

Technische Universität München

ZENTRUM MATHEMATIK

**Neuere statistische Methoden in
der Pflegeversicherung**

Diplomarbeit

von

Susanne Gschlößl

Themensteller: Prof. Dr. Czado

Betreuer: Prof. Dr. Czado

Abgabetermin: 1. Juli 2002

Hiermit erkläre ich, daß ich die Diplomarbeit selbständig angefertigt und nur die angegebenen Quellen verwendet habe.

München, 1. Juli 2002

Danksagung

Meinen Dank möchte ich vor allem Frau Professor Czado für die sehr gute und intensive Betreuung und die vielen Anregungen, die entscheidend zur Entstehung dieser Arbeit beigetragen haben, aussprechen. Desweiteren gebührt mein Dank Frau Ursula Berger für die Informationen zur Modellierung zeitabhängiger Effekte und Herrn Professor Schmidt für die Beantwortung medizinischer Fragen. Nicht zuletzt danke ich Herrn Florian Rudolph, der bei Rückfragen zu den Daten und zu seiner Diplomarbeit hilfsbereit zur Verfügung stand.

Inhaltsverzeichnis

| | | |
|----------|---|-----------|
| 1 | Einleitung | 1 |
| 2 | Theoretische Grundlagen | 3 |
| 3 | Grundlagen der Überlebenszeitanalyse | 7 |
| 3.1 | Die Überlebensfunktion | 7 |
| 3.2 | Die Hazardfunktion | 8 |
| 3.3 | Zensierung und Trunkierung | 9 |
| 3.3.1 | Rechtszensierung | 9 |
| 3.3.2 | Links- und Intervallzensierung | 10 |
| 3.3.3 | Trunkierung | 11 |
| 3.4 | Berechnung des Likelihood für zensierte Daten | 11 |
| 3.5 | Schätzen des kumulierten Hazards und der Überlebensfunktion | 12 |
| 3.5.1 | Der Nelson - Aalen - Schätzer für $\Lambda(t)$ | 12 |
| 3.5.2 | Der Breslow- und Kaplan-Meier-Schätzer für $S(t)$ | 13 |
| 4 | Das Cox Proportional Hazard Modell | 15 |
| 4.1 | Partieller Log-Likelihood | 16 |
| 4.2 | Motivation des Partial-Likelihood | 17 |
| 4.3 | Statistische Hypothesentests | 21 |
| 4.3.1 | Der partielle Likelihood-Ratio Test | 21 |
| 4.3.2 | Der Wald-Test | 21 |
| 4.3.3 | Der Score-Test | 21 |
| 4.4 | Lokale Tests | 22 |
| 4.5 | Das AIC | 23 |
| 4.6 | Der Breslow-Schätzer für den Baselinehazard | 23 |
| 4.7 | Berechnung des Likelihood mit Ties | 25 |
| 4.8 | Bestimmung der funktionalen Form stetiger Kovariablen | 28 |
| 4.8.1 | Martingalresiduen | 28 |
| 4.8.2 | Martingalresiduen zur Bestimmung der funktionalen Form | 30 |
| 4.8.3 | Fraktionale Polynome | 36 |
| 4.9 | Zeitabhängige Koeffizienten | 37 |
| 4.9.1 | Schoenfeld-Residuen | 38 |
| 4.9.2 | Überprüfung von Zeitabhängigkeit mittels Schoenfeldresiduen | 39 |

| | | |
|----------|--|------------|
| 4.9.3 | Modellierung zeitabhängiger Koeffizienten mittels fraktionaler Polynome | 41 |
| 5 | Generalisierte Lineare Modelle (GLM) | 43 |
| 5.1 | Klassische Lineare Modelle | 43 |
| 5.2 | Exponentialfamilien | 44 |
| 5.3 | Generalisierte Lineare Modelle | 46 |
| 5.3.1 | Bestimmung der Anpassungsgüte bei GLM's | 46 |
| 5.3.2 | Statistische Hypothesentests | 47 |
| 5.4 | Stückweises Exponentialmodell und Poissonmodell | 47 |
| 6 | Überlebenszeitanalyse in der Pflegeversicherung | 50 |
| 6.1 | Beschreibung der Daten | 50 |
| 6.2 | Modellierung der Überlebenszeiten unter Berücksichtigung der Diagnosen . | 54 |
| 6.3 | Modellierung von Interaktionen | 56 |
| 6.4 | Interpretation des Modells (6.4) | 63 |
| 6.5 | Bestimmung der funktionalen Form der Kovariable Alter | 70 |
| 6.5.1 | Untersuchung der Modelle mit \mathbf{Z}_{θ_m} und \mathbf{Z}_{θ_w} | 71 |
| 6.5.2 | Exponentieller Ansatz für das Alter | 73 |
| 6.5.3 | Modellierung des Alters durch fraktionale Polynome | 78 |
| 6.5.4 | Vergleich der Modelle | 80 |
| 6.6 | Untersuchung der Zeitabhängigkeit der Koeffizienten | 84 |
| 6.6.1 | Illustration: Modellierung der Zeitabhängigkeit mittels fraktionaler Polynome | 89 |
| 7 | Übergangsmodele in der Pflegeversicherung | 93 |
| 7.1 | Modellierung der Übergänge zwischen den Behandlungsarten | 93 |
| 7.2 | Modellierung der Übergänge zwischen den Behandlungsstufen | 95 |
| 7.3 | Berechnung von Übergangswahrscheinlichkeiten | 98 |
| 8 | Anpassung eines Poissonmodells | 111 |
| 9 | Zusammenfassung | 123 |
| | Literaturverzeichnis | 125 |
| | Anhang | 127 |

Kapitel 1

Einleitung

Studien haben ergeben, dass das zu erwartende Durchschnittsalter der Bevölkerung in den letzten Jahrzehnten stark angestiegen ist und auch weiter zunehmen wird. Mit einem wachsenden Anteil älterer Personen, muss auch für deren Versorgung und Pflege im Alter oder bei Krankheit gesorgt werden. Der Pflegeversicherung kommt hiermit eine immer größere Bedeutung zu. Auch die Versicherung muß sich auf diese Bevölkerungsentwicklung bei der Kalkulierung ihrer Tarife und Prämien einstellen. Als Grundlage werden dazu Sterbewahrscheinlichkeiten verschiedener Bevölkerungsgruppen, in der Regel wird nach Geschlecht und Alter unterschieden, benötigt. In der Pflegeversicherung sind darüber hinaus auch andere Faktoren, wie zum Beispiel die Art der Pflege und die Schwere des Pflegefalls für die Überlebenswahrscheinlichkeit von großer Bedeutung.

Statistische Methoden der Überlebenszeitanalyse ermöglichen es, auf Basis von Daten von Pflegebedürftigen das Überleben in Abhängigkeit dieser Faktoren zu modellieren und dann die Wahrscheinlichkeiten zu schätzen.

Dieses Thema wurde bereits im Rahmen einer Diplomarbeit von Florian Rudolph [23] aufgegriffen und behandelt. In die Modellbildung gingen die Faktoren Alter, Geschlecht und Pflegestatus ein. Nicht berücksichtigt wurden dabei jedoch die pflegeauslösenden Diagnosen der Versicherungsnehmer. Sicherlich kann ein genaueres Bild der Realität wiedergegeben werden, wenn diese Diagnosen zusätzlich in die Modellbildung mit eingehen. Dies ist ein Ziel dieser Arbeit, die wie folgt aufgebaut ist.

Im zweiten Kapitel werden für die Arbeit benötigte Begriffe aus der Wahrscheinlichkeitstheorie, wie z.B. Martingale und Zählprozesse, definiert und erklärt. Das dritte Kapitel behandelt die theoretischen Grundlagen der Überlebenszeitanalyse. Es wird auch auf Zensierungstechniken und Schätzer der Überlebensfunktion eingegangen. In Kapitel 4 wird das Proportional Hazard Modell vorgestellt, welches zur Modellierung verwendet werden soll. Besonderes Augenmerk wird auf die Überprüfung der Modellannahmen gelegt. Es wird auf verschiedene neuere Methoden zur Aufdeckung und Behebung von Verletzungen dieser Annahmen eingegangen.

Anschließend wird in Kapitel 5 eine kurze Einführung in Generalisierte Lineare Modelle gegeben. Es wird gezeigt, dass Daten aus der Überlebenszeitanalyse alternativ zum Proportional Hazard Modell auch durch ein Poissonmodell angepasst werden können.

In Kapitel 6 wird die zuvor erarbeitete Theorie auf den schon von Rudolph [23] behandelten Datensatz aus der Pflegeversicherung angewandt. Im Mittelpunkt steht hierbei

die Einbindung der Diagnosen und die Überprüfung der Modellannahmen. In Kapitel 7 werden aus dem gefundenen Modell Sterbe- und Übergangswahrscheinlichkeiten zwischen verschiedenen Pflegestufen berechnet, die von Versicherungen zur Kalkulierung benötigt werden.

Schließlich wird in Kapitel 8 eine Anpassung durch ein Poissonmodell auf einen Teildatensatz durchgeführt. Die Ergebnisse werden mit denen eines Proportional-Hazard-Modells verglichen. Obwohl im Poissonmodell eine einfache Modellierung zeitabhängiger Koeffizienten möglich ist, erlaubt das Proportional Hazard Modell eine flexiblere Modellanpassung. Abschließend werden in Kapitel 9 die erarbeiteten Ergebnisse noch einmal zusammengefaßt. Werden die Diagnosen als zusätzliche Einflußfaktoren im Modell berücksichtigt, können die Überlebenswahrscheinlichkeiten erheblich besser geschätzt werden. Auch die Anwendung der neueren Methoden zur Überprüfung der Modellannahmen trägt entscheidend zur Verbesserung des Modells bei.

Kapitel 2

Theoretische Grundlagen

In diesem Kapitel werden die theoretischen Grundlagen bereitgestellt, die zum Verständnis der Überlebenszeitanalyse und des Proportional Hazard Modells benötigt werden. Neben Begriffen aus der Wahrscheinlichkeitstheorie wird kurz auf Zählprozesse und stochastische Integrale eingegangen. Die Darstellung ist an den Büchern von Anderson et al. [2] und Fleming und Harrington [8] orientiert. Hier findet sich auch eine ausführliche Behandlung von Zählprozessen in der Überlebenszeitanalyse. Die Definitionen zur stochastischen Integration sind den Büchern von Kopp [15] und Weiszäcker und Winkler [29] entnommen. Stochastische Integrale ermöglichen eine einfache Darstellung rechtszensierter Daten. Sie finden beim Nelson-Aalen-Schätzer der kumulierten Hazardfunktion und Martingalresiduen Anwendung und werden zur Bestimmung des asymptotischen Verhaltens benötigt. Im folgenden sei $\mathcal{T} = [0, \tau)$, $0 < \tau \leq \infty$ und (Ω, \mathcal{F}, P) ein Wahrscheinlichkeitsraum.

Definition 2.1 (Filtration) Eine Filtration $\{\mathcal{F}_t : t \in \mathcal{T}\}$ ist eine wachsende Familie von Sub- σ -Algebren der σ -Algebra \mathcal{F} , d.h. es gilt

$$\mathcal{F}_s \subseteq \mathcal{F}_t \subseteq \mathcal{F} \quad \forall s < t$$

\mathcal{F}_{t-} ist die kleinste σ -Algebra, die alle Mengen in $\bigcup_{s < t} \mathcal{F}_s$ enthält, d.h. $\mathcal{F}_{t-} = \sigma(\bigcup_{s < t} \mathcal{F}_s)$.

Eine Filtration heißt **rechtsstetig**, wenn gilt

$$\mathcal{F}_s = \bigcap_{t > s} \mathcal{F}_t \quad \forall s$$

Eine Filtration heißt **vollständig**, wenn \mathcal{F} vollständig ist (d.h. \mathcal{F} enthält alle Untermengen einer P -Nullmenge) und wenn die σ -Algebra \mathcal{F}_t für jedes $t \in \mathcal{T}$ alle P -Nullmengen von \mathcal{F} enthält, d.h.

$$A \subset B \in \mathcal{F}, \quad P(B) = 0 \quad \Rightarrow \quad A \in \mathcal{F}_0.$$

Definition 2.2 (Stochastischer Prozeß) Ein stochastischer Prozeß X ist eine Familie von Zufallsvariablen $X = \{X(t) : t \in \mathcal{T}\}$, die alle auf dem selben Wahrscheinlichkeitsraum (Ω, \mathcal{F}, P) definiert sind.

Der Prozeß X heißt **adaptiert** zu einer Filtration $\{\mathcal{F}_t : t \in \mathcal{T}\}$, wenn $X(t)$ \mathcal{F}_t -meßbar

ist für jedes $t \in \mathcal{T}$.

$X(t, \omega)$ heißt **Realisierung** von $X(t)$ am Punkt $\omega \in \Omega$. Die Abbildung $t \mapsto X(t, \omega)$, $\omega \in \Omega$ heißt **Pfad** von X .

Der Prozeß X heißt **cadlag** (continu à droite, limité à gauche), wenn die Pfade $\{X(t, \omega) : t \in \mathcal{T}\}$ für fast alle $\omega \in \Omega$ rechtsstetig sind und der linksseitige Grenzwert $\lim_{h \rightarrow 0} X(t - h, \omega)$ existiert.

Bemerkungen: 1) Oft betrachtet man Filtrationen, die von einem stochastischen Prozeß X erzeugt werden, d.h. $\mathcal{F}_t = \sigma(X(s) : s \leq t)$.

2) \mathcal{F}_t enthält alle Informationen über den Prozeß X bis zum Zeitpunkt t und wird daher auch **Geschichte** des Prozesses genannt. $\mathcal{F}_{t-} = \sigma(X(s) : s \leq t)$ ist die Geschichte des Prozesses bis kurz vor t .

Definition 2.3 (Martingal) Ein Martingal ist ein cadlag adaptierter Prozeß $M = \{M(t) : t \in \mathcal{T}\}$, der integrierbar ist, d.h.

$$E(|M(t)|) < \infty \quad \forall t \in \mathcal{T}$$

und der Martingal-Eigenschaft

$$E(M(t)|\mathcal{F}_s) = M(s) \quad \forall s \leq t \tag{2.1}$$

genügt.

Der Prozeß M heißt **Submartingal**, wenn gilt :

$$E(M(t)|\mathcal{F}_s) \geq M(s) \quad \forall s \leq t \quad ,$$

M heißt **Supermartingal**, wenn gilt:

$$E(M(t)|\mathcal{F}_s) \leq M(s) \quad \forall s \leq t \quad .$$

Eine äquivalente Formulierung der Martingal-Eigenschaft ist

$$E(dM(t)|\mathcal{F}_{t-}) = 0 \quad \forall t \in \mathcal{T} \quad , \tag{2.2}$$

wobei $dM(t) = M[(t + dt)-] - M(t-)$ die Änderung des Prozesses $M(t)$ über ein kurzes Zeitintervall $[t, t + dt)$ ist. Denn gilt (2.2), dann

$$\begin{aligned} E(M(t)|\mathcal{F}_s) - M(s) &= E(M(t) - M(s)|\mathcal{F}_s) \\ &= E\left(\int_0^t dM(u)|\mathcal{F}_s\right) = \int_0^t E[E[dM(u)|\mathcal{F}_{u-}]|\mathcal{F}_s] = 0 \end{aligned}$$

Definition 2.4 (Zählprozeß) Ein Zählprozeß ist ein stochastischer Prozeß $\{N(t) : t \in \mathcal{T}\}$, adaptiert zu einer Filtration $\{\mathcal{F}_t : t \in \mathcal{T}\}$ mit

- $N(0) = 0$

- $P(N(t) < \infty) = 1$
- Pfade sind mit Wahrscheinlichkeit 1 rechtsstetig, stückweise konstant und haben nur Unstetigkeiten an Sprungstellen mit Sprunghöhe 1.

Definition 2.5 (vorhersagbarer Prozeß) Ein Prozeß X heißt **vorhersagbar**, wenn er als Funktion von $(t, \omega) \in \mathcal{T} \times \Omega$ meßbar ist bezgl. der σ -Algebra, die von der Klasse der auf $\mathcal{T} \times \Omega$ linksstetigen, adaptierten Prozesse erzeugt wird.

Einfacher ausgedrückt ist ein vorhersagbarer Prozeß ein Prozeß, dessen Verhalten zum Zeitpunkt t für jedes $t \in \mathcal{T}$ festgelegt ist durch sein Verhalten im Intervall $[0, t)$. D.h. $X(t)$ ist \mathcal{F}_t -meßbar für $\mathcal{F}_s = \sigma(X(u), 0 \leq u \leq s)$.

Definition 2.6 (gleichmässig integrierbar) Ein stochastischer Prozeß $X = \{X(t) : t \in \mathcal{T}\}$ heißt **gleichmässig integrierbar** wenn gilt,

$$\lim_{n \rightarrow \infty} \sup_{t \in \mathcal{T}} E[|X(t)| I(|X(t)| > n)] = 0.$$

Satz 2.7 (Doob-Meyer Zerlegung) X sei ein rechtsstetiges, nichtnegatives Submartingal bzgl. eines Wahrscheinlichkeitsraumes (Ω, \mathcal{F}, P) mit der rechtsstetigen Filtration $\{\mathcal{F}_t : t \in \mathcal{T}\}$. Dann existieren ein rechtsstetiges Martingal M und ein wachsender, rechtsstetiger, vorhersagbarer Prozeß A mit $E[A(t)] < \infty$ und

$$X(t) = M(t) + A(t) \quad \text{fast sicher} \quad \forall t \in \mathcal{T} \quad .$$

Ist $A(0) = 0$ fast sicher und $X = M' + A'$ eine weitere solche Zerlegung mit $A'(0) = 0$, dann gilt für jedes $t \in \mathcal{T}$

$$P(M'(t) \neq M(t)) = 0 = P(A'(t) \neq A(t)) \quad .$$

Ist X zusätzlich beschränkt, dann ist M gleichmässig integrierbar und A ist integrierbar.

Definition 2.8 (Stoppzeit) Eine Stoppzeit T ist eine Zufallsvariable mit Werten in $[0, \tau]$, so dass $\{T \leq t\} \in \mathcal{F}_t \quad \forall t \in [0, \tau)$.

Definition 2.9 (Elementarprozess) Ein Prozeß H heißt **Elementarprozess** bezüglich der Filtration $\{\mathcal{F}_t : t \in \mathcal{T}\}$, wenn er beschränkt und \mathcal{F}_t -adaptiert ist und wenn es ein $n \in \mathbb{N}$ gibt, so dass jeder Pfad von H eine linksstetige Treppenfunktion mit höchstens n Sprüngen ist. Ein Elementarprozess kann dargestellt werden in der Form

$$H = H_0 1_{\{0\}} + \sum_{i=1}^n H_i 1_{]T_i, T_{i+1}]}$$

wobei H_0 \mathcal{F}_0 -meßbar, $(T_i)_{1 \leq i \leq n}$ eine wachsende Folge von Stoppzeiten und H_i \mathcal{F}_{T_i} -meßbar ist für jedes i .

Definition 2.10 (Stochastisches Integral für Elementarprozesse) Sei H ein elementarer, vorhersagbarer Prozeß auf $[0, \infty)$ und $0 = t_0 < t_1 < \dots < t_n < t_{n+1} = \infty$ eine Partition von $[0, \infty)$. Weiter sei $H_t = H_{t_i}$ für $t \in]t_i, t_{i+1}]$ $i = 0, \dots, n$, H_{t_i} sei \mathcal{F}_{t_i} -meßbar und beschränkt und H nehme den Wert H_0 für $t = 0$ an. M sei ein L^2 -beschränktes Martingal. Dann ist das **stochastische Integral** $H \cdot M = \int H dM$ definiert durch

$$(H \cdot M)_t = \int_0^t H_s dM_s = H_0 M_0 + \sum_{i=0}^n H_{t_i} (M_{t_{i+1} \wedge t} - M_{t_i \wedge t})$$

für $t \in [0, \infty)$.

Kapitel 3

Grundlagen der Überlebenszeitanalyse

In diesem Kapitel werden die Grundlagen der Überlebenszeitanalyse behandelt. Dazu gehören die Überlebensfunktion und die Hazardfunktion. Als zugrundeliegende Literatur wurden die Bücher von Klein und Moeschberger [14], Hosmer und Lemeshow [12] und Therneau und Grambsch [26] verwendet.

3.1 Die Überlebensfunktion

Definition 3.1 (Überlebensfunktion) Sei die Zufallsvariable $T \geq 0$ die Überlebenszeit in einer Population mit Verteilungsfunktion $F(t)$ und Dichtefunktion $f(t)$. Dann ist die Wahrscheinlichkeit eines Individuums, mindestens bis zum Zeitpunkt t zu überleben, definiert als

$$S(t) = P(T > t) = \int_t^{\infty} f(u)du, \quad (3.1)$$

Bemerkung: $S(t)$ ist eine monoton fallende Funktion, die in 0 den Wert 1 annimmt und im Unendlichen gegen 0 geht. Sie kann auch als

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

formuliert werden. Daraus folgt

$$\frac{dS(t)}{dt} = -f(t). \quad (3.2)$$

Für eine diskrete Zufallsvariable T mit den Werten $t_1 \leq t_2 \leq \dots$ ist die Überlebensfunktion folgendermaßen definiert:

$$S(t) = \sum_{t_j > t} p(t_j) = \sum_{t_j > t} P(X = t_j)$$

3.2 Die Hazardfunktion

Definition 3.2 (Hazardfunktion) Die Hazardfunktion für stetiges T ist definiert durch

$$\lambda(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T \leq t + \Delta t | T \geq t)}{\Delta t} \quad (3.3)$$

Sie kann als das Risiko eines Individuums zum Zeitpunkt t im nächsten Augenblick zu sterben, gegeben Überleben bis zum Zeitpunkt t , interpretiert werden. Aus (3.3) ergibt sich mit (3.1) und (3.2) der Zusammenhang

$$\begin{aligned} \lambda(t) &= \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T \leq t + \Delta t | T \geq t)}{\Delta t} \\ &= \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T \leq t + \Delta t)}{P(T \geq t) \cdot \Delta t} \\ &= \frac{1}{P(T \geq t)} \cdot \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T \leq t + \Delta t)}{\Delta t} \\ &= \frac{1}{P(T \geq t)} \cdot f(t) = \frac{f(t)}{S(t)} \\ &= -\frac{\frac{dS(t)}{dt}}{S(t)} = -\frac{d \ln S(t)}{dt} \end{aligned}$$

Definition 3.3 (kumulierte Hazardfunktion) Die kumulierte Hazardfunktion für stetiges T ist definiert durch

$$\Lambda(t) = \int_0^t \lambda(u) du = -[\ln S(t) - \ln S(0)] = -\ln S(t) \quad (3.4)$$

Danach lässt sich $S(t)$ auch als

$$S(t) = \exp[-\Lambda(t)] = \exp\left[-\int_0^t \lambda(u) du\right] \quad (3.5)$$

darstellen. Im diskreten Fall ist die Hazardfunktion wie folgt erklärt:

$$\lambda(t_j) = P(T = t_j | T \geq t_j) = \frac{P(T = t_j)}{P(T \geq t_j)} = \frac{p(t_j)}{S(t_{j-1})} \quad (3.6)$$

mit $S(t_0) = 1$. Wegen (3.6) und

$$p(t_j) = P(T \geq t_{j-1}) - P(T \geq t_j) = S(t_{j-1}) - S(t_j), \quad j = 1, 2, \dots$$

gilt

$$\lambda(t_j) = \frac{S(t_{j-1}) - S(t_j)}{S(t_{j-1})} = 1 - \frac{S(t_j)}{S(t_{j-1})}$$

Da

$$S(t) = \prod_{t_j \leq t} \frac{S(t_j)}{S(t_{j-1})}$$

ergibt sich mit

$$S(t) = \prod_{t_j \leq t} (1 - \lambda(t_j)) \quad (3.7)$$

auch im diskreten Fall die Verbindung der Überlebensfunktion zur Hazardfunktion hergestellt.

3.3 Zensierung und Trunkierung

Eine Besonderheit bei Daten in der Überlebenszeitanalyse sind Zensierungen. Zensierte Daten sind Beobachtungen, bei denen nur Teilinformationen über das interessierende Ereignis vorhanden sind. Man weiß beispielsweise nur, dass das Ereignis, z.B. der Tod, innerhalb eines bestimmten Zeitraumes eingetreten ist. Es gibt verschiedene Formen der Zensierung, man unterscheidet zwischen Rechts-, Links- und Intervallzensierung. Dieses Kapitel orientiert sich an den Büchern von Klein und Moeschberger [14] und Miller [18].

3.3.1 Rechtszensierung

Bei dieser Art der Zensierung wird das interessierende Ereignis nur beobachtet, wenn es vor einer vorher festgelegten Zeit, z.B. noch vor dem Ende einer Studie, stattfindet. Die Überlebenszeit von Individuum i wird mit X_i bezeichnet. X_1, X_2, \dots, X_n seien unabhängig, identisch verteilt mit Dichte $f(x)$ und Überlebensfunktion $S(x)$. C_r bezeichne die fixe Zensierungszeit, der Index r steht hier für Rechtszensierung. Die exakte Überlebenszeit des i -ten Individuums kann nur dann beobachtet werden, wenn das Ereignis, der Tod, spätestens zum Zensierungszeitpunkt eintritt, d.h. wenn gilt $X_i \leq C_r$. Ist dagegen $X_i > C_r$, überlebt das Individuum das Ende der Studie und seine Überlebenszeit wird zum Zeitpunkt C_r zensiert. Die Daten lassen sich daher durch Paare von Zufallsvariablen der Form (T_i, δ_i) darstellen, wobei

$$T_i = \min(X_i, C_r)$$

die Überlebenszeit X_i annimmt, falls das Ereignis beim i -ten Individuum beobachtet werden konnte und den Wert C_r , falls die i -te Beobachtung zensiert wurde.

$$\delta_i = I(X_i \leq C_r) = \begin{cases} 1 & \text{falls das Ereignis beobachtet wurde} \\ 0 & \text{falls die Beobachtung zensiert wurde} \end{cases}$$

ist eine Indikatorvariable, die angibt, ob es sich bei T_i um eine Ereignis- oder eine Zensierungszeit handelt.

Eine Verallgemeinerung dieser Zensierungsform liegt vor, wenn Individuen zu unterschiedlichen Zeitpunkten in eine Studie eintreten und das Ende der Studie schon im voraus bekannt ist. In diesem Fall hat jedes Individuum seine eigene, feste Zensierungszeit C_{ri} . Verlässt ein Individuum die Studie vor dem Ende, so dass das interessierende Ereignis nicht

mehr beobachtet werden kann, spricht man von *zufälliger Zensierung*. Mögliche Gründe für ein vorzeitiges Ausscheiden aus der Studie können zum Beispiel der Umzug eines Patienten in eine andere Stadt oder der Tod durch eine nicht in der Studie untersuchte Ursache, z. B. durch einen Autounfall, sein.

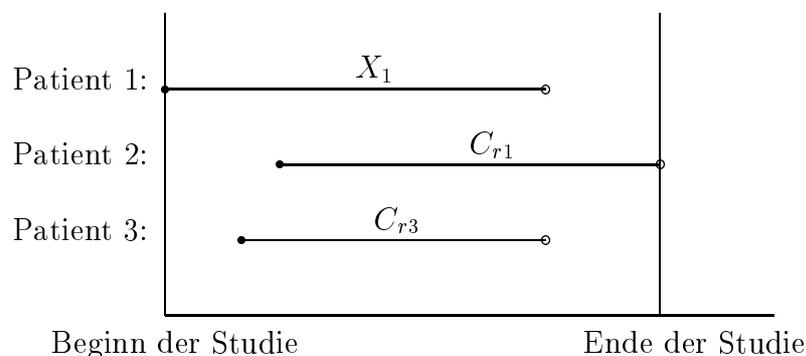


Abbildung 3.1: Verschiedene Zensierungsmöglichkeiten

In Abbildung (3.1) sind die verschiedenen Fälle dargestellt. Patient 1 wird von Studienbeginn an beobachtet und stirbt zum Zeitpunkt X_1 . Patient 2 tritt erst später in die Studie ein und ist zum Studienende immer noch am Leben, er wird daher zensiert. Ein Beispiel für zufällige Zensierung ist Patient 3. Auch er tritt verspätet in die Studie ein, verlässt sie aber auch noch vor dem Ende, ohne das interessierende Ereignis erfahren zu haben.

3.3.2 Links- und Intervallzensierung

Von Linkszensierung spricht man, wenn das Ereignis schon stattfand, bevor das Individuum in die Studie eingetreten ist. Auch in diesem Fall ist die exakte Ereigniszeit unbekannt. Ein Beispiel wäre eine Studie, die untersucht, wann Kinder eine bestimmte Fähigkeit erlernen. Kinder, die diese Fähigkeit schon zu Beginn der Studie besitzen, werden als linkszensiert betrachtet. Der Zensierungszeitpunkt, der Studienbeginn, wird mit C_l bezeichnet. Bei Linkszensierung lassen sich die Daten als Paare (T_i, ε_i) darstellen.

$$T_i = \max(X_i, C_l)$$

gibt die Laufzeit X_i an, falls das Ereignis bei Individuum i beobachtet werden konnte, und den Wert C_l , falls das Ereignis schon vor Studienbeginn stattfand. Die Indikatorvariable

$$\varepsilon_i = I(C_l \leq X_i) = \begin{cases} 1 & \text{falls das Ereignis beobachtet wurde} \\ 0 & \text{falls die Beobachtung zensiert wurde} \end{cases}$$

ist analog zur Rechtszensierung definiert.

Bei Intervallzensierung ist nur bekannt, ob das Ereignis innerhalb eines bestimmten Intervalls $(L_i, R_i]$ stattgefunden hat. Dies ist zum Beispiel der Fall, wenn Patienten nur in

regelmäßigen Abständen untersucht werden. Wird bei einer jährlichen Routineuntersuchung das interessierende Ereignis, z.B. Entstehen eines Tumors, festgestellt, so weiß man nur, dass dieses innerhalb des letzten Jahres eingetreten ist. Der genaue Zeitpunkt ist unbekannt.

3.3.3 Trunkierung

Bei Trunkierung werden nur Individuen erfasst, die eine bestimmte Bedingung erfüllt haben. Man unterscheidet wieder Rechts- und Linkstrunkierung.

Linkstrunkierung liegt vor, wenn nur Individuen beobachtet werden, die ein bestimmtes Ereignis vor dem interessierenden Ereignis erfahren haben. So kann etwa eine Studie, die das Überleben von Personen in einem Seniorenheim untersucht, nur Individuen beobachten, die schon in einem Seniorenheim leben. Alle Individuen, die vor einem Ereignis Eintritt in ein Seniorenheim sterben, sind dem Beobachter unbekannt. Im Unterschied zur Linkszensierung, wo über Patienten, die vor Studienbeginn das Ereignis erfahren, Informationen vorliegen, werden hier diese Individuen nie in der Studie berücksichtigt.

Rechtstrunkierung liegt vor, wenn nur Individuen beobachtet werden, die das Ereignis schon erfahren haben. Jedes Individuum, bei dem das Ereignis noch nicht eingetreten ist, wird nicht beobachtet. Ein Beispiel für Rechtstrunkierung wäre eine Studie über Aids, bei der nur Patienten beobachtet werden, bei denen sich die Krankheit schon entwickelt hat.

3.4 Berechnung des Likelihood für zensierte Daten

Liegen zensierte Daten vor, muss das bei der Berechnung des Likelihood berücksichtigt werden. Grundlegend ist die Annahme, dass die Überlebens- und Zensierungszeiten unabhängig voneinander sind. Ist die exakte Überlebenszeit X_i eines Individuums bekannt, so liefert diese Beobachtung zur Likelihoodfunktion den Beitrag

$$P(X = X_i) = f(X_i),$$

also die Wahrscheinlichkeit genau bis zum Zeitpunkt X_i zu überleben. Für rechtszensierte Beobachtungen weiß man nur, dass die Überlebenszeit größer als die Zensierungszeit ist, d.h. es gilt $X_i > C_r$. Der Beitrag rechtszensierter Beobachtungen zur Likelihoodfunktion ist daher

$$P(X_i > C_r) = S(C_r).$$

Linkszensierte Beobachtungen tragen den Term

$$P(X_i \leq C_l) = 1 - P(X_i > C_l) = 1 - S(C_l),$$

die Wahrscheinlichkeit, dass das Individuum das Ereignis schon vor dem Studienbeginn C_l erfahren hat, zur Likelihoodfunktion bei. Der Beitrag intervallzensierter Beobachtungen ist entsprechend

$$P(L_i < X_i \leq R_i) = S(L_i) - S(R_i).$$

Setzt man diese Informationen zusammen, so ergibt sich für die Likelihoodfunktion

$$L \propto \prod_{i \in D} f(x_i) \prod_{i \in R} S(C_r) \prod_{i \in L} [1 - S(C_l)] \prod_{i \in I} [S(L_i) - S(R_i)] \quad (3.8)$$

Dabei bezeichnet D die Menge aller Individuen, bei denen das Ereignis beobachtet wurde, R die Menge aller rechtszensierten, L die Menge aller linkszensierten und I die Menge aller intervallzensierten Beobachtungen.

Likelihood für rechtszensierte Daten

Liegt nur Rechtszensierung vor, werden die Daten durch Paare $(T_i, \delta_i), i = 1, \dots, n$ dargestellt mit der in (3.3.1) festgesetzten Bezeichnung. Gilt $\delta_i = 1$, d.h. das i -te Individuum hat das Ereignis erfahren ($i \in D$), trägt die i -te Beobachtung den Term

$$\begin{aligned} P(T_i, \delta_i = 1) &= P(T_i = X_i | \delta_i = 1) P(\delta_i = 1) = P(T_i = X_i | X_i \leq C_r) P(X_i \leq C_r) \\ &= \frac{f(t_i)}{1 - S(C_r)} (1 - S(C_r)) = f(t_i) \end{aligned}$$

zur Likelihoodfunktion bei.

Für $\delta_i = 0$ dagegen weiß man, dass das i -te Individuum mindestens bis zum Zeitpunkt C_r am Leben war ($i \in R$). Es liefert daher den Beitrag

$$\begin{aligned} P(T_i, \delta_i = 0) &= P(T_i = C_r | \delta_i = 0) P(\delta_i = 0) = P(\delta_i = 0) \\ &= P(X_i > C_r) = S(C_r) \end{aligned}$$

zum Likelihood. Insgesamt ergibt sich daraus für die i -te Beobachtung

$$P(t_i, \delta_i) = f(t_i)^{\delta_i} S(t_i)^{1-\delta_i}$$

Die Likelihoodfunktion lautet also mit $f(t) = \lambda(t)S(t)$

$$L = \prod_{i=1}^n P(t_i, \delta_i) = \prod_{i=1}^n f(t_i)^{\delta_i} S(t_i)^{1-\delta_i} = \prod_{i=1}^n \lambda(t_i)^{\delta_i} S(t_i) \quad (3.9)$$

3.5 Schätzen des kumulierten Hazards und der Überlebensfunktion

In diesem Abschnitt werden Schätzer für die kumulierte Hazardfunktion und die Überlebensfunktion vorgestellt, für den einfachsten Fall, dass keine Kovariablen vorhanden sind. Zudem sollen rechtszensierte Daten vorliegen, d.h. $T_i = \min(X_i, C_{ir}), i = 1, \dots, n$ gibt die Zeit des i -ten Individuums in der Studie bis zum Tod oder Zensierung an. Diesem Abschnitt liegen die Darstellungen bei Therneau und Grambsch [26] und Klein und Moeschberger [14] zugrunde.

3.5.1 Der Nelson - Aalen - Schätzer für $\Lambda(t)$

Der bekannteste Schätzer für die kumulierte Hazardfunktion $\Lambda(t) = \int_0^t \lambda(s) ds$ ist der Nelson - Aalen - Schätzer. $Y_i(t) = I\{T_i \geq t\}$ sei die Indikatorfunktion, die angibt, ob sich das i -te Individuum zur Zeit t noch unter Beobachtung befindet. Von Beginn der Studie

an nimmt $Y_i(t)$ den Wert 1 an, solange bis das i -te Individuum stirbt oder zensiert wird und erhält dann den Wert 0. Der Zählprozess (siehe Kapitel 2, Def. 2.4))

$N_i(t) = I\{T_i \leq t, \delta_i = 1\}$ gibt die Anzahl der Ereignisse des i -ten Individuums an. Solange das i -te Individuum lebt, ist der Wert von $N_i(t) = 0$, nach dem Tod nimmt $N_i(t)$ den Wert 1 an.

Der Nelson-Aalen - Schätzer baut auf den aggregierten Prozessen

$$\bar{Y}(t) = \sum_{i=1}^n Y_i(t) \quad \text{und} \quad \bar{N}(t) = \sum_{i=1}^n N_i(t)$$

auf. $\bar{Y}(t)$ gibt die Anzahl der Individuen an, die sich zur Zeit t in der Risikomenge befinden, $\bar{N}(t)$ zählt die Ereignisse, die bis zur Zeit t , t eingeschlossen, stattfanden.

Für kleine Intervalle der Länge h gilt wegen $\frac{\partial \Lambda(t)}{\partial t} = \lambda(t)$ die Näherung

$$\begin{aligned} \Lambda(s+h) - \Lambda(s) &\approx \lambda(s)h \\ &= P(\text{Individuum stirbt in } (s, s+h] | \text{Individuum ist zur Zeit } s \text{ in Risikomenge}). \end{aligned}$$

Ein naheliegender Schätzer für diese Wahrscheinlichkeit ist durch $\frac{\bar{N}(s+h) - \bar{N}(s)}{\bar{Y}(s)}$ gegeben, das Verhältnis der Anzahl der im Intervall $(s, s+h]$ eingetretenen Ereignisse zur Anzahl der Individuen in der Risikomenge zur Zeit s . Läßt man die Länge der Intervalle gegen 0 gehen, d.h. $h \rightarrow 0$, und summiert über alle Intervalle bis zur Zeit t , erhält man den **Nelson - Aalen - Schätzer**

$$\hat{\Lambda}(t) = \int_0^t \frac{d\bar{N}(s)}{\bar{Y}(s)}, \tag{3.10}$$

ein stochastisches Integral (siehe Kapitel 2) mit dem Zählprozess $\bar{N}(t)$ als Integrator. Eine äquivalente Formulierung von (3.10) ist durch

$$\hat{\Lambda}(t) = \sum_{t_i \leq t} \frac{d\bar{N}(t_i)}{\bar{Y}(t_i)}$$

gegeben, wobei $t_1 < t_2 < \dots$ die geordneten Todeszeitpunkte sind und $d\bar{N}(t) = \bar{N}(t) - \bar{N}(t-)$ die Anzahl der Ereignisse angibt, die genau zum Zeitpunkt t eintreten. Bis zum Eintritt des ersten Ereignisses, d.h. für $t \leq t_1$, gilt $\hat{\Lambda}(t) = 0$.

3.5.2 Der Breslow- und Kaplan-Meier-Schätzer für $S(t)$

Die Überlebensfunktion $S(t)$ ist über den Zusammenhang $S(t) = \exp[-\Lambda(t)]$ (siehe (3.5)) mit der kumulierten Hazardfunktion verbunden.

Daraus leitet sich der **Breslow-Schätzer** für $S(t)$ ab als

$$S_B(t) = \exp[-\hat{\Lambda}(t)] = \exp\left[-\sum_{t_j \leq t} d\hat{\Lambda}(t_j)\right] = \prod_{t_j \leq t} \exp[-d\hat{\Lambda}(t_j)], \tag{3.11}$$

wobei $\hat{\Lambda}(t)$ der in (3.10) definierte Nelson - Aalen - Schätzer und $d\hat{\Lambda}(t_i) = \frac{d\bar{N}(t_i)}{\bar{Y}(t_i)}$ der Zuwachs des Nelson - Aalen - Schätzers für das i -te Ereignis ist.

Ein weiterer Schätzer für die Überlebensfunktion ist der **Kaplan - Meier -Schätzer**, auch Produkt- Limit - Schätzer genannt. Er ergibt sich aus der Formel

$$S(t) = \prod_{t_j \leq t} (1 - \lambda(t_j))$$

(siehe (3.7)) und lautet

$$\hat{S}_{KM}(t) = \begin{cases} 1 & \text{wenn } t \leq t_1 \\ \prod_{t_j \leq t} (1 - d\hat{\Lambda}(t_j)) & \text{wenn } t_1 \leq t \end{cases} . \quad (3.12)$$

Die geschätzte Überlebensfunktion ist eine fallende Treppenfunktion mit Sprüngen bei den beobachteten Todeszeitpunkten. Die Höhe der Sprünge hängt neben der Anzahl der Ereignisse zu jeder Ereigniszeit t_i auch noch von der Größe der Risikomenge, also auch der Anzahl der Zensierungen, ab.

Da für kleine x der Zusammenhang $\exp[-x] \approx 1 - x$ gilt, stimmen der Breslow - Schätzer und der Kaplan - Meier - Schätzer in etwa überein, wenn $d\hat{\Lambda}(t)$ klein ist, d.h. wenn noch viele Individuen unter Beobachtung sind.

Kapitel 4

Das Cox Proportional Hazard Modell

Dieses Kapitel stellt das auf Cox zurückgehende Proportional Hazard Modell vor. Neben der Berechnung des Likelihood werden unter anderem auch Hypothesentests und Residuen für dieses Modell behandelt. Eine einfache Darstellung der Grundlagen des Proportional Hazard Modells findet sich bei Klein und Moeschberger [14]. Hosmer und Lemeshow [12] gehen vor allem auf die Modellentwicklung und die Überprüfung der Modellannahmen ein. Auch im Buch von Grambsch und Therneau [26] sind neben einer sehr guten Darstellung des Modells verschiedene Modellüberprüfungsmethoden aufgeführt. Als weitere zugrundeliegende Literatur ist das Buch von Hougaard [13] zu nennen.

Gegeben seien n Daten der Gestalt $(T_j, \delta_j, \mathbf{Z}_j)$, $j = 1, \dots, n$. Dabei ist T_j die Überlebenszeit des j -ten Patienten und δ_j die zugehörige Zensierungsvariable. Sie nimmt den Wert 1 an, falls das j -te Individuum das Ereignis Tod erfährt, anderenfalls den Wert 0. $\mathbf{Z}_j(t) = (Z_{j1}(t), \dots, Z_{jp}(t))'$ ist der Vektor der Kovariablen des j -ten Individuums, die auch von der Zeit abhängen können.

Dann wird die Hazardfunktion im Proportional Hazard Modell für ein Individuum mit dem Kovariablenvektor $\mathbf{Z}(t)$ zum Zeitpunkt t modelliert durch

$$\lambda(t | \mathbf{Z}(t)) = \lambda_0(t) \exp[\boldsymbol{\beta}'\mathbf{Z}(t)] = \lambda_0(t) \exp\left[\sum_{k=1}^p \beta_k \mathbf{Z}_k(t)\right], \quad (4.1)$$

wobei $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$ der Vektor der unbekanntenen Regressionskoeffizienten ist. $\lambda_0(t)$, der sogenannte Baselinehazard, ist eine nicht näher spezifizierte nicht-negative Funktion. Sie beschreibt die Abhängigkeit des Hazards von der Zeit, ist jedoch unabhängig von Parametern. Da Parameter nur im zweiten Faktor enthalten sind, spricht man bei diesem Modell von einem semiparametrischen Modell. Man ist vor allem daran interessiert, die Effekte der Kovariablen auf das Überleben zu untersuchen, weniger an der genauen Verteilungsfunktion der Überlebenszeit. Man beachte, dass der Log-Hazard

$$\ln \lambda(t | \mathbf{Z}(t)) = \ln \lambda_0(t) + \sum_{k=1}^p \beta_k \mathbf{Z}_k(t) \quad (4.2)$$

linear in den Kovariablen ist.

Im folgenden wollen wir zunächst nur zeitunabhängige Kovariablen $\mathbf{Z} = (\mathbf{Z}_1, \dots, \mathbf{Z}_p)$ be-

trachten. Im Proportional Hazard Modell wird die Annahme getroffen, dass die Hazardfunktionen für verschiedene Werte von Z proportional zueinander sind. Daher kommt auch der Name „proportional hazard“. Seien \mathbf{Z} bzw. \mathbf{Z}^* die Kovariablenvektoren zweier Individuen. Bildet man ihre Hazard Ratio, so erhält man

$$\frac{\lambda(t|\mathbf{Z})}{\lambda(t|\mathbf{Z}^*)} = \frac{\lambda_0(t) \exp[\boldsymbol{\beta}'\mathbf{Z}]}{\lambda_0(t) \exp[\boldsymbol{\beta}'\mathbf{Z}^*]} = \exp\left[\sum_{k=1}^p \beta_k (\mathbf{Z}_k - \mathbf{Z}_k^*)\right] \quad (4.3)$$

Da sich der einzige zeitabhängige Faktor $\lambda_0(t)$ herauskürzt, erhält man eine von der Zeit unabhängige Konstante. Sei beispielsweise \mathbf{Z}_1 die Kovariable, die das Geschlecht des Individuums angibt. Sie nehme den Wert 1 für Frauen und 0 für Männer an. Betrachtet man nun das Verhältnis der Hazardfunktionen von Frauen und Männern mit ansonsten übereinstimmenden Variablen, gibt $\exp[\beta_1]$ das relative Risiko an, welches über die Zeit konstant bleibt.

4.1 Partieller Log-Likelihood

Cox (1972) schlug eine Methode vor, die Regressionskoeffizienten ohne Kenntnis des Baselinehazard zu schätzen, die Methode des partiellen Log-Likelihood.

Es wird angenommen, dass zwischen den Todeszeiten $t_i, i = 1, \dots, n$ keine ties vorliegen, also keine zwei Todesfälle gleichzeitig auftreten. Hier bezeichnen $t_1 < t_2 < \dots < t_D$ die geordneten Todeszeitpunkte und $Z_{(i)k}$ sei die k -te Kovariable des Individuums mit Todeszeit t_i . Durch $R(t_i) = \{j \in \{1, \dots, n\} : T_j \geq t_i\}$ sei die Risikomenge zum Zeitpunkt t_i definiert, d.h. die Menge aller Individuen, die bis zum Zeitpunkt t_i noch unter Beobachtung sind. Man geht davon aus, dass jedes Todesereignis einen Beitrag zur Likelihoodfunktion liefert. Zum Zeitpunkt t_i ist die Wahrscheinlichkeit, dass das Individuum mit Kovariablenvektor $\mathbf{Z}_{(i)}$ stirbt, bedingt darauf, dass eines der noch unter Beobachtung stehenden Individuen stirbt, gegeben durch

$$\begin{aligned} & P(\text{Individuum stirbt zum Zeitpunkt } t_i | \text{ein Tod zum Zeitpunkt } t_i) \\ &= \frac{P(\text{Individuum stirbt zum Zeitpunkt } t_i | \text{Überleben bis } t_i)}{P(\text{ein Tod zum Zeitpunkt } t_i | \text{Überleben bis } t_i)} \\ &= \frac{\lambda(t_i|\mathbf{Z}_{(i)})}{\sum_{j \in R(t_i)} \lambda(t_i|\mathbf{Z}_j)} = \frac{\lambda_0(t_i) \exp[\boldsymbol{\beta}'\mathbf{Z}_{(i)}]}{\sum_{j \in R(t_i)} \lambda_0(t_i) \exp[\boldsymbol{\beta}'\mathbf{Z}_j]} \\ &= \frac{\exp[\boldsymbol{\beta}'\mathbf{Z}_{(i)}]}{\sum_{j \in R(t_i)} \exp[\boldsymbol{\beta}'\mathbf{Z}_j]} \end{aligned} \quad (4.4)$$

Man beachte, dass sich der Term $\lambda_0(t_i)$ herauskürzt und nur noch die Regressionskoeffizienten $\beta_i, i = 1, \dots, p$ einen Beitrag zu (4.4) leisten.

Der Partial Likelihood ist dann definiert als das Produkt von (4.4) über alle Todeszeitpunkte $t_i, i = 1, \dots, D$, d.h.

$$L(\boldsymbol{\beta}) = \prod_{i=1}^D \frac{\exp[\sum_{k=1}^p \beta_k Z_{(i)k}]}{\sum_{j \in R(t_i)} \exp[\sum_{k=1}^p \beta_k Z_{jk}]} \quad (4.5)$$

Durch Logarithmieren erhält man den Partial Log-Likelihood

$$\begin{aligned} LL(\boldsymbol{\beta}) &= \ln L(\boldsymbol{\beta}) \\ &= \sum_{i=1}^D \sum_{k=1}^p \beta_k Z_{(i)k} - \sum_{i=1}^D \ln \left[\sum_{j \in R(t_i)} \exp \left[\sum_{k=1}^p \beta_k Z_{jk} \right] \right]. \end{aligned} \quad (4.6)$$

Die Schätzungen für die Regressionskoeffizienten erhält man, wie beim gewöhnlichen Log-Likelihood, durch Maximieren von (4.6) nach $\boldsymbol{\beta}$. Dazu werden die sogenannten Score-Gleichungen

$$\begin{aligned} U_h(\boldsymbol{\beta}) &= \frac{\partial LL(\boldsymbol{\beta})}{\partial \beta_h} \\ &= \sum_{i=1}^D Z_{(i)h} - \sum_{i=1}^D \frac{\sum_{j \in R(t_i)} Z_{jh} \exp \left[\sum_{k=1}^p \beta_k Z_{jk} \right]}{\sum_{j \in R(t_i)} \exp \left[\sum_{k=1}^p \beta_k Z_{jk} \right]}, \quad h = 1, \dots, p \end{aligned} \quad (4.7)$$

berechnet.

Durch Nullsetzen von

$$U_i(\boldsymbol{\beta}) = 0, \quad i = 1, \dots, p$$

erhält man schließlich die ML-Schätzer. Diese nicht linearen Gleichungen können numerisch z.B. mittels des Newton-Raphson-Algorithmus, gelöst werden. Die negative Hessematrix des Partial Loglikelihood ist die sogenannte Informationsmatrix

$$I(\boldsymbol{\beta}) = [I_{kl}(\boldsymbol{\beta})]_{p \times p} = -\frac{\partial^2 LL(\boldsymbol{\beta})}{\partial \beta_k \partial \beta_l}. \quad (4.8)$$

4.2 Motivation des Partial-Likelihood

In diesem Kapitel soll gezeigt werden, dass der Partial Likelihood (Cox [6]) eine gute Methode zur Schätzung der Regressionskoeffizienten im Proportional Hazard Modell ist. Miller [18] und Fleming und Harrington [8] zeigen den Zusammenhang zwischen dem vollen Likelihood und dem Partial Likelihood für zensierte Daten ohne Ties auf. Hier seien $t_1 \leq t_2 \leq \dots \leq t_D$ wie üblich die geordneten Ereigniszeiten, weiter setzt man $t_0 = 0$ und $t_{D+1} = \infty$. Die Daten seien durch die Folge $((A_1, B_1), (A_2, B_2), \dots, (A_D, B_D))$ von Paaren von Zufallsgrößen gegeben. Dabei enthalte B_k die Informationen über alle Zensierungen innerhalb des Intervalls $[t_{k-1}, t_k)$ und die Tatsache, dass zur Zeit t_k ein Ereignis stattfindet. A_k enthalte die Information, dass genau das Individuum mit Kovariablenvektor $\mathbf{Z}_{(k)}$ zur Zeit t_k das Ereignis erfährt.

Der volle Likelihood dieser 2D Zufallsgrößen lautet dann:

$$\begin{aligned} P &(A_D, B_D, A_{D-1}, B_{D-1}, \dots, A_1, B_1) \\ &= P(A_D, B_D | A_{D-1}, B_{D-1}, \dots, A_1, B_1) P(A_{D-1}, B_{D-1}, \dots, A_1, B_1) \\ &= \left[\prod_{k=2}^D P(A_k, B_k | A_{k-1}, B_{k-1}, \dots, A_1, B_1) \right] P(A_1, B_1) \end{aligned}$$

$$\begin{aligned}
&= \left[\prod_{k=2}^D \frac{P(A_k, B_k, A_{k-1}, B_{k-1}, \dots, A_1, B_1)}{P(A_{k-1}, B_{k-1}, \dots, A_1, B_1)} \right] P(A_1, B_1) \\
&= \left[\prod_{k=2}^D \frac{P(A_k | B_k, A_{k-1}, B_{k-1}, \dots, A_1, B_1) P(B_k, A_{k-1}, B_{k-1}, \dots, A_1, B_1)}{P(A_{k-1}, B_{k-1}, \dots, A_1, B_1)} \right] P(A_1, B_1) \\
&= \left[\prod_{k=2}^D P(A_k | B_k, A_{k-1}, B_{k-1}, \dots, A_1, B_1) P(B_k | A_{k-1}, B_{k-1}, \dots, A_1, B_1) \right] P(A_1 | B_1) P(B_1) \\
&= \left[\prod_{k=2}^D P(A_k | B_k, A_{k-1}, B_{k-1}, \dots, A_1, B_1) \right] P(A_1 | B_1) \\
&\times \left[\prod_{k=2}^D P(B_k | A_{k-1}, B_{k-1}, \dots, A_1, B_1) \right] P(B_1) \tag{4.9}
\end{aligned}$$

Wegen der Annahme, dass die Überlebenszeiten unabhängig von den Zensierungszeiten sind, d.h. dass die Zensierungszeiten keine Informationen über die Verteilung der Überlebenszeiten enthalten, erscheint es plausibel, dass die B_k nur wenig Informationen über die Regressionskoeffizienten β liefern. In diesem Fall können die letzten beiden Terme in (4.9) vernachlässigt werden. Die ersten beiden Terme

$$\left[\prod_{k=2}^D P(A_k | B_k, A_{k-1}, B_{k-1}, \dots, A_1, B_1) \right] P(A_1 | B_1)$$

bilden dann den Partial Likelihood. Es gilt (siehe [8])

$$P(A_k | B_k, A_{k-1}, B_{k-1}, \dots, A_1, B_1) = \frac{\lambda(t_k, \mathbf{Z}_{(k)})}{\sum_{j \in R(t_k)} \lambda(t_k, \mathbf{Z}_j)} = \frac{\exp[\beta' \mathbf{Z}_{(k)}]}{\sum_{j \in R(t_k)} \exp[\beta' \mathbf{Z}_j]}$$

Durch Multiplizieren über alle Ereigniszeiten erhält man daraus den in (4.5) definierten Partial Likelihood.

Fleming und Harrington [8] haben den durch das Weglassen der letzten beiden Terme in (4.9) entstehenden Informationsverlust über die Regressionskoeffizienten β genauer untersucht. Sie vergleichen die Schätzung der Regressionskoeffizienten mittels des Partial Likelihoods mit der Schätzung durch den vollen Likelihood mit parametrischer Annahme für den Baselinehazard ($\lambda_0(t) = \lambda_0(t, \alpha)$) und betrachten die asymptotische relative Effizienz von β . Dazu sei $LL_v(\phi)$ der Logarithmus des Likelihoods $L_v(\phi)$ im vollen Modell (siehe (3.9)) mit $\lambda(t, \phi) = \lambda_0(t, \alpha) \exp[\beta' \mathbf{Z}]$ und Parametervektor $\phi = (\alpha, \beta)$.

$$I_v(\phi) = -\frac{\partial^2}{\partial \phi^2} LL_v(\phi) = \begin{pmatrix} I_{\alpha\alpha}(\phi) & I_{\alpha\beta}(\phi) \\ I_{\alpha\beta}(\phi)' & I_{\beta\beta}(\phi) \end{pmatrix}$$

sei die partitionierte Informationsmatrix des vollen Likelihoods mit $I_{\alpha\alpha}(\phi) = -\frac{\partial^2}{\partial \alpha^2} LL_v(\phi)$, $I_{\alpha\beta}(\phi) = -\frac{\partial^2}{\partial \alpha \partial \beta} LL_v(\phi)$ und $I_{\beta\beta}(\phi) = -\frac{\partial^2}{\partial \beta^2} LL_v(\phi)$. Die Inverse der Informations-

matrix sei durch

$$I_v^{-1}(\phi) = \begin{pmatrix} I\alpha\alpha(\phi) & I\alpha\beta(\phi) \\ I\alpha\beta(\phi)' & I\beta\beta(\phi) \end{pmatrix}$$

definiert. Wird der volle Likelihood nach α maximiert, dann gilt

$$\frac{1}{n} \left(I\beta\beta(\phi) \right)^{-1} \rightarrow \Sigma_v(\beta) \quad \text{f.s. für } n \rightarrow \infty.$$

$\Sigma_v(\beta)$ ist hier die asymptotische standardisierte Informationsmatrix für β im vollen Likelihood, wenn α durch den ML-Schätzer $\hat{\alpha}$ ersetzt ist (siehe [8], S. 313-314). $\Sigma(\beta)$ sei die asymptotische standardisierte Informationsmatrix des Partial Likelihoods, d.h. für $n \rightarrow \infty$ gilt

$$\frac{1}{n} I(\beta) \rightarrow \Sigma(\beta) \quad \text{f. s. für } n \rightarrow \infty,$$

wobei $I(\beta)$ in (4.8) definiert ist.

Die asymptotische relative Effizienz des Maximum Partial Likelihood (MPL)-Schätzers von β bzgl. des Schätzers von β mit dem vollen Likelihood, ist durch das Verhältnis der asymptotischen standardisierten Informationsmatrizen $\frac{\Sigma(\beta)}{\Sigma_v(\beta)}$ gegeben. Fleming und Harrington [8] (S. 311-315) zeigen, dass gilt:

$$\frac{\Sigma(\beta)}{\Sigma_v(\beta)} = \frac{E[\text{var}(\mathbf{Z}(T)|T, \delta = 1)|\delta = 1]}{E[\text{var}(\mathbf{Z}(T)|T, \delta = 1)|\delta = 1] + \text{var}[E(\mathbf{Z}(T)|T, \delta = 1)|\delta = 1]},$$

wobei T hier die Überlebenszeit angibt. Die asymptotische relative Effizienz ist nach oben durch 1 beschränkt und nimmt genau den Wert 1 an, wenn gilt $\text{var}[E(\mathbf{Z}(T)|T, \delta = 1)|\delta = 1] = 0$. Dieser Ausdruck gibt die Varibilität über die Ereigniszeitpunkte des über die Risikomenge zu jedem Ereigniszeitpunkt gewichteten Mittelwertes von \mathbf{Z} an. Nach Fleming und Harrington [8] ist die Varibilität klein, falls $|\beta|$ klein, \mathbf{Z} unabhängig von Zensierungen und zeitunabhängig ist. In der Regel sind diese Bedingungen erfüllt, so dass der Partial Likelihood- Ansatz effiziente Schätzer liefert.

Asymptotische Normalität

Der MPL-Schätzer $\hat{\beta}$ für β ist nach Cox [6] asymptotisch normalverteilt. Ein Beweis dazu findet sich bei Fleming und Harrington [8] (S. 287-299).

Einschub: Eine Zufallsvariable \mathbf{Z} heißt p-variater normalverteilt mit Erwartungswert $\boldsymbol{\mu}$ und Kovarianzmatrix Σ , d.h. $\mathbf{Z} \sim N_p(\boldsymbol{\mu}, \Sigma)$, falls $\mathbf{Z} \in \mathbf{R}^p$, Σ positiv definit und \mathbf{Z} die Dichte

$$f(\mathbf{Z}) = f(Z_1, \dots, Z_p) = \frac{1}{(2\pi)^{\frac{p}{2}} (\det \Sigma)^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(\mathbf{Z} - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{Z} - \boldsymbol{\mu})\right]$$

besitzt.

Fleming und Harrington zeigen zuerst folgenden Satz :

Satz 4.1 *Unter bestimmten Regularitätsbedingungen (siehe [8], S. 289-290) gilt:*

1. Der normalisierte Scorevektor $\frac{1}{\sqrt{n}}\mathbf{U}(\beta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^n U_i(\beta_0)$ ist asymptotisch p -variabel normalverteilt mit Erwartungswert 0 und Kovarianzmatrix $\Sigma(\beta_0)$, d.h.

$$\frac{1}{\sqrt{n}}\mathbf{U}(\beta_0) \stackrel{a}{\approx} N_p(0, \Sigma(\beta_0)),$$

d.h. $\frac{1}{\sqrt{n}}\mathbf{U}(\beta_0)$ konvergiert in Verteilung gegen einen $N_p(0, \Sigma(\beta_0))$ -verteilten Zufallsvektor, also

$$\frac{1}{\sqrt{n}}\mathbf{U}(\beta_0) \xrightarrow{\mathcal{L}} \mathbf{Y} \quad \text{mit} \quad \mathbf{Y} \sim N_p(0, \Sigma(\beta_0)) \quad .$$

2. Ist $\hat{\beta}$ ein konsistenter Schätzer für β_0 , dann konvergiert $\frac{1}{n}I(\hat{\beta})$ für $n \rightarrow \infty$ in Wahrscheinlichkeit gegen $\Sigma(\beta_0)$, d.h. es gilt

$$\frac{1}{n}I(\hat{\beta}) \xrightarrow{P} \Sigma(\beta_0)$$

$U_i(\beta_0)$ und $I(\hat{\beta})$ sind dabei wie in (4.7) bzw. (4.8) definiert.

Der Beweis findet sich in [8], S. 292-296. Weiter beweisen sie die Konsistenz des MPL-Schätzers $\hat{\beta}$ ([8] S. 297-298):

Satz 4.2 *Sei $\hat{\beta}$ der MPL-Schätzer von β und β_0 der wahre Wert von β im Proportional Hazard Modell. Dann gilt*

$$\lim_{n \rightarrow \infty} \hat{\beta} = \beta_0 \quad ,$$

d.h. $\hat{\beta}$ ist konsistent.

Mit diesen Ergebnissen lässt sich schließlich die asymptotische Normalität des MPL-Schätzers zeigen:

Satz 4.3 $\Sigma(\beta_0)$ sei positiv definit. Dann gilt

$$\sqrt{n}(\hat{\beta} - \beta_0) \stackrel{a}{\approx} N_p(0, \Sigma(\beta_0)^{-1})$$

Beweisskizze: Durch Taylorentwicklung von $\mathbf{U}(\hat{\beta})$ um den wahren Wert β_0 erhält man

$$\mathbf{U}(\hat{\beta}) = \mathbf{U}(\beta_0) - I(\beta^*)(\hat{\beta} - \beta_0)$$

wobei β^* auf dem Geradenstück zwischen $\hat{\beta}$ und β_0 liegt. Wegen $\mathbf{U}(\hat{\beta}) = 0$ folgt

$$\frac{1}{n}I(\beta^*)\sqrt{n}(\hat{\beta} - \beta_0) = \frac{1}{\sqrt{n}}\mathbf{U}(\beta_0)$$

Nach Satz 4.1 (1) gilt: $\frac{1}{\sqrt{n}}\mathbf{U}(\beta_0) \stackrel{a}{\approx} N_p(0, \Sigma(\beta_0))$. Da $\hat{\beta}$ konsistent ist, gilt $\beta^* \rightarrow \beta$ in Wahrscheinlichkeit für $n \rightarrow \infty$. Nach Satz 4.1 (2) konvergiert daher $\frac{1}{n}I(\beta^*)$ in Wahrscheinlichkeit gegen $\Sigma(\beta_0)$ und die Behauptung folgt.

4.3 Statistische Hypothesentests

Um die Signifikanz der Regressionskoeffizienten β zu bestimmen, stehen verschiedene statistische Tests zur Verfügung. Wir wollen im folgenden drei Tests näher betrachten. Sei $\hat{\beta} = (\hat{\beta}_1, \dots, \hat{\beta}_p)'$ der Vektor der MPL-Schätzer von β und $I(\hat{\beta})$ die Informationsmatrix, ausgewertet an $\hat{\beta}$. Getestet werden soll jeweils die Nullhypothese

$$H_0 : \beta = \beta_0$$

gegen die Alternative

$$H_1 : \beta \neq \beta_0.$$

4.3.1 Der partielle Likelihood-Ratio Test

Die Statistik des partiellen Likelihood-Ratio Tests berechnet sich als die doppelte Differenz des partiellen Log-Likelihood des geschätzten Modells, und des partiellen Log-Likelihood des Modells unter der Nullhypothese, d.h.

$$\chi_{LR}^2 = 2 \cdot [LL(\hat{\beta}) - LL(\beta_0)] \quad (4.10)$$

Rao [20] (S. 415-418) zeigt, dass diese Statistik für große Datensätze unter der Nullhypothese χ^2 -verteilt mit p Freiheitsgraden ist.

4.3.2 Der Wald-Test

Eine weitere Testmöglichkeit bietet die Wald-Statistik. Sie beruht auf der Tatsache, dass $\hat{\beta}$ asymptotisch p -variat normalverteilt ist (siehe Satz 4.3):

$$\hat{\beta} \stackrel{a}{\approx} N_p(\beta_0, I^{-1}(\hat{\beta})) \quad (4.11)$$

Aus (4.11) folgt durch Standardisieren

$$I(\hat{\beta})(\hat{\beta} - \beta_0) \stackrel{a}{\approx} N_p(0, I(\hat{\beta}))$$

Die Wald-Statistik lautet dann

$$\chi_W^2 = (\hat{\beta} - \beta_0)' I(\hat{\beta})(\hat{\beta} - \beta_0) \quad (4.12)$$

Unter der Nullhypothese ist auch sie nach Rao [20] (S. 417-418) asymptotisch χ^2 -verteilt mit p Freiheitsgraden.

4.3.3 Der Score-Test

Der Score-Test basiert auf dem Score-Vektor $U(\beta) = (U_1(\beta), \dots, U_p(\beta))'$, wobei $U_i(\beta)$, $i = 1, \dots, p$ nach (4.7) definiert sind. Für große n ist der Score-Vektor unter der Nullhypothese asymptotisch p -variat normalverteilt mit Erwartungswert 0 und Kovarianzmatrix $I(\beta)$ (siehe Satz 4.1)

$$U(\beta) \stackrel{a}{\approx} N_p(0, I(\beta)).$$

Es kann gezeigt werden (siehe Rao [20], S. 417-418), dass die Statistik

$$\chi_{SC}^2 = U(\boldsymbol{\beta}_0)^t I^{-1}(\boldsymbol{\beta}_0) U(\boldsymbol{\beta}_0) \quad (4.13)$$

unter H_0 asymptotisch χ^2 -verteilt mit p -Freiheitsgraden ist. Man beachte, dass diese Statistik ohne Kenntnis der ML-Schätzer $\hat{\boldsymbol{\beta}}$ berechnet werden kann.

Im Gegensatz zum Wald- und Score-Test, sind beim Likelihood-Ratio Test keine Matrixmultiplikationen erforderlich. Die LR- Teststatistik kann allein mittels des schon vorhandenen Outputs des Modellfits erzeugt werden und wird daher oft bevorzugt angewendet.

4.4 Lokale Tests

Oft ist man nicht nur an der gesamten Signifikanz der Regressionskoeffizienten $\boldsymbol{\beta}$ interessiert, sondern möchte Hypothesen für Untermengen der Koeffizienten untersuchen. Sei $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$, wobei $\boldsymbol{\beta}_1$ der Vektor der interessierenden q Koeffizienten und $\boldsymbol{\beta}_2$ der Vektor der verbleibenden $p-q$ Koeffizienten ist. $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2)$ sei der Vektor der zugehörigen ML-Schätzer. Die Nullhypothese lautet dann

$$H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_{10}.$$

Die Informationsmatrix I sei wie folgt partitioniert:

$$I = \begin{pmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{pmatrix}$$

Dabei enthalten I_{11} bzw. I_{22} die negativen zweiten Ableitungen des Log-Likelihood nach den $\boldsymbol{\beta}_1$ bzw. $\boldsymbol{\beta}_2$, I_{12} und I_{21} die gemischten Ableitungen.

Die Statistik des Likelihood-Ratio- Tests lautet nun

$$\chi_{LR}^2 = 2 \cdot [LL(\hat{\boldsymbol{\beta}}) - LL(\boldsymbol{\beta}_{10}, \hat{\boldsymbol{\beta}}(\boldsymbol{\beta}_{10}))], \quad (4.14)$$

wobei $\hat{\boldsymbol{\beta}}(\boldsymbol{\beta}_{10})$ den MPL-Schätzer von $\boldsymbol{\beta}_2$ bei festgehaltenem Wert $\boldsymbol{\beta}_{10}$ der ersten q Koeffizienten bezeichnet. Für große n ist diese Statistik unter der Nullhypothese χ^2 -verteilt mit q Freiheitsgraden (Rao [20], S. 418-420). Häufig will man die Signifikanz von neu ins Modell hinzugenommenen Kovariablen $\boldsymbol{\beta}_1$ testen. In diesem Fall lautet die Nullhypothese

$$H_0 : \boldsymbol{\beta}_1 = 0.$$

Für die LR-Teststatistik bildet man die doppelte Differenz des Log-Likelihood des Modells, welches die q neuen Kovariablen enthält und des Log-Likelihood des Modells ohne die betreffenden Kovariablen.

Der Waldtest besitzt im Falle eines lokalen Tests der Nullhypothese $H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_{10}$ die Statistik

$$\chi_W^2 = (\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10})' [I^{11}(\hat{\boldsymbol{\beta}})]^{-1} (\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10}) \quad (4.15)$$

Hier bezeichnet I^{11} die obere $q \times q$ -Matrix der inversen Informationsmatrix. Für den Spezialfall $q = 1$ spricht man bei der Statistik des Wald-Tests auch von der z-Statistik:

$$z = \frac{\hat{\beta}_1}{\hat{SE}(\hat{\beta}_1)} \quad (4.16)$$

Die Teststatistik des Score-Tests wird folgendermaßen formuliert:

$$\chi_{SC}^2 = U_1[\beta_{10}, \hat{\beta}_2(\beta_{10})]' [I^{11}(\beta_{10}, \hat{\beta}_2(\beta_{10}))] U_1[\beta_{10}, \hat{\beta}_2(\beta_{10})], \quad (4.17)$$

wobei $U_1[\beta_{10}, \hat{\beta}_2(\beta_{10})]$ der $q \times 1$ - Vektor der Score-Gleichungen zu β_1 ist, ausgewertet für $\beta_1 = \beta_{10}$ und die MPL-Schätzungen von β_2 für festes $\beta_1 = \beta_{10}$. Sowohl χ_W^2 als auch χ_{SC}^2 sind nach Rao [20] (S. 418-420) unter H_0 für große n χ^2 -verteilt mit q Freiheitsgraden.

4.5 Das AIC

Neben lokalen Tests gibt das Informationskriterion von Akaike (AIC) (Akaike [1]), einen Anhaltspunkt darüber, ob bestimmte Kovariablen zusätzlich mit in das Modell aufgenommen werden sollten. Bei dieser Methode wird neben dem Wert des Log-Likelihood noch die Anzahl der im Modell enthaltenen Parameter berücksichtigt. Der Wert des AIC wird durch

$$AIC = -2 \cdot LL(\beta) + kp \quad (4.18)$$

berechnet. Dabei ist p die Anzahl der Regressionsparameter des Modells und k eine feste Konstante, die im allgemeinen mit $k = 2$ angenommen wird. Diese Wahl von k ergibt sich aus der Ableitung des AIC von der Kullback-Leibler Distanz (siehe Burnham und Anderson [5] S. 43-46). $LL(\beta)$ bezeichnet wie üblich den Wert des Log-Likelihood des Modells.

Mit zunehmender Anzahl der Parameter steigt der Wert des Log-Likelihood, da das Modell immer genauer wird. Der Wert des AIC wird sich also zunächst verringern. Die Zunahme von $LL(\beta)$ wird jedoch mit größerer Parameteranzahl immer geringer, so dass das AIC ab einem gewissen Punkt wieder steigen wird. Dies ist ein Anzeichen dafür, dass zuviele unnötige Kovariablen im Modell enthalten sind. Mittels des AIC wird also versucht, ein gutes Modell zu finden, aber gleichzeitig die Anzahl der Parameter möglichst gering zu halten. Den zweiten Term kann man daher auch als eine Art Bestrafung für jede zusätzliche Kovariable interpretieren.

4.6 Der Breslow-Schätzer für den Baselinehazard

Im Abschnitt (3.5) wurden nur nichtparametrische Schätzer für die kumulierte Hazardfunktion und die Überlebensfunktion betrachtet. Nun soll ein Schätzer für den Baselinehazard $\lambda_0(t)$ hergeleitet werden, für den Fall, dass Kovariablen vorliegen.

Die Daten seien rechtszensiert, d.h. von der Form $(T_j, \delta_j, \mathbf{Z}_j)$, $j = 1, \dots, n$, wobei

$$\delta_i = \begin{cases} 1 & \text{nicht zensiert} \\ 0 & \text{zensiert} \end{cases},$$

und die Beobachtungszeit $T_j = \min(X_j, C_r)$ (siehe Abschnitt 3.3.1) sowohl eine Überlebens- als auch eine Zensierungszeit darstellen kann. Wie üblich sind die t_i , $i = 1, \dots, D$ die geordneten Todeszeitpunkte. Der Unterschied zwischen den Beobachtungszeiten T_j , $j = 1, \dots, n$ und den geordneten Todeszeitpunkten t_i , $i = 1, \dots, D$ soll in Abb. 4.1 anhand eines Beispiels mit 4 Beobachtungen ($n=4$) noch einmal veranschaulicht werden.

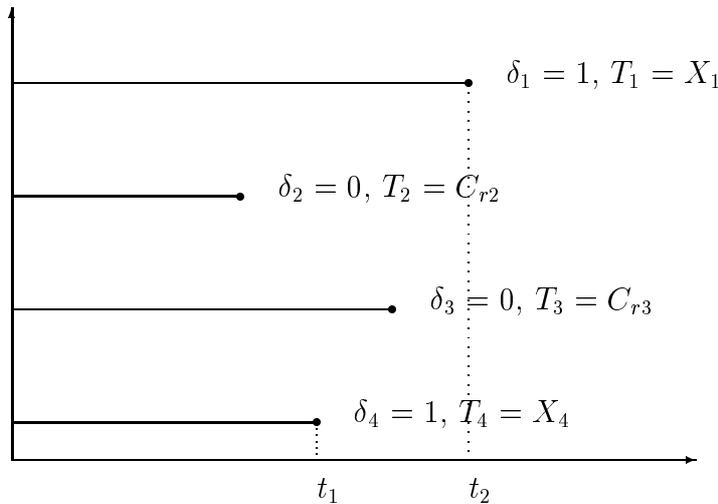


Abbildung 4.1: Unterschied zwischen Beobachtungszeiten T_j , $j = 1, \dots, n$ und geordneten Todeszeitpunkten t_i , $i = 1, \dots, D$

Bei den Beobachtungen 1 und 4 wird das Ereignis Tod zu den Zeiten $T_1 = X_1$ und $T_4 = X_4$ beobachtet, die Beobachtungen 2 und 3 werden zu den Zeitpunkten $T_2 = C_{r2}$ und $T_3 = C_{r3}$ zensiert. Es liegen also die $D=2$ geordneten Todeszeitpunkte $t_1 = T_4$ und $t_2 = T_1$ vor. Nach (3.9) kann die Likelihoodfunktion dargestellt werden als

$$L(\lambda_0, \boldsymbol{\beta}) = \prod_{j=1}^n \lambda(T_j, \mathbf{Z}_j)^{\delta_j} S(T_j, \mathbf{Z}_j) = \prod_{j=1}^n \lambda(T_j, \mathbf{Z}_j)^{\delta_j} \exp[-\Lambda(T_j, \mathbf{Z}_j)] \quad (4.19)$$

Mittels Diskretisierung der kumulierten Hazardfunktion

$$\Lambda(T_j, \mathbf{Z}_j) = \sum_{t_i \leq T_j} \lambda(t_i, \mathbf{Z}_j)$$

erhält man für das Proportional Hazard Modell (4.1)

$$L(\lambda_0, \boldsymbol{\beta}) = \prod_{j=1}^n \lambda(T_j, \mathbf{Z}_j)^{\delta_j} \exp\left[-\sum_{t_i \leq T_j} \lambda(t_i, \mathbf{Z}_j)\right] \quad (4.20)$$

$$\begin{aligned}
&= \prod_{j=1}^n \lambda_0(T_j)^{\delta_j} \exp\left[\delta_j \boldsymbol{\beta}' \mathbf{Z}_j\right] \exp\left[-\sum_{t_i \leq T_j} \lambda_0(t_i) \exp(\boldsymbol{\beta}' \mathbf{Z}_j)\right] \\
&= \left[\prod_{i=1}^D \lambda_0(t_i) \exp[\boldsymbol{\beta}' \mathbf{Z}_{(i)}] \right] \exp\left[-\sum_{j=1}^n \left\{ \sum_{t_i \leq T_j} \lambda_0(t_i) \exp(\boldsymbol{\beta}' \mathbf{Z}_j) \right\}\right] \\
&= \left[\prod_{i=1}^D \lambda_0(t_i) \exp[\boldsymbol{\beta}' \mathbf{Z}_{(i)}] \right] \exp\left[-\sum_{i=1}^D \left\{ \lambda_0(t_i) \sum_{j \in R(t_i)} \exp(\boldsymbol{\beta}' \mathbf{Z}_j) \right\}\right] \\
&= \prod_{i=1}^D \lambda_0(t_i) \exp[\boldsymbol{\beta}' \mathbf{Z}_{(i)}] \exp\left[-\lambda_0(t_i) \sum_{j \in R(t_i)} \exp(\boldsymbol{\beta}' \mathbf{Z}_j)\right],
\end{aligned}$$

wobei $R(t_i) = \{j \in \{1, \dots, n\} : T_j \geq t_i\}$. Die Gleichheit in der vorletzten Zeile gilt, da die Hazardrate bei diskreter Zeit nur für Zeitpunkte, an denen ein Ereignis stattfindet, d.h. für t_1, \dots, t_D , einen von Null verschiedenen Wert annimmt. Der Logarithmus von (4.20) lautet

$$\log L(\lambda_0, \boldsymbol{\beta}) = \sum_{i=1}^D (\log \lambda_0(t_i) + \boldsymbol{\beta}' \mathbf{Z}_{(i)} - \lambda_0(t_i) \sum_{j \in R(t_i)} \exp[\boldsymbol{\beta}' \mathbf{Z}_j]) \quad (4.21)$$

Den Breslow-Schätzer für den Baselinehazard $\hat{\lambda}_0(t_i)$ erhält man nun durch Maximieren von (4.21) nach $\lambda_0(t_i)$, d.h. $\hat{\lambda}_0(t_i)$ erfüllt

$$\frac{\partial \log L(\lambda_0, \boldsymbol{\beta})}{\partial \lambda_0(t_i)} = \frac{1}{\lambda_0(t_i)} - \sum_{j \in R(t_i)} \exp[\boldsymbol{\beta}' \mathbf{Z}_j] = 0 \quad (4.22)$$

Damit gilt

$$\hat{\lambda}_0(t_i) = \frac{1}{\sum_{j \in R(t_i)} \exp[\boldsymbol{\beta}' \mathbf{Z}_j]}$$

Daraus ergibt sich als Schätzer für den kumulierten Baselinehazard

$$\hat{\Lambda}_0(t) = \sum_{t_i \leq t} \frac{1}{\sum_{j \in R(t_i)} \exp[\boldsymbol{\beta}' \mathbf{Z}_j]}$$

oder in der Notation der Zählprozesse

$$\hat{\Lambda}_0(t) = \int_0^t \frac{\sum_{i=1}^n dN_i(s)}{\sum_{i=1}^n Y_i(s) \exp[\hat{\boldsymbol{\beta}}' \mathbf{Z}_i(s)]}. \quad (4.23)$$

Sind keine Kovariablen vorhanden, d.h. $\mathbf{Z}_i = \mathbf{0}$, $i = 1, \dots, n$, reduziert sich (4.23) zum Nelson-Aalen-Schätzer (3.10).

4.7 Berechnung des Likelihood mit Ties

Bei der Entwicklung des partiellen Likelihood von Cox wurde davon ausgegangen, dass zwischen den einzelnen Ereignissen keine ties vorliegen. In der Realität ist das jedoch

meist nicht der Fall. Oft ist dem Modell eine diskrete Zeitskala zugrundegelegt, so dass mehrere kurz aufeinanderfolgende Ereignisse dem gleichen Zeitpunkt zugeordnet werden. Es gibt verschiedene Ansätze, den Partial Likelihood in einem solchen Fall zu modifizieren: der exakten Partial Likelihood von Cox oder die Approximationen von Breslow und Efron. Eine gute Darstellung dieser drei Ansätze findet sich bei Therneau und Grambsch [26] (S. 48-53).

Seien $t_1 < t_2 < \dots < t_D$ die getrennten Ereigniszeitpunkte, d_i sei die Anzahl der Todesfälle zum Zeitpunkt t_i und D_i die Menge aller Individuen, die zum Zeitpunkt t_i sterben. $\mathbf{s}_i = \sum_{j \in D_i} \mathbf{Z}_j$ sei die Summe der Kovariablenvektoren dieser Individuen.

Die beim Auftreten von Ties entstehende Schwierigkeit soll an einem einfachen Beispiel erklärt werden. Man nehme an, dass sich zum Zeitpunkt t_i fünf nach der Ereigniszeit geordnete Individuen in der Risikomenge $R(t_i)$ befinden, von denen die ersten beiden das Ereignis erfahren. Hätte man genauere Daten, so dass bekannt wäre, welches der beiden Ereignisse zuerst stattfand, dann würde der Beitrag zur Likelihoodfunktion entweder

$$\frac{r_1}{r_1 + r_2 + r_3 + r_4 + r_5} \cdot \frac{r_2}{r_2 + r_3 + r_4 + r_5}$$

oder

$$\frac{r_2}{r_1 + r_2 + r_3 + r_4 + r_5} \cdot \frac{r_1}{r_1 + r_3 + r_4 + r_5}$$

lauten, wobei hier der Übersichtlichkeit halber die Bezeichnung $r_i = \exp[\boldsymbol{\beta}'\mathbf{Z}_i]$ eingeführt wurde. Da man diese Information bei Ties jedoch nicht besitzt, kann man nur Approximationen verwenden.

Die einfachste Approximation stammt von Breslow. Jedes der d_i Ereignisse zum Zeitpunkt t_i wird als getrennt betrachtet. Für das obige Beispiel würde der Beitrag zur Likelihoodfunktion daher

$$\frac{r_1}{r_1 + r_2 + r_3 + r_4 + r_5} \cdot \frac{r_2}{r_1 + r_2 + r_3 + r_4 + r_5}$$

sein. Allgemein bedeutet das, dass jedes Ereignis zum Zeitpunkt t_i den Beitrag

$$\frac{\exp(\boldsymbol{\beta}'\mathbf{Z}_k)}{\sum_{j \in R(t_i)} \exp(\boldsymbol{\beta}'\mathbf{Z}_j)}, \quad k \in D_i,$$

liefert. Insgesamt trägt also das Produkt

$$\prod_{k \in D_i} \frac{\exp(\boldsymbol{\beta}'\mathbf{Z}_k)}{\sum_{j \in R(t_i)} \exp(\boldsymbol{\beta}'\mathbf{Z}_j)} = \frac{\exp(\boldsymbol{\beta}'\mathbf{s}_i)}{[\sum_{j \in R(t_i)} \exp[\boldsymbol{\beta}'\mathbf{Z}_j]]^{d_i}}$$

zum Zeitpunkt t_i zum Likelihood bei. Den Partial Likelihood von Breslow erhält man nun als Produkt über alle Ereigniszeiten:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^D \frac{\exp[\boldsymbol{\beta}'\mathbf{s}_i]}{[\sum_{j \in R(t_i)} \exp[\boldsymbol{\beta}'\mathbf{Z}_j]]^{d_i}} \quad (4.24)$$

Bei dieser Approximation werden die Individuen, die das Ereignis erfahren, mehrfach im Nenner gezählt, was zur Folge hat, dass der Absolutbetrag von $\boldsymbol{\beta}$ im Vergleich zur exakten

Lösung zu klein ist.

Eine bessere Approximation, die diesen Sachverhalt berücksichtigt, aber nur wenig aufwendiger ist, liefert der Partial Likelihood von Efron [7]. Efron verwendet einen Durchschnittswert im Nenner. Für das oben genannte Beispiel tragen die Terme

$$\frac{r_1}{r_1 + r_2 + r_3 + r_4 + r_5} \cdot \frac{r_2}{0.5r_1 + 0.5r_2 + r_3 + r_4 + r_5}$$

zum Likelihood bei. Sie lassen sich wie folgt interpretieren: Beim Eintritt des ersten der beiden Ereignisse, befinden sich die Individuen, die das Ereignis erfahren beide noch in der Risikomenge, beim zweiten Ereignis, ist jedes der beiden Individuen nur noch mit einer Wahrscheinlichkeit von 0.5 in der Risikomenge enthalten. Für den Fall, dass drei Individuen zur gleichen Zeit das Ereignis erfahren, werden die Gewichte $1, \frac{2}{3}$ und $\frac{1}{3}$ gewählt. Denn im ersten Term ist jedes der drei Individuen noch mit Wahrscheinlichkeit 1 in der Risikomenge, nach dem ersten Ereignis beträgt diese Wahrscheinlichkeit nur noch $\frac{2}{3}$ und nach dem zweiten Ereignis nur noch $\frac{1}{3}$. Allgemein ergibt sich daraus der Beitrag

$$\frac{\exp[\beta' \mathbf{s}_i]}{\prod_{j=1}^{d_i} [\sum_{k \in R(t_i)} \exp[\beta' \mathbf{Z}_k] - \frac{j-1}{d_i} \sum_{k \in D_i} \exp[\beta' \mathbf{Z}_k]]}$$

und folglich der Partial Likelihood

$$L(\beta) = \prod_{i=1}^D \frac{\exp[\beta' \mathbf{s}_i]}{\prod_{j=1}^{d_i} [\sum_{k \in R(t_i)} \exp[\beta' \mathbf{Z}_k] - \frac{j-1}{d_i} \sum_{k \in D_i} \exp[\beta' \mathbf{Z}_k]]} \quad (4.25)$$

Beim exakten Partial Likelihood geht man davon aus, dass die Ties aufgrund von ungenauen Messungen der Überlebenszeiten zustande kommen und daher die genaue Reihenfolge der Ereignisse nicht bekannt ist. Es gibt daher zum Zeitpunkt t_i $d_i!$ verschiedene mögliche Reihenfolgen, in denen die Ereignisse beobachtet werden können. Sei Q_i die Menge aller möglichen d_i -Tupel von Individuen, die zum Zeitpunkt t_i aus der Risikomenge $R(t_i)$ ausgewählt werden könnten. $q = (q_1, \dots, q_{d_i})$ sei ein Element aus Q_i und $\mathbf{s}_q^* = \sum_{j=1}^{d_i} \mathbf{Z}_{q_j}$. Der exakte Partial Likelihood betrachtet das Risiko, dass aus allen möglichen d_i -Tupeln genau das d_i -Tupel der Individuen ausgewählt wird, die das Ereignis erfahren. Die Likelihoodfunktion lautet dann

$$L(\beta) = \prod_{i=1}^D \frac{\exp[\beta' \mathbf{s}_i]}{\sum_{q \in Q_i} \exp[\beta' \mathbf{s}_q^*]}. \quad (4.26)$$

Der Exakte Partial Likelihood erfordert aufgrund der Vielzahl von Auswahlmöglichkeiten einen hohen Rechenaufwand. Liegen keine ties vor, d.h. $d_i = 1, i = 1, \dots, D$, reduzieren sich sowohl die Approximationen von Breslow und Efron als auch der Exakte Partial Likelihood auf den Partial Likelihood (4.5).

4.8 Bestimmung der funktionalen Form stetiger Kovariablen

Hat man getestet, welche Kovariablen in einem Modell berücksichtigt werden müssen, ist der nächste Schritt, die funktionale Form aller stetigen Kovariablen des Modells zu überprüfen. Nach dem Proportional Hazard Modell sollten diese linear im Log-Hazard sein. So wird zum Beispiel das Risiko zwischen einem 60- und einem 65-jährigen Individuum als das gleiche wie zwischen einem 80- und einem 85-Jährigen angenommen. Diese Bedingung ist in der Realität oft verletzt, so dass geeignete Transformationen für die Kovariablen gefunden werden müssen. Martingalresiduen können dafür als Hilfsmittel verwendet werden. Sie werden bei Fleming und Harrington [8] (S. 163-165), Grambsch und Therneau [26] (S. 79-83), Klein und Moeschberger [14] (S. 333-334) als auch bei Therneau et al. [27] behandelt.

4.8.1 Martingalresiduen

Sei $\mathbf{Z}_j(t)$ der möglicherweise auch zeitabhängige Kovariablenvektor des j -ten Individuums. $N_j(t)$ nehme den Wert 1 an, wenn das Individuum j zum Zeitpunkt t das Ereignis bereits erfahren hat und 0 sonst. Weiter sei $Y_j(t)$ eine Indikatorvariable, die den Wert 1 annimmt, falls sich das j -te Individuum zur Zeit t noch unter Beobachtung befindet, anderenfalls 0, d.h. $Y_i(t) = I\{T_i \geq t\}$.

Dann sind die Martingalresiduen durch

$$\hat{M}_j(t) = N_j(t) - \int_0^t Y_j(s) \exp[\hat{\boldsymbol{\beta}}' \mathbf{Z}_j(s)] d\hat{\Lambda}_0(s), \quad j = 1, \dots, n \quad (4.27)$$

definiert, wobei $\boldsymbol{\beta}$ der Vektor der Regressionskoeffizienten und

$$\hat{\Lambda}_0(t) = \int_0^t \frac{\sum_{i=1}^n dN_i(s)}{\sum_{i=1}^n Y_i(s) \exp[\hat{\boldsymbol{\beta}}' \mathbf{Z}_i(s)]}$$

der Breslow-Schätzer der kumulierten Baseline Hazardrate ist (siehe (4.23)). $\hat{M}_j(\infty)$ sei mit \hat{M}_j bezeichnet, d.h.

$$\hat{M}_j = \hat{M}_j(\infty) = N_j(\infty) - \int_0^\infty Y_j(t) \exp[\hat{\boldsymbol{\beta}}' \mathbf{Z}_j(t)] d\hat{\Lambda}_0(t), \quad j = 1, \dots, n$$

Falls die Daten rechtszensiert und alle Kovariablen zeitunabhängig sind, gibt $N_j(\infty)$ den Endstatus des j -ten Individuums an, d.h. $N_j(\infty) = \delta_j$. $Y_j(t)$ nimmt nur den Wert 1 an solange sich das j -te Individuum unter Beobachtung befindet, also bis zur Beobachtungszeit T_j . Damit reduziert sich (4.27) in diesem Fall mit $\hat{\Lambda}_0(0) = 0$ zu

$$\begin{aligned} \hat{M}_j &= \delta_j - \int_0^{T_j} \exp\left[\sum_{k=1}^p Z_{jk} \hat{\boldsymbol{\beta}}_k\right] d\hat{\Lambda}_0(t) \\ &= \delta_j - \hat{\Lambda}_0(T_j) \exp\left[\sum_{k=1}^p Z_{jk} \hat{\boldsymbol{\beta}}_k\right], \quad j = 1, \dots, n. \end{aligned}$$

Die Martingalresiduen können daher als die Differenz zwischen der beobachteten Anzahl und der unter dem Cox Modell erwarteten Anzahl der Ereignisse über die Zeit hinweg interpretiert werden.

Satz 4.4 Die Martingalresiduen besitzen folgende Eigenschaften:

- (1) $\sum_{j=1}^n \hat{M}_j(t) = 0 \quad \forall t \geq 0$
- (2) Ausgewertet an den wahren Werten β und Λ_0 ist M_j ein Martingal, daher auch der Name Martingalresiduen.
- (3) $E(\hat{M}_i) = \text{cov}(\hat{M}_i, \hat{M}_j) = 0$ für große n , d.h. die \hat{M}_i s sind asymptotisch unkorreliert mit Erwartungswert 0.

Beweis zu (1): Ersetze $d\hat{\Lambda}_0(s)$ in (4.27) durch $d\hat{\Lambda}_0(s) = \frac{\sum_{i=1}^n dN_i(s)}{\sum_{i=1}^n Y_i(s) \exp[\hat{\beta}'\mathbf{Z}_i(s)]}$:

$$\begin{aligned} \sum_{j=1}^n \hat{M}_j(t) &= \sum_{j=1}^n N_j(t) - \sum_{j=1}^n \int_0^t Y_j(s) \exp[\hat{\beta}'\mathbf{Z}_j(s)] \frac{\sum_{i=1}^n dN_i(s)}{\sum_{i=1}^n Y_i(s) \exp[\hat{\beta}'\mathbf{Z}_i(s)]} \\ &= \sum_{j=1}^n N_j(t) - \int_0^t \sum_{j=1}^n dN_j(s) = \sum_{j=1}^n N_j(t) - \sum_{j=1}^n \int_0^t dN_j(s) = 0 \end{aligned}$$

Beweis zu (2): Nach Kapitel 2, (2.2), ist zu zeigen: $E(dM_i(t)|\mathcal{F}_s) = 0$ mit

$\mathcal{F}(t) = \sigma\{Z_i(s), N_i(s), Y_i(s+), i = 1, \dots, n, 0 \leq s \leq t\}$

Da $N_i(t)$ eine Indikatorvariable ist, gilt

$$E(dN_i(t)|\mathcal{F}_{t-}) = P(dN_i(t) = 1|\mathcal{F}_{t-}) = P(dN_i(t) = 1|Y_i(t))$$

Die letzte Gleichheit gilt, da aufgrund der Unabhängigkeit der einzelnen Beobachtungen für $N_i(t)$ nur die Geschichte der i -ten Beobachtung von Interesse ist, und zwar nur der Zustand zum Zeitpunkt $t-$. Das ist genau $Y_i(t)$. Wegen

$$P(dN_i(t) = 1|Y_i(t) = 0) = P(t \leq T_i < t + dt | t > T_i) = 0$$

und

$$P(dN_i(t) = 1|Y_i(t) = 1) = P(t \leq T_i < t + dt | t \leq T_i) = \lambda(t)dt$$

folgt

$$P(dN_i(t) = 1|Y_i(t)) = Y_i(t) \cdot \lambda(t)dt$$

Daher gilt

$$\begin{aligned} E(dM_i(t)|\mathcal{F}_{t-}) &= E(dN_i(t)|\mathcal{F}_{t-}) - E(Y_i(t) \exp[\beta'\mathbf{Z}_i(t)] \lambda_0(t) dt | \mathcal{F}_{t-}) \\ &= Y_i(t) \lambda(t) dt - Y_i(t) \lambda(t) dt = 0 \end{aligned}$$

Die Gleichheit in der letzten Zeile folgt aus der Vorhersagbarkeit und daher auch \mathcal{F}_{t-} -Meßbarkeit von $Y_i(t)$ und $\lambda(t)$.

Es gilt also $E(dM_i(t)|\mathcal{F}_s) = 0$, d.h. die Zuwächse von Martingalen haben Erwartungswert 0. Die Martingalresiduen haben daher einen konstanten Erwartungswert . Es gilt

$$E(M_i(t)) = E(M_i(0)) = N_i(0) - \int_0^t Y_i(s) \exp[\boldsymbol{\beta}'\mathbf{Z}_i(s)] d\hat{\Lambda}_0(s) = 0.$$

Beweis zu (3): Wegen $E(M_i(t)) = 0 \forall t \geq 0$ gilt $E(N_i(t)) = E\left(\int_0^t Y_i(s) \exp[\boldsymbol{\beta}'\mathbf{Z}_i(s)] d\Lambda_0(s)\right) \forall t \geq 0$. Nach Satz 4.3 gilt für $n \rightarrow \infty$: $E(\hat{\boldsymbol{\beta}}) \rightarrow \boldsymbol{\beta}$. Darüberhinaus zeigen Fleming und Harrington [8] (S.299-300), dass gilt $E(\hat{\Lambda}_0) \rightarrow \Lambda_0$ für $n \rightarrow \infty$. Mit diesen Ergebnissen folgt für $n \rightarrow \infty$

$$\begin{aligned} E(\hat{M}_i) &= E(N_i(\infty)) - E\left(\int_0^\infty Y_i(t) \exp[\hat{\boldsymbol{\beta}}'\mathbf{Z}_i] d\hat{\Lambda}_0(t)\right) \\ &\rightarrow E\left(\int_0^\infty Y_i(t) \exp[\boldsymbol{\beta}'\mathbf{Z}_i] d\Lambda_0(t)\right) - E\left(\int_0^\infty Y_i(t) \exp[\boldsymbol{\beta}'\mathbf{Z}_i] d\Lambda_0(t)\right) = 0 \end{aligned}$$

und, ebenfalls für $n \rightarrow \infty$,

$$\begin{aligned} \text{cov}(\hat{M}_i, \hat{M}_j) &= E(\hat{M}_i \hat{M}_j) \\ &= E(N_i(\infty)N_j(\infty)) - E\left(N_i(\infty) \int_0^\infty Y_j(t) \exp[\hat{\boldsymbol{\beta}}'\mathbf{Z}_j(t)] d\hat{\Lambda}_0(t)\right) \\ &\quad - E\left(N_j(\infty) \int_0^\infty Y_i(t) \exp[\hat{\boldsymbol{\beta}}'\mathbf{Z}_i(t)] d\hat{\Lambda}_0(t)\right) \\ &\quad + E\left(\int_0^\infty Y_i(t) \exp[\hat{\boldsymbol{\beta}}'\mathbf{Z}_i(t)] d\hat{\Lambda}_0(t) \int_0^\infty Y_j(t) \exp[\hat{\boldsymbol{\beta}}'\mathbf{Z}_j(t)] d\hat{\Lambda}_0(t)\right) \\ &\rightarrow E(N_i(\infty)N_j(\infty)) - E(N_i(\infty)N_j(\infty)) - E(N_j(\infty)N_i(\infty)) \\ &\quad + E(N_i(\infty)N_j(\infty)) = 0 \quad \square \end{aligned}$$

4.8.2 Martingalresiduen zur Bestimmung der funktionalen Form

Martingalresiduen bieten eine einfache graphische Methode, die funktionale Form stetiger Kovariablen zu untersuchen. Der Kovariablenvektor \mathbf{Z} sei partitioniert in den Vektor \mathbf{Z}^* , welcher die Kovariablen enthält, deren funktionale Form bekannt ist, und die Kovariable \mathbf{Z}_1 , deren funktionale Form gesucht wird. Es wird angenommen, dass \mathbf{Z}^* und \mathbf{Z}_1 unabhängig sind. Sei $f(\mathbf{Z}_1)$ die beste funktionale Form von \mathbf{Z}_1 , um den Effekt auf das Überleben zu erklären. Dann lässt sich die Hazard-Rate darstellen, als

$$\lambda(t|\mathbf{Z}^*, \mathbf{Z}_1) = \lambda_0(t) \exp[\boldsymbol{\beta}'\mathbf{Z}^*] \exp[f(\mathbf{Z}_1)] = \lambda_0(t) \exp[\boldsymbol{\beta}'\mathbf{Z}^*] h(\mathbf{Z}_1) \quad (4.28)$$

mit $h(\mathbf{Z}_1) = \exp(f(\mathbf{Z}_1))$. Therneau et al. [27] zeigen den folgenden Satz. $\hat{M}(t)$ bezeichnet dabei das Martingalresiduum des ohne die Kovariable \mathbf{Z}_1 gefitteten Modells (4.28). Der Index i von \hat{M}_i wird weggelassen, um die Notation übersichtlicher zu halten.

Satz 4.5 *Unter der Voraussetzung, dass das Modell (4.28) korrekt ist, gilt*

$$E(\hat{M}(t)|\mathbf{Z}_1) \approx \left(1 - \frac{\bar{h}(s, \mathbf{Z}^*)}{h(\mathbf{Z}_1)}\right) E(N(t)|\mathbf{Z}_1) \quad (4.29)$$

Dabei bezeichnet $\bar{h}(s, \mathbf{Z}^*) = \frac{E(\exp[f(\mathbf{Z}_1)]Y(s)|\mathbf{Z}^*)}{E(Y(s)|\mathbf{Z}^*)}$ eine Mittelung von h sowohl über die Zeit als auch über die erwartete Zusammensetzung der Risikomenge.

Beweis: Unter der Annahme, dass

$$\lambda(t|\mathbf{Z}^*, \mathbf{Z}_1) = \lambda_0(t) \exp[\boldsymbol{\beta}'\mathbf{Z}^*]h(\mathbf{Z}_1) = \lambda_0(t) \exp[\boldsymbol{\beta}'\mathbf{Z}^*] \exp f(\mathbf{Z}_1)$$

sei

$$\hat{M}(t) = N(t) - \int_0^t Y(s) \exp[\hat{\boldsymbol{\beta}}'\mathbf{Z}^*(s)]d\hat{\Lambda}_0(s)$$

das Martingalresiduum(siehe (4.27)) des ohne \mathbf{Z}_1 gefitteten Modells.

Mit (4.28) gilt

$$d\Lambda(s, \mathbf{Z}^*) = \exp[\boldsymbol{\beta}'\mathbf{Z}^*]h(\mathbf{Z}_1)d\Lambda_0(s)$$

Da ein ohne \mathbf{Z}_1 gefittetes Modell betrachtet wird, muss $h(\mathbf{Z}_1)$ durch die Mittelung $\bar{h}(s, \mathbf{Z}^*)$ ersetzt werden:

$$d\Lambda(s, \mathbf{Z}^*) = \exp[\boldsymbol{\beta}'\mathbf{Z}^*] \frac{E(\exp[f(\mathbf{Z}_1)]Y(s)|\mathbf{Z}^*)}{E(Y(s)|\mathbf{Z}^*)} d\Lambda_0(s) = \exp[\boldsymbol{\beta}'\mathbf{Z}^*]\bar{h}(s, \mathbf{Z}^*)d\Lambda_0(s) \quad (4.30)$$

Damit kann $E(\hat{M}(t)|\mathbf{Z}_1)$ geschrieben werden als

$$\begin{aligned} E(\hat{M}(t)|\mathbf{Z}_1) &= E(N(t)|\mathbf{Z}_1) - E\left(\int_0^t Y(s) \exp[\hat{\boldsymbol{\beta}}'\mathbf{Z}^*]d\hat{\Lambda}_0(s)|\mathbf{Z}_1\right) \\ &= E(N(t)|\mathbf{Z}_1) - E\left(\int_0^t Y(s) \exp[\boldsymbol{\beta}'\mathbf{Z}^*]\bar{h}(s, \mathbf{Z}^*)d\Lambda_0(s)|\mathbf{Z}_1\right) \\ &+ E\left(\int_0^t Y(s)d\Lambda(s, \mathbf{Z}^*)|\mathbf{Z}_1\right) - E\left(\int_0^t Y(s) \exp[\hat{\boldsymbol{\beta}}'\mathbf{Z}^*]d\hat{\Lambda}_0(s)|\mathbf{Z}_1\right) \\ &= E(N(t)|\mathbf{Z}_1) - E\left(\int_0^t Y(s) \exp[\boldsymbol{\beta}'\mathbf{Z}^*]\bar{h}(s, \mathbf{Z}^*)d\Lambda_0(s)|\mathbf{Z}_1\right) \\ &+ E\left(\int_0^t Y(s)\{d\Lambda(s, \mathbf{Z}^*) - \exp[\hat{\boldsymbol{\beta}}'\mathbf{Z}^*]d\hat{\Lambda}_0(s)\}|\mathbf{Z}_1\right) \\ &= \text{Term 1} + \text{Term 2} + \text{Term 3} \end{aligned} \quad (4.31)$$

Man beachte, dass sich der zweite und dritte Summand in der zweiten Gleichungszeile gegenseitig aufheben. Mit (4.30) und

$$d\hat{\Lambda}_0(s) = \frac{\sum_{i=1}^n dN_i(s)}{\sum_{j=1}^n Y_j(s) \exp[\hat{\boldsymbol{\beta}}'\mathbf{Z}_j^*]}$$

läßt sich Term 3 in (4.31) schreiben als

$$\begin{aligned} &E\left(\int_0^t Y(s)\{d\Lambda(s, \mathbf{Z}^*) - \exp[\hat{\boldsymbol{\beta}}'\mathbf{Z}^*]d\hat{\Lambda}_0(s)\}|\mathbf{Z}_1\right) \\ &= E\left(\int_0^t Y(s)\left\{\exp[\boldsymbol{\beta}'\mathbf{Z}^*]\bar{h}(s, \mathbf{Z}^*)d\Lambda_0(s) - \exp[\hat{\boldsymbol{\beta}}'\mathbf{Z}^*] \frac{\sum_{i=1}^n dN_i(s)}{\sum_{j=1}^n Y_j(s) \exp[\hat{\boldsymbol{\beta}}'\mathbf{Z}_j^*]}\right\}|\mathbf{Z}_1\right) \\ &= E\left(\int_0^t Y(s)\left\{\exp[(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'\mathbf{Z}^*] \exp[\hat{\boldsymbol{\beta}}'\mathbf{Z}^*]\bar{h}(s, \mathbf{Z}^*)d\Lambda_0(s)\right\}\right) \end{aligned}$$

$$\begin{aligned}
& - \sum_{i=1}^n dN_i(s) \frac{\exp[\hat{\boldsymbol{\beta}}' \mathbf{Z}^*]}{\sum_{j=1}^n Y_j(s) \exp[\hat{\boldsymbol{\beta}}' \mathbf{Z}_j^*]} \Big| \mathbf{Z}_1 \Big) \\
& = E \left(\int_0^t \frac{Y(s) \exp[\hat{\boldsymbol{\beta}}' \mathbf{Z}^*]}{\sum_{j=1}^n Y_j(s) \exp[\hat{\boldsymbol{\beta}}' \mathbf{Z}_j^*]} \left\{ \exp[(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})' \mathbf{Z}^*] \left(\sum_{j=1}^n Y_j(s) \exp[\hat{\boldsymbol{\beta}}' \mathbf{Z}_j^*] \right) \bar{h}(s, \mathbf{Z}^*) d\Lambda_0(s) \right. \right. \\
& - \left. \left. \sum_{i=1}^n dN_i(s) \right\} \Big| \mathbf{Z}_1 \Big) \\
& = -E \left(\sum_{i=1}^n \int_0^t \frac{Y(s) \exp[\hat{\boldsymbol{\beta}}' \mathbf{Z}^*]}{\sum_{j=1}^n Y_j(s) \exp[\hat{\boldsymbol{\beta}}' \mathbf{Z}_j^*]} \left\{ dN_i(s) - Y_i(s) \exp[(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})' \mathbf{Z}^*] \right. \right. \\
& \cdot \left. \left. \exp[\hat{\boldsymbol{\beta}}' \mathbf{Z}_i^*] \bar{h}(s, \mathbf{Z}^*) d\Lambda_0(s) \right\} \Big| \mathbf{Z}_1 \right) \tag{4.32}
\end{aligned}$$

Bickel et al. [4] zeigen, dass, wenn $\boldsymbol{\beta}$ klein ist und \mathbf{Z}_1 ignoriert wird, gilt

$$\exp[(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})' \mathbf{Z}^*] \exp[\hat{\boldsymbol{\beta}}' \mathbf{Z}_i^*] \bar{h}(s, \mathbf{Z}^*) \approx \exp[\boldsymbol{\beta}' \mathbf{Z}_i^*] \bar{h}(s, \mathbf{Z}_i^*) \quad . \tag{4.33}$$

Eingesetzt in (4.32) erhält man damit

$$\begin{aligned}
& -E \left(\sum_{i=1}^n \int_0^t \frac{Y(s) \exp[\hat{\boldsymbol{\beta}}' \mathbf{Z}^*]}{\sum_{j=1}^n Y_j(s) \exp[\hat{\boldsymbol{\beta}}' \mathbf{Z}_j^*]} \left\{ dN_i(s) - Y_i(s) \exp[(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})' \mathbf{Z}^*] \right. \right. \\
& \cdot \left. \left. \exp[\hat{\boldsymbol{\beta}}' \mathbf{Z}_i^*] \bar{h}(s, \mathbf{Z}^*) d\Lambda_0(s) \right\} \Big| \mathbf{Z}_1 \right) \\
& \approx -E \left(\sum_{i=1}^n \int_0^t \frac{Y(s) \exp[\hat{\boldsymbol{\beta}}' \mathbf{Z}^*]}{\sum_{j=1}^n Y_j(s) \exp[\hat{\boldsymbol{\beta}}' \mathbf{Z}_j^*]} \left\{ dN_i(s) - Y_i(s) \exp[\boldsymbol{\beta}' \mathbf{Z}_i^*] \right. \right. \\
& \cdot \left. \left. \bar{h}(s, \mathbf{Z}_i^*) d\Lambda_0(s) \right\} \Big| \mathbf{Z}_1 \right) \\
& \stackrel{(4.30)}{=} -E \left(\sum_{i=1}^n \int_0^t \frac{Y(s) \exp[\hat{\boldsymbol{\beta}}' \mathbf{Z}^*]}{\sum_{j=1}^n Y_j(s) \exp[\hat{\boldsymbol{\beta}}' \mathbf{Z}_j^*]} \left\{ dN_i(s) - Y_i(s) d\Lambda(s, \mathbf{Z}_i^*) \right\} \right) \tag{4.34}
\end{aligned}$$

$N_i(t) - \int_0^t Y_i(s) d\Lambda(s, \mathbf{Z}^*)$ ist für gegebenes \mathbf{Z}^* und ignoriertes \mathbf{Z}_1 ein Martingal (siehe (4.27)). Nach dem Beweis von Satz 4.4 (2) gilt daher $E \left(N_i(t) - \int_0^t Y_i(s) d\Lambda(s, \mathbf{Z}^*) \right) = 0$ und folglich auch $E \left(dN_i(t) - Y_i(t) d\Lambda(t, \mathbf{Z}^*) \right) = 0$. Term 3 konvergiert mit der Darstellung in (4.34) daher gegen 0.

Setzt man $\bar{h}(s) = E[\bar{h}(s, \mathbf{Z}^*)]$ und wählt eine feste Zeit t_0 , kann Term 2 in (4.31) geschrieben werden als

$$\begin{aligned}
& - E \left(\int_0^t Y(s) \exp[\boldsymbol{\beta}' \mathbf{Z}^*] \bar{h}(s, \mathbf{Z}^*) d\Lambda_0(s) \Big| \mathbf{Z}_1 \right) \\
& = -E \left(\int_0^t Y(s) \left\{ \bar{h}(s, \mathbf{Z}^*) + \bar{h}(t_0) \frac{h(\mathbf{Z}_1)}{h(\mathbf{Z}_1)} - \bar{h}(t_0) \right\} \exp[\boldsymbol{\beta}' \mathbf{Z}^*] d\Lambda_0(s) \Big| \mathbf{Z}_1 \right) \\
& = -E \left(\int_0^t Y(s) \frac{\bar{h}(t_0)}{h(\mathbf{Z}_1)} h(\mathbf{Z}_1) \exp[\boldsymbol{\beta}' \mathbf{Z}^*] d\Lambda_0(s) \Big| \mathbf{Z}_1 \right)
\end{aligned}$$

$$\begin{aligned}
& + E\left(\int_0^t Y(s)\{\bar{h}(t_0) - \bar{h}(s, \mathbf{Z}^*)\} \exp[\boldsymbol{\beta}'\mathbf{Z}^*]d\Lambda_0(s)|\mathbf{Z}_1\right) \\
& = -\frac{\bar{h}(t_0)}{h(\mathbf{Z}_1)}E\left(\int_0^t Y(s)h(\mathbf{Z}_1) \exp[\boldsymbol{\beta}'\mathbf{Z}^*]d\Lambda_0(s)|\mathbf{Z}_1\right) + R(t, \mathbf{Z}_1) \\
& \stackrel{(4.28)}{=} -\frac{\bar{h}(t_0)}{h(\mathbf{Z}_1)}E\left(\int_0^t Y(s)d\Lambda(s, \mathbf{Z})|\mathbf{Z}_1\right) + R(t, \mathbf{Z}_1)
\end{aligned}$$

Mit $R(t, \mathbf{Z}_1)$ wird hier der Term $E\left(\int_0^t Y(s)\{\bar{h}(t_0) - \bar{h}(s, \mathbf{Z}^*)\} \exp[\boldsymbol{\beta}'\mathbf{Z}^*]d\Lambda_0(s)|\mathbf{Z}_1\right)$ bezeichnet. $N(t) - \int_0^t Y(s)d\Lambda(s, \mathbf{Z})$ ist ein Martingal für den Kovariablenvektor \mathbf{Z} , daher gilt wie oben

$$E\left(N(t) - \int_0^t Y(s)d\Lambda(s, \mathbf{Z})\right) = 0$$

und somit

$$\begin{aligned}
& -\frac{\bar{h}(t_0)}{h(\mathbf{Z}_1)}E\left(\int_0^t Y(s)d\Lambda(s, \mathbf{Z})|\mathbf{Z}_1\right) + R(t, \mathbf{Z}_1) \\
& = -\frac{\bar{h}(t_0)}{h(\mathbf{Z}_1)}E\left(N(t)|\mathbf{Z}_1\right) + R(t, \mathbf{Z}_1)
\end{aligned} \tag{4.35}$$

Setzt man (4.35) nun in die Gleichung (4.31) ein und berücksichtigt, dass gilt Term 3 ≈ 0 , erhält man

$$\begin{aligned}
E(\hat{M}(t)|\mathbf{Z}_1) & \approx E(N(t)|\mathbf{Z}_1) - \frac{\bar{h}(t_0)}{h(\mathbf{Z}_1)}E(N(t)|\mathbf{Z}_1) + R(t, \mathbf{Z}_1) \\
& = \left(1 - \frac{\bar{h}(t_0)}{h(\mathbf{Z}_1)}\right)E(N(t)|\mathbf{Z}_1) + R(t, \mathbf{Z}_1)
\end{aligned} \tag{4.36}$$

Nun muss noch gezeigt werden, dass $R(t, \mathbf{Z}_1)$ als Restterm behandelt werden kann.

$$\begin{aligned}
& E(\hat{M}(t)|\mathbf{Z}_1) \approx \left[1 - \frac{\bar{h}(t_0)}{h(\mathbf{Z}_1)}\right]E(N(t)|\mathbf{Z}_1) + R(t, \mathbf{Z}_1) \\
\Leftrightarrow & \frac{E(\hat{M}(t)|\mathbf{Z}_1)}{E(N(t)|\mathbf{Z}_1)} \approx 1 - \frac{\bar{h}(t_0)}{h(\mathbf{Z}_1)} + \frac{R(t, \mathbf{Z}_1)}{E(N(t)|\mathbf{Z}_1)} \\
\Leftrightarrow & 1 - \frac{E(\hat{M}(t)|\mathbf{Z}_1)}{E(N(t)|\mathbf{Z}_1)} \approx \frac{\bar{h}(t_0)}{h(\mathbf{Z}_1)} \left(1 - \frac{h(\mathbf{Z}_1)R(t, \mathbf{Z}_1)}{\bar{h}(t_0)E(N(t)|\mathbf{Z}_1)}\right) \\
\Leftrightarrow & \log\left[1 - \frac{E(\hat{M}(t)|\mathbf{Z}_1)}{E(N(t)|\mathbf{Z}_1)}\right] \approx \log \bar{h}(t_0) - \log h(\mathbf{Z}_1) + \log\left(1 - \frac{h(\mathbf{Z}_1)R(t, \mathbf{Z}_1)}{\bar{h}(t_0)E(N(t)|\mathbf{Z}_1)}\right) \\
\stackrel{h(\mathbf{Z}_1) = \exp[f(\mathbf{Z}_1)]}{=} & \log \bar{h}(t_0) - f(\mathbf{Z}_1) + \log\left(1 - \frac{h(\mathbf{Z}_1)R(t, \mathbf{Z}_1)}{\bar{h}(t_0)E(N(t)|\mathbf{Z}_1)}\right) \\
= & -f(\mathbf{Z}_1) + \log\left(\bar{h}(t_0) - \frac{h(\mathbf{Z}_1)R(t, \mathbf{Z}_1)}{E(N(t)|\mathbf{Z}_1)}\right)
\end{aligned} \tag{4.37}$$

Es gilt

$$\begin{aligned}
& \log\left(\bar{h}(t_0) - \frac{h(\mathbf{Z}_1)R(t, \mathbf{Z}_1)}{E(N(t)|\mathbf{Z}_1)}\right) \\
&= \log\left(\frac{\bar{h}(t_0)E(N(t)|\mathbf{Z}_1) - h(\mathbf{Z}_1)R(t, \mathbf{Z}_1)}{E(N(t)|\mathbf{Z}_1)}\right) \\
&= \log\left(\frac{\bar{h}(t_0)E(N(t)|\mathbf{Z}_1)}{h(\mathbf{Z}_1)} - R(t, \mathbf{Z}_1)\right) - \log\left(\frac{E(N(t)|\mathbf{Z}_1)}{h(\mathbf{Z}_1)}\right) \\
&= \log(\text{-Term 2, (4.35)}) - \log\left(\frac{E(N(t)|\mathbf{Z}_1)}{h(\mathbf{Z}_1)}\right) \\
&\stackrel{\text{Term 2, (4.31)}}{=} \log\left[E\left(\int_0^t Y(s) \exp[\boldsymbol{\beta}'\mathbf{Z}^*] \bar{h}(s, \mathbf{Z}^*) d\Lambda_0(s) | \mathbf{Z}_1\right)\right] - \log\left[\frac{E\left(\int_0^t Y(s) d\Lambda(s, \mathbf{Z}) | \mathbf{Z}_1\right)}{h(\mathbf{Z}_1)}\right] \\
&= \log\left[E\left(\int_0^t Y(s) \exp[\boldsymbol{\beta}'\mathbf{Z}^*] \bar{h}(s, \mathbf{Z}^*) d\Lambda_0(s) | \mathbf{Z}_1\right)\right] \\
&\quad - \log\left[\frac{E\left(\int_0^t Y(s) h(\mathbf{Z}_1) \exp[\boldsymbol{\beta}'\mathbf{Z}^*] d\Lambda_0(s) | \mathbf{Z}_1\right)}{h(\mathbf{Z}_1)}\right] \\
&= \log\left[E\left(\int_0^t Y(s) \exp[\boldsymbol{\beta}'\mathbf{Z}^*] \bar{h}(s, \mathbf{Z}^*) d\Lambda_0(s) | \mathbf{Z}_1\right)\right] \\
&\quad - \log\left[E\left(\int_0^t Y(s) \exp[\boldsymbol{\beta}'\mathbf{Z}^*] d\Lambda_0(s) | \mathbf{Z}_1\right)\right] \\
&= \log\frac{E\left(\int_0^t Y(s) \exp[\boldsymbol{\beta}'\mathbf{Z}^*] \bar{h}(s, \mathbf{Z}^*) d\Lambda_0(s) | \mathbf{Z}_1\right)}{E\left(\int_0^t Y(s) \exp[\boldsymbol{\beta}'\mathbf{Z}^*] d\Lambda_0(s) | \mathbf{Z}_1\right)} \\
&= \log\frac{\bar{h}(s, \mathbf{Z}^*) E\left(\int_0^t Y(s) \exp[\boldsymbol{\beta}'\mathbf{Z}^*] d\Lambda_0(s) | \mathbf{Z}_1\right)}{E\left(\int_0^t Y(s) \exp[\boldsymbol{\beta}'\mathbf{Z}^*] d\Lambda_0(s) | \mathbf{Z}_1\right)} \\
&= \log \bar{h}(s, \mathbf{Z}^*) \tag{4.38}
\end{aligned}$$

Es gilt also mit (4.37)

$$\begin{aligned}
\log\left[1 - \frac{E(\hat{M}(t)|\mathbf{Z}_1)}{E(N(t)|\mathbf{Z}_1)}\right] &\approx -f(\mathbf{Z}_1) + \log \bar{h}(s, \mathbf{Z}^*) = -\log h(\mathbf{Z}_1) + \log \bar{h}(s, \mathbf{Z}^*) \\
&= \log \frac{\bar{h}(s, \mathbf{Z}^*)}{h(\mathbf{Z}_1)}. \tag{4.39}
\end{aligned}$$

Durch Exponieren erhält man

$$\begin{aligned}
1 - \frac{E(\hat{M}(t)|\mathbf{Z}_1)}{E(N(t)|\mathbf{Z}_1)} &\approx \frac{\bar{h}(s, \mathbf{Z}_1)}{h(\mathbf{Z}_1)} \\
\Leftrightarrow E(\hat{M}(t)|\mathbf{Z}_1) &\approx \left(1 - \frac{\bar{h}(s, \mathbf{Z}_1)}{h(\mathbf{Z}_1)}\right) E(N(t)|\mathbf{Z}_1).
\end{aligned}$$

□

Invertieren von

$$E(\hat{M}(t)|\mathbf{Z}_1) \approx \left(1 - \frac{\bar{h}(s, \mathbf{Z}_1)}{h(\mathbf{Z}_1)}\right) E(N(t)|\mathbf{Z}_1)$$

ergibt (siehe (4.39))

$$f(\mathbf{Z}_1) - \log(\bar{h}(s, \mathbf{Z}^*)) \approx -\log\left[1 - \frac{E(\hat{M}(t)|\mathbf{Z}_1)}{E(N(t)|\mathbf{Z}_1)}\right]$$

Mit $\bar{f} = \log(\bar{h}(s, \mathbf{Z}^*))$ erhält man

$$f(\mathbf{Z}_1) - \bar{f} \approx -\log\left[1 - \frac{\text{smooth}(E(\hat{M}(t)|\mathbf{Z}_1))}{\text{smooth}(E(N(t)|\mathbf{Z}_1))}\right], \quad (4.40)$$

wobei $\text{smooth}(E(\hat{M}(t)|\mathbf{Z}_1))$ und $\text{smooth}(E(N(t)|\mathbf{Z}_1))$ geglättete Schätzer von $E(M(t)|\mathbf{Z}_1)$ bzw. $E(N(t)|\mathbf{Z}_1)$ sind.

Nach Therneau et al. [27] kann (4.40) für die meisten Datensätze für $t = \infty$ mittels Taylor-Approximation ersetzt werden durch

$$E(\hat{M}|\mathbf{Z}_1) \approx (f(\mathbf{Z}_1) - \bar{f})c \quad . \quad (4.41)$$

Dabei bezeichnet $\hat{M} = \hat{M}(\infty)$ und $c = \frac{\text{Gesamtanzahl der Ereignisse}}{\text{Gesamtanzahl der Individuen}} = \frac{\sum_{i=1}^n N_i(\infty)}{n}$. (4.41) kann folgendermaßen plausibel gemacht werden: mittels Taylor-Entwicklung erster Ordnung um $x=0$ erhält man

$$\ln(1+x) = x + o(x^2)$$

Nach (4.29) gilt $\frac{E(\hat{M}(t)|\mathbf{Z}_1)}{E(N(t)|\mathbf{Z}_1)} \approx 1 - \frac{\bar{h}(s, \mathbf{Z}^*)}{h(\mathbf{Z}_1)}$. Unter bestimmten Bedingungen die laut Therneau et al. [27] (S.150) für die meisten Datensätze erfüllt sind liegt dieser Term in der Nähe von 0. Mit $x = -\frac{E(\hat{M}(t)|\mathbf{Z}_1)}{E(N(t)|\mathbf{Z}_1)}$ folgt daher

$$f(\mathbf{Z}_1) - \bar{f} \approx \frac{E(M(t)|\mathbf{Z}_1)}{E(N(t)|\mathbf{Z}_1)} + o\left(\left[\frac{E(M(t)|\mathbf{Z}_1)}{E(N(t)|\mathbf{Z}_1)}\right]^2\right)$$

Ersetzt man nun $E(N(t)|\mathbf{Z}_1)$ durch die relative Häufigkeit der Ereignisse c , erhält man den Zusammenhang (4.41).

Therneau et al. schlagen daher vor, ein Modell zu fitten, in dem nur die Kovariablen \mathbf{Z}^* enthalten sind und die Martingalresiduen zu berechnen. Plottet man diese Residuen gegen die Kovariable \mathbf{Z}_1 so gibt ein geglätteter Scatterplot in etwa die funktionale Form von \mathbf{Z}_1 wieder. Falls die geglättete Kurve linear ist, zeigt das, dass keine Transformation von \mathbf{Z}_1 notwendig ist.

4.8.3 Fraktionale Polynome

Royston und Altman (1994) [22] entwickelten eine Methode, um mittels fraktionaler Polynome geeignete Transformationen für stetige Kovariablen zu bestimmen. Sowohl Hosmer und Lemeshow [12] als auch Berger et al. [3] beschreiben, wie fraktionale Polynome erklärt sind.

Ein fraktionales Polynom vom Grad m für eine stetige Kovariable $x > 0$ ist bestimmt durch

$$\Phi_m(x, p) = \beta_0 + \sum_{j=1}^m \beta_j x^{p_j},$$

wobei m eine natürliche Zahl ist, β_j Regressionskoeffizienten und $p_1 \leq \dots \leq p_m$ beliebige reellwertige Potenzen sind.

Für $p_j = 0$ ist dabei der Logarithmus der Kovariable x definiert, d.h. x^0 entspricht $\ln(x)$. Durch die Wiederholung von Potenzen $p_i = \dots = p_k, i < k \leq m$ können darüber hinaus Kombinationen mit $\ln x$ mit einbezogen werden.

Beispiel: Wählt man $m = 6$ und $p = (-1, 0, 0.5, 2, 2, 2)$ so ist das dazugehörige fraktionale Polynom folgendermaßen definiert:

$$\Phi_6(x, p) = \beta_0 + \beta_1 \frac{1}{x} + \beta_2 \ln x + \beta_3 \sqrt{x} + \beta_4 x^2 + \beta_5 x^2 \ln x + \beta_6 x^2 (\ln x)^2$$

Fraktionale Polynome erlauben also eine Vielzahl unterschiedlicher Funktionen, wählt man nur positive ganzzahlige Potenzen, erhält man gewöhnliche Polynome.

Hosmer und Lemeshow [12] beschreiben, wie fraktionale Polynome eingesetzt werden können, um die funktionale Form stetiger Kovariablen zu finden. Der Einfachheit halber werden nur univariate Modelle mit einer stetigen Variable x betrachtet. Die zugehörige Hazardfunktion ist von folgender Gestalt:

$$\lambda(t, x, \beta) = \lambda_0(t) \exp(x\beta)$$

Der Log-Hazard, der nach Modellannahme linear in der Kovariable ist lautet dementsprechend:

$$\ln[\lambda(t, x, \beta)] = \ln[\lambda_0(t)] + x\beta$$

Nun ersetzt man den linearen Prädiktor $x\beta$ durch ein fraktionales Polynom

$$x\beta = \sum_{j=1}^m \beta_j x^{p_j}$$

und versucht, die Potenzen p_i optimal für das Modell zu bestimmen. Der Koeffizient β_0 wird hier nicht berücksichtigt, da er als Intercept nur in $\ln[\lambda_0(t)]$ eingehen würde.

Royston und Altman [22] raten, sich bei der Wahl der Potenzen auf die Menge

$\wp = \{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$ zu beschränken. Würde man die Potenzen als zusätzliche Parameter ins Modell einbauen und schätzen lassen, entstünde dadurch ein erheblich größerer Schätzaufwand. Nach Hosmer und Lemeshow [12] reicht es in der Regel aus, nur fraktionale Polynome ersten oder zweiten Grades zu betrachten, um geeignete Transformationen zu finden. Für $m=1$ sind also 8 Modelle, für jede Potenz $p_i \in \wp$, zu fitten. Als bestes Modell wird das Modell mit dem höchsten partiellen Log-Likelihood gewählt, der im folgenden mit $L(\hat{p}_1)$ bezeichnet wird, wo \hat{p}_1 die beste gefundene Potenz ist.

Für $m=2$ erhöht sich die Anzahl der Modelle auf 64, da nun für alle möglichen Kombinationen $(p_i, p_j) \in \wp \times \wp$ Modelle gefittet werden müssen. Wieder wird das beste Modell nach dem größten partiellen Log-Likelihood bestimmt, für den die Notation $L(\hat{p}_1, \hat{p}_2)$ eingeführt wird. Bei den beiden Modellen für $m=1$ und $m=2$ ist die erste Potenz aus Notationsgründen jeweils mit \hat{p}_1 bezeichnet, in der Praxis werden sie jedoch nur selten übereinstimmen.

Es gilt nun, zu untersuchen, ob diese beiden besten gefundenen Modelle auch wirklich signifikant besser sind, als das Modell, in das die Kovariable linear eingeht. Dazu werden verschiedene Likelihood-Ratio-Tests durchgeführt. Bezeichne $L(1)$ den partiellen Log-Likelihood des linearen Modells (d.h. $m=1$ und $p_1 = 1$).

Royston und Altman [22] schlagen vor, für jeden Term des fraktionalen Polynoms zwei Freiheitsgrade anzusetzen, einen Freiheitsgrad für den Regressionskoeffizienten und einen für die Potenz. Will man das beste Modell für $m=1$ mit dem linearen Modell vergleichen, so ist die Likelihood-Ratio-Statistik

$$\chi_{LR}^2 = 2 \cdot [L(\hat{p}_1) - L(1)]$$

χ^2 -verteilt mit 1 Freiheitsgrad für die Potenz \hat{p}_1 . Die Likelihood-Ratio-Statistik zum Vergleich des besten Modells mit $m=2$ mit dem linearen Modell

$$\chi_{LR}^2 = 2 \cdot [L(\hat{p}_1, \hat{p}_2) - L(1)]$$

ist χ^2 -verteilt mit 3 Freiheitsgraden (= 2 Freiheitsgrade für die beiden Potenzen und 1 Freiheitsgrad für den Regressionskoeffizienten β_2). Um zu testen, ob das beste Modell für $m=2$ dem besten Modell für $m=1$ vorzuziehen ist, wird die Statistik

$$\chi_{LR}^2 = 2 \cdot [L(\hat{p}_1, \hat{p}_2) - L(\hat{p}_1)]$$

betrachtet. Sie ist χ^2 -verteilt mit 2 Freiheitsgraden, jeweils einem für den Regressionskoeffizienten und die Potenz des zweiten Terms des fraktionalen Polynoms.

4.9 Zeitabhängige Koeffizienten

Eine wichtige Annahme im Cox-Modell ist die Proportional Hazard Annahme. Sie unterstellt, dass sich die Effekte der Kovariablen auf das Überleben nicht über die Beobachtungszeit hinweg verändern. Es ist jedoch durchaus möglich, dass dies nicht der Fall ist. So könnte zum Beispiel eine bestimmte Behandlung mit zunehmender Behandlungsdauer

eine bessere Wirkung erzielen.

Eine Alternative zum Proportional Hazard Modell ist die Erweiterung des Modells um zeitabhängige Koeffizienten zu

$$\lambda(t|\mathbf{Z}) = \lambda_0(t) \exp[\boldsymbol{\beta}(t)' \mathbf{Z}] \quad (4.42)$$

Im Fall $\boldsymbol{\beta}(t) = \boldsymbol{\beta}$ reduziert sich (4.42) auf das gewöhnliche Cox-Modell (4.1).

Klein und Moeschberger [14] stellen eine Methode vor, um die Proportional Hazard Annahme für eine diskrete Kovariable \mathbf{X} mit Werten $1, 2, \dots, K$ graphisch zu überprüfen. Im Proportional Hazard Modell gilt für das Verhältnis der kumulierten Hazardfunktionen von Beobachtungen mit Ausprägung $X = i$ und $X = j$, $i, j = 1, \dots, K$

$$\frac{\Lambda(t, X = i)}{\Lambda(t, X = j)} = \frac{\Lambda_0(t) \exp[\boldsymbol{\beta} \cdot i]}{\Lambda_0(t) \exp[\boldsymbol{\beta} \cdot j]} = \exp[\boldsymbol{\beta}(i - j)].$$

Klein und Moeschberger ([14], S.337-338) schlagen vor, ein nach \mathbf{X} stratifiziertes Modell zu fitten, d.h. ein Modell, welches für jedes Stratum $X = k$, $k = 1, \dots, K$ einen eigenen Baselinehazard $\lambda_{0k}(t)$, $k = 1, \dots, K$ schätzt. Für die Ratio der kumulierten Hazardraten $\Lambda_{0k}(t) = \int_0^t \lambda_{0k}(s) ds$ zweier Ausprägungen $X = i$ bzw. $X = j$ erhält man nun im stratifizierten Modell

$$\frac{\Lambda_i(t, X = i)}{\Lambda_j(t, X = j)} = \frac{\Lambda_{0i}(t) \exp[\boldsymbol{\beta} \cdot i]}{\Lambda_{0j}(t) \exp[\boldsymbol{\beta} \cdot j]} = \frac{\Lambda_{0i}(t)}{\Lambda_{0j}(t)} \exp[\boldsymbol{\beta}(i - j)].$$

Ist die Proportional Hazard Annahme erfüllt, sollte daher das Verhältnis $\frac{\Lambda_{0i}(t)}{\Lambda_{0j}(t)}$ eine Konstante ergeben, d.h. $\Lambda_{0i}(t) = c \cdot \Lambda_{0j}(t)$. Die Plots von $\log \Lambda_{0i}(t)$ und $\log \Lambda_{0j}(t)$ gegen t sollten also parallel verlaufen mit vertikalem Abstand $\log c$. Für das Auge einfacher zu überprüfen ist ein Plot der Differenz $\log \Lambda_{0i}(t) - \log \Lambda_{0j}(t)$ gegen t , der eine Konstante ergeben sollte. Diese Methode ist nur auf diskrete Kovariablen anwendbar. Mittels der Schoenfeldresiduen dagegen kann das Vorhandensein von Zeitabhängigkeit sowohl bei diskreten als auch bei stetigen Kovariablen aufgedeckt werden.

4.9.1 Schoenfeld-Residuen

Ausgangspunkt zur Definition der Schoenfeld-Residuen (Schoenfeld [24]) sind die Score-Gleichungen (4.7), die hier nochmals in der Notation der Zählprozesse aufgeführt werden:

$$\frac{\partial LL(\boldsymbol{\beta})}{\partial \beta_k} = \sum_{i=1}^D \left(Z_{(i)k} - \frac{\sum_{j=1}^n Y_j(t_i) Z_{jk}(t_i) \exp[\boldsymbol{\beta}' \mathbf{Z}_j(t_i)]}{\sum_{j=1}^n Y_j(t_i) \exp[\boldsymbol{\beta}' \mathbf{Z}_j(t_i)]} \right)$$

Die Schoenfeld-Residuen sind zu jedem Ereignis für alle p Kovariablen definiert. Unter der Annahme, dass zwischen den Ereigniszeiten keine Ties vorliegen, ist das Schoenfeld-Residuum zum i -ten Ereignis durch

$$\begin{aligned}\hat{\mathbf{r}}_i(\boldsymbol{\beta}) &:= \mathbf{Z}_{(i)} - \frac{\sum_{j=1}^n Y_j(t_i) \mathbf{Z}_j(t_i) \exp[\hat{\boldsymbol{\beta}}' \mathbf{Z}_j(t_i)]}{\sum_{j=1}^n Y_j(t_i) \exp[\hat{\boldsymbol{\beta}}' \mathbf{Z}_j(t_i)]} \\ &= \mathbf{Z}_{(i)} - \mathbf{M}(\hat{\boldsymbol{\beta}}, t_i) \quad \in \mathbf{R}^p\end{aligned}\quad (4.43)$$

gegeben. Dabei ist

$$\mathbf{M}(\boldsymbol{\beta}, t) = \frac{\sum_{i=1}^n Y_i(t) \mathbf{Z}_i(t) \exp(\boldsymbol{\beta}' \mathbf{Z}_i(t))}{\sum_{i=1}^n Y_i(t) \exp(\boldsymbol{\beta}' \mathbf{Z}_i(t))} \quad (4.44)$$

ein gewichteter Mittelwert von \mathbf{Z} über die Individuen, die sich zur Zeit t noch in der Risikomenge befinden. Die k -te Komponente des Schoenfeld-Residuums für das i -te Ereignis kann vereinfacht auch als

$$\hat{r}_{ik} = Z_{(i)k} - \frac{\sum_{j \in R(t_i)} Z_{ij} \exp[\hat{\boldsymbol{\beta}}' \mathbf{Z}_j]}{\sum_{j \in R(t_i)} \exp[\hat{\boldsymbol{\beta}}' \mathbf{Z}_j]} \quad (4.45)$$

formuliert werden. Da der ML-Schätzer $\hat{\boldsymbol{\beta}}$ durch Nullsetzen der Scoregleichungen (4.7) erhalten wurde, gilt

$$\sum_{i=1}^D \hat{r}_{ik} = 0.$$

Die Schoenfeld-Residuen finden Anwendung bei der Überprüfung der Proportional Hazard Annahme. Nach Grambsch und Therneau sind die skalierten Schoenfeld-Residuen

$$\hat{\mathbf{r}}_k^* = V(\hat{\boldsymbol{\beta}}, t_k)^{-1} \cdot \hat{\mathbf{r}}_k \quad (4.46)$$

geeigneter zur Diagnose. Hier bezeichnet $\hat{\mathbf{r}}_k = (\hat{r}_{k1}, \dots, \hat{r}_{kp})'$ den p -Vektor der Schoenfeld-Residuen zum Ereigniszeitpunkt t_k und

$$\begin{aligned}V(\boldsymbol{\beta}, t) &= \frac{\partial \mathbf{M}(\boldsymbol{\beta}, t)}{\partial \boldsymbol{\beta}} \\ &= \frac{\sum_{i=1}^n Y_i(t) \exp(\boldsymbol{\beta}' \mathbf{Z}_i(t)) \mathbf{Z}_i(t) \mathbf{Z}_i(t)'}{\sum_{i=1}^n Y_i(t) \exp(\boldsymbol{\beta}' \mathbf{Z}_i(t))} \\ &\quad - \left(\frac{\sum_{i=1}^n Y_i(t) \exp(\boldsymbol{\beta}' \mathbf{Z}_i(t)) \mathbf{Z}_i(t)}{\sum_{i=1}^n Y_i(t) \exp(\boldsymbol{\beta}' \mathbf{Z}_i(t))} \right) \left(\frac{\sum_{i=1}^n Y_i(t) \exp(\boldsymbol{\beta}' \mathbf{Z}_i(t)) \mathbf{Z}_i(t)}{\sum_{i=1}^n Y_i(t) \exp(\boldsymbol{\beta}' \mathbf{Z}_i(t))} \right)'\end{aligned}\quad (4.47)$$

die Kovarianzmatrix des Kovariablenvektors zum Zeitpunkt t .

4.9.2 Überprüfung von Zeitabhängigkeit mittels Schoenfeldresiduen

Eine graphische Methode zur Überprüfung der Zeitabhängigkeit der Koeffizienten ermöglichen die skalierten Schoenfeldresiduen. Grambsch und Therneau [9] (S.516-517) zeigen, dass folgender Satz gilt .

Satz 4.6 Sei $\mathcal{F}_t, t \geq 0$ eine rechtsstetige Filtration

$$\mathcal{F}_t = \sigma\{\mathbf{Z}_i(u), N_i(u), Y_i(u+) : 0 \leq u \leq t, i = 1, \dots, n\},$$

d.h., die Geschichte des Prozesses bis zum Zeitpunkt t ist bekannt. Die zeitabhängigen Koeffizienten seien durch

$$\beta_j(t) = \beta_j + \theta_j g_j(t), \quad (4.48)$$

gegeben, wobei $g_j(t)$ ein bezüglich \mathcal{F}_t vorhersagbarer Prozess ist, der um 0 variiert. Dann gilt:

$$E(\hat{r}_k(\boldsymbol{\beta}) | \mathcal{F}_{t_k}) \approx V(\hat{\boldsymbol{\beta}}, t_k) G(t_k) \boldsymbol{\theta}$$

wobei $\hat{\boldsymbol{\beta}}$ der Vektor der geschätzten Regressionskoeffizienten im Modell (4.1) und $\hat{r}_k(\boldsymbol{\beta})$ der Vektor der Schoenfeldresiduen des k -ten Ereignisses (siehe (4.43)) ist. Weiter ist $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)'$ und $G(t)$ eine $p \times p$ -Diagonalmatrix mit $G_{jj}(t) = g_j(t)$. $V(\boldsymbol{\beta}, t)$ ist die in (4.47) definierte Kovarianzmatrix des Kovariablenvektors in Modell (4.1) zur Zeit t .

Beweis: Angenommen, die Hypothese $H : \boldsymbol{\beta}(t) = \boldsymbol{\beta} + G(t)\boldsymbol{\theta}$ gelte, d.h. die Regressionskoeffizienten sind zeitabhängig und $\boldsymbol{\beta}$ sei bekannt. Der Vektor der Schoenfeldresiduen des k -ten Ereignisses (siehe (4.43)) lässt sich schreiben als

$$\begin{aligned} \mathbf{r}_k(\boldsymbol{\beta}) &= \mathbf{Z}_{(k)} - \mathbf{M}(\boldsymbol{\beta}, t_k) \\ &= [\mathbf{Z}_{(k)} - \mathbf{M}(\boldsymbol{\beta}(t_k), t_k)] + [\mathbf{M}(\boldsymbol{\beta}(t_k), t_k) - \mathbf{M}(\boldsymbol{\beta}, t_k)], \end{aligned} \quad (4.49)$$

wobei $\mathbf{M}(\boldsymbol{\beta}, t)$ wie in (4.44) definiert ist.

Führt man eine Taylor-Entwicklung von $\mathbf{M}(\boldsymbol{\beta}(t_k), t_k)$ um $\boldsymbol{\beta}(t_k) = \boldsymbol{\beta}$ durch, erhält man

$$\begin{aligned} \mathbf{M}(\boldsymbol{\beta}(t_k), t_k) &= \mathbf{M}(\boldsymbol{\beta}, t_k) + V(\boldsymbol{\beta}, t_k)(\boldsymbol{\beta} + G(t_k)\boldsymbol{\theta} - \boldsymbol{\beta}) + o((G(t_k)\boldsymbol{\theta})^2) \\ &= \mathbf{M}(\boldsymbol{\beta}, t_k) + V(\boldsymbol{\beta}, t_k)G(t_k)\boldsymbol{\theta} + o((G(t_k)\boldsymbol{\theta})^2). \end{aligned} \quad (4.50)$$

Dann gilt mit (4.49) und (4.50)

$$\begin{aligned} E[\mathbf{r}_k(\boldsymbol{\beta}) | \mathcal{F}_{t_k}] &= E[\mathbf{Z}_{(k)} - \mathbf{M}(\boldsymbol{\beta}(t_k), t_k) | \mathcal{F}_{t_k}] + E[\mathbf{M}(\boldsymbol{\beta}(t_k), t_k) - \mathbf{M}(\boldsymbol{\beta}, t_k) | \mathcal{F}_{t_k}] \\ &= 0 + E[\mathbf{M}(\boldsymbol{\beta}, t_k) + V(\boldsymbol{\beta}, t_k)G(t_k)\boldsymbol{\theta} + o((G(t_k)\boldsymbol{\theta})^2) - \mathbf{M}(\boldsymbol{\beta}, t_k) | \mathcal{F}_{t_k}] \\ &\approx V(\boldsymbol{\beta}, t_k)G(t_k)\boldsymbol{\theta} + o((G(t_k)\boldsymbol{\theta})^2) \\ &\approx V(\boldsymbol{\beta}, t_k)G(t_k)\boldsymbol{\theta} \end{aligned}$$

Der erste Term von $\mathbf{r}_k(\boldsymbol{\beta})$ besitzt den Erwartungswert 0, da er die Schoenfeldresiduen des wahren, zeitabhängigen Modells angibt.

Nun sei $\boldsymbol{\beta}$ unbekannt und $\hat{\boldsymbol{\beta}}$ der MPL-Schätzer von $\boldsymbol{\beta}$ unter der Nullhypothese $H_0 : \boldsymbol{\beta}(t) = \boldsymbol{\beta}$. Analog zu oben erhält man

$$E[\hat{\mathbf{r}}_k(\boldsymbol{\beta}) | \mathcal{F}_{t_k}] \approx V(\hat{\boldsymbol{\beta}}, t_k) G(t_k) \boldsymbol{\theta}$$

□

Für die skalierten Schoenfeld-Residuen $\hat{\mathbf{r}}_k^*(\boldsymbol{\beta}) = V^{-1}(\hat{\boldsymbol{\beta}}, t_k)\hat{\mathbf{r}}_k(\boldsymbol{\beta})$ folgt dann

$$E[\hat{\mathbf{r}}_k^*(\boldsymbol{\beta})|\mathcal{F}_{t_k}] \approx G(t_k)\boldsymbol{\theta} \quad (4.51)$$

oder wegen (4.48) für die i-te Komponente

$$E(\hat{r}_{ki}^*|\mathcal{F}_{t_k}) + \hat{\beta}_i \approx \beta_i(t_k) \quad (4.52)$$

Ein geglätteter Plot von $\hat{r}_{ki}^* + \hat{\beta}_i$ gegen die Zeit t sollte daher in etwa eine Gerade mit Steigung 0 ergeben, falls keine zeitliche Abhängigkeit vorliegt. Ist die Proportional Hazards Annahme verletzt, gibt der Plot Auskunft über die Art der Abhängigkeit $\beta_i(t)$.

4.9.3 Modellierung zeitabhängiger Koeffizienten mittels fraktionaler Polynome

Berger et al. [3] stellen in ihrem Artikel eine Methode vor, die es ermöglicht, Zeitabhängigkeit aufzudecken und zugleich zu modellieren. Für den univariaten Fall mit einer Kovariable \mathbf{X} läßt sich (4.42) schreiben als

$$\lambda(t|\mathbf{X}) = \lambda_0(t) \exp[\boldsymbol{\beta}(t)'\mathbf{X}] = \lambda_0(t) \exp[\beta_0\mathbf{X} + \beta_1\varphi(t)\mathbf{X}], \quad (4.53)$$

wobei $\varphi(t)$ eine Funktion der Zeit ist, mittels der die Abhängigkeit modelliert werden soll. In der Regel werden für $\varphi(t)$ einfache Funktionen wie t oder $\ln(t)$ gewählt. Setzt man $\mathbf{Z}(t) = \varphi(t)\mathbf{X}$, d.h. es gilt

$$\lambda(t|\mathbf{X}) = \lambda_0(t) \exp[\beta_0\mathbf{X} + \beta_1\mathbf{Z}(t)]. \quad (4.54)$$

Damit kann (4.54) als die Hazardfunktion eines Modells mit konstanten Koeffizienten angesehen werden, das um eine zeitabhängige Kovariable $\mathbf{Z}(t)$, die Interaktion der Kovariable \mathbf{X} mit der Zeit, erweitert wurde. Für dieses Modell können nun die Regressionskoeffizienten mittels des Partial Likelihood geschätzt werden und die Hypothesentests aus Kapitel (4.3) angewandt werden.

Berger et al. behalten diesen Vorteil bei, lassen für $\varphi(t)$ jedoch flexiblere Funktionen zu. Sie ersetzen in (4.53) $\boldsymbol{\beta}$ durch ein fraktionales Polynom (siehe Kapitel 4.8.3):

$$\boldsymbol{\beta}(t) = \Phi_m(t, p) = \beta_0 + \sum_{j=1}^m \beta_j t^{p_j} \quad (4.55)$$

Eingesetzt in (4.53) ergibt sich damit

$$\lambda(t|\mathbf{X}) = \lambda_0(t) \exp[\Phi_m(t, p)\mathbf{X}] = \lambda_0(t) \exp[\beta_0\mathbf{X} + \sum_{j=1}^m \beta_j \mathbf{Z}_j(t)], \quad (4.56)$$

mit $\mathbf{Z}_j(t) = t^{p_j} \mathbf{X}$. Die konstanten Regressionskoeffizienten β_0, \dots, β_m können wieder mittels des Partial Likelihood Ansatzes geschätzt werden. Da die Interaktionen $\mathbf{Z}_j(t) = t^{p_j} \mathbf{X}$ jedoch Funktionen von der Zeit t sind, müssen die Daten dazu erst restrukturiert werden. Zu jedem Todeszeitpunkt muss die Risikomenge neu gebildet werden, so dass die zeitabhängigen Kovariablen für alle Individuen in der Risikomenge zu jedem Ereigniszeitpunkt die entsprechenden Werte annehmen. Dies geschieht, indem man den Datensatz eines jeden Individuums in mehrere Datensätze unterteilt, zu jedem Todeszeitpunkt wird ein neuer Datensatz angepaßt. Da die Anzahl der Daten dadurch sehr stark erhöht wird, ist diese Methode jedoch nur für relativ kleine Datensätze geeignet. Das Vorgehen bei der Restrukturierung soll an einem kleinen Beispieldatensatz (Tabelle 4.1) verdeutlicht werden: Der Datensatz enthält für 3 Patienten Angaben über den Beginn und das Ende

| Patient | Behandlungsbeginn | Behandlungsende | Geschlecht | Tod |
|---------|-------------------|-----------------|------------|-----|
| 1 | 0 | 400 | w | 0 |
| 2 | 0 | 150 | w | 1 |
| 3 | 0 | 520 | m | 0 |

Tabelle 4.1: Beispieldatensatz für 3 Patienten mit Beobachtungsbeginn und -ende (in Tagen), der Kovariable Geschlecht und Todesindikator Tod

der Beobachtung (in Tagen) und das Geschlecht. Der Indikator Tod nimmt den Wert 1 an, falls der Patient stirbt, 0 anderenfalls. Es liegt nur der Todeszeitpunkt 150 Tage vor. Die nach obigem Schema restrukturierten Daten finden sich in Tabelle 4.2.

Hat man die Daten restrukturiert, kann man bei der Modellbildung wie in Kapitel 4.8.3

| Patient | Beobachtungsbeginn | Beobachtungsende | Geschlecht | Tod |
|---------|--------------------|------------------|------------|-----|
| 1 | 0 | 150 | w | 0 |
| 1 | 151 | 400 | w | 0 |
| 2 | 0 | 150 | w | 1 |
| 3 | 0 | 150 | m | 0 |
| 3 | 151 | 520 | m | 0 |

Tabelle 4.2: Durch Unterteilung zum Todeszeitpunkt 150 Tage restrukturierter Beispieldatensatz

vorgehen. Nach Berger et al. [3] genügt es auch hier, sich auf fraktionale Polynome ersten und zweiten Grades zu beschränken und nur Potenzen aus der Menge

$\wp = \{-2, -1, -0.5, 0, 0.5, 1, 2\}$ zu wählen.

Die Proportional Hazard Annahme wird durch die Hypothese: $H_0 : \beta(t) = \beta_0$ getestet. Dies entspricht der Hypothese $H_0 : \beta_1 = \dots = \beta_m = 0$, so dass ein LR-Test mit $2m$ Freiheitsgraden, je zwei Freiheitsgrade pro Term des fraktionalen Polynoms (siehe Kapitel 4.8.3), angewendet werden kann.

Kapitel 5

Generalisierte Lineare Modelle (GLM)

Eine eingehende Behandlung von GLM's findet sich in McCullagh und Nelder [17]. Habermann und Pitacco [10] und Myers [19] bieten einen kurzen Überblick über GLM's. Da GLM's eine Verallgemeinerung der Linearen Modelle darstellen, werden die klassischen Linearen Modelle zuerst kurz vorgestellt. Dann werden kurz die wesentlichen Unterschiede von GLM's angegeben und auf die Bestimmung der Anpassungsgüte und statistische Hypothesentests eingegangen.

5.1 Klassische Lineare Modelle

Ziel von statistischen Modellen ist es eine Zielvariable, auch Response genannt, in Abhängigkeit von unabhängigen Kovariablen darzustellen. Ein lineares Modell besitzt die allgemeine Form

$$Y_i = \beta_0 + \sum_{k=1}^p x_{ik}\beta_k + \epsilon_i, \quad i = 1, \dots, n,$$

oder in Matrixschreibweise

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

wobei \mathbf{Y} der $n \times 1$ -Vektor von n unabhängig normalverteilten Zielvariablen Y_i und $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)'$ der $(p+1) \times 1$ -Vektor der unbekanntenen Regressionsparameter ist. Die $n \times (p+1)$ Matrix

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}$$

enthält, neben einem Einsvektor für den Intercept β_0 , die p Kovariablenvektoren $\mathbf{x}_k = (x_{1k}, x_{2k}, \dots, x_{nk})'$, $k = 1, \dots, p$. Dabei ist x_{ij} der Wert der j -ten Kovariable von Beobachtung i . $\boldsymbol{\epsilon}$ ist der $n \times 1$ -Vektor der Fehlerterme ϵ_i mit Erwartungswert $E(\boldsymbol{\epsilon}) = \mathbf{0}$ und Varianz $Var(\boldsymbol{\epsilon}) = \sigma^2 I_n$. Neben der Varianzhomogenität wird die Annahme getroffen, dass

die Fehlerterme ϵ_i , $i = 1, \dots, n$ normalverteilt sind, d.h. es gilt

$$\epsilon_i \sim N(0, \sigma^2) \quad \text{i.i.d.} \quad i = 1, \dots, n.$$

Für den Responsevektor gilt daher aufgrund der Unabhängigkeit

$$E(\mathbf{Y}) = \boldsymbol{\mu} = \beta_0 + \sum_{j=1}^p \beta_j \mathbf{x}_j = X\boldsymbol{\beta},$$

und $Var(\mathbf{Y}) = \sigma^2 I_n$.

Der Erwartungswert für die i -te Beobachtung ist dann gegeben durch

$$E(Y_i) = \mu_i = \beta_0 + \sum_{j=1}^p \beta_j x_{ij},$$

Ein einfaches Beispiel ist die Regression auf eine Gerade. In diesem Fall sind die Y_i unabhängige, $N(\alpha + \beta x_i, \sigma^2)$ verteilte Zufallsvariablen. Die Regressionskoeffizienten erhält man durch Kleinste Quadrate-Schätzung. Da die Responsevariablen als unabhängig normalverteilt angenommen werden, stimmt diese mit der Maximum Likelihood Schätzung überein.

5.2 Exponentialfamilien

Zu den Exponentialfamilien zählen alle Verteilungen, deren Dichte auf die Form

$$f(y|\theta) = \exp\left[\frac{y \cdot \theta - b(\theta)}{a(\Phi)} + c(y, \Phi)\right] \quad (5.1)$$

gebracht werden können. Dabei ist θ der natürliche Parameter und Φ der Skalen- oder Dispersionsparameter. Oft wird $a(\Phi) = \Phi$ angenommen. $a(\cdot)$, $b(\cdot)$ und $c(\cdot)$ sind Funktionen, die die jeweilige Verteilung determinieren. Man beachte, dass $c(y, \Phi)$ nicht von θ abhängt. Man kann zeigen (siehe McCullagh und Nelder [17], S.28), dass der Erwartungswert und die Varianz von Zufallsvariablen von Exponentialfamilien leicht aus den folgenden Zusammenhängen erhalten werden .

$$\mu = E(Y) = \frac{\partial}{\partial \theta} b(\theta) \quad (5.2)$$

$$Var(Y) = \Phi \frac{\partial^2}{\partial \theta^2} b(\theta) \quad (5.3)$$

$\frac{\partial^2}{\partial \theta^2} b(\theta)$ wird als Varianzfunktion bezeichnet. Da sie über θ vom Erwartungswert μ abhängt kann sie auch als $\frac{\partial^2}{\partial \theta^2} b(\theta) = V(\mu)$ formuliert werden.

Beispiele:

- Normalverteilung

Y sei normalverteilt mit Erwartungswert $E(Y) = \mu$ und Varianz $V(Y) = \sigma^2$. Die Dichte

$$f(y|\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right]$$

lässt sich formulieren als

$$f(y|\mu) = \exp\left[\frac{y \cdot \mu - \frac{1}{2}\mu^2}{\sigma^2} - \frac{y^2}{2\sigma^2} - \frac{1}{2} \log(2\pi\sigma^2)\right]$$

In diesem Fall gilt:

$$\theta = \mu, \quad b(\theta) = \frac{\theta^2}{2}, \quad \Phi = \sigma^2, \quad c(y, \Phi) = -\frac{1}{2} \log(2\pi\sigma^2), \quad V(\mu) = 1$$

Aus (5.2) folgt:

$$E(Y) = \frac{\partial}{\partial \theta} b(\theta) = \theta = \mu \quad \text{Var}(Y) = \Phi \frac{\partial^2}{\partial \theta^2} b(\theta) = \sigma^2$$

- Poissonverteilung

Die Dichte einer mit Parameter μ poissonverteilten Zufallsvariable Y ist gegeben durch

$$f(y|\mu) = \frac{\mu^y}{y!} \exp[-\mu]$$

und kann leicht in die Form (5.1) umgeschrieben werden:

$$f(y|\mu) = \frac{\mu^y}{y!} \exp[-\mu] = \exp[y \cdot \log \mu - \mu - \log(y!)]$$

Hier gilt:

$$\theta = \log \mu, \quad b(\theta) = \mu = \exp[\theta], \quad \Phi = 1, \quad c(y, \Phi) = -\log(y!), \quad V(\mu) = \mu$$

Daraus folgt

$$E(Y) = \frac{\partial}{\partial \theta} b(\theta) = \exp[\theta]$$

$$\text{Var}(y) = \frac{\partial^2}{\partial \theta^2} b(\theta) = \exp[\theta]$$

Weitere Verteilungen die zu den Exponentialfamilien zählen, sind die Binomial- oder Gamma-verteilung.

5.3 Generalisierte Lineare Modelle

Die klassischen Linearen Modelle können verallgemeinert werden, indem man die Normalverteilungsannahme für die Responsevariablen aufhebt und stattdessen auch Verteilungen von Exponentialfamilien zulässt.

Darüberhinaus wird der lineare Prädiktor $\boldsymbol{\eta} = X\boldsymbol{\beta}$ über die sogenannte Linkfunktion g mit dem Erwartungswert $\boldsymbol{\mu}$ verbunden:

$$\boldsymbol{\eta} = g(\boldsymbol{\mu}) \quad (5.4)$$

Die Linkfunktion wird als monoton differenzierbare Funktion vorausgesetzt, so dass gilt:

$$\boldsymbol{\mu} = g^{-1}(\boldsymbol{\eta}) = g^{-1}(X\boldsymbol{\beta})$$

Gilt $\boldsymbol{\eta} = \boldsymbol{\theta}$, wobei $\boldsymbol{\theta}$ der natürliche Parameter der Exponentialfamilie ist, so spricht man vom kanonischen Link. Im Gegensatz zu den klassischen Linearen Modellen wird die Varianz nicht mehr als konstant angenommen.

Beispiele (siehe Beispiele in Abschnitt (5.1)):

- Normalverteilung:

$$\boldsymbol{\eta} = X\boldsymbol{\beta} = \boldsymbol{\mu} = \boldsymbol{\theta} \quad \Rightarrow \quad \text{kanonische Linkfunktion: } g(\boldsymbol{\mu}) = \boldsymbol{\mu}$$

- Poissonverteilung:

$$\boldsymbol{\eta} = \boldsymbol{\theta} = \log \boldsymbol{\mu} \quad \Rightarrow \quad g(\boldsymbol{\mu}) = \log(\boldsymbol{\mu})$$

ist kanonische Linkfunktion

5.3.1 Bestimmung der Anpassungsgüte bei GLM's

Ein Goodness-of-Fit- oder Diskrepanzmaß bei GLM's ist die Devianz. Liegen n Beobachtungen y_1, y_2, \dots, y_n vor, können Modelle mit maximal n Parametern $\mu_1, \mu_2, \dots, \mu_n$ angepaßt werden. Das einfachste Modell, welches nur einen zu schätzenden Parameter enthält, der den gemeinsamen Erwartungswert der Daten repräsentiert, wird Nullmodell genannt. Dem gegenüber steht das volle Modell, auch saturiertes Modell genannt, mit n Parametern. In diesem Fall stimmen die Daten mit den geschätzten Parametern überein, d.h. es gilt $\hat{\mu}_i = y_i, i = 1, \dots, n$, so daß kein Fehler mehr vorliegt. Während das Nullmodell im allgemeinen zu einfach ist, enthält das saturierte Modell keine Informationen mehr, da das Modell die Daten lediglich reproduziert.

Dennoch ist das volle Modell als Grundlage zur Bewertung kleinerer Modelle geeignet. Ein Anpassungsmaß ist die Devianz $D(\mathbf{y}, \hat{\boldsymbol{\mu}})$. Sie ist bestimmt durch die doppelte Differenz zwischen dem Log-Likelihood des vollen Modells und dem Log-Likelihood des zu bewertenden Modells :

$$\frac{D(\mathbf{y}, \hat{\boldsymbol{\mu}})}{\Phi} = -2[LL(\mathbf{y}, \hat{\boldsymbol{\mu}}, \Phi) - LL(\mathbf{y}, \mathbf{y}, \Phi)] \quad (5.5)$$

Ein weiteres Diskrepanzmaß ist die Pearson χ^2 -Statistik. Sie ist gegeben durch

$$\chi^2 = \sum_{i=1}^n \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}, \quad (5.6)$$

wobei $V(\hat{\mu}_i)$ die geschätzte Varianzfunktion ist.

Es kann gezeigt werden (Rao [20]), dass sowohl die Devianz als auch die Pearson χ^2 -Statistik asymptotisch χ^2 -verteilt mit $n-q$ Parametern sind, wobei q die Anzahl der unbekannt Parameter im Modell ist.

5.3.2 Statistische Hypothesentests

Ein Test, der die Anpassung des Modells überprüft, ist der *Residual-Devianz-Test*. Es wird die Nullhypothese

$$H_0 : \text{Modell ist gut angepaßt}$$

gegen die Alternative H_1 : nicht H_0 getestet. Die Nullhypothese wird zum Niveau α verworfen, wenn die Devianz größer als das $1 - \alpha$ -Quantil der χ_{n-q}^2 -Verteilung ist, d.h. wenn gilt: Devianz $> \chi_{n-q, 1-\alpha}^2$. Dieser Test gibt Aufschluß darüber, ob das Modell schon ausreichend spezifiziert ist, oder ob noch ein Lack-of-Fit vorliegt.

Der *Partielle Devianz-Test* testet, ob die Hinzunahme weiterer Kovariablen in das Modell signifikanten Einfluß besitzt. Es werden zwei Modelle miteinander verglichen:

Modell 1: $\boldsymbol{\eta} = X_1\boldsymbol{\beta}_1 + X_2\boldsymbol{\beta}_2$ mit $p_1 + p_2 = p$ Parametern

Modell 2: $\boldsymbol{\eta} = X_1\boldsymbol{\beta}_1$ mit p_1 Parametern

Getestet wird die Nullhypothese $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$ gegen die Alternative $H_1 : \boldsymbol{\beta}_2 \neq \mathbf{0}$. Als Teststatistik wird die Differenz zwischen der Devianz des kleineren und der Devianz des größeren Modelles betrachtet, d.h. Devianz (Modell 2) - Devianz (Modell 1). Unter der Nullhypothese ist diese Größe χ^2 -verteilt mit p_2 Parametern (siehe Rao [20]). Die Nullhypothese wird daher zum Niveau α verworfen, falls die Teststatistik größer als das $1 - \alpha$ -Quantil der χ^2 -Verteilung mit p_2 Parametern ist, d.h.

Devianz (Modell 2) - Devianz (Modell 1) $> \chi_{p_2, 1-\alpha}^2$.

5.4 Stückweises Exponentialmodell und Poissonmodell

Das stückweise Exponentialmodell ist ein Spezialfall des Proportional Hazard Modells mit einschränkenden Bedingungen an den Baselinehazard. Holford [11] und Laird und Olivier [16] zeigen die Äquivalenz dieses Modells zu einem Poissonmodell.

Im folgenden wird zuerst das stückweise Exponentialmodell vorgestellt. Es wird ein Proportional Hazard Modell der bekannten Form

$$\lambda_i(t|\mathbf{Z}_i) = \lambda_0(t) \exp[\boldsymbol{\beta}'\mathbf{Z}_i] \quad (5.7)$$

betrachtet. Die Beobachtungszeit wird in J Intervalle mit Intervallgrenzen $0 = \tau_0 < \tau_1 < \dots < \tau_J = \infty$ unterteilt. Innerhalb jedes Intervalles wird der Baselinehazard $\hat{\lambda}_0(t)$ als

konstant angenommen, d.h. es gilt

$$\lambda_0(t) = \lambda_j \quad \text{für } t \in [\tau_{j-1}, \tau_j) \quad (5.8)$$

Die Hazardrate für das i -te Individuum im Intervall j ist daher durch

$$\lambda_{ij} = \lambda_j \exp[\boldsymbol{\beta}'\mathbf{Z}_i] \quad (5.9)$$

gegeben. Durch Logarithmieren erhält man

$$\log \lambda_{ij} = \log \lambda_j + \boldsymbol{\beta}'\mathbf{Z}_i = \alpha_j + \boldsymbol{\beta}'\mathbf{Z}_i, \quad (5.10)$$

ein log-lineares Modell, in welches die kategorisierte Beobachtungszeit als ein Faktor mit den Levels $\alpha_j = \log \lambda_j, j = 1, \dots, J$ eingeht. Identifizierbarkeit ist durch die Bedingung $\alpha_1 = 0$ sichergestellt. Nach Holford [11] und Laird und Olivier [16] kann das stückweise Exponentialmodell durch ein Poisson-Regressionsmodell angepasst werden .

Es sei t_i die Überlebens- bzw. Zensierungszeit des i -ten Individuums und d_i der Indikator

$$d_i = \begin{cases} 1 & \text{Individuum } i \text{ stirbt} \\ 0 & \text{sonst} \end{cases} .$$

Nun werden Pseudobeobachtungen gebildet, indem man für jedes Individuum und jedes Intervall eine einzelne Beobachtung bildet. Die Zeit t_{ij} , die das i -te Individuum im Intervall $[\tau_{j-1}, \tau_j)$ verbringt, die sogenannte „ exposure time “, ist bestimmt durch

$$t_{ij} = \begin{cases} \tau_j - \tau_{j-1}, & t_i > \tau_j \\ t_i - \tau_{j-1}, & \tau_{j-1} < t_i < \tau_j \\ 0, & t_i < \tau_{j-1} \end{cases} . \quad (5.11)$$

Die zugehörigen Indikatoren für Todesfälle werden durch

$$d_{ij} = \begin{cases} 1 & \text{Individuum } i \text{ stirbt in Intervall } j \\ 0 & \text{sonst} \end{cases} \quad (5.12)$$

definiert. Weiter gebe $j(i)$ das Intervall an, in welchem Individuum i stirbt oder zensiert wird, d.h. das Intervall, in welches t_i fällt. Für $t_i \in j(i)$ gilt also $d_{ij} = 0$ für alle vorherigen Intervalle, d.h. $\forall j < j(i)$, und $d_{ij} = d_i$ für $j = j(i)$.

Die Parameter des stückweise Exponentialmodells können nun geschätzt werden, indem man die Todesindikatoren d_{ij} als unabhängige poissonverteilte Beobachtungen mit Erwartungswert $\mu_{ij} = t_{ij} \lambda_{ij}$ behandelt. Mit (5.9) und dem Logarithmus als Linkfunktion folgt

$$\log \mu_{ij} = \log t_{ij} + \alpha_j + \boldsymbol{\beta}'\mathbf{Z}_i.$$

Ein Vergleich mit (5.4) zeigt, dass der Term $\log t_{ij}$ bei der Modellierung des Poissonmodells subtrahiert werden muss. Dies wird praktisch umgesetzt, indem man die logarithmierte exposure time $\log t_{ij}$ als Offset wählt, was bewirkt, dass dieser Term als ein Teil des Prädiktors $\alpha_j + \boldsymbol{\beta}'\mathbf{Z}_i$ mit bekanntem Regressionskoeffizienten 1 betrachtet wird.

Hier ist anzumerken, dass die d_{ij} nicht wirklich poissonverteilt sind, da sie nur die Werte 0 oder 1 annehmen. Auch die Unabhängigkeit ist nicht gegeben, es gilt jedoch folgender Satz:

Satz 5.1 Die Likelihoodfunktionen des stückweise Exponentialmodells und des Poissonmodells $d_{ij} \sim \text{Pois}(\mu_{ij})$ sind zueinander proportional.

Beweis: Sei $\lambda_i(t)$ der Hazard und $\Lambda_i(t)$ der kumulierte Hazard des i -ten Individuums zur Zeit t , dann ist der Log-Likelihood der i -ten Beobachtung nach (4.19) und mit den vorher eingeführten Bezeichnungen gegeben durch:

$$\begin{aligned} LL_i &= d_i \log \lambda_i(t_i) - \Lambda_i(t_i) \\ &= d_{ij(i)} \log \lambda_{ij(i)} - \int_0^{t_i} \lambda_i(t) dt \\ &= d_{ij(i)} \log \lambda_{ij(i)} - \sum_{j=1}^{j(i)} t_{ij} \lambda_{ij} \\ &= \sum_{j=1}^{j(i)} [d_{ij} \log \lambda_{ij} - t_{ij} \lambda_{ij}]. \end{aligned}$$

Bei der letzten Gleichung wurde verwendet, dass $d_{ij} = 0 \quad \forall j < j(i)$. Die Log-Likelihoodfunktion für eine poissonverteilte Beobachtung d_{ij} mit Erwartungswert $\mu_{ij} = t_{ij} \lambda_{ij}$ lautet

$$\begin{aligned} LL_{ij} &= d_{ij} \log \mu_{ij} - \mu_{ij} \\ &= d_{ij} \log(t_{ij} \lambda_{ij}) - t_{ij} \lambda_{ij}. \end{aligned}$$

Es gilt nun

$$\begin{aligned} LL(\text{Poisson}) &= \sum_{i=1}^n \sum_{j=1}^{j(i)} LL_{ij} \\ &= \sum_{i=1}^n \sum_{j=1}^{j(i)} d_{ij} \log t_{ij} + d_{ij} \log \lambda_{ij} - t_{ij} \lambda_{ij} \\ &\propto \sum_{i=1}^n \sum_{j=1}^{j(i)} d_{ij} \log \lambda_{ij} - t_{ij} \lambda_{ij} = LL(\text{st. Exponential}), \end{aligned}$$

da $d_{ij} \log t_{ij}$ eine von den Parametern unabhängige Konstante ist. □

Durch die Einführung der Pseudobeobachtungen kann die Zahl der Beobachtungen stark ansteigen. Dem kann entgegengewirkt werden, indem man die Beobachtungen nach den Ausprägungen der Kovariablen gruppiert. In diesem Fall gibt d_{ij} die Anzahl der Todesfälle und t_{ij} die gesamte exposure time der Individuen mit Merkmalen \mathbf{Z}_i im Intervall j an.

Kapitel 6

Überlebenszeitanalyse in der Pflegeversicherung

In diesem Kapitel wird ein Datensatz aus der Pflegeversicherung mit Hilfe der in Kapitel 3 beschriebenen Methoden analysiert. Im folgenden Abschnitt werden die bearbeiteten Daten vorgestellt.

6.1 Beschreibung der Daten

Die Daten wurden von insgesamt 5603 Leistungsempfängern in der privaten gesetzlichen Pflegeversicherung im Zeitraum vom 1. April 1995 bis zum 31. Dezember 1998 aufgenommen. 3511 der Patienten sind Frauen, 2092 Männer. Neben Alter und Geschlecht ist von jedem Patienten der Pflegezustand bekannt, der durch die Art der Behandlung und die Pflegestufe bestimmt ist. Bei der Behandlungsart wird zwischen ambulanter und stationärer Pflegestufe unterschieden. Es gibt drei verschiedene Pflegestufen:

- Stufe 1 : erheblich pflegebedürftig
- Stufe 2 : schwer pflegebedürftig
- Stufe 3 : schwerst pflegebedürftig

Bei 5044 Leistungsempfängern sind auch die pflegeauslösenden Diagnosen bekannt. Insgesamt liegen 11 verschiedene Diagnosen vor: Tumore, Psychose, Arthrosen, Schlaganfälle, Herz-, Lungen-, Demenz-, Nierenerkrankungen, Knochenkrankheiten, Blindheit und Geburtsschäden. In Tabelle 6.1 sind die verschiedenen Variablen zusammengefaßt und mit ihrer Kodierung angegeben.

Der Zustand eines Patienten kann sich während der Dauer der Pflegebedürftigkeit ändern. Der Patient kann sowohl in eine andere Pflegestufe übergehen, als auch eine andere Behandlungsart erfahren. Die Struktur der Daten ist in Abbildung 6.1 anhand eines Beispiels mit zwei Patienten dargestellt. Patient 1, der durch dünne Linien gekennzeichnet ist, wird während des Beobachtungszeitraums pflegebedürftig und erhält ambulante Pflege in Stufe 2. Mit der Zeit verschlechtert sich sein Zustand, er geht erst in Pflegestufe 3 bei gleicher Behandlungsart über und wechselt dann in stationäre Behandlung mit Stufe 2.

| Variable | Beschreibung | Werte |
|-------------------|---|--------------------------------------|
| $Z_{Alter}(t)$ | Alter bei Eintritt in eine neue Pflegestufe | 0-108 Jahre |
| $Z_{Geschlecht}$ | Geschlecht | 0 = männlich, 1 = weiblich |
| $Z_{stat}(t)$ | Art der Behandlung zum Zeitpunkt t | 0 = ambulant, 1 = stationär |
| $Z_{Stufe2}(t)$ | Pflegestufe zum Zeitpunkt t | 0 = Stufe1 oder Stufe3, 1 = Stufe2 |
| $Z_{Stufe3}(t)$ | Pflegestufe zum Zeitpunkt t | 0 = Stufe1 oder Stufe2, 1 = Stufe3 |
| $Z_{Diagnose\ i}$ | pflegeauslösende Diagnose | 0 = nicht Diagnose i, 1 = Diagnose i |

Tabelle 6.1: Beschreibung und mögliche Werte der Variablen

Dort verweilt er einige Zeit und stirbt dann. Patient 2 (fette Linien) war schon vor Beginn des Beobachtungszeitraumes in Pflege, sein Zustand ist jedoch erst ab dem 1.4.1995, dem Beginn der Datenaufnahme, bekannt. Zu diesem Zeitpunkt befindet er sich in ambulanter Pflege in Stufe 1, über seinen vorherigen Pflegeverlauf weiß man nichts. Auch sein Zustand verschlechtert sich. Er wechselt in stationäre Pflege mit Stufe 1 und später in Stufe 3. In diesem Zustand befindet er sich auch noch zum Ende des Beobachtungszeitraumes. Da sein Tod nicht beobachtet werden konnte, wird er zum 31.12.1998 zensiert. In Tabelle

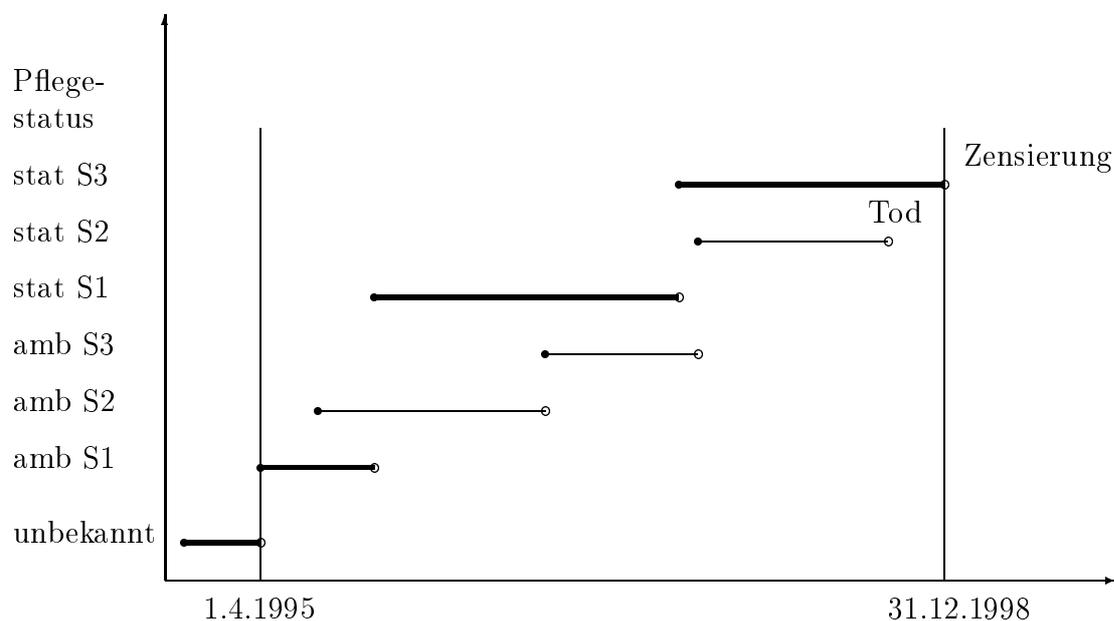


Abbildung 6.1: Beispiel zur Struktur der Daten für zwei Patienten: Patient 1 (dünne Linien), Patient 2 (fette Linien) (amb = ambulant, stat = stationär, S1 = Stufe 1, S2 = Stufe 2, S3 = Stufe 3)

6.2 sind die Häufigkeiten der verschiedenen Zustandsübergänge aufgelistet. Man erkennt deutlich, dass sich der Zustand der Patienten in den meisten Fällen verschlechtert, Ver-

| | amb S1 | amb S2 | amb S3 | stat S1 | stat S2 | stat S3 | Tod | reaktiviert | zensiert |
|---------|--------|--------|--------|---------|---------|---------|-----|-------------|----------|
| amb S1 | - | 408 | 73 | 107 | 67 | 32 | 253 | 21 | 847 |
| amb S2 | 38 | - | 274 | 9 | 197 | 56 | 535 | 1 | 786 |
| amb S3 | 2 | 18 | - | 0 | 4 | 82 | 588 | 1 | 283 |
| stat S1 | 10 | 1 | 0 | - | 103 | 25 | 77 | 1 | 212 |
| stat S2 | 1 | 5 | 1 | 6 | - | 111 | 249 | 2 | 414 |
| stat S3 | 1 | 1 | 6 | 2 | 9 | - | 418 | 0 | 355 |

Tabelle 6.2: Anzahl der Zustandsübergänge (amb = ambulant, stat = stationär, S1 = Stufe 1, S2 = Stufe 2, S3 = Stufe 3)

besserungen sind nur selten festzustellen.

Da bei der Modellbildung besonders auf den Einfluß der Diagnosen auf das Überleben eingegangen werden soll, basieren die folgenden Beschreibungen der Daten nur auf den Daten der 5044 Patienten, wo die Diagnosen bekannt sind. Unter diesen befinden sich 3176 Frauen und 1868 Männer. Nur 45.5 % der Patienten wurden durch genau eine Diagnose pflegebedürftig, bei allen anderen liegen Mehrfachdiagnosen vor. In Tabelle 6.3 ist das Vorkommen von Einzel-, Doppel- und Mehrfachdiagnosen genauer aufgelistet.

| Anzahl der Diagnosen | 1 | 2 | 3 | 4 | 5 | 6 |
|----------------------|------|------|-----|-----|-----|------|
| Anzahl der Patienten | 2296 | 1830 | 708 | 178 | 27 | 3 |
| in Prozent | 45.5 | 36.3 | 14 | 3.5 | 0.5 | 0.05 |

Tabelle 6.3: Diagnoseanzahl und Häufigkeit

Die Häufigkeit der verschiedenen Diagnosen und der jeweilige Anteil an Einzeldiagnosen sind in Tabelle 6.4 angeführt. Die Diagnosen Nierenerkrankungen, Blindheit und Geburtsschäden wurden aufgrund ihres geringen Vorkommens zur Gruppe *sonstige* zusammengefaßt.

In den Tabellen 6.5 und 6.6 sind die Häufigkeiten der verschiedenen Diagnose-Kombinationen bei den Doppeldiagnosen und den größten Gruppen der Dreifachdiagnosen zusammengestellt.

| Diagnose | Gesamtanzahl (inklusive Mehrfachdiagnosen) | Einzeldiagnosen | in Prozent |
|-----------|---|-----------------|------------|
| Tumore | 694 | 276 | 39.8 |
| Psychose | 1254 | 394 | 31.4 |
| Herz | 1922 | 378 | 19.7 |
| Schlag | 1044 | 309 | 29.6 |
| Arthrosen | 534 | 85 | 15.9 |
| Lungen | 93 | 16 | 12.9 |
| Demenz | 2151 | 587 | 27.3 |
| Knochen | 1015 | 189 | 18.6 |
| sonstige | 226 | 66 | 29.2 |

Tabelle 6.4: Häufigkeit der verschiedenen Diagnosen (mit Mehrfachdiagnosen) und Anteil an Einzeldiagnosen

| | Psychose | Herz | Schlag | Arthrosen | Lungen | Demenz | Knochen | sonstige |
|-----------|----------|------|--------|-----------|--------|--------|---------|----------|
| Tumore | 52 | 65 | 46 | 7 | 4 | 54 | 36 | 4 |
| Psychose | | 109 | 84 | 22 | 3 | 141 | 45 | 23 |
| Herz | | | 105 | 72 | 12 | 351 | 128 | 17 |
| Schlag | | | | 8 | 1 | 135 | 28 | 10 |
| Arthrosen | | | | | 1 | 66 | 31 | 5 |
| Lungen | | | | | | 6 | 4 | 0 |
| Demenz | | | | | | | 134 | 15 |
| Knochen | | | | | | | | 5 |

Tabelle 6.5: Häufigkeiten der verschiedenen Diagnose-Kombinationen bei den Doppeldiagnosen

| Diagnosen | Häufigkeit |
|----------------------------|------------|
| Psychose, Herz und Demenz | 74 |
| Herz, Schlag und Demenz | 56 |
| Herz, Arthrosen und Demenz | 54 |

Tabelle 6.6: Häufigkeiten der wichtigsten Diagnose-Kombinationen bei den Dreifachdiagnosen

6.2 Modellierung der Überlebenszeiten unter Berücksichtigung der Diagnosen

Ein Ziel dieser Arbeit ist es, die pflegeauslösenden Diagnosen mit in ein Modell für die Überlebenszeiten mit einzubeziehen. Da diese jedoch, wie schon im vorangegangenen Abschnitt erwähnt, nur bei 5044 der Pflegefälle des Datensatzes erfasst wurden, werden alle folgenden Modelle nur auf Grundlage der Daten dieser 5044 Patienten gefittet. Die Modellierung wird in Splus durchgeführt. Dort steht die Funktion `coxph` zur Schätzung der Regressionskoeffizienten im Proportional Hazard Modell zur Verfügung.

Als Ausgangsmodell wird das Endmodell von Rudolph [23] verwendet, in dem die Überlebenszeit bereits durch das Alter, das Geschlecht, die Behandlungsart und die Pflegestufe der Patienten modelliert wird. Die Hazardfunktion in diesem Modell ist durch

$$\begin{aligned} \lambda(t) = & \lambda_0(t) \exp[\beta_1 \mathbf{Z}_{Alter}(t) + \beta_2 \mathbf{Z}_{Geschlecht} + \beta_3 \mathbf{Z}_{Stat}(t) + \beta_4 \mathbf{Z}_{Stufe2}(t) + \beta_5 \mathbf{Z}_{Stufe3}(t) \\ & + \beta_6 \mathbf{Z}_{Alter}(t) \times \mathbf{Z}_{Geschlecht} + \beta_7 \mathbf{Z}_{Alter}(t) \times \mathbf{Z}_{Stat}(t) + \beta_8 \mathbf{Z}_{Geschlecht} \times \mathbf{Z}_{Stat}(t) \\ & + \beta_9 \mathbf{Z}_{Stat}(t) \times \mathbf{Z}_{Stufe2}(t) + \beta_{10} \mathbf{Z}_{Stat}(t) \times \mathbf{Z}_{Stufe3}(t)] \end{aligned} \quad (6.1)$$

gegeben. Die Interaktionen $\mathbf{Z}_{Alter}(t) \times (\mathbf{Z}_{Stufe2}(t) + \mathbf{Z}_{Stufe3}(t))$ und $\mathbf{Z}_{Geschlecht} \times (\mathbf{Z}_{Stufe2}(t) + \mathbf{Z}_{Stufe3}(t))$ sind in (6.1) nicht enthalten, da sie keinen signifikanten Beitrag zur Modellierung der Überlebenszeit liefern (siehe Rudolph [23]). Auch bei Hinzunahme der Diagnosen als Kovariablen ins Modell gewinnen diese vier Interaktionen nicht an Signifikanz und bleiben daher weiter unberücksichtigt. Die mittels der Efron-Approximation für den Partial-Likelihood gewonnenen Koeffizienten für Modell (6.1) sind in Tabelle 6.7 zu finden.

Der partielle Log-Likelihood dieses Modells liegt bei -14467.12, bei 10 Freiheitsgraden resultiert daraus ein AIC von 28954.24.

| Variable | Koeffizient | exp(Koeff.) | SE(Koeff.) | z-Statistik | p-Wert |
|---|-------------|-------------|------------|-------------|---------------------|
| \mathbf{Z}_{Alter} | 0.0243 | 1.025 | 0.0026 | 9.19 | $< 10^{-16}$ |
| $\mathbf{Z}_{Geschlecht}$ | 0.3482 | 1.417 | 0.3049 | 1.14 | $2.5 \cdot 10^{-1}$ |
| \mathbf{Z}_{stat} | -0.5254 | 0.591 | 0.4299 | -1.22 | $2.2 \cdot 10^{-1}$ |
| \mathbf{Z}_{Stufe2} | 0.7625 | 2.144 | 0.0768 | 9.93 | $< 10^{-16}$ |
| \mathbf{Z}_{Stufe3} | 1.7007 | 5.478 | 0.0770 | 22.09 | $< 10^{-16}$ |
| $\mathbf{Z}_{Alter} \times \mathbf{Z}_{Geschlecht}$ | -0.0067 | 0.993 | 0.0037 | -1.78 | $7.5 \cdot 10^{-2}$ |
| $\mathbf{Z}_{Alter} \times \mathbf{Z}_{stat}$ | 0.0149 | 1.015 | 0.0050 | 3.01 | $2.6 \cdot 10^{-3}$ |
| $\mathbf{Z}_{Geschlecht} \times \mathbf{Z}_{stat}$ | -0.4302 | 0.650 | 0.1024 | -4.20 | $2.7 \cdot 10^{-5}$ |
| $\mathbf{Z}_{stat} \times \mathbf{Z}_{Stufe2}$ | -0.2729 | 0.761 | 0.1515 | -1.80 | $7.2 \cdot 10^{-2}$ |
| $\mathbf{Z}_{stat} \times \mathbf{Z}_{Stufe3}$ | -0.6578 | 0.518 | 0.1456 | -4.53 | $6.2 \cdot 10^{-6}$ |

Tabelle 6.7: Geschätzte Koeffizienten im Modell (6.1)

Dieses Modell wird nun um die Diagnosen erweitert. Jede Diagnose wird in Form einer Indikatorvariable hinzugefügt, die wie folgt definiert ist:

$$\mathbf{Z}_{Diagnose\ i} = \begin{cases} 1 & \text{Diagnose } i \text{ wurde bei Patient festgestellt} \\ 0 & \text{sonst} \end{cases}$$

Die Hazardfunktion für dieses Modell ist von folgender Gestalt:

$$\begin{aligned}
 \lambda(t) = & \lambda_0(t) \exp[\beta_1 \mathbf{Z}_{Alter}(t) + \beta_2 \mathbf{Z}_{Geschlecht} + \beta_3 \mathbf{Z}_{Stat}(t) + \beta_4 \mathbf{Z}_{Stufe2}(t) + \beta_5 \mathbf{Z}_{Stufe3}(t) \\
 & + \beta_6 \mathbf{Z}_{Alter}(t) \times \mathbf{Z}_{Geschlecht} + \beta_7 \mathbf{Z}_{Alter}(t) \times \mathbf{Z}_{Stat}(t) + \beta_8 \mathbf{Z}_{Geschlecht} \times \mathbf{Z}_{Stat}(t) \\
 & + \beta_9 \mathbf{Z}_{Stat}(t) \times \mathbf{Z}_{Stufe2}(t) + \beta_{10} \mathbf{Z}_{Stat}(t) \times \mathbf{Z}_{Stufe3} + \beta_{11} \mathbf{Z}_{Demenz} + \beta_{12} \mathbf{Z}_{Schlag} \\
 & + \beta_{13} \mathbf{Z}_{Psychose} + \beta_{14} \mathbf{Z}_{Tumore} + \beta_{15} \mathbf{Z}_{Blindheit} + \beta_{16} \mathbf{Z}_{Knochenk} + \beta_{17} \mathbf{Z}_{Arthrosen} \\
 & + \beta_{18} \mathbf{Z}_{Herz} + \beta_{19} \mathbf{Z}_{Lungen} + \beta_{20} \mathbf{Z}_{Nieren} + \beta_{21} \mathbf{Z}_{Geburt}] \quad (6.2)
 \end{aligned}$$

Für die Haupteffekte der Diagnosevariablen wird vorerst ein Signifikanzniveau von 10 % gewählt. Tabelle 6.8 enthält die geschätzten Koeffizienten der Diagnosevariablen in Modell (6.2).

| Variable | Koeffizient | exp(Koeff.) | SE(Koeff.) | z-Statistik | p-Wert |
|--------------------------|-------------|-------------|------------|-------------|---------------------|
| \mathbf{Z}_{Demenz} | -0.1351 | 0.874 | 0.0483 | -2.799 | $5.1 \cdot 10^{-3}$ |
| \mathbf{Z}_{Schlag} | -0.2067 | 0.813 | 0.0573 | -3.606 | $3.1 \cdot 10^{-4}$ |
| $\mathbf{Z}_{Psychose}$ | -0.2056 | 0.814 | 0.0549 | -3.743 | $1.8 \cdot 10^{-4}$ |
| \mathbf{Z}_{Tumore} | 1.0504 | 2.859 | 0.0570 | 18.430 | $< 10^{-16}$ |
| $\mathbf{Z}_{Blindheit}$ | -0.1071 | 0.898 | 0.1133 | -0.945 | $3.4 \cdot 10^{-1}$ |
| $\mathbf{Z}_{Knochenk}$ | -0.818 | 0.921 | 0.0583 | -1.403 | $1.6 \cdot 10^{-1}$ |
| $\mathbf{Z}_{Arthrosen}$ | -0.0798 | 0.923 | 0.07464 | -1.069 | $2.9 \cdot 10^{-1}$ |
| \mathbf{Z}_{Herz} | 0.0784 | 1.082 | 0.0460 | 1.704 | $8.8 \cdot 10^{-2}$ |
| \mathbf{Z}_{Lungen} | 0.3391 | 1.404 | 0.1393 | 2.434 | $1.5 \cdot 10^{-2}$ |
| \mathbf{Z}_{Nieren} | 0.4271 | 1.533 | 0.7152 | 0.597 | $5.5 \cdot 10^{-1}$ |
| \mathbf{Z}_{Geburt} | -0.9219 | 0.398 | 0.8846 | -1.042 | $3.0 \cdot 10^{-1}$ |

Tabelle 6.8: Geschätzte Koeffizienten der Diagnosen im Modell (6.2)

Der Wert des AIC dieses Modells liegt mit einem partiellen Log-Likelihood von -14270.1 und 21 Freiheitsgraden bei 28582.2. Die Teststatistik des lokalen Likelihood-Ratio Test auf die Signifikanz aller Diagnosekovariablen liegt bei $2 \cdot [-14270.1 - (-14467.12)] = 394.04$. Bei 11 Freiheitsgraden ergibt dies einen p-Wert $< 10^{-16}$. Man sieht also, dass im Vergleich zu Modell (6.1) eine große Verbesserung erzielt werden konnte. Vor allem die Kovariable \mathbf{Z}_{Tumore} ist mit einem p-Wert, der kleiner als 10^{-16} ist, hochsignifikant. Ihr Koeffizient ist mit 1.05 auffallend hoch, was bedeutet, dass die Sterbewahrscheinlichkeit von Tumorpatienten in unserem Modell wesentlich höher als bei den übrigen Patienten angenommen wird. Auch die Kovariablen \mathbf{Z}_{Demenz} , \mathbf{Z}_{Schlag} und $\mathbf{Z}_{Psychose}$ scheinen mit einem p-Wert unter 1 % einen großen Einfluß auf die Überlebenszeit der Pflegepatienten zu besitzen. Lediglich die Kovariablen $\mathbf{Z}_{Blindheit}$, $\mathbf{Z}_{Knochenk}$, $\mathbf{Z}_{Arthrosen}$, \mathbf{Z}_{Nieren} und \mathbf{Z}_{Geburt} sind nicht signifikant auf einem Konfidenzniveau von 10 Prozent. Es soll nun überprüft werden, ob durch Entfernen einiger Kovariablen das Modell verbessert werden kann.

Durch iteratives Weglassen der Kovariablen zu den Diagnosen Geburt, Nieren, Blindheit, Arthrosen und Knochenkrankheiten kann das Modell sukzessive verbessert werden. Der AIC-Wert sinkt von 28582.2 auf 28578.4, der lokale Likelihood-Ratio-Test auf die Hypothese, dass die Koeffizienten dieser fünf Kovariablen gleich Null sind, führt zu einem

| Entfernen von | Log-Likelihood | df | AIC |
|--------------------------|----------------|----|----------|
| \mathbf{Z}_{Geburt} | -14271.12 | 20 | 28582.24 |
| \mathbf{Z}_{Nieren} | -14271.28 | 19 | 28580.55 |
| $\mathbf{Z}_{Blindheit}$ | -14271.74 | 18 | 28579.48 |
| $\mathbf{Z}_{Arthrosen}$ | -14272.31 | 17 | 28578.61 |
| $\mathbf{Z}_{Knochenk}$ | -14273.2 | 16 | 28578.4 |
| \mathbf{Z}_{Herz} | -14274.71 | 15 | 28579.42 |

Tabelle 6.9: Entwicklung des Log-Likelihoods und des AIC bei Entfernen wenig signifikanter Diagnosen

p-Wert von 0.29. Diese Kovariablen scheinen daher nicht signifikant zu sein. Entfernt man zusätzlich die Variable \mathbf{Z}_{Herz} , so kann das AIC zwar weiter verringert werden, der p-Wert des Likelihood-Ratio-Tests auf Signifikanz dieser Variable beträgt jedoch 0.08, so dass \mathbf{Z}_{Herz} im Modell bleibt. Im weiteren Verlauf der Arbeit werden wir nun auf diesem Modell (Modell (6.3)) aufbauen, in dem nur noch die Diagnosen Demenz, Tumore, Psychose, Herz, Schlag und Lungen enthalten sind.

$$\begin{aligned}
 \lambda(t) = & \lambda_0(t) \exp[\beta_1 \mathbf{Z}_{Alter}(t) + \beta_2 \mathbf{Z}_{Geschlecht} + \beta_3 \mathbf{Z}_{Stat}(t) + \beta_4 \mathbf{Z}_{Stufe2}(t) + \beta_5 \mathbf{Z}_{Stufe3}(t) \\
 & + \beta_6 \mathbf{Z}_{Alter}(t) \times \mathbf{Z}_{Geschlecht} + \beta_7 \mathbf{Z}_{Alter}(t) \times \mathbf{Z}_{Stat}(t) + \beta_8 \mathbf{Z}_{Geschlecht} \times \mathbf{Z}_{Stat}(t) \\
 & + \beta_9 \mathbf{Z}_{Stat}(t) \times \mathbf{Z}_{Stufe2}(t) + \beta_{10} \mathbf{Z}_{Stat}(t) \times \mathbf{Z}_{Stufe3}(t) + \beta_{11} \mathbf{Z}_{Demenz} + \beta_{12} \mathbf{Z}_{Schlag} \\
 & + \beta_{13} \mathbf{Z}_{Psychose} + \beta_{14} \mathbf{Z}_{Tumore} + \beta_{15} \mathbf{Z}_{Herz} + \beta_{16} \mathbf{Z}_{Lungen}] \quad (6.3)
 \end{aligned}$$

6.3 Modellierung von Interaktionen

Im folgenden soll überprüft werden, ob im Modell auch Interaktionen der Diagnosen mit den übrigen Kovariablen oder Interaktionen zwischen den Diagnosen berücksichtigt werden müssen. Zu Modell (6.3) werden zuerst die Interaktionen der sechs beibehaltenen Diagnosen mit \mathbf{Z}_{Alter} in einem Schritt hinzugenommen. Der Wert des AIC verbessert sich auf 28503.86 (Tabelle 6.10), ein lokaler Likelihood-Ratio-Test für diese Interaktionen liefert einen p-Wert von $1.11 \cdot 10^{-16}$. Aufgrund dieser hohen Signifikanz kann man davon ausgehen, dass mindestens eine der Interaktionen mit Alter für die Modellierung von Bedeutung ist. Betrachtet man die Koeffizienten dieser Interaktionen (Tabelle 6.11), sieht man, dass die Interaktion $\mathbf{Z}_{Alter} \times \mathbf{Z}_{Tumore}$ hochsignifikant ist, bei allen anderen Interaktionen dagegen liegt der p-Wert der z-Statistik über 0.2. Daher wird nun ein Modell gefittet, in dem nur die Tumore-Alter-Interaktion enthalten ist. Das AIC wird weiter verringert (Tabelle 6.10). Um zu testen, ob die weggelassenen Interaktionen wirklich nicht von Bedeutung für das Modell sind, führt man einen lokalen Likelihood-Ratio-Test durch. Dieser liefert für die entfernten Interaktionen einen p-Wert von 0.509, so dass man davon ausgehen kann, dass sie keinen signifikanten Einfluß besitzen. Modell (6.3) wird also nur um die Interaktion $\mathbf{Z}_{Alter} \times \mathbf{Z}_{Tumore}$ erweitert.

Als nächstes wollen wir Interaktionen zwischen der Diagnose Tumore und anderen Diagno-

| | Log-Likelihood | df | AIC |
|-----------------------------------|----------------|----|----------|
| ohne Interaktionen (Modell 6.3) | -14273.20 | 16 | 28578.40 |
| alle Alter-Diagnose-Interaktionen | -14229.93 | 22 | 28503.86 |
| nur Alter-Tumore-Interaktion | -14232.08 | 17 | 28498.16 |

Tabelle 6.10: Entwicklung des Log-Likelihoods und des AIC bei Interaktionen der Diagnosen mit Alter in Modell (6.3)

| Variable | Koeffizient | exp(Koeff.) | SE(Koeff.) | z-Statistik | p-Wert |
|---|-------------|-------------|------------|-------------|----------------------|
| $\mathbf{Z}_{Alter} \times \mathbf{Z}_{Demenz}$ | 0.0066 | 1.007 | 0.0052 | 1.269 | $2.0 \cdot 10^{-1}$ |
| $\mathbf{Z}_{Alter} \times \mathbf{Z}_{Schlag}$ | 0.0069 | 1.007 | 0.0055 | 1.259 | $2.1 \cdot 10^{-1}$ |
| $\mathbf{Z}_{Alter} \times \mathbf{Z}_{Psychose}$ | 0.0009 | 1.001 | 0.0044 | 0.215 | $8.3 \cdot 10^{-1}$ |
| $\mathbf{Z}_{Alter} \times \mathbf{Z}_{Tumore}$ | -0.0337 | 0.967 | 0.0045 | -7.482 | $7.3 \cdot 10^{-14}$ |
| $\mathbf{Z}_{Alter} \times \mathbf{Z}_{Lungen}$ | 0.0067 | 1.007 | 0.0203 | 0.332 | $7.4 \cdot 10^{-1}$ |
| $\mathbf{Z}_{Alter} \times \mathbf{Z}_{Herz}$ | 0.0044 | 1.004 | 0.0051 | 0.862 | $3.9 \cdot 10^{-1}$ |

Tabelle 6.11: Koeffizienten der Alter-Diagnose-Interaktionen im Modell (6.3)

sen untersuchen. Wieder beginnen wir damit, zu unserem Modell alle Tumor-Diagnose-Interaktionen auf einmal hinzuzufügen. Auch hier läßt ein auf 28472.16 gesunkenes AIC (Tabelle 6.12) erkennen, dass noch Verbesserungspotential vorhanden ist. Bestätigt wird das durch einen p-Wert von $9.49 \cdot 10^{-7}$, den der lokale LR-Test für die neu hinzugenommenen Interaktionen liefert.

| | Log-Likelihood | df | AIC |
|---|----------------|----|----------|
| ohne Tumore-Diagnose-Interaktionen | -14232.08 | 17 | 28498.16 |
| alle Tumore-Diagnose-Interaktionen | -14214.08 | 22 | 28472.16 |
| ohne $\mathbf{Z}_{Tumore} \times \mathbf{Z}_{Schlag}$ | -14214.74 | 21 | 28471.48 |
| ohne $\mathbf{Z}_{Tumore} \times \mathbf{Z}_{Lungen}$ | -14215.74 | 20 | 28471.48 |
| ohne $\mathbf{Z}_{Tumore} \times \mathbf{Z}_{Demenz}$ | -14218.07 | 19 | 28474.14 |

Tabelle 6.12: Entwicklung des Log-Likelihoods und des AIC bei Interaktionen der Diagnosen mit Tumore im um $\mathbf{Z}_{Alter} \times \mathbf{Z}_{Tumore}$ erweiterten Modell (6.3)

Tabelle 6.13 zeigt, dass der p-Wert der z-Statistik der Interaktionen $\mathbf{Z}_{Tumore} \times \mathbf{Z}_{Schlag}$ und $\mathbf{Z}_{Tumore} \times \mathbf{Z}_{Lungen}$ über 15 % liegt. Entfernt man diese beiden Interaktionen aus dem Modell, kann eine Verbesserung des AIC erzielt werden (siehe Tabelle 6.12).

Ein lokaler LR-Test für diese beiden Interaktionen ergibt einen p-Wert von 0.19 (Tabelle 6.14), sie können also unberücksichtigt bleiben.

Den nächst grössten p-Wert besitzt die Interaktion $\mathbf{Z}_{Tumore} \times \mathbf{Z}_{Demenz}$. Da sich das AIC bei Weglassen dieser Interaktion jedoch verschlechtert und ein lokaler LR-Test zu einem p-Wert von 0.03 führt, wird sie im Modell behalten. Unser Modell wird also um die Interaktionen der Diagnose Tumore mit den Diagnosen Psychose, Herz- und Demenzerkrankungen erweitert.

| Variable | Koeffizient | exp(Koeff.) | SE(Koeff.) | z-Statistik | p-Wert |
|--|-------------|-------------|------------|-------------|---------------------|
| $\mathbf{Z}_{Tumore} \times \mathbf{Z}_{Demenz}$ | 0.2971 | 0.743 | 0.1328 | -2.2381 | $2.5 \cdot 10^{-2}$ |
| $\mathbf{Z}_{Tumore} \times \mathbf{Z}_{Schlag}$ | -0.1741 | 0.840 | 0.1539 | -1.1309 | $2.6 \cdot 10^{-1}$ |
| $\mathbf{Z}_{Tumore} \times \mathbf{Z}_{Psychose}$ | -0.4422 | 0.643 | 0.1485 | -2.9783 | $2.9 \cdot 10^{-3}$ |
| $\mathbf{Z}_{Tumore} \times \mathbf{Z}_{Lungen}$ | -0.5523 | 0.576 | 0.4099 | -1.3476 | $1.8 \cdot 10^{-1}$ |
| $\mathbf{Z}_{Tumore} \times \mathbf{Z}_{Herz}$ | -0.5585 | 0.572 | 0.1254 | -4.4524 | $8.5 \cdot 10^{-6}$ |

Tabelle 6.13: Koeffizienten der Tumore-Diagnose-Interaktionen im um $\mathbf{Z}_{Alter} \times \mathbf{Z}_{Tumore}$ erweiterten Modell (6.3)

| | p-Wert des lokalen LR-Tests |
|---|-----------------------------|
| alle Tumore-Diagnose-Interaktionen | $9.49 \cdot 10^{-7}$ |
| ohne $\mathbf{Z}_{Tumore} \times (\mathbf{Z}_{Schlag} + \mathbf{Z}_{Lungen})$ | 0.19 |
| ohne $\mathbf{Z}_{Tumore} \times \mathbf{Z}_{Demenz}$ | 0.03 |

Tabelle 6.14: p-Werte der lokalen LR-Test für verschiedene Interaktionen mit Tumore im um $\mathbf{Z}_{Alter} \times \mathbf{Z}_{Tumore}$ erweiterten Modell (6.3)

Analog gehen wir vor, um signifikante Interaktionen der Diagnosen mit der Kovariable Geschlecht zu finden. Wieder orientieren wir uns am p-Wert der z-Statistik, um einzelne Interaktionen zu selektieren und verwenden dann das AIC und den p-Wert von lokalen Likelihood-Ratio-Tests als Gütekriterien für das Modell. Die Ergebnisse sind in den Tabellen 6.15-6.17 zusammengetragen. Man erkennt, dass zwischen der Kovariable $\mathbf{Z}_{Geschlecht}$ und den Diagnosen Tumore und Psychose signifikante Interaktionen bestehen, die nun mit in das Modell aufgenommen werden.

| | Log-Likelihood | df | AIC |
|--|----------------|----|----------|
| alle Geschlecht-Diagnose-Interaktionen | -14203.85 | 26 | 28459.7 |
| nur $\mathbf{Z}_{Geschlecht} \times (\mathbf{Z}_{Tumore} + \mathbf{Z}_{Psychose})$ | -14205.58 | 22 | 28455.16 |

Tabelle 6.15: Entwicklung des Log-Likelihoods und des AIC bei Interaktionen der Diagnosen mit Geschlecht im Modell (6.3), erweitert um $\mathbf{Z}_{Alter} \times \mathbf{Z}_{Tumore}$, $\mathbf{Z}_{Tumore} \times (\mathbf{Z}_{Psychose} + \mathbf{Z}_{Herz} + \mathbf{Z}_{Demenz})$

Die Untersuchung der Interaktionen der Diagnosen mit der Art der Behandlung ergab, dass hier die Interaktionen $\mathbf{Z}_{stat} \times (\mathbf{Z}_{Schlag} + \mathbf{Z}_{Lungen} + \mathbf{Z}_{Psychose})$ einen signifikanten Beitrag zur Modellierung der Überlebenszeit der Pflegepatienten ausmachen und neu ins Modell hinzukommen (siehe Tabellen 6.18 - 6.20). Alle drei Interaktionen besitzen positive Koeffizienten (Tabelle 6.19). Patienten in stationärer Behandlung mit den Diagnosen Schlaganfall, Lungenerkrankung oder Psychose haben daher nach unserem momentanen Modell eine höhere Sterbewahrscheinlichkeit als Patienten in ambulanter Behandlung. Nun sind noch Interaktionen der Diagnosen mit der Pflegestufe zu betrachten. Die Tabellen 6.21 - 6.23 enthalten die wichtigsten Ergebnisse der diesbezüglichen Untersuchungen. Bei Hinzunahme aller Interaktionen der Diagnosen mit der Pflegestufe konnte das AIC

| Variable | Koeffizient | exp(Koeff.) | SE(Koeff.) | z-Statistik | p-Wert |
|--|-------------|-------------|------------|-------------|---------------------|
| $\mathbf{Z}_{Geschlecht} \times \mathbf{Z}_{Tumore}$ | 0.3518 | 1.422 | 0.1225 | 2.8715 | $4.1 \cdot 10^{-3}$ |
| $\mathbf{Z}_{Geschlecht} \times \mathbf{Z}_{Schlag}$ | -0.0261 | 0.974 | 0.1129 | -0.2315 | $8.2 \cdot 10^{-1}$ |
| $\mathbf{Z}_{Geschlecht} \times \mathbf{Z}_{Demenz}$ | -0.1192 | 0.888 | 0.0989 | -1.2052 | $2.3 \cdot 10^{-1}$ |
| $\mathbf{Z}_{Geschlecht} \times \mathbf{Z}_{Herz}$ | -0.1370 | 0.872 | 0.0962 | -1.4245 | $1.5 \cdot 10^{-1}$ |
| $\mathbf{Z}_{Geschlecht} \times \mathbf{Z}_{Lungen}$ | 0.1172 | 1.124 | 0.2781 | 0.4215 | $6.7 \cdot 10^{-1}$ |
| $\mathbf{Z}_{Geschlecht} \times \mathbf{Z}_{Psychose}$ | -0.2459 | 0.782 | 0.1099 | -2.2379 | $2.5 \cdot 10^{-2}$ |

Tabelle 6.16: Koeffizienten der Geschlecht-Diagnose-Interaktionen im Modell (6.3), erweitert um $\mathbf{Z}_{Alter} \times \mathbf{Z}_{Tumore}$, $\mathbf{Z}_{Tumore} \times (\mathbf{Z}_{Psychose} + \mathbf{Z}_{Herz} + \mathbf{Z}_{Demenz})$

| | p-Wert des lokalen LR-Tests |
|--|-----------------------------|
| alle Geschlecht-Diagnose-Interaktionen | $5.7 \cdot 10^{-4}$ |
| nur $\mathbf{Z}_{Geschlecht} \times (\mathbf{Z}_{Tumore} + \mathbf{Z}_{Psychose})$ | $3.88 \cdot 10^{-5}$ |
| weggelassene Geschlecht-Diagnose-Interaktionen | 0.48 |

Tabelle 6.17: p-Werte des lokalen LR-Tests für verschiedene Geschlecht-Diagnose-Interaktionen im Modell (6.3), erweitert um $\mathbf{Z}_{Alter} \times \mathbf{Z}_{Tumore}$, $\mathbf{Z}_{Tumore} \times (\mathbf{Z}_{Psychose} + \mathbf{Z}_{Herz} + \mathbf{Z}_{Demenz})$

zwar nur um 5 Punkte verringert werden. Da ein lokaler LR-Test für die neu hinzugekommenen Interaktionen jedoch einen p-Wert von 0.0031 (Tabelle 6.23) liefert, kann man davon ausgehen, dass mindestens eine der Interaktionen signifikant ist. Tabelle 6.22 enthält die Koeffizienten der Interaktionen der Pflegestufe mit den Diagnosen Tumore und Herzkrankungen. Alle weiteren Interaktionen mit der Pflegestufe waren nicht signifikant auf einem 10%-Level und sind daher nicht in der Tabelle aufgeführt. Nimmt man nur diese vier signifikanten Interaktionen zum Modell, wird ein AIC von 28429.22 erreicht (Tabelle 6.21) und der p-Wert des lokalen LR-Tests sinkt auf einen hochsignifikanten Wert von $5.83 \cdot 10^{-5}$ (Tabelle 6.23). Für die übrigen Interaktionen dagegen erhält man nur einen p-Wert von 0.75, sie können also ohne Bedenken entfernt werden. Entfernt man nun noch die Interaktionen der Pflegestufe mit Tumoren, so bleibt der Wert des AIC nahezu unverändert (Tabelle 6.21). Der p-Wert des lokalen LR-Tests liegt mit 0.12 nur knapp über 10%, daher wird diese Interaktion vorerst noch im Modell behalten. Betrachtet man Tabelle

| | Log-Likelihood | df | AIC |
|--|----------------|----|----------|
| alle Art-Diagnose-Interaktionen | -14197.04 | 28 | 28450.08 |
| nur $\mathbf{Z}_{stat} \times (\mathbf{Z}_{Schlag} + \mathbf{Z}_{Lungen} + \mathbf{Z}_{Psychose})$ | -14197.95 | 25 | 28445.9 |
| ohne $\mathbf{Z}_{stat} \times \mathbf{Z}_{Lungen}$ | -14199.59 | 24 | 28447.18 |

Tabelle 6.18: Entwicklung des Log-Likelihoods und des AIC bei Interaktionen der Diagnosen mit Art der Behandlung im Modell (6.3), erweitert um $\mathbf{Z}_{Alter} \times \mathbf{Z}_{Tumore}$, $\mathbf{Z}_{Tumore} \times (\mathbf{Z}_{Psychose} + \mathbf{Z}_{Herz} + \mathbf{Z}_{Demenz})$, $\mathbf{Z}_{Geschlecht} \times (\mathbf{Z}_{Tumore} + \mathbf{Z}_{Psychose})$

| Variable | Koeffizient | exp(Koeff.) | SE(Koeff.) | z-Statistik | p-Wert |
|--|-------------|-------------|------------|-------------|---------------------|
| $\mathbf{Z}_{stat} \times \mathbf{Z}_{Tumore}$ | -0.0018 | 0.998 | 0.1430 | -0.0126 | $9.9 \cdot 10^{-1}$ |
| $\mathbf{Z}_{stat} \times \mathbf{Z}_{Psychose}$ | 0.3106 | 1.364 | 0.1167 | 2.6620 | $7.8 \cdot 10^{-3}$ |
| $\mathbf{Z}_{stat} \times \mathbf{Z}_{Herz}$ | -0.0747 | 0.928 | 0.0984 | -0.7589 | $4.5 \cdot 10^{-1}$ |
| $\mathbf{Z}_{stat} \times \mathbf{Z}_{Demenz}$ | 0.1068 | 1.113 | 0.1008 | 1.0595 | $2.9 \cdot 10^{-1}$ |
| $\mathbf{Z}_{stat} \times \mathbf{Z}_{Schlag}$ | 0.3038 | 1.355 | 0.1198 | 2.5351 | $1.1 \cdot 10^{-2}$ |
| $\mathbf{Z}_{stat} \times \mathbf{Z}_{Lungen}$ | 0.5857 | 1.796 | 0.3034 | 1.9305 | $5.4 \cdot 10^{-2}$ |

Tabelle 6.19: Koeffizienten der Art-Diagnose-Interaktionen im Modell (6.3), erweitert um $\mathbf{Z}_{Alter} \times \mathbf{Z}_{Tumore}$, $\mathbf{Z}_{Tumore} \times (\mathbf{Z}_{Psychose} + \mathbf{Z}_{Herz} + \mathbf{Z}_{Demenz})$, $\mathbf{Z}_{Geschlecht} \times (\mathbf{Z}_{Tumore} + \mathbf{Z}_{Psychose})$

| | p-Wert des lokalen LR-Tests |
|--|-----------------------------|
| alle Art-Diagnose-Interaktionen | $8.9 \cdot 10^{-3}$ |
| $\mathbf{Z}_{stat} \times (\mathbf{Z}_{Schlag} + \mathbf{Z}_{Lungen} + \mathbf{Z}_{Psychose})$ | $1.6 \cdot 10^{-3}$ |
| $\mathbf{Z}_{stat} \times (\mathbf{Z}_{Tumore} + \mathbf{Z}_{Herz} + \mathbf{Z}_{Demenz})$ | 0.61 |

Tabelle 6.20: p-Werte des lokalen LR-Tests für verschiedene Art-Diagnose-Interaktionen im Modell (6.3), erweitert um $\mathbf{Z}_{Alter} \times \mathbf{Z}_{Tumore}$, $\mathbf{Z}_{Tumore} \times (\mathbf{Z}_{Psychose} + \mathbf{Z}_{Herz} + \mathbf{Z}_{Demenz})$, $\mathbf{Z}_{Geschlecht} \times (\mathbf{Z}_{Tumore} + \mathbf{Z}_{Psychose})$

6.22 genauer, so fällt auf, dass vor allem die Interaktion $\mathbf{Z}_{Stufe3} \times \mathbf{Z}_{Herz}$ sehr signifikant ist, während, $\mathbf{Z}_{Stufe2} \times \mathbf{Z}_{Herz}$ mit einem p-Wert von 0.97 und einem Koeffizienten sehr nahe bei 0 kaum von Bedeutung zu sein scheint. Da es sich bei der Kovariable Pflegestufe jedoch um einen Faktor handelt, müssen beide Interaktionen im Modell enthalten bleiben.

| | Log-Likelihood | df | AIC |
|--|----------------|----|----------|
| alle Stufe-Diagnose-Interaktionen | -14183.08 | 37 | 28440.16 |
| nur $(\mathbf{Z}_{Stufe2} + \mathbf{Z}_{Stufe3}) \times (\mathbf{Z}_{Tumore} + \mathbf{Z}_{Herz})$ | -14185.61 | 29 | 28429.22 |
| nur $(\mathbf{Z}_{Stufe2} + \mathbf{Z}_{Stufe3}) \times \mathbf{Z}_{Herz}$ | -14187.75 | 27 | 28429.5 |

Tabelle 6.21: Entwicklung des Log-Likelihoods und des AIC bei Interaktionen der Diagnosen mit der Pflegestufe im Modell (6.3), erweitert um $\mathbf{Z}_{Alter} \times \mathbf{Z}_{Tumore}$, $\mathbf{Z}_{Tumore} \times (\mathbf{Z}_{Psychose} + \mathbf{Z}_{Herz} + \mathbf{Z}_{Demenz})$, $\mathbf{Z}_{Geschlecht} \times (\mathbf{Z}_{Tumore} + \mathbf{Z}_{Psychose})$, $\mathbf{Z}_{stat} \times (\mathbf{Z}_{Schlag} + \mathbf{Z}_{Lungen} + \mathbf{Z}_{Psychose})$

Bis jetzt wurden nur 2-Faktor-Interaktionen untersucht. Mögliche 3-Faktor-Interaktionen sind in Tabelle 6.24 zusammengestellt. Lediglich bei Hinzunahme der Interaktionen von \mathbf{Z}_{Alter} und \mathbf{Z}_{Tumore} mit \mathbf{Z}_{Herz} , $\mathbf{Z}_{Psychose}$, \mathbf{Z}_{Demenz} konnte eine Verringerung des AIC erzielt werden. Von diesen drei Interaktionen wiederum besitzt nur die Interaktion $\mathbf{Z}_{Alter} \times \mathbf{Z}_{Tumore} \times \mathbf{Z}_{Herz}$ mit $6.5 \cdot 10^{-3}$ einen p-Wert der z-Statistik unter 5 %. Die p-Werte der beiden anderen Interaktionen liegen über 20 %. Fittet man nun ein Modell, indem nur die Interaktion $\mathbf{Z}_{Alter} \times \mathbf{Z}_{Tumore} \times \mathbf{Z}_{Herz}$ zusätzlich berücksichtigt wird, so sinkt das AIC weiter

| Variable | Koeffizient | exp(Koeff.) | SE(Koeff.) | z-Statistik | p-Wert |
|--|-----------------------|-------------|------------|-------------|---------------------|
| $\mathbf{Z}_{Stufe2} \times \mathbf{Z}_{Tumore}$ | $3.11 \cdot 10^{-1}$ | 1.364 | 0.1763 | 1.7631 | $7.8 \cdot 10^{-2}$ |
| $\mathbf{Z}_{Stufe3} \times \mathbf{Z}_{Tumore}$ | $2.18 \cdot 10^{-1}$ | 1.244 | 0.1724 | 1.2665 | $2.1 \cdot 10^{-1}$ |
| $\mathbf{Z}_{Stufe2} \times \mathbf{Z}_{Herz}$ | $-4.57 \cdot 10^{-3}$ | 0.995 | 0.1372 | -0.0333 | $9.7 \cdot 10^{-1}$ |
| $\mathbf{Z}_{Stufe3} \times \mathbf{Z}_{Herz}$ | $-4.15 \cdot 10^{-1}$ | 0.660 | 0.1336 | -3.1101 | $1.9 \cdot 10^{-3}$ |

Tabelle 6.22: Koeffizienten der signifikanten Stufe-Diagnose-Interaktionen im Modell (6.3), erweitert um $\mathbf{Z}_{Alter} \times \mathbf{Z}_{Tumore}$, $\mathbf{Z}_{Tumore} \times (\mathbf{Z}_{Psychose} + \mathbf{Z}_{Herz} + \mathbf{Z}_{Demenz})$, $\mathbf{Z}_{Geschlecht} \times (\mathbf{Z}_{Tumore} + \mathbf{Z}_{Psychose})$, $\mathbf{Z}_{stat} \times (\mathbf{Z}_{Schlag} + \mathbf{Z}_{Lungen} + \mathbf{Z}_{Psychose})$

| | p-Wert des lokalen LR-Tests |
|---|-----------------------------|
| alle Stufe-Diagnose-Interaktionen | $3.1 \cdot 10^{-3}$ |
| alle außer $(\mathbf{Z}_{Stufe2} + \mathbf{Z}_{Stufe3}) \times (\mathbf{Z}_{Tumore} + \mathbf{Z}_{Herz})$ | 0.75 |
| $(\mathbf{Z}_{Stufe2} + \mathbf{Z}_{Stufe3}) \times (\mathbf{Z}_{Tumore} + \mathbf{Z}_{Herz})$ | $5.83 \cdot 10^{-5}$ |
| $(\mathbf{Z}_{Stufe2} + \mathbf{Z}_{Stufe3}) \times \mathbf{Z}_{Tumore}$ | 0.12 |

Tabelle 6.23: p-Werte des lokalen LR-Tests für verschiedene Stufe-Diagnose-Interaktionen im Modell (6.3), erweitert um $\mathbf{Z}_{Alter} \times \mathbf{Z}_{Tumore}$, $\mathbf{Z}_{Tumore} \times (\mathbf{Z}_{Psychose} + \mathbf{Z}_{Herz} + \mathbf{Z}_{Demenz})$, $\mathbf{Z}_{Geschlecht} \times (\mathbf{Z}_{Tumore} + \mathbf{Z}_{Psychose})$, $\mathbf{Z}_{stat} \times (\mathbf{Z}_{Schlag} + \mathbf{Z}_{Lungen} + \mathbf{Z}_{Psychose})$

auf 28424.34, ein lokaler LR-Test ergibt, dass die neue Interaktion mit einem p-Wert von 0.01 signifikant für das Modell ist. Alle übrigen 3-Faktor-Interaktionen tragen aufgrund der Verschlechterung des AIC und der hohen p-Werte nicht zum Modell bei und können daher vernachlässigt werden.

| Interaktionen | Log-Likelihood | df | AIC | p-Wert |
|--|----------------|----|----------|--------|
| ohne 3-Faktor-Interaktionen | -14185.61 | 29 | 28429.22 | |
| $\mathbf{Z}_{Alter} \times \mathbf{Z}_{Tumore} \times \mathbf{Z}_{Herz} +$ $\mathbf{Z}_{Alter} \times \mathbf{Z}_{Tumore} \times \mathbf{Z}_{Psychose} +$ $\mathbf{Z}_{Alter} \times \mathbf{Z}_{Tumore} \times \mathbf{Z}_{Demenz}$ | -14179.16 | 35 | 28428.32 | 0.04 |
| $\mathbf{Z}_{Geschlecht} \times \mathbf{Z}_{Tumore} \times \mathbf{Z}_{Psychose} +$ $\mathbf{Z}_{Geschlecht} \times \mathbf{Z}_{Tumore} \times \mathbf{Z}_{Herz} +$ $\mathbf{Z}_{Geschlecht} \times \mathbf{Z}_{Tumore} \times \mathbf{Z}_{Demenz}$ | -14182.72 | 34 | 28433.44 | 0.33 |
| $\mathbf{Z}_{Alter} \times \mathbf{Z}_{Geschlecht} \times \mathbf{Z}_{Tumore} +$ $\mathbf{Z}_{Alter} \times \mathbf{Z}_{Geschlecht} \times \mathbf{Z}_{Psychose}$ | -14184.51 | 32 | 28433.02 | 0.52 |
| $\mathbf{Z}_{Alter} \times \mathbf{Z}_{stat} \times \mathbf{Z}_{Tumore} +$ $\mathbf{Z}_{Alter} \times \mathbf{Z}_{stat} \times \mathbf{Z}_{Psychose} +$ $\mathbf{Z}_{Alter} \times \mathbf{Z}_{stat} \times \mathbf{Z}_{Schlag} +$ $\mathbf{Z}_{Alter} \times \mathbf{Z}_{stat} \times \mathbf{Z}_{Lungen}$ | -14183.52 | 37 | 28441.04 | 0.84 |
| $\mathbf{Z}_{Geschlecht} \times \mathbf{Z}_{stat} \times \mathbf{Z}_{Psychose}$ | -14185.39 | 30 | 28430.78 | 0.50 |

Tabelle 6.24: Entwicklung des Log-Likelihoods und des AIC bei Hinzunahme von 3-Faktor-Interaktionen im Modell (6.3), erweitert um $\mathbf{Z}_{Alter} \times \mathbf{Z}_{Tumore}$, $\mathbf{Z}_{Tumore} \times (\mathbf{Z}_{Psychose} + \mathbf{Z}_{Herz} + \mathbf{Z}_{Demenz})$, $\mathbf{Z}_{Geschlecht} \times (\mathbf{Z}_{Tumore} + \mathbf{Z}_{Psychose})$, $\mathbf{Z}_{stat} \times (\mathbf{Z}_{Schlag} + \mathbf{Z}_{Lungen} + \mathbf{Z}_{Psychose})$

Rufen wir uns noch einmal das momentane Modell in Erinnerung. Zum Modell 6.3 wurden folgende Interaktionen hinzugefügt:

$$\begin{aligned}
 & \mathbf{Z}_{Alter} \times \mathbf{Z}_{Tumore} \\
 & \mathbf{Z}_{Geschlecht} \times (\mathbf{Z}_{Tumore} + \mathbf{Z}_{Psychose}) \\
 & \mathbf{Z}_{stat} \times (\mathbf{Z}_{Psychose} + \mathbf{Z}_{Schlag} + \mathbf{Z}_{Lungen}) \\
 & \mathbf{Z}_{Tumore} \times (\mathbf{Z}_{Herz} + \mathbf{Z}_{Psychose} + \mathbf{Z}_{Demenz}) \\
 & (\mathbf{Z}_{Stufe2} + \mathbf{Z}_{Stufe3}) \times (\mathbf{Z}_{Herz} + \mathbf{Z}_{Tumore}) \\
 & \mathbf{Z}_{Alter} \times \mathbf{Z}_{Herz} \\
 & \mathbf{Z}_{Alter} \times \mathbf{Z}_{Tumore} \times \mathbf{Z}_{Herz}
 \end{aligned}$$

Dieses Modell enthält 31 Freiheitsgrade, das AIC beträgt 28324.34. Als Interaktionen zwischen den Diagnosen wurden bisher nur Interaktionen zwischen \mathbf{Z}_{Tumore} und den Kovariablen der restlichen Diagnosen untersucht. Dies war naheliegend, da die Kovariable \mathbf{Z}_{Tumore} im Modell (6.3) mit Abstand am signifikantesten war. Nun sollen auch noch Interaktionen zwischen den anderen Diagnosen untersucht werden. Erweitert man das momentane Modell um alle 2-Faktor-Interaktionen zwischen den Diagnosekovariablen \mathbf{Z}_{Herz} , \mathbf{Z}_{Schlag} , \mathbf{Z}_{Lungen} , \mathbf{Z}_{Demenz} und $\mathbf{Z}_{Psychose}$, deutet der p-Wert eines lokalen LR-Tests für alle neu hinzugenommenen Interaktionen von 0.02332 auf Signifikanz mindestens einer Interaktion hin. Bei Betrachtung der p-Werte der einzelnen Diagnosen, die auf der z-Statistik beruhen, erkennt man, dass die Interaktion $\mathbf{Z}_{Psychose} \times \mathbf{Z}_{Schlag}$ mit einem p-Wert von $3.7 \cdot 10^{-5}$ dem Modell hinzugefügt werden sollte. Der p-Wert aller anderen Interaktionen dagegen liegt über 10 %. Die Interaktion $\mathbf{Z}_{Psychose} \times \mathbf{Z}_{Schlag}$ wird also mit ins Modell aufgenommen.

Zu Beginn der Untersuchungen wurde die Zahl der Diagnosekovariablen von 11 auf 6 reduziert. Die 5 entfernten Diagnosen wurden aufgrund eines jeweiligen p-Wertes über 15 % selektiert. Möglicherweise ist man hier etwas zu voreilig gewesen. Es kann durchaus vorkommen, dass eine Kovariable erst durch Hinzunahme einer Interaktion signifikant für das Modell wird. Daher wollen wir an dieser Stelle noch einmal diese 5 Diagnosen mit ins Modell nehmen, diesmal mit allen Interaktionen mit \mathbf{Z}_{Alter} , $\mathbf{Z}_{Geschlecht}$, \mathbf{Z}_{stat} , \mathbf{Z}_{Stufe2} , \mathbf{Z}_{Stufe3} , \mathbf{Z}_{Tumore} und $\mathbf{Z}_{Psychose}$. Tatsächlich sind fünf dieser Interaktionen signifikant bei einem Niveau von 10 %:

$$\begin{aligned} & \mathbf{Z}_{Arthrosen} \times \mathbf{Z}_{Alter} \text{ (p-Wert: } 2.4 \cdot 10^{-2} \text{),} \\ & \mathbf{Z}_{Knochenk} \times \mathbf{Z}_{Tumore} \text{ (p-Wert: } 9.5 \cdot 10^{-2} \text{),} \\ & \mathbf{Z}_{Stufe3} \times \mathbf{Z}_{Knochenk} \text{ (p-Wert: } 9.6 \cdot 10^{-2} \text{),} \\ & \mathbf{Z}_{Psychose} \times \mathbf{Z}_{Geburt} \text{ (p-Wert: } 9.4 \cdot 10^{-2} \text{)} \end{aligned}$$

Da wir jedoch für die Interaktionen einen p-Wert von 5 % zugrunde legen, wird nur $\mathbf{Z}_{Arthrosen} \times \mathbf{Z}_{Alter}$ im Modell behalten.

Betrachtet man nun die p-Werte der einzelnen Kovariablen in diesem Modell so fällt auf, dass die Kovariablen $\mathbf{Z}_{Alter} \times \mathbf{Z}_{Geschlecht}$ und $\mathbf{Z}_{Alter} \times \mathbf{Z}_{stat}$ im Vergleich zu Ausgangsmodell (6.1) stark an Signifikanz verloren haben. Ihre p-Werte liegen nur noch bei 0.97 bzw. 0.7. Daher werden sie in einem nächsten Fit aus dem Modell entfernt. Der partielle Log-Likelihood dieses Modells beträgt nun -14170.24, bei 32 Freiheitsgraden resultiert daraus eine Verbesserung des AIC auf 28404.48. Das Entfernen der beiden oben genannten Interaktionen wird durch einen p-Wert von 0.934 des lokalen LR-Tests bestätigt. In Tabelle 6.25 sind die Koeffizienten aller Kovariablen dieses Modells zusammengestellt. Die Hazardfunktion dieses Modells lautet:

$$\begin{aligned} \lambda(t) = & \lambda_0(t) \exp[\beta_1 \mathbf{Z}_{Alter}(t) + \beta_2 \mathbf{Z}_{Geschlecht} + \beta_3 \mathbf{Z}_{Stat}(t) + \beta_4 \mathbf{Z}_{Stufe2}(t) + \beta_5 \mathbf{Z}_{Stufe3}(t) \\ & + \beta_6 \mathbf{Z}_{Geschlecht} \times \mathbf{Z}_{Stat}(t) + \beta_7 \mathbf{Z}_{Stat}(t) \times \mathbf{Z}_{Stufe2}(t) + \beta_8 \mathbf{Z}_{Stat}(t) \times \mathbf{Z}_{Stufe3}(t) \\ & + \beta_9 \mathbf{Z}_{Demenz} + \beta_{10} \mathbf{Z}_{Schlag} + \beta_{11} \mathbf{Z}_{Psychose} + \beta_{12} \mathbf{Z}_{Tumore} + \beta_{13} \mathbf{Z}_{Herz} + \beta_{14} \mathbf{Z}_{Lungen} \\ & + \beta_{15} \mathbf{Z}_{Arthrosen} + \beta_{16} \mathbf{Z}_{Alter}(t) \times \mathbf{Z}_{Tumore} + \beta_{17} \mathbf{Z}_{Alter}(t) \times \mathbf{Z}_{Arthrosen} \\ & + \beta_{18} \mathbf{Z}_{Geschlecht} \times \mathbf{Z}_{Tumore} + \beta_{19} \mathbf{Z}_{Geschlecht} \times \mathbf{Z}_{Psychose} + \beta_{20} \mathbf{Z}_{stat}(t) \times \mathbf{Z}_{Psychose} \\ & + \beta_{21} \mathbf{Z}_{stat}(t) \times \mathbf{Z}_{Schlag} + \beta_{22} \mathbf{Z}_{stat}(t) \times \mathbf{Z}_{Lungen} + \beta_{23} \mathbf{Z}_{Tumore} \times \mathbf{Z}_{Herz} \\ & + \beta_{24} \mathbf{Z}_{Alter}(t) \times \mathbf{Z}_{Herz} + \beta_{25} \mathbf{Z}_{Stufe2}(t) \times \mathbf{Z}_{Herz} + \beta_{26} \mathbf{Z}_{Stufe3}(t) \times \mathbf{Z}_{Herz} \quad (6.4) \\ & + \beta_{27} \mathbf{Z}_{Tumore} \times \mathbf{Z}_{Stufe2}(t) + \beta_{28} \mathbf{Z}_{Tumore} \times \mathbf{Z}_{Stufe3}(t) + \beta_{29} \mathbf{Z}_{Tumore} \times \mathbf{Z}_{Psychose} \\ & + \beta_{30} \mathbf{Z}_{Tumore} \times \mathbf{Z}_{Demenz} + \beta_{31} \mathbf{Z}_{Psychose} \times \mathbf{Z}_{Schlag} + \beta_{32} \mathbf{Z}_{Tumore} \times \mathbf{Z}_{Alter}(t) \times \mathbf{Z}_{Herz}] \end{aligned}$$

6.4 Interpretation des Modells (6.4)

In Abb. 6.2 ist das Modell (6.4) schematisch dargestellt. Verbindungsgeraden zwischen den einzelnen Kovariablen kennzeichnen die im Modell enthaltenen Interaktionen. Die 3-Faktor-Interaktion zwischen den Diagnosen Herz und Tumore und dem Alter ist durch

| Variable | Koeffizient | exp(Koeff.) | SE(Koeff.) | z-Statistik | p-Wert |
|--|-------------|-------------|------------|-------------|---------------------|
| \mathbf{Z}_{Alter} | 0.0384 | 1.0391 | 0.0031 | 12.200 | $< 10^{-16}$ |
| \mathbf{Z}_{Tumore} | 3.9070 | 49.7497 | 0.3670 | 10.589 | $< 10^{-16}$ |
| $\mathbf{Z}_{Arthrosen}$ | -2.1288 | 0.1190 | 0.9061 | -2.349 | $1.9 \cdot 10^{-2}$ |
| $\mathbf{Z}_{Geschlecht}$ | -0.2617 | 0.7698 | 0.0726 | -3.603 | $3.1 \cdot 10^{-4}$ |
| \mathbf{Z}_{stat} | 0.5954 | 1.8137 | 0.1529 | 3.895 | $9.8 \cdot 10^{-5}$ |
| $\mathbf{Z}_{Psychose}$ | -0.1757 | 0.8389 | 0.0976 | -1.801 | $7.2 \cdot 10^{-2}$ |
| \mathbf{Z}_{Stufe2} | 0.6822 | 1.9782 | 0.1136 | 6.003 | $1.9 \cdot 10^{-9}$ |
| \mathbf{Z}_{Stufe3} | 1.7890 | 5.9836 | 0.1113 | 16.077 | $< 10^{-16}$ |
| \mathbf{Z}_{Schlag} | -0.3485 | 0.7058 | 0.0773 | -4.510 | $6.5 \cdot 10^{-6}$ |
| \mathbf{Z}_{Lungen} | 0.1401 | 1.1504 | 0.1660 | 0.844 | $4.0 \cdot 10^{-1}$ |
| \mathbf{Z}_{Herz} | 1.0692 | 2.9130 | 0.4779 | 2.237 | $2.5 \cdot 10^{-2}$ |
| \mathbf{Z}_{Demenz} | -0.0616 | 0.9403 | 0.0527 | -1.169 | $2.4 \cdot 10^{-1}$ |
| $\mathbf{Z}_{Alter} \times \mathbf{Z}_{Tumore}$ | -0.0379 | 0.9628 | 0.0044 | -8.614 | $< 10^{-16}$ |
| $\mathbf{Z}_{Alter} \times \mathbf{Z}_{Arthrosen}$ | 0.0236 | 1.0239 | 0.0103 | 2.306 | $2.1 \cdot 10^{-2}$ |
| $\mathbf{Z}_{Geschlecht} \times \mathbf{Z}_{stat}$ | -0.3648 | 0.6943 | 0.1019 | -3.581 | $3.4 \cdot 10^{-4}$ |
| $\mathbf{Z}_{Geschlecht} \times \mathbf{Z}_{Tumore}$ | 0.4116 | 1.5093 | 0.1129 | 3.645 | $2.7 \cdot 10^{-4}$ |
| $\mathbf{Z}_{Geschlecht} \times \mathbf{Z}_{Psychose}$ | -0.2348 | 0.7907 | 0.1103 | -2.128 | $3.3 \cdot 10^{-2}$ |
| $\mathbf{Z}_{stat} \times \mathbf{Z}_{Stufe2}$ | -0.2329 | 0.7922 | 0.1529 | -1.524 | $1.3 \cdot 10^{-1}$ |
| $\mathbf{Z}_{stat} \times \mathbf{Z}_{Stufe3}$ | -0.6148 | 0.5408 | 0.1484 | -4.142 | $3.4 \cdot 10^{-5}$ |
| $\mathbf{Z}_{stat} \times \mathbf{Z}_{Psychose}$ | 0.2861 | 1.3313 | 0.1142 | 2.506 | $1.2 \cdot 10^{-2}$ |
| $\mathbf{Z}_{stat} \times \mathbf{Z}_{Schlag}$ | 0.2403 | 1.2717 | 0.1174 | 2.047 | $4.1 \cdot 10^{-2}$ |
| $\mathbf{Z}_{stat} \times \mathbf{Z}_{Lungen}$ | 0.6426 | 1.9013 | 0.3036 | 2.117 | $3.4 \cdot 10^{-2}$ |
| $\mathbf{Z}_{Tumore} \times \mathbf{Z}_{Herz}$ | -3.8934 | 0.0204 | 1.1845 | -3.286 | $1.0 \cdot 10^{-3}$ |
| $\mathbf{Z}_{Alter} \times \mathbf{Z}_{Herz}$ | -0.0089 | 0.9911 | 0.0055 | -1.620 | $1.1 \cdot 10^{-1}$ |
| $\mathbf{Z}_{Stufe2} \times \mathbf{Z}_{Herz}$ | 0.0496 | 1.0508 | 0.1350 | 0.367 | $7.1 \cdot 10^{-1}$ |
| $\mathbf{Z}_{Stufe3} \times \mathbf{Z}_{Herz}$ | -0.3673 | 0.6926 | 0.1317 | -2.788 | $5.3 \cdot 10^{-3}$ |
| $\mathbf{Z}_{Tumore} \times \mathbf{Z}_{Stufe2}$ | 0.3657 | 1.4415 | 0.1697 | 2.155 | $3.1 \cdot 10^{-2}$ |
| $\mathbf{Z}_{Tumore} \times \mathbf{Z}_{Stufe3}$ | 0.3166 | 1.3724 | 0.1656 | 1.911 | $5.6 \cdot 10^{-2}$ |
| $\mathbf{Z}_{Tumore} \times \mathbf{Z}_{Psychose}$ | -0.3499 | 0.7048 | 0.1536 | -2.279 | $2.3 \cdot 10^{-2}$ |
| $\mathbf{Z}_{Tumore} \times \mathbf{Z}_{Demenz}$ | -0.3117 | 0.7322 | 0.1345 | -2.318 | $2.0 \cdot 10^{-2}$ |
| $\mathbf{Z}_{Psychose} \times \mathbf{Z}_{Schlag}$ | 0.5159 | 1.6751 | 0.1274 | 4.051 | $5.1 \cdot 10^{-5}$ |
| $\mathbf{Z}_{Tumore} \times \mathbf{Z}_{Alter} \times \mathbf{Z}_{Herz}$ | 0.0408 | 1.0417 | 0.0143 | 2.863 | $4.2 \cdot 10^{-3}$ |

Tabelle 6.25: Koeffizienten der Kovariablen im Modell (6.4)

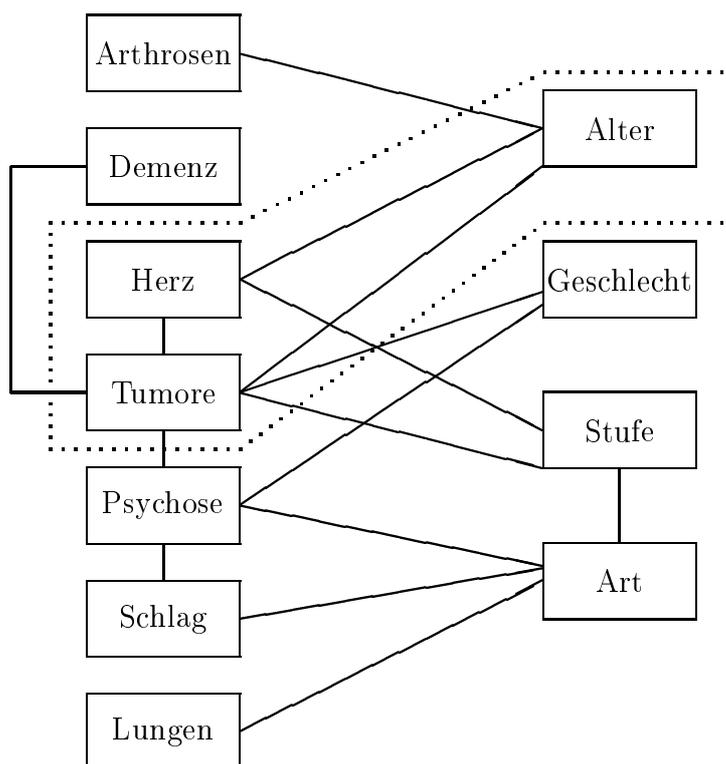


Abbildung 6.2: Kovariablen und Interaktionen in Modell (6.4)

eine Umrandung angedeutet. Wie aus Tabelle 6.25 ersichtlich ist, besitzen fast alle im Modell 6.4 enthaltenen Kovariablen einen p-Wert unter 5%. Ausnahmen bilden die Diagnosen Psychose, Lungen- und Demenzerkrankungen. Da zu diesen Diagnosen jedoch signifikante Interaktionen im Modell enthalten sind, werden sie trotzdem berücksichtigt. Des weiteren erkennt man aus Tabelle 6.25, dass die p-Werte der Interaktionen $\mathbf{Z}_{Stufe2} \times \mathbf{Z}_{stat}$ und $\mathbf{Z}_{Stufe2} \times \mathbf{Z}_{Herz}$ über 10 % liegen. Da jedoch die zugehörigen Interaktionen der Kovariable \mathbf{Z}_{Stufe3} sehr wohl signifikant sind und es sich bei der Pflegestufe um einen Faktor handelt, kann man die Interaktionen der Kovariable \mathbf{Z}_{Stufe2} nicht entfernen. Vergleicht man die Koeffizienten des momentanen Modells mit denen des Ausgangsmodells (Tabellen 6.7 und 6.8), wird deutlich, dass sich die Koeffizienten und p-Werte einiger Kovariablen wesentlich geändert haben. Die Kovariablen $\mathbf{Z}_{Geschlecht}$ und \mathbf{Z}_{stat} haben stark an Signifikanz gewonnen, ihre Koeffizienten sind nun bei etwa gleichem Betrag negativ. Diese beiden Kovariablen scheinen stark von den pflegeauslösenden Diagnosen abzuhängen.

Das Modell ist nun sehr vielschichtig, wegen der zahlreichen Interaktionen ist es nicht möglich zur Interpretation nur einzelne Koeffizienten heranzuziehen, es müssen verschiedene Überlappungen beachtet werden. Dennoch sind eindeutige Trends zu erkennen. Um diese graphisch aufzuzeigen, wurden in den Abb. 6.3- 6.5 die multiplikativen Effekte $\exp[\beta' \mathbf{Z}]$ für verschiedene Diagnosegruppen, getrennt nach Geschlecht, jeweils für die Alterswerte 60, 70 und 80 Jahre und die sechs möglichen Pflegestufen (ambulant Stufe 1, ambulant

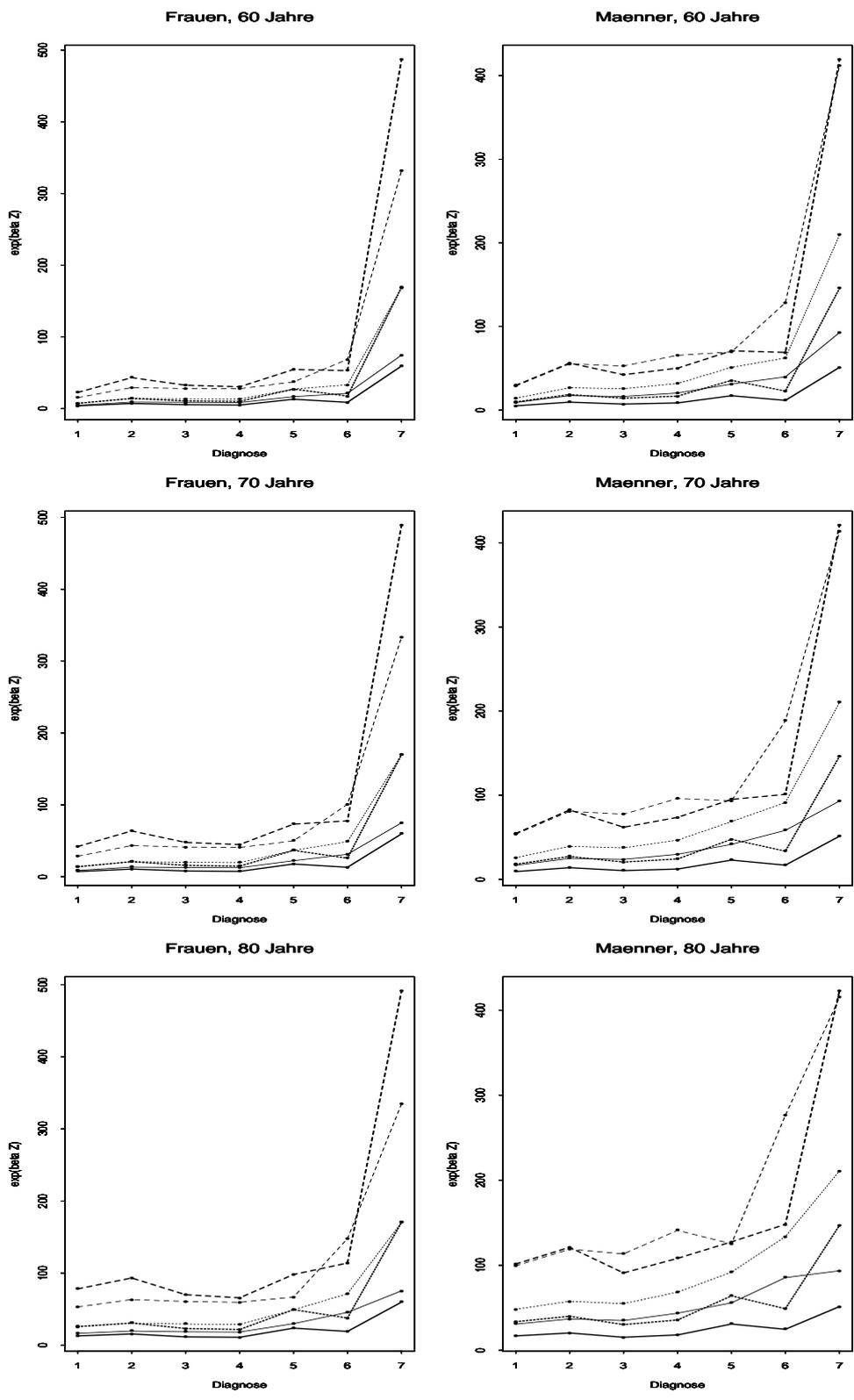


Abbildung 6.3: Multiplikatoren $\exp[\beta'Z]$ für Frauen und Männer mit exklusiven Diagnosen in Modell (6.4) mit 1: Arthrosen 2: Demenz 3: Schlag 4: Psychose 5: Herz 6: Lungen 7: Tumore, — stationär Stufe1, - - - stationär Stufe2, - - stationär Stufe3, — ambulant Stufe1, - - - ambulant Stufe2, - - - ambulant Stufe3

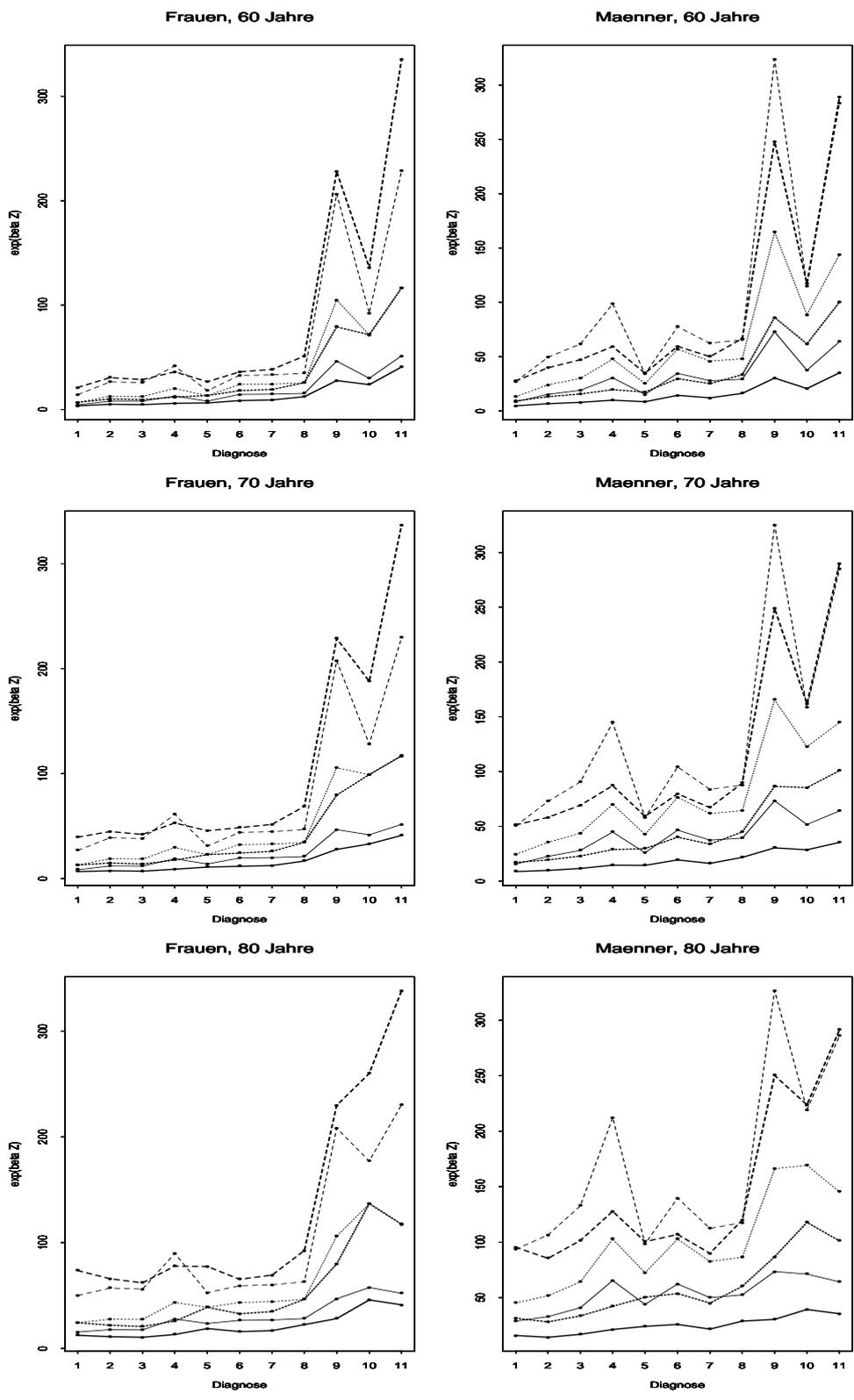


Abbildung 6.4: Multiplikatoren $\exp[\beta'Z]$ für Frauen und Männer mit Doppeldiagnosen in Modell (6.4) mit 1: Demenz und Arthrosen 2: Demenz und Schlag 3: Demenz und Psychose 4: Psychose und Schlag 5: Herz und Arthrosen 6: Psychose und Herz 7: Herz und Schlag 8: Herz und Demenz 9: Tumore und Psychose 10: Tumore und Herz 11: Tumore und Demenz, — stationär Stufe1, - - - stationär Stufe2, - - stationär Stufe3, — ambulant Stufe1, - - - ambulant Stufe2, - - ambulant Stufe3

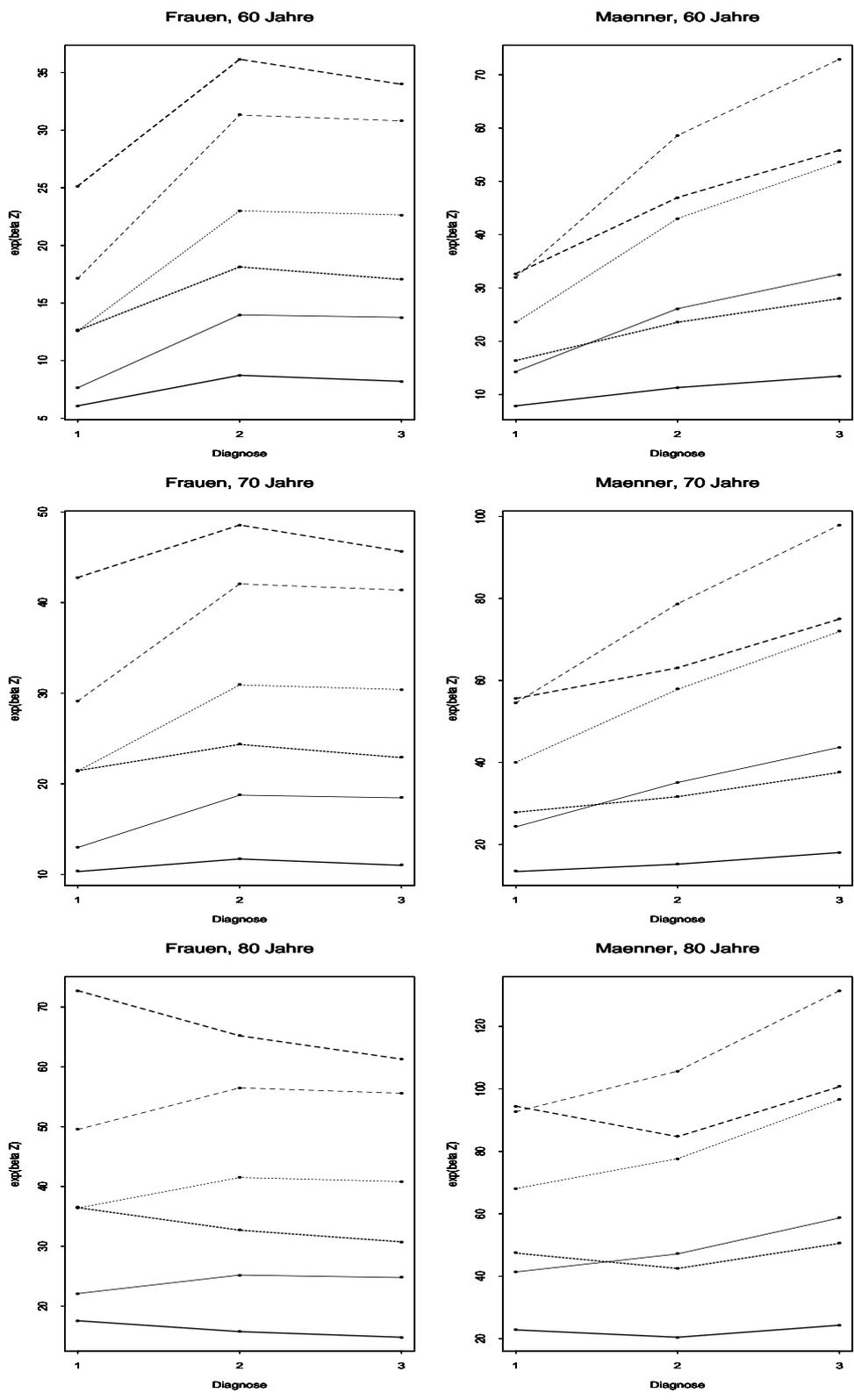


Abbildung 6.5: Multiplikatoren $\exp[\beta'Z]$ für Frauen und Männer mit Dreifachdiagnosen in Modell (6.4) mit 1: Demenz, Herz und Arthrosen 2: Demenz, Herz und Schlag, 3: Demenz, Herz und Psychose, — stationär Stufe1, - - - stationär Stufe2, - - stationär Stufe3, — ambulant Stufe1, - - - ambulant Stufe2, - - ambulant Stufe3

Stufe 2, ambulant Stufe 3, stationär Stufe 1, stationär Stufe 2, stationär Stufe 3) geplotet.

Abb. 6.3 enthält die Multiplikatoren für die sieben verschiedenen Einzeldiagnosen, die jeweils an der x-Achse abgetragen sind. In Abb. 6.4 sind die Plots für die Multiplikatoren für die Gruppen mit Doppel-, in Abb. 6.5 mit Dreifachdiagnosen zu finden.

Aus allen drei Abbildungen kann man erkennen, dass pflegebedürftige Frauen nach dem Modell eine höhere Überlebenszeit als Männer mit demselben Pflegestatus besitzen: die Multiplikatoren sind bei Frauen in fast allen Diagnosegruppen niedriger als bei Männern. Da Frauen allgemein eine längere Lebenserwartung als Männer haben, verwundert dieses Resultat nicht. Bei der Diagnose Tumore haben dagegen Männer ein geringeres Sterberisiko ($\mathbf{Z}_{Geschlecht} \times \mathbf{Z}_{Tumore}$: 0.411, $\mathbf{Z}_{Geschlecht}$: -0.26). Man sieht auch deutlich, dass mit zunehmender Pflegestufe die Sterbewahrscheinlichkeit erheblich steigt. Aus den Regressionskoeffizienten in Tabelle 6.25 erfährt man, dass ein ambulant behandelter Patient in Stufe 2, bei dem weder Tumore noch Herzkrankheiten diagnostiziert wurden, ein beinahe doppelt so hohes Risiko zu sterben hat, wie ein Patient mit den gleichen Merkmalen in Stufe 1. In der Pflegestufe 3 wäre dieses Risiko sogar sechsmal so hoch. Auch kann man prinzipiell sagen, dass stationär Behandelte dem Modell nach eine geringere Überlebenszeit besitzen. Die Plots der Multiplikatoren unterstützen diese Aussage. Die negativen Koeffizienten der Interaktionen $\mathbf{Z}_{stat} \times \mathbf{Z}_{Stufe2}$ bzw. $\mathbf{Z}_{stat} \times \mathbf{Z}_{Stufe3}$ wiederum zeigen, dass die Behandlungsart mit zunehmender Pflegestufe an Bedeutung für das Überleben verliert. In Stufe 3 spielt sie nahezu keine Rolle mehr, die Koeffizienten der Kovariablen \mathbf{Z}_{stat} und $\mathbf{Z}_{stat} \times \mathbf{Z}_{Stufe3}$ heben sich beinahe auf. Für Frauen gilt generell, dass sich die Behandlungsart weniger stark auswirkt ($\mathbf{Z}_{Geschlecht} \times \mathbf{Z}_{stat}$: -0.36), die Linien für stationäre und ambulante Pflege liegen in den einzelnen Pflegestufen näher beieinander als bei den Männern. In Stufe 1 besitzen Frauen, die keine der Diagnosen Psychose, Schlaganfall oder Lungenerkrankung besitzen, in stationärer Behandlung ein um etwa ein Viertel höheres Sterberisiko als Frauen in ambulanter Behandlung ($\exp[0.59 - 0.36] = 1.26$), in Stufe 2 wirkt sich die Behandlungsart nicht auf das Überleben aus ($\exp[0.59 - 0.23 - 0.36] = 1$) und in Stufe 3 ist das Risiko zu sterben in ambulanter Pflege sogar ca. 1.5 mal höher als in stationärer Behandlung ($\exp[0.59 - 0.36 - 0.61] = 0.68$). Ausnahmen bilden die Diagnosen Psychose, Schlaganfälle und Lungenerkrankungen. Hier besitzen stationär Behandelte aufgrund der positiven Koeffizienten der Interaktionen dieser Diagnosen mit der Art der Behandlung, in allen Pflegestufen, unabhängig vom Geschlecht, ein höheres Sterberisiko als in ambulanter Behandlung. Wie zu erwarten, steigt die Sterbewahrscheinlichkeit sowohl für Männer als auch für Frauen mit zunehmendem Alter an. Lediglich bei Tumorpatienten verschwindet der Einfluß des Alters nahezu vollständig, die Beträge der Koeffizienten der Kovariablen \mathbf{Z}_{Alter} und $\mathbf{Z}_{Alter} \times \mathbf{Z}_{Tumore}$ sind bei entgegengesetztem Vorzeichen fast gleich groß. Möglicherweise ist diese Erscheinung auf die geringen Heilungschancen bei Tumorpatienten zurückzuführen. Die Überlebenszeit bei Tumoren ist sehr kurz, wie schon der extrem hohe Koeffizient der Kovariable \mathbf{Z}_{Tumore} und der starke Anstieg der multiplikativen Effekte bei den Diagnosegruppen mit Tumor zeigt. Etwas verwunderlich sind die negativen Koeffizienten der Interaktionen $\mathbf{Z}_{Tumore} \times (\mathbf{Z}_{Herz} + \mathbf{Z}_{Psychose} + \mathbf{Z}_{Demenz})$. Nach dem Modell haben Tumorpatienten, die zusätzlich noch unter Psychose, Herz- oder Demenzerkrankungen leiden, eine höhere Überlebenswahrscheinlichkeit als Patienten, bei denen nur Tumore diagnostiziert wurden. Dieses Phänomen ist womöglich dadurch er-

klärbar, dass Patienten mit mehreren Diagnosen wahrscheinlich besser beobachtet und versorgt werden.

6.5 Bestimmung der funktionalen Form der Kovariable Alter

Nun soll überprüft werden, ob in dem momentanen Modell (6.4) die Proportional Hazard-Annahme erfüllt ist, d.h. es ist zu untersuchen, ob die Effekte aller stetigen Kovariablen, in unserem Fall nur \mathbf{Z}_{Alter} , tatsächlich linear im Log-Hazard sind, wie es das Proportional Hazard Modell unterstellt. Sollte das nicht der Fall sein, gilt es geeignete Transformationen für die Kovariable zu finden. Zuerst wollen wir mittels graphischer Tests Aufschluß darüber gewinnen, in welcher funktionaler Form \mathbf{Z}_{Alter} in das Modell eingeht. Dazu fittet man nach (4.8.2), für Männer und Frauen getrennt, Modell (6.4) ohne die Kovariable \mathbf{Z}_{Alter} und alle damit verbundenen Interaktionen und berechnet jeweils die Martingalresiduen. Diese werden dann gegen das Alter geplottet. Die Form der Graphen entspricht der funktionalen Gestalt von \mathbf{Z}_{Alter} bei Männern bzw. Frauen. In Abbildung 6.6 erkennt man, dass das Alter in beiden Graphen zuerst linear in das Modell eingeht, dann aber ist in beiden Plots ein deutlicher Knick erkennbar. Bei den Frauen ist dieser Knick jedoch ausgeprägter als bei den Männern und findet erst bei etwa 80 Jahren statt, bei Männern dagegen schon bei 60-70 Jahren (siehe Abb. 6.7). Um die genauen Knickstellen bei Männern und Frauen zu lokalisieren, wird das Modell um zwei Kovariablen θ_w und θ_m erweitert, die wie folgt definiert sind:

$$\mathbf{Z}_{\theta_w} = \begin{cases} 1 & \text{falls Patient weiblich und } \mathbf{Z}_{Alter} \geq \theta_w \\ 0 & \text{sonst} \end{cases}$$

$$\mathbf{Z}_{\theta_m} = \begin{cases} 1 & \text{falls Patient männlich und } \mathbf{Z}_{Alter} \geq \theta_m \\ 0 & \text{sonst} \end{cases}$$

Für dieses Modell wird nun der partielle Log-Likelihood in Abhängigkeit von θ_w und θ_m maximiert und dann getestet, ob das Modell dadurch signifikant verbessert werden kann. Dazu wird dieses Modell für verschiedene Werte von $\theta_w = 65 + i$ und $\theta_m = 65 + i$, $i = 0, 1, \dots, 40$ gefittet. Der partielle Log-Likelihood wird mit dem Wert -14166.55 maximal für $\theta_w = 72$ Jahre und $\theta_m = 69$ Jahre (zum Vergleich: ohne θ_w und θ_m : -14170.24). Da zur Gewinnung der optimalen Kovariablen über die Alterswerte θ_m und θ_w maximiert wurde, können diese als eine Art ML-Schätzer für das Alter angesehen werden. Sowohl \mathbf{Z}_{θ_m} als auch \mathbf{Z}_{θ_w} tragen daher 2 Freiheitsgrade zum Modell bei, jeweils einen für das Alter θ_m bzw. θ_w und einen für den Regressionskoeffizienten von \mathbf{Z}_{θ_m} bzw. \mathbf{Z}_{θ_w} . Damit liegen die p-Werte für \mathbf{Z}_{θ_m} und \mathbf{Z}_{θ_w} bei 0.23 bzw. $3.7 \cdot 10^{-2}$. Ein lokaler LR-Test auf die Signifikanz beider Kovariablen mit 4 Freiheitsgraden, ergibt einen p-Wert von 0.117. In Zukunft werden sie mit $\mathbf{Z}_{\theta_{w72}}$ und $\mathbf{Z}_{\theta_{m69}}$ bezeichnet.

Mit -14167.26 ist der partielle Log-Likelihood für $\theta_w = 91$ Jahre und $\theta_m = 88$ Jahre nur wenig niedriger. Diese Variablen werden analog mit $\mathbf{Z}_{\theta_{w91}}$ und $\mathbf{Z}_{\theta_{m88}}$ notiert. Ein LR-Test für diese beiden Alterswerte ergibt einen p-Wert von 0.202, die individuellen p-Werte liegen bei 0.15 für $\mathbf{Z}_{\theta_{m88}}$ und 0.18 für $\mathbf{Z}_{\theta_{w91}}$. Obwohl nur die Variable $\mathbf{Z}_{\theta_{w72}}$ auf einem

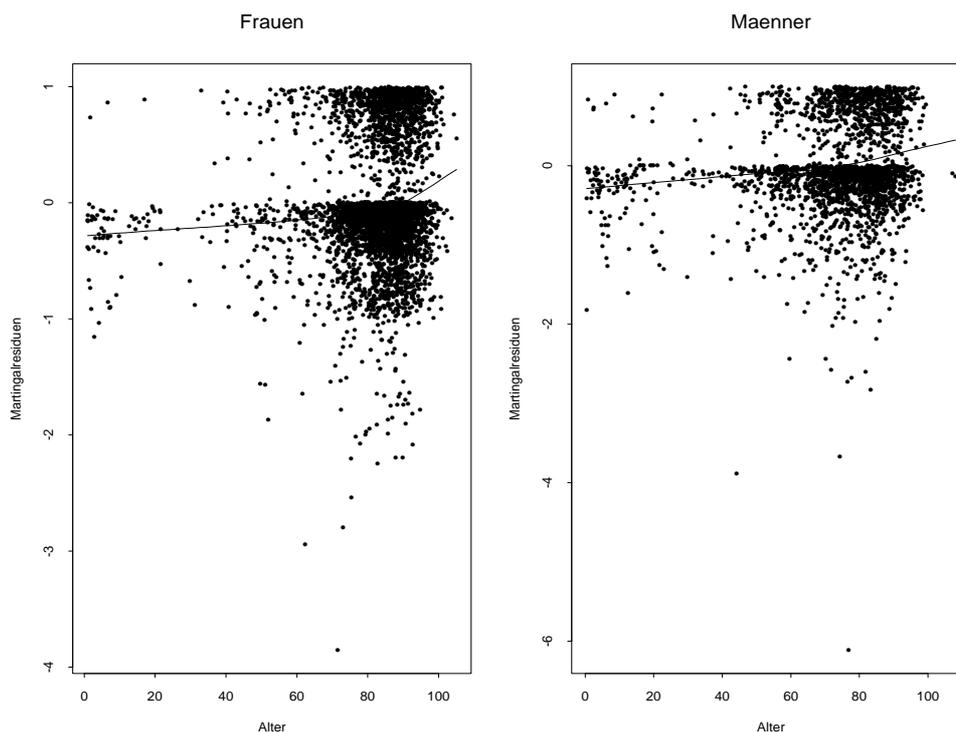


Abbildung 6.6: Martingalresiduen zur Bestimmung der funktionalen Form des Alters in Modell (6.4)

Niveau von 10 % signifikant ist, sollen im folgenden alle vier Indikatorvariablen noch näher untersucht werden. Eventuell gewinnen sie an Signifikanz, wenn Interaktionen mit den übrigen Kovariablen hinzukommen.

6.5.1 Untersuchung der Modelle mit \mathbf{Z}_{θ_m} und \mathbf{Z}_{θ_w}

Zunächst werden Interaktionen für die Kovariablen für das Alter 72 Jahre bei Frauen und 69 Jahre bei Männern untersucht. Als Ausgangsmodell wird Modell (6.4) mit \mathbf{Z}_{θ_w72} und \mathbf{Z}_{θ_m69} als zusätzlichen Kovariablen verwendet. Zu diesem Modell werden nun Interaktionen von \mathbf{Z}_{θ_w72} und \mathbf{Z}_{θ_m69} mit den übrigen Kovariablen betrachtet. Die Interaktionen mit \mathbf{Z}_{Alter} führen zu keiner Verbesserung des Modells, ein lokaler LR-Test für diese Interaktionen ergibt nur einen p-Wert von 0.95. Anders bei den Interaktionen mit den sieben im Modell enthaltenen Diagnosen. Der p-Wert des LR-Tests für alle diese Interaktionen von 0.105, deutet an, dass hier signifikante Interaktionen mit einigen Diagnosen vorliegen. Genauere Untersuchungen ergeben, dass es sich hierbei um die Interaktionen $\mathbf{Z}_{\theta_w72} \times \mathbf{Z}_{Lungen}$ und $\mathbf{Z}_{\theta_m69} \times (\mathbf{Z}_{Tumore} + \mathbf{Z}_{Lungen})$ handelt, die alle signifikant auf einem 5 %-Level sind. Der partielle Log-Likelihood konnte auf -14161.89 erhöht werden. Bei nunmehr 39 Freiheitsgraden beträgt der Wert des AIC 28401.84, er konnte also im Vergleich zu Modell (6.4) leicht verbessert werden (Tabelle 6.26). Weiter wurden Interaktionen mit der Art der Behandlung und der Pflegestufe betrachtet, von denen jedoch keine signifikant auf einem Niveau von 5 % war, so dass wir uns auf die drei gefundenen Interaktionen mit den Dia-

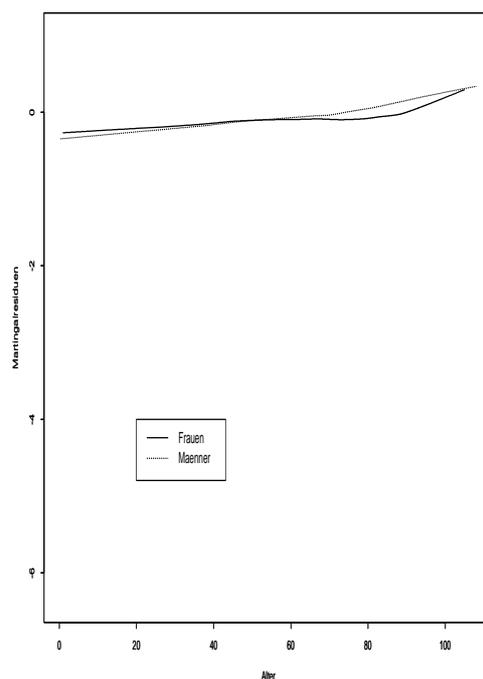


Abbildung 6.7: Vergleich der geglätteten Martingalresiduen für Männer und Frauen in Modell (6.4)

gnosen beschränken.

Vielleicht kann eine weitere Verbesserung erzielt werden, wenn man den partiellen Log-Likelihood dieses Modells erneut für verschiedene Alterswerte θ_w und θ_m maximiert. Eventuell sind durch Hinzunahme der Interaktionen mit \mathbf{Z}_{Tumore} und \mathbf{Z}_{Lungen} nun nicht mehr die Werte 72 Jahre für Frauen und 69 Jahre für Männer optimal. Wir ersetzen also in diesem Modell die Kovariablen \mathbf{Z}_{θ_w72} bzw. \mathbf{Z}_{θ_m69} durch \mathbf{Z}_{θ_w} bzw. \mathbf{Z}_{θ_m} und fitten dieses Modell für verschiedene Kombinationen der Alterswerte $\theta_w = 71 + i$, $\theta_m = 60 + i$, $i = 0, 1, \dots, 24$. Bei Frauen beginnen wir erst mit 71 Jahren, das dies das minimale Alter ist, das bei Frauen mit Lungenerkrankungen aufgetreten ist. Der partielle Log-Likelihood wird mit -14159.27 maximiert für 70 Jahre bei Männern und wieder 72 Jahre bei Frauen. Mit diesen Werten verbessert sich das AIC auf 28396.54 (Tabelle 6.26).

Ebenso gehen wir mit den Kovariablen \mathbf{Z}_{θ_m88} und \mathbf{Z}_{θ_w91} zu den Alterswerten 88 Jahre für Männer und 91 Jahre für Frauen vor. Wieder gehen wir von Modell (6.4) aus, nun um \mathbf{Z}_{θ_w91} und \mathbf{Z}_{θ_m88} erweitert. Auch hier trägt keine der Interaktionen der beiden neuen Kovariablen mit den Kovariablen \mathbf{Z}_{Alter} , \mathbf{Z}_{stat} , \mathbf{Z}_{Stufe2} und \mathbf{Z}_{Stufe3} zu einer signifikanten Verbesserung des Modells bei. Lediglich die Interaktion $\mathbf{Z}_{\theta_w91} \times \mathbf{Z}_{Tumore}$ liegt mit einem p-Wert von 0.027 unter dem geforderten 5%-Level und wird ins Modell hinzugenommen. Tabelle 6.26 enthält die Werte des partiellen Log-Likelihood und des AIC.

Wie oben wird auch hier getestet, ob eine erneute Maximierung des partiellen Log-Likelihood in Abhängigkeit von θ_w und θ_m , diesmal mit der Interaktion $\mathbf{Z}_{\theta_w} \times \mathbf{Z}_{Tumore}$, das Modell verbessert. Dazu wird dieses Modell für verschiedene Kombinationen der Alters-

| | Log-Likelihood | df | AIC | p-Wert |
|--|----------------|----|----------|----------------------|
| Modell 6.4 | -14170.24 | 32 | 28404.48 | |
| mit $\mathbf{Z}_{\theta_{w72}} + \mathbf{Z}_{\theta_{w72}} \times \mathbf{Z}_{Lungen} + \mathbf{Z}_{\theta_{m69}} \times (\mathbf{Z}_{Tumore} + \mathbf{Z}_{Lungen})$ | -14161.89 | 39 | 28401.84 | $5.7 \cdot 10^{-3}$ |
| mit $\mathbf{Z}_{\theta_{w72}} + \mathbf{Z}_{\theta_{w72}} \times \mathbf{Z}_{Lungen} + \mathbf{Z}_{\theta_{m70}} \times (\mathbf{Z}_{Tumore} + \mathbf{Z}_{Lungen})$ | -14159.27 | 39 | 28396.54 | $2.6 \cdot 10^{-3}$ |
| mit $\mathbf{Z}_{\theta_{w91}} + \mathbf{Z}_{\theta_{m88}} + \mathbf{Z}_{\theta_{w91}} \times \mathbf{Z}_{Tumore}$ | -14164.82 | 37 | 28403.64 | $5.5 \cdot 10^{-2}$ |
| mit $\mathbf{Z}_{\theta_{w80}} + \mathbf{Z}_{\theta_{m88}} + \mathbf{Z}_{\theta_{w80}} \times \mathbf{Z}_{Tumore}$ | -14163.8 | 37 | 28401.6 | $2.5 \cdot 10^{-2}$ |
| mit $\mathbf{Z}_{\theta_{w72}} + \mathbf{Z}_{\theta_{w80}} + \mathbf{Z}_{\theta_{w72}} \times \mathbf{Z}_{Lungen} + \mathbf{Z}_{\theta_{m70}} \times (\mathbf{Z}_{Tumore} + \mathbf{Z}_{Lungen}) + \mathbf{Z}_{\theta_{w80}} \times \mathbf{Z}_{Tumore}$ | -14149.8 | 42 | 28383.6 | $1.18 \cdot 10^{-5}$ |

Tabelle 6.26: Entwicklung des Log-Likelihood , AIC und p-Wert des lokalen LR-Test bei Hinzunahme verschiedener θ -Kovariablen zu Modell 6.4

werte $\theta_w = 78 + i$, $\theta_m = 75 + i$, $i = 0, 1, \dots, 24$ gefittet. Der partielle Log-Likelihood wird mit -14163.8 maximiert für die Alterswerte 80 Jahre bei Frauen und wieder 88 Jahre bei Männern , der Wert des AIC wird leicht verbessert (Tabelle 6.26). Die Kovariable $\mathbf{Z}_{\theta_{w91}}$ wird also durch $\mathbf{Z}_{\theta_{w80}}$ ersetzt.

Nun werden alle neuen Kovariablen und Interaktionen, die als signifikant erkannt wurden, in einem Modell zusammengeführt. Modell (6.4) wird um $\mathbf{Z}_{\theta_{w72}}$, $\mathbf{Z}_{\theta_{w80}}$, $\mathbf{Z}_{\theta_{m70}}$, $\mathbf{Z}_{\theta_{m88}}$ und die Interaktionen $\mathbf{Z}_{\theta_{w80}} \times \mathbf{Z}_{Tumore}$, $\mathbf{Z}_{\theta_{w72}} \times \mathbf{Z}_{Lungen}$ und $\mathbf{Z}_{\theta_{m70}} \times (\mathbf{Z}_{Tumore} + \mathbf{Z}_{Lungen})$ erweitert. Der Wert des partiellen Log-Likelihood verringert sich so auf -14149.8, auch der Wert des AIC sinkt (Tabelle 6.26). Ein lokaler LR-Test auf alle im Vergleich zu Modell (6.4) neu hinzugenommenen Kovariablen führt zu einem p-Wert von $1.18 \cdot 10^{-5}$. Die neu eingeführten Kovariablen tragen also signifikant zur Modellierung der Überlebenszeit der Pflegepatienten bei.

6.5.2 Exponentieller Ansatz für das Alter

Die vorangegangenen Untersuchungen zeigten, dass die Kovariablen $\mathbf{Z}_{\theta_{w72}}$, $\mathbf{Z}_{\theta_{w80}}$, $\mathbf{Z}_{\theta_{m70}}$ und $\mathbf{Z}_{\theta_{m88}}$ in enger Verbindung zu einzelnen Diagnosen , nämlich Tumoren und Lungenerkrankungen, und dem Geschlecht stehen. Dies und die Tatsache, dass die beiden größten Werte des partiellen Log-Likelihoods bei ganz unterschiedlichen Alterswerten auftreten, legt die Vermutung nahe, dass die funktionale Form des Alters sehr von den einzelnen Diagnosen abhängt. Daher bilden wir nun für die verschiedenen Diagnosegruppen (Einzel-, Doppel- und Dreifachdiagnosen), jeweils getrennt für Männer und Frauen , Plots der Martingalresiduen gegen das Alter. Wieder werden die Martingalresiduen von Modell (6.4) ohne \mathbf{Z}_{Alter} und die damit verbundenen Interaktionen verwendet.

In den Plots 6.8 - 6.10 sieht man deutlich, dass die funktionale Form des Alters in den verschiedenen Diagnosegruppen ganz unterschiedlich ist. Nur bei einigen Gruppen sind noch Knicke erkennbar wie in Abbildung 6.6. Viele Diagnosegruppen lassen eher auf einen exponentiellen oder quadratischen Verlauf des Alters schließen. Darüber hinaus fällt auf,

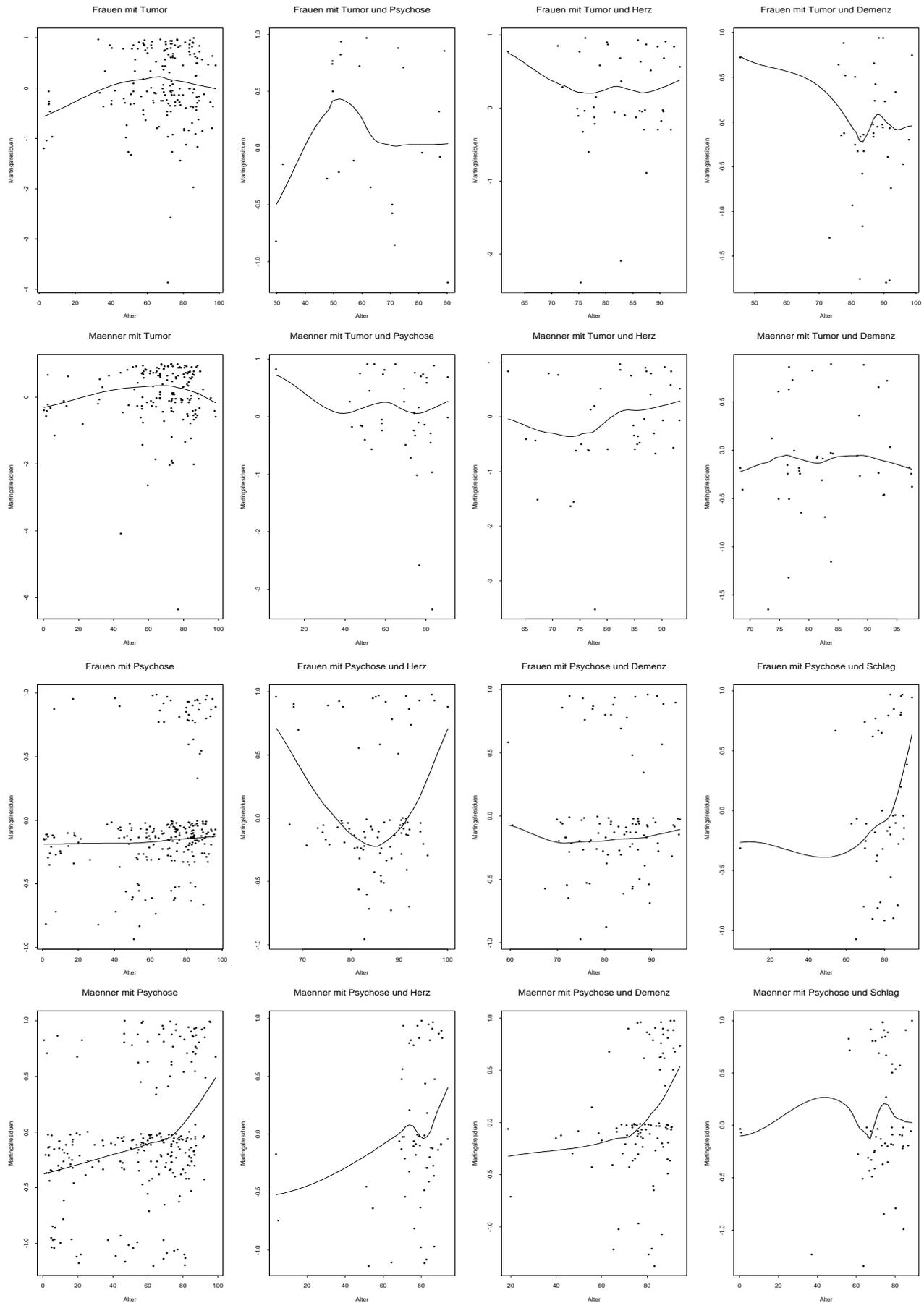


Abbildung 6.8: Martingalresiduen für die Einzel- und Doppeldiagnosengruppen für Tumore und Psychose

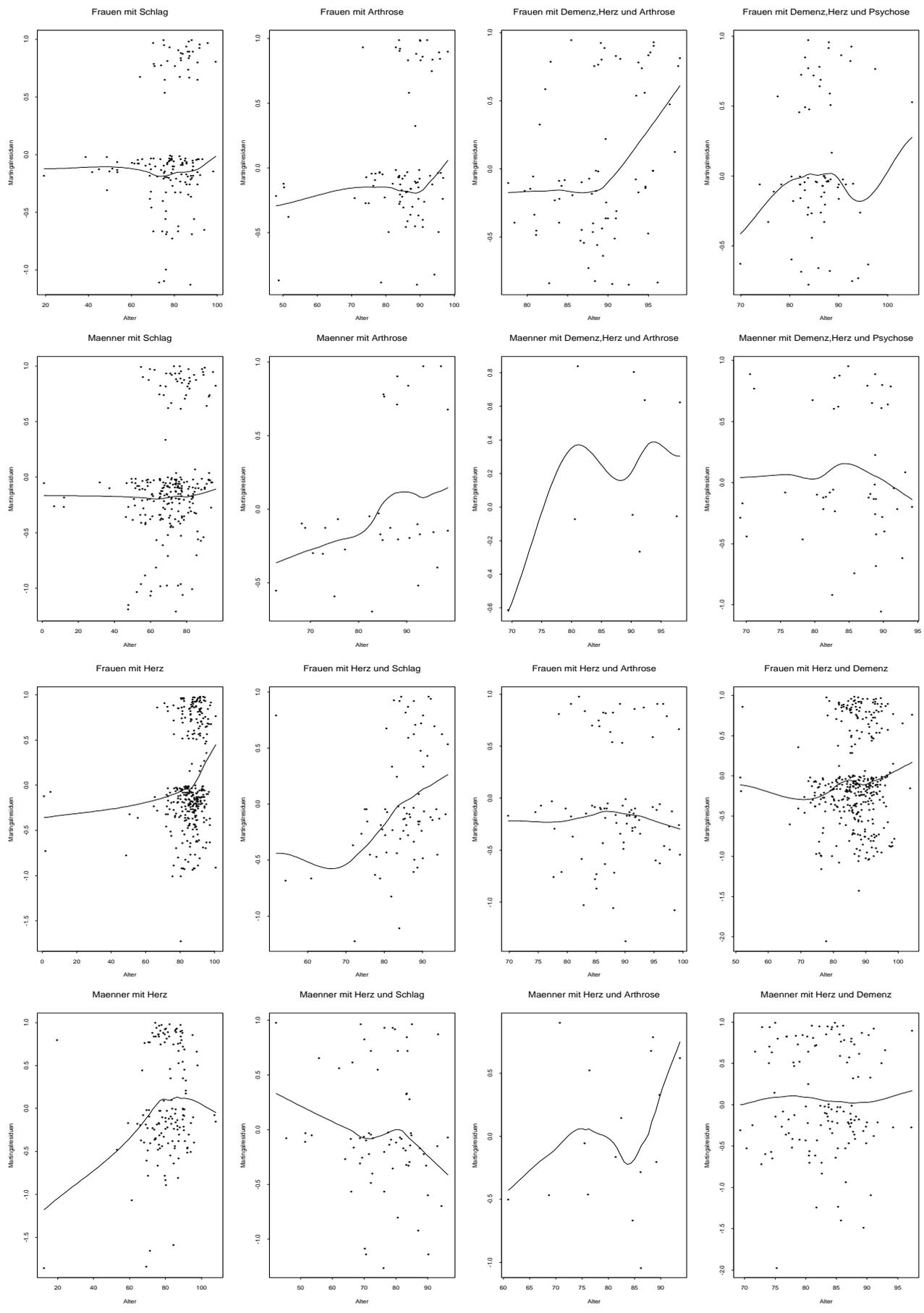


Abbildung 6.9: Martingalresiduen für die Einzel-, Doppel- und Dreifachdiagnosegruppen für Schlaganfall, Herz und Arthrosen

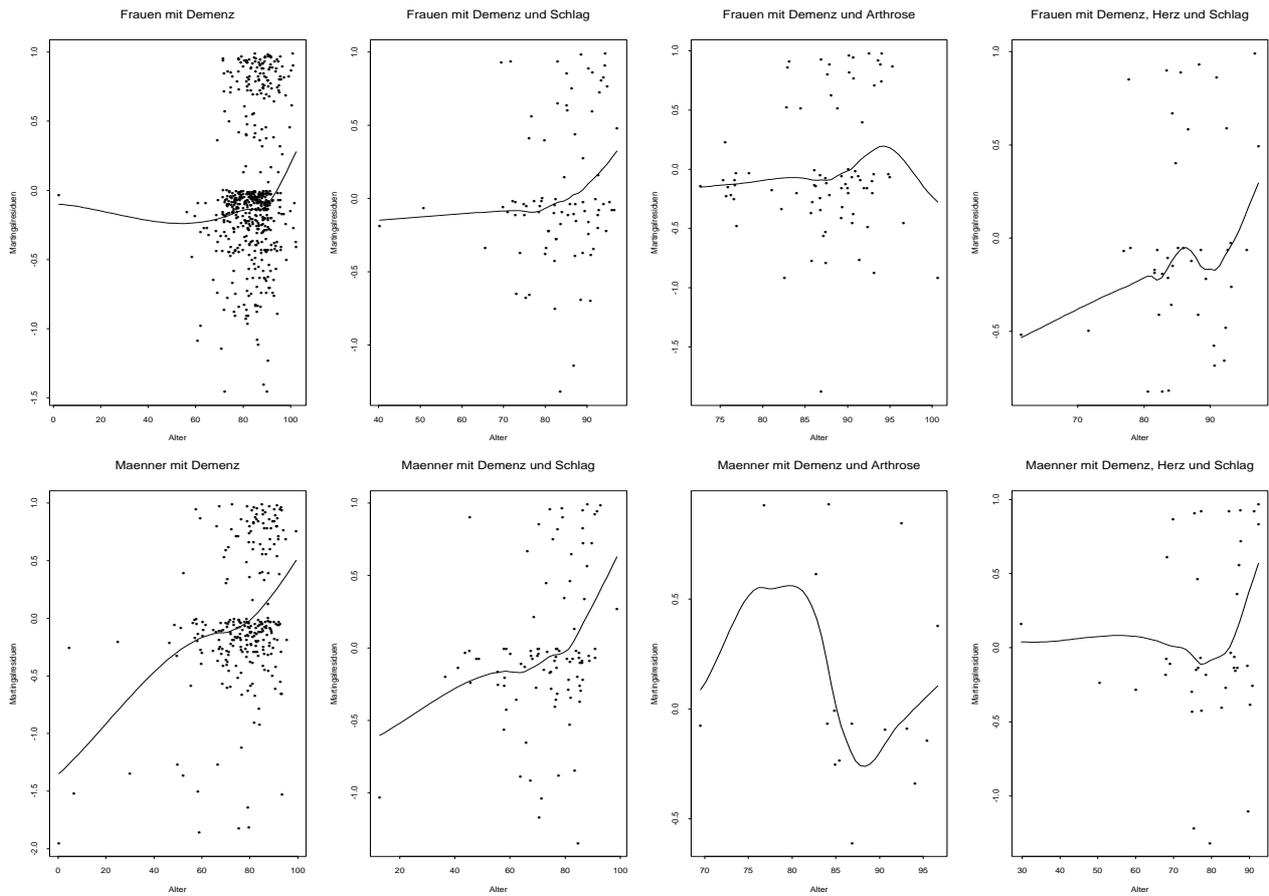


Abbildung 6.10: Martingalresiduen für die Einzel-, Doppel- und Dreifachdiagnosegruppen mit Demenz

dass sich die funktionale Form des Alters in den Einzeldiagnosen bei Hinzukommen einer weiteren Diagnose meist stark verändert. Dies kann zwar zum Teil dadurch erklärt werden, dass in den Doppel- und Dreifachdiagnosegruppen oft nur noch wenige Daten vorliegen. Dennoch sollen in den folgenden Untersuchungen exklusive Gruppen, d.h. Einzel-, Doppel- und Dreifachdiagnosegruppen getrennt, betrachtet werden. Nur Gruppen, die einen ähnlichen Verlauf des Alters vermuten lassen, werden zusammengefasst. So bilden z.B. Männer mit Psychose als einziger Diagnose, Männer mit Psychose und Herzkrankung und Männer mit Psychose und Demenzerkrankung zusammen eine Gruppe (siehe Abb. 6.8).

Es soll nun getestet werden, ob ein exponentieller Ansatz für das Alter in den verschiedenen Untergruppen signifikante Verbesserungen bringt. Modell (6.4) wird zu diesem Zweck um die Interaktion

$$\exp(c \cdot \mathbf{Z}_{Alter}) \times \mathbf{Z}_{Gruppe}$$

erweitert. Dabei ist \mathbf{Z}_{Gruppe} die Indikatorvariable der jeweils gerade untersuchten Diagnosegruppe (unterteilt nach Geschlecht).

Die Variable \mathbf{Z}_{Alter} wird durch

$$\mathbf{Z}_{Alter} \times (1 - \mathbf{Z}_{Gruppe})$$

ersetzt. Dadurch wird gewährleistet, dass der lineare Ansatz für das Alter nur für die übrigen Patienten beibehalten wird, nicht aber für die gerade untersuchte Gruppe.

Für jede Untergruppe wird dieses Modell für verschiedene Werte des Parameters c gefittet und der Wert des partiellen Log-Likelihoods betrachtet. c variiert dabei in Schritten von 0.004 von 0.002 bis 0.102. Für Frauen und Männer mit Tumor wurden aufgrund der konkaven Krümmung der Martingalplots zusätzlich auch negative Werte für c ($c = -0.102$ bis $c = -0.002$) untersucht. Dabei ergibt sich, dass lediglich bei Frauen mit Tumoren als pflegeauslösender Diagnose und Männern mit Psychose, Psychose und Herz- bzw. Demenzerkrankungen ein signifikantes Ergebnis erreicht werden kann. Für Frauen mit Tumor wird der partielle Log-Likelihood für $c = -0.102$ auf -14163.77 erhöht. Wieder werden 2 Freiheitsgrade für die neue Interaktion angesetzt, da der optimale Wert von c durch Maximieren nach c gefunden wurde. Zum Vergleich muss das Modell betrachtet werden, in welches das Alter für Frauen mit Tumor linear eingeht, d.h. $\exp(-0.102 \cdot \mathbf{Z}_{Alter}) \times \mathbf{Z}_{Tumore,w}$ muss durch $\mathbf{Z}_{Age} \times \mathbf{Z}_{Tumore,w}$ ersetzt werden. Der partielle Log-Likelihood für dieses Modell liegt bei -14168.1, daraus resultiert ein p-Wert von $3.24 \cdot 10^{-3}$. Durch den exponentiellen Ansatz wird also in dieser Gruppe das Alter signifikant besser modelliert.

In der Gruppe der Männer mit Psychose, Herz- und Demenzerkrankungen wird der Log-Likelihood für $c = 0.022$ mit -14167.76 maximiert. Beim Vergleich mit dem Modell mit einem linearen Ansatz für diese Gruppe (partieller Log-Likelihood: -14170.6), ergibt dies einen p-Wert von $1.7 \cdot 10^{-2}$.

In Abbildung 6.11 sind die durch den exponentiellen Ansatz gefundenen funktionalen Formen für das Alter für beide Gruppen graphisch dargestellt. Ein Vergleich mit den zugehörigen Martingalplots in Abb. 6.8 (letzte Zeile) zeigt, dass das Alter für Männer mit Psychose, Psychose und Herzkrankungen und Psychose und Demenzerkrankungen schon sehr gut modelliert ist. Bei Frauen mit Tumor ist die Übereinstimmung mit dem

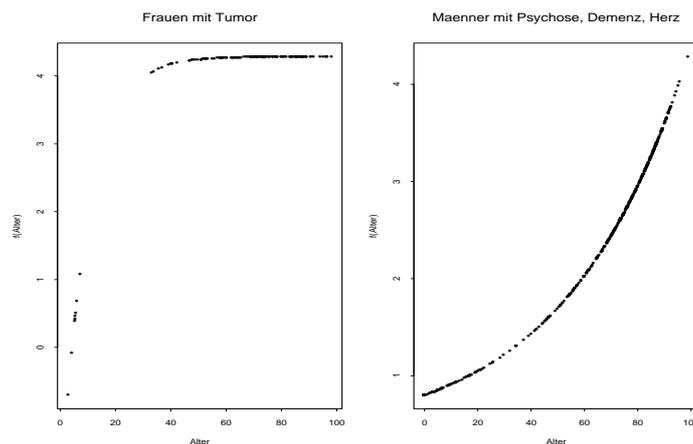


Abbildung 6.11: Exponentieller Ansatz für das Alter für Frauen mit Tumor und Männer mit Psychose, Herz- und Demenzerkrankungen

Martingalplot (siehe Abb. 6.9) nicht ganz so deutlich, der grobe Verlauf ist jedoch auch in dieser Gruppe gut getroffen.

6.5.3 Modellierung des Alters durch fraktionale Polynome

Mittels Exponentialfunktionen konnte nur bei zwei Untergruppen die funktionale Form des Alters besser bestimmt werden. Die Martingalresiduenplots ließen in einigen Untergruppen einen quadratischen Verlauf des Alters vermuten. Dieser und noch weit mehr verschiedene funktionale Formen können mittels der Methode der fraktionalen Polynome (siehe (4.8.3)) ermittelt werden. Daher soll nun, wieder für die verschiedenen Untergruppen, getrennt nach Geschlecht, das Alter durch ein fraktionales Polynom modelliert werden. Wir beschränken uns dabei auf Polynome ersten und zweiten Grades ($m=1,2$). Zunächst betrachten wir den Fall $m=1$. Dazu wird in Modell (6.4) die Kovariable \mathbf{Z}_{Alter} wie im vorangegangenen Abschnitt durch

$$\mathbf{Z}_{Alter} \times (1 - \mathbf{Z}_{Gruppe})$$

ersetzt und die folgende Interaktion neu hinzugenommen:

$$\mathbf{Z}_{Alter}^{p_1} \times \mathbf{Z}_{Gruppe}$$

Die Kovariable \mathbf{Z}_{Gruppe} ist analog zu Abschnitt (6.5.2) definiert.

Dieses Modell wird für jede Untergruppe für alle Werte $p_1 \in \wp = \{-2, -1, -0.5, 0, 0.5, 1, 2\}$ gefittet. Eine Modellierung mit $p_1 = 3$ wird zwar von Royston und Altman [22] vorgeschlagen, brachte in unserem Fall jedoch nur sehr niedrige Werte des partiellen Log-Likelihood hervor. In jeder Untergruppe wird der Wert von p_1 ermittelt, der den höchsten Wert des partiellen Log-Likelihood nach sich zieht.

Ein LR-Test für die neue Interaktion brachte hervor, dass dieser Ansatz nur bei Frauen mit Tumoren signifikante Verbesserungen bringt. In dieser Gruppe ist für $p_1 = -1$ (d.h. \mathbf{Z}_{Alter}^{-1}) der partielle Log-Likelihood mit dem Wert -14164.86 maximal.

Nun wird dieses Vorgehen für $m=2$ wiederholt. Dazu wird Modell (6.4) um die beiden Interaktionen

$$\mathbf{Z}_{Alter}^{p_1} \times \mathbf{Z}_{Gruppe}$$

$$\mathbf{Z}_{Alter}^{p_2} \times \mathbf{Z}_{Gruppe}$$

erweitert. Für jede Untergruppe wird dieses Modell gefittet, wobei p_1 und p_2 nun alle möglichen Kombinationen $(p_1, p_2) \in \varnothing \times \varnothing$ annehmen. Dabei stellt sich heraus, dass für $p_1 = p_2 = 2$ nur sehr schlechte Werte des partiellen Log-Likelihood erreicht werden.

Wiederum werden für Frauen mit Tumoren signifikante Verbesserungen beobachtet. Der partielle Log-Likelihood wurde für $p_1 = p_2 = -1$ mit dem Wert -14163.44 maximiert. Das Alter wird demnach für diese Untergruppe für $m=2$ am besten durch $\mathbf{Z}_{Alter}^{-1} + \log(\mathbf{Z}_{Alter}) \times \mathbf{Z}_{Alter}^{-1}$ beschrieben.

Auch in der Gruppe der Männern mit Psychose, Herz- bzw. Demenzerkrankungen erhält man auf einem 10 %-Level signifikante Ergebnisse. Hier wird ein maximaler Log-Likelihood von -14167.09 für $p_1 = -2$ und $p_2 = 1$ erzielt. Alter besitzt hier also die funktionale Form $\mathbf{Z}_{Alter}^{-2} + \mathbf{Z}_{Alter}$.

Für diese beiden Untergruppen werden wir nun testen ob die gefundenen fraktionalen Polynome für $m=1,2$ signifikant besser sind als ein linearer Ansatz, d.h. ein fraktionales Polynom ersten Grades mit der Potenz $p_1 = 1$.

Für die Gruppe Frauen mit Tumoren erzielt dieses Modell einen partiellen Log-Likelihood von -14168.1 (siehe Abschnitt (6.5.2)). Um es mit dem besten Modell für $m=1$ zu vergleichen, betrachten wir die LR-Statistik

$$\chi_{LR}^2 = 2 \cdot (-14164.86 + 14168.1) = 6.48,$$

die mit einem Freiheitsgrad χ^2 -verteilt ist. Das führt zu einem p-Wert von 0.011. Durch das fraktionale Polynoms ersten Grades mit der Potenz $p_1 = -1$ wird das Alter also signifikant besser modelliert als durch einen linearen Ansatz. Die LR-Statistik zum Vergleich des besten Modells für $m=2$ mit dem linearen Modell lautet

$$\chi_{LR}^2 = 2 \cdot (-14163.44 + 14168.1) = 9.32$$

und ist mit 3 Freiheitsgraden χ^2 -verteilt. Der p-Wert liegt hier bei 0.02, das Modell für $m=2$ ist also wieder signifikant besser als das lineare Modell. Das beste Modell für $m=2$ wird mit dem besten Modell für $m=1$ mit der LR-Statistik

$$\chi_{LR}^2 = 2 \cdot (-14163.44 + 14164.86) = 2.84$$

verglichen. Bei einer χ^2 -Verteilung mit 2 Freiheitsgraden ergibt sich ein p-Wert von 0.24. Das bedeutet, dass das beste Modell für $m=1$ dem besten Modell für $m=2$ vorzuziehen ist.

Analog gehen wir in der Gruppe Männer mit Psychose, Herz- und Demenzerkrankungen vor. Der partielle Log-Likelihood des linearen Modells liegt hier bei -14170.6 (siehe Abschnitt (6.5.2)). Vergleicht man dieses Modell mit dem besten Modell mit fraktionalem Polynom zweiten Grades, erhält man mit einer LR-Statistik von

$$2 \cdot (-14167.09 + 14170.6) = 7.02$$

| | Log-Likelihood | $2 \cdot \chi_{LR}^2$ vs. linear | p-Wert | Potenzen |
|---------------|----------------|----------------------------------|--------|----------|
| linear (1 df) | -14168.1 | | | 1 |
| m=1 (2 df) | -14164.86 | 3.24 | 0.011 | -1 |
| m=2 (4 df) | -14163.44 | 4.66 | 0.02 | -1, -1 |

Tabelle 6.27: Fraktionale Polynome für Alter bei Frauen mit Tumoren

| | Log-Likelihood | $2 \cdot \chi_{LR}^2$ vs. linear | p-Wert | Potenzen |
|---------------|----------------|----------------------------------|--------|----------|
| linear (1 df) | -14170.6 | | | 1 |
| m=1 (2 df) | -14169.48 | 2.24 | 0.13 | 2 |
| m=2 (4 df) | -14167.09 | 7.02 | 0.07 | -2,1 |

Tabelle 6.28: Fraktionale Polynome für Alter bei Männern mit Psychose, Herz- oder Demenzerkrankungen

und 3 Freiheitsgraden einen p-Wert von 0.07.

In den Tabellen 6.27 und 6.28 findet sich eine Zusammenfassung der Ergebnisse.

Die gefundenen optimalen fraktionale Polynome für beide Gruppen sind in Abb. 6.12 graphisch dargestellt.

6.5.4 Vergleich der Modelle

Der exponentielle Ansatz und die Modellierung des Alters mittels fraktionaler Polynome führen zu ähnlichen Ergebnissen. In der Gruppe der Männer mit Psychose, Herz- und Demenzerkrankungen liefern beide Methoden in etwa den selben Wert des partiellen Log-Likelihood. Da im Modell mit dem exponentiellen Ansatz jedoch zwei Freiheitsgrade weniger enthalten sind, ist dieses Modell vorzuziehen. Auch die Plots der funktionalen Form Abb.6.11 und Abb.6.12 zeigen, dass der exponentielle Ansatz den durch den Martingalplot vorgegebenen Verlauf des Alters besser modelliert.

Um zu überprüfen, ob das Alter durch diese Transformation nun linear in das Modell eingeht, betrachten wir für diese Untergruppe nochmals die Martingalresiduen. Dieses Mal werden die Martingalresiduen des Modells, in dem der exponentielle Term berücksichtigt ist, gegen das Alter geplottet.

In Abbildung 6.13 sieht man deutlich, dass der geglättete Plot der Martingalresiduen nun nahezu linear ist, so dass für diese Untergruppe die im Proportional Hazard Modell unterstellte Linearität der stetigen Kovariablen gegeben ist.

Bei Frauen mit Tumor erhält man ebenfalls durch beide Ansätze für den partiellen Log-Likelihood Werte in der gleichen Größenordnung. Auch in den Abbildungen 6.11 und 6.12 sind kaum Unterschiede in der Form des Alters zu erkennen. Man sieht jedoch deutlich, dass das Alter in beiden Plots ab etwa 40 Jahren linear mit Steigung 0 verläuft. Die funktionale Form wird in erster Linie von den wenigen Beobachtungen bestimmt, die bis zu einem Alter von etwa 15 Jahren vorliegen. Bei diesen Beobachtungen handelt es sich um Mädchen mit Tumor bis zu 15 Jahren. Es ist durchaus vorstellbar, dass in dieser Gruppe andere Tumorarten und auch ein anderer Krankheitsverlauf als bei Frauen über

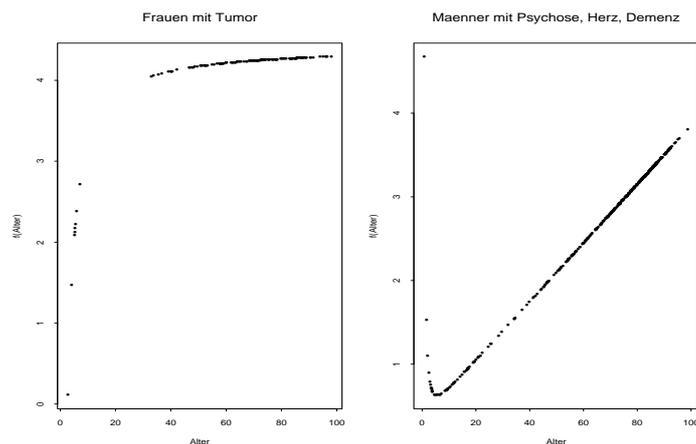


Abbildung 6.12: Fraktionale Polynome für das Alter bei Frauen mit Tumor und Männern mit Psychose, Herz- und Demenzerkrankungen

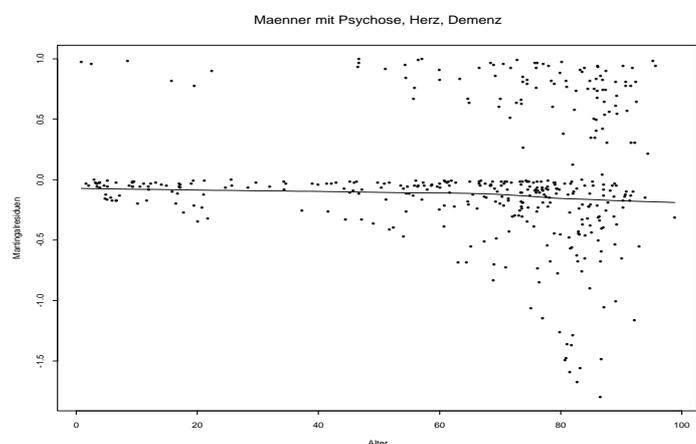


Abbildung 6.13: Martingalresiduen im Modell mit exponentiellem Ansatz für Männer mit Psychose, Herz- und Demenzerkrankungen

15 Jahren mit Tumor vorherrschen. Eventuell wird das Modell verbessert, wenn man hier zwei eigenständige Gruppen bildet.

Der Plot der Martingalresiduen gegen das Alter, getrennt für die beiden Altersgruppen (Abb. 6.14) zeigt, dass das Alter bei Frauen ab 15 Jahren tatsächlich nahezu linear verläuft. Der Plot der Martingalresiduen für Mädchen ist aufgrund der geringen Datenmenge nicht aussagefähig.

Daher wird nun ein Modell gefittet, in dem das Alter linear, jedoch für Mädchen (≤ 15 Jahre) und Frauen (> 15 Jahre) mit Tumor getrennt gefittet wird. Dazu werden die beiden Kovariablen $\mathbf{Z}_{\text{Alter}} \times \mathbf{Z}_{\text{Tumor}, \text{Mädchen}}$ und $\mathbf{Z}_{\text{Alter}} \times \mathbf{Z}_{\text{Tumor}, \text{Frauen}}$ in das Modell genommen. $\mathbf{Z}_{\text{Tumor}, \text{Mädchen}}$ und $\mathbf{Z}_{\text{Tumor}, \text{Frauen}}$ sind dabei wieder Indikatorvariablen für die beiden Gruppen, d.h.

$$\mathbf{Z}_{\text{Tumor}, \text{Mädchen}} = \begin{cases} 1 & \text{falls Patient weiblich, } \leq 15 \text{ Jahre mit Diagnose Tumor} \\ 0 & \text{sonst} \end{cases}$$

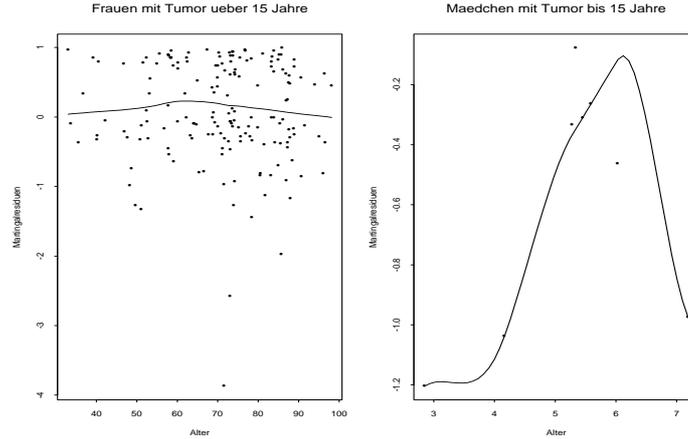


Abbildung 6.14: Martingalresiduen getrennt für Frauen mit Tumor über 15 Jahre und Mädchen mit Tumor bis zu 15 Jahre

$$\mathbf{Z}_{Tumor,Frauen} = \begin{cases} 1 & \text{falls Patient weiblich, } > 15 \text{ Jahre mit Diagnose Tumor} \\ 0 & \text{sonst} \end{cases}$$

Für dieses Modell erhält man mit -14163.57 einen genauso hohen Wert des partiellen Log-Likelihoods wie in den Modellen mit exponentiellem Ansatz und fraktionalem Polynom.

Der Betrag des Regressionskoeffizienten der Kovariable $\mathbf{Z}_{Alter} \times \mathbf{Z}_{Tumor,Frauen}$ liegt mit 0.0004 sehr nahe bei 0, das Alter hat bei Frauen mit Tumor also so gut wie keinen Einfluß auf das Überleben. Dies wird auch durch den hohen p-Wert von 0.91 dieser Kovariable bestätigt. Es genügt daher, nur die Kovariable $\mathbf{Z}_{Alter} \times \mathbf{Z}_{Tumor,Mädchen}$ im Modell zu behalten. Zusammen mit dem exponentiellem Term für die Gruppe der Männer mit Psychose, Herz- und Demenzerkrankungen $\mathbf{Z}_{P,H,D,m}$, wobei

$$\mathbf{Z}_{P,H,D,m} = \begin{cases} 1 & \text{falls Patient männlich, mit Einzeldiagnose Psychose,} \\ & \text{Doppeldiagnose Psychose und Herz oder Psychose und Demenz} \\ 0 & \text{sonst} \end{cases} ,$$

erhält man nun ein Modell mit folgender Hazardfunktion

$$\begin{aligned} \lambda(t) &= \lambda_0(t) \exp[\boldsymbol{\beta}'\mathbf{Z}(\text{Modell (6.4) ohne } \mathbf{Z}_{Alter})] \\ &\cdot \exp[\beta_1 \mathbf{Z}_{Alter} \times (1 - \mathbf{Z}_{Tumore,w}) \times (1 - \mathbf{Z}_{P,H,D,m}) \\ &+ \beta_{33} \mathbf{Z}_{Alter} \times \mathbf{Z}_{Tumore,Mädchen} + \beta_{34} \exp(0.022 \cdot \mathbf{Z}_{Alter}) \times \mathbf{Z}_{P,H,D,m}]. \end{aligned} \quad (6.5)$$

Für dieses Modell liegt der partielle Log-Likelihood bei -14161.22, der Wert des AIC beträgt bei 35 Freiheitsgraden 28392.44.

Auch die Modellierung von Knickpunkten in Abschnitt (6.5.1) führte zu signifikanten Ergebnissen . Die Hazardfunktionen des besten θ - Modells lautet

$$\begin{aligned} \lambda(t) &= \lambda_0(t) \exp[\boldsymbol{\beta}'\mathbf{Z}(\text{Modell (6.4)})] \cdot \exp[\beta_{33} \mathbf{Z}_{\theta_{w72}} + \beta_{34} \mathbf{Z}_{\theta_{w80}} + \beta_{35} \mathbf{Z}_{\theta_{m70}} \\ &+ \beta_{36} \mathbf{Z}_{\theta_{w72}} \times \mathbf{Z}_{Lungen} + \beta_{37} \mathbf{Z}_{\theta_{m70}} \times \mathbf{Z}_{Tumore} + \beta_{38} \mathbf{Z}_{\theta_{m70}} \mathbf{Z}_{Lungen} \\ &+ \beta_{39} \mathbf{Z}_{\theta_{w80}} \times \mathbf{Z}_{Tumore}], \end{aligned} \quad (6.6)$$

Vergleicht man diese beiden Modelle allein aufgrund des partiellen Loglikelihoods und des AIC (Tabelle 6.26), liefert Modell (6.6) die besseren Resultate. Der Wert des AIC liegt knapp 9 Punkte unter dem von Modell (6.5).

Die Plots der funktionalen Form des Alters in Modell (6.6) (Abbildung 6.15) deuten dagegen nur sehr grob auf den durch die Martingalplots gegebenen Verlauf des Alters hin. Bei Frauen mit Tumor zum Beispiel spiegeln die Sprungstellen bei 72 und 80 Jahren in etwa den leicht fallenden Verlauf des Alters ab etwa 70 Jahre wider. Insgesamt jedoch scheint die stetige Modellierung in Modell (6.5) ein sinnvollerer Ergebnis zu liefern.

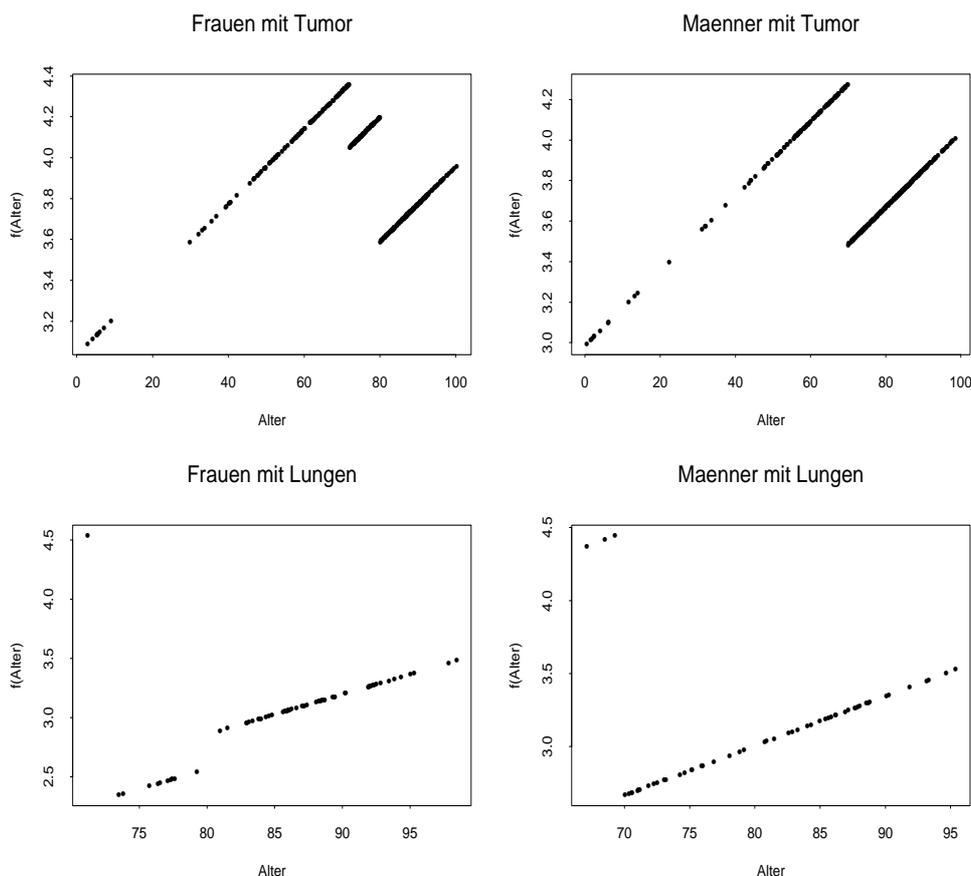


Abbildung 6.15: Funktionale Form des Alters für Männer und Frauen mit Tumor bzw. Lungenerkrankung in Modell (6.6)

Auch für Modell (6.6) werden nochmals Plots der Martingalresiduen für die einzelnen Untergruppen betrachtet. Dazu werden die Martingalresiduen von Modell (6.6) ohne den linearen Term des Alters, jedoch mit den Termen der θ -Kovariablen und θ -Interaktionen, gegen das Alter geplottet. Abb. 6.16 zeigt, dass die Plots der Martingalresiduen im θ -Modell kaum von denen in Modell (6.4) zu unterscheiden sind.

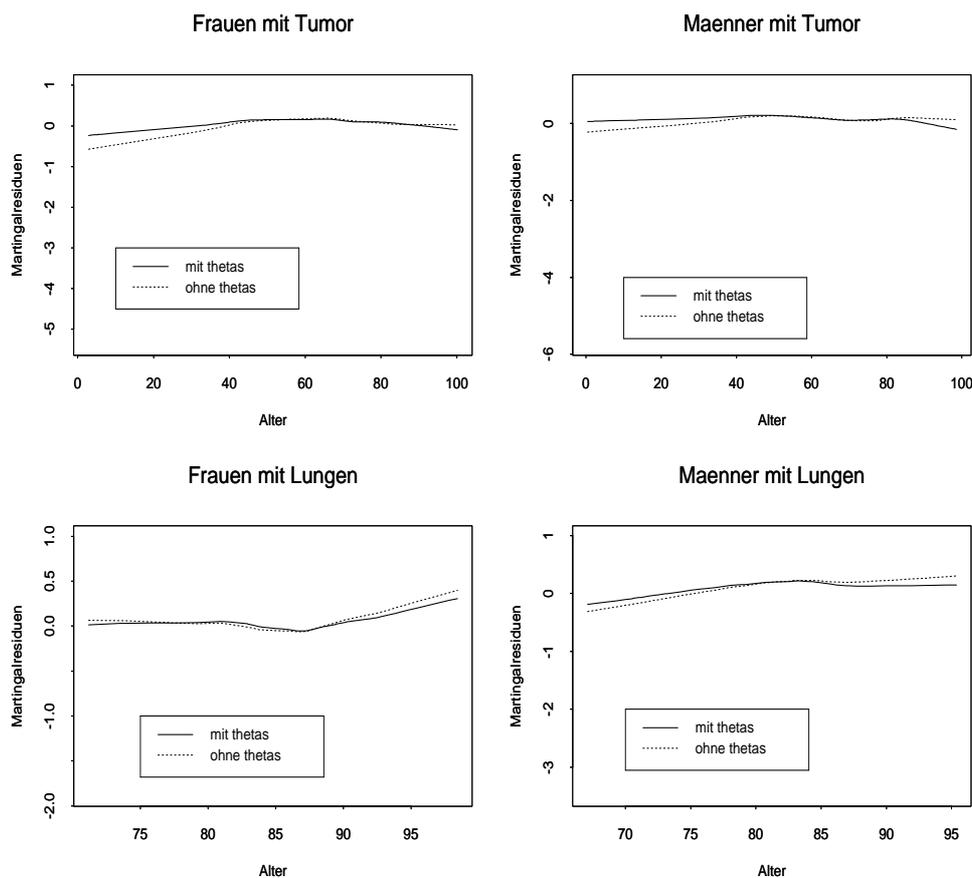


Abbildung 6.16: Martingalresiduen für Männer und Frauen mit Tumor bzw Lungenerkrankung mit und ohne θ -Knickpunkte

6.6 Untersuchung der Zeitabhängigkeit der Koeffizienten

In diesem Abschnitt soll überprüft werden, ob die Annahme der „proportional Hazards“ in unserem Modell erfüllt ist, oder ob zeitabhängige Koeffizienten vorhanden sind. Graphische Auskunft darüber gibt nach Abschnitt 4.9 ein geglätteter Plot der skalierten Schönfeldresiduen gegen die Zeit.

Abbildung 6.17 zeigt diese Plots für die einzelnen Kovariablen sowohl für Modell (6.4), als auch für die Modelle mit den θ - Variablen (6.6) und dem exponentiellem Ansatz (6.5). Alle drei Modelle weisen eine starke Zeitabhängigkeit der Kovariablen auf, bei den meisten Kovariablen unterscheiden sich die Modelle kaum. Die Splus-Routine `zph`, mittels der diese Plots durchgeführt wurden, führt zugleich für jede Kovariable Tests auf Geraden mit Steigung 0 durch, d.h. es wird die Hypothese: „Einfluß der Kovariable ändert sich nicht mit der Zeit“ getestet. Die globalen p-Werte dieser Tests liegen bei allen Modellen unter 10^{-16} und bestätigen somit die graphischen Ergebnisse.

Betrachtet man die einzelnen Plots in Abb. 6.17 genauer, sieht man, dass bei den Kovariablen Z_{stat} (3.Reihe von oben, links), Z_{Stufe2} und Z_{Stufe3} (4. Reihe von oben) die

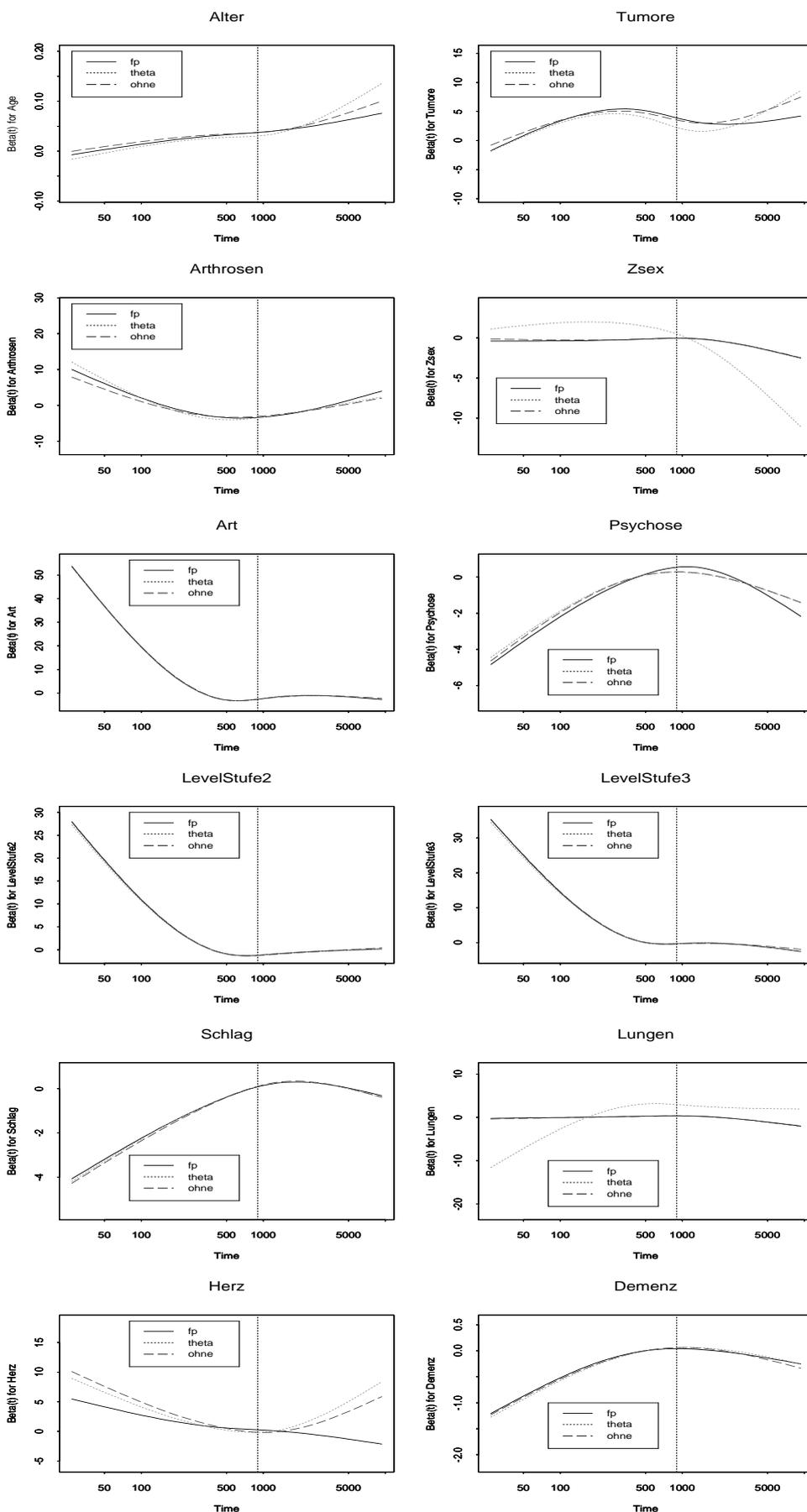


Abbildung 6.17: Plot der skalierten Schoenfeldresiduen gegen die Zeit für die Modelle (6.4), (6.6) und (6.5)

Zeitabhängigkeit vor allem in den ersten 900 Tagen der Pflege vorliegt. Dieser Zeitpunkt ist durch eine gestrichelte Linie gekennzeichnet. Für längere Behandlungszeiten sind die Plots dagegen relativ konstant. In den ersten zweieinhalb Pflegejahren scheinen die Kovariablen wie z.B. Pflegeart, Pflegestufe oder Alter also einen anderen Einfluß auf das Überleben der Patienten zu haben, als bei längeren Behandlungsdauern. Um diese Vermutung näher zu untersuchen, wird Modell (6.5) um einen globalen Wendepunkt zum Zeitpunkt 900 Tage erweitert. Dazu wird die Indikatorvariable

$$\mathbf{Z}_{t_{900}} = \begin{cases} 1 & \text{falls } t \leq 900 \text{ Tage} \\ 0 & \text{sonst} \end{cases}$$

eingeführt und das Modell (6.5), erweitert um die Interaktionen von $\mathbf{Z}_{t_{900}}$ mit allen anderen Kovariablen, erneut gefittet. Die Hazardfunktion lautet dann

$$\lambda(t) = \lambda_0(t) \exp[\boldsymbol{\beta}'\mathbf{Z}(t) + \boldsymbol{\alpha}'\mathbf{Z}_{t_{900}} \times \mathbf{Z}(t)],$$

wobei hier $\mathbf{Z}(t)$ den Kovariablenvektor aus Modell (6.4) bezeichnet.

Der Log-Likelihood von diesem Modell beträgt -14136.8, eine wesentliche Verbesserung im Vergleich zu dem Wert des Log-Likelihood von -14170.24 in Modell (6.4). Obwohl die Anzahl der Freiheitsgrade auf 63 erhöht wurde, liegt der Wert des AIC mit 28399.6 unter dem von Modell (6.4). Auch ein lokaler LR-Test für alle neuen Interaktionen zeigt mit einem p-Wert von $1.9 \cdot 10^{-4}$, dass das Modell durch den globalen Wendepunkt signifikant verbessert wurde. Denselben Ansatz führen wir bei den Modellen (6.6) und (6.5) durch und stellen auch dort eine Verbesserung der Modelle in ähnlich hoher Größenordnung fest. Die Vermutung, dass sich die Effekte der Kovariablen in den ersten 900 Tagen von denen in späterer Behandlung unterscheiden, hat sich bestätigt. Daher teilen wir den Datensatz nach der Behandlungsdauer auf in einen Datensatz, in dem nur die Patienten mit Behandlungsdauer ≤ 900 Tage, und einen, in dem nur Patienten mit Behandlungsdauer > 900 Tage berücksichtigt sind. Für beide Datensätze wird nun erneut ein Proportional Hazard Modell angepasst. Die Signifikanz der Kovariablen und Interaktionen wird wie in Abschnitt (6.3) mittels der p-Werte des Wald- bzw. LR-Tests überprüft. Auch für diese beiden Modelle soll das Alter mittels fraktionaler Polynome bzw. eines exponentiellen Ansatzes modelliert werden.

Fraktionale Polynome und exponentieller Ansatz

Im Modell für Behandlungsdauer ≤ 900 Tage hat das Alter bei Frauen mit Tumor wie in Modell (6.5) keinen signifikanten Einfluß auf das Überleben. Auch bei Mädchen mit Tumor spielt das Alter in den ersten zweieinhalb Jahren der Behandlung keine signifikante Rolle. Der Plot der Martingalresiduen für Frauen mit Tumor mit Behandlungsdauern bis zu 900 Tagen (siehe Abb. 6.18, Bild links oben) verläuft fast linear und bestätigt dieses Resultat. Zur Suche nach optimalen fraktionalem Polynomen wird wie in Abschnitt (6.5.3) vorgegangen. Es können jedoch bei keiner der Untergruppen signifikante fraktionale Polynome gefunden werden. Im Gegensatz zu Modell (6.5) verbessert auch ein exponentieller Ansatz für das Alter bei Männern mit Psychose, Herz- oder Demenzerkrankungen das Modell nicht. Der Plot der Martingalresiduen für diese Untergruppe (Abb. 6.18, Bild rechts oben) belegt, dass ein linearer Ansatz hier für das Alter gerechtfertigt ist.

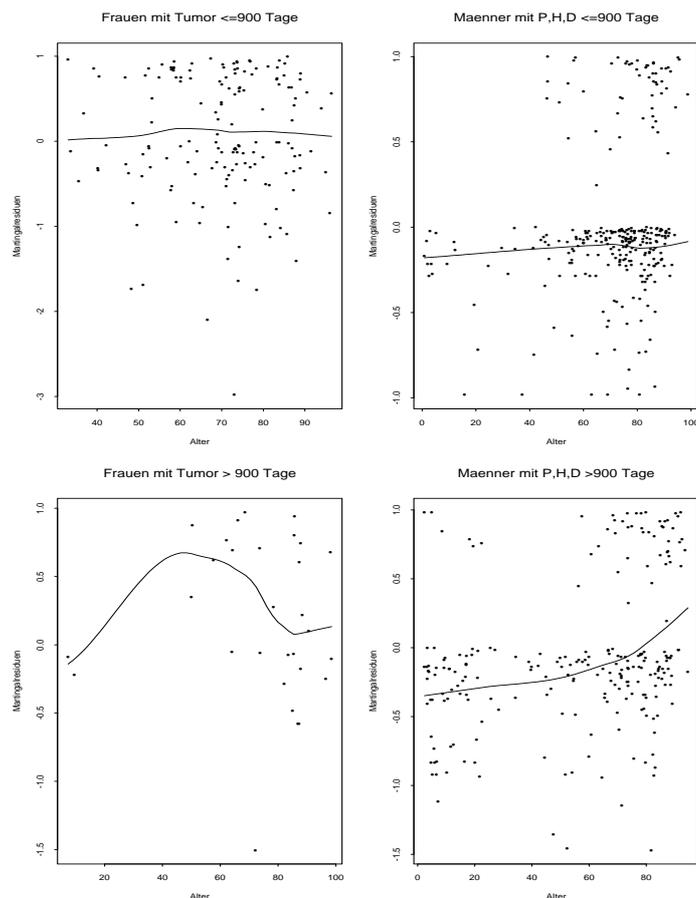


Abbildung 6.18: Plot der Martingalresiduen für Frauen mit Tumor und Männer mit Psychose, Herz oder Demenz, jeweils getrennt für Behandlungsdauern bis zu und über 900 Tage

Im Modell mit Behandlungsdauer > 900 Tage dagegen läßt der Martingalplot (Bild rechts unten) für diese Gruppe eventuell einen exponentiellen Verlauf des Alters vermuten. Untersuchungen ergeben jedoch, das ein exponentieller Ansatz bzw. die Modellierung mittels fraktionaler Polynome nicht signifikant besser als ein linearer Ansatz für das Alter sind.

Bei Frauen mit Tumor kann dagegen das signifikante fraktionale Polynome $\log(\mathbf{Z}_{Alter})$ (p-Wert: $1.4 \cdot 10^{-2}$) gefunden werden. Durch eine Altersunterteilung wie in Modell (6.5) wird in dieser Gruppe kein besseres Modell erreicht. Dies erscheint plausibel, da der zugehörige Martingalplot (Abb. 6.18, Bild links unten) zeigt, dass nur zwei Beobachtungen mit einem Alter unter 40 Jahren vorliegen.

Die Hazardfunktionen der beiden Modelle lauten also für Behandlungsdauern bis zu 900 Tagen

$$\begin{aligned} \lambda(t) = & \lambda_0(t) \exp[\beta_1 \mathbf{Z}_{Alter} \times \mathbf{Z}_{T,M\ddot{a}dchen} + \beta_2 \mathbf{Z}_{Alter} \times (1 - \mathbf{Z}_{T,w}) + \beta_3 \mathbf{Z}_{Geschlecht} \\ & + \beta_4 \mathbf{Z}_{stat} + \beta_5 \mathbf{Z}_{Stufe2} + \beta_6 \mathbf{Z}_{Stufe3} + \beta_7 \mathbf{Z}_{Tumore} + \beta_8 \mathbf{Z}_{Demenz} + \beta_9 \mathbf{Z}_{Herz} \\ & + \beta_{10} \mathbf{Z}_{Psychose} + \beta_{11} \mathbf{Z}_{Schlag} + \beta_{12} \mathbf{Z}_{Arthrosen} + \beta_{13} \mathbf{Z}_{Lungen} \end{aligned}$$

$$\begin{aligned}
 & + \beta_{14} \mathbf{Z}_{Alter} \times (1 - \mathbf{Z}_{T,w}) \times \mathbf{Z}_{Tumore} + \beta_{15} \mathbf{Z}_{Alter} \times (1 - \mathbf{Z}_{T,w}) \times \mathbf{Z}_{Arthrosen} \\
 & + \beta_{16} \mathbf{Z}_{stat} \times \mathbf{Z}_{Schlag} + \beta_{17} \mathbf{Z}_{stat} \times \mathbf{Z}_{Stufe2} + \beta_{18} \mathbf{Z}_{stat} \times \mathbf{Z}_{Stufe3} \\
 & + \beta_{19} \mathbf{Z}_{Geschlecht} \times \mathbf{Z}_{Tumore} + \beta_{20} \mathbf{Z}_{Stufe2} \times \mathbf{Z}_{Herz} + \beta_{21} \mathbf{Z}_{Stufe3} \times \mathbf{Z}_{Herz} \\
 & + \beta_{22} \mathbf{Z}_{Psychose} \times \mathbf{Z}_{Schlag} + \beta_{23} \mathbf{Z}_{Tumore} \times \mathbf{Z}_{Herz} + \beta_{24} \mathbf{Z}_{Schlag} \times \mathbf{Z}_{Arthrosen}] \quad (6.7)
 \end{aligned}$$

und

$$\begin{aligned}
 \lambda(t) = & \lambda_0(t) \exp[\beta_1 \mathbf{Z}_{Alter} \times (1 - \mathbf{Z}_{T,w}) + \beta_2 \mathbf{Z}_{Geschlecht} + \beta_3 \mathbf{Z}_{stat} + \beta_4 \mathbf{Z}_{Stufe2} \\
 & + \beta_5 \mathbf{Z}_{Stufe3} + \beta_6 \mathbf{Z}_{Tumore} + \beta_7 \mathbf{Z}_{Herz} + \beta_8 \mathbf{Z}_{Psychose} + \beta_9 \mathbf{Z}_{Schlag} \\
 & + \beta_{10} \log(\mathbf{Z}_{Alter}) \times \mathbf{Z}_{T,w} + \beta_{11} \mathbf{Z}_{Alter} \times (1 - \mathbf{Z}_{T,w}) \times \mathbf{Z}_{Stufe2} \\
 & + \beta_{12} \mathbf{Z}_{Alter} \times (1 - \mathbf{Z}_{T,w}) \times \mathbf{Z}_{Stufe3} + \beta_{13} \mathbf{Z}_{Geschlecht} \times \mathbf{Z}_{stat} \\
 & + \beta_{14} \mathbf{Z}_{Geschlecht} \times \mathbf{Z}_{Psychose} + \beta_{15} \mathbf{Z}_{stat} \times \mathbf{Z}_{Stufe2} + \beta_{16} \mathbf{Z}_{stat} \times \mathbf{Z}_{Stufe3} \\
 & + \beta_{17} \mathbf{Z}_{Psychose} \times \mathbf{Z}_{Schlag} + \beta_{18} \mathbf{Z}_{stat} \times \mathbf{Z}_{Psychose} + \beta_{19} \mathbf{Z}_{Stufe2} \times \mathbf{Z}_{Herz} \\
 & + \beta_{20} \mathbf{Z}_{Stufe3} \times \mathbf{Z}_{Herz}] \quad (6.8)
 \end{aligned}$$

für Behandlungszeiten über 900 Tage. Die Koeffizienten der beiden Modelle sind, zusammen mit den p-Werten des Wald-Tests, in den Tabellen 6.29 und 6.30 aufgeführt.

Es fällt auf, dass sich in den Modellen die Sterberate bei der Diagnose Tumor stark unterscheidet. In den ersten 900 Pflegejahren ist das Sterberisiko mit einem Koeffizienten von 3.76 wesentlich höher als in längerer Pflegedauer (Koeffizient: 0.53). Dies ist mit der kurzen Überlebenserwartung von Tumorpatienten erklärbar. Die Koeffizienten zeigen auch, dass das Alter nach 900 Tagen in Pflegestufe 1 die Überlebenszeit stark beeinflusst. Es muß jedoch beachtet werden, dass eine Interpretation allein anhand eines Koeffizientenvergleichs schwierig ist, da viele Wechselwirkungen zwischen den Kovariablen vorliegen, die zusätzlich beachtet werden müssen.

Für diese beiden Modelle wird nun erneut mittels der Routine `zph` ein Test auf die Zeitabhängigkeit durchgeführt. Der globale p-Wert von $1.21 \cdot 10^{-5}$ im Modell (6.7) mit Behandlungsdauer ≤ 900 Tage läßt darauf schließen, dass immer noch Zeitabhängigkeiten vorliegen. Betrachtet man die p-Werte der einzelnen Kovariablen, erkennt man, dass diese Verletzung der „proportional Hazard-Annahme“ auf die Kovariablen \mathbf{Z}_{Stufe2} (p-Wert $4.3 \cdot 10^{-3}$), \mathbf{Z}_{Stufe3} (p-Wert $2.6 \cdot 10^{-3}$) und \mathbf{Z}_{Demenz} (p-Wert $1.6 \cdot 10^{-3}$) zurückzuführen ist. Im Modell (6.8) mit Behandlungsdauern über 900 Tage kann man aufgrund eines globalen p-Wertes von 0.97 davon ausgehen, dass die Annahmen des Proportional Hazard Modells erfüllt sind und keine Zeitabhängigkeiten mehr vorliegen. Durch die Teilung des Datensatzes bei 900 Tagen konnte also zumindest für längere Behandlungsdauern die Zeitabhängigkeit der Effekte eliminiert werden. Die Modelle (6.7) und (6.8) modellieren die Überlebenszeit der Pflegepatienten daher besser als die Modelle (6.5) und (6.6) aus Abschnitt 6.5.4, denen der gesamte Datensatz zugrunde liegt. Die Modellierung der Sterbeintensitäten wird an dieser Stelle abgeschlossen, mit den Modellen (6.7) und (6.8) als Endmodellen.

| Variable | Koeffizient | exp(Koeff.) | SE(Koeff.) | z-Statistik | p-Wert |
|---|-------------|-------------|------------|-------------|----------------------|
| $\mathbf{Z}_{Alter} \times \mathbf{Z}_{T,M\ddot{a}dchen}$ | -1.2786 | 0.2784 | 1.5425 | -0.8290 | $4.1 \cdot 10^{-1}$ |
| $\mathbf{Z}_{Alter} \times (1 - \mathbf{Z}_{T,w})$ | 0.0299 | 1.030 | 0.0042 | 7.110 | $1.2 \cdot 10^{-12}$ |
| $\mathbf{Z}_{Geschlecht}$ | -0.4028 | 0.6685 | 0.0785 | -5.1303 | $2.9 \cdot 10^{-7}$ |
| \mathbf{Z}_{stat} | 0.4487 | 1.5663 | 0.1693 | 2.6508 | $8.0 \cdot 10^{-3}$ |
| \mathbf{Z}_{Stufe2} | 0.7268 | 2.0685 | 0.1232 | 5.8987 | $3.7 \cdot 10^{-9}$ |
| \mathbf{Z}_{Stufe3} | 1.9681 | 7.1570 | 0.1181 | 16.6641 | $< 10^{-16}$ |
| \mathbf{Z}_{Tumore} | 3.7635 | 43.0974 | 0.3757 | 10.0177 | $< 10^{-16}$ |
| \mathbf{Z}_{Demenz} | -0.1718 | 0.8421 | 0.0684 | -2.5129 | $1.2 \cdot 10^{-2}$ |
| \mathbf{Z}_{Herz} | 0.4227 | 1.5261 | 0.1404 | 3.010 | $2.6 \cdot 10^{-3}$ |
| $\mathbf{Z}_{Psychose}$ | -0.3272 | 0.7209 | 0.0878 | -3.7257 | $1.9 \cdot 10^{-4}$ |
| \mathbf{Z}_{Schlag} | -0.6296 | 0.5328 | 0.1116 | -5.6403 | $1.7 \cdot 10^{-8}$ |
| $\mathbf{Z}_{Arthrosen}$ | -3.0331 | 0.0482 | 1.2687 | -2.3906 | $1.7 \cdot 10^{-2}$ |
| \mathbf{Z}_{Lungen} | 0.5235 | 1.6880 | 0.1794 | 2.9188 | $3.5 \cdot 10^{-3}$ |
| $\mathbf{Z}_{Alter} \times (1 - \mathbf{Z}_{T,w})$ $\times \mathbf{Z}_{Tumore}$ | -0.0329 | 0.9676 | 0.0046 | -7.1408 | $9.3 \cdot 10^{-13}$ |
| $\mathbf{Z}_{Alter} \times (1 - \mathbf{Z}_{T,w})$ $\times \mathbf{Z}_{Arthrosen}$ | 0.0326 | 1.0332 | 0.0143 | 2.2789 | $2.3 \cdot 10^{-2}$ |
| $\mathbf{Z}_{stat} \times \mathbf{Z}_{Schlag}$ | 0.4344 | 1.5440 | 0.1712 | 2.5377 | $1.1 \cdot 10^{-2}$ |
| $\mathbf{Z}_{stat} \times \mathbf{Z}_{Stufe2}$ | -0.2167 | 0.8052 | 0.1997 | -1.0853 | $2.8 \cdot 10^{-1}$ |
| $\mathbf{Z}_{stat} \times \mathbf{Z}_{Stufe3}$ | -0.6523 | 0.5208 | 0.1960 | -3.3288 | $8.7 \cdot 10^{-4}$ |
| $\mathbf{Z}_{Geschlecht} \times \mathbf{Z}_{Tumore}$ | 0.4228 | 1.5263 | 0.1416 | 2.9861 | $2.8 \cdot 10^{-3}$ |
| $\mathbf{Z}_{Stufe2} \times \mathbf{Z}_{Herz}$ | 0.0099 | 1.0099 | 0.1664 | 0.0593 | $9.5 \cdot 10^{-1}$ |
| $\mathbf{Z}_{Stufe3} \times \mathbf{Z}_{Herz}$ | -0.4574 | 0.6329 | 0.1677 | -2.7271 | $6.4 \cdot 10^{-3}$ |
| $\mathbf{Z}_{Psychose} \times \mathbf{Z}_{Schlag}$ | 0.6842 | 1.9821 | 0.1826 | 3.747 | $1.8 \cdot 10^{-4}$ |
| $\mathbf{Z}_{Tumore} \times \mathbf{Z}_{Herz}$ | -0.6597 | 0.5170 | 0.1643 | -4.0167 | $5.9 \cdot 10^{-5}$ |
| $\mathbf{Z}_{Schlag} \times \mathbf{Z}_{Arthrosen}$ | 0.7320 | 2.0792 | 0.3150 | 2.3237 | $2.0 \cdot 10^{-2}$ |

Tabelle 6.29: Koeffizienten und p-Werte der Kovariablen im Modell (6.7) mit Behandlungsdauer ≤ 900 Tage

6.6.1 Illustration: Modellierung der Zeitabhängigkeit mittels fraktionaler Polynome

Die Effekte der Behandlungsstufe im Modell (6.7) mit Behandlungsdauern bis zu 900 Tagen weisen immer noch eine Abhängigkeit von der Zeit auf. In Kapitel (4.9.3)) wurde eine Methode vorgestellt, die Zeitabhängigkeit von Kovariablen mittels fraktionaler Polynome zu modellieren. Wie in Kapitel (4.9.3) erwähnt, ist dieses Verfahren jedoch nur für kleine Datensätze geeignet, da sich die Anzahl der Daten durch die notwendige Restrukturierung vervielfacht. Mit 3426 Daten ist der Datensatz mit Behandlungsdauer ≤ 900 Tage zu groß und eine Anwendung nicht möglich.

Zur Illustration der Methode, soll sie auf einen kleineren Unterdatensatz angewendet werden. Der Datensatz, der nur Männer mit einer Behandlungsdauer bis zu 300 Tagen

| Variable | Koeffizient | exp(Koeff.) | SE(Koeff.) | z-Statistik | p-Wert |
|--|-------------|-------------|------------|-------------|----------------------|
| $\mathbf{Z}_{Alter} \times (1 - \mathbf{Z}_{T,w})$ | 0.05262 | 1.054 | 0.0084 | 6.294 | $3.1 \cdot 10^{-10}$ |
| $\mathbf{Z}_{Geschlecht}$ | -0.1743 | 0.840 | 0.1034 | -1.686 | $9.2 \cdot 10^{-2}$ |
| \mathbf{Z}_{stat} | 0.7074 | 2.029 | 0.2414 | 2.931 | $3.4 \cdot 10^{-3}$ |
| \mathbf{Z}_{Stufe2} | 2.9296 | 18.721 | 0.7028 | 4.169 | $3.1 \cdot 10^{-5}$ |
| \mathbf{Z}_{Stufe3} | 2.1129 | 8.272 | 0.7280 | 2.902 | 3.710^{-3} |
| \mathbf{Z}_{Tumore} | 0.5299 | 1.699 | 0.1153 | 4.595 | $4.3 \cdot 10^{-6}$ |
| \mathbf{Z}_{Herz} | 0.09547 | 1.100 | 0.2020 | 0.472 | $6.4 \cdot 10^{-1}$ |
| $\mathbf{Z}_{Psychose}$ | -0.1450 | 0.865 | 0.1340 | -01.082 | $2.8 \cdot 10^{-1}$ |
| \mathbf{Z}_{Schlag} | -0.1180 | 0.889 | 0.0937 | -1.260 | $2.1 \cdot 10^{-1}$ |
| $\log(\mathbf{Z}_{Alter}) \times \mathbf{Z}_{T,w}$ | 0.8692 | 2.385 | 0.0940 | 9.248 | $< 10^{-16}$ |
| $\mathbf{Z}_{Alter} \times (1 - \mathbf{Z}_{T,w})$ $\times \mathbf{Z}_{Stufe2}$ | -0.0279 | 0.972 | 0.0086 | -3.249 | $1.2 \cdot 10^{-3}$ |
| $\mathbf{Z}_{Alter} \times (1 - \mathbf{Z}_{T,w})$ $\times \mathbf{Z}_{Stufe3}$ | -0.0065 | 0.994 | 0.0088 | -0.734 | $4.6 \cdot 10^{-1}$ |
| $\mathbf{Z}_{Geschlecht} \times \mathbf{Z}_{stat}$ | -0.5805 | 0.560 | 0.1453 | -3.996 | $6.4 \cdot 10^{-5}$ |
| $\mathbf{Z}_{Geschlecht} \times \mathbf{Z}_{Psychose}$ | -0.3709 | 0.690 | 0.1581 | -2.346 | $1.9 \cdot 10^{-2}$ |
| $\mathbf{Z}_{stat} \times \mathbf{Z}_{Stufe2}$ | -0.2053 | 0.814 | 0.2430 | -0.845 | $4.0 \cdot 10^{-1}$ |
| $\mathbf{Z}_{stat} \times \mathbf{Z}_{Stufe3}$ | -0.5262 | 0.591 | 0.2345 | -2.244 | $2.5 \cdot 10^{-2}$ |
| $\mathbf{Z}_{Psychose} \times \mathbf{Z}_{Schlag}$ | 0.4240 | 1.528 | 0.1761 | 2.408 | $1.6 \cdot 10^{-2}$ |
| $\mathbf{Z}_{stat} \times \mathbf{Z}_{Psychose}$ | 0.41811 | 1.519 | 0.1586 | 2.637 | $8.4 \cdot 10^{-3}$ |
| $\mathbf{Z}_{Stufe2} \times \mathbf{Z}_{Herz}$ | 0.2615 | 1.299 | 0.2302 | 1.136 | $2.6 \cdot 10^{-1}$ |
| $\mathbf{Z}_{Stufe3} \times \mathbf{Z}_{Herz}$ | -0.2079 | 0.812 | 0.2218 | -0.937 | $3.5 \cdot 10^{-1}$ |

Tabelle 6.30: Koeffizienten und p-Werte der Kovariablen im Modell (6.8) mit Behandlungsdauer > 900 Tage

beinhaltet, ist mit 431 Daten ein geeignetes Beispiel. Durch Restrukturierung wird die Zahl der Daten auf 3559 erhöht.

Es muss betont werden, dass dieser Datensatz nicht repräsentativ ist. Er beinhaltet nur Männer, die innerhalb der ersten 300 Pflage tage in einen anderen Pflegezustand übergingen, zensiert wurden oder starben. Korrekterweise hätten die Daten aller Männer, die länger als 300 Tage im gleichen Zustand verweilten, zum Zeitpunkt 300 Tage in je zwei Datensätze aufgesplittet werden müssen. Dies hätte die Anzahl der Daten jedoch wieder stark erhöht und die Durchführung der Methode erschwert.

Das Modell des vorangegangenen Abschnitts für Behandlungsdauern bis zu 900 Tagen (ohne $\mathbf{Z}_{Geschlecht}$) wird nun auf den restrukturierten Datensatz angewandt. Der Test auf Zeitabhängigkeit der Routine zph ergibt für die Kovariablen \mathbf{Z}_{Stufe2} und \mathbf{Z}_{Stufe3} einen p-Wert von 0.03 bzw. $8.8 \cdot 10^{-4}$. Die p-Werte aller anderen Kovariablen liegen über 5 %. Es genügt daher, den Effekt der Pflegestufe zeitabhängig zu modellieren. Dazu wird das Modell um die Interaktionen

$$\mathbf{t}^{p1} \times (\mathbf{Z}_{Stufe2} + \mathbf{Z}_{Stufe3})$$

bzw.

$$\mathbf{t}^{p_1} \times (\mathbf{Z}_{Stufe2} + \mathbf{Z}_{Stufe3}) + \mathbf{t}^{p_2} \times (\mathbf{Z}_{Stufe2} + \mathbf{Z}_{Stufe3})$$

erweitert und für verschiedene Werte $p_1, p_2 \in \varnothing = \{-2, -1, -0.5, 0, 0.5, 1, 2\}$ gefittet. Die Ergebnisse der besten gefundenen fraktionalen Polynome ersten und zweiten Grades sind in Tabelle 6.31 zusammengefasst. Beide fraktionalen Polynome sind signifikant, aufgrund der geringeren Anzahl von Freiheitsgraden ist jedoch das Polynom ersten Grades vorzuziehen. Die Pflegestufe wird demnach am besten modelliert durch

$$\mathbf{Z}_{Stufe2} + \mathbf{Z}_{Stufe3} + \mathbf{t}^{-0.5} \times (\mathbf{Z}_{Stufe2} + \mathbf{Z}_{Stufe3}).$$

Wendet man die Routine `zph` nun auf dieses Modell, so liefert der Test auf Zeitabhängigkeit einen globalen p-Wert von 0.82. Auch die p-Werte für \mathbf{Z}_{Stufe2} und \mathbf{Z}_{Stufe3} einzeln bestätigen mit den Werten 0.99 und 0.78, dass in diesem Modell die Proportional Hazard Annahme schließlich erfüllt ist. Die beiden oberen Plots in Abb. 6.19 zeigen für die Kovariablen \mathbf{Z}_{Stufe2} und \mathbf{Z}_{Stufe3} die mittels der skalierten Schoenfeldresiduen aufgedeckte Art der Zeitabhängigkeit im Modell, in dem noch keine Modellierung durch fraktionale Polynome vorgenommen wurde. Darunter ist jeweils das gefundene zugehörige fraktionale Polynom ersten Grades abgebildet. Bei der Kovariable \mathbf{Z}_{Stufe3} ist die Ähnlichkeit zum Plot der Schoenfeldresiduen sehr gut erkennbar, bei \mathbf{Z}_{Stufe2} gilt dies erst ab etwa 100 Tagen. Dies ist darauf zurückzuführen, dass die Pflegestufe als ein Faktor in das Modell eingeht. Die Zeitabhängigkeit muss daher sowohl für \mathbf{Z}_{Stufe3} als auch für \mathbf{Z}_{Stufe2} durch dasselbe fraktionale Polynom modelliert werden.

| | Log-Likelihood | df | p-Wert | Potenzen |
|---------|----------------|----|---------------------|----------|
| ohne FP | -719.5906 | | | |
| m=1 | -711.9711 | 3 | $1.6 \cdot 10^{-3}$ | -0.5 |
| m=2 | -711.4079 | 6 | $1.2 \cdot 10^{-2}$ | 1,1 |

Tabelle 6.31: Optimale fraktionale Polynome zur Modellierung der Zeitabhängigkeit der Pflegestufe

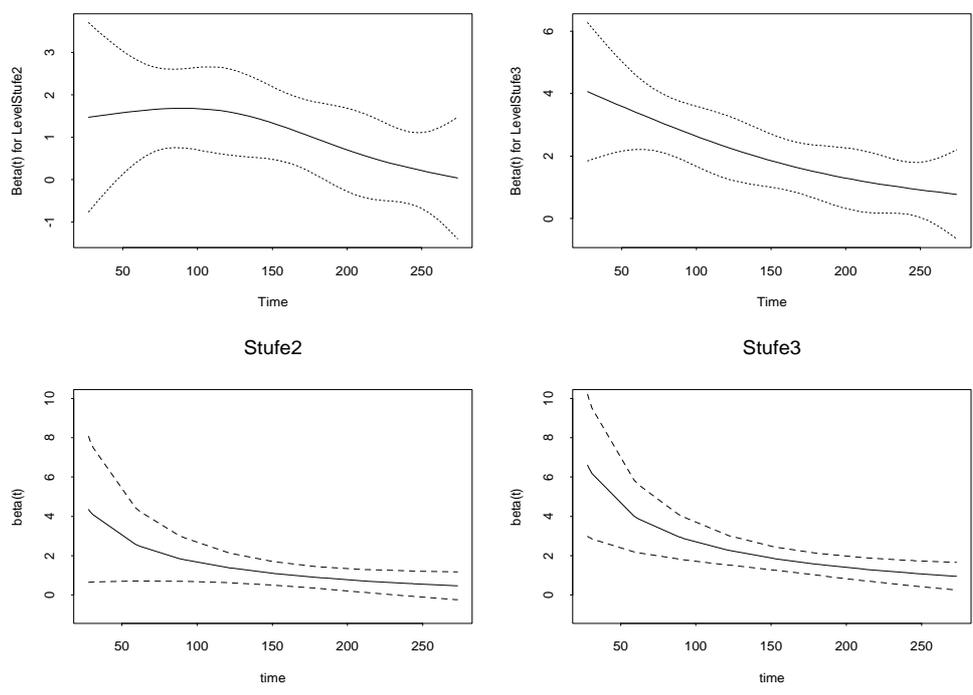


Abbildung 6.19: Plot der skalierten Schoenfeldresiduen (erste Reihe) und gefundene fraktionale Polynome (zweite Reihe) für die Kovariablen \mathbf{Z}_{Stufe2} und \mathbf{Z}_{Stufe3}

Kapitel 7

Übergangsmodelle in der Pflegeversicherung

Bisher wurde nur auf die Schätzung von Sterbeintensitäten eingegangen. In der Pflegeversicherung sind jedoch auch andere Zustandsübergänge von Interesse. Die Pflegebedürftigen können sowohl in eine andere Behandlungsart als auch eine andere Pflegestufe wechseln. Auch diese Bestandsänderungen können mittels eines Proportional Hazard Modells untersucht werden.

7.1 Modellierung der Übergänge zwischen den Behandlungsarten

In diesem Abschnitt werden die Übergänge von stationärer zu ambulanter Pflege und umgekehrt betrachtet. In Abb. 7.1 ist ein solches Modell mit drei möglichen Zuständen (Zustand 1: ambulante Pflege, Zustand 2: stationäre Pflege, Zustand 3: Tod) schematisch dargestellt. Die Sterbeintensitäten bei ambulanter (λ_{13}) bzw. stationärer Pflege (λ_{23}) wurden im vorangegangenen Abschnitt untersucht. Für Behandlungsdauern bis zu 900 Tagen wurden sie am besten durch Modell (6.7), für Behandlungsdauern über 900 Tagen durch Modell (6.8) modelliert.

Die Intensitäten λ_{12} (Übergang von ambulant nach stationär) und λ_{21} (Übergang von stationär nach ambulant) können nun wieder über ein Proportional Hazard Modell geschätzt werden. Sie lassen sich daher darstellen in der Form

$$\lambda_{12}(t, \mathbf{Z}) = \lambda_{012}(t) \exp[\boldsymbol{\beta}'\mathbf{Z}(t)]$$

bzw.

$$\lambda_{21}(t, \mathbf{Z}) = \lambda_{021}(t) \exp[\boldsymbol{\beta}'\mathbf{Z}(t)],$$

wobei in den Kovariablenvektor $\mathbf{Z}(t)$ die Variablen Alter, Geschlecht, Pflegestufe und Diagnosen eingehen. Hier ist zu beachten, dass in die Modellbildung jeweils nur ein Teildatensatz eingeht. Bei der Modellierung des Überganges von Zustand i nach Zustand j , $i, j = 1, 2$ werden nur Beobachtungen berücksichtigt, die sich im Ausgangszustand i

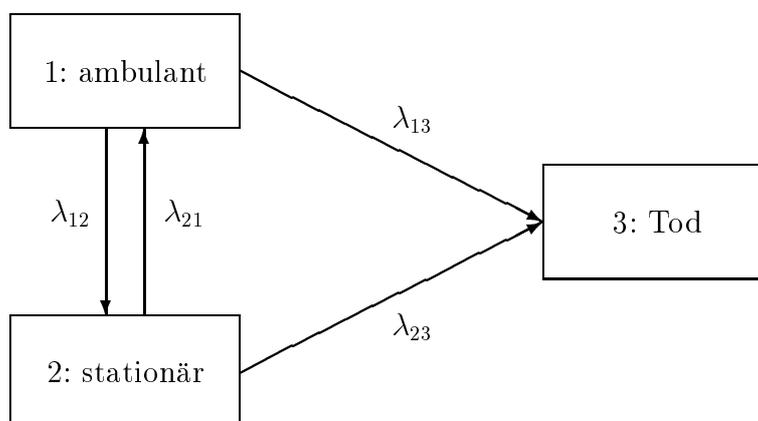


Abbildung 7.1: Übersicht der Zustandsübergänge bei Berücksichtigung der Behandlungsart

befinden. Auch Patienten, die sich zwar im Zustand i befinden, dann aber sterben, gehen nicht mit in die untersuchten Daten ein. Somit wird verhindert, dass Beobachtungen, von denen bekannt ist, dass sie sterben, als zensiert betrachtet werden.

Nach einer Modellbildungsphase ergibt sich für den Wechsel von ambulanter in stationäre Behandlung folgende Hazardfunktion

$$\begin{aligned}
 \lambda_{12}(t) = & \lambda_{012}(t) \exp[\beta_1 \mathbf{Z}_{Alter}(t) + \beta_2 \mathbf{Z}_{Geschlecht} + \beta_3 \mathbf{Z}_{Stufe2}(t) + \beta_4 \mathbf{Z}_{Stufe3}(t) \\
 & + \beta_5 \mathbf{Z}_{Demenz} + \beta_6 \mathbf{Z}_{Psychose} + \beta_7 \mathbf{Z}_{Herz} + \beta_8 \mathbf{Z}_{Blindheit} \\
 & + \beta_9 \mathbf{Z}_{Alter}(t) \times \mathbf{Z}_{Demenz}]
 \end{aligned} \tag{7.1}$$

Die zugehörigen geschätzten Koeffizienten sind in Tabelle 7.1 aufgeführt. Für den Übergang von stationärer zu ambulanter Behandlung lautet die Hazardfunktion

$$\lambda_{21}(t) = \lambda_{021}(t) \exp[\beta_1 \mathbf{Z}_{Alter} + \beta_2 \mathbf{Z}_{Geschlecht} + \beta_3 \mathbf{Z}_{Stufe2} + \beta_4 \mathbf{Z}_{Stufe3}] \tag{7.2}$$

Aus Tabelle 7.2 erkennt man, dass weder das Alter, das Geschlecht noch die Pflegestufe signifikanten Einfluß auf diesen Übergang besitzen. Dasselbe gilt für die Diagnosen, die nicht in der Tabelle aufgeführt sind. Wie schon in Abschnitt 6.1, Tabelle 6.2, festgestellt, liegen zu diesem Zustandsübergang kaum Daten vor, in der Regel sind fast nur Verschlechterungen des Zustands zu verzeichnen.

| Variable | Koeffizienten | exp(Koeff.) | SE(Koeff.) | z-Statistik | p-Wert |
|---|---------------|-------------|------------|-------------|----------------------|
| \mathbf{Z}_{Alter} | 0.0485 | 1.050 | 0.0068 | 7.10 | $1.2 \cdot 10^{-12}$ |
| $\mathbf{Z}_{Geschlecht}$ | 0.3309 | 1.392 | 0.999 | 3.31 | $9.3 \cdot 10^{-4}$ |
| \mathbf{Z}_{Stufe2} | 0.4249 | 1.529 | 0.0950 | 4.47 | $7.7 \cdot 10^{-6}$ |
| \mathbf{Z}_{Stufe3} | 0.7198 | 2.054 | 0.1365 | 5.27 | $1.3 \cdot 10^{-7}$ |
| \mathbf{Z}_{Demenz} | 3.0484 | 21.082 | 0.7638 | 3.99 | $6.6 \cdot 10^{-5}$ |
| $\mathbf{Z}_{Psychose}$ | 0.2511 | 1.285 | 0.1001 | 2.51 | $1.2 \cdot 10^{-2}$ |
| \mathbf{Z}_{Herz} | 0.1750 | 1.191 | 0.0876 | 2.00 | $4.6 \cdot 10^{-2}$ |
| $\mathbf{Z}_{Blindheit}$ | -0.5436 | 0.581 | 0.2932 | -1.85 | $6.4 \cdot 10^{-2}$ |
| $\mathbf{Z}_{Alter} \times \mathbf{Z}_{Demenz}$ | -0.027 | 0.973 | 0.0091 | -2.96 | $3.1 \cdot 10^{-3}$ |

Tabelle 7.1: Geschätzte Koeffizienten im Modell (7.1) zur Modellierung des Übergangs von ambulanter zu stationärer Pflege

| Variable | Koeffizienten | exp(Koeff.) | SE(Koeff.) | z-Statistik | p-Wert |
|---------------------------|---------------|-------------|------------|-------------|--------|
| \mathbf{Z}_{Alter} | 0.0009 | 1.001 | 0.0177 | 0.0525 | 0.960 |
| $\mathbf{Z}_{Geschlecht}$ | -0.7068 | 0.493 | 0.4547 | -1.5544 | 0.120 |
| \mathbf{Z}_{Stufe2} | -0.8472 | 0.429 | 0.4939 | -1.7153 | 0.086 |
| \mathbf{Z}_{Stufe3} | -0.4787 | 0.620 | 0.4913 | -0.9744 | 0.33 |

Tabelle 7.2: Geschätzte Koeffizienten im Modell (7.2) zur Modellierung des Übergangs von stationärer zu ambulanter Pflege

7.2 Modellierung der Übergänge zwischen den Behandlungsstufen

Analog zu Abschnitt 7.1 werden nun die Übergänge zwischen den einzelnen Pflegestufen untersucht. Dabei gehen wir von einem Modell mit vier Zuständen aus (siehe Abb. 7.2): Stufe 1, Stufe 2, Stufe 3, Tod. Schätzungen für die Sterbeintensitäten λ_{14} , λ_{24} und λ_{34} liegen bereits aus den Modellen (6.7) und (6.8) vor, die Intensitäten λ_{12} , λ_{13} und λ_{23} modelliert man durch ein Proportional Hazard Modell, in das die Variablen Alter, Geschlecht, Behandlungsart und die pflegeauslösenden Diagnosen eingehen. Aufgrund der geringen Anzahl von Beobachtungen, deren Zustand sich verbesserte (siehe Abschnitt 6.1, Tabelle 6.2), werden die Übergangintensitäten von Stufe 2 und 3 nach Stufe 1 bzw. von Stufe 3 nach Stufe 2 nicht modelliert und vereinfachend mit dem Wert 0 angenommen. Für die Stufenübergänge wurden folgende Hazardfunktionen ermittelt:

$$\lambda_{12}(t) = \lambda_{012}(t) \exp[\beta_1 \mathbf{Z}_{Alter} + \beta_2 \mathbf{Z}_{stat} + \beta_3 \mathbf{Z}_{Demenz} + \beta_4 \mathbf{Z}_{Tumore} + \beta_5 \mathbf{Z}_{Alter} \times \mathbf{Z}_{stat}] \quad (7.3)$$

$$\lambda_{13}(t) = \lambda_{013}(t) \exp[\beta_1 \mathbf{Z}_{Alter} + \beta_2 \mathbf{Z}_{Geschlecht} + \beta_3 \mathbf{Z}_{Demenz} + \beta_4 \mathbf{Z}_{Tumore} + \beta_5 \mathbf{Z}_{Psychose}] \quad (7.4)$$

$$\lambda_{23}(t) = \lambda_{013}(t) \exp[\beta_1 \mathbf{Z}_{Alter} + \beta_2 \mathbf{Z}_{Geschlecht} + \beta_3 \mathbf{Z}_{stat} + \beta_4 \mathbf{Z}_{Schlag} + \beta_5 \mathbf{Z}_{Demenz}]$$

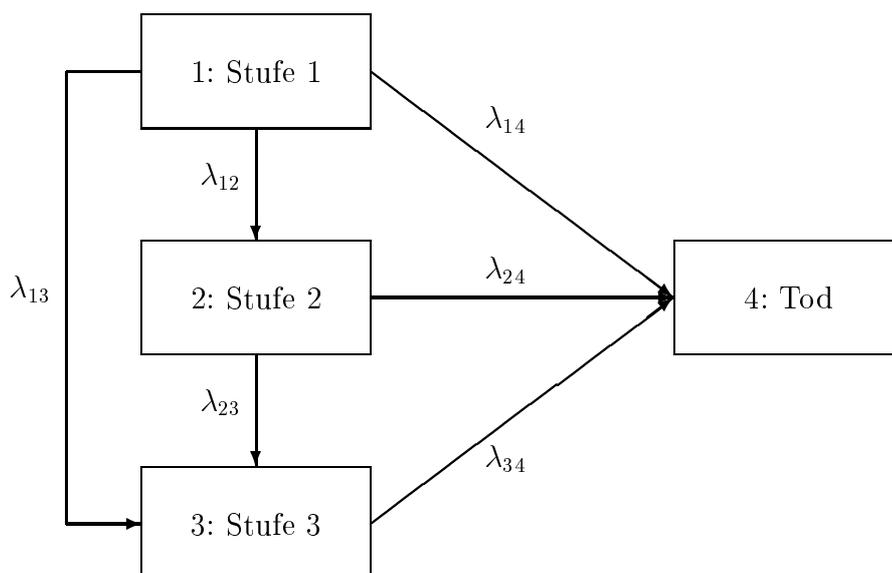


Abbildung 7.2: Übersicht der Zustandsübergänge bei Berücksichtigung der Pflegestufe

$$\begin{aligned}
 & + \beta_6 \mathbf{Z}_{Tumore} + \beta_7 \mathbf{Z}_{Herz} + \beta_8 \mathbf{Z}_{Psychose} + \beta_9 \mathbf{Z}_{Alter} \times \mathbf{Z}_{Schlag} \\
 & + \beta_{10} \mathbf{Z}_{stat} \times \mathbf{Z}_{Tumore} + \beta_{11} \mathbf{Z}_{Tumore} \times \mathbf{Z}_{Demenz} + \beta_{12} \mathbf{Z}_{Tumore} \times \mathbf{Z}_{Herz} \\
 & + \beta_{13} \mathbf{Z}_{Demenz} \times \mathbf{Z}_{Schlag} \quad (7.5)
 \end{aligned}$$

Die Tabellen 7.3 - 7.5 enthalten die dazugehörigen geschätzten Koeffizienten. Wie in Abschnitt 7.1 gingen auch hier in die Modellierung der Übergänge von Zustand i nach Zustand j ($i=1,2, j=2,3$) nur die Daten von Beobachtungen, die sich in Zustand i befinden und nicht sterben, ein.

| Variable | Koeffizienten | exp(Koeff.) | SE(Koeff.) | z-Statistik | p-Wert |
|---|---------------|-------------|------------|-------------|---------------------|
| \mathbf{Z}_{Alter} | 0.0159 | 1.016 | 0.0037 | 4.34 | $1.4 \cdot 10^{-5}$ |
| \mathbf{Z}_{stat} | -2.4192 | 0.089 | 1.1254 | -2.15 | $3.2 \cdot 10^{-2}$ |
| \mathbf{Z}_{Demenz} | 0.4761 | 1.610 | 0.0861 | 5.53 | $3.2 \cdot 10^{-8}$ |
| \mathbf{Z}_{Tumore} | 0.4307 | 1.538 | 0.1313 | 3.28 | $1.0 \cdot 10^{-3}$ |
| $\mathbf{Z}_{Alter} \times \mathbf{Z}_{stat}$ | 0.0287 | 1.029 | 0.0131 | 2.20 | $2.8 \cdot 10^{-2}$ |

Tabelle 7.3: Geschätzte Koeffizienten im Modell (7.3) zur Modellierung des Übergangs λ_{12} von Stufe 1 zu Stufe 2

| Variable | Koeffizienten | exp(Koeff.) | SE(Koeff.) | z-Statistik | p-Wert |
|---------------------------|---------------|-------------|------------|-------------|---------------------|
| \mathbf{Z}_{Alter} | 0.0383 | 1.039 | 0.0094 | 4.08 | $4.6 \cdot 10^{-5}$ |
| $\mathbf{Z}_{Geschlecht}$ | -0.3990 | 0.671 | 0.1910 | -2.09 | $3.7 \cdot 10^{-2}$ |
| \mathbf{Z}_{Demenz} | 0.5162 | 1.676 | 0.1805 | 2.86 | $4.2 \cdot 10^{-3}$ |
| \mathbf{Z}_{Tumore} | 0.8409 | 2.318 | 0.2439 | 3.45 | $5.7 \cdot 10^{-4}$ |
| $\mathbf{Z}_{Psychose}$ | 0.5549 | 1.742 | 0.1976 | 2.81 | $5.0 \cdot 10^{-3}$ |

Tabelle 7.4: Geschätzte Koeffizienten im Modell (7.4) zur Modellierung des Übergangs λ_{13} von Stufe 1 zu Stufe 3

| Variable | Koeffizienten | exp(Koeff.) | SE(Koeff.) | z-Statistik | p-Wert |
|--|---------------|-------------|------------|-------------|-----------------------|
| \mathbf{Z}_{Alter} | 0.0068 | 1.007 | 0.0039 | 1.723 | $8.5 \cdot 10^{-2}$ |
| $\mathbf{Z}_{Geschlecht}$ | -0.2029 | 0.816 | 0.1060 | -1.914 | $5.6 \cdot 10^{-2}$ |
| \mathbf{Z}_{stat} | -0.0599 | 0.942 | 0.1215 | -0.493 | $6.2 \cdot 10^{-1}$ |
| \mathbf{Z}_{Schlag} | -1.2903 | 0.275 | 0.7851 | -1.644 | $1 - 0 \cdot 10^{-1}$ |
| \mathbf{Z}_{Demenz} | 0.9219 | 2.514 | 0.1357 | 6.792 | $1.1 \cdot 10^{-11}$ |
| \mathbf{Z}_{Tumore} | 2.0857 | 8.050 | 0.1796 | 11.615 | $< 10^{-16}$ |
| \mathbf{Z}_{Herz} | 0.3566 | 1.428 | 0.1094 | 3.259 | $1.1 \cdot 10^{-3}$ |
| $\mathbf{Z}_{Psychose}$ | 0.2756 | 1.317 | 0.1115 | 2.472 | $1.3 \cdot 10^{-2}$ |
| $\mathbf{Z}_{Alter} \times \mathbf{Z}_{Schlag}$ | 0.0239 | 1.024 | 0.0099 | 2.425 | $1.5 \cdot 10^{-2}$ |
| $\mathbf{Z}_{stat} \times \mathbf{Z}_{Tumore}$ | -0.9724 | 0.378 | 0.4847 | -2.006 | $4.5 \cdot 10^{-2}$ |
| $\mathbf{Z}_{Tumore} \times \mathbf{Z}_{Demenz}$ | -0.8163 | 0.442 | 0.2964 | -2.754 | $5.9 \cdot 10^{-3}$ |
| $\mathbf{Z}_{Tumore} \times \mathbf{Z}_{Herz}$ | -0.8778 | 0.416 | 0.3117 | -2.816 | $4.9 \cdot 10^{-3}$ |
| $\mathbf{Z}_{Demenz} \times \mathbf{Z}_{Schlag}$ | -0.6395 | 0.528 | 0.2326 | -2.750 | $6.0 \cdot 10^{-3}$ |

Tabelle 7.5: Geschätzte Koeffizienten im Modell (7.5) zur Modellierung des Übergangs λ_{23} von Stufe 2 zu Stufe 3

7.3 Berechnung von Übergangswahrscheinlichkeiten

In den vorangegangenen Abschnitten wurde auf die Ermittlung der Übergangsintensitäten eingegangen. Versicherungen benötigen zur Tarif- und Prämienkalkulation jedoch Wahrscheinlichkeiten für die verschiedenen Übergänge, in Abhängigkeit von Geschlecht, Alter und Pflegedauer der Patienten. Diese können aus den Intensitäten berechnet werden.

Wir beschränken uns auf die Betrachtung der Übergangswahrscheinlichkeiten des in Abb. 7.2 dargestellten Modells.

Die einjährigen Übergangswahrscheinlichkeiten von Zustand i ($i=1,2,3$) nach Zustand j ($j=2,3,4$) mit Pflegedauer d (in Jahren) für einen Patienten mit Kovariablenvektor \mathbf{Z} sind durch

$$p_{ij}(\mathbf{Z}, d) = \sum_{d \leq t_k < d+1} \left\{ \lambda_{ij}(t_k, \mathbf{Z}) \prod_{d \leq t_l < t_k} (1 - \lambda_{ij}(t_l, \mathbf{Z})) \right\} \quad (7.6)$$

gegeben. Mit dem Produkt-Limit-Schätzer (siehe Abschnitt 3.5.2) lässt sich das Produkt in (7.6) für $j=4$ (Übergang zu Tod) darstellen als

$$\prod_{d \leq t_l < t_k} (1 - \lambda_{i4}(t_l, \mathbf{Z})) = \frac{S(t_k, \text{Stufe} = i)}{S(d, \text{Stufe} = i)} = P(d \leq T < t_k | T \geq d, \text{Stufe} = i) \quad , \quad (7.7)$$

die Überlebenswahrscheinlichkeit für einen Patienten in Pflegestufe i bis zum Zeitpunkt t_k , gegeben Überleben bis zum Zeitpunkt d . Analog lässt sich das Produkt in (7.6) für $j=2,3$ deuten als die Wahrscheinlichkeit zum Zeitpunkt t_k in Zustand j überzugehen, gegeben der Patient befindet sich mindestens seit Zeitpunkt d in Zustand i . (7.6) kann damit anschaulich folgendermaßen erklärt werden: die Übergangswahrscheinlichkeit von i nach j zum Zeitpunkt t setzt sich zusammen aus der Wahrscheinlichkeit, bis zur Zeit t im Zustand i zu verweilen, gegeben man befindet sich zu Beginn des Jahres in diesem Zustand, siehe (7.7), und der Intensität, zum Zeitpunkt t von Zustand i nach Zustand j zu wechseln. Da man an der Übergangswahrscheinlichkeit für ein Jahr interessiert ist, werden die Wahrscheinlichkeiten zu jedem Ereigniszeitpunkt innerhalb dieses Jahres addiert.

Die einjährigen Übergangswahrscheinlichkeiten, in Abhängigkeit der von den Versicherungen benötigten Kovariablen Alter x , Geschlecht s und Pflegedauer d , lassen sich also folgendermaßen schätzen:

$$\hat{p}_{ij}(x, s, d) = \sum_{d \leq t_k < d+1} \left\{ \hat{\lambda}_{ij}(t_k, \mathbf{Z}_{\text{Alter}}(t_k) = x, \mathbf{Z}_{\text{Geschlecht}} = s) \prod_{d \leq t_l < t_k} (1 - \hat{\lambda}_{ij}(t_l, \mathbf{Z}_{\text{Alter}}(t_l) = x, \mathbf{Z}_{\text{Geschlecht}} = s)) \right\} \quad , \quad (7.8)$$

für $i = 1, 2, 3$ und $j = 2, 3, 4$. Die $\hat{\lambda}_{ij}(t, \mathbf{Z})$ sind dabei die geschätzten Übergangsintensitäten von Zustand i nach Zustand j . Die Schätzungen der $\hat{\lambda}_{ij}(t, \mathbf{Z})$ beruhen auf den Modellen (6.7) und (6.8), in denen auch die Behandlungsart, die pflegeauslösenden Diagnosen und Interaktionen dieser Kovariablen enthalten sind. Diese werden von den Versicherungen jedoch nicht bei der Tarifierung berücksichtigt, so dass die Übergangswahrscheinlichkeiten nur in Abhängigkeit des Alters, des Geschlechts und der Pflegedauer der Versicherungsnehmer gesetzt werden. Um die in (7.8) benötigten $\hat{\lambda}_{ij}(t_l, \mathbf{Z}_{\text{Alter}}(t_l) = x, \mathbf{Z}_{\text{Geschlecht}} = s)$

zu erhalten, müssen deshalb die Intensitäten über die Diagnosen und die Behandlungsart gewichtet und gemittelt werden:

$$\hat{p}_{i,j}(x, s, d) = \frac{1}{n_{x,s,d,amb}^{ij} + n_{x,s,d,stat}^{ij}} \left\{ \sum_{k=1}^{22} [n_{x,s,d,amb,k}^{ij} \hat{p}_{i,j}(x, s, d, \text{ambulant, Diagnose } k) + n_{x,s,d,stat,k}^{ij} \hat{p}_{i,j}(x, s, d, \text{stationär, Diagnose } k)] \right\} \quad (7.9)$$

$n_{x,s,d,amb}^{ij}$ und $n_{x,s,d,stat}^{ij}$ geben dabei die Gesamtanzahl der Beobachtungen in ambulanter bzw. stationärer Behandlung in der Gruppe der Patienten mit Geschlecht s , Pflegestufe i , Pflegedauer d und einem Alter zwischen $x-5$ und $x+5$ Jahren ($x=60,70,80,90$) an. Für $j=2,3$, d.h. für die Übergänge nach Stufe 2 bzw. Stufe 3, sind Patienten, bei denen der Tod beobachtet wurde, nicht in diesen Zahlen enthalten, da diese Beobachtungen auch nicht bei der Modellierung der Stufenübergänge berücksichtigt wurden (siehe Abschnitt 7.2). Für die Sterbewahrscheinlichkeiten ($j=4$) werden $n_{x,s,d,amb}^{i4}$ und $n_{x,s,d,stat}^{i4}$ dagegen auf Grundlage des gesamten Datensatzes bestimmt. Mit $n_{x,s,d,amb,k}^{ij}$ und $n_{x,s,d,stat,k}^{ij}$, $k = 1, \dots, 22$ ist die Anzahl der Beobachtungen in derselben Gruppe mit Diagnosegruppe k bezeichnet.

Als Diagnosegruppen wurden dabei die 21 größten Gruppen (Einzel-, Doppel- und Dreifachdiagnosen), die auch in Kapitel 6.1 in den Tabellen 6.4, 6.5 und 6.6 aufgeführt sind, und eine Gruppe *sonstige*, in welche alle übrigen Daten fallen, verwendet. Zum Beispiel enthalten die Tabellen 7.6 und 7.7 diese Zahlen jeweils für Männer und Frauen in Pflegestufe 3 im ersten Pflegejahr. In der linken Spalte sind jeweils die 22 verschiedenen Diagnosegruppen angegeben, in der letzten Zeile sind die Größen $n_{x,s,d,amb}^{ij}$ und $n_{x,s,d,stat}^{ij}$ der jeweiligen Gruppen zu finden. Tabelle 7.8 gibt einen Überblick über die Verteilung der Todesfälle auf Geschlecht, Pflegestufe und Behandlungsart. Die Häufigkeiten der Übergänge zwischen den Pflegestufen sind in Tabelle 7.9 zusammengefaßt.

Zur Berechnung der Wahrscheinlichkeiten werden die Hazardraten aus den Modellen (6.7) und (6.8) des geteilten Datensatzes verwendet. Da die Teilung zum Zeitpunkt 900 Tage erfolgte, werden die Wahrscheinlichkeiten, innerhalb des ersten bzw. des zweiten Pflegejahres zu sterben, allein auf Basis des Modells (6.7) mit Behandlungsdauer bis zu 900 Tagen ermittelt. Die Sterbewahrscheinlichkeiten im dritten Pflegejahr ergeben sich aus der Summe der Sterbewahrscheinlichkeiten im Zeitraum [720, 900] Tage (Modell (6.7)) und im Zeitraum (900, 1080] Tage (Modell (6.8)).

Zum Vergleich werden auch die Wahrscheinlichkeiten des Ausgangsmodells (6.1), dem Modell ohne Diagnosen, berechnet. In dieses Modell geht die Behandlungsart als Kovariable ein, so dass auch hier die Wahrscheinlichkeiten gewichtet werden müssen:

$$\hat{p}_{i,j}(x, s, d) = \frac{1}{n_{x,s,d,amb}^{ij} + n_{x,s,d,stat}^{ij}} \left\{ n_{x,s,d,amb}^{ij} \hat{p}_{i,j}(x, s, d, \text{ambulant}) + n_{x,s,d,stat}^{ij} \hat{p}_{i,j}(x, s, d, \text{stationär}) \right\} \quad (7.10)$$

$n_{x,s,d,amb}^{ij}$ und $n_{x,s,d,stat}^{ij}$ sind wie oben definiert. Die geschätzten Sterbewahrscheinlichkeiten ($\hat{p}_{i,4}$, $i = 1, 2, 3$) sind in den Tabellen 7.10 und 7.11 aufgeführt.

| | Alter | | | | | | | |
|-------------------------|---------|------|---------|------|---------|------|---------|------|
| | 55 - 64 | | 65 - 74 | | 75 - 84 | | 85 - 94 | |
| Diagnosegruppe | amb | stat | amb | stat | amb | stat | amb | stat |
| Demenz | 2 | 0 | 2 | 1 | 10 | 4 | 3 | 7 |
| Schlag | 4 | 1 | 7 | 4 | 4 | 2 | 6 | 2 |
| Psychose | 4 | 2 | 4 | 1 | 5 | 1 | 2 | 0 |
| Tumore | 9 | 0 | 18 | 1 | 12 | 2 | 1 | 1 |
| Arthrosen | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Herz | 0 | 0 | 2 | 1 | 4 | 2 | 2 | 0 |
| Lungen | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Demenz, Schlag | 1 | 0 | 2 | 2 | 4 | 2 | 0 | 1 |
| Demenz, Herz | 0 | 0 | 1 | 0 | 5 | 0 | 3 | 1 |
| Demenz, Arthrosen | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Demenz, Psychose | 0 | 0 | 0 | 2 | 3 | 2 | 7 | 3 |
| Demenz, Tumore | 0 | 0 | 1 | 0 | 3 | 1 | 1 | 0 |
| Tumore, Psychose | 3 | 0 | 1 | 1 | 1 | 0 | 1 | 0 |
| Tumore, Herz | 0 | 0 | 2 | 0 | 2 | 0 | 2 | 0 |
| Psychose, Herz | 0 | 0 | 1 | 0 | 2 | 0 | 0 | 0 |
| Psychose, Schlag | 1 | 0 | 2 | 0 | 1 | 0 | 0 | 0 |
| Herz, Schlag | 1 | 0 | 1 | 0 | 4 | 0 | 0 | 0 |
| Herz, Arthrosen | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Herz, Arthrosen, Demenz | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| Herz, Schlag, Demenz | 0 | 0 | 0 | 0 | 2 | 1 | 1 | 1 |
| Herz, Psychose, Demenz | 0 | 0 | 0 | 0 | 1 | 0 | 4 | 2 |
| sonstige | 10 | 2 | 14 | 1 | 14 | 7 | 11 | 3 |
| Σ | 35 | 5 | 58 | 14 | 78 | 24 | 44 | 21 |

Tabelle 7.6: $n_{x,s=0,d=1,amb}^{34}$ und $n_{x,s=0,d=1,stat}^{34}$ für Männer in Stufe 3 im ersten Pflegejahr

Als Vergleichsgröße werden zusätzlich die empirischen Sterbewahrscheinlichkeiten betrachtet, die sich aus den Daten ergeben. Diese berechnen sich als die relativen Häufigkeiten in den nach Geschlecht, Altersklasse, Pflegestufe und Pflegejahr unterteilten Gruppen:

$$\hat{p}_{i,4}^e(x, s, d) = \frac{p}{n}, \quad (7.11)$$

wobei p die Anzahl der Todesfälle und n die Anzahl der Beobachtungen in der Gruppe mit Alter $[x-5, x+5)$, Geschlecht s und Pflegedauer d angeben. Die Tabellen 7.12 und 7.13 enthalten diese Häufigkeiten, inklusive der Angabe von p und n, für Männer und Frauen.

In Abb. 7.3 sind die Sterbewahrscheinlichkeiten innerhalb des ersten Pflegejahres graphisch dargestellt, um einen Vergleich zu erleichtern. In x-Richtung sind jeweils die vier Altersstufen abgetragen, in y-Richtung zuerst Männer, dann Frauen, jeweils von Stufe 1 bis Stufe 3. Die empirisch ermittelten Wahrscheinlichkeiten sind im Plot links oben

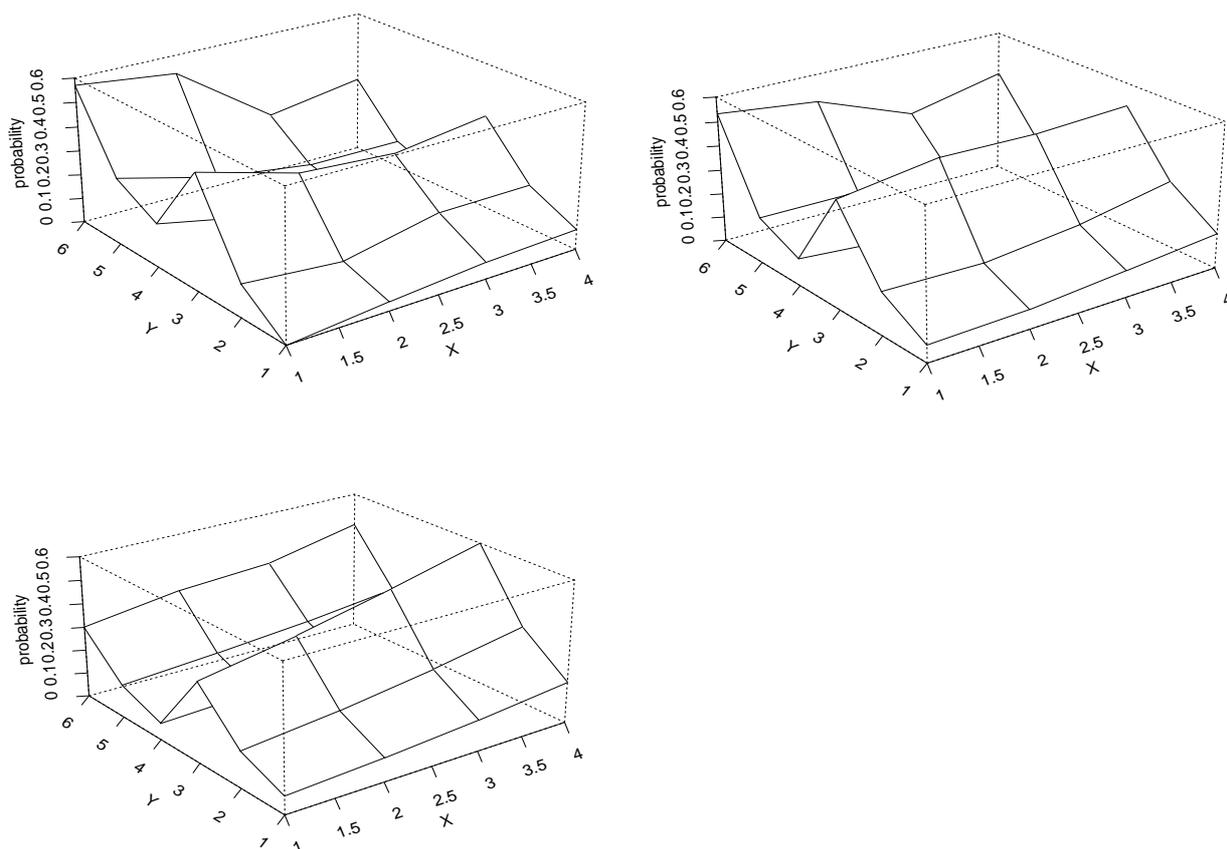


Abbildung 7.3: Vergleich der Sterbewahrscheinlichkeiten im ersten Pflegejahr: links oben: empirisch, rechts oben: Modell mit Diagnosen, links unten: Modell ohne Diagnosen, x-Achse: 1: 55-64 , 2: 65-74, 3: 75-84 , 4: 85-94 Jahre , y-Achse: 1: Männer, Stufe 1, 2: Männer, Stufe 2, 3: Männer, Stufe 3, 4: Frauen, Stufe 1, 5: Frauen, Stufe 2, 6: Frauen , Stufe 3

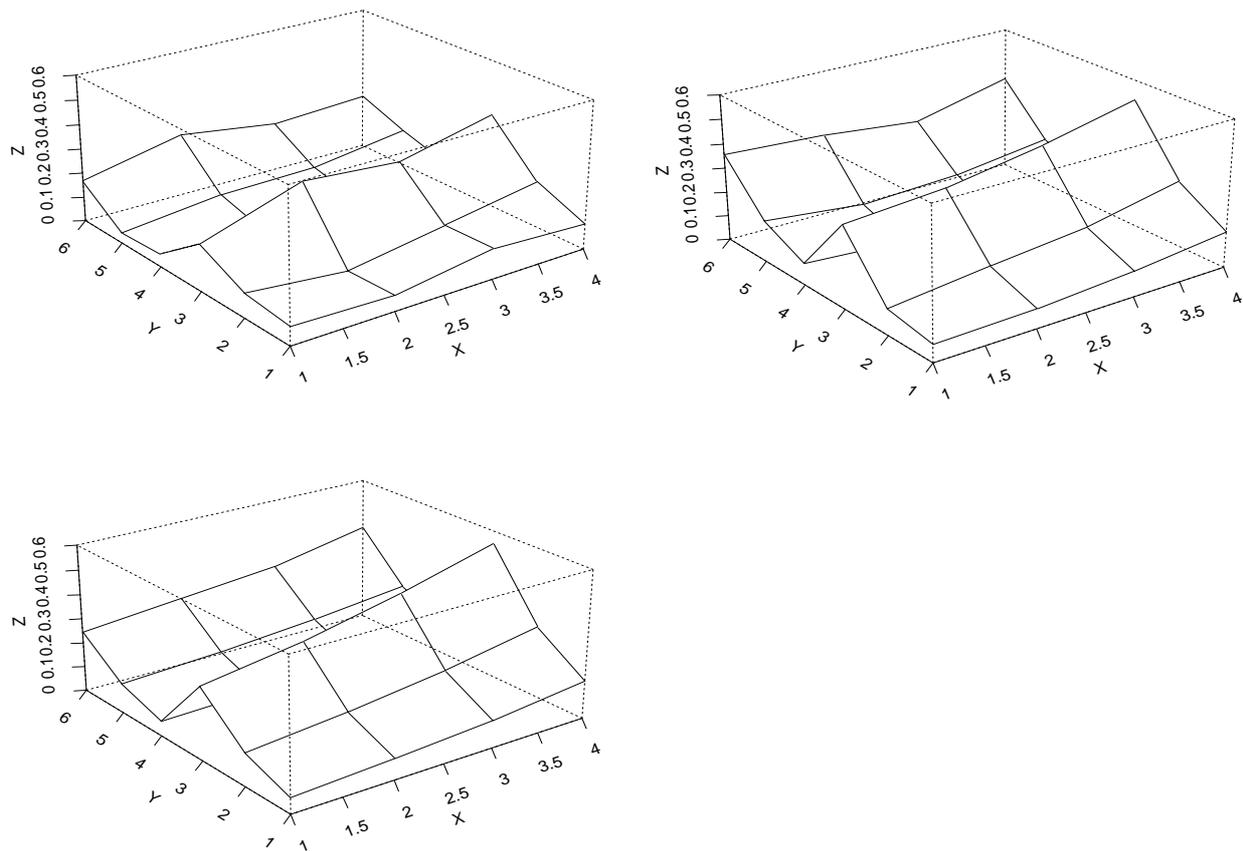


Abbildung 7.4: Vergleich der Sterbewahrscheinlichkeiten im zweiten Pflegejahr: links oben: empirisch, rechts oben: Modell mit Diagnosen, links unten: Modell ohne Diagnosen, x-Achse: 1: 55-64 , 2: 65-74, 3: 75-84 , 4: 85-94 Jahre , y-Achse: 1: Männer, Stufe 1, 2: Männer, Stufe 2, 3: Männer, Stufe 3, 4: Frauen, Stufe 1, 5: Frauen, Stufe 2, 6: Frauen , Stufe 3

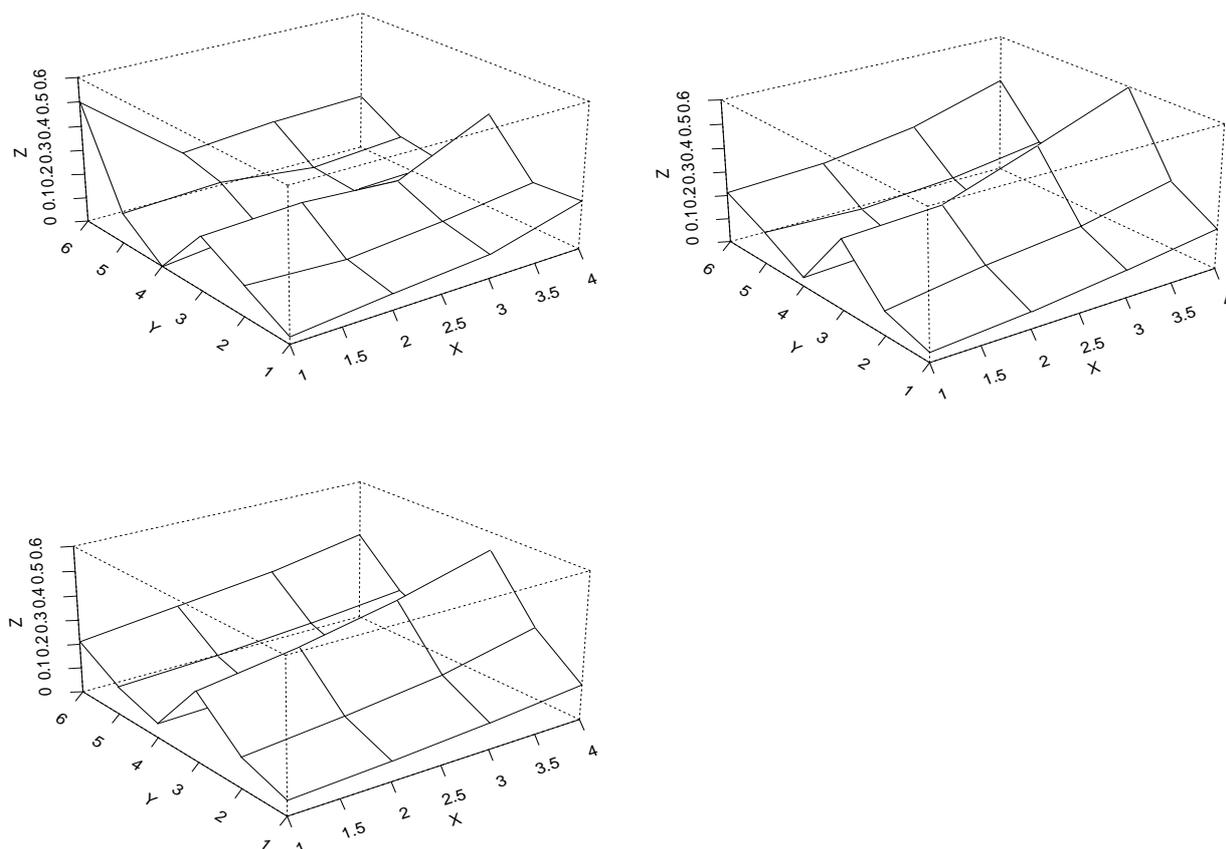


Abbildung 7.5: Vergleich der Sterbewahrscheinlichkeiten im dritten Pflegejahr: links oben: empirisch, rechts oben: Modell mit Diagnosen, links unten: Modell ohne Diagnosen, x-Achse: 1: 55-64 , 2: 65-74, 3: 75-84 , 4: 85-94 Jahre , y-Achse: 1: Männer, Stufe 1, 2: Männer, Stufe 2, 3: Männer, Stufe 3, 4: Frauen, Stufe 1, 5: Frauen, Stufe 2, 6: Frauen , Stufe 3

| Diagnosegruppe | Alter | | | | | | | |
|-------------------------|---------|------|---------|------|---------|------|---------|------|
| | 55 - 64 | | 65 - 74 | | 75 - 84 | | 85 - 94 | |
| | amb | stat | amb | stat | amb | stat | amb | stat |
| Demenz | 1 | 0 | 3 | 1 | 5 | 12 | 8 | 6 |
| Schlag | 0 | 0 | 3 | 1 | 5 | 3 | 4 | 3 |
| Psychose | 1 | 0 | 3 | 0 | 2 | 2 | 2 | 0 |
| Tumore | 3 | 2 | 15 | 1 | 4 | 1 | 3 | 0 |
| Arthrosen | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 2 |
| Herz | 0 | 0 | 0 | 0 | 6 | 1 | 8 | 1 |
| Lungen | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Demenz, Schlag | 0 | 0 | 1 | 1 | 2 | 2 | 3 | 3 |
| Demenz, Herz | 0 | 0 | 0 | 0 | 1 | 5 | 7 | 9 |
| Demenz, Arthrosen | 0 | 0 | 0 | 0 | 1 | 2 | 2 | 0 |
| Demenz, Psychose | 0 | 0 | 3 | 0 | 2 | 3 | 2 | 2 |
| Demenz, Tumore | 0 | 0 | 0 | 0 | 1 | 2 | 4 | 1 |
| Tumore, Psychose | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| Tumore, Herz | 1 | 0 | 1 | 0 | 3 | 0 | 0 | 1 |
| Psychose, Herz | 1 | 0 | 0 | 2 | 3 | 1 | 1 | 2 |
| Psychose, Schlag | 0 | 0 | 0 | 0 | 1 | 2 | 0 | 1 |
| Herz, Schlag | 0 | 1 | 2 | 0 | 3 | 1 | 1 | 0 |
| Herz, Arthrosen | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Herz, Arthrosen, Demenz | 0 | 0 | 0 | 0 | 1 | 1 | 3 | 2 |
| Herz, Schlag, Demenz | 0 | 0 | 0 | 0 | 0 | 2 | 1 | 0 |
| Herz, Psychose, Demenz | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| sonstige | 3 | 0 | 5 | 2 | 17 | 12 | 29 | 20 |
| Σ | 11 | 3 | 36 | 8 | 59 | 53 | 79 | 53 |

Tabelle 7.7: $n_{x,s=1,d=1,amb}^{34}$ und $n_{x,s=1,d=1,stat}^{34}$ für Frauen in Stufe 3 im ersten Pflegejahr

dargestellt, die Wahrscheinlichkeiten des Modells mit Diagnosen sind rechts oben, die des Modells ohne Diagnosen links unten abgebildet. Es ist deutlich erkennbar, dass bei Berücksichtigung der Diagnosen Wahrscheinlichkeiten erhalten werden, die näher an den empirischen Wahrscheinlichkeiten liegen. Vor allem in der ersten Altersgruppe (55 - 64 Jahre) und bei Frauen in Stufe 3 (y-Achse: 6) ist dies gut sichtbar. Alle drei Modelle zeigen sehr gut, dass die Sterbewahrscheinlichkeit mit zunehmender Pflegestufe deutlich ansteigt. In Stufe 3 sind, verglichen zu den empirischen, die Wahrscheinlichkeiten des Modells ohne Diagnosen jedoch deutlich schlechter angepasst. Das Modell unterstellt hier linear mit dem Alter steigende Wahrscheinlichkeiten, während die empirischen Häufigkeiten zeigen, dass hier vor allem in den Altersklassen 55 - 64 und 65 - 74 Jahre der Anteil der Todesfälle sehr hoch war. Das Modell mit Diagnosen dagegen modelliert in dieser Gruppe besser.

Auch im zweiten und dritten Pflegejahr kann man feststellen, dass das Modell mit Diagnosen Sterbewahrscheinlichkeiten liefert, die tendentiell näher an den empirischen liegen

| | Frauen | | | Männer | | |
|-----------|---------|---------|---------|---------|---------|---------|
| | Stufe 1 | Stufe 2 | Stufe 3 | Stufe 1 | Stufe 2 | Stufe 3 |
| ambulant | 155 | 306 | 276 | 98 | 229 | 312 |
| stationär | 53 | 180 | 300 | 24 | 69 | 118 |

Tabelle 7.8: Anzahl der Todesfälle in den verschiedenen Pflegestufen

| | Stufe 1 nach Stufe 2 | | Stufe 1 nach Stufe 3 | | Stufe 2 nach Stufe 3 | |
|-----------|----------------------|--------|----------------------|--------|----------------------|--------|
| | Frauen | Männer | Frauen | Männer | Frauen | Männer |
| ambulant | 319 | 156 | 63 | 42 | 178 | 152 |
| stationär | 91 | 13 | 20 | 5 | 91 | 21 |

Tabelle 7.9: Anzahl der Übergänge zwischen den Pflegestufen

als die im Modell ohne Diagnosen. Gut zu erkennen ist das z.B. bei Männern in Stufe 2 im zweiten Pflegejahr oder bei Männern in Stufe 1 im dritten Pflegejahr, siehe Tabellen 7.10, 7.11 und 7.12. Die graphische Darstellung der Wahrscheinlichkeiten in den Abb. 7.4 und 7.5 zeigt jedoch, dass die Unterschiede zwischen den Modellen meist sehr gering sind. Die Diagnosen scheinen also vor allem im ersten Pflegejahr das Überleben entscheidend zu beeinflussen. Ab dem zweiten Pflegejahr sinkt ihr Einfluß auf die Sterbewahrscheinlichkeiten. Dies erscheint plausibel, da z.B. Tumorpatienten, wie in der Modellbildung festgestellt, in der Regel sehr kurze Überlebenserwartungen besitzen und daher in späteren Pflegejahren kaum mehr ins Gewicht fallen. Bei den anderen Diagnosen kann man sich vorstellen, dass der Zustand der Patienten vor allem im ersten Pflegejahr, d.h. im unmittelbar auf das pflegeauslösende Ereignis folgenden Zeitraum, sehr stark von der jeweiligen Diagnose beeinflusst wird. Nach einer längeren Zeit in Pflege werden die Auslöser der Pflegebedürftigkeit an Bedeutung für das Überleben verlieren. Diese Vermutung kann mit den Tabellen 6.29 und 6.30 bestätigt werden. Während im Modell (6.7) für Behandlungszeiten bis zu 900 Tagen noch sieben Diagnosen signifikanten Einfluß besitzen, sind es im Modell (6.8) für längere Pflegedauern nur noch vier Diagnosen. Es fällt zudem auf, dass die Koeffizienten der Diagnosekovariablen im Modell (6.7) wesentlich höher sind, d.h. die Diagnosen wirken sich stärker auf das Überleben aus. Der Koeffizient für Z_{Tumore} zum Beispiel beträgt im Modell (6.7) 3.76, während er im Modell (6.8) nur noch bei 0.53 liegt. Insgesamt kann man daher sagen, dass sich die durch die Berücksichtigung der Diagnosen erreichte Verbesserung des Modells, zumindest im ersten Pflegejahr, auch positiv auf die resultierenden Wahrscheinlichkeiten ausgewirkt hat.

Die Wahrscheinlichkeiten für die Übergänge zwischen den Stufen wurden auf Basis der Modelle (7.3), (7.5) und (7.5) aus Kapitel 7.2 berechnet. Die Ergebnisse sind in Tabelle 7.14 aufgeführt. Zum Vergleich werden die empirischen Wahrscheinlichkeiten für die Stufenübergänge ermittelt (Tabellen 7.15 und 7.16). Die Größenordnungen der Wahrscheinlichkeiten stimmen weitgehend überein. Am geringsten sind die Übergangswahrscheinlichkeiten von Stufe 1 nach Stufe 3. Es ist nicht verwunderlich, dass das Risiko für solche

| | | Männer | | | Frauen | | |
|---------------|---------------|---------|---------|---------|---------|---------|---------|
| | Alter | Stufe 1 | Stufe 2 | Stufe 3 | Stufe 1 | Stufe 2 | Stufe 3 |
| 1. Pflegejahr | 55 - 64 Jahre | 0.078 | 0.155 | 0.339 | 0.071 | 0.140 | 0.301 |
| | 65 - 74 Jahre | 0.102 | 0.197 | 0.413 | 0.085 | 0.168 | 0.358 |
| | 75 - 84 Jahre | 0.130 | 0.252 | 0.503 | 0.103 | 0.201 | 0.381 |
| | 85 - 94 Jahre | 0.175 | 0.323 | 0.607 | 0.128 | 0.244 | 0.463 |
| 2. Pflegejahr | 55 - 64 Jahre | 0.067 | 0.134 | 0.297 | 0.061 | 0.120 | 0.245 |
| | 65 - 74 Jahre | 0.088 | 0.171 | 0.359 | 0.073 | 0.141 | 0.283 |
| | 75 - 84 Jahre | 0.113 | 0.220 | 0.446 | 0.089 | 0.173 | 0.321 |
| | 85 - 94 Jahre | 0.154 | 0.290 | 0.549 | 0.111 | 0.213 | 0.396 |
| 3. Pflegejahr | 55 - 64 Jahre | 0.063 | 0.125 | 0.284 | 0.057 | 0.115 | 0.211 |
| | 65 - 74 Jahre | 0.084 | 0.161 | 0.343 | 0.069 | 0.132 | 0.256 |
| | 75 - 84 Jahre | 0.110 | 0.209 | 0.426 | 0.084 | 0.163 | 0.302 |
| | 85 - 94 Jahre | 0.145 | 0.292 | 0.528 | 0.107 | 0.201 | 0.370 |

Tabelle 7.10: Geschätzte Sterbewahrscheinlichkeiten \hat{p}_{i4} , $i = 1, 2, 3$ basierend auf Modell (6.1) ohne Diagnosen (siehe (7.10))

| | | Männer | | | Frauen | | |
|---------------|---------------|---------|---------|---------|---------|---------|---------|
| | Alter | Stufe 1 | Stufe 2 | Stufe 3 | Stufe 1 | Stufe 2 | Stufe 3 |
| 1. Pflegejahr | 55 - 64 Jahre | 0.072 | 0.177 | 0.442 | 0.111 | 0.187 | 0.533 |
| | 65 - 74 Jahre | 0.081 | 0.168 | 0.499 | 0.081 | 0.187 | 0.488 |
| | 75 - 84 Jahre | 0.113 | 0.206 | 0.491 | 0.073 | 0.161 | 0.341 |
| | 85 - 94 Jahre | 0.145 | 0.275 | 0.509 | 0.101 | 0.199 | 0.423 |
| 2. Pflegejahr | 55 - 64 Jahre | 0.074 | 0.107 | 0.340 | 0.087 | 0.169 | 0.357 |
| | 65 - 74 Jahre | 0.081 | 0.152 | 0.370 | 0.063 | 0.125 | 0.336 |
| | 75 - 84 Jahre | 0.105 | 0.190 | 0.434 | 0.071 | 0.145 | 0.296 |
| | 85 - 94 Jahre | 0.146 | 0.267 | 0.523 | 0.101 | 0.196 | 0.389 |
| 3. Pflegejahr | 55 - 64 Jahre | 0.044 | 0.102 | 0.295 | 0.036 | 0.131 | 0.212 |
| | 65 - 74 Jahre | 0.072 | 0.162 | 0.315 | 0.058 | 0.119 | 0.231 |
| | 75 - 84 Jahre | 0.117 | 0.204 | 0.451 | 0.074 | 0.148 | 0.292 |
| | 85 - 94 Jahre | 0.170 | 0.283 | 0.595 | 0.116 | 0.204 | 0.404 |

Tabelle 7.11: Geschätzte Sterbewahrscheinlichkeiten \hat{p}_{14} , $i = 1, 2, 3$ basierend auf den Modellen (6.7) und (6.8) mit Diagnosen (siehe (7.9))

| | | Stufe 1 | | | Stufe 2 | | | Stufe 3 | | |
|---------------|---------------|---------|----|---------------|---------|----|---------------|---------|----|---------------|
| | Alter | n | p | $\frac{p}{n}$ | n | p | $\frac{p}{n}$ | n | p | $\frac{p}{n}$ |
| 1. Pflegejahr | 55 - 64 Jahre | 45 | 0 | 0.000 | 52 | 7 | 0.135 | 40 | 19 | 0.475 |
| | 65 - 74 Jahre | 105 | 4 | 0.038 | 114 | 11 | 0.097 | 72 | 26 | 0.361 |
| | 75 - 84 Jahre | 186 | 13 | 0.070 | 180 | 32 | 0.178 | 102 | 34 | 0.333 |
| | 85 - 94 Jahre | 119 | 10 | 0.084 | 139 | 25 | 0.180 | 65 | 25 | 0.385 |
| 2. Pflegejahr | 55 - 64 Jahre | 40 | 3 | 0.075 | 41 | 4 | 0.098 | 21 | 4 | 0.191 |
| | 65 - 74 Jahre | 94 | 6 | 0.064 | 120 | 7 | 0.058 | 65 | 21 | 0.323 |
| | 75 - 84 Jahre | 136 | 17 | 0.125 | 147 | 18 | 0.122 | 99 | 29 | 0.293 |
| | 85 - 94 Jahre | 96 | 10 | 0.104 | 134 | 25 | 0.187 | 79 | 30 | 0.380 |
| 3. Pflegejahr | 55 - 64 Jahre | 33 | 1 | 0.030 | 24 | 3 | 0.125 | 18 | 4 | 0.222 |
| | 65 - 74 Jahre | 60 | 4 | 0.067 | 86 | 9 | 0.105 | 58 | 14 | 0.241 |
| | 75 - 84 Jahre | 80 | 8 | 0.100 | 115 | 16 | 0.139 | 78 | 17 | 0.217 |
| | 85 - 94 Jahre | 60 | 12 | 0.200 | 102 | 19 | 0.186 | 69 | 27 | 0.391 |

Tabelle 7.12: Empirische Sterbewahrscheinlichkeiten $\hat{p}_{i4}^e, i = 1, 2, 3$ bei Männern (siehe (7.11))

| | | Stufe 1 | | | Stufe 2 | | | Stufe 3 | | |
|---------------|---------------|---------|----|---------------|---------|----|---------------|---------|----|---------------|
| | Alter | n | p | $\frac{p}{n}$ | n | p | $\frac{p}{n}$ | n | p | $\frac{p}{n}$ |
| 1. Pflegejahr | 55 - 64 Jahre | 28 | 5 | 0.179 | 22 | 6 | 0.273 | 14 | 8 | 0.571 |
| | 65 - 74 Jahre | 99 | 9 | 0.091 | 89 | 15 | 0.169 | 44 | 23 | 0.523 |
| | 75 - 84 Jahre | 396 | 25 | 0.063 | 269 | 31 | 0.115 | 112 | 28 | 0.250 |
| | 85 - 94 Jahre | 453 | 15 | 0.033 | 375 | 42 | 0.112 | 132 | 41 | 0.311 |
| 2. Pflegejahr | 55 - 64 Jahre | 19 | 1 | 0.053 | 23 | 1 | 0.043 | 12 | 2 | 0.167 |
| | 65 - 74 Jahre | 77 | 3 | 0.039 | 80 | 7 | 0.088 | 39 | 10 | 0.256 |
| | 75 - 84 Jahre | 298 | 14 | 0.047 | 262 | 25 | 0.095 | 144 | 29 | 0.201 |
| | 85 - 94 Jahre | 395 | 36 | 0.091 | 392 | 59 | 0.151 | 175 | 39 | 0.223 |
| 3. Pflegejahr | 55 - 64 Jahre | 9 | 0 | 0.000 | 16 | 2 | 0.125 | 10 | 5 | 0.500 |
| | 65 - 74 Jahre | 54 | 2 | 0.037 | 62 | 9 | 0.145 | 33 | 6 | 0.182 |
| | 75 - 84 Jahre | 187 | 16 | 0.086 | 223 | 22 | 0.099 | 129 | 28 | 0.217 |
| | 85 - 94 Jahre | 238 | 24 | 0.101 | 329 | 42 | 0.128 | 180 | 42 | 0.233 |

Tabelle 7.13: Empirische Sterbewahrscheinlichkeiten $\hat{p}_{i4}^e, i = 1, 2, 3$ bei Frauen (siehe (7.11))

sprunghaften Verschlechterungen geringer ist. Ein Vergleich mit Wahrscheinlichkeiten aus einem Modell ohne Diagnosen wird hier nicht durchgeführt. Da die betrachteten Modelle für die Stufenübergänge aus Kapitel (7.2) sehr wenige Kovariablen und vor allem kaum Diagnosen enthalten, ist kein großer Unterschied der Wahrscheinlichkeiten zu erwarten.

| | | Stufe 1 → Stufe 2 | | Stufe 1 → Stufe 3 | | Stufe 2 → Stufe 3 | |
|---------------|---------------|-------------------|--------|-------------------|--------|-------------------|--------|
| | Alter | Männer | Frauen | Männer | Frauen | Männer | Frauen |
| 1. Pflegejahr | 55 - 64 Jahre | 0.108 | 0.113 | 0.024 | 0.016 | 0.130 | 0.074 |
| | 65 - 74 Jahre | 0.128 | 0.127 | 0.034 | 0.022 | 0.122 | 0.118 |
| | 75 - 84 Jahre | 0.161 | 0.152 | 0.051 | 0.031 | 0.142 | 0.100 |
| | 85 - 94 Jahre | 0.191 | 0.178 | 0.071 | 0.042 | 0.167 | 0.104 |
| 2. Pflegejahr | 55 - 64 Jahre | 0.144 | 0.155 | 0.027 | 0.021 | 0.131 | 0.103 |
| | 65 - 74 Jahre | 0.170 | 0.167 | 0.039 | 0.025 | 0.142 | 0.110 |
| | 75 - 84 Jahre | 0.208 | 0.197 | 0.056 | 0.034 | 0.168 | 0.106 |
| | 85 - 94 Jahre | 0.245 | 0.234 | 0.080 | 0.048 | 0.192 | 0.118 |
| 3. Pflegejahr | 55 - 64 Jahre | 0.126 | 0.137 | 0.023 | 0.023 | 0.097 | 0.067 |
| | 65 - 74 Jahre | 0.149 | 0.144 | 0.035 | 0.024 | 0.111 | 0.086 |
| | 75 - 84 Jahre | 0.184 | 0.173 | 0.055 | 0.032 | 0.135 | 0.091 |
| | 85 - 94 Jahre | 0.220 | 0.208 | 0.075 | 0.045 | 0.144 | 0.097 |

Tabelle 7.14: Geschätzte Übergangswahrscheinlichkeiten $\hat{p}_{ij}, i < j$ zwischen den Pflegestufen basierend auf den Modellen (7.3), (7.4) und (7.5)

| | | Stufe 1 → Stufe 2 | | | Stufe 1 → Stufe 3 | | | Stufe 2 → Stufe 3 | | |
|---------------|---------------|-------------------|----|---------------|-------------------|----|---------------|-------------------|----|---------------|
| | Alter | n | p | $\frac{p}{n}$ | n | p | $\frac{p}{n}$ | n | p | $\frac{p}{n}$ |
| 1. Pflegejahr | 55 - 64 Jahre | 42 | 7 | 0.167 | 42 | 2 | 0.048 | 40 | 8 | 0.200 |
| | 65 - 74 Jahre | 94 | 11 | 0.117 | 94 | 4 | 0.043 | 92 | 4 | 0.044 |
| | 75 - 84 Jahre | 156 | 17 | 0.109 | 156 | 10 | 0.064 | 125 | 15 | 0.120 |
| | 85 - 94 Jahre | 93 | 14 | 0.151 | 93 | 7 | 0.075 | 87 | 76 | 0.069 |
| 2. Pflegejahr | 55 - 64 Jahre | 37 | 2 | 0.054 | 37 | 0 | 0.000 | 33 | 2 | 0.061 |
| | 65 - 74 Jahre | 84 | 13 | 0.155 | 84 | 1 | 0.012 | 101 | 14 | 0.139 |
| | 75 - 84 Jahre | 110 | 17 | 0.155 | 110 | 8 | 0.073 | 109 | 11 | 0.101 |
| | 85 - 94 Jahre | 73 | 15 | 0.205 | 73 | 3 | 0.041 | 92 | 30 | 0.077 |
| 3. Pflegejahr | 55 - 64 Jahre | 31 | 0 | 0.000 | 31 | 0 | 0.000 | 20 | 2 | 0.100 |
| | 65 - 74 Jahre | 54 | 8 | 0.148 | 54 | 1 | 0.019 | 62 | 5 | 0.081 |
| | 75 - 84 Jahre | 67 | 9 | 0.134 | 67 | 2 | 0.029 | 85 | 12 | 0.141 |
| | 85 - 94 Jahre | 43 | 6 | 0.140 | 43 | 3 | 0.070 | 67 | 10 | 0.150 |

Tabelle 7.15: Empirische Übergangswahrscheinlichkeiten für Männer

| | | Stufe 1 → Stufe 2 | | | Stufe 1 → Stufe 3 | | | Stufe 2 → Stufe 3 | | |
|---------------|---------------|-------------------|----|---------------|-------------------|----|---------------|-------------------|----|---------------|
| | Alter | n | p | $\frac{p}{n}$ | n | p | $\frac{p}{n}$ | n | p | $\frac{p}{n}$ |
| 1. Pflegejahr | 55 - 64 Jahre | 28 | 1 | 0.036 | 28 | 0 | 0.000 | 22 | 2 | 0.091 |
| | 65 - 74 Jahre | 99 | 12 | 0.121 | 99 | 1 | 0.010 | 89 | 10 | 0.112 |
| | 75 - 84 Jahre | 396 | 40 | 0.101 | 396 | 10 | 0.025 | 269 | 14 | 0.052 |
| | 85 - 94 Jahre | 453 | 57 | 0.126 | 453 | 10 | 0.022 | 375 | 21 | 0.056 |
| 2. Pflegejahr | 55 - 64 Jahre | 19 | 6 | 0.316 | 19 | 0 | 0.000 | 23 | 3 | 0.130 |
| | 65 - 74 Jahre | 77 | 11 | 0.143 | 77 | 2 | 0.026 | 80 | 6 | 0.075 |
| | 75 - 84 Jahre | 298 | 37 | 0.124 | 298 | 8 | 0.027 | 262 | 21 | 0.080 |
| | 85 - 94 Jahre | 395 | 57 | 0.144 | 395 | 13 | 0.033 | 392 | 30 | 0.077 |
| 3. Pflegejahr | 55 - 64 Jahre | 9 | 1 | 0.011 | 9 | 0 | 0.000 | 16 | 0 | 0.000 |
| | 65 - 74 Jahre | 54 | 9 | 0.167 | 54 | 2 | 0.037 | 62 | 6 | 0.097 |
| | 75 - 84 Jahre | 187 | 27 | 0.144 | 187 | 3 | 0.016 | 223 | 21 | 0.094 |
| | 85 - 94 Jahre | 238 | 32 | 0.135 | 238 | 8 | 0.034 | 329 | 11 | 0.033 |

Tabelle 7.16: Empirische Übergangswahrscheinlichkeiten für Frauen

Kapitel 8

Anpassung eines Poissonmodells

Alternativ können die Daten, wie in Kapitel 5.4 vorgestellt, durch ein Poissonmodell angepasst werden. Dazu müssen zuerst die Daten bearbeitet werden. Durch die Erzeugung von Pseudobeobachtungen kann die Zahl der Daten sehr stark ansteigen. Daher wird im folgenden nur der Übergang von Stufe 1 nach Tod ($p_{1,4}$ in Abb. 7.2) modelliert.

Es werden nur Patienten berücksichtigt, die sich in Pflegestufe 1 befinden und dann in diesem Zustand entweder sterben oder zensiert werden. Entfernt werden die Daten von Pflegebedürftigen, die von Stufe 1 in eine andere Pflegestufe wechseln, um zu verhindern, dass diese Beobachtungen zusätzlich als zensiert in das Modell eingehen.

Nach Kapitel 5.4 ist ein Poissonmodell mit dem Todesindikator als Response äquivalent zu einem stückweise Exponentialmodell. Dieses nimmt an, dass der Baselinehazard intervallweise konstant ist. Es müssen also zunächst diese Zeitintervalle bestimmt werden. Für den zugrundeliegenden Datensatz wird folgende Unterteilung verwendet:

$$[0, 301), [301, 501), [501, 831), [831, 1370), [1370, \infty) \text{ Tage} \quad (8.1)$$

Die Intervallgrenzen sind dabei so gewählt, dass in jedem Intervall etwa 300 Beobachtungen, das entspricht einem Fünftel des Datensatzes, liegen. Nun werden die Pseudobeobachtungen, wie in Kapitel 5.4 unter (5.11) und (5.12) beschrieben, gebildet: jede Beobachtung wird in eine extra Beobachtung pro Zeitintervall unterteilt.

Die Bearbeitung kann vereinfacht werden, indem man die Daten gruppiert. Die Kovariable $\mathbf{Z}_{\text{Alter}}$, die bisher metrisch in die Modellierung einging, wird daher in vier Alterskategorien unterteilt und als Faktor behandelt. Als Altersklassen, nach den Quartilen gebildet, ergeben sich:

$$[0, 75), [75, 83), [83, 88), [88, 108) \text{ Jahre}$$

Als Referenzkategorie wird die Klasse bis 75 Jahre gewählt. Die Diagnosen werden zunächst nicht als Kovariablen berücksichtigt, nur die Variablen Alter, Behandlungsart und Geschlecht gehen in die Modellierung mit ein. Aggregiert man in jedem der fünf Zeitintervalle die Daten mit gleicher Ausprägung der Kovariablen, verringert sich die Zahl der Beobachtungen auf $4 \cdot 2 \cdot 2 \cdot 5 = 80$ (Alter: 4, Art: 2, Geschlecht: 2, Zeitintervall: 5 Ausprägungen). Der Todesindikator d_{ij} für die Gruppe mit Merkmalen i in Zeitintervall j wird durch die Summe der Todesfälle $d_{ij} = \sum_{k=1}^{n_i} d_{kj}$ in jeder Gruppe ersetzt, wobei n_i die Anzahl der

Beobachtungen mit Merkmalen i ist. Zusätzlich wird die Variable $\mathbf{Z}_{exposure} = \sum_{k=1}^{n_i} t_{kj}$ eingeführt, die die gesamte Pflegedauer der Beobachtungen pro Gruppe angibt. Dabei ist t_{kj} definiert wie in (5.11).

Für diese Daten kann nun ein Poissonmodell angepasst werden mit Response $d_{ij} \sim Poi(\mu_{ij})$ und Offset $\log(\mathbf{Z}_{exposure})$. Die Zeitintervalle werden als Faktor betrachtet und gehen daher als Indikatorvariablen

$$\mathbf{Z}_{Zeitj} = \begin{cases} 1 & \text{Intervall } j \\ 0 & \text{sonst} \end{cases}, j = 1, \dots, 5,$$

in das Modell ein. Nach Kapitel 5.4 gilt mit logarithmischer Linkfunktion:

$$\log \mu_{ij} = \log(\mathbf{Z}_{exposure}) + \alpha_j + \beta' \mathbf{Z}_i$$

Der Parameter $\alpha_j = \log \lambda_j$ ist hier der logarithmierte Baselinehazard des j -ten Zeitintervalls, als Referenzgröße ist der Baselinehazard des ersten Intervalls gewählt, d.h. es gilt $\alpha_1 = 0$. Im Vektor \mathbf{Z} sind die Kovariablen Alter, Geschlecht und Behandlungsart enthalten.

Eine Untersuchung ergibt, dass alle drei Kovariablen signifikanten Einfluss besitzen, zwischen Alter und Behandlungsart treten Wechselwirkungen auf. Das Poissonmodell ermöglicht zugleich eine einfache Modellierung von zeitabhängigen Effekten. Lässt man Interaktionen zwischen der Behandlungsdauer und den übrigen Kovariablen zu, erhält man nach (5.9)

$$\log \lambda_{ij} = \alpha_j + \beta' \mathbf{Z}_i + \gamma'(\mathbf{Z}_i \times \mathbf{Z}_{Zeitj}), \quad (8.2)$$

Die Parameter γ geben hier den Effekt bei Kovariablen mit Ausprägung i im Intervall j an. Betrachtet man z.B. nur die Interaktion mit der Kovariable $\mathbf{Z}_{Geschlecht}$, so gilt:

$$\begin{aligned} \log \lambda_{ij} &= \alpha_j + \beta_1 \mathbf{Z}_{Geschlecht} + \beta_2 \mathbf{Z}_{stat} + \beta_3 \mathbf{Z}_{75-83Jahre} + \beta_4 \mathbf{Z}_{83-88Jahre} + \beta_5 \mathbf{Z}_{88-108Jahre} \\ &+ \gamma_1(\mathbf{Z}_{Geschlecht} \times \mathbf{Z}_{Zeit2}) + \gamma_2(\mathbf{Z}_{Geschlecht} \times \mathbf{Z}_{Zeit3}) + \gamma_3(\mathbf{Z}_{Geschlecht} \times \mathbf{Z}_{Zeit4}) \\ &+ \gamma_4(\mathbf{Z}_{Geschlecht} \times \mathbf{Z}_{Zeit5}) \end{aligned}$$

Der Effekt des Geschlechts kann nun über die verschiedenen Zeitintervalle hinweg variieren. In den ersten 300 Tagen der Pflege wird er durch β_1 , im Intervall [301, 501) Tage durch $\beta_1 + \gamma_1$, im Intervall [501,831) durch $\beta_1 + \gamma_2$ usw. angegeben.

In unserem Modell konnte so das Alter zeitabhängig modelliert werden. Der Log-Hazard $\log \lambda$ in diesem Modell lautet

$$\begin{aligned} \log \lambda &= \text{Intercept} + \alpha_2 \mathbf{Z}_{Zeit2} + \alpha_3 \mathbf{Z}_{Zeit3} + \alpha_4 \mathbf{Z}_{Zeit4} + \alpha_5 \mathbf{Z}_{Zeit5} + \beta_1 \mathbf{Z}_{Geschlecht} \quad (8.3) \\ &+ \beta_2 \mathbf{Z}_{stat} + \beta_3 \mathbf{Z}_{75-83Jahre} + \beta_4 \mathbf{Z}_{83-88Jahre} + \beta_5 \mathbf{Z}_{88-108Jahre} \\ &+ \beta_6(\mathbf{Z}_{75-83Jahre} \times \mathbf{Z}_{Zeit2}) + \beta_7(\mathbf{Z}_{83-88Jahre} \times \mathbf{Z}_{Zeit2}) + \beta_8(\mathbf{Z}_{88-108Jahre} \times \mathbf{Z}_{Zeit2}) \\ &+ \beta_9(\mathbf{Z}_{75-83Jahre} \times \mathbf{Z}_{Zeit3}) + \beta_{10}(\mathbf{Z}_{83-88Jahre} \times \mathbf{Z}_{Zeit3}) + \beta_{11}(\mathbf{Z}_{88-108Jahre} \times \mathbf{Z}_{Zeit3}) \\ &+ \beta_{12}(\mathbf{Z}_{75-83Jahre} \times \mathbf{Z}_{Zeit4}) + \beta_{13}(\mathbf{Z}_{83-88Jahre} \times \mathbf{Z}_{Zeit4}) + \beta_{14}(\mathbf{Z}_{88-108Jahre} \times \mathbf{Z}_{Zeit4}) \\ &+ \beta_{15}(\mathbf{Z}_{75-83Jahre} \times \mathbf{Z}_{Zeit5}) + \beta_{16}(\mathbf{Z}_{83-88Jahre} \times \mathbf{Z}_{Zeit5}) + \beta_{17}(\mathbf{Z}_{88-108Jahre} \times \mathbf{Z}_{Zeit5}) \\ &+ \beta_{18}(\mathbf{Z}_{75-83Jahre} \times \mathbf{Z}_{stat}) + \beta_{19}(\mathbf{Z}_{83-88Jahre} \times \mathbf{Z}_{stat}) + \beta_{20}(\mathbf{Z}_{88-108Jahre} \times \mathbf{Z}_{stat}) \end{aligned}$$

Die Koeffizienten der Kovariablen sind zusammen mit dem t-Wert des Wald-Tests in Tabelle 8.1 aufgeführt. Als grobe Richtlinie gilt, dass Kovariablen mit einem t-Wert betragsmäßig über 2 signifikant auf einem 5 %- Level sind. Bei 55 Freiheitsgraden und einem Wert der Devianz von 72.97 ergibt ein Residual-Devianz-Test einen p-Wert von 0.053. Bei einem Signifikanzniveau von 5 % führt dieser Wert zwar schon zu einer Verwerfung der Nullhypothese H_0 : es besteht ein Lack-of-Fit, der niedrige p-Wert deutet jedoch an, dass das Modell noch nicht sehr gut ist. Eventuell sind noch zu wenige Kovariablen berücksichtigt worden. Durch Hinzunahme der Diagnosen könnte eine Verbesserung erzielt werden.

| Variable | Koeffizient | SE(Koeff.) | t-Wert | p-Wert |
|--|-------------|------------|----------|---------------------|
| Intercept | -7.8214 | 0.2554 | -30.6213 | $< 10^{-16}$ |
| $\mathbf{Z}_{75-83Jahre}$ | 0.0411 | 0.3373 | 0.1218 | $9.0 \cdot 10^{-1}$ |
| $\mathbf{Z}_{83-88Jahre}$ | 0.0679 | 0.3499 | 0.1941 | $8.5 \cdot 10^{-1}$ |
| $\mathbf{Z}_{88-108Jahre}$ | -0.1183 | 0.3715 | -0.3186 | $7.5 \cdot 10^{-1}$ |
| \mathbf{Z}_{Zeit2} | 0.3620 | 0.3512 | 1.0078 | $3.1 \cdot 10^{-1}$ |
| \mathbf{Z}_{Zeit3} | -0.4963 | 0.4329 | -1.1464 | $2.5 \cdot 10^{-1}$ |
| \mathbf{Z}_{Zeit4} | -0.2606 | 0.4029 | -0.6470 | $5.2 \cdot 10^{-1}$ |
| \mathbf{Z}_{Zeit5} | -1.3928 | 0.4530 | -3.0748 | $2.1 \cdot 10^{-3}$ |
| \mathbf{Z}_{stat} | -0.8746 | 0.5921 | -1.4770 | $1.4 \cdot 10^{-1}$ |
| $\mathbf{Z}_{Geschlecht}$ | -0.3951 | 0.1212 | -3.2609 | $1.1 \cdot 10^{-3}$ |
| $\mathbf{Z}_{75-83Jahre} \times \mathbf{Z}_{Zeit2}$ | -0.3443 | 0.5014 | -0.6867 | $9.5 \cdot 10^{-1}$ |
| $\mathbf{Z}_{83-88Jahre} \times \mathbf{Z}_{Zeit2}$ | -0.0380 | 0.5060 | -0.0751 | $9.4 \cdot 10^{-1}$ |
| $\mathbf{Z}_{88-108Jahre} \times \mathbf{Z}_{Zeit2}$ | 0.4556 | 0.4982 | 0.9144 | $3.6 \cdot 10^{-1}$ |
| $\mathbf{Z}_{75-83Jahre} \times \mathbf{Z}_{Zeit3}$ | 0.5602 | 0.5371 | 1.0430 | $3.0 \cdot 10^{-1}$ |
| $\mathbf{Z}_{83-88Jahre} \times \mathbf{Z}_{Zeit3}$ | 1.0527 | 0.5353 | 1.9666 | $4.9 \cdot 10^{-2}$ |
| $\mathbf{Z}_{88-108Jahre} \times \mathbf{Z}_{Zeit3}$ | 1.3406 | 0.5419 | 2.4740 | $1.3 \cdot 10^{-2}$ |
| $\mathbf{Z}_{75-83Jahre} \times \mathbf{Z}_{Zeit4}$ | 0.1025 | 0.5251 | 0.1952 | $8.5 \cdot 10^{-1}$ |
| $\mathbf{Z}_{83-88Jahre} \times \mathbf{Z}_{Zeit4}$ | 0.6095 | 0.5202 | 1.1716 | $2.4 \cdot 10^{-1}$ |
| $\mathbf{Z}_{88-108Jahre} \times \mathbf{Z}_{Zeit4}$ | 1.2012 | 0.5158 | 2.3291 | $2.0 \cdot 10^{-2}$ |
| $\mathbf{Z}_{75-83Jahre} \times \mathbf{Z}_{Zeit5}$ | 0.8271 | 0.6272 | 1.3188 | $1.9 \cdot 10^{-1}$ |
| $\mathbf{Z}_{83-88Jahre} \times \mathbf{Z}_{Zeit5}$ | 1.6769 | 0.6013 | 2.7886 | $5.3 \cdot 10^{-3}$ |
| $\mathbf{Z}_{88-108Jahre} \times \mathbf{Z}_{Zeit5}$ | 1.9870 | 0.5775 | 3.4409 | $5.8 \cdot 10^{-4}$ |
| $\mathbf{Z}_{75-83Jahre} \times \mathbf{Z}_{stat}$ | 1.8711 | 0.6435 | 2.9076 | $3.6 \cdot 10^{-3}$ |
| $\mathbf{Z}_{83-88Jahre} \times \mathbf{Z}_{stat}$ | 0.9321 | .6577 | 1.4173 | $1.6 \cdot 10^{-1}$ |
| $\mathbf{Z}_{88-108Jahre} \times \mathbf{Z}_{stat}$ | 1.1566 | 0.6274 | 1.8435 | $6.5 \cdot 10^{-2}$ |

Tabelle 8.1: Koeffizienten, t-Werte und p-Werte der Kovariablen im Poissonmodell (8.3)

Zum Vergleich wird der Übergang von Stufe 1 nach Tod auch durch ein Proportional-Hazard-Modell modelliert. Auch hier werden die durch Bildung von Pseudobeobachtungen erzeugten Daten verwendet, jedoch in ungruppiertes Form, da die Todesindikatorvariable im Cox-Modell nur die Werte 1 oder 0 annehmen kann. Wieder haben alle drei Kovariablen Einfluß auf das Überleben, zwischen Alter und Behandlungsart wird ebenfalls eine

Interaktion festgestellt. Die Hazardfunktion ist durch

$$\begin{aligned} \lambda(t) = & \lambda_0(t) \exp[\beta_1 \mathbf{Z}_{75-83\text{Jahre}} + \beta_2 \mathbf{Z}_{83-88\text{Jahre}} + \beta_3 \mathbf{Z}_{88-108\text{Jahre}} \\ & + \beta_4 \mathbf{Z}_{stat} + \beta_5 \mathbf{Z}_{Geschlecht} + \beta_6 (\mathbf{Z}_{75-83\text{Jahre}} \times \mathbf{Z}_{stat}) \\ & + \beta_7 (\mathbf{Z}_{83-88\text{Jahre}} \times \mathbf{Z}_{stat}) + \beta_8 (\mathbf{Z}_{88-108\text{Jahre}} \times \mathbf{Z}_{stat})] \end{aligned} \quad (8.4)$$

gegeben. Die zugehörigen Koeffizienten finden sich in Tabelle 8.2.

| Variable | Koeffizient | exp(Koeff.) | SE(Koeff.) | z-Statistik | p-Wert |
|--|-------------|-------------|------------|-------------|---------------------|
| $\mathbf{Z}_{75-83\text{Jahre}}$ | 0.230 | 1.259 | 0.200 | 1.15 | $2.5 \cdot 10^{-1}$ |
| $\mathbf{Z}_{83-88\text{Jahre}}$ | 0.607 | 1.835 | 0.191 | 3.18 | $1.5 \cdot 10^{-3}$ |
| $\mathbf{Z}_{88-108\text{Jahre}}$ | 0.780 | 2.181 | 0.187 | 4.18 | $3.0 \cdot 10^{-5}$ |
| \mathbf{Z}_{stat} | -0.816 | 0.442 | 0.594 | -1.37 | $1.7 \cdot 10^{-1}$ |
| $\mathbf{Z}_{Geschlecht}$ | -0.377 | 0.686 | 0.122 | -3.10 | $1.9 \cdot 10^{-3}$ |
| $\mathbf{Z}_{75-83\text{Jahre}} \times \mathbf{Z}_{stat}$ | 1.804 | 6.071 | 0.644 | 2.80 | $5.1 \cdot 10^{-3}$ |
| $\mathbf{Z}_{83-88\text{Jahre}} \times \mathbf{Z}_{stat}$ | 0.925 | 2.522 | 0.659 | 1.40 | $1.6 \cdot 10^{-1}$ |
| $\mathbf{Z}_{88-108\text{Jahre}} \times \mathbf{Z}_{stat}$ | 1.167 | 3.214 | 0.629 | 1.86 | $6.3 \cdot 10^{-2}$ |

Tabelle 8.2: Koeffizienten und p-Werte der Kovariablen im Proportional-Hazard-Modell (8.4)

Ein direkter Vergleich der Koeffizienten der beiden Modelle ist aufgrund der im Poissonmodell modellierten Zeitabhängigkeit des Alters nur schwer möglich. Daher wird zusätzlich ein Poissonmodell gefittet, das nur die im Coxmodell enthaltenen Kovariablen und Interaktionen beinhaltet. Der Log-Hazard für dieses Modell lautet

$$\begin{aligned} \log \lambda = & \text{Intercept} + \alpha_2 \mathbf{Z}_{Zeit2} + \alpha_3 \mathbf{Z}_{Zeit3} + \alpha_4 \mathbf{Z}_{Zeit4} + \alpha_5 \mathbf{Z}_{Zeit5} + \beta_1 \mathbf{Z}_{Geschlecht} \quad (8.5) \\ & + \beta_2 \mathbf{Z}_{stat} + \beta_3 \mathbf{Z}_{75-83\text{Jahre}} + \beta_4 \mathbf{Z}_{83-88\text{Jahre}} + \beta_5 \mathbf{Z}_{88-108\text{Jahre}} \\ & + \beta_6 (\mathbf{Z}_{75-83\text{Jahre}} \times \mathbf{Z}_{stat}) + \beta_7 (\mathbf{Z}_{83-88\text{Jahre}} \times \mathbf{Z}_{stat}) + \beta_8 (\mathbf{Z}_{88-108\text{Jahre}} \times \mathbf{Z}_{stat}). \end{aligned}$$

In Tabelle 8.3 sind die durch dieses Modell erhaltenen Koeffizienten aufgelistet. Diese Koeffizienten liegen nun wie zu erwarten sehr nahe an denen des Proportional-Hazard-Modells.

Um einen Vergleich der Hazardraten der Modelle durchführen zu können, müssen die Baselinehazards ermittelt werden. Im Poissonmodell muss neben den Koeffizienten der Variablen \mathbf{Z}_{Zeitj} , $j = 2, \dots, 5$ auch der Intercept berücksichtigt werden. Darüber hinaus muß der Baselinehazard pro Zeitintervall mit der Länge des jeweiligen Intervalls multipliziert werden, so daß die Zeiteinheiten des Exposures und der Raten übereinstimmen. Der Baselinehazard in den einzelnen Intervallen $[\tau_{j-1}, \tau_j)$ (Intervallgrenzen siehe (8.1)) ergibt sich wegen $\log \lambda_j = \alpha_j$ somit als

$$\lambda_j = (\tau_j - \tau_{j-1}) \cdot \exp[\text{Intercept} + \alpha_j], \quad j = 1, \dots, 5$$

Eine Schätzung des Baselinehazards $\lambda_0(t)$ im Proportional-Hazard-Modell zu jedem Ereigniszeitpunkt t erhält man mittels des Breslow-Schätzers (4.23). Der Baselinehazard

| Variable | Koeffizient | SE(Koeff.) | t-Wert | p-Wert |
|---|-------------|------------|----------|----------------------|
| Intercept | -8.2944 | 0.1822 | -45.5201 | $< 10^{-16}$ |
| $\mathbf{Z}_{75-83\text{Jahre}}$ | 0.2534 | 0.1994 | 1.2709 | $2.0 \cdot 10^{-1}$ |
| $\mathbf{Z}_{83-88\text{Jahre}}$ | 0.6399 | 0.1905 | 3.3597 | $7.8 \cdot 10^{-4}$ |
| $\mathbf{Z}_{88-108\text{Jahre}}$ | 0.8123 | 0.1864 | 4.3582 | $1.3 \cdot 10^{-5}$ |
| $\mathbf{Z}_{\text{Zeit2}}$ | 0.3833 | 0.1737 | 2.2064 | $2.7 \cdot 10^{-2}$ |
| $\mathbf{Z}_{\text{Zeit3}}$ | 0.3224 | 0.1649 | 1.9551 | $5.1 \cdot 10^{-2}$ |
| $\mathbf{Z}_{\text{Zeit4}}$ | 0.2823 | 0.1663 | 1.6978 | $9.0 \cdot 10^{-1}$ |
| $\mathbf{Z}_{\text{Zeit5}}$ | -0.2135 | 0.1974 | -1.0816 | $2.8 \cdot 10^{-1}$ |
| \mathbf{Z}_{stat} | -0.8857 | 0.5918 | -1.4966 | $1.4 \cdot 10^{-1}$ |
| $\mathbf{Z}_{\text{Geschlecht}}$ | -0.3781 | 0.1210 | -3.1240 | $1.8 \cdot 10^{-3}$ |
| $\mathbf{Z}_{75-83\text{Jahre}} \times \mathbf{Z}_{\text{stat}}$ | 1.8519 | 0.6427 | 2.8814 | $4.0 \cdot 10^{-3}$ |
| $\mathbf{Z}_{83-88\text{Jahre}} \times \mathbf{Z}_{\text{stat}}$ | 0.9638 | .6569 | 1.4672 | $1.4 \cdot 10^{-1}$ |
| $\mathbf{Z}_{88-108\text{Jahre}} \times \mathbf{Z}_{\text{stat}}$ | 1.2601 | 0.6260 | 2.0130 | $0.04 \cdot 10^{-1}$ |

Tabelle 8.3: Koeffizienten, t-Werte und p-Werte der Kovariablen im Poissonmodell (8.5) ohne Zeitabhängigkeit des Alters

des j -ten Zeitintervalls ergibt sich nun aus der Summe dieser Baselinehazards über alle Todeszeitpunkte innerhalb des Intervalls $[\tau_{j-1}, \tau_j)$:

$$\lambda_j = \sum_{\tau_{j-1} \leq t_i < \tau_j} \lambda_0(t_i)$$

Die Hazardraten können nun sowohl im Poisson- als auch im Proportional-Hazard-Modell über die Formel

$$\lambda_{ij} = \lambda_j \exp[\boldsymbol{\beta}'\mathbf{Z}_i]$$

für jedes Intervall und jede Kovariablenausprägung berechnet werden. Die $\boldsymbol{\beta}$ sind dabei die Regressionskoeffizienten des jeweilig betrachteten Modells.

Zusätzlich werden die empirischen Hazardraten mit Hilfe des Nelson-Aalen-Schätzers (3.10) für jedes Zeitintervall bestimmt:

$$\lambda_{ij} = \sum_{\tau_{j-1} \leq t_l < \tau_j} \frac{d_i(t_l)}{r_i(t_l)}$$

Hier bezeichnen $d_i(t)$ und $r_i(t)$ die Anzahl der Todesfälle bzw. die Anzahl der Individuen in der Risikomenge zum Zeitpunkt t in der Gruppe der Beobachtungen mit Merkmalen \mathbf{Z}_i . In Abb. 8.1 sind die Hazardraten der verschiedenen Modelle für Frauen graphisch dargestellt. An der x-Achse sind die fünf Zeitintervalle abgetragen, an der y-Achse die verschiedenen Gruppen, unterteilt nach Alter und Behandlungsart. Die Hazardraten sind in z-Richtung geplottet.

Zunächst fällt der starke Anstieg der Hazardraten im letzten Zeitintervall auf. Ab 1370 Tagen liegen die durch die Modelle erhaltenen Sterberaten deutlich über den empirischen, die rechts oben zu finden sind. Dieser Unterschied ist eventuell darauf zurückzuführen,

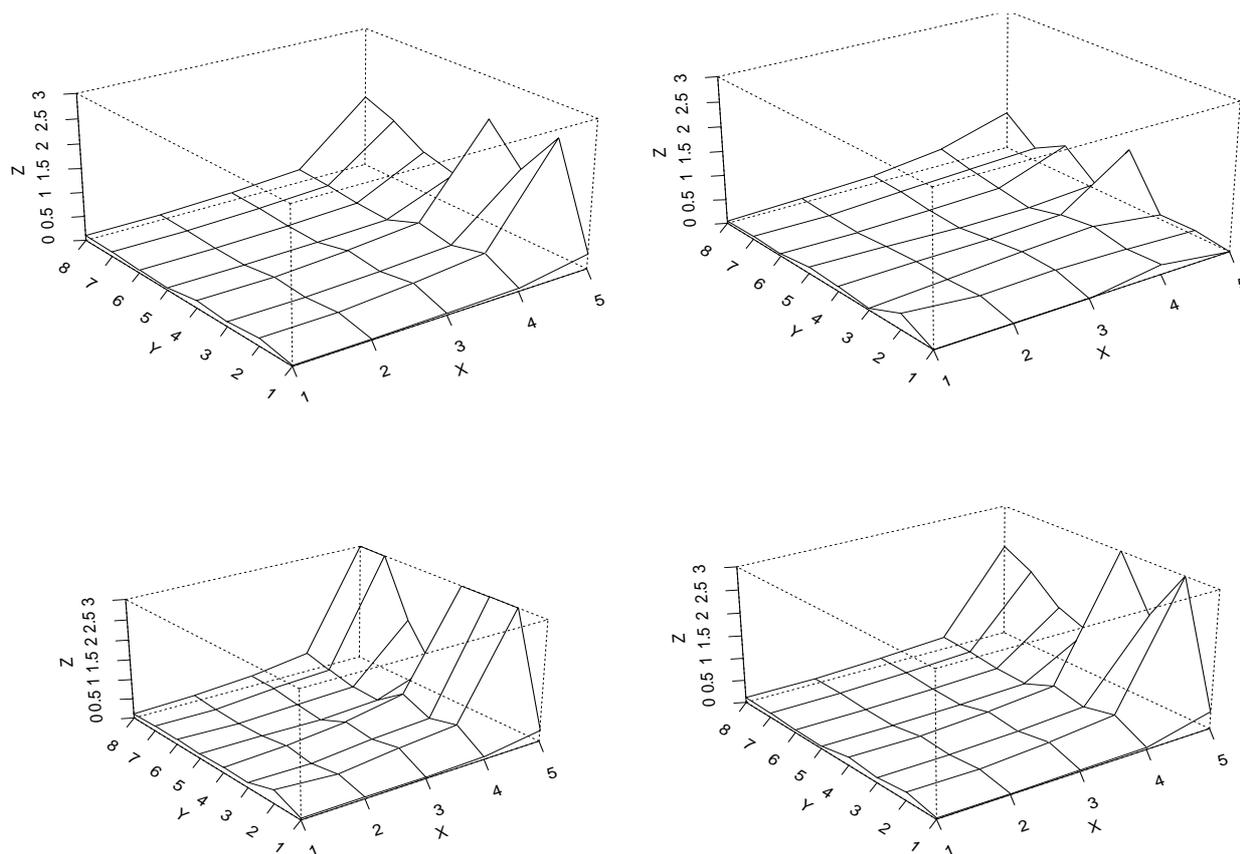


Abbildung 8.1: Hazardraten für Frauen: links oben: Coxmodell (8.4), rechts oben: empirisch, links unten: Poissonmodell mit Zeitabhängigkeit (8.3), rechts unten: Poissonmodell ohne Zeitabhängigkeit (8.5), x-Achse: 1: [0,301), 2: [301, 501), 3: [501, 831), 4: [831, 1370), 5: [1370, ∞) Tage, y-Achse: 1: stationär, <75 Jahre, 2: stationär, 75- 83 Jahre, 3: stationär, 83- 88 Jahre, 4: stationär, 88- 108 Jahre, 5: ambulant, <75 Jahre, 6: ambulant, 75- 83 Jahre, 7: ambulant, 83 - 88 Jahre, 8: ambulant, 88 - 108 Jahre

dass in diesem Zeitintervall wesentlich weniger Daten vorliegen. In Tabelle 8.4 sind für Frauen die beobachteten Todesfälle (d) und die Gesamtanzahl (n) der Individuen je Gruppe, unterteilt nach Alter, Behandlungsart und Zeitintervall, aufgeführt. Mit zunehmender Behandlungsdauer nimmt die Zahl der Beobachtungen deutlich ab. Bei der Bildung der Zeitintervalle wurde zwar darauf geachtet, dass diese in etwa gleich viele Daten enthalten. Durch die darauffolgende Erzeugung von Pseudobeobachtungen erhöhte sich jedoch vor allem in den ersten Intervallen die Zahl der Beobachtungen, während im letzten Zeitintervall keine neuen Beobachtungen entstanden.

Abb. 8.2 zeigt dieselben Plots, jedoch nur für die ersten vier Zeitintervalle, so dass ein kleinerer Maßstab gewählt werden kann. Man erkennt deutlich die große Übereinstimmung im Coxmodell (8.4) (links oben) und dem entsprechenden Poissonmodell (8.5) mit denselben Kovariablen und Interaktionen (rechts unten). Links unten sind die Hazardraten des Poissonmodells (8.3) mit Zeitabhängigkeit geplottet. Während die Hazardraten

| Dauer | [0,301) | | [301,501) | | [501,831) | | [831, 1370) | | [1370, ∞) | |
|----------------------|---------|-----|-----------|-----|-----------|----|-------------|----|-----------|----|
| Gruppe | d | n | d | n | d | n | d | n | d | n |
| stat, - 75 Jahre | 0 | 6 | 0 | 4 | 0 | 5 | 1 | 5 | 0 | 4 |
| stat, 75 - 83 Jahre | 1 | 15 | 1 | 16 | 1 | 11 | 1 | 6 | 1 | 10 |
| stat, 83 - 88 Jahre | 1 | 31 | 1 | 22 | 1 | 17 | 1 | 7 | 1 | 10 |
| stat, 88 - 108 Jahre | 1 | 25 | 1 | 33 | 1 | 30 | 1 | 19 | 1 | 2 |
| amb, - 75 Jahre | 2 | 80 | 1 | 60 | 1 | 55 | 1 | 35 | 1 | 13 |
| amb, 75 - 83 Jahre | 3 | 128 | 1 | 85 | 1 | 63 | 1 | 39 | 1 | 3 |
| amb, 83 - 88 Jahre | 1 | 120 | 1 | 111 | 1 | 66 | 1 | 36 | 1 | 14 |

Tabelle 8.4: Empirisch ermittelte Anzahl der Todesfälle (d) und der Beobachtungen (n) je Gruppe bei Frauen

im Coxmodell in jeder Gruppe linear mit der Zeit wachsen, wie es das Modell unterstellt, ist hier in einigen Gruppen eine Nichtlinearität festzustellen. Besonders deutlich wird dies in Gruppe 2 (stationär, 75 - 83 Jahre). Zu Pflegebeginn liegt die Sterberate bei etwa 0.2, im zweiten Zeitintervall fällt sie ab auf etwa 0.15 und steigt anschließend wieder stark an. Ein Vergleich mit den empirischen Hazards zeigt, dass auch hier diese Nichtlinearität vorliegt. Im letzten Zeitintervall dagegen liegen die Hazardraten des Coxmodells näher an den empirischen Raten, das Poissonmodell (8.3) unterstellt hier höhere Sterberaten. Für die Hazardraten von Männern erhält man ähnliche Ergebnisse.

Sowohl das Proportional-Hazard- als auch das Poissonmodell werden nun um die Diagnosen erweitert. Für das Proportional Hazard Modell ergibt sich die Hazardfunktion

$$\begin{aligned}
 \lambda(t) = & \lambda_0(t) \exp[\beta_1 \mathbf{Z}_{Geschlecht} + \beta_2 \mathbf{Z}_{stat} + \beta_3 \mathbf{Z}_{75-83Jahre} + \beta_4 \mathbf{Z}_{83-88Jahre} + \beta_5 \mathbf{Z}_{88-108Jahre} \\
 & + \beta_6 \mathbf{Z}_{Tumore} + \beta_7 \mathbf{Z}_{Psychose} + \beta_8 \mathbf{Z}_{Schlag} + \beta_9 \mathbf{Z}_{Herz} + \beta_{10} \mathbf{Z}_{Lungen} + \beta_{11} \mathbf{Z}_{Demenz} \\
 & + \beta_{12} (\mathbf{Z}_{Schlag} \times \mathbf{Z}_{Lungen}) + \beta_{13} (\mathbf{Z}_{Demenz} \times \mathbf{Z}_{Lungen}) + \beta_{14} (\mathbf{Z}_{Geschlecht} \times \mathbf{Z}_{Psychose}) \\
 & + \beta_{15} (\mathbf{Z}_{Geschlecht} \times \mathbf{Z}_{Herz}) + \beta_{16} (\mathbf{Z}_{75-83Jahre} \times \mathbf{Z}_{Tumore}) + \beta_{17} (\mathbf{Z}_{83-88Jahre} \times \mathbf{Z}_{Tumore}) \\
 & + \beta_{18} (\mathbf{Z}_{88-108Jahre} \times \mathbf{Z}_{Tumore}) + \beta_{19} (\mathbf{Z}_{75-83Jahre} \times \mathbf{Z}_{stat}) + \beta_{20} (\mathbf{Z}_{83-88Jahre} \times \mathbf{Z}_{stat}) \\
 & + \beta_{21} (\mathbf{Z}_{88-108Jahre} \times \mathbf{Z}_{stat})], \tag{8.6}
 \end{aligned}$$

der Log-Hazard des Poissonmodells ist gegeben durch

$$\begin{aligned}
 \log \lambda = & Intercept + \alpha_2 \mathbf{Z}_{Zeit2} + \alpha_3 \mathbf{Z}_{Zeit3} + \alpha_4 \mathbf{Z}_{Zeit4} + \alpha_5 \mathbf{Z}_{Zeit5} + \beta_1 \mathbf{Z}_{Geschlecht} + \beta_2 \mathbf{Z}_{stat} \\
 & + \beta_3 \mathbf{Z}_{75-83Jahre} + \beta_4 \mathbf{Z}_{83-88Jahre} + \beta_5 \mathbf{Z}_{88-108Jahre} + \beta_6 \mathbf{Z}_{Tumore} + \beta_7 \mathbf{Z}_{Psychose} \\
 & + \beta_8 \mathbf{Z}_{Schlag} + \beta_9 \mathbf{Z}_{Herz} + \beta_{10} \mathbf{Z}_{Lungen} + \beta_{11} (\mathbf{Z}_{Schlag} \times \mathbf{Z}_{Lungen}) \\
 & + \beta_{12} (\mathbf{Z}_{Geschlecht} \times \mathbf{Z}_{Psychose}) + \beta_{13} (\mathbf{Z}_{75-83Jahre} \times \mathbf{Z}_{Tumore}) \\
 & + \beta_{14} (\mathbf{Z}_{83-88Jahre} \times \mathbf{Z}_{Tumore}) + \beta_{15} (\mathbf{Z}_{88-108Jahre} \times \mathbf{Z}_{Tumore}) \\
 & + \beta_{16} (\mathbf{Z}_{75-83Jahre} \times \mathbf{Z}_{stat}) + \beta_{17} (\mathbf{Z}_{83-88Jahre} \times \mathbf{Z}_{stat}) + \beta_{18} (\mathbf{Z}_{88-108Jahre} \times \mathbf{Z}_{stat}) \\
 & + \beta_{19} (\mathbf{Z}_{75-83Jahre} \times \mathbf{Z}_{Zeit2}) + \beta_{20} (\mathbf{Z}_{83-88Jahre} \times \mathbf{Z}_{Zeit2}) + \beta_{21} (\mathbf{Z}_{88-108Jahre} \times \mathbf{Z}_{Zeit2}) \\
 & + \beta_{22} (\mathbf{Z}_{75-83Jahre} \times \mathbf{Z}_{Zeit3}) + \beta_{23} (\mathbf{Z}_{83-88Jahre} \times \mathbf{Z}_{Zeit3}) + \beta_{24} (\mathbf{Z}_{88-108Jahre} \times \mathbf{Z}_{Zeit3})
 \end{aligned}$$

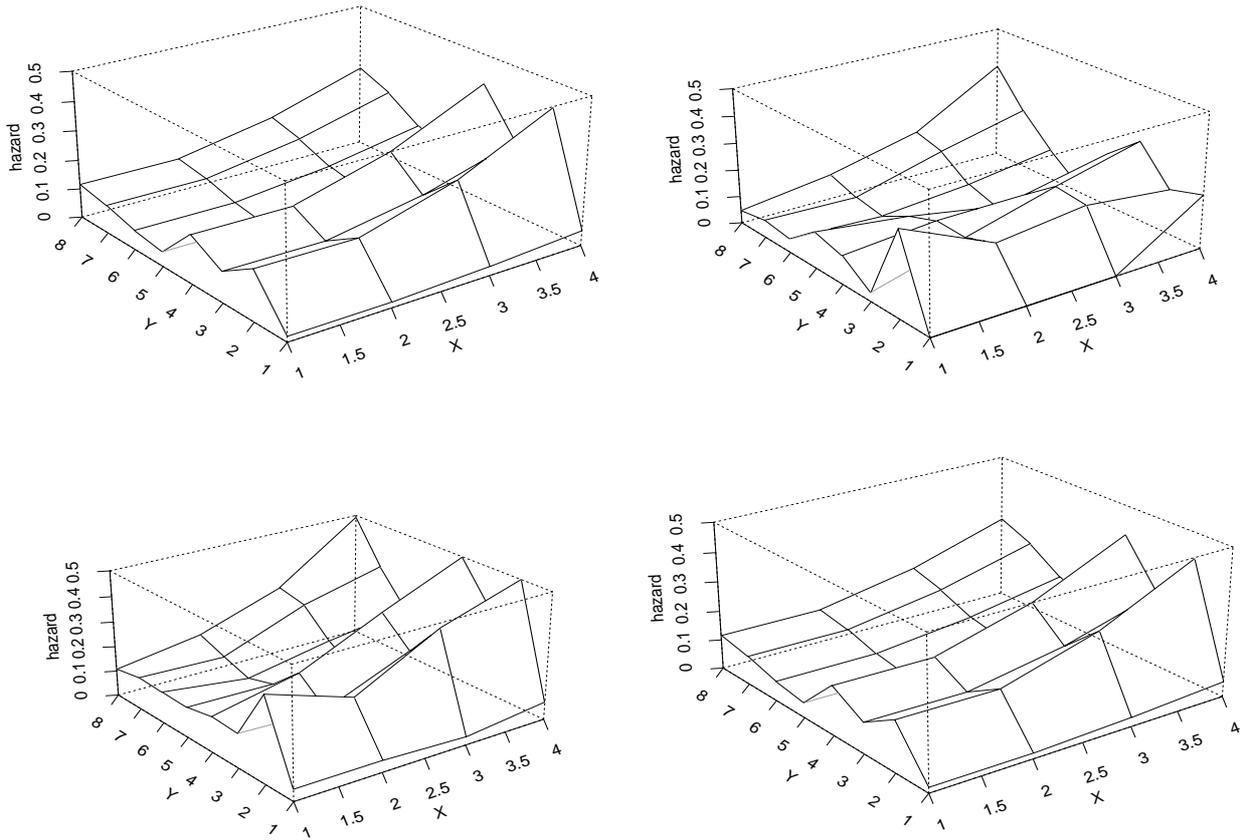


Abbildung 8.2: Hazardraten für Frauen in den ersten vier Zeitintervallen: links oben: Coxmodell (8.4), rechts oben: empirisch, links unten: Poissonmodell mit Zeitabhängigkeit (8.3), rechts unten: Poissonmodell ohne Zeitabhängigkeit (8.5), Legende siehe Abb. 8.1

$$\begin{aligned}
 & + \beta_{25}(\mathbf{Z}_{75-83\text{Jahre}} \times \mathbf{Z}_{\text{Zeit4}}) + \beta_{26}(\mathbf{Z}_{83-88\text{Jahre}} \times \mathbf{Z}_{\text{Zeit4}}) + \beta_{27}(\mathbf{Z}_{88-108\text{Jahre}} \times \mathbf{Z}_{\text{Zeit4}}) \\
 & + \beta_{28}(\mathbf{Z}_{75-83\text{Jahre}} \times \mathbf{Z}_{\text{Zeit5}}) + \beta_{29}(\mathbf{Z}_{83-88\text{Jahre}} \times \mathbf{Z}_{\text{Zeit5}}) + \beta_{30}(\mathbf{Z}_{88-108\text{Jahre}} \times \mathbf{Z}_{\text{Zeit5}}) \\
 & + \beta_{31}(\mathbf{Z}_{\text{Tumore}} \times \mathbf{Z}_{\text{Zeit2}}) + \beta_{32}(\mathbf{Z}_{\text{Tumore}} \times \mathbf{Z}_{\text{Zeit3}}) + \beta_{33}(\mathbf{Z}_{\text{Tumore}} \times \mathbf{Z}_{\text{Zeit4}}) \\
 & + \beta_{34}(\mathbf{Z}_{\text{Tumore}} \times \mathbf{Z}_{\text{Zeit5}}) + \beta_{35}(\mathbf{Z}_{\text{Schlag}} \times \mathbf{Z}_{\text{Zeit2}}) + \beta_{36}(\mathbf{Z}_{\text{Schlag}} \times \mathbf{Z}_{\text{Zeit3}}) \\
 & + \beta_{37}(\mathbf{Z}_{\text{Schlag}} \times \mathbf{Z}_{\text{Zeit4}}) + \beta_{38}(\mathbf{Z}_{\text{Schlag}} \times \mathbf{Z}_{\text{Zeit5}}). \tag{8.7}
 \end{aligned}$$

In beiden Modellen besitzen die Diagnosen Tumore, Psychose, Schlaganfälle, Herz-, und Lungenerkrankungen signifikanten Einfluß. Zwischen Schlaganfällen und Lungenerkrankungen liegen Wechselwirkungen vor. Demenzerkrankungen gehen nur in das Coxmodell ein. Auch Interaktionen zwischen dem Alter mit Tumoren und der Behandlungsart als auch zwischen Geschlecht und Psychose sind in beiden Modellen festzustellen. Während im Coxmodell zusätzlich Interaktionen zwischen Demenz- und Lungenerkrankungen als auch zwischen Geschlecht und Herzerkrankungen bestehen, werden im Poissonmodell das Alter, Tumore und Schlaganfälle zeitabhängig modelliert. Auch hier werden die Hazardraten der beiden Modelle mit den empirischen Sterberaten verglichen. Wie bei der Berechnung der Sterbewahrscheinlichkeiten, muss über die Sterberaten in den verschiedenen

Diagnosegruppen (siehe Kapitel (7.3)) gemittelt werden. Die Gewichte werden wie in Kapitel (7.3) bestimmt, jedoch nun mit den im Modell verwendeten Zeitintervallen der Pflegedauer. Auch hier liegen die Sterberaten der Modelle im letzten Zeitintervall weit über den empirisch ermittelten. Die Hazardraten für Frauen bzw. Männer in den ersten vier Zeitintervallen sind in den Abb. 8.3 und 8.4 dargestellt.

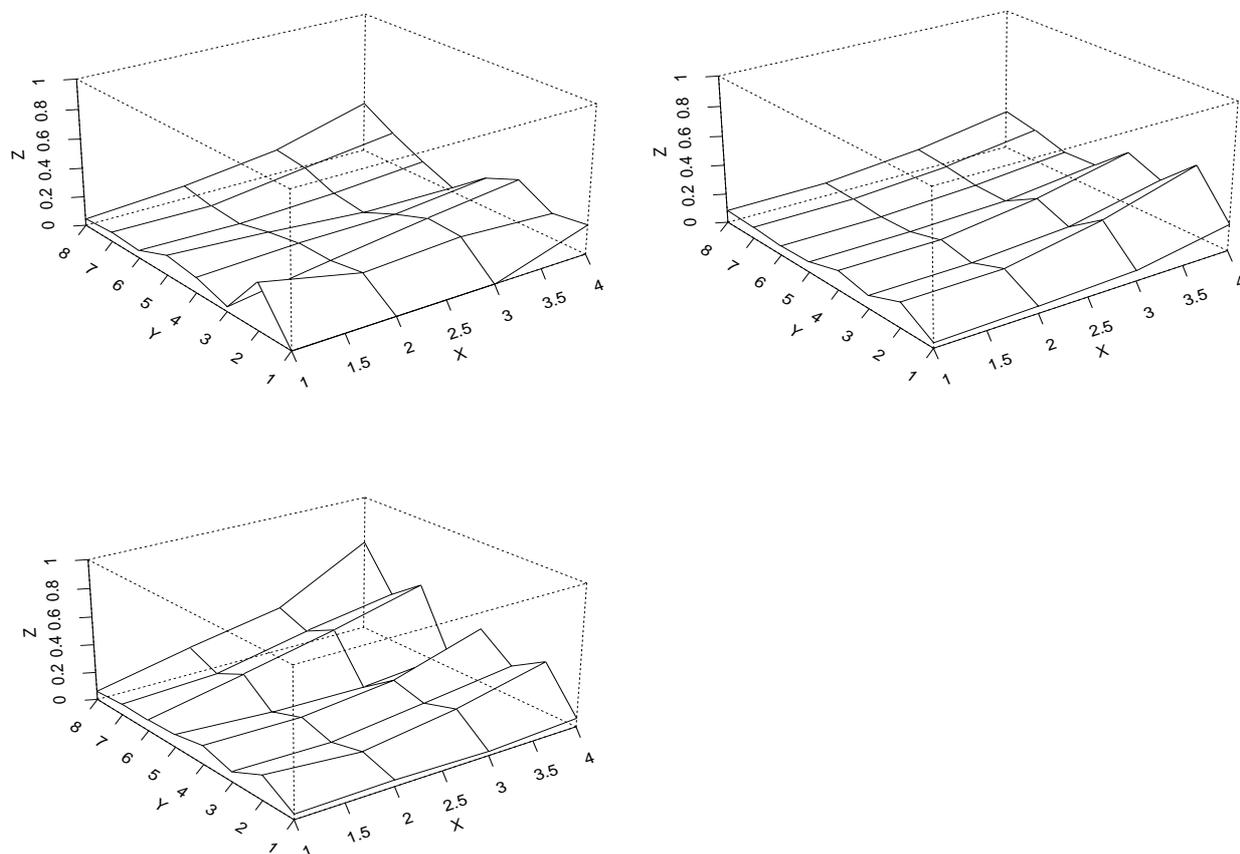


Abbildung 8.3: Hazardraten für Frauen in den ersten vier Zeitintervallen, gemittelt über die Diagnosen: links oben: empirisch, rechts oben: Coxmodell (8.6), links unten: Poissonmodell (8.7), Legende siehe Abb. 8.1

Der nichtlineare Verlauf der Raten in Gruppe 2 ist wieder im Poissonmodell besser getroffen. Bei den Frauen (siehe Abb. 8.3) ist dies nur leicht, bei den Männern (siehe Abb. 8.4) jedoch sehr deutlich erkennbar. In den Gruppen 5-7 dagegen erscheint die Modellierung des Coxmodells sowohl bei den Frauen als auch bei den Männern besser. In den Abb. 8.5 und 8.6 findet man einen Vergleich der Hazardraten für einzelne Diagnosegruppen. Abb. 8.5 stellt die Sterberaten für Frauen mit Demenz dar. Hier scheint das Poissonmodell in den Gruppen 6 und 8 die empirischen Raten besser zu treffen. Bei Männern mit Tumor (Abb. 8.6) läßt sich keine genaue Aussage treffen. Die empirischen Sterberaten sind hier mit Vorsicht zu genießen, da die Anzahl der zugrundeliegenden Daten sehr gering ist.

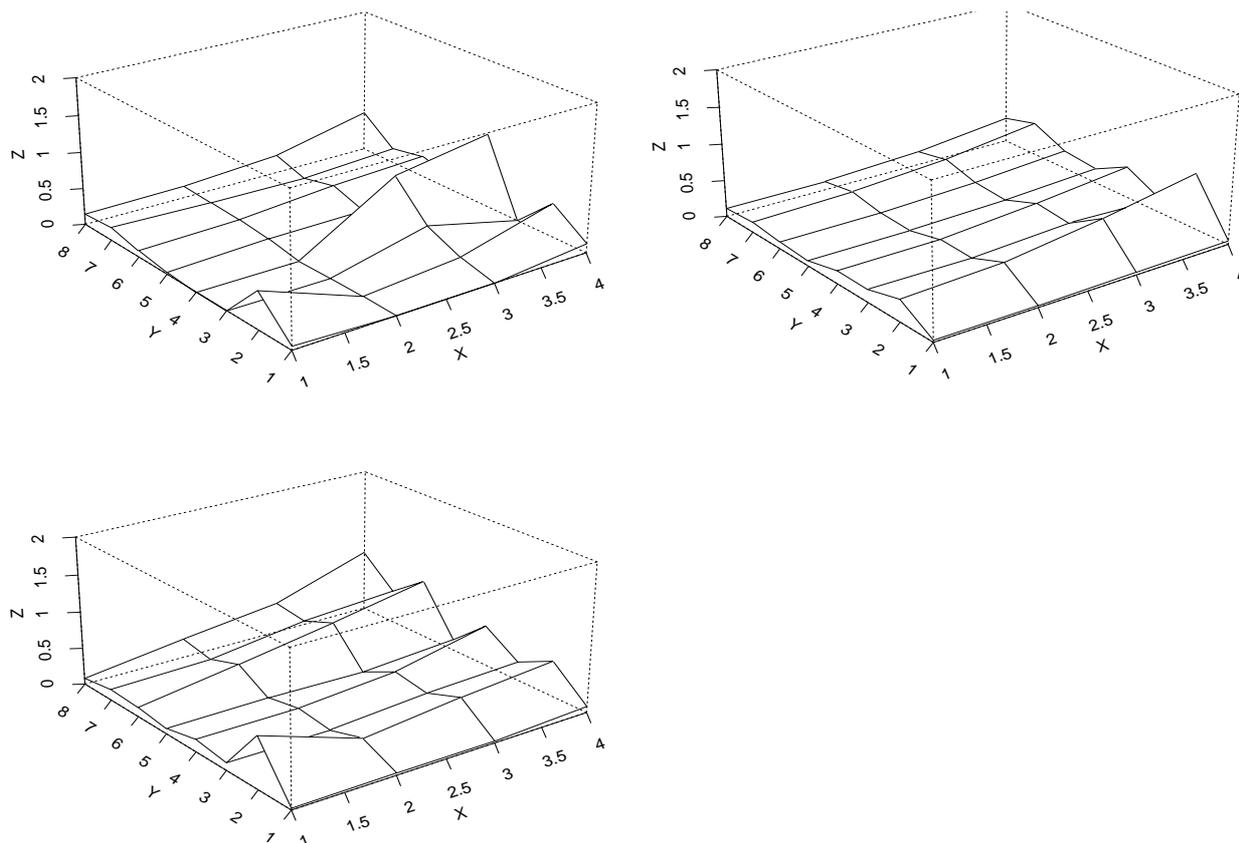


Abbildung 8.4: Hazardraten für Männer in den ersten vier Zeitintervallen, gemittelt über die Diagnosen: links oben: empirisch, rechts oben: Coxmodell (8.6), links unten: Poissonmodell (8.7), Legende siehe Abb. 8.1

Vergleich von Poisson- und Proportional Hazard Modell

Die Plots der Hazardraten zeigten, dass ein direkter Vergleich der Modelle sehr schwierig ist. Dennoch lassen sich verschiedene Vor- und Nachteile beider Modelle aufzeigen.

- Eine offensichtliche Einschränkung im Poissonmodell ist die Annahme, dass die Hazardraten intervallweise konstant sind. Grundlegend bei der Anpassung eines Poissonmodells ist die Wahl dieser Zeitintervalle. Eine Unterteilung der Pflegedauer, welche nach Bildung der Pseudobeobachtungen eine gleichmässige Aufteilung der Beobachtungen auf die Intervalle gewährleistet, hätte bei dem betrachteten Beispiel des vorangegangenen Abschnittes möglicherweise ganz andere Hazardraten zur Folge gehabt. Um dies zu untersuchen, hätten die Daten jedoch zuerst neu bearbeitet und dann ein neues Modell angepasst werden müssen. Im Coxmodell dagegen genügt es, ein einziges Modell zu fiten. In diesem Modell ist es dann ohne großen Aufwand möglich, den Einfluß verschiedener Zeitintervalle auf die Hazardraten zu testen.
- Eine wichtige Rolle spielt die Zeitunterteilung auch bei der Berechnung von Über-

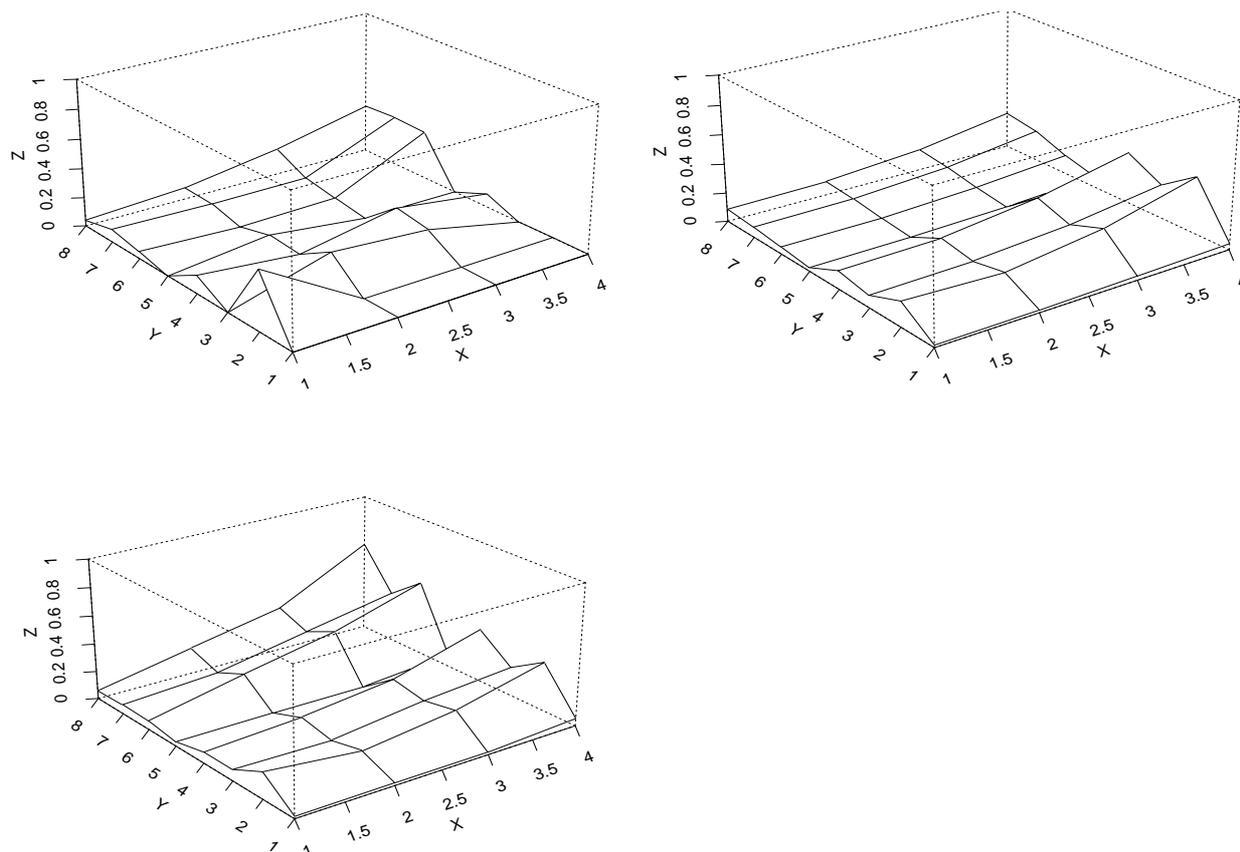


Abbildung 8.5: Hazardraten für Frauen mit Demenz in den ersten vier Zeitintervallen mit Diagnosen: links oben: empirisch, rechts oben: Coxmodell (8.6), links unten: Poissonmodell (8.7), Legende siehe Abb. 8.1

gangswahrscheinlichkeiten. Versicherungen sind an einjährigen Wahrscheinlichkeiten interessiert. Sollen diese mittels eines Poissonmodells ermittelt werden, ist es notwendig, auch die Intervallgrenzen der Pflegedauer jährlich zu wählen. Dies bedeutet eine starke Einschränkung.

- Das Proportional Hazard Modell unterstellt einen mit der Beobachtungsdauer wachsenden Baselinehazard. In den vorher gezeigten Plots wurde das sehr deutlich. Im Poissonmodell sind die Baselinehazards der verschiedenen Intervalle dagegen unabhängig voneinander, so daß auch ein nicht monotoner Verlauf möglich ist.
- Eng damit verbunden ist ein weiterer Vorteil des Poissonmodells gegenüber dem Proportional Hazard Modell : die einfach durchführbare Modellierung von zeitabhängigen Effekten. Im Coxmodell ist die Modellierung von Zeitabhängigkeit zwar auch möglich, doch erfordert sie hier wesentlich mehr Aufwand und ist bei großen Datensätzen teilweise kaum durchführbar, wie z. B. die Methode der fraktionierten Polynome. Doch auch in diesem Punkt hängt die Güte des Poissonmodells stark von den gewählten Zeitintervallen ab. Ändert sich die Art der Zeitabhängigkeit rasch,

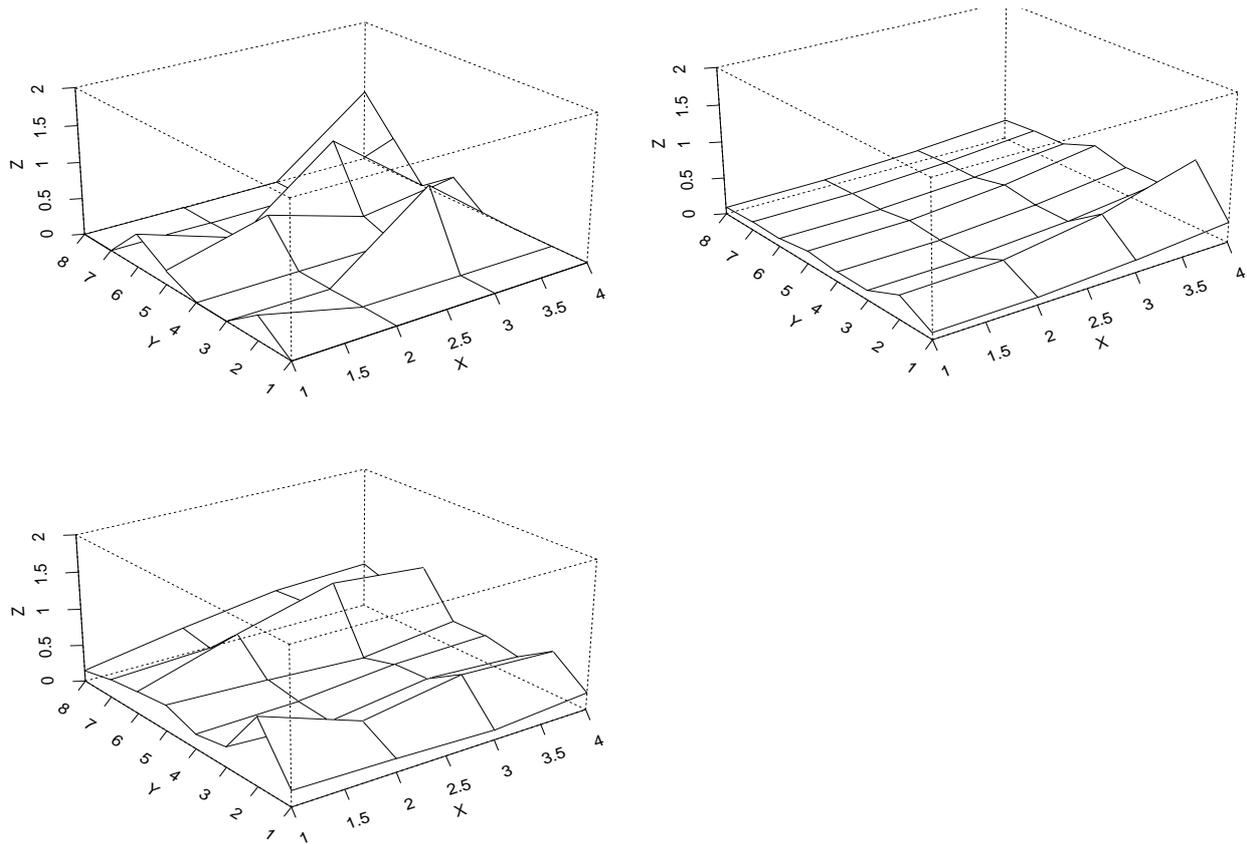


Abbildung 8.6: Hazardraten für Männer mit Tumor in den ersten vier Zeitintervallen mit Diagnosen: links oben: empirisch, rechts oben: Coxmodell (8.6), links unten: Poissonmodell (8.7), Legende siehe Abb. 8.1

kann bei kleineren Intervallen eine wesentlich bessere Modellierung erfolgen, als bei wenigen großen.

Kapitel 9

Zusammenfassung

Wir wollen die Ergebnisse dieser Arbeit noch einmal kurz zusammenfassen. Aufbauend auf der Diplomarbeit von Rudolph [23] wurde an einen Datensatz aus der Pflegeversicherung ein Proportional Hazard Modell angepaßt. Dabei ging es in erster Linie darum, die pflegeauslösenden Diagnosen als zusätzliche Kovariablen in das Modell mit einzubauen. Durch die Diagnosen und Beachtung der bestehenden Wechselwirkungen zu anderen Kovariablen wurde eine große Verbesserung des Modells erzielt.

Bei der Überprüfung der Modellannahmen wurde festgestellt, dass die funktionale Form des Einflusses des Alters genauer untersucht werden musste. Dazu wurden verschiedene Methoden auf den Datensatz angewandt. Neben einer Modellierung mittels Knickpunkte, wurde das Alter auch durch einen exponentiellen Ansatz und fraktionale Polynome modelliert. Es wurde deutlich, dass ein enger Zusammenhang zwischen dem Verlauf des Alters und den Diagnosen vorliegt, so dass diese Verfahren auf einzelne Diagnosegruppen getrennt angewandt wurden. Die sinnvollsten Ergebnisse wurden durch fraktionale Polynome und exponentielle Modellierung erhalten.

Darüberhinaus zeigt sich, dass eine starke Zeitabhängigkeit im Modell vorlag. Durch Teilung des Datensatzes wurde erreicht, dass zumindest für Pflegedauern über zweieinhalb Jahren die Annahme der proportionalen Hazards erfüllt ist. Im Modell der ersten 900 Pflegetage dagegen wiesen die Pflegestufe und die Diagnose Demenz weiterhin zeitabhängige Effekte auf. Eine zeitstetige Modellierung mittels fraktionaler Polynome konnte aufgrund der Größe des Datensatzes nicht durchgeführt werden, wurde aber zur Illustration der Methode an einem kleinen Datensatz demonstriert.

An die Modellbildung schloss sich die Berechnung einjähriger Übergangswahrscheinlichkeiten auf Basis der beiden gefundenen Modelle an. Zum Vergleich wurden sowohl die aus dem Ausgangsmodell ohne Diagnosen resultierenden Wahrscheinlichkeiten als auch die empirischen Wahrscheinlichkeiten betrachtet. Auch hier wurde deutlich, dass die Einbeziehung der Diagnosen zu einer Verbesserung beigetragen hatte.

Schließlich wurde ein Teil des Datensatzes mittels eines Poissonmodells untersucht. Die Hazardraten wurden ermittelt und mit denen eines Proportional Hazard Modells verglichen. Auch hier wurden als Vergleichsgröße die empirischen Hazardraten herangezogen. Der Vergleich erwies sich als schwierig. Vor allem in frühen Behandlungsdauern lieferte das Poissonmodell bessere Ergebnisse. Hier machte sich bemerkbar, dass einige Kovariablen zeitabhängig modelliert worden waren. In anderen Gruppen dagegen schien die Modellie-

rung des Coxmodells besser. Verbesserungen des Poissonmodells könnten eventuell durch eine andere Wahl der Zeitintervalle, innerhalb welcher der Baselinehazard als konstant angenommen wird, erreicht werden.

Auch wenn die Güte des Poissonmodells stark von der verwendeten Gruppierung der Daten abhängt und die stückweise konstanten Baselinehazards eine Einschränkung darstellen, ermöglicht es doch eine einfache Modellierung von zeitabhängigen Effekten.

Ziel einer weiteren Untersuchung der Daten könnte daher die Anwendung eines Poissonmodells auf den kompletten Datensatz sein, mit besonderem Augenmerk auf die korrekte Modellierung der immer noch vorhandenen Zeitabhängigkeit.

Literaturverzeichnis

- [1] H. Akaike. Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *2nd International Symposium on Information Theory* (pp. 267-281). Budapest: Akademiai Kiado, 1973
- [2] P. K. Andersen, O. Borgan, R. D. Gill, N. Keiding. *Statistical Models based on Counting Processes*. Springer, 1992
- [3] U. Berger, P. Gerein, K. Ulm, J. Schäfer. On the use of fractional polynomials in dynamic Cox Models, *SFB386: Discussion Paper 207*. Ludwig-Maximilians-Universität München, 2000.
- [4] P. S. Bickel, C. A. Klaassen, Y. Ritov & J. A. Wellner. *Efficient and Adaptive Estimation in Semiparametric Models*. Baltimore: Johns Hopkins University Press, 1989
- [5] K. P. Burnham, D. R. Anderson. *Model selection and inference. A practical information-theoretic approach*. Springer, 1998
- [6] D. R. Cox. Partial Likelihood. *Biometrika*, 62, 1975, pp.269-276
- [7] B. Efron. The Efficiency of Coxs Likelihood Function for Censored Data. *Journal of the American Statistical Association*, 72, 1977, pp. 557-565
- [8] T. R. Fleming and D. P. Harrington. *Counting Processes and Survival Analysis*. Wiley Series in Probability and mathematical Statistics, 1991
- [9] P. M. Grambsch and T. M. Therneau. Proportional hazards tests and diagnostics based on weighted residuals, *Biometrika*, 81, 3, 1994, pp. 515-526
- [10] S. Habermann and E. Pitacco. *Actuarial Models for Disability Insurance*. Chapman & Hall/CRC, 1999
- [11] T. R. Holford. The Analysis of Rates and of Survivorship Using Log-Linear Models. *Biometrics*, 36, 1980, pp.299-305
- [12] D.W. Hosmer and JR., S. Lemeshow. *Applied survival analysis*. Wiley Series in Probability and Statistics, New York, 1999
- [13] P. Hougaard. *Analysis of Multivariate Survival Data*. Springer, 2000
- [14] J.P. Klein and M.L. Moeschberger. *Survival Analysis*. Springer, New York, 1997

- [15] P. E. Kopp. *Martingales and stochastic integrals*. Cambridge University Press, 1984
- [16] N. Laird and D. Olivier. Covariance Analysis of Censored Survival Data Using Log-Linear Analysis Techniques. *Journal of the American Statistical Association* , Volume 76, Issue 374, 1981, pp.231-240
- [17] P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman and Hall, Second Edition.
- [18] R. G. Miller Jr. *Survival Analysis* . Wiley Series in Probability and mathematical Statistics: Applied Probability and Statistics, 1981
- [19] R. H. Myers. *Classical and modern regression with applications*. PWS-KENT Publishing Company, Boston, 1989
- [20] C. R. Rao. *Linear statistical inference and its applications*. John Wiley & Sons, 1973
- [21] A. E. Renshaw. Modelling Excess Mortality using GLIM, 1988
- [22] P. Royston and D.G. Altman. Regression using Fractional Polynomials of Continuous Covariates: Parsimonious Parametric Modelling. *Applied Statistics*, 43, 1994, pp.429-467
- [23] F. Rudolph. *Anwendungen der Überlebenszeitanalyse in der Pflegeversicherung*. Diplomarbeit, 2000
- [24] D. Schoenfeld. Partial residuals for the proportional hazards regression model, *Biometrika*, 69, 1, 1982, pp.239-241
- [25] *S-Plus 5 for UNIX*. Math Soft,Inc, 1998
- [26] T. M. Therneau, P. M. Gramsch. *Modeling Survival Data: Extending the Cox Model*. Springer, Ney York, 2000
- [27] T. M. Therneau, P. M. Gramsch, T. R. Fleming. Martingale-based residuals for survival models, *Biometrika*, 77, 1, 1990, pp.147-160
- [28] W.N. Venables, B.D. Ripley. *Modern Applied Statistics with S-Plus*. Springer, 1996
- [29] H. von Weizsäcker, G. Winkler. *Stochastic Integrals*. Friedr. Vieweg & Sohn, Braunschweig, 1990

Anhang

Die Funktionen *fracpol1.s* und *fracpol2.s* wurden zur Bestimmung der funktionalen Form des Alters in den verschiedenen Diagnosegruppen mittels fraktionaler Polynome verwendet. *fracpol1.s* fittet Modelle mit allen fraktionalen Polynomen ersten Grades, *fracpol2.s* mit allen möglichen fraktionalen Polynomen zweiten Grades. Hier sind beide Funktionen exemplarisch für die Modellierung des Alters in der Gruppe Frauen mit Tumor (tw) im Modell (6.8) für Behandlungszeiten über 900 Tage angegeben. Die Kovariable *nds* gibt dabei an, wie viele Diagnosen bei einem Patienten vorliegen.

```
fracpol1_function(daten) {
  attach(daten) result_rep(1,7) P1_cbind(-2,-1,-0.5,0,0.5,1,2)

  tw_rep(0,length(daten[,1]))
  tw[daten$Tumore==1&daten$Zsex=="w"&daten$nds==1]_1
  ntw_1-tw
  ntwAge_daten$Age*ntw

  for(i in 1:7) {
    print(i)
    p1_P1[i]
    if(p1==0) { Age1_log(daten$Age)}
    else { Age1_daten$Age^p1}

    temp_data.frame(daten,Age1)

    result[i]_coxph(Surv(LTCbegin, LTCend, Death) ~Age1:tw+
      ntwAge* Level + Zsex * (Art + Psychose) + Art * (Level + Psychose)
      + Level * Herz +Psychose * Schlag+Tumore,temp)$loglik[2]

    print(result[i])
  }

  result
}
```

```

fracpol2_function(daten)
{
attach(daten)
result_matrix(1,7,7)
P1_cbind(-2,-1,-0.5,0,0.5,1,2)
P2_cbind(-2,-1,-0.5,0,0.5,1,2)

tw_rep(0,length(daten[,1]))
tw[daten$Tumore==1&daten$Zsex=="w"&daten$nds==1]_1
ntw_1-tw
ntwAge_daten$Age*ntw

for(i in 1:7){
  for(j in 1:7){
    print(i)
    print(j)
    p1_P1[i]
    p2_P2[j]

    if(p1==0) { Age1_log(daten$Age)}
    else {Age1_daten$Age^p1}

    if(p2==0) { Age2_log(daten$Age)}
    else {Age2_daten$Age^p2}

    if(p1==p2) {Age2_Age1*log(daten$Age)}

    temp_data.frame(daten, Age1, Age2)

    result[i,j]_coxph(Surv(LTCbegin, LTCend, Death) ~Age1:tw+Age2:tw+
      ntwAge* Level + Zsex * (Art + Psychose) + Art * (Level + Psychose)
      + Level * Herz +Psychose * Schlag+Tumore,temp)$loglik[2]

    print(result[i,j])

  }
}
result
}

```

Mittels der Funktion *punkt.s* wurden aus den Sterbeintensitäten die Sterbewahrscheinlichkeiten im Modell mit Diagnosen berechnet, siehe Kapitel 7.3, (7.9). Die Unterfunktion *pfunktion.s* liefert als Ausgabe die einzelnen Sterbewahrscheinlichkeiten der 22 Diagnose-

gruppen, sowohl bei ambulanter als auch stationärer Behandlung. Die Funktion *gewich-temat.s* stellt die dazugehörigen Gewichte, die zur Mittelung der Wahrscheinlichkeiten benötigt werden zur Verfügung. Zur Berechnung der Sterbeintensitäten im Modell ohne Diagnose und der Stufenübergangswahrscheinlichkeiten wurden entsprechende Funktionen verwendet.

```

pwkt_function(fit,s, x ,l,d1,d2,steps) {
xmat_pfunction(fit,s,x,l,d1,d2)

gew_gewichtemat(x,steps,l,s,d1,d2)
na_sum(gew[,1])
ns_sum(gew[,2])

pe_1/(na+ns)*(gew[,1]%%xmat[1,]+gew[,2]%%xmat[2,])

return(pe)
}

```

Die Funktion *pfunction.s* ruft die Funktion *wkten.s* zur Berechnung der Wahrscheinlichkeiten (nach der Formel (7.8), Kapitel 7.3) der einzelnen Diagnosegruppen auf. *d1* und *d2* geben dabei die Grenzen des Zeitintervalls an, für welches die Sterbewahrscheinlichkeiten berechnet werden.

```

wkten_function(x, t, d1, d2) {
  n <- length(t)
  j <- 0
  l <- 0
  for(i in 1:n) {
    if(t[i] < d1)
      l <- l + 1
  }
  for(i in 1:n) {
    if(t[i] < d2)
      j <- j + 1
  }

  hazard_rep(0,n)
  for(i in 2:n){hazard[i]_1-x[i]/x[i-1]}
  sum_hazard[l+1]
  prod_1
  for(i in (l+2):j){
    prod <- prod * (1 - hazard[i-1])
    sum <- sum + hazard[i] * prod
  }
  return(sum)
}

```

```

pfunction_function(fit,s,x,l,d1,d2)
{
pv_matrix(0,2,22)

for(iin1:2)
{
a1_survfit(fit,newdata_data.frame(Age=x,twyoungAge=0,nphdmtwAge=x,
exp0022phdm=0,ntwAge=x,twjungAge=0,Zsex=s,Art=i-1,Tumore=0,Demenz=1,
Psychose=0,Herz=0,Arthrosen=0,Lungen=0,Schlag=0,Level=nlcdata$Level[5-1],
logAgetw=0))

pv[i,1]_wkten(a1$surv,a1$time,d1,d2)

a2_survfit(fit,newdata_data.frame(Age=x,twyoungAge=0,nphdmtwAge=x,
exp0022phdm=0,ntwAge=x,twjungAge=0,Zsex=s,Art=i-1,Tumore=0,Demenz=0,
Psychose=0,Herz=0,Arthrosen=0,Lungen=0,Schlag=1,
Level=nlcdata$Level[5-1],logAgetw=0))

pv[i,2]_wkten(a2$surv,a2$time,d1,d2)

a3_survfit(fit,newdata_data.frame(Age=x,twyoungAge=0,nphdmtwAge=x*s,
exp0022phdm=exp(0.022*x)*(1-s),ntwAge=x,twjungAge=0,Zsex=s,Art=i-1,
Tumore=0,Demenz=0,Psychose=1,Herz=0,Arthrosen=0,Lungen=0,Schlag=0,
Level=nlcdata$Level[5-1],logAgetw=0))

pv[i,3]_wkten(a3$surv,a3$time,d1,d2)

a4_survfit(fit,newdata_data.frame(Age=x,twyoungAge=0,nphdmtwAge=x*(1-s),
exp0022phdm=0,ntwAge=x*(1-s),twjungAge=0,Zsex=s,Art=i-1,Tumore=1,Demenz=0,
Psychose=0,Herz=0,Arthrosen=0,Lungen=0,Schlag=0,Level=nlcdata$Level[5-1],
logAgetw=s))

pv[i,4]_wkten(a4$surv,a4$time,d1,d2)

a5_survfit(fit,newdata_data.frame(Age=x,twyoungAge=0,nphdmtwAge=x,
exp0022phdm=0,ntwAge=x,twjungAge=0,Zsex=s,Art=i-1,Tumore=0,Demenz=0,
Psychose=0,Herz=0,Arthrosen=1,Lungen=0,Schlag=0,Level=nlcdata$Level[5-1],
logAgetw=0))

pv[i,5]_wkten(a5$surv,a5$time,d1,d2)

a6_survfit(fit,newdata_data.frame(Age=x,twyoungAge=0,nphdmtwAge=x,
exp0022phdm=0,ntwAge=x,twjungAge=0,Zsex=s,Art=i-1,Tumore=0,Demenz=0,

```

```

Psychose=0,Herz=1,Arthrosen=0,Lungen=0,Schlag=0,Level=nlcdata$Level[5-1],
logAgetw=0))

pv[i,6]_wkten(a6$surv,a6$time,d1,d2)

a7_survfit(fit,newdata_data.frame(Age=x,twyoungAge=0,nphdmtwAge=x,
exp0022phdm=0,ntwAge=x,twjungAge=0,Zsex=s,Art=i-1,Tumore=0,Demenz=0,
Psychose=0,Herz=0,Arthrosen=0,Lungen=1,Schlag=0,Level=nlcdata$Level[5-1],
logAgetw=0))

pv[i,7]_wkten(a7$surv,a7$time,d1,d2)

a8_survfit(fit,newdata_data.frame(Age=x,twyoungAge=0,nphdmtwAge=x,
exp0022phdm=0,ntwAge=x,twjungAge=0,Zsex=s,Art=i-1,Tumore=0,Demenz=1,
Psychose=0,Herz=0,Arthrosen=0,Lungen=0,Schlag=1,Level=nlcdata$Level[5-1],
logAgetw=0))

pv[i,8]_wkten(a8$surv,a8$time,d1,d2)

a9_survfit(fit,newdata_data.frame(Age=x,twyoungAge=0,nphdmtwAge=x,
exp0022phdm=0,ntwAge=x,twjungAge=0,Zsex=s,Art=i-1,Tumore=0,Demenz=1,
Psychose=0,Herz=1,Arthrosen=0,Lungen=0,Schlag=0,Level=nlcdata$Level[5-1],
logAgetw=0))

pv[i,9]_wkten(a9$surv,a9$time,d1,d2)

a10_survfit(fit,newdata_data.frame(Age=x,twyoungAge=0,nphdmtwAge=x,
exp0022phdm=0,ntwAge=x,twjungAge=0,Zsex=s,Art=i-1,Tumore=0,Demenz=1,
Psychose=0,Herz=0,Arthrosen=1,Lungen=0,Schlag=0,Level=nlcdata$Level[5-1],
logAgetw=0))

pv[i,10]_wkten(a10$surv,a10$time,d1,d2)

a11_survfit(fit,newdata_data.frame(Age=x,twyoungAge=0,nphdmtwAge=x*s,
exp0022phdm=exp(0.022*x)*(1-s),ntwAge=x,twjungAge=0,Zsex=s,Art=i-1,
Tumore=0,Demenz=1,Psychose=1,Herz=0,Arthrosen=0,Lungen=0,Schlag=0,
Level=nlcdata$Level[5-1],logAgetw=0))

pv[i,11]_wkten(a11$surv,a11$time,d1,d2)

a12_survfit(fit,newdata_data.frame(Age=x,twyoungAge=0,nphdmtwAge=x,
exp0022phdm=0,ntwAge=x,twjungAge=0,Zsex=s,Art=i-1,Tumore=1,Demenz=1,
Psychose=0,Herz=0,Arthrosen=0,Lungen=0,Schlag=0,Level=nlcdata$Level[5-1],
logAgetw=0))

```

```

pv[i, 12]_wkten(a12$surv, a12$time, d1, d2)

a13_survfit(fit, newdata_data.frame(Age=x, tyoungAge=0, nphdmtwAge=x,
exp0022phdm=0, ntwAge=x, twjungAge=0, Zsex=s, Art=i-1, Tumore=1, Demenz=0,
Psychose=1, Herz=0, Arthrosen=0, Lungen=0, Schlag=0, Level=nlcdata$Level[5-1],
logAgetw=0))

pv[i, 13]_wkten(a13$surv, a13$time, d1, d2)

a14_survfit(fit, newdata_data.frame(Age=x, tyoungAge=0, nphdmtwAge=x,
exp0022phdm=0, ntwAge=x, twjungAge=0, Zsex=s, Art=i-1, Tumore=1, Demenz=0,
Psychose=0, Herz=1, Arthrosen=0, Lungen=0, Schlag=0, Level=nlcdata$Level[5-1],
logAgetw=0))

pv[i, 14]_wkten(a14$surv, a14$time, d1, d2)

a15_survfit(fit, newdata_data.frame(Age=x, tyoungAge=0, nphdmtwAge=x*s,
exp0022phdm=exp(0.022*x)*(1-s), ntwAge=x, twjungAge=0, Zsex=s, Art=i-1,
Tumore=0, Demenz=0, Psychose=1, Herz=1, Arthrosen=0, Lungen=0, Schlag=0,
Level=nlcdata$Level[5-1], logAgetw=0))

pv[i, 15]_wkten(a15$surv, a15$time, d1, d2)

a16_survfit(fit, newdata_data.frame(Age=x, tyoungAge=0, nphdmtwAge=x,
exp0022phdm=0, ntwAge=x, twjungAge=0, Zsex=s, Art=i-1, Tumore=0, Demenz=0,
Psychose=1, Herz=0, Arthrosen=0, Lungen=0, Schlag=1, Level=nlcdata$Level[5-1],
logAgetw=0))

pv[i, 16]_wkten(a16$surv, a16$time, d1, d2)

a17_survfit(fit, newdata_data.frame(Age=x, tyoungAge=0, nphdmtwAge=x,
exp0022phdm=0, ntwAge=x, twjungAge=0, Zsex=s, Art=i-1, Tumore=0, Demenz=0,
Psychose=0, Herz=1, Arthrosen=0, Lungen=0, Schlag=1, Level=nlcdata$Level[5-1],
logAgetw=0))

pv[i, 17]_wkten(a17$surv, a17$time, d1, d2)

a18_survfit(fit, newdata_data.frame(Age=x, tyoungAge=0, nphdmtwAge=x,
exp0022phdm=0, ntwAge=x, twjungAge=0, Zsex=s, Art=i-1, Tumore=0, Demenz=0,
Psychose=0, Herz=1, Arthrosen=1, Lungen=0, Schlag=0, Level=nlcdata$Level[5-1],
logAgetw=0))

pv[i, 18]_wkten(a18$surv, a18$time, d1, d2)

a19_survfit(fit, newdata_data.frame(Age=x, tyoungAge=0, nphdmtwAge=x,
exp0022phdm=0, ntwAge=x, twjungAge=0, Zsex=s, Art=i-1, Tumore=0, Demenz=1,
Psychose=0, Herz=1, Arthrosen=1, Lungen=0, Schlag=0, Level=nlcdata$Level[5-1],
logAgetw=0))

pv[i, 19]_wkten(a19$surv, a19$time, d1, d2)

a20_survfit(fit, newdata_data.frame(Age=x, tyoungAge=0, nphdmtwAge=x,
exp0022phdm=0, ntwAge=x, twjungAge=0, Zsex=s, Art=i-1, Tumore=0, Demenz=1,
Psychose=0, Herz=1, Arthrosen=0, Lungen=0, Schlag=1, Level=nlcdata$Level[5-1],
logAgetw=0))

pv[i, 20]_wkten(a20$surv, a20$time, d1, d2)

a21_survfit(fit, newdata_data.frame(Age=x, tyoungAge=0, nphdmtwAge=x,
exp0022phdm=0, ntwAge=x, twjungAge=0, Zsex=s, Art=i-1, Tumore=0, Demenz=1,
Psychose=1, Herz=1, Arthrosen=0, Lungen=0, Schlag=0, Level=nlcdata$Level[5-1],
logAgetw=0))

pv[i, 21]_wkten(a21$surv, a21$time, d1, d2)

a22_survfit(fit, newdata_data.frame(Age=x, tyoungAge=0, nphdmtwAge=x,
exp0022phdm=0, ntwAge=x, twjungAge=0, Zsex=s, Art=i-1, Tumore=0, Demenz=0,
Psychose=0, Herz=0, Arthrosen=0, Lungen=0, Schlag=0, Level=nlcdata$Level[5-1],
logAgetw=0))

pv[i, 22]_wkten(a22$surv, a22$time, d1, d2)
}

return(pv)
}

```

Die Funktion *gewichtemat.s* stellt, wie bereits erwähnt, die erforderlichen Gewichte bereit. Sie ruft die Funktionen *gewichteamb.s* und *gewichtestat.s* auf, die die Gewichte der jeweiligen Gruppen in ambulanter bzw. stationärer Pflege berechnen. Hier ist nur die Funktion *gewichteamb.s* angegeben, in *gewichtestat.s* ist lediglich ambulant durch stationär zu ersetzen.

Die Restrukturierung der Daten zur Modellierung der Zeitabhängigkeit mittels fraktionaler Polynome erfolgte mittels der Funktion *restruct.s*. Sie wurde, in Anlehnung an die gleichnamige Funktion von Berger, zu finden unter

`www.imse.med.tu-muenchen.de/persons/berger/coxfp_lib/` ,

auf den zu bearbeitenden Datensatz abgeändert. Zur Modellierung der Zeitabhängigkeit mittels fraktionaler Polynome wurden die Funktionen *fracpol1.s* und *fracpol2.s* in leicht abgewandelter Form benützt. Für fraktionale Polynome ersten Grades ist hier die Funktion *zfracpol1.s* angegeben.

```

gewichtemat_function(alter, steps, stufe, sex, d1, d2) {
  result_matrix(0, 22, 2)

  result[, 1]_gewichteam(alter, stufe, sex, steps, d1, d2)
  result[, 2]_gewichtestat(alter, stufe, sex, steps, d1, d2)

  result
}

gewichteam_function(alter, stufe, sex, steps, d1, d2)
{
  if(sex==1){
    if(stufe==1){
      data_nltcdata[nltcdata$Level=="Stufe1"&nltcdata$Zsex=="w"&
nltcdata$Age < alter+steps/2 & nltcdata$Age >= alter-steps/2&nltcdata$Art==
"ambulant"&nltcdata$LTCbegin<d2&nltcdata$LTCend>d1,]
    }
    if(stufe==2){
      data_nltcdata[nltcdata$Level=="Stufe2"&nltcdata$Zsex=="w"&
nltcdata$Age < alter+steps/2 & nltcdata$Age >= alter-steps/2&nltcdata$Art==
"ambulant"&nltcdata$LTCbegin<d2&nltcdata$LTCend>d1,]
    }
    if(stufe==3){
      data_nltcdata[nltcdata$Level=="Stufe3"&nltcdata$Zsex=="w"&
nltcdata$Age < alter+steps/2 & nltcdata$Age >= alter-steps/2&nltcdata$Art==
"ambulant"&nltcdata$LTCbegin<d2&nltcdata$LTCend>d1,]
    }
  }

  if(sex==0){
    if(stufe==1){
      data_nltcdata[nltcdata$Level=="Stufe1"&nltcdata$Zsex=="m"&
nltcdata$Age < alter+steps/2 & nltcdata$Age >= alter-steps/2&nltcdata$Art==
"ambulant"&nltcdata$LTCbegin<d2&nltcdata$LTCend>d1,]
    }
    if(stufe==2){
      data_nltcdata[nltcdata$Level=="Stufe2"&nltcdata$Zsex=="m"&
nltcdata$Age < alter+steps/2 & nltcdata$Age >= alter-steps/2&nltcdata$Art==
"ambulant"&nltcdata$LTCbegin<d2&nltcdata$LTCend>d1,]
    }
    if(stufe==3){
      data_nltcdata[nltcdata$Level=="Stufe3"&nltcdata$Zsex=="m"&
nltcdata$Age < alter+steps/2 & nltcdata$Age >= alter-steps/2&nltcdata$Art==
"ambulant"&nltcdata$LTCbegin<d2&nltcdata$LTCend>d1,]
    }
  }
}

```

```

}
}

result_rep(0, 22)
if(dim(data)[[1]]==0){
  result[, 1:21]_rep(0, 21)
}
else{
  result[, 1]_dim(data[data$Demenz==1&data$nds==1,])[[1]]

  result[, 2]_dim(data[data$Schlag==1&data$nds==1,])[[1]]
  result[, 3]_dim(data[data$Psychose==1&data$nds==1,])[[1]]
  result[, 4]_dim(data[data$Tumore==1&data$nds==1,])[[1]]
  result[, 5]_dim(data[data$Arthrosen==1&data$nds==1,])[[1]]
  result[, 6]_dim(data[data$Herz==1&data$nds==1,])[[1]]
  result[, 7]_dim(data[data$Lungen==1&data$nds==1,])[[1]]

  result[, 8]_dim(data[data$Demenz==1&data$Schlag==1&data$nds==2,])[[1]]
  result[, 9]_dim(data[data$Demenz==1&data$Herz==1&data$nds==2,])[[1]]
  result[, 10]_dim(data[data$Demenz==1&data$Arthrosen==1&data$nds==2,])[[1]]
  result[, 11]_dim(data[data$Demenz==1&data$Psychose==1&data$nds==2,])[[1]]
  result[, 12]_dim(data[data$Demenz==1&data$Tumore==1&data$nds==2,])[[1]]
  result[, 13]_dim(data[data$Tumore==1&data$Psychose==1&data$nds==2,])[[1]]
  result[, 14]_dim(data[data$Tumore==1&data$Herz==1&data$nds==2,])[[1]]
  result[, 15]_dim(data[data$Psychose==1&data$Herz==1&data$nds==2,])[[1]]
  result[, 16]_dim(data[data$Psychose==1&data$Schlag==1&data$nds==2,])[[1]]
  result[, 17]_dim(data[data$Herz==1&data$Schlag==1&data$nds==2,])[[1]]
  result[, 18]_dim(data[data$Herz==1&data$Arthrosen==1&data$nds==2,])[[1]]

  result[, 19]_dim(data[data$Herz==1&data$Arthrosen==1&data$Demenz==1&
data$nds==3,])[[1]]
  result[, 20]_dim(data[data$Herz==1&data$Schlag==1&data$Demenz==1&
data$nds==3,])[[1]]
  result[, 21]_dim(data[data$Herz==1&data$Psychose==1&data$Demenz==1&
data$nds==3,])[[1]]
}

result[, 22]_length(data[, 1]) - sum(result[, 1:21])

result
}

```

```

restruct <-function(data) {

begin <- data[, 4]
end <- data[,5]
instatus <- data[, 7]
age <-data[,11]
inxvar <- data[, -(4:5)]           # ohne begin,end
inxvar <- inxvar[,-5]             # ohne death
inxvar <- inxvar[,-8]            #ohne Alter
todtime <- end * instatus #filtert nur Todeszeitpunkte raus.
todtime <- unique(sort(todtime)) #und davon nur einfache und sortiert

todtime <- todtime[2:length(todtime)]

newdata <-data.frame(begin,instatus,end,age,inxvar)
names(newdata) <- c("X1", "X2", "X3", "X4",names(inxvar))
X1_newdata$X1
X2_newdata$X2
X3_newdata$X3
X4_newdata$X4

for(i in 1:length(todtime)) {
j_1
while(j <=length(X1)){

if((X1[j]< todtime[i]) &&(todtime[i]<(X3[j]-1))) {
newdata <- rbind(newdata[-j,], cbind(X1[j],
X2[j]*(X1[j]==X2[j]-2),todtime[i],X4[j],inxvar[j,]),
cbind(todtime[i]+1,X2[j]*(X3[j]==X3[j])
,X3[j],X4[j]+(todtime[i]+1-X1[j])/365.25,inxvar[j,]))
}

if((X1[j]< todtime[i]) && (todtime[i]==X3[j]-1)){
newdata <- rbind(newdata[-j,], cbind(X1[j],
X2[j]*(X1[j]==X2[j]-2),todtime[i],X4[j],inxvar[j,]),
cbind(todtime[i],X2[j]*(X3[j]==X3[j])
,X3[j],X4[j]+(todtime[i]+1-X1[j])/365.25,inxvar[j,]))
}

if((X1[j]< todtime[i]) &&(todtime[i]<X3[j])) {
j_1-1}

newdata <- data.frame(newdata)
names(newdata) <- c("X1", "X2", "X3","X4", names(inxvar))
X1_newdata$X1

```

```

X2_newdata$X2
X3_newdata$X3
X4_newdata$X4
inxvar_newdata[,5:dim(newdata)[2]]

j_1-1
print(j)
print(i)
print(length(X1))

}
}
names(newdata) <- c("LTCbegin","Death","LTCend","Age",
names(newdata)[5:dim(newdata)[2]])

return(newdata)
}

zfracpoli_function(daten)
{
attach(daten)
result_rep(1,6)
P1_cbind(-2,-1,-0.5,0,0.5,1,2)

for(i in 1:7){
print(i)
p1_P1[i]

if(p1==0) { t1_log(daten$LTCend)}
else {t1_daten$LTCend^p1}
temp_data.frame(daten,t1)

result[i]_coxph(Surv(LTCbegin, LTCend, Death) ~ Demenz+ Age * (Tumore +
Arthrosen) + Art *(Schlag+ Level) + Tumore * Herz + Psychose *
Schlag +Level*Herz+Schlag*Arthrosen+Lungen+
strata(LTCend)+t1:Level,temp)$loglik[2]
print(result[i])
}

result

}

```