Dipl.-Ing. Johannes Moosbrugger BSc

Modellierung von Stornowahrscheinlichkeiten in der privaten Krankenversicherung

MASTERARBEIT

zur Erlangung des akademischen Grades eines Diplom-Ingenieurs

Operations Research und Statistik

eingereicht an der

Technische Universität Graz

Betreuer:
Univ.-Prof. Dipl.-Ing. Dr.techn. Herwig FRIEDL

Institut für Statistik

Graz, März 2016

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 Klasse der Generalisierten Additiven Modelle ermöglicht es, in Generalisierten Linearen Modellen im Allgemeinen und in logistische Regressionsmodellen im Speziellen auch nichtlineare Abhängigkeitsstrukturen abzubilden. Die Darstellung solcher nichtlinearer Zusammenhänge erfolgt dabei durch sogenannte Glatte Funktionen. In unserem Fall werden diese Glatten Funktionen durch kubische Splinefunktionen erzeugt. Wir nutzen die Modellklasse der Generalisierten Additiven Modelle, um die Wahrscheinlichkeit für eine Stornierung eines Krankenversicherungsvertrages in Abhängigkeit verschiedener Charakteristiken des jeweiligen Versicherungsnehmers und des Vertrages an sich zu modellieren. Im resultierenden Modell werden drei stetige Prädiktoren als Glatte Funktionen und zwei nominale Prädiktoren als mehrstufige Faktoren modelliert. Auf Basis der Prädiktionsgüte der betrachteten Modelle ist ersichtlich, dass eine zusätzliche Modellierung von Interaktionstermen keinen signifikanten Vorteil bringt. Die resultierende Einfachheit dieses Modells ermöglicht außerdem eine grafische Interpretation der Abhängigkeitsstrukturen zwischen Stornowahrscheinlichkeit und den stetigen Prädiktoren. Anhand dieser Ergebnisse kann das Aufkommen von Stornierungen für unterschiedliche Kalenderjahre direkt verglichen werden. Außerdem ermöglicht dieses Modell die akkurate Schätzung sowohl der Anzahl an Stornierungen wie auch der Summe der stornierten Versicherungsprämien für ein Folgejahr. Die praktische Umsetzung erfolgt dabei mit der freien Statistiksoftware R.

ABSTRACT

Generalized additive models enable us to modell nonlinear relationships in generalized linear modells and therefore also in logistic regression modells. For this purpose, parts of the linear predictor are specified in terms of a sum of smooth functions. These smooth functions are represented by a set of basis functions. In our application, cubic splines are used for this purpose. We apply the class of generalized additive regression modells to describe the dependency between the withdrawal of health insurance contracts and the specific characterisations of the insurance holder and the contract itself. The resulting modell consists of three continuous predictors which are modelled as smooth functions as well as two nominal predictors modelled as multilevel factors. An analysis of the prediction quality shows that no further interaction terms are needed. The resulting simplicity of the modell enables a graphical interpretation of the modelled dependencies. By the use of these results, withdrawal patterns of different years can be compared directly. This model also enables an accurate prediction of the number of cancelations and the sum of cancelled insurance premiums for an upcoming year. The practical implementation is done by the use of the free statistic software R.

DANKSAGUNG

An dieser Stelle möchte ich mich bei all denjenigen bedanken, die mich während der Anfertigung dieser Masterarbeit unterstützt und motiviert haben.

Zuerst gebührt mein Dank Herr Prof. Friedl, der meine Masterarbeit betreut und begutachtet hat. Für die hilfreichen Anregungen und die konstruktive Kritik bei der Erstellung dieser Arbeit möchte ich mich herzlich bedanken.

Ein besonderer Dank gilt Herr Stockreiter von der Merkur Versicherung AG. Ohne die zur Verfügung gestellten Daten hätte diese Arbeit nicht entstehen können. Mein Dank gilt ihrer Informationsbereitschaft und ihren interessanten Beiträgen und Antworten auf meine Fragen.

Abschließend möchte ich mich bei meinen Eltern Michaele und Harald bedanken, die mir mein Studium durch ihre Unterstützung ermöglicht haben und stets ein offenes Ohr für meine Sorgen hatten.

Johannes Moosbrugger,
Graz, 08.03.2016

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

Das Ziel dieser Arbeit ist die Modellierung der Stornowahrscheinlichkeit von Krankenversicherungsverträgen. Von Interesse ist dabei vor allem die Frage, wovon die Wahrscheinlichkeit für eine Vertragsbeendigung abhängt und in Folge dessen ob sich diese Abhängigkeit im Betrachtungszeitraum verändert.

Als Datenbasis unserer Modellierung dienen jährliche Beobachtungen des Bestandes an Privat-Krankenversicherungsverträgen eines Versicherungsunternehmens. Dieser Bestand setzt sich aus einzelnen Versicherungsverträgen zusammen. Ein Versicherungsvertrag kann dabei aus mehreren Tarifen bestehen. Am Ende jedes Kalenderjahres ist für jede dieser Tarifpositionen bekannt, ob sie im Laufe des betrachteten Jahres gekündigt wurde oder nicht. Zusätzlich stehen uns noch weitere Informationen bezüglich jeder Tarifposition, wie z.B. verschiedene Angaben zum Versicherungsnehmer, zur Verfügung.

Unser Ansatz besteht darin, die binäre Responsevariable (storniert bzw. nicht storniert) durch ein logistisches Regressionsmodell zu beschreiben. Im Zuge der Modellselektion sind wir damit in der Lage einzuschätzen, welche der uns zur Verfügung stehenden Informationen zu jeder Tarifposition einen signifikanten Einfluss auf die Stornowahrscheinlichkeit haben. Wir wollen dabei so vorgehen, dass für jedes Kalenderjahr ein separates Modell geschätzt wird. Die Modellselektion findet dann parallel auf allen Kalenderjahren statt. Dies bedeutet, dass wir ein Modell suchen, dessen Auswahl wir in jedem der uns zur Verfügung stehenden Beobachtungsjahre vertreten können. Dies hat zum einen den Vorteil, dass unser Modell nicht zu sehr auf die Spezifika eines einzelnen Beobachtungsjahres angepasst ist. Zum anderen ermöglicht uns dieses Vorgehen ebenfalls die geschätzten Koeffizienten unseres Modells für die verschiedenen Jahre zu vergleichen. Damit sind wir in der Lage eine Aussage darüber zu treffen, ob sich die Abhängigkeitsstrukturen zwischen der Stornowahrscheinlichkeit und den erklärenden Variablen in den verschiedenen Kalenderjahren unterscheiden.

Zu beachten ist bei diesem Vorgehen vor allem der Schritt der Modellevaluation. Ist dies bei logistischen Regressionsmodellen, aufgrund der unterschiedlichen Skalen der Responsevariable und des Linearen Prädiktors, von vornherein schon erschwert, so erfordert es bei unserer konkreten Problemstellung aufgrund des sehr geringen Stornoanteils (im Bereich weniger Prozentpunkte) individueller Lösungsansätze. Aus diesem Grund verwenden wir zum Zwecke des Modellvergleichs spezifisch an unsere Problemstellung angepasste Maße für die Prädiktionsgüte von logistischen Regressionsmodellen.

Ein weiterer Punkt, den es zu beachten gilt, ist der große Datenumfang des Versicherungsbestandes. Pro Kalenderjahr stehen uns rund 300.000 beobachtete Tarifpositionen zur Verfügung. Dies hat zur Folge, dass wir bereits zur Schätzung von mäßig komplexen Modellen auf Techniken der Rechen-Parallelisierung zurückgreifen müssen. Um komplexe

Modelle, d.h. zum Beispiel Modelle mit vielen Interaktionen, in einer praktikablen Zeit zu berechnen, bedarf es allerdings größerer Rechencluster.

Diese Masterarbeit ist wie folgt gegliedert. In Kapitel 2 erläutern wir noch einmal detailliert die Problemstellung und beschreiben die uns zur Verfügung stehenden Daten. Motiviert durch diese erste Datenanalyse wollen wir uns in Kapitel 3 die theoretischen Grundlagen zur Modellierung der Stornowarscheinlichkeiten erarbeiten. Dies bedeutet, dass wir hier die Klasse der Generalisierten Additiven Modelle vorstellen und dabei näher auf die logistischen Regressionsmodelle eingehen werden. Die eigentliche Modellierung folgt in Kapitel 4. Diese lässt sich grob in die Bereiche der Modellselektion und der Modellevaluierung unterteilen. In der Modellselektion wählen wir eine Menge an möglichen Modellen und bestimmen die relevanten Prädiktoren. Im Abschnitt der Modellevaluierung prüfen wir, wie sich diese Modelle bezüglich der von uns eingeführten Maße der Prädiktionsgüte verhalten. Nachdem wir ein Modell bestimmt haben, benutzen wir dieses um die Abhängigkeitstrukturen zwischen Stornowahrscheinlichkeit und den erklärenden Variablen zu beschreiben. Diese Modellauswertung erfolgt in Kapitel 5. In Kapitel 6 fassen wir unsere Ergebnisse zusammen und resümieren.

1.1 Notation

In dieser Arbeit machen wir Gebrauch von folgenden Notationen und Abkürzungen:

$\mathbb{E}(\cdot)$ ... die Erwartungswertfunktion

$\mathbb{V}ar(\cdot)$ ... die Varianzfunktion

$\mathbb{R}_+$ ... die Menge der nicht negativen reellen Zahlen

$\mathbb{R}^{n\times m}$ ... die Menge der reellwertigen $n \times m$ Matrizen

$\mathbb{I}$ ... die Einheitsmatrix in den durch den Kontext gegebenen Dimensionen

x' ... die Transposition von $x \in \mathbb{R}^n$

$\|x\|$ ... die Euklidsche Norm eines Vektors $x \in \mathbb{R}^n$

$\exp(\cdot)$ ... die Exponentialfunktion

$\log(\cdot)$ ... die natürliche Logarithmusfunktion

$N(\mu, \sigma^2)$ ... die Normalverteilung mit Erwartungswert μ und Varianz σ^2

LM... Lineares Modell

GLM... Generalisiertes Lineares Modell

AM... Additives Modell

GAM... Generalisiertes Additives Modell

LEF... Lineare Exponential Familie

AIC... Akaike Informationskriterium

BIC... Bayessches Informationskriterium

SSE... Summe der Fehlerquadrate

VR... Versicherungsnehmer

2 Problemstellung

In diesem Kapitel wollen wir die genaue Ausgangslage unserer Arbeit erläutern. Dabei erklären wir in Abschnitt 2.1 zunächst das genaue Anforderungsprofil an unser Projekt. In Abschnitt 2.2 analysieren wir die uns zu Verfügung stehenden Daten. Diese erste Analyse der Daten motiviert unsere Wahl der Modellklasse in Abschnitt 2.3.

2.1 Anforderungen

Ein Versicherungsunternehmen besitzt zu jedem Zeitpunkt einen gewissen Bestand an Krankenversicherungsverträgen. Dieser setzt sich aus allen laufenden Verträgen, die zwischen dem Versicherungsunternehmen und einer Privatperson im Bereich der Krankenversicherung geschlossen wurden, zusammen. Ein solcher Vertrag verpflichtet das Versicherungsunternehmen im Falle von Erkrankungen, Mutterschaft und manchmal auch nach Unfällen die Kosten für die Behandlung voll oder teilweise zu erstatten. Im Gegenzug dafür erhält das Unternehmen von dem Versicherungsnehmer die sogenannte Versicherungsprämie, eine fixe monatliche Zahlung. Diese Prämien werden dafür genutzt, die im Kollektiv aufgetretenen Kosten für Behandlungen zu erstatten sowie laufende Kosten für Verwaltung und Ähnlichem zu decken. Sie stellen somit einen zentralen Bestandteil der finanziellen Planung eines Versicherungsunternehmens dar. Die Möglichkeit einer Stornierung eines solchen Krankenversicherungsvertrages stellt daher ein finanzielles Risiko dar, welches es abzuschätzen gilt.

Wir beziehen uns in dieser Arbeit immer auf einen Betrachtungszeitraum von einem Kalenderjahr. Natürlich besitzt ein Versicherungsnehmer, unter Betrachtung der vertraglichen Kündigungsfrist, zu jedem Zeitpunkt das Recht den Krankenversicherungsvertrag zu kündigen. Zum Zwecke der Modellierung verzichten wir jedoch darauf den genauen Kündigungstermin zu betrachten und beschränken uns auf die Tatsache, ob der Vertrag am Ende des Beobachtungsjahres noch gültig ist oder nicht. Dieses Vorgehen hat mehrere Gründe. Zum einen erfordert unser konkreter Anwendungsfall nicht die Stornowahrscheinlichkeit in Abhängigkeit von Jahreszeiten oder Monaten zu modellieren. Zum anderen ist eine genaue Zuordnung des Kündigungstermines, z.B. zu einem bestimmten Monat, nur sehr schwer umsetzbar. Dies liegt daran, dass unterschiedliche Faktoren wie z.B. Kündigungsfristen, Urlaubszeiten o.ä., das genaue Datum einer Kündigung verschleiern können.

Unser Ziel ist es also ein Modell für die jährliche Stornowahrscheinlichkeit von Krankenversicherungstarifen zu erstellen. Mit Hilfe dieses Modells wollen wir feststellen, von welchen erklärenden Variablen die Wahrscheinlichkeit für eine Stornierung abhängt. Dabei soll, wenn möglich, auch die Struktur dieser Abhängigkeit geschätzt werden. Natürlich wollen wir dieses Modell anschließend auch für die Schätzung der zu erwartenden Anzahl an Stornierungen verwenden. Aufgrund des oben beschriebenen Anwendungsfeldes unserer Arbeit

in der Risikoevaluierung wäre auch eine Schätzung der stornierten monatlichen Prämien von Interesse. Mit Hilfe einer solchen Schätzung wäre es also möglich, den zu erwartenden Ausfall an monatlichen Prämienzahlungen adäquat in die finanziellen Planungen miteinzubeziehen.

2.2 Datenbasis

Als ersten Schritt der Modellierung wollen wir in diesem Abschnitt eine kurze Analyse der uns zur Verfügung stehenden Daten vornehmen. Wie bereits erwähnt, setzt sich der Bestand an Krankenversicherungen aus einzelnen Krankenversicherungsverträgen zusammen. Jeder einzelne Vertrag besteht wiederum aus einer oder mehreren Tarifpositionen. Diese Tarifpositionen stellen unsere Beobachtungen dar, d.h. für jede dieser Positionen wissen wir, ob sie am Ende des Beobachtungsjahres storniert wurde oder nicht.

Unsere Daten umfassen den Zeitraum der drei aufeinanderfolgenden Kalenderjahre 2012, 2013 und 2014. Für jedes dieser Jahre besitzen wir eine Auflistung aller zu Beginn des jeweiligen Jahres gültigen Tarifpositionen. Jede Tarifposition besitzt folgende Informationen:

- VERTRNR: Eine eindeutige Kennung die den übergeordneten Krankenversicherungsvertrag identifiziert.
- VRNR: Eine eindeutige Kennung die den Versicherungsnehmer identifiziert.
- PG: Die Klasse einer Tarifposition.
- AGE: Das Alter des Versicherungsnehmers zu Beginn des Jahres.
- SEX: Das Geschlecht des Versicherungsnehmers.
- BEGINN: Das Jahr in dem die Vertragsposition eröffnet wurde.
- GRUPPE: Binäre Variable die beschreibt, ob der übergeordnete Vertrag Teil einer Gruppenversicherung ist.
- RABATT: Binäre Variable die beschreibt, ob die Prämie einer Tarifposition in irgendeiner Form rabattiert wurde oder nicht.
- PRAEMIE: Die monatliche Prämie in Euro.
- STORNO: Binäre Variable die beschreibt, ob die Tarifposition im Laufe des Jahres storniert wurde oder nicht.

Ein besonderes Augenmerk gilt hier dem großen Datenumfang. In allen drei Kalenderjahren gab es jeweils zu Beginn des Jahres um die 300.000 offene Tarifpositionen. Wie in Abbildung 1 ersichtlich, wurden davon in jedem Kalenderjahr zwischen 8.000 und 9.000

Positionen storniert. Als Bestand wird hier und im Folgenden immer die Gesamtheit aller Tarife eines Kalenderjahres bezeichnet. Abbildung 1 ist demnach so zu interpretieren, als dass im Jahr 2012 8.740 der insgesamt 284.862 Tarife storniert wurden.

Im restlichen Verlauf dieses Abschnitts werden wir näher auf die Abhängigkeit zwischen der Anzahl an stornierten Tarife und den oben aufgelisteten erklärenden Variablen eingehen. Wir ziehen dafür nur das aktuellste Beobachtungsjahr (2014) in Betracht und verweisen für eine detailliertere Analyse und einen Vergleich der einzelnen Jahre auf unser vorangegangenes Projekt Moosbrugger (2015).

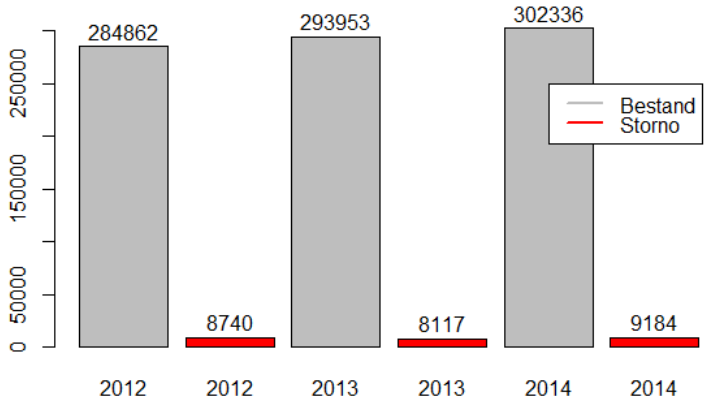

Abbildung 1: Bestand und Storno in den Jahren 2012 bis 2014

2.2.1 Das Alter des Versicherungsnehmers

Zunächst wollen wir das Alter des Versicherungsnehmers betrachten. Diese Variable ist dabei als natürliche Zahl gegeben. In allen Beobachtungsjahren sind vom Neugeborenen mit Altersklasse gleich Null bis hin zu sehr hohen Altern (das Maximum liegt bei 106 Jahren) alle Altersklassen vertreten. Naturgemäß wird der Datenbestand für hohe Altersklassen immer dünner. In Abbildung 2 links ist dieser Sachverhalt gut zu erkennen. Die meisten Versicherungsnehmer sind jünger als 75 Jahre. Für diese Darstellung haben wir die Variable Alter in Altersklassen zu je fünf Jahren zusammengefasst.

Ebenfalls in Abbildung 2 links wurde die absolute Anzahl an Stornierungen für jede dieser Altersklassen rot dargestellt. Auch hier gilt, dass die meisten der stornierten Tarife zu Versicherungsnehmern gehören, die jünger als 60 Jahre sind. Interessant ist dabei die Beobachtung, dass alle Quartile (das 25%, 50% und 75%- Quantil) der Variable Alter für die stornierten Tarife deutlich kleiner sind als die Vergleichswerte im gesamten Bestand. Die entsprechenden Werte sind in Tabelle 1 ersichtlich. Selbiges gilt auch für den Mittelwert.

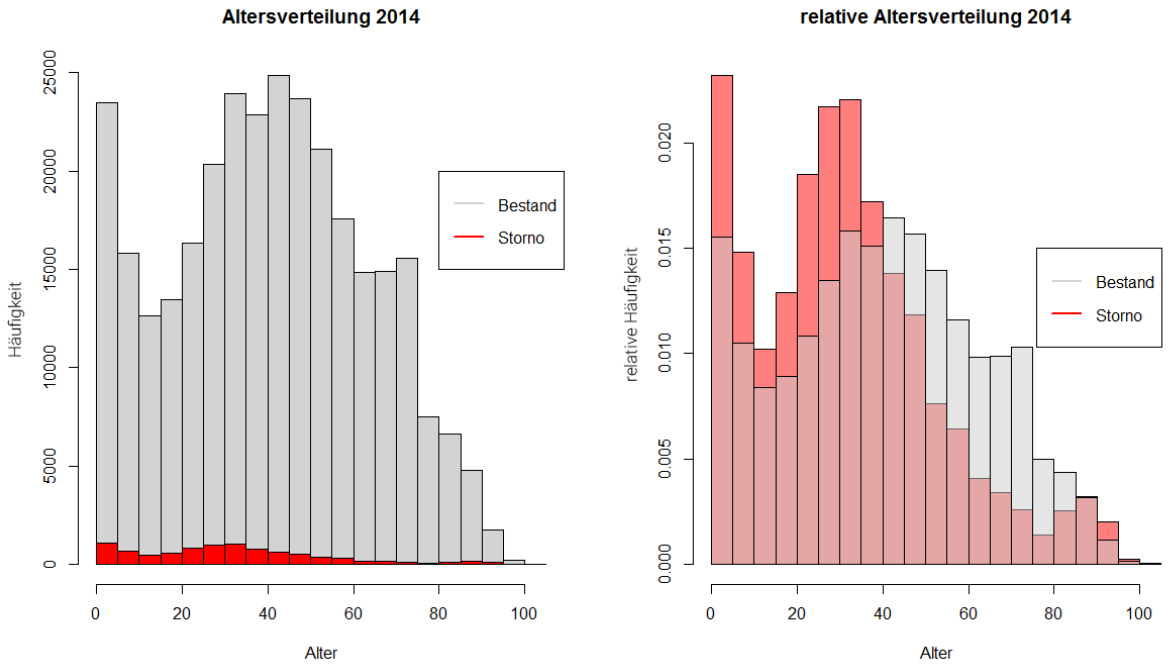

Abbildung 2: Histogramm von Alter für Bestand und Storno im Jahre 2014

Es scheint also so, als ob jüngere Versicherungsnehmer eine erhöhte Stornowahrscheinlichkeit aufweisen. Um diesen Zusammenhang näher zu betrachten haben wir in Abbildung 2 rechts die relativen Häufigkeit der Altersklassen im gesamten Bestand mit jenen der stornierten Tarife verglichen. Auch für diese Darstellung wurde die Variable Alter wieder in Altersklassen zu je fünf Jahren zusammengefasst. In dieser Abbildung der beiden Histogramme (die transparent übereinander gelegt wurden) erkennen wir gut die Unterschiede zwischen den jüngeren und den älteren Altersklassen. Hätte die Variable Alter keinen Einfluss auf die Stornowahrscheinlichkeit, so müssten sich beide Histogramme decken. Wir erkennen allerdings einen deutlich höheren Stornoanteil für die unteren Altersklassen bzw. einen deutlich geringeren Stornoanteil für die höheren Altersklassen.

Daten	Min.	1st Qu.	Median	Mean	3rd. Qu.	Max
Bestand 2014	0.00	24.00	41.00	40.88	58.00	105.00
Storno 2014	0.00	16.00	30.00	32.00	44.00	102.00

Tabelle 1: Zusammenfassung der Variable Alter für Bestand und Storno

Um den Zusammenhang zwischen dem Alter und der Stornowahrscheinlichkeit besser darstellen zu können, berechnen wir für jedes Alter (jede natürlich Zahl zwischen Null und 106) den relativen Anteil an Stornierungen. In Abbildung 3 haben wir das Alter gegen

diesen relativen Anteil an Stornierungen aufgetragen. So erkennen wir, dass der Anteil an Stornierungen in den ersten vier bis fünf Jahren anzusteigen scheint. Auch zwischen dem 20. und 25. Lebensjahr erkennen wir einen erhöhten Stornoanteil. Danach scheint dieser Anteil kontinuierlich zu sinken. Die Effekte der sehr hohen Altersklassen sind nicht zu interpretieren. Hier ist die Anzahl der Beobachtungen zu gering um daraus irgendwelche Schlüsse ziehen zu können. Es ist auf jeden Fall festzuhalten, dass der Zusammenhang zwischen dem Alter und dem relativen Anteil an Stornierungen wohl nur bedingt von linearer Natur ist.

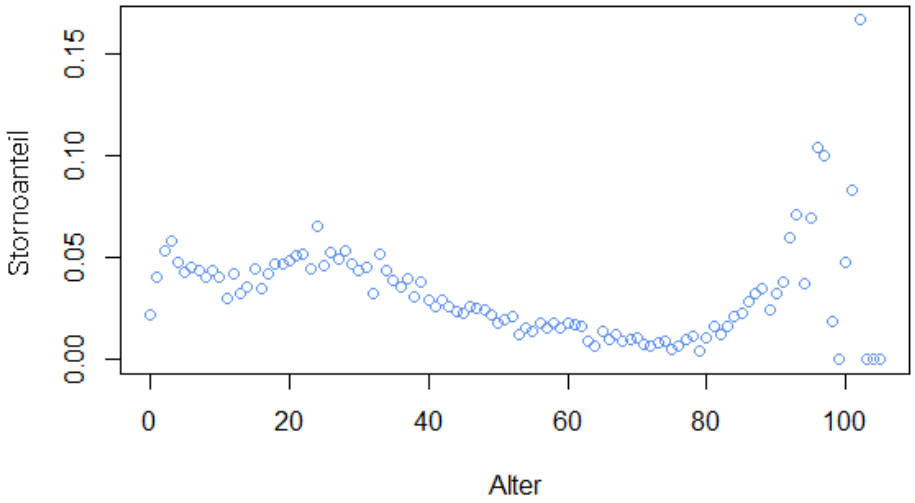

Abbildung 3: Abhängigkeit des Stornoanteils von Alter im Jahre 2014

2.2.2 Die monatliche Prämie

Ähnlich wie für das Alter werden wir auch den Prädiktor der monatlichen Prämie analysieren. Zuerst betrachten wir in Tabelle 2 wieder eine grobe Zusammenfassung der Beobachtungen. Wie wir erkennen können, gibt es Tarife für die keine Prämie gezahlt wird. Dafür gibt es zwei mögliche Gründe. Zum einen sind Neugeborene im ersten Lebensjahr prämienfrei, zum anderen besteht für den Versicherungsnehmer die Möglichkeit einen Krankenversicherungsvertrag ruhend zu stellen. Dies ist für eine maximale Zeitspanne von zwölf Monaten möglich. Ansonsten beobachten wir ein ähnliches Bild wie bereits zuvor für das Alter des Versicherungsnehmers. Stornierte Tarife weisen im Durchschnitt eine deutlich niedrigere monatliche Prämie auf als dies im gesamten Bestand der Fall ist. Demnach neigen wohl eher Versicherungsnehmer mit einer vergleichsweise niedrigeren Prämie dazu, ihren Tarif zu stornieren.

Daten	Min.	1st Qu.	Median	Mean	3rd. Qu.	Max
Bestand 2014	0.00	16.02	33.34	57.19	80.93	609.60
Storno 2014	0.00	13.60	22.34	38.17	42.22	454.20

Tabelle 2: Zusammenfassung der Variable Prämie für Storno

Wir wollen diese Beobachtung auch wieder anhand von Histogrammen der absoluten und relativen Häufigkeiten überprüfen. In Abbildung 4 sehen wir die entsprechenden Grafiken. Für diese Darstellung wurde die Variable Prämie in Klassen mit einer Größe von je 10 Euro eingeteilt. Anhand der absoluten Häufigkeiten (Abbildung 4 links) ist ersichtlich, dass sich ein Großteil der monatlichen Prämien in Größenordnungen von weniger als 100 Euro bewegt. Interessant sind auch wieder die (transparent übereinandergelegten) Histogramme der relativen Häufigkeiten (Abbildung 4 rechts). Hier erkennen wir wieder, dass wohl vermehrt Tarife mit geringen monatlichen Prämien storniert werden.

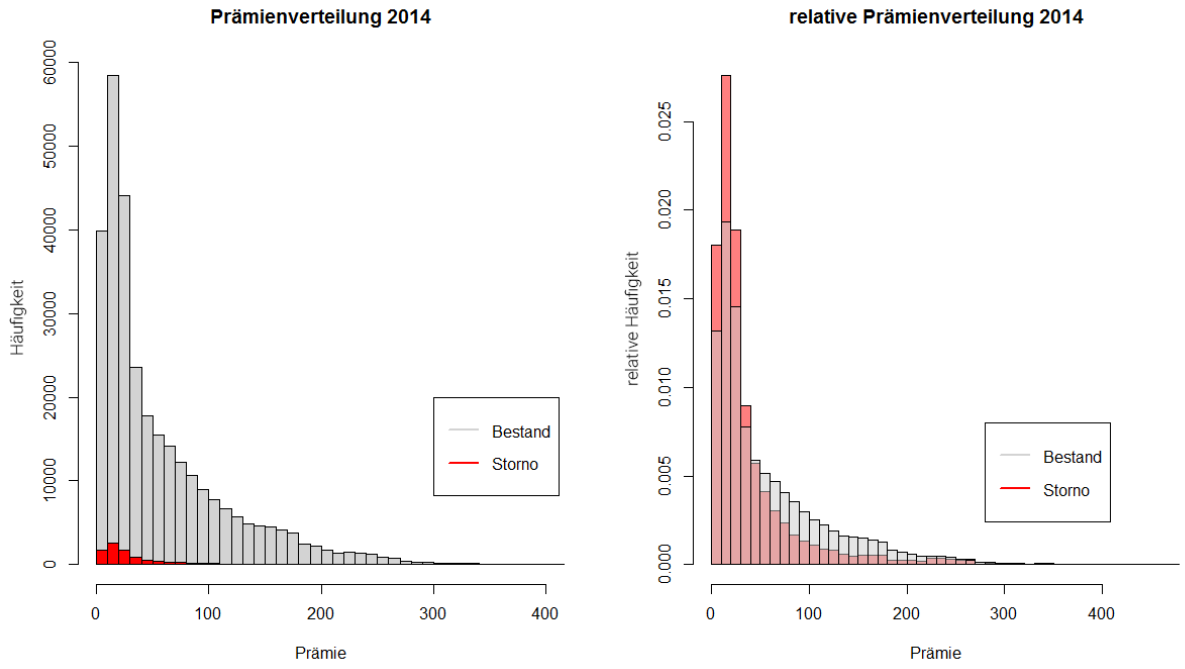

Abbildung 4: Histogramm der Prämie für Bestand und Storno im Jahre 2014

Um die Abhängigkeitsstruktur zwischen dem Auftreten von Stornierungen und der monatlichen Prämie grafisch abzubilden, unterteilen wir die Prämien in Prämienklassen, sodass eine möglichst gleichmäßige Verteilung auf diese Klassen entsteht. Für jede dieser Prämienklassen berechnen wir dann wieder den relativen Anteil an Stornierungen. Diese Darstellung ist in Abbildung 5 ersichtlich. Auch hier beobachten wir, dass zwischen

der monatlichen Prämie und dem relativen Anteil an Stornierungen wohl kein linearer Zusammenhang besteht. Deutlich sehen wir einen abnehmenden Stornoanteil mit zunehmender monatlich zu zahlender Prämie. Da aber ein überwiegender Anteil der Tarife zu den unteren Prämienklassen zuzuordnen ist, und diese Gewichtung in dieser Abbildung nicht ersichtlich ist, wollen wir auf Basis dieser Darstellung nicht zu viele Schlüsse über die Abhängigkeitsstruktur zwischen Prämie und Storno ziehen. Auch sei noch einmal erwähnt, dass die Unterteilung der Tarife in verschiedene Prämienklassen von uns willkürlich und nicht äquidistant vorgenommen wurde. Diese Unterteilung diente nur der grafische Darstellbarkeit und das Ziel dabei war lediglich eine möglichst gleichmäßige Verteilung der Tarife auf die so entstehenden Prämienklassen.

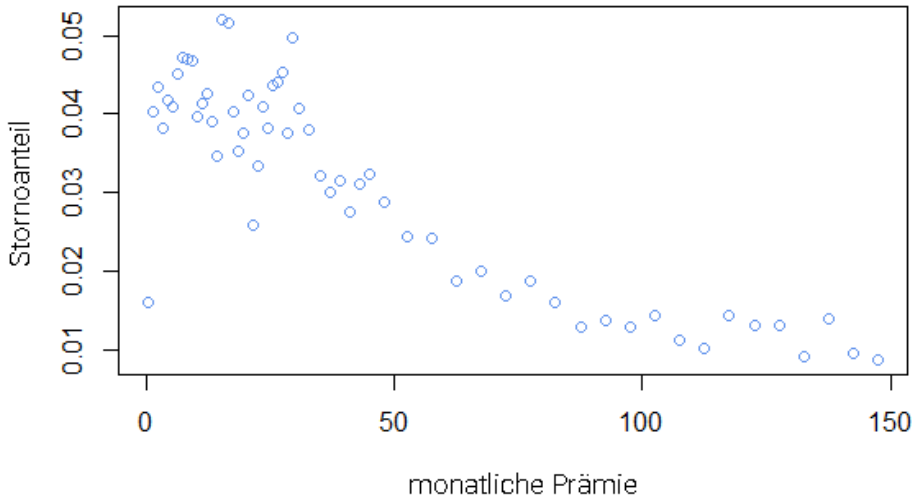

Abbildung 5: Abhängigkeit des Stornoanteils von der Prämie im Jahre 2014

2.2.3 Die Laufzeit

Die Laufzeit eines Tarifes ergibt sich aus der Differenz zwischen dem aktuellen Beobachtungsjahr und der Variable „BEGINN". Um einen ersten Überblick zu erhalten, haben wir die Beobachtungen dieser Variable in Tabelle 3 zusammengefasst. Wieder lässt sich ein deutlicher Unterschied zwischen dem gesamten Bestand und den stornierten Tarifen ausmachen. So scheint es, als dass die stornierten Tarife geringere mittlere Laufzeiten aufweisen als dies die Tarife im gesamten Bestand tun. Selbiges gilt für alle anderen Quantile die die Variable Laufzeit beschreiben. Es liegt also nahe zu vermuten, dass eher Tarife mit einer geringen Laufzeit dazu neigen storniert zu werden.

Um zu überprüfen ob tatsächlich vermehrt Tarife mit einer geringeren Laufzeit storniert werden, vergleichen wir wieder die Histogramme der relativen und absoluten Häufigkeiten.

Daten	Min.	1st Qu.	Median	Mean	3rd. Qu.	Max
Bestand 2014	1.00	5.00	14.00	21.13	34.00	89.00
Storno 2014	1.00	4.00	8.00	14.52	21.00	85.00

Tabelle 3: Zusammenfassung der Variable Laufzeit für Storno

Zum Zwecke der Übersichtlichkeit wurden hierfür die Tarife nach ihren Laufzeiten in Blöcke zu je fünf Jahren zusammengefasst. In Abbildung 6 rechts machen wir die gleiche Beobachtung wie wir es bereits in der Zusammenfassung der Beobachtungen von Laufzeit in Tabelle 3 gemacht haben. So sehen wir, dass Tarife mit einer Laufzeit von mehr als 20 Jahren im gesamten Bestand relativ gesehen häufiger vorkommen als sie dies in der Menge der stornierten Tarife tun. Umgekehrt ist der relative Anteil an Tarifen mit einer Laufzeit von weniger als zehn Jahren bei den stornierten Tarifen wesentlich höher. Den größten Unterschied machen wir für jene Tarife aus, die vor weniger als fünf Jahren abgeschlossen wurden. Hier beobachten wir z.B. im Jahr 2014, dass der Anteil dieser vor Kurzem abgeschlossenen Tarife im gesamten Bestand annähernd drei Prozentpunkte weniger beträgt als in der Menge der stornierten Tarife.

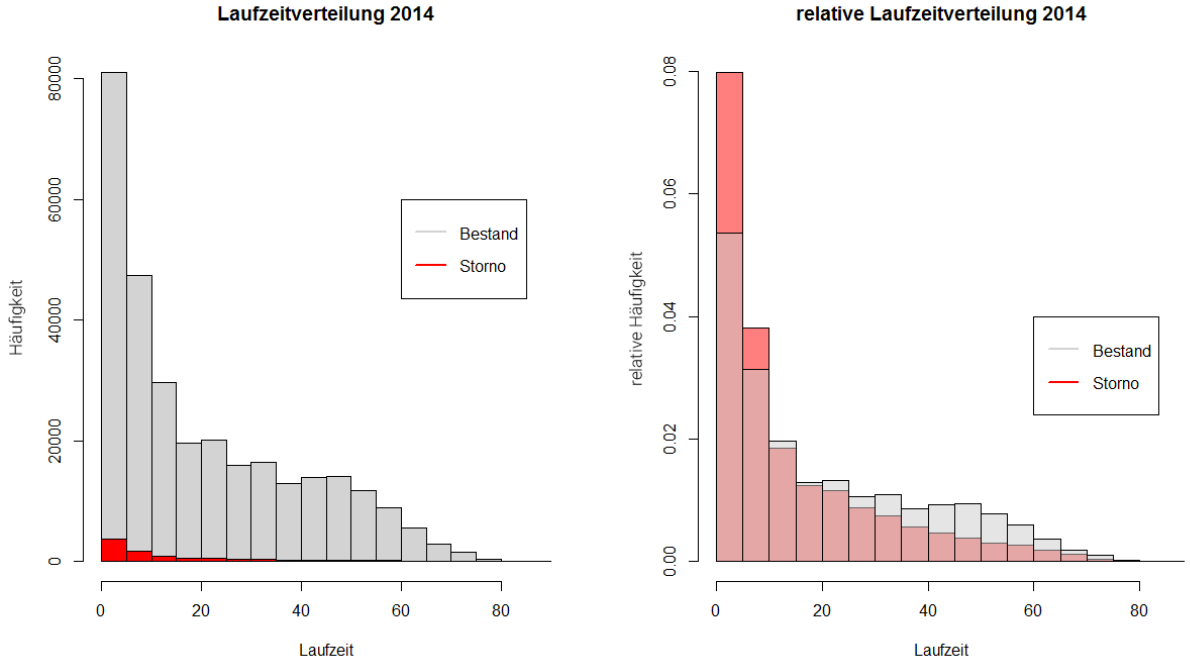

Abbildung 6: Histogramm der Laufzeit für Bestand und Storno im Jahre 2014

Nachdem wir also wieder festgestellt haben, dass die Laufzeit wohl einen zu beachtenden Einfluss auf die Stornowahrscheinlichkeit hat, wollen wir nun noch einen ersten Eindruck

von der Abhängigkeitsstruktur bekommen. In Abbildung 7 wurde der relative Anteil an Stornierungen für jede Laufzeit gegen ebendiese Laufzeit aufgetragen. Wir beobachten, dass bis zu einer Laufzeit von drei bis fünf Jahren der Anteil an Stornierung zunimmt. Für die darauffolgenden höheren Laufzeiten scheint es so, als würde dieser Anteil sukzessive abnehmen. Die Tarife mit einer sehr langen Laufzeit (mehr als 60 Jahre) wurden dabei wieder außer Acht gelassen, da wir hier wiederum nur sehr wenige Beobachtungen haben. Es scheint also auch für die Laufzeit so, als würde die Abhängigkeitsstruktur zwischen dem relativen Anteil an Stornierungen und der Laufzeit keinem linearen Zusammenhang folgen.

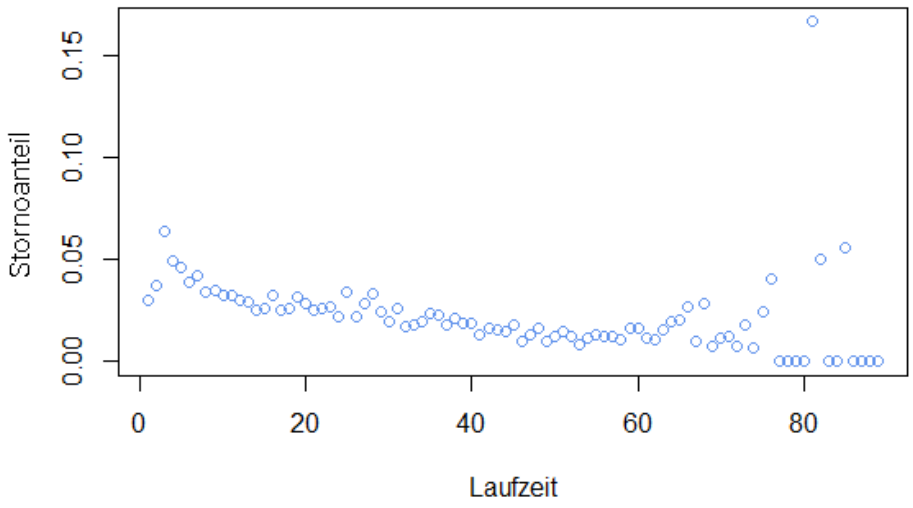

Abbildung 7: Abhängigkeit des Stornoanteils von der Laufzeit im Jahre 2014

2.2.4 Die Tarifklassen

Der gesamte Bestand an Krankenversicherungstarifen teilt sich auf vier unterschiedliche Tarifklassen auf. Da die genaue Art der Tarifklasse für unseren Anwendungszweck nicht relevant ist, belassen wir die Namen dieser Klassen bei ihren codierten Bezeichnungen „PG01", „PG02a", „PG02b" und „PG03". Die absoluten Häufigkeiten der einzelnen Tarifklassen, ersichtlich in Abbildung 8, geben einen ersten Überblick über diesen Faktor. Die Tarifklasse mit der höchsten Vorkommnis in allen Beobachtungsjahren ist demnach „PG02b". Hier beobachten wir jeweils mehr als 95.000 Tarife. Die kleinste Anzahl an Tarifen zählen wir mit etwa 60.000 Positionen in allen Jahren für „PG02a". Während „PG03" in den Jahren 2012 und 2013 mit jeweils knapp 70.000 Positionen noch die zweitmeisten Tarife umfasst, ist die zweitgrößte Tarifklasse im Jahr 2013 „PG01". Zusammengefasst ist also „PG02b" die mit Abstand größte und „PG02a" die kleinste Tarifklasse.

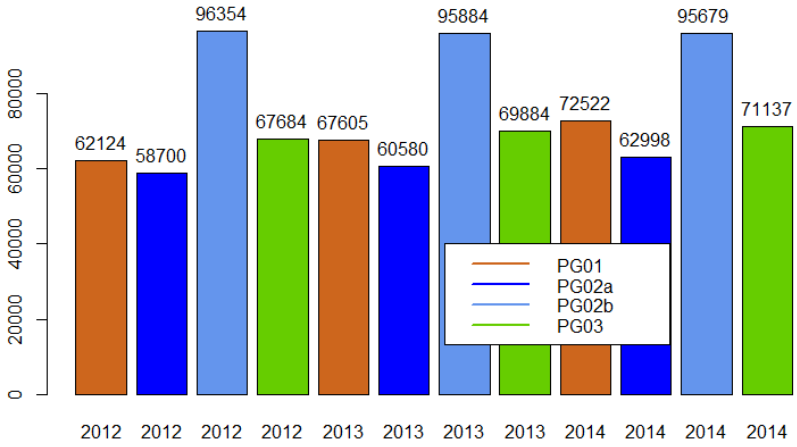

Abbildung 8: Vorkommnis der Tarife für alle Beobachtungsjahre

Um zu überprüfen, ob es bezüglich des Stornierungsverhaltens Unterschiede zwischen diesen Klassen gibt, berechnen wir für jede Klasse und jedes der drei Beobachtungsjahre den relativen Anteil an Stornierungen. Die entsprechenden Ergebnisse sind in Abbildung 9 ersichtlich. Hier sehen wir, dass die laut Abbildung 8 größte Tarifklasse „PG02b" in allen Jahren den geringsten relativen Anteil an stornierten Tarifen aufweist. Dieser liegt bei etwa 1.5%. Demgegenüber stehen die Tarifklassen „PG01" und „PG03". Hier beobachten wir in allen Jahren einen vergleichsweise hohen Anteil an Stornierungen von über 3%. Den höchsten Wert, mit mehr als 5% an stornierten Verträgen, gab es im Jahr 2014 in der Klasse „PG01". Während der Stornierungsanteil für „PG02b" und „PG03" über alle Jahre relativ konstant zu sein scheint, beobachten wir für die Tarifklasse „PG02a" einen deutlich abnehmenden Anteil. Dieser betrug im Jahr 2012 noch rund 4%. In den Folgejahren 2013 und 2014 waren es dann nur noch etwas mehr als 2% der „PG02a"-Tarife die storniert wurden.

Offensichtlich werden wir diese erklärende Variable der Tarifklasse in der Modellierung als Faktor betrachten. Die vier Stufen dieses Faktors ergeben sich aus den vier beobachteten Tarifklassen. Aufgrund der Beobachtungen, die wir anhand von Abbildung 9 gemacht haben, erwarten wir, dass dieser Faktor einen signifikanten Informationsgehalt über die Stornowahrscheinlichkeit beinhaltet. So ist zu erwarten, dass für Tarife der Tarifklasse „PG02b" eine verminderte Stornowahrscheinlichkeit geschätzt wird. Natürlich ist dies nur eine erste Vermutung. So wäre es z.B. möglich, dass die hier beobachteten Unterschiede zwischen den Tarifklassen durch die anderen Prädiktoren erklärt werden.

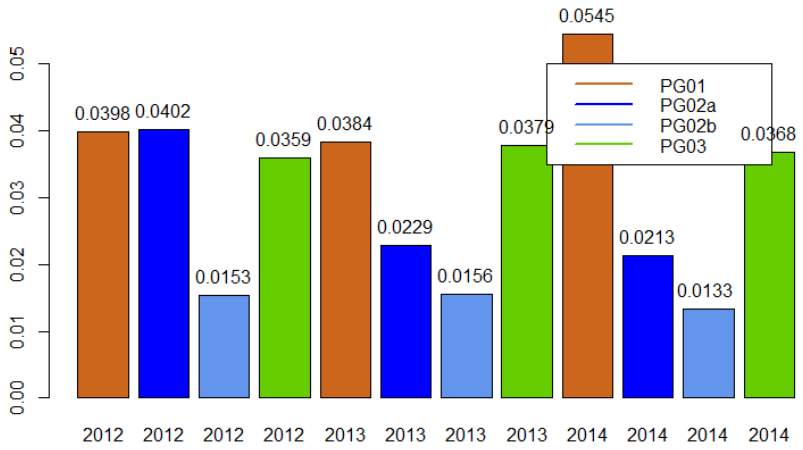

Abbildung 9: Relative Häufigkeit von Storno in den Tarifklassen

2.2.5 Tarife mit Rabatt

Die erklärende Variable Rabatt gibt an, ob die Prämie eines Tarifes in irgendeiner Form eine Rabattierung beinhaltet oder nicht. Es sei an dieser Stelle noch einmal erwähnt, dass wir auf Basis der uns zur Verfügung stehenden Daten weder wissen wie hoch dieser Rabatt ist noch warum dieser gewährt wurde. Somit unterteilt die Variable Rabatt den Bestand in zwei Teilmengen, jene Tarife mit und jene ohne Rabattierung. In Abbildung 10 sehen wir, dass die Vorkommnis von rabattierten Tarifen in allen Beobachtungsjahren ungefähr gleich groß ist. Dies sind in allen Kalenderjahren etwa 65.000 Tarifen, was einem Anteil von rund 28% entspricht.

Um einschätzen zu können, ob sich rabattierte Tarife in ihrer Stornowahrscheinlichkeit von nicht rabattierten Tarifen unterscheiden, berechnen wir für alle Beobachtungsjahre die jeweiligen relativen Anteile an Stornierungen. Die Ergebnisse sind Abbildung 11 zu entnehmen. Hierbei ist zu erkennen, dass im Jahr 2012 sowohl bei den rabattfreien wie auch bei den rabattierten Tarifen rund 3% des Bestandes storniert wurden. Anders ist die Situation in den darauffolgenden Jahren 2013 und 2014. In beiden Jahren beobachten wir, dass deutlich mehr als 3% der nicht rabattierten Tarife storniert wurden. Im Jahr 2013 betrug dieser Anteil 3.1% und im Jahr 2014 sogar annähernd 3.5%. Somit scheint es, als hätte die Wahrscheinlichkeit für eine Stornierung von Tarifen ohne Rabatt über diesen Beobachtungszeitraum von drei Jahren zugenommen. Bei den rabattierten Tarifen beobachten wir allerdings einen verminderten relativen Anteil an Stornierungen in den Jahren 2013 und 2014. Während dieser im Jahr 2012 noch bei rund 3% lag, liegt er 2013 und 2014 nunmehr bei etwas mehr als einem Prozentpunkt. Der Unterschied zu den rabattfreien Tarifen lässt vermuten, dass der zweistufige Faktor Rabatt tatsächlich einen zu beachtenden

Einfluss auf die Stornowahrscheinlichkeit hat. Auch beobachten wir hier einen Unterschied zwischen den Kalenderjahren. Während die Ergebnisse der Jahre 2013 und 2014 durchaus vergleichbar sind, weichen die Beobachtungen des Jahres 2012 doch sehr von den anderen Jahren ab.

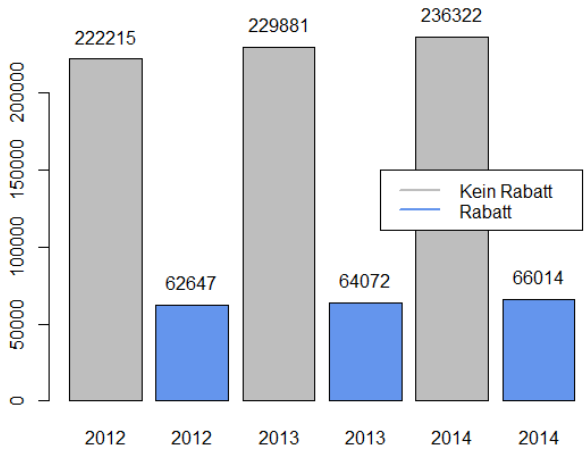

Abbildung 10: Vorkommnis von Rabatt in allen Beobachtungsjahren

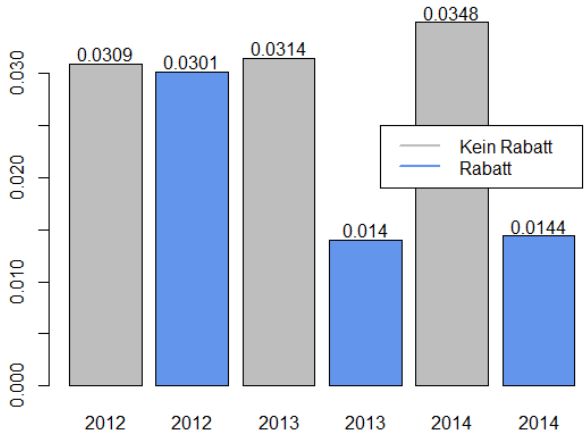

Abbildung 11: Relativer Stornoanteil von Tarifen mit/ohne Rabatt

2.2.6 Gruppenversicherungen

Im Bestand der Krankenversicherungsverträge kommt es vor, dass mehrere Verträge unter einer gemeinsamen Gruppenversicherung abgeschlossen wurden. Die genauen Details einer solchen Gruppenversicherung sind für unseren Anwendungszweck wiederum nicht relevant. Anhand der uns zur Verfügung stehenden Daten ist dabei lediglich zu erkennen, ob ein Tarif als Teil einer Gruppenversicherung abgeschlossen wurde oder nicht. Uns stehen daher weder Informationen über die Gruppengröße noch über die jeweils zur selben Gruppe gehörenden Tarife zur Verfügung. Eine solche Gruppe umfasst dabei alle Tarife der zugehörigen Verträge. Die erklärende Variable Gruppe teilt also wieder den Bestand in zwei Teilmengen, jenen Tarifen die als Teil einer Gruppenversicherung abgeschlossen wurden und jene ohne Gruppenzugehörigkeit. Um einen ersten Überblick über die Vorkommnis solcher Gruppenversicherungen zu erlangen, haben wir in Abbildung 12 die absolute Anzahl an Tarifen mit und ohne Gruppenzugehörigkeit dargestellt. Demnach scheint der Anteil an Tarifen, die zu einer Gruppenversicherung gehören, über die Jahre hinweg relativ stabil zu sein. Dabei gibt es deutlich mehr Tarife ohne Gruppenzugehörigkeit.

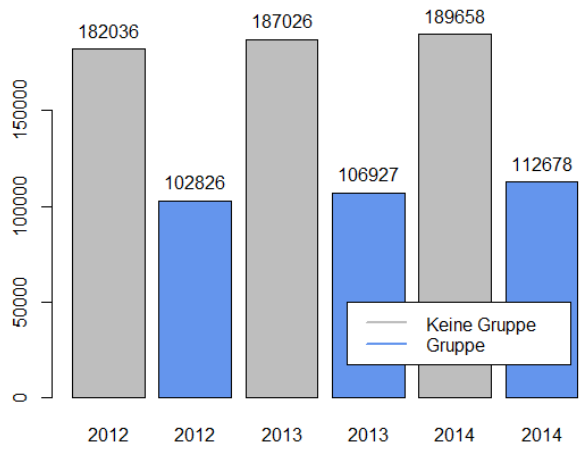

Abbildung 12: Vorkommnis von Gruppentarifen in allen Beobachtungsjahren

Wiederum sind wir an einer ersten Einschätzung des Effektes der Gruppenversicherung auf die Stornowahrscheinlichkeit interessiert. Dafür berechnen wir abermals den relativen Anteil an Stornierung für Tarife in einer Gruppe und solcher ohne Gruppenzugehörigkeit. Die entsprechenden Ergebnisse sind in Abbildung 13 dargestellt. Dabei beobachten wir, dass für das Jahr 2012 wohl kein sonderlicher Unterschied zwischen diesen beiden Teilmengen besteht. Sowohl für die Tarife mit Gruppe als auch für jene ohne Gruppe liegt der Anteil an Stornierungen hier bei rund 3%. Für die Jahre 2013 und 2014 sehen wir hingegen zwei unterschiedliche Effekte. Während im Jahre 2013 anteilsmäßig weniger Gruppentarife stor-

niert wurden, ist dies für das Jahr 2014 genau umgekehrt. Der Unterschied beträgt dabei jeweils mehr als einen Prozentpunkt. Die relativen Anteile an Stornierungen für Tarife mit und ohne Gruppenzugehörigkeit waren also in einem Jahr gleichgroß und in den beiden anderen Jahren genau entgegengesetzt. Auf Basis dieser Beobachtung scheint uns der Faktor Gruppe als nicht sonderlich relevant, da die hier gemachten Beobachtungen eher auf ein zufälliges Muster hinweisen.

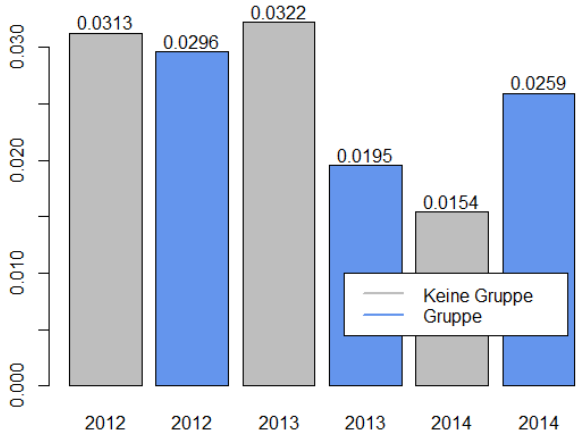

Abbildung 13: Relativer Stornoanteil von Tarifen in/ohne Gruppe

Für alle drei Beobachtungsjahre gilt, dass ausschließlich jene Tarifpositionen, die als Teil einer Gruppenversicherung abgeschlossen wurden, auch einen Rabatt beinhalten. Es existiert in dem uns zur Verfügung stehenden Datenbestand also kein Tarif, dessen Prämie eine Rabattierung beinhaltet und der nicht Teil einer Gruppenversicherung ist. Jedoch besitzt nicht jeder Tarif, der Teil einer Gruppe ist, auch automatisch eine Rabattierung. Die Interaktion zwischen Gruppe und Rabatt hat demnach nur drei Ausprägungen. Somit ist es nicht notwendig, in einem Modell, welches bereits Gruppe und Rabatt als Faktoren beinhalten, zusätzlich die Interaktion zwischen diesen beiden Faktoren zu berücksichtigen (der etwaige Koeffizient würde immer als Null geschätzt werden).

2.2.7 Das Geschlecht des Versicherungsnehmers

Die letzte uns zur Verfügung stehende erklärende Variable ist das Geschlecht des Versicherungsnehmers. Von vornherein scheint es eher unwahrscheinlich, dass dieser Prädiktor einen signifikanten Einfluss auf die Stornowahrscheinlichkeit hat. Zudem dürfte, auf Grund aktueller Gesetzesänderungen zur Geschlechterneutralität, ein solcher Effekt auch nicht in die Berechnung der Versicherungsprämie miteinbezogen werden. Als ersten Schritt wollen wir uns wieder die Vorkommnis von weiblichen und männlichen Versicherungsnehmern

betrachten. Anhand von Abbildung 14 erkennen wir, dass es in allen Beobachtungsjahren mehr Tarife mit einem weiblichen Versicherungsnehmer gibt. Dieses Verhältnis scheint über alle Beobachtungsjahre hinweg relativ stabil zu sein.

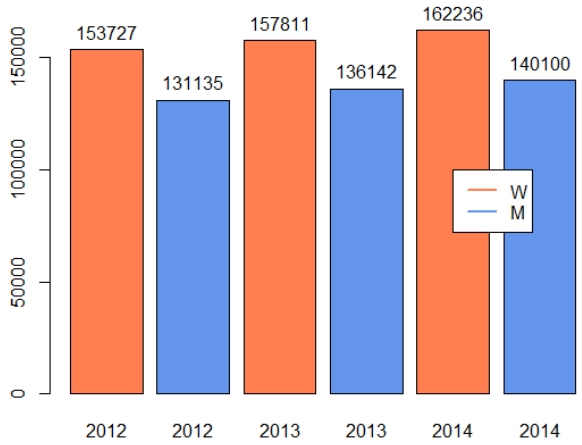

Abbildung 14: Geschlechterverteilung in den Beobachtungsjahren

Wiederum berechnen wir für jedes Beobachtungsjahr sowohl den relativen Anteil an von Frauen als auch von Männern stornierten Tarifen. Die Ergebnisse sind in Abbildung 15 ersichtlich. Auf den ersten Blick sind diese relativen Anteile in allen Jahren in etwa gleich groß. Erst bei genauerem Hinsehen fällt auf, dass in allen Jahren der relative Anteil an von Männern stornierten Tarifen etwas geringer zu sein scheint. Diese Unterschiede spielen sich aber lediglich im Promillebereich ab. Somit folgern wir aus dieser ersten groben Analyse, dass zwar Muster im Unterschied zwischen den Stornowahrscheinlichkeiten von Männern und Frauen erkennbar, jedoch sehr klein sind. Demnach scheint es naheliegend, dass das Geschlecht in der späteren Modellierung der Stornowahrscheinlichkeit wohl keine allzu große Rolle spielen wird.

2.3 Die Wahl der Modellklasse

Auf Basis der Anforderungen an unser Modell und den uns zur Verfügung stehenden Daten wollen wir zum Abschluss dieses Kapitels unsere Wahl der Modellklasse motivieren. Wie bereits in der Einleitung erwähnt, ist es das Ziel mit Hilfe eines Regressionsmodells die Stornowahrscheinlichkeit der Daten eines einzelnen Kalenderjahres zu modellieren und so die verschiedenen Beobachtungsjahre zu vergleichen. Ebenfalls sind wir daran interessiert, die Abhängigkeitsstruktur zwischen der Wahrscheinlichkeit für eine Stornierung und

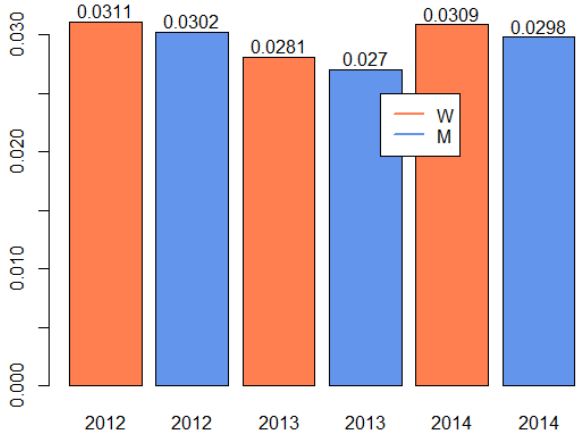

Abbildung 15: Relativer Stornoanteil von Tarifen nach Geschlecht des Versicherungsnehmers

den erklärenden Variable zu beschreiben. Da wir demnach Wahrscheinlichkeiten modellieren, liegt es nahe dies mit der Modellklasse der Generalisierten Linearen Modelle (GLMe) zu tun. Im Speziellen sind dabei natürlich logistische Regressionsmodelle gemeint. Mit Hilfe dieser Modellklasse wäre gewährleistet, dass die geschätzten Wahrscheinlichkeiten tatsächlich Wahrscheinlichkeiten, also Zahlenwerte im offenen Intervall $(0, 1)$, sind.

Fassen wir die Ergebnisse aus Abschnitt 2.2 noch einmal zusammen, so sehen wir, dass uns mit Alter, Laufzeit und Prämie drei „stetige" und mit Rabatt, Tarifklasse, Geschlecht und Gruppe vier faktorielle Prädiktoren zur Verfügung stehen. Dabei werden wir Alter und Laufzeit als stetige Prädiktoren betrachten obwohl ihre Ausprägungen nur natürliche Zahlen annehmen. Der Grund dafür ist, dass eine Interpretation als Faktoren zu einer zu großen Menge an Faktorstufen führen würde. Auf der Seite der faktoriellen Prädiktoren gibt es mit Gruppe und Geschlecht zwei erklärende Variablen, an deren Relevanz bereits die erste groben Analyse im vorherigen Abschnitt 2.2 zweifeln lässt. So gab es zwischen den relativen Anteilen an Stornierungen von Männern und Frauen nur einen sehr kleinen Unterschied. Wir haben ebenfalls gesehen, dass die Abhängigkeit zwischen dem relativen Anteil an Stornierungen und je einem der drei stetigen Prädiktoren wohl nicht von linearer Natur ist. Aus diesem Grund entscheiden wir uns für ein logistisches Generalisiertes Additives Modell. Diese Modellklasse stellt in gewisser Hinsicht eine Erweiterung zu den GLMen dar. Der für unseren Anwendungsfall maßgeblich Vorteil ist, dass in einem GAM auch Abhängigkeitsstrukturen modelliert werden können die keinem linearen Zusammenhang folgen. Durch die Möglichkeit der grafischen Darstellung dieser geschätzten nichtlinearen Abhängigkeitsstrukturen können diese einfach interpretiert und verglichen werden. Die Klasse der GAMe wird im nun folgenden Kapitel vorgestellt.

3 Generalisierte Additive Modelle

In diesem Kapitel werden wir die Klasse der Generalisierten Additiven Modelle (GAMe) vorstellen. Dafür rufen wir uns in Abschnitt 3.1 zunächst die Klassen der Linearen Modelle und der Generalisierten Linearen Modelle (GLMe) in Erinnerung. Aufbauend auf diesen Modellklassen erfolgt in Abschnitt 3.2 eine kurze Einführung in die Unterschiede zwischen Generalisierten Linearen und Generalisierten Additiven Modellen. Nachdem wir uns so über die Vorteile der GAMe bewusst werden, präsentieren wir im restlichen Teil dieses Kapitels eine kurze Herleitung dieser Modellklasse. Der zentrale Bestandteil eines GAMes sind die sogenannten Glatten Funktionen. In Abschnitt 3.3 erläutern wir diesen Begriff und gehen näher auf die Darstellung und den Grad der Glättung solcher Glatter Funktionen ein. Mit Hilfe dieser Funktionen ist es uns in Abschnitt 3.4 möglich sogenannte Additive Modelle zu definieren. Durch eine zusätzliche Verallgemeinerung erhalten wir aus diesen Additiven Modellen in Abschnitt 3.5 die Klasse der Generalisierten Additiven Modelle. Da sich unsere Arbeit mit einem konkreten Anwendungsfall dieser Modelle befasst, präsentieren wir im abschließenden Abschnitt 3.6 die Handhabung von GAMen in der von uns verwendeten Statistiksoftware R.

3.1 Motivation

Bevor wir die Klasse der GAMe genauer betrachten, soll an dieser Stelle noch einmal erklärt werden, warum wir uns für diese (vergleichsweise) komplexe Modellklasse entschieden haben. Dafür betrachten wir im Folgenden Lineare Modelle und Generalisierte Lineare Modelle. In den Abschnitten 3.1.1 und 3.1.2 werden wir die zentralen Aspekte dieser beiden Modellklassen kurz wiederholen und die Frage beantworten, warum diese für unsere Problemstellung nicht geeignet sind.

3.1.1 Lineare Modelle

Die einfachste Möglichkeit die Stornowahrscheinlichkeit in Abhängigkeit von verschiedenen Prädiktoren zu modellieren bieten Lineare Modelle, wie sie u.a. in Andres (1986) beschrieben werden. Formal ausgedrückt haben wir n Beobachtungen x_i und y_i, wobei y_i Realisationen der unabhängigen Zufallsvariablen Y_i und x_i (vektorwertig) die Prädiktoren darstellen. Der Erwartungswert dieser Zufallsvariablen Y_i sei gegeben durch $\mathbb{E}(Y_i) = \mu_i$. Die einem Linearen Modell zugrunde liegende Annahme ist, dass zwischen x und y folgender (linearer) Zusammenhang besteht:

$$Y_i = \mu_i + \epsilon_i = x_i'\beta + \epsilon_i. \tag{1}$$

Hierbei sind die ϵ_i's unabhängige und identisch normalverteilte Zufallsfehler mit $\mathbb{E}(\epsilon_i) = 0$ und $\mathbb{V}ar(\epsilon_i) = \sigma^2$. Die unbekannten (konstanten) Parameter dieses Modells sind β und σ^2. Somit ergibt sich aus obiger Modellannahme (1) die sogenannte Regressionsfunktion als

$$\mathbb{E}(Y_i) = \mathbb{E}(x_i'\beta + \epsilon_i) = x_i'\beta. \tag{2}$$

Mit einem geeigneten Schätzer für β können wir folglich die Erwartungswertfunktion der Y_i in Abhängigkeit der Prädiktoren x_i (linear) modellieren. Eine naheliegende Wahl für einen solchen Parameterschätzer ist diesen so zu wählen, dass das Modell möglichst gut die beobachteten Daten wiedergibt. Eine Möglichkeit dafür ist die Summe der Fehlerquadrate, die durch

$$SSE(\beta) = \sum_{i=1}^{n} (y_i - x_i'\beta)^2 \tag{3}$$

gegeben ist, zu minimieren. Diese Zielfunktion setzt sich also aus der Summe der quadratischen Abweichungen zwischen den beobachteten Responsevariablen und deren Modellschätzungen zusammen. Der kleinste Quadrate Schätzer $\hat{\beta}$, also die Optimallösung der Minimierung von $SSE(\beta)$, kann dabei analytisch bestimmt werden.

Für unseren Anwendungszweck (die Modellierung von Wahrscheinlichkeiten) ist dieses Modell jedoch nicht geeignet. Der schwerwiegendste Grund dafür ist, dass die durch ein Lineares Modell modellierten Stornowahrscheinlichkeiten im Allgemeinen nicht im Intervall $[0, 1]$ liegen und somit für unseren Anwendungsfall nicht von Gebrauch sind. Eine Möglichkeit diesem Problem beizukommen stellt die Klasse der GLMe dar.

3.1.2 Generalisierte Lineare Modelle

Wir wollen uns hier mit der Klasse der Generalisierten Linearen Modelle und dabei insbesondere mit Logistischen Regressionsmodellen beschäftigen. Mit Hilfe dieser Modellklasse können wir gewährleisten, dass die modellierten Wahrscheinlichkeiten tatsächlich im Intervall $[0, 1]$ liegen. Die nun folgenden Ausführungen basieren dabei lose auf McCullagh und Nelder (1989).

Ein GLM ist unter Annahme der Existenz von $\mathbb{E}(y_i) = \mu_i$ und $\mathbb{V}ar(y_i)$ durch folgende Komponenten gegeben:

- stochastische Komponente: $y_i \overset{ind}{\sim} \text{LEF}(\theta_i)$, $\mathbb{E}(y_i) = \mu_i = \mu(\theta_i)$
- systematische Komponente: $\eta_i = x_i'\beta$
- Linkfunktion: $g(\mu_i) = \eta_i$.

Hierbei beschreibt $\text{LEF}(\cdot)$ die einparametrige Lineare Exponentialfamilie von Verteilungsfunktionen. Diese Familie beinhaltet alle Verteilungen deren Dichtefunktion $f(\cdot)$ sich in Abhängigkeit eines Parameters θ als

$$f(y|\theta) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right) \tag{4}$$

darstellen lässt, wobei ϕ eine feste Größe ist und $a(\cdot)$, $b(\cdot)$ und $c(\cdot)$ bekannte Funktionen mit $a(\phi) > 0$ sind. Beispiele für Mitglieder der Linearen Exponentialfamilie sind u.a. die

Normalverteilung, die Poissonverteilung und auch die standardisierte Binomialverteilung. Die weiteren Komponenten des obigen GLMs sind die Linkfunktion, eine monotone und zweimal stetig differenzierbare Funktion, sowie der Lineare Prädiktor η_i.

Wird als Verteilung die Normalverteilung aus der LEF und als Linkfunktion die identische Abbildung gewählt, so stimmt das resultierende GLM mit einem Linearen Modell überein und kann, wie in Abschnitt 3.1.1 erwähnt, analytisch geschätzt werden. Für alle anderen GLMe kann die Parameterschätzung im Allgemeinen nur noch durch iterative Methoden vorgenommen werden.

Der für uns wesentliche Vorteil der GLMe gegenüber den weiter oben vorgestellten Linearen Modellen ist, dass eine Funktion des Erwartungswertes, und nicht der Erwartungswert selbst, linear modelliert wird. Dies ermöglicht uns, durch eine geeignete Wahl der Verteilung und der Linkfunktion, $\hat{\mu}_i = g^{-1}(\hat{\eta}_i) \in (0, 1)$ zu gewährleisten.

Die naheliegendste Wahl ist dabei das Vorkommen von Stornierungen durch eine Bernoulli-Zufallsvariablen zu modellieren. Eine Bernoulli-Zufallsvariable kann natürlich auch als binomialverteilte Zufallsvariable mit einem Versuch interpretiert werden. Durch diese Wahl der binomialen Verteilungsfunktion aus der Lineare Exponentialfamilie ergibt sich für die binären Responsevariablen y_i (Storno oder kein Storno) folgende Erwartungswert- und Varianzstruktur

$$\begin{aligned} \mathbb{E}(y_i) &= \frac{\exp(\theta_i)}{1 + \exp(\theta_i)} = \mu_i \\ \mathbb{V}ar(y_i) &= \mu_i(1 - \mu_i), \end{aligned}$$

wobei also $y_i \overset{ind}{\sim} \text{Binomial}(1, \mu_i)$ die zugrunde liegende Verteilungsannahme ist.

Als Linkfunktion wählen wir den Logitlink. Dieser stellt die sogenannte kanonische Linkfunktion zur Binomialverteilung dar, d.h. dass durch den Linearen Prädiktor η direkt der Verteilungsparameter θ modelliert wird, also $g(\mu) = \theta$. Die kanonische Linkfunktion zur Binomialverteilung ist gegeben als

$$\text{logit}(\mu) = \log\left(\frac{\mu}{1 - \mu}\right) = \eta = \theta. \tag{5}$$

Damit ergibt sich die Abhängigkeitsstruktur zwischen der zu schätzende Wahrscheinlichkeit μ und dem Linearen Prädiktor η als

$$\mu = \frac{\exp(\eta)}{1 + \exp(\eta)}. \tag{6}$$

Da $\exp(\eta) > 0$, folgt aus obiger Darstellung (6) für die Modellschätzungen $\hat{\mu}$, dass sie tatsächlich im Intervall $(0, 1)$ liegen. Somit haben wir durch die Wahl der Binomialverteilung und des Logitlinks für unser GLM sichergestellt, dass es sich bei den modellierten

Stornowahrscheinlichkeiten tatsächlich um Wahrscheinlichkeiten handelt. Der Bildbereich der Modellschätzung war allerdings nicht der einzige Kritikpunkt bei der Anwendung eines Linearen Modells. Auch die Art der modellierten Abhängigkeitsstruktur zwischen Response und Prädiktoren spielt für unseren Anwendungszweck eine bedeutende Rolle.

In Abschnitt 2.2 haben wir die dieser Arbeit zugrunde liegenden Daten vorgestellt. Als Prädiktoren stehen uns demnach mit dem Alter des Versicherungsnehmers, der Laufzeit des Versicherungsvertrages und der monatlich zu zahlenden Prämie drei stetige Variablen zur Verfügung. Indem wir, jeweils einzeln, den relativen Anteil an Stornierungen gegen diese Variablen geplottet haben, kamen wir zu dem Schluss, dass die Abhängigkeitsstrukturen wohl nicht von linearer Natur sind. In einem, wie in diesem Abschnitt definierten, GLM sehen wir allerdings, dass die Prädiktoren in der systematischen Komponente nur linear vorkommen. Dies bedeutet, dass wir mit einem logistischen Regressionsmodell zwar tatsächlich Wahrscheinlichkeiten modellieren können, wir jedoch nicht in der Lage sind einen nicht linearen Zusammenhang zwischen Response- und Prädiktorvariablen abzubilden.

Aus diesem Grund wenden wir uns im weiteren Verlauf dieses Kapitels den Generalisierten Additiven Modellen zu. Diese Modellklasse, die wir schlussendlich zur Modellierung der Stornowahrscheinlichkeiten benutzen werden, ermöglicht es uns auch Abhängigkeiten die keiner linearen Struktur folgen in einem GLM im Allgemeinen aber auch in einem logistischen Regressionsmodell im Speziellen abzubilden.

3.2 Einführung in Glatte Modelle

In diesem Abschnitt geben wir einen kurzen Einblick in die Klasse der GAMe. Dies stellt lediglich einen ersten Überblick über diese Modellklasse dar und soll die in den folgenden Abschnitten präsentierte Herleitung motivieren. Eine detailliertere Einführung in die Klasse der GAMe erfolgt dann in Abschnitt 3.5. Die in diesem und in den nächsten Abschnitten folgenden Ausführungen basieren dabei vor allem auf Wood (2006).

Ein GAM ist ein GLM dessen Linearer Prädiktor zusätzlich sogenannte Glatte Funktionen der Prädiktorvariablen beinhaltet. Aus diesem Grund setzt sich ein GAM aus denselben Komponenten wie ein GLM zusammen, d.h.

- stochastische Komponente: $y_i \overset{ind}{\sim} \text{LEF}(\theta_i)$, $\mathbb{E}(y_i) = \mu_i = \mu(\theta_i)$
- systematische Komponente: $\eta_i = x_i'\beta + f_1(x_{1i}) + f_2(x_{2i}) + \ldots$
- Linkfunktion: $g(\mu_i) = \eta_i$.

Der einzige Unterschied ist nun, dass die systematische Komponente (der Lineare Prädiktor) nun nicht mehr nur Prädiktor- mal Parametervektor enthält sondern zusätzlich auch eine Summe von Funktionen die in den Prädiktoren ausgewertet werden. Dies ermöglicht es,

komplexe Abhängigkeitsstrukturen zwischen Responsevariable und Prädiktoren zu modellieren. Wie wir in Abschnitt 2.2 gesehen haben, scheint die Abhängigkeitsstruktur zwischen den stetigen Prädiktoren Alter, Laufzeit und Prämie und den relativen Anteilen an Stornierungen keinem linearen Zusammenhang zu folgen. Somit bietet sich für diese Prädiktoren eine Modellierung mithilfe der hier vorgestellten Methoden an. Um dabei die Berechenbarkeit des Modells zu gewährleisten wird die Wahl dieser zusätzlichen Funktionen auf die Klasse der Glatten Funktionen, wie wir sie im folgenden Abschnitt präsentieren werden, eingeschränkt. Trotz dieser Einschränkung gewinnen GAMe gegenüber GLMe an zusätzlicher Flexibilität. Diese zusätzliche Flexibilität führt allerdings auch dazu, dass in der Modellschätzung zwei weitere theoretische Aspekte zu betrachten sind. Zum einen müssen diese Glatten Funktionen dargestellt werden. Dafür gilt es festzulegen, was Glatte Funktionen sind und wie diese erzeugt werden können. Zum anderen müssen wir bestimmen, wie glatt diese Funktionen tatsächlich sind und was „Glätte" in diesem Zusammenhang überhaupt bedeutet. Auf diese beiden Aspekte wird im nun folgenden Abschnitt 3.3 „Glatte Funktionen" näher eingegangen. Auf Basis dieser Ergebnisse können wir dann in Abschnitt 3.4 ein erstes sogenanntes Additives Modell definieren. Aus diesen Additiven Modellen resultieren in späterer Folge die Generalisierten Additiven Modelle.

3.3 Glatte Funktionen

Wie bereits erwähnt, spielen die sogenannten Glatten Funktionen eine zentrale Rolle für die Klasse der GAMe. Anstatt die im Linearen Prädiktor vorkommenden Funktionen beliebig zu wählen, wird der Raum der möglichen Funktionen eingeschränkt. Dieses Vorgehen hat mehrere Gründe. So lässt sich dadurch die Modellkomplexität besser kontrollieren was natürlich einen direkten Einfluss auf den zur Modellschätzung erforderlichen Rechenaufwand hat. Außerdem kann durch den Grad der Glättung auch die Anpassung eines Modells an die beobachteten Daten beeinflusst werden.

Zum Zwecke der Einführung in den Begriff der Glatten Funktionen wählen wir in der Darstellung eines GAMs (wie in Abschnitt 3.2) die Linkfunktion $g(\cdot)$ als die identische Abbildung und die Verteilungsfunktion als die Normalverteilung. Des weiteren nehmen wir vorerst an, dass sich der Lineare Prädiktor lediglich aus einer Glatten Funktion eines Prädiktors zusammensetzt. Somit ergibt sich als Modell für die Responsevariablen

$$y_i = f(x_i) + \epsilon_i. \tag{7}$$

Hierbei beschreiben die ϵ_i wiederum unabhängig identisch normalverteilte Zufallsvariablen, also $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$. Des weiteren wollen wir an dieser Stelle der Einfachheit halber annehmen, dass die Prädiktoren alle aus dem Intervall $[0, 1]$ stammen.

Unser Ziel ist es nun, diese Glatte Funktion $f(\cdot)$ so darzustellen, dass wir Modell (7) mit den bereits bekannten Methoden, wie sie für GLMe verwendet werden, schätzen können. Solche Darstellungsformen werden in Abschnitt 3.3.1 vorgestellt. Sobald wir Glatte Funktionen

in geeigneter Form darstellen können, gilt es die Glättung dieser Funktion an die Daten anzupassen. Auf diesen Schritt gehen wir in Abschnitt 3.3.2 näher ein.

3.3.1 Die Darstellung Glatter Funtionen

Damit wir ein GAM mit denselben (iterativen) Methoden wie ein GLM schätzen können, müssen wir die im Linearen Prädiktor vorkommenden Glatten Funktionen so darstellen, dass aus der systematischen Komponente wieder eine lineare Funktion der Prädiktoren wird. Die Vorgehensweise dafür wollen wir am vereinfachten Modell (7) präsentieren.

Dabei ist die grundlegende Idee, die Funktion $f(\cdot)$ durch fix definierte Basisfunktionen darzustellen. Diese Basisfunktionen erzeugen einen Raum von Funktionen aus welchem wir eine Glatte Funktion, die entweder $f(\cdot)$ entspricht oder eine Approximation davon ist, wählen. Bezeichnen wir mit $b_j(\cdot)$ für $j = 1, \ldots, m$ die $m \in \mathbb{N}$ Basisfunktionen eines solchen Raumes von Funktionen, so können wir $f(\cdot)$ darstellen als

$$f(x) = \sum_{j=1}^{m} b_j(x)\beta_j. \tag{8}$$

In dieser Darstellung werden die Basiskoeffizienten mit β_j bezeichnet. Gegeben den Basisfunktionen $b_j(\cdot)$, ist somit eine Glatte Funktion eindeutig durch diese Koeffizienten β_j definiert. Mit Hilfe dieser Repräsentation von $f(\cdot)$ können wir Modell (7) darstellen als

$$y_i = \sum_{j=1}^{m} b_j(x_i)\beta_j + \epsilon_i. \tag{9}$$

Dies entspricht aber einem Linearen Modell wie wir es bereits in Abschnitt 3.1.1 kennen gelernt haben. Die zu schätzenden Modellparameter β_j sind nun nicht die Koeffizienten der entsprechenden Prädiktoren sondern gehören zu der jeweiligen Basisfunktion die in diesen Prädiktoren ausgewertet wurde.

Für die Wahl der Basisfunktionen gibt es natürlich viele Möglichkeiten. Als geeignete Klasse von Funktionen haben sich dabei vor allem die sogenannten Splines erwiesen, wie sie in de Boor (1978) und Nürnberger (1989) beschrieben werden. Da wir im weiteren Verlauf der Modellierung von Stornowahrscheinlichkeiten kubische Splines verwenden werden, wollen wir hier näher auf diese spezielle Klasse von Splines eingehen. Dafür gehen wir wieder von unserem vereinfachten Modell aus, d.h. zu jedem Prädiktor x_i haben wir eine Beobachtung der Responsvariable y_i. Ein kubischer Spline ist im Allgemeinen eine Funktion, welche sich stückweise aus kubischen Polynomen zusammensetzt. Die Polynome sind dabei so gewählt, dass am stetigen Übergang zwischen zwei benachbarten Polynomen die ersten beiden Ableitungen dieser Polynome übereinstimmen. Diese Übergangsstellen oder Berührungspunkte werden im weiteren Verlauf als Knoten bezeichnet.

Würden wir unsere Punktepaare $\{x_i, y_i\}$ durch einen kubischen Spline interpolieren, würden wir als Knotenpunkte die Werte $\{x_i\}$ wählen. Da wir allerdings eine kubische Splinebasis definieren wollen, werden diese Knoten üblicherweise äquidistant zwischen $\min_i x_i$ und $\max_i x_i$ oder an bestimmten Quantilen der Wertemenge $\{x_i\}$ verteilt. Die Menge dieser Knotenpunkte sei gegeben als $\{\bar{x}_1, \ldots, \bar{x}_{m-2}\}$, wobei m wie oben die Dimension der Splinebasis definiert.

Gegeben einer Menge von Knotenpunkten gibt es viele äquivalente Möglichkeiten eine kubische Spline-Basis zu wählen. Ein Beispiel dafür wäre etwa die folgenden Basisfunktionen, wie sie in Gu (2002) beschrieben werden:

$$
\begin{aligned}
& b_1(x) = 1,\ b_2(x) = x \\
& b_j(x) = \frac{1}{4}\left[(\bar{x}_{j-2} - 1/2)^2 - 1/12\right] \cdot \left[(x - 1/2)^2 - 1/12\right] \\
& \qquad - \frac{1}{24}\left[(|x - \bar{x}_{j-2}| - 1/2)^4 - 1/2(|x - \bar{x}_{j-2}| - 1/2)^2 + 7/240\right] \\
& \qquad \text{für } j = 2, \ldots, m.
\end{aligned}
$$

Mit Hilfe der Darstellung von $f(\cdot)$ mit diesen Basisfunktionen kann nun Modell (7) wie ein Lineares Modell geschätzt werden.

Offen bleibt, wie die bis zu diesem Zeitpunkt allgemein definierte Dimension m der Basisfunktionen gewählt wird. Die Wahl dieser Dimension beeinflusst offensichtlich direkt wie glatt die aus diesen Basisfunktionen erzeugten Funktionen sein können. Somit wird durch die Wahl von m die Anpassung der Modellschätzungen an die beobachteten Daten festgelegt. Da dies ein zentraler Aspekt bei der Modellierung der Abhängigkeitsstruktur zwischen einem Prädiktor und der Responsevariable darstellt, wollen wir uns im folgenden Abschnitt genauer mit dem Grad der Glättung beschäftigen.

3.3.2 Der Grad der Glättung

Die bevorzugte Möglichkeit den Grad der Glättung zu beeinflussen ist nicht die Basisdimension m zu verändern sondern mit Hilfe eines Strafterms in der Parameterschätzung die effektive Glätte der Funktion zu beeinflussen. In anderen Worten bedeutet dies, dass wir zwar theoretisch sehr unglatte Funktionen zulassen, die Parameterschätzung jedoch so anpassen, dass glattere Funktionen bevorzugt werden. Zu diesem Zwecke wählen wir die Basisdimension m etwas größer als es voraussichtlich notwendig ist und fügen zur Zielfunktion der Parameterschätzung einen zusätzlichen Glätte-Term hinzu. Fassen wir die Basiskoeffizienten β_j zu einem Vektor der Dimension m zusammen, d.h.

$$\beta = (\beta_1, \ldots, \beta_m)',$$

so könnte z.B. für unser Lineares Modell (7) die Zielfunktion der Parameterschätzung anstatt, wie in Abschnitt 3.1.1 beschrieben,

$$\|y - X\beta\|^2 = \sum_{i=1}^{n} (y_i - x_i'\beta)^2, \tag{10}$$

nun durch einen Strafterm erweitert werden

$$\|y - X\beta\|^2 + \lambda \int_0^1 [f''(x)]^2 dx. \tag{11}$$

Mit einem Strafterm dieser Form wird also die Glätte einer Funktion anhand des integrierten Quadrats der zweiten Ableitung bestimmt. Zusätzlich wird dieser Term, im englischen auch *roughness penalty* genannt, durch einen frei wählbaren Parameter $\lambda \in \mathbb{R}_+$ gewichtet. Es ist leicht zu erkennen, dass für $\lambda \to \infty$ die zweite Ableitung der Funktion $f(\cdot)$ identisch Null sein muss. Dies entspricht einer linearen Funktion. Demgegenüber kann für eine Wahl von λ nahe bei Null die Funktion $f(\cdot)$ beliebig unglatt sein.

Da $f(\cdot)$ als Glatte Funktion linear in den Parametern β ist, kann der oben eingeführte Strafterm als Quadratische Form in β geschrieben werden, d.h.

$$\int_0^1 [f''(x)]^2 dx = \beta' S \beta, \tag{12}$$

wobei S eine quadratische Matrix mit bekannten Koeffizienten ist. So sind z.B. für die in Abschnitt 3.3.1 eingeführten Basisfunktionen für kubische Splines diese Koeffizienten gegeben als

$$\begin{aligned} S_{i+2,j+2} = &\frac{1}{4}\left[(\bar{x}_j - 1/2)^2 - 1/12\right] \cdot \left[(\bar{x}_i - 1/2)^2 - 1/12\right] \\ &- \frac{1}{24}\left[(|\bar{x}_i - \bar{x}_j| - 1/2)^4 - 1/2(|\bar{x}_i - \bar{x}_j| - 1/2)^2 + 7/240\right] \\ &\text{für } i, j = 1, \ldots, m-2 \end{aligned}$$

und $S_{i,j} = 0$ sonst (siehe Gu, 2002).

Mit Hilfe dieser Darstellung des Strafterms kann die Zielfunktion der Parameterschätzung nun zu

$$\|y - X\beta\|^2 + \lambda \beta' S \beta \tag{13}$$

umgeschrieben werden. Somit lässt sich der Grad der Glätte nun alleinig über die Wahl des Parameters λ steuern. Die klassische Methode zur Bestimmung eines optimalen λ ist die sogenannte Kreuzvalidierung (siehe Picard, 1984), deren grundlegende Idee wir im Folgenden präsentieren wollen.

Unsere Bestrebung ist es λ so zu wählen, dass die resultierende (geschätzte) Funktion $\hat{f}(\cdot)$ möglichst gut mit der (unbekannten) wahren Funktion $f(\cdot)$ übereinstimmt. Eine mögliche Maßzahl dafür ist der sogenannte Kreuzvalidierungswert, den wir als

$$\nu_o(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(\hat{\mu}_i^{[-i]} - y_i)^2 \tag{14}$$

definieren. Hierbei bezeichnen wir mit $\hat{\mu}_i^{[-i]} = \hat{f}^{[-i]}(x_i)$ jene Schätzung von $f(x_i) = \mathbb{E}(y_i)$, die wir von dem Modell erhalten, welches ohne die Beobachtung $\{x_i,\, y_i\}$ geschätzt wurde. Wir vergleichen also für jedes Punktepaar $\{x_i, y_i\}$ wie gut ein Modell, welches ohne die Information von x_i geschätzt wurde, y_i erklären kann und ermitteln dann den Mittelwert über alle $i = 1, \ldots, n$. Somit erhalten wir für jede Wahl von λ eine Maßzahl dafür, wie gut das resultierende Modell zu neuen Daten passt.

Eine Berechnung von $\nu_o(\lambda)$ in obiger Darstellung ist allerdings sehr aufwendig. Der Grund dafür liegt auf der Hand. Um $\nu_o(\lambda)$ berechnen zu können, müssen wir n unterschiedliche Modelle schätzen. Wollen wir dann auch noch jenen Wert von λ bestimmen, welcher $\nu_o(\lambda)$ minimiert, so müssen sehr viele Modelle geschätzt und ausgewertet werden. Um $\nu_o(\lambda)$ effizienter berechnen zu können, betrachten wir folgende äquivalente Darstellung dieser Maßzahl. Es lässt sich zeigen, dass wir $\nu_o(\lambda)$ schreiben können als

$$\nu_o(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\frac{(y_i - \hat{\mu}_i)^2}{(1 - A_{ii})^2}, \tag{15}$$

wobei A die sogenannte Hat-Matrix, gegeben als

$$A = X(X'X + \lambda S)^{-1}X', \tag{16}$$

darstellt.

Um die Korrektheit von Gleichung (15) nachzuweisen, schreiben wir zunächst die Zielfunktion der Parameterschätzung für das Modell, welches das Punktepaar $\{x_i, y_i\}$ nicht berücksichtigt, um. Wie wir bereits gesehen haben, setzt sich diese Zielfunktion aus der Summe der quadratischen Abweichungen der Modellschätzungen zu den Beobachtungen und einem Strafterm zusammen. Sie ist demnach gegeben als

$$\sum_{j=1, j\neq i}^{n}(y_j - \hat{\mu}_j^{[-i]})^2 + \text{Strafterm.} \tag{17}$$

Wir wollen die Beobachtungen y nun so modifizieren, dass wir in obiger Darstellung (17) der Zielfunktion wieder über alle $j = 1, \ldots, n$ summieren können. Durch Addition von $(\hat{\mu}_i^{[-i]} - \hat{\mu}_i^{[-i]})^2 = 0$ lässt sich (17) umschreiben zu

$$\sum_{j=1}^{n}(y_j^* - \hat{\mu}_j^{[-i]})^2 + \text{Strafterm.} \tag{18}$$

Hierbei ist y^* der Vektor der modifizierten Beobachtungen mit $y^* = y - \bar{y}^{[i]} + \bar{\mu}^{[i]}$, wobei $\bar{y}^{[i]}$ und $\bar{\mu}^{[i]}$ Vektoren sind deren jeweils i-te Koordinate gegeben ist als y_i bzw. $\hat{\mu}_i^{[-i]}$ und alle restlichen Koordinaten gleich Null sind. Somit stellt der Ausdruck (18) die Zielfunktion der Parameterschätzung eines Modells auf Basis der modifizierten Beobachtungen y^* dar.

Die Hatmatrix A dieses modifizierten Modells ist dabei wieder dieselbe wie zuvor für das nicht modifizierte Modell. Damit ergibt sich für $\hat{\mu}_i^{[-i]}$, dass

$$\hat{\mu}_i^{[-i]} = A_i y^* = A_i y - A_{ii} y_i + A_{ii} \hat{\mu}_i^{[-i]} = \hat{\mu}_i - A_{ii} y_i + A_{ii} \hat{\mu}_i^{[-i]}. \tag{19}$$

Durch einfaches Umformen dieser Gleichung erhalten wir

$$y_i - \hat{\mu}_i^{[-i]} = \frac{y_i - \hat{\mu}_i}{1 - A_{ii}}. \tag{20}$$

Somit lässt sich also die Differenz zwischen der tatsächlichen Beobachtung y_i und der Schätzung $\hat{\mu}_i^{[-i]}$ jenes modifizierten Modells, welches das Punktepaar $\{x_i, y_i\}$ nicht berücksichtigt, durch die Schätzungen des nicht modifizierten Modells und dem entsprechenden Diagonalelement der Hatmatrix beschreiben. Setzen wir diese Gleichheit (20) in die Definition (14) des Kreuzvalidierungswertes ein, so erhalten wir die behauptete Darstellung (15).

In praktischen Anwendungen werden die Gewichte $(1 - A_{ii})$ oft durch das durchschnittliche Gewicht $tr(\mathbb{I} - A)/n$ ersetzt. Damit ergibt sich der sogenannte Generalisierte Kreuzvalidierungswert

$$\nu_g(\lambda) = \frac{n \sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2}{[tr(\mathbb{I} - A)]^2}. \tag{21}$$

Dieser bietet gegenüber dem normalen Kreuzvalidierungswert $\nu_o(\lambda)$ Vorteile in der Effizienz der Berechenbarkeit und der Invarianz gegenüber Rotation der Daten durch eine orthogonale Matrix (siehe Wahba, 1990). Zu diesem Zweck wird hierbei einfach der normale Kreuzvalidierungswert für eine standardisierte Rotation berechnet. Die zugehörige Rotationsmatrix ist dabei so gewählt, dass die Diagonalelemente der resultierenden Hatmatrix alle gleich sind. Mithilfe iterativer Methoden kann nun ein λ gewählt werden, welches diesen generalisierten Kreuzvalidierungswert minimiert (siehe Abschnitt 3.5.2).

3.4 Additive Modelle

Im bisherigen Verlauf dieses Kapitels haben wir gesehen, wie wir die Abhängigkeitsstruktur in einem Linearen Modell zwischen Responsevariable und einem Prädiktor mit Hilfe einer Glatten Funktion modellieren können. Als nächsten Schritt hin zu der Klasse der GAMe werden wir in diesem Abschnitt zwei (oder mehrere) Prädiktoren, die in Form einer Glatten Funktion im Linearen Prädiktor vorkommen, betrachten. Eine beispielhafte Modellstruktur wäre demnach

$$y_i = f_1(x_i) + f_2(z_i) + \epsilon_i. \tag{22}$$

Hierbei werden nun zwei Prädiktoren x_i und z_i und die zugehörigen Funktionen $f_1(\cdot)$ und $f_2(\cdot)$ betrachtet. Der Einfachheit halber gehen wir im Folgenden wieder von der Annahme aus, dass alle Prädiktoren x_i und z_i im Intervall $[0, 1]$ liegen.

Ein Modell, wie es Gleichung (22) beschreibt, wird als sogenanntes Additives Modell bezeichnet. Der Grund für diese Bezeichnung ist die bereits implizit getroffene Annahme, dass die Abhängigkeit zwischen der Responsevariable und den Prädiktoren mit $f_1(x) + f_2(z)$ durch eine additive Struktur modelliert wird. Somit werden allgemeinere Formen, wie sie für eine Glatte Funktion in zwei Variablen $f(x, z)$ möglich sind, bereits von vornherein ausgeschlossen.

Ein weiterer Punkt, den es zu beachten gilt, ist die Identifizierbarkeit eines solchen Additiven Modells. Betrachten wir z.B. obiges Modell (22), so ist leicht ersichtlich, dass durch Addition einer beliebigen Konstante zu $f_1(\cdot)$ und Subtraktion selbiger Konstante von $f_2(\cdot)$ ein äquivalentes Modell resultiert. Aus diesem Grunde müssen bei der Modellschätzung sogenannte Identifizierbarkeitsbedingungen berücksichtigt werden. Sind diese Bedingungen jedoch erfüllt, so kann ein Additives Modell mit den bereits bekannten Methoden geschätzt werden. Dies bedeutet, dass wir für alle vorkommenden Glatten Funktionen eine Menge von Basisfunktionen definieren deren Koeffizienten wir dann durch Minimierung der Fehlerquadratsumme schätzen. Dabei wird der Parameter λ, der die Gewichtung des „Glätte"-Strafterms angibt, wiederum durch die oben vorgestellte Methode der Kreuzvalidierung bestimmt.

Benutzen wir für Modell (22) die Basisfunktionen für kubische Splines aus Abschnitt 3.3.1, so erhalten wir für die Funktionen $f_1(\cdot)$ und $f_2(\cdot)$ im Linearen Prädiktor folgende Darstellungen

$$f_1(x) = \delta_1 + x\delta_2 + \sum_{j=1}^{m_1-2} R(x, \bar{x}_j)\delta_{j+2} \quad \text{sowie}$$
$$f_2(z) = \gamma_1 + z\gamma_2 + \sum_{j=1}^{m_2-2} R(z, \bar{z}_j)\gamma_{j+2}.$$

Hierbei sind δ_j und γ_j die zu schätzenden Parameter, $\bar{x}_j$ und $\bar{z}_j$ die zuvor bestimmten Stützstellen und die $R(\cdot, \cdot)$ die in Abschnitt 3.3.1 präsentierten Basisfunktionen, d.h.

$$\begin{aligned} R(a, b) = &\frac{1}{4}\left[(b - 1/2)^2 - 1/12\right] \cdot \left[(a - 1/2)^2 - 1/12\right] \\ &- \frac{1}{24}\left[(|a - b| - 1/2)^4 - 1/2(|a - b| - 1/2)^2 + 7/240\right]. \end{aligned}$$

Die einfachste Möglichkeit die oben besprochene Identifizierbarkeitsbedingung zu erfüllen ist es, entweder δ_1 oder γ_1 gleich Null zu setzen. Wählen wir $\gamma_1 = 0$, so lässt sich Modell (22) schreiben als $y = X\beta + \epsilon$, wobei sich die i-te Zeile der Modellmatrix X ergibt als

$$X_i = [1, x_i, R(x_i, \bar{x}_1), \ldots, R(x_i, \bar{x}_{m_1-2}), z_i, R(z_i, \bar{z}_1), \ldots, R(z_i, \bar{z}_{m_2-2})]$$

und der Parametervektor β gegeben ist als

$$\beta = [\delta_1, \ldots, \delta_{m_1}, \gamma_2, \ldots, \gamma_{m_2}]'.$$

Analog zu Abschnitt 3.3.2 lassen sich natürlich auch wieder die Strafterme darstellen als

$$\int_0^1 f_1''(x)^2 dx = \beta' S_1 \beta \quad \text{sowie}$$

$$\int_0^1 f_2''(x)^2 dx = \beta' S_2 \beta,$$

wobei die Struktur von S_1 und S_2 wieder gegeben ist durch $S_{1\,i+2,j+2} = R(\bar{x}_i, \bar{x}_j)$ für $i, j = 1, \ldots, m_1 - 2$ und $S_{2\,i+2,j+2} = R(\bar{z}_i, \bar{z}_j)$ für $i, j = 1, \ldots, m_2 - 2$. Damit ergibt sich als Zielfunktion der Modellschätzung

$$\|y - X\beta\|^2 + \lambda_1 \beta' S_1 \beta + \lambda_2 \beta' S_2 \beta, \tag{23}$$

wobei nun die zwei Parameter λ_1 und λ_2 die Glätte der Funktionen $f_1(\cdot)$ und $f_2(\cdot)$ bestimmen. Erwähnenswert ist hierbei, dass für λ_1 und λ_2 bekannt und $S := \lambda_1 S_1 + \lambda_2 S_2$, die obige Zielfunktion umgeschrieben werden kann zu

$$\|y - X\beta\|^2 + \beta' S \beta = \left\| \begin{pmatrix} y \\ 0 \end{pmatrix} - \begin{pmatrix} X \\ B \end{pmatrix} \beta \right\|^2. \tag{24}$$

Hierbei ist B so gewählt, dass $B'B = S$ gilt. Mit dieser Darstellung ist ersichtlich, dass ein Additives Modell, wie wir es hier definiert haben, für eine feste Wahl der Glättungsparameter durch dieselben Methoden geschätzt werden kann wie sie auch für Lineare Modelle genutzt werden. Zu diesem Zweck müssen wir also lediglich neue Beobachtungen hinzufügen. Die Prädiktoren dieser neuen Beobachtungen entsprechen dabei den Zeilen der Matrix B. Die zugehörigen Beobachtungen der Responsevariable sind gleich Null.

Die hier definierte Klasse der Additiven Modelle stellt die Basis für die Generalisierten Additiven Modelle dar. Im nun folgenden Abschnitt wird diese Modellklasse eingeführt. Des weiteren wird dabei näher auf die Schätzung der Glättungsparameter eingegangen.

3.5 Generalisierte Additive Modelle

Wir haben bereits gesehen, wie wir mit einer Glatten Funktion in Lineare Modellen umgehen können. Durch Hinzunahme weiterer Glatter Funktionen haben wir dann die sogenannten Additiven Modellen erhalten. Von diesen Modellen wissen wir nun, dass für fest vorgegebene Glättungsparameter deren Schätzung äquivalent zu jener eines Linearen Modells mit zusätzlich hinzugefügten Beobachtungen ist. Der letzte Schritt von Additiven

Modellen hin zu den in Abschnitt 3.2 bereits vorgestellten Generalisierten Additiven Modellen ist nun vergleichbar mit jenem aus Abschnitt 3.1.2, der uns von Linearen Modellen zu Generalisierten Linearen Modellen brachte. Dies bedeutet, dass wir nunmehr eine stetige monotone Funktion (Linkfunktion) des Erwartungswertes der Responsevariablen mit Hilfe des Linearen Prädiktors modellieren. Die Verteilung dieser Responsevariablen kann wiederum eine beliebige Verteilung aus der einparametrigen Linearen Exponentialfamilie (LEF) entsprechen.

Während es uns möglich war Additive Modelle mit Hilfe der Darstellung von Glatten Funktionen durch Linearkombinationen von Basisfunktionen als Lineare Modelle zu interpretieren (und zu schätzen), ist eine ähnliche Vorgehensweise für GAMe und GLMe nicht mehr möglich. Der Grund dafür ist, dass es keine Vorgehensweise gibt, um ein GAM in ein GLM zu überführen. Dies hat zur Folge, dass sich die Methoden zur Schätzung eines GAMs von den Methoden zur Schätzung von Linearen Modellen unterscheiden.

Im Folgenden wollen wir eine solche Methode zur Schätzung von GAMen präsentieren. Die hier vorgestellte Methode entspricht dem sogenannten „penalized iteratively re-weighted least squares" Schema, kurz P-IRLS. Selbiges Schema, jedoch ohne die Strafterme, wird zur Schätzung von GLMen verwendet (siehe Björk, 1984). Dieses iterative Schema wird solange durchgeführt, bis bestimmte Konvergenzkriterien erfüllt sind. Da diese Konvergenzkriterien abhängig von der jeweiligen Implementation sind, werden wir hier nicht näher auf diese eingehen.

Es folgt eine Beschreibung des P-IRLS Iterationsschemas. Der aktuelle Iterationsschritt wird dabei mit k beschrieben. Sind die momentanen Iterate des Parametervektors und des geschätzten Erwartungswertvektors gegeben als β^k bzw. μ^k, so sieht ein Iterationsschritt des P-IRLS Schemas wie folgt aus.

1. Berechne Gewichte w_i und Pseudobeobachtungen z_i als

$$w_i = \frac{1}{V(\mu_i^k)g'(\mu_i^k)} \quad \text{und}$$
$$z_i = g(\mu_i^k)(y_i - \mu_i^k) + X_i\beta^k,$$

   wobei $V(\cdot)$ die Varianzfunktion der Responsevariablen entspricht, also $\mathbb{V}ar(Y_i) = V(\mu_i^k)\phi$ und X_i die i-te Zeile der Modellmatrix X darstellt. Dabei ist ϕ der als bekannt angenommene Dispersionsparameter der Verteilung (siehe Abschnitt 3.1.2).

2. Erhalte β^{k+1} als Optimallösung von

$$\min_\beta \left(\|\sqrt{W}(z - X\beta)\|^2 + \lambda_1\beta' S_1\beta + \lambda_2\beta' S_2\beta \right).$$

   Hierbei ist W eine Diagonalmatrix mit $W_{ii} = w_i$, den zuvor berechneten Gewichten. Mit Hilfe von β^{k+1} kann nun μ^{k+1} berechnet werden. Es folgt der nächste Iterationsschritt (Punkt 1.).

Im restlichen Teil dieses Abschnitts widmen wir uns wichtigen Details eines GAMs die bis jetzt nur oberflächlich erwähnt wurden. Dabei wollen wir vor allem auf den effektiven Freiheitsgrad eines GAMs sowie der Schätzung des Glättungsparameters λ eingehen.

3.5.1 Der effektive Freiheitsgrad

In einem klassischen GLM ergibt sich der Freiheitsgrad des resultierenden Modells als die Anzahl der Beobachtungen minus der Dimension des Parametervektors β, also der Anzahl an Prädiktoren. Wie bereits erwähnt, empfiehlt es sich im Zuge der Modellspezifikation eines GAMs die Dimension der Basisfunktionen, und somit die mögliche Flexibilität der entstehenden Glatten Funktionen, etwas höher anzusetzen als dies im ersten Moment notwendig erscheint. Der Grund dafür ist, dass damit ein größerer Raum von Funktionen entsteht der somit mehr Abhängigkeitsstrukturen abdeckt. Durch den (mit dem Parameter λ gewichteten) Strafterm resultieren im Allgemeinen aber geschätzte Funktionen die nicht die ganze mögliche Flexibilität ausnutzen. Der tatsächliche Freiheitsgrad dieser Funktion stimmt also nicht mit der Basisdimension überein. Nun wirft dieser Umstand die Frage auf, wie der Freiheitsgrad einer solchen Funktion zu messen ist.

Um den effektiven Freiheitsgrad zu motivieren, kehren wir noch einmal zu den Linearen- und Additiven Modellen zurück. Für Lineare Modelle ergibt sich die Schätzung für den Parametervektor β^{LM} als

$$\hat{\beta}^{LM} = (X'X)^{-1}X'y, \tag{25}$$

wobei X die Modellmatrix und y der Vektor der beobachteten Responsevariablen darstellt. Es lässt sich leicht zeigen, dass für Additive Modelle der Schätzer für β^{AM} gegeben ist als

$$\hat{\beta}^{AM} = (X'X + S)^{-1}X'y. \tag{26}$$

Hierbei beschreibt S das λ-fache jener quadratischen Matrix, die durch Gleichung (12) aus Abschnitt 3.3.2 definiert wird. Wir können diesen Schätzer nun umschreiben zu

$$\begin{aligned}\hat{\beta}^{AM} &= (X'X + S)^{-1}X'y \\ &= (X'X + S)^{-1}X'X(X'X)^{-1}X'y \\ &= F\hat{\beta}^{LM},\end{aligned}$$

wobei F gegeben ist als $F = (X'X + S)^{-1}X'X$. Dies bedeutet, dass F den Parameterschätzer $\hat{\beta}^{LM}$ des Linearen Modells auf den Parameterschätzer $\hat{\beta}^{AM}$ des Additiven Modells abbildet. Nun ergibt sich daraus, dass $\partial\hat{\beta}_i^{AM}/\partial\hat{\beta}_i^{LM} = F_{ii}$ gilt. Somit gibt das i-te Diagonalelement von F die Änderungsrate von $\hat{\beta}_i^{AM}$ in Abhängigkeit von $\hat{\beta}_i^{LM}$ an. Ohne einen Strafterm gilt für die obige Matrix $S = 0$ woraus $F = \mathbb{I}$ und somit $\hat{\beta}^{AM} = \hat{\beta}^{LM}$ folgt. Die Parameter der Modelle sind identisch, und somit die vorher beschriebene Änderungsrate F_{ii} gleich Eins. Aus diesem Grund interpretieren wir F_{ii} als effektiven Freiheitsgrad.

Im allgemeinen (gewichteten) Fall lässt sich zeigen, dass die Abbildung F gegeben ist als

$$F = (X'WX + S)^{-1}X'WX. \tag{27}$$

Somit können wir im obigen P-IRLS Schema die effektiven Freiheitsgrade iterativ berechnen.

3.5.2 Schätzung des Glättungsparameters

Das oben beschriebene P-IRLS Schema dient zur iterativen Berechnung des Parametervektors β, gegeben einem Glättungsparameter λ. In diesem Abschnitt wollen wir nun auf die Schätzung dieses Parameters λ eingehen. Wie bereits erwähnt, werden wir dafür Methoden der sogenannten Kreuzvalidierung verwenden. Dafür müssen wir zunächst den in Abschnitt 3.3.2 vorgestellten Generalisierten Kreuzvalidierungswert $\nu_g(\lambda)$ auf GAMe verallgemeinern.

Die Zielfunktion der Parameterschätzung in einem GAM ist gegeben durch

$$\|\sqrt{W}(z - X\beta)\|^2 + \text{Strafterm}. \tag{28}$$

Wir können nun, analog zur Vorgehensweise in Abschnitt 3.3.2, den Generalisierten Kreuzvalidierungswert (in Abhängigkeit der Gewichte w) berechnen als

$$\nu_g^w(\lambda) = \frac{n\|\sqrt{W}(z - X\beta)\|^2}{[n - tr(A)]^2}. \tag{29}$$

Da die im P-IRLS Schema verwendete Zielfunktion nur eine Approximation der Loglikelihoodfunktion ist, ist auch diese Darstellung nur eine lokale Approximation. Eine globale Version dieses Wertes erhalten wir indem wir die Deviance im Parameterschätzer $\hat{\beta}$ verwenden, also

$$\nu_g(\lambda) = \frac{nD(\hat{\beta})}{[n - tr(A)]^2}, \tag{30}$$

siehe Hastie und Tibshiranie (1990).

Im Allgemeinen haben wir nun zwei Möglichkeiten wie wir den Glättungsparameter λ minimieren können.

1. Direkte Minimierung von ν_g: Hierfür benutzen wir klassische (iterative) Methoden der Optimierung. Für jedes dabei vorkommende Iterat von λ muss der Funktionswert mit Hilfe des P-IRLS Schemas berechnet werden.

2. Iterative Minimierung von ν_g^w: Wir minimieren λ in jedem Iterationsschritt des P-IRLS Schemas. Da sie vergleichsweise recheneffizient ist, wird diese Methode auch Performance Iteration genannt.

Die Performance Iteration benötigt dabei im Normalfall nicht mehr Iterationsschritte wie zur Parameterschätzung ohnehin vonnöten gewesen wären. Es kommt allerdings vor, dass die Resultate der Performance Iteration nicht exakt mit jenen der direkten Minimierung übereinstimmen. Dies könnte in der Theorie zu Problemen führen wenn zwei Modelle anhand ihrer ν_g^w-Werte verglichen werden sollen. In der Praxis jedoch sind in den meisten Fällen zwei Modelle, die ähnliche ν_g^w-Werte aufweisen, ohnehin nicht unterscheidbar.

Ein größeres Problem stellt die Konvergenz der Performance Iterationen dar. Die einfachste Art von Fehlern, die in einem solchen P-IRLS Schema auftreten kann, ist dabei die Möglichkeit des Kreisens der Iterate. Tritt ein solches Kreisen auf, so konvergieren die Iterate natürlich nicht. Ein typischer Grund für ein solches Verhalten stellt dabei die sogenannte „Concurvity" dar. Dieser Effekt tritt auf, wenn der Lineare Prädiktor Terme wie $f_1(x_i) + f_2(z_i)$ beinhaltet, wobei die Prädiktoren z_i in Wahrheit ihrerseits Funktionen von x_i sind.

Es sei an dieser Stelle erwähnt, dass die direkte Minimierung keine dieser Nachteile aufweist. Einziges Problem dieser Methode ist der hohe dafür erforderliche Rechenaufwand. Der Grund dafür liegt auf der Hand. Für jeden Zwischenschritt muss das P-IRLS Schema bis zur Konvergenz iteriert werden um die Zielfunktion (ν_g) für das momentanen Iterat auszuwerten.

3.5.3 Tensorprodukte

Wie bereits erwähnt, beschränkten wir uns bei der Modellierung von eindimensionalen Glatten Kurven auf die Darstellung durch kubische Splines. Im späteren Verlauf dieser Arbeit werden wir allerdings auch Interaktionen zwischen zwei stetigen Prädiktoren betrachten. Um die dafür verwendeten zweidimensionalen Glatten Funktionen zu modellieren werden wir die Technik der sogenannten Tensorprodukte nutzen.

Das wesentliche Charakteristikum von Tensorprodukten ist, dass sich die Glätte der resultierenden Funktion entlang der einzelnen Koordinaten unterscheiden kann. Diese Eigenschaft ist vor allem dann von Vorteil, wenn die betrachteten Prädiktorvariablen in unterschiedliche Einheiten gemessen werden. Bei unserem konkreten Anwendungsfall wäre dies z.B. für das Alter des Versicherungsnehmers (Jahre) und der Höhe der monatlichen Prämie (Euro) der Fall.

Da wir in dieser Arbeit lediglich Tensorprodukte von zwei stetigen Prädiktoren betrachten, wollen wir an dieser Stelle auch nur den zweidimensionalen Fall motivieren. Die Verallgemeinerung für mehrdimensionale Tensorprodukte basiert dabei auf der gleichen Vorgehensweise und kann u.a. in de Boor (1978) nachgeschlagen werden.

Um ein Tensorprodukt von zwei stetigen Prädiktorvariablen x und z zu motivieren, gehen wir davon aus, dass die Basisfunktion für die Darstellung der zwei Glatten Funktionen $f_1(\cdot)$

und $f_2(\cdot)$ bereits bestimmt wurden. Somit können wir diese analog zu oben darstellen als

$$f_1(x) = \sum_{j=1}^{m_1} \alpha_j a_j(x) \quad \text{und} \quad f_2(z) = \sum_{k=1}^{m_2} \beta_k b_k(z),$$

wobei α_j und β_j die Parameter und $a_j(\cdot)$ und $b_j(\cdot)$ die Basisfunktionen repräsentieren.

Wir wollen nun eine zweidimensionale Funktion erzeugen, welche entlang der beiden Koordinaten eine unterschiedliche Glätte aufweisen kann. Die grundlegende Idee dafür ist es die Parameter α_j ihrerseits als Funktionen der zweiten Koordinate z darzustellen:

$$\alpha_j(z) = \sum_{k=1}^{m_2} \beta_{jk} b_k(z).$$

Setzen wir diese Form der Parameter α_j in obiger Darstellung der Funktion $f_1(\cdot)$ ein, so erhalten wir eine zweidimensionale Funktion $f_{1\,2}(\cdot)$ gegeben als

$$f_{1\,2}(x, z) = \sum_{j=1}^{m_1} \sum_{k=1}^{m_2} \beta_{jk} b_k(z) a_j(x).$$

Mit dieser Vorgehensweise ist sichergestellt, dass sich $f_1(\cdot) = f_{1\,2}(\cdot, z)$ glatt in z verhält. Dies kann für eine beliebige Anzahl an stetigen Prädiktoren fortgeführt werden.

Der Vollständigkeit halber wollen wir im Folgenden die Strafterme eines solchen Tensorproduktes näher betrachten. Analog zu Abschnitt 3.3.2 können wir diese Strafterme schreiben als

$$J_1(f_1) = \alpha' S_1 \alpha \quad \text{und} \quad J_2(f_2) = \beta' S_2 \beta,$$

wobei S_1 und S_2 wiederum quadratische Matrizen mit bekannten Koeffizienten und α und β die Vektordarstellungen obiger Koeffizienten sind. Für die von uns als Basis verwendeten kubischen Splines könnte als Straffunktion wieder das Integral über die zweiten Ableitungen dienen. Somit wäre

$$J_1(f_1) = \int \left(\frac{\partial^2 f_1}{\partial x^2} \right)^2 dx \quad \text{und} \quad J_2(f_2) = \int \left(\frac{\partial^2 f_2}{\partial z^2} \right)^2 dz.$$

Definieren wir $f_{1|2}(x) := f_{1\,2}(x, z)$, wobei z fixiert ist, und $f_{2|1}(z)$ analog, so wäre die nächstliegendste Möglichkeit um die Glätte von $f_{1\,2}(\cdot, \cdot)$ zu bestimmen gegeben als

$$J(f_{1\,2}) = \lambda_1 \int_z J_1(f_{1|2}) dz + \lambda_2 \int_x J_2(f_{2|1}) dx.$$

Hierbei stellen λ_1 und λ_2 die Glättungsparameter entlang der unterschiedlichen Koordinaten dar. Für die von uns genutzten kubischen Splines ergibt sich der Strafterm eines zweidimensionalen Tensorproduktes somit als

$$J(f_{1\,2}) = \int_{x,z} \left(\lambda_1 \left(\frac{\partial^2 f_{1\,2}}{\partial x^2} \right)^2 + \lambda_2 \left(\frac{\partial^2 f_{1\,2}}{\partial z^2} \right)^2 \right) dxdz.$$

Da das Ziel unserer Arbeit die Erstellung eines konkreten GAMs ist, werden wir zum Abschluss dieses Kapitels auf die praktische Handhabung dieser Modellklasse eingehen. Der nun folgende Abschnitt stellt dabei eine Übersicht über die wichtigsten Befehle in der Statistiksoftware R dar. Diese Übersicht erleichtert die Interpretation der in Kapitel 4 präsentierten Ergebnisse.

3.6 Generalisierte Additive Modelle in R

Zum Abschluss dieses Kapitels wollen wir näher auf die praktische Anwendung von GAMen eingehen. Die folgenden Ausführungen beziehen sich dabei auf das freie Softwarepaket R[1]. R ist eine Programmiersprache für statistische Problemstellungen. Die Kernfunktionalitäten können dabei durch sogenannte Pakete erweitert werden. Ein solches Paket ist *mgcv*[2] (Mixed GAM Computation Vehicle). Dieses Paket kann dafür genutzt werden, GAMe, wie wir sie in diesem Kapitel vorgestellt haben, zu schätzen. Die Bedienung orientiert sich dabei an den klassischen Funktionen *lm* und *glm*, welche zur Schätzung von Linearen Modellen bzw. GLMen genutzt werden. Ein weiterer Vorteil von *mgcv* ist, dass mit Hilfe dieses Paketes eine Parallelisierung der Berechnungen auf mehrere Rechenkerne einfach umzusetzen ist. Dies ist vor allem in Hinblick auf die von uns verwendeten Daten (mehr als 300.000 Beobachtungen pro Jahr) von großem Wert.

Im Folgenden wollen wir kurz die von uns verwendeten Basisbefehle von *mgcv* präsentieren. Wir beschränken uns dabei auf jene Funktionen und Parameter, die für unseren Anwendungsfall relevant sind. Für weitere Informationen zu *mgcv* verweisen wir auf die Dokumentation[3] dieses Paketes.

Die zentrale Funktionalität von *mgcv*, das Schätzen von GAMen, wird durch die Funktion *gam* ausgeführt. Wie bereits erwähnt, ist der Aufruf dieser Funktion sehr ähnlich zu *lm* und *glm* und lautet

```
modell<-gam(formula,family=binomial(link),data=list())
```

Mit dem Parameter *formula* wird wie gewohnt die Modellformel, also die Komponenten und Form des Linearen Prädiktors, beschrieben. Die Verteilungsfunktion der Responsevariablen sowie die Linkfunktion werden durch *family* spezifiziert. Wird keine Linkfunktion angegeben, so wird automatisch die kanonische Linkfunktion der jeweiligen Verteilung benutzt. Die Bezeichnungen der Responsevariable und der Prädiktoren, die in der Modellformel

[1] https://www.r-project.org/
[2] https://cran.r-project.org/package=mgcv
[3] https://cran.r-project.org/web/packages/mgcv/mgcv.pdf

benutzt wurden, müssen als Spaltennamen in den zugrundeliegenden Daten vorkommen. Diese Datenbasis wird als *data* angegeben. In der Modellformel ist es nun möglich, Glatte Funktionen von Prädiktoren zu berücksichtigen. Die Glatten Terme werden durch die Funktionen *s()*, *te()* und *ti()* definiert. Diese Funktionen können ausschließlich zur Definition von Glatten Funktionen in der Modellierung von GAMen genutzt werden und sind demnach keine R-Funktionen im herkömmlichen Sinn. Sie unterscheiden sich nach ihrem Anwendungszweck.

- *s()*: Wird genutzt, um eindimensionale Glatte Funktionen zu definieren. Durch zusätzliche Parameter kann z.B. die Splinebasis sowie der maximale Freiheitsgrad definiert werden.
- *te()*: Mit dieser Funktion können sogenannte Glatte Tensorprodukte definiert werden. Mit Hilfe dieser Tensorprodukte können zwei- oder mehrdimensionale Glatte Kurven geschätzt werden.
- *ti()*: Im Unterschied zu *te()* wird mit dieser Funktion lediglich die Tensorprodukt-Interaktion modelliert. So ist es z.B. möglich mit Hilfe von *s()* die Haupteffekte und mit Hilfe von *ti()* die Interaktion separat zu modellieren und auszuwerten.

In unserem Anwendungsfall kommen für eine Modellierung als Glatte Kurven die drei stetigen Prädiktoren Alter, Laufzeit und Prämie in Frage. In Kapitel 4 werden wir dabei hauptsächlich eindimensionale Glatte Kurven verwenden. Dies ist zum einen dadurch begründet, dass die Modellierung von mehrdimensionalen Glatten Kurven in unserer Anwendung keinen erkennbaren Vorteil bringen wird. Zum anderen ist die Schätzung eines solchen Tensorproduktes mit einem höheren Rechenaufwand verbunden. Ein weiterer, sehr auf unserem konkreten Anwendungsfall basierender Grund ist, dass es für drei- oder höher dimensionale Kurven keine Möglichkeit der grafischen Interpretation gibt. Da bereits die Modellierung mit eindimensionalen Glatten Kurven für Daten, mit dem Umfang wie sie in unserer Anwendung vorliegen, sehr rechenintensiv ist, präsentieren wir nun eine Möglichkeit wie diese Berechnungen parallelisiert werden können.

Um die Berechnungen eines GAM auf mehrere Rechenkerne zu verteilen, und somit die Rechenzeit drastisch zu verkürzen, bedarf es eines weiteren R-Paketes. Mit Hilfe des Paketes *parallel* ist es möglich, mehrere sogenannte Rechencluster zu erstellen. Diese können entweder durch verschiedene Rechenkerne eines einzelnen Computers oder durch verschiedene Rechner in einem Netzwerk repräsentiert werden. Nach der Initialisierung dieser Cluster exportieren wir die Daten, sodass diese auf den Recheneinheiten zur Verfügung stehen. In der Funktion *makePSOCKcluster* beschreibt der Parameter *names* entweder die (Netzwerk-) Namen der zu verwendenden Rechner oder die Anzahl an Rechenkernen die auf dem lokalen System genutzt werden sollen.

```
cl = makePSOCKcluster(names)
clusterExport(cl, data)
```

Anschließend können wir mit der Funktion *bam* (aus dem *mgcv* Package) die Berechnungen zur Schätzung eines GAMs auf diese Cluster verteilen. Im Funktionsaufruf unterscheidet sich *bam* von *gam* lediglich durch den zusätzlichen Parameter *cluster*, welcher die Information über die zur Verfügung stehenden Recheneinheiten beinhaltet.

```
modell<-bam(formula,family=binomial(link),data=list(),cluster=cl)
```

Dabei wird ausschließlich die für eine effiziente Berechnung des Iterationsschemas genutzte QR-Zerlegung[4] der Modellmatrix parallel berechnet. Da diese Zerlegung jedoch einer der dominierenden Faktoren in der Laufzeit der Modellschätzung darstellt, resultiert dadurch ein beachtlicher Zeitvorteil. Auf dem von uns genutzten Rechner konnte durch eine Parallelisierung auf vier (physische) Rechenkerne die Laufzeit nahezu geviertelt werden. Die Art und Weise wie die Berechnung eines GAMs mit Hilfe von *parallel* und *bam* parallelisiert werden kann hängt nicht von der Form des konkreten Modells ab. Aus diesem Grund werden wir im weiteren Verlauf dieser Arbeit auf die Angabe des Parameters *cluster* beim Funktionsaufruf verzichten.

Auch bei der Modellevaluierung ähnelt ein Modell geschätzt durch *gam* oder *bam* einem klassischen Regressionsmodell. So erhalten wir durch *summary(modell)* eine übersichtliche Zusammenfassung des Modells. Diese beinhaltet etwa die genaue Modellformel und die geschätzten Koeffizienten mit zugehörenden (geschätzten) Standardabweichungen. Des weiteren ist es auch möglich viele andere Funktionen die für Lineare Modelle oder Generalisierte Lineare Modelle bekannt sind zu verwenden. So erhalten wir z.B. mit *AIC(modell)* das Akaike Informationskriterium eines GAMs. Zusätzlich zu diesen klassischen Auswertungen bietet das *mgcv* Paket speziell für GAMe entworfene Funktionen zu Modellevaluation an. Sehr nützlich ist etwa *gam.check(modell)*. Durch diese Funktion wird eine Reihe von Residuenplots erstellt und wir erhalten einen informellen Test dafür, ob der von uns spezifizierte maximale Freiheitsgrad ausreichend groß bemessen war.

[4] siehe Watkins (2008)

4 Modellierung der Stornowahrscheinlichkeiten

Nachdem wir uns in Kapitel 2 mit der Datenbasis und in Kapitel 3 mit möglichen Modellklassen beschäftigt haben, soll in diesem Kapitel die tatsächliche Modellierung von Stornowahrscheinlichkeiten in der Krankenversicherung beschrieben werden. Unser Ziel ist es dabei ein Modell zu entwickeln, dessen Auswahl wir in allen uns zur Verfügung stehenden Beobachtungsjahren vertreten können. Mit Hilfe eines solch einheitlichen Modells ist es dann möglich, die Abhängigkeitsstrukturen zwischen Stornowahrscheinlichkeit und den einzelnen Prädiktoren über die Jahre hinweg zu vergleichen.

Dieses Kapitel strukturiert sich wie folgt. In Abschnitt 4.1 werden wir versuchen die Stornowahrscheinlichkeit mit Hilfe eines GLMs zu modellieren. Da uns dabei kein zufriedenstellender Kompromiss zwischen Modellkomplexität und adäquater Darstellung der nichtlinearen Abhängigkeitsstrukturen gelingen wird, beschäftigen wir uns ab Abschnitt 4.2 mit der Modellklasse der GAMe. Nachdem wir eine Menge von brauchbaren Modellen bestimmt haben, befassen wir uns in Abschnitt 4.3 mit der Modellevaluation. Diese gestaltet sich trickreich, da die meisten klassischen Methoden, wie wir sie für allgemeine GLMe und GAMe kennen, für logistische Regressionsmodelle nicht aussagekräftig sind. Aus diesem Grund werden in diesem Abschnitt individuell für unseren Anwendungszweck entworfene Kennzahlen zum Modellvergleich präsentiert.

Bevor wir mit der Modellierung beginnen, wiederholen wir an dieser Stelle noch einmal kurz die Ausgangslage. Es gilt ein Modell zu bestimmen, mit dessen Hilfe wir die Stornowahrscheinlichkeit in Abhängigkeit der Prädiktoren schätzen können. Die möglichen Prädiktoren sind dabei das Alter und das Geschlecht des Versicherungsnehmers, die Laufzeit und die monatlich zu zahlende Prämie der Tarifposition, die Tarifklasse sowie die beiden binären Beobachtungen der Gruppenzugehörigkeit und des inkludierten Rabatts einer Position. Die Daten der einzelnen Beobachtungsjahre liegen dabei als *data2012*, *data2013* und *data2014* vor. Wenn nicht explizit angegeben, beziehen sich die in diesem Kapitel präsentierten Analysen und Grafiken auf das aktuellste Beobachtungsjahr (2014). Es gilt nun in einem ersten Schritt die geeignete Modellklasse zu bestimmen.

4.1 Modellierung als GLM

Da die Responsevariablen von binärer Form sind (Stornierung oder keine Stornierung), ist es naheliegend, ein logistisches Regressionsmodell zu wählen. Für die binären Prädiktorvariablen Gruppe, Rabatt und Geschlecht ergibt sich dabei selbstredend die Modellierung als zweistufige Faktoren. In diesem ersten Schritt wollen wir auch die Laufzeit und die monatlichen Prämien eines Tarifes sowie das Alter des Versicherungsnehmers als Faktoren modellieren. Der Grund dafür ist, dass wir auf diese Art und Weise die bestmögliche (weil am wenigsten restringierte) Anpassung des Modells an die Daten erhalten. Außerdem haben wir bereits in Abschnitt 2.2 die Abhängigkeitsstruktur analysiert und gesehen, dass

eine lineare Modellierung wohl nicht adäquat wäre. Während sich die Laufzeit und das Alter (die beide als natürlich Zahlen angegeben sind) direkt als Faktorstufen interpretieren lassen, ist eine solche Einteilung in Faktorstufen bei den monatlichen Prämien durch die Daten nicht vorgegeben. Aus diesem Grund wurden die monatlichen Prämien in Klassen eingeteilt, deren Anzahl und Größe so gewählt wurden, dass eine möglichst gleichmäßige Verteilung der Tarife auf diese Klassen entsteht.

Unser erstes Modell ergibt sich nun als GLM, welches alle uns zur Verfügung stehenden Prädiktoren als Faktoren beinhaltet. Die Verteilung der Responsevariable wurde als die Binomialverteilung definiert. Die zugehörige kanonische Linkfunktion ist die logit-Funktion, wie sie bereits in Abschnitt 3.1.1 Gleichung (5) definiert wurde. Als erstes bemerken wir dabei die unpraktikabel lange Laufzeit, die zur Schätzung dieses Modells notwendig ist. Grund dafür ist natürlich die Kombination zwischen enorm vielen Beobachtungen und sehr vielen Faktorstufen.

```
glm(STORNO~ALTER_F+LAUFZEIT_F+PRAEMIE_F+PG+RABATT+GRUPPE+VRSEX,
        data2014, family=binomial())
```

Die große Anzahl an Faktorstufen ist aus vielerlei Hinsicht unvorteilhaft. Die vielen unterschiedlichen Stufen erschweren die Interpretation und die Vergleichbarkeit der Ergebnisse, was ein zentraler Aspekt unserer Modellierung ist. Da nun z.B. für jede Altersstufe eine eigene Stornowahrscheinlichkeit modelliert wird, kann die zugrunde liegende Abhängigkeit zwischen der Wahrscheinlichkeit für eine Stornierung und der Variable Alter nur schwer beschrieben werden. Auch beobachten wir sogenannten Separationseffekte. Dies sind extrem große oder kleine Werte für einzelne geschätzte Koeffizienten. Der Grund dafür ist, dass z.B. für manche Alters- oder Tarifklassen alle beobachteten Tarife nicht storniert wurden. Um dies auch im Modell abzubilden, werden die entsprechenden Koeffizienten quasi als minus oder plus unendlich (sehr große oder kleine Werte) gewählt. Des weiteren können wir auch keine Interaktionen betrachten, da z.B. eine Interaktion zwischen Alter und Laufzeit zu tausenden von zu schätzenden Koeffizienten führen würde.

Dieses Modell liefert allerdings auch einige interessante Aspekte. So erkennen wir etwa, dass das Geschlecht des Versicherungsnehmers zusätzlich zu den anderen Prädiktoren keine signifikante Information über die Stornowahrscheinlichkeit liefert (dies gilt für alle Beobachtungsjahre). Dies erkennen wir z.B. an einer „Analysis of Variance" (AnoVa) Analyse des obigen GLM (wie im Folgenden abgebildet) oder daran, dass sich der Koeffizient des Faktors Geschlecht nicht signifikant von Null unterscheidet. Da dies für alle Beobachtungsjahre gilt, entscheiden wir uns dafür den Faktor Geschlecht von nun an zu vernachlässigen.

```
             Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                        302335      82265
ALTER_F     105  2556.32     302230      79709 < 2.2e-16 ***
LAUFZEIT_F   88   478.74     302142      79230 < 2.2e-16 ***
```

PRAEMIE_F	64	694.88	302078	78535	< 2.2e−16 ***
PG	3	1471.68	302075	77064	< 2.2e−16 ***
RABATT	1	53.01	302074	77011	3.32e−13 ***
GRUPPE	1	1.15	302073	77010	0.2843
VRSEX	1	0.74	302072	77009	0.3891

Wir erkennen anhand dieser AnoVa-Tabelle auch zum ersten Mal eine charakteristische Eigenschaft von logistischen Regressionsmodellen. Durch die Hinzunahme von allen uns zur Verfügung stehenden Prädiktoren können wir die Deviance lediglich um 6.4% verringern (von 82265 auf 77009). Der Grund dafür sind die unterschiedlichen Skalen der beobachteten Responsevariablen und der von uns geschätzten Wahrscheinlichkeiten. Während wir für jeden Tarif am Ende eines Jahres wissen, ob dieser storniert wurde oder nicht ($\in \{0, 1\}$), erhalten wir durch unser Modell eine Schätzung für die Wahrscheinlichkeit einer Stornierung ($\in [0, 1]$). Dies hat auch zur Folge, dass andere klassische Methoden, wie sie zur Validierung von GLMen genutzt werden, für logistische Modelle ungeeignet sind. So sind z.B. alle Residuenplots nicht zu interpretieren. Auch die klassischen „Goodness-of-Fit" (GoF) Maße liefern keine brauchbaren Ergebnisse.

Die Reduktion der Deviance ist für unseren Anwendungsfall demnach nicht als absolute sondern als relative Maßzahl von Interesse. Dies bedeutet, dass die Deviance-Reduktion eines einzelnen Modells keine große Aussagekraft besitzt, wir aber sehr wohl zwei Modelle anhand dieser Kennzahl vergleichen können. Wir werden uns in Abschnitt 4.3 mit weiteren Methoden zur Modellevaluation beschäftigen.

Natürlich wäre es an dieser Stelle möglich auch andere GLMe zu betrachten. So könnten wir z.B. versuchen die Anzahl der Faktorstufen zu verringern um so der Modellkomplexität und dem Separationseffekt entgegenzuwirken. Dieses Vorgehen müsste aber individuell auf jedes Jahr angepasst werden. Somit wäre ein jahresübergreifender Vergleich der Modelle nur schwer umsetzbar. Auch wäre es möglich, das Alter, die Laufzeit und die monatliche Prämie als stetige Prädiktoren in ein GLM aufzunehmen. Dies widerspricht jedoch den Beobachtungen die wir in Abschnitt 2.2 gemacht haben. Dort haben wir gesehen, dass diese Abhängigkeitsstrukturen wohl nicht von linearer Natur sind. Aus diesem Grund wollen wir uns zum Zwecke der Modellselektion nun nur noch mit der Klasse der GAMe befassen.

4.2 Modellierung als GAM

Auf Basis der in Abschnitt 4.1 gemachten Beobachtungen haben wir uns bereits dafür entschieden das Geschlecht des Versicherungsnehmers nicht in unser Modell aufzunehmen. Außerdem wissen wir auch, dass die Prädiktoren Tarifklasse (PG), Rabatt und Gruppe als Faktoren modelliert werden. Der erste Schritt in der Modellierung eines GAMs wird es nun sein, die Darstellungsformen von Alter, Laufzeit und Prämie zu bestimmen. Da die Vorgehensweise für jeden dieser Prädiktoren dieselbe ist, wollen wir sie im Folgenden exemplarisch am Alter des Versicherungsnehmers demonstrieren.

4.2.1 Die Darstellung stetiger Prädiktoren

Wollen wir einen stetigen Prädiktor wie in Abschnitt 3.3 beschrieben durch eine Glatte Kurve modellieren, so gibt es zwei Vorgaben die wir treffen müssen. Dies ist zum einen die Klasse und zum anderen die Dimension der Basisfunktionen. Wir entscheiden uns hier für eine kubische Splinebasis da dies der meist verbreitete Ansatz für derartige Anwendungen ist. Es sei auch erwähnt, dass (in unserem Anwendungsfall) für Glatte Funktionen von einem einzelnen Prädiktor alle im R-Paket *mgcv* zur Verfügung stehenden Arten von Basisfunktionen zu äquivalenten Modellen führen. Anders ist dies wenn das Modell eine Glatte Funktion von zwei oder mehreren stetigen Prädiktoren beinhaltet. In diesem Fall müsste z.B. untersucht werden, ob die so entstehende mehrdimensionale Fläche in alle Richtungen dieselbe „Festigkeit" besitzen darf oder nicht. In einfacheren Worten bedeutet dies zu entscheiden, ob die Glätte der Funktion entlang allen Dimensionen mit dem gleichen Strafterm gewichtet werden soll. Da wir uns vorläufig aber nur mit Glatten Funktionen von einer Variable befassen, belassen wir es bei den uns vertrauten kubischen Basisfunktionen.

Uns bleibt nun also noch die Wahl der Dimension dieser Splinebasis. Die einfachste Vorgehensweise um dafür einen passenden Wert zu bestimmen ist ein Modell mit einer beliebig gewählten Dimension zu schätzen, und dann den effektiven mit dem maximal möglichen Freiheitsgrad zu vergleichen. Ergibt sich dabei ein effektiver Freiheitsgrad der sehr nahe an der von uns gewählten maximalen Flexibilität liegt, ist es ratsam die Dimension der Splinebasis zu erhöhen. Das R-Paket *mgcv* bietet (unter anderem) zu diesem Zweck die Funktion *gam.check*. Wie bereits in Abschnitt 3.6 erwähnt, wird durch diese Funktion unter anderem ein informeller Test für den gewählten maximalen Freiheitsgrad durchgeführt. Grob gesprochen wird durch diesen Test überprüft, ob die Residuen noch eine Struktur aufweisen, die nicht durch das Modell erklärt wurde. Die Ausgabe dieser Funktion besteht aus dem maximal möglichen Freiheitsgrad *k'*, dem effektiven Freiheitsgrad *edf* sowie dem sogenannten *k-index* und einem *p*-Wert. Der *k-index* kann so interpretiert werden, dass umso weiter dieser unter Eins liegt, umso wahrscheinlicher ist eine in den Residuen verbliebene Struktur. Ein kleiner *p*-Wert zusammen mit einem *edf* der sehr nahe an *k'* liegt deutet darauf hin, dass die gewählte maximale Basisdimension zu klein ist.

Wir wollen nun im Folgenden die Basisdimension für das Alter des Versicherungsnehmers bestimmen. Zu diesem Zweck schätzen wir ein Modell in dem wir die Basisdimension als $k = 10$ angeben. Anschließend überprüfen wir mit *gam.check* ob diese Wahl von k angebracht war.

```
modell_alter<-bam(STORNO~s(VRALTER, bs="cr",k=10),
                  family=binomial,data=data2014)

gam.check(modell_alter)
              k'    edf     k-index  p-value
s(VRALTER)  9.000 8.476     0.916     0.56
```

Wir erkennen, dass der effektive Freiheitsgrad (edf) sehr nahe an der von uns gewählten maximalen Flexibilität von $k' = k - 1 = 9$ liegt (-1 da ein Freiheitsgrad durch den Intercept verloren geht). Aus diesem Grund setzen wir $k = 20$ und schätzen das Modell erneut. Nun ergibt sich ein effektiver Freiheitsgrad von $edf = 14.8$ bei $k' = 19$. Ähnliche Werte erhalten wir für die Kalenderjahre 2012 und 2013. Somit erscheint uns für die Modellierung des Alters eine Basisdimension von $k = 20$ als angebracht

Wir wissen also nun, wie wir das Alter als stetige Variable modellieren wollen. Unsere Wahl ist eine kubische Splinebasis mit Dimension zwanzig. Durch die gleiche Vorgangsweise erhalten wir auch die Darstellungsformen für die Laufzeit und die monatliche Prämie. Für beide Prädiktoren wählen wir ebenfalls eine kubische Splinebasis. Bei der monatlichen Prämie hat diese Basis eine Dimension von dreißig, bei der Laufzeit eine Dimension von fünfundzwanzig. Es sei an dieser Stelle noch einmal erwähnt, dass diese Werte lediglich den maximal möglichen Freiheitsgrad, also nicht den effektiven Freiheitsgrad, eines Modells definieren. Des weiteren wurden diese Werte so gewählt, dass der maximale Grad an Flexibilität für alle Beobachtungsjahre ausreichend groß ist.

Um einen ersten Eindruck dieser Glatten Funktion zu erhalten, werden wir das obige GAM *modell-alter* grafisch darstellen. Da wir dabei nur eine Glatte Funktion eines stetigen Prädiktors betrachtet haben, können wir die Modellschätzungen als Wahrscheinlichkeiten grafisch darstellen. Um dabei den Bezug zu den beobachteten Stornierungen herzustellen, haben wir für jede Altersstufe (alle natürliche Zahlen von 0 bis 105) den relativen Anteil an Stornierungen berechnet. Diese Werte wurden in Abbildung 16 als blaue Kreise markiert.

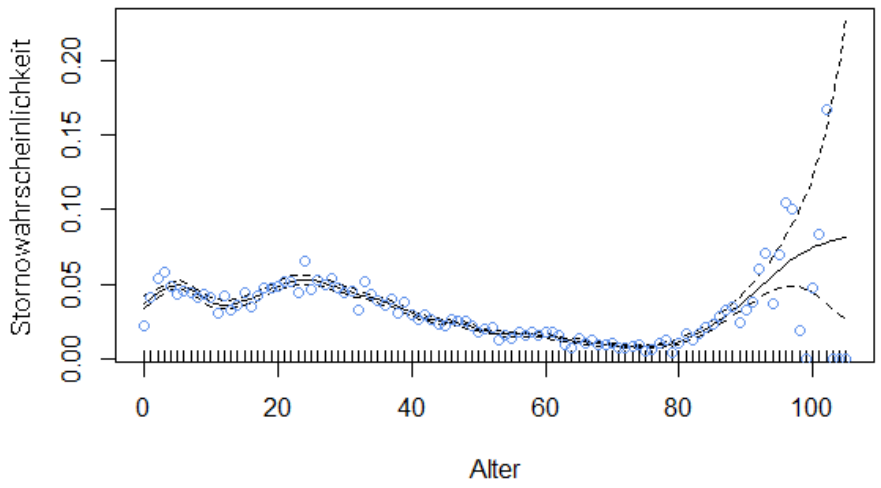

Abbildung 16: Geschätzte und beobachtete Stornowahrscheinlichkeiten in Abhängigkeit von Alter im Jahre 2014

Wir erkennen dabei gut, wie die Modellschätzung (die durchgezogene schwarze Linie) einen nichtlinearen Verlauf über das Alter beschreibt. Die gestrichelte Linie markiert dabei eine (geschätzte) Standardabweichung über und unter der Modellschätzung. Es lässt sich gut beobachten, dass das so resultierende Band bis zu den Altersstufen von 90 Jahren relativ schmal ist. Für die Altersklassen darüber ist naturgemäß die Datenbasis zu dünn. So gibt es Altersstufen über 100 Jahre in denen keine Stornierung beobachtet wurde während für andere an die 20% der Tarife storniert wurden. Dies drückt sich auch durch das immer breiter werdende Band der Standardabweichung aus.

Der Vollständigkeit halber präsentieren wir in den Abbildungen 17 und 18 die entsprechenden Schätzungen und Beobachtung für die Variablen Laufzeit und Prämie. Für die Laufzeiten der Versicherungstarife ist wieder ein ähnlicher Effekt wie beim Alter des Versicherungsnehmers zu erkennen. Das Band der geschätzten Standardabweichungen wird für sehr lange Laufzeiten zunehmend breiter. Dies ist bei den Prämien auf den ersten Blick nicht der Fall. Hier muss allerdings beachtet werden, dass die Darstellung in Abbildung 18 nur für Prämien bis 250 Euro reicht. Dies entspricht dem 99%-Quantil aller Prämien im Jahr 2014. Zusätzlich ist in Abbildung 18 zu beachten, dass die Prämien in von uns frei definierte Prämienklassen eingeteilt wurden. Diese Einteilung in Klassen, welche zur Berechnung des beobachteten Stornoanteils notwendig ist, wurde dabei so gewählt, dass eine möglichst gleichmäßige Verteilung der Tarife auf diese Klassen entsteht. All diesen Abbildungen ist gemein, dass sie auf Modellen basieren, welche jeweils nur den einzelnen stetigen Prädiktor beinhalten. Somit können die Abhängigkeitsstrukturen in einem Modell welches z.B. alle drei stetigen Prädiktoren beinhalten von diesen abweichen.

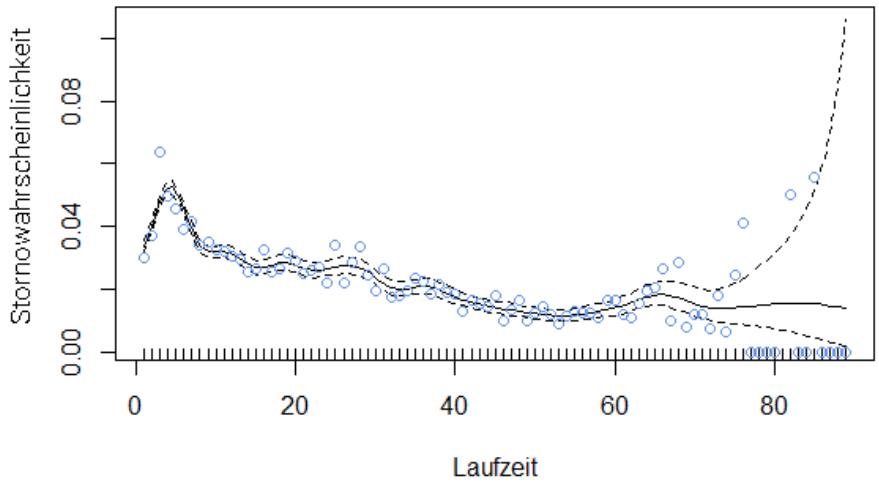

Abbildung 17: Geschätzte und beobachtete Stornowahrscheinlichkeiten in Abhängigkeit von Laufzeit im Jahre 2014

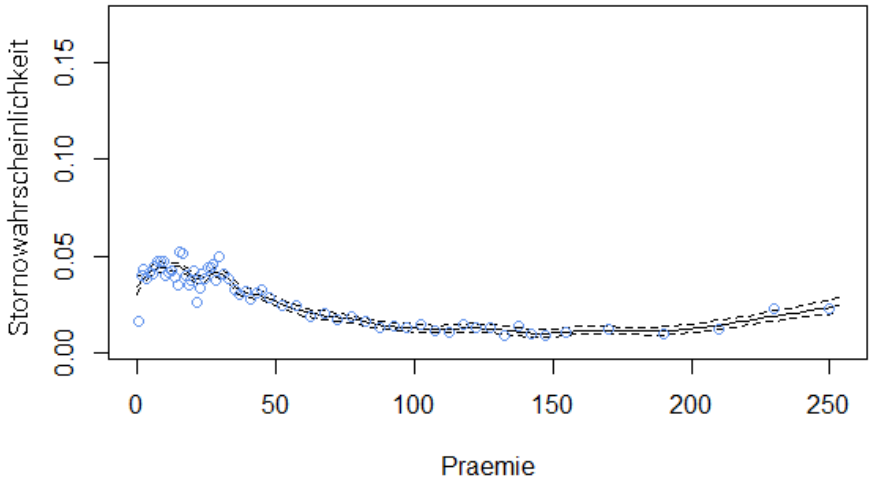

Abbildung 18: Geschätzte und beobachtete Stornowahrscheinlichkeiten in Abhängigkeit von Prämie im Jahre 2014

Die Transformation der geschätzten Werte des Linearen Prädiktors auf die Wahrscheinlichkeitsskala (wie in obigen Abbildung 16, 17 und 18) ist nur für Modelle möglich, die lediglich eine Glatte Funktion beinhalten. Enthält ein Modell zwei oder mehrere Glatte Funktionen, so müssten für eine ähnliche Darstellung jeweils alle restlichen stetigen Prädiktoren auf eine Ausprägung fixiert werden. Dadurch wäre keine anschauliche Interpretation der Modellschätzungen mehr möglich. Die geschätzten Glatten Funktionen können dann lediglich auf der Skala des Linearen Prädiktors dargestellt werden. Damit ist es aber auch nicht mehr möglich, die beobachteten Anteile an Stornierungen mit den Modellschätzungen zu vergleichen. Die Interpretation einer solchen grafischen Darstellung wird sich dann auf die Quantifizierung des Effektes eines stetigen Prädiktors auf die Wahrscheinlichkeit einer Stornierung beschränken.

4.2.2 Der Faktor Gruppe

Da wir nun wissen, wie wir die stetigen Prädiktoren modellieren, können wir jetzt das erste Modell der Stornowahrscheinlichkeiten schätzen. Die nächstliegendste Wahl für so ein Modell ist alle uns verbleibenden Prädiktoren additiv in das Modell aufzunehmen. Dieses erste Modell erhalten wir demnach durch folgenden Funktionsaufruf.

```
modell_add<-bam(STORNO~s(VRALTER, bs="cr",k=20)+
                       s(LAUFZEIT, bs="cr",k=25)+
                       s(PRAEMIE, bs="cr",k=30)+PG+RABATT+GRUPPE,
                       family=binomial,data=data2014)
```

In der *summary* dieses Modells für die Daten des Jahres 2014 erkennen wir dabei, dass sich der Koeffizient des Faktors Gruppe nicht signifikant von Null unterscheidet. Obwohl der entsprechende Koeffizient für die Jahre 2012 und 2013 sehr wohl signifikant ist, nehmen wir dies zum Anlass den Faktor Gruppe für die weitere Modellierung nicht mehr zu berücksichtigen. Der Grund dafür ist der von uns angestrebte Anwendungszweck des resultierenden Modells. In Abschnitt 2.1 haben wir bereits erwähnt, dass wir ein Modell suchen, welches wir auf den Daten der unterschiedlichen Kalenderjahre fitten können um dann anhand der resultierenden Modellschätzungen die Abhängigkeitsstrukturen zwischen Stornierungsverhalten und den Prädiktoren vergleichen zu können. Wir suchen also ein Modell, dessen Wahl zumindest auf allen uns zur Verfügung stehenden Daten vertretbar ist. Da alle anderen Prädiktoren in allen Beobachtungsjahren ausnahmslos hoch signifikant sind, erscheint uns der Faktor Gruppe aus diesem Grund als vernachlässigbar.

```
summary(modell_add)

Parametric coefficients:
             Estimate Std. Error  z value Pr(>|z|)
(Intercept) -3.05098    0.02808 -108.642  < 2e-16 ***
PGPG02a     -0.99703    0.04068  -24.507  < 2e-16 ***
PGPG02b     -1.37393    0.05245  -26.197  < 2e-16 ***
PGPG03       0.12357    0.04431    2.789  0.00529 **
RABATT=1    -0.32229    0.04838   -6.662 2.71e-11 ***
GRUPPE=N    -0.01844    0.02806   -0.657  0.51105

Approximate significance of smooth terms:
               edf Ref.df Chi.sq p-value
s(VRALTER)  12.809  15.05  972.7  <2e-16 ***
s(LAUFZEIT) 15.381  17.96  185.6  <2e-16 ***
s(PRAEMIE)   9.538  11.46  187.0  <2e-16 ***
```

Ein weiterer Grund dafür den Faktor Gruppe nicht zu betrachten, haben wir bereits in Abschnitt 2.2.6 gesehen. Dort haben wir festgestellt, dass sich im Jahr 2012 der relative Stornoanteil der Tarife die Teil einer Gruppenversicherung waren nicht sonderlich von jenem des gesamten Bestandes unterschieden hat. Ganz anders war dies in den Jahren 2013 und 2014. Im Jahr 2013 war der relative Stornoanteil der Gruppenversicherungen sehr viel geringer als im gesamten Bestand. Im Jahr 2014 beobachteten wir den genau umgekehrten Sachverhalt. Auffällig ist dies, da die Effekte der beiden anderen Faktoren (Tarifklasse und Rabatt) über die Beobachtungsjahre nicht gegenläufig waren, wie dies eben für den Faktor Gruppe der Fall ist. Somit scheint es naheliegend, dass die beobachteten Effekte für Gruppenversicherungen rein zufälliger Natur sind. Wir entschließen uns also Gruppe aus dem obigen Modell zu entfernen. Dadurch erhalten wir ein Modell, welches sich durch eine einfache Modellstruktur und die hohe Signifikanz aller Prädiktoren auszeichnet. Dieses Modell bezeichnen wir im Folgenden als Referenzmodell.

4.2.3 Das Referenzmodell

Unser Referenzmodell ergibt sich als das einfachste GAM welches die Prädiktoren Alter, Laufzeit und Prämie als Glatte Kurven und die Tarifklasse sowie Rabatt als Faktoren beinhaltet. Wie bereits im vorigen Abschnitt erwähnt, sind für dieses Modell in den Beobachtungsjahren 2012, 2013 und 2014 alle Prädiktoren hoch signifikant.

```
modell1<-bam(STORNO~s(VRALTER, bs="cr",k=20)+
                    s(LAUFZEIT, bs="cr",k=25)+
                    s(PRAEMIE, bs="cr",k=30)+PG+RABATT,
                    family=binomial, data=data2014)
```

Für das Jahr 2014 erhalten wir mit diesem Modell eine Reduktion der Deviance im Ausmaß von 5.85% (von 82265 auf 77454). Im Vergleich zu dem in Abschnitt 4.1 präsentierten GLM, welches die stetigen Prädiktoren als Faktoren modellierte, ist dies zwar ein geringerer Wert, jedoch bietet uns *modell1* demgegenüber große Vorteile. So reduziert sich zum einen die Modellkomplexität drastisch (die Anzahl der Freiheitsgrade beträgt nun 48). Bei etwa 300.000 Beobachtungen wirkt sich dies natürlich merklich auf die Rechenzeit der Modellierung aus. Zum anderen haben wir nun die stetigen Prädiktoren als Glatte Funktionen modelliert. Dies ermöglicht eine einfache Interpretation der Effekte den diese Prädiktoren auf die Stornowahrscheinlichkeit haben.

Wir wollen an dieser Stelle noch einmal die in Abschnitt 4.2.1 getroffenen Behauptungen überprüfen. Dort hieß es, dass die Wahl der Basisfunktionen im Falle von eindimensionalen Glatten Funktionen, wie sie in obigen Modell vorkommen, keinen ersichtlichen Einfluss auf das resultierende Modell hat. Dies ist auch für *modell1* der Fall, da wir auch für andere Basisfunktionen (festgelegt durch den Parameter *bs* im Funktionsaufruf) dieselben Modellschätzungen erhalten. Den zweiten Punkt, den es zu überprüfen gilt, stellt die Dimension dieser Basisfunktionen dar. Da wir die Wahl dieser Dimensionen basierend auf Modellen getroffen haben, welche jeweils nur einen stetigen Prädiktor berücksichtigten, sind diese auch für *modell1* ausreichend. Dies zeigt sich auch daran, dass die effektiven Freiheitsgrade der Glatten Funktionen nicht zu knapp bei den von uns definierten Werten liegen. Ganz im Gegenteil scheint es nun so, dass sich durch die gleichzeitige Modellierung aller stetigen Prädiktoren in einem Modell die Variabilität der einzelnen Glatten Funktionen verringert hat. So ist für das Jahr 2014 der effektive Freiheitsgrad der monatliche Prämie nur mehr bei etwa 9.5. Da eine Verringerung der in Abschnitt 4.2.1 bestimmten Dimensionen jedoch kaum Auswirkung auf die benötigte Rechenzeit hat, belassen wir es bei den gewählten Werten.

Das hier definierte Modell *modell1* soll uns im weiteren Verlauf als Vergleich (Referenz) dienen. Auf Basis dieses Modells werden wir überprüfen, ob durch Hinzunahme von Interaktionen eine Verbesserung der Modellgüte erreicht werden kann.

4.2.4 Faktorinteraktionen

Das in Abschnitt 4.2.3 präsentierte Modell *modell1* beinhaltet bereits alle von uns als relevant eingestuften Prädiktoren. Es bleibt also noch zu überprüfen, ob etwaige Interaktionen zwischen diesen Prädiktoren eine relevante Information über die Stornowahrscheinlichkeit enthalten. Dafür beschränken wir uns in diesem Abschnitt zunächst auf die Faktorinteraktionen, also jenen Wechselwirkungen, die die beiden Faktoren Tarifklasse und Rabatt betreffen.

Als Erstes können wir die Interaktion zwischen Tarifklasse und Rabatt ausschließen. Der einfache Grund dafür ist, dass für einzelne der dabei resultierenden acht Faktorstufen (vier Tarifklassen jeweils unterteilt in rabattierte und nicht rabattierte Tarife) sehr wenige oder in manchen Kalenderjahren auch gar keine Beobachtungen vorliegen.

Damit gilt es also festzustellen, ob die Interaktionen zwischen den drei stetigen Prädiktoren und jeweils einem der beiden Faktoren relevant sind. Ein solche Interaktion ist dabei so zu interpretieren, dass für jede Stufe des Faktors eine eigene Glatte Funktion für den stetigen Prädiktor geschätzt wird. Um dabei festzustellen welche Interaktion (mit Tarifklasse oder Rabatt) den höheren Informationsgehalt liefert, werden wir wie folgt vorgehen. Für jeden der drei stetigen Prädiktoren schätzen wir jeweils zwei Modelle. Beide Modelle beinhalten die Tarifklasse und Rabatt als additive Faktoren. Das eine Modell enthält den jeweiligen stetigen Prädiktor in Interaktion mit Rabatt, das andere Modell die Interaktion mit der Tarifklasse. Anschließend vergleichen wir die beiden Modelle anhand der Reduktion der Deviance, des Akaike Informationskriteriums (AIC) sowie des Bayesschen Informationskriteriums (BIC). Diese Informationskriterien sind so konstruiert, dass kleinere Werte zu bevorzugen sind. Das AIC ist dabei nichts anderes als die Summe des negativen doppelten Loglikelihoodwertes und der doppelten Anzahl an zu schätzenden Parametern. Im Unterschied dazu wird im BIC die Anzahl der Parameter nicht mit dem Wert Zwei sondern mit der logarithmierten Anzahl an Beobachtungen multipliziert. Dies hat zur Folge, dass im BIC die Anzahl an zu schätzenden Parametern stärker gewichtet ist und somit gegenüber dem AIC eher einfachere Modelle bevorzugt werden.

Für den stetigen Prädiktor Alter sind die entsprechenden Modelle demnach wie folgt gegeben. Das Modell *modella1*, welches die Interaktion zwischen Tarifklasse und Alter beinhaltet, ist dabei das komplexere Modell. Der Grund hierfür ist, dass in diesem Modell vier Glatte Funktionen (für jede der vier Tarifklassen) geschätzt werden. In *modella2* sind dies nur zwei, eine für Tarife mit Rabatt und eine für jene ohne.

```
modella1<-bam(STORNO~s(VRALTER, bs="cr",k=20,by=PG)+PG+RABATT)
modella2<-bam(STORNO~s(VRALTER, bs="cr",k=20,by=RABATT)+PG+RABATT)
```

In Tabelle 4 sind die entsprechenden Kennzahlen dieser Modelle abgebildet. Dabei wurden die Kennzahlen des jeweils zu bevorzugenden Modells gekennzeichnet. Es ist gut ersichtlich,

dass das komplexere Modell *modella1* in allen Jahren eine höhere Reduktion der Deviance aufweist. Die höchste Differenz beträgt dabei 0.17%. Wie zu erwarten war, bevorzugt das AIC das komplexere Modelle *modella1* während das BIC das einfachere Modell *modella2* präferiert. Das Jahr 2012 stellt dabei eine Ausnahme dar. Bei genauerer Betrachtung fällt jedoch auf, dass die beiden Werte des AIC in diesem Jahr sehr nahe beieinander liegen. Somit können wir für das Jahr 2012 auf Basis des AIC keine Aussage treffen.

Daten	modella1	modella2	modella1	modella2	modella1	modella2
	Deviance-Reduktion		AIC		BIC	
2014	**5.48%**	5.38%	**77841.3**	77902	78292.2	**78214**
2013	**4.06%**	3.89%	**71356.5**	71442.8	71838.2	**71726.4**
2012	**3.63%**	3.59%	75371.7	**75367.5**	75871.8	**75700.3**

Tabelle 4: Modelle mit Faktorinteraktionen für Alter

Wir wiederholen diese Vorgehensweise nun für die Laufzeit der Tarifpositionen. Die Modelle wurden wieder entsprechend definiert, wobei das als *modelll1* bezeichnete Modell die Interaktion mit PG und *modelll2* die Interaktion mit Rabatt enthält.

```
modelll1<-bam(STORNO~s(LAUFZEIT, bs="cr",k=25,by=PG)+PG+RABATT)
modelll2<-bam(STORNO~s(LAUFZEIT, bs="cr",k=25,by=RABATT)+PG+RABATT)
```

In Tabelle 5 sind die entsprechenden Kennzahlen für die Modelle der Faktorinteraktionen von Laufzeit abgebildet. Wir erkennen wieder ein ähnliches Muster wie wir es bereits für die Faktorinteraktionen des Prädiktors Alter gemacht haben. Das komplexere Modell *modelll1* scheint eine höhere Reduktion der Deviance zu bieten und wird vom AIC bevorzugt. Ausnahme ist wieder das Beobachtungsjahr 2012. In diesem Jahr wird das einfachere Modell *modelll2* von allen betrachteten Kennzahlen bevorzugt. Das BIC bevorzugt in allen Jahren die Interaktion mit Rabatt (*modelll2*).

Daten	modelll1	modelll2	modelll1	modelll2	modelll1	modelll2
	Deviance-Reduktion		AIC		BIC	
2014	**4.67%**	4.47%	**78532.2**	78649.9	79108.1	**78984.6**
2013	**3.45%**	3.23%	**71827.4**	71946.6	72431.7	**72302.7**
2012	5.12%	**5.18%**	74242.9	**74140.8**	74941.4	**74540.6**

Tabelle 5: Modelle mit Faktorinteraktionen für Laufzeit

Analog zu oben wiederholen wir die entsprechenden Modellschätzungen und Auswertungen nun für die monatliche Prämie. Die in Tabelle 6 abgebildeten Kennzahlen zeigen dasselbe

Muster wie jene für Laufzeit in Tabelle 5. Mit Ausnahme des Jahres 2012 scheint es wieder so, als ob das komplexere Modell (Interaktion mit Tarifklasse) die höhere Reduktion der Deviance aufweisen kann während das einfachere Modell (Interaktion mit Rabatt) vom BIC bevorzugt wird.

```
modellp1<-bam(STORNO~s(PRAEMIE, bs="cr",k=30,by=PG)+PG+RABATT)
modellp2<-bam(STORNO~s(PRAEMIE, bs="cr",k=30,by=RABATT)+PG+RABATT)
```

Daten	modellp1	modellp2	modellp1	modellp2	modellp1	modellp2
	Deviance-Reduktion		AIC		BIC	
2014	**3.85%**	3.68%	**79163.6**	79273.2	79500.4	**79444.5**
2013	**2.31%**	2.19%	**72624.2**	72691.7	72952.5	**72907.4**
2012	2.74%	**2.79%**	76074.3	**76038.1**	76627.7	**76579.3**

Tabelle 6: Modelle mit Faktorinteraktionen für Prämie

Ein weiterer interessanter Aspekt ergibt sich durch den Vergleich der in den Tabellen 4, 5 und 6 angeführten Deviance-Reduktionen. Anhand dieser Kennzahlen ergibt sich für jedes Beobachtungsjahr eine Reihung der stetigen Prädiktoren anhand ihres Informationsgehaltes. In den Jahren 2013 und 2014 ist demnach das Alter vor der Laufzeit und der monatlichen Prämie jener stetige Prädiktor, welcher das Stornoverhalten am besten erklärt. Diese Reihung ist dabei unabhängig davon, ob die Interaktion mit der Tarifklasse oder mit Rabatt im Modell berücksichtigt wurde. Das Jahr 2012 stellt wieder eine Ausnahme dar. Hier scheint es so, als hätte die Laufzeit den höchsten Informationsgehalt zur Stornowahrscheinlichkeit. Des weiteren erkennen wir, dass in allen Jahren die monatliche Prämie die geringste Deviance-Reduktion dieser drei stetigen Prädiktoren aufweist.

Aus den obigen Beobachtungen ziehen wir folgende Schlüsse. Für alle stetigen Prädiktoren hat sich ein ähnliches Muster ergeben. Die Interaktion mit der Tarifklasse liefert eine höhere Reduktion der Deviance während die Interaktion mit Rabatt den bessere Kompromiss zwischen Modellkomplexität und Anpassung liefert. Wir können somit ausschließen, dass z.B. für einen dieser stetigen Prädiktoren eine der beiden Faktorinteraktionen klar zu bevorzugen ist während dies für einen anderen stetigen Prädiktor nicht der Fall ist. Auch haben wir weitere Hinweise darauf gefunden, dass sich das Stornoverhalten im Jahr 2012 stärker von den anderen Daten unterscheidet wie dies für die Jahre 2013 und 2014 der Fall ist.

Für das weitere Vorgehen bedeutet dies, dass wir uns bezüglich der Faktorinteraktionen auf die folgenden Modelle *modell2* und *modell3* beschränken. Ersteres beinhaltet dabei alle stetigen Prädiktoren in Interaktion mit der Tarifklasse. Dieses Modell besitzt insgesamt also 12 Glatte Kurven. Das Modell *modell3* beinhaltet demgegenüber alle stetigen Prädiktoren in Interaktion mit Rabatt. Somit müssen für dieses Modell 6 Glatte Kurven geschätzt werden.

```
modell2<-bam(STORNO~s(VRALTER, bs="cr",k=20,by=PG)+
                    s(LAUFZEIT, bs="cr",k=25,by=PG)+
                    s(PRAEMIE, bs="cr",k=30,by=PG)+PG+RABATT,
                    family=binomial)

modell3<-bam(STORNO~s(VRALTER, bs="cr",k=20,by=RABATT)+
                    s(LAUFZEIT, bs="cr",k=25,by=RABATT)+
                    s(PRAEMIE, bs="cr",k=30,by=RABATT)+PG+RABATT,
                    family=binomial)
```

Für diese beiden Modelle können wir natürlich wieder die gleichen Kennzahlen wie zuvor berechnen. In Tabelle 7 sind die entsprechenden Werte der Deviance-Reduktion, des AIC und des BIC eingetragen. Dabei zeigt sich wieder, dass das komplexere Modell wohl die höhere Anpassung an die Daten liefert, jedoch die zusätzliche Modellkomplexität dadurch nicht gerechtfertigt zu sein scheint. An dieser Stelle wollen wir uns aber noch nicht auf ein Modell festlegen. Dies erfolgt in Abschnitt 4.3, wo wir zu diesem Zwecke der Modellauswahl individuelle Evaluierungsmethoden präsentieren werden.

Daten	modell2	modell3	modell2	modell3	modell2	modell3
	Deviance-Reduktion		AIC		BIC	
2014	**6.22%**	5.98%	**77351.8**	77484.3	78421.8	**78204.4**
2013	**4.92%**	4.67%	**70821.7**	70938	71874.7	**71618.8**
2012	**7.18%**	7.1%	72794	**72771.2**	74323.7	**73864.7**

Tabelle 7: Modelle mit Faktorinteraktionen

Wir haben also festgestellt, dass entweder eine Faktorinteraktion mit der Tarifklasse oder mit Rabatt beachtet werden soll. Ausschließen wollen wir hiermit Mischformen, also z.B. ein Modell, welches eine Interaktion zwischen Laufzeit und Rabatt sowie zwischen Alter und Tarifklasse beinhaltet. Im nun folgenden Abschnitt wollen wir uns noch mit der Möglichkeit einer Interaktion zwischen zwei stetigen Prädiktoren beschäftigen.

4.2.5 Interaktionen zwischen zwei stetigen Prädiktoren

Nachdem wir im vorigen Abschnitt mögliche Interaktionen zwischen den stetigen Prädiktoren und den Faktoren betrachtet haben, wenden wir uns in diesem Abschnitt den Interaktionen zwischen jeweils zwei der stetigen Prädiktoren zu. Unser Ziel ist es dabei das erfolgversprechendste Modell, welches eine Interaktion zwischen zwei stetigen Prädiktoren beinhaltet, zu bestimmen. Dieses Modell, oder diese Modelle, werden wir dann in der Modellevaluierung in Abschnitt 4.3 mit den bereits bestimmten Modellen *modell1*, *modell2* und *modell3* vergleichen.

Bei Interaktionen zwischen zwei oder mehreren Prädiktoren, die alle als Glatte Funktionen modelliert wurden, gilt es im Vergleich zu einfachen Glatten Funktionen zusätzliche Aspekte zu betrachten. So erhöht sich der Rechenaufwand, da z.B. für zwei stetige Prädiktoren nun nicht nur zwei eindimensionale Glatte Funktionen sondern eine Glatte Fläche geschätzt werden muss. Des weiteren muss bestimmt werden, ob diese Fläche überall gleich steif ist oder ob sich die Flexibilität in den beiden Dimensionen unterscheidet. Für unseren Anwendungsfall hat dies zur Folge, dass wir uns auf Interaktionen zwischen zwei der drei stetigen Prädiktoren beschränken werden. Das Modell mit der dreifach Interaktion wäre sehr aufwändig und eine einfache grafische Interpretation der Ergebnisse nicht mehr möglich. Bei den drei uns gegebenen stetigen Prädiktoren Alter, Laufzeit und Prämie beschränken wir uns somit auf drei Modelle die jeweils eine Interaktion zwischen zwei dieser drei Prädiktoren beinhalten. Da wir in Abschnitt 4.2.4 bereits festgestellt haben, dass die Faktorinteraktion mit Rabatt wohl effizienter als jene mit der Tarifklasse ist, behalten wir diese Interaktion in den hier betrachteten drei Modellen bei[5]. Außerdem entscheiden wir uns dafür, die Glatten Flächen durch ein sogenanntes Tensorprodukt zu modellieren. Dies hat zur Folge, dass sich die Flexibilität dieser Fläche entlang beider Koordinaten unterscheiden kann. Somit ergeben sich die drei in diesem Abschnitt zu vergleichenden Modelle als

```
modellal<-bam(STORNO~te(VRALTER,LAUFZEIT,k=6,by=RABATT)+
              s(PRAEMIE,bs="cr",k=30,by=RABATT)+PG+RABATT,
              family=binomial)

modellap<-bam(STORNO~te(VRALTER,PRAEMIE,k=6,by=RABATT)+
              s(LAUFZEIT,bs="cr",k=25,by=RABATT)+PG+RABATT,
              family=binomial)

modellpl<-bam(STORNO~te(PRAEMIE,LAUFZEIT,k=6,by=RABATT)+
              s(VRALTER,bs="cr",k=20,by=RABATT)+PG+RABATT,
              family=binomial)
```

Jedes dieser Modelle besitzt also ein Tensorprodukt, welches aus jeweils zwei der drei stetigen Prädiktoren besteht, sowie dem dritten Prädiktor als einfache Glatte Funktion. Diese Glatten Schätzungen wurden jeweils einmal für jede der beiden Faktorstufen von Rabatt geschätzt. Zusätzlich sind natürlich noch die beiden Faktoren Tarifklasse und Rabatt inkludiert. In allen drei Modellen haben wir als Kompromiss zwischen (Rechen-) Komplexität und Datenanpassung den maximalen Freiheitsgrad jeder Koordinate im Tensorprodukt als $k = 6$ definiert. Auch die Analyse mit Hilfe der Funktion *gam.check* zeigt wieder, dass der resultierende maximale Freiheitsgrad von 35 ($= 6 \cdot 6 - 1$) für alle Glatten Flächen hinreichend groß bemessen ist. Für die jeweiligen Glatten Funktionen haben wir die in Abschnitt 4.2.1 bestimmten Werte für k unverändert gelassen.

[5]In Abschnitt 4.3 werden wir sehen, dass die Interaktion mit der Tarifklasse gegenüber Rabatt tatsächlich keine relevanten Vorteile bringt

Daten	modellal	modellap	modellpl
	Deviance-Reduktion		
2014	5.74%	**6.01%**	5.95%
2013	4.34%	**4.69%**	4.51%
2012	5.9%	**7.15%**	5.35%
	AIC		
2014	77642	**77460.7**	77484.6
2013	71216.3	**70921.7**	71030.4
2012	73682.5	**72703.1**	74076.3
	BIC		
2014	78186.4	78216.6	**78088**
2013	72069.3	71597.7	**71567.5**
2012	74617.22	**73620.42**	74842.24

Tabelle 8: Modelle mit Interaktionen der stetigen Prädiktoren

In Tabelle 8 vergleichen wir diese Modelle wieder anhand ihrer Kennzahlen. In allen Beobachtungsjahren weist *modellap* die höchste Reduktion der Deviance auf. Dementsprechend wird dieses Modell auch vom AIC in allen Jahren bevorzugt. Dieses Modell beinhaltet die Interaktion zwischen Alter und Prämie. Bezüglich des BIC sehen wir, dass in den Jahren 2013 und 2014 *modellpl*, welches die Interaktion zwischen Laufzeit und Prämie beinhaltet, bevorzugt wird. Da im Jahre 2013 der Unterschied zwischen den BIC-Werten von *modellap* und *modellpl* allerdings sehr gering ist und der tatsächliche Rechenaufwand beider Modelle annähernd gleich ist, kommen wir trotzdem zu dem Schluss, dass *modellap* das Interessanteste dieser drei Modelle ist. Aus diesem Grund ist es dieses Modell welches wir mit den in den vorigen Abschnitten bestimmten Modellen *modell1*, *modell2* und *modell3* vergleichen wollen.

```
modell4<-bam(STORNO~te(VRALTER,PRAEMIE,k=6,by=RABATT)+
               s(LAUFZEIT,bs="cr",k=25,by=RABATT)+PG+RABATT,
               family=binomial)
```

Wir haben somit ein viertes Modell (*modell4*) gefunden. Dieses Modell wollen wir im nun folgenden Abschnitt zusammen mit den bereits zuvor bestimmten Modellen näher betrachten. Unser Ziel ist es dabei, auf Basis der Prädiktionsgüte eines dieser Modelle für die Modellierung der Stornowahrscheinlichkeit zu wählen.

4.3 Modellevaluation

In diesem Abschnitt gilt es nun ein konkretes Modell zu wählen. Auf Basis der klassischen Kennzahlen (AIC, BIC und Deviance) haben wir bisher vier interessante Generalisierte Additive Modelle bestimmt. Diese beinhalten alle dieselben Prädiktoren, unterscheiden sich

jedoch in der Abhängigkeitsstruktur, also der Modellformel. Wie bereits erwähnt, unterscheidet sich die Evaluierung solcher logistischer Regressionsmodelle von jener von klassischen Linearen Modellen, GLMen oder nicht logistischen GAMen. Der Grund dafür ist, dass sich die Schätzungen für die Stornowahrscheinlichkeiten, die diese Modelle erlauben, nur bedingt mit den binären Beobachtungen (storniert oder nicht storniert) vergleichen lassen. Um einen direkten Vergleich trotzdem zu ermöglichen, werden wir die Schätzungen der Stornowahrscheinlichkeiten dafür nutzen, um die Tarife als storniert oder nicht storniert zu schätzen. Dabei werden wir jene Tarife als storniert betrachten, welche eine sehr hohe geschätzte Stornowahrscheinlichkeit aufweisen. In unseren Daten bewegen sich die geschätzten Wahrscheinlichkeiten jedoch vorwiegend im einstelligen Prozentbereich. Eine solche Klassifikation der Tarife als storniert oder nicht storniert ist aus diesem Grund mit Vorsicht zu genießen.

Da wir also aufgrund der binären Beobachtungen nur bedingt Aussagen über die Anpassung der Modelle an die Daten machen können, beschäftigen wir uns in diesem Abschnitt auch mit der Bestimmung der Prädiktionsgüte. In Abschnitt 4.3.1 präsentieren wir dafür Methoden, deren Vorgehensweise der klassischen Kreuzvalidierung entsprechen. Damit werden wir in Abschnitt 4.3.3 die bisher bestimmten Modelle vergleichen und uns für ein konkretes Modell entscheiden.

4.3.1 Kreuzvalidierung

In diesem Abschnitt wollen wir Methoden zur Quantifizierung der Prädiktionsgüte unserer logistischen Regressionsmodelle vorstellen. Ziel ist es also zu bestimmen, wie gut ein Modell die Stornowahrscheinlichkeit für Daten schätzen kann, welche nicht bei der Modellschätzung zur Verfügung standen. Die offensichtlichste Möglichkeit dafür wäre natürlich ein Modell, welches z.B. auf den Daten von 2012 geschätzt wurde, zur Prädiktion der Responsevariablen des Jahres 2013 zu verwenden und die Ergebnisse mit den tatsächlichen Beobachtungen zu vergleichen. Bei dieser Vorgehensweise gilt es aber zwei Punkte zu betrachten. Zum einen müssen wir immer noch klären wie wir geschätzte Stornowahrscheinlichkeiten und beobachtete Stornierungen vergleichen können. Zum anderen haben wir Daten von „nur" drei Jahren zur Verfügung. Ein Vergleich anhand von nur drei Datenmengen erhöht allerdings die Gefahr, dass unsere Entscheidung durch zufällige Abweichungen beeinflusst wird. Aus diesem Grunde benutzen wir das Verfahren der Kreuzvalidierung.

Damit wir beim Vergleich der Prädiktionsgüten unserer Modelle verlässlichere Ergebnisse bekommen, unterteilen wir unsere Datensätze der Jahre 2012, 2013 und 2014 weiter. Die Daten jedes Jahres werden dafür zufällig in zehn gleichgroße Blöcke (zu je 10%) eingeteilt. Mit den resultierenden (dreißig) Datenmengen setzen wir zwei Testmethoden um.

- Testmethode 1: In der ersten Methode wird das jeweilige Modell mit Hilfe von neun der zehn Datenblöcke eines Jahres geschätzt. Anschließend werden die Schätzungen

dieses Modells für die Responsevariablen des zehnten Blockes mit den tatsächlichen Beobachtungen verglichen. Dieses Schema wird für jedes Jahr zehnmal durchgeführt. Somit wird jeder einzelne Datensatz neunmal zur Modellschätzung und einmal zur Modellevaluierung verwendet.

- Testmethode 2: Hier wird das jeweilige Modell unter Verwendung aller Daten eines einzelnen Kalenderjahres geschätzt. Für alle der zehn Datenblöcke des Folgejahres werden dann die Schätzungen der Responsevariablen mit den tatsächlichen Beobachtungen verglichen. Somit wird in dieser Testmethode das jeweilige Modell nur einmal auf den gesamten Daten eines Kalenderjahres geschätzt, die Prädiktionsgüte aber auf den zehn unterschiedlichen Datenblöcken des Folgejahres ausgewertet.

Der Grund für diese Vorgehensweise ist unsere Methode zum Vergleich der geschätzten Wahrscheinlichkeiten mit den tatsächlichen Stornierungen. Dieser Vergleich erfolgt aufgrund der oben beschriebenen unterschiedlichen Skalen nicht an einzelnen Tarifen sondern anhand von mehreren Tarifen zugleich. Konkret wird dabei die tatsächliche Anzahl an Stornierungen in der Kontrollmenge mit der Summe der geschätzten Stornowahrscheinlichkeiten verglichen. Da es sich bei den Stornierungen um eine binäre Variable handelt, ergibt die Summe der geschätzten Stornowahrscheinlichkeiten eine Schätzung für den Erwartungswert der Anzahl an Stornierungen. Diese Schätzung können wir nun sehr wohl mit der beobachteten Anzahl an Stornierungen vergleichen.

Für die, wie oben beschriebenen, zehn Testdatensätze eines einzelnen Jahres, bestehend aus n_j ($j \in \{1, \ldots, 10\}$) Tarifen, seien die geschätzten Stornowahrscheinlichkeiten als $\hat{p}_i$ und die beobachteten Stornierungen als $\text{STORNO}_i \in \{0, 1\}$ ($i \in \{1, \ldots, n_j\}$) gegeben. Damit lässt sich die sogenannte Maßzahl der relativen Abweichung der Stornierungen (RAS) schreiben als

$$\text{RAS}_j := \frac{\sum_{i=1}^{n_j} \hat{p}_i - \sum_{i=1}^{n_j} \text{STORNO}_i}{\sum_{i=1}^{n_j} \text{STORNO}_i} \quad j \in \{1, \ldots, 10\}. \tag{31}$$

Wir vergleichen also unsere Schätzung mit der Beobachtung und bringen die Abweichung in Relation zu der Zielgröße, der beobachteten Anzahl an Stornierungen. Dabei entspricht der Zähler nichts anderem als einer Summe von Residuen, da

$$\sum_{i=1}^{n_j} \hat{p}_i - \sum_{i=1}^{n_j} \text{STORNO}_i = \sum_{i=1}^{n_j} \left(\hat{p}_i - \text{STORNO}_i\right).$$

Aufgrund des geringen Stornoanteils sind die geschätzten Wahrscheinlichkeiten $\hat{p}_i$ in der Regel sehr klein und viele der Beobachtungen STORNO_i werden Null sein. Somit wird sich diese Summe aus vielen kleinen positiven Termen und wenigen großen negativen Termen zusammensetzen. Einzeln für sich betrachtet geben uns diese kleinen positiven oder großen

negativen Residuen wenig Aufschluss über die Modellgüte. Durch das Summieren dieser Residuen bewerten wir anhand von RAS die Prädiktionsgüte eines Modelles auf Basis des Bestandes und nicht auf Einzelvertragsebene. Da wir uns auch dafür interessieren ob ein Modell dazu neigt die Anzahl an Stornierungen zu über- oder unterschätzen verzichten wir an dieser Stelle darauf z.B. den Absolutbetrag oder die quadrierten Residuen zu betrachten.

Durch jede der oben beschriebenen Testmethoden erhalten wir also für alle zu betrachtende Modell jeweils zehn Werte dieser Kennzahl. Anhand von Lage und Streuung dieser Werte kann dann die Prädiktionsgüte bezüglich der Anzahl an Stornierungen verglichen werden.

Der konkrete Verwendungszweck unserer Ergebnisse motiviert eine weitere Kennzahl wie die oben eingeführte relative Abweichung der Stornierungen. Da eine Vertragsauflösung ein finanzielles Risiko für das Versicherungsunternehmen darstellt (die monatliche Prämie fällt weg), wollen wir die Modelle auch auf diesen Aspekt hin vergleichen. Zu diesem Zwecke vergleichen wir die Summe der stornierten monatlichen Prämien mit ihrer Schätzung. Diese Schätzung erhalten wir, indem wir den Erwartungswert der stornierten Prämien auf Basis der geschätzten Wahrscheinlichkeiten berechnen. Ist die monatliche Prämie eines Tarifes durch PRAEMIE_i gegeben, so berechnet sich die relativen Abweichung der stornierten Prämien (RAP) als

$$\text{RAP}_j := \frac{\sum_{i=1}^{n_j} \hat{p}_i \cdot \text{PRAEMIE}_i - \sum_{i=1}^{n_j} \text{STORNO}_i \cdot \text{PRAEMIE}_i}{\sum_{i=1}^{n_j} \text{STORNO}_i \cdot \text{PRAEMIE}_i} \quad j \in \{1, \ldots, 10\}. \tag{32}$$

Wieder erkennen wir im Zähler die Summe der Residuen, die nun durch die Prämien gewichtet werden. Analog zu oben lässt sich ein einzelnes (gewichtetes) Residuum nur schwer interpretieren. Auf Bestandsebene haben wir so jedoch eine weitere, durch die Prämiengewichtung spezifisch für unseren Anwendungsfall konzipierte Maßzahl für die Prädiktionsgüte, für die gilt

$$\sum_{i=1}^{n_j} \hat{p}_i \cdot \text{PRAEMIE}_i - \sum_{i=1}^{n_j} \text{STORNO}_i \cdot \text{PRAEMIE}_i = \sum_{i=1}^{n_j} (\hat{p}_i - \text{STORNO}_i) \cdot \text{PRAEMIE}_i.$$

Pro Testmethode erhalten wir auf diese Art und Weise jeweils zehn Werte von RAP für jedes zu überprüfende Modell. Analog zu den RAS-Werten werden diese wieder anhand ihrer Lage und Streuung verglichen. Mit Hilfe dieser beiden Kennzahlen (RAS und RAP) können wir nun die Prädiktionsgüte von verschiedenen Modellen auf der Bestandsebene vergleichen. Auf Basis dieser Ergebnisse werden wir uns in Abschnitt 4.3.3 für ein konkretes Modell entscheiden.

4.3.2 Klassifikation

Um das Problem der unterschiedlichen Skalen von Modellschätzungen und Beobachtungen zu umgehen, präsentieren wir in diesem Abschnitt eine Methode zur Klassifikation der Schätzungen. Dabei wollen wir eine einzelne Vertragsposition entweder als geschätzt storniert oder als geschätzt nicht storniert klassifizieren. Auf Basis einer solchen Einteilung können wir dann auf Einzelvertragsbasis die Beobachtung der Responsevariable direkt mit der so festgelegten Klasse der Schätzung vergleichen.

Um die Tarife zu klassifizieren definieren wir einen Schwellwert $c \in (0,1)$. Ein Tarif wird dann als storniert klassifiziert, wenn die zugehörige geschätzte Stornowahrscheinlichkeit $\hat{p}$ über diesem Schwellwert liegt, also $\hat{p} > c$. Umgekehrt wird der Tarif als nicht storniert klassifiziert, wenn $\hat{p}$ kleiner gleich dem Schwellwert ist, also $\hat{p} \leq c$.

Ein zentraler Aspekt dieser Vorgehensweise ist natürlich die Wahl des Schwellwerts. Um dabei ein möglichst optimales c (im Sinne von korrekt klassifzierten Tarifen) zu wählen, brauchen wir eine Maßzahl für die Güte einer solchen Klassifikation. Anschließend werden wir jenes $c \in (0,1)$ wählen, welches die so definierte Maßzahl maximiert. Für gegebene Modellschätzungen der Stornowahrscheinlichkeiten und für ein fixes $c \in (0,1)$ lässt sich der Bestand an Verträgen in vier Teilmengen, wie sie in Tabelle 9 dargestellt sind, einteilen.

	Beobachtung	
Schätzung	storniert	nicht storniert
storniert	A	C
nicht storniert	B	D

Tabelle 9: Kontingenztabelle der Verträge nach Storno

Die Zahl A beschreibt dabei die Anzahl jener Verträge, die im betrachteten Kalenderjahr storniert wurden und deren geschätzte Stornowahrscheinlichkeit größer als c war. Diese Tarife wurden also richtiger Weise als storniert eingeschätzt. Die Anzahl jener Verträge die storniert wurden aber durch die oben beschriebene Methode als nicht storniert klassifiziert wurden, wird durch B angegeben. C und D sind analog die Anzahl der falsch bzw. richtig klassifizierten nicht stornierten Verträge.

Nun können wir den Anteil an richtig klassifizierten stornierten Verträgen berechnen. Dieser Wert wird auch als Sensitivität einer Klassifikation bezeichnet. Die Sensitivität in Abhängigkeit des Schwellwerts c ergibt sich somit als

$$Sen(c) = \frac{A(c)}{A(c) + B(c)}.$$

Hierbei entsprechen $A(c)$ und $B(c)$ in ihrer Definition A und B aus Tabelle 9, wobei hier

die Abhängigkeit vom Schwellwert c explizit dargestellt wird.

Analog können wir auch den Anteil an richtig klassifizierten nicht stornierten Verträgen berechnen. Dieser Anteil wird auch als Spezifität bezeichnet und ergibt sich, wiederum in Abhängigkeit von c, als

$$Spe(c) = \frac{D(c)}{C(c) + D(c)}.$$

Offensichtlich gilt sowohl $Sen(c) \in [0, 1]$ als auch $Spe(c) \in [0, 1]$. Um nun die Güte einer Klassifikation in einer einzelnen Maßzahl zusammenzufassen benutzen wir den sogenannten Youden-Index, wie er in Youden (1950) beschrieben wird. Der Youden-Index $J(c)$ berechnet sich aus der Sensitivität $Sen(c)$ und der Spezifität $Spe(c)$ als

$$J(c) = Sen(c) + Spe(c) - 1.$$

Dieser Index nimmt nur Werte im abgeschlossenen Intervall $[0, 1]$ an, also $0 \leq J(c) \leq 1$. Die obere Schranke ist offensichtlich, da nach obiger Definition sowohl $Sen(c) \leq 1$ als auch $Spe(c) \leq 1$ gilt. Wäre der Youden-Index einer Klassifikation negativ, so könnte die getroffene Einteilung einfach umgekehrt werden, wodurch der Youden-Index der neuen Klassifikation positiv wäre. Bei dieser Umkehrung wird jeder Vertrag der zuvor als storniert geschätzt wurde, nun als nicht storniert geschätzt, und umgekehrt. Dies entspricht einer einfachen Vertauschung der beiden Zeilen aus Tabelle 9. Dass der Youden-Index durch diesen Zeilentausch sein Vorzeichen dreht, lässt sich durch folgende einfache Umformungen belegen. Es gilt

$$\begin{aligned}
J(c) = Sen(c) + Spe(c) - 1 &< 0 \\
\frac{A(c)}{A(c) + B(c)} + \frac{D(c)}{C(c) + D(c)} - 1 &< 0 \\
\frac{A(c) + B(c) - B(c)}{A(c) + B(c)} + \frac{D(c)}{C(c) + D(c)} - \frac{C(c) + D(c)}{C(c) + D(c)} &< 0 \\
1 - \frac{B(c)}{A(c) + B(c)} &< \frac{C(c)}{C(c) + D(c)} \\
0 &< \frac{B(c)}{A(c) + B(c)} + \frac{C(c)}{C(c) + D(c)} - 1.
\end{aligned}$$

Auf der rechten Seite der letzten Ungleichung steht nun nichts anderes als der Youden-Index der umgekehrte Klassifikation.

Eine Einteilung, in welcher entweder alle oder keiner der Verträge als storniert klassifiziert werden, hat einen Youden-Index von Null. Dies liegt daran, dass für eine solche Einteilung entweder die Sensitivität gleich Eins und die Spezifität gleich Null ist oder Umgekehrtes

gilt. Somit hat das Nullmodell, in dem alle Tarife die gleiche geschätzte Stornowahrscheinlichkeit haben, immer einen Youden-Index von Null. Eine perfekte Klassifikation, d.h. alle tatsächlich stornierten Tarife werden als storniert und alle nicht stornierten Tarife als nicht storniert geschätzt, hat einen Index-Wert von Eins. Generell gilt, dass höhere Werte (näher bei Eins) des Youden-Index zu bevorzugen sind.

Im nun folgenden Abschnitt werden wir den optimalen Schwellwert, d.h. jenes $c^* \in (0,1)$ für das $J(c^*)$ maximal ist, grafisch bestimmen. Da für gegebene Schätzungen der Stornowahrscheinlichkeit die Berechnung des Youden-Index nicht sehr aufwändig ist, können wir einfach für alle $c \in \{0.001, 0.002, \ldots, 1\}$ den Wert $J(c)$ berechnen und grafisch darstellen.

4.3.3 Modellvergleich

Mit den in Abschnitten 4.3.1 und 4.3.2 präsentierten Methoden der Kreuzvalidierung und der Klassifikation sind wir nun bereit verschiedene Modelle der Stornowahrscheinlichkeit zu vergleichen. In Abschnitt 4.2 haben wir vier verschiedene logistische Regressionsmodelle identifiziert die für uns von Interesse sind. Diese vier Modelle haben wir als *modell1*, *modell2*, *modell3* und *modell4* bezeichnet. Das erste Modell (*modell1*) beinhaltete alle von uns als relevant identifizierten Prädiktoren in additiver Form. In den Modellen *modell2* und *modell3* wurden jeweils die Faktorinteraktionen zwischen den stetigen Prädiktoren und der Tarifklasse oder Rabatt berücksichtigt. Das Modell *modell4* beinhaltet zusätzlich zur Faktorinteraktion mit Rabatt die Interaktion zwischen Alter und Prämie.

```
modell1<-bam(STORNO~s(VRALTER, bs="cr",k=20)+
                    s(LAUFZEIT, bs="cr",k=25)+
                    s(PRAEMIE, bs="cr",k=30)+PG+RABATT,
                    family=binomial)

modell2<-bam(STORNO~s(VRALTER, bs="cr",k=20,by=PG)+
                    s(LAUFZEIT, bs="cr",k=25,by=PG)+
                    s(PRAEMIE, bs="cr",k=30,by=PG)+PG+RABATT,
                    family=binomial)

modell3<-bam(STORNO~s(VRALTER, bs="cr",k=20,by=RABATT)+
                    s(LAUFZEIT, bs="cr",k=25,by=RABATT)+
                    s(PRAEMIE, bs="cr",k=30,by=RABATT)+PG+RABATT,
                    family=binomial)

modell4<-bam(STORNO~te(VRALTER,PRAEMIE,k=6,by=RABATT)+
                    s(LAUFZEIT, bs="cr",k=25,by=RABATT)+PG+RABATT,
                    family=binomial)
```

Als erstes wollen wir die Klassifikationseigenschaften vergleichen. Dafür werden für die Schätzung der Modelle alle Daten eines einzelnen Kalenderjahres verwendet. Anschließend stellen wir die Youden-Indizes, wie in Abschnitt 4.3.2 beschrieben, in Abhängigkeit des Schwellwerts c dar. Bei der Berechnung dieser Indizes wird dabei die Klassifikation mit der beobachteten Responsevariable für jene Tarife verglichen, welche auch für die Modellschätzung verwendet wurden. Anhand dieser grafischen Darstellung können wir dann die Güte der Klassifikationen bewerten und vergleichen.

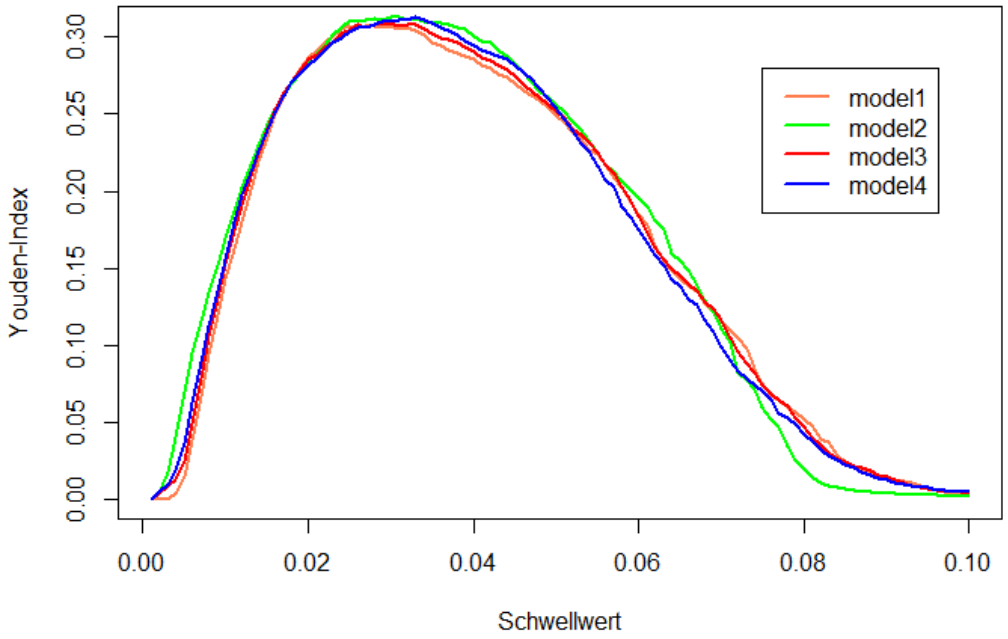

Abbildung 19: Der Youden-Index in Abhängigkeit des Schwellwerts im Jahre 2014

Die Youden-Indizes der obigen vier Modelle für die Daten des Jahres 2014 sind in Abbildung 19 dargestellt. Für eine Wahl des Schwellwertes jenseits der 10% beträgt dieser Index dabei annähernd Null. Dies liegt daran, dass für einen derart hohen Schwellwert nahezu alle Tarifpositionen als nicht storniert klassifziert werden. Wir erkennen, dass für alle Modelle das Maximum für einen Schwellwert $c^* \in [0.025, 0.035]$ angenommen wird. Dies bedeutet, dass die beste Einteilung dann erreicht wird, wenn Tarife mit einer geschätzten Stornowahrscheinlichkeit von weniger oder gleich c^* als nicht storniert und Tarife deren Schätzung über c^* liegt als storniert klassifiziert werden. Diese Beobachtungen gelten für die Daten aller betrachteter Kalenderjahre und somit auch für das Jahr 2012, welches sich bisher doch sehr von den anderen Jahren unterschied. Ebenfalls gilt für alle Jahre, dass sich die dargestellten Indizes-Kurven für alle der obigen GLMe nahezu überdecken. Somit lässt sich anhand der Klassifikation kein Modell bestimmen, welches den anderen zu bevorzugen wäre.

Da sich alle Modelle in allen Beobachtungsjahren sehr ähnlich verhalten, wollen wir die Sensitivität und Spezifität nur anhand von *modell1* im Jahre 2014 genauer betrachten. Hier wird der maximale Youden-Index von 0.307 für $c^* = 0.025$ angenommen. Für diese Wahl des Schwellwerts werden 75.83% der nicht stornierten Tarife als nicht storniert klassifiziert. Von den Tarifen die im Jahr 2014 storniert wurden sind 54.9% auch als storniert geschätzt worden. Somit wurden etwa die Hälfte der stornierten Verträge und Dreiviertel der nicht stornierten Verträge richtig klassifiziert.

Abschließend wollen wir auch noch die jahresübergreifende Klassifikation betrachten. Dafür werden wie zuvor die vier Modelle unter Verwendung aller Daten eines Jahres geschätzt. Die Youden-Indizes werden nun jedoch nicht auf Basis der Daten der Modellschätzung sondern auf Basis der Daten des Folgejahres berechnet. Dies bedeutet, dass die Modelle der Jahre 2012 und 2013 genutzt werden um für den Bestand der Jahre 2013 und 2014 die Stornowahrscheinlichkeiten zu schätzen. Anschließend werden auf Basis dieser Schätzungen die Youden-Indizes berechnet.

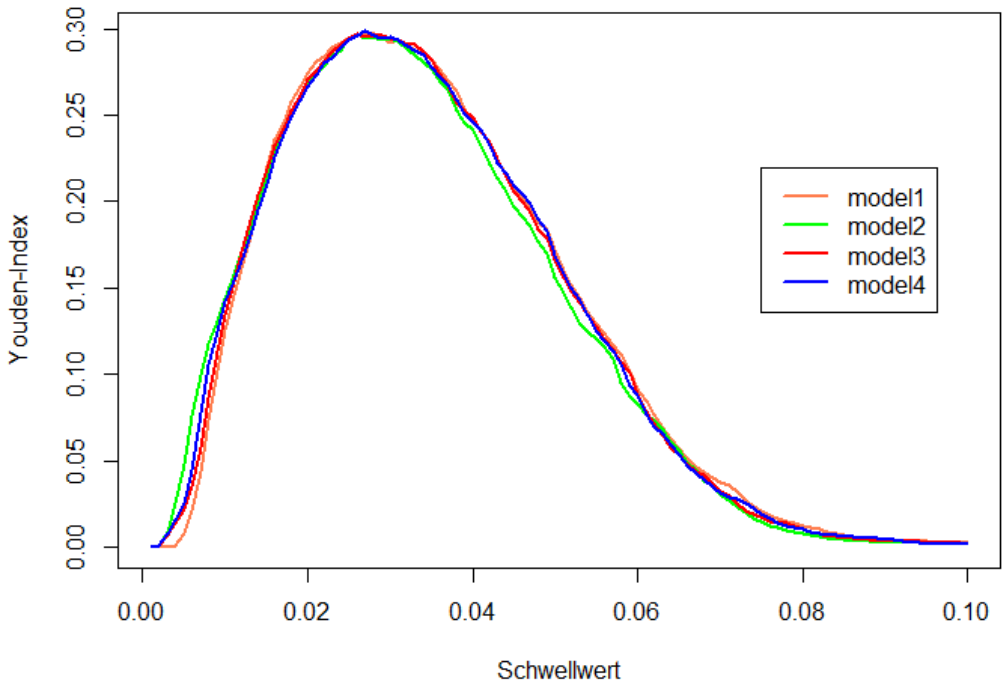

Abbildung 20: Der Youden-Index in Abhängigkeit des Schwellwerts für 2013-2014

Wir erkennen in Abbildung 20, welche sich auf die Jahre 2013 und 2014 bezieht, ein ähnliches Verhalten wie zuvor in Abbildung 19. Die betrachteten GLMe unterscheiden sich kaum bezüglich ihrer (jahresübergreifenden) Klassifikationseigenschaften. Konkret bedeutet dies für *modell1*, geschätzt auf den Daten von 2013, einen maximalen Youden-Index

von 0.297 für 2014. Das heißt, dass auf Basis der Daten von 2013, mit Hilfe von *modell1* und der in Abschnitt 4.3.2 beschriebenen Methode, 58% der in 2014 tatsächlich stornierten Verträge als storniert geschätzt wurden und 72% der in 2014 nicht stornierten Verträge tatsächlich als nicht storniert klassifiziert wurden.

Aufgrund dieser Ergebnisse und der Tatsache, dass die Schätzung als storniert für einen Vertrag mit einer Stornowahrscheinlichkeit von nur wenigen Prozentpunkten sehr gewagt erscheint, vermuten wir die Stärken dieser Modelle nicht auf Einzelvertragsbasis. Aus diesem Grund beschäftigen wir uns im restlichen Teil dieses Kapitels mit Maßzahlen, die die Prädiktionsgüte unserer Modelle auf Bestandsbasis bewerten.

Bevor wir die Prädiktionsgüte mit den vorgestellten Methoden der Kreuzvalidierung bewerten, wollen wir diese Modelle zunächst noch einmal anhand der klassischen Kennzahlen (Deviance, AIC und BIC) vergleichen. Die entsprechenden Werte sind in Tabelle 10 abgebildet. Wie zu erwarten war, liefert das komplexeste Modell (*modell2*) die höchste Reduktion der Deviance. In diesem Modell werden für alle drei stetigen Prädiktoren jeweils vier Glatte Funktionen geschätzt. Diese Komplexität drückt sich auch dadurch aus, dass dieses Modell in den Jahren 2013 und 2014 den höchsten BIC Wert aufweist. Anhand des BIC Wertes wären aufgrund der geringeren Anzahl an zu schätzenden Parametern die Modelle *modell4* und *modell1* zu bevorzugen. Durch diese Kennzahlen erhalten wir also die wenig zufriedenstellende Aussage, dass das komplexeste Modell die beste Anpassung an die Daten liefert, während die einfacheren Modelle wohl den besseren Kompromiss zwischen Komplexität und Datenanpassung bieten. Es bedarf also der Evaluierung der Prädiktionsgüte um uns für ein konkretes Modell entscheiden zu können.

Daten	modelll	modell2	modell3	modell4
	Deviance-Reduktion			
2014	5.85%	**6.22%**	5.98%	6.01%
2013	4.58%	**4.92%**	4.67%	4.69%
2012	5.16%	**7.18%**	7.1%	7.15%
	AIC			
2014	77543.82	**77351.8**	77484.3	77460.7
2013	70974.55	**70821.7**	70938	70921.7
2012	74194.94	72794	72771.2	**72703.1**
	BIC			
2014	**78018.34**	78421.8	78204.4	78216.6
2013	**71478.12**	71874.7	71618.8	71597.7
2012	74804.4	74323.7	73864.7	**73620.42**

Tabelle 10: Modellvergleich anhand klassischer Kennzahlen

Anstatt die Prädiktionsgüte all dieser Modelle auf einmal zu vergleichen, werden wir dies im restlichen Teil dieses Abschnittes paarweise tun. Dieses Vorgehen hat mehrere Gründe. So erhöht es die Übersichtlichkeit wenn wir jeweils nur zwei Modelle betrachten. Außerdem können wir auf diese Art und Weise unsere Entscheidung aus Abschnitt 4.2.1, bei Modell *modell4* die Interaktion mit Rabatt zu betrachten, besser nachvollziehen. Wir gehen dabei wie folgt vor. Zuerst wollen wir *modell2* und *modell3* vergleichen. Damit können wir feststellen, ob sich die zusätzliche Komplexität von *modell2* hinsichtlich der Prädiktionsgüte auszahlt. Das zu bevorzugende dieser beiden Modelle werden wir dann mit *modell4* vergleichen. Somit lässt sich abschätzen, ob die Modellierung von Alter und Prämie als zweidimensionale Glatte Fläche von Vorteil ist. Das so erhaltene Modell vergleichen wir dann abschließend mit unserem Referenzmodell *modell1*. Dies ist das einfachste der hier untersuchten Modelle und bietet außerdem auch die besten Voraussetzungen für eine Interpretation der Ergebnisse wie sie für unseren Anwendungsfall nötig ist.

Zusätzlich zu dem Vergleich der jeweils betrachteten zwei Modelle präsentieren wir im Folgenden auch noch die entsprechenden Kennzahlen des Nullmodells. Dieses Modell stellt das trivialste aller Regressionsmodelle dar. Es entspricht der Annahme, dass es keine vom Zufall abweichende Abhängigkeitsstruktur zwischen den Prädiktoren und der Stornowahrscheinlichkeit gibt. Die Stornierung eines Tarifes ist also ein rein zufälliges Ereignis.

```
modell0<-bam(STORNO~1,family=binomial)
```

Bei dem Nullmodell *modell0* wird zur Schätzung der Stornowahrscheinlichkeit einfach der relative Anteil an Stornierungen in der betrachteten Menge an Tarifen herangezogen. Dementsprechend ergibt sich die Schätzung für die Anzahl an stornierten Tarife eines Testdatensatzes unserer Kreuzvalidierung als relativer Anteil an Stornierungen in der betrachteten Datenmengen multipliziert mit der Anzahl an zu schätzenden Tarifen. Analog erfolgt die Schätzung für die Summe an stornierten Prämien.

4.3.3.1 Faktorinteraktion mit Tarifklasse oder Rabatt (*modell2* oder *modell3*)

Wir wollen zunächst die beiden Modelle der Faktorinteraktionen aus Abschnitt 4.2.4 vergleichen. Diese sind die Modelle *modell2* und *modell3*. Ersteres beinhaltet für jeden stetigen Prädiktor jeweils vier Glatte Funktionen, eine für jede Tarifklasse. Modell *modell3* hingegen beinhalten für jeden stetigen Prädiktor zwei Glatte Funktionen, eine für Tarife mit Rabatt und eine für jene ohne. In Tabelle 10 erkennen wir, dass das komplexere Modell *modell2* bezüglich der Reduktion der Deviance und dem AIC zu bevorzugen wäre. Demgegenüber stehen die BIC-Werte. Sie deuten darauf hin, dass *modell3* das effizientere Modell ist.

Nun werden wir anhand der in Abschnitt 4.3.1 präsentierten Methoden die Prädiktionsgüten dieser Modelle und die des Nullmodells vergleichen. Wir beginnen dabei mit den relativen Fehlern der Schätzungen für die Anzahl an Stornierungen. Dies entspricht der von uns

eingeführten Kennzahl RAS. Folgen wir der in Abschnitt 4.3.1 beschriebenen Testmethode 1, erhalten wir für jedes Modell und für jedes Beobachtungsjahr jeweils zehn unterschiedliche Werte für diese Kennzahl. Diese Werte wurden in Abbildung 21 jeweils als Boxplot zusammengefasst.

Dieser Vergleich anhand der RAS-Werte in Abbildung 21 liefert tatsächlich interessante Ergebnisse. Auffällig ist abermals der Unterschied zwischen dem Jahr 2012 und den Jahren 2013 und 2014. Im Vergleich zu 2013 und 2014 streuen die RAS-Werte im Jahr 2012 nur sehr wenig um Null. Zu beachten ist hier allerdings die Skala, die sich nur im Bereich von wenigen Prozentpunkten bewegt. Dies bedeutet, dass die betrachteten Modelle *modell2*, *modell3* und *modell0* geschätzt auf 90% der Daten eines Jahres, die Anzahl an Stornierungen in den restlichen 10% der Beobachtungen bis auf wenige Prozentpunkte genau schätzen können. Das Nullmodell macht hier keine Ausnahme. Bezüglich der Anzahl an Stornierungen im gleichen Jahr liefern uns die betrachteten GAMe also keinerlei Vorteile gegenüber dem trivialen Nullmodell.

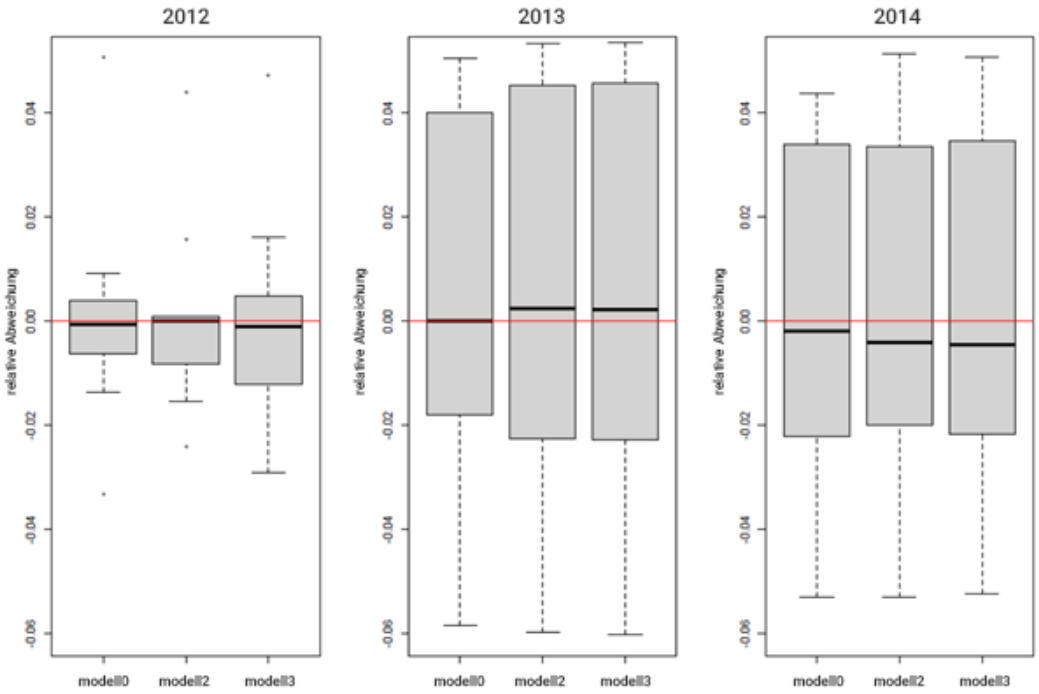

Abbildung 21: Evaluierung von *modell2* und *modell3* anhand von RAS (Testmethode 1)

Wir wollen die Prädiktionsgüte dieser Modelle nun auch anhand unserer zweiten Kennzahl RAP vergleichen. Wie in Abschnitt 4.3.1 erläutert, beschreibt RAP die relative Abweichung zwischen der geschätzten und der tatsächlichen Summe an Prämien von stornierten Tarifen. In Abbildung 22 sind die entsprechenden Ergebnisse dargestellt. Bezüglich der

stornierten Prämien fällt die Prädiktionsgüte des Nullmodells deutlich ab. In allen Beobachtungsjahren liegt der Mittelwert über 30%. Dies bedeutet, dass die Schätzung des monatlichen Prämienausfalles durch die triviale Methode teilweise bis zu 60% über dem tatsächlichen Wert liegt. Die RAP-Werte der GAMe *modell2* und *modell3* bewegen sich hingegen wieder vorwiegend im einstelligen Prozentbereich. Auffällig ist dabei auch, dass sich die Boxplots dieser zwei Modelle kaum voneinander unterscheiden. Es scheint also weder das eine noch das andere GAM die deutlich bessere Prädiktionsgüte aufzuweisen.

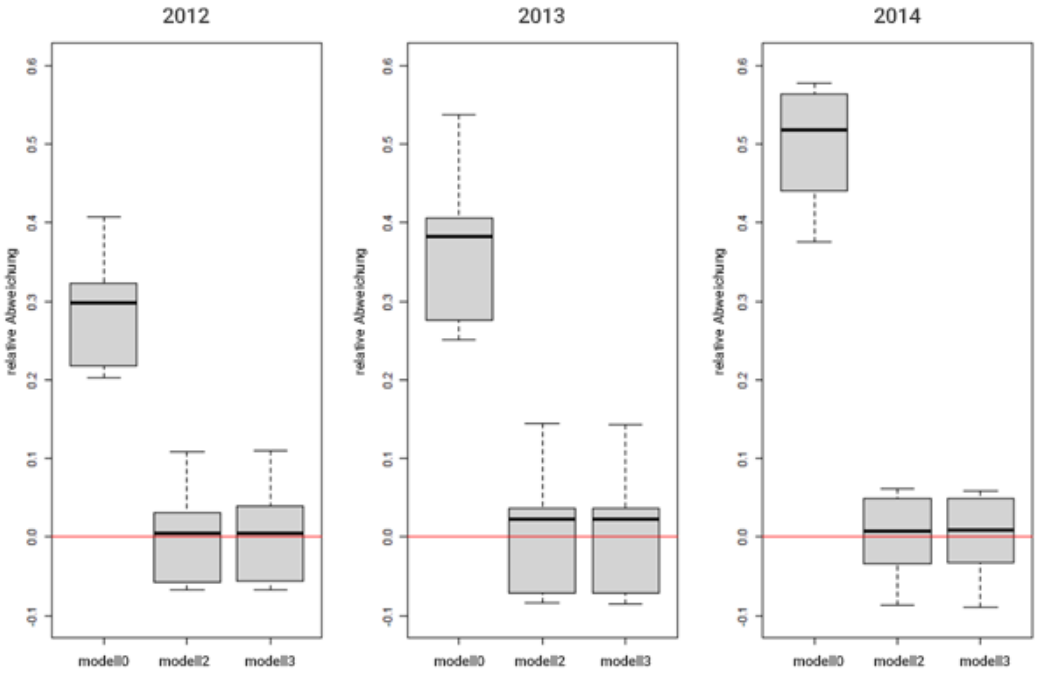

Abbildung 22: Evaluierung von *modell2* und *modell3* anhand von RAP (Testmethode 1)

Die bisherigen Auswertungen bezogen sich alle auf die in Abschnitt 4.3.1 beschriebene Testmethode 1. Es wurden also jeweils 10% der Daten eines Jahres für die Modellschätzung ausgespart. Anschließend werden die Schätzungen der Modelle, die auf den restlichen 90% der Daten basieren, mit den tatsächlichen Werten dieser Datenmenge verglichen. Nun wollen wir auch Testmethode 2 anwenden. Hier werden die Modelle jeweils auf allen Daten der Jahre 2012 und 2013 geschätzt. Anschließend werden diese Modelle benutzt um Schätzungen für jeweils 10% der Daten des darauffolgenden Jahres zu berechnen. Da wir dabei nur direkt aufeinanderfolgende Jahre betrachten, erhalten wir so für jedes Modell jeweils nur zwei Boxplots für die Kennzahlen RAS und RAP.

Die entsprechenden Ergebnisse sind in den Abbildungen 23 und 24 ersichtlich. Dabei erkennen wir ein ähnliches Bild wie wir es bereits zuvor für Testmethode 1 beobachten haben.

Der große Unterschied liegt allerdings in den Skalen. Während wir für den jahresinternen Vergleich (Testmethode 1) zuvor noch Werte im einstelligen Prozentbereich beobachtet haben, ist dies nun vor allem für die Daten von 2012-2013 nicht mehr der Fall. Nichtsdestotrotz gilt abermals, dass das Nullmodell zwar bezüglich RAS vertretbare Ergebnisse liefert, jedoch bei den RAP-Werten drastisch abfällt. Außerdem scheint es wieder so, als gäbe es zwischen der Prädiktionsgüte von *modell2* und *modell3* keine großen Unterschiede. Zu beachten ist außerdem, dass diese beiden Abbildungen den nächsten Hinweis darauf geben, dass sich die Daten des Jahres 2012 von den anderen Jahren unterscheiden. So sehen wir in Abbildung 24, dass die RAP-Werte der GAMe *modell2* und *modell3* relativ eng um Null schwanken. Es war also möglich, mit Hilfe der Daten aus 2013 die Summe der stornierten Prämien des Jahres 2014 mit einer Abweichung von weniger als 10% zu schätzen. Selbiges gilt für die RAS-Werte, wobei hier auch das Nullmodell *modell0* vergleichbare Werte liefert. Für die Jahre 2012 und 2013 beobachten wir in Abbildung 23 ein anderes Verhalten. Hier bewegen sich die RAP-Werte der GAMe zwischen 10% und 20% und auch die Absolutbeträge der RAS-Werte sind deutlich höher als jene von 2013-2014.

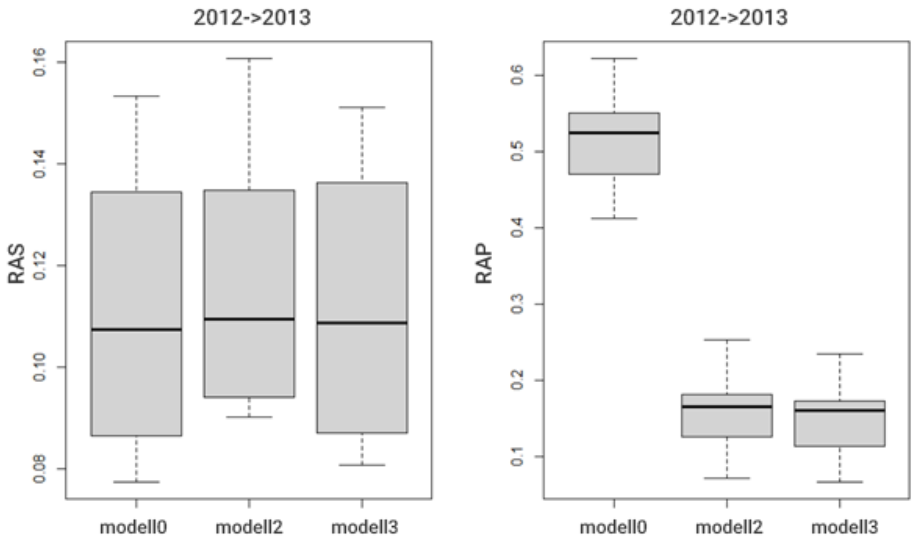

Abbildung 23: Evaluierung von *modell2* und *modell3* anhand von RAS und RAP (Testmethode 2, 2012 und 2013)

Zusammenfassend ergibt sich für diese beiden Modelle folgendes Bild. Das komplexere Modell *modell2*, das insgesamt 12 Glatte Funktionen beinhaltet, liefert die bessere Anpassung an die Daten. Anhand des BIC ist die zusätzliche Modellkomplexität, die diesen Vorteil in der Datenanpassung bringt, aber zu hoch. Auch unsere beiden Kennzahlen RAS und RAP, die die Prädiktionsgüte quantifizieren, sehen keinen Vorteil im komplexeren *modell2*. Somit fällt unsere Wahl zwischen diesen beiden Modell auf *modell3*. Dieses Modell beinhaltet insgesamt nur sechs Glatte Funktionen und ist somit weniger aufwändig zu schätzen.

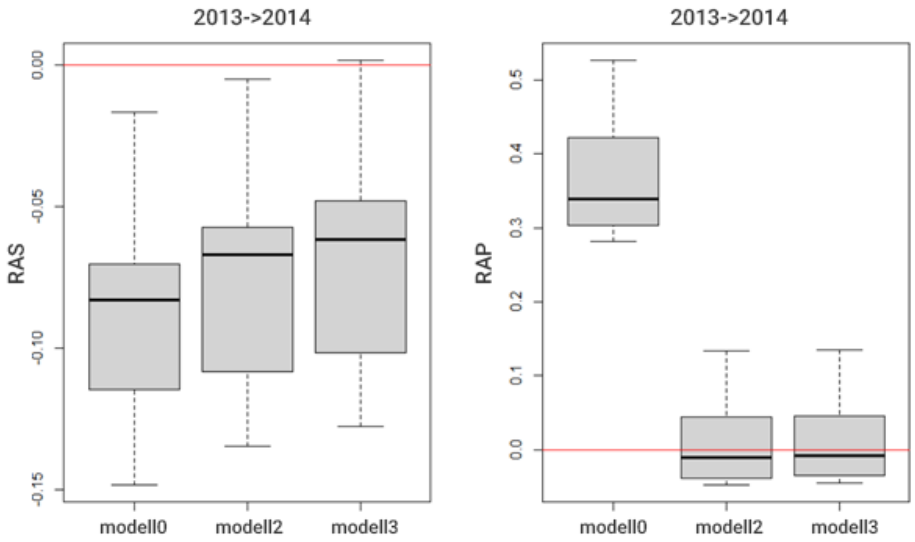

Abbildung 24: Evaluierung von *modell2* und *modell3* anhand von RAS und RAP (Testmethode 2, 2013 und 2014)

4.3.3.2 Interaktion zwischen Alter und Prämie (*modell3* oder *modell4*)

Nachdem wir also im vorherigen Abschnitt festgestellt haben, dass die Faktorinteraktion mit Rabatt (*modell3*) effizienter ist als jene mit der Tarifklasse (*modell2*), gilt es nun zu überprüfen ob die Betrachtung einer zusätzlichen Interaktion zwischen zwei stetigen Prädiktoren von Vorteil ist. In Abschnitt 4.2.5 haben wir dafür die Interaktion zwischen Alter und Prämie als aussichtsreichsten Kandidaten identifiziert. Modell *modell4* enthält neben den zwei Faktoren Rabatt und Tarifklasse zwei Glatte Flächen die die Wechselwirkung zwischen Alter und Prämie beschreiben und zwei Glatte Kurven für die Laufzeit, jeweils für beide Stufen von Rabatt. In diesem Abschnitt wollen wir also die Modelle *modell3* und *modell4* vergleichen.

Anhand der Kennzahlen in Tabelle 10 erkennen wir, dass *modell4* eine etwas bessere Datenanpassung liefert als dies *modell3* tut (die Reduktion der Deviance ist in allen Jahren größer). Dementsprechend bevorzugt auch das AIC dieses Modell, wobei die jeweiligen Werte sehr nahe beieinander liegen. Ähnliches gilt für die BIC-Werte. Mit Ausnahme des Jahres 2014 weist *modell4* die geringeren Werte auf. Die Kennzahlen liegen aber wieder sehr nahe beieinander. Anhand der klassischen Kennzahlen können wir uns also für keines dieser beiden Modelle entscheiden. Aus diesem Grund betrachten wir nun wieder deren Prädiktionsgüte mit Hilfe der RAS- und RAP-Werte.

In Abbildung 25 sind die RAS-Werte für die Modelle *modell0*, *modell3* und *modell4* nach Testmethode 1 der Jahre 2012, 2013 und 2014 abgebildet. Hierbei ist zu beachten, dass die für die Berechnung dieser Kennzahlen notwendige Einteilung eines Jahresbestandes

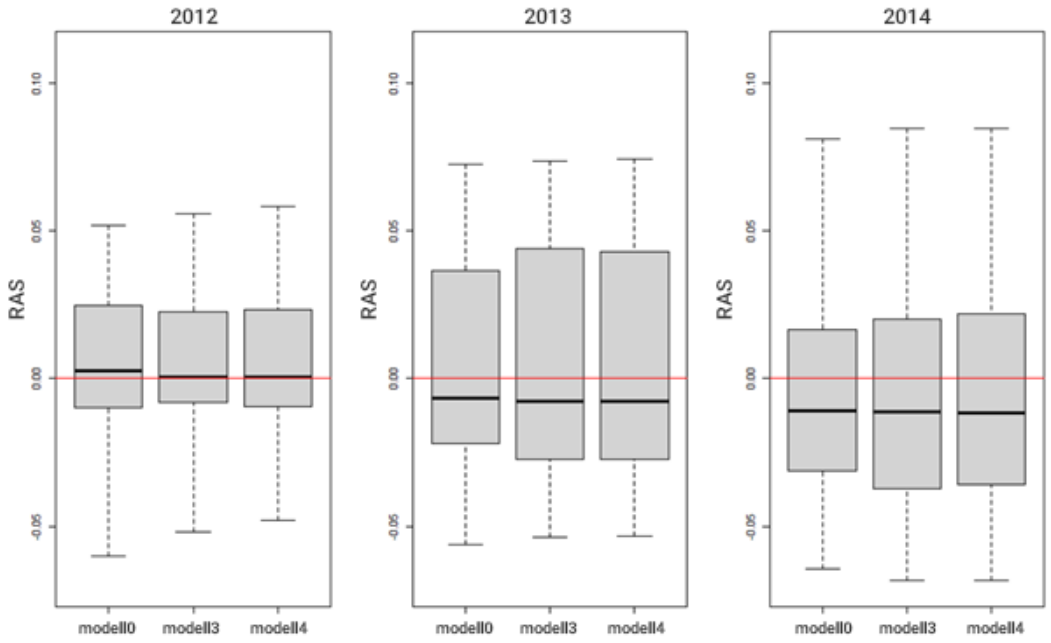

Abbildung 25: Evaluierung von *modell3* und *modell4* anhand von RAS (Testmethode 1)

von Tarifen in zehn gleichgroße Teilmengen zufällig erfolgt. Aus diesem Grund müssen die hier dargestellten Boxplots für *modell0* und *modell3* nicht mit jenen aus Abbildung 21 übereinstimmen (was sie auch nicht tun). Dennoch gilt wieder für alle Modelle, dass die Schätzungen für die Anzahl an Stornierungen in den jeweils ausgesparten 10% der Daten nur einzelne Prozentpunkte von der tatsächlichen Beobachtung abweichen. Auch hier weist das Nullmodell ein ähnliches Verhalten wie die beiden GAMe auf. Betrachten wir nur die komplexeren GAMe, so können wir zwischen *modell3* und *modell4* wiederum kaum einen Unterschied bezüglich der geschätzten Anzahl an Stornierungen ausmachen.

Wiederum wiederholen wir diese Analyse auch für die RAP-Werte. Die Ergebnisse sind in Abbildung 26 zusammengefasst. Da abermals die Einteilung der Jahresbestände an Tarifen in zehn gleichgroße Teilmengen zufällig erfolgt, gilt auch hier, dass die Boxplots von *modell0* und *modell3* nicht mit jenen aus Abbildung 22 zwangsläufig übereinstimmen müssen. Nichtsdestotrotz sehen wir ähnliche Ergebnisse. Das Nullmodell überschätzt im Mittel jedes Jahr die stornierten Prämien um mehr als 30 Prozentpunkte. Demgegenüber stehen die GAMe. Der Mittelwerte der relativen Abweichung zwischen Schätzung und Beobachtung liegt hier immer sehr nahe bei Null. Abermals können wir anhand dieser Kennzahlen kaum zwischen diesen beiden Modellen unterscheiden.

Der Vergleich zwischen *modell3* und *modell4* verhält sich sehr ähnlich zu jenem zwischen *modell2* und *modell3*. Die zusätzliche Modellkomplexität von *modell4* macht sich in der Datenanpassung bemerkbar. So beobachten wir in allen Jahren eine etwas höhere Reduktion

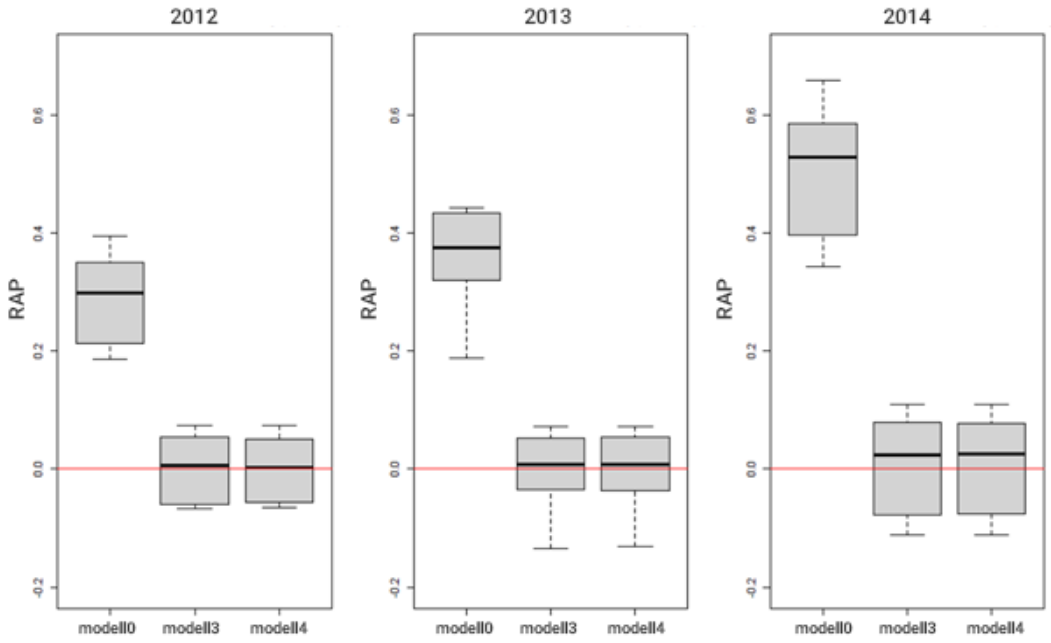

Abbildung 26: Evaluierung von *modell3* und *modell4* anhand von RAP (Testmethode 1)

der Deviance. Praktische Relevanz scheint dieser Vorteil aber nicht zu haben. So unterscheiden sich diese Modelle in ihrer Prädiktionsgüte kaum. Anhand der von uns eingeführten Kennzahlen RAS und RAP kommen wir zu dem Schluss, dass die zusätzliche Betrachtung einer Interaktion zwischen den zwei stetigen Prädiktoren Alter und Prämie keinen relevanten Vorteil bringt. Daher entscheiden wir uns hier für *modell3*. Auf einen Vergleich dieser Modelle anhand von der in Abschnitt 4.3.1 präsentierten Testmethode 2 verzichten wir an dieser Stelle. Der Grund dafür ist, dass die Ergebnisse jenen aus Abschnitt 4.3.3.1 wieder sehr ähnlich sind und somit keine neuen Einsichten liefern.

4.3.3.3 Interaktionen (*modell1* oder *modell3*)

Den letzten Modellvergleich, den wir nun vornehmen werden, ist jener zwischen den Modellen *modell1* und *modell3*. Damit beantworten wir also die Frage, ob es überhaupt einen Vorteil bringt die Faktorinteraktion mit Rabatt zu betrachten oder ob es ausreicht, alle relevanten Prädiktoren in additiver Form zu modellieren. Das als *modell1* bezeichnete Modell entspricht dabei unserem Referenzmodell aus Abschnitt 4.2.3. Es ist das einfachste der hier betrachteten Modelle und ermöglicht so die effizienteste Berechnung und eine schöne Interpretation (siehe Kapitel 5). Um zu bestimmen ob eine höhere Modellkomplexität (*modell3*) erforderlich ist, entspricht unsere Vorgehensweise dabei jener aus den vorangegangen Abschnitten. Dies bedeutet, dass wir zuerst die klassischen Kennzahlen und danach die Prädiktionsgüte anhand der RAS- und RAP-Werte dieser beiden Modelle vergleichen werden.

Die klassischen Kennzahlen entnehmen wir abermals Tabelle 10. Dabei ist ersichtlich, dass das nunmehr komplexere *modell3* die höhere Reduktion der Deviance und somit die höhere Datenanpassung in allen Beobachtungsjahren liefert. Während der Unterschied in den Jahren 2013 und 2014 nur sehr gering ist, beträgt er im Jahr 2012 annähernd 2%. Mit Ausnahme des Jahres 2012 wird das einfachere *modell1* jedoch vom BIC bevorzugt und weist dabei sogar den geringsten BIC-Wert aller betrachteten Modelle auf. Es scheint also wieder so, als ob das komplexere Modell die bessere Datenanpassung liefert, während das einfachere Modell effizienter zu sein scheint. Somit stellt sich wiederum die Frage, ob sich eine erhöhte Modellkomplexität auch bezüglich der Prädiktionsgüte auszahlt.

Den Vergleich der Prädiktionsgüten beginnen wir wieder anhand der jahresinternen Testmethode 1 aus Abschnitt 4.3.1. In Abbildung 27 sehen wir die entsprechenden Boxplots der RAS-Werte. Noch auffälliger als in den vorherigen Modellvergleichen (Abbildungen 21 und 25) können wir kaum Unterschiede zwischen den Kennzahlen der einzelnen Modelle ausmachen. Wieder scheint das Nullmodell bezüglich der Anzahl an Stornierungen dieselbe Prädiktionsgüte wie die komplexeren GAMe zu haben. Die hier betrachteten Modelle zeigen in allen Beobachtungsjahren ein sehr ähnliches Verhalten. Die Mittelwerte der RAS-Werte liegen entweder alle relativ genau auf der Null (2012) oder geschlossen im positiven (2013) oder negativen (2014) Bereich. Auch die RAP-Werte in Abbildung 28 unterscheiden sich kaum von den bisherigen Modellvergleichen (Abbildungen 22 und 26).

Bei der Schätzung der stornierten Prämien sind die GAMe dem Nullmodell also ganz klar überlegen. Während das Nullmodell *modell0* die stornierten Prämien wieder um bis zu 60% überschätzt, liegen die relativen Abweichungen der Modelle *modell1* und *modell3* vorwiegend im niedrigen einstelligen Prozentbereich. Zwischen den beiden GAMe *modell1* und *modell3* erkennen wir allerdings abermals keinen relevanten Unterschied. Somit scheint das simpelste der von uns betrachteten Modelle schon auszureichen um die Abhängigkeitsstruktur zwischen der Stornowahrscheinlichkeit und den Prädiktoren so zu beschreiben, dass damit brauchbare Schätzungen für die Anzahl an Stornierungen und die stornierten Prämien gemacht werden können.

Interessant ist auch ein direkter Vergleich der RAP-Werte des Nullmodells in den Abbildungen 21, 25 und 27. Die Daten der entsprechenden Boxplots von *modell0* unterscheiden sich lediglich durch die jeweils zufällig bestimmte Einteilung der Beobachtungen in zehn gleichgroße Blöcke. In allen drei Abbildungen beobachten wir, dass sich die Lage der Daten über die drei Beobachtungsjahre verändert. So liegt der Mittelwert der RAP-Werte im Jahr 2012 jeweils bei rund 30%, im Jahr 2013 bei etwas weniger als 40% und im Jahr 2014 teilweise über 50%. Das Nullmodell basiert auf der Annahme, dass es keine Abhängigkeitsstruktur zwischen der Stornowahrscheinlichkeit und den hier betrachteten Prädiktoren gibt. Da die oben erwähnten Boxplots dafür sprechen, dass die besten Prädiktionsergebnisse die das Nullmodell liefert im Jahr 2012 zu beobachten sind, könnte dies als Hinweis darauf gedeutet werden, dass die im Jahr 2012 beobachteten Stornierungen in ihrem Auftreten tatsächlich von den anderen Jahren abweichen.

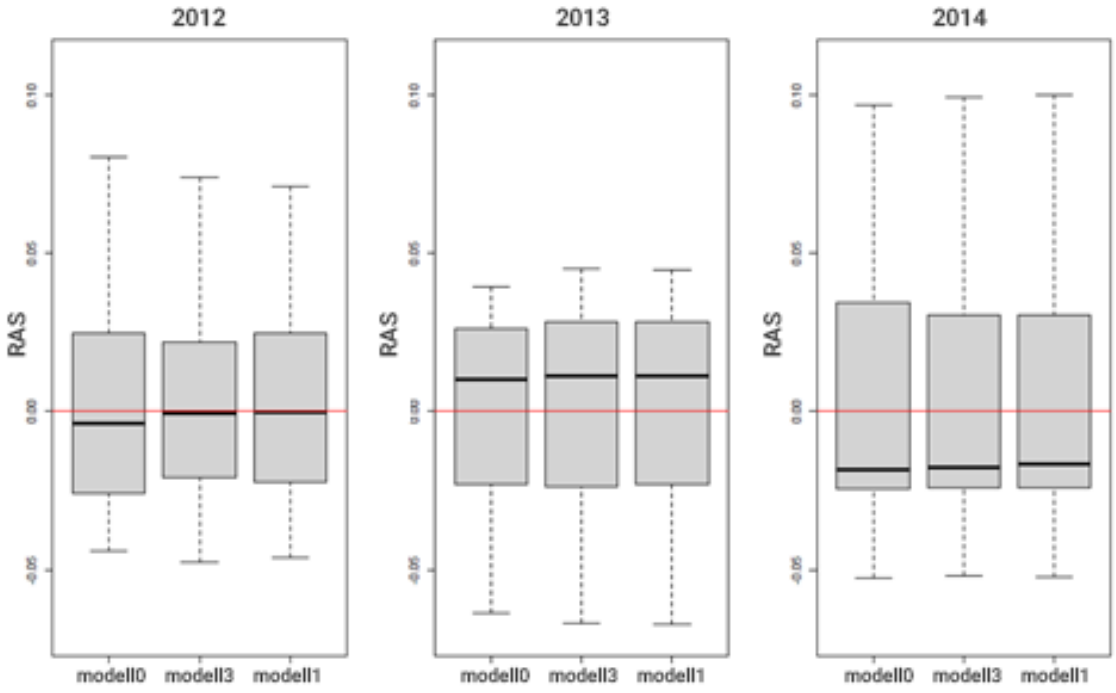

Abbildung 27: Evaluierung von *modell1* und *modell3* anhand von RAP (Testmethode 1)

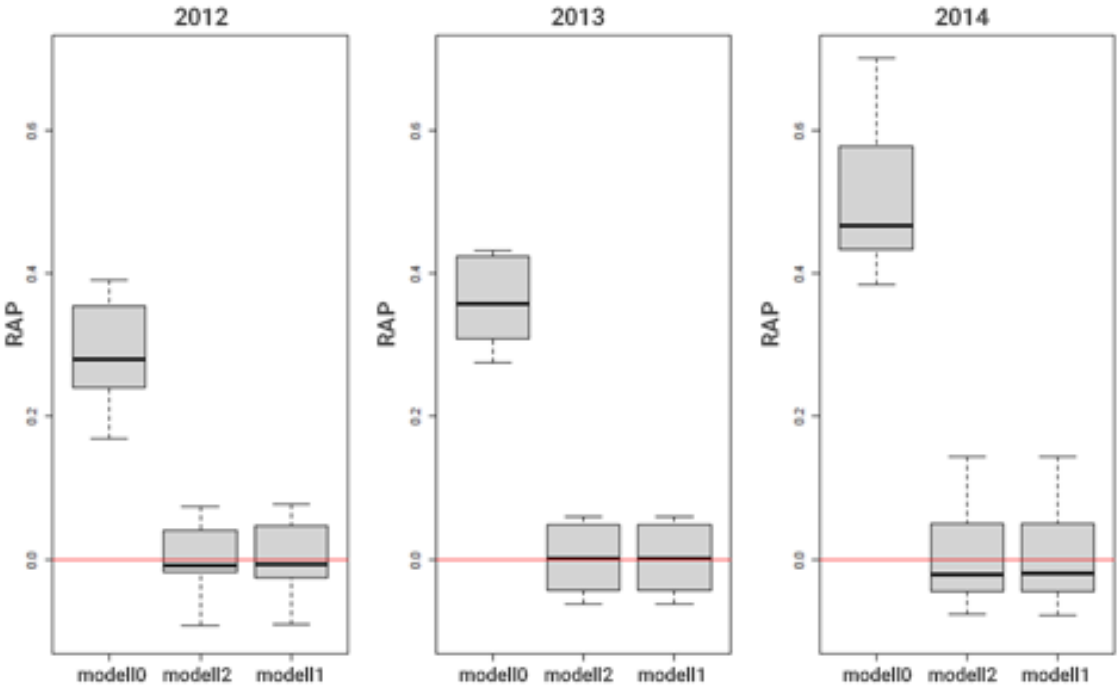

Abbildung 28: Evaluierung von *modell1* und *modell3* anhand von RAP (Testmethode 1)

Die Prädiktionsgüte dieser Modelle wollen wir hier auch wieder anhand von Testmethode 2 aus Abschnitt 4.3.1 vergleichen. Für das Nullmodell und das GAM mit der Rabattinteraktion (modell3) haben wir dies bereits in Abschnitt 4.3.3.1 getan. Dort haben wir festgestellt, dass auch für diese Testmethode die Prädiktionsgüte des Nullmodells bezüglich der Anzahl an Stornierungen gleichauf mit jener der GAMe ist. Bezüglich der stornierten Prämien gilt aber auch hier, dass die GAMe die deutlich besseren Ergebnisse erzielen. Eine weitere Beobachtung war, dass die Daten der Jahre 2013 und 2014 sehr viel besser zueinander passen wie dies mit den Daten des Jahres 2012 der Fall ist.

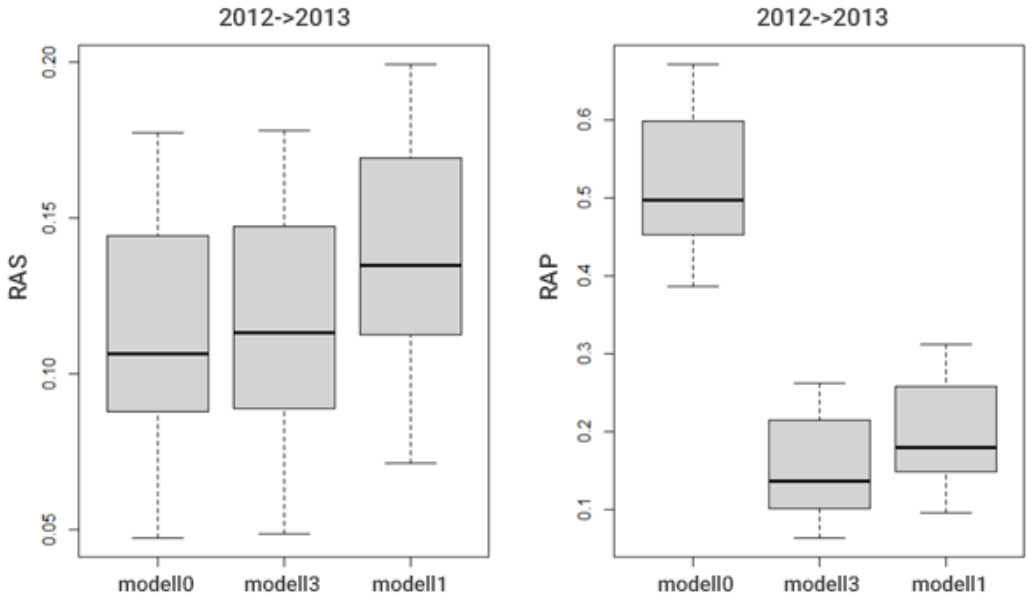

Abbildung 29: Evaluierung von *modell1* und *modell3* anhand von RAS und RAP (Testmethode 2, 2012 und 2013)

In Abbildung 29 sehen wir die RAS- und RAP-Werte nach Testmethode 2 für die Jahre 2012 und 2013. Dabei erkennen wir, dass die Mittelwerte der relativen Fehler in den Schätzung der Anzahl an Stornierungen (RAS) jeweils über 10% liegen. Wenn wir diese Werte mit jenen aus Abbildung 27 vergleichen, erkennen wir, dass (erwartungsgemäß) Modelle, die auf den Daten des gleichen Beobachtungsjahres geschätzt wurden, bessere Ergebnisse in der Prädiktion liefern als dies Modelle tun, die auf den Daten des vorangegangenen Jahres basieren. Selbiges gilt für die RAP-Werte. Während diese in Abbildung 28 noch relativ eng um Null schwankten, befinden sie sich nun im zweistelligen Prozentbereich. Dies gilt allerdings nur für die Jahre 2012 und 2013. In Abbildung 30 sehen wir die entsprechenden Ergebnisse für die Jahre 2013 und 2014. Im linken Teil dieser Grafik ist ersichtlich, dass die RAS-Boxplots der GAMe *modell1* und *modell3* die mit rot markierte Null zumindest überdecken. Die RAS-Mittelwerte dieser Modelle sind nun auch wieder im einstellige Pro-

zentbereich. Noch besser sind die Ergebnisse bezüglich der RAP-Werte. Diese schwanken für die beiden GAMe wieder relativ eng um die Null während die RAP-Werte des Nullmodells im mittleren zweistelligen Prozentbereich realisierten. Somit gilt auch für Modell *modell1*, dass es, geschätzt auf den Daten des Jahres 2013, sehr gut zu den Beobachtungen des Jahres 2014 passt. Wird es jedoch auf Basis der Daten von 2012 geschätzt, so sind die Vorhersagen für das Jahr 2013 von schlechterer Qualität. Dies ist abermals ein Hinweis darauf, dass die Daten des Jahres 2012 von den anderen beiden Beobachtungsjahren 2013 und 2014 abweichen.

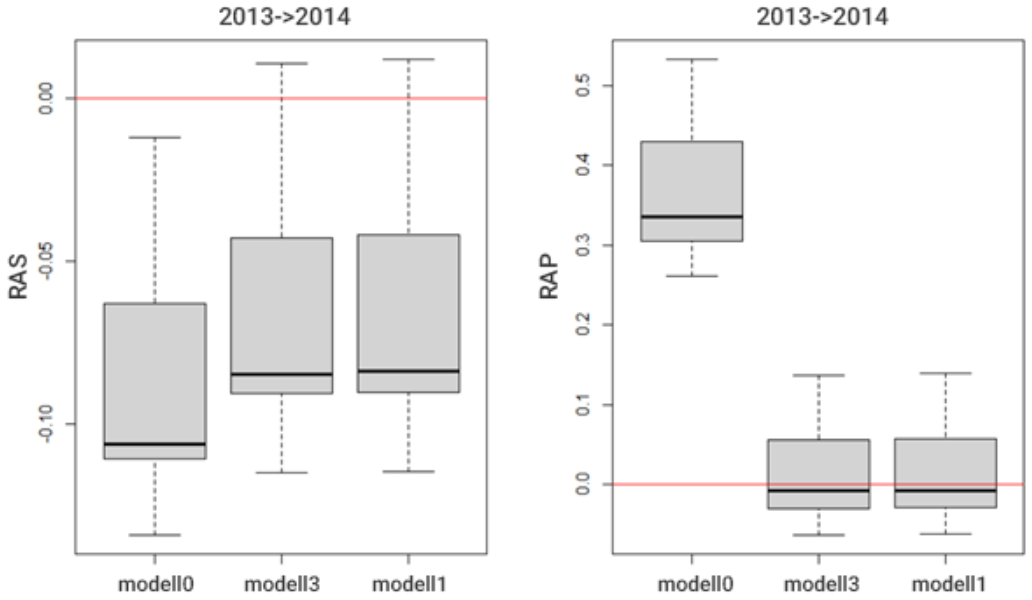

Abbildung 30: Evaluierung von *modell1* und *modell3* anhand von RAS und RAP (Testmethode 2, 2013 und 2014)

Auf Basis der in diesem Abschnitt gemachten Beobachtungen entscheiden wir uns schlussendlich für *modell1*. Dieses Modell beinhaltet zwei faktorielle sowie drei stetige Prädiktoren. Die faktoriellen Prädiktoren der Tarifklasse und des Rabatts besitzen jeweils vier bzw. zwei Faktorstufen. Die stetigen Prädiktoren Alter, Laufzeit und Prämie wurden als Glatte Funktionen modelliert. Die Klasse der Basisfunktionen dieser Glatten Funktionen wurde als die der kubischen Splines definiert. Der jeweilige maximale Freiheitsgrad ergab sich durch die gesonderte Modellierung der Stornowahrscheinlichkeit in Abhängigkeit jeweils einer dieser stetigen Prädiktoren. Anhand der klassischen Kennzahlen von Regressionsmodellen, wie sie in Tabelle 10 dargestellt sind, haben wir zwar erkannt, dass die anderen GAMe die bessere Anpassung an die Daten erreichen, jedoch die damit verbundene zusätzliche Komplexität in der Modellierung nicht rechtfertigbar war. Bereits das Bayessche Informationskriterium (BIC) zeichnete in den Jahren 2013 und 2014 Modell *modell1* als zu bevorzugendes Modell aus. Zusätzlich haben die von uns eingeführten Methoden RAS und RAP zur Evaluie-

rung der Prädiktionsgüte gezeigt, dass *modell1* den komplexeren Modellen diesbezüglich in nichts nachsteht. Somit fällt unsere Wahl auf dieses Modell und im nun folgenden Abschnitt wollen wir uns mit der Auswertung dieses Modells beschäftigen.

5 Modellauswertung

Im vorangegangenen Kapitel 4 haben wir unsere Wahl der Generalisierten Additiven Modelle als Modellklasse im Allgemeinen und von *modell1* aus Abschnitt 4.2.3 als konkretes Modell im Speziellen motiviert. Dieses Modell zeichnet sich dadurch aus, dass die drei Prädiktoren des Alters des Versicherungsnehmers, der Laufzeit des Tarifes und der monatlichen Versicherungsprämie als Glatte Funktionen modelliert sind. Des weiteren beinhaltet es die Klasse des jeweiligen Tarifes und die Rabattierung als Faktoren mit jeweils vier bzw. zwei Faktorstufen. Die drei Glatten Funktionen werden durch eine kubische Splinebasis dargestellt. Die entsprechenden Basisdimensionen sind so gewählt, dass sie eine ausreichende Flexibilität der Kurven zulassen.

```
modell1<-bam(STORNO~ s(VRALTER, bs="cr",k=20)+
                     s(LAUFZEIT, bs="cr",k=25)+
                     s(PRAEMIE, bs="cr",k=30)+PG+RABATT,
                     family=binomial)
```

Im Prozess der Modellauswahl orientierten wir uns dabei an klassischen Kennzahlen von Regressionsmodellen (Reduktion der Deviance, Akaike- und Bayessches Informationskriterium) sowie an den von uns in Abschnitt 4.3.1 eingeführten Methoden zur Evaluierung der Prädiktionsgüte. Das Ergebnis ist ein GAM mit dessen Hilfe wir nun die Daten verschiedener Jahre vergleichen können. Außerdem ermöglicht uns ein solches Modell nicht nur die Anzahl an Stornierungen sondern auch die Summe der stornierten Prämien für ein Folgejahr akkurat zu schätzen. Die so erhaltenen Ergebnisse wollen wir nun betrachten.

In Tabelle 11 sehen wir den Vergleich zwischen der geschätzten und der tatsächlichen Anzahl an Stornierungen. Die Schätzungen bekommen wir also von obigen Modell *modell1*, welches jeweils auf Basis der Vorjahresdaten geschätzt wurde. So finden wir z.B. in der Spalte „2012→2013" die tatsächliche Anzahl an Stornierungen sowie die relative Abweichung der jeweiligen Modellschätzungen. Dabei führen wir zu Vergleichszwecken auch die Ergebnisse des Nullmodells an. Dieses schätzt die Anzahl an Stornierungen im Folgejahr einfach durch den relativen Anteil an Stornierungen im aktuellen Basisjahr.

	Daten	
Auswertung	**2012→2013**	**2013→2014**
Anzahl an Stornierungen	8117	9184
Nullmodell	+11.1% (9019)	−9.1% (8348)
modell1	+13.5% (9211)	−7.1% (8538)

Tabelle 11: Prädiktionsgüte bezüglich der Anzahl an Stornierungen

Wie wir sehen können, liefern beide Modelle bezüglich der Anzahl an Stornierungen etwa dieselbe Qualität an Schätzungen. Die Anzahl an Stornierungen im Jahr 2013 werden auf Basis der Daten von 2012 jeweils um etwas mehr als 10% überschätzt. Von 2013 auf 2014 wird diese Anzahl um etwas weniger als 10% unterschätzt. Anhand unseres GAM *modell1* erkennen wir schon hier, dass die Daten von 2013 und 2014 sehr viel besser zusammen passen als dies die Daten von 2012 und 2013 tun. Der relative Fehler in der Schätzung der Anzahl an Stornierung hat sich fast halbiert. Diese Ergebnisse passen auch sehr gut zu den Beobachtungen die wir in Abschnitt 4.3.3 und dort in den Abbildungen 29 und 30 gemacht haben. Dort wurden die Daten des Folgejahres zufällig in zehn gleichgroße Teilmengen unterteilt. Die abgebildeten Boxplots fassen die zehn relativen Abweichungen zwischen geschätzter und beobachteter Anzahl an Stornierungen zusammen. Die so bestimmten Mittelwerte entsprechen in etwa den Werten aus Tabelle 11.

Das zweite Kriterium zur Evaluation der Prädiktionsgüte war die Schätzung der Summe der stornierten Prämien. Eine solche Schätzung ist vor allem deshalb interessant, weil sie Aufschluss über das finanzielle Risiko für das Versicherungsunternehmen, welches durch den Ausfall der monatlichen Zahlungen von stornierten Tarifen entsteht, gibt. Im Zuge der Modellselektion haben wir bereits festgestellt, dass in der Schätzung dieser stornierten Prämien ein großer Vorteil der GAMe gegenüber dem Nullmodell liegt.

	Daten	
Auswertung	**2012→2013**	**2013→2014**
Summe stornierter Prämien	339403.5	350517.2
Nullmodell	+51.1% (512783.2)	+36.2% (477411.6)
modell1	+18.9% (403886.3)	+1.1% (354288.5)

Tabelle 12: Prädiktionsgüte bezüglich der Summe an stornierten Prämien

In Tabelle 12 sehen wir, dass im Jahr 2013 Tarife, deren monatlichen Prämien sich auf rund 340.000 Euro aufsummierten, storniert wurden. Die Schätzung des Nullmodells, welches sich aus dem Produkt des relativen Anteil an Stornierungen des Jahres 2012 und der Summe aller Prämien der Tarife aus 2013 ergibt, liegt dabei mit 500.000 Euro um mehr als 50% über dem tatsächlichen Wert. Demgegenüber steht die Schätzung des GAM mit etwa 400.000 Euro. Dieser Wert liegt nur knapp 19% über der Beobachtung. Wieder scheint es so, als würden die Daten von 2013 und 2014 wesentlich besser zusammen passen. Die Schätzung des Nullmodells liegt hier nur mehr 36.2% zu hoch. Das GAM *modell1* schafft es gar die Summe der stornierten Prämien des Jahres 2014 auf Basis der Daten von 2013 auf 1.1% genau zu schätzen. Durch die Modellierung der Stornowahrscheinlichkeit als GAM erhalten wir also durchaus brauchbare Schätzungen für die Anzahl an Stornierungen und die Summe der stornierten Prämien.

Nun wollen wir anhand unseres Modells *modell1* die Abhängigkeitsstruktur zwischen der Stornowahrscheinlichkeit und den uns zur Verfügung stehenden Prädiktoren analysieren. Dafür vergleichen wir die geschätzten Koeffizienten der Faktoren bzw. die geschätzten Glatten Funktionen der stetigen Prädiktoren für die unterschiedlichen Jahre. Damit ist es uns möglich, Veränderungen in der Abhängigkeitsstruktur über die Jahre festzustellen bzw. diese Abhängigkeitsstruktur überhaupt zu beschreiben. Während für die drei Glatten Funktionen ein sehr schöner grafischer Vergleich möglich ist, vergleichen wir die Koeffizienten der zwei Faktoren Tarifklasse und Rabatt direkt anhand ihrer Schätzungen.

	2012		2013		2014	
Koeffizient	Schätzung	Std.Fehler	Schätzung	Std.Fehler	Schätzung	Std.Fehler
Intercept	−3.41	0.026	−3.46	0.025	−3.06	0.022
PG02a	−0.26	0.040	−0.54	0.042	−1.00	0.040
PG02b	−1.11	0.052	−0.75	0.052	−1.38	0.052
PG03	+0.44	0.044	+0.57	0.045	+0.12	0.043
Rabatt = 1	+0.27	0.035	−0.47	0.043	−0.31	0.042

Tabelle 13: Vergleich der geschätzten Faktorkoeffizienten

In Tabelle 13 erkennen wir noch einmal, dass alle Stufen der im Modell vorkommenden Faktoren Tarifklasse und Rabatt hoch signifikante Koeffizienten besitzen. Dies bedeutet, dass sich die entsprechenden Schätzungen der Koeffizienten signifikant von Null unterscheiden. Der Intercept entspricht hier der Tarifklasse PG01 ohne Rabatt. Mit Ausnahme der Koeffizienten für die Rabattstufe „Rabatt = 1" behalten alle Koeffizienten über alle Beobachtungsjahre dasselbe Vorzeichen bei. Dies bedeutet, dass die Tarife der Tarifklasse PG03 in allen Jahren eine höhere geschätzte Stornowahrscheinlichkeit aufweisen als dies die Tarife der anderen Klassen tun. Demgegenüber besitzen die Tarife der Tarifklassen PG02 in allen Jahren die geringsten geschätzten Stornowahrscheinlichkeiten, wobei sich hier PG02b noch einmal von PG02a in Form einer verringerte Wahrscheinlichkeit absetzt. Dies stimmt auch mit den ersten Analysen aus Abschnitt 2.2.4 überein. Generell ist zu beachten, dass sich diese Werte nicht als absolute Wahrscheinlichkeiten interpretieren lassen. Viel mehr gibt ihr Vorzeichen und der Absolutbetrag an, in welche Richtung und wie sehr der entsprechende Faktor die Wahrscheinlichkeit für eine Stornierung beeinflusst. Ein positiver Wert bedeutet dabei, dass die Wahrscheinlichkeit für eine Stornierung gegenüber dem Basislevel (Intercept) höher ist, während ein negativer Wert für eine verringerte Wahrscheinlichkeit steht.

Wir rufen uns nun die erste grobe Analyse des Faktors Rabatt aus Abschnitt 2.2.5 und dabei vor allem Abbildung 11 in Erinnerung. Dort haben wir gesehen, dass Tarife deren Prämie eine Rabattierung beinhaltet in den Jahren 2013 und 2014 einen niedrigeren Anteil an Stornierungen aufwiesen als dies im gesamten Bestand dieser Jahre der Fall war. Für das

Jahr 2012 waren diese Anteile in etwa gleichgroß. Anhand der geschätzten Koeffizienten in Tabelle 13 sehen wir ein ähnliches Bild. Die Vorzeichen der Koeffizienten für „Rabatt = 1" sind in den Jahren 2013 und 2014 negativ und für das Jahr 2012 positiv. Auch der Absolutbetrag ist in den beiden späteren Jahren größer. Somit haben in diesen Jahren Tarife mit Rabatt eine verminderte Stornowahrscheinlichkeit während im Jahr 2012 eine Rabattierung der Prämie zu einer erhöhten Stornowahrscheinlichkeit führte.

Wir wenden uns nun den stetigen Prädiktoren zu. Wie wir bereits in Abschnitt 4.2.1 gesehen haben, lassen sich die geschätzten Glatten Kurven grafisch darstellen. Im Unterschied zu den Abbildungen 16, 17 und 18 können wir diese Schätzungen jedoch nicht mehr auf der Skala der Responsevariable, also als Wahrscheinlichkeit, abbilden. Der Grund dafür ist, dass in *modell1* mehrere Prädiktoren als Glatte Funktionen modelliert sind. Somit müssten wir zur Abbildung auf der Skala der Responsevariable die jeweils zwei anderen Prädiktoren fixieren. Damit wäre aber kein umfassender Überblick des Effektes eines einzelnen stetigen Prädiktors auf die Stornowahrscheinlichkeit mehr möglich. In den nun folgenden Abbildungen sind deshalb die Glatten Funktionen auf der Skala des Linearen Prädiktors abgebildet. Somit sind die Werte wiederum nicht direkt als Wahrscheinlichkeit zu interpretieren. Wieder lässt sich aber durch das Vorzeichen und den Absolutbetrag der Effekt des entsprechenden Prädiktors auf die Stornowahrscheinlichkeit ablesen.

In Abbildungen 31 sehen wir die Glatten Funktionen des stetigen Prädiktors Alter für alle Beobachtungsjahre. Die jeweiligen durchgezogenen farbigen Kurven markieren die Punktschätzungen. Durch die gestrichelten Kurven ist jeweils der Bereich einer (geschätzten) Standardabweichung über und unter dieser Punktschätzung markiert. Die so gebildete Fläche zwischen diesen Kurven wurden in der entsprechenden Farbe markiert. Die horizontal durchgezogene schwarze Linie markiert die Null. Positive Werte sprechen für eine erhöhte Stornowahrscheinlichkeit im Vergleich zum Basislevel. Umgekehrt stehen negative Werte für eine verringerte Wahrscheinlichkeit. Wiederum gilt, dass die Kurven ab einem Alter von 80 Jahren nur bedingt von Interesse sind. Der Grund dafür ist, dass dort die Datenbasis zu dünn ist und somit die Schätzungen wenig Aussagekraft haben.

Auffällig ist zunächst, dass sich die beiden Bänder der Jahre 2013 und 2014 (rot und grün) gegenseitig sehr viel mehr ähneln als sie dies mit dem Band des Jahres 2012 (blau) tun. Somit unterscheidet sich die Abhängigkeit zwischen dem Stornoverhalten und dem Alter des Versicherungsnehmers im Jahr 2012 von den anderen Jahren. Dies passt zu den bisherigen Beobachtungen. In allen Jahren gilt aber, dass die höchste Stornowahrscheinlichkeit die Altersgruppe der 20 bis 25 Jährigen aufweist. Auch haben alle Jahre gemein, dass für Versicherungsnehmer mit Alter 25+ die Wahrscheinlichkeit für eine Stornierung mit zunehmenden Alter abnimmt. Erst für sehr hohe Altersklassen beobachten wir wieder eine Steigerung, jedoch ist diese Aussage aufgrund der oben genannten ausdünnenden Datenbasis mit Vorsicht zu genießen. Mit Ausnahme des Jahres 2013 beobachten wir auch in den ersten zehn bis fünfzehn Lebensjahren eine sinkende Wahrscheinlichkeit für Stornierungen.

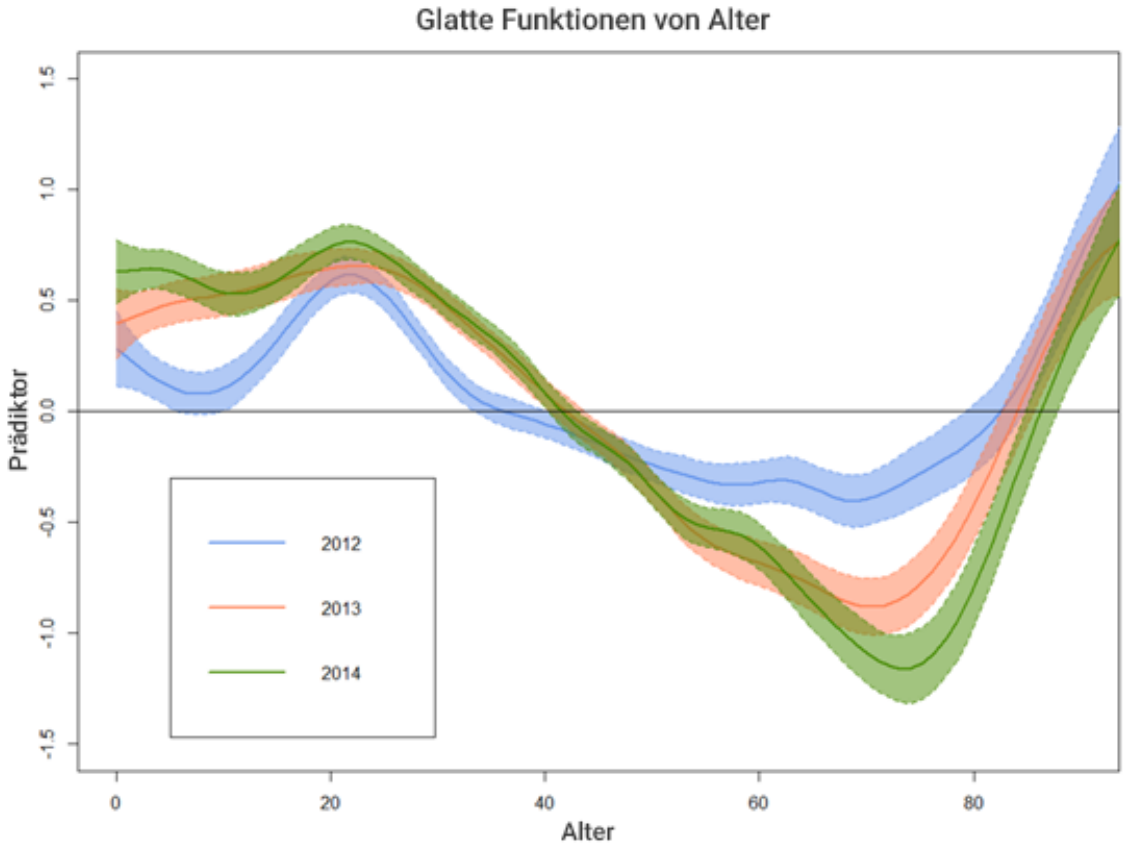

Abbildung 31: Vergleich der Glatten Funktionen für den Prädiktor Alter

Analog zu oben wollen wir nun den Einfluss der Höhe der monatlichen Prämie auf das Auftreten von Stornierungen untersuchen. Die entsprechenden Glatten Kurven und die dazugehörenden Schätzungen der Standardabweichung sind in Abbildung 32 ersichtlich. Wieder markiert die horizontale schwarze Linie die Null und somit jene Bereiche, in welchen ein positiver oder negativer Effekt auf die Stornowahrscheinlichkeit vorliegt. Es ist wieder eindeutig zu erkennen, dass sich das blaue Band (2012) markant von den anderen beiden Bändern unterscheidet. Somit gilt auch für die Abhängigkeitsstruktur zwischen Prämie und Storno, dass sich das Jahr 2012 von den anderen Jahren abhebt. Abgesehen davon scheint es aber für die monatlich vom Versicherungsnehmer zu zahlende Prämie einen klaren Effekt zu geben. Mit steigender Prämie steigt die Wahrscheinlichkeit für eine Stornierung. Dies gilt, mit Abstrichen, sogar für das Jahr 2012. Dort scheint es allerdings so, dass es eine erhöhte Anzahl an Stornierungen von Tarifen mit einer monatlich Prämie von um die 130 Euro gegeben hat. In unserer ersten groben Analyse des Prädiktors Prämie in Abschnitt 2.2.2 haben wir den relativen Anteil an Stornierungen für von uns eingeteilte Prämienklassen berechnet und in Abbildung 5 gegen die Prämie aufgetragen. Dort hatte es den Anschein, als würden eher Tarife mit einer geringen monatliche Prämie zur Stornierung neigen. Allerdings haben wir bereits dort darauf hingewiesen, dass die maßgebliche

Einteilung in die Prämienklassen von uns willkürlich vorgenommen wurde. Augenscheinlich hat dieses Vorgehen und die Tatsache, dass dort dieser Prädiktor separiert betrachtet wurde, eine Verfälschung der tatsächlichen Abhängigkeitsstruktur zur Folge.

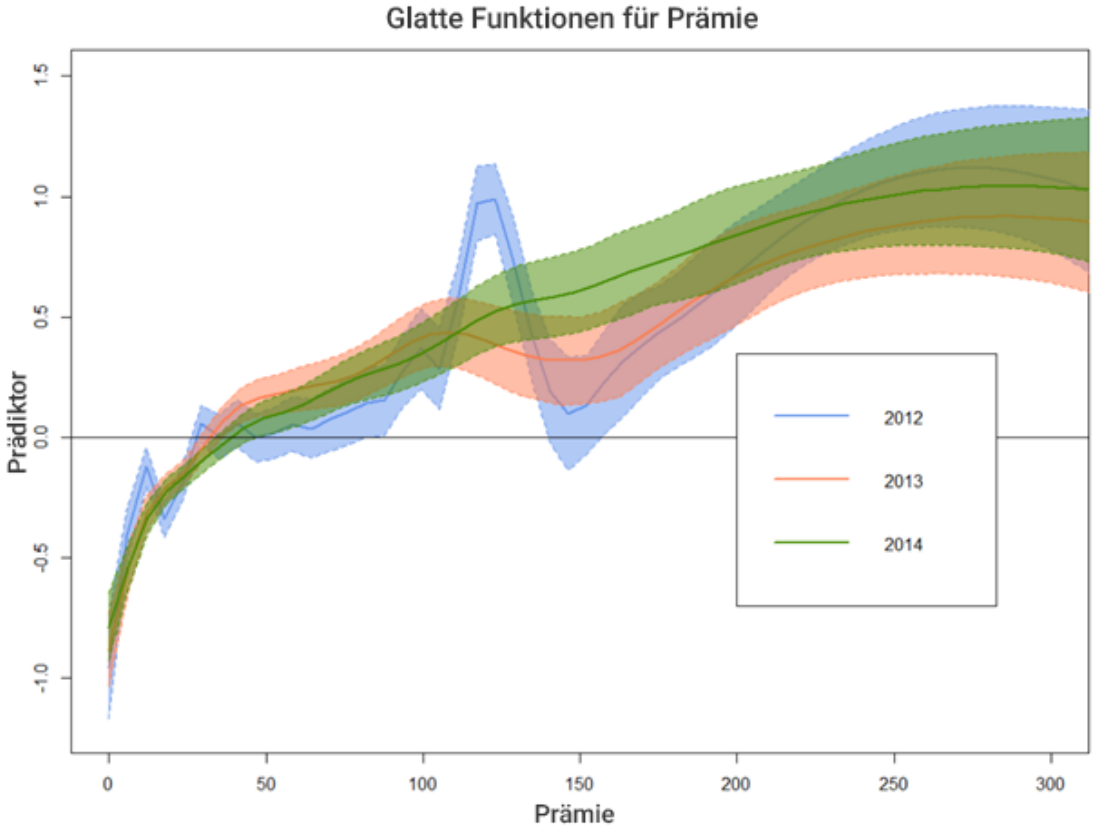

Abbildung 32: Vergleich der Glatten Funktionen für den Prädiktor Prämie

An dieser Stelle sei angemerkt, dass die Glatten Funktionen der Prämie nur für Prämien bis zu 300 Euro abgebildet sind. Es gibt zwar durchaus Tarife die eine noch höhere monatliche Prämie aufweisen, jedoch sind diese so selten, dass den entsprechenden Schätzungen nicht allzu viel Vertrauen geschenkt werden darf. Diesen Sachverhalt erkennen wir auch schon daran, dass die Bänder in Abbildung 32 nach rechts hin immer breiter werden.

Der dritte stetige Prädiktor ist die Laufzeit des Versicherungsvertrages. Die geschätzten Glatten Funktionen sind in Abbildung 33 ersichtlich. Im Vergleich zu den Glatten Funktionen von Alter und Prämie unterscheidet sich hier das Jahr 2012 (blau) nicht mehr auffällig von den anderen beiden Jahren. Lediglich die Stärke der Effekte ist im Jahr 2012 größer. Die grundsätzliche Art des Effektes bleibt aber über alle Jahre gleich. So erkennen wir die höchsten Stornowahrscheinlichkeiten für Tarife mit einer Laufzeit von weniger als zehn Jahren. Das Maximum scheint irgendwo bei um die fünf Jahre zu liegen.

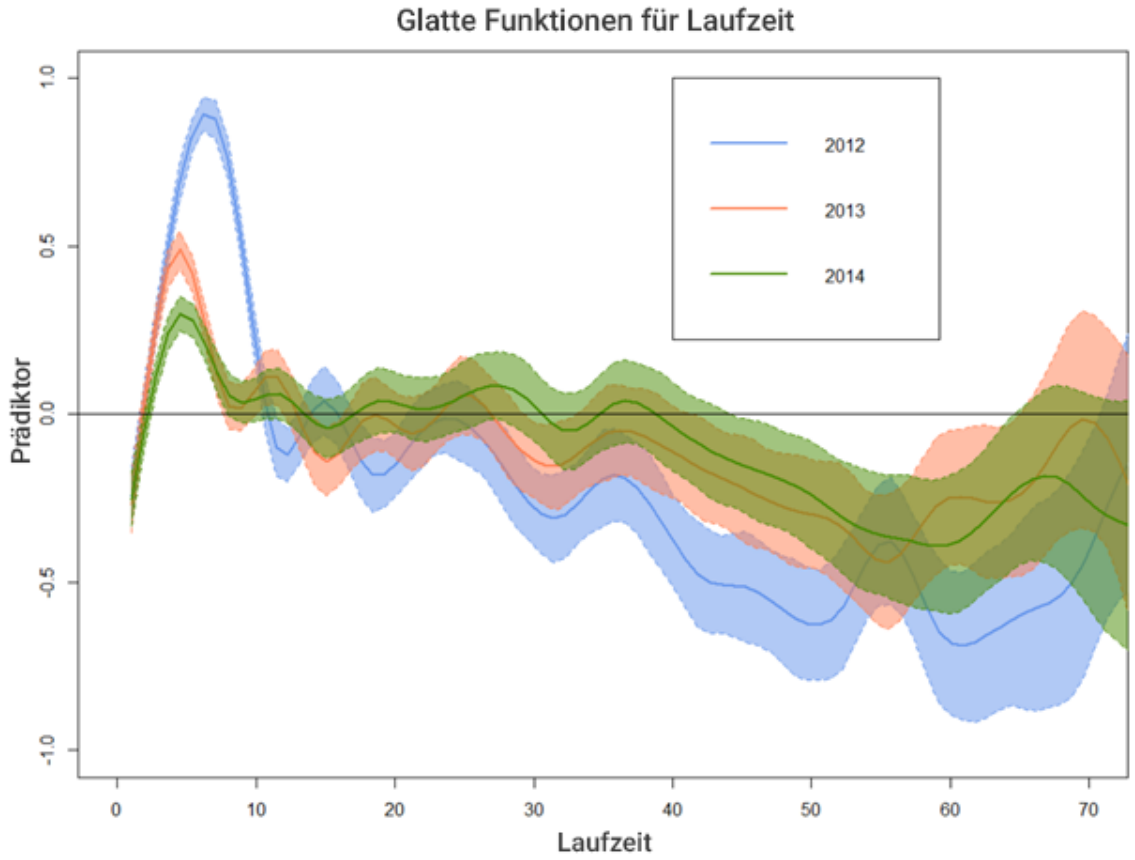

Abbildung 33: Vergleich der Glatten Funktionen für den Prädiktor Laufzeit

Für Tarife die eine Laufzeit von mehr als zehn Jahren haben, beobachten wir eine auffällig periodische Auf-und-ab-Bewegung, wobei die Wahrscheinlichkeit für eine Stornierung mit zunehmender Laufzeit abzunehmen scheint. So gibt es z.B. in allen Beobachtungsjahren für eine Laufzeit zwischen 35 und 40 Jahren eine erhöhte Stornowahrscheinlichkeit. Wieder gilt, dass für zunehmende Laufzeiten weniger Beobachtungen zur Verfügung stehen. Dies drückt sich abermals dadurch aus, dass die abgebildeten Bänder nach rechts hin breiter werden.

Durch die Modellierung als GAM ist es uns also möglich das Stornierungsverhalten in unterschiedlichen Datenmengen wie etwa verschiedenen Kalenderjahren direkt miteinander zu vergleichen. Durch die Wahl einer vergleichsweise einfachen Modellformel beinhalten alle in diesem Modell vorkommenden Prädiktoren in allen Jahren einen signifikanten Informationsgehalt. Dadurch ist auch gewährleistet, dass dieses Modell nicht zu sehr an die Spezifikas eines einzelnen Jahres angepasst ist.

6 Conclusio

Das Ziel dieser Arbeit war es, die Vorkommnis von Stornierungen in der privaten Krankenversicherung zu modellieren. Dabei sollte erklärt werden, wie die Wahrscheinlichkeit für eine Stornierung von den Charakteristiken des Versicherungsnehmers bzw. des Versicherungsvertrages abhängt. Ebenfalls von Interesse war ein Vergleich dieser Abhängigkeitsstruktur über verschiedene Zeiträume hinweg. Als Datenbasis diente uns zu diesem Zweck eine Auflistung aller offener Tarifpositionen von drei aufeinanderfolgenden Beobachtungsjahren. Diese Beobachtungsjahre waren die Kalenderjahre 2012, 2013 und 2014. Für jede beobachtete Tarifposition war bekannt, ob im vorliegenden Kalenderjahr eine Stornierung des jeweiligen Vertrages erfolgt ist oder nicht. Zusätzlich standen uns für jede Tarifposition weitere Informationen über den Versicherungsnehmer und den Versicherungsvertrag zur Verfügung. Die Informationen bezüglich des Versicherungsnehmers umfassten das Alter sowie das Geschlecht der jeweiligen Person. Zu jeder Tarifposition waren die Laufzeit, die Tarifklasse, die Gruppenzugehörigkeit, ein etwaige Rabattierung sowie die Höhe der monatlichen Prämie bekannt.

Die dieser Arbeit zugrunde liegende Idee war es, die Stornowahrscheinlichkeit mit Hilfe eines Regressionsmodells in Abhängigkeit der oben erwähnten erklärenden Variablen zu modellieren. Ein solches Modell wird dann jeweils auf Basis der Daten eines einzelnen Kalenderjahres geschätzt. Anhand dieser unterschiedlichen Schätzungen könnte dann die Abhängigkeitsstruktur zwischen Stornowahrscheinlichkeit und Prädiktoren über die verschiedenen Jahre hinweg verglichen werden. Außerdem bietet ein solches Modell die Möglichkeit, die Anzahl an Stornierungen eines Folgejahres auf Basis der Daten des aktuellen Kalenderjahres zu schätzen.

Da es also die Absicht war Wahrscheinlichkeiten zu modellieren, lag es nahe dies mit der Klasse der logistischen Regressionsmodelle zu tun. Die logistischen Regressionsmodelle sind ein Spezialfall der Generalisierten Linearen Modelle und demnach sowohl in der Theorie als auch für praktische Anwendungen sehr gut ausgearbeitet. Eine erste grobe Analyse der Daten hat jedoch gezeigt, dass eine lineare Modellierung der Abhängigkeitsstruktur in unserem Anwendungsfall nicht angebracht zu sein scheint. Aus diesem Grund fiel unsere Wahl der Modellklasse schlussendlich auf die Klasse der Generalisierten Additiven Modelle. Diese Modelle stellen dahingehend eine Erweiterung zu den klassischen Generalisierten Linearen Modellen dar, als dass es nun möglich ist nichtlineare Abhängigkeitsstrukturen zu modellieren. Einzig die Quantifizierung der Modellanpassung an die Daten ist aufgrund der unterschiedlichen Skalen von Modellschätzungen und beobachteten Responsevariablen für diese Klasse der Regressionsmodelle schwierig. Aus diesem Grund fokussierten wir uns bei der Wahl eines konkreten Modelles auf die Prädiktionsgüte.

Im Laufe der Modellselektion hat sich schnell herausgestellt, dass sowohl das Geschlecht des Versicherungsnehmers sowie die Gruppenzugehörigkeit der einzelnen Tarife wohl keinen

signifikanten Einfluss auf die Stornowahrscheinlichkeit haben. Im schlussendlichen Modell wurden die drei Prädiktoren Alter, Laufzeit und Prämie durch stetige, nichtlineare Glatte Funktionen modelliert. Die beiden Prädiktoren Tarifklasse und Rabatt sind als vier- bzw. zweistufige Faktoren im Modell abgebildet. Im Zuge der Modellevaluation hat sich herausgestellt, dass ein solches Modell den besten Kompromiss zwischen Komplexität und Datenanpassung bzw. Prädiktionsgüte liefert. So ist es z.B. möglich, die Summe der stornierten Prämien des Jahres 2014 auf Basis der Daten des Jahres 2013 auf 1.1% genau zu schätzen.

Mit Hilfe dieses Modells war es uns dann möglich die Daten im Hinblick auf das Stornierungsverhalten auszuwerten. Auffällig war dabei die Tatsache, dass die Daten der Jahre 2013 und 2014 sehr viel besser zueinander passen als dies für das Jahr 2012 der Fall ist. Nichtsdestotrotz war allen Jahren gemein, dass die Tarifklasse PG03 die höchsten Stornowahrscheinlichkeiten aufwies. Im Gegensatz dazu ist die Wahrscheinlichkeit einer Stornierung für Tarife der Klasse PG02b am geringsten. Auf Basis der geschätzten Glatten Funktionen konnten klare Strukturen in der Abhängigkeit zwischen Stornierung und Alter, Laufzeit und Prämie festgestellt werden. So stelle sich heraus, dass eine höhere monatliche Prämie zu einer höheren Stornowahrscheinlichkeit führt. Für das Alter konnte ein etwas komplexeres Muster ausfindig gemacht werden. So liegt vor allem in den ersten Lebensjahren sowie um das zwanzigste Lebensjahr eine erhöhte Wahrscheinlichkeit für eine Stornierung vor. Für ältere Versicherungsnehmer nimmt die Stornowahrscheinlichkeit sukzessive ab. Bei der Laufzeit beobachteten wir die höchsten Wahrscheinlichkeiten für eine Stornierung bei etwa fünf Jahren.

Abschließend wollen wir noch die Vor- und Nachteile unserer Vorgehensweise ansprechen. Die Modellierung der Stornowahrscheinlichkeit durch ein Generalisiertes Additives Modell liefert für jedes Kalenderjahr eine sehr schöne, einfach zu interpretierende Darstellung der Abhängigkeitsstrukturen. Durch den direkten Vergleich der Modelle der verschiedenen Jahre können wir zumindest Unterschiede in den modellierten Strukturen feststellen. Eine Quantifizieren als Wahrscheinlichkeit des Effektes jedes einzelnen Prädiktors ist jedoch nicht möglich. Auch kommt die zeitliche Komponente in dieser Vorgehensweise nur durch die unterschiedliche Modellschätzungen zum Ausdruck. Der Effekt des Kalenderjahres oder eine Abhängigkeit von der Zeit lässt sich so also nicht schätzen.

Literatur

Andres, J. (1986). *Verallgemeinerte lineare Modelle*. Beltz, Weinheim.

Björk, A. (1984). Generalized Least Squares Problems. *SIAM*.

de Boor (1978). A Practical Guide to Splines. *New York: Springer*.

Gu, C. (2002). Smoothing Spline ANOVA Models. *New York: Springer*.

Hastie, T. und Tibshiranie, R. (1990). *Generalized Additive Models*. Chapman & Hall/CRC, Boca Raton.

McCullagh, P. und Nelder, J. (1989). *Generalized Linear Models*. Chapman & Hall/CRC, Boca Raton. 2nd Edition.

Moosbrugger, J. (2015). Einlesen großer Datenmengen von Excel in R und eine erste Datenanalyse. *Projektbericht, Institut für Statistik, TU Graz*.

Nürnberger, G. (1989). *Approximation by Spline Functions*. Springer, Berlin Heidelberg.

Picard, R. und Cook, D. (1984). Cross-validation of regression models. *Journal of the American Statistical Association*, 79(387):575–583.

Wahba, G. (1990). Spline Models for Observational Data. *Philadelphia: SIAM*.

Watkins, D. S. (2008). The QR algorithm revisited. *SIAM Review*, 50(1):133–145.

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

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, 3:32–35.