TECHNISCHE UNIVERSITÄT WIEN
Vienna University of Technology

Diplomarbeit

Prämienkalkulation von Versicherungsprodukten mit Verallgemeinerten Linearen Modellen

Ausgeführt am

Institut für Stochastik und Wirtschaftsmathematik

unter der Leitung von

Herr Ao.Univ.Prof. Dipl.-Ing. Dr.techn. Peter Grandits

durch

Kevin Hochwarter, BSc

Maria Tusch Straße 27/10/13, 1220 Wien

Inhaltsverzeichnis

Abbildungsverzeichnis

Kapitel 1

Einleitung

Tarifierung[1] ist:

"Kalkulation der Nettorisikoprämie (Risikoprämie, Prämienkalkulation) eines Versicherungsvertrags nicht auf individueller Basis, sondern zum Zweck der Verringerung des Einflusses von Zufallsschwankungen in den Beobachtungsdaten auf der Basis der Einbettung von Risiken in ein Tarifkollektiv."

Wie jedes Unternehmen sind auch Versicherungen täglich damit beschäftigt den größtmöglichen Gewinn zu erwirtschaften und ihre Strategien dahingehend zu optimieren. Dabei gibt es eine Menge an verschiedenen Faktoren und Wege, dieses Ziel zu realisieren. Einen wesentlichen Prozess stellt dabei die Preisgestaltung der Versicherungsprodukte, die letztlich an den Kunden verkauft werden wollen, dar.

Das Ziel ist es, die höchstmögliche Prämie bei angemessener Versicherungsleistung zu erwirtschaften, die für den Kunden aber noch lukrativ erscheint. Natürlich hat nicht nur der Preis Einfluss auf das Kaufverhalten eines Versicherungsnehmers, jedoch ist der Preis in der Entscheidung oftmals nicht unwesentlich.

Gerade in Versicherungssparten, die einen sogenannten gesättigten Markt betreffen, wie es in Österreich beispielsweise bei der Kraftfahrzeug-Haftpflicht oder Haushaltsversicherung der Fall ist, gewinnt die Preisoptimierung sowie die genaue Selektion der Risiken (dem Versicherungsnehmer) immer mehr an Bedeutung. Doch genau diese Selektion wird auch aufgrund der immer strenger werdenden Datenschutzbestimmungen für Versicherer immer schwieriger.

Diese Arbeit setzt sich mit der Erstellung einer Scoring Tarifstruktur auseinander, die es möglich macht, jeden Kunden anhand von statistisch erfassten Parametern in eine Risikoklasse einzuordnen und somit eine angemessene Prämie anzubieten. Die Analyse der Daten und die Erstellung dieser Risikoklassen erfolgt mit Hilfe von Verallgemeinerten Linearen Modellen.

[1] vgl. [10] Definition

Verallgemeinerte Lineare Modelle stellen eine Erweiterung des klassischen Linearen Modelles dar, was nichts anderes als eine lineare Regression darstellt. Es handelt sich dabei um ein Analyseverfahren bei dem versucht wird, die beobachteten abhängigen Variablen durch unabhängige Variablen zu schätzen. In der Versicherung bedeutet das, dass versucht wird, das Schadenverhalten bzw. die Schadenfreiheit mit Hilfe aller zur Verfügung stehenden Variablen zu erklären um daraus die Prämie für zukünftige Verträge besser validieren zu können. Dabei ist es auch wichtig, herauszufinden welche erhobenen Parameter wesentlichen Einfluss auf das Schadenverhalten haben und welche nicht. Man spricht dabei von der Trennung der systematischen Komponenten vom Zufall.

Kapitel 2

Verallgemeinerte Lineare Modelle

Bei Linearen Modellen, die einer klassischen linearen Regression entsprechen, wird angenommen, dass die Responsevariablen (Zielvariablen) Y_i für $i = 1, \ldots, N$ stochastisch unabhängig und normalverteilte Größen mit Erwartungswert μ_i und konstanter Varianz σ^2 besitzten.

Eine lineare Regression ist ein statistisches Analysverfahren, dessen Ziel es ist, eine beobachtete Variable (Zielvariable) durch eine oder mehrere unabhängige Variablen zu beschreiben.

Allgemein formuliert, hat ein klassisches Lineares Modell die Form:

$$y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots x_{iK}\beta_K + \epsilon_i = x_i^T\beta + \epsilon_i, \qquad i = 1, 2, \cdots, N$$

Die kompakte Notation lautet:

$$y = X\beta + \epsilon$$

Dabei handelt es sich bei y um die Zielvariablen, X ist die Designmatrix der unabhängigen Variablen, β der Regressionskoeffizient mit dem die Daten geschätzt werden und ϵ ist der Fehlerterm. Die Zielvariablen y_i sind unanbhängig normalverteilt mit Varianz σ^2. Fasst man die Fehlerterme ϵ_i zusammen, so ist der Erwartunsgwert des Fehlerterms gleich 0.

Ein Lineares Modell entspricht somit der Form

$$E(Y_i) = \mu_i = x_i^T\beta \qquad Y_i \sim N(\mu_i, \sigma^2) \qquad \epsilon_i \sim N(0, \sigma^2)$$

das mit den unabhängigen Zufallsvariablen $Y_1, ..., Y_N$ die Basis vieler Datenanalysen darstellt. Jedoch ist die Annahme einer Normalverteilung nicht immer angemessen. Dazu betrachte man zum Beispiel Modelle für absolute Häufigkeit oder relative Anteile.

Weiterentwicklungen in der Statistik erlauben uns aber nun die Linearen Modelle anzupassen und in folgenden Punkten zu verallgemeinern:
Zum einen muss die Responsevariable nicht nur einer Normalverteilung entsprechen und zum anderen muss der Zusammenhang zwischen Responsevariable Y_i und der erklärenden Variable x_i (auch unabhängige Variable oder Regressor genannt) nicht linear sein.

Außerdem darf die Varianz vom Erwartungswert abhängen.

Der Fortschritt in der Verteilungsannahme für Verallgemeinerte Lineare Modelle (generalized linear model - GLM) wird im folgenden Unterkapitel genauer beschrieben. Anstatt der Normalverteilung kann für ein GLM auf eine Verteilung aus der Exponentialfamilie zurückgegriffen werden.

Ein weiterer Fortschritt liegt in der Erweiterung der numerischen Methoden um den Parameter β des vorhin beschriebenen Linearen Modelles zu schätzen. Dabei existiert eine nichtlineare Funktion bezüglich dem Erwartungswert und der linearen Komponente der Form

$$g(\mu_i) = x_i^T \beta.$$

Die Funktion g wird auch Linkfunktion genannt.

2.1 Exponentialfamilie von Verteilungen

In Verallgemeinerten Linearen Modellen betrachtet man eine Zufallsvariable Y, deren Dichte aus der Exponentialfamilie stammt.

Definition 2.1.1
Die Verteilung einer Zufallsvariable Y gehört zur Exponentialfamilie, wenn sich die Dichte- bzw. Wahrscheinlichkeitsfunktion in folgender Form darstellen lässt

$$f(y, \theta, \phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right) \forall y \in \mathbb{R}.$$

Die Funktionen $a(.)$, $b(.)$ und $c(.)$ sind messbare Funktionen mit $a(\phi) > 0$. Bei den übrigen Variablen handelt es sich um den kanonischen Parameter θ und den Dispersionsparameter ϕ. Kann ϕ als feste Größe betrachtet werden, handelt es sich bei $f(y, \theta, \phi)$ um eine einparametrige Exponentialfamilie.

Nun werden einige wichtige Verteilungen der Exponentialfamilie erläutert.

Normalverteilung

Für die Verteilung $Y \sim N(\mu, \sigma^2)$ ist die Dichtefunktion durch

$$\begin{aligned} f(y, \mu, \sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) \\ &= \exp\left(\frac{y\mu - \frac{\mu^2}{2}}{\sigma^2} - \frac{y^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right), y \in \mathbb{R} \end{aligned}$$

gegeben. Setzt man in die Definition 2.1.1 die Parameter $\theta = \mu$ und $\phi = \sigma^2$, so führt dies zur Exponentialfamilie mit folgenden Funktionen:

$$\begin{aligned} a(\phi) &= \phi \\ b(\theta) &= \frac{\theta^2}{2} \\ c(y, \phi) &= -\frac{y^2}{2\phi} - \frac{1}{2}\log(2\pi\phi) \end{aligned}$$

Die Normalverteilung wird für Modelle mit stetigen oder kontinuierlichen Daten verwendet, die eine symmetrische Verteilung haben. Sie wird aus mehreren Gründen bevorzugt verwendet. Viele Naturphänomene, wie zum Beispiel Körpergröße oder der Blutdruck von Menschen, sind weitgehend durch eine Normalverteilung beschrieben. Außerdem ist der Durchschnitt oder ein zufälliger Ausschnitt von Daten, die nicht normalverteilt sind, annäherungsweise normalverteilt. Dieses Ergebnis liefert der Zentrale Grenzwertsatz.

Poissonverteilung

Für die Verteilung $Y \sim P(\mu)$ ist die Dichtefunktion durch

$$f(y,\mu) = \frac{\mu^y}{y!}\exp^{-\mu} = exp(y\log\mu - \mu - \log y!,) \qquad\qquad y = 1,2,3,...$$

gegeben. Mit den Parametern $\theta = \log\mu$ und konstantem $\phi = 1$ erhält man die Exponentialfamilie mit folgenden Funktionen:

$$\begin{aligned} a(\phi) &= \phi \\ b(\theta) &= \exp(\theta) \\ c(y,\phi) &= -\log y! \end{aligned}$$

Die Poissonverteilung wird für Modelle herangezogen, die sich mit der Anzahl von Gegebenheiten beschäftigen. In der Praxis handelt es sich dabei oft um die Anzahl von Ereignissen in einem bestimmten Zeitintervall, wenn die Eintrittswahrscheinlichkeit sehr gering ist und die Ereignisse unabhängig voneinander sind. Wie wir später sehen werden, wird für die Modellierung von Schadenanzahlmodellen eines Versicherungsportfolios die Annahme einer Poissonverteilung verwendet.

Wenn eine Zufallsvariable poissonverteilt ist, so sind der Erwartungswert und die Varianz gleich. Bei der Modellierung von offensichtlich poissonverteilten Daten, kommt es oft vor, dass die Verteilung eine höhere Varianz hat. Man spricht dann von einer "overdispersion", einer Abweichung von der Poissonverteilung und muss die Modelle gegebenenfalls anpassen.

Gammaverteilung

Für die Verteilung $Y \sim \Gamma(a,\lambda)$ ist die Dichtefunktion durch

$$f(y,a,\lambda) = \exp(-\lambda y)y^{a-1}\frac{\lambda^a}{\Gamma(a)}, \qquad a,\lambda,y > 0$$

gegeben. Damit gilt, dass der Erwartungswert die Form $E(y) = \frac{a}{\lambda}$ und die Varianz die Form $V(y) = \frac{a}{\lambda^2}$ hat. Durch die Reparametrisierung von $\mu = \frac{\nu}{\lambda}$ mit $\nu = a$ erhält man für den Erwartungswert $E(y) = \mu$ und für die Varianz $V(y) = \frac{\mu^2}{\nu}$. Die Dichtefunktion kann man wie folgt darstellen:

$$\begin{aligned} f(y,\mu,\nu) &= \exp(-\frac{\nu}{\mu}y)(\frac{\nu}{\mu})^\nu y^{\nu-1}\frac{1}{\Gamma(\nu)} \\ &= \exp\left((-\frac{\nu}{\mu}y) + \nu\log\nu - \nu\log\mu + (\nu-1)\log y - \log\Gamma(\nu)\right) \\ &= \exp\left(\frac{y(-\frac{1}{\mu}) + \log(\frac{1}{\mu})}{\frac{1}{\nu}} + \nu\log\nu + (\nu-1)\log y - \log\Gamma(\nu)\right), \qquad \mu,\nu,y > 0 \end{aligned}$$

Mit $\theta = -\frac{1}{\mu}$ und $\phi = \frac{1}{\nu}$ erhält man die Exponentialfamilie mit folgenden Funktionen:

$$\begin{aligned} a(\phi) &= \phi \\ b(\theta) &= -\log(-\theta) \\ c(y,\phi) &= \frac{1}{\phi}\log\frac{1}{\phi} + \left(\frac{1}{\phi} - 1\right)\log y - \log\Gamma\left(\frac{1}{\phi}\right) \end{aligned}$$

Für Versicherungsunternehmen eignet sich die Annahme einer Gammaverteilung für Modelle die sich mit der Schadenshöhe von Einzelschäden auseinandersetzen.

2.2 Parametrisierung

Das folgende Kapitel beschäftigt sich mit der Parametrisierung der Klasse der Verallgemeinerten Linearen Modelle. Im Wesentlichen kann man drei Komponenten unterscheiden[1]:

1. stochastische Komponente: $y_i \sim$ Exponentialfamilie (θ_i) mit $E(y_i) = \mu_i = \mu(\theta_i)$
2. systematische Komponente: $\eta_i = x_i^\top \beta$
3. Linkfunktion: $g(\mu_i) = \eta_i$

Dabei bezeichnet $y = (y_1, \cdots, y_n)^\top$ den Zufallsvektor bestehend aus n unabhängigen Komponenten y_i mit $E(y_i) = \mu_i$ und $Var(y_i) = a_i \phi V(\mu_i)$ (siehe 2.3 Parameterschätzung). Des weiteren bezeichnet der Vektor $x_i = (x_{i1}; \cdots, x_{ik})^\top$ die Kovariablen bzw. den Vektor der bekannten erklärenden Variablen.

Im folgenden Abschnitt wird näher auf die oben angeführten Komponenten eingegangen[2].

Stochastische Komponente

Die Beobachtungen des Zufallsvektors y_i, $i = 1, \ldots, n$ haben eine Dichtefunktion aus der Exponentialfamilie von Verteilungen mit Erwartungswert μ_i. Allgemein formuliert, kann die Dichtefunktion als

$$f_{Y_i}(y_i, \theta_i, \phi) = exp\left(\frac{y_i\theta_i - b(\theta_i)}{\frac{a(\phi)}{\omega_i}} + c(y_i, \phi, \omega_i)\right)$$

mit den bekannten Gewichten ω_i und $\sum_{i=1}^{n} \omega_i = 1$ geschrieben werden. Damit hängt die Verteilung der Zufallsvariablen von einem n-dimensionalen Parameter $\theta = (\theta_i, \cdots, \theta_n)$ bei gegebenen Dispersionsparamter ϕ ab.

Da bei Betrachtung von n Beobachtungen sowie n Parametern die Erstellung einer aussagekräftigen Statistik nicht möglich ist, ist es das Ziel, dass man eine Abhängigkeit der Verteilung rein von k exogenen Größen $\beta_1, \cdots, \beta_k$ erreicht. Um dieses Ziel zu erreichen, benötigt man die oben erwähnten weiteren Komponenten.

Systematische Komponente

Zu jeder Beobachtung y_i des Zufallsvektors y existieren verschiedene unabhängige Merkmale x_{ij}, $j = 1, \ldots, k$. Die Linearkombination dieser Kovariablen wird zu einem linearen Prädiktor $\eta = (\eta_1, \cdots, \eta_n)$ zusammengefasst. Jede Kovariable hat einen speziellen Einfluss auf die abhängige Variable, gemessen durch β_j, $j = 1, \ldots, k$. Für den linearen Prädiktor gilt:

[1] vgl. [5] Seite 19
[2] vgl. [4] Seite 7-8

$$\eta := \sum_{j=1}^{k} \beta_j x_{ij} \qquad bzw. \qquad \eta := X\beta \qquad mit \qquad X = (X_1, \cdots, X_n)^\top \in \mathbb{R}^{n \times k}$$

Die Matrix X wird auch als Design Matrix bezeichnet.

Linkfunktion

Die Linkfunktion $g(.) : \mathbb{R} \to \mathbb{R}$ verbindet die stochastische Komponente mit der systematischen Komponente durch eine Transformation des Erwartungswertes. Diese Funktion wird außerdem als monoton und differenzierbar vorausgesetzt. Es gilt:

$$g(\mu_i) = \eta_i = \sum_{j=1}^{k} x_{ij}\beta_j = x_i^t\beta, \qquad\qquad i = 1, \cdots, n$$

Das bedeutet, dass der Mittelwert μ_i der i-ten Beobachtung von den unbekannten Parametern $\beta_1, \cdots, \beta_k$ abhängt. Eine Linkfunktion ist kanonisch, falls $\eta_i = \theta_i \; \forall \; i = 1, \cdots, n$.

Grundlegende Unterschiede zwischen dem Verallgemeinerten Linearen Modell und dem Linearen Modell sind unter anderem:

1. keine allgemeine Additivität bezüglich der nicht beobachtbaren Fehlerterme ϵ_i wie beim einfachen Modell
2. die Varianz kann vom Erwartungswert abhängen
3. eine Funktion des Erwartungswertes wird linear modelliert

2.3 Parameterschätzung

Im folgenden Abschnitt werden wir uns mit der Schätzung des Modellparameters auseinandersetzen. Dies erfolgt mit der gängigen Maximum Likelihood-Methode.

Da für die Bestimmung des Erwartungswertes und der Varianz der Verteilungen aus der Exponentialfamilie die Betrachtung der Log-Likelihoodfunktionen wesentlich ist, folgen zunächst einige wichtige Sätze.

Satz 2.3.1
Für die Ableitung der Log-Likelihood Funktion $l(y,\theta) = \log f(y,\theta)$ gilt unter den Regularitätsbedingungen:

1. $E\left(\frac{\partial l(y,\theta)}{\partial \theta}\right) = 0$
2. $E\left(\frac{\partial l(y,\theta)}{\partial \theta}\right)^2 - E\left(-\frac{\partial^2 l(y,\theta)}{\partial \theta^2}\right) = 0$

Mit

$$\frac{\partial l(y,\theta)}{\partial \theta} = \frac{1}{f(y,\theta)}\frac{\partial f(y,\theta)}{\partial \theta} \qquad und \qquad \int_{\mathbb{R}} f(y,\theta)dy = 1$$

folgt 1., da

$$E\left(\frac{\partial l(y,\theta)}{\partial \theta}\right) = E\left(\frac{\partial f(y,\theta)}{\partial \theta}\frac{1}{f(y,\theta)}\right) = \int_{\mathbb{R}} \frac{\partial f(y,\theta)}{\partial \theta}dy = \frac{\partial}{\partial \theta}\int_{\mathbb{R}} f(y,\theta)dy = 0.$$

Punkt 2. gilt aufgrund der Kettenregel

$$\begin{aligned} E\left(-\frac{\partial^2 l(y,\theta)}{\partial \theta^2}\right) &= E\left(-\frac{\partial^2 f(y,\theta)}{\partial \theta^2}\frac{1}{f(y,\theta)} + \frac{\partial f(y,\theta)}{\partial \theta}\frac{\partial f(y,\theta)}{\partial \theta}\frac{1}{f(y,\theta)^2}\right) \\ &= -\int_{\mathbb{R}} \frac{\partial^2 f(y,\theta)}{\partial \theta^2}dy + \int_{\mathbb{R}} \frac{\partial l(y,\theta)}{\partial \theta}\frac{\partial l(y,\theta)}{\partial \theta}f(y,\theta)dy \\ &= E\left(\frac{\partial l(y,\theta)}{\partial \theta}\right)^2. \end{aligned}$$

Mit dem Satz 2.3.1 ergeben sich folgende Resultate für den Erwartungswert bzw. für die Varianz von Exponentialfamilien.

Für den Erwartungswert gilt:

$$E\left(\frac{\partial l(y,\theta)}{\partial \theta}\right) = \frac{E(y - b'(\theta))}{a(\phi)} = 0 \Leftrightarrow E(y) = b'(\theta) = \mu$$

Für die Varianz gilt:

$$\begin{aligned} 0 = E\left(\frac{\partial^2 l(y,\theta)}{\partial\theta^2}\right) + E\left(\frac{\partial l(y,\theta)}{\partial\theta}\right)^2 \\ \Leftrightarrow 0 = -\frac{b''(\theta)}{a(\phi)} + \frac{var(y)}{a^2(\phi)} \\ \Leftrightarrow var(y) = a(\phi)b''(\theta) \end{aligned}$$

Somit ist die Varianz ein Produkt von zwei Funktionen. Die Funktion $b(.)$, welche Kumulantenfunktion genannt wird, ist eine von θ und damit vom Erwartungswert abhängige Funktion, wobei es sich bei der Funktion $a(\phi)$ um eine vom Erwartungswert unabhängige Funktion handelt. Außerdem wird vorausgesetzt, dass die Funktion $b(.)$ zweimal differenzierbar und somit invertierbar ist. Durch folgende Umformung

$$b'(\theta) = \mu \Rightarrow \theta = b'^{-1}(\mu)$$

erhält man die Varianzfunktion

$$V(\mu) = b''(b'^{-1}(\mu)) = b''(\theta).$$

Diese beschreibt den Einfluss des Erwartungswertes μ auf die Varianz von y.

Satz 2.3.2 Wedderburn (1974)
Für eine Beobachtung y mit $E(y) = \mu$ und $var(y) = \phi V(\mu)$ hat die Log-Likelihoodfunktion $l(y,\mu) = \log f(y,\mu)$ die Eigenschaft

$$\frac{\partial l(y,\mu)}{\partial\mu} = \frac{y-\mu}{\phi V(\mu)},$$

dann und nur dann, wenn die Dichte bzw. Wahrscheinlichkeitsfunktion von y in der Form

$$\exp\left(\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\right)$$

geschrieben werden kann, wobei θ eine Funktion von μ und ϕ unabhängig von μ ist.

Satz 2.3.3
Liegt eine n-elementige Zufallsstichprobe $y_1, \ldots, y_n$ aus der Exponentialfamilie vor, so ist der Maximum Likelihood-Schätzer für μ die Nullstelle der Scorefunktion

$$\sum_{i=1}^{n} \frac{\partial l(y_i, \theta)}{\partial \mu} = \sum_{i=1}^{n} \frac{\partial l(y_i, \theta)}{\partial \theta} \frac{\partial \theta}{\partial \mu} = \sum_{i=1}^{n} \frac{y_i - b'(\theta)}{a(\phi)} \frac{\partial \theta}{\partial \mu}.$$

Mit $b'(\theta) = \mu$ und wegen

$$\frac{\partial \mu}{\partial \theta} = \frac{\partial b'(\theta)}{\partial \theta} = b''(\theta) = V(\mu)$$

vereinfacht sich die Score-Funktion zu

$$\sum_{i=1}^{n} \frac{\partial l(y_i, \theta)}{\partial \mu} = \sum_{i=1}^{n} \frac{y_i - \mu}{a(\phi) V(\mu)} = \sum_{i=1}^{n} \frac{y_i - \mu}{var(y)}.$$

2.3.1 Maximum Likelihood-Schätzung

Das Ziel wird es nun sein, einen geeigneten Schätzer für den Parametervektor β zu konstruieren.

Es werden die Zufallsvariablen $y_1, \ldots, y_n$ mit folgender Dichte betrachtet:

$$f_{Y_i}(y_i, \theta_i, \phi) = exp\left(\frac{y_i \theta_i - b(\theta_i)}{\frac{a(\phi)}{\omega_i}} + c(y_i, \phi, \omega_i) \right)$$

Bei $\omega_1, \ldots, \omega_n$ handelt es sich um bekannte Volumsmaße. Für die Likelihoodfunktion gilt somit

$$L(y, \theta, \phi) = \prod_{i=1}^{n} f_{Y_i}(y_i, \theta_i, \phi).$$

In weiterer Folge wird man durch das Maximieren der Maximum Likelihoodfunktion versuchen, einen geeigneten Maximum Likelihood-Schätzer zu finden. Dabei sucht man den Parameter $\hat{\beta}$, für den die Wahrscheinlichkeit der beobachteten Stichprobe maximal ist.

Wie schon zu Beginn des Kapitels erwähnt, ist die Betrachtung der Log-Likelihoodfunktion im Zusammenhang mit Exponentialfamilien wesentlich. Vor allem ist die Maximierung der Log-Likelihoodfunktion oft leichter, als bei der Likelihoodfunktion.

Die Log-Likelihoodfunktion ist gegeben durch:

$$l(y, \theta, \phi) = \ln L(y, \theta, \phi) = \ln \prod_{i=1}^{n} f_{Y_i}(y_i, \theta_i, \phi)$$

Falls $y_1, \ldots, y_n$ unabhängige Responses sind und die y_i aus der selben Exponentialfamilie stammen mit Parameter (θ_i, ϕ_i) (dabei beschreibt der Vektor $\theta = (\theta_1, \ldots, \theta_n)^\top$ die unbekannten Parameter, welche geschätzt werden sollen) und $\phi = (\phi_1, \ldots, \phi_n)^\top$ aus bekannten Komponenten besteht, dann gilt für die Dichte:

$$l(y, \theta, \phi) = \sum_{i=1}^{n} \left(\frac{(y_i \theta_i - b(\theta_i))}{\phi} + c(y_i, \phi) \right).$$

Mit dem Satz 2.3.3 und der allgemeinen Annahme $\mu = \mu(\beta)$ folgt die Score-Gleichung

$$\frac{\partial l(y, \theta(\beta))}{\partial \beta_j} = \sum_{i=1}^{n} \frac{y_i - \mu_i}{\phi_i V(\mu_i)} \frac{\partial \mu_i}{\partial \beta_j} = 0, \qquad j = 1, \ldots, k.$$

Mit $\mu_i = b'(\theta_i)$ und der Linkfunktion $g(\mu_i) = \eta_i = \sum_{j=1}^{k} x_{ij}\beta_j$ bzw. $\mu_i = g^{-1}(\eta_i)$ gilt

$$\frac{\partial \mu_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial g(\mu_i)} x = \frac{x}{g'(\mu_i)}.$$

Zusammen ergibt das folgende Score-Funktion:

$$\frac{\partial l(y, \theta(\beta))}{\partial \beta_j} = \sum_{i=1}^{n} \frac{y_i - \mu_i}{\phi_i V(\mu_i)} \frac{x_{ij}}{g'(\mu_i)} = 0, \qquad j = 1, \ldots, k.$$

Falls diese Score Gleichung eine eindeutige Lösung $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_k)$ besitzt, ist dies der Maximum Likelihood-Schätzer.

Spezialfall: Kanonische Linkfunktion

Den speziellen Link $g(\mu) = \theta$ nennt man die kanonische Linkfunktion. Dabei wird der Parameter θ direkt durch den linearen Prädiktor η modelliert. In diesem Fall ist g(.) die Inverse von $b'(.)$ und wegen $\mu = b'(\theta)$ folgt

$$g'(\mu) = \frac{\partial g(\mu)}{\partial \mu} = \frac{\partial \theta}{\partial \mu} = \frac{1}{b''(\theta)} = \frac{1}{V(\mu)}$$

Die Score Gleichung vereinfacht sich dadurch zu

$$\frac{\partial l(y,\theta(\beta))}{\partial \beta_j} = \sum_{i=1}^{n} \frac{y_i - \mu_i}{\phi_i} x_{ij} = 0, \qquad j = 1, \ldots, k.$$

Beide Gleichungssysteme können nur iterativ gelöst werden.

2.3.2 Lösungsverfahren

Mit der Newton-Raphson Methode ist es möglich, die vorher gewonnenen Score Funktionen zu lösen.

Die Newton-Raphson Methode liefert folgende Iterationsvorschrift für den k+1-ten Schritt[3]:

$$\beta^{k+1} = \beta^k + \left(-\frac{\partial^2 l(y,\theta(\beta))}{\partial\beta\partial\beta^\top} \right)^{-1} \frac{\partial l(y,\theta(\beta))}{\partial\beta}, \qquad k = 0, 1 \ldots,$$

wobei beide Ableitungen der rechten Seite obiger Iterationsvorschrift an der Stelle β^k betrachtet werden. In Matrixnotation folgt für den Score-Vektor

$$\frac{\partial l(y,\theta(\beta))}{\partial\beta} = X^\top DW(y-\mu),$$

mit $D = diag(d_i)$ und $W = diag(\omega_i)$, wobei

$$\begin{aligned} d_i &= g'(\mu_i), \\ \frac{1}{\omega_i} &= \phi V(\mu_i)(g'(\mu_i))^2. \end{aligned}$$

Als negative Hessematrix der Log-Likelihood Funktion resultiert somit

$$-\frac{\partial^2 l(y,\theta(\beta))}{\partial\beta\partial\beta^\top} = -X^\top \left(\frac{\partial DW}{\partial\eta^\top}(y-\mu) - DW\frac{\partial\mu}{\partial\eta^\top} \right) X = X^\top \left(W - \frac{\partial DW}{\partial\eta^\top}(y-\mu) \right) X,$$

wegen $\frac{\partial\mu}{\partial\eta} = D^{-1}$. Des weiteren ist

$$\begin{aligned} \frac{\partial d_i\omega_i}{\partial\eta_i} &= -\frac{\phi_i V'(\mu_i)\frac{\partial\mu_i}{\partial\eta_i} g'(\mu_i) + \phi_i V(\mu_i) g''(\mu_i)\frac{\partial\mu_i}{\partial\eta_i}}{(\phi_i V(\mu_i) g'(\mu_i))^2} \\ &= -\frac{V'(\mu_i)g'(\mu_i) + V(\mu_i)g''(\mu_i)}{\phi_i V^2(\mu_i) g'^3(\mu_i)}. \qquad (*) \end{aligned}$$

Fasst man die Elemente

[3] vgl. [5] Seite 21

$$\omega_i^* = \omega_i - \frac{\partial d_i \omega_i}{\partial \eta_i}(y_i - \mu_i)$$

zusammen zur Diagonalmatrix W^*, für die $E(W^*) = W$ gilt, so resultiert als Newton-Raphson Vorschrift

$$\beta^{k+1} = \beta^k + (X^\top W * X)^{-1} X^\top DW(y - \mu), \qquad t = 0, 1, \ldots$$

Mit sogenannten Pseudobeobachtungen (adjusted dependent variates)

$$z = X\beta + W^{*-1} DW(y - \mu)$$

kann die Newton-Raphson Vorschrift in eine Iterative (Re)Weighted Least Squares Notation (IWLS- oder IRLS-Prozedur) umgeschrieben werden

$$\beta^{k+1} = (X^\top W * X)^{-1} X^\top W * z,$$

wobei hier die rechte Seite in β^k betrachtet wird.
Für die kanonische Linkfunktionen $g'(\mu) = \frac{1}{V(\mu)}, g''(\mu) = -\frac{V'(\mu)}{V^2(\mu)}$ verschwinden die Ableitungen in (*), da

$$\frac{\partial d_i \omega_i}{\partial \eta_i} = \frac{\frac{V'(\mu_i)}{V(\mu_i)} - \frac{V'(\mu_i)}{V(\mu_i)}}{\frac{\phi_i}{V(\mu_i)}} = 0$$

und es gilt $W^* = W$. Die Pseudobeobachtungen vereinfachen sich dadurch zu

$$z = X\beta + D(y - \mu) = X\beta + V^{-1}(y - \mu),$$

mit $V = diag(V(\mu_i))$, weshalb als Iterationsvorschrift

$$\beta^{k+1} = (X^\top W X)^{-1} X^\top W z$$

folgt.

Um auch für nicht-kanonische Linkfunktionen ein einfaches Schema zu haben, ist es üblich, anstelle der beobachteten negativen Hessematrix deren Erwartungswert (Informationsmatrix) zu verwenden. Da $E(X^\top W * X) = X^\top W X$ gilt, folgt bei der Fisher Scoring Technik wiederum als Iterationsvorschrift

$$\beta^{k+1} = (X^\top W X)^{-1} X^\top W z$$

mit den Pseudobeobachtungen

$$z = X\beta + D(y - \mu).$$

Dafür ist $E(z) = X\beta$ und $var(z) = Dvar(y)D = W^{-1}$. Dies bedeutet, dass die Gewichte W gerade die reziproken Varianzen der Pseudobeobachtungen beschreiben.

Im Allgemeinen existieren keine analytischen Lösungen für einen Maximum Likelihood-Schätzer in Verallgemeinerten Linearen Modellen. Aufgrund dessen können nur asymptotische Eigenschaften hergeleitet werden. Nach Fahrmeir und Kaufmann (1985), siehe Referenz [1], können die asymptotischen Eigeschaften nur dann sichergestellt werden, wenn die Regularitätsbedingungen gelten. Unter gewissen Voraussetzungen erhält man einen asymptotisch normalverteilten Maximum Likelihood-Schätzer.

2.4 Anpassungsgüte des Modells

Jedes Modell muss am Ende plausibilisiert werden. Dabei ist es wesentlich, ob die verwendeten Daten hinreichend gut beschrieben wurden. Ziel der Modellierung muss es sein, durch die Analyse der gesamten Daten bzw. Merkmale diejenigen auszuschließen, die keinen Einfluss auf die Zielvariable haben.

Im Grunde genommen wird die ideale Modellanpassung zwischen dem Nullmodell und dem saturierten (volles) Modell gesucht. Das Nullmodell behinhaltet nur den Mittelwert aller Beobachtungen, wohingegen das saturierte Modell alle Parameter beinhaltet. Das saturierte Modell liefert eine perfekte Modellanpassung, jedoch ist es nicht aussagekräftig, da es alle Daten nur wiederholt und nicht auswertet. Als Gütekriterium der Modellanpassung dienen unter anderem die verallgemeinerte Pearson Statistik sowie die Devianz. Des Weiteren dienen die Informationskriterien AIC (Akaike Informationskriterium) und BIC (Bayessches Informationskriterium) als Hilfe für die Variablenselektion.

2.4.1 Pearson-Statistik

Für die Pearson Statistik gilt

$$X^2 = \frac{1}{n} \sum_{i=1}^{n} \frac{(y_i - \hat{\mu_i})^2}{a_i var(\hat{\mu_i})}$$

wobei es sich bei $\hat{\mu_i}$ um einen geschätzten Wert für μ_i handelt, der mit Hilfe von $\hat{\beta_i}$ ermittelt wurde, und $var(\mu_i)$ ist die zugehörige geschätzte Varianzfunktion. Die einzelnen Summanden der Pearson Statistik

$$r_i = \frac{y_i - \hat{\mu_i}}{\sqrt{a_i var(\hat{\mu_i})}}$$

nennt man Pearson Residuen.Im Fall von normalverteilten Daten liegt eine exakte χ^2_{n-k} -Verteilung vor, denn dann besteht die Pearson Statistik aus einer Summe von quadrierten standardisierten normalverteilten Zufallsvariablen. In anderen Fällen ist die Pearson Statistik asymptotisch χ^2_{n-k} -verteilt. Im Allgemeinen kann man festhalten, dass Modelle mit einem größeren Wert der Pearson Statistik eine schlechtere Modellanpassung haben als Vergleichsmodelle mit einem kleinerem Wert.

2.4.2 Devianz

Wie zu Beginn des Kapitels schon erwähnt wurde, ist die Richtlinie einer guten Anpassung eines Modells oftmals mit dem saturierten Modell verbunden. In vielen Fällen haben die saturierten Modelle eine perfekte Modellanpassung und haben die Eigenschaft

$$-2 \log Likelihood(saturiertesModell) = 0$$

Die Devianz ist definiert als die Differenz zwischen dem saturiertem Modell und dem gefitteten Modell.
Für die skalierte Devianz gilt:

$$\frac{D(y,\mu)}{\phi} = -2\ln\left(\frac{\log f(y,\mu)}{\log f(y,y)}\right) = 2\log f(y,y) - 2\log f(y,\mu) = 2(\log f(y,y) - \log f(y,\mu)).$$

Nun gilt für den Maximum Likelihood-Schätzer $\hat{\mu}$, dass dieser die Devianz minimiert.

2.4.3 Akaike Informationskriterium (AIC)

Das Akaike Informationskriterium ist definiert als

$$AIC = -2l(\hat{\theta}) + 2k = -2l(y,\hat{\beta},\phi) + 2k.$$

Dabei ist l die Log-Likelihood Funktion, $\hat{\theta}$ der geschätzte Maximum Likelihood Parameter und k die Anzahl der zu schätzenden Parameter im Modell, der als Strafterm für die Komplexität interpretiert werden kann. Der Strafterm verhindert ein Overfitting, da die Erhöhung der Anzahl der Parameter im Modell die Anpassungsgüte fast immer erhöht. Im Allgemeinen versucht man Modelle mit einem möglichst kleinen AIC Wert zu finden.

Nachteil: Der Strafterm $2k$ ist von der Stichprobengröße n unabhängig. Dadurch kommt es zum Effekt, dass das AIC bei großen Stichproben eher Modelle mit vielen Parametern begünstigt.

2.4.4 Bayessches Informationskriterium

Das Bayessche Informationskriterium ist definiert als

$$BIC = -2l(\hat{\theta}) + k\log n = -2l(y,\hat{\beta},\phi) + k\log n.$$

Hier wächst der Strafterm logarithmisch zur Größe der Stichprobe. Das BIC bestraft Modelle mit mehreren Parametern ab einer Stichprobengröße von acht Beobachtungen höher als das AIC.

Kapitel 3

Versicherung

Die Aufgabe einer Versicherung stellt das Angebot von Versicherungsverträgen bzw. Versicherungsprodukten dar, die im Falle eines Schadensereignisses den Versicherungsnehmer vor finanziellen Schwierigkeiten schützen.

Bei Abschluss eines Versicherungsvertrages gehen beide Parteien gewisse Pflichten ein. Zum einen hat der Versicherungsnehmer die Pflicht alle zum Vertragsabschluss relevanten Informationen wahrheitsgetreu anzugeben, sowie nach Abschluss die vereinbarte Versicherungsprämien zu bezahlen. Das Versicherungsunternehmen hat daraufhin die Pflicht die im Rahmen des Versicherungsvertrages festgesetzte Leistungen im Schadensfall zu erbringen.

Damit das Versicherungsunternehmen in der Lage ist den Schadensleistungen nachzukommen, muss es ausreichend Prämien einnehmen. Dazu gehört der Aufbau eines Portfolios, das durch geeignete Auswahl an Risiken - in dem Fall besteht das Risiko aus dem Versicherungsnehmer - mehr Prämien einnimmt, als an Schadenszahlungen wieder abgibt.

Die Schwierigkeit liegt dabei in der Ungewissheit zum Zeitpunkt der Prämienfestsetzung, ob bzw. wann ein möglicher Schaden eintritt und wie hoch dieser sein wird. Je nach Versicherungssparte und deren Deckungssummen kann die Schwankungsbreite enorm ausfallen. Je größer ein Portfolio ist, desto effizienter gleichen sich große und kleine Risiken aus, man spricht hierbei vom Ausgleich im Kollektiv.

Ziel der Tarifierung ist es nun, die ideale Prämie zu finden, die für den Kunden noch attraktiv erscheint und gleichzeitig den Ertrag des Versicherungsunternehmens maximiert.

Um die Prämien zu optimieren, versucht man als Versicherer anhand von Tarifmerkmalen die ideale Prämie für jeden Kunden zu kalkulieren. Die Tarifmerkmale lassen sich in drei Gruppen unterscheiden:

1. personenbezogene Merkmale: Alter, Familienstand etc.
2. Merkmale des zu versichernden Objektes: Alter, Leistung der Fahrzeuge, etc.
3. geografische Merkmale: Zulassungsbezirk eines Fahrzeuges, Ort des Gebäudes etc.

Um all diese Merkmale in die Prämienkalkulation mit einzubeziehen, führt man eine Tarifkalkulation auf Datenbasis des Bestandes der Versicherungssparte bzw. des Versicherungsproduktes durch. Zusätzlich können natürlich auch externe Daten miteinbezogen werden.

Für die Tarifierung sind folgende Kennzahlen wesentlich:

Schadenbedarf

Der Schadenbedarf gibt für ein Kollektiv den durchschnittlichen Aufwand je Risiko (für ein Jahr) an:

$$\text{Schadenbedarf} = \frac{\sum Aufwand}{\sum Jahreseinheiten}$$

Für die Berechnung des Schadenbedarfes gibt es zwei verschiedene Möglichkeiten. Die oben genannte traditionelle Methode und die Zerlegung in Schadenhäufigkeit und Schadendurchschnitt:

$$\text{Schadenbedarf} = \text{Schadenhäufigkeit} * \text{Schadendurchschnitt}$$

$$\text{Schadenhäufigkeit} = \frac{\sum \text{Anzahl der Schäden}}{\sum Jahreseinheiten}$$

$$\text{Schadendurchschnitt} = \frac{\sum Aufwand}{\sum \text{Anzahl der Schäden}}$$

Beide Methoden haben ihre Vor- und Nachteile, auf die später eingegangen wird. Schadenbedarf bezieht sich im Allgemeinen auf eine Exposureeinheit, welche im obigen Fall durch die Jahreseinheiten dargestellt wird.

Schadenhäufigkeit

Die Schadenhäufigkeit (oft auch als Schadenfrequenz bezeichnet) gibt die Zahl der Schäden bezogen auf das Exposuremaß an. Das kann die Zahl aller Schäden pro Exposureeinheiten oder auch die Zahl der Großschäden bezogen auf die Zahl der Schäden sein. Eine allgemeinere Formel ist:

$$\text{Schadenhäufigkeit} = \frac{\sum \text{Anzahl der Schäden}}{\sum \text{zeitlich abgegrenztes Risikomaß}}$$

Schadendurchschnitt

Die Schadendurchschnitte geben den Mittelwert der Schadenhöhen an. Dies kann für alle Schäden, die kupierten Basisschäden, aber auch für die kupierten Überschäden erfolgen.

Schadensatz

In Versicherungssparten mit der Exposureeinheit Versicherungssumme (VSU) wie z.B. Wohngebäude oder Unfall ist der Schadensatz eine wichtige Kennzahl. Er gibt für ein Kollektiv den durchschnittlichen Schadenbedarf je versicherter Summe und je Jahreseinheiten an.

$$\text{Schadensatz} = \frac{\sum Aufwand}{\sum VSU * Jahreseinheiten}$$

Schadenquote

Eine für die Praxis oft notwendige Zielgröße stellt auch die Schadenquote dar, die aber aufgrund von Rabattkontingenten, der Vermischung von Tarifgenerationen oder des Prämienzyklus eine nicht immer objektive Kennzahl darstellt.

$$\text{Schadenquote} = \frac{\sum Schadenaufwand}{\sum \text{zeitlich abgegrenzter Bestandsbeitrag}}$$

3.1 Datengrundlage

Bevor man sich mit den Tarifmodellen beschäftigen kann, ist es wichtig, sich über die Datenbasis und deren Qualität Gedanken zu machen. Auf die Analyse der Datenqualität wird hier nicht weiter eingegangen, da der Fokus dieser Arbeit auf der Modellierung liegt. Folgende Modellannahmen sind jedoch für die spätere Modellierung wesentlich.

Die Vertragsunabhängigkeit ist für die Tarifierung wesentlich. In der Praxis ist das jedoch nicht immer gegeben. Als Beispiel dafür betrachte man ein Hagelschaden Kumul Ereignis, bei dem mehrere Verträge eines Portfolios demselben Schaden zugrunde liegen. Solche Ereignisse sollte man aus der Datenbasis für die Tarifgestaltung ausschließen bzw. mit geeigneten Modellen für Naturkatastrophen bewerten.

Außerdem sollten alle Risiken bzw. deren Schadenfrequenz und Schadenhöhe zeitunabhängig sowie homogen sein.

3.2 Multiplikatives Modell

Für Multiplikative Modelle gilt[1]:

In der Praxis weisen viele Verträge keinen einzigen Schaden auf, was dazu führt, dass man nach Methoden sucht, die eine erwartete reine Prämie ausgeben, welche glatter über den Einzelverträgen variiert, relativ stabil über die Zeit ist und nicht anfällig auf zufällige Fluktuationen reagiert. Dies erfüllt das Multiplikative Modell, welches sowohl für die reine Prämie als auch für die Schadenfrequenz und Schadenhöhe angewandt werden kann.

Sind N Tariffaktoren mit einer Anzahl von n_i Ausprägungen für Tarifmerkmal i gegeben (der Einfachheit halber beginnen wir mit 2 Tariffaktoren, also $N = 2$), dann haben

[1] vgl. [4] Seite 25

wir in der Tarifzelle (i, j), die einem Einzelvertrag entspricht, als Exposure ω_{ij} und als Response X_{ij}, sodass für den Zielwert $Y_{ij} = \frac{X_{ij}}{\omega_{ij}}$ gilt. Mit dem Exposure $\omega_{ij} = 1$ gilt für den Erwartungswert $E(Y_{ij}) = \mu_{ij}$. Somit lautet das Multiplikative Modell für den Fall mit zwei Tarifmerkmalen

$$\mu_{ij} = \gamma_0 \gamma_{1i} \gamma_{2j}.$$

Dabei handelt es sich bei γ_{1i} für $i = 1, \ldots, n_1$ um Parameter die den verschiedenen Ausprägungen des ersten Tarifmerkmals entsprechen und bei γ_{2j} für $j = 1, ..., n_2$ um die für das zweite Merkmal. Der sogenannte Basiswert ist γ_0. Man wählt nun eine Referenzzelle, vorzugsweise mit großem Exposure. Setzt man in dem Beispiel mit zwei Merkmalen (1, 1) als Referenzzelle, so gilt $\gamma_{11} = \gamma_{21} = 1$. Nun kann γ_0 als Basiswert für die Polizze in der Referenzzelle interpretiert werden. Die übrigen Parameter messen den relativen Unterschied in Bezug auf die Referenzzelle und werden Relativitäten genannt. Die Multiplikativitätsannahme bedeutet, das zwischen den Tariffaktoren keine Interaktion existiert. Eine Erweiterung der Formel für den Fall von N Tariffaktoren sieht wie folgt aus:

$$\mu_{i1}, \cdots, \mu_{iN} = \gamma_0 \gamma_{1i_1} \cdots \gamma_{Ni_N}.$$

Man passt also den Basiswert an und die übrigen Parameter kontrollieren wieviel als Prämie berechnet wird. Im folgenden Kapitel wird mit einem gegebenen Datenbestand und der Hilfe eines Multiplikativen Modelles eine Tarifstruktur erstellt.

Kapitel 4

Tarifierung

Im Kapitel der Tarifierung beschäftigen wir uns mit der Erstellung einer Tarifstruktur. Dabei werden wir mit Hilfe der Verallgemeinerten Linearen Modelle ein Schadenfrequenz- sowie ein Schadendurchschnittsmodell modellieren, um auf die gewünschte Zielgröße, den Schadenbedarf, zu kommen.

Die Modellierung der Daten wird mit Hilfe des von Willis Towers Watson entwickelten Pricing Tools Emblem durchgeführt. Dabei gliedert sich der Prozess in drei Hauptaufgaben:

1. Daten einlesen und strukturieren: Emblem File Converter
2. Modellierung der Modelle: Emblem Modeller
3. bei indirekter Modellierung des Schadenbedarfes werden ein Schadenfrequenz und ein Schadenhöhenmodell verknüpft: Modell Combiner.

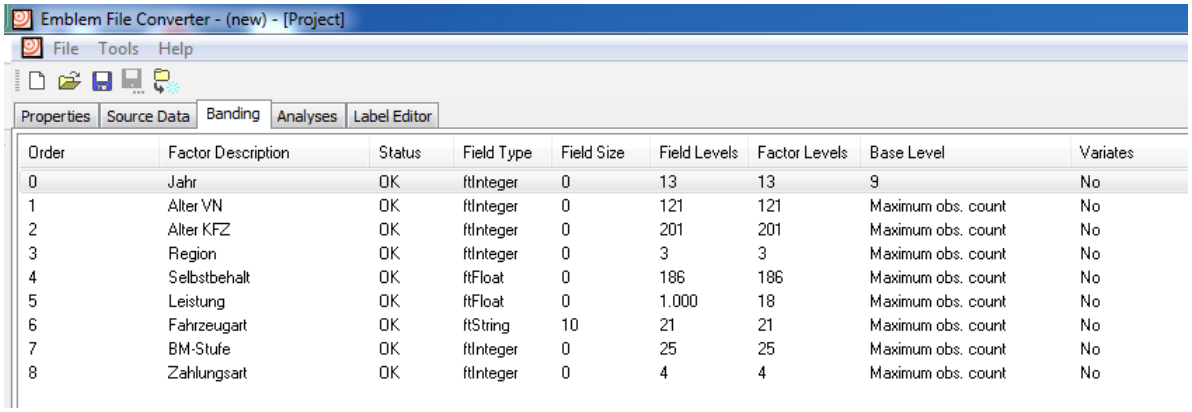
Als Resultat erhält man eine kalkulierte Basisnettoprämie sowie Auf- und Abschlagsfaktoren für die im Modell miteinbezogenen Merkmale, den Tariffaktoren.

4.1 Daten

Die folgende verwendete Datenbasis wurde von der HDI Versicherung AG zur Verfügung gestellt und enthält einen Datenauszug der KFZ-Haftpflicht Sparte. Bevor wir uns mit der Modellierung beschäftigen, betrachten wir die Datenbasis genauer.

4.1.1 Datenstruktur

Zu Beginn wird die Datenbasis mit dem oben erwähnten Emblem File Converter eingelesen. Anschließend hat man noch die Möglichkeit, Änderungen an der Datenstruktur vorzunehmen und wählt alle Merkmale, die man in der Modellierung betrachten möchte. In folgender Tabelle werden alle in der Modellierung verwendeten Merkmale aufgelistet.



Order	Factor Description	Status	Field Type	Field Size	Field Levels	Factor Levels	Base Level	Variates
0	Jahr	OK	ftInteger	0	13	13	9	No
1	Alter VN	OK	ftInteger	0	121	121	Maximum obs. count	No
2	Alter KFZ	OK	ftInteger	0	201	201	Maximum obs. count	No
3	Region	OK	ftInteger	0	3	3	Maximum obs. count	No
4	Selbstbehalt	OK	ftFloat	0	186	186	Maximum obs. count	No
5	Leistung	OK	ftFloat	0	1.000	18	Maximum obs. count	No
6	Fahrzeugart	OK	ftString	10	21	21	Maximum obs. count	No
7	BM-Stufe	OK	ftInteger	0	25	25	Maximum obs. count	No
8	Zahlungsart	OK	ftInteger	0	4	4	Maximum obs. count	No

Abbildung 4.1: Datenbasis KFZ-Haftpflicht

Die Spaltenbezeichnungen haben folgende Bedeutungen:

Order: Laufnummer des Merkmals

Factor Description: Name des Merkmals

Status: erfolgreich importierte Daten

Field Type: Datentyp

Field Size: wie viele Zeichen des Strings übernommen werden

Field Levels: Anzahl der Verschiedenen Ausprägungen des Merkmals

Factor Levels: Anzahl der Unterteilungen/Gruppierungen der Ausprägungen (kann somit maximal der Anzahl des Field Levels entsprechen)

Base Level: gibt an welches Factor Level als Referenzwert gewählt wurde

Merkmale die in das Tool übernommen werden

Jahr: Geschäftsjahr

Alter VN: Alter des Versicherungsnehmers zum Jahresende

Alter KFZ: Alter des Fahrzeuges zum Jahresende

Region: Die Zulassungsbezirke wurden grob in drei Regionen unterteilt. Region 0 bezieht sich auf urbane Regionen wie Wien. Region 1 beinhaltet suburbane Regionen. Region 2 beschränkt sich auf den ländlichen Bereich.

Selbstbehalt: Höhe des vertraglichen Selbstbehaltes

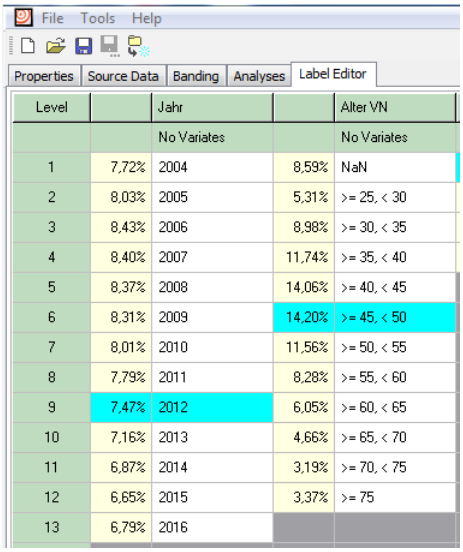
Leistung: kW des Fahrzeuges

BM-Stufe: Bonus Malus Stufe des Versicherungsnehmers

Zahlungsart: 1/2/4/12 für jährliche, halbjährliche, vierteljährliche oder monatliche Zahlungsweise.

Wie man sehen kann, ist die Zahl der Factor Levels bei vielen Merkmalen sehr hoch. In der Praxis werden die Field Levels in Gruppen zusammengefasst (soweit möglich) und in weiterer Folge als solche modelliert. Zum Beispiel teilt man das Alter in Altersgruppen. Je nach Größe der Datenbasis sind Altersgruppen in 2, 5 oder 10er Schritten sinnvoll.

Nach Anpassung der Factor Levels einiger Merkmale, ergibt sich beispielsweise folgende Einteilung:



Level		Jahr		Alter VN
		No Variates		No Variates
1	7,72%	2004	8,59%	NaN
2	8,03%	2005	5,31%	>= 25, < 30
3	8,43%	2006	8,98%	>= 30, < 35
4	8,40%	2007	11,74%	>= 35, < 40
5	8,37%	2008	14,06%	>= 40, < 45
6	8,31%	2009	14,20%	>= 45, < 50
7	8,01%	2010	11,56%	>= 50, < 55
8	7,79%	2011	8,28%	>= 55, < 60
9	7,47%	2012	6,05%	>= 60, < 65
10	7,16%	2013	4,66%	>= 65, < 70
11	6,87%	2014	3,19%	>= 70, < 75
12	6,65%	2015	3,37%	>= 75
13	6,79%	2016		

Abbildung 4.2: Daten nach Anpassung der Factor Level

In der Übersicht des Label Editors erhält man gleich zu Beginn einen guten Überblick über die Datenverteilung. Eines ist nämlich klar, je mehr Informationen pro Merkmal bzw. Merkmal Level vorhanden sind, desto effizienter wird die Modellierung. Beispielsweise wird noch vor der Modellierung das Merkmal der Fahrzeugart angepasst, da aufgrund der geringen Verteilung von Informationen der Fahrzeugarten, ausgenommen der PKWs, diese statistisch irrelevant wären.

Im letzten Schritt werden die benötigten Dateien für die Modellierung der Modelle erstellt.

Anmerkung: Um Effekte von Großschäden zu minimieren wurde der Einzelschadenaufwand bei einer Höhe von 100.000 EUR kupiert.

4.1.2 Emblem

Die erstellten Dateien können nun mit dem Emblem Modeller geladen werden. Der Emblem Modeller stellt das zentrale Tool der Modellierung dar. Es basiert auf Verallgemeinerten Linearen Modellen und bietet somit eine ideale Grundlage der Tarifgestaltung.

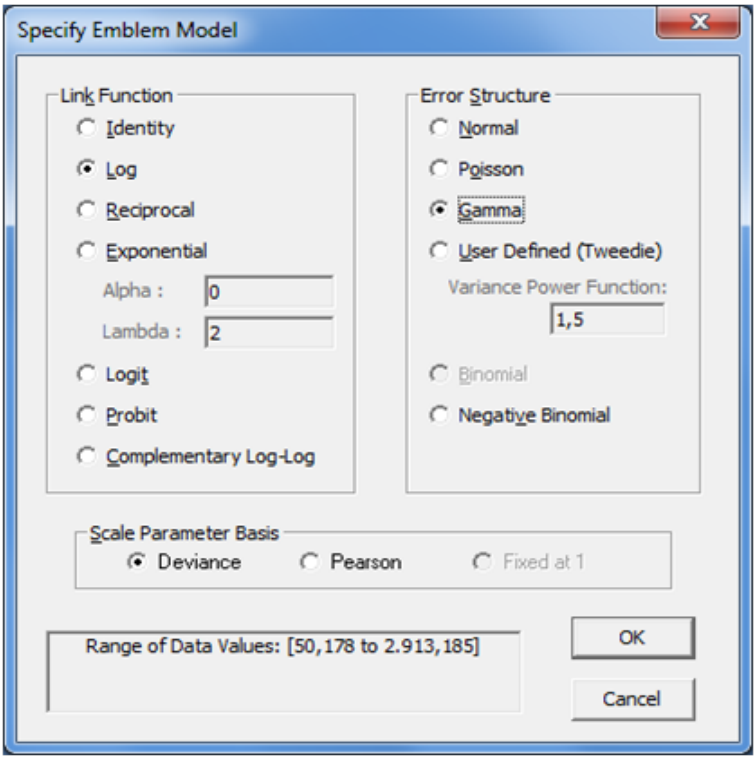


Abbildung 4.3: Verteilungsannahme und Linkfunktion

Wie man in Abbildung 4.3 sehen kann, hat man die Möglichkeit, je nach Modell, die Linkfunktion und Verteilungsannahme selbst zu bestimmen.

Allgemeine Übersicht des Cockpits von Emblem:

Als kurze Einführung in das Tool Emblem wird nun der Hauptbildschirm der Software dargestellt und die einzelnen Elemente kurz erläutert.

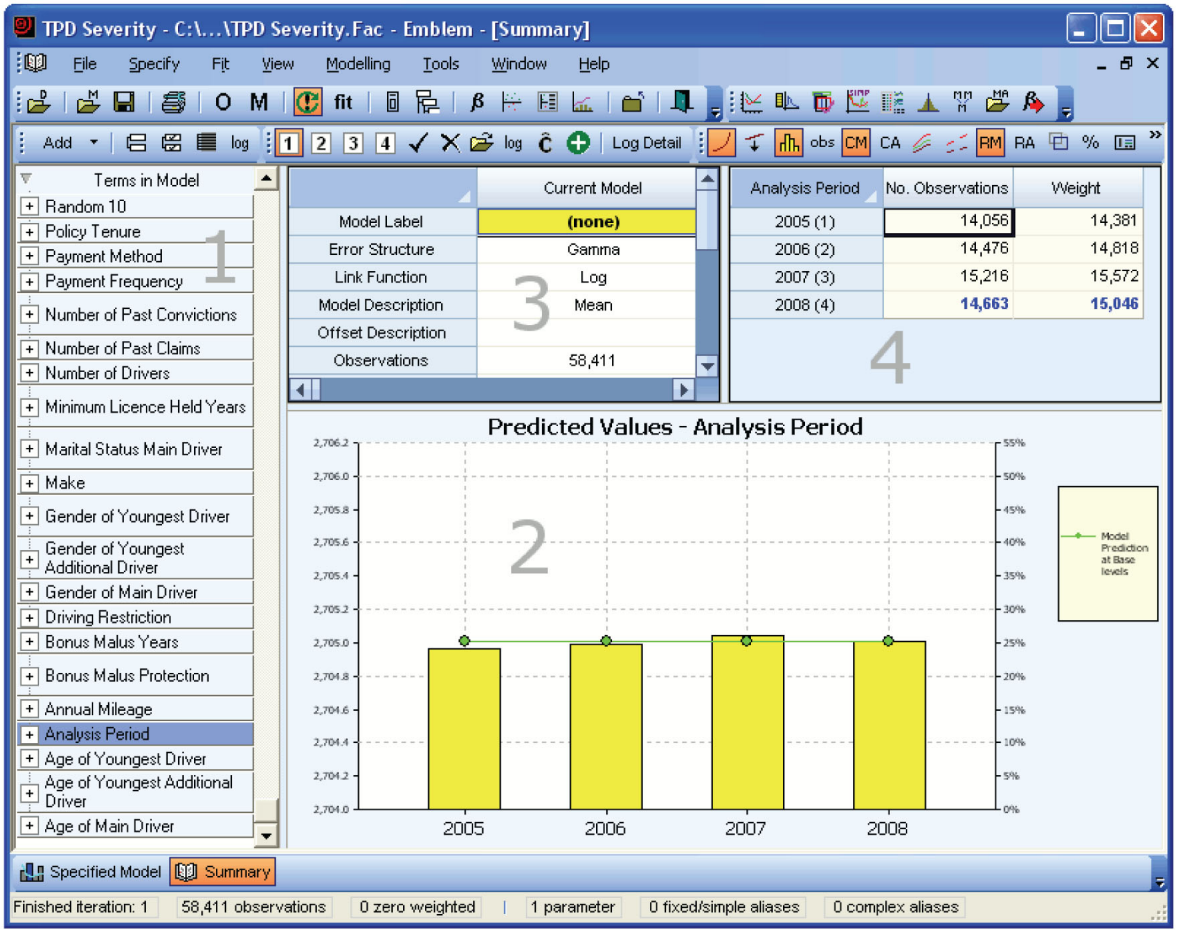


Abbildung 4.4: Emblem Hauptbildschirm (Testdaten)

Der Bereich mit der Nummer 1 wird Factor Tree genannt und beinhaltet alle im Vorfeld ausgewählten Faktoren. Nummer 2 zeigt den sogenannten Summery Graph und gibt die graphische Darstellung des gesamten Modells oder eines ausgewählten Faktors wieder. Auf der X-Achse werden die Merkmale aufgelistet, wohingegen auf der linken Y-Achse die Durchschnittsschadenhöhe und auf der rechten Seite der prozentuelle Anteil der Beobachtungen pro Merkmal dargestellt werden (für Balkendiagramm). Feld 3 zeigt eine kurze Zusammenfassung des gesamten Modells. Man erkennt, dass ein Gamma Modell mit einer Log Linkfunktion ausgewählt wurde und das Gesamtmodell insgesamt 58.411 Beobachtungen beinhaltet. Außerdem erkennt man im Bereich der Nummer 4 die Details des ausgewählten Merkmales. Hier zum Beispiel die gewählte Analysis Period, die im Schadenjahr 2005 eine Anzahl von 14.056 Beobachtungen mit einem Gewicht von 14.381 besitzt.

4.2 Schadenfrequenz Modell

Der Schadenbedarf soll durch die Kombination eines Schadenfrequenz- und eines Schadendurchschnittsmodells modelliert werden. Zunächst beginnt man mit der Analyse des Schadenfrequenz Modells. Dafür wählt man als Verteilungsannahme, wie schon zu Beginn der Arbeit erwähnt, die Poissonverteilung und die logarithmische Linkfunktion.

Betrachtet man zunächst die Ausgangsdaten im Modell.

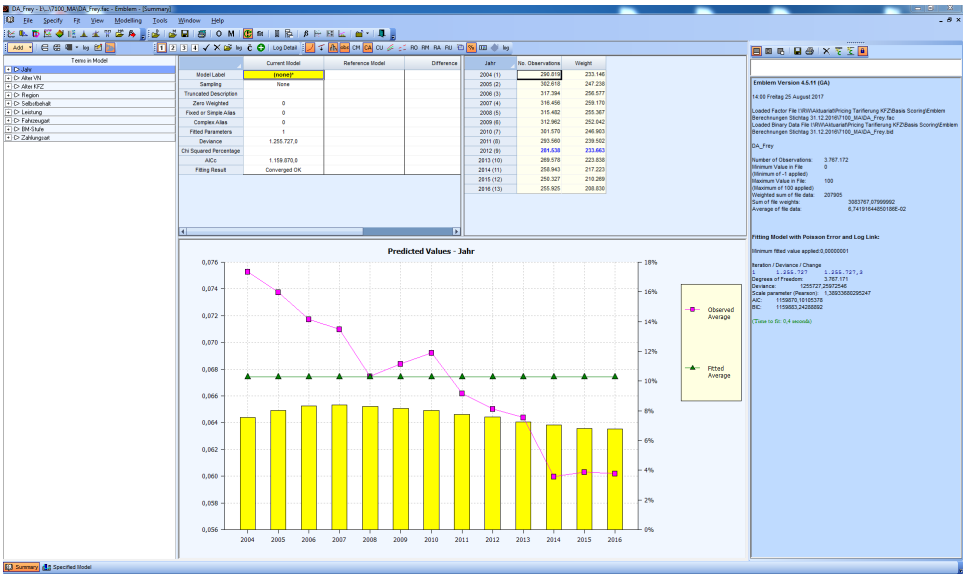


Abbildung 4.5: Cockpit Schadenfrequenz Modell

In Abbildung 4.5 wird das Emblem Cockpit dargestellt. Im Factor Tree, siehe Abbildung 4.6, sind die oben besprochenen Merkmale aufgelistet.

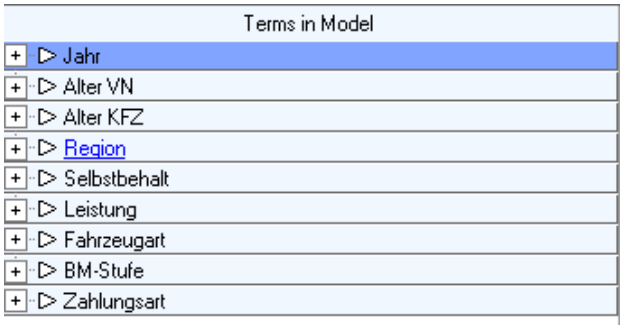


Abbildung 4.6: Factor Tree

Wie man dem Summary Graph, der zurzeit den Schadenfrequenzverlauf der Geschäftsjahre anzeigt, entnehmen kann, entwickelt sich das Problem der Schadenhäufigkeit in eine positive Richtung. Die pinke Linie zeigt hier den beobachteten Durchschnittswert. Es sei noch angemerkt, dass man der grünen Linie (fitted average) entnehmen kann, dass sich zurzeit noch keine Merkmale im Modell befinden.

Jahr	No. Observations	Weight
2004 (1)	290.819	233.146
2005 (2)	302.618	247.238
2006 (3)	317.394	256.577
2007 (4)	316.456	259.170
2008 (5)	315.482	255.367
2009 (6)	312.962	252.042
2010 (7)	301.570	246.903
2011 (8)	293.560	239.502
2012 (9)	281.538	233.663
2013 (10)	269.578	223.838
2014 (11)	258.943	217.223
2015 (12)	250.327	210.269
2016 (13)	255.925	208.830

Abbildung 4.7: Merkmalsübersicht - Jahr

Der Abbildung 4.7 kann man die Anzahl der Beobachtungen und deren Gewicht entnehmen. Diese Ansicht kann für jedes Merkmal aufgerufen werden.

Es gibt eine Vielzahl von Möglichkeiten um mit der Modellierung zu beginnen. Für diese Arbeit wurde im ersten Schritt ein Full Model erstellt, welches speziell für die Betrachtung der Devianzveränderung als Referenzmodell betrachtet wird. Dafür muss jedes Merkmal in das Modell aufgenommen werden. Die Devianz zeigt uns somit den idealen Wert.

	Current Model	Reference Model	Difference
Model Label	(none)*		
Sampling	None		
Truncated Description	Jahr + Alter VN + ...		
Zero Weighted	0		
Fixed or Simple Alias	0		
Complex Alias	0		
Fitted Parameters	81		
Deviance	1.179.879,0		
Chi Squared Percentage			
AICc	1.192.538,0		
Fitting Result	Converged OK		

Abbildung 4.8: Full Model

Die Aufgabe besteht nun darin, dass Modell um die Merkmale zu verringern, die keinen signifikanten Einfluss auf jenes haben. Merkmale, die am Ende als wesentlich klassifiziert werden, bilden dann die Grundlage der Tarifstruktur. Es wird nun näher auf die Entscheidungskriterien für die Aufnahme eines Merkmals in das Modell eingegangen.

4.2.1 Analyse der Devianz

Das Full Model wurde bereits erstellt und als Referenz Modell für die weiteren Anpassungen hinterlegt. Zunächst wird durch Reduktion der Parameter deren Einfluss auf die Devianz analysiert, um einen ersten Eindruck über die relevanten Merkmale zu erhalten. Der Ausgangswert der Devianz, siehe Tabelle 4.8, liegt bei 1.179.879.

Als Beispiel eines relevanten Merkmales betrachten wir das Merkmal der Leistung:

	Current Model	Reference Model
Model Label	(none)*	(none)*
Sampling	None	None
Truncated Description	Alter KFZ + Region	Jahr + Alter VN + ...
Zero Weighted	0	0
Fixed or Simple Alias	0	0
Complex Alias	0	0
Fitted Parameters	69	81
Deviance	1.203.344,0	1.179.879,0
Chi Squared Percentage		
AICc	1.183.091,0	1.192.569,0
Fitting Result	Converged OK	Converged OK

Abbildung 4.9: Leistung aus Modell ausgeschlossen

Die Devianz weist einen deutlich höheren Wert im Vergleich zum Full Model auf, was darauf schließen lässt, dass das Merkmal einen großen Einfluss auf die Schadenfrequenz hat. Im Gegensatz dazu beträgt die Devianz nach Herausnahme des Merkmals der Zahlungsart 1.180.164, somit ist es nicht signifikant und kann nach diesem Kriterium aus dem Modell ausgeschlossen werden.

4.2.2 Zeitliche Konsistenz

Neben den statistischen Kennzahlen, gibt es jedoch auch noch andere wesentliche Faktoren die für die Entscheidungsfindung eine wesentliche Rolle spielen. Oft erscheinen Merkmale statistisch relevant, jedoch zeigt sich nach Analyse der zeitlichen Konsistenz, dass der Verlauf sehr zufallsorientiert ist.

Betrachtet man hierfür das Merkmal der Regionen und bildet eine sogenannte Interaction mit dem Merkmal Jahr. Somit bekommt man eine grafische Darstellung, wie sich die Schadenfrequenz der einzelnen Regionen in den jeweiligen Geschäftsjahren verhalten hat.

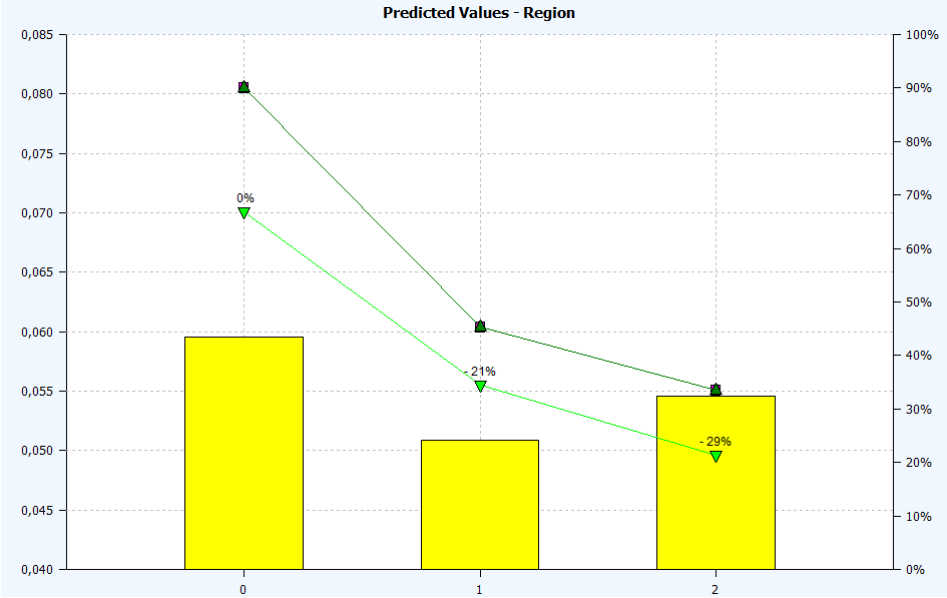


Abbildung 4.10: Darstellung der Regionen

Der Summery Graph zeigt einen deutlichen Trend. Die Schadenfrequenz von Region 0 ist deutlich höher als im Vergleich zu Region 1. Auch zwischen Region 1 und Region 2 zeigt sich noch eine Verbesserung, die aber geringer ausfällt.

Der oben angesprochen Verlauf pro Geschäftsjahr ergibt folgendes Bild:

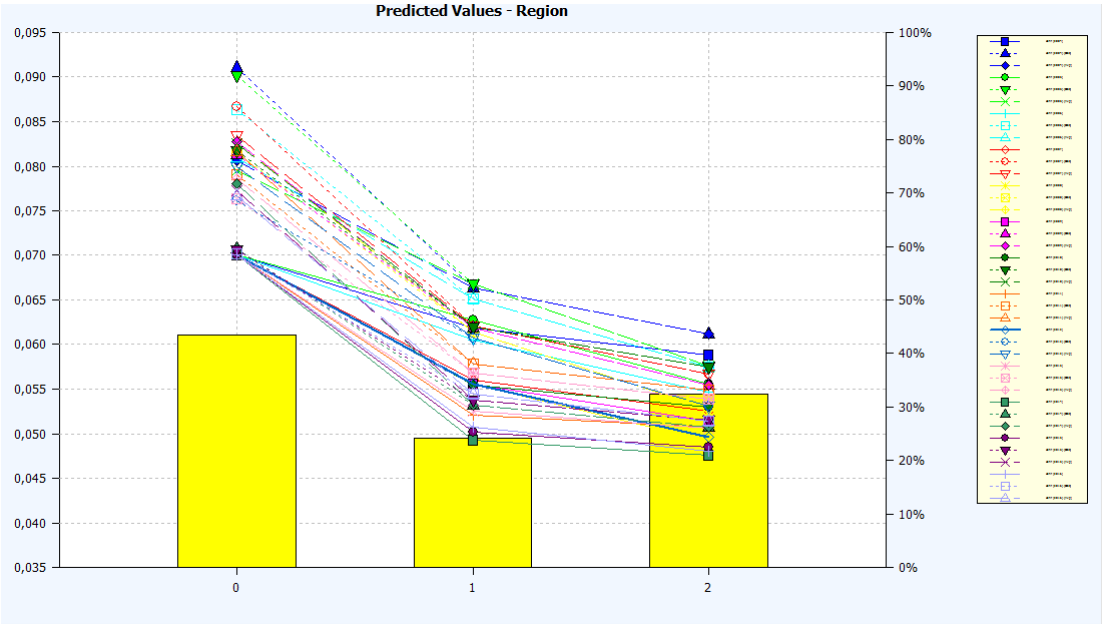


Abbildung 4.11: Interaction Region - Jahr

Der im Modell beobachtete und geschätzte Trend zeigt eine starke zeitliche Konsistenz. Nahezu alle Geschäftsjahre zeigen den beobachteten Verlauf. Aufgrund dessen sollte das Merkmal im Modell berücksichtigt werden.

4.2.3 Glättung von Strukturbrüchen

Um die Parameteranpassung des Modelles zu verbessern, können kleine Strukturbrüche im Verlauf durch das Legen einer Anpassungskurve geglättet werden. Die Anpassung der Merkmale kann auch zusätzlich verbessert werden, indem man Merkmalsausprägungen, die als Ausreißer identifiziert werden oder aufgrund der sehr geringen Bestandsgröße zu irrelevanten Ergebnissen führen, zu einer Gruppe zusammenfasst.

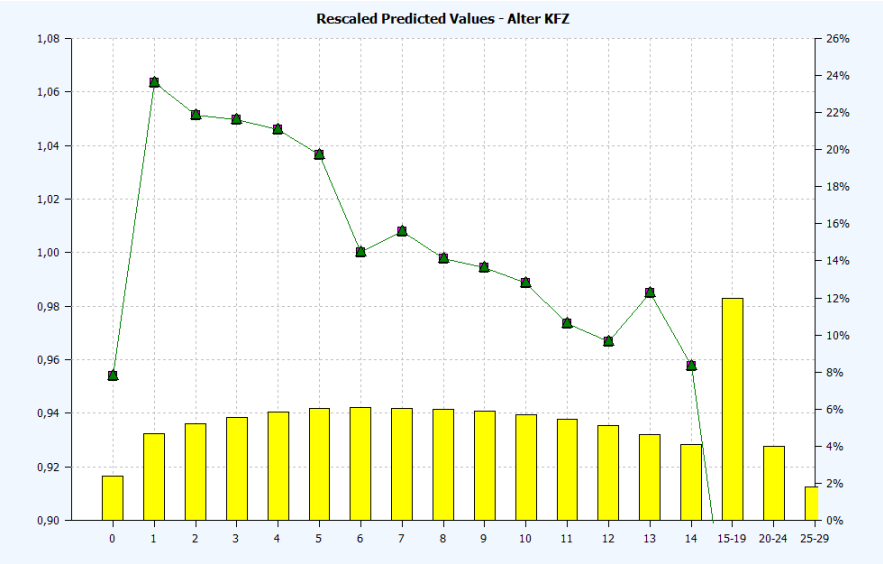


Abbildung 4.12: Alter des KFZ

Die Abbildung 4.12 zeigt im Bereich der Merkmalsausprägung des Alter KFZ 6 einen ungewöhnlichen Peak nach unten. Um diesen Effekt etwas abzuschwächen versucht man eine Approximation durch das Legen einer Kurve durchzuführen.

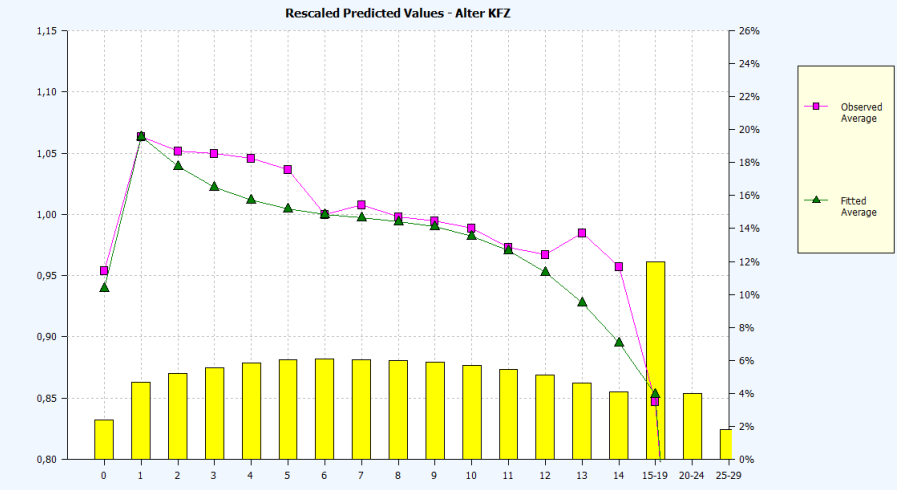


Abbildung 4.13: Alter KFZ

4.2.4 Statistik

In Emblem gibt es auch die Möglichkeit auf verschiedenste statistische Kennzahlen zurückzugreifen, die bei der Interpretation und der Entscheidungsfindung, ob gewisse Parameter signifikant sind oder nicht, helfen.

Zu den oben genannten statistischen Kennzahlen gehören unter anderem die Varianz/-Kovarianzmatrix, Ch-Square, Phi-Koeffizient, Cramers V und Standardfehler.

Variance / Covariance	1-Mean	2-Alter KFZ (0)	3-Alter KFZ (1)	4-Alter KFZ (2)	5-Alter KFZ (3)	6-Alter KFZ (4)	7-Alter KFZ (5)	8-Alter KFZ (7)	9-Alter KFZ (8)	10-Alter KFZ (9)	11-Alter KFZ (10)	12
1-Mean	**0,00017**	-0,00009	-0,00009	-0,00009	-0,00009	-0,00009	-0,00009	-0,00010	-0,00010	-0,00010	-0,00010	
2-Alter KFZ (0)	-0,00009	**0,00035**	0,00010	0,00010	0,00010	0,00010	0,00010	0,00009	0,00009	0,00009	0,00009	
3-Alter KFZ (1)	-0,00009	0,00010	**0,00021**	0,00010	0,00010	0,00010	0,00010	0,00009	0,00009	0,00009	0,00009	
4-Alter KFZ (2)	-0,00009	0,00010	0,00010	**0,00020**	0,00010	0,00010	0,00010	0,00009	0,00009	0,00009	0,00009	
5-Alter KFZ (3)	-0,00009	0,00010	0,00010	0,00010	**0,00019**	0,00010	0,00009	0,00009	0,00009	0,00009	0,00009	
6-Alter KFZ (4)	-0,00009	0,00010	0,00010	0,00010	0,00010	**0,00019**	0,00009	0,00009	0,00009	0,00009	0,00009	
7-Alter KFZ (5)	-0,00009	0,00010	0,00010	0,00010	0,00009	0,00009	**0,00019**	0,00009	0,00009	0,00009	0,00009	
8-Alter KFZ (7)	-0,00010	0,00009	0,00009	0,00009	0,00009	0,00009	0,00009	**0,00019**	0,00009	0,00009	0,00009	
9-Alter KFZ (8)	-0,00010	0,00009	0,00009	0,00009	0,00009	0,00009	0,00009	0,00009	**0,00019**	0,00010	0,00010	
10-Alter KFZ (9)	-0,00010	0,00009	0,00009	0,00009	0,00009	0,00009	0,00009	0,00009	0,00010	**0,00019**	0,00010	
11-Alter KFZ (10)	-0,00010	0,00009	0,00009	0,00009	0,00009	0,00009	0,00009	0,00009	0,00010	0,00010	**0,00020**	
12-Alter KFZ (11)	-0,00009	0,00009	0,00009	0,00009	0,00009	0,00009	0,00009	0,00010	0,00010	0,00010	0,00010	
13-Alter KFZ (12)	-0,00009	0,00009	0,00009	0,00009	0,00009	0,00009	0,00009	0,00010	0,00010	0,00010	0,00010	
14-Alter KFZ (13)	-0,00009	0,00009	0,00009	0,00009	0,00009	0,00009	0,00009	0,00010	0,00010	0,00010	0,00010	
15-Alter KFZ (14)	-0,00009	0,00009	0,00009	0,00009	0,00009	0,00009	0,00009	0,00010	0,00010	0,00010	0,00010	
16-Alter KFZ (15-19)	-0,00009	0,00009	0,00009	0,00009	0,00009	0,00009	0,00009	0,00010	0,00010	0,00010	0,00010	
17-Alter KFZ (20-24)	-0,00010	0,00009	0,00009	0,00009	0,00009	0,00009	0,00009	0,00010	0,00010	0,00010	0,00010	
18-Alter KFZ (25-29)	-0,00010	0,00010	0,00009	0,00009	0,00009	0,00009	0,00009	0,00010	0,00010	0,00010	0,00010	
19-Alter KFZ (30+)	-0,00009	0,00010	0,00009	0,00009	0,00009	0,00009	0,00009	0,00010	0,00010	0,00010	0,00010	
20-Region (1)	-0,00001	-0,00000	-0,00000	-0,00000	-0,00000	-0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	
21-Region (2)	-0,00001	-0,00000	-0,00000	-0,00000	-0,00000	-0,00000	0,00000	-0,00000	0,00000	0,00000	0,00000	
22-Leistung (= 0)	-0,00003	-0,00001	-0,00000	-0,00000	0,00000	0,00000	0,00000	-0,00000	-0,00000	-0,00000	-0,00000	
23-Leistung (> 0, <= 20)	-0,00003	-0,00001	-0,00000	-0,00000	-0,00000	-0,00000	-0,00000	0,00000	0,00000	0,00000	0,00000	
24-Leistung (> 20, <= 40)	-0,00002	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	-0,00000	-0,00000	-0,00000	-0,00000	
25-Leistung (> 40, <= 60)	-0,00002	-0,00000	-0,00000	-0,00000	-0,00000	-0,00000	-0,00000	0,00000	0,00000	-0,00000	-0,00000	
26-Leistung (> 80, <= 100)	-0,00002	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	
27-Leistung (> 100)	-0,00002	-0,00000	-0,00000	-0,00000	-0,00000	-0,00000	-0,00000	0,00000	0,00000	0,00000	0,00000	
28-AV CF (Group 1)	-0,00005	-0,00000	-0,00000	-0,00000	-0,00000	-0,00000	-0,00000	0,00000	0,00000	0,00000	0,00000	
29-AV CF (>= 30, < 35)	-0,00005	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	-0,00000	0,00000	0,00000	0,00000	
30-AV CF (>= 35, < 40)	-0,00005	0,00000	0,00000	0,00000	0,00000	0,00000	0,00000	-0,00000	-0,00000	0,00000	0,00000	
31-AV CF (>= 40, < 45)	-0,00005	-0,00000	0,00000	0,00000	0,00000	-0,00000	-0,00000	-0,00000	-0,00000	-0,00000	-0,00000	
32-AV CF (>= 50, < 55)	-0,00005	-0,00000	-0,00000	-0,00000	-0,00000	-0,00000	-0,00000	0,00000	-0,00000	-0,00000	-0,00000	
33-AV CF (>= 55, < 60)	-0,00005	-0,00000	-0,00000	-0,00000	-0,00000	-0,00000	-0,00000	0,00000	0,00000	0,00000	0,00000	
34-AV CF (>= 60, < 65)	-0,00005	-0,00000	-0,00000	-0,00000	-0,00000	-0,00000	-0,00000	0,00000	0,00000	0,00000	0,00000	
35-AV CF (>= 75)	-0,00005	-0,00000	-0,00000	-0,00000	-0,00000	-0,00000	-0,00000	0,00000	0,00000	0,00000	0,00000	

Abbildung 4.14: Varianz/Kovarianz Matrix

Chi-Square | Continuity Adjusted Chi-Square | Likelihood-Ratio Chi-Square | Phi Coefficient | Contingency Coefficient | Cramer's V

Factor (#Levels)	Jahr (13)	Alter VN (12)	Alter KFZ (19)	Region (3)	Selbstbehalt (4)	Leistung (7)	Fahrzeugart (21)	BM-Stufe (25)	Zahlungsart (4)
Jahr (13)	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
Alter VN (12)	0,063	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
Alter KFZ (19)	0,045	0,031	0,000	0,000	0,000	0,000	0,000	0,000	0,000
Region (3)	0,012	0,031	0,011	0,000	0,000	0,000	0,000	0,000	0,000
Selbstbehalt (4)	0,208	0,068	0,320	0,028	0,000	0,000	0,000	0,000	0,000
Leistung (7)	0,047	0,048	0,173	0,032	0,185	0,000	0,000	0,000	0,000
Fahrzeugart (21)	0,037	0,066	0,104	0,035	0,223	0,486	0,000	0,000	0,000
BM-Stufe (25)	0,148	0,080	0,052	0,053	0,193	0,167	0,136	0,000	0,000
Zahlungsart (4)	0,061	0,088	0,094	0,037	0,107	0,173	0,244	0,212	0,000

Abbildung 4.15: Cramers V

Das fertige Schadenfrequenzmodell enthält nun folgende Paramter:

Parameter Number	Name	Value	Standard Error	Normal Probability (%)	Alias Indicator (%)	Weight	Weight (%)	Exp(Value)
1	Mean	-2,6417	0,01303	0,0		3.083.767	100,0	0,0712
2	Alter KFZ (0)	0,0316	0,01871	95,4		73.040	2,4	1,0321
3	Alter KFZ (1)	0,0609	0,01454	100,0		144.108	4,7	1,0628
4	Alter KFZ (2)	0,0536	0,01416	100,0		160.716	5,2	1,0550
5	Alter KFZ (3)	0,0529	0,01393	100,0		171.276	5,6	1,0543
6	Alter KFZ (4)	0,0487	0,01377	100,0		180.033	5,8	1,0499
7	Alter KFZ (5)	0,0378	0,01368	99,7		185.908	6,0	1,0385
-	Alter KFZ (6)					187.323	6,1	
8	Alter KFZ (7)	0,0101	0,01376	76,9		186.723	6,1	1,0102
9	Alter KFZ (8)	-0,0021	0,01384	44,1		184.897	6,0	0,9979
10	Alter KFZ (9)	-0,0103	0,01392	22,9		181.457	5,9	0,9897
11	Alter KFZ (10)	-0,0223	0,01407	5,6		175.540	5,7	0,9779
12	Alter KFZ (11)	-0,0441	0,01431	0,1		167.678	5,4	0,9568
13	Alter KFZ (12)	-0,0574	0,01459	0,0		157.458	5,1	0,9442
14	Alter KFZ (13)	-0,0432	0,01491	0,2		143.374	4,6	0,9577
15	Alter KFZ (14)	-0,0688	0,01563	0,0		125.874	4,1	0,9336
16	Alter KFZ (15-19)	-0,1489	0,01242	0,0		369.262	12,0	0,8617
17	Alter KFZ (20-24)	-0,5504	0,02010	0,0		123.495	4,0	0,5767
18	Alter KFZ (25-29)	-1,0068	0,03811	0,0		55.726	1,8	0,3654
19	Alter KFZ (30+)	-1,3796	0,03904	0,0		109.880	3,6	0,2517
-	Region (0)					1.340.760	43,5	
20	Region (1)	-0,2352	0,00642	0,0		743.245	24,1	0,7904
21	Region (2)	-0,3038	0,00601	0,0		999.762	32,4	0,7380
22	Leistung (= 0)	-2,6313	0,05828	0,0		101.833	3,3	0,0720
23	Leistung (> 0, <= 20)	-1,3487	0,02304	0,0		210.336	6,8	0,2596
24	Leistung (> 20, <= 40)	-0,4619	0,01115	0,0		320.484	10,4	0,6301
25	Leistung (> 40, <= 60)	-0,0982	0,00666	0,0		778.389	25,2	0,9064
-	Leistung (> 60, <= 80)					853.647	27,7	
26	Leistung (> 80, <= 100)	0,1119	0,00705	100,0		511.008	16,6	1,1184
27	Leistung (> 100)	0,1663	0,00819	100,0		308.071	10,0	1,1809
28	AV CF (Group 1)	0,1422	0,00847	100,0		660.755	21,4	1,1528
29	AV CF (>= 30, < 35)	-0,0744	0,01101	0,0		265.066	8,6	0,9283
30	AV CF (>= 35, < 40)	-0,0494	0,01024	0,0		355.891	11,5	0,9518
31	AV CF (>= 40, < 45)	0,0026	0,00968	60,8		431.978	14,0	1,0026
-	AV CF (>= 45, < 50)					442.331	14,3	
32	AV CF (>= 50, < 55)	-0,0141	0,01025	8,4		364.401	11,8	0,9860
33	AV CF (>= 55, < 60)	-0,0194	0,01137	4,4		263.896	8,6	0,9808
34	AV CF (>= 60, < 65)	-0,0053	0,01259	33,7		193.762	6,3	0,9947
35	AV CF (>= 75)	0,4219	0,01360	100,0		105.688	3,4	1,5249

Abbildung 4.16: Parameterübersicht

4.3 Durchschnittsschaden Modell

Widmen wir uns nun dem zweiten relevanten Modell, dem Durchschnittsschaden Modell. Dieses Modell enthält alle Informationen der Gesamtschadenhöhe des jeweiligen Schadenfalles. Als Verteilungsannahme wählen wir die Gammaverteilung. Verglichen mit dem Schadenfrequenz Modelles fällt auf, dass wir nur einen Bruchteil an Beobachtungen im Vergleich zum Frequenz Modell haben. Genau dies macht ein Versicherungsgeschäft überhaupt möglich. Würde jeder Versicherungsnehmer einen Schaden verursachen, würde das Geschäft nicht mehr tragfähig sein und man müsste als Versicherungsunternehmen Prämien einfordern, die niemand bereit wäre zu bezahlen.

Die folgende Grafiken zeigen die Anzahl der Beobachtungen der beiden Modelle:

Jahr	No. Observations	Weight
2004 (1)	290.819	233.146
2005 (2)	302.618	247.238
2006 (3)	317.394	256.577
2007 (4)	316.456	259.170
2008 (5)	315.482	255.367
2009 (6)	312.962	252.042
2010 (7)	301.570	246.903
2011 (8)	293.560	239.502
2012 (9)	281.538	233.663
2013 (10)	269.578	223.838
2014 (11)	258.943	217.223
2015 (12)	250.327	210.269
2016 (13)	255.925	208.830

Abbildung 4.17: Anzahl der Beobachtungen des Schadenfrequenz Modells

Jahr	No. Observations	Weight
2004 (1)	16.302	17.553
2005 (2)	16.920	18.236
2006 (3)	17.174	18.405
2007 (4)	17.215	18.398
2008 (5)	16.094	17.229
2009 (6)	16.116	17.243
2010 (7)	15.925	17.094
2011 (8)	14.861	15.855
2012 (9)	14.197	15.195
2013 (10)	13.526	14.415
2014 (11)	12.256	13.030
2015 (12)	11.943	12.681
2016 (13)	11.842	12.571

Abbildung 4.18: Anzahl der Beobachtungen des Durchschnittsschaden Modells

Betrachten wir zunächst den Verlauf des Durchschnittsschadens der letzten Jahre:

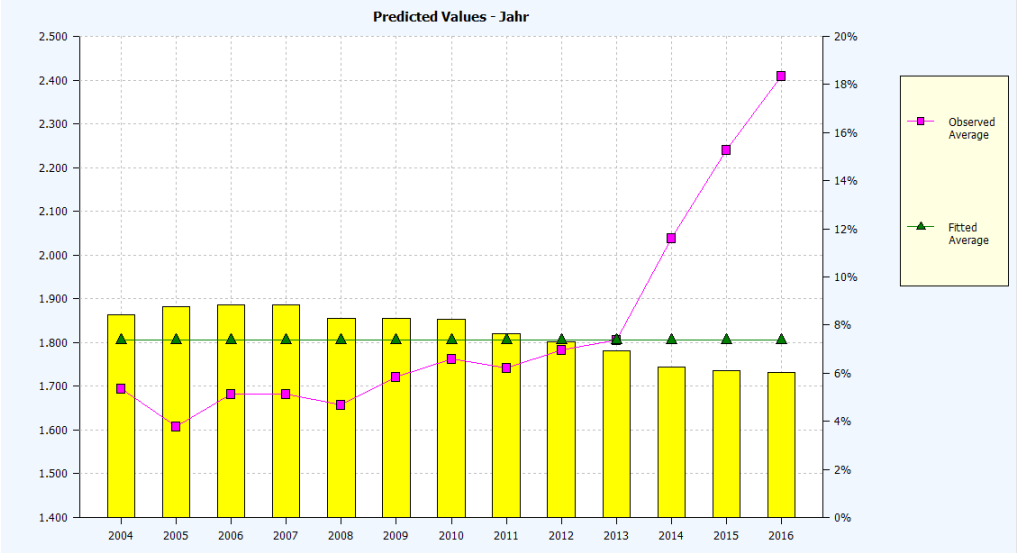


Abbildung 4.19: Verlauf des Durchschnittsschadens pro Jahr

Denkt man an das Schadenfrequenz Modell zurück, das einen stark abfallenden Verlauf, also eine Verbesserung der Schadenhäufigkeit zeigt, erkennt man hier das Hauptproblem des Versicherungsgeschäftes. Die Versicherer haben mit den immer höher werdenen Kosten zu kämpfen. Das spiegelt sich natürlich nicht nur in der KFZ-Haftplicht wieder. Gerade in der KFZ-Industrie wird viel für die Sicherheit getan. Verschiedenste Assistenzsysteme tragen dazu bei, dass die Unfallhäufigkeit stark reduziert werden kann. Jedoch sind es auch genau diese Innovationen, die neben den steigenden Preisen für Material und Arbeitsstunden, einen enormen Einfluss auf die steigenden Kosten im Schadensfall haben.

Zusätzlich sei angemerkt, dass die Daten ab einer Einzelschadenhöhe von 100.000 EUR kupiert worden sind. Dabei handelt es sich um die Großschadenkupierung. Großschäden sind sehr zufallsbehaftet und werden dadurch als Ausreißer behandelt. Die Großschadenproblematik sollte somit isoliert betrachtet werden, worauf in dieser Arbeit aber nicht näher eingegangen wird.

Die Analyse des Durchschnittsschaden Modells verläuft im Grunde identisch zu der Analyse für das Schadenfrequenz Modell. Nachfolgend wird ein Ergebnisunterschied der beiden Modelle dargestellt, um einen Eindruck des Schadenverhaltens zu bekommen.

4.3.1 Alter KFZ

Ein eindeutiges Beispiel für den Unterschied im Verlauf der beiden Modelle bietet das Merkmal Alter KFZ. In den beiden folgenden Grafiken werden wir sehen, wie unterschiedlich sich der Verlauf in der Schadenfrequenz und im Durchschnittsschaden zeigt.

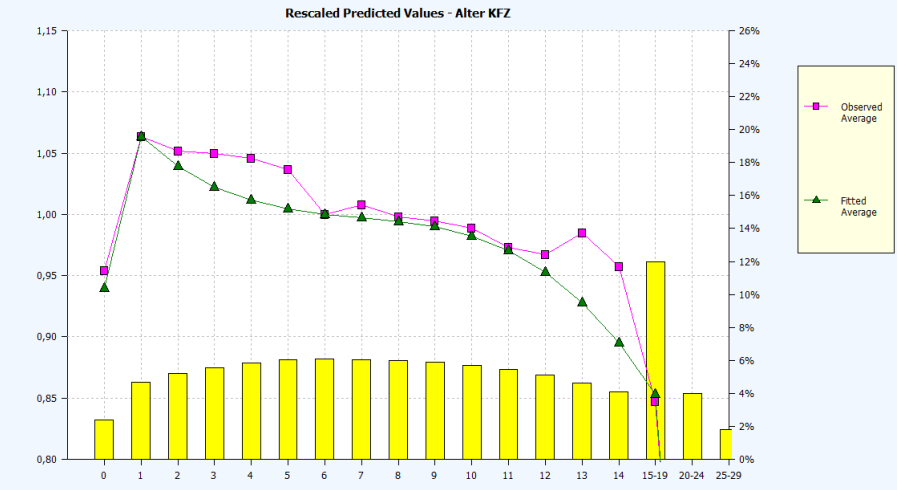


Abbildung 4.20: Frequenz Alter KFZ

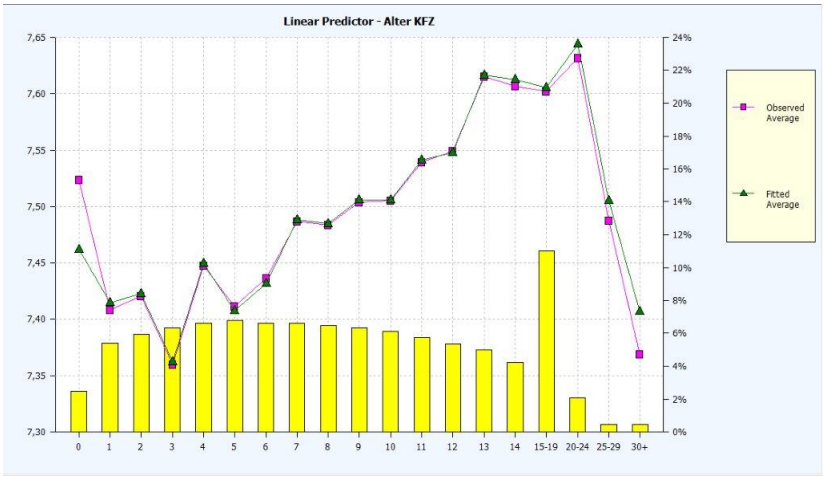


Abbildung 4.21: Durchschnittsschaden Alter KFZ

Es zeigt sich eine mit dem Alter abnehmende Schadenfrequenz, jedoch steigt mit dieser Entwicklung der Durchschnittsschaden.

Wie man zum Abschluss des nächsten Kapitels (anhand der Prämienkalkulationstabelle) erkennen wird, erhalten Fahrzeuge mit dem Alter zwischen 7 und 14 Jahren aufgrund des höchsten Schadenbedarfs den höchsten Prämienzuschlag.

4.4 Schadenbedarf

Wurden die Modelle für die Schadenfrequenz und die Durchschnittsschadenhöhe erfolgreich modelliert, ist es möglich, mit dem von Willis Towers Watson entwickelten Modell Combiner beide Modelle zusammenzuführen. Durch das Zusammenführen beider Modelle erhält man ein Modell um die gesuchte Zielgröße, den Schadenbedarf, zu modellieren. In der Praxis wird für diese Art von Modell entweder die Gammaverteilung oder die Tweedie-Verteilung als Verteilungsannahme getroffen.

Nach Erstellung des kombinierten Schadenbedarf Modelles werden dieselben Analysen, wie in den Kapiteln davor erläutert, durchgeführt. Hat man dieses Tarifmodell erfolgreich modelliert wird durch Export der geschätzten Parameter folgendes Datenblatt erstellt:

Base	119,7032								
Alter VN		**Alter KFZ**		**Region**		**Leistung**		**BM-Stufe**	
Alter VN		**Alter KFZ**		**Region**		**Leistung**		**BM-Stufe**	
NaN	1,0833	0	1,0514	0	1,0000	= 0	0,1703	0	1,0000
>= 25, < 30	1,1880	1	1,0504	1	0,7906	> 0, <= 20	0,1718	1	1,4967
>= 30, < 35	0,8966	2	1,0460	2	0,7381	> 20, <= 40	0,5791	2	1,5201
>= 35, < 40	0,8691	3	0,9838			> 40, <= 60	0,9039	3	1,6950
>= 40, < 45	0,9574	4	1,0700			> 60, <= 80	1,0000	4	1,6979
>= 45, < 50	1,0000	5	1,0178			> 80, <= 100	1,1304	5	2,0325
>= 50, < 55	1,0186	6	1,0000			> 100	1,2789	6	2,0795
>= 55, < 60	0,9534	7	1,0648					7	2,5606
>= 60, < 65	0,8955	8	1,0502					8	2,6943
>= 65, < 70	1,0611	9	1,0646					9	3,8275
>= 70, < 75	1,1728	10	1,0480					10	3,5199
>= 75	1,6674	11	1,0622					11	3,5164
		12	1,0491					12	3,5138
		13	1,1346					13	3,5361
		14	1,1017					14	3,5191
		15-19	1,0133					15	3,5134
		20-24	0,6733					16	3,5274
		25-29	0,4272					17	3,4932
		30+	0,2947						

Abbildung 4.22: Tariffaktoren

Mit Base wird die Basisprämie dargestellt, die für unser Tarifmodell 119,70 EUR beträgt. Des weiteren erhält man für jedes wesentliche Tarifmerkmal die jeweiligen Zu- bzw. Abschlagsfaktoren der Basisprämie.

Als Beispiel für eine Nettoprämienkalkulation wollen wir nun folgenden Versicherungskunden betrachten. Eine 27-jährige Person, die ein Fahrzeug mit dem Alter von 10 Jahren und einer Leistung von 75 kW sowie einem Zulassungsbezirk in Region 1 und einer Bonus Malus Stufe von 3 hat, würde folgende Prämie zahlen:

$$119,70 * 1,1880 * 1,048 * 0,7906 * 1 * 1,695 = 199,71 EUR$$

Interessant sind noch die Nettoprämien der beiden Extrema des besten beziehungsweise vermeintlich schlechtesten Kunden. Die Bandbreite liegt hier zwischen 13 EUR (Leistung 0 und 0-20 wird als Datenfehler angesehen) und knappen 950 EUR.

Kapitel 5

Mindestprämie

Um sicherzugehen, dass das jeweilige Versicherungsprodukt auch rentabel ist, ist es notwendig, die Durchschnittsprämie des Bestandes im Auge zu behalten. Ein guter Vergleichswert bietet hier die Mindestprämie. Die Mindestprämie setzt sich zusammen aus dem durchschnittlichen Schadenbedarf, allen Kosten (Verwaltungskosten, Schadenbearbeitungskosten, Provisionen), Rückversicherungsabgaben, Rückversicherungsprovisionen, Spätschadenzuschlag, sowie einem Risikoaufschlag.

Ich möchte zum Abschluss noch auf die Kalkulation des Risikoaufschlages eingehen. Diesen kann man unter anderem mit der Gausschen-Fehlerfortpflanzung berechnen. Dieser Wert gibt einen Risikoaufschlag für einen möglichen Fehler in der Schadenbedarfsanalyse zurück.

5.1 Gaussche Fehlerfortpflanzung

In der Statistik ist die Ausbreitung der Unsicherheit (oder Ausbreitung des Irrtums) die Wirkung von Unsicherheiten der Variablen (oder Fehler) auf die Unsicherheit einer darauf basierenden Funktion. In der Versicherung liegt die Fehlerquelle vor allem in der Reservehöhenermittlung, da sich das Ausmaß des Schadens oft erst nach einiger Zeit zeigt und daher nur grob geschätzt werden kann.

Vereinfachung

Die Vernachlässigung von Korrelationen oder die Annahme von unabhängigen Variablen ergibt folgende Varianzformel um die Fehlerausbreitung zu berechnen,:

$$s_f = \sqrt{\left(\frac{\partial f}{\partial x}\right)^2 s_x^2 + \left(\frac{\partial f}{\partial y}\right)^2 s_y^2 + \left(\frac{\partial f}{\partial z}\right)^2 s_z^2 \cdots}$$

Wobei s_f die Standardabweichung der Funktion f, s_x die Standardabweichung von x etc. darstellen.

Es folgt ein Beispiel für die Berechnung der Fehlerfortpflanzung:

Stichtag	Summe von Aufwand	Summe von abgegr.Prämie	Anzahl von JE	Summe von Schadenanzahl	Schadenbedarf
20071231	1 721 874,56	3 530 082,06	13 871,84	906,63	124,13
20081231	1 503 608,41	3 343 109,46	13 865,42	824,94	108,44
20091231	1 672 914,52	3 155 122,35	13 853,96	854,10	120,75
20101231	1 890 880,46	3 023 354,74	13 548,40	871,68	139,56
20111231	1 585 249,63	2 911 667,51	13 486,97	799,95	117,54
20121231	1 534 701,86	2 846 517,00	13 166,06	762,69	116,57
20131231	1 572 515,33	2 743 663,64	12 929,31	758,76	121,62
20141231	1 464 886,88	2 692 619,78	12 678,17	712,48	115,54
Erwartungswert	1 618 328,96	3 030 767,07	13 425,02	811,40	120,52

Abbildung 5.1: Daten Übersicht

STD. Ábweichung Schadenbedarf	Var. Koeff. Schadenbedarf	STD. Ábweichung Aufwand	Var. Koeff. Aufwand	STD. Ábweichung JE	Var. Koeff. JE
8,45676507	7%	129844,8572	8,02%	428,2544833	3%

Abbildung 5.2: Standardabweichungen und Varationskoeffizienten

Formel	Aufwand/Polizzen	x/y		
				FEHLER
	Fehler:	Ableitung nach x	"1/y"	8,63%
		Ableitumg mach y	"-x/y^2"	€ 10,41

Berechnung				
	X	1 464 886,88		
	y	12 678		
	Delta X	117 533,60		
	Delta Y	404		
	Fehlerentwicklung	9,98	"+/-" an Schadenbedarf in EUR	8,63%

Abbildung 5.3: Ergebnis

Zusätzlich zum durchschnittlichen Schadenbedarf, allen genannten Kosten und Zuschlägen, sollte somit für den gegebenen Bestand ein Risikoaufschlag von 10,41 EUR eingeplant bzw. verdient werden.

5.2 Schlussfolgerung

Verallgemeinerte Lineare Modelle liefern für die Tarifierung wichtige Ergebnisse und können den Erfolg eines Versicherungsunternehmens steigern. Es ist eine gängige Methode um den Schadenbarf des Bestanden zu analysieren. Wichtig ist es auch mögliche Tarifgenerationen isoliert voneinander zu betrachten. Dabei ist es erst möglich den Erflog oder den Einfluss der umgesetzten tariflichen Änderungen effizient beurteilen zu können.

In der Praxis werden die vorhin kalkulierten Tarifmultiplikatoren jedoch nicht 1:1 an den Versicherungsnehmer weitergegeben. Die minimale Prämie von 13EUR, die man in der Berechnung für den vermeintlich besten Kunden anbieten würde, ist wirtschaftlich gesehen deutlich zu niedrig. Der endgültige Preis hängt oft nicht nur von den kalkulierten Faktoren sondern auch von vertrieblichen Strategien sowie von Experteneinschätzungen der Unternehmen ab. Wichtig ist, dass alle Berechnungen laufen überprüft und evaluiert werden um auf mögliche Änderungen im Bestand reagieren zu können.

Literaturverzeichnis

[1] [Anette J. Dobson and Adrian G. Barnett (2008)] : An Introduction to Generalized Linear Models - Third Edition, Verlag Taylor and Francis Group

[2] [DAV-Arbeitsgruppe Tarifierungsmethodik (2011)] : Aktuarielle Methoden der Tarifgestaltung in der Schaden-/Unfallversicherung, Verlag Versicherungswirtschaft GmbH Karlsruhe

[3] [Hans Jürgend Andreß (1986)] : Verallgemeinerte Lineare Modelle, Verlag Friedrich Vieweg und Sohn Verlagsgesellschaft mbH Braunschweig

[4] [LPatricia Siedlok (2011)] : Die Tarifierung in der Autohaftpflichtversicherung mittels verallgemeinerter linearer Modelle, Mathematisches Institut für Statistik Fachbereich für Mathematik und Informatik Westfälische Wilhelms Universität
URL: http://www.uni-muenster.de/Stochastik/paulsen/Abschlussarbeiten/Diplomarbeiten/Siedlok.pdf
(Zugriff 01.08.2017)

[5] [Herwig Friedl (2013)]: Generalisierte Lineare Modelle, Institut für Statistik Technische Universität Graz
URL: http://www.stat.tugraz.at/courses/files/GLM.pdf
(Zugriff 20.08.2017)

[6] [Ludwig Fahrmeir and Heinz Kaufmann (1985)] : Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models, Verlag The Annals of Statistics
URL: https://projecteuclid.org/euclid.aos/1176346597
(Zugriff 20.08.2017)

[7] [Michael Schomaker (2006)] : Neuere Ansätze für Kriterien zur Modellselektion bei Regressionsmodellen unter Berücksichtigung der Problematik fehlender Daten
URL: http://www.rappenantilope.de/Dokumente/diplom.pdf
(Zugriff: 10.08.2017)

[8] [Paul E. Johnson (2016)] : Residuals and Analysis of fit
URL: http://pj.freefaculty.org/guides/stat/Regression-GLM/GLM2-SigTests/GLM-2-guide.pdf
(Zugriff: 12.08.2017)

[9] [Wikipedia] : Propagation of uncertainty
URL: https://en.wikipedia.org/wiki/Propagation_of_uncertainty
(Zugriff: 26.08.2017)

[10] [Gabler Wirtschaftslexikon] : Tarifierung
URL: http://wirtschaftslexikon.gabler.de/Definition/tarifierung.html
(Zugriff: 26.08.2017)