Alexander Wolf, BSc

Verweildaueranalysen zur Modellierung von Stornowahrscheinlichkeiten

MASTERARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

Masterstudium Finanz- und Versicherungsmathematik

eingereicht an der

Technischen Universität Graz

Betreuer

Ao. Univ.-Prof. Dipl.-Ing. Dr.techn. Herwig Friedl

Institut für Statistik

Graz, April 2017

EIDESSTATTLICHE ERKLÄRUNG

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommenen Stellen als solche kenntlich gemacht habe. Das in TUGRAZonline hochgeladene Textdokument ist mit der vorliegenden Masterarbeit identisch.

Datum

Unterschrift

ZUSAMMENFASSUNG

Survivalmodelle sind Regressionsmodelle mit denen Lebensdauern bzw. Überlebensdauern modelliert werden können. Diese Klasse von Modellen ist eingebettet in die Theorie der Survival Analysis. In dieser Arbeit wurde ein Survivalmodell entwickelt, welches die Vertragslaufzeit von Versicherungsnehmern einer Krankenversicherung modelliert. Durch dieses Modell konnten Merkmale gefunden werden, welche Auswirkungen auf das Stornorisiko haben. Die Modellierung der Vertragslaufzeit erlaubt es, die Stornowahrscheinlichkeit über den Verlauf der Vertragsjahre zu gewinnen.

ABSTRACT

Survival models are regression models which are able to model survival times. This class of models is embedded in the theory of Survival Analysis. In this thesis a survival model has been developed, which describes the term of the contract of health insurance policy holders. With this model, characteristics can be identified which have an impact on the cancellation risk. Modeling the term of the contract allows us to get the cancellation probability over the course of the contract years

Inhaltsverzeichnis

1. Einleitung und Motivation

Die Ereigniszeitanalyse bzw. Survival Analysis beschäftigt sich mit der statistischen Analyse von Lebensdauern bzw. Überlebensdauern. Beispiele für solche Lebensdauern sind unter anderem die Überlebenszeit eines Patienten nach diagnostizierter Krebserkrankung, Dauer einer Ehe oder Funktionsdauer einer Glühbirne. Der Ursprung der Ereigniszeitanalyse liegt in der Medizinischen Forschung in der Untersuchung von Überlebenszeiten, jedoch haben sich im Laufe der Zeit auch weitere Anwendungsgebiete in der Biologie, Epidemiologie, Wirtschaft oder Ingenieurwissenschaft aufgetan.

Ein Aspekt welcher die Ereigniszeitanalyse von anderen statistischen Disziplinen unterscheidet ist das Auftreten von Zensierung. Eine Zensierung tritt auf, wenn eine gewisse Information über den Ausfallszeitpunkt existiert jedoch der exakte Zeitpunkt des Ausfalls nicht bekannt ist. Es existieren mehrere Typen von Zensierung wobei alle diese Typen gemeinsam haben, dass sie die statische Analyse der Daten erschweren.

In Kapitel 2 wird die Lebensdauer formal definiert und danach die vier charakterisierenden Funktionen der Lebensdauer, das sind Survival-, Hazard-, erwartete Restlebensdauer- und Dichtefunktion, eingeführt. Weiters werden die verschiedenen Zensierungstypen vorgestellt sowie die Likelihoodfunktionen der Lebensdauern für diese Zensierungstypen hergeleitet.

In Kapitel 3 wird im ersten Teil ein Schätzer der Survivalfunktion, der sogenannte Kaplan-Meier Schätzer, hergeleitet und dafür Konfidenzintervalle besprochen. Im zweiten Teil wird der Log Rank Test vorgestellt anhand dessen die Survivalfunktion für mehrere Gruppen auf Unterschiede überprüft werden kann.

In Kapitel 4 werden parametrische Survivalmodelle behandelt. Im ersten Teil werden Verteilungen besprochen welche zur Modellierung von Lebensdauern

besonders vorteilhaft sind. Im zweiten Teil wird mit dem Accelerated-Failure-Time Modell (AFT-Modell) ein parametrisches Survivalmodell besprochen. Für das AFT-Modell werden die Likelihoodfunktion und die Maximum Likelihood Schätzer (MLE) hergeleitet sowie Methoden zur Überprüfung der Modellanpassung besprochen.

In Kapitel 5 wird das Proportional Hazard (PH) Modell eingeführt und diskutiert. Dieses Modell zeichnet sich dadurch aus, dass keinerlei Verteilungsannahmen für die Lebensdauern getroffen werden müssen. Für das PH Modell wird die Likelihoodfunktion und weiters der Maximum Partial Likelihood Schätzer (MPLE) hergeleitet. Anschließend werden einige Erweiterungen für das PH Modell vorgestellt. Das Kapitel wird mit einem Abschnitt zum Thema Diagnostik abgeschlossen. In diesem Abschnitt werden spezielle Residuen eingeführt und definiert um danach damit die Modellanpassung überprüfen zu können.

Nachdem in den vorangegangenen Kapiteln dieser Masterarbeit die Theorie der Ereigniszeitanalyse ausgeführt wurde, widmet sich das Kapitel 6 der praktischen Anwendung der Ereigniszeitanalyse. Es wird hier die Verweildauer krankenversicherter Personen innerhalb einer Versicherungspolizze mittels eines Survivalmodells modelliert. Weiters sollen mithilfe dieses Modells die Stornowahrscheinlichkeiten geschätzt werden.

Das Kapitel 7 widmet sich der Auswertung des in Kapitel 6 entwickelten Stornierungs-Modells. Hierbei wird der Einfluss von Alter, monatlicher Prämie, Beginnjahr und Gruppe/Rabatt auf die Stornowahrscheinlichkeit bzw. Vertragslaufzeit diskutiert. Abschließend werden danach die Ergebnisse dieser Arbeit nochmals in Kapitel 8 kurz zusammengefasst.

2. Grundlagen

2.1. Grundbegriffe der Survival Analysis

Es wird nun die Lebensdauer formal definiert. Die Ausführungen in diesem Abschnitt orientieren sich dabei an Lawless (2003), Liu (2012) und Glomb (2007). Die **Lebensdauer** wird als eine Zufallsvariable $T \geq 0$ definiert welche die Dauer bzw. Wartezeit (engl. time to event) bis zum Eintritt eines wohldefinierten Ereignisses beschreibt. Weitere gängige Bezeichnungen für die Lebensdauer T sind **Lebenszeit** (engl. survival time), **Wartezeit** oder **Verweildauer**.

Beispiele für Lebensdauern mit dazugehörigen Ereignissen sind:

Beispiel 1: Sei T die Überlebenszeit eines Patienten nach einer Operation, wobei der Tod des Patienten das Ereignis repräsentiert.

Beispiel 2: Sei T die Lebensdauer einer Glühbirne. Das Ereignis/der Ausfall ist in diesem Fall der Defekt der Glühbirne.

Beispiel 3: Sei T die Remissionszeit für Krebspatienten. Das Ereignis ist somit das neuerliche Auftreten von Krebszellen.

Bemerkung 1.
Der Eintritt des Ereignisses wird auch oft mit Ausfall bezeichnet.

2.1.1. Stetige Lebenszeiten

In diesem Abschnitt werden die vier charakterisierenden Funktionen der Lebenszeit, für den Fall dass T eine stetige Lebenszeit ist, hergeleitet.

Die **Survivalfunktion** $S(t)$ beschreibt die Wahrscheinlichkeit, dass ein Individuum t Zeiteinheiten überlebt (d.h. das Ereignis tritt erst nach t Zeiteinheiten auf). Somit kann die Survivalfunktion $S(t)$ definiert werden durch

$$S(t) := \Pr(T \geq t)\,. \tag{2.1}$$

Die Survivalfunktion $S(t)$ ist eine stetig monoton fallende Funktion. Aus (2.1) folgt außerdem $S(0) = 1$ und $\lim_{t\to\infty} S(t) = 0$. Synonym verwendete Bezeichnungen für die Survivalfunktion sind Survivorfunktion, Überlebensfunktion oder Zuverlässigkeitsfunktion (vgl. Glomb, 2007).

Da T eine stetige Zufallsvariable ist, kann die Survivalfunktion umformuliert werden zu

$$S(t) = 1 - F(t) = \int_t^\infty f(u)du\,, \tag{2.2}$$

mit $F(t)$ der stetigen Verteilungsfunktion von T und $f(t)$ der stetigen Dichtefunktion von T.

Die **Hazardfunktion** $h(t)$ ist definiert durch

$$h(t) := \lim_{\Delta t\to 0} \frac{\Pr\left(t \leq T \leq t + \Delta t | T \geq t\right)}{\Delta t}\,. \tag{2.3}$$

Sie gibt das momentane Ausfallsrisiko pro Zeiteinheit an, unter der Bedingung dass das Individuum bis zum Zeitpunkt t noch lebt. Man beachte, dass die Hazardfunktion keine Wahrscheinlichkeit ist sondern eine Rate, welche nur nicht-negative Werte annehmen kann.

Weiters kann die Wahrscheinlichkeit, für ein Individuum welches bis zum Zeitpunkt t lebt und unmittelbar danach ausfällt, durch $h(t)\Delta t$ approximiert werden (siehe (2.3)). Die Hazardfunktion ist unter anderem auch bekannt als die conditional failure rate in reliability, die Intensitätsfunktion bei Stochastischen Prozessen und als altersspezifische Ausfallsrate in der Epidemiologie.

Da T eine stetige Zufallsvariable ist gilt,

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d\log(S(t))}{dt}\,, \tag{2.4}$$

wegen

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \frac{\Pr\left(t \leq T \leq t + \Delta t | T \geq t\right)}{\Delta t} = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \frac{\Pr(t \leq T \leq t + \Delta t\,, T \geq t)}{\Pr(T \geq t)} \\ &= \lim_{\Delta t \to 0} \frac{1}{\Delta t} \frac{\Pr(t \leq T \leq t + \Delta t)}{\Pr(T \geq t)} = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \frac{F(t + \Delta t) - F(t)}{S(t)} \\ &= \lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{\Delta t} \frac{1}{S(t)} = \frac{f(t)}{S(t)}\,. \end{aligned}$$

Eine eng verwandte Funktion der Hazardfunktion ist die **kumulative Hazardfunktion** $H(t)$, welche definiert ist durch

$$H(t) := \int_0^t h(u)du = -\log(S(t))\,. \tag{2.5}$$

Daher kann die Survivalfunktion durch die kumulative Hazardfunktion $H(t)$ ausgedrückt werden als

$$S(t) = \exp\left(-H(t)\right)\,. \tag{2.6}$$

In der Abbildung 2.1 sind drei verschiedene Hazardfunktionen mit zugehörigen Survivalfunktionen abgebildet. In dieser Abbildung ist gut zu erkennen, dass die Hazardfunktion im Gegensatz zu der Survivalfunktion verschiedenste Formen annehmen kann während die Survivalfunktion immer monoton fallend sein muss.

Die **erwartete Restlebensdauer** (mean residual life function) $mrl(t)$ ist die vierte charakterisierende Funktion der Survival Analysis. Diese Funktion gibt an, welche restliche Lebensdauer einem Individuum im Alter t im Mittel noch verbleibt. Die erwartete Restlebendauer $mrl(t)$ ist definiert als

$$mrl(t) := \mathrm{E}\left[T - t | T \geq t\right], \quad t \geq 0\,. \tag{2.7}$$

Weiters gelten folgende Beziehungen falls T stetig verteilt ist:

$$\begin{aligned} mrl(t) &= \frac{1}{S(t)} \int_t^\infty S(u)du, \quad t \geq 0, \\ \mu = \mathrm{E}[T] &= mrl(0) = \int_0^\infty S(u)du \quad \text{und} \\ \mathrm{Var}[T] &= 2\int_0^\infty tS(t)dt - \left(\int_0^\infty S(t)dt\right)^2\,. \end{aligned}$$

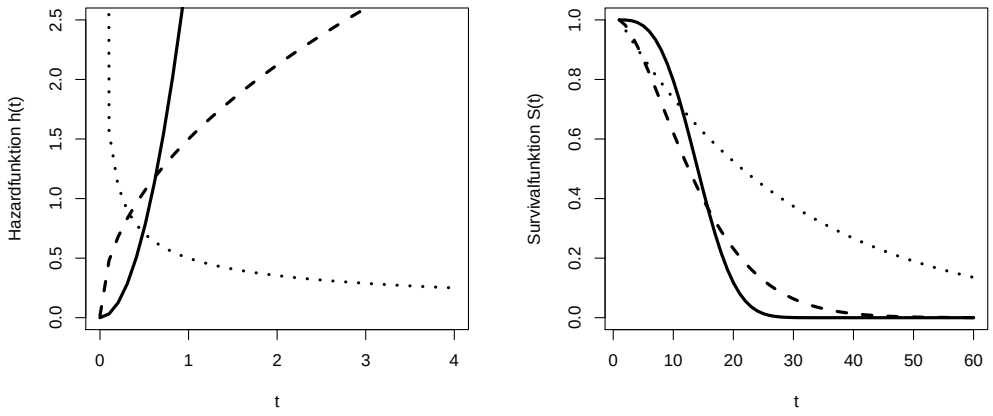

Abbildung 2.1.: Beispiele unterschiedlicher Hazardfunktionen mit zugehörigen Survivalfunktionen

Zusammenfassend kann gesagt werden, dass die Lebenszeit T äquivalent durch die Dichtefunktion, die Survivalfunktion, die Hazardfunktion und die erwartete Restlebensdauer charakterisiert werden kann. Daher wird die Lebenszeit in Regressionsmodellen immer mithilfe einer der vier charakterisierenden Funktionen modelliert (meistens wird hierfür die Hazardfunktion oder die Survivalfunktion verwendet).

Bemerkung 2.
In dieser Arbeit wird fast ausschließlich von stetigen Lebenszeiten T_i ausgegangen. Falls diskrete Lebenszeiten behandelt werden, wird dies explizit erwähnt.

2.1.2. Diskrete Lebenszeiten

In manchen Fällen treten Lebenszeiten nur in diskreter Form auf. Daher wird in diesem Fall T als diskrete Zufallsvariable modelliert. Angenommen T kann nur die Werte $t_1, t_2, \ldots$ annehmen, mit $0 \leq t_1 < t_2 < \cdots$, dann sei die Wahrscheinlichkeitsfunktion

$$f(t_j) = \Pr(T = t_j) \quad j = 1, 2, \ldots . \tag{2.8}$$

Die Survivalfunktion ist somit

$$S(t) = \Pr(T \geq t) = \sum_{t_j \geq t} f(t_j)\,. \tag{2.9}$$

Betrachtet man $S(t)$ als eine Funktion von $t \geq 0$ dann ist $S(t)$ eine monoton fallende Treppenfunktion, mit $S(0) = 1$ und $S(\infty) = 0$. Die diskrete Hazardfunktion ist definiert als

$$\begin{aligned} h(t_j) &= \Pr(T = t_j | T \geq t_j) \\ &= \frac{f(t_j)}{S(t_j)} \quad j = 1, 2 \ldots\,. \end{aligned} \tag{2.10}$$

Ähnlich wie im stetigen Fall charakterisieren die Wahrscheinlichkeitsfunktion, die Survivalfunktion und die Hazardfunktion die Verteilung von T. Wegen $f(t_j) = S(t_j) - S(t_{j+1})$ folgt

$$h(t_j) = 1 - \frac{S(t_{j+1})}{S(t_j)}\,.$$

Weiters gilt

$$\begin{aligned} S(t_{r+1}) &= \frac{S(t_2)}{S(t_1)} \frac{S(t_3)}{S(t_2)} \cdots \frac{S(t_r)}{S(t_{r-1})} \frac{S(t_{r+1})}{S(t_r)} \\ &= \prod_{j=1}^{r} \frac{S(t_{j+1})}{S(t_j)}\,, \end{aligned}$$

wegen $S(t_1) = 1$ und somit

$$S(t) = \prod_{t_j < t} (1 - h(t_j)). \tag{2.11}$$

Ein diskretes Gegenstück zu der stetigen Funktion $H(t)$ kann auf zwei Arten definiert werden: zum einen über $-\log S(t)$, mit $S(t)$ aus (2.11) oder zum anderen über $\sum_{t_j \leq t} h(t_j)$, wobei beide Zugangsweisen nicht denselben Wert liefern.

Mithilfe von Riemann Stieltjes Integralen und Produkt Integralen ist es gelungen ein allgemeineres Konzept für die Survival Analysis zu formulieren. Mithilfe dieses Konzeptes können diskrete, stetige und gemischte Lebensverteilungen innerhalb eines Konzeptes eingebettet werden. Für mehr Information zu diesem verallgemeinerten Konzept siehe Abschnitt 1.2.3 in Lawless (2003).

2.2. Zensierung und Trunkierung (Stutzung)

Die bisher besprochenen Begriffe können auch in anderen Gebieten der Statistik gefunden werden. Die grundlegende Eigenschaft, welche Survival Analysis von anderen Gebieten der Statistik unterscheidet, ist die Zensierung. Im Wesentlichen tritt Zensierung dann auf, wenn eine gewisse Information über den Ausfallszeitpunkt existiert jedoch der exakte Zeitpunkt des Ausfalls nicht bekannt ist (vgl. Klein und Moeschberger, 2003). Es existieren drei Zensierungsarten, nämlich Rechtszensierung (right censoring), Linkszensierung (left censoring) und Intervalzensierung (interval censoring).

Allgemein gilt, dass die Rechtszensur die am häufigst auftretende Zensierungsart ist (im speziellen die Rechtszensierung Typ I). Zusätzlich sind die meisten Ergebnisse für rechtszensierte Lebenszeiten theoretisch leichter herleitbar und somit existieren für diese Art der Zensierung auch die meisten Arbeiten bzw. Ergebnisse.

2.2.1. Rechtszensierung

Man spricht von einer **rechtszensierten Lebenszeit** T, wenn der Ausfall noch nicht beobachtet wurde und dieser rechts des *Zensurzeitpunktes* c liegt. Somit wird statt der exakten Lebenszeit nur

$$Y = \min(T, c)\,, \tag{2.12}$$

beobachtet. Zum Beispiel wurde in einer Studie nicht der Tod aller Patienten abgewartet sondern nach einer gewissen Dauer abgebrochen.

Rechtszensierung kann weiters in Typ I und Typ II Rechtszensur unterteilt werden, wobei der Typ I die gängigere Variante ist.

Typ I Rechtzensur

Die **Typ I Rechtszensierung** tritt auf wenn jedes Individuum einen fixen (deterministischen) Zensurzeitpunkt c_i besitzt (vgl. Lawless, 2003). Somit

wird Y_i tatsächlich beobachtet wenn $T_i \leq c_i$ gilt. Formal ausgedrückt bedeutet dies

$$Y_i = \min(T_i, c_i)\,.$$

Ein typisches Beispiel für Typ I Zensierung ist, wenn eine Studie eine vorher festgelegte Zeitdauer besitzt.

Typ II Rechtszensur

Diese Art der Rechtszensierung tritt genau dann auf, wenn nur die r kleinsten Lebenszeiten $T_{(1)}, \ldots, T_{(r)}$ von n möglichen beobachtet werden. Ein großer Nachteil bei dieser Art der Zensierung ist, dass das Ende der Studie bis zum Schluss unbekannt ist (vgl. Lawless (2003)).

In den Abbildungen 2.2 und 2.3 werden Typ I und Typ II Rechtszensierung visuell aufbereitet. Im Beispiel der Abbildung 2.2 ist der Zensurzeitpunkt für alle Individuen gleich und im voraus festgelegt. Im Beispiel der Abbildung 2.3 tritt die Zensur erst dann auf wenn 4 Individuen ausgefallen sind.

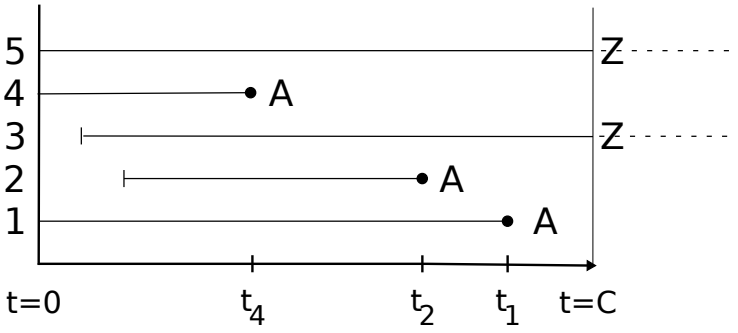

Abbildung 2.2.: Typ I Rechtszensierung. Mit A werden Ausfälle markiert, mit Z werden zensierte Lebenszeiten markiert und t_1, t_2, t_4 bezeichnen die beobachteten Ausfallszeitpunkte.

2.2.2. Linkszensierung

Ist eine Lebenszeit **linkszensiert**, bedeutet dies, dass der Ausfall vor dem Beginn der Studie erfolgte. Daher liegt der Ausfallszeitpunkt links des Stu-

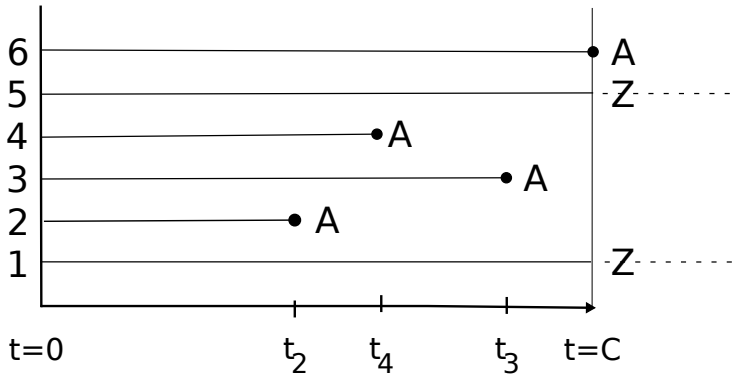

Abbildung 2.3.: Typ II Rechtszensierung mit $r = 4$. Mit A werden Ausfälle markiert, mit Z werden zensierte Lebenszeiten markiert und t_2, t_3, t_4 bezeichnen die beobachteten Ausfallszeitpunkte. Die Zensur setzt sofort ein, nachdem 4 Individuen ausgefallen sind (hier fällt Individuum 6 als Viertes aus). Daher gilt, $c = t_6$.

dienbeginns bzw. Beobachtungstarts, es ist jedoch nicht bekannt wann genau dieser stattgefunden hat. Der exakte Ausfallszeitpunkt ist somit nur genau dann bekannt, wenn der Ausfallszeitpunkt nach dem Zensierungspunkt c stattfindet, dh. $T \geq c$ gilt. Somit wird anstatt der exakten Lebenszeit T nur

$$Y = \max(T, c) \ , \tag{2.13}$$

beobachtet, mit T der exakten Lebenszeit und c dem Studienbeginn bzw. dem Beobachtungsstart.

Beispiel 1.
Ein Beispiel für Linkszensierung ist die Studie zur Untersuchung der Zeit bis zum ersten Marihuana Konsum von High School Schülern in Kalifornien. Angenommen ein Schüler gibt bei der Befragung an, bereits Marihuana konsumiert zu haben, aber kann sich nicht mehr an den Zeitpunkt des Beginns erinnern. Somit wäre dann die Zeit des ersten Marihuana Konsums für diesen Schüler linkszensiert (vgl. Klein und Moeschberger, 2003).

Mithilfe der Abbildung 2.4 soll die Linkszensierung noch einmal veranschau-

licht werden. Bei den Individuen 2 und 4 tritt das Ereignis bereits auf, bevor die Studie überhaupt begonnen hat.

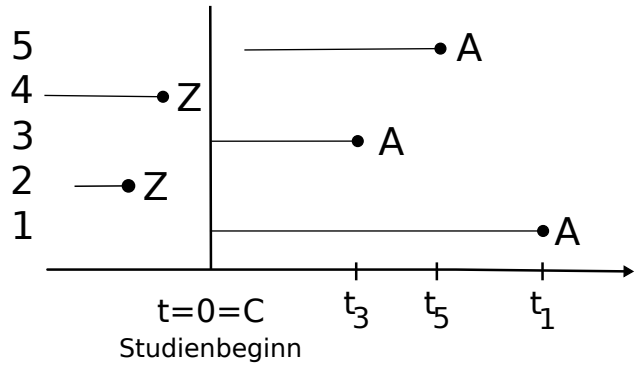

Abbildung 2.4.: Linkszensierung. Mit A werden Ausfälle markiert, mit Z werden zensierte Lebenszeiten markiert und t_1, t_3, t_5 bezeichnen die beobachteten Ausfallszeitpunkte.

2.2.3. Intervallzensierung

Die Intervallzensierung ist die allgemeinste Art der Zensierung. Diese Zensierung tritt dann auf wenn für Individuen nur bekannt ist, dass das Ereignis innerhalb eines Intervalls aufgetreten ist, dh.

$$Y = T \in (c^L, c^R] \,. \tag{2.14}$$

Intervallzensierung ist ein typisches Produkt bei Studien von regelmäßigen Nachuntersuchungen. Wird hier eine Krankheit diagnostiziert, so ist lediglich bekannt, dass der Zeitpunkt des Ausbruchs zwischen den letzten beiden Untersuchungen liegt (vgl. Glomb, 2007).

Die Intervalzensierung ist eine Verallgemeinerung der Links- und Rechtszensierung. Sei die linke Intervallgrenze $c^L = 0$ und die rechte Intervallgrenze gleich c erhält man Rechtszensierung. Umgekehrt ergibt sich Linkszensierung falls die rechte Intervallgrenze unendlich ist und die linke Intervallgrenze gleich c ist. (vgl. Klein und Moeschberger, 2003).

2.2.4. Trunkierung (Stutzung)

Eine weitere Eigenschaft die viele Survival Daten besitzen, aber oft mit Zensierung verwechselt wird, ist die **Trunkierung** (engl. truncation) oder auch Stutzung genannt. Trunkierung bedeutet, dass lediglich die Individuen beobachtet werden deren Ausfallszeitpunkt innerhalb von (S^L, S^R) fallen. Das wiederum hat zur Folge, dass Individuen deren Ausfallszeitpunkte nicht in dieses Intervall fallen, nicht beobachtet werden.

Der Unterschied zwischen Zensierung und Trunkierung ist, dass obwohl ein Individuum zensiert wurde eine gewissen Information über das Individuum besteht (Ausfallszeitpunkt liegt links von c, Ausfallszeitpunkt liegt rechts von c, Ausfallszeitpunkt liegt in $(c^L, c^R]$), während es bei trunkierten (gestutzten) Daten absolut keine Information über Individuen gibt deren Ausfallszeitpunkte außerhalb von (S^L, S^R) liegen (vgl. Klein und Moeschberger, 2003). Es existieren zwei Arten von Trunkierung (Stutzung) nämlich Linkstrunkierung und Rechtstrunkierung.

Linkstrunkierung

Gelte für die rechte Intervallgrenze bei einer Trunkierung $s^R = \infty$, dann nennt man diese Trunkierungsform **Linkstrunkierung** (engl. left truncation). Wir beobachten T genau dann wenn $s^L < T$ gilt.

Beispiel 2.
Die Lebensdauern von Bewohnern eines Altersheimes sollen untersucht werden. In der Studie werden nur Individuen beobachtet, die noch im Seniorenheim leben. Alle Personen die bereits vor dem Beginn der Studie verstarben, sind linkstrunkiert und werden somit in der Studie nicht berücksichtigt (vgl. Klein und Moeschberger, 2003).

Rechtstrunkierung

Gelte für die linke Intervallgrenze bei einer Trunkierung $s^L = 0$, dann nennt man diese Trunkierungsform **Rechtstrunkierung** (engl right truncation). Die Lebenszeit T wird genau dann beobachtet wenn $T < s^R$ gilt.

Beispiel 3.
In einer Studie wurde das Alter von Schülern erhoben in dem sie zu rauchen begonnen haben. Man erhält hier rechtstrunkierte Daten, da die Schüler die nach ihrer Schulzeit begonnen haben zu rauchen, nicht inkludiert sind.

Falls Daten trunkiert sind, hat dies einen Einfluss auf die Likelihoodfunktion. In diesem Fall muss eine bedingte Verteilung verwendet werden.

Daten können auch gleichzeitig trunkiert und zensiert sein. Die in der Praxis am häufigsten auftretende Kombination sind rechtszensierte und linkstrunkierte Daten.

2.3. Likelihoodfunktion für zensierte und trunkierte Daten

In diesem Abschnitt werden die Likelihoodfunktionen für die in Abschnitt 2.2 beschriebenen Zensierungs- und Trunkierungsarten hergeleitet. Im ersten Teil des Abschnitts wird die Likelhoodfunktion für rechtszensierte Lebenszeiten des Typ I behandelt und danach wird diese Herangehensweise auf linkszensierte, intervallzensierte und trunkierte Lebenszeiten angewandt.

Die Ausführungen in diesen Abschnitt folgen Liu (2012), Klein und Moeschberger (2003) und Glomb (2007). Um die Likelihoodfunktion für Rechtszensierung Typ I herleiten zu können, muss zuerst etwas an Notation eingeführt werden. Es wird angenommen, dass n individuelle Lebenszeiten $T_1, \ldots, T_n$ studiert werden. Außerdem werden nicht alle Lebenszeiten T_i beobachtet sondern

$$Y_i = \min(T_i, c_i)\,.$$

Zusätzlich wird noch die *Rechtszensur-Indikatorvariable* R_i eingeführt, mit

$$R_i = I(T_i = Y_i) = I(T_i \leq c_i)\,. \tag{2.15}$$

Mithilfe der Rechtszensur-Indikatorvariable R_i kann man nun zwischen rechtszensierten und tatsächlich beobachteten Lebenszeiten unterscheiden. Somit erhalten wir die drei Zufallsvariablen T_i, Y_i und R_i für das i-te Individuum.

Die Likelihoodfunktion für Typ I Rechtszensierung, gegeben den drei Zufallsvariablen, kann als die gemeinsame Wahrscheinlichkeitsverteilung von (Y_i, R_i) formuliert werden (vgl. Liu, 2012). Die **gemeinsame Dichtefunktion** von (Y_i, R_i) für stetige und diskrete rechtszensierte Lebenszeiten des Typ I lautet

$$f(y_i, r_i) = f_{T_i}(y_i)^{r_i} \Pr(T_i > c_i)^{1-r_i} . \tag{2.16}$$

Für den Fall, dass die Lebenszeit T_i diskret verteilt ist, gilt

$$\begin{aligned} \Pr\big((Y_i, R_i) = (y_i, 1)\big) &= \Pr\left(T_i = y_i\right) = f_{T_i}(y_i) \\ \Pr\big((Y_i, R_i) = (c_i, 0)\big) &= \Pr\left(T_i > c_i\right) \end{aligned}$$

mit $f_{T_i}(y_i) = \Pr(T_i = y_i)$. Für den Fall, dass die Lebenszeit T_i stetig ist, betrachte man die Approximation $\Pr(Y_i \in [y_i, y_i + \varepsilon])$ mit $\varepsilon > 0$. Es gilt nun

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

Somit ist die gemeinsame Wahrscheinlichkeitsverteilung von (Y_i, R_i) für stetige und diskrete Lebenszeiten äquivalent (vgl. Herleitung der gemeinsamen Dichtefunktion in Glomb, 2007).

Unter der Annahme, dass die n Lebenszeiten $T_1, \ldots, T_n$ unabhängig sind, erhält man die folgende **Likelihoodfunktion für rechtszensierte Typ I Daten**

$$L = \prod_{i=1}^{n} f_{T_i}(y_i)^{r_i} S_{T_i}(y_i+)^{1-r_i} , \tag{2.17}$$

wobei im allgemeinen $S_{T_i}(y_i+) = \Pr(T_i > y_i)$ gilt. Falls jedoch $S_{T_i}(t)$ in y_i stetig ist, gilt $S_{T_i}(y_i+) = S_{T_i}(y_i)$ (vgl. Lawless, 2003).

Bemerkung 3.
In der Literatur wird die Rechtszensur-Indikatorvariable auch oft andersrum definiert, d.h. $R_i = I(T_i \geq C_i)$.

Um die Likelihoodfunktion für linkszensierte Daten (siehe Abschnitt 2.2.2) herleiten zu können wird eine andere Indikatorvariable benötigt und zwar

$$L_i = I(T_i = Y_i) = I(T_i \geq c) \tag{2.18}$$

und es ist

$$Y_i = \max(T_i, c_i)\,.$$

Unter der Annahme, dass die n Lebenszeiten $T_1, \ldots, T_n$ unabhängig und stetig sind, erhält man die **Likelihoodfunktion für linkszensierte Daten**

$$L = \prod_{i=1}^{n} f_{T_i}(y_i)^{l_i} F_{T_i}(y_i)^{1-l_i}\,. \tag{2.19}$$

Die Herleitung von L für linkszensierte Daten läuft äquivalent zu der Herleitung im rechtszensierten Typ I Fall ab.

Für intervallzensierte Lebenszeiten (siehe Abschnitt 2.2.3) mit $Y_i \in (c_i^L, c_i^R]$ ist die Wahrscheinlichkeit einer solchen Lebenszeit gleich $\Pr(Y_i \in (c_i^L, c_i^R]) = F_{T_i}(c_i^R) - F_{T_i}(c_i^L)$. Unter der Annahme, dass die n Lebenszeiten $T_1, \ldots, T_n$ unabhängig und stetig sind, erhält man die **Likelihoodfunktion für intervallzensierte Daten**

$$L = \prod_{i=1}^{n} \left(F_{T_i}(c_i^R) - F_{T_i}(c_i^L)\right). \tag{2.20}$$

Um die Likelihoodfunktion für trunkierte Daten (siehe Abschnitt 2.2.4) herzuleiten, muss die gemeinsame Verteilung von Y darauf bedingt werden, dass der Ausfall in (s^L, s^R) erfolgt. d.h. $\Pr(Y \in (s^L, s^R)) = \frac{f_T(y)}{F_T(s^R) - F_T(s^L)}$. Unter der Annahme, dass die n Lebenszeiten $T_1, \ldots, T_n$ unabhängig und stetig sind, erhält man die **Likelihoodfunktion für trunkierte Daten**

$$L = \prod_{i=1}^{n} \frac{f_{T_i}(y_i)}{F_{T_i}(s^R) - F_{T_i}(s^L)}\,.$$

3. Nicht-parametrische Schätzer

3.1. Kaplan-Meier und Nelson-Aalen Schätzer

In diesem Abschnitt werden Schätzer für die Survivalfunktion hergeleitet. Die Ausführungen folgen hierbei in großen Teilen Liu (2012).

Die Survivalfunktion $S(t)$ ist per Definition (siehe (2.1)) die Wahrscheinlichkeit, dass ein Individuum älter als t Zeiteinheiten wird. Angenommen die n Beobachtungen $y_1, \ldots, y_n$ der Lebenszeiten $T_1, \ldots, T_n$ sind weder zensiert noch trunkiert, dann kann durch die **empirische Survivalfunktion**

$$\widehat{S}_{\mathrm{ES}}(t) = \frac{\text{Anzahl Ausfallzeitpunkte} > t}{n} \tag{3.1}$$

die Survivalfunktion auf sehr einfache Art und Weise geschätzt werden (vgl. Collett, 2003). Diese Treppenfunktion sinkt somit jeweils zu einem Ausfallzeitpunkt um d/n Einheiten, mit der Anzahl an Ausfällen d zu diesem Ausfallzeitpunkt.

Sind jedoch die beobachteten Lebenszeiten zensiert, muss die empirische Survivalfunktion (3.1) modifiziert werden, da nicht mehr klar ist, wie viele Ausfallzeitpunkte größer als t sind. Der von Kaplan und Meier (1958) eingeführte Schätzer behebt dieses Problem. Um diesen Schätzer herleiten zu können, müssen im ersten Schritt die verschiedenen beobachteten Ausfall- und Zensurzeitpunkte geordnet werden, dh.

$$y_{(1)} < y_{(2)} < \cdots < y_{(\acute{n}-1)} < y_{(\acute{n})} \,.$$

Mit $\acute{n}$ wird die Anzahl voneinander verschiedener Ausfall- und Zensurzeitpunkte bezeichnet (d.h. wenn alle Individuen zu unterschiedlichen Zeitpunkten ausfallen bzw. zensiert sind, gilt $\acute{n} = n$). Die Survivalfunktion $S(t)$ wird nun mithilfe einer diskreten Wahrscheinlichkeitsverteilung, welche auf den Massezeitpunkten $y_{(1)}, \ldots, y_{(\acute{n})}$ definiert ist, geschätzt.

In (2.11) wurde gezeigt, dass die Survivalfunktion $S(t)$ für diskret verteilte Lebenszeiten durch

$$S(t) = \prod_{y_i < t} \big(1 - h(y_i)\big) \tag{3.2}$$

ausgedrückt werden kann. Um die Survivalfunktion schätzen zu können, muss nun noch ein Schätzer für $h(t_i)$ gefunden werden. Solch ein passender Schätzer ist

$$\widehat{h}(y_i) = \frac{d_i}{n_i} \quad \text{für } i = 1, \ldots, \acute{n}\,, \tag{3.3}$$

mit der Anzahl der zum i-ten Ausfallzeitpunkt gefährdeten Individuen n_i und der Anzahl an Ausfällen d_i. Durch Einsetzen von $\widehat{h}(y_i)$ aus (3.3) in (3.2) erhält man den **Kaplan-Meier Schätzer** $\widehat{S}_{\mathrm{KM}}$ für die Survivalfunktion und zwar

$$\widehat{S}_{\mathrm{KM}}(t) = \prod_{y_{(i)} \leq t} \left(1 - \frac{d_i}{n_i}\right) = \prod_{y_{(i)} \leq t} \frac{n_i - d_i}{n_i}\,. \tag{3.4}$$

Es kann weiters gezeigt werden, dass der Kaplan-Meier Schätzer konsistent und asymptotisch normalverteilt ist (siehe Kapitel 3.2.4 in Lawless, 2003). Weiters kann gezeigt werden, dass der Kaplan-Meier Schätzer und die empirische Survivalfunktion für nicht-zensierte Daten übereinstimmen.

Bemerkung 4.
Der Kaplan-Meier Schätzer kann auch als nicht-parametrischer Maximum Likelihood Schätzer (MLE) der Survivalfunktion $S(t)$ hergeleitet werden. Aus diesem Grund wird er auch oft als Product Limit Estimator (PL) bezeichnet.

Der Kaplan-Meier Schätzer kann wegen (2.5) als Schätzer der kumulativen Hazardfunktion verwendet werden und zwar

$$\widehat{H}(t) = -\log\left(\prod_{y_{(i)} \leq t} \frac{n_i - d_i}{n_i}\right)\,. \tag{3.5}$$

Durch die Approximation $\log(1+x) \approx x$ für kleines x kann der Schätzer für die kumulative Hazardfunktion (3.5) weiter umgeformt werden zu

$$\begin{aligned}\widehat{H}(t) &= -\sum_{y_{(i)} \leq t} \log\left(1 - \frac{d_i}{n_i}\right) \\ &\approx -\sum_{y_{(i)} \leq t} \left(-\frac{d_i}{n_i}\right) = \sum_{y_{(i)} \leq t} \frac{d_i}{n_i}\,. \end{aligned} \tag{3.6}$$

Der Schätzer in (3.6) wurde von Nelson (1969) bzw. Aalen (1978) eingeführt und wird daher oft als **Nelson-Aalen Schätzer** bezeichnet. Weiters kann der Nelson-Aalen Schätzer mithilfe von (2.6) umgeformt werden zu

$$\widehat{S}_{\mathrm{NA}}(t) = \exp\left(-\sum_{y_{(i)} \leq t} \frac{d_i}{n_i}\right)\,, \tag{3.7}$$

einem weiteren Schätzer der Survivalfunktion.

Der Kaplan Meier Schätzer und der Nelson-Aalen Schätzer sind Grundwerkzeuge jeder deskriptiven Survival Datenanalyse. Für die beiden Schätzer müssen außer der Unabhängigkeit keinerlei Verteilungsannahmen getroffen werden (daher sind sie nicht-parametrisch) und beruhen somit rein auf den empirischen Daten.

Beispiel 4.
In diesem kurzen Beispiel wird der Kaplan-Meier Schätzer und der Nelson-Aalen Schätzer für die Beispieldaten 5, 17, 20+, 24, 32, 35+, 40, 46, 47, 50, 59, 74 berechnet (das + Symbol kennzeichnet rechtszensierte Lebenszeiten).

Anhand der Abbildung 3.1 kann man sehen, dass die Treppenfunktion des Kaplan-Meier Schätzers zum Zeitpunkt jedes beobachteten Ausfalls fällt, jedoch nicht bei Auftreten eines zensierten Ereignisses. Der Nelson-Aalen Schätzer in Abbildung 3.2 wirkt wegen der exp()-Funktion in (3.7) glatter, dennoch liefert auch er eine ähnliche Schätzung der Survivalfunktion wie der Kaplan-Meier Schätzer.

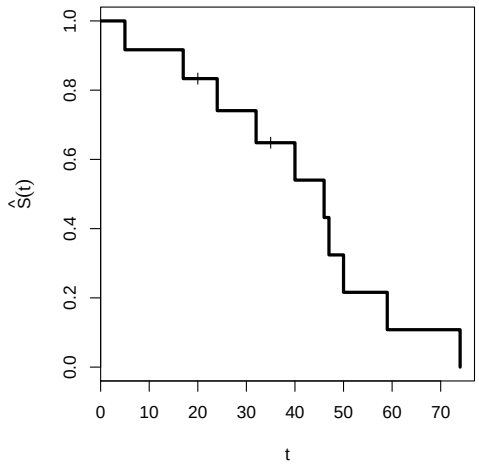

Abbildung 3.1.: Kaplan-Meier Schätzer für Beispiel 4. Das Symbol + kennzeichnet rechtszensierte Lebenszeiten.

R-Code 3.1: Kaplan-Meier- und Nelson-Aalen-Schätzer

```
> y <- c(5, 17, 20, 24, 32, 35, 40, 46, 47, 50, 59, 74)
> r  <- c(1, 1, 0, 1, 1, 0, 1, 1, 1, 1, 1, 1)
> bsp.surv <- Surv(y, r)
> bsp.surv
[1]  5  17  20+ 24  32  35+ 40  46  47  50  59  74
>
> # KAPLAN-MEIER Schaetzer
> bsp.km <- survfit(bsp.surv ~ 1, conf.int = F)
> summary(bsp.km)
 time n.risk n.event survival std.err
    5     12       1    0.917  0.0798
   17     11       1    0.833  0.1076
   24      9       1    0.741  0.1295
   32      8       1    0.648  0.1426
   40      6       1    0.540  0.1544
   46      5       1    0.432  0.1568
   47      4       1    0.324  0.1503
   50      3       1    0.216  0.1335
   59      2       1    0.108  0.1014
```

```
 74      1       1    0.000     NaN
> plot(survfit(bsp.surv ~ 1), xlab = "t",
+ ylab = expression(hat(S)(t)), conf.int = F, lwd=3)
>
> # NELSON-AALEN Schaetzer fuer Survivalfunktion
> na_schaetzer <- exp(-cumsum( bsp.km$n.ev / bsp.km$n.risk))
> plot(times, na_schaetzer,type="l", xlab = "t",
+ ylab = expression(hat(S)(t)), lwd = 3)
>
> # NESLON-AALEN-Schaetzer fuer kum. Hazardfunktion
> na_kum_hazard <- cumsum(bsp.km$n.ev / bsp.km$n.risk)
> plot(times, na_kum_hazard, type="l", xlab="t",
+ ylab = expression(hat(H)(t)), lwd = 3)
```

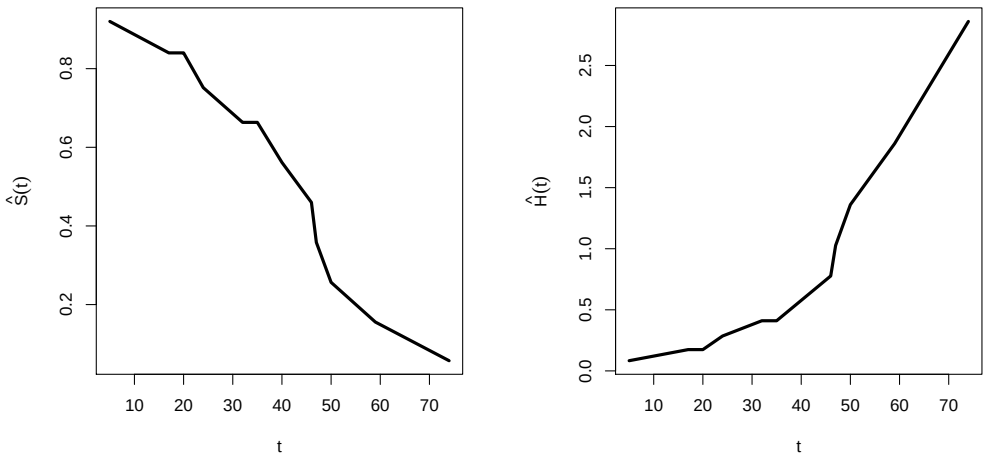

Abbildung 3.2.: Nelson-Aalen Schätzer für Beispiel 4. Links ist der Schätzer für die Survivalfunktion abgebildet und rechts der Schätzer für die kumulative Hazardfunktion.

3.1.1. Varianzschätzung der Survivalfunktion

Es gibt eine Vielzahl an Möglichkeiten die Varianz des Kaplan-Meier Schätzers zu schätzen. Die hier vorgestellte Variante wird als Greenwood Formel bezeichnet und wird mithilfe der Delta Methode hergeleitet (siehe Anhang Satz A.1.1).

Die Herleitung beginnt mit der Betrachtung von $\log(\widehat{S}_{\mathrm{KM}}(t))$, mit $\widehat{S}_{\mathrm{KM}}(t)$ definiert in (3.4) also

$$\log(\widehat{S}_{\mathrm{KM}}(t)) = \sum_{y_{(i)} \leq t} \log\left(\frac{n_i - d_i}{n_i}\right) = \sum_{y_{(i)} \leq t} \log\left(\widehat{s}(y_{(i)})\right),$$

mit $\widehat{s}(y_{(i)}) := \frac{n_i - d_i}{n_i}$ der bedingten Survivalwahrscheinlichkeit im Intervall $I_i = (y_{(i-1)}, y_{(i)})$. Angenommen die Anzahl an Überlebenden zum Zeitpunkt $y_{(i)}$ folgt einer Binomialverteilung, d.h. $(n_i - d_i) \sim \mathrm{Bin}(n_i, \pi_i)$, dann kann $\widehat{s}(y_{(i)})$ als Schätzer von π_i angesehen werden. Somit ist ein Schätzer für die Varianz von $\widehat{s}(y_{(i)})$ gleich

$$\widehat{\mathrm{Var}}[\widehat{s}(y_{(i)})] = \frac{\widehat{s}(y_{(i)})\left(1 - \widehat{s}(y_{(i)})\right)}{n_i}$$

(vgl. Hosmer und Stanley, 1999). Unter der Verwendung der Delta Methode hat die Varianz von $\log(\widehat{s}(y_{(i)}))$ folgende Form:

$$\begin{aligned}\widehat{\mathrm{Var}}[\log(\widehat{s}(y_{(i)}))] &\approx \left(\frac{1}{\widehat{s}(y_{(i)})}\right)^2 \frac{\widehat{s}(y_{(i)})(1 - \widehat{s}(y_{(i)}))}{n_i} \\ &\approx \frac{1 - \widehat{s}(y_{(i)})}{\widehat{s}(y_{(i)}) n_i}.\end{aligned}$$

Da $\widehat{s}(y_{(i)})n_i = (n_i - d_i)$ gilt, wird nun der Nenner und Zähler mit n_i erweitert, wodurch

$$\begin{aligned}\widehat{\mathrm{Var}}[\log \widehat{s}(y_{(i)})] &\approx \frac{n_i(1 - \widehat{s}(y_{(i)}))}{\widehat{s}(y_{(i)}) n_i^2} \\ &\approx \frac{d_i}{(n_i - d_i) n_i}\end{aligned}$$

folgt. Aufgrund der Unabhängigkeit der verschiedenen Ausfallszeitpunkte $y_{(i)}$ kann nun die Varianz von $\log(\widehat{S}(t))$ geschätzt werden, in dem alle Varianzen von $\log(\widehat{s}(y_{(i)}))$ für $y_{(i)} \leq t$ aufsummiert werden, also

$$\widehat{\mathrm{Var}}[\log \widehat{S}_{\mathrm{KM}}(t)] \approx \sum_{y_{(i)} \leq t} \frac{d_i}{(n_i - d_i)n_i} \,. \tag{3.8}$$

Im letzten Schritt muss $\widehat{\mathrm{Var}}[\log \widehat{S}_{\mathrm{KM}}(t)]$ wieder rücktransformiert werden zu $\widehat{\mathrm{Var}}[\widehat{S}_{\mathrm{KM}}(t)]$. Diese Rücktransformation erfolgt wieder mithilfe der Delta Methode, unter Verwendung der exp() Funktion. Dies ergibt folgenden Ausdruck für die **Varianz des Kaplan-Meier Schätzers** (Greenwood Formel)

$$\widehat{\mathrm{Var}}[\widehat{S}_{\mathrm{KM}}(t)] \approx (\widehat{S}_{\mathrm{KM}}(t))^2 \sum_{y_{(i)} \leq t} \frac{d_i}{(n_i - d_i)n_i} \,. \tag{3.9}$$

Die **Varianz des Nelson-Aalen Schätzers** kann ebenfalls äquivalent wie die Greenwood Formel, unter Verwendung der Delta Methode, hergeleitet werden und lautet

$$\widehat{\mathrm{Var}}[\widehat{H}(t)] \approx \sum_{y_{(i)} \leq t} \frac{d_i}{n_i^2} \,. \tag{3.10}$$

3.1.2. Lokale Konfidenzintervalle und Konfidenzbänder

Da der Kaplan-Meier Schätzer $\widehat{S}_{\mathrm{KM}}(t)$ asymptotisch normalverteilt ist gilt

$$Z(t) = \frac{\widehat{S}_{\mathrm{KM}}(t) - S(t)}{\widehat{\sigma}_s(t)} \xrightarrow{D} N(0,1) \,, \tag{3.11}$$

mit $\widehat{\sigma}_s^2(t) = \widehat{\mathrm{Var}}[\widehat{S}_{KM}(t)]$ der Varianz aus (3.9). Somit kann folgendes lokales (bzw. punktweises) Konfidenzintervall, zum Niveau $1 - \alpha$, für die Survivalfunktion formuliert werden:

$$\widehat{S}_{KM}(t) \pm z_{1-\alpha/2}\widehat{\sigma}_s(t) \,, \tag{3.12}$$

mit $z_{1-\alpha/2}$ dem $1 - \alpha/2$ Quantil der Standard Normalverteilung. Jedoch ist dieses Konfidenzintervall nur bedingt brauchbar. Während normalverteilte

Variablen Werte aus $(-\infty, \infty)$ annehmen können, kann $S(t)$ nur Werte von 0 bis 1 annehmen. Somit kann es vorkommen, dass die Intervallgrenzen Werte außerhalb von $(0, 1)$ annehmen.

Dieses Problem kann mithilfe von Transformation umgangen werden, wie Logit Transformation, arcus sinus Transformation und log-log Transformation. Bei diesen Methoden wird im ersten Schritt $\widehat{S}_{KM}(t)$ transformiert, im zweiten Schritt wird ein Konfidenzintervall für die Transformation berechnet und im dritten Schritt wird dieses Konfidenzintervall wieder rücktransformiert.

Die meist verwendete Methode ist die **log-log Transformation**, welche hier nun vorgestellt wird. Durch diese Transformation wird $\widehat{S}(t) := \widehat{S}_{\mathrm{KM}}(t)$ auf einen Wertebereich von $(-\infty, \infty)$ transformiert. Dafür muss zuerst $\widehat{v}(t)$ definiert werden mit

$$\widehat{v}(t) := \log\Big(-\log \widehat{S}(t)\Big) \,. \tag{3.13}$$

Aus (2.5) ist bekannt, dass $\log(S(t))$ die kumulative Hazardfunktion darstellt. Deswegen wird $\widehat{v}(t)$ auch als **logarithmierte Hazardfunktion** bezeichnet. Mithilfe der Delta Methode, unter Verwendung von (3.8), erhält man für die Varianz von $\widehat{v}(t)$ folgenden Ausdruck:

$$\begin{aligned}\widehat{\mathrm{Var}}[\widehat{v}(t)] &\approx \left(\frac{1}{\log \widehat{S}(t)}\right)^2 \widehat{\mathrm{Var}}[\log \widehat{S}(t)] \\ &\approx \left(\frac{1}{\log \widehat{S}(t)}\right)^2 \sum_{y_{(i)} \le t} \frac{d_i}{(n_i - d_i) n_i} \,.\end{aligned}$$

Das ergibt das punktweise Konfidenzintervall für $\log\left(-\log S(t)\right)$ zum Niveau $1 - \alpha$

$$\log\Big(-\log \widehat{S}_{\mathrm{KM}}(t)\Big) \pm z_{1-\alpha/2}\widehat{\sigma}_v(t), \quad \text{mit} \quad \widehat{\sigma}_v^2(t) := \widehat{\mathrm{Var}}[\widehat{v}(t)] = \left(\frac{\widehat{\sigma}_s(t)}{\log \widehat{S}(t)}\right)^2 .$$

Das Konfidenzintervall von $\log(-\log S(t))$ kann nun durch Rücktransformation mithilfe von $\exp(-\exp(x))$ umgeformt werden zu einem Konfidenzintervall für $S(t)$, d.h.

$$\widehat{S}_{\mathrm{KM}}(t) \exp\big(-\exp(\pm z_{1-\alpha/2}\widehat{\sigma}_v(t))\big)$$

Das **punktweise Konfidenzintervall für** $S(t)$ zum Niveau $1-\alpha$ lautet somit

$$\left(\widehat{S}(t)^{1/\hat{\theta}},\, \widehat{S}(t)^{\hat{\theta}}\right) \quad \text{mit} \quad \hat{\theta} = \exp\left(z_{1-\alpha/2}\widehat{\sigma}_v(t)\right) . \tag{3.14}$$

Beispiel 5 (Fortführung von Beispiel 4).
Für den in Beispiel 4 berechneten Kaplan-Meier Schätzer wird nun ein Konfidenzintervall zum Niveau $\alpha = 0.05$ erstellt.

```
> bsp.konf_int <- survfit(bsp.surv~1, conf.int = .95)
> summary(bsp.konf_int)
Call: survfit(formula = bsp.surv ~ 1, conf.int = .95)
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    5     12       1    0.917  0.0798        0.7729        1.000
   17     11       1    0.833  0.1076        0.6470        1.000
   24      9       1    0.741  0.1295        0.5259        1.000
   32      8       1    0.648  0.1426        0.4211        0.998
   40      6       1    0.540  0.1544        0.3084        0.946
   46      5       1    0.432  0.1568        0.2121        0.880
   47      4       1    0.324  0.1503        0.1306        0.804
   50      3       1    0.216  0.1335        0.0644        0.725
   59      2       1    0.108  0.1014        0.0171        0.680
   74      1       1    0.000     NaN            NA           NA
> plot(bsp.konf_int, xlab="t",
+ ylab = expression(hat(S)(t)), lwd = 3)
```

3.2. Log Rank Test

Im vorangegangenen Kapitel wurde versucht die Survivalfunktion $S(t)$ zu schätzen und punktweise Konfidenzintervalle dafür anzugeben. Nun wird versucht die Survivalfunktionen von mehrerer Gruppen zu vergleichen und auf Unterschiede zu überprüfen. Der bekannteste Test hierfür ist der sogenannte Log Rank Test. Die Ausführungen zur Herleitung des Log Rank Tests orientieren sich hierbei an Liu (2012) und Klein und Moeschberger (2003).

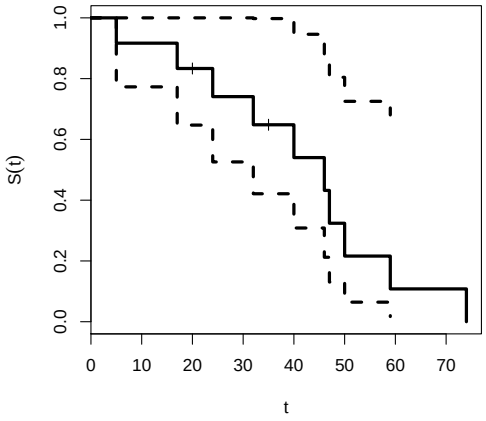

Abbildung 3.3.: Punktweises Konfidenzintervall der Survivalfunktion zum Niveau $1 - \alpha = 0.95$ basierend auf den Daten aus Beispiel 5.

3.2.1. Log Rank Test für zwei Gruppen

Angenommen die Daten lassen sich in zwei Gruppen G_1 und G_2 unterteilen, so soll nun überprüft werden, ob sich die Survivalfunktionen dieser beiden Gruppen statistisch unterscheiden. Somit testet man die Hypothese

$$\begin{aligned} H_0 &: S_1(t) = S_2(t) \quad \forall t \geq 0 \quad \text{gegen} \\ H_1 &: S_1(t) \neq S_2(t) \quad \text{für zumindest einen Zeitpunkt } t\,, \end{aligned}$$

wobei $S_1(t)$ bzw. $S_2(t)$ die Survivalfunktionen für Gruppe G_1 bzw. G_2 sind.

Für die $\acute{n}$ sortierten beobachteten Ausfallzeitpunkte $y_{(1)} < y_{(2)} < \cdots < y_{(\acute{n})}$, mit $\acute{n}$ der Anzahl an unterschiedlichen Ausfallzeitpunkten (d.h. $\acute{n} \leq n$), sei n_1 bzw. n_2 die beobachtete Anzahl an Individuen in G_1 bzw. G_2. Weiters sei n_{1i} sowie n_{2i} die beobachtete Anzahl an gefährdeten Individuen zum i-ten Ausfallzeitpunkt und d_{1i} sowie d_{2i} die Anzahl an beobachteten Ausfälle zum i-ten Ausfallzeitpunkt für G_1 und G_2 (siehe Tabelle 3.1).

Es kann die Anzahl an Ausfällen d_{1i} zum i-ten Ausfallzeitpunkt $y_{(i)}$ für G_1 als **hypergeometrisch verteilte Zufallsvariable** (siehe Definition A.1.1)

Gruppe	Ausfall		Anzahl unter Risiko
	Ja	Nein	
G_1	d_{1i}	$n_{1i} - d_{1i}$	n_{1i}
G_2	d_{2i}	$n_{2i} - d_{2i}$	n_{2i}
Insgesamt:	$d_i = d_{1i} + d_{2i}$	$n_i - d_i$	$n_i = n_{1i} + n_{2i}$

Tabelle 3.1.: Anzahl an Ausfällen und Nichtausfällen zum i-ten Ausfallzeitpunkt $y_{(i)}$ für zwei Gruppen.

angesehen werden, mit Wahrscheinlichkeitsfunktion

$$\Pr(D_{1i} = d_{1i}; n_i, n_{1i}, d_i) = \frac{\binom{d_i}{d_{i1}}\binom{n_i - d_i}{n_{1i} - d_{1i}}}{\binom{n_i}{n_{1i}}}$$

und Erwartungswert und Varianz (siehe Satz A.1.2)

$$\begin{aligned} \mathrm{E}[D_{1i}] &= \frac{d_i n_{1i}}{n_i}\,, \\ \mathrm{Var}[D_{1i}] &= \frac{d_i(n_i - d_i)n_{1i}n_{2i}}{n_i^2(n_i - 1)}\,. \end{aligned}$$

Die Idee hinter dem Log Rank Test ist, die Differenz zwischen beobachteten und erwarteten Ausfällen pro Zeitpunkt zu untersuchen, also

$$\tilde{D} = \sum_{i=1}^{\acute{n}} \big(D_{1i} - \mathrm{E}[D_{1i}]\big)$$

mit deren Erwartungswert

$$\mathrm{E}[\tilde{D}] = \sum_{i=1}^{\acute{n}} \big(\mathrm{E}[D_{1i}] - \mathrm{E}[D_{1i}]\big) = 0$$

und deren Varianz

$$\mathrm{Var}[\tilde{D}] = \sum_{i=1}^{\acute{n}} \mathrm{Var}[D_{1i}]\,.$$

Es kann gezeigt werden, dass $\tilde{D}$ durch eine Normalverteilung approximiert werden kann falls die Anzahl an unterschiedlichen Ausfallzeitpunkten $\acute{n}$ nicht zu klein ist, somit gilt

$$Z_{\text{logrank}} = \frac{\tilde{D}}{\sqrt{\text{Var}[\tilde{D}]}} \sim N(0,1) \tag{3.15}$$

bzw.

$$Q_{\text{logrank}} = \frac{\tilde{D}^2}{\text{Var}[\tilde{D}]} \sim \chi^2_1 \,. \tag{3.16}$$

Falls H_0 gilt, muss $\tilde{D}$ die Summe der Differenzen zwischen beobachteten und erwarteten Ausfällen pro Zeitpunkt klein sein. Umgekehrt deutet ein großer Wert von $\tilde{D}$ auf eine Verletzung von H_0. Mithilfe von Z_{logrank} bzw. Q_{logrank} und den Quantilen der Normalverteilung bzw. χ^2_1-Verteilung kann somit die Nullhypothese H_0 überprüft werden (vgl. Collett, 2003).

3.2.2. Log Rank Test für drei oder mehr Gruppen

Die Hypothesen für diesen Test lauten:

$$\begin{aligned} &H_0 : S_1(t) = S_2(t) = \cdots = S_k(t), \quad \forall t \geq 0 \quad \text{gegen} \\ &H_1 : \exists\, t \geq 0 : S_i(t) \neq S_j(t) \quad 1 \leq i, j \leq k \,. \end{aligned}$$

Der Log Rank Test für mehrere Gruppen basiert auf den gleichen Überlegungen wie der Log Rank Test für zwei Gruppen. Es wird hier ebenso die Differenz zwischen beobachteten und erwarteten Ausfällen zu den verschiedenen Ausfallzeitpunkten untersucht. In Tabelle 3.2 wird die Notation für mehrere Gruppen vorgestellt.

Um eine übersichtlichere Forumulierung des Log Rank Tests zu erhalten, wird dieser in Matrix Vektor Schreibweise formuliert. Sei $\boldsymbol{O}_i$ der Vektor der zum i-ten Ausfallzeitpunkt $y_{(i)}$ beobachteten Ausfälle $\boldsymbol{O}_i = (d_{1i}, \ldots, d_{(k-1)i})^\intercal$. Die Verteilung des Vektors $\boldsymbol{O}_i$ kann als **multivariat hypergeometrisch verteilt** angenommen werden (vgl. Definition A.1.2, Liu, 2012). Für den

Gruppe	Ausfall		Anzahl unter Risiko
	Ja	Nein	
G_1	d_{1i}	$n_{1i} - d_{1i}$	n_{1i}
G_2	d_{2i}	$n_{2i} - d_{2i}$	n_{2i}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
G_k	d_{ki}	$n_{ki} - d_{ki}$	n_{ki}
Insgesamt:	$d_i = \sum_{j \leq k} d_{ji}$	$n_i - d_i$	$n_i = \sum_{j \leq k} n_{ji}$

Tabelle 3.2.: Anzahl an Ausfällen und Nichtausfällen zum i-ten Ausfallzeitpunkt $y_{(i)}$ für k Gruppen.

Vektor $\boldsymbol{E}_i = \mathrm{E}(\boldsymbol{O}_i)$ mit den erwarteten Anzahlen an Ausfällen zum Zeitpunkt $y_{(i)}$ gilt

$$\boldsymbol{E}_i = \left(\frac{d_i n_{1i}}{n_i}, \ldots, \frac{d_i n_{(k-1)i}}{n_i} \right)^\intercal . \tag{3.17}$$

Die Kovarianzmatrix von $\boldsymbol{O}_i$ (vgl. Satz A.1.3) lautet

$$\boldsymbol{V}_i = \begin{pmatrix} v_{11i} & v_{12i} & \cdots & v_{1(k-1)i} \\ v_{21i} & v_{22i} & \cdots & v_{2(k-1)i} \\ \vdots & & \ddots & \vdots \\ v_{(k-1)1i} & \cdots & & v_{(k-1)(k-1)i} \end{pmatrix} ,$$

mit

$$v_{lli} = \frac{n_{li}(n_i - n_{li}) d_i (n_i - d_i)}{n^2 (n_i - 1)} \quad \text{und} \quad v_{lmi} = \frac{n_{li} n_{mi} d_i (n_i - d_i)}{n^2 (n_i - 1)} .$$

Die **Log Rank Teststatistik** lautet somit

$$Q_{\text{logrank}} = \tilde{\boldsymbol{D}}^\intercal \boldsymbol{V}^{-1} \tilde{\boldsymbol{D}} \tag{3.18}$$

mit

$$\tilde{D}_j = \sum_{i=1}^{\acute{n}} (\boldsymbol{O}_{ij} - \boldsymbol{E}_{ij}) \quad \text{und} \quad V_j = \sum_{i=1}^{\acute{n}} V_{ij} , \quad \text{für } j = 1, \ldots, k . \tag{3.19}$$

Da $\tilde{\boldsymbol{D}} = (\tilde{D}_1, \ldots, \tilde{D}_k)$ durch eine multinomiale Normalverteilung mit $k-1$ Freiheitsgraden approximiert werden kann und $\mathrm{E}[\tilde{\boldsymbol{D}}] = \mathbf{0}$ gilt, ist Q_{logrank} asymptotisch χ^2_{k-1} verteilt. Falls die Anzahl an beobachteten und erwarteten Ausfällen pro Zeitpunkt und Gruppe stark voneinander abweichen (d.h. $\tilde{D}_j$, $j = 1, \ldots, k$ ist groß), deutet dass auf eine Verletzung von H_0 hin. Somit kann die Nullhypothese mithilfe von Q_{logrank} und den Quantilen der χ^2_{k-1}-Verteilung überprüft werden.

Der Log Rank Test beinhaltet keine Gewichtung der Gruppen bei der Berechnung der Teststatistik. Da beim Vergleich von mehreren Gruppen die Gruppen untereinander zumeist unterschiedlich große Anzahl an Individuen haben, kann es zu Verfälschungen kommen. Um diese Ungleichheiten unter den Gruppen auszugleichen wurde der Log Rank Test zu einer gewichteten Teststatistik erweitert (vgl. Liu, 2012).

In Anlehnung an (3.19) wird eine neue gewichtete Differenz der erwarteten und beobachteten Ausfälle formuliert und zwar

$$\tilde{\boldsymbol{D}}_{\mathrm{w},j} = \sum_{i=1}^{\acute{n}} w(y_{(i)}) \left(\boldsymbol{O}_{ij} - \boldsymbol{E}_{ij}\right), \qquad j = 1, \ldots, k\,, \tag{3.20}$$

mit dem Gewichten $w(y_{(i)})$ zum Ausfallszeitpunkt $y_{(i)}$. Der **gewichtete Log Rank Test** lautet somit

$$Q_{\mathrm{w}} = \tilde{\boldsymbol{D}}_{\mathrm{w}}^{\intercal} \boldsymbol{V}_{\mathrm{w}}^{-1} \tilde{\boldsymbol{D}}_{\mathrm{w}} \tag{3.21}$$

mit

$$\boldsymbol{V}_{\mathrm{w},j} = \sum_{i=1}^{\acute{n}} w(y_{(i)})^2 \boldsymbol{V}_{ij}\,, \qquad j = 1, \ldots, k\,.$$

und ist asymptotisch χ^2_{k-1} verteilt. Abhängig von der Wahl der Gewichte ergeben sich unterschiedliche Teststatistiken. In der Tabelle 3.3 sind die gängigsten Teststatistiken mit den zugehörigen Gewichten aufgelistet.

Test	**Gewicht $w(y_{(i)})$**
Log Rank	$w(y_{(i)}) = 1$
Gehan Wilcoxon	$w(y_{(i)}) = n_i$
Fleming & Harrington	$w(y_{(i)}) = \widehat{S}(y_{(j-1)})^p(1 - \widehat{S}(y_{(j-1)}))^q$ mit $p, q \geq 0$
Taron Waron	$w(y_{(i)})) = \sqrt{n_i}$

Tabelle 3.3.: Einige Teststatistiken mit den dazugehörigen Gewichten $w(y_{(i)})$.

Bemerkung 5 (Warum nur $k-1$ Elemente verwendet werden). *Da die Komponenten der Vektoren $\tilde{\boldsymbol{D}}$ und $\tilde{\boldsymbol{D}}_{\mathrm{w}}$ linear abhängig sind (d.h. $\sum_{j=1}^{k} \tilde{D}_j = 0$ bzw. $\sum_{j=1}^{k} \tilde{D}_{\mathrm{w},j} = 0$), werden nur $k-1$ Komponenten des Vektors betrachtet werden (vgl. Klein und Moeschberger, 2003). Zum Beispiel wird auch für den Log Rank Test für zwei Gruppen aus Abschnitt 3.2.1 nur die erste Gruppe betrachtet.*

4. Parametrische Modelle

4.1. Gängige Wahrscheinlichkeitsverteilungen

Für die Survival Analysis existieren Wahrscheinlichkeitsverteilungen welche die Verteilung von Lebenszeiten besonders gut beschreiben. Zu diesen Verteilungen gehören u.a. die Exponential-Verteilung, die Weibull-Verteilung, die Log-Normal-Verteilung und die Log-Logistik-Verteilung. Diese Verteilungen werden nun genauer diskutiert und auf die Eignung zur Beschreibung von Lebenszeiten überprüft. Die Ausführungen in diesem Abschnitt folgen in großen Teilen Lawless (2003), Glomb (2007) und Tableman und Kim (2004).

Bemerkung 6.
In diesem Abschnitt werden Verteilungen für die Lebenszeit T betrachtet. Es wird somit Zensierung und Trunkierung nicht berücksichtigt.

4.1.1. Log-Lokation-Skalen-Modell

Die in den Abschnitten 4.1.3, 4.1.5 und 4.1.6 vorgestellten Verteilungen sind alle Teil der Log-Lokations-Skalen-Verteilungsfamilie. Aus diesem Grund wird in diesem Abschnitt die Log-Lokations-Skalen-Verteilungsfamilie nun genauer betrachtet.

Eine Zufallsvariable W stammt aus der parametrischen Lokation-Skalen-Familie falls sich die Zufallsvariable, mithilfe von Lokationsparameter u und Skalenparameter b, schreiben lässt als

$$W = u + bZ \qquad u \in \mathbb{R},\, b > 0\,, \tag{4.1}$$

mit Z der standardisierten Lokation-Skalen-Zufallsvariable mit $u = 0$ und $b = 1$. Die dazugehörige Dichte- und Survivalfunktion lautet

$$f_W(w) = \frac{1}{b} f_Z\left(\frac{w-u}{b}\right) \quad w, u \in \mathbb{R},\, b > 0 \tag{4.2}$$

und

$$S_W(w) = S_Z\left(\frac{w-u}{b}\right). \tag{4.3}$$

Zu der Lokation-Skalen-Familie gehören die Extremwert-, Logistik- und Normalverteilung. Die standardisierten Survivalfunktionen für diese Verteilungen lauten

$S_Z(z) = \exp(-\exp(z))$	Extremwert-Verteilung
$S_Z(z) = 1 - \Phi(z)$	Normal-Verteilung
$S_Z(z) = (1 + \exp(z))^{-1}$	Logistik-Verteilung

Eine Zufallsvariable T stammt aus der Log-Lokation-Skalen-Familie falls sich die Zufallsvariable schreiben lässt als

$$\log(T) = W = u + bZ \quad u \in \mathbb{R},\, b > 0\,, \tag{4.4}$$

mit der Zufallsvariable W aus der Lokation-Skalen-Familie. Für die Dichte- und Survivalfunktion von $T = \exp(W)$ gilt nun

$$f_T(t) = \frac{1}{bt} f_Z\left(\frac{\log(t)-u}{b}\right) \tag{4.5}$$

und

$$S_T(t) = S_Z\left(\frac{\log(t)-u}{b}\right). \tag{4.6}$$

Zu der Familie der Log-Lokation-Skalen-Verteilungen gehören Weibull-, Log-Normal- und Log-Logistik-Verteilungen. In der Tabelle 4.1 ist nochmals der Zusammenhang der einzelnen Verteilungen aus der Lokations-Skalen-Famile und der Log-Lokations-Skalen-Familie zusammengefasst.

T	$\Longleftrightarrow$	$W = \log(T)$
Weibull-Verteilung	$\Longleftrightarrow$	Extremwert-Verteilung
Log-Normal-Verteilung	$\Longleftrightarrow$	Normal-Verteilung
Log-Logistik-Verteilung	$\Longleftrightarrow$	Logistik-Verteilung

Tabelle 4.1.: Zusammenhang zwischen den Lokations-Skalen-Modellen und den Log-Lokations-Skalen-Modellen.

4.1.2. Exponential-Modell

Das Exponential-Modell hat die Dichtefunktion

$$f(t) = \lambda \exp(-\lambda t) \quad t \geq 0, \lambda > 0\,, \tag{4.7}$$

die Survivalfunktion und Hazardfunktion

$$S(t) = \exp(-\lambda t) \quad \text{und} \quad h(t) = \lambda \quad t \geq 0, \lambda > 0\,. \tag{4.8}$$

Die Exponential-Verteilung war eine der ersten verwendeten Verteilung der Survival Analysis. Jedoch ist sie aufgrund ihrer konstanten Hazardfunktion, welche aus der Gedächtnislosigkeit der Exponential-Verteilung resultiert, nur bedingt brauchbar.

Die Beziehung zwischen der Survivalfunktion und der kumulativen Hazardfunktion lautet

$$\log(H(t)) = \log\big(-\log(S(t)\big) = \log(\lambda) + \log(t)$$

oder mithilfe von $\log(t)$ ausgedrückt

$$\log(t) = -\log(\lambda) + \log\big(-\log(S(t)\big)\,. \tag{4.9}$$

Folglich weist der Plot $\log(t)$ gegen $\log\left(-\log(S(t)\right)$ einer exponentialverteilten Lebenszeit eine Steigung von 1 und einen Intercept von $-\log(\lambda)$ auf.

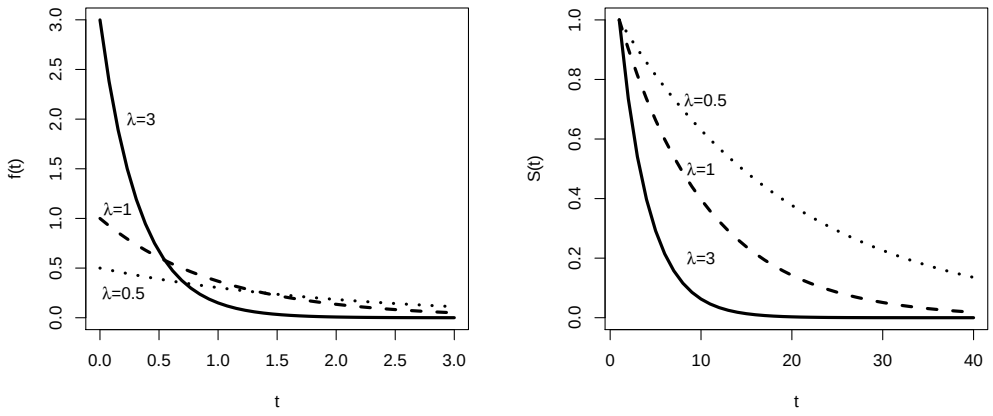

Abbildung 4.1.: Links zu sehen ist die Dichtefunktion für ein Exponential-Modell und rechts ist die Survivalfunktion zu sehen.

4.1.3. Weibull-Modell

Das Weibull-Modell hat die Dichtefunktion

$$f(t) = \lambda\beta(\lambda t)^{\beta-1} \exp\left(-(\lambda t)^{\beta}\right) \quad \lambda, \beta, t > 0\,, \tag{4.10}$$

die Survivalfunktion

$$S(t) = \exp\left(-(\lambda t)^{\beta}\right) \quad \lambda, \beta, t > 0 \tag{4.11}$$

und die Hazardfunktion

$$h(t) = \lambda\beta(\lambda t)^{\beta-1} \quad \lambda, \beta, t > 0\,. \tag{4.12}$$

Die Hazardfunktion einer Weibull-Verteilung ist monoton steigend für $\beta > 1$, monoton fallend für $\beta < 1$ und konstant λ für $\beta = 1$, daher wird β auch als Gestaltparameter (engl. shape parameter) bezeichnet. In Abbildung 4.2 ist die Hazardfunktion und Survivalfunktion für verschiedene β Werte zu sehen. Der Parameter λ wird oft als Skalenparameter bezeichnet, da er für unterschiedliche Werte nicht die Form der Funktion ändert sondern die Position auf der horizontalen (t) Achse.

Bemerkung 7.
Die Weibull-Verteilung ist eine Exponential-Verteilung für $\beta = 1$.

Die Dichte-, Hazard- und Survivalfunktion des Weibull-Modells weisen eine äußerst simple Form auf. Aufgrund dieser Eigenschaften ist das Weibull-Modell auch das wohl meist verwendete parametrische Survival Modell.

Die Beziehung zwischen der Survivalfunktion und der kumulativen Hazardfunktion lautet

$$\log\big(H(t)\big) = \log\big(-\log(S(t)\big) = \beta\big(\log(\lambda) + \log(t)\big)\,,$$

oder anders ausgedrückt

$$\log(t) = -\log(\lambda) + \sigma \log\big(-\log(S(t)\big)\,, \tag{4.13}$$

mit $\sigma = 1/\beta$. Folglich muss der Plot von $\log(t)$ gegen $\log\big(-\log(S(t)\big)$ einer weibullverteilten Lebenszeit einer Gerade mit Steigung $\sigma = 1/\beta$ und Intercept $-\log(\lambda)$ entsprechen.

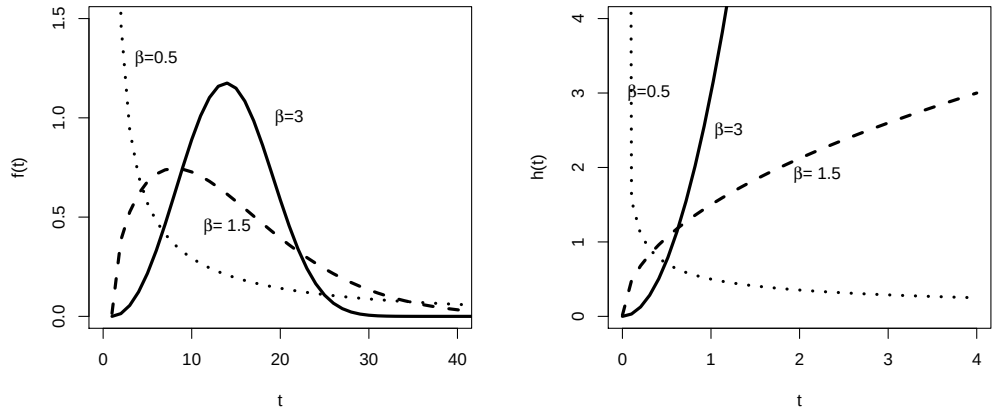

Abbildung 4.2.: Links ist die Dichtefunktion und rechts ist die Hazardfunktion für das Weibull-Modell dargestellt.

4.1.4. Extremwert-Modell (Gumbel-Modell)

Das Extremwert-Modell hat die Dichtefunktion

$$f(w) = b^{-1} \exp\left(\frac{w-u}{b} - \exp\left(\frac{w-u}{b}\right)\right) \quad w, u \in \mathbb{R}, b > 0 \tag{4.14}$$

und Survivalfunktion

$$S(w) = \exp\left(-\exp\left(\frac{w-u}{b}\right)\right) \quad w, u \in \mathbb{R}, b > 0\,. \tag{4.15}$$

Die Extremwert-Verteilung ist vor allem aufgrund der Beziehung mit der Weibull-Verteilung interessant. Ist nämlich die Zufallsvariable T weibullverteilt, so folgt $W = \log(T)$ einer Extremwert-Verteilung mit $u = -\log\lambda$ und $b = 1/\beta$ (siehe Abschnitt 4.1.1). Da bei der Analyse von Lebenszeiten oft mit den logarithmierten Lebenszeiten gearbeitet wird, ist diese Verteilung auch für die Survival Analysis relevant.

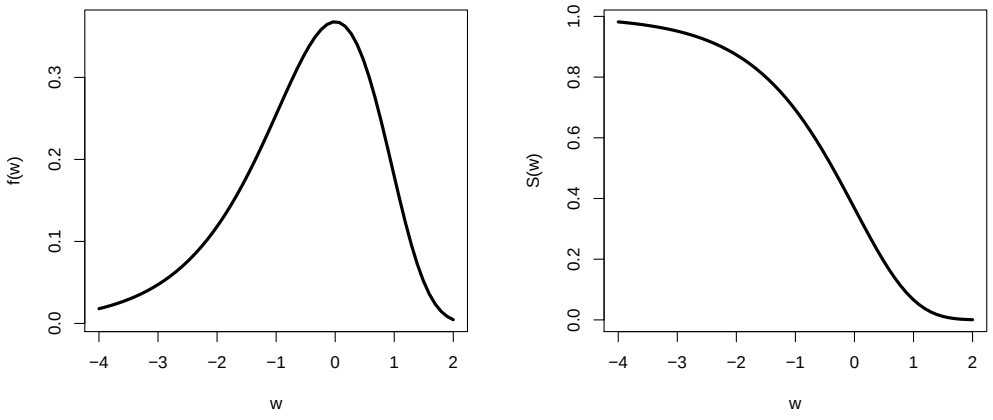

Abbildung 4.3.: Dichtefunktion (links) und Survivalfunktion (rechts) für die Standard Extremwert-Verteilung EV(0,1).

4.1.5. Log-Normal-Modell

Das Log-Normal-Modell hat die Dichtefunktion

$$f(t) = \frac{1}{(2\pi)^{1/2}\sigma t} \exp\left(-\frac{1}{2}\left(\frac{\log(t)-\mu}{\sigma}\right)^2\right) \quad \mu \in \mathbb{R},\, \sigma^2, t > 0 \qquad (4.16)$$

und Survivalfunktion

$$S(t) = 1 - \Phi\left(\frac{\log(t)-\mu}{\sigma}\right) \quad \mu \in \mathbb{R},\, \sigma^2, t > 0. \qquad (4.17)$$

Eine Lebenszeit T ist log-normalverteilt, falls $W = \log(T)$ einer Normalverteilung folgt (siehe Abschnitt 4.1.1), mit Erwartungswert μ und Varianz σ^2. Folglich hat W die Form $W = \mu + \sigma Z$, wobei Z eine standard-normalverteilte Zufallsvariable ist.

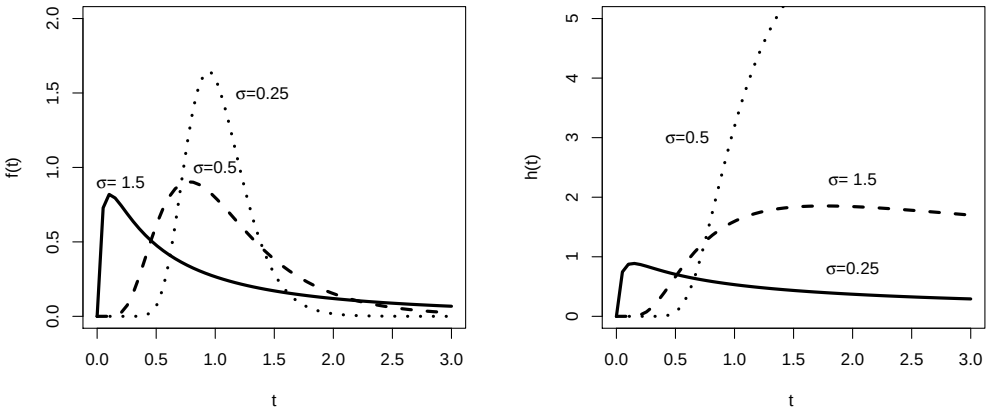

Abbildung 4.4.: Dichtefunktion (links) und Survivalfunktion (rechts) für die Log-Normal-Verteilung, mit $\mu = 0$ und $\sigma = 0.25, 0.5, 1.5$

Die Hazardfunktion hat den Wert 0 für $t = 0$ und steigt an bis sie ihr Maximum erreicht hat, danach ist diese Funktion monton fallend (siehe Abbildung 4.4).

Durch eine Umformung der Survivalfunktion von Log-normalverteilten Zufallsvariablen erhält man folgende Beziehung

$$\Phi^{-1}\big(1 - S(t)\big) = \Phi^{-1}\left(1 - 1 + \Phi\left(\frac{\log(t) - \mu}{\sigma}\right)\right) = \frac{\log(t) - \mu}{\sigma}$$

oder anders ausgedrückt

$$\log(t) = \mu + \sigma\Phi^{-1}\big(1 - S(t)\big)\,. \tag{4.18}$$

Folglich muss der Plot $\log(t)$ gegen $\Phi^{-1}\big(1-S(t)\big)$ bei einer log-normalverteilten Zufallsvariable einer Gerade mit Steigung σ und Intercept μ entsprechen.

4.1.6. Log-Logistik-Modell

Das Log-Logistik-Modell hat die Dichtefunktion

$$f(t) = (\lambda\alpha)(t\lambda)^{\alpha-1}\big(1 + (t\lambda)^{\alpha}\big)^{-2} \quad t, \alpha, \beta > 0\,, \tag{4.19}$$

die Survivalfunktion

$$S(t) = \big(1 - (t\lambda)^{\alpha}\big)^{-1} \quad t, \alpha, \beta > 0 \tag{4.20}$$

und die Hazardfunktion

$$h(t) = (\lambda\alpha)(t\lambda)^{\alpha-1}\big(1 + (t\lambda)^{\alpha}\big)^{-1} \quad t, \alpha, \beta > 0\,. \tag{4.21}$$

Eine Lebenszeit T heißt Log-Logistikverteilt, falls $W = \log(T)$ einer Logistik-Verteilung folgt (siehe Abschnitt 4.1.1).

Diese Verteilung ist sehr beliebt, da sie wie die Weibull-Verteilung eine einfache algebraische Form für die Survival- und Hazardfunktion aufweist. Die Hazardfunktion ist abgesehen vom Nenner $(1 + (t/\lambda)^{\alpha})^{-2}$ identisch mit der Hazardfunktion der Weibull-Verteilung.

Für die Gestaltparameter Werte $\alpha < 1$ fällt die Hazardfunktion monoton, wobei hier $\lim_{t\to 0-} h(t) = \infty$ gilt. Andererseits ähnelt die Form der Hazardfunktion für Gestaltparameter Werte $\alpha > 1$ stark einer Log-Normal-Verteilung Hazardfunktion, da die Funktion bis $t = (\alpha-1)^{1/\alpha}/\lambda$ ansteigt und

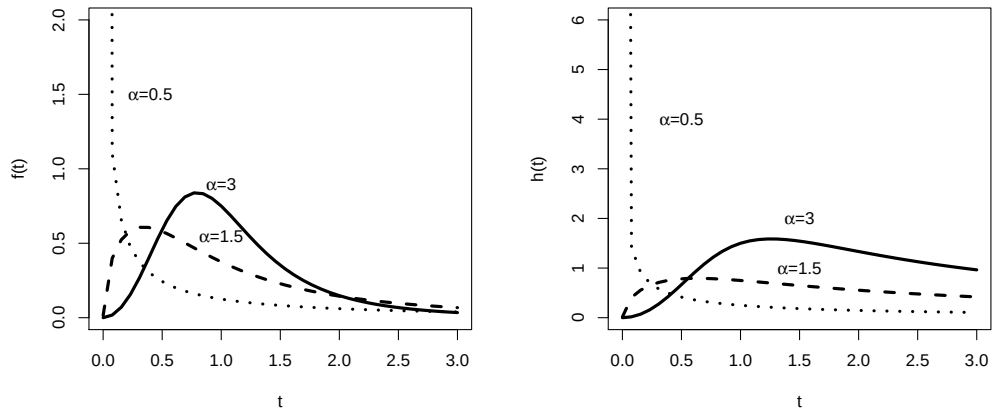

Abbildung 4.5.: Dichtefunktion (links) und Hazardfunktion (rechts) für Log-Logistik-Verteilung, mit $\lambda = 1$ und $\alpha = 3, 1.5, 0.5$.

danach wieder abfällt. Aus diesem Grund und der einfachen algebraischen Form wird die Log-Logistik-Verteilung auch oft anstelle einer Log-Normal-Verteilung verwendet.

Die Chance (odds) nach dem Zeitpunkt t noch zu leben ist

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

Somit besteht eine lineare Beziehung zwischen $\log t$ und der logarithmierten Chance (log-odds) nach t noch zu leben. Diese Beziehung lautet

$$\log t = \mu + \sigma \left(- \log \left(\frac{S(t)}{1 - S(t)} \right) \right) , \tag{4.22}$$

mit $\mu = \log \lambda$ und $\sigma = 1/\alpha$. Somit kann die Modellanpassungsgüte mithilfe des Plots $\log(t)$ gegen $-\log\left(\frac{S(t)}{1-S(t)}\right)$ beurteilt werden.

4.2. AFT-Modelle

Mit dem **Accelerated-Failure-Time-Modell (AFT-Modell)** wird nun ein parametrisches Modell vorgestellt, welches zum einen den Einfluss von Kovariablen $\boldsymbol{X}^\intercal = (X_1, \ldots, X_p)$ auf die Lebenszeit berücksichtigt und zum anderen eine Verteilungsannahme für die beobachteten Lebenszeiten trifft. Die am häufigst angenommenen Verteilungen für AFT-Modelle stammen aus der Log-Lokations-Skalen Verteilungsfamilie und lauten Weibull-, Log-Normal- und Log-Logistik-Verteilung (siehe Abschnitt 4.1). Die Ausführungen in diesem Abschnitt folgen Glomb (2007), Klein und Moeschberger (2003) und Collett (2003).

Bei AFT-Modellen wirken sich die Kovariablen eines Individuums multiplikativ auf die Zeitskala aus. Somit beeinflussen die Kovariablen die Geschwindigkeit in welcher ein Individuum auf der Zeitskala voranschreitet. Durch diese Art der Modellierung ist das Modell und seine Parameter sehr gut interpretierbar (vgl. Collett, 2003).

Für ein AFT-Modell lässt sich die Survivalfunktion für ein Individuum mit Kovariablenvektor $\boldsymbol{X}$ zum Zeitpunkt t ($t \geq 0$), schreiben als Basis Survivalfunktion (engl. baseline survival function) S_0 zum Zeitpunkt $t \cdot \exp(\boldsymbol{\gamma}^\intercal \boldsymbol{X})$, d.h.

$$S(t|\boldsymbol{X}) = S_0\big(t \cdot \exp(\boldsymbol{\gamma}^\intercal \boldsymbol{X})\big)\,. \tag{4.23}$$

Der Faktor $\exp(\boldsymbol{\gamma}^\intercal \boldsymbol{X})$ wird **acceleration factor** genannt, da er abhängig vom Kovariablenvektor den Wert der Zeitskala für die Basis Survivalfunktion verändert (der acceleration factor bestimmt die Geschwindigkeit in welcher ein Individuum auf der Zeitskala voranschreitet).

Mithilfe von (2.4) gilt für die Hazardfunktion eines AFT-Modells

$$h(t|\boldsymbol{X}) = \exp(\boldsymbol{\gamma}^\intercal \boldsymbol{X}) h_0\big(t \cdot \exp(\boldsymbol{\gamma}^\intercal \boldsymbol{X})\big)\,, \tag{4.24}$$

mit $h_0(t)$ der Basis Hazardfunktion (vgl. Klein und Moeschberger, 2003).

Beispiel 6 (Kleinbaum und Klein, 2005).
Seien $S_M(t)$ und $S_H(t)$ die Survivalfunktionen für Menschen und Hunde. Es wird oft angenommen, dass Hunde 7-mal so schnell altern als Menschen. Umgelegt auf das AFT-Modell bedeutet dies, dass die Survivalwahrscheinlichkeit

eines Hundes für t Jahre gleich der Survivalwahrscheinlichkeit eines Menschen für $7t$ Jahre ist, d.h.

$$S_H(t) = S_M(7t)\,.$$

Somit ist der acceleration factor für Hunde gleich 7.

Eine parametrische Beziehung zwischen Kovariablenvektor und Lebenszeit lässt sich auch mit einem **log-linearen Modell** beschrieben, d.h.

$$W = \log(T) = \beta_0 + \boldsymbol{\beta}^\intercal \boldsymbol{X} + bZ\,, \tag{4.25}$$

mit dem Vektor der Regressionskoeffizienten $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$, dem Intercept β_0, der Zufallsvariabe Z und dem unbekannten Variabilitätsparameter b. Hierbei wird angenommen, dass Z aus der Lokation-Skalen Verteilungsfamilie stammt (siehe Abschnitt 4.1.1) und somit auch W.

Somit gilt für die Survivalfunktion im log-linearen Modell des i-ten Individuums mit Lebenszeit T_i und zugehörigen Kovariablenvektor $\boldsymbol{X}_i$

$$S(t|\boldsymbol{X}_i) = \Pr(T_i \geq t) = \Pr\big(\exp(\beta_0 + \boldsymbol{\beta}^\intercal \boldsymbol{X}_i + bZ) \geq t\big)\,. \tag{4.26}$$

Ein Zusammenhang des AFT-Modells und des log-linearen Modells ist gegeben durch

$$\begin{aligned} S(t|\boldsymbol{X}_i) &= \Pr\big(\exp(\beta_0 + \boldsymbol{\beta}^\intercal \boldsymbol{X}_i + bZ) \geq t\big) \\ &= \Pr\big(\exp(\beta_0 + bZ) \geq t \cdot \exp(-\boldsymbol{\beta}^\intercal \boldsymbol{X}_i)\big) \\ &= S_0\big(t \cdot \exp(-\boldsymbol{\beta}^\intercal \boldsymbol{X}_i)\big)\,, \end{aligned} \tag{4.27}$$

mit S_0 der Survivalfunktion der Zufallsvariable $\exp(\beta_0 + bZ)$, somit gilt $S_0(t) = \Pr(\exp(\beta_0 + bZ) \geq t)$. In Folge entspricht das log-linere Modell (4.25) dem AFT-Modell (4.23) für $\boldsymbol{\gamma} = -\boldsymbol{\beta}$ (vgl. Klein und Moeschberger, 2003).

4.2.1. Inferenz

Bisher wurden für das log-lineare Survivalmodell die tatsächlichen Lebenszeiten T modelliert, d.h.

$$W = \log(T) = \beta_0 + \boldsymbol{\beta}^\intercal \boldsymbol{X} + bZ\,,$$

mit Z aus der Log-Lokation-Skalen Famlie. Ab sofort wird angenommen, dass rechtszensierte Lebenszeiten $Y_1, \ldots, Y_n$ mit $Y_i = \min(T_i, c_i)$, $i = 1, \ldots, n$, und dazugehörigen Rechtszensur-Indikatorvariablen $R_i = I(T_i = Y_i)$ beobachtet werden (siehe Abschnitt 2.3). Somit gilt für das log-lineare Modell für rechtszensierte Lebenszeiten

$$W_i = \log(Y_i) = \beta_0 + \boldsymbol{\beta}^\intercal \boldsymbol{X}_i + bZ\,. \tag{4.28}$$

Ziel ist es nun eine Likelihoodfunktion für das log-lineare Survival Modell mit rechtszensierten Lebenszeiten herzuleiten.

Aus Abschnitt 2.3 ist bekannt, dass für rechtszensierte Lebenszeiten die Likelihoodfunktion folgende Form aufweist

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f_{T_i}(y_i|\boldsymbol{\theta})^{r_i} S_{T_i}(y_i|\boldsymbol{\theta})^{1-r_i} \quad \text{für } \boldsymbol{\theta} = (\beta_0, \boldsymbol{\beta}^\intercal, b)^\intercal\,. \tag{4.29}$$

Laut (4.5) und (4.6) gilt für die Survival- und Dichtefunktion von Zufallsvariablen aus der Log-Lokations-Skalen Familie

$$\begin{aligned} f_T(t|\boldsymbol{\theta}) &= \frac{1}{bt} f_Z\left(\frac{\log(t) - u}{b}\right) \quad \text{und} \\ S_T(t|\boldsymbol{\theta}) &= S_Z\left(\frac{\log(t) - u}{b}\right)\,. \end{aligned} \tag{4.30}$$

Da angenommen wurde, dass W aus der Log-Lokation-Skalen Familie stammt bzw. $Y = \exp(W)$ aus der Lokation-Skalen Familie, folgt aus (4.30) für die Dichte- und Survivalfunktion von Y

$$\begin{aligned} f_T(y|\boldsymbol{\theta}) &= \frac{1}{by} f_Z\left(\frac{\log(y) - u(\boldsymbol{X})}{b}\right) \quad \text{und} \\ S_T(y|\boldsymbol{\theta}) &= S_Z\left(\frac{\log(y) - u(\boldsymbol{X})}{b}\right)\,, \end{aligned} \tag{4.31}$$

mit $u(\boldsymbol{X}) := \beta_0 + \boldsymbol{\beta}^\intercal \boldsymbol{X}$ ($u(\boldsymbol{X})$ fungiert als Lokationsparameter).

Wird nun die Dichte- und Survivalfunktion $f_T(y|\boldsymbol{\theta})$ und $S_T(y|\boldsymbol{\theta})$ (4.31) in die Likelihoodfunktion (4.29) eingesetzt, kann nun die **Likelihoodfunktion**

für die logarithmierten rechtszensierten Lebenszeiten (y_i, r_i) folgendermaßen formuliert werden:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left(\frac{1}{by_i} f_Z \left(\frac{\log(y_i) - u(\boldsymbol{X}_i)}{b} \right) \right)^{r_i} S_Z \left(\frac{\log(y_i) - u(\boldsymbol{X}_i)}{b} \right)^{1-r_i} . \tag{4.32}$$

Folglich gilt für die log-Likelihoodfunktion, wenn man $z_i := (\log(y_i) - u(\boldsymbol{X}_i))/b$ setzt, gleich

$$\begin{aligned} l(\boldsymbol{\theta}) &= \log L(\boldsymbol{\theta}) \\ &= \sum_{i=1}^{n} \big(-r_i \log(by_i) + r_i f_Z(z_i) + (1 - r_i) \log(S_Z(z_i)) \big) \\ &= c - r \log(b) + \sum_{i=1}^{n} \big(r_i f_Z(z_i) + (1 - r_i) \log(S_Z(z_i)) \big) , \end{aligned} \tag{4.33}$$

mit $r := \sum_{i=1}^{n} r_i$ und $c := \sum_{i=1}^{n} r_i \log(y_i)$.

Die folgenden Ausführungen zur Herleitung der ersten und zweiten partiellen Ableitung von $l(\boldsymbol{\theta})$ orientieren sich stark an Glomb (2007) und Lawless (2003). Für die ersten partiellen Ableitungen von $z_i := z_i(\beta_0, \boldsymbol{\beta}, b)$ für $\beta_j \neq \beta_0$, $j = 1, \ldots, p$, gilt

$$\frac{\partial z_i}{\partial \beta_0} = -\frac{1}{b}, \quad \frac{\partial z_i}{\partial \beta_j} = -\frac{1}{b} x_{ij} \quad \text{und} \quad \frac{\partial z_i}{\partial b} = -\frac{1}{b} z_i . \tag{4.34}$$

Nun können mithilfe der in (4.33) hergeleiteten Log-Likelihoodfunktion und den partiellen Ableitungen aus (4.34) die Scorefunktionen für β_0, β_j und b

berechnet werden. Damit resultieren

$$\begin{aligned} U_{\beta_0}(\boldsymbol{\theta}) = \frac{\partial l(\boldsymbol{\theta})}{\partial \beta_0} &= \sum_{i=1}^{n} \left(r_i \frac{\partial \log f_Z(z_i)}{\partial z_i} \frac{\partial z_i}{\partial \beta_0} + (1 - r_i) \frac{\partial \log S_Z(z_i)}{\partial z_i} \frac{\partial z_i}{\partial \beta_0} \right) \\ &= -\frac{1}{b} \sum_{i=1}^{n} \left(r_i \frac{\partial \log f_Z(z_i)}{\partial z_i} + (1 - r_i) \frac{\partial \log S_Z(z_i)}{\partial z_i} \right) , \end{aligned} \tag{4.35}$$

$$\begin{aligned} U_{\beta_j}(\boldsymbol{\theta}) = \frac{\partial l(\boldsymbol{\theta})}{\partial \beta_j} &= \sum_{i=1}^{n} \left(r_i \frac{\partial \log f_Z(z_i)}{\partial z_i} \frac{\partial z_i}{\partial \beta_j} + (1 - r_i) \frac{\partial \log S_Z(z_i)}{\partial z_i} \frac{\partial z_i}{\partial \beta_j} \right) \\ &= -\frac{1}{b} \sum_{i=1}^{n} \left(r_i \frac{\partial \log f_Z(z_i)}{\partial z_i} + (1 - r_i) \frac{\partial \log S_Z(z_i)}{\partial z_i} \right) x_{ij} , \end{aligned} \tag{4.36}$$

$$\begin{aligned} U_b(\boldsymbol{\theta}) = \frac{\partial l(\boldsymbol{\theta})}{\partial b} &= \sum_{i=1}^{n} \left(r_i \frac{\partial \log f_Z(z_i)}{\partial z_i} \frac{\partial z_i}{\partial b} + (1 - r_i) \frac{\partial \log S_Z(z_i)}{\partial z_i} \frac{\partial z_i}{\partial b} \right) \\ &= -\frac{r}{b} - \frac{1}{b} \sum_{i=1}^{n} \left(r_i \frac{\partial \log f_Z(z_i)}{\partial z_i} + (1 - r_i) \frac{\partial \log S_Z(z_i)}{\partial z_i} \right) z_i \end{aligned} \tag{4.37}$$

(vgl. Glomb, 2007; Lawless, 2003). In Matrix-Schreibweise lässt sich der Scorevektor schreiben als

$$U(\boldsymbol{\theta}) = \begin{pmatrix} U_{\beta_0}(\boldsymbol{\theta}) \\ U_{\boldsymbol{\beta}}(\boldsymbol{\theta}) \\ U_b(\boldsymbol{\theta}) \end{pmatrix} . \tag{4.38}$$

mit $U_{\boldsymbol{\beta}}(\boldsymbol{\theta}) = \big(U_{\beta_1}(\boldsymbol{\theta}), \ldots, U_{\beta_p}(\boldsymbol{\theta})\big)^\intercal$. Die ML-Schätzer (Maximum-Likelihood-Estimator MLE) von $\boldsymbol{\theta} = (\beta_0, \boldsymbol{\beta}^\intercal, b)^\intercal$ lassen sich durch Lösen des Gleichungssystems $U(\boldsymbol{\theta}) = \mathbf{0}$ finden. In der `survreg`-Funktion in `R` wird der MLE für die unbekannten Parameter $\boldsymbol{\theta}$ mithilfe des Newton-Raphson-Verfahrens bestimmt (siehe Therneau, 2015).

Weiters sind die zweiten partiellen Ableitungen gleich

$$\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \beta_0^2} = \frac{1}{b^2} \sum_{i=1}^{n} \left(r_i \frac{\partial^2 \log f_Z(z_i)}{\partial z_i^2} + (1 - r_i) \frac{\partial^2 \log S_Z(z_i)}{\partial z_i^2} \right), \tag{4.39}$$

$$\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \beta_0 \partial \beta_j} = \frac{1}{b^2} \sum_{i=1}^{n} \left(r_i \frac{\partial^2 \log f_Z(z_i)}{\partial z_i^2} + (1 - r_i) \frac{\partial^2 \log S_Z(z_i)}{\partial z_i^2} \right) x_{ij}, \tag{4.40}$$

$$\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \beta_j \partial \beta_k} = \frac{1}{b^2} \sum_{i=1}^{n} \left(r_i \frac{\partial^2 \log f_Z(z_i)}{\partial z_i^2} + (1 - r_i) \frac{\partial^2 \log S_Z(z_i)}{\partial z_i^2} \right) x_{ij} x_{ik}, \tag{4.41}$$

$$\begin{aligned} \frac{\partial^2 l(\boldsymbol{\theta})}{\partial \beta_0 \partial b} = & -\frac{2}{b^2} \sum_{i=1}^{n} \left(r_i \frac{\partial \log f_Z(z_i)}{\partial z_i} + (1 - r_i) \frac{\partial \log S_Z(z_i)}{\partial z_i} \right) z_i \\ & - \frac{1}{b^2} \sum_{i=1}^{n} \left(r_i \frac{\partial^2 \log f_Z(z_i)}{\partial z_i^2} + (1 - r_i) \frac{\partial^2 \log S_Z(z_i)}{\partial z_i^2} \right), \end{aligned} \tag{4.42}$$

$$\begin{aligned} \frac{\partial^2 l(\boldsymbol{\theta})}{\partial b^2} = & \frac{r}{b^2} + \frac{2}{b^2} \sum_{i=1}^{n} \left(r_i \frac{\partial \log f_Z(z_i)}{\partial z_i} + (1 - r_i) \frac{\partial \log S_Z(z_i)}{\partial z_i} \right) z_i \\ & + \frac{1}{b^2} \sum_{i=1}^{n} \left(r_i \frac{\partial^2 \log f_Z(z_i)}{\partial z_i^2} + (1 - r_i) \frac{\partial^2 \log S_Z(z_i)}{\partial z_i^2} \right) z_i^2, \end{aligned} \tag{4.43}$$

$$\begin{aligned} \frac{\partial^2 l(\boldsymbol{\theta})}{\partial \beta_j \partial b} = & \frac{1}{b^2} \sum_{i=1}^{n} \left(r_i \frac{\partial \log f_Z(z_i)}{\partial z_i} + (1 - r_i) \frac{\partial \log S_Z(z_i)}{\partial z_i} \right) x_{ij} \\ & + \frac{1}{b^2} \sum_{i=1}^{n} \left(r_i \frac{\partial^2 \log f_Z(z_i)}{\partial z_i^2} + (1 - r_i) \frac{\partial^2 \log S_Z(z_i)}{\partial z_i^2} \right) z_i x_{ij} \end{aligned} \tag{4.44}$$

(vgl. Glomb, 2007; Lawless, 2003).

Die $(p+2) \times (p+2)$ Informationsmatrix lautet schließlich

$$\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}) = -\mathrm{E} \begin{pmatrix} \frac{\partial^2 l}{\partial \beta_0^2} & \frac{\partial^2 l}{\partial \beta_0 \partial \beta_1} & \cdots & \frac{\partial^2 l}{\partial \beta_0 \partial \beta_p} & \frac{\partial^2 l}{\partial \beta_0 \partial b} \\ \frac{\partial^2 l}{\partial \beta_1 \partial \beta_0} & \frac{\partial^2 l}{\partial \beta_1^2} & \cdots & \frac{\partial^2 l}{\partial \beta_1 \partial \beta_p} & \frac{\partial^2 l}{\partial \beta_1 \partial b} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \cdots & \cdots & \cdots & \frac{\partial^2 l}{\partial \beta_p^2} & \cdots \\ \frac{\partial^2 l}{\partial b \partial \beta_0} & \cdots & \cdots & \cdots & \frac{\partial^2 l}{\partial b^2} \end{pmatrix} . \tag{4.45}$$

Für große Stichproben kann die gemeinsame Verteilung von $\hat{\boldsymbol{\theta}}$ durch eine $(p+2)$-dimensionale Normalverteilung approximiert werden, d.h.

$$\sqrt{n}\big(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\big) \xrightarrow{D} N_{p+2}\big(\mathbf{0}, \boldsymbol{\mathcal{I}}^{-1}(\hat{\boldsymbol{\theta}})\big) \tag{4.46}$$

(siehe Lawless, 2003).

Da $(\hat{\beta}_0, \hat{\boldsymbol{\beta}}^\intercal, \hat{b})^\intercal$ die Lösungen für $\partial l(\boldsymbol{\theta})/\partial \beta_0 = 0$, $\partial l(\boldsymbol{\theta})/\partial \boldsymbol{\beta} = \mathbf{0}$ und $\partial l(\boldsymbol{\theta})/\partial b = 0$ sind, vereinfachen sich die partiellen Ableitungen (4.43),(4.44) und (4.42) in der geschätzten Varianz-Kovarianzmatrix $\boldsymbol{\mathcal{I}}^{-1}(\hat{\boldsymbol{\theta}})$, somit gilt

$$\begin{aligned}
\left.\frac{\partial l(\boldsymbol{\theta})}{\partial \beta_0 \partial b}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} &= \frac{1}{b^2} \sum_{i=1}^{n} \left(r_i \frac{\partial^2 \log f_Z(z_i)}{\partial z_i^2} + (1 - r_i) \frac{\partial^2 \log S_Z(z_i)}{\partial z_i^2} \right) \tilde{z}_i \,, \\
\left.\frac{\partial l(\boldsymbol{\theta})}{\partial \beta_j \partial b}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} &= \frac{1}{b^2} \sum_{i=1}^{n} \left(r_i \frac{\partial^2 \log f_Z(z_i)}{\partial z_i^2} + (1 - r_i) \frac{\partial^2 \log S_Z(z_i)}{\partial z_i^2} \right) \tilde{z}_i x_{ij} \,, \\
\left.\frac{\partial l(\boldsymbol{\theta})}{\partial b^2}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} &= -\frac{r}{b^2} + \frac{1}{b^2} \sum_{i=1}^{n} \left(r_i \frac{\partial^2 \log f_Z(z_i)}{\partial z_i^2} + (1 - r_i) \frac{\partial^2 \log S_Z(z_i)}{\partial z_i^2} \right) \tilde{z}_i{}^2 \,,
\end{aligned}$$

mit $\tilde{z}_i = (\log(y_i) - \hat{\beta}_0 - \hat{\boldsymbol{\beta}})/\hat{b}$ (vgl. Glomb, 2007).

Es wird nun kurz das Testen von Hypothesen in verschachtelten (nested) loglinearen Modellen behandelt. Sei $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)^\intercal$ der Vektor der unbekannten Parameter, wobei $\boldsymbol{\theta}_1$ und $\boldsymbol{\theta}_2$ Vektoren sind mit Dimensionen m und $((p+2) - m)$. Nun soll die Hypothese $H_0 : \boldsymbol{\theta}_2 = \boldsymbol{\theta}_{02}$ gegen $H_1 : \boldsymbol{\theta}_2 \neq \boldsymbol{\theta}_{02}$ getestet werden, wobei $\boldsymbol{\theta}_{02}$ ein fester Vektor der Dimension $((p+2) - m)$ ist.

Durch den Likelihood-Ratio Test kann nun die Hypothese H_0 überprüft werden. Die Likelihood-Ratio Teststatistik lautet

$$\Lambda_{\text{LR}} = 2\big(l(\check{\boldsymbol{\theta}}) - l(\bar{\boldsymbol{\theta}})\big) \tag{4.47}$$

und folgt unter H_0 einer χ^2_m-Verteilung. Hierbei entspricht $\bar{\boldsymbol{\theta}} = (\bar{\boldsymbol{\theta}}_1, \bar{\boldsymbol{\theta}}_2)$ dem MLE unter H_0 und $\check{\boldsymbol{\theta}} = (\check{\boldsymbol{\theta}}_1, \check{\boldsymbol{\theta}}_2)$ dem uneingeschränkten MLE (vgl. Medeiros, da Silva-Junior, Valencia und Ferrari, 2014).

4.3. Beispiel Kehlkopfkrebs Studie

Mithilfe eines Beispiels soll nun das AFT-Modell besser veranschaulicht werden. Die Daten für das Beispiel stammen aus Abschnitt 1.8 von Klein und Moeschberger (2003) und sind im Paket `KMsurv` unter dem Namen `larynx` enthalten.

Die Daten gehören zu einer Studie, welche die Überlebenszeit von 90 männlichen Patienten mit diagnositiziertem Kehlkopfkrebs (engl cancer of the larynx) zwischen 1970 und 1978 in einem Spital in der Niederlande untersuchte. Bei der Studie wurde zum einen das Alter (`age`) des Patienten bei der Diagnose und zum anderen das Stadium des Kehlkopfkrebs bei der Diagnose (`stage`) aufgenommen. Hierbei bezeichnet Stadium I den am wenigsten fortgeschrittenen Fall und Stadium IV den am meist fortgeschrittenen Fall (siehe Abschnitt 1.8 in Klein und Moeschberger, 2003).

Das log-lineare Modell für die rechtszensierten Kehlkopfdaten mit den Haupteffekten `age` und `stage` ist gegeben durch

$$W = \log(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + bZ\,,$$

mit $X_j\,, j = 1, 2, 3$, den Indikatorvariablen für Stadium II, III und IV und X_4 dem Alter. Weiters wird angenommen, dass die Zufallsvariable Z einer Extremwert-Verteilung folgt womit die Lebenszeit Y dann Weibullverteilt ist (siehe Abschnitt 4.1.1). Im `R`-Code 4.1 wird nun das Weibull AFT Modell an die Kehlkopfkrebs Daten angepasst.

R-Code 4.1: Weibull AFT Modell

```
> library(survival)
> library(KMsurv)
> data("larynx")
> attach(larynx)
>
> # MODELLANPASSUNG
> larynx.weib <- survreg(Surv(time,delta) ~ as.factor(stage) + age,
+ dist = "weibull")
> summary(larynx.weib)
Call:
survreg(formula = Surv(time, delta) ~ as.factor(stage) + age,
    dist = "weibull")
                     Value Std. Error      z        p
(Intercept)         3.5288     0.9041  3.903 9.50e-05
as.factor(stage)2 -0.1477     0.4076 -0.362 7.17e-01
as.factor(stage)3 -0.5866     0.3199 -1.833 6.68e-02
as.factor(stage)4 -1.5441     0.3633 -4.251 2.13e-05
age               -0.0175     0.0128 -1.367 1.72e-01
Log(scale)        -0.1223     0.1225 -0.999 3.18e-01

Scale= 0.885
Weibull distribution
Loglik(model)= -141.4   Loglik(intercept only)= -151.1
  Chisq= 19.37 on 4 degrees of freedom, p= 0.00066
Number of Newton-Raphson Iterations: 5
n= 90
>
> # LIKELHOOD RATIO TEST nachgerechnet
> larynx.weib2 <- survreg(Surv(time,delta) ~ 1, dist = "weibull")
> larynx.loglik.diff <- larynx.weib$loglik[2] -
+                           larynx.weib2$loglik[2]
> 1 - pchisq(2 * larynx.loglik.diff, 4)
[1] 0.0006636971
```

Die Lebenszeit eines Patienten im Stadium IV ist, wie in R-Code 4.1 ersichtlich, hier nun um den Faktor

$$\exp(\gamma_3) = \exp(-\hat{\beta}_3) = \exp\big(-(-1.5441)\big) \approx 4.68$$

kürzer verglichen mit einem Patienten im Stadium I und dem selben Alter. Somit lautet der accelaration-factor $\exp(-\hat{\beta}_3) \approx 4.68$ aufgrund des Zusammenhangs zwischen dem log-linearen Modell und dem AFT-Modell siehe (4.27).

Laut Abschnitt 4.1.1 kann die Survivalfunktion einer Weibullverteilten Lebensdauer Y mithilfe der Survivalfunktion zu Z, $S_Z(z) = \exp\big(-\exp(z)\big)$, folgendermaßen dargestellt werden

$$S(y|\boldsymbol{X}) = \exp\left(-\exp\left(\frac{\log y - \beta_0 - \beta_1 X_1 - \beta_2 X_2 - \beta_3 X_3 - \beta_4 X_4}{b}\right)\right)$$

(vgl. Glomb, 2007).

Mithilfe der ML-Schätzer aus dem `R`-Code 4.1 gilt für die geschätzte Survivalfunktion $\widehat{S}(y|\boldsymbol{X})$ eines Kehlkopfkrebspatienten

$$\widehat{S}(y|\boldsymbol{X}) = \exp\Bigg(-\exp\Bigg(\frac{\log y - 3.53 + 0.15X_1 + 0.59X_2}{0.885} + \frac{1.54X_3 + 0.02X_4}{0.885}\Bigg)\Bigg). \tag{4.48}$$

Die Hazardfunktion für dieses Beispiel lautet

$$\begin{aligned} h(y|\boldsymbol{X}) &\overset{(2.4)}{=} -\frac{d\log\big(S(y|\boldsymbol{X})\big)}{dy} \\ &\overset{(4.48)}{=} \frac{1}{by}\exp\big(\log y - \beta_0 - \beta_1 X_1 - \beta_2 X_2 - \beta_3 X_3 - \beta_4 X_4\big). \end{aligned} \tag{4.49}$$

Mithilfe des MLE aus dem `R`-Code 4.1 gilt für die geschätzte Hazardfunktion $\widehat{h}(y|\boldsymbol{X})$ eines Kehlkopfkrebspatienten

$$\widehat{h}(y|\boldsymbol{X}) = \frac{1}{0.89y}\exp\big(\log y - 3.53 + 0.15X_1 + 0.59X_2 + 1.54X_3 + 0.02X_4\big). \tag{4.50}$$

In Abbildung 4.6 wird die geschätzte Hazardfunktion (4.50) und die geschätzte Survivalfunktion (4.48) eines 30-jährigen Patienten für die vier verschiedenen Stadien des Kehlkopfkrebs abgebildet.

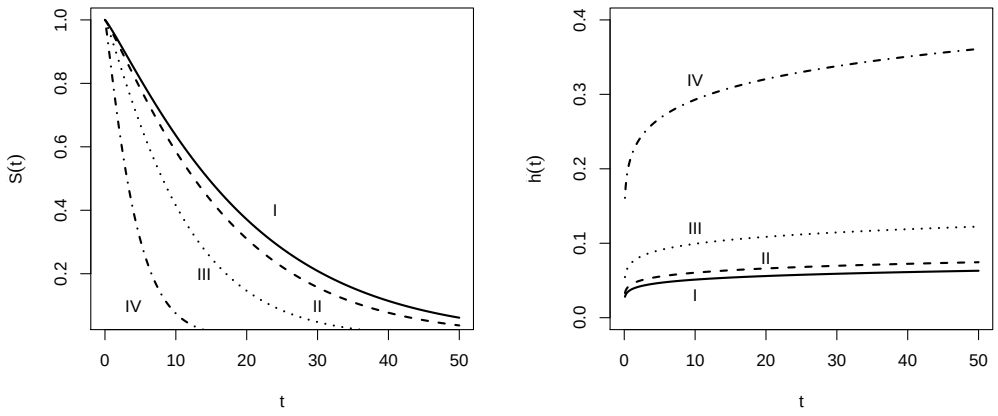

Abbildung 4.6.: Die geschätzte Survivalfunktion (links) und Hazardfunktion (rechts) für einen 30-jährigen Patienten der Kehlkopfkrebs Studie in den vier verschiedenen Stadien, d.h. für Stadium I gilt $\boldsymbol{X} = (0, 0, 0, 30)$, für Stadium II gilt $\boldsymbol{X} = (1, 0, 0, 30)$ usw..

R-Code 4.2: Weibull AFT Modell (Fortsetzung)

```
> # P L O T von S(t)
> surv.func.w.0 <- function(y){exp(-exp((log(y)-3.5288 +
+                                        30*0.0175)/.885))}
> surv.func.w.1 <- function(y){exp(-exp((log(y)-3.5288 + 0.1477 +
+                                        30*0.0175)/.885))}
> surv.func.w.2 <- function(y){exp(-exp((log(y)-3.5288 + 0.5866 +
+                                        30*0.0175)/.885))}
> surv.func.w.3 <- function(y){exp(-exp((log(y)-3.5288 + 1.5441 +
+                                        30*0.0175)/.885))}
> y<- seq(0,50,0.1)
> plot(y, surv.func.w.0(y), type="l", col="black", xlab="t", lwd = 2,
+ ylab=expression(hat(S)(t)))
> lines(y, surv.func.w.1(y), lwd = 2, lty = 2)
> lines(y, surv.func.w.2(y), lwd = 2, lty = 3)
> lines(y, surv.func.w.3(y), lwd = 2, lty = 4)
>
```

```
> # P L O T von h(t)
> haz.func.w.0 <- function(y){exp((log(y)-3.5288 +
+                                  30*0.0175)/.885)/(.885*y)}
> haz.func.w.1<- function(y){exp((log(y)-3.5288 + 0.1477 +
+                                  30*0.0175)/.885)/(.885*y)}
> haz.func.w.2 <- function(y){exp((log(y)-3.5288 + 0.5866 +
+                                 30*0.0175)/.885)/(.885*y)}
> haz.func.w.3 <- function(y){exp((log(y)-3.5288 + 1.5441 +
+                                 30*0.0175)/.885)/(.885*y)}
>
> plot(y, haz.func.w.0(y), type="l", col="black", xlab="t", lwd = 2,
+       ylab=expression(hat(h)(t)),ylim = c(0,.4))
> lines(y, haz.func.w.1(y), lwd = 2, lty = 2)
> lines(y, haz.func.w.2(y), lwd = 2, lty = 3)
> lines(y, haz.func.w.3(y), lwd = 2, lty = 4)
```

4.4. Diagnostik

In diesem Abschnitt werden graphische Methoden besprochen, um die Anpassungsgüte und Verteilungsannahme eines AFT-Modells überprüfen bzw. beurteilen zu können. Die Ausführungen in diesem Abschnitt folgen in großen Teilen Collett (2003), Klein und Moeschberger (2003) und Glomb (2007).

4.4.1. Überprüfung der Verteilungsannahme

Bevor die Anpassungsgüte eines AFT-Modells untersucht wird, sollte überprüft werden, ob die Verteilungsannahme für dieses Modell überhaupt korrekt ist. Bei den hier vorgestellten Methoden wird versucht eine Funktion der Survivalfunktion $S(t)$ zu finden, welche linear in $\log(t)$ ist.

Im Abschnitt 4.1 wurden für die Weibull-, Log-Logistik- und Log-Normal-Verteilung solche Methoden hergeleitet anhand derer die Verteilungsannahme überprüft werden kann. In der Tabelle 4.2 sind diese Ergebnisse nochmals zusammengefasst. Falls ein Plot aus der Tabelle einer Geraden gleicht, spricht nichts gegen die jeweilige Verteilungsannahme.

Modell	**Plot** $\log(t) \Longleftrightarrow H(t)$	
Weibull	$H(t) = \log\big(-\log(S(t)\big)$	(4.13)
Log-Normal	$\Phi^{-1}\big(1 - S(t)\big)$	(4.18)
Log-Logistik	$-\log\left(\frac{S(t)}{1-S(t)}\right)$	(4.22)

Tabelle 4.2.: Graphische Methode, um die Verteilungsannahme zu überprüfen.

Beispiel 7.
Für die Kehlkopfkrebs Studie aus Abschnitt 4.3 wird die Weibull Verteilungsannahme überprüft. Anhand der Abbildung 4.7 ist zu erkennen, dass die Weibull Verteilungsannahme plausibel ist.

R-Code 4.3: Überprüfung der Weibull Verteilungsannahme

```
> larynx.km <- survfit((Surv(time,delta)) ~ 1)
>
> # P L O T   log(t) gegen H(t)
> plot(log(larynx.km$time), log(-log(larynx.km$surv)), pch=19,
+     col = "grey75", xlab= "ln t", ylab="ln(-ln(S(t)))")
> weib.lm <-lm( log(-log(larynx.km$surv)) ~ log(larynx.km$time))
> abline(weib.lm, lwd = 2, col = "black")
```

4.4.2. Cox Snell Residuen

Für das log-lineare Survivalmodell (AFT-Modell in log-linearer Darstellung) mit rechtszensierten Lebenszeiten

$$W_i = \log(Y_i) = \beta_0 + \boldsymbol{\beta}^\intercal \boldsymbol{X}_i + bZ$$

(siehe (4.28)), soll nun die Modellanpassungsgüte mithilfe von Cox Snell Residuen überprüft werden. Die Cox Snell Residuen sind die gängigsten Größen um die Modellanpassung für AFT-Modelle zu überprüfen. Für ein AFT-Modell sind die **Cox Snell Residuen** definiert als

$$r_{Ci} := \widehat{H}(y_i|\boldsymbol{X}_i) = -\log\big(\widehat{S}(y_i|\boldsymbol{X}_i)\big)\,, \quad i = 1, \ldots, n \tag{4.51}$$

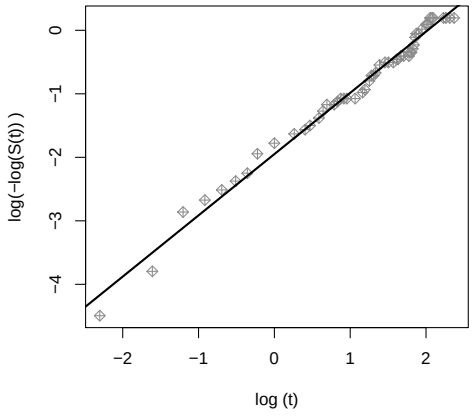

Abbildung 4.7.: Überprüfung der Weibull-Verteilungsannahme für die Kehlkopfkrebsstudie.

mit $\widehat{H}(y_i|\boldsymbol{X}_i)$ der geschätzten kumulativen Hazardfunktion eines AFT-Modells und $\widehat{S}(y_i|\boldsymbol{X}_i)$ der geschätzten Survivalfunktion.

Da $F(T) \sim U(0,1)$ gilt, folgt

$$S(T) = 1 - F(T) \sim U(0,1)\,.$$

Für die Verteilung von $\log S(T)$ gilt deshalb

$$\begin{aligned}\Pr\left(-\log S(T) \leq x\right) &= \Pr\left(S(T) \geq \exp(-x)\right) \\ &= 1 - \exp(-x) \sim Exp(1)\,,\end{aligned}$$

womit $\log S(T)$ standard-exponentialverteilt ist.

Falls ein AFT-Modell ausreichend gut an die Daten angepasst wurde, dann liefert ein Schätzer der Survivalfunktion $\widehat{S}(y)$, basierend auf dem angepassten AFT-Modell, fast die selben Werte wie die wahre Survivalfunktion $S(y)$, d.h. für ein ausreichend gut angepasstes AFT-Modell gilt, $S(y_i) \approx \widehat{S}(y_i)$, $i = 1, \ldots, n$.

Somit kann angenommen werden, dass für ein ausreichend gut angepasstes AFT-Modell der Schätzer der Survivalfuntion $\widehat{S}(y_i)$, $i = 1, \ldots, n$, ähnliche

Eigenschaften wie die wahre Survivalfunktion $S(y_i)$ aufweist. Demzufolge können nun $-\log\big(\widehat{S}(y_i)\big)$ als n Beobachtungen einer Exponentialverteilung mit Erwartungswert 1 angesehen werden (Exp(1)) und somit auch die Cox Snell Residuen, weil $r_{Ci} = -\log\big(\widehat{S}(y_i)\big)$ gilt.

Falls die beobachtete Überlebenszeit y eines Individuums rechtszensiert ist, dann ist das dazugehörige Cox Snell Residuum ebenso rechtszensiert. Dadurch sind die Cox Snell Residuen eine zensierte Stichprobe aus einer Exp(1)-verteilten Population und ein Test dieser Verteilungsannahme liefert einen Test der Modellanpassung des AFT-Modells.

Die Expontial-Verteilungsannahme der Cox Snell Residuen kann mithilfe von (4.9) überprüft werden. Daher wird $r_{Ci} = -\log\big(\widehat{S}(y_i)\big)$ gegen $\widehat{H}(r_{Ci}) = -\log\big(\widehat{S}(r_{Ci})\big)\,, i = 1, \ldots, n,$ aufgetragen. Gleicht dieser Plot einer Geraden mit Steigung 1 dann scheinen die Cox Snell Residuen epxonentialverteilt und somit ist das AFT-Modell ausreichend gut an die Daten angepasst.

Beispiel 8.
Nachdem die Weibull-Verteilungsannahme für die Kehlkopfkrebs Studie aus Abschnitt 4.3 in Beispiel 7 überprüft wurde, wird nun die Modellanpassung des Weibull AFT Modells überprüft. Das in Abschnitt 4.3 aufgestellte Modell lautet

$$W = \log(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + bZ\,,$$

mit $X_j\,, j = 1, 2, 3$ den Indikatorvariablen für Stadium II, III und IV und X_4 dem Alter. Die Cox Snell Residuen für dieses Modell sind dann

$$\begin{aligned} r_{Ci} &= -\log(\widehat{S}\big(y|\boldsymbol{X}_i)\big) \\ &\overset{(4.48)}{=} \exp\left(\frac{\log y - \widehat{\beta}_0 + \widehat{\beta}_1 X_{1i} + \widehat{\beta}_2 X_{2i} + \widehat{\beta}_3 X_{3i} + \widehat{\beta}_4 X_{4i}}{\widehat{b}}\right) \quad i = 1, \ldots, 90. \end{aligned}$$

R-Code 4.4: Überprüfung der Modellanpassung des Weibull AFT Modells

```
> # Berechnen der Cox Snell Residuen
> larynx.weib <- survreg(Surv(time,delta) ~ as.factor(stage) + age,
+          dist = "weibull")
> hat.b <- larynx.weib$scale
```

```
> reg.linear <- larynx.weib$linear.predictor
> cs.resid <- exp( (log(time) - reg.linear)/hat.b)
> cs.fit <- survfit(Surv(cs.resid, delta)~1)
>
> # Plot der Cox Snell Residuen
> plot(cs.fit$time, -log(cs.fit$surv), col = "grey75", pch = 19,
+ ylim =c(0,2.5), xlab = expression(r[Ci]),
+ ylab = expression(paste("H ", (r[Ci]), sep = " ")))
>
> abline(lm(-log(cs.fit$surv) ~ cs.fit$time),lwd = 2)
```

In der Abbildung 4.8 ist zu erkennen, dass die meisten Cox Snell Residuen sehr nahe einer Geraden mit Steigung 1 liegen. Somit kann von einer ausreichend guten Modellanpassung des Weibull AFT Modells für die Kehlkopfkrebs Studie ausgegangen werden.

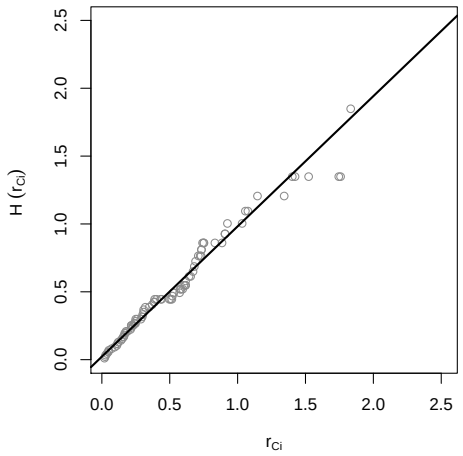

Abbildung 4.8.: Cox Snell Residuen zum Weibull-AFT-Modell.

5. Proportional Hazard Modell

In parametrischen Modellen siehe Abschnitt 4 und Abschnitt 4.2 wird eine Verteilungsannahme für die Lebenszeiten getroffen. Oft ist es jedoch sehr schwierig wenn nicht sogar unmöglich solch eine korrekte Verteilung für die Lebenszeiten zu finden. Vielfach ist auch die Form der Survivalfunktionskurve bzw. Hazardfunktionskurve gar nicht von Interesse sondern lediglich der Effekt der Kovariablen auf das Überlebensrisiko.

Als Resultat dieser Überlegungen wurde das Proportional Hazard Modell (Cox Modell bzw. PH Modell) von Cox (1972) eingeführt. Der Grundgedanke des Proportional Hazard Modells ist, dass jedes Individuum einer Population einem unbekannten nicht spezifizierten Grundrisiko, mit Baseline Hazard $h_0(t)$ bezeichnet, ausgesetzt ist und sich die Kovariablen multiplikativ auf dieses Grundrisiko auswirken. Für das Proportional Hazard Modell sind keinerlei Verteilungsannahmen notwendig und die Hazardfunktion kann sämtliche Formen annehmen (auch die Form einer Treppenfunktion ist möglich). Die Eigenschaften des Proportional Hazard Modells haben dazu geführt, dass es das am weitest verbreitete Survival Modell ist.

5.1. Definition und Eigenschaften

In dem von Cox (1972) formulierten **Proportional Hazard (PH) Modell** wird die Hazardfunktion eines Individuums mit Kovariablenvektor $\boldsymbol{X} = (X_1, \ldots, X_p)^\intercal$ zu jedem Zeitpunkt $t \geq 0$ modelliert als

$$h(t|\boldsymbol{X}) = h_0(t) \exp\left(\boldsymbol{\beta}^\intercal \boldsymbol{X}\right) , \tag{5.1}$$

mit dem Vektor der unbekannten Parameter $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\intercal$ und der unbekannten nur von t abhängigen Baseline Hazardfunktion $h_0(t)$. Die Baseline

Hazardfunktion $h_0(t)$ stellt die Hazardfunktion für $\boldsymbol{X} = \mathbf{0}$ dar. Es gilt zu beachten, dass der lineare Prädiktor $\boldsymbol{\beta}^\intercal \boldsymbol{X} = \beta_1 X_1 + \cdots + \beta_p X_p$ keinen Intercept β_0 enthält, da dieser bereits in der Baseline Hazardfunktion $h_0(t)$ enthalten ist.

Bemerkung 8.
Da für das Proportional Hazard Modell die Baseline Hazardfunktion $h_0(t)$ nicht spezifiziert ist, wird das Modell auch oft als semi-parametrisches Survivalmodell bezeichnet.

Man betrachte zwei Individuen mit Kovariablenvektor $\boldsymbol{X}$ und $\boldsymbol{X}^*$, dann ist das Verhältnis deren Hazardfunktionen (engl. hazard ratio) in einem PH Modell gleich

$$\frac{h(t|\boldsymbol{X})}{h(t|\boldsymbol{X}^*)} = \frac{h_0(t)\exp\left(\boldsymbol{\beta}^\intercal \boldsymbol{X}\right)}{h_0(t)\exp\left(\boldsymbol{\beta}^\intercal \boldsymbol{X}^*\right)} = \exp\big(\boldsymbol{\beta}^\intercal(\boldsymbol{X} - \boldsymbol{X}^*)\big) \quad \forall t \geq 0 \tag{5.2}$$

(vgl. Kleinbaum und Klein, 2005). Somit sind die Hazardfunktionen $h(t|\boldsymbol{X})$ und $h(t|\boldsymbol{X}^*)$ in einem Proportional Hazard Modell zu jedem Zeitpunkt t proportional zueinander (daher der Name Proportional Hazard Modell). Das Verhältnis der Hazardfunktion (5.2) ist das relative Risiko eines Individuums mit Kovariablenvektor $\boldsymbol{X}$ auszufallen im Vergleich zu einem Individuum mit Kovariablenvektor $\boldsymbol{X}^*$. Aus diesem Grund werden die Kovariablen X_i, $i = 1, \ldots, p$ auch oft als *Risikofaktoren* bezeichnet.

Für stetige Lebensdauern gilt laut (2.6) für die Survivalfunktion $S(t) = \exp(-H(t))$. Mithilfe dieser Beziehung kann nun die Survivalfunktion im PH Modell hergeleitet werden

$$\begin{aligned} S(t|\boldsymbol{X}) &= \exp\big(-H(t|\boldsymbol{X})\big) = \exp\left(-\int_0^t h_0(u)\exp(\boldsymbol{\beta}^\intercal \boldsymbol{X})du\right) \\ &= \left[\exp\left(-\int_0^t h_0(u)du\right)\right]^{\exp(\boldsymbol{\beta}^\intercal \boldsymbol{X})} \\ &= S_0(t)^{\exp(\boldsymbol{\beta}^\intercal \boldsymbol{X})}, \end{aligned} \tag{5.3}$$

mit der Baseline Survivalfunktion $S_0(t) := \exp\left(-\int_0^t h_0(u)du\right)$.

5.2. Inferenz für bindungsfreie Lebenszeiten

Betrachtet man das PH Modell zeigt sich, dass der Parametervektor $\boldsymbol{\beta}$ den Effekt von $\boldsymbol{X}$ spezifiziert. Somit sollte man sich bei der Inferenz auf $\boldsymbol{\beta}$ konzentrieren und die Baseline Hazardfunktion $h_0(t)$ als nuisance Parameter (Störparameter) betrachten.

Da die Baseline Hazardfunktion $h_0(t)$ im PH Modell völlig unspezifiziert bleibt, kann ein gewöhnlicher Maximum Likelihood Schätzer (MLE) nicht zur Schätzung von $\boldsymbol{\beta}$ verwendet werden und eine andere Methode wird benötigt. In Cox (1972, 1975) wurde mit dem partiellen Likelihood (siehe Definition A.1.3) eine Methode vorgestellt, welche den nuisance Paramater $h_0(t)$ zur Gänze aus den Gleichungen zur Schätzung der Parameter $\boldsymbol{\beta}$ entfernt.

Die Herleitung der partiellen Likelihood in diesem Abschnitt folgt in großen Teilen Höhle (WS 2008/2009) und Liu (2012). Für n rechtszensierte, stetige und unabhängige Lebenszeiten $T_1, \ldots, T_n$ mit rechtszensierten Lebenszeiten $Y_1, \ldots, Y_n$ und dazugehöriger Rechtszensur-Indikatorvariable R_i ist die Likelihoodfunktion laut (2.17) gleich

$$L = \prod_{i=1}^{n} f_{T_i}(y_i)^{r_i} S_{T_i}(y_i)^{1-r_i} \, .$$

Wegen $h(t) = f(t)/S(t)$ laut (2.4) kann die gemeinsame Dichtefunktion umgeschrieben werden zu

$$L = \prod_{i=1}^{n} h_{T_i}(y_i)^{r_i} S_{T_i}(y_i)$$

und wegen $S(t) = \exp(-H(t))$ (siehe (2.6)) kann die Likelihoodfunktion nochmals umgeformt werden zu

$$L = \prod_{i=1}^{n} h_{T_i}(y_i)^{r_i} \exp\left(- \int_0^{y_i} h_{T_i}(u) du \right) . \tag{5.4}$$

Nun wird die Darstellung der Likelihoodfunktion in (5.4) dazu verwendet eine Likelihoodfunktion für das PH Modell herzuleiten. Diese Likelihoodfunktion

wird dann so lange umgeformt bis man eine partielle Likelihoodfunktion für $\boldsymbol{\beta}$ erhält.

Für das PH Modell gilt $h_{T_i}(y_i) = h(y_i|\boldsymbol{x}_i) = h_0(y_i)\exp(\boldsymbol{\beta}^\intercal \boldsymbol{x}_i)$. Eingesetzt in (5.4) ergibt das die Likelihoodfunktion

$$\begin{aligned} L(\boldsymbol{\beta}, h_0(t)) &= \prod_{i=1}^{n} \left[\left(h_0(y_i)\exp(\boldsymbol{\beta}^\intercal \boldsymbol{x}_i)\right)^{r_i} \exp\left(-\int_0^{y_i} h_0(u)\exp(\boldsymbol{\beta}^\intercal \boldsymbol{x}_i)du\right)\right] \\ &\overset{(5.3)}{=} \prod_{i=1}^{n} \left[\left(h_0(y_i)\exp(\boldsymbol{\beta}^\intercal \boldsymbol{x}_i)\right)^{r_i} S_0(y_i)^{\exp(\boldsymbol{\beta}^\intercal \boldsymbol{x}_i)}\right] . \end{aligned} \tag{5.5}$$

Um die Likelihoodfunktion für das PH Modell herleiten zu können müssen die Zensurzeitpunkte sortiert werden $y_{(1)} < y_{(2)} < \cdots < y_{(n)}$ und es wird zusätzlich angenommen, dass alle Zensurzeitpunkte voneinander verschieden (bindungsfrei) sind. Somit lautet die Likelihoodfunktion (5.5) für bindungsfreie Lebenszeiten

$$L(\boldsymbol{\beta}, h_0(t)) = \prod_{i=1}^{n} \left[\left(h_0(y_{(i)})\exp(\boldsymbol{\beta}^\intercal \boldsymbol{x}_{(i)})\right)^{r_i} S_0(y_{(i)})^{\exp(\boldsymbol{\beta}^\intercal \boldsymbol{x}_{(i)})}\right] , \tag{5.6}$$

wobei $\boldsymbol{x}_{(i)}$ den Kovariablenvektores des i-ten Individuums mit Ausfallszeitpunkt $y_{(i)}$ bezeichnet.

Weiters wird $\mathcal{R}(t) := \{i : y_i \geq t\}$ eingeführt, die Menge der Individuen welche unmittelbar vor dem Zeitpunkt t noch unter Beobachtung stehen. Wird nun (5.6) mit $\left(\sum_{j\in\mathcal{R}(y_{(i)})} h_0(y_{(i)})\exp(\boldsymbol{\beta}^\intercal \boldsymbol{x}_j)\right)^{r_i}$ erweitert, so erhält man folgenden

Ausdruck für die Likelhoodfunktion des PH Modells

$$\begin{aligned} L(\boldsymbol{\beta}, h_0(t)) =& \prod_{i=1}^{n} \left(\frac{h_0(y_{(i)}) \exp(\boldsymbol{\beta}^\intercal \boldsymbol{x}_{(i)})}{\sum_{j \in \mathcal{R}(y_{(i)})} h_0(y_{(i)}) \exp(\boldsymbol{\beta}^\intercal \boldsymbol{x}_j)} \right)^{r_i} \\ & \cdot \prod_{i=1}^{n} \left(\sum_{j \in \mathcal{R}(y_{(i)})} h_0(y_{(i)}) \exp(\boldsymbol{\beta}^\intercal \boldsymbol{x}_j) \right)^{r_i} \cdot \prod_{i=1}^{n} S_0(y_{(i)})^{\exp(\boldsymbol{\beta}^\intercal \boldsymbol{x}_{(i)})} \\ =& \prod_{i=1}^{n} \left(\frac{\exp(\boldsymbol{\beta}^\intercal \boldsymbol{x}_{(i)})}{\sum_{j \in \mathcal{R}(y_{(i)})} \exp(\boldsymbol{\beta}^\intercal \boldsymbol{x}_j)} \right)^{r_i} \\ & \cdot \prod_{i=1}^{n} \left(\sum_{j \in \mathcal{R}(y_{(i)})} h_0(y_{(i)}) \exp(\boldsymbol{\beta}^\intercal \boldsymbol{x}_j) \right)^{r_i} \cdot \prod_{i=1}^{n} S_0(y_{(i)})^{\exp(\boldsymbol{\beta}^\intercal \boldsymbol{x}_{(i)})} . \end{aligned} \tag{5.7}$$

Bemerkung 9.
Der Ausdruck $\left(\sum_{j \in \mathcal{R}(y_{(i)})} h_0(y_{(i)}) \exp(\boldsymbol{\beta}^\intercal \boldsymbol{x}_j) \right)^{r_i}$ *in* (5.7) *beschreibt die Summe der Hazardfunktionen aller Individuen in* $\mathcal{R}(y_{(i)})$ *zum Zeitpunkt* $y_{(i)}$.

Betrachtet man das erste Produkt der Likelhoodfunktion (5.7) genauer erkennt man, dass es die Baseline Hazardfunktion $h_0(t)$ gar nicht beinhaltet und somit als partielle Likelihoodfunktion (siehe Definition A.1.3) angesehen werden kann mit nuisance Parameter $\phi = h_0(t)$, d.h. die partielle Likelihoodfunktion von $\boldsymbol{\beta}$ lautet

$$L_p(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left(\frac{\exp(\boldsymbol{\beta}^\intercal \boldsymbol{x}_{(i)})}{\sum_{j \in \mathcal{R}(y_{(i)})} \exp(\boldsymbol{\beta}^\intercal \boldsymbol{x}_j)} \right)^{r_i} . \tag{5.8}$$

Da für rechtszensierte Lebenszeiten (d.h. $r_i = 0$) der Faktor in obiger Likelihoodfunktion gleich 1 ist, kann **die partielle Likelihoodfunktion von $\boldsymbol{\beta}$** für rechtszensierte, stetige unabhängige und bindungsfreie Lebenszeiten in einem **Proportional Hazard (PH) Modell** geschrieben werden als

$$L_p(\boldsymbol{\beta}) = \prod_{i=1}^{k} \frac{\exp(\boldsymbol{\beta}^\intercal \boldsymbol{x}_{(i)})}{\sum_{j \in \mathcal{R}(y_{(i)})} \exp(\boldsymbol{\beta}^\intercal \boldsymbol{x}_j)} , \tag{5.9}$$

mit der Anzahl an beobachteten Ausfällen $k := \sum_{i=1}^{n} r_i$ und den dazugehörigen sortierten beobachten (d.h. nicht zensierten) Ausfallszeitpunkten

$y_{(1)} < y_{(2)} < \cdots < y_{(k)}$. Man beachte, dass die rechtszensierten Ausfallszeitpunkte die partielle Likelihoodfunktion $L_p(\boldsymbol{\beta})$ beeinflussen, da zu jedem Ausfallszeitpunkt $y_{(i)}$ die dazugehörige Risikomenge $\mathcal{R}(y_{(i)})$ die rechtszensierten Ausfallszeitpunkte enthält.

Bemerkung 10.
Es existieren noch alternative Ansätze zur Herleitung der partiellen Likelihoodfunktion. Einer dieser Ansätze beruht auf der bedingten Wahrscheinlichkeit, dass ein Individuum im Zeitintervall $\big(y_{(i)}, (y_{(i)} + \Delta)\big)$ *ausfällt, d.h.*

$$\frac{\Pr\left(\textit{Individuum } i \textit{ mit Kovariablenvektor } \boldsymbol{x}_i \textit{ fällt aus in } \big(y_{(i)}, (y_{(i)} + \Delta)\big)\right)}{\sum_{j \in \mathcal{R}(y_{(i)})} \Pr\left(\textit{Individuum } j \textit{ fällt aus in } \big(y_{(i)}, (y_{(i)} + \Delta)\big)\right)}. \tag{5.10}$$

Nun wird (5.10) *erweitert mit* $1/\Delta$*, d.h.*

$$\frac{\Pr\left(\textit{Individuum } i \textit{ mit Kovariablenvektor } \boldsymbol{x}_i \textit{ fällt aus in } \big(y_{(i)}, (y_{(i)} + \Delta)\big)\right)/\Delta}{\sum_{j \in \mathcal{R}(y_{(i)})} \Pr\left(\textit{Individuum } j \textit{ fällt aus in } \big(y_{(i)}, (y_{(i)} + \Delta)\big)\right)/\Delta}. \tag{5.11}$$

Die bedingte Wahrscheinlichkeit wird zur Hazardrate für $\Delta \to 0$ *(vgl.* (2.3)*), d.h. aus* (5.11) *wird für* $\Delta \to 0$

$$\frac{\textit{Hazardfunktion von Individuum } i \textit{ mit Kovariablenvektor } \boldsymbol{x}_i}{\sum_{j \in \mathcal{R}(y_{(i)})} \textit{Hazardfunktion zum Zeitpunkt } y_{(i)} \textit{ für Individuum } j}.$$

Somit kann die bedingte Wahrscheinlichkeit für das i*-te Individuum zum Zeitpunkt* $y_{(i)}$ *im PH Modell geschrieben werden als*

$$\frac{h_0(y_{(i)}) \exp(\boldsymbol{\beta}^\intercal \boldsymbol{x}_{(i)})}{\sum_{j \in \mathcal{R}(y_{(i)})} h_0(y_{(i)}) \exp(\boldsymbol{\beta}^\intercal \boldsymbol{x}_j)} = \frac{\exp(\boldsymbol{\beta}^\intercal \boldsymbol{x}_{(i)})}{\sum_{j \in \mathcal{R}(y_{(i)})} \exp(\boldsymbol{\beta}^\intercal \boldsymbol{x}_j)}$$

und somit ist die partielle Likelihoodfunktion für das PH Modell gleich

$$L_p(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left(\frac{\exp(\boldsymbol{\beta}^\intercal \boldsymbol{x}_{(i)})}{\sum_{j \in \mathcal{R}(y_{(i)})} \exp(\boldsymbol{\beta}^\intercal \boldsymbol{x}_j)} \right)^{r_i} \tag{5.12}$$

(vgl. Liu, 2012).

Logarithmiert man (5.9) so erhält man die log partielle Likelihoodfunktion von $\boldsymbol{\beta}$

$$l_p(\boldsymbol{\beta}) = \sum_{i=1}^{k} \boldsymbol{\beta}^\intercal \boldsymbol{x}_{(i)} - \sum_{i=1}^{k} \log \left(\sum_{j \in \mathcal{R}(y_{(i)})} \exp(\boldsymbol{\beta}^\intercal \boldsymbol{x}_j) \right) . \tag{5.13}$$

Äquivalent zum MLE ist der Maximum partial Likelihood Schätzer (MPLE) definiert als

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} L_p(\boldsymbol{\beta}) = \arg\max_{\boldsymbol{\beta}} l_p(\boldsymbol{\beta}) . \tag{5.14}$$

Somit ist die Score Statistik für die h-te Kovariable ($h = 1, \ldots, p$) gleich

$$U(\beta_h) = \frac{\partial l_p(\boldsymbol{\beta})}{\partial \beta_h} = \sum_{i=1}^{k} \left[x_{(i)h} - \frac{\sum_{j \in \mathcal{R}(y_{(i)})} x_{jh} \exp(\boldsymbol{\beta}^\intercal \boldsymbol{x}_j)}{\sum_{j \in \mathcal{R}(y_{(i)})} \exp(\boldsymbol{\beta}^\intercal \boldsymbol{x}_j)} \right] . \tag{5.15}$$

Der MPLE $\hat{\boldsymbol{\beta}}$ kann durch Lösen des Gleichungssystems

$$\boldsymbol{U}(\boldsymbol{\beta}) = (U(\beta_1), \ldots, U(\beta_p))^\intercal = \mathbf{0}$$

gefunden werden.

In `R` wird das Newton-Raphson Verfahren genutzt um das Gleichungssystem $\boldsymbol{U}(\boldsymbol{\beta}) = \mathbf{0}$ zu lösen. Gestartet mit $\hat{\boldsymbol{\beta}}^{(0)}$ wird das Verfahren iterativ angewandt

$$\hat{\boldsymbol{\beta}}^{(s+1)} = \hat{\boldsymbol{\beta}}^{(s)} + \boldsymbol{\mathcal{I}}^{-1}(\hat{\boldsymbol{\beta}}^{(s)}) \boldsymbol{U}(\hat{\boldsymbol{\beta}}^{(s)}) , \quad s = 0, 1, \ldots$$

bis $\hat{\boldsymbol{\beta}}^{(s+1)}$ konvergiert, mit $\boldsymbol{\mathcal{I}}(\hat{\boldsymbol{\beta}})$ der beobachteten Fishermatrix (vgl. Therneau und Grambsch, 2000).

In Andersen und Gill (1982) wurde gezeigt, dass der MPLE $\hat{\boldsymbol{\beta}}$ ein konsistenter Schätzer für $\boldsymbol{\beta}$ ist und multivariat normalverteilt ist, d.h.

$$\hat{\boldsymbol{\beta}} \sim N_p\big(\boldsymbol{\beta}, \boldsymbol{\mathcal{I}}^{-1}(\hat{\boldsymbol{\beta}})\big) . \tag{5.16}$$

Nun wird die Likelihood Ratio Teststatistik vorgestellt, anhand derer die Relevanz der Parameter $\boldsymbol{\beta}$ überprüft werden kann. Zuerst soll die einfache Nullhypothese $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$ gegen $H_1 : \boldsymbol{\beta} \neq \boldsymbol{\beta}_0$ getestet werden. Die Likelihood Ratio Teststatistik beruht für PH Modelle auf der Tatsache, dass der MPLE $\hat{\boldsymbol{\beta}}$ asymptotisch normalverteilt ist (vgl. (5.16)). Für die **Likelihood**

Ratio Teststatistik zur Überprüfung der einfachen Nullhypothese H_0 gilt

$$\Lambda_{\mathrm{LR}} = 2\big(l_p(\hat{\boldsymbol{\beta}}) - l_p(\boldsymbol{\beta}_0)\big) \sim \chi^2_p \tag{5.17}$$

unter H_0 (vgl. Klein und Moeschberger, 2003).

Zumeist ist man auch daran interessiert zu testen, ob eine Teilmenge des Parameters $\boldsymbol{\beta}$ relevant ist. Dafür wird die Nullhypothese $H_0 : \boldsymbol{\beta}_2 = \boldsymbol{\beta}_{02}$ überprüft, mit $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)^\intercal$ und $\boldsymbol{\beta}_1$ dem q-dimensionalen Vektor der freien Parameter sowie $\boldsymbol{\beta}_2$ dem $p-q$-dimensionalen Vektor der festgelegten restlichen Parameter. Für die Likelihood Ratio Teststatistik zur Überprüfung der Nullhypothese H_0 gilt

$$\Lambda_{\mathrm{LR}} = 2\big(l_p(\hat{\boldsymbol{\beta}}) - l_p(\hat{\boldsymbol{\beta}}_1, \boldsymbol{\beta}_{02})\big) \sim \chi^2_{p-q} \tag{5.18}$$

unter H_0, mit $\hat{\boldsymbol{\beta}}_1$ dem MPLE von $\boldsymbol{\beta}_1$ und $\hat{\boldsymbol{\beta}}$ dem MPLE von $\boldsymbol{\beta}$ (vgl. Klein und Moeschberger, 2003).

5.3. Inferenz für gebundene Lebenszeiten

In Abschnitt 5.2 wurde bei der Herleitung der partiellen Likelihoodfunkton von $\boldsymbol{\beta}$ angenommen, dass die n sortierten Ausfallszeitunkte voneinander verschieden bzw. bindungsfrei sind (engl. no ties), d.h. $y_{(1)} < \cdots < y_{(n)}$.

Jedoch ist diese Annahme nicht praxisnah, da in den meisten Studien Lebenszeiten mit dem exakt gleichen Ausfallszeitpunkt auftreten. Einige Statistiker haben daraufhin Methoden bzw. Ansätze entwickelt um für gebundene Lebenszeiten die Parameter im PH Modell schätzen zu können. Die hier vorgestellten Methoden beruhen auf der Herleitung der partiellen Likelhoodfunktion über die bedingten Wahrscheinlichkeiten (siehe Bemerkung 10).

Seien $y_{(1)} < \cdots < y_{(\acute{n})}$ die $\acute{n}$ verschiedenen sortierten Ausallszeitpunkte. Sei d_i die Anzahl an Ausfällen zum Zeitpunkt $y_{(i)}$ und $\mathcal{D}_i$ die Menge aller Individuen welche zum Zeitpunkt $y_{(i)}$ ausfallen (d.h. $|\mathcal{D}_i| = d_i$ für $1 \leq i \leq \acute{n}$). Sei $\boldsymbol{s}_i$ die Summe über alle Kovariablenvektoren $\boldsymbol{X}_j$ der Individuen welche zum Zeitpunkt $y_{(i)}$ ausfallen, d.h. $\boldsymbol{s}_i := \sum_{j \in \mathcal{D}_i} \boldsymbol{X}_j$. Weiters sei $\mathcal{R}(t) := \{i : y_i \geq t\}$ die Menge der Individuen welche unmittelbar vor dem Zeitpunkt t noch unter Beobachtung stehen.

Bemerkung 11.
Da hier angenommen wurde, dass den Daten ein zeitstetiges Modell zugrunde liegt (PH Modell), entstehen Bindungen durch Gruppieren von Ausfallzeitpunkten wegen zu ungenauer Messmethoden. Somit haben d_i Ausfälle zum Zeitpunkt $y_{(i)}$ tatsächlich voneinander verschiedene Ausfallszeitpunkte (Glomb, 2007).

Die erste vorgestellte Methode ist die Breslow Approximation. Diese Approximation hat die einfachste Form der hier vorgestellten Methoden und war lange Zeit die einzig implementierte Methode in allen Statistik Programmen und ist auch nach wie vor die default Methode in fast allen Statistik Programmen (auch in `R`). Die Breslow Approximation für die partielle Likelihood lautet

$$L_p^{\text{Breslow}}(\boldsymbol{\beta}) = \prod_{i=1}^{\acute{n}} \frac{\exp(\boldsymbol{\beta}^\intercal \boldsymbol{s}_i)}{\left(\sum_{j \in \mathcal{R}(y_{(i)})} \exp(\boldsymbol{\beta}^\intercal \boldsymbol{x}_j)\right)^{d_i}} \,. \tag{5.19}$$

In der Breslow Approximation werden gebundene Lebenszeiten als Lebenszeiten mit dem exakt selben Ausfallszeitpunkt angesehen (vgl. Bemerkung 11 und Bemerkung 10). Diese Approximation ist die ungenaueste der hier vorgestellten Methoden jedoch mit der kürzesten Laufzeit (vgl. Klein und Moeschberger, 2003).

Die Efron Approximation für die partielle Likelihood lautet

$$L_p^{\text{Efron}}(\boldsymbol{\beta}) = \prod_{i=1}^{\acute{n}} \frac{\exp(\boldsymbol{\beta}^\intercal \boldsymbol{s}_i)}{\prod_{j=1}^{d_i} \left(\sum_{k \in \mathcal{R}(y_{(k)})} \exp(\boldsymbol{\beta}^\intercal \boldsymbol{x}_k) - \frac{j-1}{d_i} \sum_{k \in \mathcal{D}_i} \exp(\boldsymbol{\beta}^\intercal \boldsymbol{x}_k)\right)} \tag{5.20}$$

(vgl Borucka, 2014).

Bei Stichproben mit wenig gebundenen Lebenszeiten liefern die Breslow Approximation und die Efron Approximation nahezu identische Werte für die partielle Likelihoodfunktion (vgl. Klein und Moeschberger, 2003).

Die exakte partielle Likelihood Methode (engl. exact partial likelhood) berücksichtigt sämtliche Kombinationsmöglichkeiten bzw. Permutationen der Reihenfolge der d_i Ausfälle zum Ausfallszeitpunkt $y_{(i)}$. Somit gibt es zum Zeitpunkt $y_{(i)}$ genau $d_i!$ mögliche Kombinationen der Reihenfolge. Sei nun $\mathcal{Q}_i$ die Menge der $d_i!$ Kombinationsmöglichkeiten der Ausfallreihenfolge, mit $P =$

$(p_1, \ldots, p_{d_i})$ für $P \in \mathcal{Q}_i$. Weiters sei $\mathcal{R}(y_{(i)}, P, k)$ definiert als $\mathcal{R}(y_{(i)}, P, k) := \mathcal{R}(y_{(i)}) \backslash \{p_1, \ldots, p_k\}$ (vgl. Liu, 2012).

Der Beitrag zur partiellen Likelihoodfunktion zum Zeitpunkt $y_{(i)}$ nach der exakten partiellen Methode lautet nun

$$\frac{\exp(\boldsymbol{\beta}^\intercal \boldsymbol{s}_i)}{d_i! \sum_{P \in \mathcal{Q}_i} \prod_{k=1}^{d_i} \left(\sum_{l \in \mathcal{R}(y_{(i)}, P, k)} \exp(\boldsymbol{\beta}^\intercal \boldsymbol{x}_l) \right)} . \tag{5.21}$$

Die partielle Likelihoodfunktion mit der exakten partiellen Likelihood Methode lautet somit

$$L_p^{\text{Exakt}}(\boldsymbol{\beta}) = \prod_{i=1}^{\acute{n}} \frac{\exp(\boldsymbol{\beta}^\intercal \boldsymbol{s}_i)}{d_i! \sum_{P \in \mathcal{Q}_i} \prod_{k=1}^{d_i} \left(\sum_{l \in \mathcal{R}(y_{(i)}, P, k)} \exp(\boldsymbol{\beta}^\intercal \boldsymbol{x}_l) \right)} . \tag{5.22}$$

Diese Methode benötigt aufgrund der Berücksichtigung aller Permutationen mit Abstand am meisten Laufzeit (vgl. Liu, 2012).

Bemerkung 12.
Bei der Formulierung der exakten partiellen Likelihood Methode wird meistens der Faktor $\prod_{i=1}^{\acute{n}} \frac{1}{d_i!}$ *weggelassen.*

5.4. Beispiel Kehlkopfkrebs Studie

Mithilfe eines Beispiels soll nun das Proportional Hazard Modell besser veranschaulicht werden. Die Daten für das Beispiel stammen aus der Kehlkopfkrebs Studie, welche auch in Abschnitt 4.3 verwendet und auch dort genauer erklärt wurden.

Das postulierte Proportional Hazard Modell

$$h(t|\boldsymbol{X}) = h_0(t) \exp\big(\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4\big) ,$$

mit X_i, $i = 1, 2, 3$, den Indikatorvariablen für Stadium II, III und IV und X_4 dem Alter. In R-Code 5.1 wurde dieses Modell an die Kehlkopfkrebs Studiendaten angepasst.

R-Code 5.1: Proportional Hazard Modell Beispiel

```
> library(survival)
> library(KMsurv)
> data("larynx")
> attach(larynx)
>
> # MODELL ANPASSEN
> larynx.coxph <- coxph(Surv(time,delta) ~ as.factor(stage) + age,
+                           method = "breslow")
> summary(larynx.coxph)
  n= 90, number of events= 50

                     coef exp(coef) se(coef)     z Pr(>|z|)
as.factor(stage)2 0.14004   1.15032  0.46249 0.303   0.7620
as.factor(stage)3 0.64238   1.90100  0.35611 1.804   0.0712 .
as.factor(stage)4 1.70598   5.50678  0.42191 4.043 5.27e-05 ***
age               0.01903   1.01921  0.01426 1.335   0.1820
---
                  exp(coef) exp(-coef) lower .95 upper .95
as.factor(stage)2     1.150     0.8693    0.4647     2.848
as.factor(stage)3     1.901     0.5260    0.9459     3.820
as.factor(stage)4     5.507     0.1816    2.4086    12.590
age                   1.019     0.9811    0.9911     1.048

Likelihood ratio test= 18.31  on 4 df,   p=0.001072
Wald test            = 21.15  on 4 df,   p=0.0002958
Score (logrank) test = 24.78  on 4 df,   p=5.573e-05
>
> # LIKELHOOD RATIO TEST fuer H_0 = Nullmodell nachrechnen
> larynx.coxph2 <- update(larynx.coxph, formula = . ~ -1)
> larynx.loglik.diff <- larynx.coxph$loglik[2] - larynx.coxph2$loglik
> 1 - pchisq(2*larynx.loglik.diff, df = length(coef(larynx.coxph)))
[1] 0.001072204
```

Mithilfe des Verhältnisses der Hazardfunktionen kann der geschätzte Effekt der Kovariablen untersucht werden. Aus (5.2) ist bekannt, dass für dieses

Verhältnis

$$\frac{h(t\,|\boldsymbol{X})}{h(t\,|\boldsymbol{X}^*)} = \exp\big(\boldsymbol{\beta}^\intercal(\boldsymbol{X}-\boldsymbol{X}^*)\big) \quad \forall t \geq 0$$

gilt. Dieses Verhältnis der Hazardfunktionen kann im R-Code 5.1 in der Spalte `exp(coef)` von `summary(larynx.coxph)` abgelesen werden. Für Patient A im Stadium I und Alter 30 und Patient B im Stadium IV und Alter 30 gilt für das Verhältnis der Hazardfunktionen

$$\frac{h(t\,|\mathrm{B})}{h(t\,|\mathrm{A})} = \frac{h(t\,|\text{ Stadium IV, Alter=30})}{h(t\,|\text{ Stadium I, Alter=30})} = \exp(\beta_3) = 5.507\,.$$

Somit hat der Patient B zu jedem Zeitpunkt t das 5.5-fach höhere Risiko zu versterben als der Patient A.

Umgekehrt sieht man, dass das Verhältnis der Hazardfunktionen zwischen einem Patienten C im Stadium II und Alter 30 und Patient A lediglich $\frac{h(t\,|\mathrm{C})}{h(t\,|\mathrm{A})} = \exp(\beta_2) = 1.067$ beträgt. Somit hat Patient C zu jedem Zeitpunkt t ein um bloß 6.7 Prozent höheres Risiko zu versterben als Patient A.

Man sollte das Verhältnis der Hazardfunktionen auch immer hinterfragen. So kann ein Verhältnis der Hazardfunktionen mit dem Wert 5.5 zum Beispiel von 5500/1000 oder auch 0.0055/0.001 stammen (vgl. Liu, 2012). Aus diesem Grund ist es ratsam auch die geschätzte Survivalfunktion für ein PH Modell zu betrachten. In R kann die Survivalfunktion des PH Modells mithilfe der `survfit` Funktion geschätzt werden (siehe R-Code 5.2).

R-Code 5.2: Proportional Hazard Modell Beispiel (Fortsetzung)

```
> larynx$stadium.factor <- factor(larynx$stage,
+        labels = c("StageI","StageII","StageIII","StageIV"))
> attach(larynx)
> larynx.coxph <- coxph(Surv(time,delta) ~ stadium.factor,
+                          data = larynx)
>
> # SCHAETZUNG VON S(t) FUER DAS PH MODELL
> plot(survfit(larynx.coxph,
+       newdata = data.frame(stadium.factor = "StageI", age=30),
+       conf.int = F),lwd=3, xlab="Jahr t",
+       ylab = expression(hat(S)(t)), col = "darkgreen")
>
```

```
> lines(survfit(larynx.coxph,
+       newdata = data.frame(stadium.factor = "StageII", age=30),
+       conf.int = F), lwd = 3, col = "gold2")
> lines(survfit(larynx.coxph,
+       newdata = data.frame(stadium.factor = "StageIII", age=30),
+       conf.int = F), lwd = 3, col = "darkorange2")
> lines(survfit(larynx.coxph,
+     newdata = data.frame(stadium.factor = "StageIV", age=30),
+     conf.int = F),  lwd = 3, col = "firebrick2")
```

Anhand der Abbildung 5.1 soll noch einmal der Effekt des Stadiums auf das Ausfallrisiko veranschaulicht werden. In der Abbildung 5.1 sind nun die Survivalkurven für die verschiedenen Stadien eines 30-jährigen Patienten der Kehlkopfkrebs Studie aufgetragen.

In der Abbildung 5.1 erkennt man, dass die Survivalkurven von Stadium I und Stadium II nahezu identisch sind. Ein erster Hinweis auf diesen Umstand lieferte bereits der p-Wert 0.7620 von `as.factor(stage)2` in R-Code 5.2.

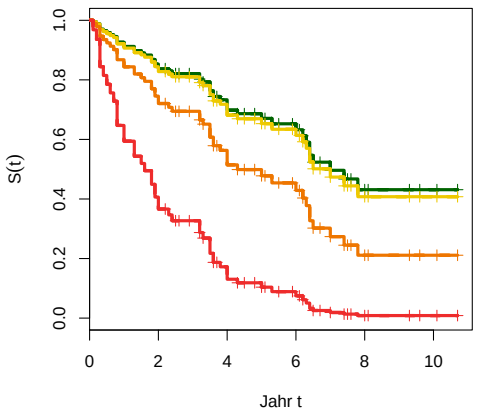

Abbildung 5.1.: Geschätzte Survivalkurven unter dem Modell für 30-jährige Kehlkopfkrebs Patienten im Stadium I (grün), Stadium II (gelb), Stadium II (orange) und Stadium IV (rot).

5.5. Erweiterungen des PH Modells

In diesem Abschnitt werden drei Erweiterungen für das PH Modell vorgestellt. Zuerst wird ausgeführt wie ein PH Modell mit Splines aussieht. Danach wird das Stratifizierte PH Modell diskutiert, welches für jedes Stratum eine eigene Baseline Hazardfunktion schätzt. Abschließend wird das PH Modell für zeitabhängige Kovariablen diskutiert.

5.5.1. PH Modell mit Splines

Die Ausführungen in diesem Abschnitt folgen in großen Teilen Sleeper und Harrington (1990). Angenommen die Beziehung zwischen der stetigen Kovariable Z und der Hazardfunktion $h_0(t)$ ist nicht linear. Falls eine Funktion g existiert welche diese Beziehung beschreibt, dann kann ein verallgemeinertes PH Modell formuliert werden als

$$h(t|Z) = h_0(t)\exp\big(g(Z)\big) \quad \forall t \geq 0\,, \tag{5.23}$$

mit der unbekannten Baseline Hazardfunktion $h_0(t)$ der Hazardfunktion für $g(Z) = 0$ und der stetigen Kovariable Z.

Da jedoch im allgemeinen die Funktionen g unbekannt ist, muss auf eine Approximationen dieser Funktionen zurückgegriffen werden. Eine solche Approximation ist die durch Basisfunktionen erzeugte glatte Funktion $s(Z)$ möglich. Das PH Modell mit einer glatten Funktion lautet

$$h(t|Z) = h_0(t)\exp\big(s(Z)\big) \quad \forall t \geq 0\,, \tag{5.24}$$

wobei die glatte Funktion $s(Z)$ eine Linearkombination der Basisfunktionen $b_j(.)$ für $j = 1, \ldots, m$ ist d.h.

$$s(Z) = \sum_{j=1}^{m} \beta_j b_j(Z).$$

Das PH Modell kann weiters verallgemeinert werden zu

$$h(t|\boldsymbol{X}, \boldsymbol{Z}) = h_0(t)\exp\big(\boldsymbol{\beta}^\intercal \boldsymbol{X} + s_1(Z_1) + \cdots + s_q(Z_q)\big) \quad \forall t \geq 0\,, \tag{5.25}$$

für Kovariablen $\boldsymbol{X} = \big(X_1, \ldots, X_p\big)^\intercal$ und $\boldsymbol{Z} = \big(Z_1, \ldots, Z_q\big)^\intercal$, glatten Funktionen $s_1(.), \ldots, s_q(.)$ und der unbekannten Baseline Hazardfunktion $h_0(t)$.

Für mehr Informationen zu Splines und PH Modellen wird auf Sleeper und Harrington (1990) und Therneau und Grambsch (2000) verwiesen.

5.5.2. Zeitabhängiges PH Modell

Bisher wurde angenommen, dass sich die Werte der Kovariablen über die ganze Studien- bzw. Beobachtungsdauer nicht verändern. Diese Annahme ist korrekt für solch Kovariablen wie Geschlecht, Ethnizität oder Alter bei der ersten Ehe. Jedoch ist diese Annahme problematisch für solch Kovariablen wie Blutdruck, Ehestatus oder Beschäftigungsverhältnis, welche sich über den Lauf der Zeit immer wieder verändern können (vgl. Liu, 2012). Eine *zeitabhängige Kovariable* ist nun definiert als Kovariable, welche im Lauf der Beobachtungs- bzw. Studienzeit ihren Wert verändern kann.

Das **zeitabhängige PH Modell** ist eine Erweiterung des Proportional Hazard Modells, welches auch zeitabhängige Kovariablen zulässt. Dieses Modell lautet

$$h(t|\boldsymbol{X}) = h_0(t) \exp\big(\boldsymbol{\beta}^\intercal \boldsymbol{X}(t)\big) \quad \forall t \geq 0\,, \tag{5.26}$$

mit dem unbekannten Paramatervektor $\boldsymbol{\beta} = \big(\beta_1, \ldots, \beta_p\big)^\intercal$, der unbekannten Baseline Hazardfunktion $h_0(t)$ und dem Kovariablenvektor $\boldsymbol{X}(t)$, wobei der Kovariablenvektor $\boldsymbol{X}(t)$ sowohl zeitabhängige als auch zeitunabhängige Kovariablen enthält, d.h. $\boldsymbol{X}(t) = \big(X_1, \ldots, X_b, X_{b+1}(t), \ldots, X_p(t)\big)^\intercal$.

Die partielle Likelihoodfunktion für ein zeitabhängiges PH Modell mit bindungsfreien Ausfallszeitpunkten ist praktisch äquivalent zu der partiellen Likelhoodfunktion des klassischen PH Modells (siehe (5.9)). Die partielle Likelihoodfunktion lautet

$$L_p(\boldsymbol{\beta}) = \prod_{i=1}^{k} \frac{\exp\big(\boldsymbol{\beta}^\intercal \boldsymbol{x}_{(i)}(y_{(i)})\big)}{\sum_{j \in \mathcal{R}(y_{(i)})} \exp\big(\boldsymbol{\beta}^\intercal \boldsymbol{x}_j(y_{(i)})\big)}\,, \tag{5.27}$$

mit den sortierten bindungsfreien Ausfallszeitpunkten $y_{(1)} < \cdots < y_{(k)}$, der Menge der Individuen welche unmittelbar vor dem Zeitpunkt t noch unter

Beobachtung stehen $\mathcal{R}(t) := \{i : y_i \geq t\}$ und der Wert des (i)-ten Kovariablenvektors ausgewertet zum $y_{(i)}$-ten Zeitpunkt $\boldsymbol{x}_{(i)}(y_{(i)})$ (vgl. Klein und Moeschberger, 2003). Die weitere Herleitung der log-Likehoodfunktion, der Likelihood Ratio Teststatistik und der Likelihoodfunktion für gebundene Lebenszeiten ist äquivalent zum klassischen PH Modell.

Im klassischen PH Modell wurde die Hazardfunktion in zwei Teile aufgeteilt, der erste Teil beschreibt das zeitabhängige Grundrisiko (Baseline Hazardfunktion) $h_0(t)$ und der zweite Teil beschreibt den multiplikativen bzw. proportionalen Effekt der Risikofaktoren $\exp\big(\boldsymbol{\beta}^\intercal \boldsymbol{X}\big)$. Für das zeitabhängige PH Modell ist der zweite Teil der Hazardfunktion jedoch nicht mehr zeitunabhängig weswegen in diesem Modell nicht mehr die proportionale Hazard Annahme gilt.

Man betrachte zwei Individuen mit zeitabhängigen Kovariablenvektoren $\boldsymbol{X}(t)$ und $\boldsymbol{X}^*(t)$. Das dazugehörige Verhältnis der Hazardfunktionen bei einem zeitabhängigen PH Modell lautet dafür

$$\frac{h\big(t|\boldsymbol{X}(t)\big)}{h\big(t|\boldsymbol{X}^*(t)\big)} = \exp\big(\boldsymbol{\beta}^\intercal(\boldsymbol{X}(t) - \boldsymbol{X}^*(t))\big)\,. \tag{5.28}$$

Da sich beim Verhältnis der Hazardfunktionnen in (5.28) die zeitabhängigen Kovariablen von $\boldsymbol{X}(t)$ und $\boldsymbol{X}^*(t)$ über die Zeit verändern, ist $(\boldsymbol{X}(t) - \boldsymbol{X}^*(t))$ eine von t abhängige Funktion und somit nicht mehr konstant. Das beweist, dass für das zeitabhängige PH Modell die proportionale Hazardfunktion Annahme nicht mehr gültig ist (vgl. Liu, 2012).

5.5.3. Stratifiziertes Proportional Hazard (PH) Modell

Manchmal ist die Proportionale Hazard Annahme für eine Kovariable nicht zutreffend. In diesen Fall wird versucht nach dieser Kovariable zu stratifizieren. Als Folge ist diese Kovariable nicht mehr als Kovariable anzusehen sondern als Stratifizierungsfaktor.

Angenommen der Kovariablenvektor $\boldsymbol{X} = (X_1, \ldots, X_p)^\intercal$ erfüllt die Proportionale Hazard Annahme und die Kovariable Z verletzt diese Annahme. Zudem wird angenommen, dass Z genau s verschiedene Werte annimmt (d.h. Z ist eine faktorielle Kovariable).

Mit dem **PH Modell** wurde das PH Modell dahingehend erweitert, sodass nach der Kovariable Z stratifiziert werden kann. Dieses Modell ist definiert als

$$h_m(t|\boldsymbol{X}, Z) = h_{0m}(t)\exp(\boldsymbol{\beta}^\intercal \boldsymbol{X}) \quad \forall t \geq 0\,, \tag{5.29}$$

mit der unbekannten Baseline Hazardfunktion $h_{0m}(t)$ für jedes Stratum $m = 1, \ldots, s$ und dem unbekannten Paramatervektor $\boldsymbol{\beta} = \big(\beta_1, \ldots, \beta_p\big)^\intercal$(vgl. Klein und Moeschberger, 2003).

Bemerkung 13.
Lediglich die Baseline Hazardfunktionen $h_{0m}(t)$ unterscheiden sich für verschiedene Strata m, hingegen bleiben die Parameter $\boldsymbol{\beta}$ für unterschiedliche Strata gleich.

Bemerkung 14.
Angenommen es wird nach den Kovariablen $\boldsymbol{Z} = (Z_1, Z_2)^\intercal$ stratifiziert mit Wertebereichen $Z_1 = \{a, b, c\}$ und $Z_2 = \{1, 2\}$. Dann existieren für das stratifzierte PH Modell $s = 6 = 3 \cdot 2$ verschiedene Strata. Die Strata sind Kombinationen aus Werten von Z_1 und Z_2 und lauten: $\{a, 1\}, \{a, 2\}, \{b, 1\}, \{b, 2\}$, $\{c, 1\}$ und $\{c, 2\}$. Für $\boldsymbol{Z} = (Z_1, \ldots, Z_k)^\intercal$ mit $k > 2$ werden die Strata nach dem gleichem Prinzip gebildet wie im Fall $k = 2$.

Der Unterschied zum klassischen PH Modell ist, dass im stratifizierte PH Modell für jedes Stratum eine eigene Baseline Hazardfunktion angenommen wird. Dies hat zur Folge, dass die Proportionalität der Hazardfunktionen für verschiedene Strata nicht mehr gültig ist. Somit gilt für das Verhältnis der Hazardfunktionen für Kovariablenvektoren $\boldsymbol{X}$ und $\boldsymbol{X}^*$ mit verschiedenen Strata $j \neq k$

$$\frac{h(t|\boldsymbol{X}, Z = j)}{h(t|\boldsymbol{X}^*, Z = k)} = \frac{h_{0j}(t)\exp\left(\boldsymbol{\beta}^\intercal \boldsymbol{X}\right)}{h_{0k}(t)\exp\left(\boldsymbol{\beta}^\intercal \boldsymbol{X}^*\right)} = \frac{h_{0j}(t)}{h_{0k}(t)}\exp\big(\boldsymbol{\beta}^\intercal(\boldsymbol{X} - \boldsymbol{X}^*)\big)\,, \forall t \geq 0\,,$$

d.h. das Verhältnis ist nicht mehr konstant.

Die partielle Likelihoodfunktion des stratifizierten PH Modells für das Stratum $m = 1, \ldots, s$ lautet gleich wie im klassischen PH Modell (5.8), d.h.

$$L_{mp}(\boldsymbol{\beta}) = \prod_{i=1}^{k_m} \frac{\exp(\boldsymbol{\beta}^\intercal \boldsymbol{x}_{(i_m)})}{\sum_{j \in \mathcal{R}(y_{(i_m)})} \exp(\boldsymbol{\beta}^\intercal \boldsymbol{x}_{j_m})}\,,$$

mit der Anzahl an beobachteten Ausfällen im Stratum k_m und den dazugehörigen sortierten beobachteten Ausfallszeitpunkten $y_{(1_m)} < \cdots < y_{(k_m)}$ im Stratum m. Folglich gilt für die dazugehörige log partielle Likelihoodfunktion für das Stratum m

$$l_{mp}(\boldsymbol{\beta}) = \sum_{i=1}^{k_m} \boldsymbol{\beta}^\intercal \boldsymbol{x}_{(i_m)} - \sum_{i=1}^{k_m} \log \left(\sum_{j \in \mathcal{R}(y_{(i_m)})} \exp(\boldsymbol{\beta}^\intercal \boldsymbol{x}_{j_m}) \right) .$$

Die vollständige stratifzierte partielle Likelihoodfunktion ist das Produkt aller partiellen Likelihoodfunktionen der m verschiedenen Strata, d.h.

$$L_p^{\text{strat}}(\boldsymbol{\beta}) = \prod_{m=1}^{s} L_{mp}(\boldsymbol{\beta}) .$$

Folglich gilt dann für die stratifizierte log partielle Likelihoodfunktion

$$l_p^{\text{strat}}(\boldsymbol{\beta}) = \sum_{m=1}^{s} l_{mp}(\boldsymbol{\beta}) = \sum_{m=1}^{s} \log L_{mp}(\boldsymbol{\beta}) \tag{5.30}$$

(vgl. Liu, 2012).

Für ein stratifiziertes PH Modell wird angenommen, dass sich die Kovariablen $\boldsymbol{X}$ in den verschiedenen Strata identisch verhalten. Um diese Annahme überprüfen zu können wird nun ein Likelihood Ratio Test entwickelt. Sei $l_p^{\text{strat}}(\boldsymbol{\beta}) = \sum_{m=1}^{s} l_{mp}(\boldsymbol{\beta})$ die log partielle Likelihoodfunktion (5.30) des stratifizierten PH Modells. Weiters sei $l_p^m(\boldsymbol{\beta}^m)$ die teilweise log partielle Likelihoodfunktion eines PH Modells welches nur an die Daten des m-ten Stratums angepasst wurde. Folglich ist die log partielle Likelihoodfunktion eines PH Modells mit unterschiedlichen Parametern pro Stratum gleich $\sum_{m=1}^{s} l_p^m(\boldsymbol{\beta}^m)$.

Der Likelihood Ratio Test überprüft nun ob sich die Kovariablen in jedem Strata gleich verhalten, womit die Hypothese $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}^m, m = 1, \ldots, s$ geprüft wird. Für die Likelihood Ratio Teststatistik Λ_{LR} zur Überprüfung der Nullhypothese gilt

$$\Lambda_{LR} = 2 \left(\sum_{m=1}^{s} l_p^m(\hat{\boldsymbol{\beta}}^m) - l_p^{\text{strat}}(\hat{\boldsymbol{\beta}}) \right) \sim \chi^2_{p(s-1)} \tag{5.31}$$

unter H_0, mit $\hat{\boldsymbol{\beta}}^m$ dem MLE von $l_p^m(\boldsymbol{\beta}^m)$ und $\hat{\boldsymbol{\beta}}$ dem MLE von $l_p^{\text{strat}}(\boldsymbol{\beta})$ (vgl. Klein und Moeschberger, 2003).

5.6. Diagnostik

5.6.1. Residuen

Wie üblich in der Regressionsanalyse wird die Anpassungsgüte mithilfe von Residuen graphisch untersucht. Aufgrund von Zensierung gestaltet sich die Beurteilung der Anpassungsgüte jedoch als äußerst schwierig. Aus diesem Grund wurden eigens für die Survival Analysis, bzw. sogar für das Proportional Hazard Modell, spezielle geeignete Residuen eingeführt. Es werden nun einige dieser speziellen Residuen diskutiert, welche später dazu benutzt werden, um die Modellanpassung, die funktionale Form der Kovariablen und die proportionale Hazard Annahme zu untersuchen. Die Ausführungen in diesem Abschnitt basieren zu großen Teilen auf Collett (2003), Liu (2012) und Klein und Moeschberger (2003).

Es wird angenommen, dass n Lebenszeiten beobachtet wurden, wobei diese Lebenszeiten rechtszensiert sein können, d.h. es werden die rechtszensierten Ausfallszeitpunkte $y_1, \ldots, y_n$ beobachtet. Weiters wird angenommen, dass ein Proportional Hazard Modell an die Lebenszeiten angepasst wurde für den Kovariablenvektor $\boldsymbol{X} = (X_1, \ldots, X_p)^\intercal$. Die angepasste Hazardfunktion für das i-te Individuum, $i = 1, \ldots, n$, lautet somit

$$\widehat{h}_i(t|\boldsymbol{x}_i) = \widehat{h}_0(t) \exp\left(\hat{\boldsymbol{\beta}}^\intercal \boldsymbol{x}_i\right),$$

mit der geschätzten Baseline Hazardfunktion $\widehat{h}_0(t)$ und dem geschätzten Parametervektor $\hat{\boldsymbol{\beta}}$.

Cox Snell Residuen

Die Cox Snell Residuen sind die meist verwendeten Residuen in der Survival Analysis und wurden in allgemeiner Form in Cox und Snell (1968) erstmals eingeführt.

Das Cox Snell Residuum, für das i-te Residuum $i = 1, \ldots, n$ im PH Modell, ist definiert als

$$r_{Ci} := \widehat{H}(y_i|\boldsymbol{x}_i) = \widehat{H}_0(y_i) \exp\left(\hat{\boldsymbol{\beta}}^\intercal \boldsymbol{x}_i\right), \tag{5.32}$$

mit dem Schätzer der kumulativen Baseline Hazardfunktion $\widehat{H}_0(y_i)$ ausgewertet zum Ausfallszeitpunkt y_i dieses Individuums.

Falls das Proportional Hazard Modell korrekt ist und der geschätzte Parametervektor $\hat{\boldsymbol{\beta}}$ nahe dem wahren Parametervektor $\boldsymbol{\beta}$ liegt, dann sollten die Residuen r_{Ci}, $i = 1, \ldots, n$, einer zensierten Stichprobe aus einer Exp(1)-verteilten Population gleichen (siehe Herleitung der Cox Snell Residuen für das AFT-Modell in Abschnitt 4.4). Somit sollte für ein ausreichend gut angepasstes PH Modell der Plot von r_{Ci} gegen $\widehat{H}(r_{Ci}) = -\log\big(\widehat{S}(r_{Ci})\big)$, $i = 1, \ldots, n$, einer Geraden mit Steigung 1 entsprechen (siehe Exponentialverteilung Abschnitt 4.9).

Man beachte, dass die Cox Snell Residuen grundlegend verschiedene Eigenschaften als die Residuen der Linearen Regression aufweisen. So sind Cox Snell Residuen nicht symmetrisch um 0 verteilt und nicht-negativ!

Martingal Residuen

Das i-te Martingal Residuum $i = 1, \ldots, n$ im PH Modell ist definiert als

$$r_{Mi} := R_i - r_{Ci} = R_i - \widehat{H}_0(y_i) \exp\big(\widehat{\boldsymbol{\beta}}^\intercal \boldsymbol{x}_i\big)\,, \tag{5.33}$$

mit der Rechtszensur-Indikatorvariable R_i und dem Cox-Snell Residuum r_{Ci}.

Das Martingal Residuum r_{Mi} kann als Differenz zwischen der beobachteten Anzahl an Ausfällen (0 oder 1) für ein Individuum i in $[0, y_i]$ und der erwarteten Anzahl an Ausfällen basierend auf dem angepassten PH Modell angesehen werden.

Die Herleitung der Martingal Residuen erfolgt über die Martingal Theorie und der Formulierung des PH Modells als Zählprozess. Die detaillierte Herleitung kann unter anderem in Therneau und Grambsch (2000) und Liu (2012) nachgeschlagen werden.

Für den wahren Parametervektor $\boldsymbol{\beta}$ eingesetzt in (5.33), ist r_{Mi} ein Martingal (d.h. $\mathrm{E}(r_{Mi}) = 0$). Weiters haben Martingal Residuen den Wertebereich $(-\infty, 1]$, wobei für zensierte Lebenszeiten ($r_i = 0$) die Martingal Residuen negative Werte annehmen und für nicht zensierte Lebenszeiten ($r_i = 1$) die Martingal Residuen positive Werte annehmen. Zudem sind Martingal Residuen unkorreliert für große Stichproben und es gilt $\sum_{i=1}^{n} r_{Mi} = 0$.

Deviance Residuen

Ein großer Nachteil der Martingal Residuen ist, dass sie nicht symmetrisch um 0 verteilt sind wie die Residuen der Linearen Regression. Folglich sind Plots von Martingal Residuen schwer interpretierbar.

Die Deviance Residuen für ein PH Modell sind definiert durch

$$r_{Di} = \text{sign}(r_{Mi}) \left(-2\big[r_{Mi} + R_i \log(R_i - r_{Mi})\big]\right)^{1/2} , \tag{5.34}$$

mit der Vorzeichenfunktion sign, dem Martingal Residuum r_{Mi} und der Rechtszensur-Indikatorvariable R_i. Somit sind die Deviance Residuen transformierte Martingal Residuen, welche symmetrisch um 0 verteilt sind falls das Modell korrekt ist.

5.6.2. Überprüfung der Modellanpassung

In diesem Abschnitt wird die Untersuchung der Modellanpassung mithilfe der Cox Snell Residuen und der Deviance Residuen diskutiert. Es wird in beiden Fällen das Modell für die Kehlkopfkrebs Studie aus Beispiel 5.4 untersucht, für das gilt

$$h(t|\boldsymbol{X}) = h_0(t) \exp\big(\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4\big) ,$$

mit X_i, $i = 1, 2, 3$, den Indikatorvariablen für Stadium II, III und IV und X_4 dem Alter.

Überprüfung der Modellanpassung mit Cox Snell Residuen

Es wird nun das Kehlkopfkrebs Beispiel darauf untersucht, ob das Modell ausreichend gut an die Daten angepasst wurde.

Dafür werden in R-Code 5.3 die Cox Snell Residuen r_{Ci} gegen $-\log(\widehat{S}(r_{Ci}))$ abgebildet (siehe Abschnitt 5.6.1). Anhand von Abbildung 5.2 sieht man, dass die Cox Snell Residuen eine Gerade mit Steigung 1 bilden und daher das Modell als gut angepasst angesehen werden kann.

In der Abbildung 4.8 wurden ebenfalls Cox Snell Residuen aufgetragen, jedoch für ein Weibull AFT Modell. Bei der Betrachtung der beiden Abbildungen 5.2 und 4.8 der Cox Snell Residuen scheinen beide Modelle eine in etwa gleich gute Modellanpassung aufzuweisen.

R-Code 5.3: Überprüfung der Modellanpassung mit Cox Snell Residuen

```
> data("larynx")
> larynx.coxph <- coxph(Surv(time,delta) ~ as.factor(stage) + age,
+                                          data = larynx)
> # Berechnen der Cox Snell RESIDUEN
> r.cs <- delta - residuals(larynx.coxph, type = "martingale")
> r.surv <- survfit(Surv(r.cs, delta) ~ 1, data = larynx)
>
> # Plot der Cox Snell RESIDUEN
> plot(r.surv$time, -log(r.surv$surv), col = "grey55", pch = 21,
+       xlab = expression(r[Ci]),
+       ylab = expression(paste("H ", (r[Ci]), sep = " ")))
> abline(0, 1, lwd = 1.5)
```

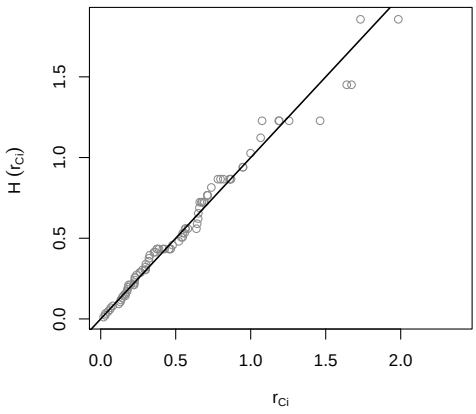

Abbildung 5.2.: Cox Snell Residuen des PH Modells.

Überprüfung der Modellanpassung mit Deviance Residuen

Die kommenden Ausführungen in diesem Abschnitt orientieren sich stark an Collett (2003). Das Martingal Residuum r_{Mi} kann als Differenz zwischen der beobachteten Anzahl an Ausfällen (0 oder 1) für Individuum i in $[0, y_i]$ und der erwarteten Anzahl an Ausfällen basierend auf dem angepassten Proportional Hazard Modell angesehen werden. Somit heben die Martingal Residuen Individuen hervor, welche gemäß dem angepassten PH Modell zu früh (große negative Martingal Residuen) oder zu spät (Martingal Residuen nahe 1) ausgefallen sind.

Somit sollten, mithilfe eines Plots der Martingal Residuen, Ausreißer gefunden werden. Leider sind die Martingal Residuen jedoch asymmetrisch um 0 verteilt, was die Aufgabe erheblich erschwert. Da die Deviance Residuen transformierte Martingal Residuen sind, welche symmetrisch um 0 verteilt sind, werden Deviance Residuen zur Auffindung von Ausreißern verwendet.

Da der lineare Prädiktor $\widehat{\boldsymbol{\beta}}^\intercal \boldsymbol{x}_i$ das geschätzte Ausfallrisiko für ein Individuum i mit Kovariablenvektor $\boldsymbol{x}_i$ angibt, ist es sinnvoll r_{Di} und $\widehat{\boldsymbol{\beta}}^\intercal \boldsymbol{x}_i$ miteinander zu vergleichen.

Bemerkung 15.
Der lineare Prädiktor $\widehat{\boldsymbol{\beta}}^\intercal \boldsymbol{x}_i$ wird im Zusammenhang mit PH Modellen auch Risk Score genannt.

Für eine niedrige bis mittlere Anzahl an zensierten Beobachtungen sollte der Plot $\widehat{\boldsymbol{\beta}}^\intercal \boldsymbol{x}_i$ gegen r_{Di} einem normalverteilten Weißen Rauschen ähneln. Für eine große Anzahl an zensierten Beobachtungen liegt ein Großteil der Residuen nahe bei 0.

Es wird nun wieder das Kehlkopfkrebs Beispiel aufgegriffen und dafür mithilfe der Deviance Residuen auf Ausreißer untersucht. Im linken Bild der Abbildung 5.3 ist der Plot der Deviance Residuen zu sehen und im rechten Bild der Abbildung 5.3 ist der lineare Prädiktor $\widehat{\boldsymbol{\beta}}^\intercal \boldsymbol{x}_i$ gegen r_{Di} aufgetragen.

R-Code 5.4: Überprüfung der Modellanpassung mit Deviance Residuen

```
1 > data("larynx")
2 > larynx.coxph <- coxph(Surv(time,delta) ~ as.factor(stage) +  age,
3 +                                  data = larynx)
```

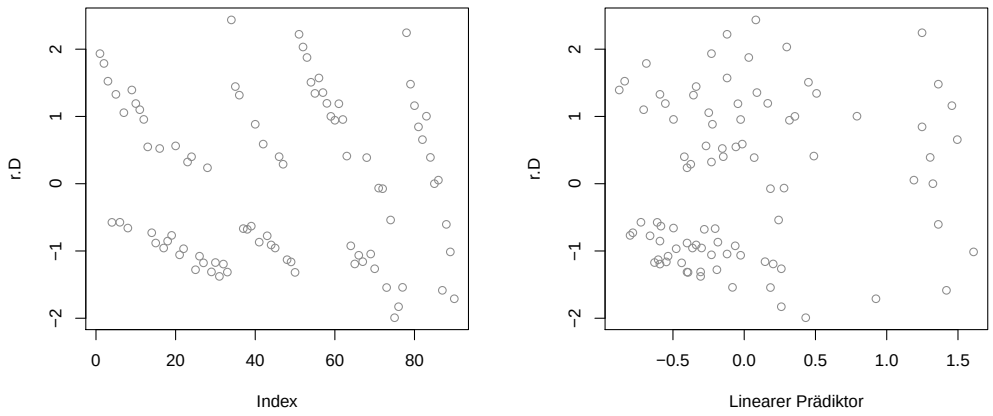

Abbildung 5.3.: Index Plot der Deviance Residuen r_{Di} (links), Plot des linearen Prädiktors $\widehat{\boldsymbol{\beta}}^{\mathsf{T}}\boldsymbol{x}_i$ gegen r_{Di} (rechts).

```
> # ERSTELLEN DER DEVIANCE RESIDUEN
> r.d <-  residuals(larynx.coxph, type = "deviance")
> plot(r.d, col = "grey55", pch = 21)
> plot( larynx.coxph$linear.predictors, r.d,col = "grey55", pch = 21)
```

Grundsätzlich sind in der Abbildung 5.3 keine Ausreißer zu erkennen. Interessehalber betrachten wir nun trotzdem das größte Deviance Residuum mit Wert 2.44 und Index 34 bzw. linearem Prädiktor 0.0815. Der dazugehörige Kehlkopfkrebs Patient war 86 Jahre Alt, im Stadium II und verstarb bereits nach 0.2 Jahren. Somit verstarb der Patient gemäß dem angepassten PH Modell viel zu früh, womit dieses hohe Deviance Residuum erklärt werden kann.

5.6.3. Überprüfen der funktionalen Form der Kovariable

Mithilfe von Martingal Residuen r_{Mi} (siehe Abschnitt 5.6.1) kann die funktionale Form des Effekts von metrischen Kovariablen im Prädiktor bestimmt werden. Mögliche funktionale Formen für eine Kovariable $X_j \in \boldsymbol{X}$ sind

- die lineare Form $\beta_j X_j$
- oder nichtlineare Form von X_j, d.h. $f(X_j)$.

Um die funktionale Form zu bestimmen, wird im ersten Schritt ein Null PH Modell (PH Modell ohne Kovariablen) an die Daten angepasst. Im zweiten Schritt werden die Martingal Residuen r_{Mi} des angepassten PH Modells gegen die Werte der Kovariable X_j abgebildet (vgl. Höhle, WS 2008/2009).

Beispiel 9.
Zur Veranschaulichung wird die funktionale Form anhand eines Beispiels diskutiert. Die Daten aus dem Beispiel stammen aus Klein und Moeschberger (2003) und beinhalten die Daten von Knochenmarktransplantationen von 43 Patienten mit Morbus Hodgkin oder Non-Hodgkin-Lymphom. Die Daten beinhalten die zwei Transplantationsformen autogenisch oder allogenisch sowie die Kovariablen Wartezeit bis zur Transplantation und den Karnofksy-Index (quantifiziert die Lebensqualität der Krebspatienten).

Im R-Code 5.5 wird nun untersucht welche funktionale Form die Kovariable Wartezeit bis zur Transplantation aufweist. Anhand von Abbildung 5.4 ist klar zu erkennen, dass die Kovariable Wartezeit bis zur Transplantation keine lineare Form aufweist. In Klein und Moeschberger (2003) wurde vorgeschlagen die Kovariable als Indikatorvariable mit

$$X = \begin{cases} 0, \text{ falls Wartezeit} < 70 \\ 1, \text{ falls Wartezit } \geq 70 \end{cases}$$

zu verwenden.

R-Code 5.5: Überprüfung der funktionalen Form

```
> data("hodg")
> # ANPASSEN DES MODELLS
> m.hodg <- coxph(Surv(time, delta) ~ gtype * dtype + score,
+                                       data = hodg)
>
> # BERECHNEN DER MARTINGAL RESIDUEN
> r.m <-  residuals(m.hodg, type = "martingal")
> plot(x <- hodg$wtime, r.m, col = "grey55", pch = 21,
+         xlab = "Wartezeit (Monate)", ylab = expression(r[Ci]))
```

```
10 > lines(lowess(hodg$wtime, r.m), lwd = 2)
```

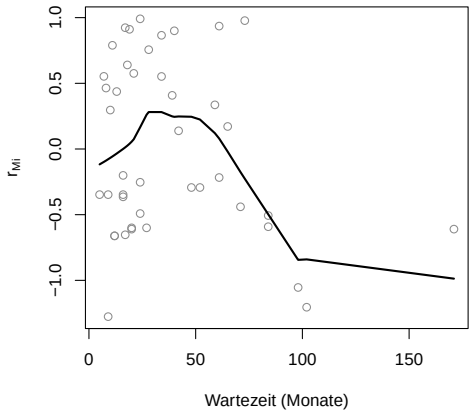

Abbildung 5.4.: Wartezeit bis zur Transplantation gegen Martingal Residuen r_{Mi}.

5.6.4. Überprüfung der Proportional Hazard Annahme

Für ein PH Modell

$$h(t|\boldsymbol{X}) = h_0(t) \exp\left(\boldsymbol{\beta}^\intercal \boldsymbol{X}\right) \quad \forall t \geq 0\,,$$

ist die zentrale Annahme, dass die Hazardfunktionen zu jedem Zeitpunkt t proportional zueinander sind. Somit ist das Verhältnis der Hazardfunktionen zweier Individuen, mit Kovariablenvektoren $\boldsymbol{X}$ und $\boldsymbol{X}^*$, zu jedem Zeitpunkt t konstant θ, d.h.

$$\begin{aligned} \frac{h(t|\boldsymbol{X})}{h(t|\boldsymbol{X}^*)} = \exp\left(\boldsymbol{\beta}^\intercal(\boldsymbol{X} - \boldsymbol{X}^*)\right) &= \theta\,, \quad \text{bzw.} \\ h(t|\boldsymbol{X}) &= \theta h(t|\boldsymbol{X}^*)\,. \end{aligned}$$

Aus diesem Grund sollte bevor ein PH Modell überhaupt angepasst wird die proportionale Hazardfunktion Annahme überprüft werden. Die Ausführungen im kommenden Abschnitt folgen dabei in großen Teilen Kleinbaum und Klein (2005).

Graphische Beurteilung der proportionalen Hazard Annahme

Mit der hier vorgestellten Methode kann graphisch beurteilt werden, ob die proportionale Hazard Annahme für eine Kovariable gültig ist. Die Methode beruht darauf, dass durch die $\log(-\log)$-Transformation von $S(t|\boldsymbol{X})$ eine lineare Beziehung zwischen $\log\big(H(t|\boldsymbol{X})\big)$ und $\boldsymbol{\beta}^\intercal \boldsymbol{X}$ entsteht.

Laut (5.3) gilt die Beziehung

$$S(t|\boldsymbol{X}) = S_0(t)^{\exp(\boldsymbol{\beta}^\intercal \boldsymbol{X})}\,, \tag{5.35}$$

mit der Baseline Survivalfunktion $S_0(t) := \exp\Big(-\int_0^t h_0(u)du\Big)$. Logarithmiert man $S(t|\boldsymbol{X})$ aus (5.35), so erhält man laut (2.5) die kumulierte Hazardfunktion, d.h.

$$H(t|\boldsymbol{X}) = -\log\big(S(t|\boldsymbol{X})\big) = -\log\big(S_0(t)\big)\cdot\exp(\boldsymbol{\beta}^\intercal \boldsymbol{X})\,. \tag{5.36}$$

Durch nochmaliges Logarithmieren von (5.36) erhält man die lineare Beziehung

$$\log\big(H(t|\boldsymbol{X})\big) = \log\big(-\log\big(S(t|\boldsymbol{X})\big)\big) = -\log\big(\log\big(S_0(t)\big)\big) + \boldsymbol{\beta}^\intercal \boldsymbol{X}\,. \tag{5.37}$$

Folglich müssen die Graphen des Plots t gegen $\log\big(-\log\big(S(t|\boldsymbol{X})\big)\big)$, für verschiedenen Werte der Kovariable $\boldsymbol{X}$, parallel zueinander sein, falls die proportionale Hazard Annahme gültig ist.

Beispiel 10.
Es wird nun das Kehlkopfkrebs Beispiel aus Abschnitt 5.4 aufgegriffen, um die graphische Beurteilung des proportionalen Hazards für den Faktor des Krebsstadiums durchzuführen. Aus diesem Grund wird die Variable für das Alter X_4 im Modell weggelassen. Das abgeänderte PH Modell lautet nun

$$h(t|\boldsymbol{X}) = h_0(t)\exp\big(\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3\big)\,,$$

mit X_i, $i = 1, 2, 3$, den Indikatoren für Stadium II, III und IV. Es wird nun t gegen $\log\Big(-\log\big(\widehat{S}(t|\boldsymbol{X})\big)\Big)$ in einem Plot aufgetragen, für die Indikatorvariablen X_i, $i = 1, 2, 3$, des Kehklkopfkrebs Stadiums und $\widehat{S}(t|\boldsymbol{X})$ dem Kaplan-Meier Schätzer.

Falls die Annahme einer proportionalen Hazardfunktionen korrekt ist, gelten laut (5.37) folgende Aussagen:

- der Abstand der Graphen für Stadium I und Stadium II beträgt konstant β_1,
- der Abstand der Graphen für Stadium I und Stadium III beträgt konstant β_2,
- der Abstand der Graphen für Stadium II und Stadium III beträgt konstant $(\beta_3 - \beta_2)$,
- usw.

R-Code 5.6: Graphische Überprüfung der proportionalen Hazard Annahme

```
1 > larynx.stage.km <- survfit(Surv(time, delta) ~ as.factor(stage))
2 > plot(larynx.stage.km, fun = "cloglog", col =
3 +   c("darkgreen", "gold2", "darkorange2", "firebrick2"), lwd = 3)
```

In der Abbildung 5.5 sind die verschiedenen Graphen zur Beurteilung der proportionalen Hazardfunktion Annahme für die Stadien der Kehlkopfkrebs Studie abgebildet. Die Graphen in Abbildung 5.5 für Stadium IV, Stadium III und Stadium II wirken parallel. Einzig der Graph für Stadium I kreuzt den Graph von Stadium II etwas. Da aber diese Graphen sowieso sehr nah zusammen liegen, wird die proportionale Hazardfunktionen Annahme hier nicht verworfen.

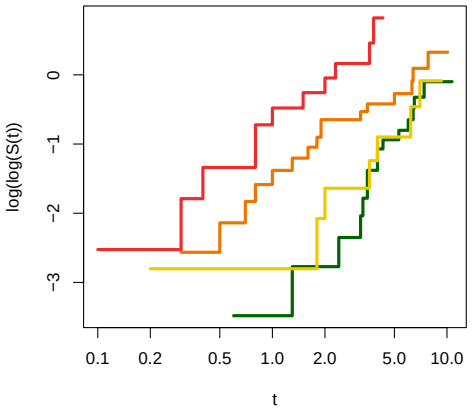

Abbildung 5.5.: $\log\Big(-\log\big(\widehat{S}(t|\boldsymbol{X})\big)\Big)$ für Kehlkopfkrebs Patienten im Stadium I (grün), Stadium II (gelb), Stadium II (orange) und Stadium IV (rot).

6. Stornierungs-Modelle

In den vorangegangenen Abschnitten wurde die Theorie der Survival Analysis aufbereitet. Nun werden diese theoretischen Konzepte eingesetzt die Vertragslaufzeiten von Krankenversicherungspolizzen zu modellieren, um dadurch Stornowahrscheinlichkeiten schätzen zu können.

Es wird nun kurz die Aufgabenstellung motiviert. Versicherungsunternehmen mit einer Krankenversicherungssparte halten eine Vielzahl an Krankenversicherungspolizzen. Eine Krankenversicherungspolizze ist ein Vertrag, welcher zwischen einem Versicherungsnehmer und einer Versicherung abgeschlossen wird. Er regelt die teilweise oder volle Kostenerstattung für die Behandlung von Krankheiten, Mutterschaft oder Unfällen des Versicherten gegen eine monatliche Versicherungsprämie an die Versicherung.

Die Kennzahl Stornowahrscheinlichkeit ist für eine Krankenversicherung essentiell. Sie gibt das Risiko an, mit welcher eine Krankenversicherungspolizze storniert wird.

Es gilt nun ein Modell zu finden, welches die Stornowahrscheinlichkeiten beschreibt. Der klassische Zugang zur Lösung dieses Problems wäre die Modellierung der Stornowahrscheinlichkeiten pro Kalenderjahr mithilfe eines logistischen Modell (siehe Moosbrugger, 2016). Der Ansatz dieser Arbeit ist es, die Stornowahrscheinlichkeit über den Verlauf der Versicherungslaufzeit von Krankenversicherungspolizzen zu modellieren.

Hauptziel ist es nun ein Modell zu finden, welches Antworten auf folgende Fragen zu finden:

- Existieren Faktoren, welche Einfluss auf die Stornowahrscheinlichkeit und dadurch auf die Vertragslaufzeit haben? Wie wirken sich diese Faktoren aus?

- Wie hoch ist die Wahrscheinlichkeit, dass ein Versicherungsnehmer nach t Jahren noch nicht storniert hat?

6.1. Beschreibung der Daten

Der Datensatz für das Beispiel wurde von der Merkur Versicherung AG zur Verfügung gestellt. Er umfasst ca. 280000 Krankenversicherungspolizzen, welche im Zeitraum 2001 bis 2014 beobachtet wurden. Somit beinhaltet der Datensatz sämtliche Polizzen, die im Zeitraum 2001 bis 2014 abgeschlossen wurden. Zusätzlich beinhaltet der Datensatz aber auch den Bestand an Polizzen zum Stichtag 1. Januar 2001. Das bedeutet, dass auch sämtliche Polizzen die vor 2001 abgeschlossen aber erst nach 2001 storniert wurden erfasst sind. Dadurch können Polizzen analysiert werden deren Beginnjahr bis ins Jahr 1967 reichen.

Im folgenden werden die verschiedenen erfassten Merkmale einer Polizze genauer erklärt:

- `PRAEMIE_MONATLICH`: Die monatlich zu zahlende Prämie.
- `BEGINN_JAHR`: Das Jahr in dem die Polizze abgeschlossen wurde.
- `SEX`: Das Geschlecht des Versicherungsnehmers.
- `GRUPPE_RABATT`: Polizzen können Teil einer Gruppenversicherung sein und in diesem Fall auch einen Rabatt erhalten. Polizzen die nicht Teil einer einer Gruppenversicherung sind, können jedoch keinen Rabatt erhalten.

  Der Faktor `GRUPPE_RABATT` beschreibt nun diese Beziehung zwischen

Rabatt und Gruppenversicherung und hat deswegen die Ausprägungen

$$\texttt{GRUPPE\_RABATT} = \begin{cases} \texttt{Keine Gruppe}, & \text{Polizze ist nicht Teil einer Gruppenversicherung.} \\ \texttt{Gruppe}, & \text{Polizze ist nicht rabattiert, jedoch Teil einer Gruppenversicherung.} \\ \texttt{Rabatt}, & \text{Polizze ist Teil einer Gruppenversicherung und ist rabattiert.} \end{cases},$$

- `DELTA`: Ist die Rechts-Indikatorvariable mit

$$\texttt{DELTA} = \begin{cases} 1, & \text{falls die Polizze storniert wurde,} \\ 0, & \text{sonst.} \end{cases}$$

  Polizzen können aus genau zwei Gründen rechtszensiert werden (d.h. `DELTA = 0`):

  1. Die Polizze wurde bis zum Ende des Beobachtungsraum 2014 nicht storniert

  2. Die Polizze wurde aufgelöst, weil der Versicherungsnehmer verstarb.

- `LAUFZEIT`: Die beobachtete Laufzeit der Polizze.

- `ALTER_STORNO`: Gibt das Alter an, zu dem der Versicherungsnehmer das letzte mal beobachtet wurde. `ALTER_STORNO` kann somit das Alter zum Zeitpunkt des Stornos, Todes oder Ende 2014 sein.

Bei den hier untersuchten Daten handelt es sich um rechtszensierte Daten. Von den insgesamt ca. 280000 zu untersuchenden Versicherungspolizzen sind ca. 230000 rechtszensiert. Es gilt diesen hohen Anteil an rechtszensierten Daten immer zu beachten da, er später eventuell der Grund für Probleme sein kann.

6.2. Deskriptive Analysis

Es wird nun versucht die Datensitutation der 280000 Polizzen bestmöglich darzustellen. Dafür wird der beobachtete relative Stornoanteil (d.h. Prozentsatz der stornierten Polizzen) in Abhängigkeit von den Merkmalen Beginnjahr, Stornoalter und monatlicher Prämie in den Abbildungen 6.1, 6.2 und 6.3 dargestellt. In diesen Abbildungen wird der relative Stornoanteil noch weiters für die Merkmale `SEX` und `GRUPPE_RABATT` aufgeteilt.

6.2.1. `BEGINN_JAHR`

In der Abbildung 6.1 ist der relative Stornoanteil in Abhängigkeit von Beginnjahr abgebildet. Allgemein ist für den relativen Stornoanteil in Abhängigkeit von Beginnjahr in Abbildung 6.1 ein Anstieg bis zum Jahr 2001 zu erkennen, gefolgt von einem Abfall. Dieses Verhalten ist durch die spezielle Datensituation zu erklären, welche abgeschlossene Krankenversicherungspolizzen im Zeitraum 2001 bis 2014 beobachtet sowie den Bestand an Krankenversicherungspolizzen zum Stichtag 1. Januar 2001.

Der Anstieg des relativen Stornoanteils kann wie folgt erklärt werden: Betrachtet man eine Polizze mit `BEGINN_JAHR = 1968`, dann hat diese Polizze bereits 33 Jahre Vertragslaufzeit bis zum Beginn der Untersuchung und somit ein geringes Restrisiko in verbleibenden Zeit storniert zu werden. Im Gegensatz dazu hat eine Polizze mit `BEGINN_JAHR = 2001` ein viel höheres Stornorisiko, da gerade in den ersten Vertragsjahren die Versicherungspolizze vermehrt storniert wird.

Der Abfall des relativen Stornoanteils ab 2001 lässt folgendermaßen erklären. Betrachtet man eine Polizze mit `BEGINN_JAHR = 2001`, dann kann die Polizze in den Jahren bis 2014 storniert werden. Es besteht somit die Möglichkeit die Polizze 13 Jahre lang zu stornieren. Für eine Polizze mit `BEGINN_JAHR = 2012` besteht die Möglichkeit nur 2 Jahre lang zu stornieren. Folglich ist der relative Stornoanteil für Polizzen mit `BEGINN_JAHR = 2001` deutlich höher als für Polizzen mit `BEGINN_JAHR = 2012` bzw. allgemein für Polizzen mit `BEGINN_JAHR > 2001`.

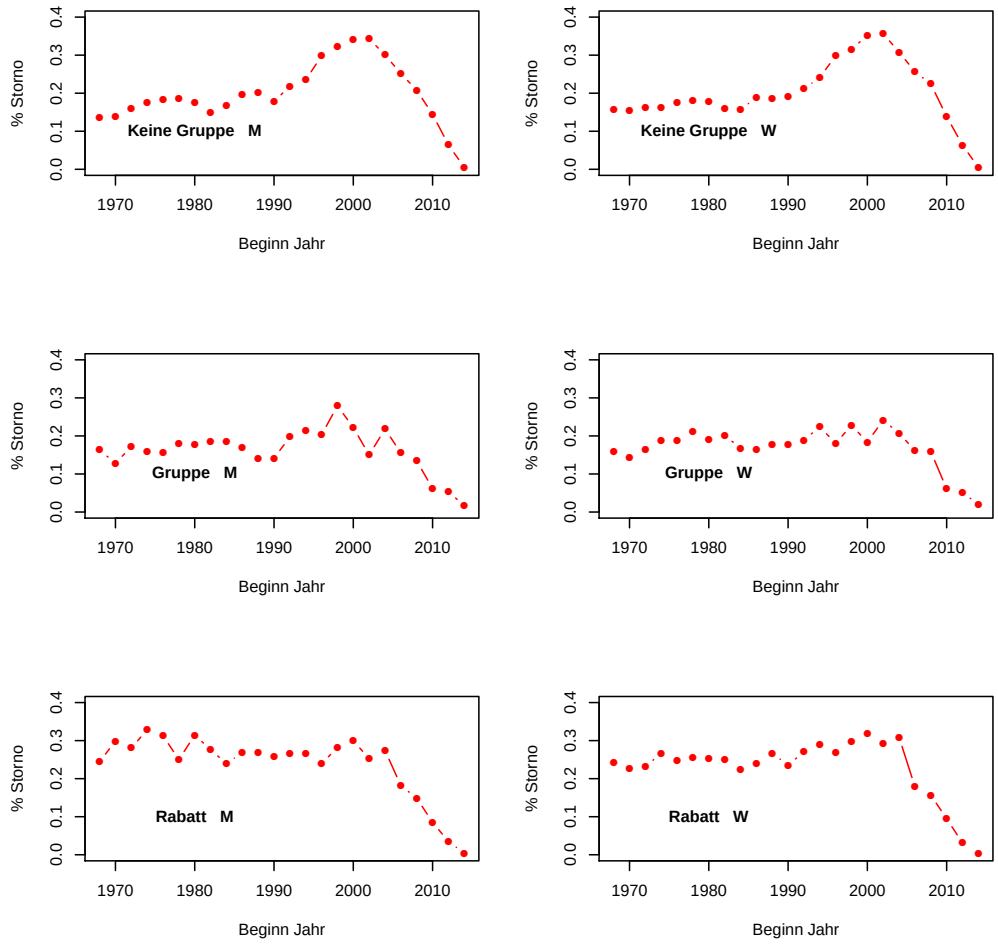

Abbildung 6.1.: Beobachteter Stornoanteil in Abhängigkeit von Beginnjahr.

Für die relativen Stornoanteile über `BEGINN_JAHR` in Abbildung 6.1 ist bezüglich Geschlecht (`SEX`) kein Unterschied beobachtbar. Auch bezüglich des relativen Stornoanteils des Merkmals `GRUPPE_RABATT` sind keine großen Abweichungen feststellbar.

6.2.2. PRAEMIE_MONATLICH

In der Abbildung 6.2 ist der relative Stornoanteil in Abhängigkeit von der monatlichen Prämie abgebildet. Für die relativen Stornoanteile sind bezüglich Geschlecht (SEX) keine Unterschiede und bezüglich GRUPPE_RABATT nur leichte Unterschiede feststellbar. Der Stornoanteil über PRAEMIE_MONATLICH in Abbildung 6.2 weist eine nur schwer interpretierbare Form auf.

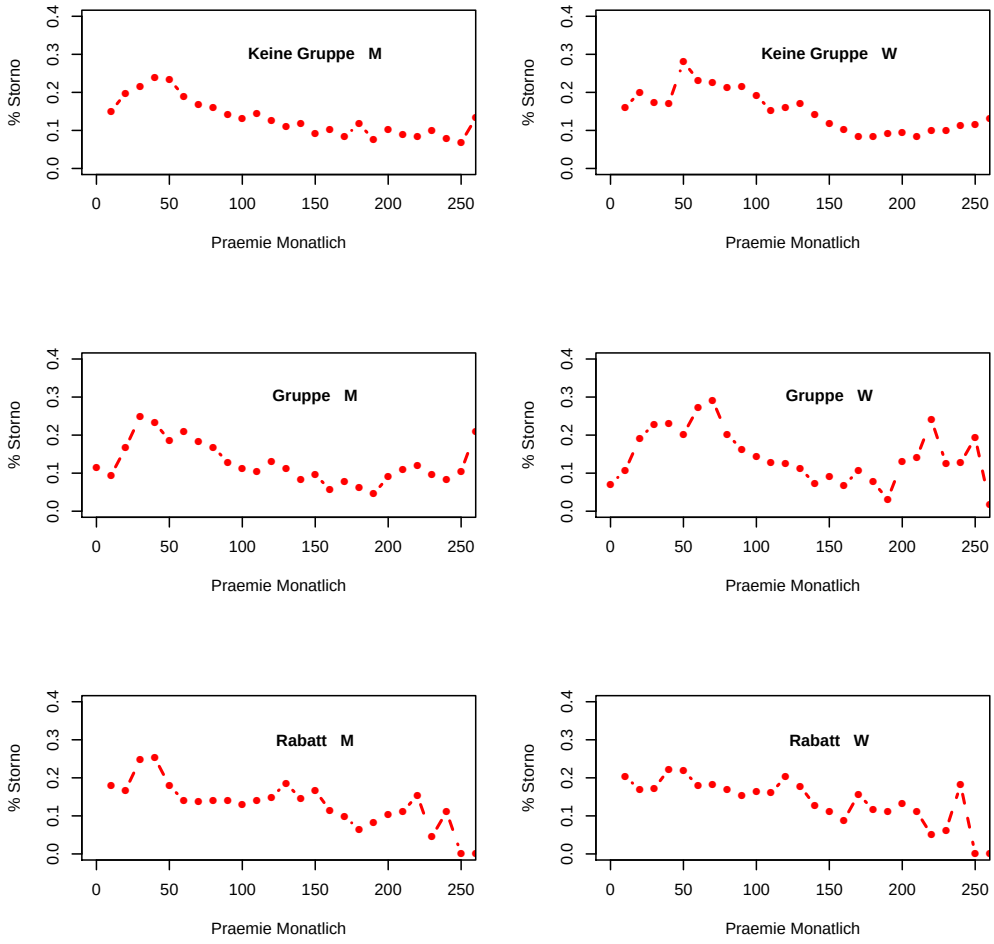

Abbildung 6.2.: Beobachteter Stornoanteil über der monatlichen Prämie.

6.2.3. ALTER_STORNO

In der Abbildung 6.3 ist der relative Stornoanteil in Abhängigkeit des Stornoalters abgebildet. Der relative Stornoanteil über ALTER_STORNO hat um Anfang 20 ein Maximum. Ein Grund hierfür liegt darin, dass in diesem Alter die Prämien nicht mehr die Eltern bezahlen, sondern der jeweilige Versicherungsnehmer.

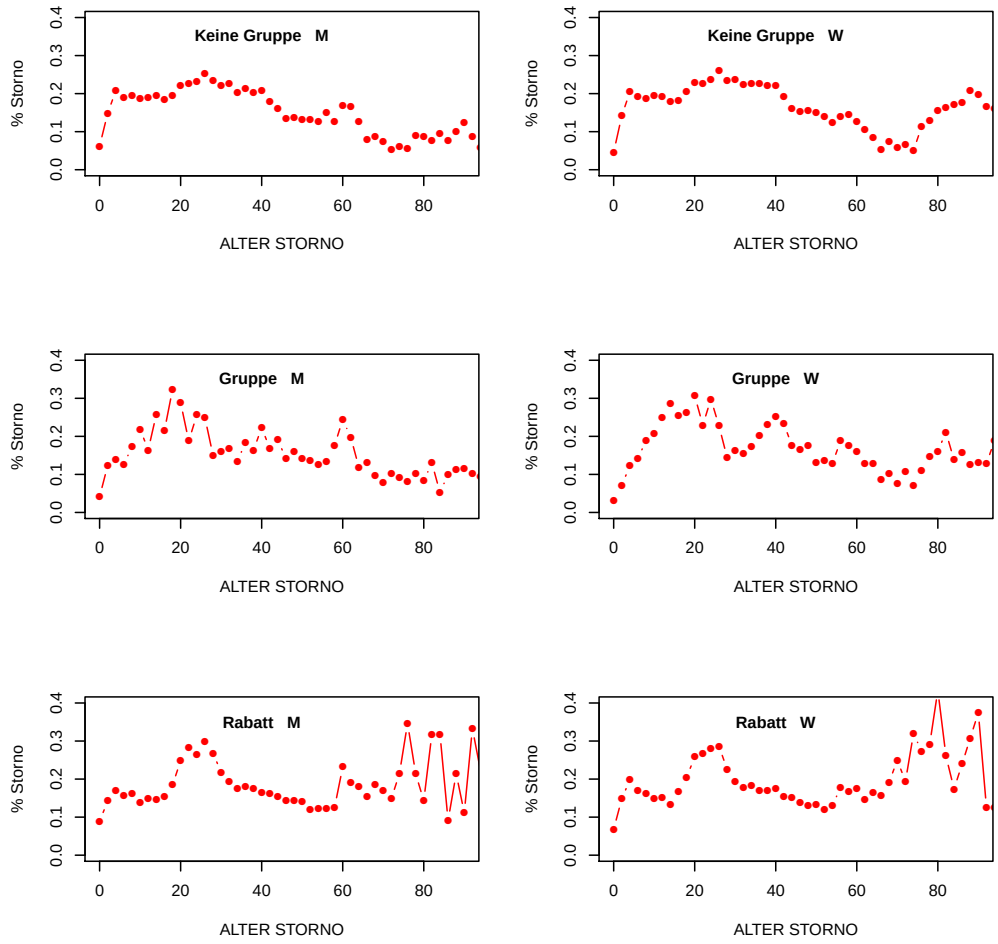

Abbildung 6.3.: Beobachteter Stornoanteil über ALTER_STORNO.

Ein weiteres lokales Maximum ist um `ALTER_STORNO=55` erkennbar. In dieses Alter fällt für viele Versicherungsnehmer der Pensionsantritt. Offenbar entscheiden sich viele Versicherungsnehmer in den ersten Pensionsjahren gegen eine Krankenversicherung.

Für die relativen Stornoanteile über `ALTER_STORNO` in Abbildung 6.3 sind bezüglich Geschlecht (`SEX`) keine Unterschiede und bezüglich `GRUPPE_RABATT` nur kleine Unterschiede feststellbar.

Allgemein ist anhand der Abbildungen 6.1, 6.2 und 6.3 für die Merkmale `BEGINN_JAHR`, `ALTER_STORNO` und `PRAEMIE_MONATLICH` feststellbar, dass der relative Stornoanteil für diese Merkmale als nicht konstant angesehen werden kann. Dies deutet daraufhin, dass diese Merkmale in Survival Modellen mithilfe von glatten Funktionen modelliert werden sollten.

6.3. Deskriptive Survival Analysis

Mithilfe der deskriptiven Survival Analysis wird nun ein genaueres Bild von den Stornierungen über die Vertragslaufzeit gezeichnet. Mit der Survivalfunktion $S(t)$ wird hierbei die Wahrscheinlichkeit bezeichnet, nach t Vertragsjahren die Versicherungspolizze nicht storniert zu haben.

In der Abbidlung 6.4 wurde der Kaplan Meier Schätzer der Stornierungswahrscheinlichkeit in Abhängigkeit von der Vertragslaufzeit abgebildet. Es ist zu erkennen, dass vor allem in den ersten 10 Vertragsjahren die Polizze häufiger storniert wird. Weiters erkennt man, dass ab ca. 20 Vertragsjahren die Stornorate pro Jahr als „konstant“ angesehen werden kann. Allgemein scheitn selbst nach über 40 Jahren Vertragslaufzeit, die Wahrscheinlichkeit die Polizze nicht storniert zu haben größer, als 50% zu sein.

In Abbildung 6.5 ist der Kaplan Meier Schätzer der Stornierungswahrscheinlichkeit über die Vertragslaufzeit bezüglich des Merkmals `SEX` abgebildet.In dieser Abbildung ist kein Unterschied in den geschätzten Survivalfunktionen für männliche und weibliche Versicherungsnehmer festzustellen. Auch der Log Rank Test für `SEX` mit p-Wert 0.22 (siehe `R`-Code 6.1) unterstützt die Hypothese, dass die Survivalfunktionen für Frauen und Männer als identisch anzusehen sind.

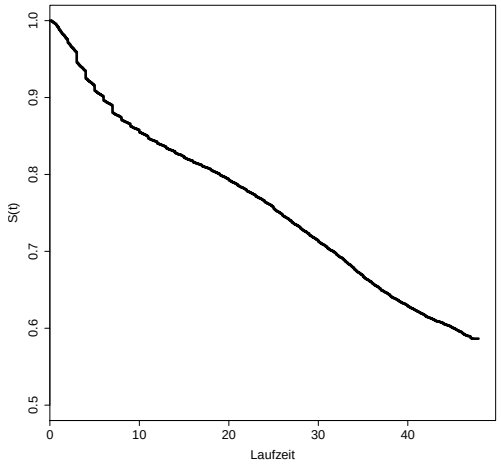

Abbildung 6.4.: Kaplan-Meier Schätzer für die Survivalfunktion der Stornierung.

Für die geschätzten Survivalfunktionen des Kaplan Meier Schätzers bezüglich des Merkmals `GRUPPE_RABATT` in Abbildung 6.6 sind große Unterschiede festzustellen. Auch der Log Rank Test mit p-Wert 0 (siehe `R`-Code 6.1) lehnt die Hypothese von identischen Survivalfunktionen ab.

Das Stornorisiko in Abbildung 6.6 für `GRUPPE` (d.h. Gruppenversicherung ohne Rabatt) ist deutlich am niedrigsten. Die Survivalfunktionen für `Rabatt` (d.h. Gruppenversicherung mit Rabatt) und `Keine Gruppe` (d.h. keine Gruppenversicherung kein Rabatt) sind bis zum 20. Vertragsjahr nahezu identisch, erst danach sind Abweichungen feststellbar, wobei `Keine Gruppe` die höhere Survivalwahrscheinlichkeit aufweist.

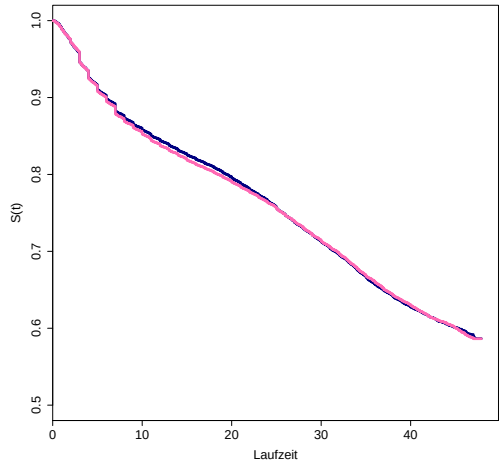

Abbildung 6.5.: Kaplan-Meier Schätzer für die Survivalfunktion der Stornierung für Frauen (pink) und Männer (blau).

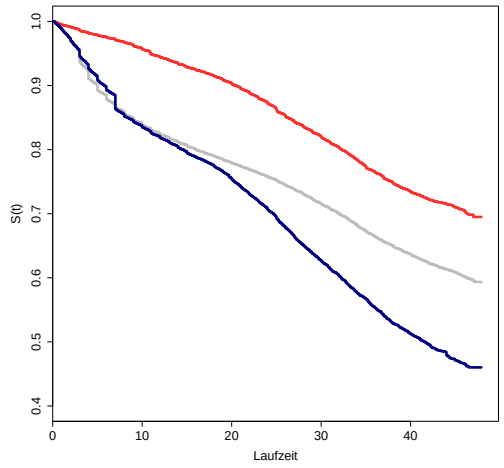

Abbildung 6.6.: Kaplan-Meier Schätzer für die Survivalfunktion der Stornierung für `GRUPPE_RABATT` mit `Gruppe` (rot) `Rabatt` (blau) und `Keine Gruppe` (grau).

R-Code 6.1: Deskriptive Analysis

```
> attach(polizzen)
> polizzen.surv <- Surv(polizzen$LAUFZEIT, polizzen$DELTA )
> polizzen.km   <- survfit(polizzen.surv~1, type = "kaplan-meier")
>
> # KAPLAN MEIER SCHAETZER
> plot(polizzen.km, lwd = 2, main = "Kaplan Meier",
+   conf.int = F, xlab = "Laufzeit", ylab = "S(t)")
>
> # SEX:   KAPLAN MEIER UND LOG RANK
> sex.km <- survfit(polizzen.surv ~ SEX, type = "kaplan-meier")
> plot(sex.km   , col = c("navy","hotpink"),
+   lwd = 3, xlab = "Laufzeit", ylab = "S(t)")
> survdiff(polizzen.surv ~ SEX)
Call:
survdiff(formula = polizzen.surv ~ SEX)
           N Observed Expected (O-E)^2/E (O-E)^2/V
SEX=M 132959    23391    23526     0.778      1.49
SEX=W 145105    26059    25924     0.706      1.49

 Chisq= 1.5  on 1 degrees of freedom, p= 0.222
>
> # GRUPPE_RABATT:   KAPLAN MEIER UND LOG RANK
> gruppe.rabatt.km <- survfit(polizzen.surv ~ GRUPPE_RABATT)
> ggsurvplot(gruppe.rabatt.km, risk.table = TRUE)
> survdiff(polizzen.surv ~ GRUPPE_RABATT)
Call:
survdiff(formula = polizzen.surv ~ GRUPPE_RABATT)
                                N Observed Expected (O-E)^2/E
                                   (O-E)^2/V
GRUPPE_RABATT=Keine Gruppe 158589    28752    27480      58.9
    133
GRUPPE_RABATT=Gruppe        34960     5568     9715    1769.9
    2255
GRUPPE_RABATT=Rabatt        84515    15130    12256     674.1
    917

 Chisq= 2582  on 2 degrees of freedom, p= 0
```

6.4. Modell

In diesem Abschnitt wird nun versucht ein geeignetes PH Modell zu finden, welches die Stornierungswahrscheinlichkeit über die Vertragslaufzeit der Polizzen ausreichend genau beschreibt. Die Auswahl der Modellklasse ist sehr entscheidend, weswegen hier nun die Hauptgründe aufgelistet werden welche zur Wahl des PH Modells führten:

- **Verteilung**: Bei der Analyse des Datensatzes konnte keine geeignete Verteilung für die Versicherungslaufzeiten gefunden werden. Dadurch ist ein PH Modell vorzuziehen, da es im Gegensatz zum AFT Modell keine Verteilungsannahme braucht.
- **Verbreitung:** Das PH Modell ist das mit Abstand meist verwendete Survivalmodell. Dementsprechend gibt es auch für dieses Modell die meiste Literatur, die meisten Artikel und auch die meisten Pakete bzw. Lösungen in `R`.
- **Funktionelle Form:** Die deskriptive Analyse hat gezeigt, dass eine nichtlineare Abhängigkeitsstruktur zwischen Stornowahrscheinlichkeit und einigen Kovariablen vorherrscht. Für PH Modelle können Kovariablen als Funktionen oder auch als glatte Funktionen sehr einfach modelliert werden.

Im ersten Schritt wird nun das PH Modell `m1` in `R`-Code 6.2 mit den gegebenen Kovariablen `BEGINN_JAHR, PRAEMIE_MONATLICH, ALTER_STORNO`, `SEX` und `GRUPPE_RABATT` an die Daten angepasst. Sämtliche Kovariablen von `m1` scheinen signifikant zu sein, was auch der Likelihood Ratio Test von `anova(m1)` bestätigt.

R-Code 6.2: Volles Modell

```
> m1 <- coxph(polizzen.surv ~ ALTER_STORNO  + PRAEMIE_MONATLICH +
+                   BEGINN_JAHR + SEX + GRUPPE_RABATT)
> m1
                          coef exp(coef)  se(coef)     z       p
ALTER_STORNO          -0.019086  0.981095  0.000437 -43.65 < 2e-16
PRAEMIE_MONATLICH     -0.001735  0.998266  0.000171 -10.18 < 2e-16
BEGINN_JAHR            0.082668  1.086181  0.000728 113.50 < 2e-16
SEXW                   0.064907  1.067060  0.009195   7.06 1.7e-12
```

```
GRUPPE_RABATTGruppe -0.172735  0.841361  0.015122   -11.42 < 2e-16
GRUPPE_RABATTRabatt -0.050369  0.950878  0.010171    -4.95 7.3e-07

Likelihood ratio test=41047  on 6 df, p=0
n= 278064, number of events= 49450
>
> anova(m1)
Terms added sequentially (first to last)
                    loglik     Chisq Df Pr(>|Chi|)
NULL               -581450
ALTER_STORNO       -569431 24038.134  1  < 2.2e-16 ***
PRAEMIE_MONATLICH  -569400    61.778  1  3.886e-15 ***
BEGINN_JAHR        -561021 16758.549  1  < 2.2e-16 ***
SEX                -560997    47.980  1  4.306e-12 ***
GRUPPE_RABATT      -560927   140.377  2  < 2.2e-16 ***
```

Interpretation von `m1` bzw. `anova(m1)`:

- Anhand der χ^2-Werte 24038 und 16759 in `anova(m1)` für `ALTER_STORNO` und `BEGINN_JAHR` sieht man, dass diese beiden Kovariablen den mit Abstand größten Einfluss bzw. Effekt auf die Vertragslaufzeit der Polizzen haben.

- Der Koeffizient von `PRAEMIE_MONATLICH` mit Wert -0.0017 wirkt sehr klein. Für `PRAEMIE_MONATLICH = 100` beträgt der Effekt auf die Hazardfunktion jedoch $\exp(-0.0017 * 100) = 0.84$. D.h. das Stornorisiko für eine monatliche Prämie von 100 ist um ca. 16% niedriger, verglichen mit einer monatlichen Prämien von 0.

- Das Modell `m1` besagt, dass das Stornorisiko für Frauen (`W`) um 6.7% Prozent höher ist als für Männer (`M`), wegen $\exp(0.064) = 1.067$.

- Weiters besagt das Modell `m1`, dass das Stornorisiko für `Gruppe` ca. 16% niedriger (wegen $\exp(-0.17) = 0.84$) und für `Rabatt` ca. 5% niedriger (wegen $\exp(-0.05) = 0.95$) ist, im Vergleich zu `Keine Gruppe`.

- Allgemein scheinen alle Kovariablen signifikant zu sein. Man bedenke jedoch, dass die χ^2-Werte in `anova(m1)` von `PRAEMIE_MONATLICH`,

SEX und GRUPPE_RABATT mit 61.7, 47.9 und 140.3 sehr niedrig sind im Verhältnis zu Log-Likelihodfunktionswerten von <-500000.

Das Modell m1 sieht nun auf den ersten Blick recht zufriedenstellend aus. Jedoch ist die Annahme, dass sich die stetigen Kovariablen PRAEMIE_MONATLICH, BEGINN_JAHR und ALTER_STORNO linear auf die Hazardfunktion der Vertragslaufzeiten auswirken zu hinterfragen. Aus diesem Grund wird die funktionale Form für diese drei Kovariablen im nächsten Abschnitt näher studiert.

6.4.1. Überprüfung der funktionalen Formen

Bisher wurde angenommen, dass sich im Modell die stetigen Kovariablen PRAEMIE_MONATLICH, BEGINN_JAHR und ALTER_STORNO linear auf die Hazardfunktion der Vertragslaufzeiten auswirken. Die Abbildungen 6.1, 6.2 und 6.3 aus Abschnitt 6.2 geben jedoch dazu Anlass, die funktionale Form der Kovariablen PRAEMIE_MONATLICH, BEGINN_JAHR und ALTER_STORNO genauer zu untersuchen. Es wird nämlich vermutet, dass eine nichtlineare Abhängigkeitstruktur für diese Kovariablen vorherrscht.

Zur Überprüfung der funktionalen Form wird jeweils ein PH Modell geschätzt, wobei hier eine der drei stetigen Kovariablen mittels glatter Funktionen durch kubische Splinefunktionen `pspline(, degree = 3)` modelliert wird. Danach wird mittels Likelihood Ratio Test überprüft, ob das neue geschätzte Modell signifikant besser ist, als das Modell m1. Zusätzlich wird auch noch graphisch überprüft, ob die resultierende funktionale Form der Kovariablen sich von einer linearen Form unterscheidet.

Bemerkung 16.
Glatte Funktionen von Kovariablen können in R mithilfe von `pspline()` modelliert werden.

ALTER_STORNO
Als erstes wird nun die funktionale Form der Kovariable ALTER_STORNO unter die Lupe genommen. Bereits die Abbildung 6.3 der deskriptiven Analyse deutete daraufhin, dass eine lineare Form der Kovariable ALTER_STORNO nicht angebracht sein könnte. Sowohl der Plot der geschätzten funktionellen Form in Abbdilung 6.7 sowie der Likelihood Ratio Test in R-Code 6.3 mit p-Wert 2.2e-16 lehnen eine lineare Form der Kovariable ALTER_STORNO ab.

R-Code 6.3: Funktionelle Form für Alter

```
# ALTER_STORNO
> m.alter_spline <- coxph(polizzen.surv ~  BEGINN_JAHR +
+           PRAEMIE_MONATLICH + GRUPPE_RABATT + SEX +
+           pspline(ALTER_STORNO, df = 0, caic = T, degree = 3))
> anova(m.alter_spline, m1)
Analysis of Deviance Table
   loglik  Chisq     Df P(>|Chi|)
1 -559032
2 -560901 3736.7 15.574 < 2.2e-16 ***
> termplot(m.alter_spline, se=T, terms=5, ylabs="Log hazard",
+         col.term=1, col.se=1)
```

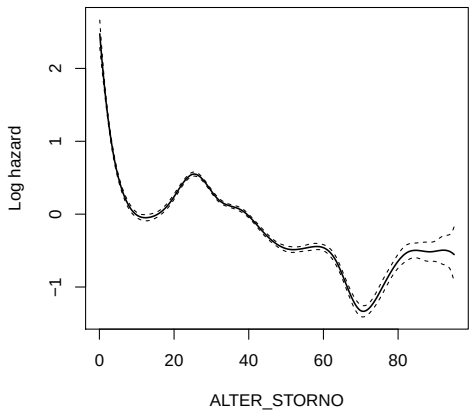

Abbildung 6.7.: Geschätzte funktionale Form für die Kovariable ALTER_STORNO.

PRAEMIE_MONATLICH

Weiters wird nun die funktionale Form der Kovariable PRAEMIE_MONATLICH untersucht. Auch hier spricht die deskriptive Analyse in Abbildung 6.2 gegen eine lineare Form der Kovariable. Diese Vermutung wird bestätigt, denn der Plot der geschätzten funktionellen Form in Abbildung 6.8 sowie der Likelihood Ratio Test in R-Code 6.4 mit p-Wert 2.2e-16 lehnen eine lineare Form der Kovariable PRAEMIE_MONATLICH ab.

R-Code 6.4: Funktionelle Form für monatliche Prämie

```
1
2  > #PRAEMIE_MONATLICH
3  > m.preamie_spline <- coxph(polizzen.surv ~  BEGINN_JAHR   +
4  +     ALTER_STORNO + GRUPPE_RABATT + SEX +
5  +     pspline(PRAEMIE_MONATLICH, df = 0, caic = T, degree = 3))
6  > anova(m.preamie_spline, m1)
7  Analysis of Deviance Table
8     loglik  Chisq     Df P(>|Chi|)
9  1 -560486
10 2 -560901 830.16 12.704 < 2.2e-16 ***
11 > termplot(m.preamie_spline, se=T, terms=5,
12 +     ylabs="Log hazard", col.term=1, col.se=1)
```

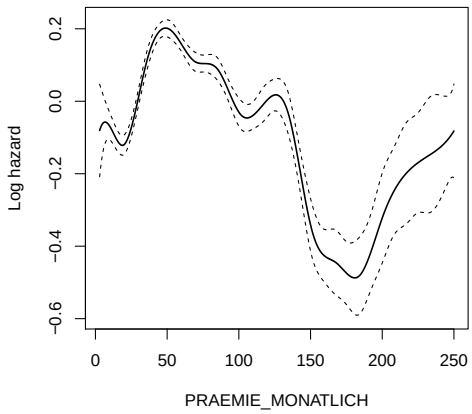

Abbildung 6.8.: Geschätzte funktionale Form für die Kovariable PRAEMIE_MONATLICH.

BEGINN_JAHR

Abschließend wird auch noch die funktionelle Form des Kovariable Beginnjahr überprüft. Bei der Betrachtung der Abbildung 6.1 in der deskriptiven Analyse ist auch hier die lineare Form fraglich. Sowohl der Plot der geschätzten funktionellen Form in Abbdilung 6.9 sowie der Likelihood Ratio Test in

R-Code 6.5 mit p-Wert 2.2e-16, lehnen hier eine lineare Form der Kovariable BEGINN_JAHR ab.

R-Code 6.5: Funktionelle Form für Beginnjahr

```
>  #BEGINN_JAHR
> m.beginn_jahr <- coxph(polizzen.surv ~  PRAEMIE_MONATLICH  +
+       ALTER_STORNO + SEX + GRUPPE_RABATT +
+       pspline(BEGINN_JAHR, df = 0, caic = T, degree = 3))
> anova(m.beginn_jahr, m1)
Analysis of Deviance Table
   loglik  Chisq     Df P(>|Chi|)
1 -558309
2 -560901 5183.9 13.783 < 2.2e-16 ***
> termplot(m.beginn_jahr, se=T, terms=5,
+       ylabs="Log hazard", col.term=1, col.se=1)
```

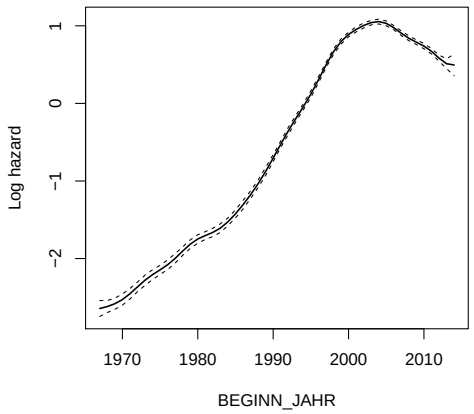

Abbildung 6.9.: Geschätzte funktionale Form für die Kovariable BEGINN_JAHR.

Zusammenfassend kann gesagt werden, dass der Effekt der stetigen Kovariablen PRAEMIE_MONATLICH, BEGINN_JAHR und ALTER_STORNO auf die Hazardfunktion als nicht-linear angesehen werden muss. Als Folge dessen werden

diese Kovariablen ab sofort nur mehr mithilfe von glatten Funktionen modelliert.

6.4.2. Modell Auswahl

Es wird nun ein PH Modell, bezeichnet mit `m2` (siehe R-Code 6.6), an die Daten angepasst bei dem die Kovariablen `PRAEMIE_MONATLICH`, `BEGINN_JAHR` und `ALTER_STORNO` mittels glatter Funktionen modelliert werden.

R-Code 6.6: PH Modell mit Splines

```
> m2 <- coxph(polizzen.surv ~  pspline(ALTER_STORNO, df = 0,
+                              caic = T, degree = 3) +
+                              pspline(PRAEMIE_MONATLICH, df = 0,
+                              caic = T, degree = 3) +
+                              pspline(BEGINN_JAHR, df = 0,
+                              caic = T, degree = 3) +
+                              GRUPPE_RABATT + SEX)
> m2
                                coef  se(coef)     Chisq   DF       p
pspline(ALTER_STORNO, df  -1.92e-02  4.72e-04  1.66e+03  1.0 < 2e-16
pspline(ALTER_STORNO, df                        9.94e+03 15.0 < 2e-16
pspline(PRAEMIE_MONATLICH -1.27e-03  1.94e-04  4.31e+01  1.0 5.3e-11
pspline(PRAEMIE_MONATLICH                       3.41e+02 12.4 < 2e-16
pspline(BEGINN_JAHR, df =  8.11e-02  8.28e-04  9.59e+03  1.0 < 2e-16
pspline(BEGINN_JAHR, df =                       9.35e+03 16.0 < 2e-16
GRUPPE_RABATTGruppe       -7.88e-02  1.54e-02  2.61e+01  1.0 3.3e-07
GRUPPE_RABATTRabatt       -9.08e-02  1.02e-02  7.87e+01  1.0 < 2e-16
SEXW                       4.01e-02  9.37e-03  1.83e+01  1.0 1.9e-05

Degrees of freedom for terms= 16.0 13.4 17.0  2.0  1.0
Likelihood ratio test=54696  on 49.4 df, p=0  n= 276829
>
anova(m2)
Analysis of Deviance Table
Terms added sequentially (first to last)

                          loglik      Chisq      Df Pr(>|Chi|)
```

```
NULL                        -581425
pspline(ALTER_STORNO,     -567752 27346.1875 17.0000  < 2.2e-16 ***
pspline(PRAEMIE_MONATLICH-567598   308.5395 17.0000  < 2.2e-16 ***
pspline(BEGINN_JAHR       -554129 26945.6121 16.0000  < 2.2e-16 ***
GRUPPE_RABATT             -554086    87.0249  2.0000  < 2.2e-16 ***
SEX                       -554077     18.299  0.9984 1.881e-05
```

Anhand von `anova(m2)` aus R-Code 6.6 ist ersichtlich, dass durch Hinzunahme von `SEX` in das Modell die Deviance um 18.3 abnimmt und die Likelihood Ratio Teststatistik einen p-Wert von 1.881e-05 für $H_0 : \beta_{SEX} = 0$ liefert (siehe (5.18)). Die Kovariable `SEX` scheint nicht relevant zu sein, da eine Reduktion der Deviance von nur 18.3 in Relation zu den Log-Likelihoodfunktionswerten von <-500000 als zu gering angesehen wird.

Als Konsequenz wurde die Kovariable `SEX` aus dem finalen Modell entfernt. Die Tatsache, dass `SEX` nicht signifikant ist, ist wenig überraschend wenn man bedenkt, dass bereits der Log-Rank Test keine Unterschiede in den Survivalkurven für `SEX` festmachen konnte (siehe Abschnitt 6.3) und auch die Plots der deskriptiven Analyse keinen Unterschied für `SEX` lieferten (siehe Abschnitt 6.2).

Als nächstes wäre noch interessant, ob eine signifikante Interaktion von Gruppe/Rabatt mit Stornoalter, monatlicher Prämie und Beginnjahr besteht. Leider unterstützt die Funktion `coxph()` in `R` keine Interaktionen im Zusammenhang mit `pspline()`. Womit die Modellierung und Überprüfung dieser Interaktionen hinfällig wird.

Folglich beinhaltet nun das finale Modell `m3` zur Modellierung der Vertragslaufzeiten von Krankenversicherungen die Kovariablen für Stornoalter, monatliche Prämie, Beginnjahr und Gruppe/Rabatt, wobei jeweils Stornoalter, monatliche Prämie und Beginnjahr mittels glatter Funktionen modelliert werden.

R-Code 6.7: Finale Modell

```
> m3 <- coxph(polizzen.surv ~ pspline(ALTER_STORNO, df = 0,
                              caic = T, degree = 3) +
                              pspline(PRAEMIE_MONATLICH, df = 0,
                              caic = T, degree = 3) +
```

```
                              pspline(BEGINN_JAHR, df = 0, caic = T,
                              degree = 3) + GRUPPE_RABATT)
> m3
                                coef  se(coef)     Chisq   DF       p
pspline(ALTER_STORNO, df  -1.98e-02  4.66e-04  1.80e+03  1.0 < 2e-16
pspline(ALTER_STORNO, df                       9.98e+03 15.0 < 2e-16
pspline(PRAEMIE_MONATLICH -1.02e-03  1.89e-04  2.94e+01  1.0 6.0e-08
pspline(PRAEMIE_MONATLICH                      3.54e+02 12.4 < 2e-16
pspline(BEGINN_JAHR, df =  8.12e-02  8.28e-04  9.61e+03  1.0 < 2e-16
pspline(BEGINN_JAHR, df =                      9.36e+03 16.0 < 2e-16
GRUPPE_RABATTGruppe       -7.81e-02  1.54e-02  2.56e+01  1.0 4.2e-07
GRUPPE_RABATTRabatt       -8.91e-02  1.02e-02  7.59e+01  1.0 < 2e-16
```

Nachdem nun das finale Modell ausgewählt wurde, muss nun im kommenden Abschnitt die proportionale Hazard Annahme und die Modell Anpassung eingehend überprüft werden.

6.4.3. Überprüfung der PH Annahme

In diesem Abschnitt wird die PH Annahme für die faktoriellen Kovariable `GRUPPE_RABATT` untersucht. Dazu wird die graphische Methode aus dem Abschnitt 5.6.4 herangezogen. Es werden daher die logarithmierten Kaplan-Meier Survivalkurven daraufhin überprüft, ob sie parallel zueinander laufen.

Auf Basis der Abbildung 6.10 bleibt die proportionale Hazard Annahme für die Kovariable `GRUPPE_RABATT` aufrecht. Obwohl der Graph für `Keine Gruppe` zwischen den beiden anderen Graphen hin und her pendelt, ist dies nicht ausreichend um die PH Annahme abzulehnen.

6.4.4. Modellanpassung

In diesem Abschnitt wird die Modellanpassung für das PH Modell `m3` studiert. Zu diesem Zweck werden die beiden auf Cox Snell Residuen und Deviance Residuen basierenden Methoden aus Abschnit 5.6.2 auf das Modell angewendet.

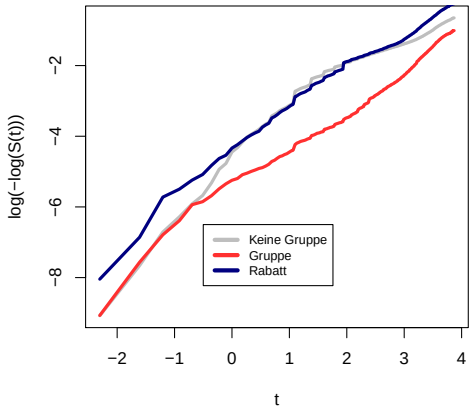

Abbildung 6.10.: Überprüfung der PH Annahme für die Kovariable `GRUPPE_RABATT` mit `Gruppe` (rot), `Rabatt` (blau) und `Keine Gruppe` (grau).

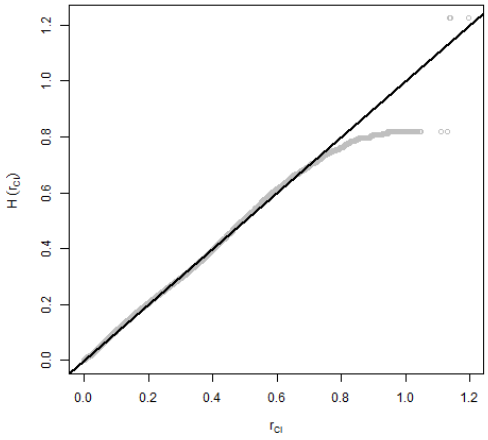

Abbildung 6.11.: Plot der Cox Snell Residuen r_{Ci} des Modells `m3` .

Bei einer ausreichend guten Modellanpassung eines Modells sollte der Plot der Cox Snell Residuen r_{Ci} gegen $\widehat{H}(r_{Ci})$ einer Geraden mit Steigung 1 gleichen.

In Abbildung 6.11 wurde dieser Plot für die Cox Snell Residuen von Modell `m3` ausgeführt. Die Abbildung zeigt, dass r_{Ci} gegen $\widehat{H}(r_{Ci})$ der Geraden mit Steigung 1 gleicht, womit eine ausreichend gute Modellanpassung für das Modell `m3` gewährleistet ist.

Bei der zweiten Methode zur Überprüfung der Modellanpassung werden die Deviance Residuen r_{Di} gegen den linearen Prädiktor $\widehat{\boldsymbol{\beta}}^{\intercal}\boldsymbol{x}_i$ aufgetragen. Falls das Modell ausreichend gut an die Daten angepasst ist, sollten keine Ausreißer erkennbar sein. Bei der Betrachtung der Abbildung 6.12 sind keinerlei Ausreißer zu erkennen. Weiters ist zu beobachten, dass ein Großteil der Deviance Residuen um 0 liegen. Der Grund dafür liegt in der großen Anzahl an zensierten Versicherungspolizzen.

In Summe kann durch den Deviance Plot aus Abbildung 6.12 auf eine ausreichend guten Anpassung des Modells `m3` an die Versicherungspolizzen geschlossen werden.

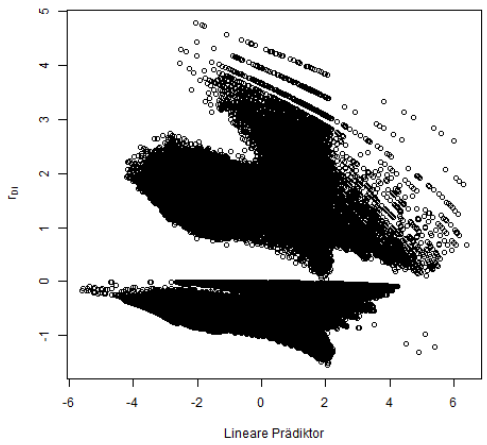

Abbildung 6.12.: Plot der Deviance Residuen r_{Di} gegen den linearen Prädiktor $\widehat{\boldsymbol{\beta}}^{\intercal}\boldsymbol{x}_i$ des Modells `m3` .

7. Modellauswertung

Im vorangegangen Abschnitt wurde die Wahl des Proportional Hazard (PH) Modell `m3` zur Modellierung der Vertragslaufzeit von Krankenversicherungspolizzen motiviert. In diesem Modell werden die drei stetigen Prädiktoren des Alters (`ALTER_STORNO`), der monatlichen Prämie (`PRAEMIE_MONATLICH`) und dem Beginnjahr der Versicherungspolizze (`BEGINN_JAHR`) mithilfe von glatten Funktionen modelliert. Zusätzlich enthält das Modell noch einen Faktor (`GRUPPE_RABATT`) welcher angibt, ob eine Gruppenversicherung, keine Gruppenversicherung oder eine rabattierte Polizze vorliegt.

```
1 > m3 <- coxph(polizzen.surv ~ pspline(ALTER_STORNO, df = 0,
2                               caic = T, degree = 3) +
3                               pspline(PRAEMIE_MONATLICH, df = 0,
4                               caic = T, degree = 3) +
5                               pspline(BEGINN_JAHR, df = 0, caic = T,
6                               degree = 3) + GRUPPE_RABATT)
```

Es wird nun der Einfluss der Prädiktoren im Modell auf die Vertragslaufzeit eingehend studiert. Dafür wird die durch das Modell geschätzte Survivalfunktion $S(t)$, Tabellen mit den Quantilen der Survivalfunktion sowie die geschätzten funktionalen Formen der Prädiktoren betrachtet.

Bemerkung 17.
Mit der Survivalfunktion $S(t)$ wird in diesem Zusammenhang die Wahrscheinlichkeit bezeichnet, nach t Jahren Vertragslaufzeit die Polizze nicht storniert zu haben.

7.1. Einfluss des Alters

Um den Effekt des Alters (`ALTER_STORNO`) zu studieren, wird die Survivalfunktion für verschieden Werte ausgewertet, wobei die verbleibenden Prädiktoren fixiert werden. Anhand der Abbildung 7.1 erkennt man, dass allgemein mit zunehmenden Alter das Stornorisiko abnimmt. Weiters wurden in der Tabelle 7.1 für die Survivalkurven aus Abbildung 7.1 die 25%, 50% und 75% Quantile aufgelistet.

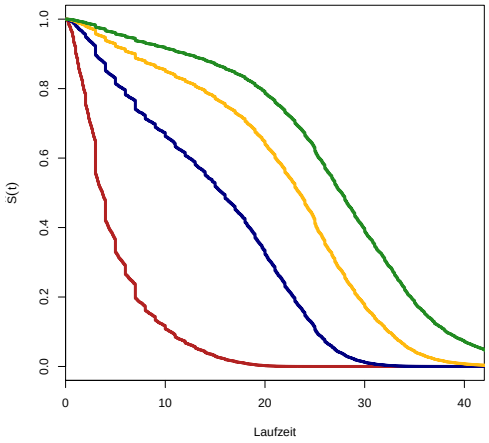

Abbildung 7.1.: Geschätzte Survivalfunktion in Abhängigkeit von `ALTER_STORNO` für die Werte 3 (rot), 20 (blau), 50 (gelb) und 70 (grün).

Mithilfe des Plots der glatten Funktion des Alters (`ALTER_STORNO`) wird nun der direkte Effekt auf die logarithmierte Hazardfunktion $h_0(t)$ der Vertragslaufzeit dargestellt. Anhand der Abbildung 7.2 sieht man, dass um die Geburt das Stornorisiko am höchsten ist. Weiters ist feststellbar, dass das Stornorisiko mit zunehmendem Alter abnehmend ist. Jedoch sind um 20, 60 und 70 Jahren deutliche Anstiege in der Funktion und somit im Stornorisiko feststellbar. Zudem erkennt man, dass die Konfidenzintervalle für die glatten Funktion des Alters ab 80 immer breiter werden. Der Grund darin liegt in der immer kleiner werdenden Anzahl an Versicherten mit zunehmendem Alter.

	Quantile		
Merkmal	0.25	0.50	0.75
`ALTER_STORNO = 3`	2.2	4	7
`ALTER_STORNO = 20`	7.1	16.1	22
`ALTER_STORNO = 50`	16.7	23.9	28.7
`ALTER_STORNO = 70`	22.2	28.3	34.2

Tabelle 7.1.: Quantile der Survivalfunktion unter dem Modell `m3` für verschiedene Werte von `ALTER_STORNO`.

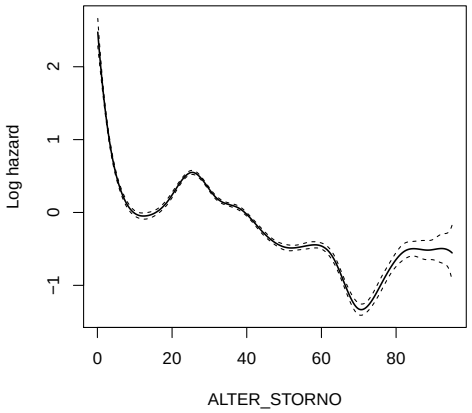

Abbildung 7.2.: Glatte Funktion für die Kovariable `ALTER_STORNO`.

7.2. Einfluss von Beginnjahr

Der Effekt des Beginnjahres wird studiert, in dem die Survivalfunktion für verschiedene Werte ausgewertet wird (siehe Abbildung 7.3), wobei die restlichen Prädiktoren fixiert werden. Zusätzlich wird auch die glatte Funktion des Prädiktors betrachtet (siehe Abbildung 7.4) um den direkten Effekt auf die logarithmierte Hazardfunktion $h_0(t)$ beobachten zu können.

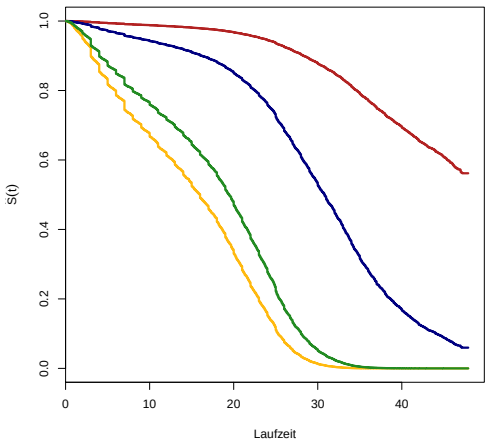

Abbildung 7.3.: Geschätzte Survivalfunktion in Abhängigkeit von `BEGINN_JAHR` für die Jahre 1975 (rot), 1990 (blau), 2000 (gelb) und 2010 (grün).

	Quantile		
Merkmal	0.25	0.50	0.75
`BEGINN_JAHR = 1975`	34.2	45.5	
`BEGINN_JAHR = 1990`	24.3	30.8	36.9
`BEGINN_JAHR = 2000`	7	16	21.9
`BEGINN_JAHR = 2010`	10.7	19.5	24.7

Tabelle 7.2.: Quantile der Survivalfunktion unter dem Modell `m3` für verschiedene Werte von `BEGINN_JAHR`.

Die Tabelle 7.1 zeigt wie stark die Vertragslaufzeiten für verschiedene Beginnjahre abnehmen kann. So haben 50% aller Polizzen mit Beginnjahr = 1975 eine Vertragslaufzeit $\geq$ 45.5 Jahren im Vergleich zu $\geq$ 16 Jahren bei Beginnjahr = 2000.

Sowohl die Abbildung 7.3 als auch die Abbildung 7.4 zeigen, dass das Stornorisiko bis zum Jahr 2001 stetig ansteigt und danach bis zum Ende der

Beobachtung 2014 wieder abfällt.

Generell ist für Polizzen mit Beginnjahr 1968 das Stornorisiko am niedrigsten. Eine Erklärung dafür liegt in der speziellen Art des Datensatzes, welcher zum einem alle abgeschlossenen Polizzen im Zeitraum 2001 bis 2014 beobachtet und zum anderem den Bestand an Polizzen zu Beginn des Jahres 2001 beinhaltet.

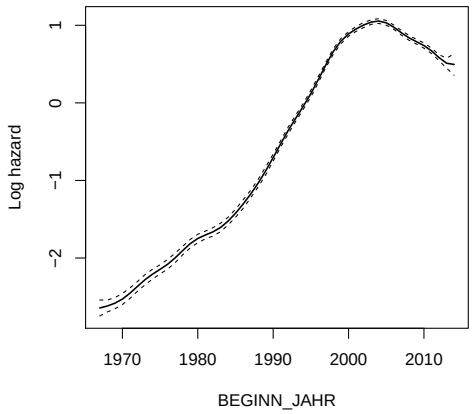

Abbildung 7.4.: Glatte Funktion für die Kovariable `BEGINN_JAHR`.

Der Anstieg der glatten Funktion kann wie folgt erklärt werden: Betrachtet man eine Polizze mit `BEGINN_JAHR = 1968`, dann hat diese Polizze bereits 33 Jahre Vertragslaufzeit bis zum Beginn der Untersuchung und somit ein geringes Restrisiko in der verbleibenden Zeit storniert zu werden. Im Gegensatz dazu hat eine Polizze mit `BEGINN_JAHR = 2001` ein viel höheres Stornorisiko, da gerade in den ersten Vertragsjahren die Versicherungspolizze vermehrt storniert wird.

Der Abfall der glatten Funktion ab 2001 kann wie folgt erklärt werden: Betrachtet man eine Polizze mit `BEGINN_JAHR = 2001`, dann kann die Polizze in den Jahren bis 2014 storniert werden. Es besteht somit die Möglichkeit die Polizze 13 Jahre lang zu stornieren. Für eine Polizze mit `BEGINN_JAHR = 2012` besteht die Möglichkeit nur 2 Jahre lang zu stornieren. Dadurch ist das Stornorisiko für Polizzen mit `BEGINN_JAHR = 2001` deutlich höher

als für Polizzen mit `BEGINN_JAHR = 2012` bzw. allgemein für Polizzen mit `BEGINN_JAHR > 2001`.

7.3. Einfluss der monatlichen Prämie

Um den Effekt der monatlichen Prämie (`PRAEMIE_MONATLICH`) auf die Vertragslaufzeit zu studieren, wird der Plot der geschätzten Survivalfunktion (siehe Abbildung 7.5) und der Plot der glatten Funktion (siehe Abbildung 7.6) betrachtet. Weiters werden in der Tabelle 7.3 für die Survivalkurven aus Abbildung 7.5 die 25%, 50% und 75% Quantile aufgelistet.

Anhand Abbildung 7.6 erkennt man, dass das höchste Risiko zu stornieren bei einer monatlichen Prämie von 60 auftritt. Zudem ist ein Auf- und Absteigen der glatten Funktion für die monatliche Prämie zu beobachten. Dieses Verhalten spiegelt sich in auch den geschätzten Survivalkurven in Abbildung 7.5 wieder. Weiters ist zu beobachten, dass das Band der Konfidenzintervalle für sehr hohe Prämien (>150) und sehr kleine Prämien (<5) immer breiter wird, da sowohl sehr hohe, wie sehr niedrige Prämien weniger oft vorkommen.

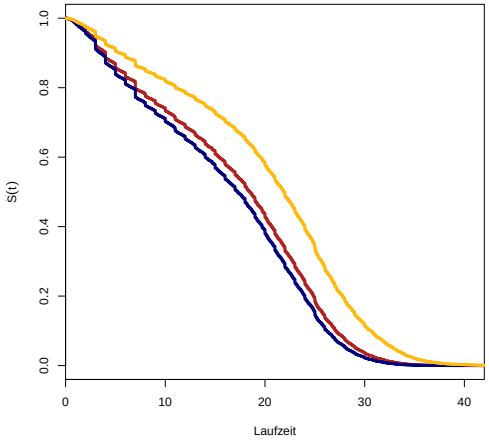

Abbildung 7.5.: Geschätzte Survivalfunktion unter `PRAEMIE_MONATLICH` für die Prämien 25€ (rot), 50€ (blau) und 150€ (gelb).

Der Effekt der monatlichen Prämie auf die Survivalfunktion bzw. Vertragslaufzeit in der Tabelle 7.3 ist um einiges geringer als der Effekt von Beginnjahr oder Alter (vergleiche Tabelle 7.1 und Tabelle 7.2). Da sich die Werte der Quantilie für unterschiedliche Werte von `PRAEMIE_MONATLICH` nicht so stark ändern.

	Quantile		
Merkmal	0.25	0.50	0.75
`PRAEMIE_MONATLICH = 25`	9.3	18.5	23.8
`PRAEMIE_MONATLICH = 50`	8	17.3	22.9
`PRAEMIE_MONATLICH = 150`	14	21.9	26.7

Tabelle 7.3.: Quantile der Survivalfunktion unter dem Modell `m3` für verschiedene Werte von `PRAEMIE_MONATLICH`.

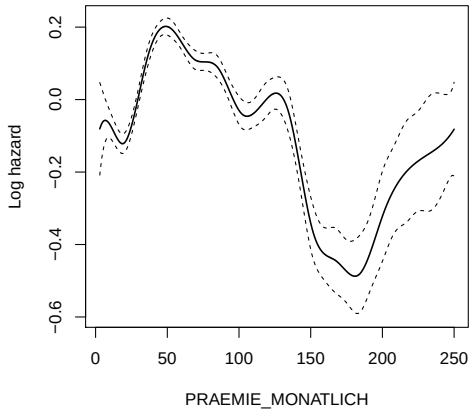

Abbildung 7.6.: Glatte Funktion für die Kovariable `PRAEMIE_MONATLICH`.

7.4. Einfluss von Gruppenversicherung und Rabatt

In diesem Abschnitt wird der Einfluss von Gruppenversicherung und Rabatt auf die Vertragslaufzeit besprochen. Eine Polizze kann nur dann einen Rabatt erhalten wenn sie Teil einer Gruppenversicherung ist. Mithilfe des dreistufigen Faktors `GRUPPE_RABATT` wurde die Beziehung zwischen Gruppenversicherung und Rabatt kodiert.

Im Modell `m3` wurde der Unterschied von `Gruppe` und `Rabatt` zu `Keine Gruppe` modelliert (siehe Tabelle 7.4). Allgemein gilt, dass das Stornorisiko für `Gruppe` und `Rabatt` im Vergleich zu `Keine Gruppe` abnimmt.

Für nur gruppenversicherte Polizzen (`Gruppe`) besagt das Modell, dass das Risiko nach t Jahren die Versicherungspolizze zu stornieren 0.92-fach bzw. 8% kleiner ist als im Vergleich zu Polizzen die nicht Teil einer Gruppenversicherung sind (siehe Tabelle 7.4).

Weiters besagt das Modell für rabattierte Polizze (`Rabatt`), dass das Risiko nach t Jahren die Versicherungspolizze zu stornieren 0.91-fach bzw. 9% kleiner ist als im Vergleich zu Polizzen die nicht Teil einer Gruppenversicherung sind (siehe Tabelle 7.4).

	Koeffizient	exp(Koeffizient)
`Gruppe`	-0.078	0.924
`Rabatt`	-0.089	0.914

Tabelle 7.4.: Koeffizienten der Kovariable `GRUPPE_RABATT`.

In Abbildung 7.7 sind nochmals die geschätzten Survivalkurven in Abähngigkeit von `GRUPPE_RABATT` abgebildet. Die Abbildung zeigt, dass das Stornorisiko für `Gruppe` sowie `Rabatt` im Vergleich zu `Keine Gruppe` abnimmt. Weiters erkennt man in der Abbildung auch, dass die Survivalkurven für `Gruppe` sowie `Rabatt` praktisch identisch sind und somit auch das Stornorisiko.

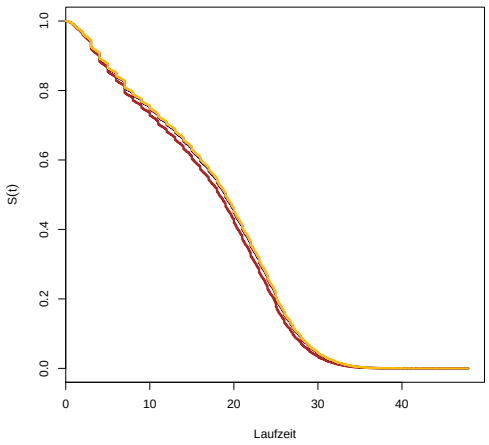

Abbildung 7.7.: Geschätzte Survivalfunktion unter `GRUPPE_RABATT` für `Keine Gruppe` (rot), `Gruppe` (blau) und 150 € (gelb).

8. Conclusio

Das Ziel dieser Arbeit war es, die Vertragslaufzeit bis zur Stornierung von Versicherungsnehmern in der privaten Krankenversicherung zu modellieren. Hierbei sollten Eigenschaften des Versicherungsnehmers gefunden werden, von denen die Versicherungslaufzeit abhängt. Der zu analysierende Datensatz umfasst sämtliche im Zeitraum 2001 bis 2014 abgeschlossenen und stornierten Versicherungspolizzen. Zusätzlich beinhaltet der Datensatz auch den Gesamtbestand an Versicherungspolizzen zum Zeitpunkt 1. Januar 2014. Dadurch konnten Versicherungsverträge analysiert werden, deren Beginnjahr bis ins Jahr 1967 reichen. Weiters standen zu jeder Polizze Informationen zum Versicherungsnehmer sowie zur Polizze selbst zur Verfügung. Die Informationen beinhalteten Alter und Geschlecht des Versicherungsnehmers sowie die monatliche Prämie, das Beginnjahr, die Laufzeit, die Gruppenzugehörigkeit und etwaige Rabattierung der Polizze.

Der Zugang um die Vertragslaufzeit modellieren zu können, erfolgte über die Theorie der Survival Analysis. Denn im Gegensatz zu Linearen oder Generalisiert Linearen Modellen, welche Größen wie Wahrscheinlichkeiten, Anzahlen etc. modellieren, beschäftigt sich die Survival Analysis mit der Analyse und Modellierung von Zeitdauern. Die Idee die hinter dieser Arbeit steckt war, die Vertragslaufzeit mittels eines Regressionsmodells der Survival Analysis (kurz Survivalmodell) zu modellieren.

Bei der Auswahl eines geeigneten Survivalmodells zur Modellierung der Vertragslaufzeiten fiel die Wahl auf das Proportional Hazard Modell. Dieses Modell zeichnet sich dadurch aus, dass einerseits keinerlei Verteilungsannahmen für die Vertragslaufzeiten getroffen werden müssen und andererseits das Modell sehr einfach zu interpretieren ist.

Die Modellauswahl ergab, dass das Geschlecht wohl keinen Einfluss auf die Vertragslaufzeit hat. Des weiteren konnte festgestellt werden, dass eine nicht-

lineare Abhängigkeitsstruktur zwischen der Vertragslaufzeit und einigen Prädiktoren besteht. Daraufhin wurden diese Prädiktoren durch stetige glatte Funktionen, welche durch kubische Splinefunktionen erzeugt werden, modelliert.

Das finale Modell beinhaltet schlussendlich die durch stetige glatte Funktionen modellierten Prädiktoren Alter, monatliche Prämie und Beginnjahr. Weiters enthält das Modell auch noch einen dreistufigen Faktor für die Beziehung zwischen Gruppenversicherung und Rabatt.

Durch das Modell konnten die Vertragslaufzeit bis zur Stornierung und die Effekte darauf ausgewertet werden. Es konnte festgestellt werden, dass die Vertragslaufzeit von Polizzen welche innerhalb einer Gruppenversicherung sind oder einen Rabatt erhalten im Schnitt später storniert werden als Polizzen die nicht Teil einer Gruppenversicherung sind. Bei Betrachtung der geschätzten glatten Funktionen konnten Muster entdeckt werden, wie die Vertragslaufzeit von Alter, monatlicher Prämie und Beginnjahr abhängt. Für das Alter konnte hierbei eine etwas komplexere Abhängigkeitsstruktur ausgemacht werden. In den ersten Lebensjahren sowie um die zwanzigsten Lebensjahre konnte ein erhöhtes Stornorisiko und dadurch verkürzte Vertragslaufzeiten festgestellt werden. Generell nimmt jedoch das Stornorisiko mit zunehmendem Alter ab und somit die Vertragslaufzeit zu. Für monatliche Prämie wird das höchste Stornorisiko bei einer Prämie von 60 € beobachtet und für Beginnjahr das niedrigste Stornorisiko im Jahr 1968.

Generell bietet die Modellierung der Vertragslaufzeit durch ein PH Modell ein sehr leicht zu interpretierendes Modell der Abhängigkeiten. Zudem kann durch die Modellierung der Vertragslaufzeit, die Wahrscheinlichkeit zu stornieren über den Verlauf der Vertragsjahre gewonnen werden.

A. Anhang

A.1. Definitionen und Sätze

Satz A.1.1 (Delta Methode).
Sei T_n eine Folge von Zufallsvariablen die

$$\sqrt{n}\,(T_n - \theta) \xrightarrow{D} N(0, \sigma^2)$$

genügt. Sei g eine reellwertige Funktion mit $g'(\theta) \neq 0$. Dann gilt

$$\sqrt{n}\,(g(T_n) - g(\theta)) \xrightarrow{D} N\left(0, \sigma^2 g'(\theta)^2\right) .$$

Falls g zweimal stetig differenzierbar mit $g'(\theta) = 0$ und $g''(\theta) \neq 0$, dann gilt

$$n\,(g(T_n) - g(\theta)) \xrightarrow{D} \sigma^2 \frac{g''(\theta)}{2} \chi_1^2 .$$

Definition A.1.1 (Hypergeometrische Verteilung).
Eine Zufallsvariable X ist hypergeometrisch verteilt, falls die Wahrscheinlichkeitsfunktion die Form

$$\Pr(X = k; N, n, M) = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}} \quad k = 0, 1, \ldots, n \tag{A.1}$$

hat, mit der Populationsgröße N, der Gesamtanzahl an vorhandenen Erfolgen in der Population M, der Anzahl der Ziehungen ohne Zurücklegen n und der Anzahl an beobachteten Erfolgen k.

Satz A.1.2 (Momente der hypergeometrischen Verteilung).
Sei X eine ***hypergeometrisch verteilte Zufallsvariable*** *wie in Definition A.1.1, dann gilt*

$$\mathrm{E}[X] = \frac{nM}{N} \tag{A.2}$$

und

$$\mathrm{Var}[X] = \frac{nM(N-M)(N-n)}{N^2(N-1)}\,. \tag{A.3}$$

Definition A.1.2 (Multivariate hypergeometrische Verteilung).
Seien X_i hypergeometrisch verteilt für $i = 1, \ldots, n$. Dann ist die gemeinsame Wahrscheinlichkeitsfunktion von $\boldsymbol{X} = (X_1, \ldots, X_n)$ mit den Parametern $(M_1, \ldots, M_n)$

$$\Pr(X_1 = k_1, X_2 = k_2, \ldots, X_n = k_n) = \frac{\binom{M_1}{k_1} \cdots \binom{M_n}{k_n}}{\binom{N}{k}}\,, \tag{A.4}$$

mit $M_1 + M_2 + \cdots + M_n = N$ und $k_1 + k_2 + \cdots + k_n = k$. Man nennt $\boldsymbol{X}$ multivariat hypergeometrisch verteilt.

Satz A.1.3 (Momente der multivariaten hypergeometrischen Verteilung).
Sei $\boldsymbol{X} = (X_1, \ldots, X_n)$ multivariat hypergeometrisch verteilt wie in der Definition A.1.2, so gilt für $i, j = 1, \ldots, n$

$$\mathrm{E}[X_i] = \frac{kM_i}{N}\,, \tag{A.5}$$

$$Var[X_i] = \frac{nM_i(N-M_i)(N-k)}{N^2(N-1)} \quad \textit{und} \tag{A.6}$$

$$Cov[X_i, X_j] = -k\frac{M_i}{N}\frac{M_j}{N}\left(1 - \frac{k-1}{N-1}\right) \quad \textit{für } i \neq j\,. \tag{A.7}$$

Definition A.1.3 (Partielle Likelihood).
Sei $\boldsymbol{Y}$ ein Zufallsvektor mit Dichtefunktion $f_{\boldsymbol{Y}}(\boldsymbol{y}|\boldsymbol{\theta})$, mit $\boldsymbol{\theta} = (\phi, \boldsymbol{\beta})$ wobei der Vektor $\boldsymbol{\beta}$ der Parameter von Interesse ist und ϕ ein nuisance Parameter ist. Kann $\boldsymbol{Y}$ in $\boldsymbol{Y} = (\boldsymbol{W}, \boldsymbol{V})$ zerlegt werden, so dass die gemeinsame Dichte durch

$$f_{\boldsymbol{Y}}(\boldsymbol{y}|\boldsymbol{\theta}) = f_{\boldsymbol{W}|\boldsymbol{V}}(\boldsymbol{w}|\boldsymbol{v}, \boldsymbol{\theta}) \cdot f_{\boldsymbol{V}}(\boldsymbol{v}|\boldsymbol{\theta}) \tag{A.8}$$

gegben ist, dann nennt man den ersten Faktor $f_{\boldsymbol{W}|\boldsymbol{V}}$ ***partielle Likelihood*** *falls er nur von* $\boldsymbol{\beta}$ *nicht aber von* ϕ *abhängt, d.h.*

$$f_{\boldsymbol{W}|\boldsymbol{V}}(\boldsymbol{w}|\boldsymbol{v},\boldsymbol{\theta}) = f_{\boldsymbol{W}|\boldsymbol{V}}(\boldsymbol{w}|\boldsymbol{v},\boldsymbol{\beta})$$

(vgl. Cox, 1975).

B. Literaturverzeichnis

Literatur

Aalen, O. (1978). Nonparametric inference for a family of counting processes. *Annals of Statistics*, *6*, 534–45.

Andersen, P. K. und Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study. *Annals of Statistics*, *10*, 1100–1120.

Borucka, J. (2014). Methods for handling tied events in the Cox Proportional Hazard Model. *STUDIA OECONOMICA POSNANIENSIA*, *2*, 91–106.

Collett, D. (2003). *Modelling Survival Data in Medical Research* (2. Aufl.). New York: Chapman & Hall/CRC.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society, Series B*, *34*, 187–220.

Cox, D. R. (1975). Partial Likelihood. *Biometrika*, *62*, 269–276.

Cox, D. R. und Snell, E. J. (1968). A general definition of residuals. *Journal of the Royal Statistical Society. Series B*, *30*, 248–275.

Glomb, P. (2007). *Statistische Modelle und Methoden in der Analyse von Lebenszeitdaten* (Unveröffentlichte Diplomarbeit). Institut für Mathematik, Universität Oldenburg.

Höhle, M. (WS 2008/2009). *Analyse von Lebensdauern* (Vorlesungsskript). Institut fur Statistik, Ludwig-Maximilians-Universität München.

Hosmer, D. W. J. und Stanley, L. (1999). *Applied Survival Analysis.* Chichester: Wiley.

Kaplan, E. L. und Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, *53*, 457–81.

Klein, J. P. und Moeschberger, M. L. (2003). *Survival Analysis Techniques for Censored and Truncated Data.* New York: Springer.

Kleinbaum, D. G. und Klein, M. (2005). *Survival Analysis - A Self-Learning Text* (3. Aufl.). New York: Springer.

Lawless, J. F. (2003). *Statistical Models and Methods for Lifetime Data*. New Jersey: Wiley.

Liu, X. (2012). *Survival Analysis: Models and Applications*. Chichester: Wiley.

Medeiros, F. M. C., da Silva-Junior, A. H. M., Valencia, D. M. und Ferrari, S. L. (2014). Testing inference in accelerated failure time models. *International Journal of Statistics and Probability*, *3*, 121–131.

Moosbrugger, J. (2016). *Modellierung von Stornowahrscheinlichkeiten in der privaten Krankenversicherung* (Unveröffentlichte Diplomarbeit). Institut für Statistik, Technische Universität Graz.

Nelson, W. (1969). Nonparametric inference for a family of counting processes. *Journal of Quality Technology*, *1*, 27–52.

Sleeper, L. A. und Harrington, D. P. (1990). Regression splines in the Cox model with application to covariate effects in liver disease. *Journal of the American Statistical Association*, *85*, 941–949.

Tableman, M. und Kim, J. S. (2004). *Survival Analysis Using S: Analysis of Time-to-Event Data*. New York: Chapman & Hall/CRC.

Therneau, T. M. (2015). A package for survival analysis in S [Software-Handbuch]. Zugriff auf `http://CRAN.R-project.org/package=survival` (version 2.38)

Therneau, T. M. und Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model*. New York: Springer.