Christina Bilek, BSc

Modellierung und Projektion von Mortalitätsraten privatversicherter Personen

MASTERARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieurin

Masterstudium Finanz- und Versicherungsmathematik

eingereicht an der

Technischen Universität Graz

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

Institut für Statistik

Graz, August 2017

EIDESSTATTLICHE ERKLÄRUNG

AFFIDAVIT

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.

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly indicated all material which has been quoted either literally or by content from the sources used. The text document uploaded to TUGRAZonline is identical to the present master's thesis.

Datum/Date	Unterschrift/Signature

Zusammenfassung

Die Modellierung und Projektion von Mortalitäten ist vor allem im Versicherungssektor ein aktuelles Thema, da die Bestandsentwicklung maßgeblich vom Sterblichkeitsverlauf des Kollektivs abhängt. Im Rahmen der Arbeit werden verschiedene Modellierungsansätze aufgezeigt, wobei vor allem auf Penalisierte Generalisierte Lineare Modelle unter der Annahme einer Poissonverteilung eingegangen wird, die eine flexible Modellierung des Zusammenhangs einer oder mehrerer Prädiktoren auf die Response zulassen. Werden kubische Basis-Splines als Regressionsbasis verwendet, so liegt meist eine raue Schätzung vor, was durch das Einführen eines Strafterms im Modell verhindert werden kann. Somit bildet diese Modellklasse einen attraktiven Zugang für die Modellierung von Mortalitätsoberflächen. Neben P-Splines ist das Lee-Carter Modell ein weit verbreitetes Modell im Zusammenhang mit Sterblichkeitsanalysen. Auch auf diese Möglichkeit, Mortalitätsraten zu beschreiben, soll im Zuge der Arbeit näher eingegangen werden. Illustriert wird der theoretische Hintergrund anhand eines Datensatzes, der aus einer Privatversicherung stammt. Alle in der Arbeit angeführten Beispiele sind mit der Statistiksoftware R analysiert worden.

Abstract

The modeling and prediction of death rates is of high importance in the insurance sector, as the development of the portfolio depends on the mortality of the collective. Within the present thesis different methods for the modeling of mortality rates shall be introduced, with a particular focus on penalized generalized linear models under the assumption of a Poisson distribution. This model class allows for a flexible modeling of the relation of one or more predictors and the response. The use of cubic B-splines as regression base is mostly accompanied by a rough estimation result. In order to avoid wiggly fits, a penalty is introduced in the model. Hence, penalized generalized linear models provide an attractive way of modeling mortality surfaces. Apart from P-splines, the Lee-Carter model is a well-known model in the context of mortality analysis. This approach shall also be discussed within the thesis. The theoretical background is illustrated with a data set stemming from a private insurance. All examples shown in the thesis were analyzed using the statistical software R.

Danksagung

An dieser Stelle möchte ich mich bei all jenen bedanken, die mich beim Verfassen dieser Masterarbeit unterstützt haben.

Ein besonderer Dank gebührt Herrn Prof. Friedl für die engagierte Betreuung dieser Arbeit und die zahlreichen Anregungen, die das Entstehen der Arbeit vorangetrieben haben.

Außerdem möchte ich mich bei Herrn Stockreiter der Merkur Versicherung AG bedanken, der die Masterarbeit erst möglich gemacht hat und sie durch interessante Gespräche und neue Aspekte bereichert hat.

Ein herzliches Dankeschön an meine Eltern und Großeltern. Sie sind mir sowohl mit moralischer, als auch finanzieller Unterstützung stets zur Seite gestanden. Natürlich möchte ich mich auch bei meinem Bruder bedanken, der mich ebenfalls mental unterstützt hat.

Mein Dank gilt auch meinen Studienkollegen und Freunden, die mich durch das Studium begleitet haben und ohne die meine Studienzeit nur halb so schön gewesen wäre.

Zudem möchte ich mich bei meinen Kollegen der Merkur Versicherung AG sowie der Europäischen Zentralbank für die stets motivierenden Worte in dieser herausfordernden Phase bedanken.

Christina Bilek,
Graz, 09.08.2017

Inhaltsverzeichnis

Kapitel 1

Einleitung

Sterbewahrscheinlichkeiten bilden in vielen Kalkulationsbereichen von Versicherungen eine entscheidende Eingangsgröße. Da die Sterblichkeit einem ständigen Wandel unterliegt und auch die Sterblichkeit des Kollektivs eines Versicherungsunternehmens von jener der Gesamtbevölkerung signifikant abweichen kann, ist es von großem Interesse die Mortalitäten vom Bestand des Versicherungsunternehmens zu analysieren.

Ziel dieser Arbeit ist es, unterschiedliche Herangehensweisen zur Modellierung und Glättung von Sterbewahrscheinlichkeiten sowohl in einer Dimension, also auch in zwei Dimensionen aufzuzeigen und diese miteinander zu vergleichen. Außerdem ist eine Projektion der Sterbewahrscheinlichkeiten in die Zukunft vorgesehen.

Das zweite Kapitel widmet sich dem Datensatz, der die Datengrundlage der Arbeit bildet. Wichtige Merkmale werden herausgearbeitet und ein Überblick über die Struktur und den Umfang der Daten gegeben. Im Anschluss stehen verschiedene Möglichkeiten Sterblichkeiten zu beschreiben im Mittelpunkt. Das dritte Kapitel behandelt das Konzept der Linearen Regression in ihren Grundzügen, bevor das Hauptaugenmerk auf die allgemeineren Generalisierten Linearen Modelle gelegt wird. Aus dieser Modellklasse werden schließlich einige Modelle, die vor allem für Mortalitäten von Relevanz sind, herausgegriffen und näher besprochen.

Im Zusammenhang mit der Modellierung von Sterberaten ist das Lee-Carter Modell weit verbreitet und soll im vierten Kapitel im Zentrum der Ausführungen stehen. Dabei ist neben dem theoretischen Hintergrund des Modells auch die praktische Umsetzung anhand der gegebenen Daten illustriert und die Projektion der Mortalität in die Zukunft wird ebenfalls thematisiert. Außerdem ist nicht nur das klassische Lee-Carter Modell Gegenstand des Kapitels, sondern auch eine Variante, die eine für das Ziel Mortalitätsraten zu beschreiben wesentliche Verallgemeinerung hinsichtlich der Verteilungsannahme bietet.

Einen zum Lee-Carter Modell konträren Modellierungsansatz stellen die von Eilers und Marx (1996) eingeführten glatten Modelle mit Penalisierung dar. Diese sogenannten P-Spline Modelle werden im Kapitel fünf besprochen und ihre Anwendung auf die Daten, sowohl im Hinblick auf die univariate (Sterblichkeit in Abhängigkeit einer erklärenden Variable), als auch die bivariate Modellierung illustriert. Des Weiteren soll das von Camarda (2012) entwickelte R-Paket MortalitySmooth vorgestellt werden, da es ein maßgeschnei-

dertes Tool zur ein- und zweidimensionalen Modellierung von P-Spline Modellen bietet. Darüber hinaus wird auch die Vorhersage von Mortalitäten diskutiert.

Abschließend ist zum Überblick der vorgestellten Modelle eine Zusammenfassung der wesentlichen Ergebnisse sowie eine Diskussion der Vor- und Nachteile ausgewählter Modellierungsansätze im letzten Kapitel der Arbeit enthalten. Zur Abrundung der Thematik werden die geglätteten Sterberaten für den vorliegenden Datensatz und die Sterblichkeit der österreichischen Gesamtbevölkerung einem Vergleich unterzogen. Dies soll zum einen zeigen wie unterschiedlich die Entwicklung verschiedener Populationen sein kann und zum anderen die Wichtigkeit der Mortalitätsanalyse betonen, die in unterschiedlichsten Wissenschaftsbereichen relevant ist und Aktuare oder Demographen immer wieder vor neue Herausforderungen stellt.

Kapitel 2

Aufgabenstellung und Datenanalyse

In diesem Kapitel soll das Thema der vorliegenden Arbeit vorgestellt und die zur Verwendung kommenden Daten ersten Analysen unterzogen werden. Darüber hinaus werden verschiedene Sterblichkeitsmaße besprochen.

2.1 Motivation der Aufgabenstellung

Ein Versicherungsunternehmen (auch Versicherer) bietet dem Versicherungsnehmer (auch Versicherten) Schutz vor gewissen Risiken. Welches Risiko bzw. welche Risiken im konkreten Fall versichert sind, ist Gegenstand des Versicherungsvertrages. So können in der Personenversicherung beispielsweise das Risiko des Todes, der Krankheit oder der Invalidität Vertragsbestandteil sein.

Im Falle des Eintritts des Versicherungsfalles leistet das Versicherungsunternehmen eine oder mehrere Geldzahlungen. Im Gegenzug dazu muss der Versicherungsnehmer Prämien (auch Beiträge) bezahlen. Die Höhe der Versicherungsprämie ist zum einen von der Art der Leistung, die vom Versicherer im Versicherungsfall zu erbringen ist und zum anderen von der Art und Eintrittswahrscheinlichkeit des versicherten Risikos abhängig. Um auch in unerwartet schlechten Risiko-, Kosten- und Kapitalanlagesituationen eine gewisse Sicherheit zu haben, ziehen Versicherer bei der Kalkulation der Prämie Rechnungsgrundlagen[1] erster Ordnung heran, die schon Sicherheitszuschläge beinhalten.

Aus diesem Grund wird die Sterblichkeit, die das zentrale Thema dieser Arbeit bilden wird, also überschätzt (bzw. im Falle einer Erlebensfallversicherung unterschätzt) und somit ausreichende Reserven geschaffen sowie garantiert, dass den Leistungsverpflichtungen nachgekommen werden kann. Demnach bezahlt der Versicherungsnehmer aber eine höhere Prämie, als sie seinem tatsächlichen Risikoprofil entspricht und das Versicherungsunternehmen erwirtschaftet Gewinne, die jedoch keinen Bestandteil des Unternehmensgewinnes

[1] Unter Rechnungsgrundlagen verstehen wir die zentralen Eingangsparameter in der Beitrags- und Deckungsrückstellungsberechnung in der Lebensversicherung und privaten Krankenversicherung. Zu den Rechnungsgrundlagen zählen der Zins, die Sterblichkeit, die Kosten und in der Krankenversicherung zusätzlich die Stornowahrscheinlichkeit.

bilden und daher eine Rückerstattung an den Versicherungsnehmer, abhängig von der Vertragsart und den Vertragsbedingungen, vorgesehen ist.

Folglich ist es für das Versicherungsunternehmen essentiell auch die tatsächlichen Sterblichkeitsverläufe seines Kollektivs zu kennen. Diese sind ein Element der Rechnungsgrundlagen zweiter Ordnung, die vor allem auch im Zuge der Rechnungslegung nach Solvency II[2] zu verwenden sind.

Das Kernthema der vorliegenden Arbeit ist die Modellierung der tatsächlichen Sterbewahrscheinlichkeit basierend auf Daten einer Privatversicherung, welche schließlich als Rechnungsgrundlagen zweiter Ordnung Verwendung finden sollen. Außerdem wird aus diesen Ergebnissen die Sterblichkeit für ein Jahr im Voraus prognostiziert. Es ist überdies auch von Interesse die geschätzte Sterblichkeit des Kollektivs der Privatversicherung mit den von der Statistik Austria veröffentlichten Sterbetafeln[3] für Gesamtösterreich zu vergleichen, um den Unterschied des Sterblichkeitsverlaufs privatversicherter Personen und der österreichischen Gesamtbevölkerung zu illustrieren und die Notwendigkeit der bestandsbasierten Sterblichkeitsanalyse zu verdeutlichen.

2.2 Datenanalyse

Im nachstehenden Abschnitt wird der Datensatz, der die Grundlage dieser Arbeit bildet, beschrieben und einer ersten Analyse unterzogen. Außerdem sollen für die weiteren Ausführungen wesentliche und in der Praxis verwendete Sterblichkeitsmaße diskutiert werden sowie die ungeglättete Mortalität illustriert werden.

2.2.1 Beschreibung des Datensatzes

Es stehen die beobachteten Sterbe- sowie Stornofälle des Kollektivs über einen Zeitraum von fünfzehn Jahren zur Verfügung, genauer vom Jahr 2001 bis zum Jahr 2015. Die Stornofälle sind hinsichtlich der Sterblichkeitsanalyse relevant, da eine Person, die innerhalb eines Beobachtungsjahres storniert, ab dem Zeitpunkt der Stornierung aus dem Versicherungsbestand ausscheidet und keine Angaben über das Überleben bzw. Nicht-Überleben des Beobachtungsjahres gemacht werden können.

Das Format der Rohdaten sowie die entsprechenden R-Skripte, die der weiterführenden Datenaufbereitung dienen, sind im Anhang A einsehbar. In Tabelle 2.1 sind die verfügbaren Informationen der bereits weiterverarbeiteten Rohdaten angeführt. Jeder Zeile im Datensatz entspricht eine Person, die solange keine Stornierung oder kein Todesfall vorliegt auch im Folgejahr wieder im Datensatz enthalten ist. Wichtig ist, dass jede Person innerhalb eines Beobachtungsjahres nur einmal aufscheint, auch wenn sie möglicherweise mehrere Versicherungsverträge besitzt, da es andernfalls bei der Modellierung der Sterblichkeit zu

[2]Solvency II ist ein EU-weites Projekt, das das Versicherungsaufsichtsrecht einheitlich regelt. Zentraler Gegenstand ist die Marktwertbilanz, die Versicherungsunternehmen nun vorlegen müssen und deren Berechnung auf möglichst realistischen Annahmen basieren soll, also auf den Rechnungsgrundlagen zweiter Ordnung.

[3]Die Sterbetafel gibt getrennt nach Geschlecht an, wie sich ein Kollektiv von bestimmten Personen auf Grund des Ablebens verringert.

Abhängigkeiten kommen würde.

Bezeichnung	Beschreibung
VRNR	Mit Hilfe dieser Nummer sind die Personen im Datensatz eindeutig identifizierbar.
GEBDATUM	Enthält das Geburtsdatum.
GESCHLECHT	Das Geschlecht ist ein binäres Merkmal mit den möglichen Ausprägungen männlich („M“) und weiblich („W“).
BEWEGUNGJ	Das Bewegungsjahr gibt an, in welchem Jahr über den Beobachtungszeitraum von fünfzehn Jahren der Sterbefall oder die Stornierung bzw. keines der beiden Ereignisse beobachtet wurde (kein Eintrag).
STERBEM	Beinhaltet das Sterbemonat, falls diese Person innerhalb des Bewegungsjahres verstorben ist und andernfalls keinen Eintrag.
STORNOM	Beinhaltet das Monat, in dem storniert wurde bzw. falls keine Stornierung vorliegt ist kein Eintrag enthalten.
ALTER	Gibt das Alter der jeweiligen Person im entsprechenden Bewegungsjahr an und berechnet sich aus der Differenz zwischen dem Bewegungsjahr und dem Geburtsjahr.
UEBERLSTAT	Gibt den Überlebensstatus an. Überlebt die Person das Bewegungsjahr, so wird UEBERLSTAT auf 0 gesetzt, andernfalls auf 1.
STORNOSTAT	Ist wie UEBERLSTAT ein binäres Merkmal, das im Falle einer Stornierung innerhalb des Bewegungsjahres auf 1 gesetzt wird und ansonsten auf 0.

Tabelle 2.1: Inhalt des Datensatzes.

2.2.2 Eine erste Datenanalyse

Einen ersten Eindruck über die vorliegende Datensituation soll Abbildung 2.1 geben. Es ist ersichtlich, dass die interessierende Größe, die Anzahl der Sterbefälle innerhalb eines Jahres, sich über den Beobachtungszeitraum nicht signifikant ändert, wenngleich der Bestand vor allem ab dem Jahr 2010 erkennbar wächst. Es ist außerdem ersichtlich, dass die Anzahl der Stornierungen deutlich über jener der Sterbefälle liegt. Die konkreten Zahlen der Sterbe- und Stornierungsfälle pro Jahr sind in Tabelle 2.2 angeführt.

Abgesehen von der Anzahl der Stornierungen und Sterbefälle ist auch das Merkmal „Geschlecht“ im Hinblick auf die Modellierung der Sterblichkeit von Interesse. In Abbildung 2.2 ist die Anzahl der Männer sowie die der Frauen über den beobachteten Zeitraum dargestellt. Der Anteil der Frauen am Gesamtbestand ist erkennbar höher als jener der Männer und auch aus dem zu erwartenden Mortalitätsunterschied von Frauen und Männern ist es

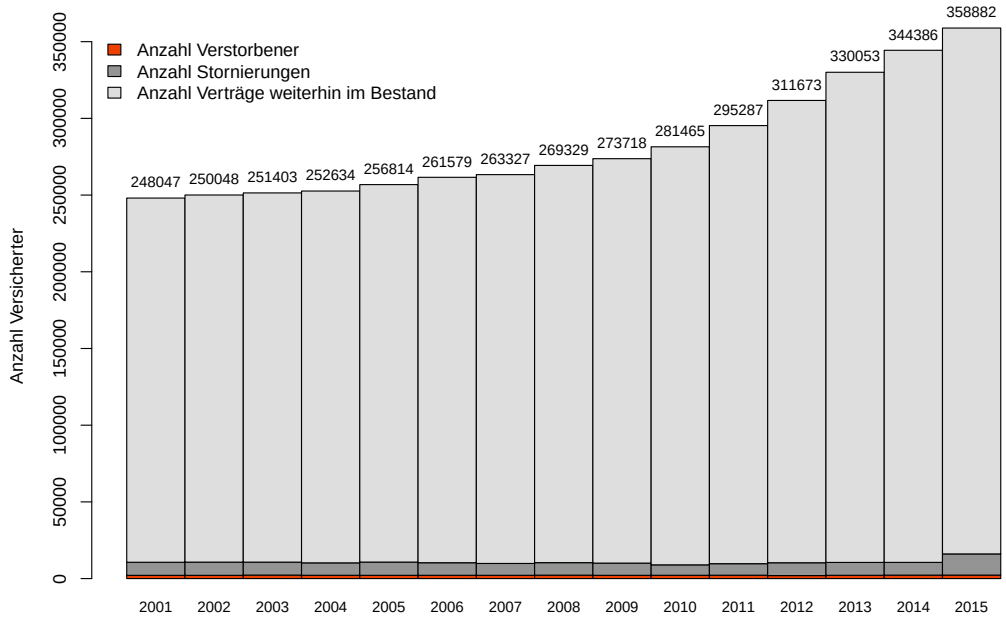

Abbildung 2.1: Entwicklung der Sterblichkeit, des Stornoverhaltens und des Bestandes vom Jahr 2001 bis zum Jahr 2015.

Jahr	**Storniert**	**Verstorben**	**Jahr**	**Storniert**	**Verstorben**
2001	8530	2113	2009	7963	2124
2002	8609	2120	2010	6803	2154
2003	8436	2284	2011	7521	2197
2004	8097	2161	2012	8348	1971
2005	8628	2132	2013	8393	2149
2006	8197	2147	2014	8302	2258
2007	7725	2142	2015	13805	2272
2008	8156	2236			

Tabelle 2.2: Anzahl der Stornierungen und Sterbefälle nach Beobachtungsjahr.

naheliegend, bei der Modellierung der Sterblichkeit anhand des Geschlechts zu differenzieren.

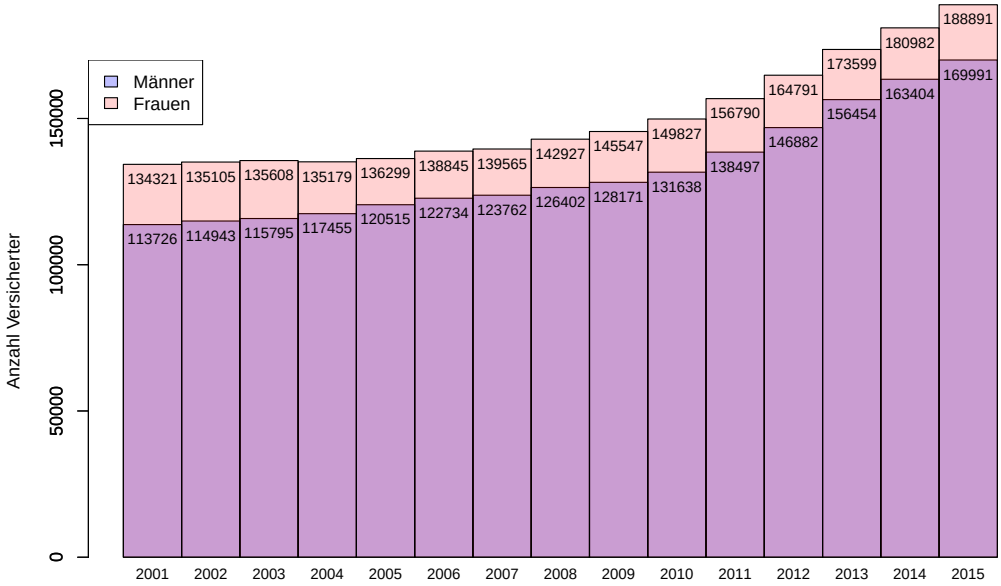

Abbildung 2.2: Anzahl der Personen im Kollektiv nach Beobachtungsjahr und Geschlecht. Da die Anzahl der Frauen die Anzahl der Männer im Bestand übersteigt, wird die hellblaue Färbung für Männer vom Hellrot der Frauen überlagert.

In weiterer Folge ist hinsichtlich der Mortalitätsanalyse ebenso das Alter ein relevantes Merkmal, das nun näheren Betrachtungen unterzogen wird. Das Alter der Versicherten, das sich als Differenz des Bewegungsjahres und des Geburtsjahres berechnet[4], nimmt für Frauen Werte im Intervall $[0, 112]$ an und für Männer im geringfügig kleineren Intervall $[0, 108]$. Die Verteilung des Alters in ausgewählten Beobachtungsjahren ist in Abbildung 2.3 dargestellt. Es lässt sich davon ableiten, dass die Altersverteilung über die Jahre 2001 bis 2015 hinweg einer ähnlichen Struktur folgt. In den unteren Alterssegmenten, wie auch in den oberen Altersbereichen sind im Vergleich zum Altersbereich $[30, 80]$, in dem sich die meisten versicherten Personen befinden, wenige Versicherungsnehmer im Bestand. Diese Altersstruktur ist bei Frauen und Männern gleichermaßen zu beobachten. Lediglich im Beobachtungsjahr 2015 sind kleinere Abweichungen von der Verteilung der Vorjahre feststellbar. Diese betreffen in erster Linie den Altersbereich von 0 bis 10 Jahren, wo im Vergleich zu den Vorjahren doppelt so viele Versicherte enthalten sind. Die meisten Versicherungsnehmer weisen ein Alter von 30 bis 40 Jahren auf, wohingegen in den übrigen Beobachtungsjahren das Alterssegment $[40, 50]$ am stärksten besetzt ist. Erkennbar ist auch hier das Wachsen des Bestandes ab dem Jahr 2010, wodurch vor allem in den unteren Alterssegmenten die Zahl der Versicherten gestiegen ist.

[4]Nähere Erläuterungen zur Berechnung des Alters der Versicherten sind im Anhang A zu finden.

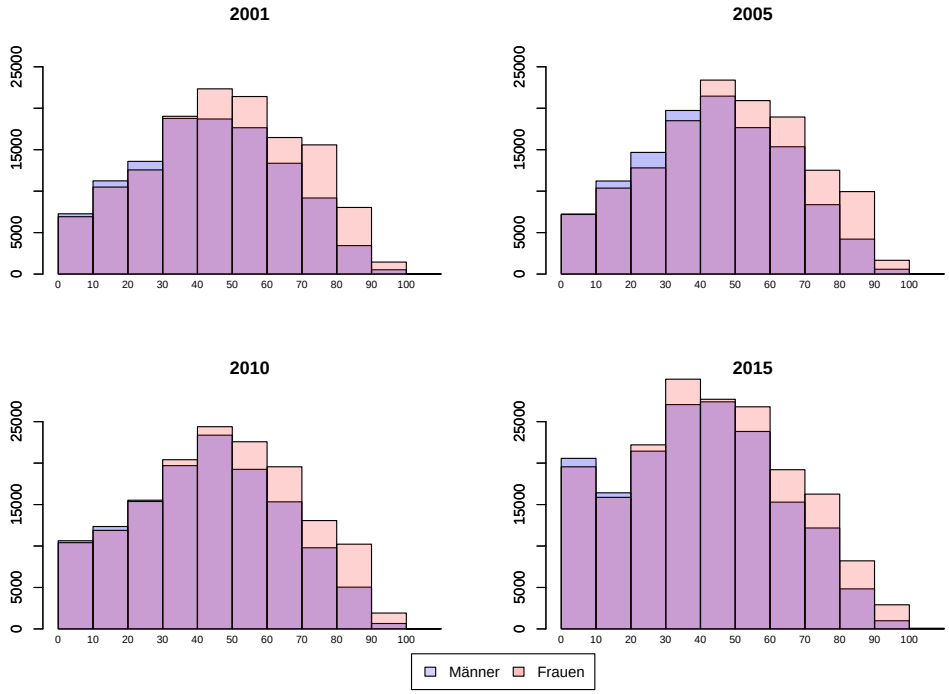

Abbildung 2.3: Verteilung des Alters nach Beobachtungsjahr und Geschlecht.

Nun ist es von Interesse die Anzahl der beobachteten Sterbefälle nach Alter und Beobachtungsjahr zu betrachten, um beurteilen zu können, in welchem Altersbereich genügend Sterbefälle festgestellt wurden und somit eine zuverlässige Modellierung möglich ist. Es ist erkennbar, dass vor allem im Alterssegment von 0 bis 40 Jahren wenige bis gar keine Beobachtungen zu Sterbefällen vorhanden sind, weshalb eine Modellierung in diesem Altersbereich nicht sinnvoll ist (siehe Abbildung 2.4). Die Jahre 2001, 2005, 2010 und 2015 sind auch hier als repräsentativ für alle anderen Beobachtungsjahre anzusehen, da die Verteilung in den übrigen Jahren keine abweichende Struktur aufweist. Wiederum ist im Jahr 2015 im Alterssegment $[80, 100]$ eine geringfügige Abweichung der Verteilung der Sterbefälle von den restlichen betrachteten Jahren ersichtlich. Die Anzahl der Sterbefälle nimmt einerseits innerhalb des Bereichs $[80, 90]$ gegenüber den Vorjahren ab und nimmt hingegen im Intervall $[90, 100]$ fast auf das Doppelte zu. Bezüglich dem Geschlecht übersteigen die Sterbefälle der Männer jene der Frauen, was sich ab einem Alter von 80 Jahren umkehrt, da von da an mehr Sterbefälle von Frauen zu beobachten sind.

Eine genauere Aussage über die Veränderung der Sterblichkeit über den Beobachtungszeitraum liefert Abbildung 2.5, wo die Sterbefälle innerhalb eines Jahres in Relation zum Gesamtbestand in diesem Jahr gesetzt werden. Bei einem Vergleich des linken und rechten Bildes ist erkennbar, dass Frauen eine geringere Sterbewahrscheinlichkeit aufweisen als Männer. In der linken Grafik steigt die Sterblichkeit ab einem Alter von 60 Jahren sichtlich an, bei Frauen ist dies hingegen ab ca. 75 Jahren der Fall. Betrachten wir den Verlauf über die Beobachtungsjahre, so ist in beiden Bildern zu sehen, dass gegenüber 2001

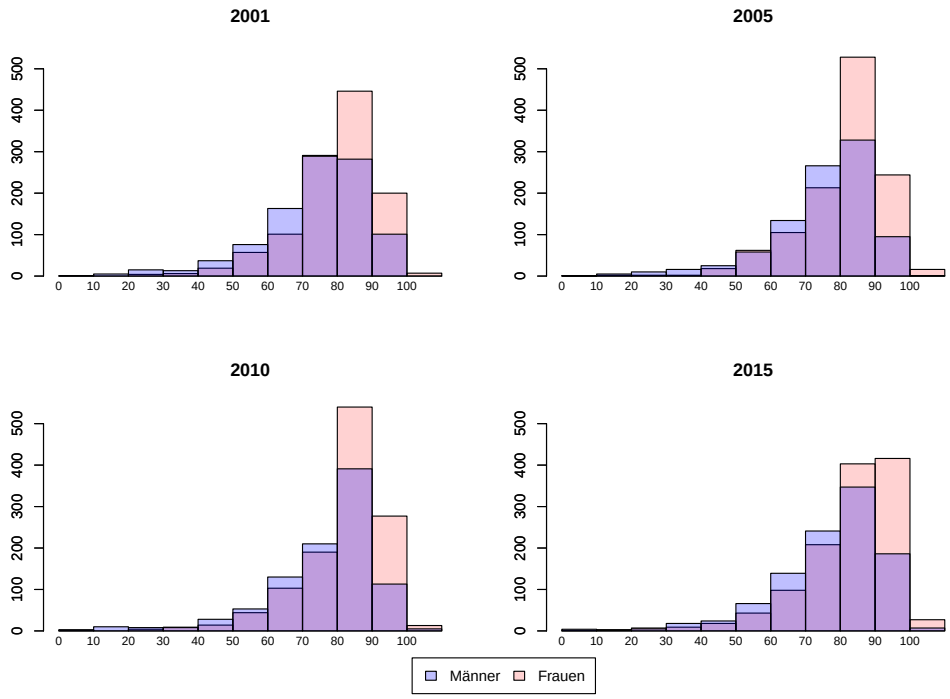

Abbildung 2.4: Anzahl der Sterbefälle nach Alter, Beobachtungsjahr und Geschlecht.

die Sterbewahrscheinlichkeit im Jahr 2015 niedriger ist, zumindest bis zu einem Alter von ca. 95 Jahren. Die beiden auffälligen Datenpunkte $(99, 0)$ und $(100, 0)$ im linken Bild sind aus den für diese beiden Alter fehlenden Beobachtungen zu Todesfällen in den jeweiligen Jahren zu erklären.

Wie schon oben erwähnt, beeinflussen Stornierungen ebenso die Schätzung. Bei der Berechnung der in Abbildung 2.5 abgebildeten Sterbewahrscheinlichkeiten wurde auf Grund von der nicht möglichen Beobachtung der Sterblichkeit nach einer Stornierung angenommen, dass der Versicherungsnehmer das Beobachtungsjahr, in dem storniert wurde, überlebt. Wird in Abbildung 2.6 die Verteilung der Stornofälle nach dem Alter betrachtet, dann ist erkennbar, dass Stornierungen sowohl bei Frauen, als auch bei Männern, wie zu erwarten ist, eher in den unteren Alterssegmenten $[0, 50]$ erfolgen. Da für die Modellierung der Sterblichkeit in dieser Arbeit der Altersbereich $[40, 100]$ relevant ist, fällt der Großteil der Stornierungen nicht in das interessierende Altersintervall. Dennoch sollten Stornierungen nach Möglichkeit bei der Modellierung der Mortalität Berücksichtigung finden, um die Realität bestmöglich widerzuspiegeln.

2.2.3 Maße für die Sterblichkeit und Analyse der rohen Sterblichkeit

Dieses Unterkapitel widmet sich der Mortalitätsanalyse der Rohdaten anhand unterschiedlicher Sterblichkeitsmaße, die in der Folge auch vorgestellt werden sollen und wofür die in Tabelle 2.3 angeführte Notation vereinbart wird. Die Definitionen dieses Abschnitts basie-

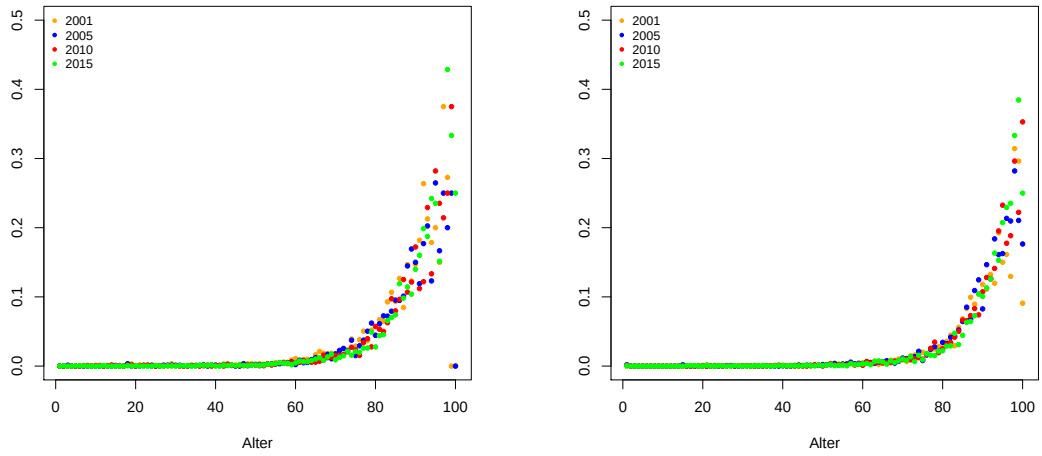

Abbildung 2.5: Sterbefälle in Relation zum Gesamtbestand nach Alter in den Jahren 2001, 2005, 2010 und 2015 sowie getrennt nach Geschlecht (Männer linkes Bild, Frauen rechtes Bild).

ren auf Feichtinger (1973), Milbrodt und Helbig (1999), Thatcher, Kannisto und Vaupel (1999) und Currie (2013).

Notation	**Beschreibung**
x	Alter einer Person
t	Kalenderjahr mit $t \in \{2001, \ldots, 2015\}$; ein Kalenderjahr läuft vom Zeitpunkt t bis $t+1$
T_x	Sterbealter einer x-jährigen Person
$m_{x,t}$	rohe Mortalitätsrate im Alter x und Beobachtungsjahr t
$\mu_{x,t}$	Sterbeintensität einer zum Zeitpunkt t x-jährigen Person
$q_{x,t}$	einjährige Sterbewahrscheinlichkeit einer im Beobachtungsjahr t x-jährigen Person
$D_{x,t}$	beobachtete Sterbefälle im Alter x und Beobachtungsjahr t
$E_{x,t}$	Anzahl der x-Jährigen, die im Beobachtungsjahr t dem Mortalitätsrisiko unterliegen; berechnet mittels verlebter Personenjahre
$E'_{x,t}$	Anzahl der x-Jährigen, die am Anfang des Beobachtungsjahres t dem Mortalitätsrisiko unterliegen

Tabelle 2.3: Verwendete Notationskonvention.

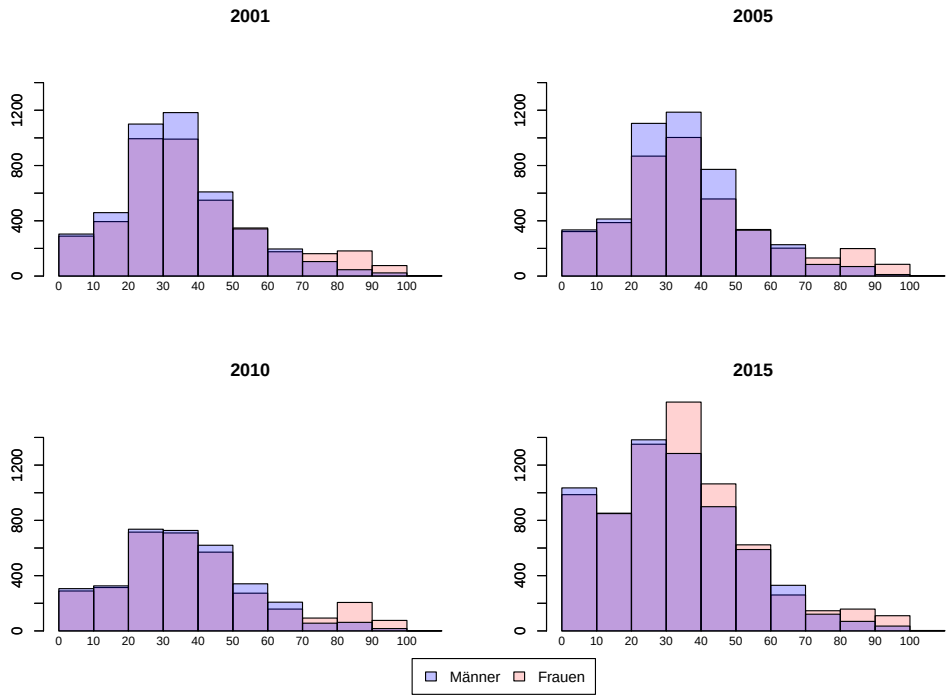

Abbildung 2.6: Anzahl der Stornierungen nach Alter, Beobachtungsjahr und Geschlecht.

Aus den Rohdaten können die rohen Sterberaten

$$m_{x,t} = \frac{D_{x,t}}{E_{x,t}} \tag{2.1}$$

berechnet werden. Der Zähler in Darstellung (2.1), also die Anzahl der Todesfälle im Referenzzeitraum (ein Jahr), ist aus den Bestandsdaten einfach zu ermitteln. Schwierigkeiten bereitet hingegen der Nenner, da wie schon oben erwähnt, nicht alle Personen im gesamten Beobachtungszeitraum, auf Grund von Stornierungen und dem Ableben innerhalb des Referenzzeitraums, dem Risiko des Todes unterliegen.

Gemäß demographischen Methoden (vgl. Feichtinger, 1973) werden daher die verlebten Personenjahre im Beobachtungszeitraum herangezogen oder als Approximation die sich zur Jahresmitte im Versichertenbestand befindlichen Personen als Bezugsgröße verwendet. Für Letzteres muss angenommen werden, dass Stornierungen und Todesfälle über das Jahr gleichverteilt sind, was aus den gegebenen Daten jedoch nicht bestätigt werden kann. Die Abbildung 2.7 zeigt, dass diese Annahme zumindest für das Jahr 2001 nicht gerechtfertigt zu sein scheint und auch der Chi-Quadrat-Anpassungstest zur Nullhypothese H_0: *Anzahl der Stornierungen und Todesfälle ist über die Monate eines Jahres gleichverteilt* mit den in Tabelle 2.4 dargestellten p-Werten untermauert diese These. Dies liegt in erster Linie an den Stornierungen, die erfahrungsgemäß am Beginn des Jahres erfolgen und nicht an den Sterbefällen, die wie in der Abbildung erkennbar, über die Monate hinweg keinen größeren Schwankungen unterliegen. Daher werden für die Berechnung der rohen

Sterberaten die verlebten Personenjahre, also eine Gewichtung jeder Person im Bestand mit dem Zeitraum, in dem sie im Referenzjahr Teil des Kollektivs ist und somit dem Mortalitätsrisiko unterliegt, zur Verwendung kommen. Ein Versicherungsnehmer, der im Beobachtungsjahr im Juni storniert oder verstirbt, bekommt demnach ein Gewicht von 0.5 zugewiesen.

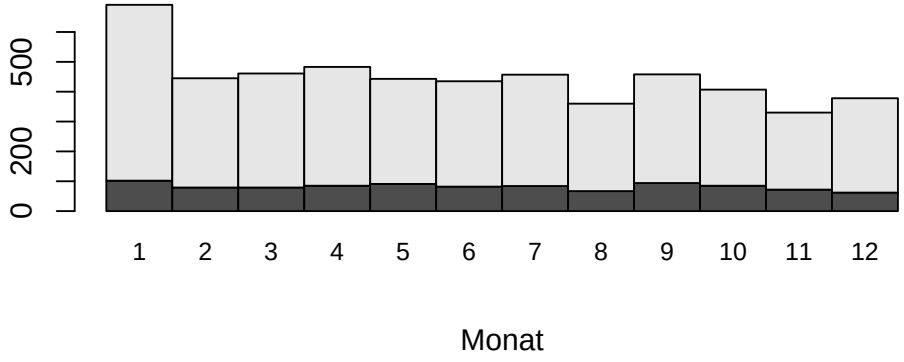

Abbildung 2.7: Anzahl der Sterbefälle (dunkelgrau) und Stornierungen (hellgrau) in den einzelnen Monaten im Jahr 2001.

Jahr	**p-Wert**	**Jahr**	**p-Wert**
2001	8,48E-37	2009	4,8E-25
2002	9,2E-16	2010	1,93E-24
2003	1,33E-19	2011	2,12E-06
2004	7,42E-20	2012	7,7E-82
2005	2,25E-11	2013	4,13E-12
2006	2,95E-19	2014	0,000136
2007	7,14E-21	2015	4,47E-12
2008	1,79E-39		

Tabelle 2.4: p-Werte des Chi-Quadrat-Anpassungstests für alle Jahre des Beobachtungszeitraums.

Auch die Sterbewahrscheinlichkeiten $q_{x,t}$ können aus den Daten mittels

$$q_{x,t} = \frac{D_{x,t}}{E'_{x,t}} \tag{2.2}$$

geschätzt werden. Der Nenner unterschiedet sich hierbei von jenem in (2.1) und bezieht sich auf die Anzahl der Versicherten, die am Anfang des beobachteten Jahres dem Mortalitätsrisiko ausgesetzt sind. Daraus wird ersichtlich, dass bei Sterberaten mögliche Veränderungen im betrachteten Versichertenbestand Berücksichtigung finden, bei Sterbewahrscheinlichkeiten jedoch nicht. Dennoch gibt es zwischen $m_{x,t}$ und $q_{x,t}$ einen Zusammenhang, der in diesem Abschnitt noch näheren Betrachtungen unterzogen werden soll.

Während sich die Mortalitätsrate $m_{x,t}$ und die Sterbewahrscheinlichkeit $q_{x,t}$ jeweils auf die den Berechnungen zugrunde gelegten Referenzzeiträumen bezieht (hier ein Jahr), handelt es sich bei der Sterbeintensität $\mu_{x,t}$ um die momentane Sterberate. Ist dx ein kleines Zeitintervall, so gibt $\mu_{x,t}dx$ näherungsweise die Wahrscheinlichkeit an, dass eine zum Zeitpunkt t x-jährige Person in $(x, x + dx)$ verstirbt. Definiert ist die Sterbeintensität durch (vgl. Milbrodt & Helbig, 1999)

$$\mu_{x,t} = \lim_{dx \to 0} \frac{P(x < T_x \leq x + dx | T_x > x)}{dx}. \tag{2.3}$$

Daraus lässt sich eine Beziehung zur Sterbewahrscheinlichkeit herleiten. Dazu nehmen wir an, dass die Sterbeintensität $\mu_{x,t}$ innerhalb eines Jahres, also im Intervall $[t, t + 1)$ und bezüglich dem Alter und damit in $[x, x + 1)$ konstant ist (vgl. Currie, 2013), d.h.

$$\mu_{x+u,t+v} \approx \mu_{x,t}, \qquad 0 \leq u, v < 1. \tag{2.4}$$

Aus Definition (2.3) lässt sich die Beziehung $\mu_{x,t} = -\frac{d}{dx} \log(1 - q_{x,t})$ herleiten (siehe Milbrodt & Helbig, 1999) und durch Integration und Annahme (2.4) ergibt sich weiters

$$q_{x,t} \approx 1 - e^{-\mu_{x,t}}. \tag{2.5}$$

Dies ist eine wichtige Beziehung zwischen der Sterbeintensität und der Sterbewahrscheinlichkeit, die vor allem dann von Bedeutung ist, wenn Modelle betrachtet werden, die für $\mu_{x,t}$ formuliert sind und umgekehrt und Vergleichbarkeit geschaffen werden soll.

Außerdem ergibt sich aus der Annahme der konstanten Sterbeintensität (2.4) ein Zusammenhang zwischen der Mortalitätsrate und der Sterbeintensität. Es gilt näherungsweise $\mu_{x,t} = m_{x,t}$ (vgl. Finan, 2011), wodurch die Mortalitätsrate als Schätzer für die Sterbeintensität herangezogen werden kann. Während die Sterbewahrscheinlichkeit ausschließlich Werte in $[0, 1]$ annehmen kann, können sowohl die Mortalitätsrate als auch die Sterbeintensität Werte größer als 1 annehmen, wie es in Abbildung 2.8 zu sehen ist.

Nun sollen die rohen Mortalitätsraten aus den Daten berechnet und grafisch dargestellt werden. Dazu müssen die Originaldaten in eine für die weiteren Analysen passende Form gebracht werden, wofür es notwendig ist, getrennt nach Geschlecht eine Matrix, die die Anzahl der Sterbefälle nach Alter (Zeilen der Matrix) und Beobachtungsjahr (Spalten der Matrix) enthält, zu erstellen. Mit den konkreten Daten ist das die Matrix $\mathbf{D}^M = (D^M_{x,t}) \in \mathbb{R}^{109\times 15}$ für Männer bzw. für Frauen die Matrix $\mathbf{D}^W = (D^W_{x,t}) \in \mathbb{R}^{113\times 15}$. Analog dazu

benötigen wir pro Geschlecht eine Matrix, die die Anzahl der dem Risiko des Todes unterliegenden Personen (auch Exposures) enthält, also eine Matrix $\mathbf{E}^M = (E^M_{x,t}) \in \mathbb{R}^{109\times 15}$ für Männer und für Frauen die Matrix $\mathbf{E}^W = (E^W_{x,t}) \in \mathbb{R}^{113\times 15}$. Die Aufbereitung dieser Matrizen in R kann im Anhang A in Listing A.3 nachvollzogen werden.

Die Abbildung 2.8 zeigt die Mortalitätsraten für Männer und Frauen über den Beobachtungszeitraum 2001 bis 2015. Es ist erkennbar, dass die rohen Mortalitätsraten sehr volatil sind und daher eine Glättung vorteilhaft ist, um bessere Schätzer mit geringerer Varianz zu erhalten. Anhand der Abbildungen ist ersichtlich, dass wie schon oben erwähnt, vor allem im Altersbereich $[0, 40]$ teilweise keine Beobachtungen vorhanden sind (weiße Felder) und daher keine zuverlässige Schätzung der Mortalitätsrate in diesem Bereich möglich sein wird. Außerdem ist zu sehen, dass die Sterblichkeit, wie erwartet, mit dem Alter deutlich zunimmt und bis zu einem Alter von 60 Jahren nahe bei Null liegt. Auch hier ist die Sterblichkeit der Frauen sichtlich geringer als jene der Männer. Eine Veränderung der Mortalitätsraten über den Beobachtungszeitraum hinweg ist vor allem in den Kontur-Diagrammen jeweils im rechten Bild der Abbildung 2.8 ersichtlich. Insbesonders bei den männlichen Versicherten ist ein Rückgang der Sterblichkeit bis hin zum Jahr 2015 deutlich. Bei den Frauen ist dieser weniger stark ausgeprägt, ist aber dennoch bis zum Alter von ca. 95 Jahren zu beobachten. Danach nimmt die Sterblichkeit im Jahr 2015 sowohl bei Frauen, als auch bei Männern gegenüber den restlichen Beobachtungsjahren zu.

Wie anhand der angeführten Abbildungen ersichtlich ist, hängen Mortalitätsraten vom Alter (x) und der Zeit (t) ab, sind also zweidimensional. Aus diesem Grund wird es sinnvoll sein, auch eine zweidimensionale Glättung, in beide Dimensionen Zeit und Alter, vorzunehmen. Basierend auf dieser Datenanalyse werden daher in weiterer Folge Modelle in Betracht gezogen, die hinsichtlich dem Geschlecht differenzieren und die Sterbewahrscheinlichkeit oder die Mortalitätsrate unter Annahme einer geeigneten Verteilung in Abhängigkeit vom Alter und dem Beobachtungsjahr darstellen.

Davor steht im folgenden Kapitel die Klasse der Generalisierten Linearen Modelle im Vordergrund, die zunächst für die eindimensionale Modellierung (in Abhängigkeit vom Alter) Anwendung findet und schließlich zur zweidimensionalen Glättung hinführen soll.

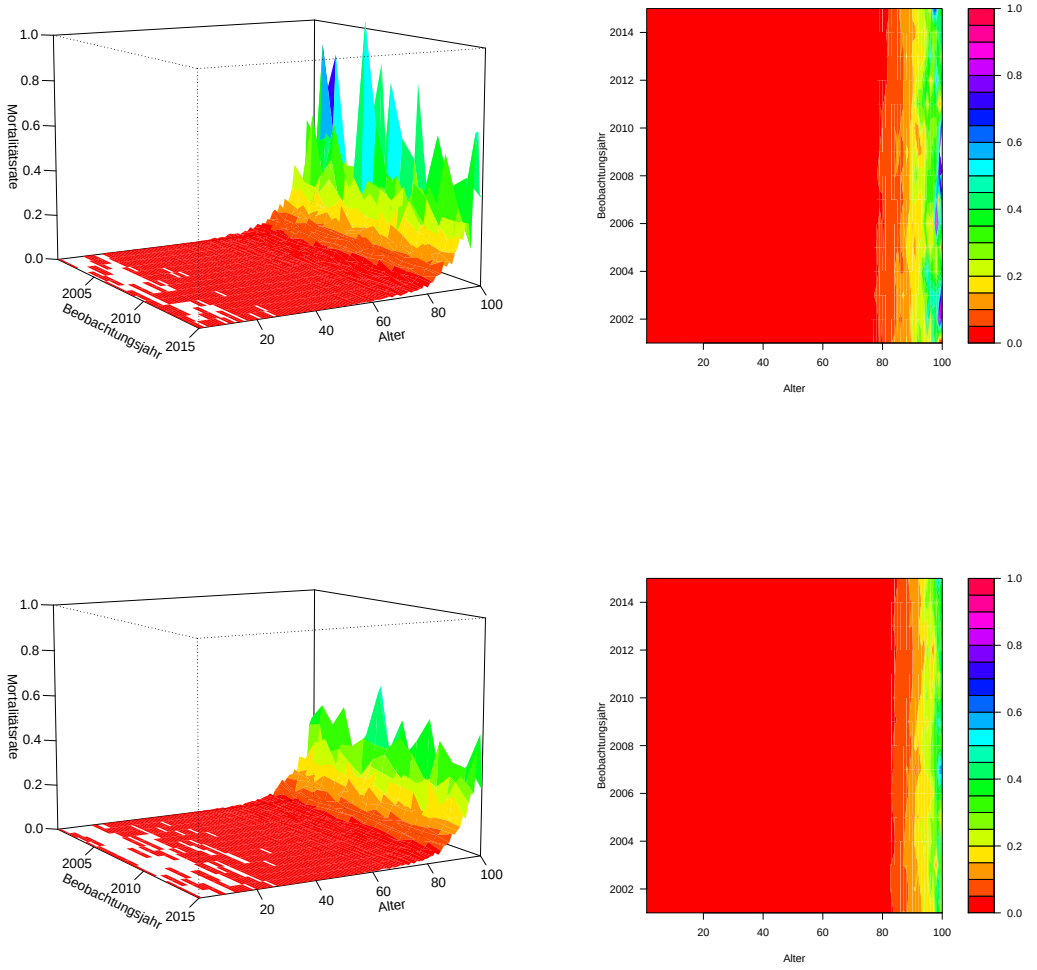

Abbildung 2.8: Mortalitätsraten der Männer (oben) und Frauen (unten) im Bestand im Altersbereich $[1, 100]$ über den gesamten Beobachtungszeitraum.

Kapitel 3

Mortalitätsmodelle in Abhängigkeit vom Alter

In diesem Kapitel sollen zunächst die theoretischen Hintergründe geeigneter Modelle für die Modellierung von Mortalitätsraten bzw. Sterbewahrscheinlichkeiten dargelegt werden. Im Vordergrund steht hierbei noch die eindimensionale Modellierung der Mortalität in Abhängigkeit vom Alter. Die im Rahmen dieses Kapitels vorgestellten Modelle werden schließlich auch auf die vorliegenden Daten angewandt.

Im ersten Unterkapitel soll der einfachste Zugang einen Zusammenhang zwischen zwei oder mehreren Merkmalen zu beschreiben, also die Lineare Regression, kurz vorgestellt werden. Danach erfolgt im zweiten Unterabschnitt eine Verallgemeinerung der Linearen Regression und damit werden die Generalisierten Linearen Modelle eingeführt. Der letzte Abschnitt widmet sich unterschiedlichen Modellierungsansätzen. In diesem Zusammenhang werden die Poisson-Regression, das Quasi-Poisson-Modell sowie das Negativ-Binomial-Modell als Verallgemeinerung des Poisson-Modells und abschließend die Logistische Regression diskutiert.

3.1 Das Lineare Regressionsmodell

In diesem Abschnitt wird das Multiple Lineare Regressionsmodell (vgl. Weisberg, 2005) eingeführt, das eine Möglichkeit bietet eine interessierende oder abhängige Variable y (Response-Variable) durch p unabhängige oder erklärende Variablen $\mathbf{x}$ (Prädiktor-Variablen, Kovariablen) zu beschreiben.

Formal sei eine Stichprobe aus einer Population vom Umfang n vorliegend, wobei die i-te Beobachtung die Form $\{x_{i1}, \ldots, x_{ip}, y_i\}, i \in \{1, \ldots, n\}$, besitzt. Die Response-Variable y wird linear in $\mathbf{x}$ modelliert:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i,$$

dabei sind die ϵ_i's Fehlerterme, die unabhängig seien und einer Normalverteilung mit Erwartung $\mu = 0$ und konstanter Varianz σ^2 unterliegen. Zu schätzen sind die unbekannten Parameter $\boldsymbol{\beta}$ und σ^2.

In Matrixschreibweise erhalten wir für das Multiple Lineare Regressionsmodell die Form

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

mit der $n \times (p+1)$-Designmatrix $\mathbf{X}$, dem Responsevektor $\mathbf{y}$, dem Parametervektor $\boldsymbol{\beta}$ und dem Vektor der Fehlerterme $\boldsymbol{\epsilon}$:

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

Wie schon oben erwähnt gilt $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$, woraus sofort $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n)$ folgt.

Die zu schätzenden Parameter $\boldsymbol{\beta}$ werden über die Kleinste-Quadrate-Methode bestimmt, indem

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

in $\boldsymbol{\beta}$ minimiert wird, woraus als Schätzer $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$ resultiert.

Die *fitted values* haben somit die Form $\hat{\boldsymbol{\mu}} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$ und fassen wir anschließend noch $\mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$ zur $n \times n$-Hat-Matrix zusammen, die den Vektor der Responses auf die fitted values abbildet, dann ergibt sich

$$\hat{\boldsymbol{\mu}} = \mathbf{H}\mathbf{y}.$$

Eine bekannte Eigenschaft ist, dass die Spur der Hat-Matrix $\operatorname{tr}(\mathbf{H}) = \operatorname{tr}(\mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top) = \operatorname{tr}((\mathbf{X}^\top \mathbf{X})^{-1}(\mathbf{X}^\top \mathbf{X})) = \operatorname{tr}(\mathbf{I}_{p+1}) = p + 1$ der Anzahl der Parameter im Modell entspricht.

Da die fitted values von den tatsächlichen Werten abweichen können, geben die Residuen

$$r_i = y_i - \hat{\mu}_i$$

Auskunft über die Abweichung und die Adäquatheit des Modells an der Stelle x_i.

Die im Linearen Regressionsmodell getroffenen Annahmen sind nicht für jede Datensituation geeignet. Für das Ziel der vorliegenden Arbeit, Mortalitätsraten bzw. Sterbewahrscheinlichkeiten zu modellieren, sind Lineare Modelle nicht passend. Einerseits ist die Annahme der Normalverteilung der Responses nicht gerechtfertigt (siehe Abbildung 2.4) und andererseits können die mit einem Linearen Modell geschätzten Erwartungswerte im Wertebereich $(-\infty, +\infty)$ liegen, was insbesonders für die Schätzung von Sterblichkeitsmaßen nicht sinnvoll ist.

Als erklärende Variable ist hinsichtlich der Mortalitätsanalyse, wie im vorangegangenen Kapitel besprochen, vor allem das Alter der Versicherten von Relevanz. Wie in Abbildung 2.5 zu erkennen, ist der unter dem Linearen Modell postulierte lineare Zusammenhang zwischen der erwarteten Sterblichkeit und dem Alter keineswegs gegeben.

Um den Anforderungen und den der Arbeit zugrunde liegenden Daten besser Rechnung tragen zu können, werden in weiterer Folge Modelle vorgestellt, die eine Verallgemeinerung der Linearen Modelle darstellen und eine flexiblere Modellierung erlauben.

3.2 Generalisierte Lineare Modelle

Auf Grund der im vorangegangenen Abschnitt angeführten Restriktionen des Linearen Regressionsmodells, wird nun die Modellklasse der Generalisierten Linearen Modelle eingeführt, die gegenüber den Linearen Modellen hinsichtlich der Verteilungsannahme und der Modellierung des Erwartungswertes deutlich mehr Flexibilität zulassen.

3.2.1 Die Lineare Exponentialfamilie

Bevor das GLM definiert wird, muss der Begriff der linearen Exponentialfamilie besprochen werden. Eine Zufallsvariable y entstammt der linearen Exponentialfamilie, falls ihre Wahrscheinlichkeits- oder Dichtefunktion die folgende Gestalt annimmt (vgl. McCullagh & Nelder, 1989):

$$f(y|\theta) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right), \tag{3.1}$$

wobei die Funktionen $a(\cdot), b(\cdot)$ und $c(\cdot)$ bekannt sind und $a(\phi) > 0$ gilt. Ist ϕ eine feste, bekannte Größe, so sprechen wir von einer einparametrigen linearen Exponentialfamilie mit Parameter θ.

Für alle Mitglieder der linearen Exponentialfamilie gilt für die ersten beiden Momente

$$\begin{aligned} \mathrm{E}(y) &= \mu = b'(\theta), \\ \mathrm{Var}(y) &= a(\phi)b''(\theta) = a(\phi)V(\mu). \end{aligned}$$

Es ist ersichtlich, dass sich die Varianz als Produkt zweier Funktionen ergibt. Eine, die ausschließlich vom Parameter θ und somit auch von μ abhängt und eine, die alleinig vom Dispersionsparameter ϕ abhängig ist. Die Funktion $b''(\theta)$ in Termen von μ betrachtet wird mit $V(\mu)$ bezeichnet. Wir nehmen im Folgenden an, dass $a_i(\phi) = a_i\phi$, wobei der Dispersionsparameter ϕ konstant ist und die Gewichte a_i hingegen für jede Beobachtung unterschiedlich sein können.

Sehr viele bekannte Verteilungen, wie die Normalverteilung, Gammaverteilung, Binomialverteilung und die Poissonverteilung sind Mitglieder der linearen Exponentialfamilie. In der Folge sollen zwei für die vorliegende Arbeit relevante Verteilungen besprochen werden.

- **Poissonverteilung**: Es sei $y \sim Poi(\mu)$, dann gilt für die Wahrscheinlichkeitsfunktion
$$f(y|\mu) = \frac{\mu^y}{y!}e^{-\mu} = \exp(y\log\mu - \mu - \log y!), \qquad y \in \{0,1,2,\dots\}.$$
Wir erhalten mit $\theta = \log\mu$ und festem Dispersionsparameter $\phi = 1$ die Darstellung wie in (3.1), wobei $a = 1$, $b(\theta) = \exp(\theta)$ und $c(y,\phi) = -\log y!$. Die Poissonverteilung ist somit Mitglied der Exponentialfamilie.

- **Standardisierte Binomialverteilung**: Es sei $my \sim B(m, \pi)$ (y bezeichnet die relative Häufigkeit eines Erfolges bei m Versuchen), dann erhalten wir für die Wahrscheinlichkeitsfunktion die Form

$$\begin{aligned} f(y|m,\pi) = P(mY = my) &= \binom{m}{my}\pi^{my}(1-\pi)^{m-my} \\ &= \binom{m}{my}\left(\frac{\pi}{1-\pi}\right)^{my}(1-\pi)^{m} \\ &= \exp\left(\frac{y\log\frac{\pi}{1-\pi} - \log\frac{1}{1-\pi}}{1/m} + \log\binom{m}{my}\right), \end{aligned}$$

  $y \in \{0, \frac{1}{m}, \frac{2}{m}, \ldots, 1\}$.

  Wiederum ergibt sich mit $\theta = \log\frac{\pi}{1-\pi}$ und festem Dispersionsparameter $\phi = 1$ die Darstellung (3.1), mit $a = \frac{1}{m}$, $b(\theta) = \log(1 + e^{\theta})$ und $c(y, \phi) = \log\binom{m}{my}$. Auch die standardisierte Binomialverteilung gehört also der linearen Exponentialfamilie an.

3.2.2 Das Generalisierte Lineare Modell

Generalisierte Lineare Modelle (GLMe) setzen sich aus drei Modellkomponenten zusammen (vgl. McCullagh & Nelder, 1989):

- **Zufallskomponente**: Es wird angenommen, dass die Responses $y_i, i \in \{1, \ldots, n\}$, unabhängig sind und einer Verteilung mit Parameter θ_i entstammen, die Mitglied der linearen Exponentialfamilie ist. Der Erwartungswert der Response wird mit $\mathrm{E}(y_i) = \mu_i = \mu(\theta_i)$ bezeichnet.

- **Systematische Komponente**: Die Kovariablen $\mathbf{x}_l = (x_{1l}, \ldots, x_{nl})^{\top}$, $l \in \{1, \ldots, p\}$, ergeben den linearen Prädiktor $\boldsymbol{\eta}$ mit $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$, wobei $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^{\top}$ den zu schätzenden Parametervektor und $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)^{\top}$ den Vektor der linearen Prädiktoren darstellt. Die $n \times (p+1)$-Designmatrix $\mathbf{X}$ fasst die Kovariablen $\mathbf{x}_l$ zusammen und hat die Gestalt

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \ldots & x_{np} \end{pmatrix}.$$

- **Link-Komponente**: Die Verknüpfung der Zufalls- und systematischen Komponente erfolgt über $g(\mu_i) = \eta_i$. Das heißt, der Erwartungswert wird nicht so wie im Linearen Modell linear modelliert, sondern eine Funktion des Erwartungswertes $g(\mu)$, die monoton und zweimal stetig differenzierbar ist, wird linear modelliert.

Wird die Link-Funktion so spezifiziert, dass $\theta_i = \eta_i$ für alle i hält, also auf Grund der Beziehung $\theta_i = b'^{-1}(\mu_i) = b'^{-1}(g^{-1}(\eta_i))$ die Wahl $g(\cdot) = b'^{-1}(\cdot)$ betrachtet wird, dann sprechen wir von einer kanonischen Link-Funktion. Für die beiden im Abschnitt 3.2.1 besprochenen Verteilungen ergeben sich folgende kanonische Link-Funktionen:

- **Poissonverteilung**: $\eta = \log(\mu)$,
- **Standardisierte Binomialverteilung**: $\eta = \log(\frac{\mu}{1-\mu})$.

Über einen Iterativen (Re)Weighted Least Squares Algorithmus (IWLS), der auf der Newton-Raphson-Methode oder der Fisher-Scoring-Methode basiert, erfolgt die Berechnung des Maximum-Likelihood-Schätzers für den Parametervektor $\boldsymbol{\beta}$, da die Likelihood-Gleichungen im Allgemeinen nichtlinear in $\boldsymbol{\beta}$ sind. Auf die genaue Herleitung des Algorithmus wird an dieser Stelle nicht eingegangen, diese kann in McCullagh und Nelder (1989) eingesehen werden.

3.2.3 Bewertung der Modellgüte und Modellvergleich

Die Güte der Anpassung eines GLMs kann mittels der Deviance $\mathrm{D}(\mathbf{y}, \hat{\boldsymbol{\mu}})$ gemessen werden, die die Likelihood-Quotienten-Teststatistik zu den Hypothesen H_0: $\boldsymbol{\mu} = g^{-1}(\mathbf{x}^\top \boldsymbol{\beta})$ gegen H_1: $\boldsymbol{\mu} \neq g^{-1}(\mathbf{x}^\top \boldsymbol{\beta})$ darstellt und als negativer doppelter Logarithmus des Quotienten zweier maximierter Likelihood-Funktionen definiert ist:

$$\frac{1}{\phi}\mathrm{D}(\mathbf{y}, \hat{\boldsymbol{\mu}}) = -2\left(l(\hat{\boldsymbol{\mu}}|\mathbf{y}) - l(\mathbf{y}|\mathbf{y})\right) \overset{approx}{\sim} \chi^2_{n-(p+1)} \ ,$$

falls ϕ bekannt ist. Dabei bezeichnet $l(\hat{\boldsymbol{\mu}}|\mathbf{y})$ die maximierte Log-Likelihood-Funktion des unter H_0 betrachteten Modells und $l(\mathbf{y}|\mathbf{y})$ die maximierte Log-Likelihood-Funktion eines Modells ohne vorgegebener Struktur bezüglich dem Erwartungswert (volles oder saturiertes Modell), dessen Likelihood-Funktion folglich für $\hat{\mu}_i = y_i$ für alle i maximal wird.

Für die beiden im Abschnitt 3.2.1 betrachteten Verteilungen ergibt sich für die Deviance:

- **Poissonverteilung**:

$$\mathrm{D}(\mathbf{y}, \hat{\boldsymbol{\mu}}) = 2\sum_{i=1}^{n}\left(y_i \log\frac{y_i}{\hat{\mu}_i} - (y_i - \hat{\mu}_i)\right),$$

- **Standardisierte Binomialverteilung**:

$$\mathrm{D}(\mathbf{y}, \hat{\boldsymbol{\mu}}) = 2\sum_{i=1}^{n} m_i\left((1-y_i)\log\frac{1-y_i}{1-\hat{\mu}_i} + y_i \log\frac{y_i}{\hat{\mu}_i}\right).$$

Wie oben ersichtlich, kann die Deviance für die Poissonverteilung bzw. die standardisierte Binomialverteilung durch $\mathrm{D}(\mathbf{y}, \hat{\boldsymbol{\mu}}) = \sum_{i=1}^{n} d_i$ ausgedrückt werden, was zur Definition der Deviance Residuen

$$r_i^D = \mathrm{sign}(y_i - \hat{\mu}_i)\sqrt{d_i}$$

führt.

Ein weiteres Gütemaß der Modellanpassung ist die Pearson Statistik X^2, die durch

$$X^2 = \sum_{i=1}^{n}\frac{(y_i - \hat{\mu}_i)^2}{a_i V(\hat{\mu}_i)}$$

definiert ist und die auch als Schätzer für den manchmal unbekannten Dispersionsparameter ϕ mittels

$$\hat{\phi}_P = \frac{1}{n-(p+1)} X^2$$

herangezogen werden kann. Außerdem leiten sich aus den Summanden der Pearson Statistik die Pearson Residuen

$$r_i^P = \frac{(y_i - \hat{\mu}_i)}{\sqrt{a_i V(\hat{\mu}_i)}}$$

ab. Anstatt der Pearson Statistik kann auch die gemittelte Deviance

$$\hat{\phi}_D = \frac{1}{n-(p+1)} \mathrm{D}(\mathbf{y}, \hat{\boldsymbol{\mu}})$$

zur Schätzung von ϕ verwendet werden.

Für den Modellvergleich ineinander geschachtelter Modelle (*nested models*) kann der Likelihood-Quotienten-Test (LQ-Test) zu den Hypothesen H_0: $\beta_q = \cdots = \beta_p = 0$ gegen H_1: $\beta_0, \ldots, \beta_p$ *beliebig* $(q < p)$ mit der Teststatistik

$$lq = \frac{1}{\phi}\left(D(\mathbf{y}, \hat{\boldsymbol{\mu}}) - D(\mathbf{y}, \tilde{\boldsymbol{\mu}})\right) = -2(l(\hat{\boldsymbol{\mu}}|\mathbf{y}) - l(\tilde{\boldsymbol{\mu}}|\mathbf{y})),$$

wobei $l(\hat{\boldsymbol{\mu}}|\mathbf{y})$ und $l(\tilde{\boldsymbol{\mu}}|\mathbf{y})$ die unter dem in H_0 bzw. H_1 betrachteten Modell maximierten Log-Likelihood-Funktionen darstellen, verwendet werden. Unter H_0 und bei bekanntem Dispersionsparameter ist lq asymptotisch $\chi^2_{p-q+1}-$verteilt.

Zum Vergleich von Modellen mit unterschiedlichen Prädiktoren und Parametern ist Akaike's Informationskriterium (AIC) ein häufig verwendetes Kriterium

$$\mathrm{AIC} = -2l(\hat{\boldsymbol{\mu}}|\mathbf{y}) + 2k,$$

und k bezeichnet die Anzahl der im Modell betrachteten Parameter. Falls ϕ bekannt ist, nimmt k demnach den Wert $p+1$ an. Ist ϕ unbekannt, gilt $k = p+2$. Alternativ kann auch das Bayesianische Informationskriterium (BIC) zur Modellwahl verwendet werden

$$\mathrm{BIC} = -2l(\hat{\boldsymbol{\mu}}|\mathbf{y}) + \log(n)k,$$

das auf Grund seiner Definition Modellkomplexität höher bestraft als das AIC.

3.3 Modelle zur Modellierung der Mortalität

3.3.1 Poisson-Regression für Anzahlen und Raten

Ein sehr häufig verwendetes Modell, um Zählvariablen zu modellieren ist das loglineare Poisson-Modell (vgl. McCullagh & Nelder, 1989 und Agresti, 2002). Dabei wird angenommen, dass die Response-Variablen y_i die Werte aus $\{0, 1, 2, \ldots\}$ annehmen, einer Poissonverteilung mit Erwartungswert $\mu_i > 0$ folgen und $\mathbf{X}$ sei die Designmatrix, die die Kovariablen-Vektoren zusammenfasst. Das loglineare Poisson-Modell hat schließlich folgende Gestalt

$$\log \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}, \tag{3.2}$$

womit sichergestellt wird, dass der Erwartungswert stets positiv ist.

Für das Ziel dieser Arbeit, Sterberaten bzw. Sterbewahrscheinlichkeiten zu modellieren ist es jedoch sinnvoller nicht reine Zählvariablen zu betrachten, sondern diese wie im Abschnitt 2.2.3 beschrieben, in Relation zu den verlebten Personenjahren zu setzen. Im Falle der Mortalitätsanalyse entsprechen die Responses der Anzahl der Sterbefälle $D_{x,t}$ innerhalb eines Beobachtungsjahres t und im Alter x. Des Weiteren seien die verlebten Personenjahre als Approximation der innerhalb des beobachteten Jahres dem Mortalitätsrisiko unterliegenden Personen mit $E_{x,t}$ gegeben und werden als feste Größe angesehen. $D_{x,t}$ ist poissonverteilt mit Erwartungswert $\mu_{x,t}E_{x,t}$. Der Einfachheit halber wird in der Folge das Beobachtungsjahr t nicht in der Notation berücksichtigt, da in diesem Abschnitt ohnehin die eindimensionale Glättung im Vordergrund steht und das beobachtete Jahr als fest anzusehen ist.

Das loglineare Modell für die erwartete Mortalitätsrate hat nun die Form

$$\begin{aligned} \log \frac{\mathrm{E}(D_x)}{E_x} &= \log \mu_x = \beta_0 + \beta_1 x \\ \log \mathrm{E}(D_x) &= \log E_x + \log \mu_x = \beta_0 + \beta_1 x + \log E_x. \end{aligned} \tag{3.3}$$

Der Term $\log E_x$ ist eine feste Größe und wird als *Offset* bezeichnet, dessen Koeffizient bei der Parameterschätzung konstant bei Eins gehalten wird.

Das soeben vorgestellte Poisson-Modell für die Sterbeintensität korrespondiert auch zu dem von Gompertz im Jahre 1825 erkannten Verlauf der momentanen Sterberate in Abhängigkeit vom Alter. Wird μ_x auf der logarithmischen Skala aufgetragen, so ist auch anhand der vorliegenden Daten festzustellen, dass zumindest im Altersbereich $[45, 95]$ ein linearer Zusammenhang zwischen dem Alter und der Sterbeintensität vorliegt (siehe Abbildung 3.1). Daraus leitet sich das Gompertz-Modell in seiner einfachsten Form

$$\mu_x = e^{\beta_0 + \beta_1 x} \tag{3.4}$$

ab und stimmt sichtlich mit (3.3) überein.

Für die Spezifikation des Modells in R wird das loglineare Modell mit dem Alter als erklärende Variable an dieser Stelle noch in Matrix-Notation eingeführt. Dazu sei n_a die Anzahl der betrachteten Alter, $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_{n_a})^\top$ der Prädiktorvektor, $\mathbf{e} = (e_1, \ldots, e_{n_a})^\top$ der Vektor, der die Anzahl der dem Mortalitätsrisiko ausgesetzten Personen enthält, $\mathbf{x} = (x_1, \ldots, x_{n_a})^\top$ der Vektor der erklärenden Variablen (Alter) und $\boldsymbol{\beta} = (\beta_0, \beta_1)^\top$ sei der Parametervektor. Mit $\mathbf{1}_{n_a}$ wird der Vektor der Länge n_a bezeichnet, der nur aus Einsen besteht. Dann erhalten wir

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \log \mathbf{e}, \qquad \mathbf{X} = [\mathbf{1}_{n_a} : \mathbf{x}]. \tag{3.5}$$

Im Folgenden halten wir das Jahr 2015 fest und verwenden die Funktion glm, um das obige Modell in R zu fitten. Dazu betrachten wir weibliche Versicherte in einem Altersbereich von 45 bis 95 Jahren. Die entsprechenden Werte für die Anzahl der Verstorbenen und der dem Risiko des Todes unterliegenden Personen im gewählten Altersbereich bzw. Jahr können über die Funktion data.select, deren Code im Appendix A in Listing A.4 einsehbar

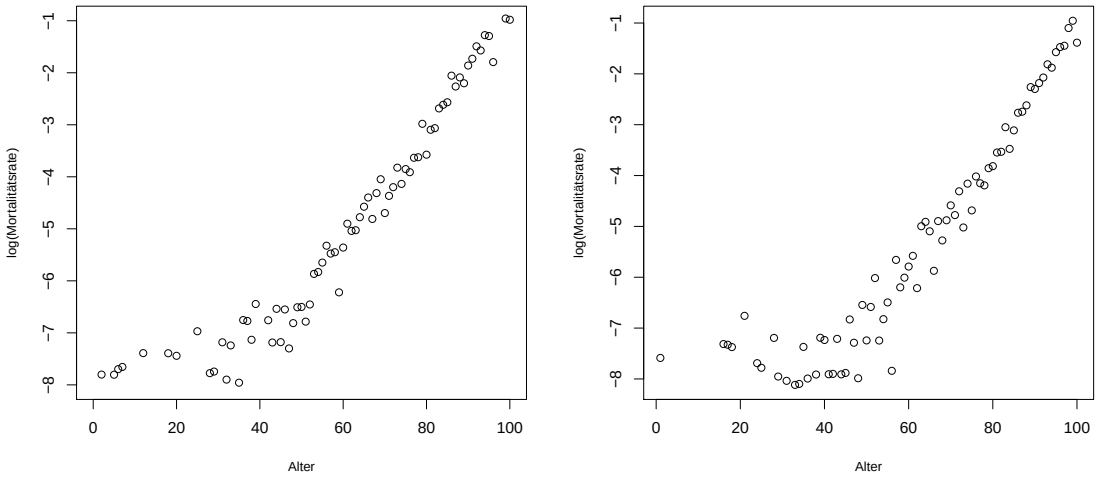

Abbildung 3.1: Beobachtete Mortalitätsraten männlicher (links) bzw. weiblicher (rechts) Versicherter im Jahr 2015 auf der log-Skala aufgetragen.

ist, selektiert werden.

```
> sex<-"W"
> Jahr<-2015
> Alter<-seq(45,95,1)
> source("C:/Matrizen_Aufbereitung.R")
> deaths<-data.select(Jahr,Alter,sex,"Deaths")
> exposures<-data.select(Jahr,Alter,sex,"Exposures")
```

Die unter diesem Modell geschätzten Mortalitätsraten sind in Abbildung 3.2 veranschaulicht und der summary-Befehl angewandt auf das angepasste Poisson-Modell model_poi liefert das folgende Output:

```
> model_poi<-glm(deaths~Alter+offset(log(exposures)),
+ family=poisson(link="log"))
> summary(model_poi)

Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) -13.394865   0.245184  -54.63   <2e-16 ***
Alter         0.123465   0.002958   41.74   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 2536.68  on 50  degrees of freedom
Residual deviance:   71.99  on 49  degrees of freedom
AIC: 299.47
```

Daran ist zu erkennen, dass der Effekt des Alters hochsignifikant ist. Allerdings weisen eine Deviance von 71.99 bei 49 Freiheitsgraden sowie der kritische Wert der Chiquadrat-Verteilung $\chi^2_{49;0.95} = 66.34$ darauf hin, dass das gewählte Modell nicht optimal ist. Außerdem ergibt die Pearson Statistik dividiert durch die Freiheitsgrade sowie die gemittelte Deviance als Schätzer für den Dispersionsparameter ϕ einen Wert größer als Eins, weshalb Überdispersion zu vermuten ist.

```
> pearson_res<-residuals(model_poi, "pearson")
> X2<-sum(pearson_res^2)
> X2/df.residual(model_poi)
[1] 1.5278
> model_poi$deviance/df.residual(model_poi)
[1] 1.469189
```

In weiterer Folge soll getestet werden, ob die fehlende Modellanpassung auf wirkliche Überdispersion (*real overdispersion*) zurückzuführen ist und somit ein zum Poisson-Modell alternatives Modell sinnvoll ist oder in scheinbarer Überdispersion (*apparent overdispersion*) begründet ist, also Interaktionsterme fehlen oder nichtlineare Effekte berücksichtigt werden müssen (vgl. Hilbe, 2011, S.141-142).

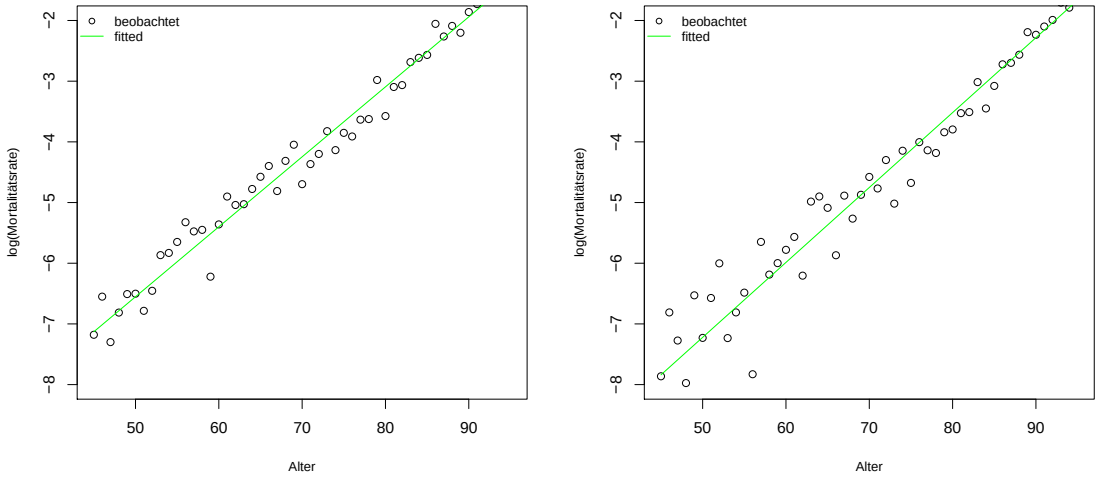

Abbildung 3.2: Beobachtete und geschätzte Mortalitätsraten männlicher (links) bzw. weiblicher (rechts) Versicherter im Jahr 2015 auf der log-Skala aufgetragen.

Wir wollen als ersten Schritt prüfen, ob apparent overdispersion vorliegend ist und untersuchen, ob der Einfluss des Alters nichtlinear ist, weshalb wir einen quadratischen Term in das Modell aufnehmen. Die Designmatrix ändert sich dementsprechend in

$$\mathbf{X} = [\mathbf{1}_{n_a} : \mathbf{x} : \mathbf{x}^2].$$

```
> model_poi_sq<-glm(deaths~Alter+I(Alter^2)+
+ offset(log(exposures)),family=poisson(link="log"))
> summary(model_poi_sq)

Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.5032272  1.1724312  -7.253 4.09e-13 ***
Alter       -0.0078540  0.0312944  -0.251    0.802
I(Alter^2)   0.0008594  0.0002054   4.184 2.86e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 2536.683  on 50  degrees of freedom
Residual deviance:   55.629  on 48  degrees of freedom
AIC: 285.11
```

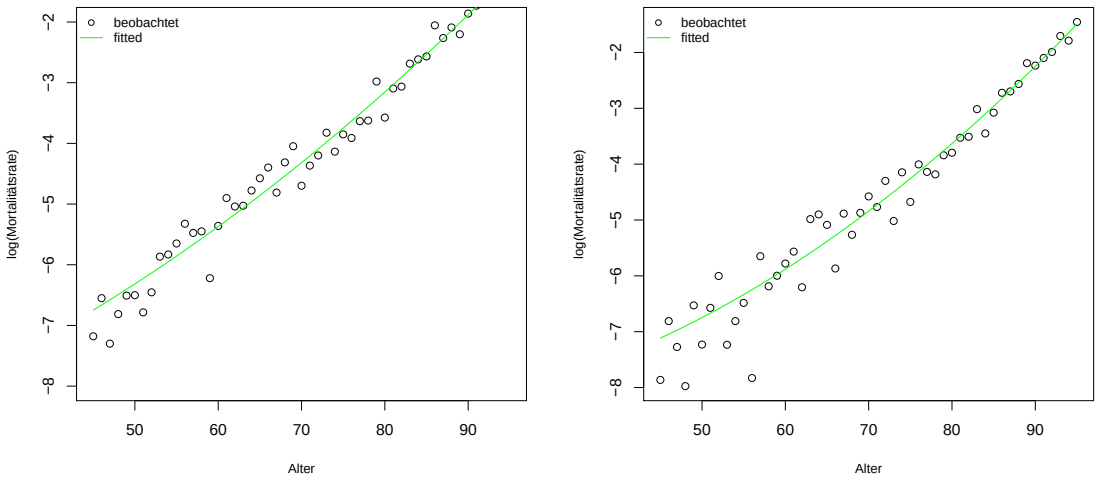

Abbildung 3.3: Unter dem Poisson-Modell mit quadratischem Alterseffekt geschätzte Mortalitätsraten männlicher (links) und weiblicher (rechts) Versicherter im Jahr 2015 auf der log-Skala aufgetragen.

Wie anhand des obigen R-Outputs und der Abbildung 3.3 zu erkennen ist, hat sich die Modellanpassung gegenüber dem Poisson-Modell, das lediglich den Haupteffekt (Alter) enthält, deutlich verbessert. Der quadratische Effekt des Alters ist hochsignifikant und auch die Pearson Statistik dividiert durch die Freiheitsgrade hat sich auf 1.14 verringert.

3.3.2 Modellierung von Überdispersion

Das Phänomen der Überdispersion, also dass die tatsächliche Varianz die unter dem Poisson-Modell erklärte Varianz übersteigt, kommt insbesonders bei Zähldaten häufig

vor. Grund dafür ist beispielsweise nicht beobachtbare Heterogenität in den Daten, die aus nicht vorhandenen bzw. nicht messbaren erklärenden Größen resultiert (siehe Denuit, Maréchal, Pitrebois & Walhin, 2007, S.79-80). Damit ist die beim Poisson-Modell grundlegende Annahme $\mathrm{Var}(y) = \mathrm{E}(y) = \mu$ in diesem Fall nicht gerechtfertigt. Als Folge von Überdispersion werden die Standardfehler der Parameterschätzer unterschätzt und damit die Signifikanz der Regressionsparameter überschätzt. Im Folgenden werden zwei Modelle betrachtet, die der vorherrschenden Überdispersion Rechnung tragen.

Quasi-Poisson-Modell

Zunächst betrachten wir die Quasi-Likelihood-Methode, die auf dem GLM-Regelwerk beruht und eine Parameterschätzung, die rein nur auf einer Spezifikation der Erwartungswert-Varianz-Beziehung der Responses basiert, zulässt.

Dazu fixieren wir den Dispersionsparameter nicht wie unter dem Poisson-Modell bei $\phi = 1$, sondern schätzen ihn durch die Daten. Zusätzlich nehmen wir an, dass die Varianz proportional zum Erwartungswert ist, also

$$\mathrm{Var}(y) = \phi \mathrm{E}(y).$$

Die Spezifikation der Beziehung zwischen dem Erwartungswert und der Varianz und die anschließende Anwendung eines modifizierten IWLS-Algorithmus liefern als Ergebnis den Maximum-Quasi-Likelihood-Schätzer für $\boldsymbol{\beta}$. Da das Bestimmen der Parameterschätzer über den IWLS-Algorithmus nicht vom Dispersionsparameter abhängt (McCullagh & Nelder, 1989, S.328), ändern sich die geschätzten Werte für den Parametervektor gegenüber dem Poisson-Modell nicht. Die Standardfehler von $\hat{\boldsymbol{\beta}}$ hängen hingegen von ϕ ab und ändern sich dementsprechend um den Faktor $\hat{\phi}^{1/2}$.

Wie schon oben erwähnt, bildet die mittlere Pearson Statistik X^2 einen Schätzwert für den Dispersionsparameter ϕ. Im Beispiel zum Poisson-Modell (model_poi_sq) ergibt sich $X^2 = 54.80$ bei $n = 51$ Beobachtungen und $p = 3$ Parametern. Damit erhöhen sich die Standardfehler im obigen R-Output um den Faktor $\hat{\phi}_P^{1/2} = \sqrt{54.80/48} = 1.07$. Es ist ersichtlich, dass mit $\hat{\phi}_P = 1.14$ der Schätzer des Dispersionsparameters nahe bei Eins liegt und sich aus diesem Grund die Standardfehler nur geringfügig ändern und damit einhergehend sich an der Signifikanz der Parameter keine starken Änderungen ergeben. Dieses Verhalten kann auch im nachstehenden R-Output beobachtet werden:

```
> pr<-residuals(model_poi_sq, "pearson")
> sum(pr^2)
[1] 54.79576

> model_quasi<-glm(deaths~Alter+I(Alter^2)+
+ offset(log(exposures)),family=quasipoisson)
> summary(model_quasi)

Coefficients:
Estimate Std. Error t value Pr(>|t|)
```

```
12 (Intercept) -8.5032272  1.2526856  -6.788 1.55e-08 ***
13 Alter       -0.0078540  0.0334366  -0.235 0.815290
14 I(Alter^2)   0.0008594  0.0002194   3.916 0.000284 ***
15 ---
16 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

18 (Dispersion parameter for quasipoisson family taken to be
       1.141588)

20 Null deviance: 2536.683  on 50  degrees of freedom
21 Residual deviance:   55.629  on 48  degrees of freedom
22 AIC: NA
```

Negativ-Binomial-Modell

Eine weitere Möglichkeit Überdispersion zu modellieren bietet das Negativ-Binomial-Modell, bei dem für die Response y eine Negativ-Binomial-Verteilung angenommen wird. Es gibt unterschiedliche Möglichkeiten die Negativ-Binomial-Verteilung zu definieren. In diesem Abschnitt wird auf das in der Literatur häufig als NB2-Modell (vgl. Cameron & Trivedi, 1998) bezeichnete Modell eingegangen. Dabei gehen wir von einem Poisson-Modell aus und fügen einen multiplikativen Zufallseffekt τ hinzu, der die unbeobachtete Heterogenität in den Daten modellieren soll. Die dazu folgenden Ausführungen orientieren sich an Cameron und Trivedi (1998), Zeileis, Kleiber und Jackman (2008) und Hilbe (2011).

Die Negativ-Binomial-Verteilung kann aus einer Poissonverteilung, deren Parameter selbst eine Zufallsvariable ist und einer Gammaverteilung unterliegt, hergeleitet werden. Es sei dazu τ eine Zufallsvariable, die einer Gammaverteilung $\gamma(\Theta, \Theta)$ mit $\Theta > 0$ folgt und für den Erwartungswert und die Varianz somit

$$\mathrm{E}(\tau) = 1 \qquad \text{und} \qquad \mathrm{Var}(\tau) = \frac{1}{\Theta}$$

gilt. Angenommen $y|\tau \sim Poi(\mu\tau)$, so kann daraus die marginale Wahrscheinlichkeitsfunktion von y hergeleitet werden, die, wie in der nachstehenden Herleitung zu sehen ist, der Negativ-Binomial-Verteilung $NB(\Theta, \frac{\Theta}{\mu+\Theta})$ entspricht. Unter diesem Modell folgt für $y \in \{0, 1, 2, \dots\}$

$$\begin{aligned} P(Y = y) &= \int_0^\infty \frac{e^{-\mu\tau}(\mu\tau)^y}{y!} \frac{\Theta^\Theta}{\Gamma(\Theta)} \tau^{(\Theta-1)} e^{-\Theta\tau} d\tau \\ &= \frac{\Theta^\Theta \mu^y}{y!\Gamma(\Theta)} \frac{\Gamma(y+\Theta)}{(\mu+\Theta)^{y+\Theta}} \\ &= \frac{\Gamma(y+\Theta)}{\Gamma(y+1)\Gamma(\Theta)} \left(\frac{\Theta}{\mu+\Theta}\right)^\Theta \left(\frac{\mu}{\mu+\Theta}\right)^y \\ &= \binom{y+\Theta-1}{y} \left(\frac{\Theta}{\mu+\Theta}\right)^\Theta \left(\frac{\mu}{\mu+\Theta}\right)^y . \end{aligned} \tag{3.6}$$

Es ist zu beachten, dass im zweiten Schritt das Integral gleich Eins ist, da der Integrand die Dichtefunktion der Gammaverteilung $\gamma(\Theta + y, \mu + \Theta)$ ist. Gegenüber dem Poisson-Modell

haben wir im Negativ-Binomial-Modell also nur einen zusätzlichen Parameter, nämlich den Überdispersionsparameter Θ. Für den Erwartungswert von y ergibt sich

$$\mathrm{E}(y) = \mathrm{E}(\mathrm{E}(y|\tau)) = \mathrm{E}(\mu\tau) = \mu\mathrm{E}(\tau) = \mu$$

und für die Varianz auf Grund der Varianzzerlegung

$$\begin{aligned}\mathrm{Var}(y) &= \mathrm{E}(\mathrm{Var}(y|\tau)) + \mathrm{Var}(\mathrm{E}(y|\tau)) \\ &= \mathrm{E}(\mu\tau) + \mathrm{Var}(\mu\tau) = \mu + \mu^2\frac{1}{\Theta} \\ &= \mu(1 + \mu/\Theta). \end{aligned} \tag{3.7}$$

Für festes Θ kann das Negativ-Binomial-Modell in das Regelwerk der GLMe eingebettet werden und deren Theorie zur Schätzung der Parameter angewandt werden. Ist der Überdispersionsparameter nämlich bekannt, so kann (3.6) weiter umgeschrieben werden in

$$P(Y = y) = \exp\left(y\log\left(\frac{\mu}{\mu+\Theta}\right) - (-\Theta)\log\left(\frac{\Theta}{\mu+\Theta}\right) + \log\binom{y+\Theta-1}{y}\right),$$

sodass die Form (3.1) erkennbar ist. Die Negativ-Binomial-Verteilung gehört also für bekanntes Θ zur einparametrigen linearen Exponentialfamilie mit $\phi = 1$, $\theta = \log\left(\frac{\mu}{\mu+\Theta}\right)$, $b(\theta) = (-\Theta)\log\left(1 - e^{\theta}\right)$ und $c(y,\phi) = \log\binom{y+\Theta-1}{y}$.

In der Praxis ist jedoch Θ meist unbekannt und muss aus den Daten geschätzt werden, weshalb es sich in diesem Fall beim Negativ-Binomial-Modell nicht mehr um einen Spezialfall eines GLMs handelt. Dennoch können GLM-Methoden für die Maximum-Likelihood-Schätzung verwendet werden. Das R-Paket MASS stellt die Funktion glm.nb zur Verfügung, durch deren Anwendung $\boldsymbol{\beta}$ iterativ geschätzt wird, indem in den Iterationsschritten Θ festgehalten wird und umgekehrt. Somit ergeben sich die Maximum-Likelihood-Schätzer sowohl für $\boldsymbol{\beta}$, als auch für Θ. Als Link-Funktion dient der Log-Link, der sich vor allem für Vergleichszwecke zum Poisson-Modell für Anzahlen als Link eignet.

Wird die Substitution $\alpha = 1/\Theta$ angewandt, so erhalten wir in (3.7) die Darstellung $\mathrm{Var}(y) = \mu + \mu^2\alpha$, wodurch sich das NB2-Modell mit Log-Link und $\alpha = 0$ zum Poisson-Modell reduziert. Um zu testen, ob das Poisson-Modell für die Datensituation adäquat ist, kann ein LQ-Test herangezogen werden und die Nullhypothese H_0: $\alpha = 0$ gegen die Alternative H_1: $\alpha > 0$ getestet werden. Die LQ-Teststatistik ist definiert durch

$$LQ = -2(l_{poi} - l_{nb2}), \tag{3.8}$$

wobei l_{poi} das Maximum der Log-Likelihood-Funktion des Poisson-Modells bezeichnet und l_{nb2} jenes der Log-Likelihood-Funktion des allgemeineren Negativ-Binomial-Modells. Da sich bei diesem Test die Problematik eines beschränkten Parameterraums ergibt, weil α den Wert Null nicht unterschreiten darf, handelt es sich bei der LQ-Teststatistik um keine χ^2_1-verteilte Größe. Simulationen haben ergeben, dass eine 50:50 Mischung von Null und einer χ^2_1-Verteilung der asymptotischen Verteilung der LQ-Teststatistik entspricht. Eine konservative Sichtweise stellt dennoch die Betrachtung von LQ als χ^2_1-verteilte Größe dar,

womit die Nullhypothese verworfen wird, falls $LQ > \chi^2_{1-\delta,1}$, wobei $1-\delta$ das Konfidenzniveau bezeichnet.

Wir wollen nun wieder ein größeres Augenmerk auf die vorliegenden Daten werfen und anhand des oben beschriebenen LQ-Tests überprüfen, ob ein Negativ-Binomial-Modell dem Poisson-Modell (3.3) vorzuziehen ist. Dazu betrachten wir wiederum weibliche Versicherte im Altersintervall $[45, 95]$ und berechnen die Teststatistik (3.8):

```
> library(MASS)
> model_nb<-glm.nb(deaths~Alter+offset(log(exposures)),
+ link=log)
> summary(model_nb)

Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) -13.203694   0.275569  -47.91   <2e-16 ***
Alter         0.121103   0.003408   35.54   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(58.4249) family taken
    to be 1)

Null deviance: 1924.597  on 50  degrees of freedom
Residual deviance:   54.863  on 49  degrees of freedom
AIC: 299.06

Theta:  58.4
Std. Err.:  45.9

2 x log-likelihood:  -293.064
```

```
> logLik(model_poi)
'log Lik.' -147.7348 (df=2)
> logLik(model_nb)
'log Lik.' -146.5319 (df=3)
> as.numeric(-2*(logLik(model_poi)-logLik(model_nb)))
[1] 2.405984
> qchisq(1-0.05, 1)
[1] 3.841459
```

Aus der berechneten LQ-Teststatistik ist abzulesen, dass α bei einem Konfidenzniveau von 95% nicht signifikant verschieden von Null ist. Wie zuvor beschrieben, handelt es sich dabei um eine konservative Betrachtungsweise, weshalb wir nun den kritischen Wert $\chi^2_{1-2\delta,1}$ für $\delta = 0.05$ betrachten (vgl. Cameron & Trivedi, 1998, S.78):

```
> qchisq(1-0.1,1)
[1] 2.705543
```

Zu einem Konfidenzniveau von 90% kann $\alpha = 0$ gerade nicht verworfen werden. In Abbildung 3.4 sind die mit dem NB2-Modell geschätzten Mortalitätsraten abgebildet.

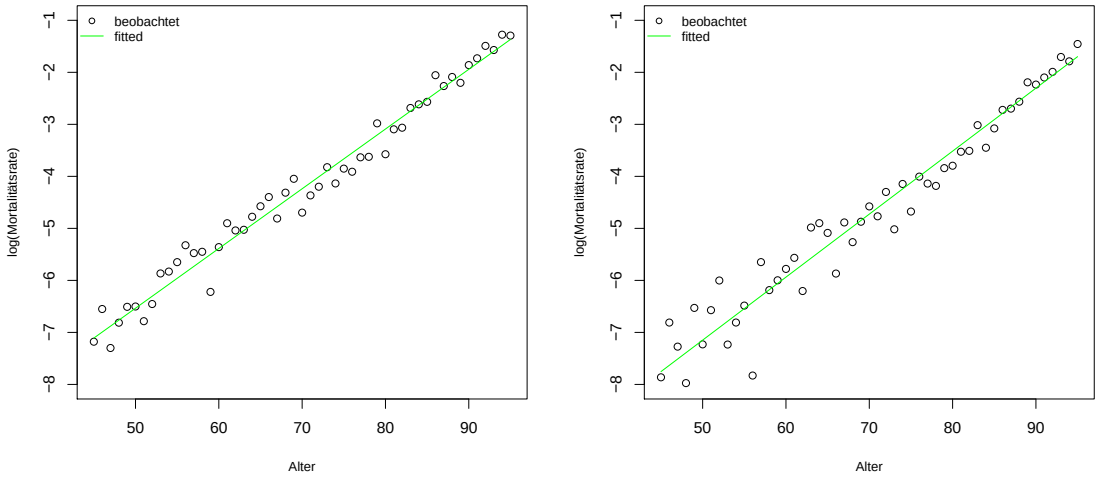

Abbildung 3.4: Unter dem Negativ-Binomial-Modell geschätzte Mortalitätsraten männlicher (links) bzw. weiblicher (rechts) Versicherter im Jahr 2015 auf der log-Skala aufgetragen.

3.3.3 Logistische Regression

Abgesehen vom loglinearen Modell, ist auch das logistische Regressionsmodell insbesonders für binär kodierte Response-Variablen, wie sie im Falle der Sterblichkeitsanalyse vorliegen, im Allgemeinen besonders gut geeignet (vgl. McCullagh & Nelder, 1989). Bei Mortalitätsuntersuchungen kann die Response y_i lediglich zwei Werte annehmen, nämlich $y_i = 1$ (Person i verstirbt) oder $y_i = 0$ (Person i überlebt). Interessiert sind wir an der Sterbewahrscheinlichkeit, also an $P(y_i = 1) = \pi_i$ und wollen diese in Abhängigkeit vom Alter x modellieren.

Nehmen wir an, dass wie im Linearen Modell ein linearer Zusammenhang zwischen der Sterbewahrscheinlichkeit und dem Alter besteht, also $\pi_i = \beta_0 + \beta_1 x_i$, so ergibt sich wie schon oben erwähnt das Problem, dass π_i auch negative Werte bzw. Werte größer Eins annehmen kann und somit die Wahrscheinlichkeiten nicht notwendigerweise im Intervall $[0, 1]$ liegen. Aus diesem Grund wird eine Link-Funktion $g(\cdot)$ gewählt, die das Intervall $[0, 1]$ auf $(-\infty, +\infty)$ abbildet, wodurch das Modell die Gestalt $g(\pi_i) = \eta_i = \beta_0 + \beta_1 x_i$ aufweist. Als Link-Funktionen werden in diesem Kapitel der Logit-Link $g(\pi) = \log\left(\frac{\pi}{1-\pi}\right)$, der, wie im Abschnitt 3.2.2 beschrieben, die kanonische Link-Funktion bildet und die komplementäre Log-Log Funktion $g(\pi) = \log(-\log(1 - \pi))$ verwendet. Für kleine Werte von π_i liefern der Logit-Link und der komplementäre Log-Log Link sehr ähnliche Ergebnisse (siehe Abbildung 3.5).

Um die Sterbewahrscheinlichkeit mittels logistischer Regression zu schätzen, halten wir im

Folgenden wieder das Jahr fest, weshalb es sich in diesem Zusammenhang wie schon im vorangegangenen Abschnitt um eine eindimensionalen Glättung handelt (nur in Richtung des Alters). Die Anzahl der Todesfälle sei dazu binomialverteilt, also $D_{x,t} \sim B(E'_{x,t}, q_{x,t})$, wobei bei diesem Ansatz mit $E'_{x,t}$ die Anzahl der am Jahresanfang im Bestand befindlichen Versicherten betrachtet wird und nicht wie im Poisson-Modell die verlebten Personenjahre als Bezugsgröße dienen. Nun kann das im Abschnitt 3.3.1 vorgestellte Poisson-Modell (3.3) mit der Beziehung zwischen Sterbeintensität und Sterbewahrscheinlichkeit (2.5) auf die Binomialverteilung übertragen werden

$$\begin{aligned} q_x &\approx 1 - e^{-\mu_x} \\ &= 1 - e^{-e^{\beta_0 + \beta_1 x}}, \end{aligned}$$

womit wir für die Sterbewahrscheinlichkeit folgendes Modell erhalten:

$$\log(-\log(1 - q_x)) \approx \beta_0 + \beta_1 x.$$

Dabei handelt es sich bei der Link-Funktion $\eta_x = \log(-\log(1 - q_x))$ nicht um die kanonische Link-Funktion, sondern um die komplementäre Log-Log Transformation.

Auch hier kann ebenso der Befehl glm in R verwendet werden, um das Modell anzupassen. Die Modellmatrix $\mathbf{X}$ weist in diesem Zusammenhang dieselbe Gestalt auf, wie jene im Poisson-Modell (3.5) und als Gewichte gehen $E'_{x,t}$ ein. Ausführlichere Darstellungen zur Spezifikation eines logistischen Modells in R sind in Venables und Ripley (2002, S.190) zu finden.

```
> rates<-deaths/exposures
> model_cloglog<-glm(rates~Alter,family=binomial(link=
+ "cloglog"),weights=exposures)
> summary(model_cloglog)

Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) -13.383939   0.244639  -54.71   <2e-16 ***
Alter         0.123210   0.002951   41.76   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 2534.354  on 50  degrees of freedom
Residual deviance:   71.087  on 49  degrees of freedom
AIC: 296.9
```

```
> model_logit<-glm(rates~Alter, family=binomial(link="logit"),
+ weights=exposures)
> summary(model_logit)

Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) -13.511205   0.250796  -53.87   <2e-16 ***
```

```
Alter          0.125132   0.003042   41.14   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 2534.35  on 50  degrees of freedom
Residual deviance:   74.78  on 49  degrees of freedom
AIC: 300.59
```

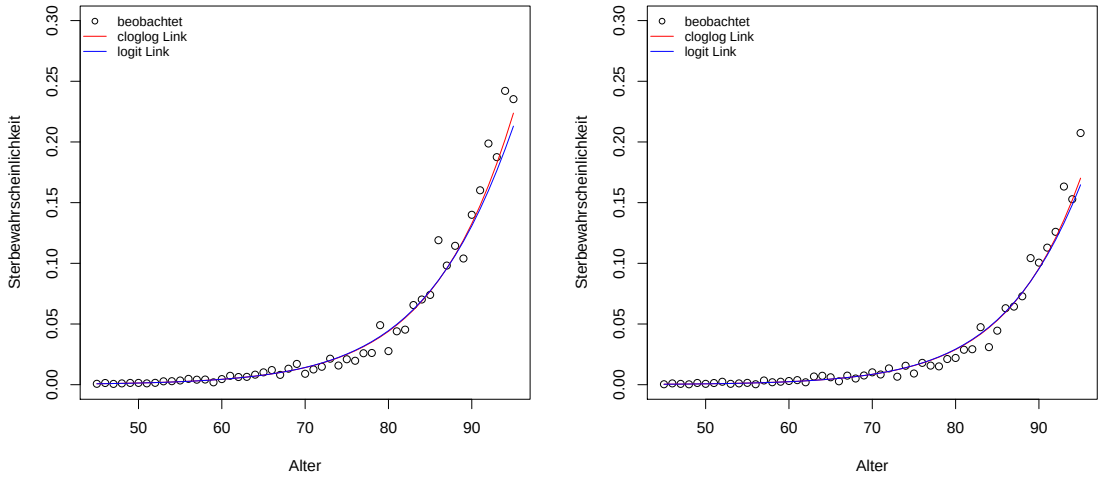

Abbildung 3.5: Beobachtete und geschätzte Sterbewahrscheinlichkeiten männlicher (links) bzw. weiblicher (rechts) Versicherter im Jahr 2015.

Betrachten wir Akaike's Informationskriterium, so liefert das logistische Modell mit Logit-Link ein AIC = 300.59 und jenes mit dem komplementären Log-Log Link ein AIC = 296.9, weshalb wir uns für das Regressionsmodell mit komplementärem Log-Log Link entscheiden, um die Sterbewahrscheinlichkeit zu schätzen.

Kapitel 4

Das Lee-Carter Modell

Im vorangegangenen Kapitel wurden unterschiedliche Ansätze zur Modellierung von Mortalitätsraten bzw. Sterbewahrscheinlichkeiten in Abhängigkeit vom Alter besprochen. Wie schon im Abschnitt 2.2.3 diskutiert, hängen Mortalitätsraten sowohl vom Alter, als auch von der Zeit ab. Daher gehen wir in diesem Kapitel von der eindimensionalen Glättung von Sterbemaßen zur zweidimensionalen Modellierung über und verwenden für die Schätzung die gesamte Historie der vorhandenen Daten in Richtung des Alters und der Zeit.

Ein weit verbreitetes Modell zur Modellierung und Prognose von Mortalitätsraten stellt das Lee-Carter Modell dar. Dieses wird Gegenstand des aktuellen Kapitels sein, zunächst in der von Lee und Carter diskutierten Form und anschließend werden wir auf eine Verallgemeinerung hinsichtlich der zugrunde liegenden Verteilungsannahme übergehen. Des Weiteren bildet auch die Vorhersage der Mortalität in beiden Ansätzen einen zentralen Teil des Kapitels.

4.1 Das klassische Lee-Carter Modell

Das von Lee und Carter im Jahr 1992 vorgestellte Modell (siehe Lee & Carter, 1992) zur Schätzung und Prognose von Mortalitätsraten $\mu_{x,t}$ hat die folgende Form:

$$\log \mu_{x,t} = \alpha_x + \beta_x \kappa_t, \tag{4.1}$$

wobei $x \in \{1, \ldots, n_a\}$, $t \in \{1, \ldots, n_t\}$, und n_t die Anzahl der betrachteten Jahre widerspiegelt.

Bei α_x und β_x handelt es sich um altersspezifische Konstanten und κ_t stellt einen zeitabhängigen Index dar. Genauer gibt α_x das allgemeine Sterblichkeitsniveau im Alter x an und β_x reflektiert die Sensitivität der Mortalitätsrate gegenüber Veränderungen des Index κ_t.

Das Lee-Carter Modell ist unterbestimmt, weshalb die Restriktionen

$$\sum_{x=1}^{n_a} \beta_x = 1, \qquad \sum_{t=1}^{n_t} \kappa_t = 0 \tag{4.2}$$

eingeführt werden, um eine eindeutige Lösung für Gleichung (4.1) zu erhalten. Unter der

obigen Annahme, dass sich die κ_t's auf Null summieren, folgt für α_x sofort

$$\alpha_x = \frac{1}{n_t} \sum_{t=1}^{n_t} \log \mu_{x,t}.$$

Die noch zu schätzenden Parameter β_x und κ_t können nicht über herkömmliche Methoden bestimmt werden, da sich auf der rechten Seite in (4.1) ausschließlich zu schätzende unbekannte Größen befinden. Die von Lee und Carter in diesem Zusammenhang verwendete Singulärwertzerlegung (SVD, *Singular Value Decomposition*) bietet einen Zugang, der die Berechnung einer Kleinsten-Quadrate-Lösung zulässt (siehe Knüsel, 2008). Dabei wird die SVD auf die $n_a \times n_t$-Matrix $\mathbf{Z}$ mit Rang r und den Einträgen $z_{x,t} = [\log \mu_{x,t} - \alpha_x]$ angewandt, sodass sich die folgende Zerlegung ergibt:

$$\mathbf{Z} = \mathbf{U}\mathbf{S}\mathbf{V}^\top \qquad \text{mit} \quad \mathbf{S} = \left(\begin{array}{c|c} \tilde{\mathbf{S}} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} \end{array}\right), \tag{4.3}$$

wobei $\mathbf{U} \in \mathbb{R}^{n_a \times n_a}$ und $\mathbf{V} \in \mathbb{R}^{n_t \times n_t}$ orthogonal sind, $\mathbf{S} \in \mathbb{R}^{n_a \times n_t}$ und $\tilde{\mathbf{S}}$ eine Diagonalmatrix mit den monoton fallenden Einträgen $s_1 \geq s_2 \geq \cdots \geq s_r > 0$ darstellt. Der Vektor $\mathbf{u}_j = (u_{1,j}, \ldots, u_{n_a,j})^\top$ heißt j-ter Links-Singulärvektor von $\mathbf{Z}$ und analog dazu heißt $\mathbf{v}_j = (v_{1,j}, \ldots, v_{n_t,j})^\top$ j-ter Rechts-Singulärvektor von $\mathbf{Z}$.

Wir wollen $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{n_a})^\top \in \mathbb{R}^{n_a}$ und $\boldsymbol{\kappa} = (\kappa_1, \ldots, \kappa_{n_t})^\top \in \mathbb{R}^{n_t}$ derart wählen, dass

$$g(\boldsymbol{\beta}, \boldsymbol{\kappa}) = \|\mathbf{Z} - \boldsymbol{\beta}\boldsymbol{\kappa}^\top\|^2$$

minimiert wird. Da für alle $\boldsymbol{\kappa}$

$$g(\boldsymbol{\beta}, \boldsymbol{\kappa}) = \begin{cases} g((1, 0, \ldots, 0)^\top, \mathbf{0}), & \text{wenn } \boldsymbol{\beta} = \mathbf{0}, \\ g(\boldsymbol{\beta}/\|\boldsymbol{\beta}\|, \boldsymbol{\kappa}\|\boldsymbol{\beta}\|), & \text{wenn } \boldsymbol{\beta} \neq \mathbf{0}, \end{cases}$$

gilt, nehmen wir im Folgenden $\|\boldsymbol{\beta}\|^2 = \beta_1^2 + \cdots + \beta_{n_a}^2 = 1$ an.

Unter Verwendung der SVD (4.3) erhält das Lee-Carter Modell die Darstellung

$$\mathbf{U}\mathbf{S}\mathbf{V}^\top = \boldsymbol{\beta}\boldsymbol{\kappa}^\top.$$

Multiplizieren wir von links mit $\mathbf{U}^\top$ und von rechts mit $\mathbf{V}$, so erhalten wir

$$\mathbf{U}^\top\mathbf{U}\mathbf{S}\mathbf{V}^\top\mathbf{V} = \mathbf{U}^\top\boldsymbol{\beta}\boldsymbol{\kappa}^\top\mathbf{V}.$$

Setzen wir $\tilde{\boldsymbol{\beta}} = \mathbf{U}^\top\boldsymbol{\beta}$ und $\tilde{\boldsymbol{\kappa}} = \mathbf{V}^\top\boldsymbol{\kappa}$, dann ergibt sich unter Berücksichtigung der Orthogonalität von $\mathbf{U}$ und $\mathbf{V}$ die folgende Form des Lee-Carter Modells:

$$\mathbf{S} = \tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\kappa}}^\top$$

und es gilt weiters

$$\|\tilde{\boldsymbol{\beta}}\|^2 = \tilde{\boldsymbol{\beta}}^\top\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}^\top\mathbf{U}\mathbf{U}^\top\boldsymbol{\beta} = \boldsymbol{\beta}^\top\boldsymbol{\beta} = \|\boldsymbol{\beta}\|^2.$$

Das transformierte Kleinste-Quadrate-Problem lässt sich nach obigen Überlegungen folgendermaßen formulieren:

$$\text{Minimiere } \|\mathbf{S} - \tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\kappa}}^\top\|^2 \text{ unter der Nebenbedingung } \|\tilde{\boldsymbol{\beta}}\|^2 = 1.$$

Es gilt mit $h = \min\{n_a, n_t\}$

$$\begin{aligned}
\|\mathbf{S} - \tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\kappa}}^\top\|^2 &= \sum_{i=1}^{h}(s_i - \tilde{\beta}_i\tilde{\kappa}_i)^2 + \sum_{i=1}^{n_a}\sum_{\substack{j=1\\ j\neq i}}^{n_t}\tilde{\beta}_i^2\tilde{\kappa}_j^2 \\
&= \sum_{i=1}^{h} s_i^2 - 2\sum_{i=1}^{h} s_i\tilde{\beta}_i\tilde{\kappa}_i + \sum_{i=1}^{h}\tilde{\beta}_i^2\tilde{\kappa}_i^2 + \sum_{i=1}^{n_a}\sum_{\substack{j=1\\ j\neq i}}^{n_t}\tilde{\beta}_i^2\tilde{\kappa}_j^2 \\
&= \sum_{i=1}^{h} s_i^2 - 2\sum_{i=1}^{h} s_i\tilde{\beta}_i\tilde{\kappa}_i + \underbrace{\sum_{i=1}^{n_a}\tilde{\beta}_i^2}_{=1}\sum_{j=1}^{n_t}\tilde{\kappa}_j^2 \\
&= \sum_{i=1}^{h}\underbrace{(\tilde{\kappa}_i - s_i\tilde{\beta}_i)^2}_{\geq 0} + \sum_{i=h}^{n_t}\underbrace{\tilde{\kappa}_i^2}_{\geq 0} + \sum_{i=1}^{h} s_i^2 - \sum_{i=1}^{h}\underbrace{s_i^2}_{\leq s_1^2}\tilde{\beta}_i^2 \\
&\geq \sum_{i=2}^{h} s_i^2,
\end{aligned}$$

wobei im letzten Schritt im Falle von $\min\{n_a, n_t\} = n_t$ verwendet wird, dass $\sum_{i=1}^{n_t}\tilde{\beta}_i^2 \leq \sum_{i=1}^{n_a}\tilde{\beta}_i^2 = 1$.

Die Quadratsumme wird somit minimal für die Wahl $\tilde{\boldsymbol{\beta}} = (1, 0, \ldots, 0)^\top \in \mathbb{R}^{n_a}$ und $\tilde{\boldsymbol{\kappa}} = (s_1, 0, \ldots, 0)^\top \in \mathbb{R}^{n_t}$, da damit $\|\mathbf{S} - \tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\kappa}}^\top\|^2 = \sum_{i=2}^{h} s_i^2$ gilt. Nach Rücktransformation auf die Originalgrößen ergeben sich schlussendlich die Kleinsten-Quadrate-Schätzer für $\boldsymbol{\beta}$ und $\boldsymbol{\kappa}$:

$$\hat{\boldsymbol{\beta}} = \mathbf{U}\tilde{\boldsymbol{\beta}} = \mathbf{u}_1 \quad \text{und} \quad \hat{\boldsymbol{\kappa}} = \mathbf{V}\tilde{\boldsymbol{\kappa}} = s_1\mathbf{v}_1.$$

Es handelt sich also beim ersten Links-Singulärvektor $\mathbf{u}_1$ um einen Schätzer von $\boldsymbol{\beta}$ und der erste Rechts-Singulärvektor $\mathbf{v}_1$ multipliziert mit s_1 liefert einen Schätzer für $\boldsymbol{\kappa}$. Dabei sind α_x, $\hat{\beta}_x$ und $\hat{\kappa}_t$ derart bestimmt, dass sie

$$\sum_{x,t}\left(\log\mu_{x,t} - \alpha_x - \beta_x\kappa_t\right)^2$$

minimieren.

Die Schätzer müssen zunächst noch normiert werden, um die Restriktion $\sum_{x=1}^{n_a}\beta_x = 1$ zu erfüllen, womit sich die Vektoren

$$\hat{\boldsymbol{\beta}} = \frac{\mathbf{u}_1}{\sum_{i=1}^{n_a} u_{i,1}},$$

$$\hat{\boldsymbol{\kappa}} = s_1\left(\sum_{i=1}^{n_a} u_{i,1}\right)\mathbf{v}_1$$

als optimale Parameterschätzer ergeben. Das nachstehend angeführte Listing zeigt das in R implementierte Lee-Carter Modell.

```
> t<-seq(2001,2015,1)
> x<-seq(51,95,1)
> M<-deaths/exposures
> alpha_LC<-vector(mode="double", length=length(x))

> for (i in 1:length(x)){
+   alpha_LC[i]<-1/length(t)*sum(log(M[i,]))}

> Z<-log(M)-alpha_LC
> S<-svd(Z)

> beta_LC<-S$u[,1]/sum(S$u[,1])
> kappa_LC<-S$d[1]*sum(S$u[,1])*S$v[,1]

> rates<-matrix(0 ,nrow=length(x),ncol=length(t))

> for (j in 1:length(x)){
+   for (i in 1:length(t)){
+     rates[j,i]<-exp(alpha_LC[j]+beta_LC[j]*kappa_LC[i])}}
```

In den Abbildungen 4.2 bis 4.4 des nächsten Unterabschnitts, der sich einer Variante des Lee-Carter Modells widmet, sind die geschätzten Parameter α_x, $\hat{\beta}_x$ und $\hat{\kappa}_t$ für den Altersbereich $[51, 95]$ und den Zeitraum 2001 bis 2015 dargestellt. Die Wahl des Altersbereichs beruht auf der dünn besetzten Matrix der Anzahl der Sterbefälle $\mathbf{D}$, die für das Altersintervall $[1, 50]$ Nullen enthält und aus diesem Grund $\log \mu_{x,t}$ nicht berechenbar ist.

Anhand der Abbildungen ist ersichtlich, dass das allgemeine Sterblichkeitsniveau bei Männern etwas höher ist als jenes der Frauen (siehe Grafik 4.2). Des Weiteren ist bei beiden Geschlechtern im Altersintervall $[51, 70]$ der Schätzer $\hat{\beta}_x$ im Vergleich zum restlichen Segment $[71, 95]$ volatiler. Das bedeutet, dass insbesonders im niedrigen Altersbereich die Mortalitätsraten sensitiver gegenüber Änderungen im zeitabhängigen Index κ_t sind. In der in Abbildung 4.4 gezeigten Entwicklung des Parameterschätzers für κ_t über den Beobachtungszeitraum ist zu sehen, dass die Mortalitätsraten gegenüber dem Jahr 2001 sinken.

4.1.1 Projektion im klassischen Lee-Carter Modell

Nun sind mit Hilfe der vorliegenden Daten die Modell-Parameter α_x, β_x und κ_t geschätzt worden. Im zweiten Schritt des Lee-Carter Modells werden die Werte für κ_t unter der Verwendung eines ARIMA-Modells (*Autoregressive Integrated Moving Average Model*) extrapoliert, wobei die beiden Parameter α_x und β_x als konstant über die Zeit hinweg angesehen werden. Lee und Carter haben ein ARIMA(0,1,0)-Modell, also einen Random Walk mit Drift, zur Modellierung von κ_t verwendet.

Das ARIMA(p,d,q)-Modell mit Drift für eine Zeitreihe $x_t, t \in \mathbb{N}$, kann in folgender Form beschrieben werden (siehe Shumway & Stoffer, 2011):

$$\left(1 - \sum_{i=1}^{p} \phi_i B^i\right) (1 - B)^d x_t = \delta + \left(1 + \sum_{j=1}^{q} \theta_j B^j\right) w_t, \tag{4.4}$$

wobei $\phi_1, \ldots, \phi_p$ sowie $\theta_1, \ldots, \theta_q$ mit $\phi_p \neq 0$ und $\theta_q \neq 0$ die zu schätzenden Parameter bezeichnen und $w_t \overset{iid}{\sim} N(0, \sigma_w^2)$ stellt die Fehlerterme dar. Außerdem ist der Backshift-Operator B , durch $B^k x_t = x_{t-k}$ definiert. Ist $\mathrm{E}((1-B)^d x_t) = \mu$, dann erhalten wir für δ die Beziehung $\delta = \mu(1 - \phi_1 - \cdots - \phi_p)$.

Aus den Schritten nach der Box und Jenkins Methode (siehe Box & Jenkins, 1970) zur Identifikation eines passenden ARIMA(p,d,q)-Modells geht hervor, dass für weibliche Versicherte ein ARIMA(0,1,0)-Modell für κ_t optimal ist. Demnach wird κ_t als stochastischer Prozess der Form

$$\kappa_t = \kappa_{t-1} + \delta + w_t$$

modelliert und δ spiegelt in diesem Zusammenhang die durchschnittliche jährliche Änderung von κ_t wider. Die Modellierung von κ_t für weibliche Versicherte ist im nachstehenden R-Code und Output ersichtlich. Dazu wird das von Stoffer (2016) entwickelte R-Paket astsa verwendet.

```
> kappa_ts<-ts(kappa_LC, start=2001, end=2015, frequency=1)
> library("astsa")
> ARIMAfit <- sarima(kappa_ts, 0, 1, 0)
> ARIMAfit$ttable
         Estimate     SE t.value p.value
constant  -0.8559 1.1406 -0.7504  0.4654
```

Für Männer hingegen stellt sich heraus, dass ein ARIMA(1,1,0)-Modell am besten für die vorliegende Datensituation geeignet ist. Dieses hat die Gestalt

$$\kappa_t = \kappa_{t-1} + \delta + \phi_1(\kappa_{t-1} - \kappa_{t-2}) + w_t.$$

```
> ARIMAfit_m <- sarima(kappa_ts, 1, 1, 0)
> ARIMAfit_m$ttable
         Estimate     SE t.value p.value
ar1        -0.7531 0.1520 -4.9535  0.0003
constant   -0.8284 0.6247 -1.3263  0.2076
```

Die prognostizierten Mortalitätsraten ergeben sich mit Hilfe der Extrapolation des Zeitindex κ_t, wie im nachfolgenden R-Code nachvollzogen werden kann.

```
> n<-5
> fore<-sarima.for(kappa_ts, n, 0, 1, 0)
> rates_pred<-matrix(0 ,nrow=length(x),ncol=n)

> for (j in 1:length(x)){
+   for (i in 1:n){
+    rates_pred[j,i]<-exp(alpha_LC[j]+beta_LC[j]*fore$pred[i])}}
```

Die Abbildung 4.1 zeigt die unter dem Lee-Carter Modell geschätzten Mortalitätsraten sowie die zu erwartenden Raten für das Zeitintervall $[2016, 2020]$ für Männer bzw. Frauen. Wie es basierend auf den geschätzten Parametern der jeweiligen ARIMA-Modelle zu erwarten ist (siehe obige R-Codes), sinkt die Mortalität bis zum Jahr 2020 kontinuierlich.

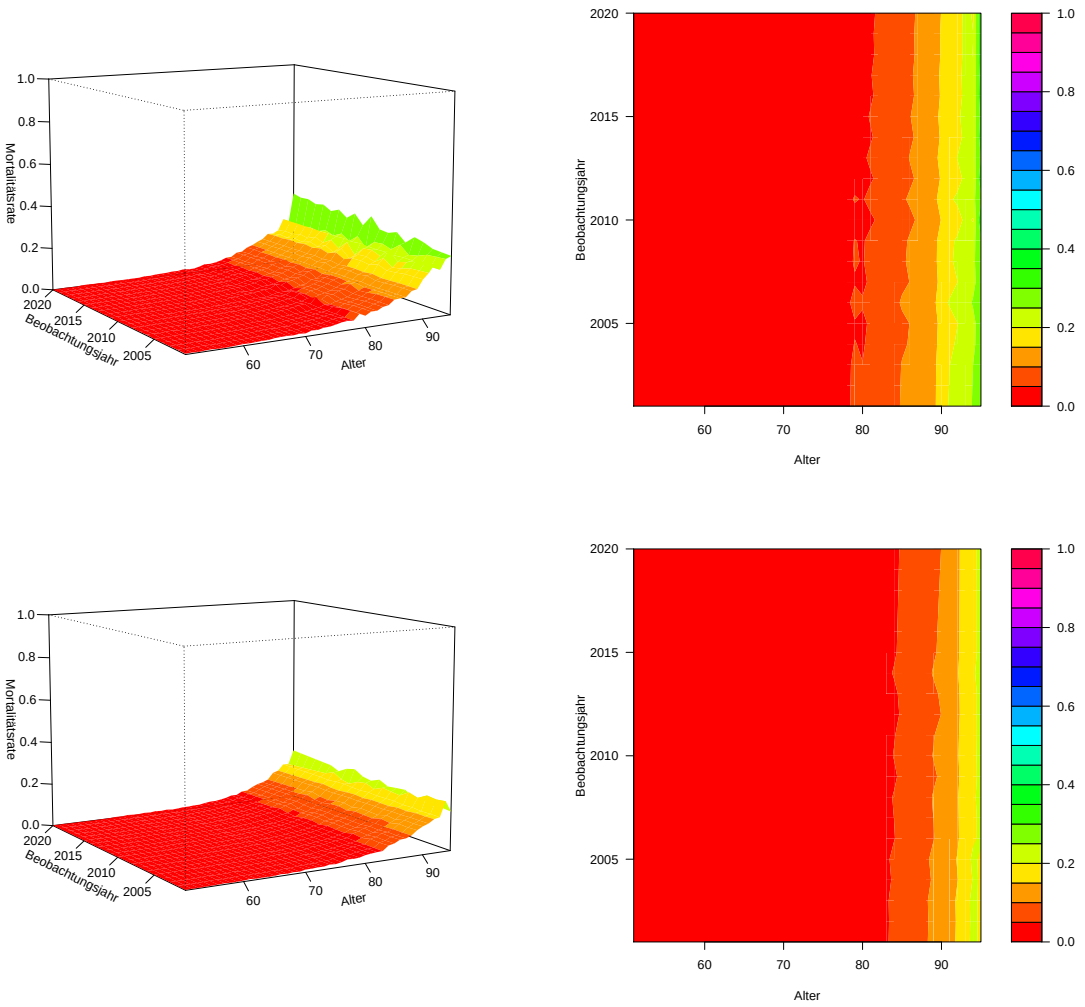

Abbildung 4.1: Unter dem Lee-Carter Modell berechnete Mortalitätsraten für Männer (oben) und Frauen (unten) im Zeitraum $[2001, 2015]$ und prognostizierte Raten bis zum Jahr 2020 durch die Verwendung des unter dem ARIMA-Modell extrapolierten Index κ_t.

4.2 Lee-Carter Methode unter Anwendung einer Poissonverteilung für die Anzahl der Verstorbenen

Im klassischen Lee-Carter Modell wird keine Verteilungsannahme für die Parameterschätzung zugrunde gelegt. Ein modernerer und auch im Hinblick auf Mortalitätsraten sinnvoller Zugang ist die Annahme einer Verteilung für die Anzahl der Sterbefälle, wie es in den nachstehenden Ausführungen betrachtet wird.

Wie schon zuvor besprochen, ist es für die Anzahl der Sterbefälle naheliegend eine Poissonverteilung anzunehmen, also $D_{x,t} \sim Poi(\mu_{x,t}E_{x,t})$ und $\mu_{x,t} = \exp(\alpha_x + \beta_x\kappa_t)$ (siehe Wilmoth, 1993 oder Brouhns, Denuit & Vermunt, 2002). Die Bedeutung der Parameter α_x, β_x und κ_t ist nach wie vor dieselbe wie unter dem klassischen Lee-Carter Modell. Wir erhalten somit für die Mortalitätsrate $\mu_{x,t}$ ein Modell der Form

$$\log \mu_{x,t} = \log \frac{\mathrm{E}(D_{x,t})}{E_{x,t}} = \alpha_x + \beta_x\kappa_t,$$

das auf Grund des multiplikativen Terms $\beta_x\kappa_t$ nicht in das GLM-Regelwerk passt. Es handelt sich hierbei um ein Generalisiertes Nichtlineares Modell, weshalb wir das von Turner und Firth (2015) entwickelte gnm-Paket verwenden, um das Modell in R anzupassen.

Für die Modellspezifikation müssen die in den Variablen deaths bzw. exposures enthaltenen Matrizen vektorisiert werden (siehe nachstehendes Listing) und die dazu korrespondierenden Faktor-Variablen alter_f sowie jahr_f sind die mittels $\mathbf{1}_{n_a} \otimes \mathbf{x}$ bzw. $\mathbf{z} \otimes \mathbf{1}_{n_t}$ definierten Vektoren, wobei $\mathbf{x} = (x_1, \ldots, x_{n_a})^\top$ den Altersvektor und $\mathbf{z} = (z_1, \ldots, z_{n_t})^\top$ den Zeitvektor, der die Beobachtungsjahre enthält, darstellen. Mit der Funktion Mult kann schließlich der multiplikative Term spezifiziert werden und $\log E_{x,t}$ geht als Offset in das Modell ein. Auch in diesem Abschnitt modellieren wir Mortalitätsraten für Männer und Frauen in einem Altersbereich zwischen 51 und 95 Jahren, um Vergleichbarkeit zu schaffen.

```
1 > d<-c(deaths)
2 > e<-c(exposures)
3 > alter_f<-as.factor(kronecker(rep(1,length(t)),x))
4 > jahr_f<-as.factor(kronecker(t,rep(1,length(x))))

6 > lc_modell_gnm<-gnm(d ~ alter_f + Mult(alter_f,jahr_f), offset=
7 + log(e), family=quasipoisson(link="log"))
```

Um eine eindeutige Lösung für die Schätzer von $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ und $\boldsymbol{\kappa}$ zu erhalten, werden diese im Lee-Carter Modell unter den Restriktionen (4.2) berechnet. Im Modell lc_modell_gnm wird hingegen eine zufällige Parametrisierung für die Koeffizienten geliefert (siehe Turner & Firth, 2015), sodass die folgende Anpassung von den im Modell berechneten Schätzern $\hat{\boldsymbol{\alpha}}_m$, $\hat{\boldsymbol{\beta}}_m$ und $\hat{\boldsymbol{\kappa}}_m$ notwendig ist (siehe Currie, 2013):

$$\begin{aligned}
\hat{\boldsymbol{\alpha}} &= \hat{\boldsymbol{\alpha}}_m + \bar{\kappa}_m\hat{\boldsymbol{\beta}}_m, \\
\hat{\boldsymbol{\kappa}} &= n_a\bar{\beta}_m(\hat{\boldsymbol{\kappa}}_m - \bar{\kappa}_m\mathbf{1}_{n_t}), \\
\hat{\boldsymbol{\beta}} &= \hat{\boldsymbol{\beta}}_m/(n_a\bar{\beta}_m),
\end{aligned}$$

mit $\bar{\beta}_m = \sum_x \hat{\beta}_{x,m}/n_a$ und $\bar{\kappa}_m = \sum_t \hat{\kappa}_{t,m}/n_t$. In den Abbildungen 4.2 bis 4.4 sind die unter der Annahme einer Poissonverteilung geschätzten Parameter (als LC Variante bezeichnet) gemeinsam mit den Schätzungen unter dem klassischen Lee-Carter Modell (als klass. LC bezeichnet) dargestellt.

Verglichen mit den berechneten Schätzern im klassischen Lee-Carter Modell verlaufen die Schätzer von β_x und κ_t, unter Berechnung nach der in diesem Abschnitt betrachteten Variante, weniger volatil. Das schon im Unterabschnitt 4.1 beschriebene Verhalten bleibt aber auch hier erhalten. So ist ein deutliches Fallen des Mortalitätsindex κ_t über den Zeitraum $[2001, 2015]$ hinweg zu beobachten und auch die erhöhte Sensitivität gegenüber Änderungen im Mortalitätsindex im niedrigen Altersbereich geht aus Abbildung 4.3 hervor.

4.2.1 Projektion im Lee-Carter Modell mit Poissonverteilung

In weiterer Folge wird nun der Parameter κ_t unter dem ARIMA-Modell extrapoliert. Es stellt sich heraus, dass das ARIMA(0,1,0)-Modell sowohl für Männer, als auch für Frauen passend ist.

In Abbildung 4.5 ist für Männer und Frauen im Altersintervall $[51, 95]$ der prognostizierte Mortalitätsindex κ_t für einen Zeithorizont von fünf Jahren ersichtlich.

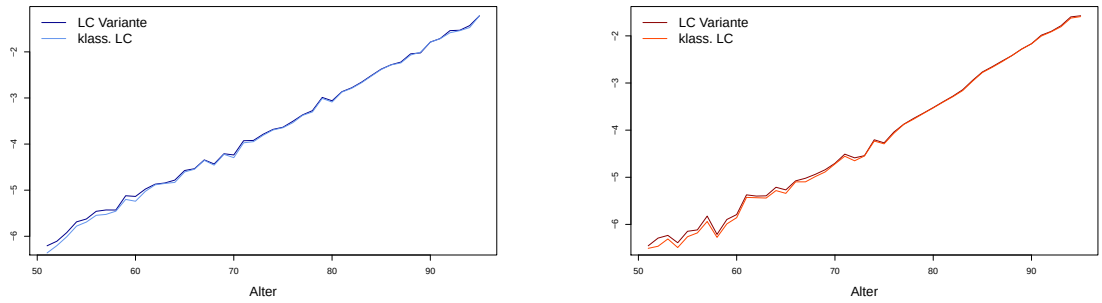

Abbildung 4.2: Parameterschätzer für α_x im Altersbereich $[51, 95]$ für Männer (links) und Frauen (rechts).

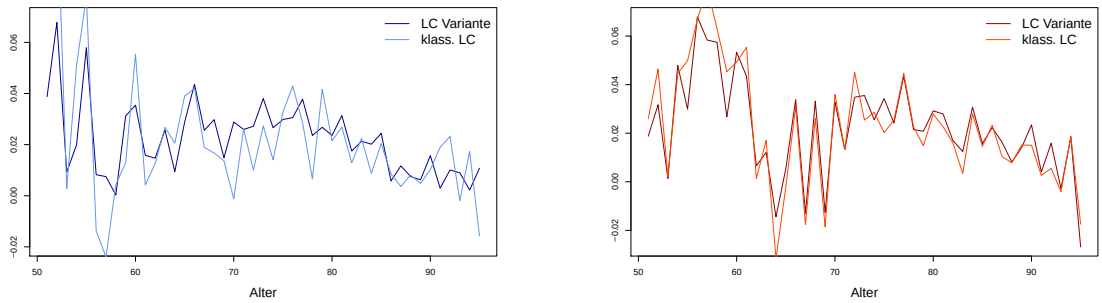

Abbildung 4.3: Parameterschätzer für β_x im Altersbereich $[51, 95]$ für Männer (links) und Frauen (rechts).

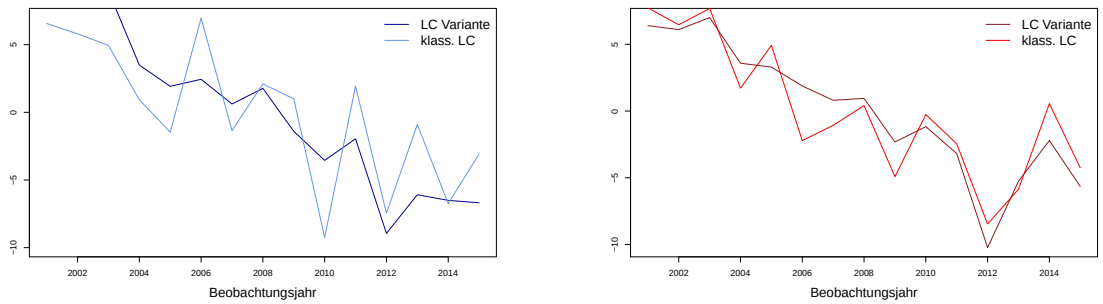

Abbildung 4.4: Parameterschätzer für κ_t im Altersbereich $[51, 95]$ für Männer (links) und Frauen (rechts).

Es ist deutlich zu erkennen, dass für beide Geschlechter der Mortalitätsindex gegenüber dem Jahr 2001, das den Anfang des Beobachtungszeitraums darstellt, abnimmt und damit einhergehend auch die Mortalitätsraten über den Beobachtungszeitraum sinken werden, wie es in Abbildung 4.6 für Männer und Frauen illustriert ist. Anhand des in Abbildung 4.5 eingezeichneten Konfidenzintervalls ist jedoch ersichtlich, dass sich dieses für weiter in der Zukunft liegende Jahre vergrößert und damit die Prognose mit zunehmender Unsicherheit behaftet ist. Daher ist die gewählte Vorhersage für fünf Jahre nicht als stichhaltig anzusehen, sondern soll vielmehr der Veranschaulichung dienen.

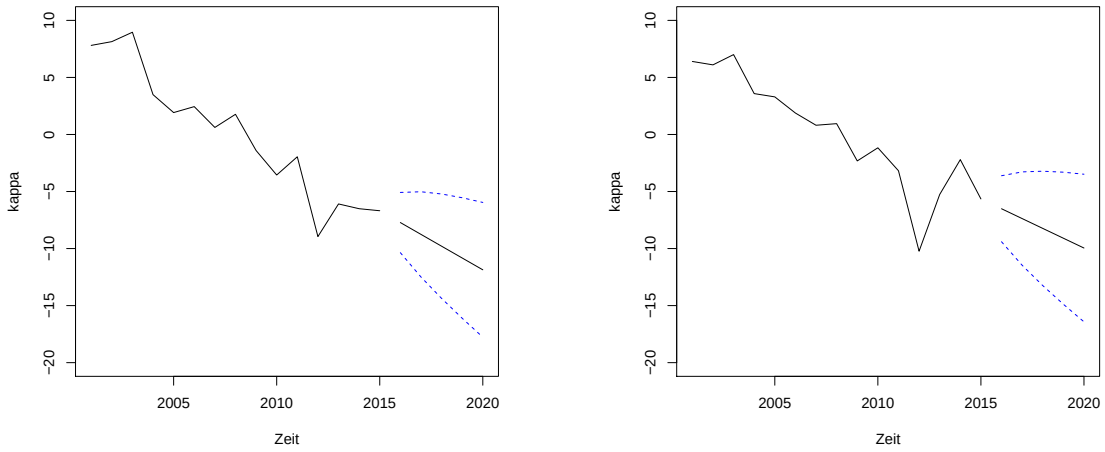

Abbildung 4.5: Prognostizierter Mortalitätsindex κ_t plus bzw. minus dem zweifachen Standardfehler für Männer (links) und Frauen (rechts) im ARIMA(0,1,0)-Modell.

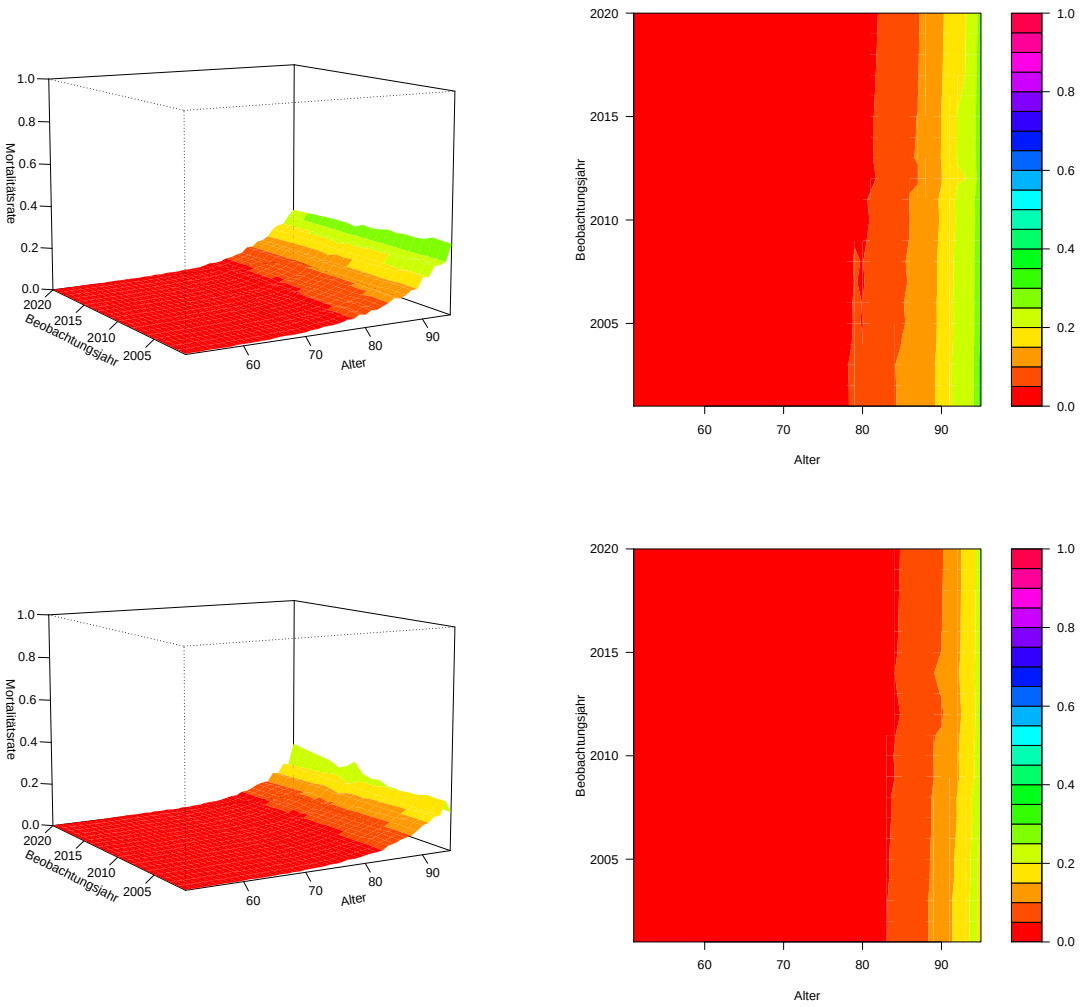

Abbildung 4.6: Unter der besprochenen Variante des Lee-Carter Modells berechnete Mortalitätsraten für Männer (oben) und Frauen (unten) im Zeitraum $[2001, 2015]$ und prognostizierte Raten bis zum Jahr 2020 durch die Verwendung des unter dem ARIMA-Modell extrapolierten Index κ_t.

Kapitel 5

Nichtparametrische Modelle

Bisher wurden Modelle betrachtet, denen Annahmen über die Form des funktionalen Zusammenhangs der Kovariablen und der Response zugrunde liegen. Außerdem ist im Falle des zuvor diskutierten Lee-Carter Modells die Anzahl der zu schätzenden Parameter unverhältnismäßig groß, da sowohl für jedes Alter, als auch jedes Beobachtungsjahr ein individueller Parameter geschätzt wird.

Aus diesem Grund werden nichtparametrische Regressionsmodelle eingeführt, die eine flexiblere Modellierung des Alters- bzw. Beobachtungsjahreinflusses auf die Sterberaten zulassen. Da für diesen Zugang ein Penalisierungsansatz zur Verwendung kommen wird, werden zunächst Basis-Splines (B-Splines) und anschließend penalisierte Splines (P-Splines) vorgestellt. Dies erfolgt zunächst anhand eines einfachen eindimensionalen Modells, bedarf aber danach einer Verallgemeinerung in Bezug auf die Verteilungsannahme sowie die Dimension, da im Hinblick auf Mortalitätsdaten zweidimensionale Daten (in Richtung des Alters und in Richtung des Beobachtungsjahres) vorliegend sind. Das Einbringen praktischer Aspekte unter Anwendung des Datensatzes soll den theoretischen Hintergrund abrunden. Bevor abschließend die Prognose der Sterberaten im eindimensionalen sowie zweidimensionalen Fall diskutiert wird, steht im vorletzten Abschnitt die Modelldiagnose im Zentrum. In diesem Zusammenhang wird auch ein Vergleich zum Lee-Carter Modell angestellt.

5.1 Einführung in Penalisierte Splines

Das Ziel dieses Abschnitts ist es, eine Einführung in den nichtparametrischen Modellierungsansatz zu geben, der eine flexiblere Modellierung ermöglicht. Dazu betrachten wir zunächst ein Modell, das den glatten Zusammenhang der Response y_i und eines Prädiktors $x_i, i \in \{1, \ldots, n\}$, repräsentiert (siehe Wood, 2006):

$$y_i = f(x_i) + \epsilon_i, \tag{5.1}$$

wobei f eine glatte Funktion, die zu schätzen ist, darstellt und $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$.

Für die Schätzung der glatten Funktion f wählen wir eine Basis, also einen Funktionenraum dem f entstammt. Dafür stehen unterschiedliche Basen zur Auswahl. Wir konzen-

trieren uns im Folgenden auf Basis-Splines (B-Splines), da Penalisierte Splines (P-Splines), die wir im Anschluss diskutieren werden, auf B-Splines aufbauen.

5.1.1 B-Splines

Bei B-Spline Basisfunktionen handelt es sich um zusammengesetzte Polynomstücke vom Grad q, die an den sogenannten Knoten miteinander verbunden werden (siehe De Boor, 2001). Die l gewählten Knoten teilen den Wertebereich der Prädiktor-Variable $[k_1, k_l] = [\min_i(x_i), \max_i(x_i)]$ in $l-1$ Teilintervalle. Dabei kann die Knotenwahl äquidistant, aber auch quantilbasiert erfolgen, sodass in jenem Bereich, in dem viele Datenpunkte vorliegen, auch eine größere Anzahl an Knoten gegeben ist.

Für die Anwendung des untenstehenden Algorithmus zur Berechnung einer B-Spline Basis ist eine erweiterte Knotenmenge zu spezifizieren. Die Anzahl der zu definierenden Knoten beträgt $l + 2q$, das heißt wir betrachten die Knoten $(k_{-(q-1)}, \ldots, k_{l+q})$, wobei $k_j \leq k_{j+1}, j \in \{-q+1, \ldots, l+q-1\}$. Die Knoten $(k_2, \ldots, k_{l-1})$ werden als innere Knoten bezeichnet und die äußeren Knoten $(k_{-(q-1)}, \ldots, k_0)$ sowie $(k_{l+1}, \ldots, k_{l+q})$ garantieren ein regelmäßiges Verhalten innerhalb des interessierenden Intervalls $[k_1, k_l]$. Die Anzahl der definierten B-Spline Basisfunktionen ist $d = l + q - 1$. Allgemeine Eigenschaften von B-Splines vom Grad q seien im Folgenden zusammengefasst und können in Abbildung 5.1 verifiziert werden:

- Die B-Spline Basisfunktion setzt sich aus $q+1$ stückweisen Polynomen vom Grad q zusammen;
- In den Knoten stimmen die Ableitungen bis zur $(q-1)$-ten Ordnung überein und sind stetig;
- Die Polynomstücke werden in q inneren Knoten verbunden;
- B-Spline Basisfunktionen sind lokal, d.h. sie sind nur über einen Bereich von $q+2$ Knoten positiv und ansonsten gleich Null;
- Jede B-Spline Basisfunktion überlappt sich mit $2q$ benachbarten Basisfunktionen;
- B-Spline Basisfunktionen sind nach oben beschränkt;
- Für ein gegebenes x sind $q+1$ B-Splines von Null verschieden;

De Boor (2001, S. 90) definiert einen rekursiven Algorithmus zur Berechnung des i-ten B-Splines $b_i^q(x)$ vom Grad q:

$$b_i^q(x) = \frac{x - k_i}{k_{i+q} - k_i} b_i^{q-1}(x) + \frac{k_{i+q+1} - x}{k_{i+q+1} - k_{i+1}} b_{i+1}^{q-1}(x),$$

$$b_i^{-1}(x) = \begin{cases} 1 & k_i \leq x < k_{i+1}, \\ 0 & \text{sonst.} \end{cases}$$

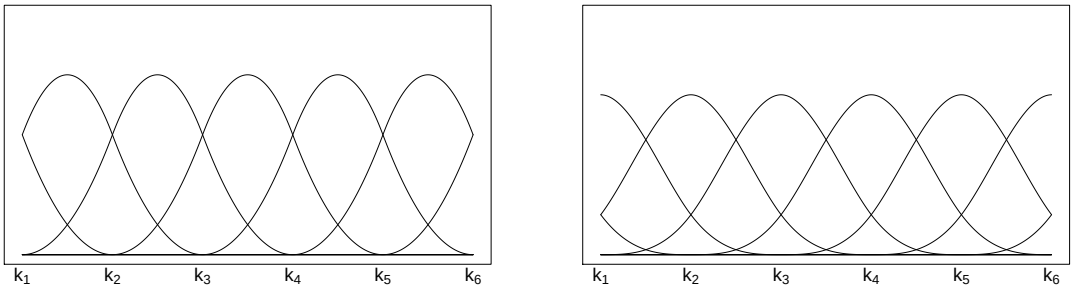

Abbildung 5.1: B-Spline Basis mit $l = 6$ äquidistant gewählten Knoten vom Grad $q = 2$ (links) und $q = 3$ (rechts).

Die zu schätzende Funktion f wird als Linearkombination der d Basisfunktionen repräsentiert

$$f(x) = \sum_{j=1}^{d} b_j^q(x)\beta_j,$$

wobei β_j die zu schätzenden Parameter bezeichnet.

Kehren wir zum Modell (5.1) zurück und betrachten Datenpaare $(x_i, y_i), i \in \{1, \ldots, n\}$, sowie die Matrix $\mathbf{B} \in \mathbb{R}^{n \times d}$ mit den Einträgen $B_{i,j} = b_j^q(x_i), j \in \{1, \ldots, d\}$, und die zu schätzenden Parameter $\boldsymbol{\beta} \in \mathbb{R}^d$, dann können wir Modell (5.1) in Matrixschreibweise umformulieren:

$$\mathbf{y} = \mathbf{B}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n).$$

Es ist also ein Lineares Modell (siehe Abschnitt 3.1) vorliegend und eine Schätzung für $f(x)$ ergibt sich über die Kleinste-Quadrate-Methode durch Minimieren von

$$(\mathbf{y} - \mathbf{B}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{B}\boldsymbol{\beta}) \tag{5.2}$$

in $\boldsymbol{\beta}$. Als Lösung erhalten wir wie schon im Abschnitt 3.1 besprochen $\hat{\boldsymbol{\beta}} = (\mathbf{B}^\top \mathbf{B})^{-1} \mathbf{B}^\top \mathbf{y}$.

Zur Illustration dieses Vorgangs, wird im linken Bild von Abbildung 5.2 die geschätzte glatte Funktion und die mit β_j gewichteten B-Spline Basisfunktionen $b_j^q(x)$ dargestellt. Datengrundlage bildet der im Zusammenhang mit nichtparametrischer Regression äußerst bekannte *motorcycle* Datensatz, der in der R-Bibliothek MASS verfügbar ist. Er besteht aus $n = 133$ Beobachtungen der Kopfbeschleunigung ($\mathbf{y}$) in Abhängigkeit von der Zeit ($\mathbf{x}$) nach einem Aufprall infolge eines simulierten Motorradunfalls. Wie im rechten Bild von Abbildung 5.2 gezeigt, führt eine große Anzahl an Knoten und damit verwendeter B-Splines zu einer größeren Flexibilität der geschätzten Funktion als durch die Daten gerechtfertigt wird, während eine geringere Anzahl eine glattere Schätzung ergibt.

Vor der Schätzung muss also der Grad q sowie die Knoten festgelegt werden, die beide einen Einfluss auf die Glattheit der Kurve haben und sich die Frage stellt, welche Wahl nun

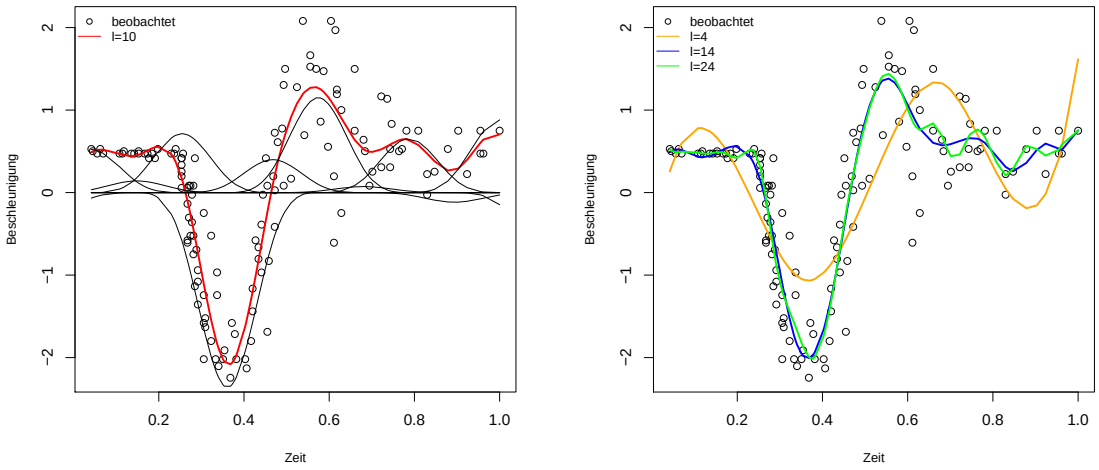

Abbildung 5.2: B-Spline Basis mit $l = 10$ inneren Knoten vom Grad $q = 3$ (links) sowie Schätzung mittels B-Splines vom Grad $q = 3$ und unterschiedlicher Knotenanzahl (rechts).

eine optimal angepasste Kurve liefert. Um dem Glättungsproblem zu begegnen, betrachten wir im Folgenden den von Eilers und Marx (1996) vorgestellten Penalisierungsansatz *difference penalty*. In groben Zügen erläutert, wird die Funktion $f(x)$ mittels B-Splines und einer ausreichend großen Anzahl an Knoten approximiert. Zusätzlich wird ein Strafterm eingeführt, der eine zu große Variation von $f(x)$ bestraft.

5.1.2 P-Splines

Im Falle des Penalisierungsansatzes von Eilers und Marx (1996) wird im zu minimierenden Ausdruck (5.2) ein additiver Strafterm hinzugefügt, der durch den Glättungsparameter $\lambda \geq 0$ gewichtet ist:

$$(\mathbf{y} - \mathbf{B}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{B}\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}^\top \mathbf{D}_c^\top \mathbf{D}_c \boldsymbol{\beta}, \tag{5.3}$$

wobei $\mathbf{D}_c\boldsymbol{\beta} = \Delta^c\boldsymbol{\beta}$, also die Matrix $\mathbf{D}_c$ Differenzen der Ordnung c von $\boldsymbol{\beta}$ erstellt. Betrachten wir beispielsweise Differenzen zweiter Ordnung, dann ergibt sich $\Delta^2\beta_j = \Delta(\Delta\beta_j) = \Delta(\beta_j - \beta_{j-1}) = \beta_j - 2\beta_{j-1} + \beta_{j-2}$ und damit die $(d-2) \times d$-Matrix

$$\mathbf{D}_2 = \begin{pmatrix} 1 & -2 & 1 & & \\ & 1 & -2 & 1 & \\ & & \ddots & \ddots & \ddots \\ & & & 1 & -2 & 1 \end{pmatrix}.$$

Als Lösung erhalten wir gemäß dem Bilden der Ableitung der Zielfunktion (5.3) nach $\boldsymbol{\beta}$ und anschließendem Nullsetzen

$$\hat{\boldsymbol{\beta}} = (\mathbf{B}^\top \mathbf{B} + \lambda \mathbf{D}_c^\top \mathbf{D}_c)^{-1} \mathbf{B}^\top \mathbf{y}.$$

Die Steuerung der Glattheit der Schätzung erfolgt bei diesem Ansatz nicht mehr über die Anzahl und die Positionierung von Knoten, sondern über einen einzelnen Parameter, was einen klaren Vorteil darstellt. Durch die passende Wahl des Glättungsparameters, die im nächsten Abschnitt detailliert besprochen wird, kann einer zu starken Anpassung der geschätzten Kurve an die Daten entgegengewirkt werden. Für $\lambda \to 0$ erhält der Bestrafungsterm eine geringe Gewichtung und die Schätzung wird demzufolge jener in (5.2) ähneln, wohingegen für $\lambda \to \infty$ die Schätzung durch den Strafterm dominiert wird.

Wie sich die unterschiedliche Wahl von λ auf die Glattheit auswirkt und welcher Unterschied zwischen einer penalisierten und nicht-penalisierten ($\lambda = 0$) Schätzung besteht, soll in Abbildung 5.3 anhand der motorcycle Daten veranschaulicht werden.

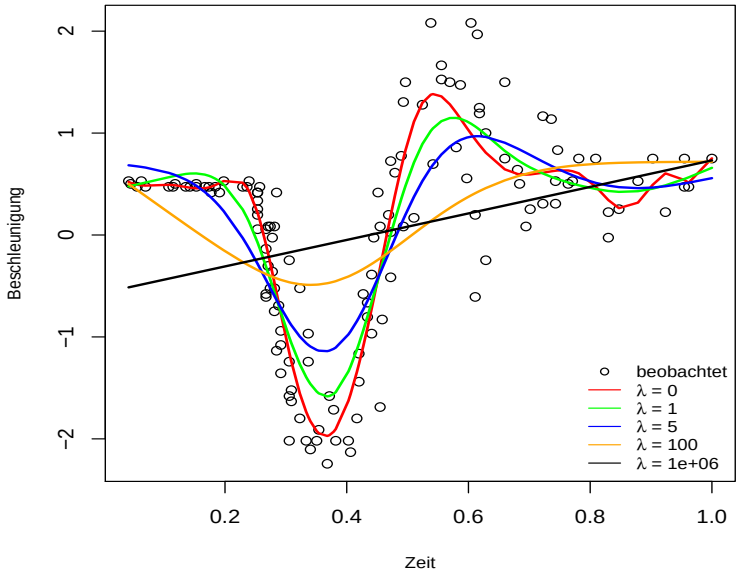

Abbildung 5.3: P-Splines vom Grad $q = 3$ mit $l = 15$ inneren Knoten und unterschiedlichen Werten für λ.

5.2 Das eindimensionale P-Spline Modell und die Anwendung auf Mortalitätsdaten

Um die Theorie der P-Splines auch für Mortalitätsdaten, die als poissonverteilt angesehen werden, anwendbar zu machen, ist es notwendig eine Verallgemeinerung hinsichtlich der Verteilungsannahme zu treffen. Zudem wird es für das Modellieren sowohl in Richtung des Alters, als auch des Beobachtungsjahres, erforderlich sein eine Basis für eine zweidimensionale Regression zu erstellen, was später diskutiert wird.

Wie schon im Abschnitt 3.2.2 besprochen, wird in einem GLM ein linearer Prädiktor $\boldsymbol{\eta} = \mathbf{B}\boldsymbol{\beta}$ mit der Link-Funktion $\boldsymbol{\eta} = g(\boldsymbol{\mu})$ definiert, wobei wie zuvor $\mu_i = \mathrm{E}(y_i)$. Die

Responses y_i seien poissonverteilt und als Link-Funktion spezifizieren wir den kanonischen Link $\boldsymbol{\eta} = g(\boldsymbol{\mu}) = \log(\boldsymbol{\mu}) = \mathbf{B}\boldsymbol{\beta}$. Der Strafterm wird zur Schätzung des Parametervektors $\boldsymbol{\beta}$ von der Log-Likelihood-Funktion $l(\boldsymbol{\beta}|\mathbf{y})$ subtrahiert und wir erhalten die penalisierte Log-Likelihood-Funktion

$$l^*(\boldsymbol{\beta}|\mathbf{y}) = l(\boldsymbol{\beta}|\mathbf{y}) - \frac{\lambda}{2}\boldsymbol{\beta}^\top \mathbf{D}_c^\top \mathbf{D}_c \boldsymbol{\beta},$$

wobei der Faktor $\frac{1}{2}$ so gewählt wurde, dass er nach Differentiation verschwindet.

Im nächsten Schritt wollen wir l^* maximieren. Die penalisierte Log-Likelihood-Funktion hat die Form

$$l^*(\boldsymbol{\beta}|\mathbf{y}) = \sum_{i=1}^{n} \left(\frac{y_i\theta_i - b(\theta_i)}{a_i\phi} + c(y_i, \phi) \right) - \frac{\lambda}{2} \sum_{j=c+1}^{d} (\Delta^c \beta_j)^2.$$

Unter Anwendung der Kettenregel

$$\frac{\partial l^*}{\partial \beta_j} = \frac{\partial l^*}{\partial \theta} \frac{\partial \theta}{\partial \mu} \frac{\partial \mu}{\partial \eta} \frac{\partial \eta}{\partial \beta_j}$$

erhalten wir für die Ableitung von l^* nach β_j den Ausdruck

$$\frac{\partial l^*}{\partial \beta_j} = \sum_{i=1}^{n} \left(\frac{y_i - b'(\theta)}{a_i\phi} \frac{1}{V(\mu_i)} \frac{\partial \mu_i}{\partial \eta_i} b_{ij} \right) - \frac{\lambda}{2} \sum_{j=c+1}^{d} \frac{\partial (\Delta^c \beta_j)^2}{\partial \beta_j}.$$

Unter der Annahme poissonverteilter Responses und dem Log-Link, folgt somit für $\frac{\partial \eta_i}{\partial \mu_i} = \frac{\partial \log(\mu_i)}{\partial \mu_i} = \frac{1}{\mu_i}$, $a_i = 1$, $\phi = 1$ und $V(\mu_i) = b''(\theta_i) = \mu_i$ (siehe Abschnitt 3.2.1 und 3.2.2) und damit

$$\frac{\partial l^*}{\partial \beta_j} = \sum_{i=1}^{n} (y_i - \mu_i) b_{ij} - \frac{\lambda}{2} \sum_{j=c+1}^{d} \frac{\partial (\Delta^c \beta_j)^2}{\partial \beta_j}$$

bzw. in Matrixschreibweise und mit $\mathbf{P} = \lambda \mathbf{D}_c^\top \mathbf{D}_c$

$$\frac{\partial l^*}{\partial \beta_j} = \mathbf{B}^\top (\mathbf{y} - \boldsymbol{\mu}) - \mathbf{P}\boldsymbol{\beta}.$$

Die Maximum-Likelihood-Schätzung erfolgt in diesem Fall mit einem leicht abgeänderten IWLS-Algorithmus. Bevor wir die Iterationsvorschrift der Newton-Raphson-Methode anwenden, berechnen wir die zweite Ableitung

$$\frac{\partial^2 l^*}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} = -\mathbf{B}^\top \mathbf{W} \mathbf{B} - \mathbf{P},$$

wobei $\mathbf{W} = \text{diag}(\boldsymbol{\mu})$. Das Newton-Raphson-Verfahren ergibt schließlich

$$\begin{aligned}
\boldsymbol{\beta}^{(t+1)} &= \boldsymbol{\beta}^{(t)} + \left(\mathbf{B}^\top \mathbf{W}^{(t)} \mathbf{B} + \mathbf{P} \right)^{-1} \left(\mathbf{B}^\top (\mathbf{y} - \boldsymbol{\mu}^{(t)}) - \mathbf{P}\boldsymbol{\beta}^{(t)} \right) \\
&= \left(\mathbf{B}^\top \mathbf{W}^{(t)} \mathbf{B} + \mathbf{P} \right)^{-1} \mathbf{B}^\top (\mathbf{y} - \boldsymbol{\mu}^{(t)}) + \left(\mathbf{B}^\top \mathbf{W}^{(t)} \mathbf{B} + \mathbf{P} \right)^{-1} \left(\mathbf{B}^\top \mathbf{W}^{(t)} \mathbf{B} \right) \boldsymbol{\beta}^{(t)}.
\end{aligned}$$

Setzen wir schließlich noch $\mathbf{z} = \mathbf{W}^{-1}(\mathbf{y} - \boldsymbol{\mu}) + \mathbf{B}\boldsymbol{\beta}$ als Pseudobeobachtungen, dann ist sofort die Ähnlichkeit zum Standardverfahren für das Schätzen von GLMen ersichtlich:

$$\boldsymbol{\beta}^{(t+1)} = \left(\mathbf{B}^\top \mathbf{W}^{(t)} \mathbf{B} + \mathbf{P}\right)^{-1} \mathbf{B}^\top \mathbf{W}^{(t)} \mathbf{z}^{(t)}. \tag{5.4}$$

Der einzige Unterschied liegt in der Modifikation von $\mathbf{B}^\top\mathbf{W}\mathbf{B}$ durch das Addieren des Strafterms $\mathbf{P}$. Aus der obigen Darstellung lässt sich die Form der Hat-Matrix

$$\mathbf{H}_\lambda = \mathbf{B}\left(\mathbf{B}^\top \mathbf{W}\mathbf{B} + \mathbf{P}\right)^{-1} \mathbf{B}^\top \mathbf{W} \tag{5.5}$$

ableiten, wobei die Matrix $\mathbf{W}$ die Gewichte der letzten Iteration nach Konvergenz enthält.

5.2.1 Wahl des Glättungsparameters λ

Wie schon oben erwähnt, ist die Balance zwischen Modellkomplexität und der Fähigkeit des Modells die beobachteten Daten zu reproduzieren über den Glättungsparameter steuerbar, dessen optimale Wahl hier besprochen wird.

Ein Qualitätskriterium für das verwendete Modell ist der mittlere quadratische Fehler (MSE, *mean squared error*)

$$\mathrm{MSE}(x_i, \lambda) = \mathrm{E}\left((\hat{f}(x_i) - f(x_i))^2\right) = \mathrm{Var}(\hat{f}(x_i)) + b_\lambda(x_i)^2,$$

der sich bekanntlich als die Summe des quadrierten Bias und der Varianz schreiben lässt und im Allgemeinen für jedes x_i unterschiedliche Werte annimmt. In diesem Zusammenhang gilt für den Bias

$$\begin{aligned}
\mathbf{b}_\lambda &= \mathrm{E}(\hat{\mathbf{f}}) - \mathbf{f} = \mathrm{E}(\mathbf{B}\hat{\boldsymbol{\beta}}) - \mathbf{f} = \mathbf{B}\mathrm{E}(\hat{\boldsymbol{\beta}}) - \mathbf{f} \\
&= \mathbf{B}\left(\mathbf{B}^\top \mathbf{W}\mathbf{B} + \mathbf{P}\right)^{-1} \mathbf{B}^\top \mathbf{W}\mathbf{W}^{-1}\mathrm{E}(\mathbf{y} - \boldsymbol{\mu}) + \mathbf{B}\left(\mathbf{B}^\top \mathbf{W}\mathbf{B} + \mathbf{P}\right)^{-1} \mathbf{B}^\top \mathbf{W}\mathbf{B}\boldsymbol{\beta} - \mathbf{f} \\
&= \mathbf{H}_\lambda \mathbf{B}\boldsymbol{\beta} - \mathbf{f} = (\mathbf{H}_\lambda - \mathbf{I})\boldsymbol{f}
\end{aligned}$$

und für die Varianz

$$\begin{aligned}
\mathrm{Var}(\hat{\mathbf{f}}) &= \mathrm{Var}(\mathbf{B}\hat{\boldsymbol{\beta}}) = \mathbf{B}\mathrm{Var}(\hat{\boldsymbol{\beta}})\mathbf{B}^\top \\
&= \mathbf{B}\left(\mathbf{B}^\top \mathbf{W}\mathbf{B} + \mathbf{P}\right)^{-1} \mathbf{B}^\top \mathbf{W}\mathrm{Var}\left(\mathbf{W}^{-1}(\mathbf{y} - \boldsymbol{\mu})\right) \mathbf{W}^\top \mathbf{B}\left(\mathbf{B}^\top \mathbf{W}\mathbf{B} + \mathbf{P}\right)^{-1} \mathbf{B}^\top \\
&= \mathbf{B}\left(\mathbf{B}^\top \mathbf{W}\mathbf{B} + \mathbf{P}\right)^{-1} \mathbf{B}^\top \mathbf{W}\mathbf{B}\left(\mathbf{B}^\top \mathbf{W}\mathbf{B} + \mathbf{P}\right)^{-1} \mathbf{B}^\top.
\end{aligned}$$

Um ein einzelnes Kriterium zu bekommen, bilden wir den Mittelwert aller MSEs der x-Werte (siehe Hastie & Tibshirani, 1990):

$$\begin{aligned}
\mathrm{MSE}(\lambda) &= \frac{1}{n}\sum_{i=1}^{n} \mathrm{MSE}(x_i, \lambda) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{Var}(\hat{f}(x_i)) + \frac{1}{n}\sum_{i=1}^{n} b_\lambda(x_i)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathrm{Var}(\hat{f}(x_i)) + \frac{1}{n}\sum_{i=1}^{n} b_\lambda(x_i)^2 = \frac{\mathrm{tr}(\mathrm{Var}(\hat{\mathbf{f}}))}{n} + \frac{\mathbf{b}_\lambda^\top \mathbf{b}_\lambda}{n}.
\end{aligned}$$

Für ein passendes Modell sollten sowohl der Bias, als auch die Varianz klein sein, da somit die Schätzungen $\hat{f}(x_i)$ wie gewünscht nahe bei $f(x_i)$ liegen und darüberhinaus nicht zu variabel sind. Eine simultane Minimierung beider Größen ist jedoch nicht möglich. So reduziert ein glatteres Schätzverfahren (λ wird größer) die Varianz, führt aber zur gleichen Zeit zu einer stärkeren Verzerrung. Umgekehrt resultiert eine Verkleinerung von λ in einem geringeren Bias (weniger glatte Schätzung), lässt aber die Varianz wachsen. Aus diesem Grund muss ein passender Kompromiss über die Auswahl des Glättungsparameters gefunden werden.

In der Literatur werden verschiedene Ansätze zur Wahl des Glättungsparameters verwendet. Dies kann entweder basierend auf der (generalisierten) Kreuzvalidierung oder einem Informationskriterium, AIC oder BIC, erfolgen. Bei der Kreuzvalidierung (CV, *Cross-validation*) (siehe Wood, 2006, S. 169ff) wird ein Datenpunkt (x_i, y_i) entfernt und anhand der verbleibenden $n-1$ Beobachtungen eine Schätzung bei gegebenem λ berechnet. Anschließend wird der fitted value für den davor entfernten Datenpunkt $\hat{\mu}_i^{(-i)}$ vorhergesagt. Durch Wiederholung dieses Vorgangs für alle Beobachtungen kann der Vorhersagefehler

$$\mathrm{CV}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{\mu}_i^{(-i)}\right)^2$$

bestimmt werden. Das optimale λ minimiert $\mathrm{CV}(\lambda)$. Um nicht n Schätzungen des nichtparametrischen Modells vornehmen zu müssen und somit auch die Laufzeit zur Berechnung von $\mathrm{CV}(\lambda)$ zu verbessern, kann gezeigt werden, dass $\mathrm{CV}(\lambda)$ mit dem Vektor der fitted values $\hat{\boldsymbol{\mu}}$ und den Diagonalelementen der Hat-Matrix $\boldsymbol{H}_\lambda$ berechenbar ist:

$$\mathrm{CV}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{\mu}_i}{1 - h_{ii}}\right)^2.$$

Dennoch ist einerseits die Berechnung von $\mathrm{CV}(\lambda)$ numerisch aufwändig und andererseits ist ein Invarianzproblem bezüglich einer Rotation der Daten mit einer orthogonalen Matrix gegeben. Genauer sei $\mathbf{Q}$ eine orthogonale Matrix, dann sind alle Aussagen die auf Basis der Minimierung der Zielfunktion $(\mathbf{y} - \mathbf{B}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{B}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^\top\mathbf{D}_c^\top\mathbf{D}_c\boldsymbol{\beta}$ und jene basierend auf $(\mathbf{Qy} - \mathbf{QB}\boldsymbol{\beta})^\top(\mathbf{Qy} - \mathbf{QB}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^\top\mathbf{D}_c^\top\mathbf{D}_c\boldsymbol{\beta}$ äquivalent, wobei $\mathbf{Q}$ die passende Dimension habe. Im Allgemeinen gilt dies aber nicht für den Wert der Kreuzvalidierung. Daher werden für die generalisierte Kreuzvalidierung (GCV, *generalized cross-validation*) die Diagonalelemente h_{ii} durch deren Mittelwert ersetzt:

$$\mathrm{GCV}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{\mu}_i}{1 - \frac{\mathrm{tr}(\mathbf{H}_\lambda)}{n}}\right)^2.$$

Wie oben erwähnt, ist die Wahl des Glättungparameters neben dem Kreuzvalidierungsverfahren auch über die schon im Abschnitt 3.2.3 eingeführten Informationskriterien durchführbar. Dazu benötigen wir die Dimension des Modells, die im Falle von GLMen über die Anzahl der Parameter gegeben ist. Im Falle von nichtparametrischen Modellen führen wir den Begriff der *effective dimension* ein, die nach Hastie und Tibshirani (1990) als die Summe der Diagonalelemente der Hat-Matrix definiert wird:

$$\mathrm{ED}(\boldsymbol{\beta}, \lambda) = \mathrm{tr}(\mathbf{H}_\lambda).$$

Diese Definition ist angelehnt an die schon im Linearen Modell besprochene Eigenschaft der Hat-Matrix die Anzahl der Parameter zu reproduzieren. Für P-Spline Modelle bewährt sich analog dazu die effective dimension zur Quantifizierung der Modellkomplexität. Betrachten wir ein nicht-penalisiertes Modell, so liefert $\text{tr}(\mathbf{H}_\lambda)$ die Anzahl der verwendeten Basisfunktionen, wohingegen diese bei Penalisierungsansätzen durch den Strafterm reduziert wird. Die effective dimension wird also kleiner je größer λ wird. Es gilt, dass sich $\text{tr}(\mathbf{H}_\lambda)$ für immer größer werdendes λ der verwendeten Differenzenordnung c annähert (siehe Eilers & Marx, 1996). Für $\mathbf{Q}_B := \mathbf{B}^\top\mathbf{W}\mathbf{B}$ ergibt sich

$$\begin{aligned}
\text{tr}(\mathbf{H}_\lambda) &= \text{tr}\left(\mathbf{B}\left(\mathbf{B}^\top\mathbf{W}\mathbf{B}+\mathbf{P}\right)^{-1}\mathbf{B}^\top\mathbf{W}\right) \\
&= \text{tr}\left(\left(\mathbf{B}^\top\mathbf{W}\mathbf{B}+\mathbf{P}\right)^{-1}\mathbf{B}^\top\mathbf{W}\mathbf{B}\right) \\
&= \text{tr}\left(\mathbf{Q}_B^{1/2}(\mathbf{Q}_B+\mathbf{P})^{-1}\mathbf{Q}_B^{1/2}\right) \\
&= \text{tr}\left((\mathbf{I}+\mathbf{Q}_B^{-1/2}\mathbf{P}\mathbf{Q}_B^{-1/2})^{-1}\right) \\
&= \text{tr}\left((\mathbf{I}+\lambda\mathbf{Q}_B^{-1/2}\mathbf{D}_c^\top\mathbf{D}_c\mathbf{Q}_B^{-1/2})^{-1}\right) = \sum_{i=1}^{n}\frac{1}{1+\lambda\gamma_i},
\end{aligned}$$

wobei im ersten Schritt die Eigenschaft der zyklischen Vertauschung $\text{tr}(ABC) = \text{tr}(BCA)$ verwendet wird. Im letzten Schritt kommt für eine Matrix A mit den Eigenwerten δ_i die Beziehung $\text{tr}(A) = \sum_{i=1}^{n}\delta_i$ zur Anwendung und dass die Eigenwerte von A^{-1} gleich $1/\delta_1, 1/\delta_2, \ldots, 1/\delta_n$ sind. Im obigen Ausdruck bezeichnet $\gamma_i, i \in \{1,\ldots,n\}$, die Eigenwerte von $\mathbf{Q}_B^{-1/2}\mathbf{D}_c^\top\mathbf{D}_c\mathbf{Q}_B^{-1/2}$. Da c Eigenwerte von $\mathbf{D}_c^\top\mathbf{D}_c$ gleich Null sind, ist ebenso die Anzahl der Eigenwerte der Matrix $\mathbf{Q}_B^{-1/2}\mathbf{D}_c^\top\mathbf{D}_c\mathbf{Q}_B^{-1/2}$, die Null sind, gleich c. Wird nun λ groß, so tragen jene Summanden, in denen $\gamma_i = 0$ zur Summe bei und $\text{tr}(\mathbf{H}_\lambda)$ nähert sich dem Wert c.

Die Abbildung 5.4 zeigt den Verlauf der effective dimension für einen Strafterm mit Differenzen zweiter bzw. dritter Ordnung. Wie oben beschrieben nähert sich für großes λ die effective dimension der Ordnung c an. Außerdem ist die effective dimension durch die Anzahl der verwendeten Basisfunktionen (im Falle des vorliegenden Beispiels kommen $d = 13$ Basisfunktionen zur Anwendung) nach oben beschränkt.

Zur Wahl des optimalen Glättungsparameters wird von Eilers und Marx (1996) Akaike's Informationskriterium vorgeschlagen.

$$\text{AIC}(\lambda) = \text{D}(\mathbf{y}, \boldsymbol{\beta}, \lambda) + 2\text{ED}(\boldsymbol{\beta}, \lambda),$$

wobei die Deviance bei poissonverteilten Responses im Abschnitt 3.2.3 gegeben ist. Von Currie (2013) wird als Auswahlkriterium das Bayesianisches Informationskriterium

$$\text{BIC}(\lambda) = \text{D}(\mathbf{y}, \boldsymbol{\beta}, \lambda) + \log(n)\text{ED}(\boldsymbol{\beta}, \lambda)$$

bevorzugt, da dieses höhere Modellkomplexität stärker bestraft als das AIC.

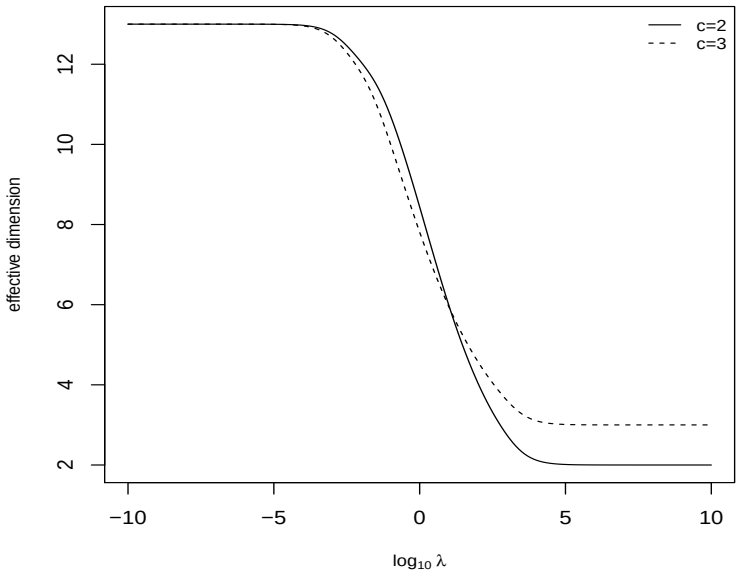

Abbildung 5.4: Verlauf der effective dimension in Abhängigkeit von λ und für $c = 2$ und $c = 3$.

5.2.2 Glättung in einer Dimension am Beispiel von Mortalitätsdaten

Kehren wir zum vorliegenden Datensatz zurück, um die oben eingeführte Theorie anhand eines Beispiels zu veranschaulichen. Dazu verwenden wir wiederum Mortalitätsdaten weiblicher Versicherter im Altersintervall $[45, 95]$ des Beobachtungsjahres 2015. Um Mortalitätsraten zu modellieren, betrachten wir das Modell

$$\boldsymbol{\eta} = \log \boldsymbol{\mu} = \mathbf{B}\boldsymbol{\beta} + \log \mathbf{e}.$$

Für praktische Aspekte geben Eilers und Marx (2002) ein „Rezept“ für die Wahl des Settings im P-Spline Modell an. Die Knoten sollen über den Definitionsbereich von x gleichverteilt sein und je ein Knoten für alle vier bis fünf Beobachtungen gesetzt werden. Allgemein hat die Knotenanzahl keinen maßgeblichen Einfluss auf die Schätzung, solange ausreichend viele Knoten spezifiziert werden und damit eine flexible Modellierung möglich ist. Außerdem soll eine quadratische oder kubische B-Spline Basis mit einem Strafterm ebenso von zweiter oder dritter Ordnung gewählt werden.

Ein für die Modellierung von Mortalitätsraten maßgeschneidertes R-Paket ist das von Carlo G. Camarda entwickelte MortalitySmooth Paket (siehe Camarda, 2012 und Camarda, 2015), welches eine Glättung von Sterberaten mit P-Splines zulässt. Das Herzstück für die eindimensionale Modellierung bildet die Funktion Mort1Dsmooth, für die als Input der Altersvektor, der Vektor der je Alter beobachteten Sterbefälle und als Offset der Vektor der dem Sterberisiko ausgesetzten Personen dienen. Gemäß der für die praktische Anwendung von Eilers und Marx gegebenen Daumenregel werden in der Standardoption kubische B-Splines und ein quadratischer Strafterm mit einem Knoten für alle fünf Be-

obachtungen verwendet. Diese Voreinstellung kann jedoch vom Benutzer geändert werden.

Das nachstehende Listing zeigt, wie für die vorliegende Datensituation mittels der Funktion Mort1Dsmooth das eindimensionale P-Spline Modell angepasst werden kann. Der optimale Glättungsparameter wird dabei, je nachdem welches Kriterium über die Option method spezifiziert ist, über die Minimierung des BICs (Standardeinstellung) oder des AICs (method=2) bestimmt. Der Benutzer hat jedoch auch die Möglichkeit λ selbst anzugeben (method=3). Im Paket MortalitySmooth wird intern die Funktion cleversearch, welche im R-Paket svcm (siehe Heim, 2009) verfügbar ist, aufgerufen. Diese ermöglicht die Minimierung einer gegebenen Funktion über ein vom Benutzer definiertes Parametergitter. Für den vorliegenden Datensatz muss das Gitter, über dem das optimale λ gesucht wird, jedoch verfeinert werden, da die Voreinstellung mit TOL2 $= 0.5$ auf der logarithmischen Skala zu ungenau ist.

```
1  > sex<-"W"
2  > t<-2015
3  > a<-seq(45,95,1)
4  > deaths<-data.select(t,a,sex,"Deaths")
5  > exposures<-data.select(t,a,sex,"Exposures")

7  > library('MortalitySmooth')
8  > model_bic<-Mort1Dsmooth(x=a, y=deaths, offset=log(exposures),
9  + control=list(TOL2=0.0001))
10 > model_aic<-Mort1Dsmooth(x=a, y=deaths, offset=log(exposures),
11 + control=list(TOL2=0.0001), method=2)
```

In den Modellen model_bic und model_aic werden kubische B-Splines ($q = 3$), ein Strafterm zweiter Ordnung ($c = 2$) sowie $l = 11$ Knoten bei einer Anzahl von 51 Beobachtungen verwendet. Somit gehen $d = 13$ Basisfunktionen ins Modell ein und es sind 13 Parameter zu schätzen. Die effective dimension beträgt nach der Wahl des Glättungsparameters mittels BIC (model_bic) ED $= 3.25$, wobei $\lambda = 353.83$ bei einem BIC $= 66.91$ und einer Deviance von 54.11. Im Vergleich dazu liegt die effective dimension des korrespondierenden nicht-penalisierten B-Spline Modells ($\lambda = 0$) bei 13 und weist eine Deviance von 48.24 auf. Hinsichtlich des AICs als Optimalitätskriterium (model_aic) erhalten wir ein optimales $\lambda = 148.18$ bei einer effective dimension von 3.78 und einer Deviance D $= 52.67$. Dabei wird deutlich, dass der optimale Glättungsparameter unter dem BIC stets größer ist und somit eine glattere Schätzung liefert, als unter der Verwendung des AICs. Dies ist auch in Abbildung 5.5, wo die unter den beiden Kriterien geschätzten Mortalitätsraten dargestellt sind, minimal erkennbar.

Zur Wahl des optimalen Glättungsparameters illustriert Abbildung 5.6 die Werte des jeweiligen Informationskriteriums für unterschiedliche Werte von λ. Jenes λ, das das AIC bzw. das BIC minimiert, ist an der vertikalen Linie ersichtlich.

Alternativ können P-Splines auch mit der Funktion gam angepasst werden, die im von Wood (2017) entwickelten R-Paket mgcv verfügbar ist. Der Code zur Modellierung der Mortalitätsraten mit obigen Daten und kubischen B-Splines sowie einer Penalisierung zweiter Ordnung unter Verwendung von $d = 13$ Basisfunktionen ist nachstehend verfügbar. Als

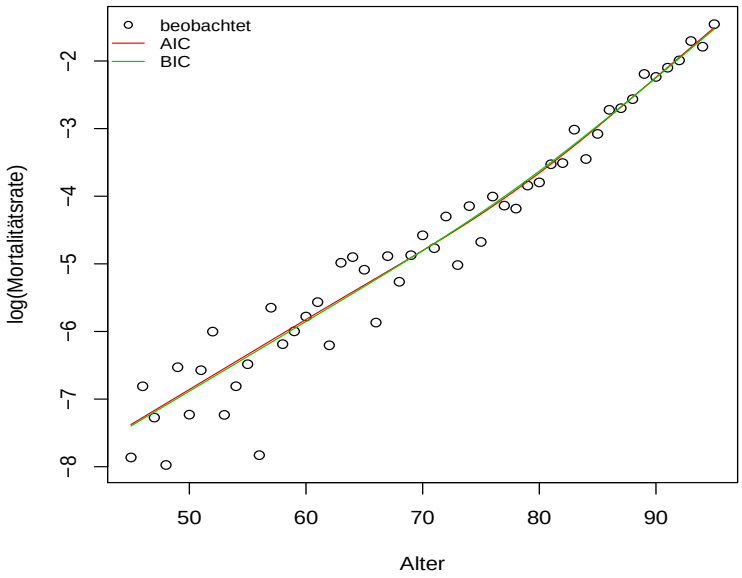

Abbildung 5.5: Glatte Mortalitätsraten unter Verwendung des unter dem AIC (rot) bzw. BIC (grün) optimierten Glättungsparameters.

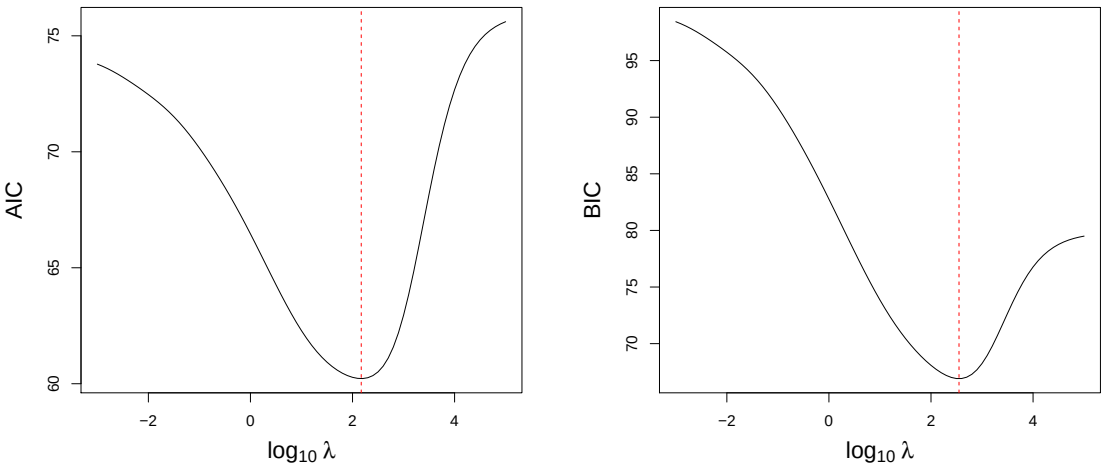

Abbildung 5.6: Verlauf des AICs (links) und des BICs (rechts) für gegebenes λ. Die rote Linie kennzeichnet jeweils für welches λ das AIC ($\lambda = 148.18$) bzw. das BIC ($\lambda = 353.83$) minimiert wird.

Optimalitätskriterium für den Glättungsparameter wird in diesem Fall das GCV verwendet und führt zu einer Schätzung, die jener unter Verwendung des AICs im Bild 5.5 ähnelt.

```
> library('mgcv')
> model_gam<-gam(deaths~s(a,bs='ps',m=c(2,3),k=13) +
+ offset(log(exposures)), family='poisson')
> log(fitted(model_gam)/exposures)
```

5.3 Das zweidimensionale P-Spline Modell und die Anwendung auf Mortalitätsdaten

Wir gehen nun zur Modellierung von Mortalitätsraten in Richtung des Alters und des Beobachtungsjahres über und verwenden dazu die gesamte Historie vorhandener Daten. Allgemein betrachtet stellt das Modell der Form

$$\mathrm{E}(\mathbf{y}) = f(\mathbf{x}_a, \mathbf{x}_t)$$

also den Zusammenhang der Response $\mathbf{y}$ mit den beiden Prädiktoren $\mathbf{x}_a$ und $\mathbf{x}_t$ dar.

Im Hinblick auf die Mortalitätsdaten gilt in der Folge die nachstehende Notation, $\mathbf{Y} = (y_{ij}) \in \mathbb{R}^{n_a \times n_t}$ sei die Matrix der beobachteten Sterbefälle und $\mathbf{E} = (e_{ij}) \in \mathbb{R}^{n_a \times n_t}$, $i \in \{1, \ldots, n_a\}, j \in \{1, \ldots, n_t\}$, sei die Matrix der dem Risiko des Todes unterliegenden Personenzahlen.

Analog zum eindimensionalen Fall erzeugen wir eine B-Spline Basismatrix, wobei die Verwendung des Kronecker-Produktes zweier Matrizen es erlauben wird, vom eindimensionalen Fall des P-Spline Modells auf den zweidimensionalen zu verallgemeinern. Sei dazu $\mathbf{B}_a \in \mathbb{R}^{n_a \times d_a}$ die auf dem Alter basierende Regressionsmatrix und $\mathbf{B}_t \in \mathbb{R}^{n_t \times d_t}$ die B-Spline Matrix für das Beobachtungsjahr. Im zweidimensionalen Modell hat die Modellmatrix $\mathbf{B}$ schließlich die folgende Form:

$$\mathbf{B} = \mathbf{B}_t \otimes \mathbf{B}_a,$$

wobei das Kronecker-Produkt der $n_t \times d_t$-Matrix $\mathbf{B}_t = (b^t_{ij})$ und der $n_a \times d_a$-Matrix $\mathbf{B}_a = (b^a_{ij})$ eine $n_a n_t \times d_a d_t$-Matrix der Gestalt

$$\mathbf{B}_t \otimes \mathbf{B}_a = \begin{pmatrix} b^t_{11}\mathbf{B}_a & b^t_{12}\mathbf{B}_a & \ldots & b^t_{1d_t}\mathbf{B}_a \\ b^t_{21}\mathbf{B}_a & b^t_{22}\mathbf{B}_a & \ldots & b^t_{2d_t}\mathbf{B}_a \\ \vdots & \vdots & \ddots & \vdots \\ b^t_{n_t1}\mathbf{B}_a & b^t_{n_t2}\mathbf{B}_a & \ldots & b^t_{n_td_t}\mathbf{B}_a \end{pmatrix}$$

ergibt.

Zur Veranschaulichung ist in Abbildung 5.7 das Kronecker-Produkt von zwei kubischen B-Splines dargestellt. Für die Modellierung werden jedoch über dem Alter-Jahr Raster eine

Vielzahl von B-Splines verwendet. Einen Eindruck wie das Kronecker-Produkt mehrerer B-Splines aussieht, soll Abbildung 5.8 geben. Jeder Hügel entspricht dem Kronecker-Produkt zweier eindimensionaler B-Splines, einer in Richtung des Alters und einer in Richtung des Beobachtungsjahres.

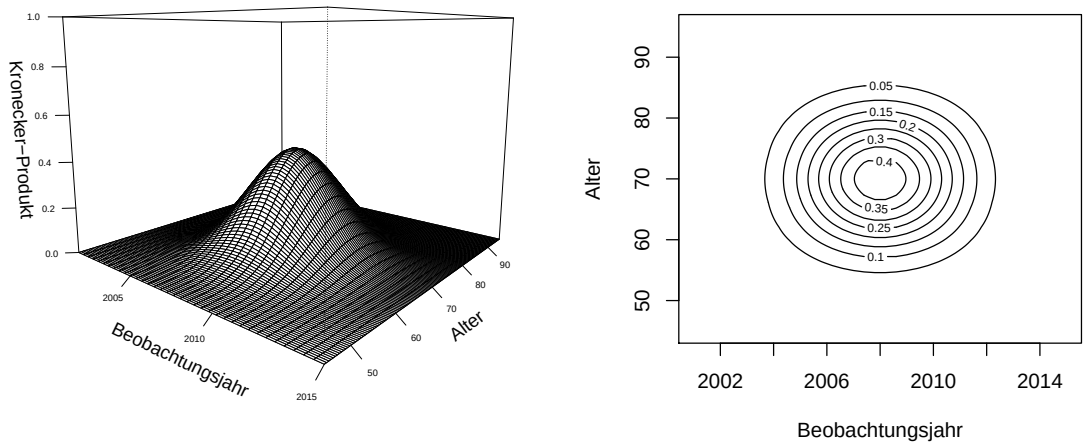

Abbildung 5.7: Kronecker-Produkt zweier eindimensionaler kubischer B-Splines, dargestellt über das Altersintervall $[45, 95]$ und die Beobachtungsjahre $[2001, 2015]$.

Ziel ist es, eine glatte Oberfläche für die Mortalitätsraten $\boldsymbol{\mu}$ zu erhalten. Für das nachstehende bivariate Modell werden die Daten derart aufbereitet, dass $\mathbf{y} = \text{vec}(\mathbf{Y})$ bzw. $\mathbf{e} = \text{vec}(\mathbf{E})$ (der $n_a n_t$-dimensionale Vektor $\mathbf{y}$ beinhaltet demnach die Sterbefälle in $\mathbf{Y}$ in Spaltenreihenfolge; Analoges gilt für $\mathbf{e}$). Des Weiteren enthält der $d_a d_t$-dimensionale Vektor $\boldsymbol{\gamma}$ die korrespondierenden zu schätzenden Regressionskoeffizienten. Wir betrachten unter der Annahme poissonverteilter Responses das Modell

$$\boldsymbol{\eta} = \log(\boldsymbol{\mu}) = (\mathbf{B}_t \otimes \mathbf{B}_a)\boldsymbol{\gamma} + \log \mathbf{e} = \mathbf{B}\boldsymbol{\gamma} + \log \mathbf{e}. \tag{5.6}$$

Arrangieren wir die Einträge in $\boldsymbol{\gamma}$ in Form einer $d_a \times d_t$-Matrix $\boldsymbol{G}$, wobei wiederum $\boldsymbol{\gamma} = \text{vec}(\mathbf{G}) = (\gamma_{1,1}, \ldots, \gamma_{d_a,1}, \ldots, \gamma_{1,d_t}, \ldots, \gamma_{d_a,d_t})^\top$ und die Spalten bzw. Zeilen von $\mathbf{G}$ mit

$$\mathbf{G} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_{d_t}), \quad \mathbf{G}^r = (\boldsymbol{\gamma}_1^r, \ldots, \boldsymbol{\gamma}_{d_a}^r) \tag{5.7}$$

bezeichnet werden, dann kann der lineare Prädiktor der j-ten Spalte von $\mathbf{Y}$ wegen der Definition des Kronecker-Produktes folgendermaßen dargestellt werden:

$$\sum_{l=1}^{d_t} b_{jl}^t \mathbf{B}_a \boldsymbol{\gamma}_l.$$

Die linearen Prädiktoren der Spalten von $\mathbf{Y}$ sind also als Linearkombination von d_t Glättungen in Altersrichtung auffassbar. Analog gilt auch für den linearen Prädiktor der i-ten Zeile von $\mathbf{Y}$

$$\sum_{l=1}^{d_a} b_{il}^a \mathbf{B}_t^\top \boldsymbol{\gamma}_l^r,$$

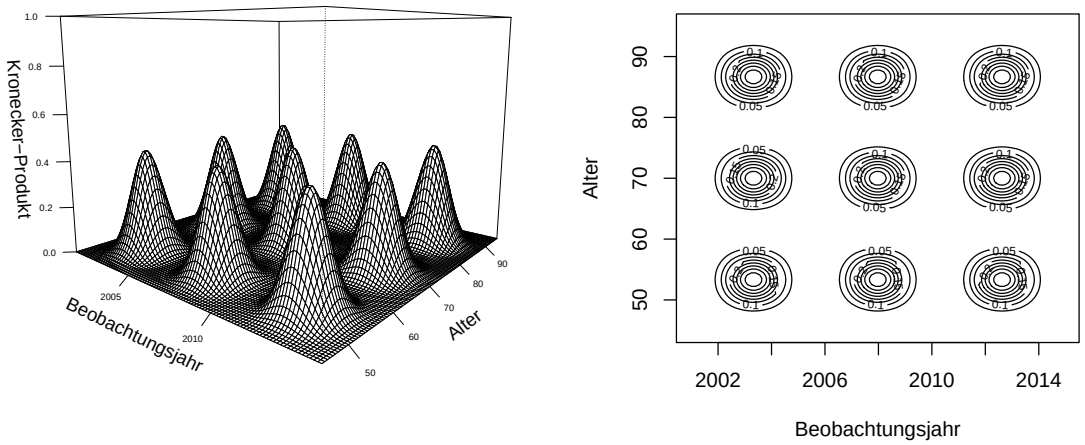

Abbildung 5.8: Ausgewählte Kronecker-Produkte eindimensionaler, kubischer B-Splines über dem Altersintervall $[45, 95]$ und den Beobachtungsjahren $[2001, 2015]$.

weshalb es naheliegend ist, für jede der Zeilen und Spalten von $\mathbf{G}$ eine Penalisierung anzuwenden. Im eindimensionalen Fall hat die quadrierte Differenzenbildung von Koeffizienten, die zu benachbarten Basisfunktionen gehören, den Strafterm gebildet. Für die Verallgemeinerung ins Zweidimensionale ist zuvor der Begriff der Nachbarschaft zu definieren (vgl. Fahrmeir, Kneib, Lang & Marx, 2013). In Abbildung 5.9 sind mögliche räumliche Nachbarn eingezeichnet.

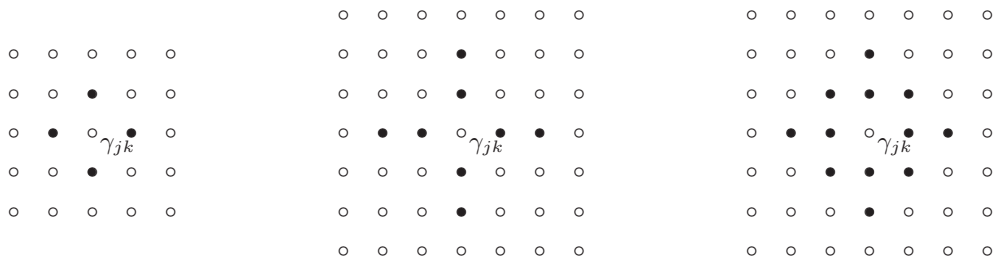

Abbildung 5.9: Mögliche räumliche Nachbarn sind als gefüllte Punkte dargestellt (aus Fahrmeir et al., 2013, S. 375).

Gehen wir zunächst auf den einfachsten Fall mit vier Nachbarn ein, so definieren wir die Penalisierung ähnlich wie im eindimensionalen Fall über die quadrierten Differenzen zwischen $\gamma_{j,k}$ und seinen vier Nachbarn (linkes Bild in Abbildung 5.9). Dazu seien $\mathbf{D}_1^a$ und $\mathbf{D}_1^t$ die zum Alter bzw. zum Beobachtungsjahr gehörenden eindimensionalen Differenzenmatrizen erster Ordnung ($c = 1$) der Dimension $(d_a - 1) \times d_a$ bzw. $(d_t - 1) \times d_t$. Zeilenweise gebildete Differenzen sind über die Anwendung der erweiterten Differenzenmatrix $\mathbf{I}_{d_t} \otimes \mathbf{D}_1^a$ auf den Parametervektor $\boldsymbol{\gamma}$

$$\boldsymbol{\gamma}^\top(\mathbf{I}_{d_t} \otimes \mathbf{D}_1^a)^\top(\mathbf{I}_{d_t} \otimes \mathbf{D}_1^a)\boldsymbol{\gamma} = \sum_{j=1}^{d_t}\sum_{i=2}^{d_a}(\gamma_{ij} - \gamma_{i-1,j})^2$$

zu erhalten und analog gilt für die spaltenweisen quadratischen Differenzen

$$\boldsymbol{\gamma}^\top(\mathbf{D}_1^t \otimes \mathbf{I}_{d_a})^\top(\mathbf{D}_1^t \otimes \mathbf{I}_{d_a})\boldsymbol{\gamma} = \sum_{i=1}^{d_a}\sum_{j=2}^{d_t}(\gamma_{ij} - \gamma_{i,j-1})^2.$$

Allgemein führt dies zu den zeilenweisen Straftermen

$$\boldsymbol{\gamma}^\top(\mathbf{I}_{d_t} \otimes \mathbf{D}_c^a)^\top(\mathbf{I}_{d_t} \otimes \mathbf{D}_c^a)\boldsymbol{\gamma} = \boldsymbol{\gamma}^\top(\mathbf{I}_{d_t}^\top \otimes \mathbf{D}_c^{a\top})(\mathbf{I}_{d_t} \otimes \mathbf{D}_c^a)\boldsymbol{\gamma} = \boldsymbol{\gamma}^\top(\mathbf{I}_{d_t}\mathbf{I}_{d_t} \otimes \mathbf{D}_c^{a\top}\mathbf{D}_c^a)\boldsymbol{\gamma},$$

und zu den spaltenweisen Penalisierungen von $\mathbf{G}$

$$\boldsymbol{\gamma}^\top(\mathbf{D}_c^t \otimes \mathbf{I}_{d_a})^\top(\mathbf{D}_c^t \otimes \mathbf{I}_{d_a})\boldsymbol{\gamma} = \boldsymbol{\gamma}^\top(\mathbf{D}_c^{t\top}\mathbf{D}_c^t \otimes \mathbf{I}_{d_a})\boldsymbol{\gamma},$$

mit der Matrix $\mathbf{D}_c^a$ von Differenzen der Ordnung c, korrespondierend zu den Zeilen von $\mathbf{G}$, und der Matrix $\mathbf{D}_c^t$, gehörend zu den Spalten von $\mathbf{G}$. Die Schätzung der unbekannten Regressionsparameter erfolgt analog wie im eindimensionalen Fall über die Maximierung der penalisierten Likelihood-Funktion

$$\begin{aligned} l^*(\boldsymbol{\gamma}|\mathbf{y}) &= l(\boldsymbol{\gamma}|\mathbf{y}) - \frac{\lambda_a}{2}\boldsymbol{\gamma}^\top(\mathbf{I}_{d_t} \otimes \mathbf{D}_c^{a\top}\mathbf{D}_c^a)\boldsymbol{\gamma} - \frac{\lambda_t}{2}\boldsymbol{\gamma}^\top(\mathbf{D}_c^{t\top}\mathbf{D}_c^t \otimes \mathbf{I}_{d_a})\boldsymbol{\gamma} \\ &= l(\boldsymbol{\gamma}|\mathbf{y}) - \frac{1}{2}\boldsymbol{\gamma}^\top(\lambda_a\mathbf{I}_{d_t} \otimes \mathbf{D}_c^{a\top}\mathbf{D}_c^a + \lambda_t\mathbf{D}_c^{t\top}\mathbf{D}_c^t \otimes \mathbf{I}_{d_a})\boldsymbol{\gamma} \\ &= l(\boldsymbol{\gamma}|\mathbf{y}) - \frac{1}{2}\boldsymbol{\gamma}^\top\mathbf{P}\boldsymbol{\gamma}, \end{aligned}$$

mit der Strafmatrix $\mathbf{P} = \lambda_a\mathbf{I}_{d_t} \otimes \mathbf{D}_c^{a\top}\mathbf{D}_c^a + \lambda_t\mathbf{D}_c^{t\top}\mathbf{D}_c^t \otimes \mathbf{I}_{d_a} \in \mathbb{R}^{d_a d_t \times d_a d_t}$ und den beiden Glättungsparametern λ_a und λ_t in Alters- und Jahresrichtung. In Analogie zum eindimensionalen Fall werden die Parameter mittels IWLS-Algorithmus (5.4) geschätzt.

Für die Auswahl der optimalen Glättungsparameter können auch im zweidimensionalen Fall die Informationskriterien AIC oder BIC herangezogen werden. Als effektive Parameterzahl dient wie bei der eindimensionalen Modellierung $\mathrm{tr}(\mathbf{H}_\lambda)$, wobei $\mathbf{H}_\lambda$ in (5.5) gegeben ist und im Zweidimensionalen $\mathbf{B} = \mathbf{B}_t \otimes \mathbf{B}_a$ gilt. Die beiden Informationskriterien sind folgendermaßen definiert:

$$\begin{aligned} \mathrm{AIC}(\lambda_a, \lambda_t) &= \mathrm{D}(\mathbf{y}, \boldsymbol{\gamma}, \lambda_a, \lambda_t) + 2\mathrm{ED}(\boldsymbol{\gamma}, \lambda_a, \lambda_t), \\ \mathrm{BIC}(\lambda_a, \lambda_t) &= \mathrm{D}(\mathbf{y}, \boldsymbol{\gamma}, \lambda_a, \lambda_t) + \log(n_a n_t)\mathrm{ED}(\boldsymbol{\gamma}, \lambda_a, \lambda_t), \end{aligned}$$

und sollen minimiert werden.

5.3.1 Das Generalisierte Lineare Array Modell

Im zweidimensionalen Modell ist die Schätzung der Regressionskoeffizienten wie im eindimensionalen Modell über (5.4) möglich. Im Falle des vorliegenden Datensatzes, der für den interessierenden Altersbereich $[45, 95]$ und einen Zeithorizont von fünfzehn Jahren von moderater Größe ist, ist das Kronecker-Produkt $\mathbf{B} = \mathbf{B}_t \otimes \mathbf{B}_a$ noch von überschaubarer Dimension. Verwenden wir wie im nachstehenden Beispiel $d_a = 13$ und $d_t = 6$ Basisfunktionen, so belauft sich die Dimension der Regressionsmatrix auf $(n_a n_t) \times (d_a d_t) = 765 \times 78$. Werden jedoch der Altersbereich und der Beobachtungszeitraum vergrößert, kann es sehr schnell zu einem Speicher- und Laufzeitproblem kommen. Eine Lösung dafür bietet der von Eilers, Currie und Durban (2006) vorgestellte Algorithmus, der die spezielle Struktur der als „Rechteck“ arrangierten Daten und der als Kronecker-Produkt definierten Regressionsmatrix ausnutzt. Das Modell wird als *Generalized Linear Array Model* (GLAM) bezeichnet und das dahinterliegende Prinzip soll an dieser Stelle kurz erläutert werden, da das R-Paket MortalitySmooth im zweidimensionalen Setting intern damit arbeitet.

Sind die Koeffizienten $\boldsymbol{\gamma}$ in Form einer Matrix, wie in (5.7), angegeben, dann lässt sich das Modell (5.6) umschreiben in

$$\boldsymbol{\eta} = \log(\mathbf{E}) + \mathbf{B}_a \mathbf{G} \mathbf{B}_t^\top,$$

womit deutlich wird, dass die Regressionsmatrix auch ohne der Verwendung des Kronecker-Produktes erzeugt werden kann. Ein ähnlicher Ansatz kommt für das ebenso zeitaufwändig zu berechnende Produkt $\mathbf{B}^\top \mathbf{W} \mathbf{B}$ zur Anwendung:

$$(\mathbf{B}_t \otimes \mathbf{B}_a)^\top \mathbf{W} (\mathbf{B}_t \otimes \mathbf{B}_a) \equiv F(\mathbf{B}_a)^\top \tilde{\mathbf{W}} F(\mathbf{B}_t),$$

mit

$$F(\mathbf{X}) = (\mathbf{X} \otimes \mathbf{1}^\top) \odot (\mathbf{1}^\top \otimes \mathbf{X}),$$

für eine $n \times d$-Matrix $\mathbf{X}$ und der $n_a \times n_t$-Gewichtsmatrix $\tilde{\mathbf{W}}$, wobei $\mathrm{vec}(\tilde{\mathbf{W}})$ den Diagonalelementen von $\mathbf{W}$ entspricht. Weiters ist $\mathbf{1}$ ein Vektor der Länge d mit Einselementen und $\odot$ bezeichnet das elementweise Produkt zweier Matrizen. Somit führt $F(\mathbf{X})$ zu einer Matrix der Dimension $n \times d^2$.

Dieser Zugang beinhaltet zwei wesentliche Vorteile, einerseits das Vermeiden der Berechnung des Kronecker-Produktes und andererseits wird die Anzahl der Multiplikationen reduziert. Somit haben wir einen Algorithmus in der Hand, der die computertechnische Umsetzung auch höher dimensionaler Modelle ermöglicht. Genauere Ausführungen dazu sind sowohl in der oben genannten Quelle, als auch in Currie, Durban und Eilers (2006) zu finden.

5.3.2 Glättung in zwei Dimensionen am Beispiel von Mortalitätsdaten

Wir wollen nun die Mortalitätsraten weiblicher Versicherter im Altersintervall $[45, 95]$ und über den Beobachtungszeitraum $[2001, 2015]$ sowohl in Richtung des Alters, als auch in Richtung des Jahres unter dem Modell (5.6) glätten. Auch für die zweidimensionale Modellierung von Mortalitätsraten beinhaltet das R-Paket MortalitySmooth die dafür passende Funktion Mort2Dsmooth. Über das Argument control sind jene Parameter, die die Schätzung der optimalen Glättungsparameter steuern, kontrollierbar. Wie auch im eindimensionalen Fall, verfeinern wir das Gitter, über dem λ gesucht wird und ändern den

Bereich, in dem der Glättungsparameter in Richtung des Jahres ermittelt wird, da die Voreinstellung RANGEy $= [10^{-4}, 10^{6}]$, wie auch in Abbildung 5.10 ersichtlich, für die vorliegende Datensituation nicht adäquat ist.

```
> sex<-"W"
> t<-seq(2001,2015,1)
> x<-seq(45,95,1)
> deaths<-data.select(t,x,sex,"Deaths")
> exposures<-data.select(t,x,sex,"Exposures")

> library("MortalitySmooth")
> model2d_bic<-Mort2Dsmooth(x=x, y=t, Z=deaths, offset=log(
+ exposures), control=list(TOL2=0.001,RANGEy=c(10^4,10^8)))
> model2d_aic<-Mort2Dsmooth(x=x, y=t, Z=deaths, offset=log(
+ exposures), control=list(TOL2=0.001,RANGEy=c(10^4,10^8)),
+ method=2)
```

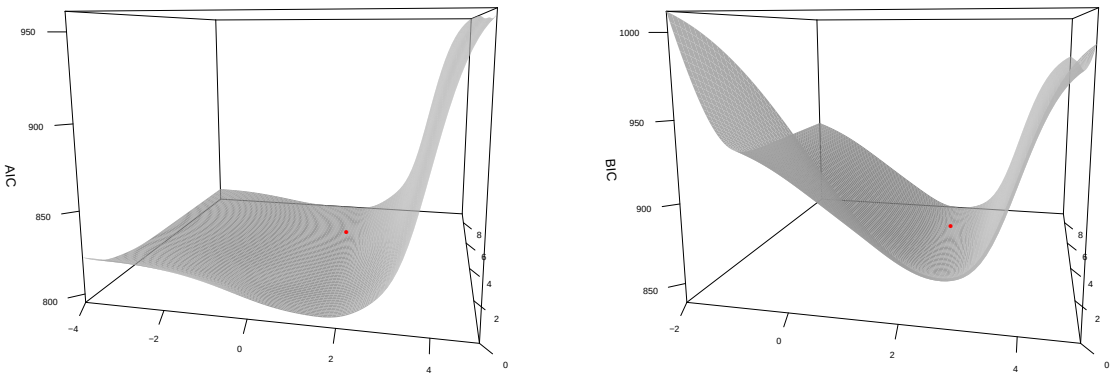

Abbildung 5.10: Oberfläche des AICs (links) und des BICs (rechts) für verschiedene Werte von λ_a und λ_t auf der log-Skala aufgetragen. Die optimalen Glättungsparameter (rot) für das AIC sind $\lambda_a = 19.36$ und $\lambda_t = 1\,902\,000$ sowie für das BIC $\lambda_a = 133.35$ und $\lambda_t = 6\,742\,200$.

In den Modellen model2d_bic und model2d_aic werden kubische B-Splines ($q = 3$), ein Strafterm zweiter Ordnung ($c = 2$) sowie $l = 11$ Knoten in Richtung des Alters (bei $n_a = 51$) und $l = 4$ Knoten in Richtung des Jahres (bei $n_t = 15$) verwendet. Somit gehen $d_a = 13$ und $d_t = 6$ Basisfunktionen ins Modell ein und es sind 78 Parameter zu schätzen, was sich auf eine effective dimension von ED $= 8.11$ nach der Wahl des Glättungsparameters mittels BIC (model2d_bic) reduziert. Dabei ist $\lambda_a = 133.35$ und $\lambda_t = 6\,742\,200$ bei einem BIC $= 837.76$ und einer Deviance von 783.92. Hinsichtlich des AICs als Optimalitätskriterium (model_aic) erhalten wir die optimalen Glättungsparameter $\lambda_a = 19.36$ und $\lambda_t = 1\,902\,000$ bei einer effective dimension von 11.27 und einer Deviance D $= 772.27$.

Anhand der Werte ist auch gut zu sehen, dass sich in der Größenordnung sehr unterschiedliche Werte für λ_a und λ_t ergeben, weshalb es durchaus Sinn macht zwei verschiedene Glättungsparameter im Modell zu haben.

Nachdem die optimalen Parameter, die den Grad der Glättung steuern, geschätzt wurden, können die geglätteten Mortalitätsoberflächen dargestellt werden (siehe Abbildung 5.11). Mit der Annahme poissonverteilter Sterbefälle, liefert das P-Spline Modell eine sichtlich glattere Schätzung im Vergleich zu den Mortalitätsraten im Lee-Carter Modell. Deutlich zu sehen ist auch der Rückgang der Sterblichkeit im betrachteten Zeitraum von fünfzehn Jahren (siehe Kontur-Diagramme in Abbildung 5.11).

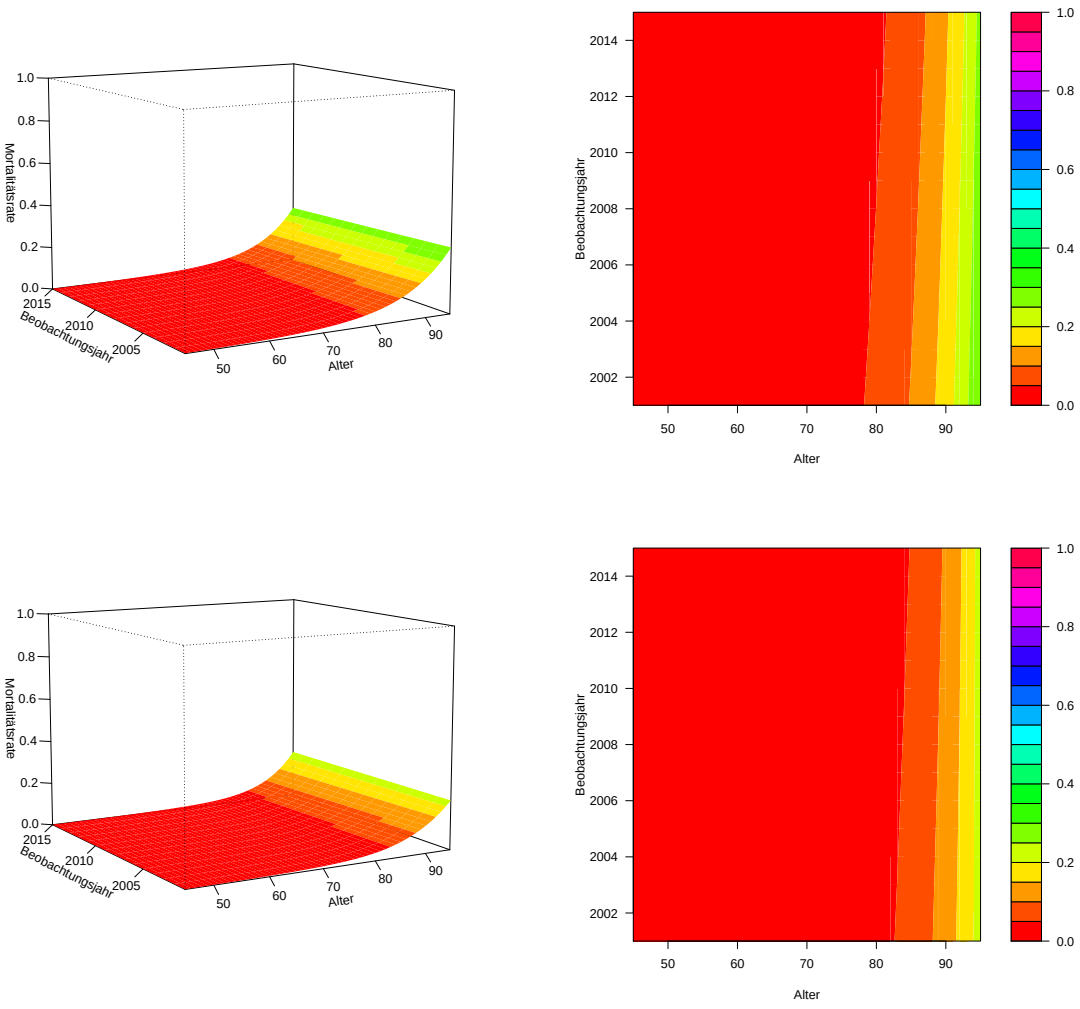

Abbildung 5.11: Unter dem bivariaten P-Spline Modell berechnete Mortalitätsraten für Männer (oben) und Frauen (unten) im Zeitraum $[2001, 2015]$ und im Altersbereich $[45, 95]$.

5.4 Modelldiagnose

Ein wesentlicher Vorteil von P-Spline Modellen ist, dass das Regelwerk der GLMe auch dafür zugänglich ist (weitere Ausführungen zu den Vorteilen von P-Splines sind im Abschnitt 5.6 zu finden). Insbesonders für die Modelldiagnose liefern Residuen Auskunft über die Erfüllung der Modellannahmen und über die Modellanpassung. Zur Analyse von Mortalitätsmodellen gehen wir wie zuvor von poissonverteilten Responses aus, weshalb wir für die im Abschnitt 3.2.3 definierten Pearson Residuen

$$r_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}$$

erhalten sowie für die Deviance Residuen

$$r_i^D = \text{sign}(y_i - \hat{\mu}_i)\sqrt{d_i},$$

wobei d_i im Falle der Poissonverteilung durch $d_i = \left(2\left(y_i \log \frac{y_i}{\hat{\mu}_i} - (y_i - \hat{\mu}_i)\right)\right)$ gegeben ist.

Für das im vorangegangenen Kapitel besprochene Lee-Carter Modell unter der Annahme einer Poissonverteilung können die Deviance Residuen ähnlich wie oben in der Form

$$r_{x,t}^D = \text{sign}(\mathbf{D}_{x,t} - \hat{\mathbf{D}}_{x,t}) \left(2\left(\mathbf{D}_{x,t} \log \frac{\mathbf{D}_{x,t}}{\hat{\mathbf{D}}_{x,t}} - (\mathbf{D}_{x,t} - \hat{\mathbf{D}}_{x,t})\right)\right)^{\frac{1}{2}}$$

dargestellt werden (vgl. Liu & Braun, 2010).

Residuen sind auch für den Vergleich zweier unterschiedlicher Modelle geeignet. Dazu sind in der Grafik 5.12 die Deviance Residuen in Abhängigkeit der beiden Prädiktoren, Alter und Beobachtungsjahr, einmal für das Lee-Carter Modell und einmal für das bivariate P-Spline Modell dargestellt. Wiederum stellt das R-Paket MortalitySmooth eine passende Funktion zur Berechnung der Residuen zur Verfügung:

```
1 > model_fem<-Mort2Dsmooth(x=x, y=t, Z=deaths,
2 + offset=log(exposures), lambdas=c(133.35,6742200), method=3)
3 > res_dev<-residuals(model_fem, type="deviance")
```

Sowohl die Residuen im Lee-Carter Modell, als auch jene im P-Spline Modell befinden sich größtenteils um den Wert Null und weisen ähnliche systematische Muster auf. Auf Grund der Vorteile des P-Spline Ansatzes, wie die glattere Schätzung, die geringere Anzahl zu schätzender Parameter und die Fähigkeit die langjährige Mortalitätsentwicklung abzubilden, stellt das zweidimensionale P-Spline Modell hinsichtlich der Anwendung auf Mortalitätsdaten einen robusten Modellierungsansatz dar.

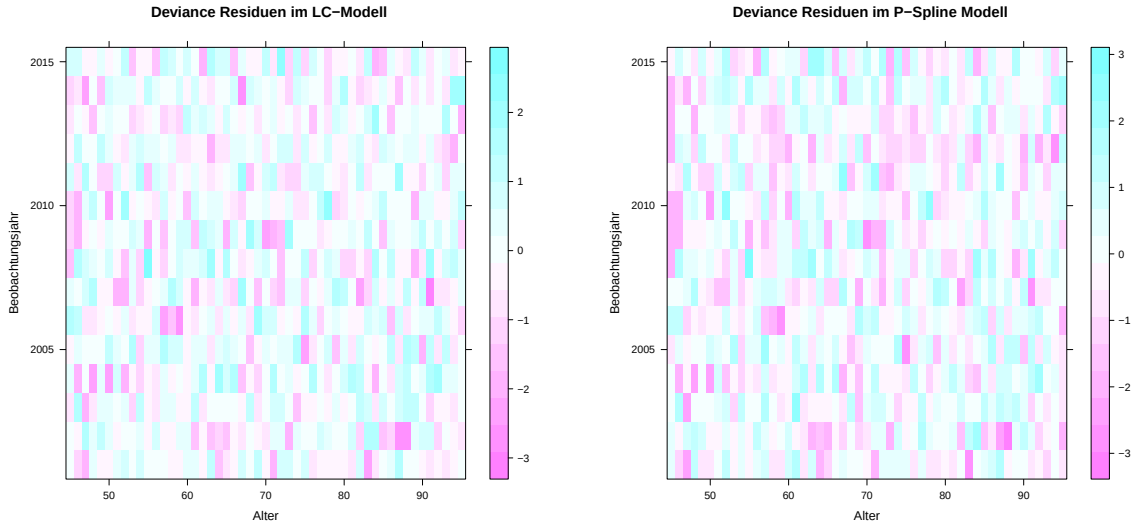

Abbildung 5.12: Deviance Residuen für das Lee-Carter Modell (links) und das P-Spline Modell (rechts) unter Verwendung von Daten weiblicher Versicherter im Alter $[45, 95]$ über die gesamte Historie, also im Bereich $[2001, 2015]$.

5.5 Prognose von Mortalitätsraten im P-Spline Modell

Wie bisher besprochen, sind P-Spline Modelle sehr gut für die Modellierung von Mortalitätsraten geeignet. Kurz zusammengefasst, wird für die Modellierung zunächst eine B-Spline Basis konstruiert und diese mit einem Strafterm versehen, mit dem Ziel eine möglichst glatte Oberfläche zu erhalten. Die Schätzung der Regressionsparameter erfolgt unter Verwendung des leicht modifizierten Fisher-Scoring-Algorithmus durch Maximierung der penalisierten Log-Likelihood-Funktion. In diesem Abschnitt wird zu sehen sein, dass die Schätzung der Regressionsparameter sowie die Prognose der Mortalitätsraten simultan erfolgen kann (siehe Currie, Durban & Eilers, 2004 und Currie, Durban & Eilers, 2003).

Es besteht ein wesentlicher Unterschied zwischen der Prognose mit penalisierten Splines und jener, die im Zusammenhang des Lee-Carter Modells besprochen wurde und Methoden der Zeitreihenanalyse verwendet. Im Gegensatz zum P-Spline Modell ist dabei eine simultane Schätzung und Prognose nicht möglich. Folglich sind bei der Prognose mit Zeitreihenmethoden zwei Stufen notwendig, die im Kapitel 4.1 beschrieben wurden.

5.5.1 Prognose im univariaten Modell

Ziel ist es, unter dem verwendeten Modell eine Prognose der Mortalität durchzuführen. Dafür wird ein bestimmtes Alter x festgehalten und die Mortalität über den Beobachtungszeitraum betrachtet sowie anschließend in die Zukunft extrapoliert. Im eindimensionalen Fall sind die Daten wie zuvor in Vektorform gegeben und umfassen einerseits den n_t-dimensionalen Vektor der Sterbefälle $\mathbf{y}$ und den ebenfalls n_t-dimensionalen Vektor der Exposures $\mathbf{e}$. Des Weiteren definieren wir eine B-Spline Regressionsmatrix $\mathbf{B}_{t_1}$. Da die Extrapolation für n_{t_p} Jahre in die Zukunft das zentrale Thema dieses Abschnitts bildet, muss die zuvor erzeugte Regressionsmatrix erweitert werden. Dazu erzeugen wir die Regressionsmatrix $\mathbf{B}$ für insgesamt $n_t + n_{t_p}$ Jahre, indem die der Matrix $\mathbf{B}_{t_1}$ zugrunde liegende

Knotenmenge vergrößert wird und auch jenen Bereich der in der Zukunft liegenden Jahre umfasst.

Als nächsten Schritt führen wir fiktive Werte für die Anzahl der Sterbefälle und Exposures, die zu den Jahren im Intervall $[t_{n_{t+1}}, t_{n_t+n_{t_p}}]$ gehören, ein und erweitern damit $\mathbf{y}$ und $\mathbf{e}$. Um zu gewährleisten, dass die soeben einbezogenen Dummy Variablen nicht zur Parameterschätzung beitragen, muss noch zusätzlich ein Gewichtsvektor definiert werden, der den beobachteten Daten das Gewicht Eins zuweist und die fiktiven Beobachtungen mit Nullen gewichtet.

Nach obiger Aufbereitung der Daten kann schließlich die Maximierung der penalisierten Log-Likelihood-Funktion erfolgen, die im Abschnitt 5.2 beschrieben ist. Sei dazu $\mathbf{V} = \text{blockdiag}(\mathbf{I}_{n_t}, \mathbf{0})$ die Matrix der Gewichte, wobei $\mathbf{0}$ eine $n_{t_p} \times n_{t_p}$-Matrix mit Nulleinträgen ist. Dann kann der penalisierte Scoring-Algorithmus gemäß (5.4) durch das Gewichten der Regressionsmatrix $\mathbf{B}$ in der Form

$$\left(\mathbf{B}^\top \mathbf{V}\mathbf{W}^{(t)}\mathbf{B} + \mathbf{P}\right)\boldsymbol{\beta}^{(t+1)} = \mathbf{B}^\top \mathbf{V}\mathbf{W}^{(t)}\mathbf{B}\boldsymbol{\beta}^{(t)} + \mathbf{B}^\top \mathbf{V}(\mathbf{y} - \boldsymbol{\mu}^{(t)})$$

geschrieben werden. Wie anhand dieses Ausdrucks ersichtlich ist, steuert der Strafterm die Prognose und beeinflusst deren Struktur.

Mit diesem Ansatz wird sichergestellt, dass die geschätzten Parameter, die Deviance, der optimale Glättungsparameter und die fitted values sowohl unter Einbeziehung der fehlenden Daten für den zu prognostizierenden Bereich, als auch nur unter Berücksichtigung der beobachteten Daten dieselben sind. Diese Invarianzeigenschaft ist somit eine Parallele zur Prognose mit Methoden aus der Zeitreihenanalyse, wo diese durch den zweistufigen Aufbau gewährleistet ist.

Für die praktische Umsetzung kann wiederum das in R verfügbare Paket MortalitySmooth verwendet werden. Es stellt die Funktion predict.Mort1Dsmooth zur Verfügung, die ein bereits mit MortalitySmooth gefittetes Modell als Argument nimmt und des Weiteren der Bereich, der Gegenstand der Prognose sein soll, zu definieren ist. Alternativ können aber wie oben beschrieben, die Modellschätzung und die Vorhersage in einem Schritt erfolgen, wobei daraus dieselben Ergebnisse resultieren (vgl. nachstehendes R-Output). Im Hinblick auf die vorliegenden Daten fixieren wir für demonstrative Zwecke das Alter 77 und modellieren den Zeitraum $[2001, 2015]$ für weibliche Versicherte. Die Prognose soll für den Bereich $[2016, 2020]$ erstellt werden, also beträgt $n_{t_p} = 5$.

```
> sex<-"W"
> t<-seq(2001,2015)
> a<-77
> deaths<-data.select(t,a,sex,"Deaths")
> exposures<-data.select(t,a,sex,"Exposures")

> model_t<-Mort1Dsmooth(x=t, y=deaths, offset=log(exposures),
+ control=list(TOL2=0.001, RANGE=c(10^4,10^8))))
> t_erw<-t[1]:2020
> prog<-predict.Mort1Dsmooth(model_t,newdata=t_erw)
> t(prog)
```

```
 -3.573463 -3.615582 -3.6577 -3.699819 -3.741938 -3.784056
 -3.826175 -3.868293 -3.910412 -3.95253 -3.994649 -4.036767
 -4.078886 -4.121004 -4.163123 -4.205241 -4.247359 -4.289478
 -4.331596 -4.373715
```

```
> deaths_erw<-c(deaths,rep(2,5))
> exposures_erw<-c(exposures,rep(4,5))
> weights<-c(rep(1,length(deaths)),rep(0,5))

> model<-Mort1Dsmooth(x=t_erw, y=deaths_erw,
+ offset=log(exposures_erw),w=weights,control=list(TOL2=0.001,
+ RANGE=c(10^4,10^8)))
> t(model$logmortality)
 -3.573463 -3.615582 -3.6577 -3.699819 -3.741938 -3.784056
 -3.826175 -3.868293 -3.910412 -3.95253 -3.994649 -4.036767
 -4.078886 -4.121004 -4.163123 -4.205241 -4.247359 -4.289478
 -4.331596 -4.373715
```

In Abbildung 5.13 wird verdeutlicht, dass für jenen Bereich, wo Beobachtungen vorliegen die Ordnung des Strafterms c eine untergeordnete Rolle spielt. Im Vorhersagebereich hingegen gibt es in Abhängigkeit von c einen sehr deutlichen Unterschied in den prognostizierten Werten. Bei der Vorhersage mit P-Splines werden die Regressionskoeffizienten extrapoliert. Je nach Ordnung des Strafterms erfolgt diese Extrapolation konstant ($c = 1$), linear ($c = 2$) oder quadratisch ($c = 3$). Da weder eine konstante, noch eine quadratische Extrapolation sinnvoll erscheint, werden wir in weiterer Folge mit $c = 2$ arbeiten.

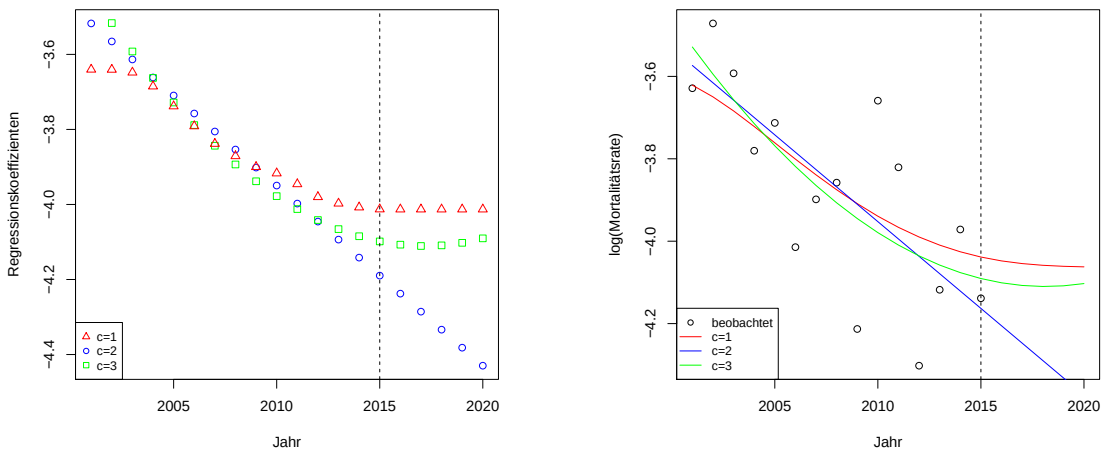

Abbildung 5.13: Geglättete Mortalitätsraten von weiblichen Versicherten im Alter von 77 Jahren über den Zeitraum $[2001, 2015]$ mit einer Prognose für $[2016, 2020]$ unter Verwendung verschiedener Werte für die Ordnung des Strafterms (rechts) und Abbildung der Regressionskoeffizienten (links - zur besseren Veranschaulichung wurde hier die Anzahl der Basisfunktionen gleich der Anzahl der Jahre gewählt).

Außerdem sei an dieser Stelle angemerkt, dass eine weit in die Zukunft reichende Prognose nicht sinnvoll ist, da diese mit großer Unsicherheit behaftet ist. In Grafik 5.14 ist zusätzlich zu den geglätteten und prognostizierten Mortalitätsraten das 95%-Konfidenzintervall abgebildet, das mit zunehmenden Jahren in der Zukunft deutlich größer wird. Aus diesem Grund wird lediglich die Prognose für ein Jahr im Voraus als zuverlässig angesehen. Ausschließlich für demonstrative Zwecke wird hier $n_{t_p} = 5$ gewählt.

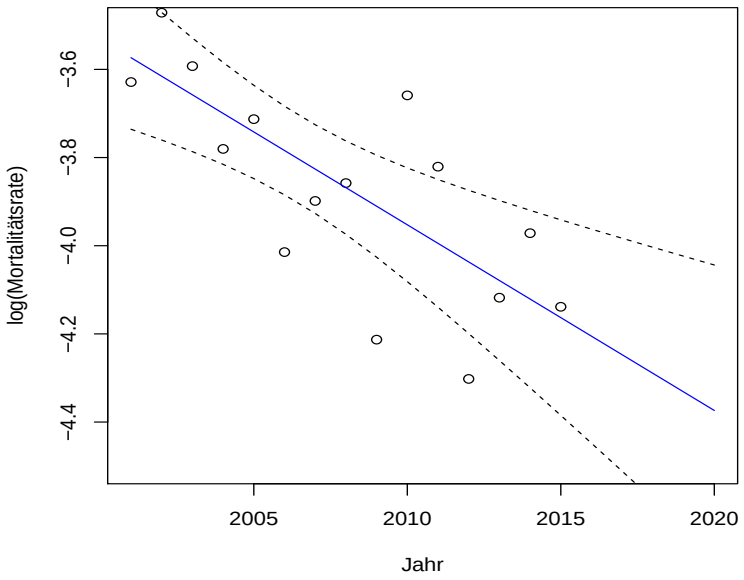

Abbildung 5.14: Geglättete Mortalitätsraten weiblicher Versicherter im Alter von 77 Jahren über den Zeitraum $[2001, 2015]$ mit einer Prognose für $[2016, 2020]$ und dem 95%-Konfidenzintervall.

5.5.2 Prognose im bivariaten Modell

Die vorangehenden Ausführungen widmen sich der Vorhersage von Mortalitätsraten im eindimensionalen Fall, das heißt ein Alter wird fixiert und die Sterblichkeit in Abhängigkeit von der Zeit modelliert und prognostiziert. Es könnte angenommen werden, dass wenn diese vorgestellte eindimensionale Prognose-Methodik auf alle verschiedenen Alter angewandt wird, die Vorhersage für alle Alter simultan möglich ist. Das ist jedoch nicht richtig, da mit dieser Vorhersagemethode im Hinblick auf Mortalitätsoberflächen negative Aspekte einhergehen. In Abbildung 5.15 sind im linken Bild für ausgewählte Alter die Prognosen bis zum Jahr 2030 unter dem eindimensionalen P-Spline Modell dargestellt. Dabei fällt auf, dass sich die prognostizierten Mortalitätsraten für das Alter 65 und 70 überschneiden, was nicht gewünscht ist. Grund dafür ist, dass es keinen Parameter zur Kontrolle der Vorhersage über die verschiedenen Alter hinweg gibt.

Dem kann durch den Übergang zum zweidimensionalen P-Spline Modell entgegengewirkt werden. Wie im eindimensionalen Fall werden auch hier fiktive Daten für die Anzahl der

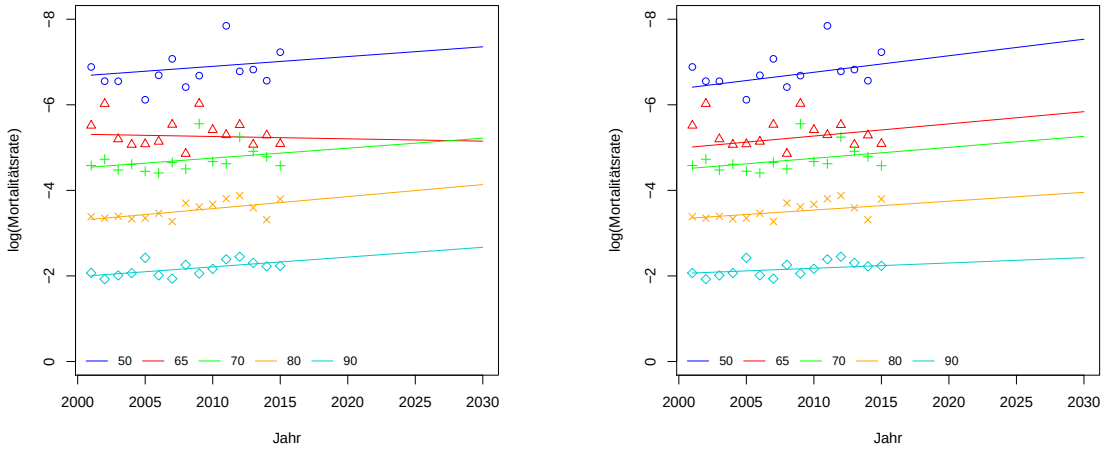

Abbildung 5.15: Beobachtete und unter Verwendung des eindimensionalen P-Spline Modells geglättete Mortalitätsraten bzw. deren Prognose bis zum Jahr 2030 (links) und unter dem zweidimensionalen Modell geglättete und vorhergesagte Mortalitätsraten bis zum Jahr 2030 (rechts).

Sterbefälle und der dem Risiko des Todes unterliegenden Personen im Bereich der Vorhersage konstruiert. Außerdem werden wiederum Gewichte, die den tatsächlich beobachteten Daten ein Gewicht von Eins zuweisen und die künstlichen Datenpunkte mit Nullen gewichten, eingeführt. Betten wir dies in das zweidimensionale P-Spline Modell ein, so definieren wir wieder eine um den zu prognostizierenden Jahresbereich erweiterte B-Spline Basismatrix $\mathbf{B}_t \in \mathbb{R}^{(n_t+n_{t_p})\times d_t}$ und wie schon zuvor bilden wir anschließend das Kronecker-Produkt $\mathbf{B} = \mathbf{B}_t \otimes \mathbf{B}_a$, um die Regressionsmatrix zu erhalten. Des Weiteren definieren wir die Matrix der Gewichte $\mathbf{V} = \text{blockdiag}(\mathbf{I}_{n_a(n_t+n_{t_p})}, \mathbf{0})$, wobei $\mathbf{0}$ eine $n_{t_p} \times n_{t_p}$-Matrix mit Nulleinträgen ist und beziehen sie folgendermaßen in den penalisierten Scoring Algorithmus ein:

$$\left(\mathbf{B}^\top \mathbf{V}\mathbf{W}^{(t)}\mathbf{B} + \mathbf{P}\right)\boldsymbol{\beta}^{(t+1)} = \mathbf{B}^\top \mathbf{V}\mathbf{W}^{(t)}\mathbf{B}\boldsymbol{\beta}^{(t)} + \mathbf{B}^\top \mathbf{V}(\mathbf{y} - \boldsymbol{\mu}^{(t)}).$$

Wie auch im eindimensionalen Fall, kann die Vorhersage unter Verwendung des R-Paketes MortalitySmooth und der für die zweidimensionale Prognose geeigneten Funktion predict.Mort2Dsmooth erfolgen.

```
> sex<-"W"
> t<-seq(2001,2015)
> a<-seq(45,95,1)
> deaths<-data.select(t,a,sex,"Deaths")
> exposures<-data.select(t,a,sex,"Exposures")

> t_erw<-t[1]:2020
> newdata<-list(x=a,y=t_erw)
> model2D_2<-Mort2Dsmooth(x=a, y=t, Z=deaths, offset=
+ log(exposures),pord=c(2,2),control=list(TOL2=0.001,
+ RANGEy=c(10^4,10^8)))
> prog2D_2<-predict.Mort2Dsmooth(model2D_2,newdata=newdata)
```

In Abbildung 5.15 sind in der rechten Grafik die vorhergesagten Mortalitätsraten unter dem zweidimensionalen P-Spline Modell illustriert. Es wurde deutliche Abhilfe für die sich überschneidenden Prognosen im Alter 65 und 70 geschaffen. Die Altersstruktur bleibt also erhalten und die einzelnen Mortalitätsverläufe sind auch im Prognosebereich getrennt. Wie im eindimensionalen Modell, wirkt sich die gewählte Ordnung c stark auf die Prognose aus und es wurden wiederum Strafterme zweiter Ordnung verwendet. Die Invarianzeigenschaften wie sie im eindimensionalen Fall genannt wurden halten hier jedoch nicht mehr.

In der Grafik 5.16 sind die Mortalitätsoberflächen für weibliche und männliche Versicherte, unter Einbeziehung einer Vorhersage über einen fünfjährigen Zeithorizont, dargestellt. Auch hier gilt, dass weit in der Zukunft liegende Prognosen unzuverlässig sind und wir daher nur Prognosen für das Folgejahr als aussagekräftig ansehen. Vor allem in den Kontur-Diagrammen ist der Rückgang der Mortalität deutlich, der sich in der Zukunft fortsetzt.

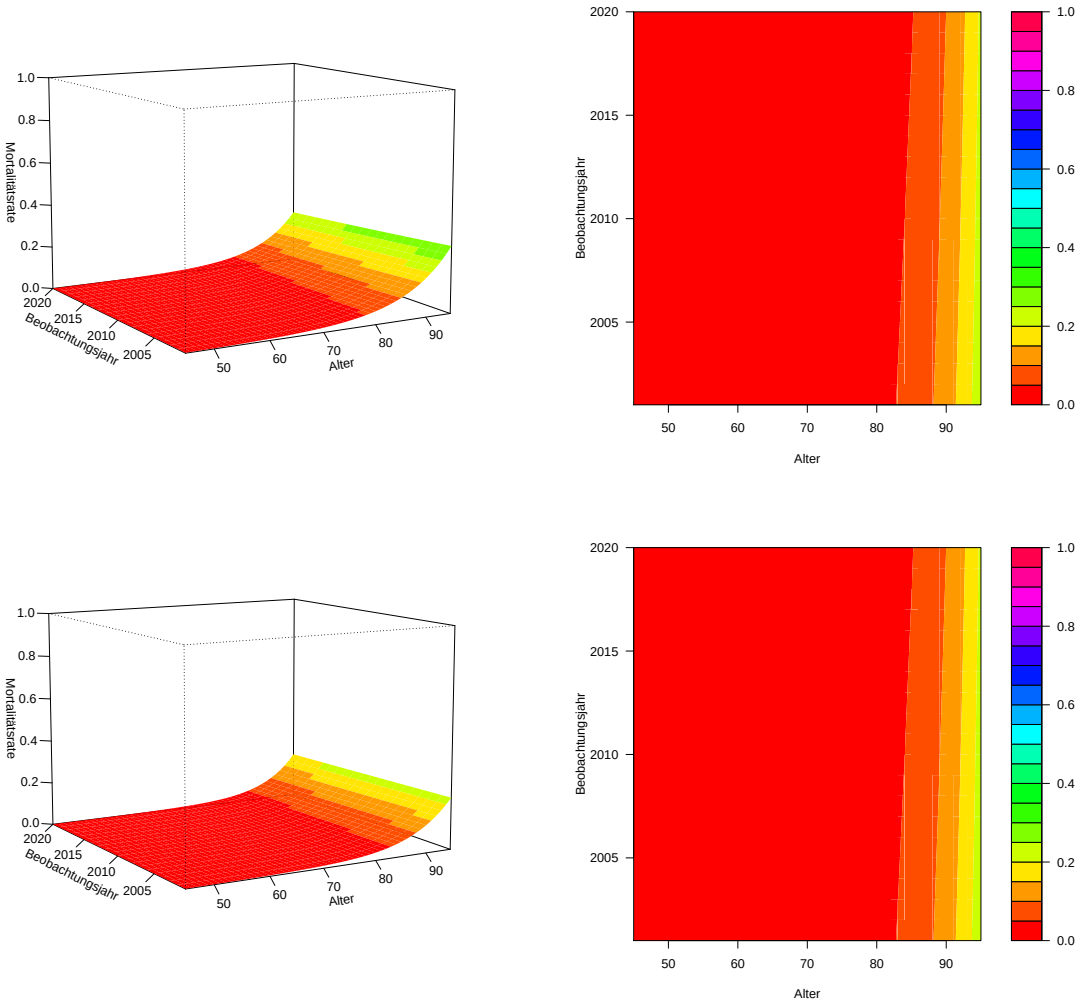

Abbildung 5.16: Geglättete Mortalitätsraten männlicher (oben) und weiblicher Versicherter im Altersintervall $[45, 95]$ Jahren über den Zeitraum $[2001, 2015]$ mit einer Prognose für $[2016, 2020]$.

5.6 Vorteile von P-Spline Modellen

Die vorangegangenen Abschnitte haben sich nichtparametrischen Modellen mit einer Penalisierung gewidmet, die vor allem in der Analyse von Mortalitätsdaten weit verbreitet sind. Die Gründe dafür sind vielfältig und sollen im abschließenden Teil des Kapitels kurz erörtert werden. Eine umfassendere Ausarbeitung zu dieser Thematik ist in Marx und Eilers (1998) zu finden.

Wie schon zu Beginn dieses Kapitels besprochen, liegt ein klarer Vorteil von nichtparametrischen Regressionsmodellen gegenüber anderen Modellklassen in der flexiblen Modellierungsmöglichkeit, die keinen starren funktionalen Zusammenhang der Response und der Prädiktoren vorschreibt. Generalisierte Additive Modelle (GAMe), deren Prädiktor von glatten Funktionen der erklärenden Variablen gebildet werden, jedoch keinen Strafterm beinhalten, bieten eine Verallgemeinerung der GLMe durch die Relaxierung der im GLM vorgegebenen linearen Form des Prädiktors (das einfachste GAM ist durch (5.1) gegeben).

Für die Modellierung der Mortalitätsraten wurde jedoch im Anschluss eine Penalisierung eingeführt. Die Idee ist, eine große Anzahl gleichverteilter Knoten und damit viele Basisfunktionen zu verwenden, wobei der Strafterm der Regulierung dient und overfitting vermieden wird. Außerdem spielt damit die Platzierung und Anzahl der Knoten eine untergeordnete Rolle und machen eine Optimierung der Knotenzahl überflüssig (siehe Eilers, Marx & Durban, 2015).

Im Gegensatz zu GAMen, die über den local scoring Algorithmus, der einen backfitting Schritt beinhaltet, geschätzt werden, wird das Backfitting im P-Spline Modell redundant. Stattdessen kommt, wie oben gezeigt, die iterative Methode des penalisierten GLM-Fisher-Scoring Algorithmus zur Anwendung.

Außerdem ist mit der Ähnlichkeit von P-Spline Modellen zu GLMen auch die Modelldiagnose auf die Methoden des GLM-Regelwerks zurückzuführen, womit der Modellvergleich und die Modellanalyse erleichtert werden. Ein weiterer Vorteil ist, dass die Parameterzahl im P-Spline Modell gegenüber dem GAM relativ gering ist, was eine Prognose, wie sie im vorangegangenen Abschnitt besprochen worden ist, einfacher macht.

Kapitel 6

Zusammenfassung der Ergebnisse

In den vorangegangenen Kapiteln werden unterschiedlichste Modelle zur Modellierung von Mortalitätsraten in einer Dimension, wie auch in zwei Dimensionen diskutiert. Es sollen nun die zwei relevantesten Modelle, nämlich das Lee-Carter Modell und das P-Spline Modell nochmals gegenübergestellt und die Ergebnisse für den vorliegenden Datensatz zusammengefasst werden. Da es sich bei den der Arbeit zugrunde liegenden Daten um Bestandsdaten einer Versicherung handelt, ist es noch von Interesse die modellierten Mortalitätsraten der österreichweiten, von der Statistik Austria veröffentlichten, Sterblichkeit gegenüberzustellen.

6.1 Vergleich des Lee-Carter Modells mit dem P-Spline Modell

In diesem Abschnitt sollen die Ergebnisse der beiden relevanten Modelle für die Modellierung in zwei Dimensionen, das Lee-Carter Modell und das P-Spline Modell, zusammengefasst und diskutiert werden.

Das Lee-Carter Modell sowie das zweidimensionale P-Spline Modell finden beide in der Modellierung und Vorhersage von Mortalitätsraten Verwendung, mit jedoch zwei konträren Zugängen. Während im Lee-Carter Modell die Sterblichkeit als eine Zeitreihe gesehen wird, wird sie im P-Spline Modell als glatte Oberfläche betrachtet. Auch hinsichtlich der Struktur handelt es sich um zwei vollkommen verschiedene Modellierungsansätze. Das Lee-Carter gibt eine funktionale Struktur für die Mortalitätsraten an, wohingegen im P-Spline Modell eine weitaus flexiblere Modellierung zugelassen wird. Hinsichtlich der Prognose von Mortalitätsraten sind außerdem Unterschiede vorhanden. Im Lee-Carter Modell werden zunächst die Modellparameter auf Grundlage der Daten geschätzt und in einem zweiten Schritt erfolgt die Extrapolation mit einem ARIMA-Modell. Demgegenüber wird bei P-Splines die Schätzung und Projektion in einem Schritt ausgeführt.

Für den Modellvergleich werden wir im Folgenden die beiden Modelle auf die der Arbeit zugrunde gelegten Daten anwenden. Tabelle 6.1 gibt einen Überblick über den vorherrschenden Umfang des Datensatzes, der gleichermaßen für Männer und Frauen gegeben ist. Außerdem ist das Setting des P-Spline Modells zu spezifizieren, das in Tabelle 6.2 aufgelistet ist.

Prädiktor	**Intervall**	**Beobachtungen**
Alter	[45,95]	51
Jahr	[2001,2015]	15
Gesamt		765

Tabelle 6.1: Datengrundlage für die Modellierung der Sterblichkeit.

Input	**Alter**	**Jahr**
Ordnung (c)	2	2
Grad (q)	3	3
Basisfunktionen (d_a, d_t)	13	6

Tabelle 6.2: Input für das P-Spline Modell.

Die Ergebnisse für männliche bzw. weibliche Versicherte unter Verwendung des P-Spline und des Lee-Carter Modells (beide Modelle unter Annahme einer Poissonverteilung) sind in Tabelle 6.3 bzw. 6.4 zusammengefasst. Die beiden Glättungsparameter ergeben sich unter Verwendung des BICs als Auswahlkriterium.

	P-Spline	**Lee-Carter**
Glättungsparameter	$\lambda_a = 1\,096.5$	
	$\lambda_t = 6\,742\,175$	
ED	5.67	115
Deviance	851.96	765.67
BIC	889.62	1 499.26
AIC	863.3	965.67

Tabelle 6.3: Ergebnisse des P-Spline und Lee-Carter Modells für männliche Versicherte.

Anhand der tabellarisch dargestellten Resultate ist zu erkennen, dass das P-Spline Modell hinsichtlich der Anzahl der Parameter klar im Vorteil liegt. Wie schon zuvor diskutiert, müssen im Lee-Carter Modell $2n_a + n_t - 2$ Parameter geschätzt werden, was zu einer unverhältnismäßig großen Anzahl an Parametern führt. Dadurch ist auch zu erklären, dass sowohl für weibliche, als auch männliche Versicherte das BIC sowie das AIC im P-Spline Modell kleiner sind als im Lee-Carter Modell. Durch die Penalisierung wird auch sichergestellt, dass das P-Spline Modell zu einer merklich glatteren Mortalitätsoberfläche

	P-Spline	**Lee-Carter**
Glättungsparameter	$\lambda_a = 133.35$	
	$\lambda_t = 6\,742\,200$	
ED	8.11	115
Deviance	783.92	636.66
BIC	837.76	1 400.25
AIC	800.14	866.66

Tabelle 6.4: Ergebnisse des P-Spline und Lee-Carter Modells für weibliche Versicherte.

führt. Wird jedoch die Deviance als Vergleichskriterium herangezogen, so ist diese im Falle des Lee-Carter Modells geringer. Dies ist darauf zurückzuführen, dass wie erwähnt das P-Spline Modell eine glattere Schätzung liefert und nur wenige bis gar keine Schwankungen im Modell enthalten sind.

6.2 Vergleich mit österreichweiten Mortalitätsdaten

Abschließend steht in diesem Abschnitt die Gegenüberstellung der bestandsspezifischen Mortalität einer Privatversicherung und jener, die für die österreichische Gesamtbevölkerung über denselben Zeitraum beobachtet wurde, im Mittelpunkt. Letztere Daten stammen aus den jährlichen Sterbetafeln, die von der Statistik Austria veröffentlicht werden.

Im Unterschied zu den allgemeinen und ausgeglichenen Sterbetafeln, die ebenfalls von der Statistik Austria um die Volkszählungsjahre herausgegeben werden, handelt es sich bei den jährlichen Sterbetafeln um rohe, nicht geglättete Werte, die Zufallsschwankungen unterliegen können. Diese Sterbetafeln werden, im Gegensatz zu den allgemeinen Sterbetafeln, nicht für versicherungsmathematische oder juridische Zwecke herangezogen. Für das Ziel, die gesamtösterreichische Mortalitätsentwicklung mit jener eines Versicherungsbestandes zu vergleichen, ist diese dennoch geeignet, da sie jährliche Beobachtungen über den gesamten Beobachtungszeitraum $[2001, 2015]$ widerspiegelt. Die für die nachfolgende Analyse verwendeten Daten stammen von der Statistik Austria (2017) und haben die in Tabelle 6.5 angegebene Form. In der Darstellung der Tabelle wurden für diese Arbeit irrelevante Daten entfernt. Des Weiteren umfasst die jährliche Sterbetafel das Altersintervall $[0, 99]$ getrennt für Männer, Frauen und Unisex. An dieser Stelle sind zur Illustration jedoch nur die ersten Zeilen der Sterbetafel für Männer abgebildet.

Die jährliche Sterbetafel enthält, nach dem Geschlecht getrennt, Auskunft über die Sterbewahrscheinlichkeit q_x im jeweiligen Jahr und im Alter x, die Anzahl der Überlebenden l_x, die Anzahl der Verstorbenen d_x im Altersintervall $[x, x + 1]$ und die fernere Lebenserwartung e_x. Dabei wird zum Alter $x = 0$ von $l_0 = 100\,000$ neugeborenen Personen ausgegangen und l_x weist dann die Anzahl der noch Lebenden aus. Die Sterbewahrscheinlichkeit errechnet sich aus dem Quotienten der Verstorbenen im Intervall $[x, x+1]$ und der Anzahl der dem Sterberisiko ausgesetzten Personen, die 100 000 beträgt.

$\mathbf{x}$	$\mathbf{q_x}$	$\mathbf{l_x}$	$\mathbf{d_x}$	$\mathbf{e_x}$
0	0,00330	100 000	330	78,63
1	0,00021	99 670	21	77,89
2	0,00021	99 649	21	76,91
3	0,00007	99 627	7	75,92
4	0,00014	99 620	14	74,93
5	0,00012	99 606	12	73,94
6	0,00007	99 594	7	72,95

Tabelle 6.5: Von der Statistik Austria veröffentlichte jährliche Sterbetafel für Männer im Jahr 2015.

Für den Vergleich der österreichweiten und der bestandsspezifischen Sterblichkeit werden die in der jährlichen Sterbetafel ausgewiesenen Sterbewahrscheinlichkeiten q_x mit der Anzahl der dem Sterberisiko unterliegenden Personen (also mit 100 000) multipliziert, um die Anzahl der Sterbefälle im jeweiligen Jahr und nach Geschlecht zu erhalten. Die Exposures sind wie erwähnt mit 100 000 gegeben. Damit sind nun auch die Daten aus der österreichweiten Sterbetafel derart aufbereitet, dass sie in das P-Spline Modell eingespeist und geglättet werden können.

In Abbildung 6.1 wird die Entwicklung der Mortalität des Versicherungsbestandes dem mittels P-Splines geglätteten Sterblichkeitsverlauf der österreichischen Bevölkerung für ausgewählte Alter gegenübergestellt. Es ist deutlich zu erkennen, dass die Sterblichkeit im Kollektiv der Privatversicherung zwar denselben Verlauf wie jene, die auf ganz Österreich bezogen ist, aufweist, jedoch klar unter der Mortalität von Gesamtösterreich liegt.

Um zu verdeutlichen, dass diese Aussage nicht nur für die in den Grafiken 6.1 ausgewiesenen Alter zutrifft, ist in Abbildung 6.2 sowohl für Frauen, als auch Männer die Differenz zwischen der geglätteten Mortalitätsrate von Österreich und des Versicherungsbestandes in Form eines Levelplots dargestellt. Es ist ersichtlich, dass im unteren Altersbereich noch wenig Unterschied gegeben ist, während im oberen Alterssegment die Differenz stark wächst um im Alter von 95 Jahren wieder etwas zu sinken. Diese Beobachtungen sind vor allem bei Männern besonders ausgeprägt zu sehen. Bei Frauen sind die Differenzen in den Mortalitätsraten im Vergleich geringer, aber dennoch ist derselbe Verlauf, wenn auch nicht so stark, erkennbar.

Der angestellte Vergleich lässt somit die Folgerung zu, dass sich die Mortalität einer Privatversicherung deutlich von der österreichweiten Sterblichkeit unterscheidet. Dabei ist nicht der Verlauf der Mortalität über den Beobachtungszeitraum signifikant verschieden, sondern vielmehr die Höhe der Sterberate, die sich für die Daten der Privatversicherung entlang der y-Achse nach unten verschiebt. Durch diese Illustration soll nochmals betont werden, wel-

che Wichtigkeit der spezifischen und regelmäßigen Mortalitätsanalyse des Bestandes einer Versicherung zukommt, da einerseits ein wesentlicher Unterschied zur Gesamtbevölkerung besteht und andererseits die Sterblichkeit einem ständigen Wandel unterliegt.

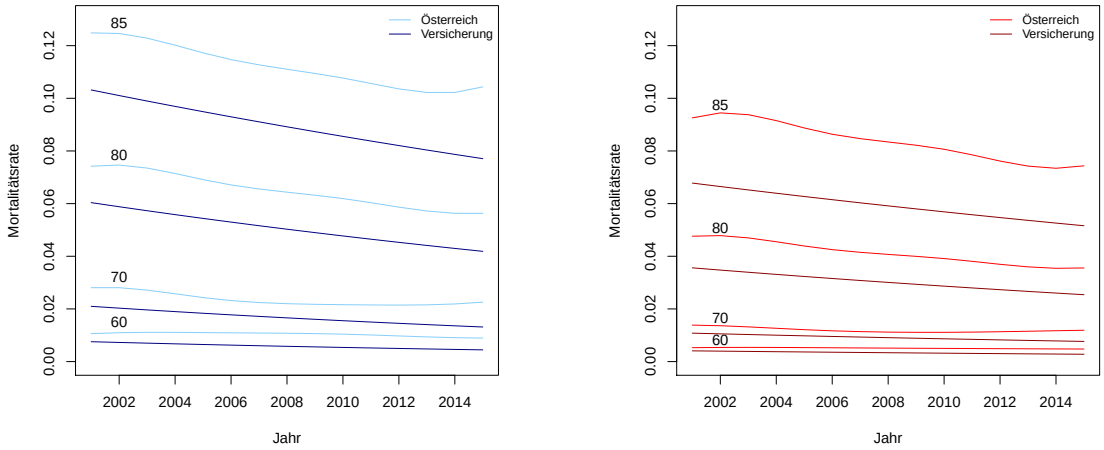

Abbildung 6.1: Vergleich der österreichweiten und bestandsspezifischen Mortalitätsrate über den Zeitraum $[2001, 2015]$ in ausgewählten Altern für Männer (links) und Frauen (rechts).

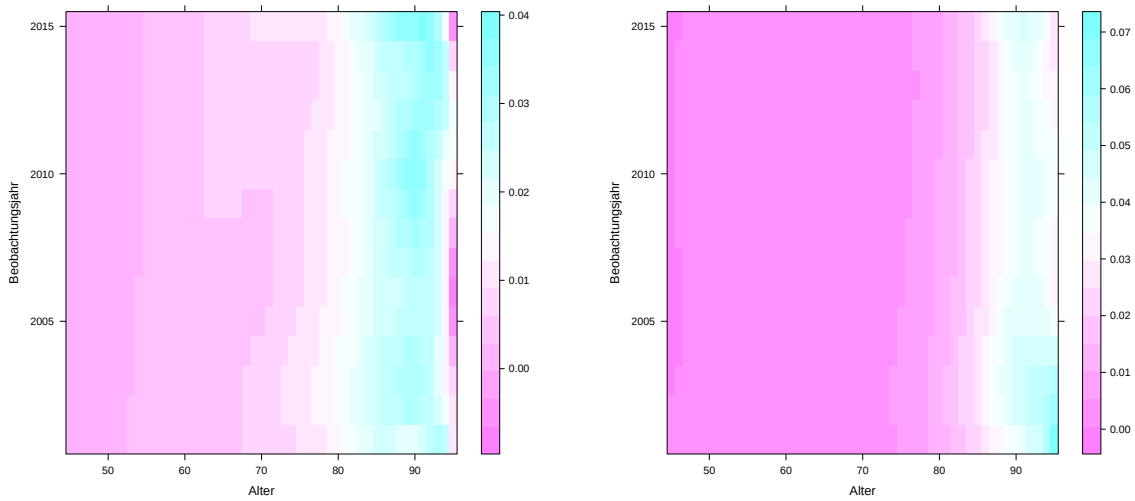

Abbildung 6.2: Differenz zwischen der österreichweiten und bestandsspezifischen Mortalitätsrate über den Zeitraum $[2001, 2015]$ und den Altersbereich $[45, 95]$ für Männer (links) und Frauen (rechts).

Conclusio

Die vorliegende Arbeit hat zum Ziel, den Mortalitätsverlauf über einen gegebenen Zeitraum und für unterschiedliche Altersklassen zu modellieren. Dazu ist ein Datensatz einer privaten Versicherung für eine Historie, die von 2001 bis 2015 reicht, vorliegend. Bei der näheren Analyse der gegebenen Daten hat es sich herauskristallisiert, dass vor allem für den unteren Altersbereich sehr wenige bis gar keine Beobachtungen vorhanden sind und somit eine Modellierung für das Intervall $[45, 95]$ als sinnvoll erachtet wurde.

Auf Grund der Natur von Mortalitätsdaten wurde sehr schnell deutlich, dass wir mit Linearen Regressionsmodellen sowie Generalisierten Linearen Modellen sehr schnell an Grenzen stoßen, da keine ausreichende Flexibilität hinsichtlich des funktionalen Zusammenhangs der Mortalitätsrate und der erklärenden Größen (Beobachtungsjahr und Alter) zugelassen wird. Im Laufe der Arbeit wurden schließlich für die Mortalitätsanalyse zwei relevante Modelle vorgestellt, das Lee-Carter Modell und das P-Spline Modell. Bei den genannten Modellen handelt es sich um zwei konträre Zugangsweisen Mortalitäten zu glätten und in die Zukunft zu projizieren.

Beide oben genannten Modelle legen die Annahme, dass die Anzahl der Sterbefälle einer Poissonverteilung folgt zugrunde. P-Splines sind für die Modellierung von Mortalitätsraten unter der Verwendung kubischer B-Splines und Straftermen zweiter Ordnung sehr gut geeignet. Außerdem sprechen der flexible Modellierungsansatz und die geringe effective dimension klar für das P-Spline Modell. Auch die relativ einfache Extrapolation, die gemeinsam mit der Parameterschätzung erfolgt, ist ein weiterer Vorteil dieser Modellklasse. Das Lee-Carter Modell hingegen bringt durch seine vorgegebene funktionale Struktur, die große Anzahl zu schätzender Parameter und die zweistufige Vorgehensweise bei der Projektion einige Nachteile mit sich und wird daher nicht für das finale Modell unter Betracht gezogen.

Abschließend wurde ein Vergleich der geglätteten Mortalität für den Versichertenbestand und jener der österreichischen Gesamtbevölkerung angestellt. Da sich die Sterblichkeiten vor allem in den oberen Altersbereichen wesentlich unterscheiden, lässt dies den Schluss zu, dass privatversicherte Personen eine im Vergleich zur Gesamtbevölkerung geringere Sterblichkeit aufweisen. Dies kann auf einem stärkeren Körperbewusstsein begründet sein und der Wichtigkeit der körperlicher Gesundheit beigemessen wird, wenn bewusst der Schritt zu einer privaten Versicherung gesetzt wird. Außerdem obliegt bei derartigen Versicherungen ebenso dem Versicherungsunternehmen die Selektion der Versicherten bei Abschluss des Vertrages, wodurch auf Grund der vorliegenden Krankheitshistorie und dem aktuellen Gesundheitszustand Anträge von Seiten des Versicherers abgelehnt werden können.

Literatur

Agresti, A. (2002). *Categorical Data Analysis* (Zweite Aufl.). Hoboken, New Jersey: Wiley-Interscience.

Box, G. und Jenkins, G. (1970). *Time Series Analysis: Forecasting and Control*. San Francisco: Holden-Day.

Brouhns, N., Denuit, M. und Vermunt, J. K. (2002). A Poisson log-bilinear regression approach to the construction of projected lifetables. *Insurance: Mathematics and Economics*, *31*, 373-393.

Camarda, C. G. (2012). MortalitySmooth: An R package for smoothing Poisson counts with P-splines. *Journal of Statistical Software*, *50*, 1-24.

Camarda, C. G. (2015). *Smoothing and forecasting Poisson counts with P-splines*. Zugriff auf `https://cran.r-project.org/web/packages/MortalitySmooth/index.html`

Cameron, A. C. und Trivedi, P. K. (1998). *Regression Analysis of Count Data*. Cambridge: Cambridge University Press.

Currie, I. (2013). Fitting models of mortality with generalized linear models and non-linear models. *Working Paper*, *Department of Actuarial Mathematics and Statistics, and the Maxwell Institute for Mathematical Sciences, Heriot-Watt University, Edinburgh*.

Currie, I., Durban, M. und Eilers, P. (2003). Using P-splines to extrapolate two-dimensional Poisson data. *Proceedings of 18th International Workshop on Statistical Modelling*, 97-102.

Currie, I., Durban, M. und Eilers, P. (2004). Smoothing and forecasting mortality rates. *Statistical Modelling*, *4*, 279-298.

Currie, I., Durban, M. und Eilers, P. (2006). Generalized linear array models with applications to multidimensional smoothing. *Journal of the Royal Statistical Society, Series B*, *68*, 259-280.

De Boor, C. (2001). *A Practical Guide to Splines*. New York: Springer.

Denuit, M., Maréchal, X., Pitrebois, S. und Walhin, J.-F. (2007). *Actuarial Modelling of Claim Counts*. West Sussex: Wiley.

Eilers, P., Currie, I. und Durban, M. (2006). Fast and compact smoothing on large multidimensional grids. *Computational Statistics and Data Analysis*, *50*, 61-76.

Eilers, P. und Marx, B. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, *11*, 89-102.

Eilers, P. und Marx, B. (2002). Generalized linear additive smooth structures. *Journal of Computational and Graphical Statistics*, *11*, 758-783.

Eilers, P., Marx, B. und Durban, M. (2015). Twenty years of P-splines. *SORT - Statistics and Operations Research Transactions*, *39*, 149-186.

Fahrmeir, L., Kneib, T., Lang, S. und Marx, B. (2013). *Regression. Models, Methods and Applications*. Berlin, Heidelberg: Springer.

Feichtinger, G. (1973). *Bevölkerungsstatistik*. Berlin: De Gruyter.

Finan, M. B. (2011). *A Reading of the Theory of Life Contingency Models: A Preparation for Exam MLC/3L*. Arkansas: Arkansas Tech University.

Hastie, T. und Tibshirani, R. (1990). *Generalized Additive Models*. London: Chapman and Hall.

Heim, S. (2009). *svcm: 2d and 3d space-varying coefficient models.* Zugriff auf `https://cran.r-project.org/web/packages/svcm/index.html`

Hilbe, J. M. (2011). *Negative Binomial Regression* (Zweite Aufl.). Cambridge: Cambridge University Press.

Knüsel, L. (2008). Singulärwert-Zerlegung und Methode der kleinsten Quadrate. *Technical Report Number 031*, *University of Munich*.

Lee, R. D. und Carter, L. R. (1992). Modeling and forecasting U.S. mortality. *Journal of the American Statistical Association*, *87*, 659-671.

Liu, X. und Braun, J. (2010). Investigating mortality uncertainty using the block bootstrap. *Journal of Probability and Statistics*, *2010*, 1-15.

Marx, B. und Eilers, P. (1998). Direct generalized additive modeling with penalized likelihood. *Computational Statistics and Data Analysis*, *28*, 193-209.

McCullagh, P. und Nelder, J. (1989). *Generalized Linear Models* (Zweite Aufl.). London: Chapman and Hall.

Milbrodt, H. und Helbig, M. (1999). *Mathematische Methoden der Personenversicherung.* Berlin: De Gruyter.

Shumway, R. H. und Stoffer, D. S. (2011). *Time Series Analysis and Its Applications* (Dritte Aufl.). New York: Springer.

Statistik Austria. (2017). *Jährliche Sterbetafeln 1947 bis 2015 für Österreich.* Zugriff auf `http://www.statistik.at/web_de/statistiken/menschen_und_gesellschaft/bevoelkerung/sterbetafeln/index.html`

Stoffer, D. (2016). *Applied statistical time series analysis.* Zugriff auf `https://cran.r-project.org/web/packages/astsa/index.html`

Thatcher, A. R., Kannisto, V. und Vaupel, J. W. (1999). *The force of mortality at ages 80 to 120.* Odense University Press. Zugriff auf `http://www.demogr.mpg.de/Papers/Books/Monograph5/start.htm`

Turner, H. und Firth, D. (2015). *Generalized nonlinear models in R: An overview of the gnm package.* Zugriff auf `https://cran.r-project.org/web/packages/gnm/index.html`

Venables, W. N. und Ripley, B. D. (2002). *Modern Applied Statistics with S* (Vierte Aufl.). New York: Springer.

Weisberg, S. (2005). *Applied Linear Regression* (Dritte Aufl.). Hoboken, New Jersey: Wiley-Interscience.

Wilmoth, J. R. (1993). Computational methods for fitting and extrapolating the Lee-Carter model of mortality change. *Technical Report*, *Department of Demography, University of California, Berkeley*.

Wood, S. (2006). *Generalized Additive Models: An Introduction with R.* Boca Raton: Chapman and Hall/CRC.

Wood, S. (2017). *Mixed GAM computation vehicle with GCV/AIC/REML smoothness estimation.* Zugriff auf `https://cran.r-project.org/web/packages/mgcv/index.html`

Zeileis, A., Kleiber, C. und Jackman, S. (2008). Regression models for count data in R. *Journal of Statistical Software*, *27*, 1-25.

Appendices

Anhang A

Datenaufbereitung

Die Rohdaten liegen in Form einer Textdatei, die 4 248 645 Zeilen umfasst, vor. In jeder Zeile sind Angaben über je eine Person im Bestand (Geburtsdatum, Geschlecht, Beobachtungsjahr, Monat der Stornierung bzw. Sterbemonat), die mit Hilfe einer eindeutigen Nummer identifiziert werden kann, enthalten.

Bei der betrachteten Personengesamtheit „Versicherte Personen einer Versicherungsgesellschaft“ handelt es sich um eine offene Personengesamtheit (vgl. Milbrodt & Helbig, 1999). Das bedeutet, dass sowohl Neueintritte in die Population durch einen Abschluss eines Versicherungsvertrages, als auch Abwanderungen durch Stornierung bzw. das Ableben auftreten können. In der vorliegenden Arbeit wird das Stornieren aller im Bestand vorhandenen Versicherungen eines Versicherungsnehmers als „Stornierung“ bezeichnet, wird hingegen nur eine Teilstornierung vorgenommen, so wird dies nicht als Stornierung erfasst, da die jeweilige Person weiterhin Teil des Versichertenbestandes ist und eine Beobachtung über ihr Ableben gemacht werden kann.

Die der Arbeit zugrunde liegenden Rohdaten haben die in Tabelle A.1 angeführte Form. Das Output in Tabelle A.1 wird mittels `head(data)` in R erzeugt.

BESTANDJ	VERTNR	VRNR	GEBDATUM	GESCHLECHT	BEWEGUNGJ	STERBEM	STORNOM
2000	543031	628	19661022	M	2001	NA	NA
2000	768616	629	19831029	M	2001	NA	NA
2000	768616	634	19860801	M	2001	NA	NA
2000	787569	639	19590915	M	2001	NA	NA
2000	454430	1025	19380108	M	2001	NA	NA
2000	689295	1061	19490810	W	2001	NA	NA

Tabelle A.1: Format der Rohdaten.

Die Rohdaten müssen für die weitere Datenanalyse modifiziert werden. Das in Listing A.1 angeführte R-Skript liest die Originaldaten mit dem Dateinamen „TOD_STORNO_2001_2015.txt“ ein, berechnet das Alter der versicherten Personen im Beobachtungsjahr als

Differenz des Geburtsjahres und des Bewegungsjahres[1] und fügt dem ursprünglichen Datensatz zwei neue Spalten, in denen der Überlebensstatus (1-Beobachtungsjahr nicht überlebt, 0-Beobachtungsjahr überlebt) sowie der Stornierungsstatus (1-Stornierung im Beobachtungsjahr, 0-keine Stornierung/Teilstornierung im Beobachtungsjahr) ersichtlich sind, hinzu. Außerdem wird eine Spalte, die die jeweilige Gewichtung nach der Zeit, die der Versicherte dem Mortalitätsrisiko im Beobachtungsjahr unterliegt, an die Tabelle angefügt.

Listing A.1: Aufbereitung der Rohdaten mit R

```
> library(lubridate)
> data<-read.table("TOD_STORNO_2001_2015.txt", header= TRUE,
+ fill= TRUE, sep=";", dec=",")

> date<-as.Date(as.character(data$GEBDATUM), "%Y%m%d")
> data[,data$GEBDATUM]<-date

> attach(data)
> ALTER<-vector(mode="integer", length=nrow(data))
> ALTER<-BEWEGUNGJ-year(GEBDATUM)

> index<-which(!is.na(STERBEM))
> UEBERLSTAT<-vector(mode="integer", length=nrow(data))
> UEBERLSTAT[index]<-1
> UEBERLSTAT[-index]<-0

> STORNOSTAT<-vector(mode="integer", length=nrow(data))
> index_storno<-which(!is.na(STORNOM))
> STORNOSTAT[index_storno]<-1
> STORNOSTAT[-index_storno]<-0

> GEWICHT<-vector(mode="numeric",length=nrow(data))
> GEWICHT[index_storno]<-STORNOM[index_storno]/12
> GEWICHT[index]<-STERBEM[index]/12
> GEWICHT[-c(index,index_storno)]<-1

> data_new<-cbind(data,ALTER,UEBERLSTAT,STORNOSTAT,GEWICHT)
```

Durch Ausführen des in Listing A.1 angeführten R-Codes werden die Rohdaten also in eine für die weiteren Berechnungen passende Form gebracht und haben schließlich die in Tabelle A.2 ersichtliche Struktur.

Des Weiteren ist es interessant eine Kontingenztafel zu erstellen, die nach Beobachtungsjahr, Alter und Geschlecht die Anzahl der Überlebenden, der Sterbefälle und der Stornierungen sowie die Gesamtzahl der betrachteten Personen angibt.

[1] Diese vereinfachte Berechnungsweise des Alters wird gewählt, da es bei der zusätzlichen Betrachtung von Monaten zu Ungenauigkeiten kommen kann. Dies liegt daran, dass bei der Erfassung von Sterbefällen und Stornierungen zeitliche Verzögerungen auftreten können und das tatsächliche Sterbemonat mit dem in den Daten eingetragenen Sterbemonat nicht zwingend übereinstimmen muss.

BESTANDJ	VERTNR	VRNR	GEBDATUM	GESCHLECHT	BEWEGUNGJ	STERBEM	STORNOM	ALTER	UEBERLSTAT	STORNOSTAT	GEWICHT
2000	543031	628	1966-10-22	M	2001	NA	NA	35	0	0	1
2000	768616	629	1983-10-29	M	2001	NA	NA	18	0	0	1
2000	768616	634	1986-08-01	M	2001	NA	NA	15	0	0	1
2000	787569	639	1959-09-15	M	2001	NA	NA	42	0	0	1
2000	454430	1025	1938-01-08	M	2001	NA	NA	63	0	0	1
2000	689295	1061	1949-08-10	W	2001	NA	NA	52	0	0	1

Tabelle A.2: Format der bearbeiteten Rohdaten.

Listing A.2: Erstellung der Kontingenztafel

```
1  > Tod<-aggregate(UEBERLSTAT~ALTER+GESCHLECHT+BEWEGUNGJ, data=
2  + data_new,sum)
3  > Storno<-aggregate(STORNOSTAT~ALTER+GESCHLECHT+BEWEGUNGJ,data=
4  + data_new,sum)
5  > Ueberl<-aggregate(UEBERLSTAT~ALTER+GESCHLECHT+BEWEGUNGJ,data=
6  + data_new,
7  + FUN=function(UEBERLSTAT){length(which(UEBERLSTAT==0))})
8  > Tabelle<-merge(merge(Tod,Storno,by=c("BEWEGUNGJ","GESCHLECHT",
9  + "ALTER"),sort=F),Ueberl,
10 + by=c("BEWEGUNGJ","GESCHLECHT","ALTER"), sort=F)

12 > Tabelle$Summe<-rowSums(Tabelle[,c(Tabelle$DEATHS,
13 + Tabelle$UE_ja)])

15 > Tabelle<-Tabelle[,c(Tabelle$Jahr, Tabelle$ALTER,
16 + Tabelle$GESCHLECHT, Tabelle$UE_ja, Tabelle$DEATHS,
17 + Tabelle$EXPOSURES, Tabelle$ST_ja)]

19 > names(Tabelle)<-c("Jahr","Alter","Geschlecht","UE_ja",
20 + "Deaths", "Exposures","ST_ja")
```

In dieser Kontingenztabelle werden die Gewichtungen, wie sie im Abschnitt 2.2.3 beschrieben werden, noch nicht betrachtet. Das heißt, es handelt sich bei den Exposures um die Anzahl der Versicherten, die sich am Anfang des Beobachtungsjahres im Bestand befinden. Sollen Gewichtungen berücksichtigt werden, so müssen Zeile 3 und 4 in Listing A.2 durch Ueberl_gewicht<−aggregate(GEWICHT∼ALTER+ GESCHLECHT+BEWEGUNGJ, data= data_new, FUN=sum) ersetzt werden.

Für die Datenanalyse und auch die weitere Verwendung der Daten in den Modellen ist es notwendig, die Anzahl der Sterbefälle sowie die Anzahl der unter Risiko stehenden Personen nach Geschlecht, Alter und Jahr in Form von Matrizen darzustellen. Das in Listing A.3 abgebildete R-Skript liest die Kontingenztabelle „Kontingenztabelle_2001_2015.txt“, die wie in Listing A.2 aufbereitet wurde, ein und erstellt anschließend mit Hilfe der Funktion acast aus der Library reshape2 die geschlechtsspezifischen Matrizen.

Listing A.3: Erstellung der Datenmatrizen

```
1  > library(reshape2)

3  > data<-read.table("Kontingenztabelle_2001_2015.txt",
4  + header= TRUE, sep=" ")
```

```
> D_M<-acast(subset(data, data$Geschlecht=="M"), Alter~Jahr,
+ value.var='Deaths', fun.aggregate=sum, margins=FALSE)
> E_M<-acast(subset(data, data$Geschlecht=="M"), Alter~Jahr,
+ value.var='Exposures', fun.aggregate=sum, margins=FALSE)

> D_W<-acast(subset(data, data$Geschlecht=="W"), Alter~Jahr,
+ value.var='Deaths', fun.aggregate=sum, margins=FALSE)
> E_W<-acast(subset(data, data$Geschlecht=="W"), Alter~Jahr,
+ value.var='Exposures', fun.aggregate=sum, margins=FALSE)
```

Die Funktion data.select, die in Listing A.4 angeführt ist und im R-Skript „Matrizen_Aufbereitung.R“ enthalten ist, dient der automatisierten Matrizenerstellung je nach gewünschtem Altersbereich, Beobachtungszeitraum und Geschlecht.

Listing A.4: Funktion zur Datenselektion

```
> data.select<-function(year,age,sex,type){

+ if (sex=="M" & type=="Deaths"){
+ return(D_M[which(rownames(D_M) %in% age),
+ which(colnames(D_M) %in% year)])}

+ else if (sex=="M" & type=="Exposures"){
+ return(E_M[which(rownames(E_M) %in% age),
+ which(colnames(E_M) %in% year)])}

+ else if (sex=="W" & type=="Deaths"){
+ return(D_W[which(rownames(D_W) %in% age),
+ which(colnames(D_W) %in% year)])}

+ else
+ return(E_W[which(rownames(E_W) %in% age),
+ which(colnames(E_W) %in% year)])}
```