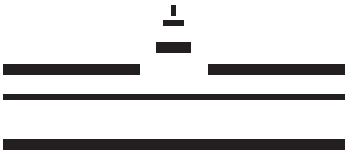

Westfälische
Wilhelms-Universität
Münster

Diplomarbeit

Die Tarifierung in der Autohaftpflichtversicherung mittels verallgemeinerter linearer Modelle

von
Patricia Siedlok

betreut von
PD Dr. Volkert Paulsen

Mathematisches Institut für Statistik
Fachbereich für Mathematik und Informatik
Westfälische Wilhelms Universität

Inhaltsverzeichnis

Hiermit erkläre ich, dass ich die Diplomarbeit selbstständig angefertigt und nur die angegebenen Quellen verwendet habe.

Münster, den 01. Juli 2011

1 Einleitung

Die Deregulierung der Versicherungswirtschaft führte zu vielen Veränderungen in den Tarifen der Kraftfahrzeug-Haftpflichtversicherung. Der Entfall der behördlichen Aufsicht brachte Tarifvariationen sowie Produktinnovationen mit sich, denen häufig stochastische Modelle für eine exakte Kalkulation fehlten.
Als Grundlage für die Bestimmung von unternehmensindividuellen Tarifen stehen Daten aus vorhandenen sowie potentiellen Versicherungsverträgen zur Verfügung. Es werden seit geraumer Zeit Modelle diskutiert, mit denen Risiken beurteilt werden, sodass durch vorher sowie nachher bestimmbare Merkmale der individuelle Gesamtschaden, die Schadenanzahl sowie Schadenhöhe erklärt werden können. In den 60er Jahren fanden häufig additive sowie multiplikative Modelle für die Erklärung von Einflüssen verschiedener Merkmale auf den Gesamtschaden eine Anwendung. Die Schätzung der Modellparameter erfolgte u.A. nach dem Bailey-Simon - oder Marginalsummenverfahren. Mitte der 80er Jahre wurden schließlich Modelle vorgestellt, bei denen einen Transformation der erklärenden Variablen stattgefunden hat, sodass man das klassische lineare Modell auf transformierte Schadengrößen anwenden konnte.
Zur gleichen Zeit wurde mit Modellen gearbeitet, die geeignete Verteilungsannahmen über die Schadenanzahl oder Schadenhöhe verwendeten. Es wurde zusätzlich eine spezielle funktionale Abhängigkeit der erwarteten Schadenanzahl oder Schadenhöhe von einer Linearkombination erklärender Größen unterstellt. Bei diesen Modellen handelt es sich um Mitglieder der Klasse der verallgemeinerten linearen Modelle, die 1972 bereits von McCullagh und Nelder vorgestellt wurden.
Sie werden im Rahmen der Tarifierung eingesetzt, um Faktoren zu bestimmen, die einen Einfluss auf den erwarteten Gesamtschaden, die durchschnittliche Schadenhöhe und erwartete Schadenanzahl haben. Zudem besteht die Möglichkeit geeignete Klassenbreiten der Einflussfaktoren zu ermitteln.

Im ersten Kapitel wird die Theorie hinter den verallgemeinerten linearen Modellen erläutert. Dazu wird zunächst auf die Verteilungsfamilie sowie deren Eigenschaften eingegangen. Anschließend stellen wir die Komponenten der verallgemeinerten linearen Modelle vor und zeigen daraufhin, auf welche Art die Parameter geschätzt werden. Es folgen Goodness of fit - Maße sowie Hypothesentests, zuletze werden verschiedene Residuen vorgestellt.

Im darauffolgenden Kapitel betrachten wir das Versicherungsmodell. Es werden zunächst allgemein Tariffaktoren eingeführt sowie Annahmen über das Modell getroffen. Das multiplikative Modell wird ebenfalls erklärt. Desweiteren wird gezeigt, wie man den Schadenbedarf als Produkt von Schadenfrequenz und Schadenhöhe

modellieren kann.

Im letzten Kapitel wird schließlich eine empirische Analyse durchgeführt. Zu Beginn wird das Datenmaterial vorgestellt und anschließend Modelle für die Schadenfrequenz, die Schadenhöhe und den Schadenbedarf entwickelt. Die Merkmale werden dabei auf ihre Signifikanz geprüft. Darüber hinaus führen wir eine Residualanalyse durch, um die Modellanpassung zu überprüfen. Zum Schluss werden die Indizes für die Ab- sowie Zuschläge bestimmt, sodass die Prämie berechnet werden kann. Bei der verwendeten Statistiksoftware handelt es sich um **R**.

2 Verallgemeinerte lineare Modelle

2.1 Verteilungsfamilie

In verallgemeinerten linearen Modellen betrachtet man eine Zufallsvariable Y, deren Dichte aus der Exponentialfamilie stammt.

Definition 2.1
Eine Zufallsvariable Y besitzt eine Dichte aus der Exponentialfamilie, falls sie mit dem kanonischen Parameter $\theta \in \mathbb{R}$ und Dispersionsparameter $\phi \in \mathbb{R}_+$ dargestellt werden kann als

$$f_Y(y, \theta, \phi) = \exp\left(\frac{y\theta - b(\theta)}{\phi} + c(y, \phi)\right) \ \forall\ y \in \mathbb{R} \tag{1}$$

mit messbaren Funktionen b, c.

Im Gegensatz zur bekannten Exponentialfamilie, bei der $\phi := 1$ gesetzt ist, betrachten wir den allgemeineren Fall $\phi > 0$ für alle Beobachtungen und als Oberbegriff dient *exponential dispersion model*, was wir im Folgenden mit EDM abkürzen.

Der Dispersionsparameter bzw. Skalenparameter ϕ wird als bekannt vorausgesetzt bzw. kann konsistent geschätzt werden, siehe [6]. Die Bestimmung der Funktionen $b(\cdot)$ und $c(\cdot)$, welche den jeweiligen Verteilungstyp aus der Exponentialfamilie spezifizieren, erfolgt durch einen Vergleich mit der Dichtefunktion.

Beispiel 2.2 (Normalverteilung)
Für eine normalverteilte Zufallsvariabele Y gilt

$$\begin{aligned} f_Y(y, \theta, \phi) &= \frac{1}{2\pi\sigma^2} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) \\ &= \exp\left(\frac{(y\mu - \frac{\mu^2}{2})}{\sigma^2} - \frac{1}{2}\left(\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2)\right)\right). \end{aligned}$$

Dies bedeutet für die Parameter und die Funktionen $b(\cdot)$ und $c(\cdot)$

$$\begin{aligned} \theta &= \mu \\ \phi &= \sigma^2 \\ b(\theta) &= \frac{\theta^2}{2} \\ c(y, \phi) &= -\frac{1}{2}\left(\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2)\right). \end{aligned}$$

Demnach ist die Normalverteilung ein Mitglied der Exponentialfamilie.

Eine Zusammenstellung der Parameter und Funktionen einiger für uns relevanter Verteilungen aus der Klasse der EDM ist in Tabelle 1 aufgeführt.

Verteilung	θ	ϕ	$b(\theta)$	$c(y,\phi)$
Normal $\mathcal{N}(\mu,\sigma^2)$	μ	σ^2	$\frac{\theta^2}{2}$	$-\frac{1}{2}\left\{\frac{y^2}{\sigma^2}+\log(2\pi\sigma^2)\right\}$
Poisson $\mathcal{P}(\lambda)$	$\ln(\lambda)$	1	e^θ	$-\log y!$
Gamma $\mathcal{G}(\alpha,\beta)$	$-\frac{1}{\alpha}$	$\frac{1}{\beta}$	$-\ln(-\theta)$	$\beta\log(\beta y)-\log y-\log\Gamma(\beta)$

Tabelle 1: Zusammenstellung der Parameter und Funktionen für einige Verteilungen aus dem EDM

Die Bestimmung von Erwartungswert sowie Varianz der Verteilungen aus dem EDM erfordert eine Betrachtung der Log-Likelihoodfunktionen.

Satz 2.3
Sei $l(Y,\theta,\phi)=\log f_Y(Y,\theta,\phi)$ die Log-Likelihoodfunktion von θ und ϕ. Dann gilt unter den Regularitätsbedingungen:

(i)

$$E_{\theta,\phi}\left(\frac{\partial l(Y,\theta,\phi)}{\partial\theta}\right)=0 \tag{2}$$

(ii)

$$E_{\theta,\phi}\left(\frac{\partial^2 l(Y,\theta,\phi)}{\partial\theta^2}\right)+E_{\theta,\phi}\left(\frac{\partial l(Y,\theta,\phi)}{\partial\theta}\right)^2=0 \tag{3}$$

Beweis:
(i)

$$\begin{aligned}
E_{\theta,\phi}\left(\frac{\partial l(Y,\theta,\phi)}{\partial\theta}\right) &= E_{\theta,\phi}\left(\frac{\partial \log f_Y(Y,\theta,\phi)}{\partial\theta}\right)\\
&= E_{\theta,\phi}\left(\frac{1}{f_Y(Y,\theta,\phi)}\frac{\partial f_Y(Y,\theta,\phi)}{\partial\theta}\right)\\
&= \int f_Y(Y,\theta,\phi)\frac{1}{f_Y(Y,\theta,\phi)}\frac{\partial f_Y(Y,\theta,\phi)}{\partial\theta}dy\\
&= \frac{\partial}{\partial\theta}\int f_Y(Y,\theta,\phi)dy\\
&= \frac{\partial}{\partial\theta}1\\
&= 0
\end{aligned}$$

(ii)

$$\begin{aligned}
& E_{\theta,\phi}\left(\frac{\partial^2 l(Y,\theta,\phi)}{\partial\theta^2}\right) + E_{\theta,\phi}\left(\frac{\partial l(Y,\theta,\phi)}{\partial\theta}\right)^2\\
= \; & E_{\theta,\phi}\left(\frac{\partial}{\partial\theta}\left(\frac{1}{f_Y(Y,\theta,\phi)}\frac{\partial f_Y(Y,\theta,\phi)}{\partial\theta}\right)\right) + E_{\theta,\phi}\left(\left[\frac{\partial \log f_Y(Y,\theta,\phi)}{\partial\theta}\right]^2\right)\\
= \; & E_{\theta,\phi}\left(\frac{1}{f_Y(Y,\theta,\phi)}\frac{\partial^2 f_Y(Y,\theta,\phi)}{\partial\theta^2}\right) - E_{\theta,\phi}\left(\frac{1}{f_Y(Y,\theta,\phi)^2}\left(\frac{\partial f_Y(Y,\theta,\phi)}{\partial\theta}\right)^2\right)\\
& + E_{\theta,\phi}\left(\frac{1}{f_Y(Y,\theta,\phi)^2}\left(\frac{\partial f_Y(Y,\theta,\phi)}{\partial\theta}\right)^2\right)\\
= \; & \int \frac{1}{f_Y(Y,\theta,\phi)} f_Y(Y,\theta,\phi)\frac{\partial^2 f_Y(Y,\theta,\phi)}{\partial\theta^2}dy\\
= \; & \frac{\partial^2}{\partial\theta^2}\int f_Y(Y,\theta,\phi)dy\\
= \; & 0
\end{aligned}$$

□

Damit lassen sich Schlussfolgerungen für den Erwartungswert und die Varianz einer Zufallsvariablen Y ziehen. Es gilt

$$E_{\theta,\phi}\left(\frac{Y-b'(\theta)}{\phi}\right) = 0 \Leftrightarrow E_{\theta,\phi}(Y) = b'(\theta) =: \mu$$

sowie

$$\begin{aligned} E_{\theta,\phi}\left(-\frac{b''(\theta)}{\phi}\right) + E_{\theta,\phi}\left(\frac{Y-\mu}{\phi}\right)^2 &= 0 \\ \Leftrightarrow -\frac{b''(\theta)}{\phi} + \frac{Var_{\theta,\phi}(Y)}{\phi} &= 0 \\ \Leftrightarrow Var_{\theta,\phi}(Y) &= b''(\theta)\phi \end{aligned}$$

Die Varianz wird also als Produkt des Skalenparameters und einer von θ und damit auch vom Erwartungswert abhängigen Funktion gebildet. Es wird vorausgesetzt, dass die Funktion $b(\cdot)$ zweimal differenzierbar und folglich invertierbar ist. Durch eine Umformung der Art

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

gelangt man zur sogenannten *Varianzfunktion*

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

Diese beschreibt den Einfluss des Erwartungswertes μ auf die Varianz von Y. In Tabelle 2 sind die Varianzfunktionen einiger Verteilungen aufgeführt.

Verteilung	Normal	Poisson	Gamma
$\nu(\mu)$	1	μ	μ^2

Tabelle 2: Varianzfunktionen

2.2 Komponenten der verallgemeinerten linearen Modelle

Wir wollen in diesem Abschnitt die Komponenten eines verallgemeinerten linearen Modells vorstellen. Wie der Name bereits vermuten lässt, handelt es sich um eine Verallgemeinerung der klassischen linearen Regression. Im Gegensatz zu linearen Modellen erfolgt keine Beschränkung auf eine normalverteilte Responsevariable. Zudem muss kein linearer Zusammenhang zwischen der erklärenden Variable und der Zielvariable bestehen, desweiteren darf die Varianz darf vom Erwartungswert abhängen.

Es sei $\mathbf{y} = (y_1, ..., y_n)^T$, ein Vektor bestehend aus n Beobachtungen, welche Realisationen von unabhängigen Zufallsvariablen Y_i, i=1,...,n, sind. Weiter seien $\mathbf{x_i} = (x_{i1}, ..., x_{ir})$, i=1,...,n, Kovariablen.
Es gibt drei wesentliche Charakteristika von verallgemeinerten linearen Modellen, von denen wir die erste bereits vorgestellt haben.

a) **Stochastische Komponente**

Die Beobachtungen y_i, i=1,...,n haben eine Dichtefunktion aus der in (1.1) vorgestellten Klasse von Verteilungen mit Erwartungswert μ_i, also

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

mit bekannten Gewichten ω_i, sodas $\sum_{i=1}^{n} \omega_i = 1$. Damit hängt die Verteilung der Zufallsvariablen $\mathbf{Y} = (Y_1, .., Y_n)$ von einem n-dimensionalen Parameter $\boldsymbol{\theta} = (\theta_1, ..., \theta_n)$ bei gegebenem Dispersionsparameter ϕ ab.

Da bei Vorliegen von n Beobachtungen sowie n Parametern keine vernünftige Statistik möglich ist, möchte man eine Abhängigkeit der Verteilung allein von r exogenen Größen $\beta_1, ..., \beta_r$ erreichen. Dies geschieht mithilfe einer Design Matrix, sowie einer Linkfunktion, welche Bestandteile der nächsten beiden Komponenten sind.

b) **Systematische Komponente**

Es existieren zu jeder Beobachtung y_i, i=1,...,n der abhängigen Variablen verschiedene unabhängige Merkmale x_{ij}, j=1,...,r. Die Linearkombination dieser Kovariablen wird zu einem *linearen Prädiktor* $\boldsymbol{\eta} = (\eta_1, ..., \eta_r)$ zusammengefasst. Jede Kovariable hat einen speziellen Einfluss auf die abhängige Variable, gemessen durch β_j, j=1,...,r. Für den linearen Prädiktor gilt somit

$$\eta_i := \sum_{j=1}^{r} \beta_j x_{ij} \text{ bzw. } \boldsymbol{\eta} := \mathbf{X}\boldsymbol{\beta} \text{ mit } \mathbf{X} = \begin{pmatrix} \mathbf{x_1} \\ \vdots \\ \mathbf{x_n} \end{pmatrix} \in \mathbb{R}^{n \times r}$$

Die Matrix $\mathbf{X}$ bezeichnet man auch als Design Matrix.

Bei Merkmalen mit qualitativen Ausprägungen müssen diese binär codiert werden, dies geschieht mithilfe von sogenannten *Dummy-Variablen*, siehe [4]. Betrachtet man z.B. ein Modell mit zwei Merkmalen α, β mit I, J Ausprägungen, so wird

jeder der insgesamt $I + J$ Ausprägungen eine Dummy Variable zugeordnet. Diese gibt Auskunft darüber, ob eine betreffende Ausprägung eines Merkmals vorliegt oder nicht. Da der Rang der durch die Codierung entstandenen Design Matrix kleiner als die Anzahl der zu schätzenden Parameter ist, muss eine Reparametrisierung des Modells erfolgen. Eine Möglichkeit bietet die *Sum-to-zero*-Beschränkung, siehe [6], S. 33. Hierbei wird das Modell um die Restrikion

$$\sum_{i=1}^{I} \alpha_i = \sum_{j=1}^{J} \beta_j = 0$$

erweitert, d.h. die Summe aller Parameter eines Merkmals muss gleich 0 sein. Die Design Matrix wird bei der Codierung um eine Zeile pro Merkmal erweitert, welche diese Restriktion erfüllt. Vorteil dieser Parametrisierung ist, dass das Modell in der Form

$$\eta = \mu + \alpha_i + \beta_j$$

dargestellt werden. Die Haupteffekte α_i und β_j stellen in Abhängigkeit von der jeweiligen Ausprägung einen Auf- bzw. Abschlag dar.
Eine Veranschaulichung der Dummy Variablen wird in Beispiel 3.4 gegeben.

c) **Linkfunktion**

Diese Funktion verknüpft die stochastische mit der systematischen Komponente durch eine Transformation des Erwartungswertes. Die Funktion $g(\cdot)\colon \mathbb{R} \to \mathbb{R}$ heißt *Linkfunktion*. Sie wird als monoton und differenzierbar vorausgesetzt. Es gilt

$$g(\mu_i) = \eta_i = \sum_{j=1}^{r} x_{ij}\beta_j = \mathbf{x}_i^t \boldsymbol{\beta} \ , \ i = 1, ..., n,$$

d.h. der Mittelwert μ_i der i-ten Beobachtung hängt von den Parametern $\beta_1, ..., \beta_r$ ab. Man nennt die Linkfunktion *kanonisch*, falls $\eta_i = \theta_i \ \forall \ i = 1, ..., n$.

Beispiel 2.4 (Kanonische Linkfunktion der Poissonverteilung)
Betrachte unabhängige Zufallsvariablen $Y_i \sim Poi(\lambda)$, i=1,...,n. Es gilt $b(\theta) = \exp(\theta)$, somit ist $\mu = b'(\theta) = \exp(\theta)$. Folglich ist die Bedingung $\eta = \theta$ genau dann erfüllt, wenn $\eta = \ln(\mu) = \ln\exp(\theta) = \theta$. Die kanonische Linkfunktion für die Poissonverteilung ist also gegeben durch den Log-Link $g(\mu) = \ln(\mu)$.

Die für uns relevanten Verteilungen besitzen die folgenden in der Tabelle aufgeführten kanonischen Linkfunktionen.

Verteilung	Kanonische Linkfunktion
Normal $\mathcal{N}(\mu, \sigma^2)$	$\eta_i = \log(\mu_i)$
Poisson $\mathcal{P}(\lambda)$	$\eta_i = \log(\mu_i)$
Gamma $\mathcal{G}(\alpha, \beta)$	$\eta_i = \frac{1}{\mu_i}$

Tabelle 3: Kanonische Linkfunktionen für einige Verteilungen aus dem EDM

Lemma 2.5
In einem verallgemeinerten linearen Modell mit einer kanonischen Linkfunktion ist $(\sum_{i=1}^n x_{i1}y_i, ..., \sum_{i=1}^n x_{ir}y_i)$ *eine suffiziente Statistik für* $\boldsymbol{\beta} = (\beta_1, ..., \beta_r)^t$.

Beweis:
Der Beweis wird mit dem folgenden Faktorisierungslemma geführt. □

Lemma 2.6 (Faktorisierungslemma)
Sei $T(\mathbf{X})$ *eine Statistik mit Bild* I *und seien* $g(t, \theta)$ *sowie* h *Funktionen auf* $\mathbb{R}^n$ *für* $t \in I$, $\theta \in \Theta$. *Genau dann ist* $T(\mathbf{X})$ *suffizient, wenn* $p(\mathbf{x}, \theta) = g(T(\mathbf{X}), \theta) \cdot h(\mathbf{x})$.

Beweis:
siehe [19], S.43. □

Durch die kanonische Linkfunktion besitzen demnach alle unbekannten Parameter linearer Struktur suffiziente Statistiken, sofern die Zielvariablen eine Verteilung aus der Gruppe der EDM besitzen und der Skalenparameter bekannt ist. Trotzdem ist die Linkfunktion größtenteils ein Artefakt, um die numerischen Schätzmethoden zu vereinfachen. Denn hat ein Modell einen linearen Bestandteil, so ermöglicht dies z.B. die Anwendung eines IWLS (iterative weighted least squares) - Algorithmus. Nach McCullagh und Nelder, vgl.[2], kann es trotz der vorteilhaften Eigenschaften der kanonischen Linkfunktion sinnvoll sein, eine andere Linkfunktion zu wählen. Kriterien zur Auswahl einer passenden Linkfunktion sind eine gute Modellanpassung, gute Interpretierbarkeit der Parameter des linearen Prädiktors sowie die Existenz einer einfachen suffizienten Statistik.

Zusammenfassend kann man sagen, dass die Verallgemeinerungen zum linearen Modell gegeben sind durch die unterschiedliche Verteilungsannahme sowie den im Allgemeinen nicht linearen Einfluss des linearen Prädiktors auf den Erwartungswert.

2.3 Parameterschätzung

In dem nächsten Abschnitt beschäftigen wir uns mit der Schätzung der Modellparameter. Diese wird mit der Maximum Likelihood Methode durchgeführt.

Dabei muss der Verteilungstyp bekannt sein und die Design-Matrix $\mathbf{X}$ vollen Rang haben. Wie wir sehen werden, ist der zu schätzende Parameter anstelle des Erwartungswertes μ_i, i=1,...,n, der Zufallsvariablen $Y_1, ..., Y_n$ der Regressionsparameter $\boldsymbol{\beta} = (\beta_1, ..., \beta_r)$, d.h. Ziel wird sein $\boldsymbol{\beta}$ aus den Beobachtungen $y_1, ..., y_n$ zu schätzen.

Wir betrachten wieder Zufallsvariablen Y_i, i=1,..,n mit Dichten der Form

$$f_{Y_i}(y_i, \theta_i, \phi) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{\frac{\phi}{\omega_i}} + c(y_i, \phi, \omega_i)\right) \tag{4}$$

mit bekannten Volumenmaßen ω_i, i=1,..,n.
Für die Likelihoodfunktion gilt

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

Um einen Maximum Likelihood Schätzer zu erhalten, sucht man durch das Maximieren der Likelihoodfunktion einen Wert $\hat{\theta}$, für welchen die beobachtete Stichprobe maximale Wahrscheinlichkeit besitzt. Da es häufig leichter ist, die Log Likelhoodfunktion anstatt der Likelihoodfunktion zu maximieren, betrachten wir

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

Für eine Dichte der Form (4) lautet die Log-Likelihoodfunktion des Parametervektors $\boldsymbol{\theta}$ somit

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

Aufgrund der Abhängigkeit der θ_i, i=1,...,n, allein von β_j, j=1,...,r, genügt es, die Log-Likelihoodfunktion bezüglich β_j zu maximieren.
Um einen Maximum Likelihood Schätzer für $\boldsymbol{\beta}$ zu finden, benötigen wir zunächst die Likelihoodfunktion von $\boldsymbol{\beta}$. Diese erhalten wir durch Anwendung der Kettenregel

$$\begin{aligned}
\frac{\partial l}{\partial \beta_j} &= \sum_{i=1}^{n} \frac{\partial l}{\partial \theta_i}\frac{\partial \theta_i}{\partial \beta_j} \\
&= \frac{1}{\phi}\sum_{i=1}^{n}(\omega_i y_i - \omega_i b'(\theta_i))\frac{\partial \theta_i}{\partial \beta_j} \\
&= \frac{1}{\phi}\sum_{i=1}^{n}(\omega_i y_i - \omega_i b'(\theta_i))\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j}
\end{aligned}$$

Aus der bereits hergeleiteten Beziehung $\mu_i = b'(\theta_i)$ sowie der Linkfunktion $g(\mu_i) = \eta_i = \sum_{j=1}^{r} x_{ij}\beta_j$, also $\mu_i = g^{-1}(\eta_i)$, folgt

$$\frac{\partial l}{\partial \beta_j} = \frac{1}{\phi}\sum_{i=1}^{n}(\omega_i y_i - \omega_i b'(\theta_i))\frac{1}{\upsilon(\mu_i)}\frac{1}{g'(\mu_i)}x_{ij} \tag{5}$$

$$= \frac{1}{\phi}\sum_{i=1}^{n}\omega_i\frac{y_i - \mu_i}{\upsilon(\mu_i)g'(\mu_i)}x_{ij}. \tag{6}$$

Man erhält eine sogenannte *Score Funktion.*

Definition 2.7
Die Score Funktion bzw. der Score Vektor ist definiert als der Vektor der Ableitungen der Log Likelihood Funktion nach $\boldsymbol{\beta}$. Es gilt

$$s(\boldsymbol{\beta}) = \frac{\partial l}{\partial \boldsymbol{\beta}}. \tag{7}$$

Da ein Maximum Likelihood Schätzer als Maximum der Likelihood Funktion definiert ist, müssen wir die Score Funktion gleich Null setzen, dabei spielt der Skalenparameter ϕ keine Rolle mehr. Es muss also gelten

$$\sum_{i=1}^{n}\omega_i\frac{y_i - \mu_i}{\upsilon(\mu_i)g'(\mu_i)}x_{ij} = 0 \ , \ j = 1, ..., r \tag{8}$$

Besitzt dieses Gleichungssystem eine eindeutige Lösung $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, ..., \hat{\beta}_r)$, so ist dies der Maximum Likelihood Schätzer. Er hängt dann nur noch von den Funktionen $\upsilon(\mu_i)$ sowie $g(\mu_i)$ ab. Würde man von gegebenen Funktionen υ und g ausgehen, könnte man den Schätzer $\hat{\boldsymbol{\beta}}$ berechnen ohne explizit ein stochastisches Modell festgelegt zu haben, vgl. [21].
Wir führen nun den Begriff der Fisher Informationsmatrix ein, um anschließend ein Verfahren vorzustellen, wie man einen Maximum Likelihood Schätzer erhält.

Definition 2.8
Die beobachtete Fisher Informationsmatrix ist definiert als

$$\mathbf{I}_{Obs}(\boldsymbol{\beta}) = -\left(\frac{\partial^2 l}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^t}\right),$$

wobei es sich bei $\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^t$ um eine Abkürzung für die vektorwertige Differentiation handelt.
Die erwartete Fisher Informationsmatrix ist definiert durch die Kovarianzmatrix des Score Vektors

$$\mathbf{I}(\boldsymbol{\beta}) := Cov_{\boldsymbol{\beta}}(s(\boldsymbol{\beta})) = E_{\boldsymbol{\beta}}(s(\boldsymbol{\beta})s(\boldsymbol{\beta}^t)).$$

Wir zeigen nun, welcher Zusamenhang zwischen diesen beiden Informationsmatrizen besteht.

Satz 2.9
Die erwartete Fisher Informationsmatrix lässt sich durch die folgende Beziehung aus der beobachteten Fisher Informationsmatrix berechen

$$\mathbf{I}(\boldsymbol{\beta}) = E_{\boldsymbol{\beta}}(\mathbf{I}_{Obs}(\boldsymbol{\beta})) = -E_{\boldsymbol{\beta}}\left(\frac{\partial^2 l}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^t}\right).$$

Beweis:
Da die erwartete Fisher Informationsmatrix als Kovarianzmatrix des Score Vektors definiert ist, wollen wir zeigen, dass $E_{\boldsymbol{\beta}}(\mathbf{I}_{Obs}(\boldsymbol{\beta})) = Cov_{\boldsymbol{\beta}}(s(\boldsymbol{\beta}))$ gilt. Wir betrachten zunächst die folgende Beziehung

$$\begin{aligned}
\frac{\partial^2 l}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^t} &= \frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^t}\ln L(\boldsymbol{\beta}) = \frac{\partial}{\partial\boldsymbol{\beta}^t}\left[\frac{\frac{\partial}{\partial\boldsymbol{\beta}}L(\boldsymbol{\beta})}{L(\boldsymbol{\beta})}\right] \\
&= \frac{\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^t}L(\boldsymbol{\beta})L(\boldsymbol{\beta}) - \frac{\partial}{\partial\boldsymbol{\beta}}L(\boldsymbol{\beta})\frac{\partial}{\partial\boldsymbol{\beta}^t}L(\boldsymbol{\beta})}{(L(\boldsymbol{\beta}))^2} \\
&= \frac{\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^t}L(\boldsymbol{\beta})}{L(\boldsymbol{\beta})} - \underbrace{\frac{\frac{\partial}{\partial\boldsymbol{\beta}}L(\boldsymbol{\beta})\frac{\partial}{\partial\boldsymbol{\beta}^t}L(\boldsymbol{\beta})}{L(\boldsymbol{\beta})L(\boldsymbol{\beta})}}_{\frac{\partial}{\partial\boldsymbol{\beta}}\ln L(\boldsymbol{\beta})\frac{\partial}{\partial\boldsymbol{\beta}^t}\ln L(\boldsymbol{\beta})}
\end{aligned}$$

Damit erhalten wir:

$$\begin{aligned}
E_{\boldsymbol{\beta}}(\mathbf{I}_{Obs}(\boldsymbol{\beta})) &= -\int\left[\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^t}\ln L(\boldsymbol{\beta})\right]L(\boldsymbol{\beta})dy \\
&= -\int\left[\frac{\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^t}L(\boldsymbol{\beta})}{L(\boldsymbol{\beta})} - \frac{\partial}{\partial\boldsymbol{\beta}}\ln L(\boldsymbol{\beta})\frac{\partial}{\partial\boldsymbol{\beta}^t}\ln L(\boldsymbol{\beta})\right]L(\boldsymbol{\beta})dy \\
&= -\int\frac{\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^t}L(\boldsymbol{\beta})}{d}y + \int\underbrace{\frac{\partial}{\partial\boldsymbol{\beta}}\ln L(\boldsymbol{\beta})}_{s(\boldsymbol{\beta})}\underbrace{\frac{\partial}{\partial\boldsymbol{\beta}^t}\ln L(\boldsymbol{\beta})}_{s(\boldsymbol{\beta}^t)}L(\boldsymbol{\beta})dy
\end{aligned}$$

An dieser Stelle nutzen wir aus, dass unter den Regularitätsbedingungen (siehe [8]) eine Vertauschung von Integration und Differentiation zulässig ist. Darüber hinaus handelt es sich bei $L(\boldsymbol{\beta})$ um eine Dichte und damit ist das Integral über

$L(\boldsymbol{\beta})$ gleich 1. Wir führen unsere Rechnung wie folgt weiter:

$$\begin{aligned} &= 0 + \int s(\boldsymbol{\beta}) s(\boldsymbol{\beta})^t L(\boldsymbol{\beta}) dy \\ &= E_{\boldsymbol{\beta}}(s(\boldsymbol{\beta}) s(\boldsymbol{\beta}^t)) \\ &= Cov_{\boldsymbol{\beta}}(s(\boldsymbol{\beta})) \text{ , da } E_{\boldsymbol{\beta}}(s(\boldsymbol{\beta})) = 0 \end{aligned}$$

□

Die *Newton-Raphson Methode* bietet eine Möglichkeit das Gleichungssystem bestehend aus den Score Funktionen zu lösen, deren numerische Umsetzung mit einer iterativen gewichteten Kleinste Quadrate Schätzung verglichen werden kann, siehe [6]. In diesem Verfahren lautet der $k+1$-te Iterationsschritt

$$\hat{\boldsymbol{\beta}}^{\boldsymbol{k+1}} = \hat{\boldsymbol{\beta}}^{\boldsymbol{k}} + \mathbf{I}^{-\mathbf{1}}_{\mathbf{Obs}}(\hat{\boldsymbol{\beta}}^{\boldsymbol{k}}) s(\hat{\boldsymbol{\beta}}^{\boldsymbol{k}}).$$

Dabei wird zu dem Lösungsvektor $\hat{\boldsymbol{\beta}}$ des vorherigen Iterationsschritts ein Korrekturterm hinzugefügt, welcher aus dem Produkt der inversen beobachteten Informationsmatrix und des Score Vektors, jeweils an der Stelle $\hat{\boldsymbol{\beta}}$ ausgewertet, besteht. Sobald der Korrekturterm verschwindet, d.h. der Score Vektor gegen den Wert 0 strebt, bricht das ganze Verfahren ab.

Aufgrund der Tatsache, dass i.A. keine analytischen Lösungen für einen Maximum Likelihood Schätzer in einem verallgemeinerten linearen Modell existieren, können nur asymptotische Eigenschaften gefolgert werden. Diese Eigenschaften sind jedoch nur dann gewährleistet, sofern die Regularitätsbedingungen gelten, siehe [8]. Unter gewissen Voraussetzungen ist der Maximum Likelihood Schätzer asymptotisch normalverteilt, siehe [1], S. 51f.

2.4 Goodness of fit

Bei der Modellüberprüfung stellt man sich mitunter die Frage, ob die Daten hinreichend gut beschrieben werden durch die Klassifikationsmerkmale, sofern Link- und Varianzfunktion als gegeben angenommen werden. Dabei interessiert unter Anderem, ob eine unabhängige Variable überhaupt einen Einfluss auf die Zielvariable hat. Das Ziel ist, durch eine angemessene Datenreduktion alle relevanten Datenstrukturen zu berücksichtigen, sodass die Zielvariable $\mathbf{y} = (y_1, ..., y_n)^t$ bestehend aus n Beobachtungen durch möglichst wenige unabhängige Merkmale x_{ij}, i=1,...,n, j=1,...,r, erklärt werden kann. Formal möchten wir zwei geschachtelte Modelle miteinander vergleichen, ein komplexes sowie ein vereinfachtes Modell. Zunächst betrachten wir zwei extreme Grenzfälle als Modelle.

a) Das einfachste Modell ist ein sogenanntes *Nullmodell*. Es beinhaltet lediglich einen Parameter, nämlich den Mittelwert μ für alle Beobachtungen y_i, i=1,...,n, als beste Prognose. Somit bleiben viele unter Umständen relevante Datendetails unberücksichtigt und es führt zu einer minimalen Datenanpassung.

b) Der andere Grenzfall ist ein vollständiges Modell, genannt *saturiertes Modell*, welches jede Beobachtung als relevant einstuft und demnach die Anzahl der Parameter der Anzahl der Beobachtungen entspricht. Wir bekommen somit für r=n einen perfekten Fit, indem wir $\hat{\mu}_i = y_i \ \forall \ i = 1, ..., n$ setzen.

Für die Praxis ist das Nullmodell zu pauschal, während das saturierte Modell nicht informativ ist, da es lediglich die Daten wiederholt und nicht zusammenfasst bzw. auswertet.
Es muss also eine Modellanpassung zwischen den beiden Extrema gefunden werden. Für die Güte der Anpassung sind zwei Maße in häufiger Verwendung: die *verallgemeinerte Pearson Statistik* χ^2 und die *Devianz*. Es wird eine Auskunft darüber erteilt, wie gut das betrachtete Modell die Daten beschreibt. Darüber hinaus wird das Akaike Informationskriterium ebenfalls zur Variablenselektion hinzugenommen.

2.4.1 Pearson Statisik

Definition 2.10
Die Pearson Statistik ist von der Form

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)}{Var(Y_i)} = \frac{1}{\phi} \sum_{i=1}^{n} \omega_i \frac{(y_i - \hat{\mu}_i)^2}{\upsilon(\hat{\mu}_i)}.$$

Bei $\hat{\mu}_i$ handelt es sich um den mithilfe von $\hat{\beta}_i$ geschätzten Wert für μ_i und bei $\upsilon(\hat{\mu}_i)$ um die geschätzte Varianzfunktion. Im Fall von normalverteilten Daten liegt eine exakte χ^2_{n-r}-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-r}-verteilt, siehe [10], S.243. Allgemein kann man sagen, dass je größer der Wert der Pearson Statistik ist, desto schlechter der Fit des Modells ausfällt.

Ein weiteres Maß für die Güte der Anpassung, welches hier betrachtet werden soll, ist das Akaike Informationskriterium. Mit diesem können Variablen selektiert werden um das Modell besser an die Daten anzupassen.

2.4.2 Akaike Informationskriterium

Definition 2.11
Das Akaike Informationskriterium (AIC) ist definiert als

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

Dabei bezeichnet k die Anzahl der zu schätzenden Parameter im Modell.

Es wird nach dem Modell mit dem kleinsten AIC-Wert gesucht, denn dieses stellt den besten Kompromiss zwischen Modellanpassung und Modellkomplexität dar, vgl. [6], S. 50. Der AIC-Wert sinkt genau dann, wenn der Log Likelihood Wert stärker abnimmt als der Strafterm 2k zunimmt.

Im nächsten Abschnitt wollen wir die Devianz vorstellen.

2.4.3 Devianz

Zunächst betrachten wir die log Likelihoodfunktion

$$l(\boldsymbol{\theta}; \phi, \mathbf{y}) = \frac{1}{\phi} \sum_{i=1}^{n} \omega_i (y_i \theta_i - b(\theta_i)) + c(y_i, \phi, \omega_i)$$

als Funktion des geschätzten Mittelwertvektors $\hat{\boldsymbol{\mu}}$ mithilfe der Beziehungen $\mu_i = b'(\theta_i)$ und $\theta_i = h(\mu_i)$ mit $h = (b')^{-1}$, also

$$l(\hat{\boldsymbol{\mu}}, \phi, \mathbf{y}) = \frac{1}{\phi} \sum_{i=1}^{n} \omega_i (y_i h(\hat{\mu}_i) - b(h(\hat{\mu}_i)) + c(y_i, \phi, \omega_i).$$

Somit haben wir eine Darstellung der log Likelihoodfunktion in der Mittelwertparametrisierung. Die log Likelihoodfunktion $l(\hat{\boldsymbol{\mu}}, \phi, \mathbf{y})$ ist konkav in $\hat{\boldsymbol{\mu}}$ und besitzt in $\mathbf{y} = \hat{\boldsymbol{\mu}}$ ein globales Maximum, denn es gilt

$$\begin{aligned}
\frac{\partial}{\partial \mu_k} l(\hat{\boldsymbol{\mu}}, \phi, \mathbf{y}) &= \frac{1}{\phi}(y_k h'(\hat{\mu}_k)) - b'(h(\hat{\mu}_k)) h'(\hat{\mu}_k) \omega_k \\
&= \frac{1}{\phi}(y_k - \hat{\mu}_k) h'(\hat{\mu}_k) \omega_k \\
&= \frac{y_k - \hat{\mu}_k}{\theta_k(\hat{\mu}_k)\phi} \cdot \omega_k
\end{aligned}$$

Dabei gilt

$$h'(\hat{\mu}_k) = \frac{1}{b''((b')^{-1}(\hat{\mu}_k))} = \frac{1}{\theta_k(\hat{\mu}_k)}.$$

Es folgt, dass $l(\mathbf{y}, \mathbf{y}, \phi) > l(\hat{\boldsymbol{\mu}}, \mathbf{y}, \phi)\ \forall\ \mathbf{y} \in \mathbb{R}^n$ ist und diese Differenz kann zur Beurteilung der Güte des Modellspezifikation herangezogen werden.

Definition 2.12
Die Devianz ist definiert als

$$D^* = 2[l(\mathbf{y}, \mathbf{y}, \phi) - l(\hat{\boldsymbol{\mu}}, \mathbf{y}, \phi)] = \frac{2}{\phi} \sum_{i=1}^{n} \omega_i (y_i h(y_i) - b(h(y_i)) - y_i h(\hat{\mu}_i) + b(h(\hat{\mu}_i))).$$

Multiplikation von D^* mit ϕ ergibt die unskalierte Devianz $D = D^* \phi$.

Die Devianz dient somit als Maß für die Abweichung zwischen den geschätzten und gefitteten Werten. Durch die 2 am Anfang wird eine Normalisierung hergestellt, im Falle einer Normalverteilung stimmt die Devianz mit der Quadratsumme der Residuen überein.

Beispiel 2.13 (Devianz der Poissonverteilung)

$$\begin{aligned} D(\mathbf{y}, \hat{\boldsymbol{\mu}}) &= 2 \sum_{i=1}^{n} \omega_i [y_i \log(y_i) - y_i] - 2 \sum_{i=1}^{n} \omega_i [y_i \log(\hat{\mu}_i) - \mu_i] \\ &= 2 \sum_{i=1}^{n} [\omega_i y_i log \left(\frac{y_i}{\hat{\mu}_i} \right) + \omega_i (\hat{\mu}_i - y_i)] \end{aligned}$$

Weitere Devianzen für von uns betrachtete Verteilungen sind:

$$D(\mathbf{y}, \hat{\boldsymbol{\mu}}) = \sum_{i=1}^{n} \omega_i (y_i - \hat{\mu}_i)^2 \quad \text{(Normalverteilung)}$$

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

Wir wollen nun mithilfe der Devianz zwei geschachtelte Modelle miteinander vergleichen. Sie indiziert uns, wie die Modellanpassung des einfacheren Modells im Vergleich zum saturierten Modell ist, welches eine perfekte Anpassung besitzt. Für diesen Vergleich berechnet man die Devianz für jedes Modell und subtrahiert die Devianz des saturierten Modells D_s von der Devianz des reduzierten Modells D_r. Es gilt das folgende Lemma:

Lemma 2.14 (D-Subtraktion)
Betrachte zwei Modelle H_r und H_s mit $H_s \subset H_r$. Sei $\hat{\mu}^{(r)}$ der Maximum Likelihood Schätzer unter H_r und $\hat{\mu}^{(s)}$ der unter H_s. Dann lautet die Statistik für den Test von H_s gegen H_r:

$$D^*(\mathbf{y}, \hat{\boldsymbol{\mu}}^{(s)}) - D^*(\mathbf{y}, \hat{\boldsymbol{\mu}}^{(r)})$$

Es sollen nun zwei Hypothesentests vorgestellt werden, deren Durchführung mit der Devianz und ihrer asymptotischen χ^2-Verteilung ermöglicht wird.

2.5 Residual Deviance und Partial Deviance Test

Der *Residual Deviance Test* testet die allgemeine Hypothese

H: Modell passt gegen K: Modell passt nicht

Die Nullhypothese wird genau dann verworfen, d.h. die Anpassung des Modells an die Daten ist nicht ausreichend, wenn

$$\text{Devianz} > \chi^2_{n-r,1-\alpha}.$$

Dabei bezeichnet wieder n die Anzahl der Beobachtungen und r die Anzahl der zu schätzenden Parameter. Bei α handelt es sich um das Signifikanzniveau und bei $\chi^2_{n-r,1-\alpha}$ um das $(1-\alpha)$-Quantil der χ^2-Verteilung mit $n-r$ Freiheitsgraden.

Während der Residual Deviance Test das komplette Modell überprüft, betrachtet der *Partial Deviance Test* nur eine oder mehrere Kovariablen und misst deren Signifikanz. Dazu wählt man zwei Modelle M_1 und M_2, die man miteinander vergleichen möchte. Das Modell M_2 muss im Vergleich zu M_1 diejenigen Parameter zusätzlich enthalten, deren Einfluss getestet werden soll. Nun wird der Regressionsparameter in zwei Bestandteile $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)^t$ zerlegt. Das Modell M_1 besteht aus den Parametern, welche in $\boldsymbol{\beta}_1$ enthalten sind, M_2 enthält zusätzlich die Parameter aus $\boldsymbol{\beta}_2$. Getestet wird im Partial Deviance Test

H : $\boldsymbol{\beta}_2 = 0$ gegen K: $\boldsymbol{\beta}_2 \neq 0$.

Hier wird die Nullhypothese genau dann verworfen, falls gilt

$$\text{Devianz}(M_1) - \text{Devianz}(M_2) > \chi^2_{r_2-r_1,1-\alpha}.$$

Dabei bezeichnet r_1 die Anzahl der Parameter im Modell M_1, analog r_2 die des Modells M_2. Bei $\chi^2_{r_2-r_1,1-\alpha}$ handelt es sich um das $(1-\alpha)$-Quantil der der χ^2-Verteilung mit $r_2 - r_1$ Freiheitsgraden. Muss die Nullhypothese abgelehnt werden, bedeutet dies, dass die zu den Parametern aus $\boldsymbol{\beta}_2$ gehörenden Kovariablen einen signifikanten Einfluss haben.

Die in diesem Abschnitt vorgestellten Gütemaße, Pearson Statistik und Devianz, geben einen ersten Überblick und lassen eine Vorentscheidung zu, ob ein

Modell passend ist oder nicht. Die endgültige Überprüfung der Modellannahmen beinhaltet eine Analyse der Residuen, welche im nächsten Abschnitt behandelt werden soll.

2.6 Residuen

In klassischen linearen Modellen sind Residuen als Differenz zwischen beobachteten und geschätzten Werten definiert, d.h.

$$r_i = (y_i - \hat{\mu}_i),\ i = 1, ..., n.$$

Dabei gilt

$$(y_i - \hat{\mu}_i) \sim N(0, \sigma^2),\ i = 1, ..., n.$$

Die Residuen sind in großen Stuchproben symmetrisch um den Wert 0 verteilt mit konstanter Varianz σ^2 . Wie in linearen Modellen werden Residuen auch hier verwendet um den Fit eines Modells zu überprüfen und eventuelle Ausreißer aufzudecken, also Muster zu erkennen, die nicht verträglich mit den Modellannahmen sind. In den verallgemeinerten linearen Modellen brauchen wir eine erweiterte Definition der Residuen, welche man auf die Verteilungen anwenden kann, die die Normalverteilung in linearen Modellen ersetzen können. Ziel wird sein, so nah wie möglich an normalverteilte Residuen zu kommen, um z.B. die grafischen Mittel aus den linearen Modellen nutzen zu können. Die Unabhängigkeit von den Gauss-Markov Annahmen, welche besagen, dass die Residuen Mittelwert Null und konstante Varianz haben, hat hier zur Folge, dass wir komplexere stochastische Strukturen betrachten müssen.
Wir stellen zwei verschieden Arten von Residuen für verallgemeinerte lineare Modelle vor.

2.6.1 Pearson Residuen

Pearson Residuen r_{P_i}, i=1,...,n, sind definiert als

$$r_{P_i} = \frac{y_i - \hat{\mu}_i}{\sqrt{\frac{v(\hat{\mu}_i)}{\omega_i}}}.$$

Die bekannten Residuen werden also mit der gewichteten geschätzten Standardabweichung skaliert. Der Name Pearson Residuum entstammt der Tatsache, dass die Summe der quadrierten Residuen der unskalierten Pearson Statistik entspricht:

$$\sum_{i=1}^{n} r_{P_i}^2 = \phi X^2.$$

Standardisierte Pearson Residuen sind definiert durch

$$r_{P_i}(\text{Standard}) = \frac{r_{P_i}}{\sqrt{\phi(1-h_i)}},$$

wobei h_i ein Korrekturwert ist.

In Idealsituationen sind Pearson Residuen normalverteilt, allerdings können sie auch sehr verzerrt sein.

Für poissonverteilte Daten sind Pearson Residuen gegeben durch

$$r_{P_i} = \frac{y_i - \hat{y}_i}{\sqrt{\frac{\hat{\mu}_i}{\omega_i}}},$$

da die Varianz mit dem Erwartungswert übereinstimmt.

2.6.2 Devianzresiduen

Verwendet man als Maß der Diskrepanz die Devianz, so trägt jede Einheit eine Größe d_i zu dem Maß bei, sodass

$$D = \sum_{i=1}^{n} d_i$$

mit

$$d(y, \mu) = 2[yh(y) - b(h(y)) - yh(\mu) + b(h(\mu))].$$

Das bedeutet, die Devianz wird aus der Summe der Devianzen der einzelnen Beobachtungen zusammengesetzt. Das Devianzresiduum wird definiert als

$$r_{D_i} = \sqrt{\omega_i d(y, \hat{\mu}_i)} \times sign(y_i - \hat{\mu}_i)$$

mit

$$sign(x) = \begin{cases} -1 & \text{für } x < 0 \\ 0 & \text{für } x = 0 \\ 1 & \text{für } x > 0 \end{cases}$$

Dieses Maß wächst mit $y_i - \hat{\mu}_i$ und es gilt

$$\sum_{i=1}^{n} r_{D_i}^2 = \phi D^*.$$

Die Standardisierten Devianzresiduen werden wie im Fall der Pearson Residuen durch eine Division mit $\sqrt{\phi(1-h_i)}$ gebildet.

Devianzresiduen lauten im Falle poissonverteilter Daten:

$$r_{D_i} = sign(y_i - \hat{\mu}_i)\sqrt{2\omega_i\left[y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i)\right]}.$$

Hat man die Residuen letzendlich bestimmt, so stellt man sie grafisch z.B. gegen die Anzahl der Beobachtungen in einem Index Plot dar um Ausreißer zu entdecken. Eine andere Möglichkeit ist die Residuen gegen die μ_i zu plotten, um die Eignung der Varianzfunktion zu prüfen.

2.7 Schätzung des Skalenparameters

Falls des Skalenparameter ϕ unbekannt ist, gibt es verschiedene Möglichkeiten, einen Schätzer anzugeben.

2.7.1 Pearson Schätzer

Um einen näherungsweise erwartungstreuen Schätzer $\hat{\phi}$ zu finden, verwendet man die im Abschnitt 2.4.1. erwähnte Eigenschaft der Pearson Statistik der asymptotischen χ^2-Verteilung

$$X^2 \sim \chi^2_{n-r}.$$

Wobei r die Anzahl der geschätzten β-Parameter und (n-r) de Anzahl der Freiheitsgrade bezeichnet. Somit ist

$$E(X^2) \approx n - r$$

und wir bekommen einen konsistenten Schätzer durch

$$\hat{\phi}_X = \frac{\phi X^2}{n-r} = \frac{1}{n-r}\sum_{i=1}^{n} \omega_i \frac{(y_i - \hat{\mu}_i)^2}{\upsilon(\hat{\mu}_i)}.$$

2.7.2 Devianzschätzer

Ein weiterer Schätzer für den Skalenparameter auf der Grundlage der Devianz ist gegeben durch

$$\hat{\phi}_D = \frac{\phi D^*(\mathbf{y}, \hat{\boldsymbol{\mu}})}{n-r} = \frac{D(\mathbf{y}, \hat{\boldsymbol{\mu}})}{n-r}.$$

Nach McCullagh und Nelder, [2] S. 296, reagiert der Devianzschätzer sehr empfindlich auf Fehlerrundungen bei gammaverteilten Beobachtungen in einer kleinen Stichprobe und sie empfehlen demnach die Nutzung des Pearson Schätzers.

Eine weitere Möglichkeit bietet die Verwendung des Kleinsten Quadrate Schätzers.

2.7.3 Kleinste Quadrate Schätzer

Bei gegebenem Regressionsvektor $\boldsymbol{\beta} \in \mathbb{R}^r$ gelten für den Erwartungswert und die Varianz

$$\begin{aligned} E_{\beta,\phi} Y_i &= \mu_i(\boldsymbol{\beta}) \\ Var_{\beta,\phi} Y_i &= \frac{\phi \upsilon(\mu_i(\boldsymbol{\beta}))}{\omega_i}, \ i = 1, ..., n. \end{aligned}$$

Damit können wir ein Gleichungssystem aufstellen, sodass für i=1,...,n gilt:

$$(Y_i - \mu_i(\boldsymbol{\beta}))^2 = \frac{\phi \upsilon(\mu_i(\boldsymbol{\beta}))}{\omega_i} + U_i(\mu_i(\boldsymbol{\beta}), \phi) \text{ mit } \ E_{\beta,\phi} U_i(\mu_i(\boldsymbol{\beta}), \phi) = 0.$$

Der Kleinste Quadrate Schätzer ergibt sich dann durch

$$\hat{\phi}_{KQS} = \frac{1}{n} \sum_{i=1}^{n} \frac{(y_i - \mu_i(\boldsymbol{\beta}))^2}{\upsilon(\mu_i(\boldsymbol{\beta}))} \cdot \omega_i.$$

Ist der Regressionsvektor $\boldsymbol{\beta}$ hingegen unbekannt, wird in der obigen Gleichung $\boldsymbol{\beta}$ durch $\hat{\boldsymbol{\beta}}$ ersetzt und wir erhalten

$$\hat{\phi}_{KQS} = \frac{1}{n} \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\upsilon(\hat{\mu}_i)} \cdot \omega_i$$

als Schätzer für ϕ.
Man kann zeigen, dass $\hat{\phi}_{KQS}$ unter bestimmten Voraussetzungen konsistent ist, siehe [7], S. 59ff.

2.8 Modellunabhängige Gütemaße

Neben den in Kapitel 2.4 vorgestellten Gütemaßen wie der Devianz und der Pearson Statistik gibt es weitere Möglichkeiten verschiedene Modelle miteinander zu vergleichen. Die modellunabhängigen Gütemaße bieten sich vor allem bei dem Vergleich von Modellen mit unterschiedlichen Varianzfunktionen an, siehe [6], S. 52. Allgemein anwendbare Gütemaße sind der absolute Fehler (AF), der quadratische

Fehler (QF) sowie der relative quadratische Fehler (RQF), die wie folgt definiert sind:

$$\begin{aligned} AF &= \sum_{i=1}^{n} \omega_i |y_i - \hat{\mu}_i| \\ QF &= \sum_{i=1}^{n} \omega_i (y_i - \hat{\mu}_i)^2 \\ RQF &= \sum_{i=1}^{n} \omega_i \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}. \end{aligned}$$

Der Vorteil dieser Gütemaße liegt darin, dass sie lediglich von den beobachteten (y_i) und prognostizierten Werten ($\hat{\mu}_i$) abhängen und somit unabhängig von modelltheoretischen Überlegungen anwendbar sind, vgl [6].

3 Versicherungsmodell

3.1 Grundlagen

3.1.1 Begriffe

Versicherungsverträge, die zwischen Versicherungsunternehmen und Versicherungsnehmer entstehen, erfüllen den Zweck, den Versicherten vor finanziellen Folgen aus Schadenfällen zu schützen. Damit dieser Schutz gewährleistet werden kann, muss der Versicherungsnehmer eine Prämie an das Versicherungsunternehmen entrichten. Das Versicherungsunternehmen muss dementsprechend die Höhe der Prämie so bestimmen, dass es im Schadenfall in der Lage ist, das abgegebene Leistungsversprechen halten zu können. Dabei zeigt sich das Problem, dass bei der Festsetzung der Prämie eine Ungewissheit besteht, ob in der Zukunft Versicherungsleistungen fällig werden und in welcher Höhe diese dann sein werden. Bei der Autohaftpflichtversicherung kennt man insbesondere die Häufigkeit der Schäden sowie deren Ausmaß nicht. Demnach ist es für das Versicherungsunternehmen von großer Bedeutung, eine ausreichend große Prämie anzusetzen, welche dennoch in ihrer Höhe interessant für den Versicherungsnehmer ist, das heißt niedrig im Verhältnis zur Höhe der versicherten Schäden. Als Maßnahme bildet das Versicherungsunternehmen Gemeinschaften von Versicherungsnehmern, Risiken werden zu einem Portfolio zusammengefasst. Die geschieht mit der Erwartung, dass in einem Bestand von Risiken die Anzahl der Schäden sowie die durchschnittliche Schadenbelastung pro Risiko gering ist. Große und kleine Risiken eines Portfolios sollten einander weitgehend ausgleichen. Es handelt sich also um einen *Ausgleich im Kollektiv*.
Eine Möglichkeit in der Autohaftpflichtversicherung ein Portfolio zu bilden, besteht in der Betrachtung der Ausprägungen von Tarifmerkmalen wie z.B. der Regionalklasse und Typklasse des Fahrzeugs. Ein Portfolio sollte möglichst ähnliche Risiken besitzen, den Idealfall mit identischen Risiken bezeichnet man als homogenes Portfolio.

Die Tarifmerkmale gehören häufig einer der folgenden Kategorien an.

- Eigenschaften des Versicherten, z.B Alter, Geschlecht
- Eigenschaften des versicherten Objekts, z.B. Alter, Typklasse
- Eigenschaften der geografischen Region, z.B. Bevölkerungsdichte

Oft teilt man Tarifmerkmale wie Alter in Intervalle und berechnet Prämien anhand von zusammengefassten Merkmalen.

Um einen Tarif zu erhalten, führt man eine Tarifanalysis durch, dabei verwendet man unternehmenseigene Daten von Versicherungsverträgen, welche manchmal vom statistischen Bundesamt ergänzt werden.

Zunächst sollen einige grundlegende Begrifflichkeiten geklärt werden.

Die *Schadenfrequenz* ist die Anzahl der Schäden geteilt durch die Laufzeit.

Bei der *Schadenschwere bzw. Schadenhöhe* handelt es sich um den Quotienten aus der totalen Schadenhöhe und der Anzahl der Schäden, z.B. durchschnittliche Kosten pro Schaden.

Der *Schadenbedarf* wird gebildet aus dem Quotienten der totalen Schadenhöhe und der Laufzeit, z.B. durchschnittliche Kosten pro Vertragsjahr. Folglich kann der Schadenbedarf dargestellt werden als Produkt aus der Schadenfrequenz und der Schadenschwere, also $SB = SF \times SH$.

Bei diesen Kennzahlen handelt es sich um die Beziehung des Ergebnisses einer Zufallsvariablen zu einem Volumenmaß.

3.1.2 Modellannahmen

Um eine Grundlage für die statistische Modellierung zu haben, werden nun einige Modellannahmen getroffen.

Annahme 3.1 (Vertragsunabhängigkeit)
Betrachte n unterschiedliche Policen. Sei X_i die Ausgangsvariable für den i-ten Versicherungsvertrag. Dann sind $X_1, ..., X_n$ unabhängig.

In der Realität ist diese Annahme nicht immer erfüllt, betrachtet man zum Beispiel Situationen, in denen mehrere Versicherungsnehmer den gleichen Schaden erleiden, wie in Sturm- oder Hagelmomenten. Solche Katastrophen werden jedoch auf eine andere Weise als auf die in dieser Diplomarbeit behandelt und werden daher nicht weiter thematisiert.

Annahme 3.2 (Zeitunabhängigkeit)
Betrachte n disjunkte Zeitintervalle. Sei X_i die Ausgangsvariable im i-ten Zeitintervall. Dann sind $X_1, ..., X_n$ unabhängig.

Dies bedeutet, dass sowohl die Anzahl der Schäden als auch die Schadenhöhe von einer Zeitperiode zur nächsten unabhängig sind. Auch hierbei gibt es in der Realität Ausnahmen, da man z.B. nach einem Unfall vorsichtiger fährt und die Wahrscheinlichkeit damit sinkt, erneut einen Unfall zu verursachen.

Aus beiden Annahmen resultiert die Tatsache, dass die individuellen Schadenkosten unabhängig voneinander sind. Für eine Tarifanalyse möchte man wie bereits beschrieben mithilfe von Tarifmerkmalen homogene Gruppen von Versicherungsverträgen haben in Form von Tarifzellen um dieselbe Prämie innerhalb einer Tarifzelle verlangen zu können.

Annahme 3.3 (Homogenität)
Betrachte zwei beliebige Versicherungsverträge aus einer Tarifzelle. Sei X_i die Ausgangsvariable für Versicherungsvertrag i. Dann besitzen X_1 und X_2 dieselbe Wahrscheinlichkeitsverteilung.

Auch diese Annahme ist in der Realität nicht immer bzw. nur in wenigen Fällen erfüllt. Daher genügt es, wenn sie beinahe homogen sind und löst das Problem in der Autohaftpflichtversicherung mir Bonus/Malus - Systemen.
Um einen Tarif festzusetzen, sind nun verschiedene Schritte notwendig.
Zunächst werden die Tariffaktoren bestimmt. Dabei handelt es sich um die Merkmale, welche die individuelle Gesamtschadenverteilung beeinflussen und am besten die zugrunde liegende Verteilung erklären können. Anschließend werden die Tariffaktoren in Klassen unterteilt. Zu einer Klasse eines Tariffaktors gehören diejenigen Merkmalsausprägungen, welche in ihrem Einfluss auf die Schadenverteilung keinen großen Unterschied aufweisen. Zuletzt wird die Risikoprämie geschätzt.

3.1.3 Multiplikative Modelle

In diesem Abschnitt soll das häufig verwendete multiplikative Modell vorgestellt werden.
In der Praxis weisen viele Zellen keinen einzigen Schaden auf, was dazu führt, dass man nach Methoden sucht, die eine erwartete reine Prämie ausgeben, welche glatter über den Zellen 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.
Seien N Tariffaktoren gegeben mit einer Anzahl von n_i Ausprägungen für Tarifmerkmal i. Der Einfachheit halber beginnen wir mit 2 Tariffaktoren, also $N = 2$. Dann haben wir in der Tarifzelle (i, j) als Exposure ω_{ij} und als Response X_{ij}, sodass für den Zielwert $Y_{ij} = \frac{X_{ij}}{\omega_{ij}}$ gilt. Unter der 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 $\{\gamma_{1i}; i = 1, ..., n_1\}$ um Parameter die den verschiedenen Ausprägungen des ersten Tarifmerkmals entsprechen und bei $\{\gamma_{2j}; j = 1, ..., n_2\}$ um die für das zweite Merkmal. Der sogenannte Basiswert ist γ_0. Man wählt nun eine Referenzzelle, vorzugsweise mit großer 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 Policen in der Referenzzelle interpretiert werden. Die übrigen Parameter messen den relativen Unterschied in Bezug auf die Referenzzelle und werden *Relativitäten* genannt.
Die Multiplikativitätsannahme bedeutet, dass zwischen den Tariffaktoren keine Interaktion existiert.
Eine Erweiterung der Formel für den Fall von N Tariffaktoren sieht wie folgt aus:

$$\mu_{i_1,i_2,...,i_N} = \gamma_0\gamma_{1i_1}\gamma_{2i_2}\cdots\gamma_{Ni_N}.$$

Man passt also den Basiswert an und die übrigen Parameter kontrollieren, wieviel als Prämie berechnet wird. Wir werden in der Empirischen Analyse auf das multiplikative Modell zurückgreifen. Dabei werden zunächst die Relativitäten bestimmt und am Schluss der Basiswert.
Wir wollen an einem Beispiel aufzeigen, wie das multiplikative Modell in den verallgemeinerten linearen Modellen angewendet werden kann.

Beispiel 3.4
Seien zwei Merkmale gegeben, eins mit zwei Ausprägungen und das andere Merkmal mit drei Ausprägungen. Sei μ_{ij} der Erwartungswert der Response in Tarifzelle (i,j), d.h. für das erste Merkmal liegt die Ausprägung i vor, für das zweite Merkmal die Ausprägung j. In linearen Modellen wird eine additive Struktur für den Mittelwert angenommen, d.h.

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

Wir setzen die Zelle (1,1) als Referenzzelle, d.h. es gilt $\gamma_{11} = \gamma_{21} = 0$ und demnach $\mu_{11} = \gamma_0$. Im nächsten Schritt schreiben wir das Modell um, sodass wir eine Listenform bekommen. Die Parameter werden wie folgt umbenannt:

$$\begin{aligned}\beta_1 &= \gamma_0\\ \beta_2 &= \gamma_{12}\\ \beta_3 &= \gamma_{22}\\ \beta_4 &= \gamma_{23}\end{aligned}$$

Mit diesen Parametern sind die Erqartungswerte in den Tarifzellen wie folgt. Nun führen wir sogenannte Dummy-Variablen durch folgende Relation ein

$$x_{ij} = \begin{cases}1, & \textit{wenn } \beta_j \textit{ in } \mu_i \textit{ enthalten ist}\\ 0, & \textit{sonst.}\end{cases}$$

i	Tarifzelle		μ_i			
1	1	1	β_1			
2	1	2	β_1		$+\beta_3$	
3	1	3	β_1			$+\beta_4$
4	2	1	β_1	$+\beta_2$		
5	2	2	β_1	$+\beta_2$	$+\beta_3$	
6	2	3	β_1	$+\beta_2$		$+\beta_4$

Tabelle 4: Parametrisierung in Listenform

In unserem Beispiel sind die Werte für die Dummy-Variablen in der folgenden Tabelle aufgelistet. Mit diesen Variablen erhalten wir für den Mittelwert

i	Tarifzelle		x_{i1}	x_{i2}	x_{i3}	x_{i4}
1	1	1	1	0	0	0
2	1	2	1	0	1	0
3	1	3	1	0	0	1
4	2	1	1	1	0	0
5	2	2	1	1	1	0
6	2	3	1	1	0	1

Tabelle 5: Dummy Variablen

$$\mu_i = \sum_{j=1}^{4} x_{ij}\beta_j \text{ für } i = 1, .., 6.$$

Nun wollen wir statt dem additiven ein multiplikatives Modell betrachten. Für den Mittelwert gilt demnach

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

Durch Logarithmieren folgt

$$\log(\mu_{ij}) = \log(\gamma_0) + \log(\gamma_{1i}) + \log(\gamma_{2j}).$$

Wie beim additiven Modell müssen wir eine Referenzzelle angeben. Wir wählen erneut die Zelle (1,1) und setzen $\gamma_{11} = \gamma_{21} = 1$. Die Parameter werden anschließend

umbenannt zu

$$\begin{aligned}\beta_1 &= \log\gamma_0\\ \beta_2 &= \log\gamma_{12}\\ \beta_3 &= \log\gamma_{22}\\ \beta_4 &= \log\gamma_{23}.\end{aligned}$$

Mithilfe der Dummy-Variablen aus Tabelle 5 folgt

$$\log(\mu_{ij}) = \sum_{i=1}^{4} x_{ij}\beta_j \text{ für } i = 1, ..., 6.$$

Somit können wir für eine logarithmische Linkfunktion $g(\mu_i) = \log(\mu_i)$ *den Mittelwert mit einer linearen Struktur durch* $g(\mu_i) = \eta_i = \sum_{j=1}^{4} x_{ij}\beta_j$ *verknüpft.*

3.2 Schadenbedarfsanalyse

In diesem Kapitel soll gezeigt werden, wie man mithilfe von verallgemeinerten linearen Modelle die Schadenbedarfsanalyse modellieren kann. Der Schadenbedarf gilt in der Regel als die interessierende Zielgröße in einem verallgemeinerten linearen Modell. Es gibt im Wesentlichen zwei Möglichkeiten, den Schadenbedarf zu modellieren:

1. Die Zielgröße Schadenbedarf direkt modellieren.

2. Die Zielgröße Schadenbedarf indirekt über die separate Betrachtung der Schadenfrequenz und der Schadenhöhe modellieren und die Ergebnisse anschließend multiplikativ zusammenführen.

In der direkten Modellierung wird eine Verteilungsfunktion gesucht, die die Gesamtschadenverteilung gut approximiert. Häufig wird die Gammaverteilung zur direkten Modellierung verwendet.

Der Vorteil der direkten Methode liegt im geringen Aufwand, denn es muss nur ein verallgemeinertes lineares Modell angepasst werden. Nachteilig ist jedoch, dass kein dominierendes Verfahren in der direkten Modellierung besteht und die Anwendung dieser Methode in der Versicherungspraxis demnach eine untergeordnete Rolle spielt. Im Gegensatz dazu ist die indirekte Modellierung in häufiger Verwendung und aus diesem Grund wollen wir hier den Fokus auf die indirekte Modellierung legen.

Sei H_i die Höhe des i-ten eingetretenen Schadens auf und N die in einem Zeitraum eingetretene Anzahl an Schäden. Dies ist eine Zufallsvariable auf $\mathbb{N}_0$, die unabhängig ist von $H_1, H_2, ...$ Das Paar

$$< N, \{H_i\}_{i\in\mathbb{N}} >$$

heißt *kollektives Modell* für einen Bestand von Risiken, wenn die Folge $\{H_i\}_{i\in\mathbb{N}}$ unabhängig und identisch verteilt und unabhängig von N ist, vgl [12], S. 164.
Für den zufälligen Gesamtschaden S gilt:

$$S = \sum_{i=1}^{N} H_i.$$

Bemerkung 3.5
Sei $R = Vert(N)$ die Schadenfrequenzverteilung und $Q = Vert(H_i)$ die Schadenhöheverteilung. Dann gilt für die Verteilung des Gesamtschadens

$$\textit{Vert(S)} = \sum_{k=0}^{\infty} R(k) Q^{*(k)}$$

Beweis:
Es gilt für $B \in \mathfrak{B}$, $B \subset (0, \infty)$

$$\begin{aligned} P(S \in B) &= \sum_{k=1}^{\infty} P(S \in B, N = k) \\ &= \sum_{k=1}^{\infty} P(\sum_{i=1}^{k} H_i \in B, N = k) \\ &= \sum_{k=1}^{\infty} P(\sum_{i=1}^{k} H_i \in B) \cdot P(N = k) \\ &= \sum_{k=1}^{\infty} Q^{*(k)}(B) R(k) \\ &= \sum_{k=0}^{\infty} Q^{*(k)}(B) R(k), \text{ da } 0 \notin B \end{aligned}$$

□

Der Vorteil der indirekten Modellierung liegt darin, dass man die Art wie die erklärenden Variablen auf die einzelnen Komponenten der interessierenden Zielgröße

Schadenbedarf wirken, besser nachvollziehen kann, vgl.[9].

Im Folgenden wollen wir die separate Modellierung der Schadenfrequenz und Schadenhöhe näher betrachten.

3.2.1 Schadenhäufigkeit

Neben anderen möglichen Modellen wie dem Logit oder Probit Modell, wird das Poisson Modell häufig in der Versicherungspraxis zur Modellierung der Schadenfrequenz verwendet, vgl [6].

3.2.1.1 Poisson Modell

Bezeichne mit $N(t)$ die Schadenanzahl eines individuellen Versicherungsvertrages in einem Zeitraum $[0,t]$ mit $N(0) = 0$. Der stochastische Prozess $\{N(t), t \geq 0\}$ wird Schadenprozess genannt. Unter den Annahmen 3.1, 3.2 und 3.3 sowie der Annahme, dass Schäden nicht clustern, handelt es sich bei dem Schadenprozess um einen Poisson Prozess, siehe [20]. Somit folgt die Anzahl der Schäden für einen beliebigen Zeitraum einer Poissonverteilung und aufgrund der vorausgesetzten Vertragsunabhängigkeit gilt dies ebenfalls auf dem agreggierten Level für alle Verträge in einer Tarifzelle.

Sei nun X_i die Anzahl der Schäden in einer Tarifzelle i, i=1,...,n, mit Laufzeit ω_i und μ_i der Erwartungswert für $\omega_i = 1$, also für eine einjährige Vertragsdauer. Folglich ist X_i poissonverteilt mit Erwartungswert $E(X_i) = \omega_i \mu_i$ und Dichtefunktion:

$$f_{X_i}(x_i, \mu_i) = e^{-\omega_i \mu_i} \frac{(\omega_i \mu_i)^{x_i}}{x_i!}$$

mit $x_i = 0, 1, 2, ...$

Für die Dichtefunktion der Schadenhäufigkeit Y_i, welche als Quotient aus der Anzahl der Schäden und der Vertragslaufzeit definiert ist, gilt

$$\begin{aligned} f_{Y_i}(y_i, \mu_i) &= P(Y_i = y_i) \\ &= P\left(\frac{X_i}{\omega_i} = y_i\right) \\ &= P(X_i = \omega_i y_i) \\ &= e^{-\omega_i \mu_i} \frac{(\omega_i \mu_i)^{\omega_i \mu_i}}{(\omega_i \mu_i)!} \\ &= \exp[\omega_i [y_i \log(\mu_i) - \mu_i] + c(y_i, \omega_i)] \end{aligned}$$

mit

$$c(y_i, \omega_i) = \omega_i y_i \log(\omega_i) - \log(\omega_i y_i!).$$

Als Likelihoodfunktion erhält man folglich:

$$l(\boldsymbol{\mu}, \mathbf{y}) = \sum_i \omega_i [y_i \log(\mu_i) - \mu_i]$$

Damit lässt sich die Devianz berechnen:

$$\begin{aligned} D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) &= 2 \sum_i \omega_i [y_i \log(\mu_i) - \mu_i] - 2 \sum_i \omega_i [y_i \log(\hat{\mu}_i) - \hat{\mu}_i] \\ &= 2 \sum_i [\omega_i y_i \log \left(\frac{y_i}{\hat{\mu}_i} \right) + \omega_i (\hat{\mu}_i - y_i)]. \end{aligned}$$

Der Parameter der Poissonverteilung einer Zelle i, λ_i, wird mit dem linearen Prädiktor verbunden mithilfe der kanonischen Linkfunktion, welche für der Poissonverteilung der Logarithmus ist. Es gilt

$$\lambda_i = \prod_{j=1}^{r} \exp(x_{ij} \beta_j) = \exp(\mathbf{x}_i^t \boldsymbol{\beta}),$$

also

$$\log(\lambda_i) = \mathbf{x}_i^t \boldsymbol{\beta}.$$

Daraus folgt für die erwartete Schadenanzahl einer Tarifierungszelle i mit n_i Risiken:

$$\mu_i = n_i \lambda_i = n_i \exp(\mathbf{x}_i^t \boldsymbol{\beta}).$$

In dem soeben betrachteten Modell verhält sich die Varianz parallel zum Mittelwert. Diese Annahme ist jedoch in vielen Fällen zu restriktiv, da der Varianzschätzer in dem Poissonmodell zu optimistisch bestimmt werden würde. *Overdispersion* liegt vor, wenn die Varianz der Beobachtungen in einer Tarifzelle größer ist als die Varianz der Poissonverteilung, bzw. allgemein wenn gilt $Var(Y) > E(Y)$, siehe [2].
Mögliche Ursachen sind u.a. die Nichtberücksichtigung von Effekten erklärender Variablen, eine zufällige Variation zwischen Kunden und versicherten Objekten, eine falsch gewählte Link Funktion oder speziell im Poissonmodell die Tatsache, dass die zugrundeliegende Beobachtungseinheit wie Zeit oder Volumen nicht fix sondern variabel ist.

Obwohl Overdispersion keinen Einfluss auf den $\boldsymbol{\beta}$-Parameter hat, ist es wichtig sie zu berücksichtigen, wenn man Konfidenzintervalle berechnen möchte. Täte man dies nicht, wären als Ergebnis die Konfidenzintervalle zu knapp.

Wir wollen nun eine mögliche Modellierung von Overdispersion vorstellen.

Sei $\Lambda_1, \Lambda_2, ...$ eine Folge von unabhängigen Zufallsvarialben mit Verteilungen auf $(0, \infty)$. Seien $X_1, X_2, ...$ unabhängige Zufallsvariablen, sodass für $\Lambda_i = \lambda_i$ die Zufallsvariablen X_i poissonverteilt sind mit Erwartungswert λ_i.
Aus den Eigenschaften der Poissonverteilung folgt

$$E(X_i \mid \Lambda_i) = \Lambda_i$$

sowie

$$Var(X_i \mid \Lambda_i) = \Lambda_i$$

Daraus folgt

$$E(X_i) = E[E(X_i \mid \Lambda_i)] = E(\Lambda_i)$$

und für die Varianz

$$\begin{aligned} Var(X_i) &= E[Var(X_i \mid \Lambda_i)] + Var[E(X_i \mid \Lambda_i)] \\ &= E(\Lambda_i) + Var(\Lambda_i) \end{aligned}$$

Nimmt man eine Erwartungswert-Varianz-Beziehung für die $\{\Lambda_i\}$ an, so indiziert dies ebenfalls eine für die $\{X_i\}$.Eine mögliche Annahme über die Beziehung zwischen $E(\Lambda_i)$ und $Var(\Lambda_i)$ lautet:

$$Var(\Lambda_i) = \nu E(\Lambda_i)$$

für ein $\nu > 0$.

Bezeichnet X_i die Anzahl der Schäden mit Erwartungswert $\omega_i \mu_i$, gilt

$$Var(X_i) = (1 + \nu)\omega_i \mu_i$$

Umgeformt zur Schadenfrequenz $Y_i = \frac{X_i}{\omega_i}$, gilt für die Varianz

$$\begin{aligned} Var(Y_i) &= \frac{Var(X_i)}{\omega_i^2} \\ &= \frac{(1 + \nu)\mu_i}{\omega_i} \end{aligned}$$

Setzt man nun $\phi = 1 + \nu$, erhält man

$$Var(Y_i) = \frac{\phi \mu_i}{\omega_i},$$

was der Varianz einer Poissonverteilung mit zusätzlichem Dispersionsparameter ϕ entspricht.

Um eine Verteilung für X_i zu spezifizieren, müssen Verteilungsannahmen für Λ_i getroffen werden. Eine populäre Wahl ist die Gammaverteilung. Sind die Λ_i also gammaverteilt und ist $Var(\Lambda_i) = \nu E(\Lambda_i)$ erfüllt, so gilt für die Dichtefunktion von X_i:

$$P(X_i) = \frac{\Gamma(\frac{\omega_i \mu_i}{\nu + x_i})}{\Gamma(\frac{\omega_i \mu_i}{\nu}) x_i!} \left(\frac{1}{1+\nu}\right)^{\frac{\omega_i \mu_i}{\nu}} \left(\frac{\nu}{1+\nu}\right)^{x_i}$$

Dies ist ein Beispiel der Negativ Binomial Verteilung, welche ebenfalls zur Exponentialfamilie gehört. Sie erlaubt im Gegensatz zur Poissonverteilung höhere Flexibiliät bei der Modellierung der Varianz und sie besitzt zwei Parameter.

3.2.2 Schadenhöhe

In der Versicherungspraxis exisiert kein dominierendes Modell zur Analyse der Schadenhöhe, wie das Poissonmodell es für die Schadenfrequenz ist. Eine wichtige Schadensummenverteilung stellt jedoch die Gammaverteilung dar, die ebenfalls zur Familie der Exponentialverteilungen gehört. Sie erfüllt die Bedingung an eine Verteilung, die im Falle der Schadenhöhe positiv sowie rechtsschief sein sollte.

3.2.2.1 Gamma Modell

Wir konzentrieren uns nun auf eine Tarifzelle und lassen deswegen den Index i zunächst weg. Es sollen folgende Notationen gelten:
Die Anzahl der Schäden soll durch ω beschrieben werden, X bezeichne die totalen Schadenkosten in der Tarifzelle und die Schadenschwere wird dargestellt durch

$$Y = \frac{X}{\omega}.$$

Die Kosten eines individuellen Schadens seien also gammaverteilt.
Eine mögliche Parametrisierung ist die mit einem Indexparameter α, einem Skalenparameter β sowie der Dichtefunktion $G(\alpha, \beta)$ gegeben durch

$$f(X) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \ , \ x > 0$$

mit Ewartungswert $\frac{\alpha}{\beta}$ und Varianz $\frac{\alpha}{\beta^2}$.
X ist die Summe von ω unabhängigen gammaverteilten Zufallsvariablen, d.h. $X \sim G(\omega\alpha, \beta)$. Demnach ist die Schadenhöhe $Y \sim G(\omega\alpha, \omega\beta)$ mit Erwartungswert $\frac{\alpha}{\beta}$ und Dichtefunktion

$$f_Y(y) = \omega f_X(\omega y) = \frac{(\omega\beta)^{\omega\alpha}}{\Gamma(\omega\alpha)} y^{\omega\alpha-1} e^{-\omega\beta y} \ , \ y > 0 \tag{9}$$

Nun transformieren wir (10) in die EDM-Form. Dazu reparametrisieren wir $y = \frac{\alpha}{\beta}$ und $\phi = \frac{1}{\alpha}$ und erhalten einen neuen Parameterraum mit $\mu, \phi > 0$. Für die Dichtefunktion ergibt sich

$$\begin{aligned} f_Y(y) &= f_Y(y, \mu, \phi) \\ &= \frac{1}{\Gamma(\frac{\omega}{\phi})} \left(\frac{\omega}{\mu\phi}\right)^{\frac{\omega}{\phi}} y^{(\frac{\omega}{\phi}-1)} e^{-\frac{\omega y}{\mu\phi}} \\ &= \exp\left\{ \frac{-\frac{y}{\mu} - \log(\mu)}{\frac{\phi}{\omega}} + c(y, \phi, \omega) \right\} \end{aligned} \tag{10}$$

für $y > 0$ und

$$c(y, \phi, \omega) = \log\left(\frac{\omega y}{\phi}\right) \frac{\omega}{\phi} - \log(y) - \log\Gamma\left(\frac{\omega}{\phi}\right)$$

Daraus folgt für den Erwartungswert und die Varianz:

$$\begin{aligned} E(Y) &= \frac{\frac{\omega}{\phi}}{\frac{\omega}{\mu\phi}} = \mu \\ Var(Y) &= \frac{\frac{\omega}{\phi}}{\frac{\omega^2}{\mu^2\phi^2}} = \frac{\mu^2\phi}{\omega} \end{aligned}$$

Möchten wir nun zeigen, dass die Gammaverteilung ein EDM handelt, genügt es einen Parameterwechsel in der Art $\theta = -\frac{1}{\mu}$ zu vollziehen. Dieser neue Parameter nimmt Werte in der offenen Menge $\{\theta < 0\}$ an. Es gilt dann für die Dichtefunktion mit wieder eingeführtem Index i:

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

Dies ist ein EDM mit $b(\theta_i) = -\log(-\theta_i)$.

Als Devianz ergibt sich:

$$D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = 2 \sum_i \omega_i \left(\frac{y_i}{\hat{\mu}_i} - 1 - \log\left(\frac{y_i}{\hat{\mu}_i}\right) \right)$$

Die Log Likelihood Funktion für das Gamma Modell mit einer logarithmischen Linkfunktion in Abhängigkeit von $\boldsymbol{\beta}$ lautet

$$l(\boldsymbol{\beta}) = -\frac{1}{\phi}\sum_i \omega_i[y_i \exp(-\mathbf{x}_i^t\boldsymbol{\beta}) + \mathbf{x}_i^t\boldsymbol{\beta}].$$

Dabei wurden die von $\boldsymbol{\beta}$ unabhängigen Terme weggelassen.

Im nächsten Kapitel führen wir anhand eines anonymisierten Datensatzes eine empirische Analyse mit dem Ziel der konkreten Prämienberechnung durch.

4 Empirische Analyse

4.1 Datengrundlage

Der in dieser Diplomarbeit verwendete Datensatz wurde von einem großen deutschen Versicherungsunternehmen zur Verfügung gestellt. Er beinhaltet aggregierte Daten aus der Kraftfahrthaftpflichtversicherung aus den Jahren 2005 bis 2007, welche bereits von Großschäden bereinigt wurden. Insgesamt sind 1.879.051 Beobachtungen vorhanden, die von 13 Variablen beschrieben werden.

Zu Beginn sollen die Tarifmerkmale erläutert werden, welche in dem Datensatz betrachtet werden.

4.1.1 Erläuterung der Tarifmerkmale

Die Tarifierungsmerkmale werden in der nachfolgenden Tabelle vorgestellt.

Merkmal	Anzahl Ausprägungen
Tarifgruppe	3
DA	2
Regionalklasse	12
Kilometerklasse	8
Typklasse	16
Schadenfreiheitsklasse	29
Fahrzeugalter	7
EFPA	7
Fahrzeugabstellort	5
Wohneigentum	7

Tabelle 6: Tarifierungsmerkmale

Die *Tarifgruppe* unterteilt die Versicherungsnehmer in Agrarier (A), Beamte (B) sowie in die Gruppe der Versicherten mit Normaltarif (N).

Unter dem Merkmal DA versteht man die *Deckungsart*, welche 2 Ausprägungen besitzt. Diese sind zum Einen die gesetzliche Mindestdeckung und zum Anderen ein Wert von 100 Mio. Euro.

In der *Kilometerklasse* wird beschrieben, wieviele Kilometer der Versicherungsnehmer im Durchschnitt pro Jahr fährt. Die Unterteilung der Klassen ist in Tabelle 5

dargestellt.

Kilometerklasse	Gefahrene Kilometer
1	bis 1.000
7	1.000 bis 7.000
10	7.000 bis 10.000
13	10.000 bis 13.000
16	13.000 bis 16.000
21	16.000 bis 21.000
26	21.000 bis 26.000
31	26.000 bis 31.000

Tabelle 7: Kilometerklasse

Die *Typklasse* stuft die auf dem Markt befindlichen Fahrzeuge in 16 verschiedene Gruppen ein. Dabei erfolgt die Einstufung auf Basis der realen Schäden und deren Kosten, die die Versicherer in den drei Vorjahren für die jeweiligen Fahrzeugtypen aufgewendet haben. Das Typklassenverzeichnis der deutschen Autoversicherer enthielt im Jahr 2008 mehr als 19.000 unterschiedliche Fahrzeuge, die anhand der vom Kraftfahrt-Bundesamt vergebenen Hersteller- und Typschlüsselnummer (HSN/TSN), eindeutig identifiziert werden können, siehe [11].

Die *Schadenfreiheitsklasse* gibt Aufschluss über die individuelle Schadenerfahrung eines Versicherungsnehmers. Hat dieser in einer Versicherungsperiode keinen Schaden verursacht, führt dies zu einer Einstufung in eine höhere Schadenfreiheitsklasse, was mit einem Prämienrabatt verbunden ist. Andersrum führt ein eingetretener Schaden zu einer Herabstufung und es muss ein Prämienzuschlag entrichtet werden. Man kann grob sagen, dass die Schadenfreiheitsklasse angibt, wieviele Jahre ein Fahrer unfallfrei ist. Als Beispiel würde eine Schadenfreiheitsklasse von 10 bedeuten, dass ein Fahrer während der letzten 10 Jahre keinen Schaden verursacht hat. Diese muss allerdings nicht zutreffend sein, da man bei einem Schaden nur einige Klassen herabgestuft wird und nicht etwa den gesamten Schadenfreiheitsrabatt verliert, siehe [6], S.85. Beginnt ein Vertrag ohne Übernahme des Schadenverlaufs des Versicherten, beginnt dieser in der Schadenfreiheitsklasse 0. Die einzelnen Rückstufungen im Schadenfall sind tabellarisch im Anhang aufgeführt.

Ein weiteres Merkmal ist die *Regionalklasse*. In den unterschiedlichen Regionalklassen erfolgt eine Zusammenfassung der Zulassungsbezirke, bei denen ein

ähnlicher Schadenverlauf zu erkennen ist. Dabei spielen das Fahrverhalten der Autofahrer, die vorhandenen Straßenverhältnisse, die Zahl der zugelassenen Fahrzeuge, sowie die Bestandszusammensetzung eine Rolle, vgl. [11].

Das *Fahrzeugalter* ist in 7 Altersklassen unterteilt. Bei Neufahrzeugen wird zusätzlich unterschieden, ob das Zulassungsdatum mit dem Versicherungsbeginn übereinstimmt, sodass eine Merkmalsausprägung von -1 vorliegen würde, ansonsten gilt für Neuwagen die Ausprägung 0.

Bei dem Merkmal *EFPA*, welches als Abkürzung für Einzelfahrer - Partner gebraucht wird, stellt sich die Frage, wer der Fahrzeugführer ist und ob es noch zusätzliche Nutzer gibt. Es erfolgt darüber hinaus eine Unterteilung nach dem Geschlecht und ob das Fahrzeug privat oder von Firmen genutzt wird.

Weiter gibt es die Merkmale *Fahrzeugabstellort* sowie *Wohneigentum*, bei welchem gefragt wird, ob Wohneigentum vorhanden ist und ob eine Versicherung in dem hier betrachteten Versicherungsunternehmen abgeschlossen wurde.

Der Datensatz umfasst neben diesen Merkmalen die Anzahl von Schäden, den Aufwand sowie die Jahreseinheiten (JE). Die Jahreseinheiten sind definiert als Verweildauer des Versicherungsvertrages in Tagen geteilt durch 360, siehe ?

Wir wollen zunächst die Schadenfrequenz und anschließend die Schadenhöhe analysieren um den Schadenbedarf schließlich als Produkt darstellen zu können. Aufgrund der enormen Beanspruchung der Statistiksoftware, musste eine Selektion der Daten durchgeführt werden. Hierzu wurde aus dem kompletten Datensatz zufällig ein Teil ausgewählt. Mit den Befehlen

```
>randomize<-sample(c(1:1879051),500000,replace=FALSE)
>sort(randomize)
>agg<-daten[sort(randomize), ]
```

erhalten wir einen verkleinerten Datensatz namens *agg*, bestehend aus 500.000 aggregierten Daten, welchen wir im Folgenden analysieren werden.

4.2 Schadenfrequenzanalyse

4.2.1 Poisson Modell

Wie im ersten Teil bereits beschrieben wurde, verwenden wir für die Schadenfrequenz das Poissonmodell mit einer logarithmischen Linkfunktion. Zunächst werden alle Merkmale in das Modell miteinbezogen, wir werden sie erst im weiteren Verlauf auf ihre Signifikanz untersuchen. Interaktionseffekte werden nicht berücksichtigt. Der Befehl, mit dem in **R** ein verallgemeinertes lineares Modell für unseren Teildatensatz *agg* erzeugt wird, lautet:

>glm(SCHADEN/JE ∼ TG + DAf + TYKLf + RKLf + KMKLf
+ SFR + Fzgalter + EFPA + FZGAOS + WOHNEIG.ART.SCHL,
family = poisson(link = log),weights=JE)

Die Tarifmerkmale mussten hierzu zunächst in Faktoren umbenannt werden. Es findet eine Gewichtung mit den Jahreseinheiten statt, sodass wir die jährliche Schadenanzahl untersuchen können.
Mit *summary* wird das Modell anschaulich dargestellt, unter Anderem sind die geschätzten Regressionskoeffizienten, die Standardfehler sowie die z-Werte enthalten. Die letzteren werden als Quotient aus dem Koeffizienten und dem dazugehörigen Standardfehler gebildet. Die folgende Tabelle zeigt einen Auszug des Modells.

```
Call:
glm(formula = SCHADEN/JE ~ TG + DAf + TYKLf + RKLf + KMKLf +
    SFR + Fzgalter + EFPA + FZGAOS + WOHNEIG.ART.SCHL,
    family = poisson(link = log), weights = JE)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7483  -0.3563  -0.2747  -0.1910   5.2427

Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -2.297649   0.213205 -10.777  < 2e-16
TGB                              0.002251   0.026178   0.086 0.931479
TGN                              0.084779   0.023813   3.560 0.000371
DAf4                             0.367150   0.189181   1.941 0.052291
TYKLf11                          0.419793   0.327320   1.283 0.199662
TYKLf12                          0.692157   0.188855   3.665 0.000247
TYKLf13                          0.710565   0.179529   3.958 7.56e-05
```

```
TYKLf14                    0.838787   0.177194   4.734 2.20e-06
TYKLf15                    0.887082   0.177182   5.007 5.54e-07
TYKLf16                    0.958313   0.177141   5.410 6.31e-08
TYKLf17                    1.000531   0.176979   5.653 1.57e-08
TYKLf18                    1.013181   0.177219   5.717 1.08e-08
TYKLf19                    1.077074   0.177417   6.071 1.27e-09
TYKLf20                    1.179719   0.178637   6.604 4.00e-11
TYKLf21                    1.247149   0.180155   6.923 4.43e-12
TYKLf22                    1.266511   0.183788   6.891 5.53e-12
TYKLf23                    1.370891   0.182670   7.505 6.15e-14
TYKLf24                    1.458122   0.192243   7.585 3.33e-14
TYKLf25                    1.730281   0.224114   7.721 1.16e-14
RKLf1                      0.032092   0.033282   0.964 0.334920
RKLf2                      0.060902   0.033466   1.820 0.068789
RKLf3                      0.091562   0.036094   2.537 0.011187
RKLf4                      0.082383   0.035052   2.350 0.018756
RKLf5                      0.121845   0.034161   3.567 0.000361
RKLf6                      0.142705   0.033286   4.287 1.81e-05
RKLf7                      0.181525   0.033647   5.395 6.85e-08
RKLf8                      0.238794   0.037122   6.433 1.25e-10
RKLf9                      0.273322   0.042054   6.499 8.07e-11
RKLf10                     0.331481   0.060632   5.467 4.57e-08
RKLf11                     0.464177   0.062122   7.472 7.90e-14
KMKLf7                     0.092966   0.020040   4.639 3.50e-06
KMKLf10                    0.147658   0.021487   6.872 6.33e-12
KMKLf13                    0.166879   0.022937   7.276 3.45e-13
KMKLf16                    0.262025   0.022676  11.555  < 2e-16
KMKLf21                    0.377134   0.034660  10.881  < 2e-16
KMKLf26                    0.499446   0.035956  13.891  < 2e-16
KMKLf31                    0.678727   0.039763  17.069  < 2e-16

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 159463  on 499999  degrees of freedom
Residual deviance: 152894  on 499913  degrees of freedom
AIC: Inf
```

Als erstes wird das benutzte Modell nochmal angegeben und dann Eigenschaften der Devianzresiduen wie Minimum, Maximum, Median, und die Quartile der Daten angegeben. In der Tabelle befinden sich dann die geschätzten Werte des Parametervektors $\hat{\boldsymbol{\beta}}$. Im Anschluss werden der angenommene Wert für den Di-

spersionsparameter sowie die Null- und Residuendevianz mit zugehöriger Anzahl an Freiheitsgraden und der AIC-Wert genannt.

Anhand der p-Werte kann man bereits einen ersten Überblick über die Signifikanz von Merkmalen erhalten. Wie bereits erläutert, wird die Hypothese $H_j : \beta_j = 0$ gegen die Alternative $K_j : \beta_j \neq 0$ getestet. Ein kleiner p-Wert lässt auf Signifikanz des Merkmals schließen. Größere p-Werte indizieren, dass das Merkmal keinen großen Einfluss auf die Schadenfrequenz hat.

Dies können wir mithilfe des im 2. Kapitel vorgestellten Partial Deviance Tests überprüfen. Wir testen beispielhaft die Merkmale Deckungsart und die Schadenfreiheitsklasse auf ihre Signifikanz. Für beide Merkmale wird ein verallgemeinertes lineares Modell sowohl mit als auch ohne dieses Merkmal erzeugt. Ist dies geschehen, berechnet man die Differenz der Devianzen von beiden Modellen. Wie im letzten Kapitel beschrieben, folgt die Teststatistik asymptotisch einer χ^2-Verteilung unter der Nullhypothese, dass das einfachere Modell gilt. Man erhält die Freiheitsgrade als Differenz der Freiheitsgrade des kompletten Modells zu seiner Vereinfachung.

Für die Deckungsart ergibt sich als Devianz für das Modell ohne die Miteinbeziehung des Merkmals ein Wert von 152.900, für das Modell mit DA beträgt der Wert 152.894. Somit entpricht die Differenz der Devianzen dem Wert 1. Entsprechend vergleicht man die Freiheitsgrade und wir erhalten ebenfalls eine Differenz von 1. Ist nun die Differenz der Devianzen größer als der Wert der χ^2 Verteilung mit einer Anzahl an Freiheitsgraden, die der Differenz der Freiheitsgrade aus beiden Modellen entspricht, so wird die Nullhypothese verworfen, welche lautet, dass das Merkmal nicht signifikant ist. In unserem Fall gilt

$$\Delta\text{Devianz} = 6 > 3,84 = \chi^2_{1;0,95},$$

d.h. die Hypothese wird verworfen, das Merkmal DA scheint Einfluss zu haben.

Analog erhalten wir für das Merkmal Schadenfreiheitsklasse ΔDevianz $= 1.606$ und $\Delta\text{df} = 28$. Es gilt

$$\Delta\text{Devianz} = 1.606 > 41,34 = \chi^2_{28;0,95},$$

also wird auch hier die Nullhypothese verworfen, denn das Merkmal Schadenfreiheitsklasse ist hoch signifikant.

In **R** lässt sich die Überprüfung der Merkmale anhand der Devianzanalyse mit dem Befehl *anova()* durchführen. Man kann damit die sequentielle Hinzufügung

jedes einzelnen Merkmals testen, indem man ihr den χ^2-Test zum Test auf Unterschiede zwischen den Modellen als Argument übergibt.
Mit dem Befehl

>anova(glm01, test="Chisq")

erhalten wir den folgenden Output:

```
Analysis of Deviance Table

Model: poisson, link: log

Response: SCHADEN/JE

Terms added sequentially (first to last)


                 Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                            499999     159463
TG                2    244.8    499997     159218 < 2.2e-16 ***
DAf               1      5.8    499996     159212   0.01608 *
TYKLf            15    358.0    499981     158854 < 2.2e-16 ***
RKLf             11    279.6    499970     158575 < 2.2e-16 ***
KMKLf             7    328.6    499963     158246 < 2.2e-16 ***
SFR              28   3415.3    499935     154831 < 2.2e-16 ***
Fzgalter          6    692.3    499929     154139 < 2.2e-16 ***
EFPA              6   1152.8    499923     152986 < 2.2e-16 ***
FZGAOS            4     25.6    499919     152960 3.832e-05 ***
WOHNEIG.ART.SCHL  6     66.0    499913     152894 2.716e-12 ***
---
Signif. codes:  0 Ô***Õ 0.001 Ô**Õ 0.01 Ô*Õ 0.05 Ô.Õ 0.1 Ô Õ 1
```

Hierbei handelt es sich um eine sogenannte Typ I - Analyse. Man beginnt mit dem Nullmodell und fügt schrittweise eine Variable hinzu. Als Resultat erhält man die Verbesserung der Modellanpassung durch die hinzugenommene Variable. Wir können erkennen, dass die Deckungsart den geringsten Einfluss auf die Schadenfrequenz besitzt. Mit einem Wert von 5,8 besitzt sie die kleinste Devianz. Es folgen die Merkmale Fahrzeugabstellort und Wohneigentum. Im Gegensatz dazu, hat die Schadenfreiheitsklasse mit einer Devianz von 3415,3 den größten Einfluss auf die Schadenfrequenz, gefolgt von dem Merkmal EFPA mit einem Wert von 1152,8.

Eine andere Möglichkeit die Merkmale nach ihrer Signifikanz zu beurteilen bietet die Typ III - Analyse. Bei dieser Methode beginnt man mit dem maximalen Modell und entfernt sukzessive ein Merkmal. Anschließend kann man beobachten, wie sich die Modellanpassung durch die Herausnahme des Merkmals verschlechtert.
Erzeugt wird eine solche Analyse mit dem *drop1()* - Befehl. Die Typ III - Analyse sieht in unserem Fall wie folgt aus:

```
Single term deletions

Model:
SCHADEN/JE ~ TG + DAf + TYKLf + RKLf + KMKLf + SFR + Fzgalter +
    EFPA + FZGAOS + WOHNEIG.ART.SCHL
                Df Deviance AIC
<none>               152894 Inf
TG               2   152932 Inf
DAf              1   152897 Inf
TYKLf           15   153377 Inf
RKLf            11   153076 Inf
KMKLf            7   153368 Inf
SFR             28   154487 Inf
Fzgalter         6   153378 Inf
EFPA             6   154054 Inf
FZGAOS           4   152905 Inf
WOHNEIG.ART.SCHL 6   152960 Inf
```

In dieser Tabelle sehen wir, dass das Merkmal Deckungsart bei Eliminiation nur eine sehr geringe Steigerung der Devianz mit sich führt und die Schadenfreiheitsklasse im Gegensatz die größte Wirkung auf die Schadenfrequenz besitzt.
Wir schauen uns nun zu Zwecken der Modelldiagnose die Residuen an.
Dazu bestimmen wir zunächst die Devianzresiduen und plotten sie gegen die angepassten Werte mit dem Befehl

```
>res01<-residuals(glm01,type="deviance")
> plot(predict(glm01),res01,xlab="Fitted Values",ylab="Residuals",
ylim = max(abs(res01)) * c(-1,1))
> abline(h=0,lty=2).
```

In dem Residualplot werden die gefitteten Werte auf der Abszisse und die Residuen auf der Ordinate abgetragen.
In diesem Plot möchte so wenig Struktur und Muster wie möglich sehen. Wenn

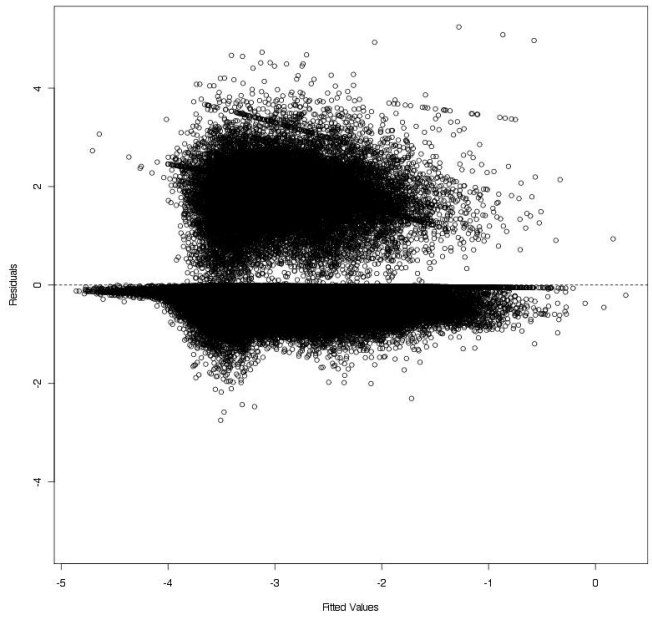

Abbildung 1: Devianzresiduen vs. fitted values, Schadenfrequenz

die Residuen mit wachsenden Werten der prognostizierten Werte systematisch steigen, fallen oder streuen, liegt die Vermutung einer falschen Varianzfunktion und/oder fehlender Kovariablen nahe. Liegen hingegen nur einige Werte außerhalb, können Ausreißer vorhanden sein, was eine Prüfung des Einflusses auf die Parameterschätzung nahe legt, siehe [13].

Mit dem Befehl

>plot(glm01b)

erhalten wir vier verschiedene Grafiken.
Die Grafik oben links zeigt die vorhergesagten Werte dargestellt gegen die Residuen., in dem kein Muster auftreten sollte. Rechts ist ein Normal QQ-Plot abgebildet, in diesem sollte der Verlauf bei normalverteilten Fehlern linear sein. Es fällt auf dass Beobachtung 395.179 weit von den anderen Residuen entfernt liegt. Bei Betrachtung des Datensatzes sehen wir, dass diese Beobachtungsnummer eine Schadenfrequenz von 8 Schäden in 90 Jahreseinheiten aufweist. Dies ist eine ungewöhnlich hohe Anzahl an Schäden und wird hier als Ausreißer dargestellt. Unten links sieht man die Scale Location. Bei diesem Plot handelt es sich um eine

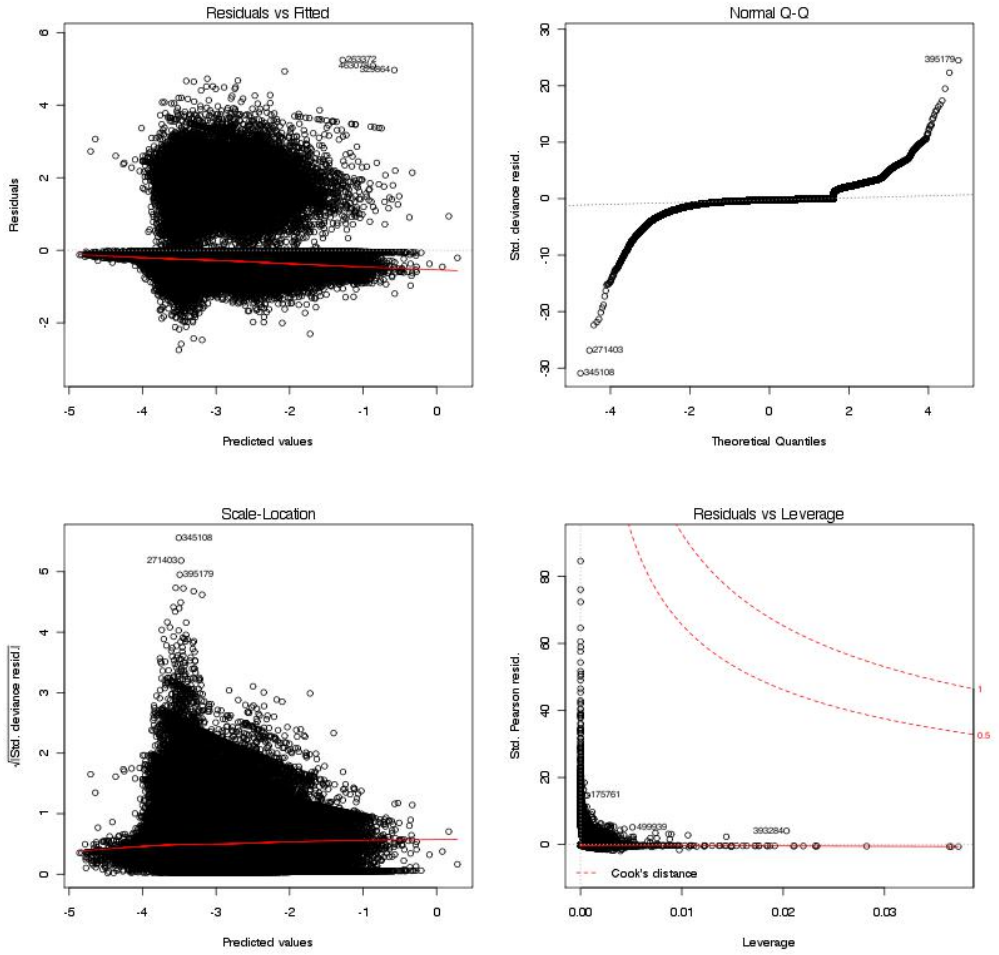

Abbildung 2: Ausgabe plot(glm01)

Wiederholung des ersten Plots, jedoch auf einer anderen Skala. Man betrachtet die Quadratwurzel der standardisierten Devianzresiduen, sodass alle Werte positiv sind. Bei einer systematischen Steigerung der Standardisierten Devianzresiduen mit steigenden vorhergesagten Werten, könnte auf ein Problem, wie z.B. eine mit dem Mittelwert steigende Varianz, geschlossen werden, siehe [16]. In unserem Fall ist dies nicht zu beobachten. Der letzte Plot unten rechts zeigt Standardisierte Pearson Residuen als eine Funktion des Leverage nebst der Cook‘s Distance. Diese versucht Leverage und die Residuen in einem einzelnen Maß zu kombinieren. Die

Cooks's Distance ist definiert als

$$D_i = \frac{(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})^T (\mathbf{X}^T\mathbf{X})(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})}{p\hat{\sigma}^2}.$$

Dabei bezeichnet $\hat{\beta}_{(i)}$ den Schätzer $\boldsymbol{\beta}$ ohne die i-te Beobachtung und p die Anzahl der Parameter. Man kann zeigen, dass

$$D_i \approx \frac{r_i^2}{p}\frac{h_{ii}}{1-h_{ii}}.$$

mit studentisierten Residuen r_i und $h_{ii} = \mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i$ als Maß für Leverage, wobei $\mathbf{x}_i$ für die i-te Zeile von $\mathbf{X}$ steht, vgl. [17]. Ziel dieses Plots ist die Beobachtungen mit einem großen Einfluss auf die Parameterschätzung hervorzuheben. Punkte mit hoher Leverage, sowie solche die am nächsten zur Cook's Distance sind, können z.B. aus dem Modell genommen werden, um zu schauen wie groß ihr Einfluss auf die Schätzer und die Standardabweichungen ist, vgl. [16].
Obwohl unsere Daten annahmenkonform modelliert wurden, können wir systematische Muster erkennen, die gegen eine korrekte Anpassung sprechen würden. Es ist jedoch nicht möglich die genaue Ursache für eventuelle Fehler herauszufinden, sodass die Residuen mit Vorsicht interpretiert werden müssen, siehe [13].

Wir präsentieren als Ergebnis die *summary* des Modells im Anhang. Hätte ein Versicherungsnehmer in jedem Merkmal die erste Ausprägungsklasse, so würde man davon ausgehen, dass er jährlich $\exp(-2,297649) = 0,100495$ Schäden verursacht, was einem Schaden in durchschnittlich ca. 10 Jahren entspricht.

4.3 Schadenintensitätsanalyse

Bei der Analyse der Schadenintensität fällt auf, dass nur eine geringe Anzahl von Verträgen einen Schaden aufweist. In unserem Datensatz wurden in circa $5,5\%$ der Verträge ein oder mehr als ein Schaden beobachtet. Genauer sieht die Anzahl der schwach besetzten Zellen wie folgt aus:

Schadenanzahl pro Zelle	0	1	2	3	4
Anzahl der Zellen	472.615	25.454	1.649	215	40

Tabelle 8: Anzahl der Schäden

Dies erschwert eine Analyse der Schadenhöhe. Desweiteren kommt hinzu, dass die

Schadenhöhe mit vielen Faktoren zusammenwirkt, welche nicht zwangsweise mit den Merkmalen erfasst werden (vgl. [14]). Es muss zusätzlich angemerkt werden, dass unser Datensatz im Vorfeld von Großschäden bereinigt wurde und in unserer Analyse demnach keine Rolle spielt. Wir wollen nun die Schadenhöhe mithilfe des Gammamodells mit logarithmischer Linkfunktion modellieren. Auch hier berücksichtigen wir Interaktionseffekte nicht. Dazu verwenden wir den folgenden Befehl, um ein verallgemeinertes lineares Modell zu erzeugen:

```
glm02<-glm(AUFWAND/SCHADEN~TG+DAf+TYKLf+RKLf+KMKLf+SFR+Fzgalter
+EFPA+FZGAOS+WOHNEIG.ART.SCHL,fam=Gamma(link="log"),
wei=SCHADEN,na.action=na.omit)
```

Anhand der *summary* erhalten wir wieder einen ersten Überblick über die Anpassung. Diese sieht in einem Auszug wie folgt aus:

```
Call:
glm(formula = AUFWAND/SCHADEN ~ TG + DAf + TYKLf + RKLf + KMKLf +
    SFR + Fzgalter + EFPA + FZGAOS + WOHNEIG.ART.SCHL,
    family = Gamma(link = "log"), weights = SCHADEN, na.action = na.omit)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.7531  -1.0871  -0.6360  -0.0333  11.8288

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                      7.9481264  0.6117769  12.992  < 2e-16
TGB                             -0.1871452  0.0748734  -2.499  0.01244
TGN                             -0.2290537  0.0683084  -3.353  0.00080
DAf4                             0.3162965  0.5414939   0.584  0.55915
TYKLf11                          0.7071053  0.9341170   0.757  0.44907
TYKLf12                          0.5908571  0.5381901   1.098  0.27228
TYKLf13                          0.5542859  0.5120465   1.082  0.27904
TYKLf14                          0.5819322  0.5050368   1.152  0.24923
TYKLf15                          0.4545811  0.5047453   0.901  0.36780
TYKLf16                          0.4861572  0.5048148   0.963  0.33554
TYKLf17                          0.5641667  0.5043552   1.119  0.26332
TYKLf18                          0.4503723  0.5050318   0.892  0.37252
TYKLf19                          0.5015073  0.5056857   0.992  0.32133
TYKLf20                          0.6189726  0.5091072   1.216  0.22407
TYKLf21                          0.6044824  0.5136768   1.177  0.23930
TYKLf22                          0.4691809  0.5240439   0.895  0.37063
```

```
TYKLf23                      0.5204955  0.5209845   0.999  0.31777
TYKLf24                      0.3699252  0.5482282   0.675  0.49983
TYKLf25                      0.4451656  0.6400389   0.696  0.48673
RKLf1                       -0.0719374  0.0952580  -0.755  0.45014
RKLf2                       -0.0989034  0.0957709  -1.033  0.30175
RKLf3                       -0.1011257  0.1032423  -0.979  0.32734
RKLf4                       -0.0555186  0.1002670  -0.554  0.57978
RKLf5                        0.0010020  0.0977639   0.010  0.99182
RKLf6                       -0.0223031  0.0952617  -0.234  0.81489
RKLf7                       -0.0565817  0.0962897  -0.588  0.55679
RKLf8                       -0.0005864  0.1063305  -0.006  0.99560
RKLf9                        0.0087106  0.1204301   0.072  0.94234
RKLf10                      -0.0127668  0.1738778  -0.073  0.94147
RKLf11                       0.1579865  0.1777862   0.889  0.37421

(Dispersion parameter for Gamma family taken to be 8.171105)

    Null deviance: 38481  on 27384  degrees of freedom
Residual deviance: 37228  on 27298  degrees of freedom
  (472615 observations deleted due to missingness)
AIC: 548575

Number of Fisher Scoring iterations: 8
```

Für die Entscheidung über die Aufnahme bzw. Herausnahme von Merkmalen gehen wir wie bei der Schadenfrequenzanalyse vor. Zuerst betrachten wir die Devianzanalyse.

```
Analysis of Deviance Table

Model: Gamma, link: log

Response: AUFWAND/SCHADEN

Terms added sequentially (first to last)


               Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                          27384      38481
TG              2    72.21     27382      38408   0.01205
DAf             1     2.83     27381      38406   0.55623
```

```
TYKLf             15   163.60     27366      38242   0.17112
RKLf              11    35.20     27355      38207   0.96005
KMKLf              7   115.14     27348      38092   0.04957
SFR               28   360.41     27320      37731   0.02711
Fzgalter           6    59.13     27314      37672   0.29954
EFPA               6   318.15     27308      37354 7.369e-07
FZGAOS             4    67.37     27304      37287   0.08301
WOHNEIG.ART.SCHL   6    58.14     27298      37228   0.31030
```

Die Devianzanalyse verdeutlicht, dass es bei der Schadenhöhe eine Rolle spielt, wer der Fahrzeugführer ist. Der p-Wert ist sehr klein, und die Devianz ist mit einem Wert von 318,15 der zweitgrößte Wert nach der Devianz der Schadenfreiheitsklasse. Im Gegensatz dazu scheinen die Regionalklasse sowie die Deckungsart keinen erheblichen Einfluss auf die Schadenhöhe zu haben. Im nächsten Schritt betrachten wir die Typ III - Analyse mithilfe des *drop1* - Befehls. Wir erhalten:

```
Single term deletions

Model:
AUFWAND/SCHADEN ~ TG + DAf + TYKLf + RKLf + KMKLf + SFR + Fzgalter +
    EFPA + FZGAOS + WOHNEIG.ART.SCHL
                 Df Deviance    AIC
<none>                37228 548575
TG                2    37324 548582
DAf               1    37232 548573
TYKLf            15    37329 548557
RKLf             11    37283 548559
KMKLf             7    37317 548572
SFR              28    37491 548551
Fzgalter          6    37257 548566
EFPA              6    37547 548602
FZGAOS            4    37294 548575
WOHNEIG.ART.SCHL  6    37287 548570
```

Aus der Typ III - Analyse lässt sich die Erkenntnis ziehen, dass mit einer Elimination des Merkmals Typklasse der Wert des Akaike Kriteriums am meisten gesenkt werden kann. Dies lässt auf eine bessere Modellanpassung ohne die Typklasse schließen. Anders als in der Typ I - Analyse scheint hier die Herausnahme der Schadenfreiheitsklasse ebenfalls zu einer Verbesserung der Anpassung zu führen. Eine Anwendung der *step* - Funktion liefert:

```
> step02<-step(glm02)
Start:  AIC=548574.6
AUFWAND/SCHADEN ~ TG + DAf + TYKLf + RKLf + KMKLf + SFR + Fzgalter +
    EFPA + FZGAOS + WOHNEIG.ART.SCHL

                   Df Deviance    AIC
- SFR              28    37491 548551
- TYKLf            15    37329 548557
- RKLf             11    37283 548559
- Fzgalter          6    37257 548566
- WOHNEIG.ART.SCHL  6    37287 548570
- KMKLf             7    37317 548572
- DAf               1    37232 548573
<none>                   37228 548575
- FZGAOS            4    37294 548575
- TG                2    37324 548582
- EFPA              6    37547 548602

Step:  AIC=548766.4
AUFWAND/SCHADEN ~ TG + DAf + TYKLf + RKLf + KMKLf + Fzgalter +
    EFPA + FZGAOS + WOHNEIG.ART.SCHL
```

Nun schauen wir uns die Residuen für das Modell mit allen Merkmalen an.

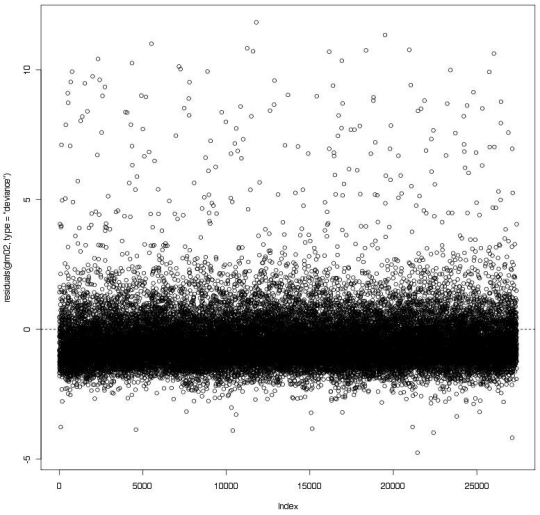

Abbildung 3: Devianzresiduen, Schadenhöhe

Die Residuen werden gegen den Indexwert geplottet, das bedeutet gegen die Beobachtungsnummer. Man sieht, dass sie hauptsächlich um den Wert 0 streuen, es jedoch auch Ausreißer nach oben und vereinzelt nach unten gibt. Würde die Variabiliät der Devianzresiduen steigen, so würde dies auf eine unangemessen Verteilung deuten. Wir können bis auf Ausreißer nach oben, welche für Beobachtungen mit hohen Schadenwerten deuten, kein Muster erkennen.
Die zusammenfassende Darstellung von den vier verschiedenen Residuen Plots, wie wir sie bei der Schadenfrequenzanalyse bereits gesehen haben, sieht wie folgt aus:

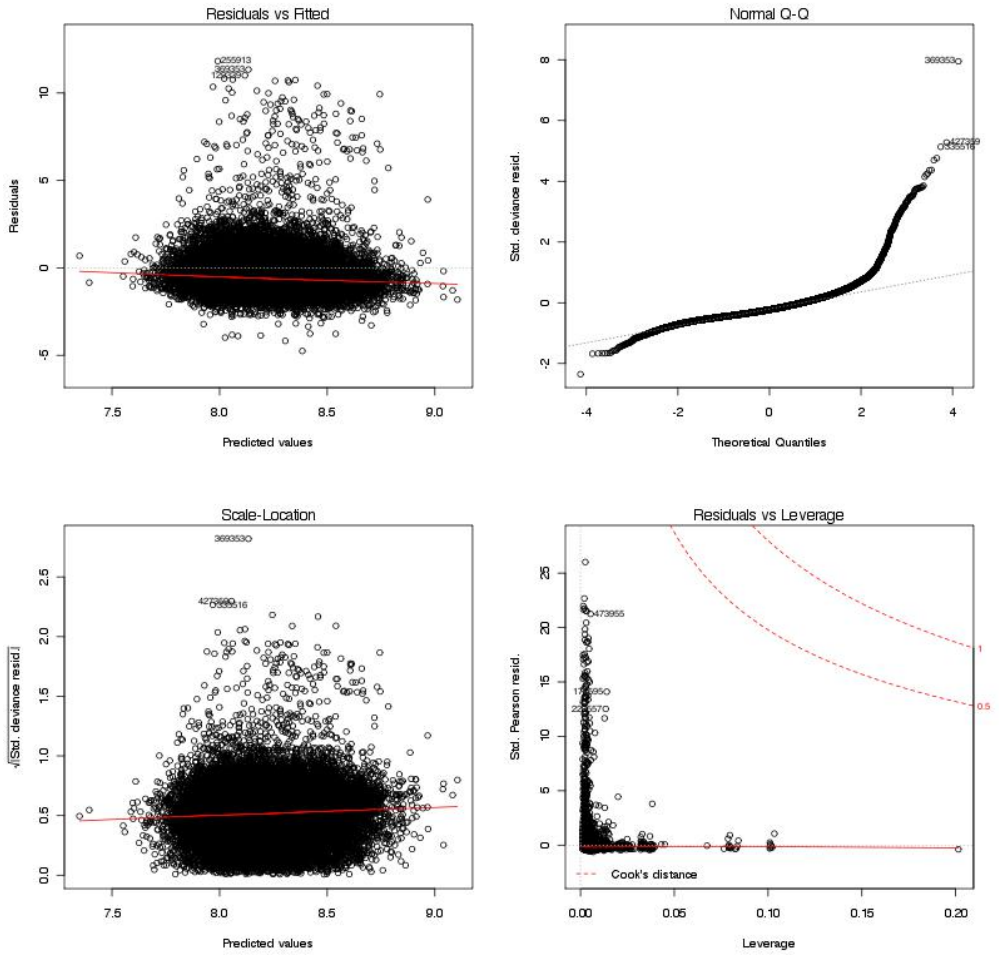

Abbildung 4: Zusammenfassung Residuen Schadenhöhe

Insgesamt ist die Anpassung des Modells für die Schadenhöhe sehr gut. Sowohl der Plot für die vorhergesagten Werte als auch der Plot für die Scale Location ist frei

von systematischen Mustern. In dem Normal QQ-Plot fallen einige Beobachtungen aus dem Rahmen, vor allem das Residuum mit Beobachtungsnummer 369.353 ist weit entfernt von den übrigen Residuen. Betrachten wir unseren Datensatz, so stellen wir fest, dass diese Beobachtung einen sehr hohen Schadenwert in Höhe von 273.600 Euro für einen einzigen Schaden besitzt. Obwohl der Datensatz von Größschäden bereinigt wurde, gibt es dennoch Beobachtungen, die knapp unter der Kupierungsgrenze liegen und sich als Ausreißer in Residualplots zeigen. Dies zeigt sich auch letzten Plot der Residuen gegen Leverage. Beobachtung 473.955 besitzt eine Schadenhöhe von 253.364 Euro für insgesamt vier verschiedene Schadenfälle. Trotzdem bleibt die Anpassung gut und wir präsentieren die abschließende *Summary* des Verallgemeinerten linearen Modells für die Schadenhöhe Anhang.

4.4 Schadenbedarfsanalyse

Den Schadenbedarf wollen wir indirekt als Produkt von Schadenfrequenz und Schadenhöhe darstellen. Da es sich bei den Koeffizienten um Exponenten der Exponentialfunktion handelt, werden sie summiert, sodass wir für den Schadenbedarf die folgenden Tabelle erhalten. Mit dem Befehl

>summary(glm01)$coefficients[,1] + summary(glm02)$coefficients[,1]

erhalten wir:

(Intercept)	5.650477148
TGB	-0.184894299
TGN	-0.144274794
DAf4	0.683446626
TYKLf11	1.126898024
TYKLf12	1.283013806
TYKLf13	1.264851122
TYKLf14	1.420718968
TYKLf15	1.341663165
TYKLf16	1.444470505
TYKLf17	1.564697749
TYKLf18	1.463553464
TYKLf19	1.578581062
TYKLf20	1.798691453
TYKLf21	1.851631112
TYKLf22	1.735692282
TYKLf23	1.891386182
TYKLf24	1.828047166
TYKLf25	2.175446682
RKLf1	-0.039845155
RKLf2	-0.038001101
RKLf3	-0.009563258
RKLf4	0.026864582
RKLf5	0.122847201
RKLf6	0.120402266
RKLf7	0.124942971
RKLf8	0.238207819
RKLf9	0.282032366
RKLf10	0.318714598
RKLf11	0.622163986
KMKLf7	0.063462546

KMKLf10	0.161188231
KMKLf13	0.117067460
KMKLf16	0.185651516
KMKLf21	0.437389261
KMKLf26	0.705745405
KMKLf31	0.656872937
SFR1	-0.708215993
SFR1/2	-0.364980908
SFR10	-1.467876798
SFR11	-1.394854920
SFR12	-1.660530897
SFR13	-1.517249167
SFR14	-1.656464157
SFR15	-1.642567903
SFR16	-1.615239532
SFR17	-1.498850487
SFR18	-1.394635143
SFR19	-1.627832091
SFR2	-0.871031784
SFR20	-1.720751972
SFR21	-1.904618401
SFR22	-1.605061587
SFR23	-1.678505043
SFR24	-1.707079554
SFR25	-1.733958552
SFR3	-1.039733097
SFR4	-1.203328693
SFR5	-1.366051124
SFR6	-1.306425403
SFR7	-1.364003214
SFR8	-1.372895156
SFR9	-1.374126254
SFRM	-0.414706399
SFRS	-0.238631140
Fzgalter0	-0.055900006
Fzgalter1	-0.049537984
Fzgalter3	0.033591497
Fzgalter6	0.188564956
Fzgalter10	0.410214679
Fzgalter20	0.565001648

EFPAEF maennl. VN	-0.657848140
EFPAEF weibl. ^= VN	-0.276920692
EFPAEF weibl. VN	-0.860660905
EFPAkein EF/PA	-0.249895165
EFPAPartner	-1.001849642
EFPAPartner Firmen	-1.099141433
FZGAOSE-/D-Garage	0.039669359
FZGAOSGrundstueck	0.047618066
FZGAOSsonst. Garagen/Parkhaus	-0.163310483
FZGAOSStrassenrand	-0.049577236
WOHNEIG.ART.SCHL1-2-FH, vers.	0.072590256
WOHNEIG.ART.SCHLETW	0.192694371
WOHNEIG.ART.SCHLkeine Angaben	0.072001055
WOHNEIG.ART.SCHLMFH	0.059185410
WOHNEIG.ART.SCHLMFH, vers.	0.551508877
WOHNEIG.ART.SCHLnein	0.119438192

Um nun eine Prämienberechnung durchführen zu können, muss für jedes Merkmal eine Normierung stattfinden, sodass für jede Merkmalsausprägung ein Indexwert zur Verfügung steht. Dieser gibt an, inwiefern sich die Basisprämie im Bezug auf eine Merkmalsausprägung verändert. Der Indexwert kann sowohl zu einer Verbesserung als auch zu einer Verschlechterung der Beitragshöhe führen. Die Normierung bewirkt, dass sich Ab- und Zuschläge im Mittel ausgleichen.
Zunächst wird der Prozentanteil der Versicherungsverträge jeder Merkmalsausprägung ermitelt, hierbei ist zu beachten, dass eine Gewichtung mit den Jahreseinheiten erfolgen muss. Anschließend wird der Schätzwert (Estimate) aus dem Produkt der Schadenfrequenz-und Schadenhöheanalyse exponenziert und im nächsten Schritt mit dem Prozentwert multipliziert. Den normierten Wert für eine Merkmalsausprägung i aus der Menge aller Ausprägungen {1,...,n} dieses Merkmals erhalten wir durch

$$Index_i = \frac{exp(Estimate_i)}{\sum_{j=1}^{n} Anteil_j \times exp(Estimate_j)}.$$

Der normierte Wert für KMKL2 ergibt sich zum Beispiel durch $1,0656 : 1,1616 = 0,9174$. Geht man auf diese Weise für alle anderen Merkmale vor, so erhält man als Ergebnis die folgenden Werte für die Merkmale, wobei die normierten Werte die Indexwerte darstellen.

Level	Anteil	Estimate	exp(Estimate)	Anteil x exp(Estimate)	Normierter Wert
KMKL1	14,1%	0,0000	1,0000	0,1408	0,8609
KMKL7	28,7%	0,0635	1,0656	0,3061	0,9174
KMKL10	19,8%	0,1612	1,1749	0,2329	1,0115
KMKL13	14,8%	0,1171	1,1242	0,1666	0,9678
KMKL16	15,2%	0,1857	1,2041	0,1825	1,0366
KMKL21	3,2%	0,4374	1,5487	0,0500	1,3332
KMKL26	2,6%	0,7057	2,0253	0,0516	1,7435
KMKL31	1,6%	0,6569	1,9288	0,0311	1,6605
				1,1616	

Tabelle 9: Indizes Kilometerklasse

Level	Anteil	Estimate	exp(Estimate)	Anteil x exp(Estimate)	Normierter Wert
TGA	8,2%	0,0000	1,0000	0,0815	1,1506
TGB	22,1%	-0,1849	0,8312	0,1835	0,9564
TGN	69,8%	-0,1443	0,8656	0,6041	0,9960
				0,8691	

Tabelle 10: Indizes Tarifgruppe

Level	Anteil	Estimate	exp(Estimate)	Anteil x exp(Estimate)	Normierter Wert
DA1	99,9%	0,0000	1,0000	0,9994	0,9994
DA4	0,1%	0,6835	1,9808	0,0012	1,9796
				1,0006	

Tabelle 11: Indizes DA

Level	Anteil	Estimate	exp(Estimate)	Anteil x exp(Estimate)	Normierter Wert
TYKL10	0,25%	0,0000	1,0000	0,0025	0,2212
TYKL11	0,07%	1,1269	3,0861	0,0022	0,6826
TYKL12	0,80%	1,2830	3,6075	0,0289	0,7980
TYKL13	3,30%	1,2649	3,5427	0,1169	0,7836
TYKL14	13,17%	1,4207	4,1400	0,5452	0,9158
TYKL15	15,90%	1,3417	3,8255	0,6083	0,8462
TYKL16	15,97%	1,4445	4,2397	0,6771	0,9378
TYKL17	19,40%	1,5647	4,7812	0,9276	1,0576
TYKL18	12,83%	1,4636	4,3215	0,5544	0,9560
TYKL19	9,66%	1,5786	4,8482	0,4683	1,0724
TYKL20	3,82%	1,7987	6,0418	0,2308	1,3364
TYKL21	2,17%	1,8516	6,3700	0,1382	1,4090
TYKL22	1,02%	1,7357	5,6729	0,0579	1,2548
TYKL23	1,15%	1,8914	6,6286	0,0762	1,4662
TYKL24	0,41%	1,8281	6,2221	0,0255	1,3763
TYKL25	0,69%	2,1755	8,8066	0,0608	1,9480
				4,5208	

Tabelle 12: Indizes Typklasse

Level	Anteil	Estimate	exp(Estimate)	Anteil x exp(Estimate)	Normierter Wert
RKL0	4,40%	0,0000	1,0000	0,0440	0,9311
RKL1	15,51%	-0,0398	0,9610	0,1491	0,8948
RKL2	14,41%	-0,0380	0,9627	0,1387	0,8964
RKL3	8,18%	-0,0096	0,9904	0,0810	0,9222
RKL4	9,96%	0,0269	1,0273	0,1023	0,9565
RKL5	11,27%	0,1228	1,1307	0,1274	1,0528
RKL6	13,93%	0,1204	1,1279	0,1571	1,0502
RKL7	11,97%	0,1249	1,1330	0,1356	1,0549
RKL8	5,72%	0,2382	1,2690	0,0726	1,1816
RKL9	3,04%	0,2820	1,3258	0,0403	1,2345
RKL10	0,91%	0,3187	1,3753	0,0125	1,2805
RKL11	0,72%	0,6222	1,8630	0,0134	1,7346
				1,074	

Tabelle 13: Indizes Regionalklasse

Level	Anteil	Estimate	exp(Estimate)	Anteil x exp(Estimate)	Normierter Wert
Fahrzeugalter-1	29,35%	0,0000	1,0000	0,2935	0,9157
Fahrzeugalter0	13,37%	-0,0559	0,9456	0,1264	0,8659
Fahrzeugalter1	13,38%	-0,0495	0,9517	0,1273	0,8714
Fahrzeugalter3	15,76%	0,0336	1,0342	0,1630	0,9470
Fahrzeugalter6	14,45%	0,1886	1,2076	0,1745	1,1058
Fahrzeugalter10	13,32%	0,4102	1,5071	0,2007	1,3800
Fahrzeugalter20	0,38%	0,5650	1,7594	0,0067	1,6110
				1,0921	

Tabelle 14: Indizes Fahrzeugalter

Level	Anteil	Estimate	exp(Est)	Anteil x exp(Est)	Normierter Wert
EFPA EF männl. $\neq$ VN	1,88%	0,0000	1,0000	0,0188	2,2124
EFPA EF männl. VN	21,31%	-0,6578	0,5180	0,1104	1,1460
EFPA EF weibl $\neq$ VN	2,13%	-0,2769	0,7581	0,0161	1,6772
EFPA EF weibl. VN	13,22%	-0,8607	0,4229	0,0559	0,9356
EFPA kein EF/PA	6,15%	-0,2499	0,7789	0,0479	1,7232
EFPA Partner	54,68%	-1,0019	0,3672	0,2008	0,8124
EFPA Partner Firmen	0,63%	-1,0991	0,3332	0,0021	0,7372
				0,452	

Tabelle 15: Indizes EFPA

Level	Anteil	Estimate	exp(Est)	Anteil x exp(Est)	Normierter Wert
Carport	8,83%	0,0000	1,0000	0,0833	0,9724
E-/D-Garage	63,05%	0,0397	1,0405	0,6560	1,0118
Grunstück	17,94%	0,0476	1,0488	0,1882	1,0198
sonst. Garagen/Parkhaus	0,91%	-0,1633	0,8493	0,0077	0,8258
Straßenrand	9,27%	-0,0496	0,9516	0,0882	0,9253
				1,0284	

Tabelle 16: Indizes Fahrzeugabstellort

Level	Anteil	Estimate	exp(Est)	Anteil x exp(Estimate)	Normierter Wert
1-2-FH	43,96%	0,0000	1,0000	0,4396	0.9389
1-2-FH, vers.	20,57%	0,0726	1,0753	0,2212	1,0096
ETW	4,87%	0,1927	1,2125	0,0590	1,1384
keine Angaben	1,54%	0,0720	1,0747	0,0166	1,0090
MFH	0,63%	0,0592	1,0610	0,0067	0,9962
MFH, vers.	0,27%	0,5515	1,7359	0,0047	1,6298
nein	28,16%	0,1194	1,1268	0,3173	1,0579
				1,0651	

Tabelle 17: Indizes Wohneigentum

Level	Anteil	Estimate	exp(Est)	Anteil x exp(Est)	Normierter Wert
SFR 0	0,05%	0,0000	1,0000	0,0005	4,3384
SFR $\frac{1}{2}$	0,87%	-0,3650	0,6942	0,0060	3,0117
SFR 1	1,02%	-0,7082	0,4925	0,0050	2,1367
SFR 2	4,77%	-0,8710	0,4185	0,0120	1,8156
SFR 3	4,19%	-1,0397	0,3536	0,0148	1,5341
SFR 4	3,94%	-1,2033	0,3002	0,0118	1,3024
SFR 5	4,30%	-1,3661	0,2551	0,0110	1,1067
SFR 6	4,54%	-1,3064	0,2708	0,0123	1,1748
SFR 7	4,52%	-1,3640	0,2556	0,0116	1,1089
SFR 8	4,03%	-1,3729	0,2534	0,0102	1,0993
SFR 9	3,77%	-1,3741	0,2531	0,0095	1,0980
SFR 10	3,80%	-1,4679	0,2304	0,0088	0,9996
SFR 11	3,71%	-1,3949	0,2479	0,0092	1,0755
SFR 12	3,50%	-1,6605	0,1900	0,0067	0,8243
SFR 13	3,29%	-1,5172	0,2193	0,0072	0,9514
SFR 14	3,14%	-1,6565	0,1908	0,0060	0,8278
SFR 15	2,91%	-1,6426	0,1935	0,0056	0,8395
SFR 16	2,82%	-1,6152	0,1989	0,0056	0,8629
SFR 17	2,85%	-1,4989	0,2234	0,0064	0,9692
SFR 18	2,72%	-1,3946	0,2479	0,0067	1,0755
SFR 19	2,61%	-1,6278	0,1964	0,0051	0,8521
SFR 20	2,73%	-1,7208	0,1789	0,0049	0,7761
SFR 21	2,71%	-1,9046	0,1489	0,0040	0,6460
SFR 22	2,94%	-1,6051	0,2009	0,0059	0,8716
SFR 23	2,64%	-1,6785	0,1867	0,0049	0,8100
SFR 24	2,54%	-1,7071	0,1814	0,0046	0,7870
SFR 25	18,99%	-1,7340	0,1766	0,0335	0,7662
SFR M	0,01%	-0,4147	0,6605	0,0000	2,8665
SFR S	0,09%	-0,2386	0,7877	0,0007	3,4171
				0,2305	

Tabelle 18: Indizes Schadenfreiheitsklasse

Hat man nun die Indizes für alle Merkmale berechnet, so erhält man den neuen Intercept-Wert aus dem Produkt des ehemaligen exp(Intercept)-Wertes mit den summierten Werten aus "*Anteil* x *exp(Estimate)*" für jedes Merkmal.

$$\begin{aligned}
\exp(INT) \quad & \cdot 1,1616 \cdot 0,8691 \cdot 1,0006 \cdot 4,5208 \cdot 1,074 \cdot 1,0921 \cdot 0,452 \\
& \cdot 1,0284 \cdot 1,0651 \cdot 0,2305 \\
= \quad & \exp(5,6505) \cdot 1,1616 \cdot 0,8691 \cdot 1,0006 \cdot 4,5208 \cdot 1,074 \cdot 1,0921 \cdot 0,452 \\
& \cdot 1,0284 \cdot 1,0651 \cdot 0,2305 \\
= \quad & 173,8648
\end{aligned}$$

4.4.1 Beispiel zur Beitragsermittlung

Wir wollen uns nun zwei Beispiele anschauen, in denen wir zwei Versicherungsnehmer mit unterschiedlichen Merkmalsausprägungen miteinander vergleichen und ihren Beitrag berechnen. Die Versicherungsnehmer (A und B) sollen die folgenden Eigenschaften besitzen. Es sei hierbei erwähnt, dass die Kombination von diesen

Merkmal	Versicherter A	Versicherter B
Tarifgruppe	B	N
Deckungsart	1	1
Typklasse	14	19
Kilometerklasse	13	21
Regionalklasse	2	8
Fahrzeugalter	1	10
EFPA	weibl. VN	männl. VN
Abstellort	Carport	Straßenrand
Wohneigentum	1-2 FH	nein
Schadenfreiheitsklasse	20	3

Tabelle 19: Merkmalsausprägungen zweier Versicherungsnehmer

Ausprägungen nicht realitätsnah sein mag, es soll hier lediglich der Extremfall von einem sehr niedrigen und einem hohen Beitrag gezeigt werden. Versicherungsnehmer B hat in unserem Beispiel eine ungünstige Merkmalskonstellation, was seine Beitragshöhe betrifft. Als Beitrag für Versicherungsnehmer A ergibt sich:

$$\begin{aligned}
\text{Beitrag(A)} \quad &= \quad 173,8648 \cdot 0,9564 \cdot 0,9994 \cdot 0,9158 \cdot 0,9678 \cdot 0,8964 \cdot 0,8714 \\
& \quad \cdot 0,9356 \cdot 0,9724 \cdot 0,9389 \cdot 0,7761 \\
&= \quad 76,27
\end{aligned}$$

Somit ist für diesen Versicherungsnehmer eine jährliche Nettoprämie in Höhe von 76,27 Euro fällig. Analog folgt für Versicherungsnehmer B:

$$\begin{aligned} \text{Beitrag(B)} &= 173,8648 \cdot 0,9960 \cdot 0,9994 \cdot 1,0724 \cdot 1,3332 \cdot 1,1816 \cdot 1,3800 \\ &\quad \cdot 1,1460 \cdot 0,9253 \cdot 1,0579 \cdot 1,5341 \\ &= 694,35 \end{aligned}$$

Somit muss B eine viel höhere Beitragsprämie zahlen als A. Grund hierfür ist die für die Beitragsermittlung günstige Konstellation an Merkmalsausprägungen. Besonderes Gewicht liegt in diesem Fall auf der Ermäßigung aus der Schadenfreiheitsklasse 20 für Versicherungsnehmer A. Negativ fällt bei Versicherungsnehmer B die Schadenfreiheitsklasse 3 auf.

5 Schlussbemerkung

In dieser Diplomarbeit wurde ein Teilbereich aus dem Prozess der Tarifierung in der Kraftfahrzeughaftpflichtversicherung vorgestellt. Weitere wichtige Bereiche wie die Großschädenproblematik oder die Kalkulation der Bruttoprämien wurden ausgeschlossen. Desweiteren spielt in der Versicherungspraxis die Erfahrungstarifierung eine wesentliche Rolle. Die in dieser Arbeit präsentierten Methoden zur Merkmalsauswahl wie die Devianzanalyse geben eine Hilfestellung, jedoch wird die Entscheidung über die Miteinbeziehung von Merkmalen in der Praxis häufig aufgrund von Erfahrungswerten getroffen.

Die verallgemeinerten linearen Modelle bieten bei hinreichend großem Stichprobenumfang eine gute Möglichkeit, den Schadenbedarf zu analysieren und zu prognostizieren. Auch die Schadenanzahl und die Schadenhöhe können analysiert werden, sodass Einflussgrößen bestimmt werden können. Die Analyse der Schadenhöhe ist jedoch schwieriger als die der Schadenfrequenz. Der Grund hierfür ist, dass die Merkmale die Schadenhöhe nicht so gut erklären können, wie die Schadenfrequenz. Dies sieht man auch an den viel kleineren p-Werten in dem Modell für Schadenfrequenz im Vergleich zu den p-Werten aus dem Schadenhöhemodell. Während in der Schadenfrequenz die Merkmale als hoch signifikant eingestuft werden, geschieht dies bei der Schadenhöhe nicht.

Anzumerken ist, dass in der Statistiksoftware **R** immer die erste Ausprägung eines Merkmals in den Intercept aufgenommen wird. Dies kann unter Umständen zu verfälschten Ergebnissen führen, wenn die erste Merkmalsausprägung nur von sehr wenigen Versicherten erfüllt wird. Sinnvoller wäre, die Merkmalsausprägung mit dem höchstem Prozentanteil in den Verträgen als Intercept zu wählen.

Desweiteren besteht die Möglichkeit a-priori Informationen in ein verallgemeinertes Modell miteinzubeziehen. Dies kann zum Beispiel dann von Nutzen sein, wenn man die empfohlenen Indexwerte für die Schadenfreiheitsklasse des Gesamtverbands der Deutschen Versicherungswirtschaft in seinen Tarif miteinbeziehen möchte und keine eigenen Indexwerte berechnen möchte.

In Zukunft werden die Versicherer ihre Tarife erneut anpassen müssen, da nach einem Beschluss des Europäischen Gerichtshofes vom 01.03.2011 Versicherungskonzerne künftig Unisex-Tarife anbieten müssen (Rechtssache C-236/09), vgl [18]. Die bislang übliche Berücksichtigung des Geschlechts als Risikofaktor für Versicherungsbeiträge diskriminiere Frauen und sei deswegen ungültig. Dabei spielt das Geschlecht häufig eine zentrale Rolle, wie wir in der empirischen Analyse gese-

hen haben. Der Indexwert für weibliche Versicherungsnehmer beträgt in unserem Fall 0,94, der für männliche Versicherungsnehmer 1,15. Frauen zahlen weniger für die Kfz-Versicherung, da sie weniger Unfälle verursachen. Der Gesamtverband der Deutschen Versicherungswirtschaft erklärte, im Schnitt würde es zu einer Anhebung der Beiträge führen, weil der Geschlechtermix als neues Risiko in die Kalkulation eingehe, siehe [11].

Tabellenverzeichnis

Abbildungsverzeichnis

Literatur

[1] Esbjörn Ohlsson, Björn Johansson: *Non-Life Insurance Pricing with Generalized Linear Models*, Springer-Verlag, 2010

[2] P. McCullagh, J.A. Nelder: *Generalized Linear Models*, Second Edition, Chapman & Hall, 1989

[3] George H. Dunteman, Moon-Ho R. Ho: *An Introduction to Generalized Linear Models*, Sage Publications, 2006

[4] Annette Dobson: *An Introduction to Generalized Linear Models* , Second Edition, Chapman & Hall, 2002

[5] James K. Lindsey: *Applying Generalized Linear Models*, Springer-Verlag, 1997

[6] Olaf Kruse: *Modelle zur Analyse und Prognose des Schadenbedarfs in der Kraftfahrzeug-Haftpflichtversicherung*, VVW, 1997

[7] Joachim Thomas Walter: *Zur Anwendung von Verallgemeinerten Linearen Modellen zu Zwecken der Tarifierung in der Kraftfahrzeug-Haftpflichtversicherung*, VVW, 1998

[8] Adrian Colin Cameron, P.K. Trivedi: *Regression Analysis of count data*, Cambridge University Press, 1998

[9] Deutsche Aktuarvereinigung: *Methoden in der Tarifierung*, Schriftenreihe Versicherungs- und Finanzmathematik der DGVFM, 2011

[10] Helmut Pruscha: *Angewandte Methoden der Mathematischen Statistik*, Teubner Skripten zur Mathematischen Stochastik, 1989

[11] Gesamtverband der deutschen Versicherungswirtschaft, `http://www.gdv.de`

[12] Klaus D. Schmidt: *Versicherungsmathematik*, 2. Auflage, Springer Verlag, 2006

[13] L. Fahrmeir, G. Tutz: *Multivariate statistical modelling based on generalized linear models*, Springer Verlag, 2nd edition, 2001

[14] K. Sticker: *Analyse der Tarifstruktur für die Haftpflichtversicherung von Personenwagen* Verlag Versicherungswirtschaft, 1983

[16] Michael J. Crawley: *The R Book*, John Wiley & Sons Ltd. , 2007

[17] JProf. Dr. Hajo Holzmann: *Skript zur Statistik 2*, Institut für Stochastik der Universität Karlsruhe (TH), Wintersemester 2007/08

[18] Urteil des Gerichtshofes 1. März 2011 `http://curia.europa.eu/jurisp/cgi-bin/gettext.pl?lang=de&num=79889698C19090236&doc=T&ouvert=T&seance=ARRET`

[19] Peter J. Bickel, Kjell A. Doksum: *Mathematical Statistics*,2nd edition, Prentice Hall, 2001

[20] R. E. Beard, T. Pentikäinen, E. Pesonen: *Risk Theory*, Chapman & Hall, 1984

[21] Christian Hipp: *Risikotheorie: Stochastische Methoden und Statistische Verfahren Teil 2*, Universität Kiel

Anhang

aus der SF-Klasse	1 Schaden	2 Schäden	3 Schäden	4 und mehr Schäden
	in die SF-Klasse			
SF 25	SF 22	SF 4	SF 2	SF M
SF 24	SF 11	SF 4	SF 2	SF M
SF 23	SF 10	SF 4	SF 2	SF M
SF 22	SF 10	SF 4	SF 2	SF M
SF 21	SF 10	SF 4	SF 2	SF M
SF 20	SF 9	SF 3	SF 2	SF M
SF 19	SF 9	SF 3	SF 2	SF M
SF 18	SF 7	SF 3	SF 2	SF M
SF 17	SF 7	SF 2	SF 1	SF M
SF 16	SF 6	SF 2	SF 1	SF M
SF 15	SF 6	SF 2	SF 1	SF M
SF 14	SF 6	SF 2	SF 1	SF M
SF 13	SF 5	SF 2	SF 1	SF M
SF 12	SF 5	SF 1	SF $\frac{1}{2}$	SF M
SF 11	SF 5	SF 1	SF $\frac{1}{2}$	SF M
SF 10	SF 4	SF 1	SF $\frac{1}{2}$	SF M
SF 9	SF 4	SF 1	SF $\frac{1}{2}$	SF M
SF 8	SF 4	SF 1	SF $\frac{1}{2}$	SF M
SF 7	SF 3	SF $\frac{1}{2}$	S	M
SF 6	SF 3	SF $\frac{1}{2}$	S	M
SF 5	SF 2	SF $\frac{1}{2}$	S	M
SF 4	SF 2	SF $\frac{1}{2}$	S	M
SF 3	SF 1	S	M	M
SF 2	SF $\frac{1}{2}$	S	M	M
SF 1	S	M	M	M
SF $\frac{1}{2}$	S	M	M	M
S	M	M	M	M
0	M	M	M	M
M	M	M	M	M

Tabelle 20: Rückstufung SFR

```
Call:
glm(formula = SCHADEN/JE ~ TG + DAf + TYKLf + RKLf + KMKLf +
    SFR + Fzgalter + EFPA + FZGAOS + WOHNEIG.ART.SCHL, family = poisson(link = log)
    weights = JE)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7483  -0.3563  -0.2747  -0.1910   5.2427

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                    -2.297649   0.213205 -10.777  < 2e-16 ***
TGB                             0.002251   0.026178   0.086 0.931479
TGN                             0.084779   0.023813   3.560 0.000371 ***
DAf4                            0.367150   0.189181   1.941 0.052291 .
TYKLf11                         0.419793   0.327320   1.283 0.199662
TYKLf12                         0.692157   0.188855   3.665 0.000247 ***
TYKLf13                         0.710565   0.179529   3.958 7.56e-05 ***
TYKLf14                         0.838787   0.177194   4.734 2.20e-06 ***
TYKLf15                         0.887082   0.177182   5.007 5.54e-07 ***
TYKLf16                         0.958313   0.177141   5.410 6.31e-08 ***
TYKLf17                         1.000531   0.176979   5.653 1.57e-08 ***
TYKLf18                         1.013181   0.177219   5.717 1.08e-08 ***
TYKLf19                         1.077074   0.177417   6.071 1.27e-09 ***
TYKLf20                         1.179719   0.178637   6.604 4.00e-11 ***
TYKLf21                         1.247149   0.180155   6.923 4.43e-12 ***
TYKLf22                         1.266511   0.183788   6.891 5.53e-12 ***
TYKLf23                         1.370891   0.182670   7.505 6.15e-14 ***
TYKLf24                         1.458122   0.192243   7.585 3.33e-14 ***
TYKLf25                         1.730281   0.224114   7.721 1.16e-14 ***
RKLf1                           0.032092   0.033282   0.964 0.334920
RKLf2                           0.060902   0.033466   1.820 0.068789 .
RKLf3                           0.091562   0.036094   2.537 0.011187 *
RKLf4                           0.082383   0.035052   2.350 0.018756 *
RKLf5                           0.121845   0.034161   3.567 0.000361 ***
RKLf6                           0.142705   0.033286   4.287 1.81e-05 ***
RKLf7                           0.181525   0.033647   5.395 6.85e-08 ***
RKLf8                           0.238794   0.037122   6.433 1.25e-10 ***
RKLf9                           0.273322   0.042054   6.499 8.07e-11 ***
RKLf10                          0.331481   0.060632   5.467 4.57e-08 ***
```

RKLf11	0.464177	0.062122	7.472	7.90e-14	***
KMKLf7	0.092966	0.020040	4.639	3.50e-06	***
KMKLf10	0.147658	0.021487	6.872	6.33e-12	***
KMKLf13	0.166879	0.022937	7.276	3.45e-13	***
KMKLf16	0.262025	0.022676	11.555	< 2e-16	***
KMKLf21	0.377134	0.034660	10.881	< 2e-16	***
KMKLf26	0.499446	0.035956	13.891	< 2e-16	***
KMKLf31	0.678727	0.039763	17.069	< 2e-16	***
SFR1	-1.042377	0.114183	-9.129	< 2e-16	***
SFR1/2	-0.700154	0.112856	-6.204	5.51e-10	***
SFR10	-1.660147	0.111930	-14.832	< 2e-16	***
SFR11	-1.639162	0.112015	-14.633	< 2e-16	***
SFR12	-1.756096	0.112877	-15.558	< 2e-16	***
SFR13	-1.711220	0.113064	-15.135	< 2e-16	***
SFR14	-1.797175	0.113847	-15.786	< 2e-16	***
SFR15	-1.790571	0.114306	-15.665	< 2e-16	***
SFR16	-1.848369	0.114958	-16.079	< 2e-16	***
SFR17	-1.711620	0.114031	-15.010	< 2e-16	***
SFR18	-1.793957	0.114912	-15.612	< 2e-16	***
SFR19	-1.775623	0.115078	-15.430	< 2e-16	***
SFR2	-1.197852	0.109320	-10.957	< 2e-16	***
SFR20	-1.871217	0.115621	-16.184	< 2e-16	***
SFR21	-1.855189	0.115630	-16.044	< 2e-16	***
SFR22	-1.823833	0.114761	-15.892	< 2e-16	***
SFR23	-1.790830	0.115271	-15.536	< 2e-16	***
SFR24	-1.898185	0.116421	-16.305	< 2e-16	***
SFR25	-1.782209	0.109154	-16.327	< 2e-16	***
SFR3	-1.361700	0.110009	-12.378	< 2e-16	***
SFR4	-1.417039	0.110406	-12.835	< 2e-16	***
SFR5	-1.488970	0.110459	-13.480	< 2e-16	***
SFR6	-1.514198	0.110500	-13.703	< 2e-16	***
SFR7	-1.544130	0.110710	-13.948	< 2e-16	***
SFR8	-1.567707	0.111266	-14.090	< 2e-16	***
SFR9	-1.629351	0.111787	-14.575	< 2e-16	***
SFRM	-0.441147	0.333962	-1.321	0.186518	
SFRS	-0.762568	0.154500	-4.936	7.99e-07	***
Fzgalter0	-0.025083	0.021320	-1.177	0.239385	
Fzgalter1	0.007996	0.020857	0.383	0.701451	
Fzgalter3	0.046520	0.019377	2.401	0.016362	*
Fzgalter6	0.162818	0.019269	8.450	< 2e-16	***

```
Fzgalter10                          0.373081   0.019175  19.457  < 2e-16 ***
Fzgalter20                          0.314012   0.093997   3.341 0.000836 ***
EFPAEF maennl. VN                  -0.584831   0.031954 -18.302  < 2e-16 ***
EFPAEF weibl. ^= VN                -0.185173   0.042932  -4.313 1.61e-05 ***
EFPAEF weibl. VN                   -0.542429   0.033673 -16.109  < 2e-16 ***
EFPAkein EF/PA                     -0.144352   0.034453  -4.190 2.79e-05 ***
EFPAPartner                        -0.721081   0.030966 -23.286  < 2e-16 ***
EFPAPartner Firmen                 -0.700228   0.081618  -8.579  < 2e-16 ***
FZGAOSE-/D-Garage                   0.023794   0.021533   1.105 0.269162
FZGAOSGrundstueck                   0.029035   0.023913   1.214 0.224666
FZGAOSsonst. Garagen/Parkhaus       0.018748   0.060571   0.310 0.756920
FZGAOSStrassenrand                  0.079862   0.026741   2.986 0.002822 **
WOHNEIG.ART.SCHL1-2-FH, vers.      -0.037601   0.016760  -2.244 0.024863 *
WOHNEIG.ART.SCHLETW                 0.103014   0.027481   3.749 0.000178 ***
WOHNEIG.ART.SCHLkeine Angaben      -0.013894   0.044362  -0.313 0.754132
WOHNEIG.ART.SCHLMFH                -0.006467   0.074385  -0.087 0.930716
WOHNEIG.ART.SCHLMFH, vers.          0.205338   0.105294   1.950 0.051161 .
WOHNEIG.ART.SCHLnein                0.090689   0.015030   6.034 1.60e-09 ***
---
Signif. codes:  0 Ô***Õ 0.001 Ô**Õ 0.01 Ô*Õ 0.05 Ô.Õ 0.1 Ô Õ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 159463  on 499999  degrees of freedom
Residual deviance: 152894  on 499913  degrees of freedom
AIC: Inf

Number of Fisher Scoring iterations: 6
```

Summary Schadenhöhe

```
Call:
glm(formula = AUFWAND/SCHADEN ~ TG + DAf + TYKLf + RKLf + KMKLf +
    SFR + Fzgalter + EFPA + FZGAOS + WOHNEIG.ART.SCHL, family = Gamma(link = "log"),
    weights = SCHADEN, na.action = na.omit)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.7531  -1.0871  -0.6360  -0.0333  11.8288

Coefficients:
```

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	7.9481264	0.6117769	12.992	< 2e-16	***
TGB	-0.1871452	0.0748734	-2.499	0.01244	*
TGN	-0.2290537	0.0683084	-3.353	0.00080	***
DAf4	0.3162965	0.5414939	0.584	0.55915	
TYKLf11	0.7071053	0.9341170	0.757	0.44907	
TYKLf12	0.5908571	0.5381901	1.098	0.27228	
TYKLf13	0.5542859	0.5120465	1.082	0.27904	
TYKLf14	0.5819322	0.5050368	1.152	0.24923	
TYKLf15	0.4545811	0.5047453	0.901	0.36780	
TYKLf16	0.4861572	0.5048148	0.963	0.33554	
TYKLf17	0.5641667	0.5043552	1.119	0.26332	
TYKLf18	0.4503723	0.5050318	0.892	0.37252	
TYKLf19	0.5015073	0.5056857	0.992	0.32133	
TYKLf20	0.6189726	0.5091072	1.216	0.22407	
TYKLf21	0.6044824	0.5136768	1.177	0.23930	
TYKLf22	0.4691809	0.5240439	0.895	0.37063	
TYKLf23	0.5204955	0.5209845	0.999	0.31777	
TYKLf24	0.3699252	0.5482282	0.675	0.49983	
TYKLf25	0.4451656	0.6400389	0.696	0.48673	
RKLf1	-0.0719374	0.0952580	-0.755	0.45014	
RKLf2	-0.0989034	0.0957709	-1.033	0.30175	
RKLf3	-0.1011257	0.1032423	-0.979	0.32734	
RKLf4	-0.0555186	0.1002670	-0.554	0.57978	
RKLf5	0.0010020	0.0977639	0.010	0.99182	
RKLf6	-0.0223031	0.0952617	-0.234	0.81489	
RKLf7	-0.0565817	0.0962897	-0.588	0.55679	
RKLf8	-0.0005864	0.1063305	-0.006	0.99560	
RKLf9	0.0087106	0.1204301	0.072	0.94234	
RKLf10	-0.0127668	0.1738778	-0.073	0.94147	
RKLf11	0.1579865	0.1777862	0.889	0.37421	
KMKLf7	-0.0295030	0.0573596	-0.514	0.60701	
KMKLf10	0.0135299	0.0614238	0.220	0.82566	
KMKLf13	-0.0498113	0.0656156	-0.759	0.44778	
KMKLf16	-0.0763738	0.0648194	-1.178	0.23871	
KMKLf21	0.0602550	0.0991593	0.608	0.54342	
KMKLf26	0.2062994	0.1028656	2.006	0.04492	*
KMKLf31	-0.0218542	0.1130104	-0.193	0.84666	
SFR1	0.3341610	0.3270572	1.022	0.30692	
SFR1/2	0.3351727	0.3232398	1.037	0.29978	

SFR10	0.1922706	0.3208062	0.599	0.54895	
SFR11	0.2443066	0.3211304	0.761	0.44680	
SFR12	0.0955649	0.3236531	0.295	0.76779	
SFR13	0.1939707	0.3241036	0.598	0.54952	
SFR14	0.1407113	0.3263226	0.431	0.66632	
SFR15	0.1480036	0.3276520	0.452	0.65148	
SFR16	0.2331291	0.3295648	0.707	0.47933	
SFR17	0.2127700	0.3270038	0.651	0.51527	
SFR18	0.3993221	0.3294180	1.212	0.22544	
SFR19	0.1477907	0.3298180	0.448	0.65409	
SFR2	0.3268205	0.3131113	1.044	0.29659	
SFR20	0.1504650	0.3315473	0.454	0.64996	
SFR21	-0.0494292	0.3312020	-0.149	0.88136	
SFR22	0.2187717	0.3289618	0.665	0.50603	
SFR23	0.1123251	0.3303790	0.340	0.73387	
SFR24	0.1911056	0.3336943	0.573	0.56685	
SFR25	0.0482500	0.3129710	0.154	0.87748	
SFR3	0.3219672	0.3152195	1.021	0.30707	
SFR4	0.2137104	0.3164252	0.675	0.49943	
SFR5	0.1229190	0.3165514	0.388	0.69779	
SFR6	0.2077724	0.3167555	0.656	0.51187	
SFR7	0.1801268	0.3173865	0.568	0.57036	
SFR8	0.1948121	0.3190067	0.611	0.54141	
SFR9	0.2552243	0.3205388	0.796	0.42590	
SFRM	0.0264407	0.9557060	0.028	0.97793	
SFRS	0.5239365	0.4422176	1.185	0.23611	
Fzgalter0	-0.0308171	0.0610190	-0.505	0.61353	
Fzgalter1	-0.0575340	0.0595988	-0.965	0.33438	
Fzgalter3	-0.0129285	0.0554588	-0.233	0.81567	
Fzgalter6	0.0257471	0.0553111	0.465	0.64158	
Fzgalter10	0.0371340	0.0556802	0.667	0.50483	
Fzgalter20	0.2509896	0.2687195	0.934	0.35030	
EFPAEF maennl. VN	-0.0730173	0.0911856	-0.801	0.42328	
EFPAEF weibl. ^= VN	-0.0917475	0.1227835	-0.747	0.45493	
EFPAEF weibl. VN	-0.3182315	0.0961095	-3.311	0.00093	***
EFPAkein EF/PA	-0.1055433	0.0982513	-1.074	0.28273	
EFPAPartner	-0.2807685	0.0885394	-3.171	0.00152	**
EFPAPartner Firmen	-0.3989138	0.2334172	-1.709	0.08746	.
FZGAOSE-/D-Garage	0.0158753	0.0616036	0.258	0.79664	
FZGAOSGrundstueck	0.0185831	0.0684516	0.271	0.78603	

```
FZGAOSsonst. Garagen/Parkhaus     -0.1820589  0.1734630  -1.050  0.29393
FZGAOSStrassenrand                -0.1294391  0.0766018  -1.690  0.09108 .
WOHNEIG.ART.SCHL1-2-FH, vers.      0.1101912  0.0479644   2.297  0.02161 *
WOHNEIG.ART.SCHLETW                0.0896807  0.0786067   1.141  0.25393
WOHNEIG.ART.SCHLkeine Angaben      0.0858950  0.1264776   0.679  0.49706
WOHNEIG.ART.SCHLMFH                0.0656528  0.2130382   0.308  0.75795
WOHNEIG.ART.SCHLMFH, vers.         0.3461711  0.3014811   1.148  0.25088
WOHNEIG.ART.SCHLnein               0.0287496  0.0431110   0.667  0.50486
---
Signif. codes:  0 Ô***Õ 0.001 Ô**Õ 0.01 Ô*Õ 0.05 Ô.Õ 0.1 Ô Õ 1

(Dispersion parameter for Gamma family taken to be 8.171105)

    Null deviance: 38481  on 27384  degrees of freedom
Residual deviance: 37228  on 27298  degrees of freedom
  (472615 observations deleted due to missingness)
AIC: 548575

Number of Fisher Scoring iterations: 8
```