```

Literaturverzeichnis

[And-82] Andersen, P. K. & Gill, R. D. (1982). Cox's Regression Model for Counting Processes: A Large Sample Study. *Annals of Statistics*, **10**, 1100–1120.

[And-93] Andersen, P. K., Borgan, Ø., Gill, R. D. & Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer, New York.

[AnH-85] Andrews, D. F. & Herzberg, A. M. (1985). *Data*. Springer, New York.

[Arn-92] Arnold, B. C., Balakrishnan, N., Nagaraja, H. N. (1992). *A First Course in Order Statistics*. John Wiley and Sons, New York.

[Ber-64] Berretoni, J. N. (1964). Practical Applications of the Weibull Distribution. *Industrial Quality Control*, **21**, 71–79.

[Bor-90] Borgan, Ø. & Leistøl, K. (1990). A Note on Confidence Bands for the Survival Curve Based on Transformations. *Scandinavian Journal of Statistics*, **17**, 35–41.

[Bre-74] Breslow, N. (1974). Covariance Analysis of Censored Data. *Biometrics*, **30**, 89–99.

[Col-03] Collet, D. (2003). *Modelling Survival Data in Medical Research*. 2nd ed. Chapmann and Hall, Boca Raton – London – New York – Washington, D.C.

[Cox-68] Cox, D. R. & Snell, E. J. (1968) A General Definition of Residuals (with Discussion). *Journal of the Royal Statistical Society, A*, **30**, 248–275.

[Cox-72] Cox, D. R. (1972). Regression Models and Life Tables. *Journal of the Royal Statistical Society, B*, **34**, 103–110.

[Cox-75] Cox, D. R. (1975). Partial Likelihood. *Biometrika*, **62**, 269–279.

[Cox-84] Cox, D. R. & Oakes, D. (1984). *Analysis of Survival Data*. Chapman and Hall, London.

[Dav-52] Davis, D. J. (1952). An Analysis of Some Failure Time Data. *Journal of the American Statistical Association*, **47**, 113–150.

[Dol-71] Doll, R. (1971). The Age Distribution of Cancer: Implications for Models of Carcinogens. *Journal of Royal Statistical Society*, A, **134**, 133–166.

[Eps-54] Epstein, B. & Sobel, M. (1954). Some Theorems Relevant to Life Testing from an Exponential Distribution. *Annals of Mathematical Statistics*, **25**, 373–381.

[Eps-58] Epstein, B. (1958). The Exponential Distribution and Its Role in Life Testing. *Industrial Quality Control*, **15**, 2–7.

[Fle-91] Fleming, T. R. & Harrington, D. P. (1991). *Counting Processes and Survival Analysis*. John Wiley and Sons, New York.

[Gil-01] Gill, R. D. (2001). Product Integration. Mathematical Institute, University of Utrecht, Netherlands, EURANDOM, Eindhoven, Netherlands. `http://www.math.uu.nl/people/gill/Preprints/prod_int_0.pdf`

[Gre-26] Greenwood, M. (1926). The Natural Duration of Cancer. *Reports on Puplic Health and Medical Subjects*, **33**, 1–26, Her Majesty's Stationery Office, London.

[Haf-01] Hafner, R. (2001). *Nichtparametrische Verfahren der Statistik*. Springer, Wien – New York.

[Kab-96] Kaballo, W. (1996). *Einführung in die Analysis I*. Spektrum, Heidelberg – Berlin – Oxford.

[Kal-80] Kalbfleisch, J. D. & Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data*. John Wiley and Sons, New York.

[Kal-02] Kalbfleisch, J. D. & Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. 2nd ed. John Wiley and Sons, New York.

[Kao-59] Kao, J. H. K. (1959). A Graphical Estimation of Mixed Weibull Parameters in Life Testing Electron Tubes. *Technometrics*, **1**, 389–407.

[Kap-58] Kaplan, E. L. & Meier, P. (1958). Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, **53**, 457–481.

[Kle-97] Klein, J. P. & Moeschberger, M. L. (1997). *Survival Analysis*. Springer, New York.

[Kra-75] Krall, J. M., Ulthoff, V. A. & Harley, J. B. (1975). A Step-up Procedure for Selecting Variables Associated with Survival. *Biometrics*, **31**, 49–57.

[Law-03] Lawless, J. F. (2003). *Statistical Models and Methods for Lifetime Data*. 2nd ed. John Wiley and Sons, New Jersey.

[Lea-87] Leathem, A. J. & Brooks, S.A. (1987). Predictive Value of Lecitin Binding on Breast Cancer Recurrance and Survival. *The Lancet*, **I**, 1054–1056.

[Lee-03] Lee, E. T. & Wenyu Wang, J. (2003). *Statistical Methods for Survival Data Analysis*. 3rd ed. John Wiley and Sons, New Jersey.

[Lin-96] Lindsey, J. K. (1996). *Parametric Statistical Inference*. Oxford University Press, New York.

[Mue-05] Müller, Ch. (2005). *Einführung in die Stochastik*. Skript zur Vorlesung im Sommersemester 2005. Universität Oldenburg.

[Paw-01] Pawitan, Y. (2001). *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford University Press, New York.

[Pet-72] Peto, R. (1972). Discussion on Professor Cox's Paper. *Journal of the Royal Statistical Society, B*, **34**, 103–110.

[Pfe-05] Pfeifer, D. (2005). *Analytische Prinzipien der Stochastik*. Skript zur Vorlesung im Wintersemester 2004/2005.

[Psc-04] Pschyrembel, W. [Hrsg.] (2004). *Klinisches Wörterbuch*. 260. Auflage. W. de Gruyter, Berlin.

[Rao-73] Rao, C. R. (1973). *Linear Statistical Inference and Its Applications*. 2nd ed. John Wiley and Sons, New York.

[Ros-33] Rosen, P. & Rammler, B. (1933). The Laws Governing the Fitness of Powdered Coal. *Journal of Inst. Fuels*, **6**, 29–36.

[Sac-06] Sachs, L. & Hedderich, J. (2006). *Angewandte Statistik – Methodensammlung mit R*. 12. Auflage. Springer, Berlin.

[Sch-95] Schervish, M. J. (1995). *Theory of Statistics*. Springer, New York.

[Tsi-81] Tsiatis, A. (1981). A Large Sample Study of Cox's Regression Model. *Annals of Statistics*, **9**, 93–108.

[VAC-67] Veteran's Administration Cooperative Urological Research Group (1967). Treatment and Survival of Men with Cancer of the Prostate. *Surgical Gynecology and Obstetrics*, **124**, 1011–1017.

[Vol-1887] Volterra, V. (1887). Sulle equazione differzeniali lineari. Accademia dei lincei 3.

[Wei-51] Weibull, W. (1951). A Statistical Distribution of Wide Applicability. *Journal of Applied Mechanics*, **18**, 293–297.

[WHO-87] World Health Organisation (1987). Special Programme of Research, Development and Research Training in Human Reproduction. Vaginal Bleeding Patterns – The Problem and an Example Data Set.

[Wit-95] Witting, H., Müller-Funk, U. (1995). *Mathematische Statistik II*. B. G. Teubner, Stuttgart.

Erklärung

Hiermit versichere ich, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel und Quellen benutzt habe.

Oldenburg, 15. Juni 2007 _____