Technische Universität München

Zentrum Mathematik

Modellwahl bei der KFZ Haftpflicht-Versicherung mit Hilfe von GLMs

Diplomarbeit

von

Ivonne Siegelin

Themenstellerin:	Prof. Dr. C. Czado, Dr. G. Sussmann (VKB)
Betreuer:	Prof. Dr. C. Czado, Dr. G. Sussmann (VKB)
Abgabetermin:	14.11.2008

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

Garching, den 14. November 2008

Danksagung

Die vorliegende Diplomarbeit ist in Zusammenarbeit mit dem Zentrum für Mathematik der Technischen Universität München und dem Aktuariat Komposit der Versicherungskammer Bayern entstanden.

Mein besonderer Dank gilt Frau Prof. Dr. Claudia Czado, für die Betreuung meiner Diplomarbeit seitens der Technischen Universität München. Ich bedanke mich für die fachlichen Diskussionen und Hilfestellungen sowie für die Beantwortung meiner vielen Fragen.

Weiterhin möchte ich mich bei den Mitarbeitern des Aktuariats Komposit für die fachliche, sowie die persönliche Unterstützung bedanken. Besonders bedanke ich mich bei Herrn Dr. Gerald Sussmann, der es mir ermöglicht hat meine Diplomarbeit in seiner Abteilung zu schreiben. Er hat mich während meiner Diplomarbeit betreut und mir immer wieder wichtige Hinweise und interessante Anregungen geliefert.

Zu guter Letzt spreche ich meiner Familie und meinen Freunden den herzlichsten Dank für ihre Unterstützung aus. Vor allem meinen Eltern danke ich dafür, dass sie mir das Studium an der Technischen Universität München ermöglicht und mich stets in meinem Vorhaben bestärkt haben.

Inhaltsverzeichnis

1 Einleitung

Der Vergleich von statistischen Modellen ist in vielen Bereichen der Wissenschaft von ebenso großer Bedeutung, wie für zahlreiche Problemstellungen aus der Wirtschaft. Die Statistik stellt eine große Anzahl an Methoden und Verfahren zur Verfügung um genestete Modelle zu vergleichen. Für den Vergleich von nicht genesteten Modellen hingegen bietet die Statistik nur wenige Instrumente. Zwei dieser Instrumente, der Vuong Test und der Distribution-Free Test, werden in dieser Arbeit anhand eines realen Datensatzes untersucht.

Im Mittelpunkt des praktischen Teils dieser Diplomarbeit steht das Ersetzen von Merkmalen eines Kraftfahrt Haftpflicht Tarifs für private PKW und der Vergleich des ursprünglichen mit dem neu entstanden Tarif.

In dieser Arbeit wird kein grundlegend neuer Tarif entwickelt, lediglich, wie bereits erwähnt, Tarifmerkmale ausgetauscht. Daher werden nicht alle Merkmale, die in den Tarif eingehen, betrachtet. Aber das vollständige Ignorieren der Merkmale würde die Ergebnisse dieser Arbeit beeinflussen. Aus diesem Grund werden die Merkmale, die nicht ersetzt werden, insoweit in das Modell einbezogen, dass sie im Modell als Offset verwendet werden.

Nach einem Überblick über die notwendigen mathematischen Grundlagen wird der reale Datensatz betrachtet. Der Datensatz wird auf zwei Weisen verdichtet. Für die beiden so entstandenen Datensätze wird eine explorative Analyse durchgeführt.

Anschließend werden, entsprechend der Problemstellung dieser Arbeit, vier verschiedene Gruppen von Modellen an die Daten angepasst. Dann wird die Anpassungsgüte dieser Modelle untersucht, es werden Hypothesentests und eine Residuenanalyse durchgeführt. Das Modell, das je Gruppe am besten an die Daten angepasst ist, wird illustriert.

Im Anschluss erfolgt eine erneute Modellwahl unter Verwendung des Vuong Tests und des Distribution-Free Tests. Es wird zunächst die Verteilung ausgewählt, die die Modelle

besser an die Daten anpasst. Danach werden genestete und nicht genestete Modelle verglichen. Abschließend werden Simulationen durchgeführt, um herauszufinden in wieweit man diesen Tests vertrauen kann.

2 Theoretische Grundlagen

Das folgende Kapitel gibt einen Überblick über die für diese Arbeit notwendigen theoretischen Grundlagen. Im ersten Abschnitt werden diskrete und stetige Verteilungen mit ihren charakteristischen Eigenschaften betrachtet. Der zweite Abschnitt geht auf die exponentielle Familie ein.

Danach werden Generalisierte Lineare Modelle vorgestellt und die beiden für diese Arbeit relevanten Modelle, das Poisson-Modell mit Offset und Gewicht und das Gamma-Modell mit Offset und Gewicht, betrachtet.

Den Abschluss dieses Kapitels bildet die Vorstellung des Vuong Tests und des Distribution-Free Tests, zwei Tests zum Vergleich nicht genesteter Modelle.

2.1 Wahrscheinlichkeitsverteilungen

Es werden nun die Verteilungen vorgestellt, die im Verlauf der Arbeit untersucht werden. Neben der Wahrscheinlichkeitsfunktion oder der Dichte wird auf die Faltungseigenschaften der einzelnen Verteilungen eingegangen.

2.1.1 Poisson-Verteilung

Die Poisson-Verteilung ist eine diskrete Verteilung. Sie wird zur Analyse von Zähldaten verwendet, zum Beispiel, um die Anzahl der Schäden bei einer Sachversicherung zu modellieren.

Im Folgenden werden die Wahrscheinlichkeitsfunktion und die Additivität der Poisson-Verteilung beschrieben.

Definition 2.1 (Wahrscheinlichkeitsfunktion der Poisson-Verteilung)

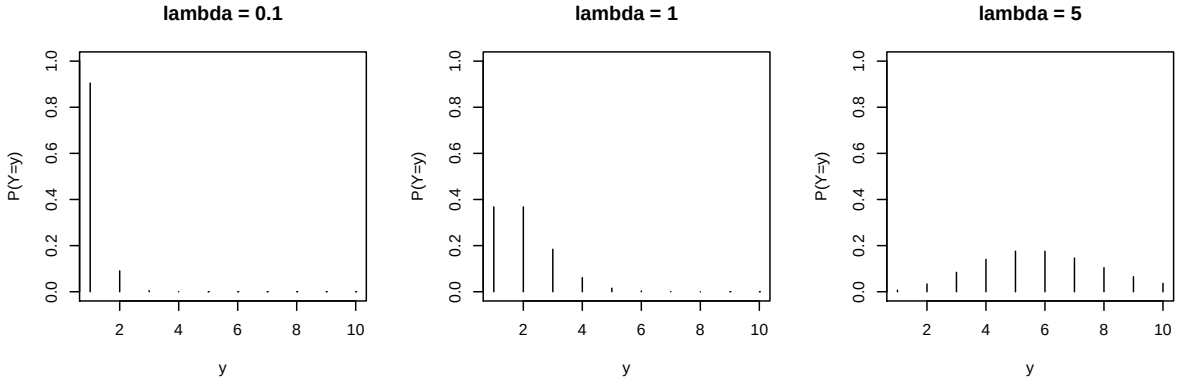
Eine diskrete, nicht-negative Zufallsvariable Y heißt Poisson-verteilt mit Parameter λ ($Y \sim Poi(\lambda)$), falls

$$P(Y = y) = \exp(-\lambda)\frac{\lambda^y}{y!} \qquad \textit{mit } \lambda > 0 \quad \textit{und} \quad y = 0, 1, 2, ...,$$

siehe Bickel and Doksum (1977), Seite 456, gilt

Abbildung 2.1 zeigt die Wahrscheinlichkeitsfunktion der Poisson-Verteilung für verschiedene Werte von λ. Dabei gilt, je kleiner der Wert für λ, desto linkssteiler ist die Wahrscheinlichkeitsfunktion. Für größere Werte von λ wird die Wahrscheinlichkeitsfunktion symmetrischer und lässt sich durch die Dichte der Normalverteilung approximieren, siehe Fahrmeir and Künstler (2002), Abschnitt 7.2.

Abbildung 2.1: Darstellung der Wahrscheinlichkeitsfunktion der Poisson-Verteilung für verschiedene Werte von λ



Eine Eigenschaft der Poisson-Verteilung ist, dass Erwartungswert und Varianz gleich sind. Es gilt:

$$\lambda = E(Y) = Var(Y)$$

Die Summe von Poisson-verteilten Zufallsvariablen hat ebenfalls eine spezielle Eigenschaft:

Satz 2.2 (Additivität der Poisson-Verteilung)

Seien Y und Z unabhängige Zufallsvariablen mit $Y \sim Poi(\lambda_1)$ und $Z \sim Poi(\lambda_2)$, dann gilt:

$$Y + Z \sim Poi(\lambda_1 + \lambda_2).$$

Beweis:

$$\begin{aligned} P_{Y+Z} &= \sum_{z=0}^{y} \frac{\lambda_1^{y+z}}{(y+z)!} \exp(-\lambda_1) \frac{\lambda_2^z}{z!} \exp(-\lambda_2) \\ &= \frac{\exp(-(\lambda_1+\lambda_2))}{y!} \sum_{z=0}^{y} \binom{y}{z} \lambda_1^{-y+z} \lambda_2^{-z} \\ &= \frac{\exp(-(\lambda_1+\lambda_2))(\lambda_1+\lambda_2)^y}{y!}. \end{aligned}$$

Somit ist die Summe zweier Poisson-verteilter Zufallsvariablen Poisson-verteilt.

2.1.2 Zero-Inflated-Poisson-Verteilung

Zur Abbildung von Nullüberschüssen eignet sich die Zero-Inflated-Poisson-Verteilung (ZIP-Verteilung), welche ebenfalls eine diskrete Verteilung ist. Im Folgenden werden die Wahrscheinlichkeitsfunktion und die Faltung zweier ZIP-verteilter Zufallsvariablen betrachtet.

Definition 2.3 (Wahrscheinlichkeitsfunktion der ZIP-Verteilung)

Sei Y eine Zufallsvariable mit diskreter Verteilung. Dann ist Y Zero-Inflated-Poisson-verteilt mit Parametern λ und ω ($Y \sim ZIP(\lambda, \omega)$), falls

$$P(Y = y|\lambda, \omega) = \begin{cases} \omega + (1-\omega)\exp(-\lambda) & \textit{für } y = 0 \\ (1-\omega)\exp(-\lambda)\frac{\lambda^y}{y!} & \textit{für } y = 1, 2, ... \end{cases}$$

gilt, siehe Lambert (1992).

Eine ZIP-verteilte Zufallsvariable Y verhält sich somit wie eine Poisson-verteilte Zufallsvariable mit erhöhter Eintrittswahrscheinlichkeit des Ereignisses Null. Unter dieser Verteilung nimmt Y den Wert Null mit Wahrscheinlichkeit $\omega \in [0, 1]$ an, und mit Wahrscheinlichkeit $(1-\omega)$ ist Y Poisson-verteilt. Der Parameter ω wird als Auswahlparameter bezeichnet.

Seien nun Y_1 und Y_2 unabhängige Zufallsvariablen mit $Y_1 \sim ZIP(\lambda_1, \omega)$ und $Y_2 \sim ZIP(\lambda_2, \omega)$, dann gilt für die Summe der Zufallsvariablen $Y_1 + Y_2$:

$$P(Y_1+Y_2|\lambda_1,\lambda_2,\omega) = \begin{cases} \omega^2 + \omega(1-\omega)e^{-\lambda_1-\lambda_2} + (1-\omega)^2 e^{-\lambda_1-\lambda_2} & \text{für } y_1, y_2 = 0 \\ \omega(1-\omega)e^{-\lambda_2}\frac{\lambda_2^{y_2}}{y_2!} + (1-\omega)^2 e^{-\lambda_1-\lambda_2}\frac{\lambda^{y_2}}{y_2!} & \text{für } y_1 = 0, y_2 = 1,2,... \\ (1-\omega)^2 e^{-\lambda_1-\lambda_2}\frac{\lambda^{y_1+y_2}}{y_1!y_2!} & \text{für } y_1, y_2 = 1,2,... \end{cases}$$

Somit ist die Summe zweier ZIP-verteilter Zufallsvariablen nicht Zero-Inflated-Poisson-verteilt.

2.1.3 Gamma-Verteilung

Die Gamma-Verteilung, welche eine stetige Verteilung ist, wird beispielsweise zur Modellierung von Schadenhöhen verwendet. Es werden nun die Dichte und die Additivität der Gamma-Verteilung beschrieben.

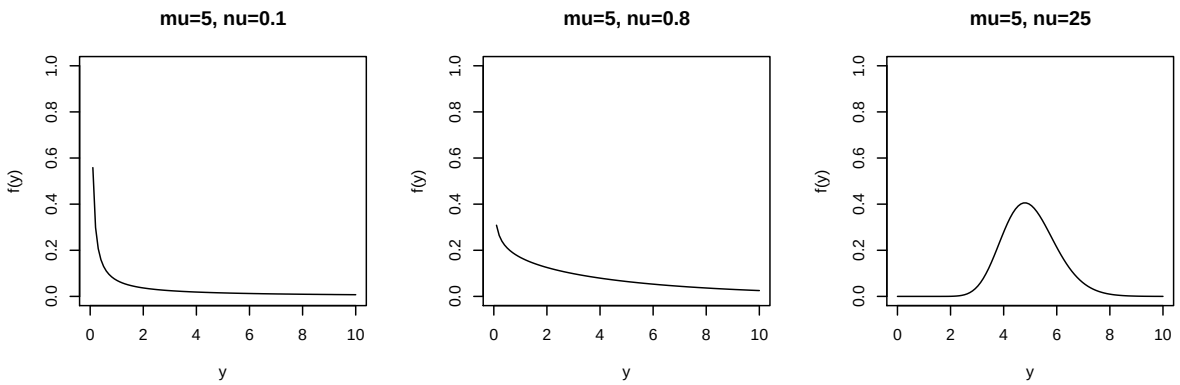
Definition 2.4 (Dichte der Gamma-Verteilung)

Eine stetige, nicht-negative Zufallsvariable Y heißt Gamma-verteilt mit Skalenparameter μ und Formparameter ν ($Y \sim \Gamma(\mu,\nu)$), falls

$$f(y) = \frac{1}{\Gamma(\nu)}\left(\frac{\nu}{\mu}\right)^{\nu} y^{(\nu-1)}\exp(-\frac{\nu y}{\mu}) \quad \textit{mit} \quad y > 0,\ \mu > 0,\ \nu > 0$$

gilt, siehe Fahrmeier and Kneib (2007), Seite 460.

Abbildung 2.2: Darstellung der Dichte der Gamma-Verteilung für $\mu = 5$ und verschiedene Werte von ν



Die Dichte der Gamma-Verteilung für verschiedene Werte des Formparameters ν zeigt Abbildung 2.2. Man kann erkennen, dass für $0 < \nu < 1$ die Hauptmasse der Dichte nahe Null liegt. Je größer der Formparameter ν wird, desto symmetrischer wird die Dichte, siehe Bickel and Doksum (1977), Seite 14.

Für Erwartungswert und Varianz der Gamma-Verteilung gilt:

$$E(Y) = \mu, \qquad Var(Y) = \frac{\mu^2}{\nu}.$$

Um die Additivtätseigenschaften von Gamma-verteilten Zufallsvariablen zu untersuchen, wird die Momenterzeugende Funktion benötigt.

Definition 2.5 (Momenterzeugende Funktion der Gamma-Verteilung)

Sei Y eine Gamma-verteilte Zufallsvariable mit $Y \sim \Gamma(\mu, \nu)$, dann ist die zugehörige Momenterzeugende Funktion gegeben durch:

$$m_Y(t) = E(e^{tY}) = \frac{1}{\left(1 - \frac{\mu t}{\nu}\right)^{\nu}}$$

siehe Mack (2002), Abschnitt 1.3.3.

Satz 2.6 (Additivität der Gamma-Verteilung)

Seien Y_1 und Y_2 unabhängige Zufallsvariablen mit $Y_1, Y_2 \sim \Gamma(\mu, \nu)$, dann gilt für

$Z = Y_1 + Y_2$ und $X = Z/2$:

$$Z \sim \Gamma(2\mu, 2\nu) \quad \textit{und} \quad X \sim \Gamma(\mu, 2\nu).$$

Beweis:

Da Y_1 und Y_2 unabhängig sind, gilt:

$$m_Z(t) = m_{Y_1}(t) m_{Y_2}(t).$$

Also folgt, dass die Summe zweier Gamma-verteilter Zufallsvariablen Gamma-verteilt ist mit

$$m_Z(t) = \frac{1}{\left(1-\frac{\mu t}{\nu}\right)^{\nu}} \frac{1}{\left(1-\frac{\mu t}{\nu}\right)^{\nu}} = \frac{1}{\left(1-\frac{\mu t}{\nu}\right)^{2\nu}} = \frac{1}{\left(1-\frac{2\mu t}{2\nu}\right)^{2\nu}}.$$

Weiter ist X Gamma-verteilt mit Skalenparameter μ und Formparameter 2ν, $X \sim \Gamma(\mu, 2\nu)$, da

$$m_X(t) = \frac{1}{\left(1-\frac{2\mu}{2\nu}\frac{t}{2}\right)^{2\nu}} = \frac{1}{\left(1-\frac{\mu t}{2\nu}\right)^{2\nu}}$$

gilt.

2.1.4 Zero-Inflated-Gamma-Verteilung

Es wird nun die Dichte der Zero-Inflated-Gamma-Verteilung (ZIG-Verteilung) betrachtet und die Verteilung der Summe zweier ZIG-verteilter Zufallsvariablen.

Definition 2.7 (Dichte der ZIG-Verteilung)

Sei Y eine Zufallsvariable mit stetiger Verteilung. Dann ist Y Zero-Inflated-Gamma-verteilt mit Parameter μ, ν und ω ($Y \sim ZIG(\mu, \nu, \omega)$), falls

$$f_Y(y) = \begin{cases} \omega & \text{für } y = 0 \\ (1-\omega)\frac{1}{\Gamma(\nu)}\left(\frac{\nu}{\mu}\right)^{\nu} y^{(\nu-1)} \exp\left(-\frac{\nu y}{\mu}\right) & \text{für } y > 0 \end{cases}$$

gilt, siehe Belasco (2002), Seite 75.

Die Gamma-Verteilung ist nur für positive Werte definiert. Will man dennoch Nullen erzeugen, so kann man wieder auf die Klasse der Zero-Inflated-Verteilungen zurückgreifen, in diesem Fall auf die Zero-Inflated-Gamma-Verteilung (ZIG-Verteilung). Auch hier gibt es einen Auswahlparameter $\omega \in [0, 1]$. Es gilt, dass die Zufallsvariable Y den Wert Null mit Wahrscheinlichkeit ω annimmt. Mit Wahrscheinlichkeit $(1-\omega)$ ist die Zufallsvariable Y Gamma-verteilt mit den Parametern μ und ν.

Seien nun Y_1, Y_2 zwei unabhängige Zufallsvariablen mit $Y_1 \sim ZIG(\mu_1, \nu, \omega)$ und $Y_2 \sim ZIG(\mu_2, \nu, \omega)$, dann gilt für die Summe der Zufallsvariablen $Y_1 + Y_2$:

$$f_{Y_1+Y_2}(y_1,y_2) = \begin{cases} \omega^2 & \text{für } y_1, y_2 = 0 \\ \omega(1-\omega)\frac{1}{\Gamma(\nu)}\left(\frac{\nu}{\mu_1}\right)^{\nu} y_1^{(\nu-1)} \exp\left(-\frac{\nu y_1}{\mu_1}\right) & \text{für } y_1 > 0, y_2 = 0 \\ (1-\omega)^2 \frac{1}{\Gamma(\nu)^2}\left(\frac{\nu}{\mu_1}\frac{\nu}{\mu_2}\right)^{\nu} (y_1\ y_2)^{(\nu-1)} e^{\left(-\nu\left(\frac{y_1}{\mu_1}+\frac{y_2}{\mu_2}\right)\right)} & \text{für } y_1, y_2 > 0. \end{cases}$$

Somit ist die Summe von zwei ZIG-verteilten Zufallsvariablen nicht Zero-Inflated-Gamma-verteilt.

2.1.5 Pareto-Verteilung

Abschließend wird die Dichte der Pareto-Verteilung und die Verteilung der Summe zweier Pareto-verteilter Zufallsvariablen angegeben.

Definition 2.8 (Dichte der Pareto-Verteilung)

Eine stetige, nicht-negative Zufallsvariable Y heißt Pareto-verteilt mit Parametern a und k ($Y \sim Par(a,k)$), falls

$$f_Y(y) = \frac{ak^a}{y^{a+1}}$$

für $a > 0, k > 0, y \geq k$ gilt, siehe Bickel and Doksum (1977), Seite 82.

Seien Y_1 und Y_2 unabhängige Zufallsvariablen mit $Y_1 \sim Par(a_1,k_1)$ *und* $Y_2 \sim Par(a_2,k_2)$, dann gilt für die Summe der Zufallsvariablen $Y_1 + Y_2$:

$$\begin{aligned} f_{Y_1+Y_2}(y) &= \int_{k_2}^{\infty} f(y-z)g(z)dz \\ &= \int_{k_2}^{\infty} \frac{a_1 k_1^{a_1}}{(y-z)^{a_1+1}} \frac{a_2\ k_2^{a_2}}{z^{a_2+1}} dz \\ &= \frac{a_1 a_2\ k_1^{a_1} k_2^{a_2}}{y^{a_1+1}} \int_{k_2}^{\infty} \frac{-1}{z^{a_1+a_2+2}}\ dz \\ &= \frac{a_1 a_2\ k_1^{a_1} k_2^{a_2}}{y^{a_1+1}} \left(\frac{-1}{z^{a_1+a_2+1} a_1 + a_2 + 1} \Big|_{k_2}^{\infty} \right) \\ &= \frac{a_1 a_2 k_1^{a_1}}{k_2^{a_1+1} y^{a_1+1}(a_1+a_2+1)}. \end{aligned}$$

Somit ist die Summe von zwei Pareto-verteilten Zufallsvariablen nicht Pareto-verteilt.

2.2 Exponentielle Familie

In diesem Abschnitt wird die Klasse der exponentiellen Familie definiert und es werden einige Eigenschaften vorgestellt. Anschließend wird die Zugehörigkeit der Poisson- und Gamma-Verteilung zur exponentiellen Familie nachgewiesen, siehe Lindsey (1997), Abschnitt 1.2.

Definition 2.9 (Einparametrige exponentielle Familie)

Die Dichte $f(y;\theta,\phi)$ einer Zufallsvariable Y gehört zur einparametrigen exponentiellen Familie, falls sie sich mit natürlichem oder kanonischem Parameter θ und Skalen- oder Dispersionsparameter ϕ in folgender Form darstellen lässt:

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

Die Funktionen $a(\cdot)$, $b(\cdot)$ und $c(\cdot)$ spezifizieren dabei die jeweilige Verteilung.

Man bezeichnet

$$l(y;\theta,\phi) := \log f(y;\theta,\phi) = \frac{\theta y - b(\theta)}{a(\phi)} + c(y,\phi)$$

als Log-Likelihood Funktion.

Für den Erwartungswert der abgeleiteten Log-Likelihood-Funktion einer exponentiellen Familie gilt Folgendes:

Satz 2.10

Unter den Regularitätsbedingungen gilt:

$$E\left(\frac{\partial l(y,\theta,\phi)}{\partial\theta}\right) = 0 \tag{2.1}$$

$$E\left(\frac{\partial^2 l(y,\theta,\phi))}{\partial\theta^2}\right) + E\left(\left[\frac{\partial l(y,\theta,\phi)}{\partial\theta}\right]^2\right) = 0. \tag{2.2}$$

Beweis:

Zu (2.1):

$$E\left(\frac{\partial l(y,\theta,\phi)}{\partial\theta}\right) \;=\; E\left(\frac{\partial \log f(y,\theta,\phi)}{\partial\theta}\right) = E\left(\frac{1}{f(y,\theta,\phi)}\frac{\partial f(y,\theta,\phi)}{\partial\theta}\right)$$

$$
\begin{aligned}
&= \int f(y,\theta,\phi)\frac{1}{f(y,\theta,\phi)}\frac{\partial f(y,\theta,\phi)}{\partial\theta}\ dy \\
&= \frac{\partial}{\partial\theta}\int f(y,\theta,\phi)\ dy = \frac{\partial}{\partial\theta}\ 1 = 0
\end{aligned}
$$

Zu (2.2):

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

Mit Hilfe von Satz 2.10 kann man Erwartungswert und Varianz der Zufallsvariable Y direkt aus der Parametrisierung für exponentielle Familien ableiten. Es gilt:

$$
\begin{aligned}
E(Y) &= b'(\theta) \\
Var(Y) &= b''(\theta)\ a(\phi).
\end{aligned}
$$

Denn

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

und

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

$$
\begin{aligned}
\Leftrightarrow\quad & \frac{-b''(\theta)}{a(\phi)}) + \frac{1}{a(\phi)^2}E((Y-b'(\theta))^2) = 0 \\
\Leftrightarrow\quad & E((Y-E(Y))^2) = b''(\theta)\ a(\phi) \\
\Leftrightarrow\quad & Var(Y) = b''(\theta)\ a(\phi).
\end{aligned}
$$

Man bezeichnet $V(\theta) := b''(\theta)$ als *Varianzfunktion.*

Die beiden folgenden Beispiele weisen nach, dass die Gamma-Verteilung und die Poisson-Verteilung zur Klasse der exponentiellen Familie gehören.

Beispiel 2.1 (Poisson-Verteilung als exponentielle Familie)

Die Wahrscheinlichkeitsfunktion der Poisson-Verteilung kann formuliert werden als

$$f(y;\theta,\phi) = \exp\left(-\lambda + y\log(\lambda) - \log(y!)\right)$$

mit

$$
\begin{aligned}
a(\phi) &= 1 \\
\theta &= \log(\lambda) \\
b(\theta) &= \lambda = e^{\log(\lambda)} = e^{\theta} \\
c(y,\phi) &= -\log(y!).
\end{aligned}
$$

Somit gehört die Poisson-Verteilung zur exponentiellen Familie.

Für den Erwartungswert und die Varianz der Zufallsvariable Y gilt:

$$E(Y) = b'(\theta) = \lambda \quad \textit{und} \quad Var(Y) = b''(\theta)\ a(\phi) = e^{\theta}1 = \lambda.$$

Beispiel 2.2 (Gamma-Verteilung als exponentielle Familie)

Die Dichte der Gamma-Verteilung kann formuliert werden als

$$f(y;\theta,\phi) = exp\left\{\nu\left(-\frac{y}{\mu} - \log(\mu)\right) + \nu\log(y) - \nu\log(\nu y) - \log\Gamma(\nu)\right\}$$

mit

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

Somit gehört auch die Gamma-Verteilung zur exponentiellen Familie.

Für den Erwartungswert und die Varianz der Zufallsvariable Y gilt:

$$E(Y) = b'(\theta) = -\frac{1}{\theta} = \mu \quad und \quad Var(Y) = b''(\theta)a(\phi) = \frac{\mu^2}{\nu}.$$

Im folgenden Abschnitt werden die Komponenten des Generalisierten Linearen Modells beschrieben.

2.3 Generalisiertes Lineares Modell

In diesem Abschnitt wird das Generalisierte Lineare Modell (GLM) vorgestellt. Wie die Bezeichnung *Generalisiertes Lineares Modell* bereits andeutet, stellt diese Modellklasse eine Verallgemeinerung des klassischen linearen Modells dar. Hier wird die Beschränkung auf eine normalverteilte Responsevariable ebenso verallgemeinert wie der lineare Zusammenhang zwischen Zielvariable und erklärender Variable, siehe Nelder and McCullagh (1991).

Für die Responsevariable werden nun alle Verteilungen zugelassen, die der exponentiellen Familie angehören, wie zum Beispiel die Binomial-, Poisson-, Negativ-Binomial- oder Gamma-Verteilung. Eine Linkfunktion ermöglicht die Transformation des Erwartungswerts. Dies bietet im Vergleich zu den klassischen Linearen Modellen ebenfalls mehr Freiheiten.

Zu Beginn des Kapitels werden die Komponenten eines GLMs aufgeführt. Im nächsten Schritt wird beschrieben wie Modelle an die Daten angepasst werden und welche Hypothesentests geeignet sind, um die Modelle zu testen. Daraufhin wird aufgezeigt, durch welche Residuen die Modellannahmen überprüft werden können.

Die Poisson- und die Gamma-Verteilung werden in den folgenden Kapiteln dieser Arbeit als Verteilung der Zielvariable diskutiert. Aus diesem Grund wird dieser Abschnitt mit einer detaillierten Beschreibung der Poisson-Regression mit Offset und Gewicht und der Gamma-Regression mit Offset und Gewicht abgeschlossen.

2.3.1 Komponenten des Generalisierten Linearen Modells

Ein Generalisiertes Lineares Modell besteht aus einer zufälligen Komponente, einer systematischen Komponente und einer zugehörigen Linkfunktion, siehe Myers and Montgomery (2001).

Man bezeichnet die Realisierung einer Zufallsvariable Y_i als einen Vektor $\mathbf{y} = (y_1, \ldots, y_n)^T$ bestehend aus n Beobachtungen. Weiterhin sind n Kovariablenvektoren $\mathbf{x_i}$ gegeben mit $\mathbf{x_i} = (x_{i1}, \ldots, x_{in})$ für $i = 1, \ldots, n$.

Die *zufällige Komponente* eines GLMs sind die Zufallsvariablen $Y_i, i = 1, \ldots, n$. Die Y_i sind unabhängig und die Verteilung gehört zur Klasse der exponentiellen Familie. Für den Erwartungswert gilt:

$$E\left(Y_i\right) =: \mu_i.$$

Der lineare Prädiktor $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_p)$ ist die *systematische Komponente* im Modell, welche definiert ist durch

$$\boldsymbol{\eta} := \boldsymbol{X}\beta \quad \text{mit} \quad \boldsymbol{X} = \begin{pmatrix} \mathbf{x_1} \\ \vdots \\ \mathbf{x_n} \end{pmatrix} \in \mathbb{R}^{n \times p}.$$

Die Verbindung von zufälliger und systematischer Komponente wird durch eine Transformation des Erwartungswerts geschaffen. Dazu beschreibt man den linearen Prädiktor $\boldsymbol{\eta}$ als Funktion des Erwartungswerts.

Diese Funktion $g : \mathbb{R} \to \mathbb{R}$ ist bekannt und wird als *Linkfunktion* bezeichnet, d.h. es gilt:

$$g(\mu_i) = \eta_i = \mathbf{x_i^T}\boldsymbol{\beta}.$$

Die Linkfunktion g heißt *kanonisch*, wenn $\theta_i = \eta_i$ ist.

2.3.2 Schätzung der Modellparameter

Dieses Kapitel befasst sich mit der Schätzung von Modellparametern, siehe Lindsey (1997), Fahrmeir and Tutz (2001) und Bickel and Doksum (1977). Es wird beschrieben, welche Methode bei GLMs verwendet wird, um die Modellparameter zu schätzen. Ebenso wird auf das nummerische Verfahren verwiesen, das bei der Schätzung Anwendung findet. Zudem wird beschrieben, welche Eigenschaften der so entstandene Schätzer hat.

Die *Maximum-Likelihood-Methode* ist die theoretische Basis für die Parameterschätzung bei GLMs. Bei dieser Methode wird vorausgesetzt, dass der Typ der Verteilung bekannt ist. Anstatt μ_i, der Erwartungswert der Zufallsvariable Y_i, wird $\boldsymbol{\beta}$ mit $\eta_i = \boldsymbol{x}_i^T\boldsymbol{\beta} = g(\mu_i)$ geschätzt.

Die zugehörigen Schätzer $\hat{\boldsymbol{\beta}}$ werden als *Maximum-Likelihood-Schätzer* bezeichnet. Um den ML-Schätzer zu erhalten, wird die Log-Likelihood-Funktion

$$L(\boldsymbol{\beta}) := \prod_{i=1}^{n} l(y_i, \mu_i, \phi) = \sum_{i=1}^{n} \log f(y_i, \mu_i, \phi)$$

maximiert.

Die Ableitung der Log-Likelihood Funktion wird als *Score-Funktion* bezeichnet.

Definition 2.11 (Score-Funktion)

Die Score-Funktion ist der Vektor der Ableitungen der Log-Likelihood Funktion nach $\boldsymbol{\beta}$ und wird definiert als

$$S(\boldsymbol{\beta}) := \frac{\partial L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}.$$

Gehört die Dichte f_i zur exponentiellen Familie, dann gilt:

$$S(\boldsymbol{\beta}) = \frac{\partial L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \frac{1}{a(\phi)} \left[y_i - \frac{\partial b(\beta_i)}{\partial \beta_i} \right] x_i.$$

Um das Maximum zu finden wird die Score-Funktion gleich Null gesetzt. Man findet also einen Schätzer für den Vektor der Regressionsparameter $\boldsymbol{\beta}$ durch das Lösen des Systems von *Score-Gleichungen* :

$$\sum_{i=1}^{n} \frac{1}{a(\phi)} \left[y_i - \mu_i\right] x_i = 0.$$

Da keine analytische Lösung dieser Gleichungen existiert, wird die *Newton-Raphson-Methode* bzw. das *Fisher-Scoring-Verfahren* verwendet.

Der vollständige Lösungsalgorithmus wird als *Iterative Weighted Least Squares* (IWLS) Algorithmus bezeichnet, detaillierte Beschreibung siehe Lindsey (1997), Seite 199ff.

Der so entstandene Schätzer $\hat{\boldsymbol{\beta}}$ hat die Eigenschaften, dass er konsistent und asymptotisch normalverteilt für $\boldsymbol{\beta}$ ist, siehe Fahrmeir and Tutz (2001), Satz 2.4, mit

$$\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, I^{-1}(\boldsymbol{\beta})).$$

Dabei ist $I(\boldsymbol{\beta})$ folgendermaßen definiert:

Definition 2.12 (Fisher-Informationsmatrix)

Falls die Regularitätsbedingungen gelten, heißt

$$I(\boldsymbol{\beta}) := E\left(S(\boldsymbol{\beta})S(\boldsymbol{\beta})^T\right) = Cov(S(\boldsymbol{\beta}))$$

die erwartete Fisher-Informationsmatrix, wobei

$$I(\boldsymbol{\beta})_{OBS} = -\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_s \partial \beta_r}$$

als die beobachtete Fisher-Informationsmatrix bezeichnet wird. Somit ergibt sich:

$$I(\boldsymbol{\beta}) = -E\left(\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_s \partial \beta_r}\right)$$

siehe Nelder and McCullagh (1991), Seite 306.

Unter gewissen Voraussetzungen gilt auch, dass der Maximum-Likelihood Schätzer (ML-Schätzer) für Verteilungen der exponentiellen Familie eindeutig ist, siehe Bickel and Doksum (1977).

Satz 2.13 (ML-Schätzung in der exponentiellen Familie)

Sei die Dichte oder die Wahrscheinlichkeitsfunktion einer Verteilung parametrisierbar als

$$f(y,\theta) = \exp\{c(\theta)T(x) + d(\theta) + S(x)\} \quad \textit{für} \quad x \in A, \theta \in \Theta,$$

wobei $T(x)$ eine suffiziente Statistik ist. Sei C das Innere vom Bild $c(\theta)$ und sei $c(\theta)$ injektiv. Falls

$$E(T(\boldsymbol{X})) = T(\boldsymbol{x})$$

eine Lösung $\hat{\theta}(\boldsymbol{x})$ besitzt mit $c(\hat{\theta}(\boldsymbol{x})) \in C$, dann ist $\hat{\theta}(\boldsymbol{x})$ der eindeutige ML-Schätzer von θ.

Eine Statistik $T(x)$ heißt *suffizient* für θ genau dann, wenn $X|T(X) = t$ unabhängig von θ ist.

2.3.3 Goodness of Fit

Um die Anpassungsgüte des Modells an die Daten zu überprüfen gibt es verschiedene Goodness of Fit Maße, wie das Akaike Informationskriterium und die Devianz.

Bei GLMs kann man die Log-Likelihood-Funktion auf zwei Weisen formulieren. Es gilt mit $\eta_i = g(\mu_i),\ \theta_i = h(\mu_i)$ für $i = 1, ..., n$:

$$\begin{aligned} l(\boldsymbol{\beta}, \boldsymbol{y}, \phi) &= \sum_{i=1}^{n} \left[\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} - c(y_i, \phi) \right] \\ &= \sum_{i=1}^{n} \left[\frac{y_i h(\mu_i) - b(h(\mu_i))}{a(\phi)} - c(y_i, \phi) \right] \\ &= l(\boldsymbol{\mu}, \boldsymbol{y}, \phi). \end{aligned}$$

Das Akaike Informationskriterium wird zur Variablenselektion verwendet. Es bestimmt welches Modell besser an die Daten angepasst ist.

Definition 2.14 (Akaike Informationskriterium)

Der AIC-Wert (Akaike Information Criterion) ist definiert durch

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

siehe de Jong and Heller (2008), Seite 62f.

Dabei ist $\hat{\boldsymbol{\beta}}$ der Vektor der geschätzten Regressionsparameter und k die Anzahl der Parameter im Modell.

Je kleiner der AIC-Wert ist, desto besser ist das Modell. Einerseits gilt, je mehr Parameter in das Modell aufgenommen werden desto kleiner wird der Wert des negativen Log-Likelihood. Anderseits wird für jeden zusätzlichen Parameter der Strafterm $2k$ größer und somit sinkt der Wert des AIC nur dann, wenn der Log-Likelihood stärker abnimmt als der Strafterm zunimmt.

Ein weiteres Kriterium für die Modellwahl ist die Devianz. Auch hier ist es möglich anhand des Wertes der Devianz zu entscheiden, welches Modell die Daten besser anpasst.

Der Wert der Devianz berechnet sich aus:

$$-2\ [\ l(\hat{\boldsymbol{\beta}}, \boldsymbol{y}, \phi) - l(\boldsymbol{y}, \boldsymbol{y}, \phi)\ ],$$

wobei $l(\boldsymbol{y}, \boldsymbol{y}, \phi)$ der maximale Log-Likelihood ist, der auch als Log-Likelihood des saturierten Modells bezeichnet werden kann, siehe Lindsey (1997).

In der Schreibweise der exponentiellen Familie erhält man für die Devianz:

$$-2\ [\ l(\hat{\boldsymbol{\beta}}, \boldsymbol{y}, \phi) - l(\boldsymbol{y}, \boldsymbol{y}, \phi)\ ] = 2\sum_{i=1}^{n} \frac{y_i(\hat{\theta}_i - \tilde{\theta}_i) - b(\hat{\theta}_i) + b(\tilde{\theta}_i)}{a_i(\phi)},$$

wobei der Hut (ˆ) für das saturierte Modell steht. Für $a_i(\phi) = \phi/w_i$ ist die Devianz in der GLM-Terminologie gegeben durch:

Definition 2.15 (Devianz)

Der Wert der Devianz wird berechnet mit:

$$\frac{D(\boldsymbol{y}, \hat{\boldsymbol{\mu}})}{\phi} := 2\sum_{i=1}^{n} \nu\ w_i\ \left[\ y_i(\hat{\theta}_i - \tilde{\theta}_i) - b(\hat{\theta}_i) + b(\tilde{\theta}_i)\ \right]$$

mit Gewichten w_i, siehe Lindsey (1997), Seite 210.

Somit ist die Maximierung des Log-Likelihoods äquivalent zur Minimierung der Devianz.

2.3.4 Partial Devianz Test und Residual Devianz Test

Dieser Abschnitt beschäftigt sich mit dem Residual Devianz Test und dem Partial Devianz Test, siehe Myers and Montgomery (2001).

Der Residual Devianz Test misst wie gut das Modell M an die beobachteten Daten angepasst ist. Hierzu wird der Quotient aus dem Log-Likelihood des Modells M und dem Log-Likelihood des saturierten Modells gebildet.

Sei nun M ein Modell mit Devianz $D(\boldsymbol{y}, \hat{\boldsymbol{\mu}})/\phi$, wobei $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta},\ \boldsymbol{\beta} \in \mathbb{R}^p,\ \boldsymbol{X} \in \mathbb{R}^{n \times p}$.

Die Hypothese des **Residual Devianz Tests**

ist gegeben durch:

$$\boldsymbol{H}:\ \boldsymbol{\eta} = g(\boldsymbol{\mu}) = \boldsymbol{X}\boldsymbol{\beta} \quad \text{versus} \quad \boldsymbol{K}:\ \text{nicht } \boldsymbol{H}$$

Die Nullhypothese $\boldsymbol{H}$ wird genau dann zum Niveau α verworfen, wenn

$$\frac{D(\boldsymbol{y}, \hat{\boldsymbol{\mu}})}{\phi} > \chi^2_{n-p,1-\alpha}.$$

Will man zwei genestete Modelle vergleichen, dann verwendet man den Partial Devianz Test. Unter genesteten Modellen versteht man Modelle M_1 und M_2 mit $M_2 \subseteq M_1$, d.h. die Kovariablen von Modell M_2 sind auch in Modell M_1 enthalten.

Seien nun M_1 und M_2 genestete Modelle, wobei $D_{M_1}(\boldsymbol{y}, \hat{\boldsymbol{\mu}}_1)/\phi_{M_1}$ die Devianz von Modell M_1 ist mit $\boldsymbol{\eta} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2,\ \boldsymbol{\beta}_1 \in \mathbb{R}^{p_1},\ \boldsymbol{\beta}_2 \in \mathbb{R}^{p_2},\ \boldsymbol{X}_1 \in \mathbb{R}^{n \times p_1},\ \boldsymbol{X}_2 \in \mathbb{R}^{n \times p_2}$. $D_{M_2}(\boldsymbol{y}, \hat{\boldsymbol{\mu}}_2)/\phi_{M_2}$ ist die Devianz von Modell M_2 mit $\boldsymbol{\eta} = \boldsymbol{X}_1\boldsymbol{\beta}_1$.

Die Hypothese des **Partial Devianz Tests**

ist gegeben durch:

$$\boldsymbol{H}: \ \boldsymbol{\beta}_2 = 0 \quad \text{versus} \quad \boldsymbol{K}: \ \boldsymbol{\beta}_2 \neq 0$$

Die Nullhypothese $\boldsymbol{H}$ wird zum Niveau α genau dann verworfen, wenn

$$\frac{D_{M_1}(\boldsymbol{y}, \hat{\boldsymbol{\mu}}_1) - D_{M_2}(\boldsymbol{y}, \hat{\boldsymbol{\mu}}_2)}{\phi_{M_1}} > \chi^2_{p_2, 1-\alpha}.$$

Ist ϕ_{M_1} unbekannt, dann wird $\hat{\phi}_{M_1}$ verwendet.

Ein Schätzer $\hat{\phi}$ für den Dispersionsparameter ϕ ist gegeben durch:

$$\hat{\phi} = \frac{1}{n-p} \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{V(\mu_i)},$$

wobei $V(\mu_i)$ die Varianzfunktion an der Stelle μ_i ist.

2.3.5 Pearson Residuen und Devianz Residuen

Residuen sind ein wichtiges statistisches Werkzeug um herauszufinden wie gut ein Modell die Daten anpasst. Beim Generalisierten Linearen Modell werden die Residuen unter anderem genutzt um Ausreißer zu finden.

Es werden nun zwei Typen von standardisierten Residuen beschrieben, die Pearson Residuen und die Devianz Residuen, siehe Lindsey (1997).

Die *Pearson Residuen* r_i^P werden berechnet als Abweichung zwischen den Daten und dem Erwartungswert gewichtet mit der Wurzel der Varianzfunktion. Die Pearson Residuen werden definiert durch:

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

Die *Devianz Residuen* r_i^D können beschrieben werden als der Anteil einer einzelnen Beobachtung an der Devianz. Für die Devianz Residuen gilt:

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

mit

$$sign(x) = \begin{cases} 1, & x \geq 0 \\ -1, & x < 0 \end{cases}$$

und $D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = \sum_{i=1}^{n} d_i$.

Wenn die beobachteten Daten keine einheitlich Varianz besitzen, betrachtet man anstatt der Pearson Residuen die Standardisierten Pearson Residuen und anstatt der Devianz Residuen die Standardisierten Devianz Residuen. Wenn w_i das Gewicht der Beobachtung y_i bezeichnet, dann werden die *Standardisierten Pearson Residuen* $\tilde{r}_i^P$ definiert durch:

$$\tilde{r}_i^P = \frac{r_i^P}{\sqrt{w_i}} = \frac{y_i - \hat{\mu}_i}{\sqrt{w_i\ V(\hat{\mu}_i)}}.$$

Für die *Standardisierten Devianz Residuen* $\tilde{r}_i^D$ gilt:

$$\tilde{r}_i^D = \frac{r_i^D}{\sqrt{w_i}} = sign(y_i - \hat{\mu}_i)\ \sqrt{\frac{d_i}{w_i}}.$$

2.3.6 Poisson-Regression mit Offset und Gewicht

Bisher ist man davon ausgegangen, dass die Rate, mit der die Ereignisse bei Poisson-verteilten Zufallsvariablen Y_i eintreten, konstant ist. Wenn aber die Rate von einer zusätzlichen Variable S_i abhängt, dann ist die Wahrscheinlichkeitsfunktion der Zufallsvariable Y_i, $i = 1, ..., n$ gegeben durch:

$$P(Y_i = y_i)^{S_i} = \left[\exp(-\lambda_i S_i)\frac{(\lambda_i S_i)^{y_i}}{y_i!}\right]^{S_i} \quad \text{mit } \lambda_i, S_i > 0 \ \text{ und } \ y_i = 0, 1, 2, ... \ .$$

Stellt man die Wahrscheinlichkeitsfunktion in der Form der exponentiellen Familie dar, dann gilt für i=1,...,n:

$$f(y_i;\theta_i,\phi) = \exp\left\{S_i\left[-\lambda_i S_i + y_i \log(\lambda_i S_i) - \log(y_i!)\right]\right\}$$

mit

$$\begin{aligned} a(\phi) &= \frac{1}{S_i} \\ \theta_i &= \log(\lambda_i S_i) \\ b(\theta_i) &= \lambda_i S_i = e^{\log(\lambda_i S_i)} = e^{\theta_i} \\ c(y_i,\theta_i) &= -\log(y_i!). \end{aligned}$$

Dabei wird S_i als *Gewicht* bezeichnet und dient zur Stabilisierung der Varianz, da durch das Gewicht S_i gilt:

$$Var(Y_i) = b''(\theta_i)a(\phi) = \lambda_i,\ i = 1,...,n.$$

Aus der Darstellung der Wahrscheinlichkeitsfunktion $f(y_i;\theta_i,\phi)$ als exponentielle Familie und mit $\lambda_i = \exp(\boldsymbol{x}_i^T\boldsymbol{\beta})$ folgt, dass

$$\mu_i = E(Y_i) = b'(\theta_i) = \exp(\theta_i) = S_i\lambda_i = S_i\exp(\boldsymbol{x}_i^T\boldsymbol{\beta}),\ i = 1,...,n.$$

Somit ergibt sich für den linearen Prädiktor $\boldsymbol{\eta}$, dass

$$\eta_i = \boldsymbol{x}_i^T\boldsymbol{\beta} = \log(\mu_i) - \log(S_i) = \theta_i - \log(S_i),\ i = 1,...,n.$$

Man bezeichnet den zweiten Term $\log(S_i)$ als *Offset* und setzt ihn als bekannt voraus. Der Offset kann als zusätzliche Regressionsvariable betrachtet werden. Somit erhält man für $S_i = 1$ die kanonische Linkfunktion $g(\mu_i) = \log(\mu_i)$. Ist $S_i \neq 1$, dann gilt:

$$g(\mu_i) = \log(\mu_i) - \log(S_i),\ i = 1,...,n.$$

Anstelle des Log-Likelihood betrachtet man den *Gewichteten Log-Likelihood*. Der Gewichtete Likelihood ist wie folgt definiert:

Definition 2.16 (Gewichteter Likelihood)

Seien Y_i mit $i = 1,...,n$ unabhängige Zufallsvariablen mit zugehörigen Gewichten w_i und Wahrscheinlichkeitsfunktion oder Dichte f_i, dann ist der Gewichtete Likelihood definiert

als

$$WL(y_i, w_i) = \prod_{i=1}^{n} f_i(y_i|\theta_i)^{w_i}$$

siehe Hu and Zidek (2002).

Somit kann man die Gewichteten Log-Likelihood-Funktion der Poisson-Regression mit Offset und Gewicht für $i = 1, ..., n$ darstellen als:

$$W(\boldsymbol{\beta}) = wl(\boldsymbol{y}; \boldsymbol{\beta}, \phi) = \sum_{i=1}^{n} \left(S_i \left[-S_i \exp(\boldsymbol{x}_i^T \boldsymbol{\beta}) + y_i \log(S_i \exp(\boldsymbol{x}_i^T \boldsymbol{\beta})) - \log(y_i!) \ \right] \right).$$

In diesem Fall ist S_i sowohl der Offset als auch das Gewicht.

Für die ML-Schätzung von $\boldsymbol{\beta}$ benötigt man die Score-Gleichungen. Die Score-Funktion für dieses Modell ist für $i = 1, ..., n$ gegeben durch:

$$\begin{aligned} S_i(\boldsymbol{\beta}) &= \frac{\partial W(\boldsymbol{\beta})}{\partial \beta_j} \\ &= \sum_{i=1}^{n} \left(S_i \left[-S_i \exp(\boldsymbol{x}_i^T \boldsymbol{\beta}) \ x_{ij} + y_i \ \frac{S_i \exp(\boldsymbol{x}_i^T \boldsymbol{\beta})}{S_i \exp(\boldsymbol{x}_i^T \boldsymbol{\beta})} \ x_{ij} \right] \right) \\ &= \sum_{i=1}^{n} S_i \left(y_i - S_i \exp(\boldsymbol{x}_i^T \boldsymbol{\beta}) \right) x_{ij}. \end{aligned}$$

Somit erhält man folgende Score-Gleichungen:

$$S(\boldsymbol{\beta}) = \begin{pmatrix} S_1(\boldsymbol{\beta}) \\ \vdots \\ S_n(\boldsymbol{\beta}) \end{pmatrix} = \boldsymbol{X}^T (\boldsymbol{Y} - \boldsymbol{\lambda}) = 0.$$

Für die Poisson-Regression mit Offset und Gewicht ist, wie bei allen Regressionsverfahren, die Wahl eines geeigneten Modells von zentraler Bedeutung. In Abschnitt 2.3.3 wurden bereits zwei Modellwahlkriterien beschrieben, der AIC-Wert und die Devianz. Für dieses Modell berechnet man den AIC-Wert folgendermaßen:

$$AIC = -2\sum_{i=1}^{n} S_i \left[y_i \log(\hat{\lambda}_i S_i) - \hat{\lambda}_i S_i - \log(y_i!) \right] + 2p,$$

wobei p die Anzahl der geschätzten Parameter ist. Die Devianz berechnet man aus:

$$\begin{aligned} D(\boldsymbol{y}, \hat{\boldsymbol{\beta}}) &= -2\left[wl(\boldsymbol{y}, \hat{\boldsymbol{\beta}}, \phi) - wl(\boldsymbol{y}, \boldsymbol{y}, \phi) \right] \\ &= -2\sum_{i=1}^{n} S_i \left[y_i \log(\hat{\lambda}_i S_i) - \hat{\lambda}_i S_i - y_i \log(y_i) + y_i \right] \\ &= -2\sum_{i=1}^{n} S_i \left[y_i \log\left(\frac{\hat{\lambda}_i S_i}{y_i} \right) - (\hat{\lambda}_i S_i - y_i) \right]. \end{aligned}$$

Der Residual Devianz Test und der Partial Devianz Test können ebenfalls zur Wahl eines Modells verwendet werden.

Die Hypothese des **Residual Devianz Tests**

für die Poisson-Regression mit Offset und Gewicht ist gegeben durch:

$$\boldsymbol{H}: \ \boldsymbol{\eta} = g(\boldsymbol{\mu}) = \boldsymbol{X}\boldsymbol{\beta} \quad \text{versus} \quad \boldsymbol{K}: \ \text{nicht } \boldsymbol{H}$$

Die Teststatistik $\boldsymbol{L}$ berechnet man mit

$$\boldsymbol{L} = D(\boldsymbol{y}, \hat{\boldsymbol{\lambda}}) = 2\sum_{i=1}^{n} S_i \left[(\hat{\lambda}_i S_i - y_i) - y_i \log\left(\frac{\hat{\lambda}_i S_i}{y_i} \right) \right].$$

Die Nullhypothese $\boldsymbol{H}$ wird zum Niveau α genau dann verworfen, wenn die Teststatistik $\boldsymbol{L}$ größer ist als $\chi^2_{n-p,1-\alpha}$.

Die Hypothese des **Partial Devianz Tests**

ist gegeben durch:

$$\boldsymbol{H}: \ \boldsymbol{\beta}_2 = 0 \quad \text{versus} \quad \boldsymbol{K}: \ \boldsymbol{\beta}_2 \neq 0.$$

Die Teststatistik $\boldsymbol{P}$ berechnet man durch

$$\boldsymbol{P} = D(\boldsymbol{y}, \hat{\boldsymbol{\lambda}}_1) - D(\boldsymbol{y}, \hat{\boldsymbol{\lambda}}),$$

wobei $\hat{\boldsymbol{\lambda}}$ der geschätzte Erwartungswertvektor im Modell $\boldsymbol{X}$ ist mit $\boldsymbol{X} = (\boldsymbol{X}_1,\ \boldsymbol{X}_2)\,,\ \boldsymbol{X} \in \mathbb{R}^p,\ \boldsymbol{X}_1 \in \mathbb{R}^{p_1},\ \boldsymbol{\lambda}^T = (\boldsymbol{\lambda}_1,\ \boldsymbol{\lambda}_2)^T\,,\ \boldsymbol{\lambda}_1 \in \mathbb{R}^{p_1},\ \boldsymbol{\lambda}_2 \in \mathbb{R}^{p_2}$. Es ist $\hat{\boldsymbol{\lambda}}_1$ der geschätzte Erwartungswertvektor im reduzierten Modell $\boldsymbol{X}_1$.

Die Nullhypothese $\boldsymbol{H}$ wird zum Niveau α genau dann verworfen, wenn die Teststatistik $\boldsymbol{P}$ größer ist als $\chi^2_{p_1,1-\alpha}$.

Für die Poisson-Regression mit Offset und Gewicht gibt es verschiedene Residuen. Die Pearson Residuen sind in diesem Fall wie folgt definiert:

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

Für die Devianz Residuen gilt:

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

wobei für die Devianz Komponenten d_i, $i = 1, ..., n$, gilt:

$$D(\boldsymbol{y}, \hat{\boldsymbol{\lambda}}) = \sum_{i=1}^{n} d_i.$$

Da die Varianz bei diesem Modell nicht einheitlich ist, betrachtet man, um die Residuen vergleichen zu können, die Standardisierten Residuen. Mit Gewichten S_i erhält man die Standardisierten Pearson Residuen:

$$\tilde{r}_i^P = \frac{r_i^P}{\sqrt{S_i}} = \frac{y_i - \hat{\lambda}_i}{\sqrt{S_i\ \hat{\lambda}_i}}.$$

Für die Standardisierten Devianz Residuen gilt:

$$\tilde{r}_i^D = \frac{r_i^D}{\sqrt{S_i}} = sign(SB_i - \hat{\lambda}_i)\ \sqrt{\frac{d_i}{S_i}}.$$

2.3.7 Gamma-Regression mit Offset und Gewicht

Ähnlich wie bei der Poisson-Regression mit Offset und Gewicht betrachtet man nun Gamma-verteilte Zufallsvariablen $Y_i,\ i = 1, ..., n$. Die Varianz dieser Zufallsvariablen sei ebenfalls nicht einheitlich und der Erwartungswert von Y_i sei von einer zusätzlichen Variable S_i abhängig. Die Dichte einer Zufallsvariable Y_i ist dann gegeben als

$$f(y_i) = \left[\frac{1}{\Gamma(\nu)}\left(\frac{\nu}{\mu_i S_i}\right)^{\nu} y_i^{(\nu-1)} \exp\left(-\frac{\nu y_i}{\mu_i S_i}\right)\right]^{w_i} \quad y_i,\ \mu_i,\ S_i,\ w_i,\ \nu > 0.$$

Stellt man die Dichte in der Form der exponentiellen Familie dar, dann gilt für $i = 1, ..., n$:

$$f(y_i; \theta_i, \phi) = \exp\left\{w_i\nu\left(-\frac{y_i}{\mu_i S_i} - \log(\mu_i S_i)\right) + w_i\nu\log(y_i) - w_i\nu\log(\nu y_i) - w_i\log\Gamma(\nu)\right\}$$

mit

$$\begin{aligned} a(\phi) &= \frac{1}{w_i\nu} \\ \theta_i &= -\frac{1}{\mu_i S_i} \\ b(\theta_i) &= \log\left(-\frac{1}{\theta_i}\right) = -\log(-\theta_i) \\ c(y_i, \phi) &= \nu\log(y_i) - \nu\log(\nu y_i) - \log\Gamma(\nu). \end{aligned}$$

Aus der Darstellung der Dichte $f(y_i; \theta, \phi)$ als exponentielle Familie und $\log(\mu_i) = \boldsymbol{x}_i^T\boldsymbol{\beta}$ folgt, dass

$$\tilde{\mu}_i = E(Y_i) = b'(\theta_i) = -\frac{1}{\theta_i} = \mu_i S_i = S_i \exp(\boldsymbol{x}_i^T\boldsymbol{\beta}), \quad \text{für } i = 1, ..., n.$$

Somit ergibt sich für den linearen Prädiktor $\boldsymbol{\eta}$, dass

$$\eta_i = \boldsymbol{x}_i^T\boldsymbol{\beta} = \log(\tilde{\mu}_i) - \log(S_i),\ i = 1, ..., n.$$

Der zweite Term $\log(S_i)$ wird auch bei der Gamma-Regression als Offset bezeichnet und als bekannt vorausgesetzt. Es wird hier nicht die kanonische Linkfunktion $\theta_i = -\frac{1}{\mu_i}$ verwendet, sondern

$$g(\tilde{\mu}_i) = \log(\tilde{\mu}_i) - \log(S_i),\ i = 1, ..., n.$$

Anstelle des Log-Likelihood betrachtet man auch hier den Gewichteten Log-Likelihood. Die Gewichtete Log-Likelihood Funktion der Gamma-Regression mit Offset und Gewicht wird dargestellt als

$$\begin{aligned} W(\boldsymbol{\beta}) &= wl(\boldsymbol{y}; \boldsymbol{\beta}, \phi) \\ &= \sum_{i=1}^{n} w_i \nu \left(-\frac{y_i}{S_i} \frac{1}{\exp(\boldsymbol{x}_i^T \boldsymbol{\beta})} - \log \left(\frac{\nu}{\exp(\boldsymbol{x}_i^T \boldsymbol{\beta})} S_i \right) \right) \\ &+ \sum_{i=1}^{n} w_i (\nu - 1) \log(y_i) - \sum_{i=1}^{n} w_i \log \Gamma(\nu). \end{aligned}$$

Für die ML-Schätzung von $\boldsymbol{\beta}$ benötigt man die Score Gleichungen. Die Score Funktion für dieses Modell ist für $i = 1, ..., n$ gegeben durch:

$$\begin{aligned} S_i(\boldsymbol{\beta}) &= \frac{\partial W(\boldsymbol{\beta})}{\partial \beta_j} \\ &= \sum_{i=1}^{n} w_i \nu \left(\frac{y_i}{S_i \exp(\boldsymbol{x}_i^T \boldsymbol{\beta})} x_{ij} - x_{ij} \right) \\ &= \sum_{i=1}^{n} w_i \nu \left(\frac{y_i}{S_i \exp(\boldsymbol{x}_i^T \boldsymbol{\beta})} - 1 \right) x_{ij}. \end{aligned}$$

Bei der Gamma-Regression mit Offset und Gewicht ist der Momentenschätzer $\hat{\phi}$ für den Dispersionsparameter $\nu = \frac{1}{\phi}$ gegeben durch:

$$\hat{\phi} = \frac{1}{n-p} \sum_{i=1}^{m} w_i \left(\frac{y_i - \hat{\mu}_i}{\hat{\mu}_i} \right)^2,$$

wobei $\hat{\mu}_i$ der Schätzer für den Erwartungswert $\tilde{\mu}_i$ von Y_i ist und p die Anzahl der geschätzten Parameter im Modell.

Für die Gamma-Regression mit Offset und Gewicht ist, wie bei allen Regressionsverfahren, die Wahl eines geeigneten Modells von zentraler Bedeutung. In Abschnitt 2.3.3 wurden

bereits zwei Modellwahlkriterien beschrieben, der AIC-Wert und die Devianz. Für dieses Modell berechnet man den AIC-Wert folgendermaßen:

$$AIC = -2\sum_{i=1}^{n} w_i \left\{ \nu \left(-\frac{y_i}{\hat{\mu}_i S_i} + \log\left(\frac{\nu}{\hat{\mu}_i S_i}\right)\right) - (\nu - 1)\log(y_i) - \log\Gamma(\nu) \right\} + 2p,$$

wobei p Anzahl der geschätzten Parameter ist. Der Wert der Devianz ist gegeben durch:

$$\begin{aligned} \frac{D(\boldsymbol{y}, \hat{\boldsymbol{\mu}})}{\phi} &= -2\left[wl(\boldsymbol{y}; \hat{\boldsymbol{\beta}}, \phi) - wl(\boldsymbol{y}; \boldsymbol{y}, \phi) \right] \\ &= -2\sum_{i=1}^{n} w_i\ \nu \left[\log\left(\frac{y_i}{\hat{\mu}_i}\right) - \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i} \right]. \end{aligned}$$

Der Residual Devianz Test und der Partial Devianz Test können ebenfalls zur Wahl eines Modells verwendet werden.

Die Hypothese des **Residual Devianz Tests**

für die Gamma-Regression mit Offset und Gewicht ist gegeben durch:

$$\boldsymbol{H}:\ \boldsymbol{\eta} = g(\boldsymbol{\mu}) = \boldsymbol{X\beta} \quad \text{versus} \quad \boldsymbol{K}:\ \text{nicht } \boldsymbol{H}$$

Die Teststatistik $\boldsymbol{L}$ berechnet man mit

$$\boldsymbol{L} = \frac{D(\boldsymbol{y}, \hat{\boldsymbol{\mu}})}{\phi} = 2\sum_{i=1}^{n} w_i\ \nu \left[\frac{y_i - \hat{\mu}_i}{\hat{\mu}_i} - \log\left(\frac{y_i}{\hat{\mu}_i}\right) \right].$$

Die Hypothese $\boldsymbol{H}$ wird zum Niveau α genau dann verworfen, wenn die Teststatistik $\boldsymbol{L}$ größer ist als $\chi^2_{n-p,1-\alpha}$.

Die Hypothese des **Partial Devianz Tests**

ist gegeben durch:

$$\boldsymbol{H}:\ \boldsymbol{\beta}_2 = 0 \quad \text{versus} \quad \boldsymbol{K}:\ \boldsymbol{\beta}_2 \neq 0$$

Die Teststatistik $\boldsymbol{P}$ berechnet man mit:

$$\boldsymbol{P} = \frac{D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}_1) - D(\boldsymbol{y}, \hat{\boldsymbol{\mu}})}{\phi_{\boldsymbol{X}}},$$

wobei $\hat{\boldsymbol{\mu}}$ der geschätzte Erwartungswertvektor im Modell $\boldsymbol{X}$ ist mit $\boldsymbol{X} = (\boldsymbol{X}_1,\ \boldsymbol{X}_2)\,,\ \boldsymbol{X} \in \mathbb{R}^p,\ \boldsymbol{X}_1 \in \mathbb{R}^{p_1},\ \boldsymbol{\mu}^T = (\boldsymbol{\mu}_1,\ \boldsymbol{\mu}_2)^T\,,\ \boldsymbol{\mu}_1 \in \mathbb{R}^{p_1},\ \boldsymbol{\mu}_2 \in \mathbb{R}^{p_2}$. Es ist $\hat{\boldsymbol{\mu}}_1$ der geschätzte Erwartungswertvektor im reduzierten Modell $\boldsymbol{X}_1$.

$\boldsymbol{H}$ wird zum Niveau α genau dann verworfen, wenn die Teststatistik $\boldsymbol{P}$ größer ist als $\chi^2_{p_1,1-\alpha}$.

Für die Gamma-Regression mit Offset und Gewicht sind verschiedene Residuen definiert. Die Pearson Residuen sind in diesem Fall wie folgt definiert:

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

Für die Devianz Residuen gilt:

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

wobei für die Devianz Komponenten d_i, $i = 1, ..., n$ gilt:

$$D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = \sum_{i=1}^{n} d_i.$$

Da die Varianz bei diesem Modell nicht einheitlich ist, betrachtet man die Standardisierten Residuen. Mit Gewichten w_i erhält man die Standardisierten Pearson Residuen:

$$\hat{r}_i^P = \sqrt{\frac{1}{\phi\ w_i}}\ r_i^P = \sqrt{\frac{1}{\phi\ w_i}}\ \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i}.$$

Für die Standardisierten Devianz Residuen gilt:

$$\hat{r}_i^D = \sqrt{\frac{1}{\phi\ w_i}}\ r_i^D = \sqrt{\frac{1}{\phi\ w_i}}\ sign(y_i - \hat{\mu}_i)\sqrt{d_i}.$$

2.4 Modellwahl bei nicht genesteten Modellen

Im folgenden Abschnitt werden zwei Tests zum Vergleich von nicht genesteten Modellen präsentiert, der Vuong Test und der Distribution-Free Test.

Beide Tests können sowohl zum Vergleich nicht genesteter Modelle als auch zum Vergleich genesteter Modelle verwendet werden. Unter nicht genesteten Modellen versteht man Folgendes:

Definition 2.17 (Nicht genestete Modelle)

Zwei Modelle heißen nicht genestet, wenn es nicht möglich ist ein Modell durch die Bedingung des Parametervektors mit linearen Restriktionen auf das andere Modell zu reduzieren, siehe Clarke (2001).

Somit heißen zwei GLMs nicht genestet, wenn sie keine gemeinsamen Variablen haben.

Für den Vergleich nicht genesteter Modelle gibt es verschiedene Methoden. Neben dem Cox Test und Bayes Faktoren gibt es den Vuong Test und den Distribution-Free Test. Sowohl der Cox Test als auch die Bayes Faktoren zum Vergleich nicht genesteter Modelle sind schwierig zu berechnen.

Der Vuong Test und der Distribution-Free Test basieren auf dem Kullback-Leibler Informationskriterium ($KLIC$). Das $KLIC$ misst den Abstand zwischen einer gegebenen und einer wahren Verteilung und wird definiert durch:

$$KLIC = E_0 \left[\log h_0 \left(Y_i | X_i\right)\right] - E_0 \left[\log f \left(Y_i | X_i, \hat{\boldsymbol{\beta}}\right)\right],$$

siehe Vuong (1989), wobei h_0 die Dichte des wahren, aber unbekannten Modells ist und $\hat{\boldsymbol{\beta}}$ der Schätzer für $\boldsymbol{\beta}$.

Das beste Modell ist somit das Modell, das die obige Gleichung minimiert. Daher sollte ein Modell dem Anderen genau dann vorgezogen werden, wenn der Log-Likelihood des einen Modells signifikant kleiner ist als der des anderen Modells.

2.4.1 Vuong Test zur Wahl nicht genesteter Modelle

Es seien nun zwei nicht genestete Modelle gegeben mit

$$F_{\theta_1} = \{f(\boldsymbol{y}|\boldsymbol{x}, \theta_1), \theta_1 \in \Theta)\} \text{ und } G_{\theta_2} = \{g(\boldsymbol{y}|\boldsymbol{z}, \theta_2), \theta_2 \in \Theta)\}\,.$$

Man will nun entscheiden, welches der beiden Modelle das bessere ist. Dazu werden nun zuerst drei Hypothesen definiert, siehe Vuong (1989):

$$\begin{aligned}
H \quad &: E\left(\log \frac{f(Y_i|X_i;\hat{\boldsymbol{\beta}})}{g(Y_i|Z_i;\hat{\boldsymbol{\gamma}})}\right) = 0 \quad \text{(Die Modelle } F_{\theta_1} \text{ und } G_{\theta_2} \text{ sind gleich gut.)} \\
H_f \quad &: E\left(\log \frac{f(Y_i|X_i;\hat{\boldsymbol{\beta}})}{g(Y_i|Z_i;\hat{\boldsymbol{\gamma}})}\right) > 0 \quad (F_{\theta_1} \text{ ist das bessere Modell.}) \\
H_g \quad &: E\left(\log \frac{f(Y_i|X_i;\hat{\boldsymbol{\beta}})}{g(Y_i|Z_i;\hat{\boldsymbol{\gamma}})}\right) < 0 \quad (G_{\theta_2} \text{ ist das bessere Modell.}).
\end{aligned}$$

Der Erwartungswert der Hypothese ist unbekannt, aber Vuong zeigt, siehe Vuong (1989), dass unter generellen Bedingungen

$$\frac{1}{n}\boldsymbol{LR}_n\left(\hat{\boldsymbol{\beta}}_n, \hat{\boldsymbol{\gamma}}_n\right) \to E\left(\log \frac{f(Y_i|X_i;\hat{\boldsymbol{\beta}})}{g(Y_i|Z_i;\hat{\boldsymbol{\gamma}})}\right)$$

gilt. Folglich ist die Log-Likelihood Ratio Statistik ein konsistenter Schätzer für den Erwartungswert.

Aus dem folgenden Theorem kann man zusammen mit der obigen Aussage einen Hypothesentest konstruieren.

Satz 2.18 (Modellwahl Tests für nicht genestete Modelle)

Seien F_{θ_1} und G_{θ_2} nicht genestete Modelle, dann gilt

$$\textit{unter } \boldsymbol{H}: \quad \frac{1}{\sqrt{n}w_n}\boldsymbol{LR}_n\left(\hat{\boldsymbol{\beta}}_n, \hat{\boldsymbol{\gamma}}_n\right) \xrightarrow{D} \mathcal{N}(0,1) \tag{2.3}$$

$$\textit{unter } \boldsymbol{H}_f: \quad \frac{1}{\sqrt{n}w_n}\boldsymbol{LR}_n\left(\hat{\boldsymbol{\beta}}_n, \hat{\boldsymbol{\gamma}}_n\right) \xrightarrow{a.s.} \infty \tag{2.4}$$

$$\textit{unter } \boldsymbol{H}_g: \quad \frac{1}{\sqrt{n}w_n}\boldsymbol{LR}_n\left(\hat{\boldsymbol{\beta}}_n, \hat{\boldsymbol{\gamma}}_n\right) \xrightarrow{a.s.} -\infty \tag{2.5}$$

siehe Vuong (1989).

Die Hypothese des **Vuong Tests**

ist somit gegeben durch:

$$\boldsymbol{H}: E\left(\log\frac{f(Y_i|X_i;\hat{\boldsymbol{\beta}})}{g(Y_i|Z_i;\hat{\boldsymbol{\gamma}})}\right) = 0 \quad \text{versus} \quad \boldsymbol{K}: E\left(\log\frac{f(Y_i|X_i;\hat{\boldsymbol{\beta}})}{g(Y_i|Z_i;\hat{\boldsymbol{\gamma}})}\right) \neq 0 \quad i = 1, ..., n.$$

Diese Art von Hypothesentests werden als Modellwahl Tests (Model Selection Tests) bezeichnet.

Die Teststatistik $\boldsymbol{V}$ berechnet man als

$$\boldsymbol{V} = \frac{\boldsymbol{LR}_n\left(\hat{\boldsymbol{\beta}}_n, \hat{\boldsymbol{\gamma}}_n\right)}{\sqrt{n}\hat{w}_n}$$

mit

$$\begin{aligned}\boldsymbol{LR}_n\left(\hat{\boldsymbol{\beta}}_n, \hat{\boldsymbol{\gamma}}_n\right) &= L_n^f(\hat{\boldsymbol{\beta}}_n) - L_n^g(\hat{\boldsymbol{\gamma}}_n) \\ &= \log\sum_{i=1}^n f(Y_i|X_i;\hat{\boldsymbol{\beta}}) - \log\sum_{i=1}^n g(Y_i|Z_i;\hat{\boldsymbol{\gamma}})\end{aligned}$$

und

$$\hat{w}_n^2 = \frac{1}{n}\sum_{i=1}^n\left(\log\frac{f(Y_i|X_i;\hat{\boldsymbol{\beta}})}{g(Y_i|Z_i;\hat{\boldsymbol{\gamma}})}\right)^2 - \left(\frac{1}{n}\sum_{i=1}^n\log\frac{f(Y_i|X_i;\hat{\boldsymbol{\beta}})}{g(Y_i|Z_i;\hat{\boldsymbol{\gamma}}}\right)^2.$$

Die Nullhypothese $\boldsymbol{H}$ wird zum Niveau α genau dann verworfen, wenn die Teststatistik $\boldsymbol{V}$ größer ist als $z_{1-\alpha}$.

Der Vuong Test kann also wie folgt beschrieben werden: wenn die Nullhypothese wahr ist, dann ist der durchschnittliche Wert der Log-Likelihood Ratio Statistik gleich Null. Ist $\boldsymbol{H}_f$

wahr, dann sollte der durchschnittliche Wert der Log-Likelihood Ratio Statistik deutlich größer als Null sein. Ist das Gegenteil $\boldsymbol{H}_g$ wahr, dann sollte dieser Wert signifikant kleiner als Null sein.

Wie auch für viele andere Tests gilt für den Vuong Test, dass die Log-Likelihoods durch die Anzahl der Parameter der beiden Modelle beeinflusst sind, und somit eine Korrektur notwendig ist. Die Statistik $\boldsymbol{LR}_n\left(\hat{\boldsymbol{\beta}}_n, \hat{\boldsymbol{\gamma}}_n\right)$ wird verändert in

$$\widehat{\boldsymbol{LR}}_n\left(\hat{\boldsymbol{\beta}}_n, \hat{\boldsymbol{\gamma}}_n\right) = \boldsymbol{LR}_n\left(\hat{\boldsymbol{\beta}}_n, \hat{\boldsymbol{\gamma}}_n\right) - \left(\frac{p-q}{2}\log(n)\right),$$

wobei p und q die Anzahl der im Modell f und g geschätzten Parameter sind.

2.4.2 Distribution-Free Test zur Wahl nicht genesteter Modelle

Im Gegensatz zum Vuong Test betrachtet der Distribution-Free Test nicht, siehe Clarke (2007), ob die Log-Likelihood Statistik signifikant verschieden von Null ist, sondern er betrachtet die Vorzeichen der Differenzen der individuellen Log-Likelihoods.

Der Distribution-Free Test, siehe Clarke (2003), ist dem Wilcoxon Rang-Vorzeichentest ähnlich und gehört zu der Klasse der Paired Sign Tests.

Die Hypothese des **Distribution-Free Tests**

ist gegeben durch:

$$\boldsymbol{H} : P\left(\log\frac{f(Y_i|X_i;\hat{\boldsymbol{\beta}})}{g(Y_i|Z_i;\hat{\boldsymbol{\gamma}})} > 0\right) = 0.5 \quad \text{vs} \quad \boldsymbol{K} : P\left(\log\frac{f(Y_i|X_i;\hat{\boldsymbol{\beta}})}{g(Y_i|Z_i;\hat{\boldsymbol{\gamma}})} > 0\right) \neq 0.5$$

für $i = 1, .., n$.

Der Distribution-Free Test untersucht folglich, ob die mittleren Log-Likelihood Ratios signifikant verschieden von Null sind. Sind beide Modelle dem wahren Modell gleich nahe, dann sollte die Hälfe der Log-Likelihood Ratios kleiner, die andere größer als Null sein.

Ist f im Vergleich zu g das „bessere“ Modell, dann sollte mehr als die Hälfte der Log-Likelihoods kleiner Null sein. Die Umkehrung gilt, falls g „besser“ als f ist.

Man erhält folgende Teststatistik $\boldsymbol{B}$:

$$\boldsymbol{B} = \sum_{i=1}^{n} \boldsymbol{I}_{(0,\infty)}(d_i)$$

mit

$$d_i = \log \frac{f(Y_i|X_i;\hat{\boldsymbol{\beta}})}{g(Y_i|Z_i;\hat{\boldsymbol{\gamma}})}.$$

Somit ist $\boldsymbol{B}$ die Anzahl der positiven Differenzen der individuellen Log-Likelihoods. $\boldsymbol{B}$ ist unter $\boldsymbol{H}$ Binomialverteilt mit den Parametern k und $p = 0.5$. Somit wird $\boldsymbol{H}$ genau dann verworfen, wenn $\boldsymbol{H} < \frac{k}{2}$

Für diesen Test gilt ebenfalls, dass er durch die Anzahl der geschätzten Koeffizienten beeinflusst wird. Der Korrekturfaktor der verwendet wird, ist an den Korrekturfaktor des Akaike Informationskriteriums und des Schwarz Informationskriteriums angelehnt.

Es wird ein Korrekturfaktor für jede Differenz der individuellen Log-Likelihoods verwendet, siehe Clarke (2003) und zwar

$$\frac{p-q}{2n} \log(n),$$

wobei p und q die Anzahl der geschätzten Koeffizienten in den Modellen f und g sind.

Schließlich erhält man die modifizierte Teststatistik $\hat{\boldsymbol{B}}$:

$$\hat{\boldsymbol{B}} = \sum_{i=1}^{n} \boldsymbol{I}_{(0,\infty)}(d_i) - \frac{p-q}{2n} \log(n).$$

Clarke zeigt (siehe Clarke (2007)), dass der Distribution-Free Test besonders gut für spitze Verteilungen arbeitet.

3 Kraftfahrt Haftpflicht Datensatz

In diesem Kapitel wird ein realer Datensatz untersucht. Bei den betrachteten Daten handelt es sich um Kraftfahrt Haftpflicht Datensätze für private Personenkraftwagen (PKW).

In den folgenden Abschnitten werden die Daten vorgestellt. Es wird auf die Herkunft der Daten eingegangen und die im Datensatz enthaltenen Merkmale beschrieben. Des Weiteren wird das Vorgehen bei der Verdichtung und der Selektion der Daten erläutert.

Im nächsten Kapitel werden einige Modelle an die Daten angepasst. Es werden Modellierungsvarianten mit verschiedenen Variablen und unterschiedlichen Verteilungsannahmen betrachtet.

In den nachfolgenden Kapiteln wird die Anpassungsgüte untersucht. Der Schwerpunkt liegt hierbei auf der Modellwahl mit Hilfe des Vuong und des Distribution-Free Tests. Den Abschluss dieser Arbeit bildet die Untersuchung der Power dieser beiden Tests durch Simulationen.

3.1 Herkunft der Daten

Der Datensatz, der in dieser Arbeit untersucht wird, wurde von der Versicherungskammer Bayern (VKB) zur Verfügung gestellt. Er besteht aus Kraftfahrt Haftpflicht Datensätzen für private PKW. Es wird der Zeitraum Januar 2006 bis Oktober 2007 betrachtet.

Insgesamt liegen 3,1 Millionen Beobachtungen (Zeilen) vor, die durch 27 Variablen (Spalten) beschrieben werden.

3.2 Daten Selektion

Die Daten werden nun selektiert, damit sie für die Modellierung verwendet werden können. Die Selektion erfolgt nach bestimmten Kriterien.

Zum einen werden nur Datensätze ausgewählt, bei welchen der Versicherungsnehmer ein Fahrzeug mindestens drei Jahre bei der VKB versichert hatte.

Zum anderen müssen die Datensätze aus den aktuellen Tarifgenerationen stammen. Diese Einschränkung wird notwendig, da sich die Ausprägungen der Merkmale zwischen den vorigen und den aktuellen Tarifgenerationen unterscheiden.

Tabelle 3.1: Anzahl der Beobachtungen vor und nach der Datenselektion insgesamt und nach Entstehungsjahren

	KH Rohdaten	**KH Gesamt**	**KH 2006**	**KH 2007**
Beobachtungen	3.139.149	517.560	234.906	282.654

Tabelle 3.1 enthält die Anzahl der Beobachtungen vor und nach der Datenselektion. Die Anzahl der Beobachtungen nach der Datenselektion ist sowohl für den gesamten Beobachtungszeitraum angegeben als auch aufgeteilt nach Entstehungsjahren.

Wie aus der Tabelle ersichtlich bleiben nach der Selektion von den ursprünglichen 3,1 Millionen Beobachtung (KH Rohdaten) nur 517.560 (KH Gesamt) übrig. Gruppiert man die Beobachtungen nach ihrem Entstehungsjahr, so liegen für das Entstehungsjahr 2006 (KH 2006) 234.906 Beobachtungen und für 2007 (KH 2007) 282.654 Beobachtungen vor.

3.3 Beschreibung der Daten

Es werden nun in Tabelle 3.2 die für diese Arbeit relevanten Merkmale vorgestellt. Auf eine ausführliche Erläuterung der weiteren im Datensatz enthaltenen Merkmale, die in den aktuellen VKB Tarif eingehen, wird verzichtet, da sie für diese Arbeit nicht verwendet werden.

Tabelle 3.2: Beschreibung ausgewählter Merkmale des KH Datensatzes

Name	**Beschreibung**	**Ausprägungen**
A	13 Klassen	A1, A2,..., A13
B	16 Klassen	B10, B11,..., B25
a	4 Klassen	a0,a1,a2,a3
b	5 Klassen	b0, b1, b2, b3, b4
c	19 Klassen	c0, c1,..., c18
SB	Der Schadenbedarf (SB) ist das Verhältnis von der gesamten Schadensumme zur Summe der Beiträge.	$SB \geq 0$
n	Die Anzahl der Schäden n gibt an, wie viele Schäden ein Versicherungsnehmer in einem Jahr verursacht hat.	$n \in \mathbb{N}$
D	Der durchschnittliche Schaden D ist das Verhältnis von der Summe der Schäden zur Anzahl der Schäden. D wird auf einen Cent gesetzt, wenn kein Schaden vorhanden ist.	in Euro
u	Eine Jahreseinheit u ist ein Vertrag eines Versicherungsnehmers, der ein vollständiges Kalenderjahr läuft.	$0 < u \leq 1$
Bt	Der Beitrag Bt ist die vom Versicherungsnehmer entrichtete Prämie. Diese Prämie ist auf die Jahreseinheit u bezogen.	in Euro
$\hat{R}$	Der Schaden $\hat{R}$ ist der vom Versicherungsnehmer verursachte Schaden. Er wird bei 20 Tausend Euro kupiert.	in Euro

Die Merkmale A, B, a, b und c aus der obigen Tabelle werden in den Modellen als Variablen verwendet.

Der Schadenbedarf SB soll mit Hilfe dieser Variablen erklärt werden. Somit stellt das Merkmal Schadenbedarf die abhängige Variable - den Response - dar.

Bei dem in Tabelle 3.2 beschriebenen Merkmal Beitrag Bt ist zu beachten, dass das Merkmal um Merkmal A und B bereinigt ist. Somit enthält das Merkmal Beitrag Bt alle

Informationen der restlichen Merkmale, die in den aktuellen VKB Tarif eingehen.

Der vom Versicherungsnehmer jährlich verursachte Schaden $\hat{R}$ wird auf 20 Tausend Euro kupiert. Die Kupierung der Schäden soll eine Verzerrung durch Großschäden vermeiden.

3.4 Datenmanipulation

Betrachtet man die pro Datensatz entstandene Schadensumme, dann stellt man fest, dass 95 Prozent der Beobachtungen im Jahr 2006 Null Euro Schaden haben. Will man die Daten mit der Gamma-Verteilung modellieren, ist dies nicht möglich, da die Gamma-Verteilung nur für positive Werte definiert ist.

Der Schaden eines Versicherungsnehmers $\hat{R}$ kann aber, wie bereits erwähnt, den Wert Null annehmen und in diesen Fällen ist der Response - der Schadenbedarf - ebenfalls Null. Somit ist es notwendig den Response zu manipulieren.

Die Manipulation erfolgt durch Addition einer Variable x mit $x \in \mathbb{R}_+$ auf den vom Versicherungsnehmer verursachten Schaden $\hat{R}$. Für die Variable x werden die Werte 1, 10 und 100 Cent gewählt. Dabei erfolgt eine Manipulation von $\hat{R}$ nur dann, wenn $\hat{R}$ den Wert Null hat. Man erhält nun $R(x)$ mit $R(x) = \hat{R} + x$.

Tabelle 3.3: Bezeichnung des manipulierten Schadenbedarfs und der manipulierten Schäden

Manipulation	**Bezeichnung**	
0 Cent	SB	R
1 Cent	SB1	R1
10 Cent	SB10	R10
100 Cent	SB100	R100

Da der vom Versicherungsnehmer verursachte Schaden $\hat{R}$ manipuliert wird, muss folglich auch der Schadenbedarf SB angepasst werden. Durch die Manipulation mit unterschiedlichen Werten für x werden neue Bezeichnungen für den manipulierten Schadenbedarf eingeführt. Die Bezeichnung des manipulierten Schadens erfolgt analog, siehe Tabelle 3.3.

3.5 Verdichtung der Daten

Der ursprüngliche KH Datensatz wird nun auf zwei Weisen verdichtet. In den beiden folgenden Abschnitten werden die beiden verdichteten Datensätze vorgestellt.

In dieser Arbeit wird der Schadenbedarf SB auf zwei Arten modelliert. Zum einen wird der aktuelle VKB Tarif nachgebildet, indem der Schadenbedarf durch die Merkmale B und A modelliert wird.

In der alternativen Modellierungsvariante wird der Schadenbedarf durch die Merkmale a, b und c modelliert.

Die Daten des KH Datensatzes sind die Daten auf Einzelebene, pro Zeile wird ein Versicherungsnehmer betrachtet. Für die Berechnung der Modelle ist es jedoch notwendig die Daten zuerst zu verdichten. Verdichtung bedeutet, dass die Einzeldatensätze zusammengefasst werden. Die Zusammenfassung erfolgt entsprechend der im Modell enthaltenen Variablen.

Jede Merkmalskombination ist nach der Verdichtung höchstens einmal vertreten. Es wird für jede Merkmalskombination die Anzahl der Schäden n, der vom Versicherungsnehmer verursachte Schaden $\hat{R}$ und die Jahreseinheiten u aufsummiert. Um den durchschnittlichen Beitrag Bt pro Zelle zu erhalten, wird der Beitrag aufsummiert und durch die entsprechenden Jahreseinheiten dividiert.

Da zwei alternative Modellierungen des Schadenbedarfs betrachtet werden, entstehen nun auch zwei aggregierte Datensätze, die im Folgenden als S Datensatz und V Datensatz bezeichnet werden. Man erhält für die aggregierten Datensätze folgende Anzahl an Zellen:

Tabelle 3.4: Anzahl der Zellen in den verdichteten Datensätzen

	S Datensatz	**V Datensatz**
KH 2006	320	184
KH 2007	322	184

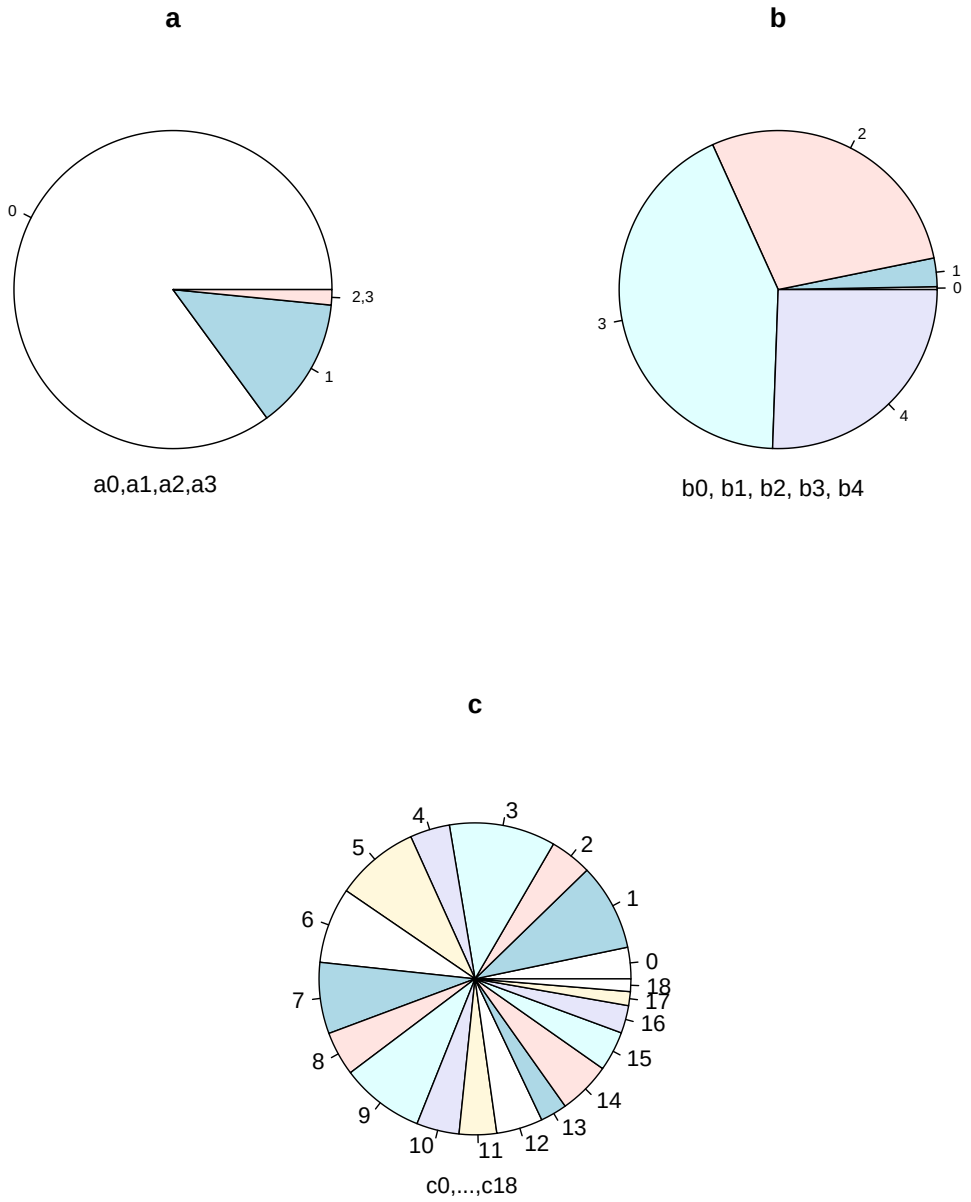
Im S Datensatz bleiben für das Entstehungsjahr 2006 noch 320 Zellen übrig und es bleiben für das Entstehungsjahr 2007 noch 322 Zellen. Der V Datensatz enthält 184 Zellen sowohl für das Entstehungsjahr 2006 als auch für 2007.

Es werden nun die beiden aggregierten Datensätze vorgestellt.

3.5.1 Explorative Analyse des S Datensatzes

Dieser Abschnitt untersucht den S Datensatz (KH 2006). Zu Beginn werden anhand einer Grafik die Anteile der Merkmale b, a und c am ursprünglichen KH Datensatz dargestellt.

Abbildung 3.1: Anteile der Merkmale a, b und c am unverdichteten KH Datensatz



Es folgt ein Ausschnitt des neu entstandenen S Datensatzes. Abschließend werden die

Lagemaße des S Datensatzes beschrieben.

Der S Datensatz ist durch die Verdichtung des KH Datensatzes (KH 2006) entstanden. Dieser wurde nach den Merkmalen b, a und c aggregiert. Abbildung 3.1 betrachtet nun die ursprünglichen Anteile dieser Merkmale am KH Datensatz.

Aus der obigen Darstellung wird deutlich, dass die Ausprägungen b0 und b1 des Merkmals b, einen sehr geringen Anteil am Gesamtdatensatz haben. Den größten Anteil hat die Klasse b3. Mehr als 80 Prozent aller Beobachtungen des Merkmals a haben die Ausprägung a0. Für das Merkmal c gilt, dass alle Klassen etwa gleich viele Beobachtungen enthalten.

Wie in Tabelle 3.4 zu erkennen ist, enthält der nach der Verdichtung entstandene S Datensatz 320, bzw. 322 Zellen. Somit ist nicht jede Merkmalskombination vertreten. Wäre jede Merkmalskombination vertreten, so würde der S Datensatz aus 380 Zeilen bestehen. Die fehlenden Zeilen sind auf die dünn besetzten Klassen a2 und a3 des Merkmals a und auf die dünn besetzten Klassen b0 und b1 des Merkmals b zurückzuführen.

Im Folgenden Ausschnitt des S Datensatzes (KH 2006) finden sich Merkmalskombinationen, die bei fast keinem Vertrag auftreten. Zum Beispiel hat Zeile Nummer 6 nur 0.8 Jahreseinheiten.

Ausschnitt des S Datensatzes

Nr	c	a	b	S_n	S_B	S_J	S_D	S_SB	S_SB1	S_SB10	S_SB100
1	0	0	0	0	782	7.2	0.01	0.00	0.00	0.0004	0.0043
2	0	0	1	7	582	117.3	1268.71	0.13	0.13	0.1304	0.1341
3	0	0	2	36	552	875.5	2163.26	0.16	0.16	0.1615	0.1651
4	0	0	3	61	568	1371.2	2791.07	0.22	0.22	0.2191	0.2227
5	0	0	4	30	628	517.1	1817.89	0.17	0.17	0.1682	0.1712
6	0	1	0	0	537	0.8	0.01	0.00	0.00	0.0005	0.0046
7	0	1	1	1	624	20.4	2785.40	0.22	0.22	0.2196	0.2230
8	0	1	2	8	614	121.5	1854.06	0.20	0.20	0.1991	0.2025
9	0	1	3	19	667	207.5	3008.08	0.41	0.41	0.4131	0.4161
10	0	1	4	9	696	92.6	2204.73	0.31	0.31	0.3080	0.3109

Um die aggregierten Merkmale des S und des V Datensatzes unterscheiden zu können, wird der Buchstabe S bzw. V den Merkmalen vorangestellt.

Tabelle 3.5 fasst die Lagemaße des S Datensatzes (KH 2006) zusammen. Wie man erkennen kann enthalten mehr als 25 Prozent der Zellen weniger als fünf Jahreseinheiten.

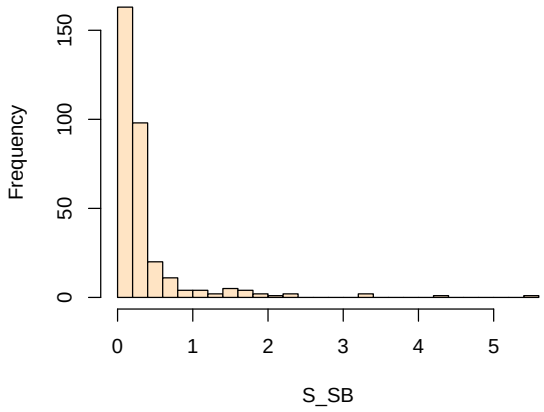
Diese Zellen sind somit nicht aussagekräftig im Bezug auf den Schadenbedarf. Dies muss im Folgenden bei der Modellierung berücksichtigt werden. Betrachtet man die restlichen Lagemaße des Merkmals Jahreseinheiten S_J, dann enthält eine Zelle maximal 4.420, aber mindestens 0,1 Jahreseinheiten. Durchschnittlich sind 348 Jahreseinheiten enthalten.

Tabelle 3.5: Lagemaße des S Datensatzes (KH 2006)

	c	**a**	**b**	**S_n**	**S_B**	**S_J**	**S_D**	**S_SB**	**S_SB1**	**S_SB10**	**S_SB100**
Minimum	0.00	0.00	0.00	0.0	513	0.01	0.01	0.000	0.000	0.000	0.001
1st Qua.	4.00	0.00	1.00	0.0	588	4.6	0.01	0.000	0.000	0.000	0.005
Median	8.00	1.00	2.00	2.5	635	26.1	1784	0.194	0.194	0.194	0.197
Mean	8.70	1.30	2.28	18.8	664	348	1830	0.321	0.321	0.322	0.325
3rd Qua.	13.00	2.00	3.00	19.0	700	268	2720	0.304	0.304	0.304	0.307
Maximum	18.00	3.00	4.00	240.0	1183	4420	20000	5.568	5.572	5.568	5.569

Die Anzahl der Schäden S_n pro Zelle variiert von Null bis 240. Durchschnittlich gab es 18,8 Schäden pro Zeile. Der maximale durchschnittliche Schaden S_D liegt bei 20.000 Euro und der minimale bei einem Cent. Ein Schaden verursacht im Durchschnitt Kosten in Höhe von 1.830 Euro. Für all diese Merkmale gilt, dass die Abweichung zwischen dem dritten Quantil und dem Maximum, im Vergleich zur Abweichung zwischen dem Mittelwert und dem dritten Quantil, sehr groß ist.

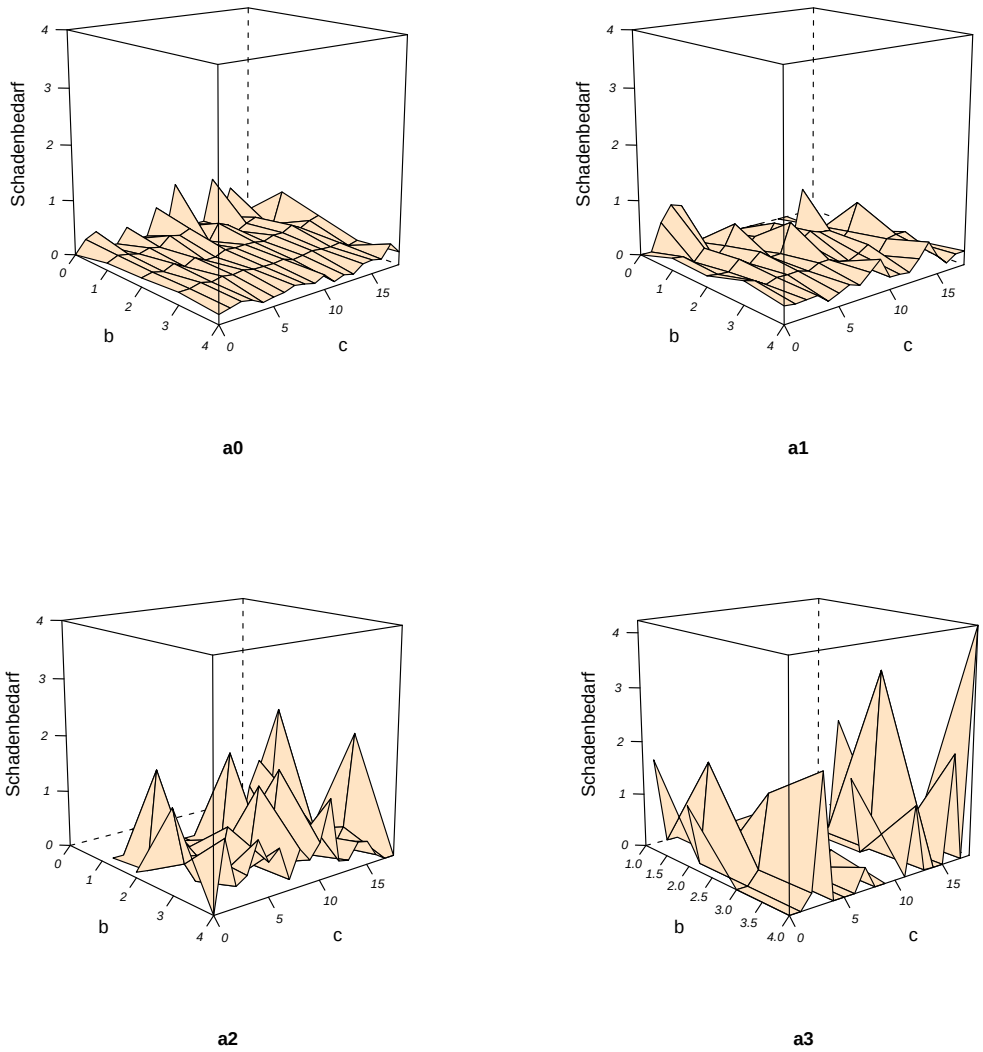
Abbildung 3.2: Histogramm von S_SB



Betrachtet man die Lagemaße des Schadenbedarfs, dann verändern sich diese nur geringfügig durch die Manipulation mit einer Konstanten. Für den Schadenbedarf S_SB gilt, dass 75 Prozent der Zellen einen Schadenbedarf kleiner als 0,304 haben.

Abbildung 3.2 zeigt das Histogramm von S_SB. Man kann sehen, dass fast alle Beobachtungen einen Schadenbedarf unter 0,5 haben. Die Histogramme von S_SB1, S_SB10 und S_SB100 bieten dasselbe Bild, siehe Anhang. Auch hier liegt der Schadenbedarf für mehr als 250 der 320 Zellen unter 0,5.

Abbildung 3.3: Merkmal b und c gegen den Schadenbedarf aufgeteilt nach Merkmal a



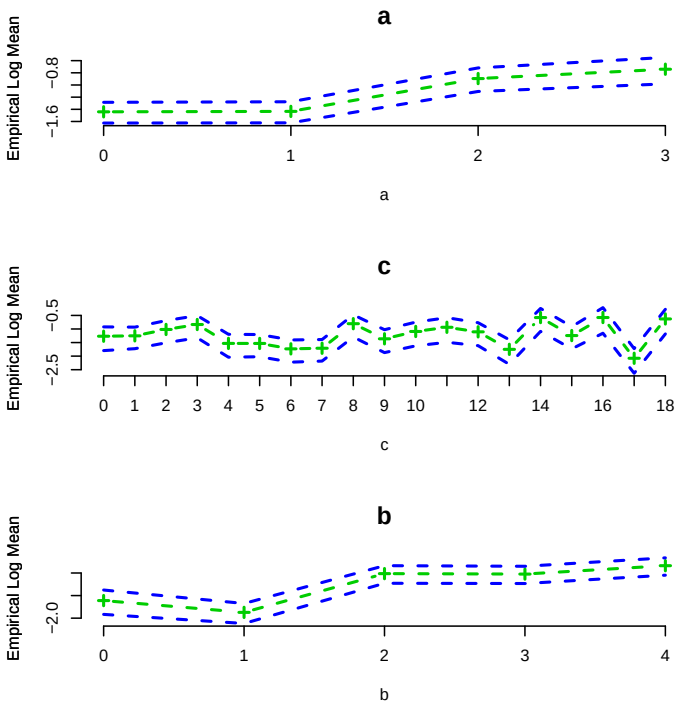
In Abbildung 3.3 ist der Schadenbedarf gegen das Merkmal b und c angetragen, wobei nach dem Merkmal a gruppiert wurde.

In dem Fall a0 schwankt der Schadenbedarf kaum. Nur für die Klassen b0 und b1 sind Schwankungen zu erkennen.

Für a1 schwankt der Schadenbedarf der Merkmale b und c stärker. Insgesamt liegt die angepasste Ebene bei a1 auf einem höheren Niveau als bei a0.

Die Schwankung ist für a2 und a3 umso stärker. Dies wird besonders bei den letzten beiden Abbildungen deutlich. Für die Ausprägungen a2 und a3 gibt es nicht für alle Kombinationen der Merkmale b und c eine Beobachtung des Schadenbedarfs. Aus diesem Grund sind die eingebetteten Ebenen nicht vollständig.

Abbildung 3.4: Empirical Log Mean der Merkmale a, b und c



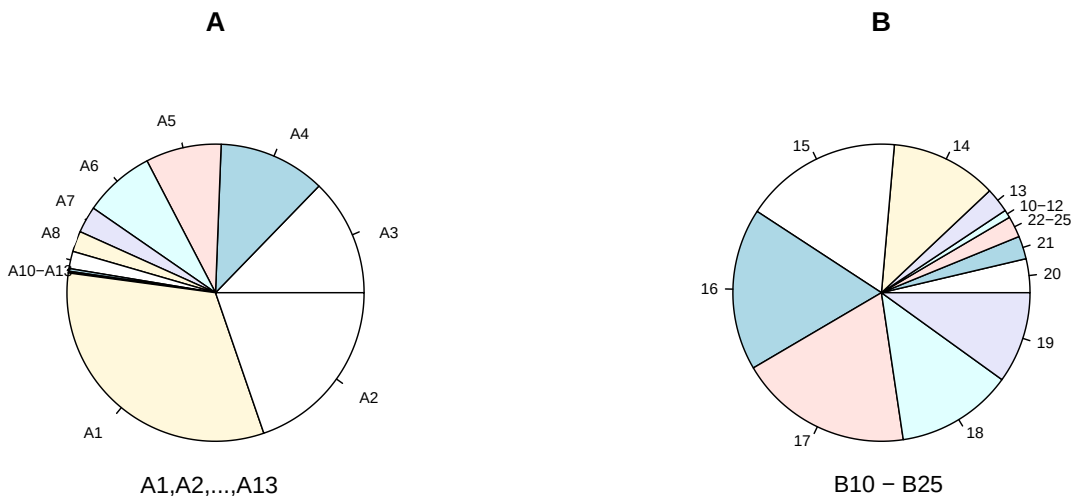
Einen Überblick über die Abhängigkeit der einzelnen Merkmale vom Schadenbedarf gibt Abbildung 3.4. Dort ist der Empirical Log Mean der Merkmale a, b und c angegeben.

3.5.2 Explorative Analyse des V Datensatzes

In diesem Abschnitt wird der V Datensatz (KH 2006) betrachtet. Es werden zuerst die Anteile der Merkmale B und A am ursprünglichen KH Datensatz graphisch dargestellt. Anschließend wird ein Ausschnitt des V Datensatzes gezeigt. Danach werden die Lagemaße des V Datensatzes beschrieben.

Auch der V Datensatz ist durch die Verdichtung des KH Datensatzes (KH 2006) entstanden. Es wurden jedoch andere Merkmale als beim S Datensatz aggregiert. Um den V

Abbildung 3.5: Anteile der Merkmale B und A am unverdichteten KH Datensatz



Datensatz zu erhalten wurden die Merkmale A und B gruppiert. In Abbildung 3.5 werden die Anteile dieser beiden Merkmale am KH Datensatz dargestellt.

Betrachtet man das Merkmal B, dann erkennt man, dass mehr als 50 Prozent der Beobachtungen die Ausprägungen B15, B16 und B17 haben. Die Ausprägungen B10 bis B12 und B22 bis B25 haben hingegen einen sehr geringen Anteil am Gesamtdatensatz. Für das Merkmal A gilt: Mehr als die Hälfte aller Datensätze haben Ausprägungen A1 oder A2. Die Ausprägung A10 und die darüber liegenden sind fast nicht vertreten.

Der V Datensatz besteht aus 184 Zeilen, siehe Tabelle 3.4, somit sind auch bei diesem Datensatz nicht alle Merkmalskombinationen vorhanden. Ein Datensatz mit allen Merkmalskombinationen würde aus 208 Zeilen bestehen. Wie bereits erwähnt sind die Ausprägung A10 und die darüber liegenden Ausprägungen des Merkmals A kaum besetzt. Dies führt zum Verlust an Zeilen im V Datensatz.

Es folgt der Ausschnitt des V Datensatzes (KH 2006). Wie beim S Datensatz vorher, finden sich auch hier Merkmalskombinationen, die bei fast keinem Vertrag auftreten. Beobachtung Nummer 10 hat beispielsweise nur 0,2 Jahreseinheiten.

Erneut wird zur Unterscheidung der Merkmale der Buchstabe des Datensatzes vorangestellt. Für den Schadenbedarf SB schreibt man V_SB.

Ausschnitt des V Datensatzes

Nr	B	A	V_n	V_B	V_J	V_D	V_SB	V_SB1	V_SB10	V_SB100
1	10	1	0	590	22.28	0.01	0.000	0.000	0.0004	0.0037
2	10	2	0	605	13.31	0.01	0.000	0.000	0.0004	0.0035
3	10	3	0	578	10.16	0.01	0.000	0.000	0.0005	0.0049
4	10	4	0	595	11.83	0.01	0.000	0.000	0.0005	0.0045
5	10	5	1	578	13.56	2874.10	0.371	0.370	0.3672	0.3706
6	10	6	1	641	9.61	0.01	0.000	0.000	0.0003	0.0034
7	10	7	0	531	1.11	0.01	0.000	0.000	0.0010	0.0102
8	10	8	0	532	5.63	0.01	0.000	0.000	0.0003	0.0033
9	10	9	0	652	2.97	0.01	0.000	0.000	0.0005	0.0052
10	10	10	0	393	0.20	0.01	0.000	0.000	0.0025	0.0251

Tabelle 3.6 fasst die Lagemaße des V Datensatzes (KH 2006) zusammen. Man kann beobachten, dass 25 Prozent der Zeilen weniger als 7,33 Jahreseinheiten enthalten. Dies muss im nachfolgenden Kapitel insoweit Berücksichtigung bei der Modellierung finden, dass diese Zeilen weniger Einfluss auf das Modell haben. Betrachtet man die übrigen Lagemaße der Jahreseinheiten V_J, dann enthält eine Zeile maximal 7.100, aber mindestens 0,03 Jahreseinheiten. Durchschnittlich sind 605 Jahreseinheiten pro Zeile enthalten.

Tabelle 3.6: Lagemaße des V Datensatzes (KH 2006)

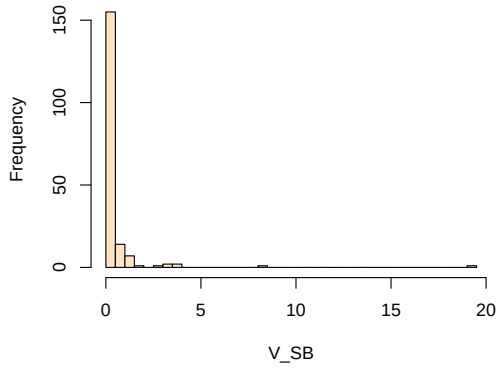
	B	**A**	**V_n**	**V_B**	**V_J**	**V_D**	**V_SB**	**V_SB1**	**V_SB10**	**V_SB100**
Minimum	10	1	0.0	393	0.03	0.01	0.000	0.000	0.000	0.001
1st Quantil	14	3	1.0	584	7.33	497.31	0.078	0.078	0.078	0.082
Median	17	5	5.5	607	62.9	2205.95	0.221	0.221	0.221	0.222
Mean	17	8	32.8	632	605	2105.66	0.501	0.501	0.501	0.501
3rd Quantil	21	9	30.5	654	486	2828.19	0.384	0.384	0.384	0.388
Maximum	25	13	324.0	1351	7100	20000.00	19.26	19.26	19.27	19.27

Im V Datensatz variiert die Anzahl der Schäden S_n von Null bis 324. Die durchschnittliche Schadenanzahl liegt bei 32,8. Der durchschnittliche Schaden V_D schwankt von einem Cent bis 20.000 Euro und liegt im Mittel bei 2.105,66 Euro. Die Abweichung zwischen dem dritten Quantil und dem Maximum ist für all diese Merkmale sehr groß im Vergleich zur Abweichung zwischen dem Mittelwert und dem dritten Quantil.

Wie beim vorhergehenden Datensatz verändern sich die Lagemaße des Schadenbedarfs nur geringfügig durch die Manipulation mit einer Konstanten. Der durchschnittliche Schaden-

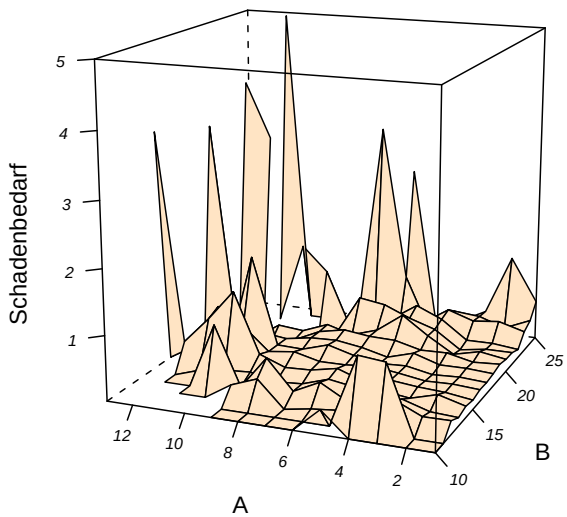
bedarf S_SB liegt bei 0,501. Maximal wird ein Schadenbedarf von 19,3 und minimal Null erreicht.

Abbildung 3.6: Histogramm von V_SB



Das Histogramm von V_SB, siehe Abbildung 3.6, zeigt, dass fast alle Zeilen einen Schadenbedarf kleiner als eins haben. Dasselbe Bild ergibt sich aus den Histogrammen für die Manipulation des Schadenbedarfs V_SB1, V_SB10 und V_SB100. Auch hier ist der Schadenbedarf von mehr als 160 Zellen kleiner als eins.

Abbildung 3.7: Merkmale A und B gegen den Schadenbedarf

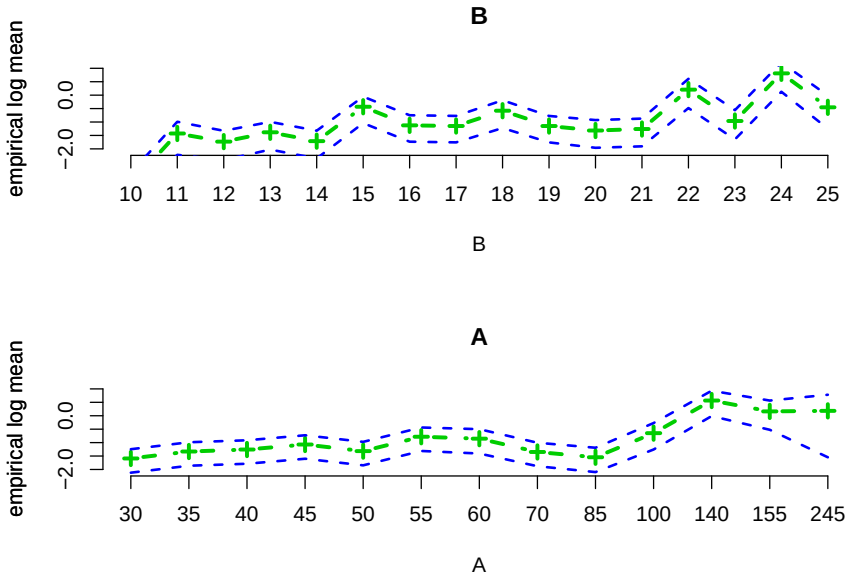


In Abbildung 3.7 wird der Schadenbedarf gegen die Merkmale A und B angetragen.

Da auch hier nicht für alle Kombinationen der Merkmale B und A Beobachtungen des Schadenbedarfs vorliegen, ist diese Ebene ebenfalls nicht vollständig. Speziell für A10 und darüber fehlen Beobachtungen.

In der obigen Abbildung ist zu erkennen, dass für B11 und B12 und für B22 bis B25, ein erhöhter Schadenbedarf vorliegt.

Abbildung 3.8: Empirical Log Mean der Merkmale A und B



Betrachtet man die Empirical Log Means in Abbildung 3.8, dann bietet sich ein ähnliches Bild. Auch hier ist erst ab B21 ein steigender Schadenbedarf zu erkennen.

Für das Merkmale A ist zwischen den Ausprägungen A1 und A9 ebenfalls keine Steigung des Schadenbedarfs zu erkennen. Erst ab A10 steigt der Schadenbedarf.

4 Anpassung von Generalisierten Linearen Modellen an S und V Datensatz

In diesem Kapitel werden verschiedene statistische Modelle an den S und den V Datensatz (KH 2006) angepasst. Zuerst wird beschrieben, welche Anforderungen an die verwendeten Modelle gestellt werden. Daraufhin werden zwei Modelltypen ausgewählt.

Diese beiden Modelltypen, das Gamma-Modell mit Offset und Gewicht und das Poisson-Modell mit Offset und Gewicht, werden anschließend vorgestellt. Die beiden Modelle werden an die manipulierten Daten angepasst. Des Weiteren werden die Auswirkungen der Manipulation der Daten auf die Modelle beschrieben. Es folgt der R Output des Modells V3G1.

Man untersucht den Goodness of Fit der vollen und der reduzierten Modelle. Die Modelle werden abschließend mit dem Residual Devianz Test und dem Partial Devianz Test verglichen.

4.1 Problemstellung

Die an die statistischen Modelle gestellten Anforderungen werden nun beschrieben.

Ziel dieser Arbeit ist es, wie bereits in Abschnitt 3.5 erwähnt, zwei Kraftfahrt Haftpflicht Tarife zu erstellen. Ein Tarif soll den aktuellen VKB Tarif nachempfinden, ein zweiter Tarif mit anderen Merkmalen soll diesem gegenüberstehen. Danach wird entschieden, welcher der beiden Tarife die Daten besser anpasst.

An die beiden Tarife werden bestimmte Bedingungen gestellt. Eine Bedingung ist, dass der entstehende Tarif eine multiplikative Struktur aufweisen soll. Diese Anforderung be-

gründet sich dadurch, dass sich ein multiplikativer Tarif in der Praxis leichter implementieren lässt.

Es wird mit verschiedenen Aggregationsstufen gearbeitet, dem unaggregierten KH Datensatz und den aggregierten S und V Datensätzen. Dies stellt besondere Anforderungen an die gewählte Verteilung. Es muss möglich sein, Verteilungsaussagen sowohl über den aggregierten Datensatz, als auch über die Beobachtungen auf Einzeldatenebene zu treffen.

In Abschnitt 2.1 wurden verschiedene Verteilungen bezüglich ihrer Faltungseigenschaften untersucht. Es wurde festgestellt, dass die Summe zweier Pareto-verteilter Zufallsvariablen nicht Pareto-verteilt ist. Dasselbe gilt für die Summe von Zero-Inflated-Poisson oder Zero-Inflated-Gamma verteilten Zufallsvariablen. Jedoch gilt für die Summe von Gamma-verteilten Zufallsvariablen, dass diese ebenfalls Gamma-verteilt ist. Auch die Summe zweier Poisson-verteilter Zufallsvariablen ist Poisson-verteilt.

Bei dieser Problemstellung kommen somit nur die Gamma- und die Poisson-Regression in Frage.

Wendet man sich wieder der Forderung nach einem multiplikativem Tarif zu, dann muss die zugehörige Linkfunktion betrachtet werden. Die kanonische Linkfunktion der Gamma-Regression ist nicht geeignet. Wählt man aber die Linkfunktion $g(\mu) = \log(\mu) = \boldsymbol{x}^T\boldsymbol{\beta}$, so entsteht ein multiplikativer Tarif.

Für die Poisson-Regression entsteht unter Verwendung der kanonischen Linkfunktion $g(\mu) = \log(\mu) = \boldsymbol{x}^T\boldsymbol{\beta}$ ein multiplikativer Tarif. Somit erfüllen sowohl die Gamma-Regression als auch die Poisson-Regression die geforderten Kriterien.

Bei der explorativen Datenanalyse des S und des V Datensatzes hat man bereits festgestellt, dass nicht alle Zeilen die gleiche Anzahl an Jahreseinheiten haben. Dies muss bei der Modellierung durch unterschiedliche Gewichtung der Zellen berücksichtigt werden.

Ebenso muss die Tatsache, dass das Merkmal Beitrag die Informationen der übrigen Merkmale des VKB Tarifs enthält, Berücksichtigung finden.

Es werden nun die Modelle mit diesen beiden Verteilungen für den Schadenbedarf vorgestellt. In Tabelle 4.1 findet man eine Zusammenfassung der künftig verwendeten Bezeichnungen.

Tabelle 4.1: Zusammenfassung der im Folgenden verwendeten Bezeichnungen

Name	**Mathematische Definition**	**Erläuterung**
m		Anzahl der Zellen
k_i		Anzahl der Beobachtungen in Zelle i
n_i		Anzahl der Schäden in Zelle i, Zufallsvariable
S_i	$S_i := \sum_{j=1}^{k_i} R_{ij} = D_i n_i$	Gesamtschadensumme der Zelle i
R_{ij}		Schadenrisiko der j-ten Beobachtung in Zelle i, Zufallsvariable
u_{ij}		Jahreseinheit der j-ten Beobachtung in Zelle i
J_i	$J_i := \sum_{j=1}^{k_i} u_{ij}$	Summe der Jahreseinheiten in Zelle i
B_i	$B_i := \sum_{j=1}^{k_i} \frac{B_{ij}}{J_i}$	Durchschnittlicher Beitrag in Zelle i
D_i	$D_i := \frac{S_i}{n_i}$	Durchschnittlicher Schaden in Zelle i
D_{ij}		Durchschnittlicher Schaden der j-ten Beobachtung in Zelle i
SB_i	$SB_i := \frac{S_i}{B_i J_i} = \frac{D_i}{B_i J_i} n_i$	Schadenbedarf in Zelle i
SB_{ij}	$SB_{ij} := \frac{R_{ij}}{B_i J_i}$	Schadenbedarf der j-ten Beobachtung in Zelle i
μ_i		Erwartungswert für Zelle i bei Gamma-Verteilung
λ_i		Erwartungswert für Zelle i bei Poisson-Verteilung

4.2 Gamma-Modell mit Offset und Gewicht

In diesem Abschnitt wird das Gamma Modell mit Offset und Gewicht vorgestellt. Es wird zur Modellierung des Schadenbedarfs SB verwendet.

Man betrachtet zuerst die Verteilung der Daten auf Einzeldatenebene. Das j-te Schadenrisiko R_{ij} in Zelle i ist unabhängig für alle $j = 1, ..., k_i,\ i = 1, ..., m$ und Gamma-verteilt mit Skalenparameter $\mu_i u_{ij}$ und Formparameter νu_{ij} ($R_{ij} \sim \Gamma(\mu_i u_{ij}, \nu u_{ij})$). Somit ist sowohl der Skalen-, als auch der Formparameter abhängig von den Jahreseinheiten u_{ij}.

Der Gesamtschaden S_i für Zelle i $\left(S_i = \sum_{j=1}^{k_i} R_{ij}\right)$ entsteht durch das Summieren der Einzelrisiken R_{ij}. Auch die S_i sind für alle $i = 1, ..., m$ unabhängig. Mit den Additivitäts-

eigenschaften der Gamma-Verteilung aus Abschnitt 2.1 folgt, dass S_i Gamma-verteilt ist mit Skalenparameter $\mu_i J_i$ und Formparameter νJ_i ($S_i \sim \Gamma(\mu_i J_i, \nu J_i)$). Die Variable J_i entsteht als Summe der Jahreseinheiten u_{ij}.

Nun untersucht man den Schadenbedarf SB_{ij} des unaggregierten KH Datensatzes. Der Schadenbedarf der j-ten Beobachtung in Zelle i SB_{ij} ist ebenfalls Gamma-verteilt, siehe Abschnitt 2.1, mit Skalenparameter $\frac{\mu_i u_{ij}}{B_i J_i}$ und Formparameter νu_{ij} $\left(SB_{ij} \sim \Gamma\left(\frac{\mu_i u_{ij}}{B_i J_i}, \nu u_{ij}\right)\right)$ und unabhängig für alle $j = 1, ..., k_i,\ i = 1, ..., m$.

Betrachtet man den Schadenbedarf SB_i des aggregierten S oder V Datensatzes, dann ist der Schadenbedarf SB_i für Zelle i Gamma-verteilt. Mit Hilfe von Abschnitt 2.1 folgt, dass SB_i Gamma-verteilt ist mit Skalenparameter $\frac{\mu_i}{B_i}$ und Formparameter νJ_i $\left(SB_i \sim \Gamma\left(\frac{\mu_i}{B_i}, \nu J_i\right)\right)$.

An den Schadenbedarf SB_i soll nun ein Gamma-verteiltes Generalisiertes Lineares Modell angepasst werden. Da die Varianz nicht homogen ist, wird J_i als Gewicht verwendet. Der Beitrag B_i ist bekannt und muss somit nicht geschätzt werden. Er geht als Offset in das Modell ein. Aufgrund der Forderung eines multiplikativen Tarifs wird der Logarithmus als Linkfunktion verwendet.

Insgesamt erhält man für die Gamma-Regression den Erwartungswert:

$$E(SB_i) = \tilde{\mu}_i = \exp(\log(\mu_i) - \log B_i) = \exp(\boldsymbol{x}_i^T \boldsymbol{\beta} - \log B_i).$$

Zur Berechnung der Modelle wird die Software ***R*** verwendet. Der zugehörige R Code für die Gamma-Regression mit Offset und Gewicht sieht wie folgt aus.

R Code für die Gamma-Regression mit Offset und Gewicht

```
glm (  S_SB  ~ offset ( -log(S_B)) + a + c + b,
               family = Gamma ( link = "log" ),
               weights = S_J )
```

In den nächsten Abschnitten dieser Arbeit beziehen sich alle verwendeten Bezeichnung der Gamma-Modelle mit Offset und Gewicht auf Tabelle 4.2.

Tabelle 4.2: Benennung der Gamma-Modelle mit Offset und Gewicht für S und V Datensatz

Name	**Response**	**Kovariablen**	**Verteilung**	**Link**	**Offset**	**Gewicht**
S0G	S_SB1		Gamma	log	$-\log(B_i)$	J_i
S1G	S_SB1	a	Gamma	log	$-\log(B_i)$	J_i
S2G	S_SB1	c	Gamma	log	$-\log(B_i)$	J_i
S3G	S_SB1	b	Gamma	log	$-\log(B_i)$	J_i
S4G	S_SB1	c + b	Gamma	log	$-\log(B_i)$	J_i
S5G	S_SB1	a + b	Gamma	log	$-\log(B_i)$	J_i
S6G	S_SB1	a + c	Gamma	log	$-\log(B_i)$	J_i
S7G	S_SB1	a + c + b	Gamma	log	$-\log(B_i)$	J_i
V0G	V_SB1		Gamma	log	$-\log(B_i)$	J_i
V1G	V_SB1	A	Gamma	log	$-\log(B_i)$	J_i
V2G	V_SB1	B	Gamma	log	$-\log(B_i)$	J_i
V3G	V_SB1	A + B	Gamma	log	$-\log(B_i)$	J_i

4.3 Poisson-Modell mit Offset und Gewicht

Das Poisson-Modell mit Offset und Gewicht wird nun vorgestellt. Es wird zur Modellierung des Schadenbedarfs SB verwendet.

Sei n_i, die Anzahl der Parameter in Zelle i, Poisson-verteilt mit Parameter λ_i ($n_i \sim Poi(\lambda_i)$) und seien die n_i unabhängig für alle $i = 1, ..., m$.

Dann kann man mit der Additivitätseigenschaft der Poisson-Verteilung aus Abschnitt 2.1 zeigen, dass der Gesamtschaden S_i in Zelle i Poisson-verteilt ist. Der Gesamtschaden S_i ist das Produkt aus der Anzahl der Schäden und der durchschnittlichen Schadenhöhe ($S_i = D_i n_i$). Für den Erwartungswert und die Varianz gilt: $E(S_i) = D_i \lambda_i$ und $Var(S_i) = D_i^2 \lambda_i$.

Untersucht man den Schadenbedarf SB_{ij} des unaggregierten KH Datensatzes, dann ist der Schadenbedarf SB_{ij} der j-ten Beobachtung in Zelle i ebenfalls Poisson-verteilt, siehe Abschnitt 2.1, und unabhängig für alle $i = 1, ..., m,\ j = 1, ..., k_i$. Der Erwartungswert ist in diesem Fall $E(SB_{ij}) = \frac{D_{ij}}{J_i B_i}\lambda_i$ und die Varianz ist $Var(SB_{ij}) = \left(\frac{D_{ij}}{J_i B_i}\right)^2 \lambda_i$. Hierbei ist D_{ij} der durchschnittliche Schaden von Beobachtung j in Zelle i.

Der Schadenbedarf SB_i des aggregierten S und V Datensatzes wird ebenfalls betrachtet. Es gilt, dass auch der Schadenbedarf SB_i der Zelle i Poisson-verteilt und für alle $i = 1, ..., m$ unabhängig ist. Der Erwartungswert ist gegeben durch $E(SB_i) = \frac{D_i}{J_i B_i}\lambda_i$ und die Varianz durch $Var(SB_i) = \left(\frac{D_i}{J_i B_i}\right)^2 \lambda_i$.

Es soll in diesem Fall erneut ein Generalisiertes Lineares Modell an den Schadenbedarf SB_i angepasst werden. Jedoch wird hier die Poisson-Verteilung zu Grunde gelegt. Da die Varianz ebenfalls nicht homogen ist wird $\frac{D_i}{J_i B_i}$ als Gewicht verwendet. Da nur der Parameter λ_i geschätzt werden soll geht $\frac{D_i}{J_i B_i}$ auch als Offset in das Modell ein. Es wird die kanonische Linkfunktion, der Logarithmus, verwendet um einen Tarif mit multiplikativer Struktur zu erhalten.

Somit gilt nun für den Erwartungswert der Poisson-Regression:

$$E(SB_i) = \tilde{\lambda}_i = \exp\left(\boldsymbol{x}_i^T \boldsymbol{\beta} + \log\left(\frac{D_i}{J_i B_i}\right)\right).$$

Die Regressionsparameter werden ebenfalls mit der Software ***R*** berechnet. Es folgt der in diesem Fall verwendete ***R*** Code.

R Code für die Poisson-Regression mit Offset und Gewicht

```
glm (  S_SB  ~ offset ( log( S_D / (S_J S_B) ) ) + a + c + b,
               family = Poisson,
               weights = S_D / (S_J S_B) )
```

In Tabelle 4.3 sind die Beziehungen der Poisson-Modelle mit Offset und Gewicht, die in den folgenden Abschnitten dieser Arbeit verwendet werden, zusammengefasst.

Tabelle 4.3: Benennung der Poisson-Modelle mit Offset und Gewicht für S und V Datensatz

Name	**Response**	**Kovariablen**	**Verteilung**	**Link**	**Offset**	**Gewicht**
S0P	S_SB1		Poisson	log	$\log\left(\frac{D_i}{J_iB_i}\right)$	$\frac{J_iB_i}{D_i}$
S1P	S_SB1	a	Poisson	log	$\log\left(\frac{D_i}{J_iB_i}\right)$	$\frac{J_iB_i}{D_i}$
S2P	S_SB1	c	Poisson	log	$\log\left(\frac{D_i}{J_iB_i}\right)$	$\frac{J_iB_i}{D_i}$
S3P	S_SB1	b	Poisson	log	$\log\left(\frac{D_i}{J_iB_i}\right)$	$\frac{J_iB_i}{D_i}$
S4P	S_SB1	c + b	Poisson	log	$\log\left(\frac{D_i}{J_iB_i}\right)$	$\frac{J_iB_i}{D_i}$
S5P	S_SB1	a + b	Poisson	log	$\log\left(\frac{D_i}{J_iB_i}\right)$	$\frac{J_iB_i}{D_i}$
S6P	S_SB1	a + c	Poisson	log	$\log\left(\frac{D_i}{J_iB_i}\right)$	$\frac{J_iB_i}{D_i}$
S7P	S_SB1	a + c + b	Poisson	log	$\log\left(\frac{D_i}{J_iB_i}\right)$	$\frac{J_iB_i}{D_i}$
V0P	V_SB1		Poisson	log	$\log\left(\frac{D_i}{J_iB_i}\right)$	$\frac{J_iB_i}{D_i}$
V1P	V_SB1	A	Poisson	log	$\log\left(\frac{D_i}{J_iB_i}\right)$	$\frac{J_iB_i}{D_i}$
V2P	V_SB1	B	Poisson	log	$\log\left(\frac{D_i}{J_iB_i}\right)$	$\frac{J_iB_i}{D_i}$
V3P	V_SB1	A + B	Poisson	log	$\log\left(\frac{D_i}{J_iB_i}\right)$	$\frac{J_iB_i}{D_i}$

4.4 Modellanpassung an die manipulierten Daten

In diesem Abschnitt wird untersucht, wie sich die Manipulation des Schadenbedarfs SB auf die Schätzer auswirkt.

Da der Schadenbedarf SB manipuliert wurde, ist es an dieser Stelle notwendig die Benennung der Modelle aus den Tabellen 4.2 und 4.3 zu verfeinern.

Der Betrag mit dem der Schadenbedarf manipuliert wurde, wird an die Bezeichnung des Modells angefügt. Wird der Schadenbedarf zum Beispiel mit einem Cent manipuliert, dann wird aus dem Modell S6G das Modell S6G1, siehe Tabelle 4.4.

Tabelle 4.4: Bezeichnung der Modelle nach der Manipulation der Daten

Manipulation	**Bezeichnung**					
0 Cent	S6G	S6P	S7G	S7P	V3G	V3P
1 Cent	S6G1	S6P1	S7G1	S7P1	V3G1	V3P1
10 Cent	S6G10	S6P10	S7G10	S7P10	V3G10	V3P10
100 Cent	S6G100	S6P100	S7G100	S7P100	V3G100	V3P100

In Tabelle 4.5 werden die manipulierten V3G Modelle untersucht. V3G Modelle sind Gamma-verteilte Modelle, die an den V Datensatzes mit den Kovariablen A und B angepasst wurden.

Tabelle 4.5: Gegenüberstellung der Regressionsparameter der manipulierten V3G Modelle

	V3G1	**V3G10**	**Differenz**	**V3G100**	**Differenz**
Intercept	3.032	3.041	−0.008	3.121	−0.088
A1	0.000	0.000	0.000	0.000	0.000
A2	0.178	0.178	0.000	0.175	0.003
A3	0.150	0.150	0.000	0.148	0.002
A4	0.443	0.443	0.001	0.437	0.006
A5	0.425	0.424	0.001	0.419	0.006
A6	0.653	0.653	0.001	0.645	0.008
A7	0.640	0.639	0.001	0.632	0.008
A8	0.565	0.564	0.001	0.558	0.007
A9	0.554	0.554	0.001	0.549	0.006
A10	1.709	1.708	0.001	1.693	0.016
A11	1.250	1.249	0.001	1.239	0.011
A12	1.599	1.597	0.001	1.584	0.015
A13	1.509	1.508	0.001	1.497	0.012
B10	0.000	0.000	0.000	0.000	0.000
B11	2.048	2.041	0.007	1.975	0.073
B12	1.583	1.576	0.007	1.514	0.069
B13	1.329	1.323	0.006	1.266	0.064
B14	1.516	1.510	0.006	1.448	0.068

B15	1.471	1.465	0.006	1.404	0.067
B16	1.594	1.587	0.007	1.525	0.069
B17	1.576	1.569	0.007	1.507	0.069
B18	1.703	1.696	0.007	1.632	0.071
B19	1.817	1.810	0.007	1.745	0.072
B20	1.848	1.841	0.007	1.775	0.072
B21	1.765	1.759	0.007	1.695	0.071
B22	1.611	1.605	0.006	1.543	0.068
B23	2.143	2.136	0.007	2.067	0.077
B24	2.160	2.153	0.007	2.084	0.076
B25	2.808	2.801	0.008	2.726	0.082

Betrachtet man die Abweichung der mit einem Cent und mit 10 Cent manipulierten Regressionsparameter in Tabelle 4.5, dann besteht kaum ein Unterschied zwischen den Regressionsparametern der verschiedenen Modelle. Auch der Unterschied der Parameter bei der Manipulation des Schadenbedarfs mit einem Cent oder mit 100 Cent fällt gering aus.

Insgesamt verändert also die Manipulation des Schadenbedarfs die Ergebnisse der Gamma-Regression kaum. Im weiteren Verlauf dieser Arbeit wird bei den Gamma Modellen mit dem durch einen Cent manipulierten Schadenbedarf gearbeitet.

Die Gruppe der S6P Modelle wird in Tabelle 4.6 untersucht. Bei diesen Modellen werden die Merkmale a und c des aggregierten S Datensatzes mit der Poisson-Verteilung modelliert.

Tabelle 4.6: Gegenüberstellung der Regressionsparameter der manipulierten S6P Modelle

	S6P	**S6P1**	**Differenz**	**S6P10**	**Differenz**	**S6P100**	**Differenz**
Intercept	3.319	3.462	−0.143	4.106	−0.788	5.698	−2.379
a0	0.000	0.000	0.000	0.000	0.000	0.000	0.000
a1	−1.473	−1.314	−0.158	−0.709	−0.764	−0.179	−1.294
a2	−3.169	−2.152	−1.017	−0.609	−2.560	0.191	−3.360

a3	−4.798	−2.520	−2.278	−0.682	−4.116	0.164	−4.962
c0	0.000	0.000	0.000	0.000	0.000	0.000	0.000
c1	1.000	0.844	0.157	0.349	0.651	0.039	0.961
c2	0.271	0.046	0.225	−0.825	1.096	−1.688	1.959
c3	1.373	1.275	0.098	0.989	0.384	0.829	0.544
c4	0.268	0.322	−0.054	0.428	−0.160	0.464	−0.196
c5	1.123	0.977	0.145	0.519	0.604	0.232	0.891
c6	1.026	1.064	−0.038	1.127	−0.101	1.140	−0.113
c7	0.971	0.758	0.212	−0.006	0.977	−0.629	1.599
c8	0.517	0.397	0.120	0.037	0.480	−0.174	0.690
c9	1.088	0.955	0.133	0.539	0.549	0.280	0.808
c10	0.575	0.402	0.174	−0.215	0.791	−0.704	1.280
c11	0.245	0.086	0.159	−0.473	0.718	−0.907	1.152
c12	0.501	0.515	−0.014	0.538	−0.037	0.541	−0.040
c13	−0.072	−0.155	0.083	−0.389	0.317	−0.517	0.445
c14	0.660	0.432	0.229	−0.503	1.164	−1.613	2.274
c15	0.345	0.337	0.008	0.300	0.045	0.272	0.073
c16	0.021	0.155	−0.134	0.445	−0.424	0.572	−0.551
c17	−0.611	−0.396	−0.215	−0.005	−0.606	0.134	−0.745
c18	−0.788	−0.504	−0.285	−0.007	−0.781	0.173	−0.961

Betrachtet man die Tabelle, dann ist eine Differenz zwischen den Regressionsparametern der ursprünglichen Daten und der mit einem Cent, 10 Cent und 100 Cent manipulierten Daten zu erkennen. Dieser Unterschied fällt jedoch in den meisten Fällen gering aus.

Somit kommt man auch hier zu dem Schluss, dass die Manipulation des Schadenbedarfs die Ergebnisse der Poisson-Regression kaum beeinflusst. Im weiteren Verlauf dieser Arbeit wird mit dem durch einen Cent manipulierten Schadenbedarf gearbeitet.

Weitere Tabellen mit Gegenüberstellungen von Regressionsparametern der manipulierten Modelle findet man im Anhang.

4.5 R Output des Modells V3G1

Es folgt der ***R*** Output des V3G1 Modells.

R Output des Modells V3G1

```
Call:
glm (formula = V_SB1 ~ offset(-log(V_B)) + A + B,
    family = Gamma(link = "log"), weights = V_J)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-23.62   -6.23   -2.42    2.20   18.27

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
Intercept   3.0324     0.6136    4.94  2.0e-06 ***
A2          0.1778     0.0493    3.61  0.00041 ***
A3          0.1497     0.0575    2.60  0.01015 *
A4          0.4433     0.0598    7.41  7.6e-12 ***
A5          0.4249     0.0693    6.13  6.7e-09 ***
A6          0.6534     0.0717    9.12  3.6e-16 ***
A7          0.6401     0.1110    5.77  4.2e-08 ***
A8          0.5649     0.1247    4.53  1.2e-05 ***
A9          0.5543     0.1421    3.90  0.00014 ***
A10         1.7091     0.3453    4.95  1.9e-06 ***
A11         1.2501     0.5167    2.42  0.01670 *
A12         1.5986     1.0671    1.50  0.13615
A13         1.5094     3.0893    0.49  0.62583
B11         2.0482     1.1781    1.74  0.08408 .
B12         1.5825     0.6480    2.44  0.01572 *
B13         1.3293     0.6225    2.14  0.03430 *
B14         1.5165     0.6152    2.47  0.01478 *
B15         1.4714     0.6144    2.39  0.01782 *
B16         1.5939     0.6144    2.59  0.01038 *
B17         1.5756     0.6143    2.56  0.01127 *
B18         1.7028     0.6150    2.77  0.00631 **
B19         1.8174     0.6155    2.95  0.00364 **
B20         1.8478     0.6198    2.98  0.00333 **
B21         1.7654     0.6232    2.83  0.00522 **
B22         1.6115     0.6431    2.51  0.01324 *
B23         2.1432     0.6374    3.36  0.00097 ***
B24         2.1602     0.6795    3.18  0.00178 **
B25         2.8083     0.7972    3.52  0.00056 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 34.1)

    Null deviance: 17694.0  on 183  degrees of freedom
Residual deviance:  8541.6  on 156  degrees of freedom
AIC: -310549

Number of Fisher Scoring iterations: 2
```

Bei Modell V3G1 wird der Schadenbedarf V_SB1 des V Datensatzes mit Offset und Gewicht mit der Gamma-Regression modelliert. Die Kovariablen A und B sind Dummie kodiert. Weitere ***R*** Outputs finden sich im Anhang.

Es fällt bei der Betrachtung des ***R*** Outputs auf, dass die Ausprägungen A1 (Merkmal A) und B10 (Merkmal B) nicht vorhanden sind. Diese sind nicht aufgeführt, weil sie den Wert Null haben. Der Wert dieser beiden Ausprägungen ist Null, da die Regressionsparameter auf der logarithmischen Skala angegeben sind. Man muss den Wert der Exponentialfunktion an der Stelle der Regressionsparameter berechnen um μ_i zu erhalten.

4.6 Goodness of Fit für die vollen und reduzierten Modelle

Die Anpassungsgüte der verschiedenen Modelle an den S Datensatz (KH 2006) und den V Datensatz (KH 2006) wird in diesem Abschnitt untersucht. Die Untersuchung erfolgt getrennt nach Datensätzen und separiert nach Poisson-verteilten und Gamma-verteilten Modellen. Somit werden vier Gruppen untersucht.

Diese getrennte Betrachtung ist notwendig, da mit dem AIC-Wert und der Devianz nur genestete Modelle verglichen werden können. Es werden nun für jede der vier Gruppen die Modelle mit minimalem AIC-Wert und Devianz Wert gesucht. Diese Modelle sind am besten an die Daten angepasst, siehe Abschnitt 2.3.3.

4.6.1 Anpassungsgüte der Gamma-verteilten Modelle an den S Datensatz

Zuerst wird die Anpassungsgüte der verschiedenen Gamma-Modelle an den S Datensatz untersucht.

In Tabelle 4.7 ist neben dem AIC-Wert und der Devianz auch die Anzahl der Freiheitsgrade (df) der jeweiligen Modelle aufgeführt. Man berechnet den AIC-Wert und den Wert der Devianz entsprechend der in Abschnitt 2.3.7 angegebenen Formeln.

Man erkennt, dass der AIC-Wert für Modell S7G1, das volle Gamma-verteilte Modell des S Datensatzes, den kleinsten Wert annimmt. Somit würde man dieses Modell den anderen

Tabelle 4.7: AIC-Wert, Devianz und Log-Likelihood der Gamma-Modelle des S Datensatzes

Modell	**Kovariablen**	**Log-Likelihood**	**AIC**	**df**	**unskalierte Devianz**	$\hat{\phi}$	**Devianz**
S0G1		102535	-205066	319	22474	118	709
S1G1	a	116879	-233748	316	17499	41.2	552
S2G1	c	105424	-210809	301	21372	114	674
S3G1	b	103261	-206511	315	22192	113	700
S4G1	c + b	106206	-212364	297	21083	111	665
S5G1	a + b	117697	-235375	312	17250	34.4	544
S6G1	a + c	120525	-241004	298	16416	36.6	517
S7G1	a + c + b	121456	-242858	294	16151	31.7	509

Modellen vorziehen. Doch muss auch bemerkt werden, dass sich der AIC-Wert des S7G1 Modells und des S6G1 Modells nur geringfügig unterscheiden.

Betrachtet man die Devianz, dann ist der Wert des vollen Modells am kleinsten. Aber auch hier unterscheiden sich die Werte der Devianz für die Modelle S7G1 und S6G1 kaum.

Dennoch wird in dieser Gruppe das Modell S7G1 als das Modell gewählt, welches am besten an die Daten angepasst ist.

4.6.2 Anpassungsgüte der Poisson-verteilten Modelle an den S Datensatz

Die Anpassungsgüte der Poisson-Modelle an den S Datensatz wird in Tabelle 4.8 betrachtet.

Die Tabelle beinhaltet neben dem AIC-Wert, den Wert der Devianz auch die Anzahl der Freiheitsgrade (df). Die Formeln um den AIC-Wert und den Wert der Devianz zu berechnen sind in Abschnitt 2.3.6 angegeben.

Wertet man die Ergebnisse der Tabelle aus, dann bietet sich ein ähnliches Bild wie im vorherigen Abschnitt. Obwohl anstelle der Gamma-Verteilung die Poisson-Verteilung verwendet wird, ist der AIC-Wert beim vollen Modell am niedrigsten. Aber auch hier gilt, dass sich der AIC-Wert des vollen Modells und des S6P1 Modells nur geringfügig unterscheiden.

Tabelle 4.8: AIC-Wert, Devianz und Log-Likelihood der Poisson-Modelle des S Datensatzes

Modell	Kovariablen	Log-Likelihood	AIC	df	Devianz
S0P1		-30050	60101	319	12005
S1P1	a	-27260	54527	316	6425
S2P1	c	-29399	58836	301	10703
S3P1	b	-29221	58451	315	10347
S4P1	c + b	-28480	57006	297	8865
S5P1	a + b	-25888	51791	312	3681
S6P1	a + c	-26488	53020	298	4882
S7P1	a + c + b	-25075	50202	294	2056

Der Wert der Devianz ist für das Modell S7P1 am geringsten. Dies führt zu dem Schluss, dass das Modell S7P1 am besten an die Daten angepasst ist.

Insgesamt entscheiden sich beim S Datensatz beide Vergleichskriterien für das Modell mit allen Variablen, unabhängig welche Verteilung zugrunde liegt.

4.6.3 Anpassungsgüte der Gamma-verteilten Modelle an den V Datensatz

Als nächstes wird nun die Anpassungsgüte der unterschiedlichen Gamma-Modelle an den V Datensatz betrachtet.

Tabelle 4.9: AIC-Wert, Devianz und Log-Likelihood der Gamma-Modelle des V Datensatzes

Modell	Kovariablen	Log-Likelihood	AIC	df	unskalierte Devianz	$\hat{\phi}$	Devianz
V0G1		120396	-240788	183	15816	126	503
V1G1	A	140280	-280532	171	11144	55.6	355
V2G1	B	121532	-243030	168	15504	139	495
V3G1	A + B	155303	-310549	156	8542	34.1	250

Neben dem AIC-Wert und der Devianz hält Tabelle 4.9 die Anzahl der Freiheitsgrade (df) fest. Die Werte des AIC und der Devianz werden wie in Abschnitt 2.3.7 angegeben

berechnet.

Betrachtet man die AIC-Werte der Modelle, dann hat das Modell mit allen Kovariablen, V3G1, den niedrigsten AIC-Wert.

Das Modell V3G1 hat auch den niedrigsten Wert der Devianz. Für diese Gruppe ist somit das volle Modell am besten an die Daten angepasst.

4.6.4 Anpassungsgüte der Poisson-verteilten Modelle an den V Datensatz

In Tabelle 4.10 wird die Anpassungsgüte der Poisson-Modelle, die an den V Datensatz angepasst wurden, betrachtet.

Tabelle 4.10: AIC-Wert, Devianz und Log-Likelihood der Poisson-Modelle des V Datensatzes

Modell	Kovariablen	Log-Likelihood	AIC	df	Devianz
V0P1		-56725	113451	183	10847
V1P1	A	-37260	74527	171	6663
V2P1	B	-29399	58836	168	6763
V3P1	A + B	-29220	58451	156	1288

In der obigen Tabelle sind unter anderem der AIC-Wert, die Devianz und die Anzahl der Freiheitsgrade (df) für die verschiedenen Modelle enthalten. Man berechnet den Wert des AIC und der Devianz mit den Formeln, die in Abschnitt 2.3.6 angegeben sind.

Es überrascht nicht, dass auch hier das volle Modell V3P1 den niedrigsten AIC-Wert hat.

Der Wert der Devianz ist ebenfalls minimal für das Modell mit allen Kovariablen. Somit ist auch in diesem Fall das volle Modell das Modell mit der besten Anpassung an die Daten.

Betrachtet man den V Datensatz, dann entscheiden sich beide Modellwahlkriterien für das Modell mit allen Variablen, unabhängig von der zugrunde gelegten Verteilung.

4.7 Modellvergleich mit dem Residual Devianz Test und dem Partial Devianz Test

Es werden nun zwei Hypothesentests, der Residual Devianz Test und der Partial Devianz Test, betrachtet. Wie im Abschnitt vorher will man entscheiden, welches Modell am besten an die Daten angepasst ist.

Auch hier erfolgt die Betrachtung getrennt nach Datensätzen und separiert nach Poisson-Verteilungsannahme und Gamma-Verteilungsannahme. Die Auswahl des besten Modells erfolgt nach den in Abschnitt 2.3.4 angegeben Kriterien.

4.7.1 Residual Devianz Test und Partial Devianz Test für die Gamma-verteilten Modelle des S Datensatzes

Zuerst wird untersucht, welches der Gamma-Modelle des S Datensatzes die Hypothesentests auswählen.

Tabelle 4.11 enthält die Ergebnisse des Residual Devianz Tests und des Partial Devianz Tests. Die Teststatistiken der beiden Tests werden wie in Abschnitt 2.3.7 angegeben berechnet.

Tabelle 4.11: Residual Devianz Test und Partial Devianz Test für die Gamma-Modelle des S Datensatzes

H* vs *K **Residual Devianz**	**Residual Devianz**	**p-Wert**	***H* vs *K*** **Partial Devianz**	**Partial Devianz**	**p-Wert**
H: S0G1 vs ***K***: not S0G1	709	1	***H***: S7G1 vs ***K***: S0G1	200	1
H: S1G1 vs ***K***: not S1G1	552	1	***H***: S7G1 vs ***K***: S1G1	13	0.9766
H: S2G1 vs ***K***: not S2G1	674	1	***H***: S7G1 vs ***K***: S2G1	165	1
H: S3G1 vs ***K***: not S3G1	700	1	***H***: S7G1 vs ***K***: S3G1	191	1
H: S4G1 vs ***K***: not S4G1	665	1	***H***: S7G1 vs ***K***: S4G1	145	1
H: S5G1 vs ***K***: not S5G1	544	1	***H***: S7G1 vs ***K***: S5G1	35	0.9998
H: S6G1 vs ***K***: not S6G1	517	1	***H***: S7G1 vs ***K***: S6G1	8	0.0009
H: S7G1 vs ***K***: not S7G1	509	1			

Um die Partial Devianz zu erhalten wurde durch den geschätzten Dispersionsparameter des vollen Modells $\hat{\phi}_{S7G1}$ geteilt. Betrachtet man den Residual Devianz Test, dann würde

man das volle Modell S7G1 wählen.

Der Partial Devianz Test hingegen trifft eine andere Entscheidung. Er entscheidet zugunsten des Modells S6G1. Trotzdem wird für diese Gruppe das Modell S7G1 als das beste Modell gewählt, da der AIC-Wert und der Wert der Devianz für dieses Modell am niedrigsten sind.

4.7.2 Residual Devianz Test und Partial Devianz Test für die Poisson-verteilten Modelle des S Datensatzes

Als nächstes werden in Tabelle 4.12 der Residual Devianz Test und Partial Devianz Test der Poisson-Modelle, die an den S Datensatz angepasst werden, betrachtet. Die Werte der Teststatistiken werden entsprechend der in Abschnitt 2.3.6 angegebenen Formeln berechnet.

Tabelle 4.12: Residual Devianz Test und Partial Devianz Test für die Poisson-Modelle des S Datensatzes

H* vs *K **Residual Devianz**	**Residual Devianz**	**p-Wert**	***H* vs *K*** **Partial Devianz**	**Partial Devianz**	**p-Wert**
H: S0P1 vs ***K***: not S0P1	12005	1	***H***: S7P1 vs ***K***: S0P1	9949	1
H: S1P1 vs ***K***: not S1P1	6425	1	***H***: S7P1 vs ***K***: S1P1	4369	1
H: S2P1 vs ***K***: not S2P1	10703	1	***H***: S7P1 vs ***K***: S2P1	8647	1
H: S3P1 vs ***K***: not S3P1	10347	1	***H***: S7P1 vs ***K***: S3P1	8291	1
H: S4P1 vs ***K***: not S4P1	8865	1	***H***: S7P1 vs ***K***: S4P1	6809	1
H: S5P1 vs ***K***: not S5P1	3681	1	***H***: S7P1 vs ***K***: S5P1	1625	1
H: S6P1 vs ***K***: not S6P1	4882	1	***H***: S7P1 vs ***K***: S6P1	2825	1
H: S7P1 vs ***K***: not S7P1	2056	1			

Der Residual Devianz Test wählt das volle Modell als bestes Modell. Der Partial Devianz Test entscheidet sich ebenfalls für das Modell S7P1. Somit ist das Modell S7P1 am besten an die Daten angepasst.

4.7.3 Residual Devianz Test und Partial Devianz Test für die Gamma-verteilten Modelle des V Datensatzes

Als nächstes werden die Gamma-Modelle des V Datensatzes mit Hilfe von Hypothesentests betrachtet.

In Tabelle 4.13 sind die Ergebnisse des Residual Devianz Tests und des Partial Devianz Tests enthalten. Die Formeln um die Werte der Teststatistiken des Residual Devianz Tests und des Partial Devianz Tests zu berechnen sind in Anschnitt 2.3.7 angegeben.

Tabelle 4.13: Residual Devianz Test und Partial Devianz Test für die Gamma-Modelle des V Datensatzes

H* vs *K **Residual Devianz**	**Residual Devianz**	**p-Wert**	***H* vs *K*** **Partial Devianz**	**Partial Devianz**	**p-Wert**
H: V0G1 vs ***K***: not V0G1	503	1	***H***: V3G1 vs ***K***: V0G1	253	1
H: V1G1 vs ***K***: not V1G1	355	1	***H***: V3G1 vs ***K***: V1G1	105	1
H: V2G1 vs ***K***: not V2G1	495	1	***H***: V3G1 vs ***K***: V2G1	245	1
H: V3G1 vs ***K***: not V3G1	250	0.9999			

Um die Partial Devianz zu erhalten wird durch den geschätzten Dispersionsparameter des vollen Modells $\hat{\phi}_{V3G1}$ geteilt.

Auch hier wählt man das Modell V3G1 als bestes Modell, wenn man den Residual Devianz Test betrachtet. Der Partial Devianz Test entscheidet ebenfalls, dass das Modell V3G1 die Daten am besten anpasst.

Somit kommt man zu dem Schluss, dass in dieser Gruppe das volle Modell am besten an die Daten angepasst ist.

4.7.4 Residual Devianz Test und Partial Devianz Test für die Poisson-verteilten Modelle des V Datensatzes

Abschließend werden die Poisson-Modelle des V Datensatzes unter Zuhilfenahme von Hypothesentests betrachtet.

Die Resultate des Residual Devianz Tests und des Partial Devianz Tests sind in Tabelle 4.14 zusammengefasst. Die Formeln zur Berechnung der Werte der Teststatistiken des Residual Devianz Tests und des Partial Devianz Tests sind in Abschnitt 2.3.6 angegeben.

Tabelle 4.14: Residual Devianz Test und Partial Devianz Test für die Poisson-Modelle des V Datensatzes

H* vs *K **Residual Devianz**	**Residual Devianz**	**p-Wert**	***H* vs *K*** **Partial Devianz**	**Partial Devianz**	**p-Wert**
H: V0P1 vs ***K***: not V0P1	10847	1	***H***: V3P1 vs ***K***: V0P1	9559	1
H: V1P1 vs ***K***: not V1P1	6663	1	***H***: V3P1 vs ***K***: V1P1	5375	1
H: V2P1 vs ***K***: not V2P1	6763	1	***H***: V3P1 vs ***K***: V2P1	5475	1
H: V3P1 vs ***K***: not V3P1	1288	1			

Basierend auf dem Residual Devianz Test wird das Modell mit allen Kovariablen ausgewählt. Der Partial Devianz Test trifft dieselbe Entscheidung. Auch hier ist das volle Modell V3P1 das beste Modell.

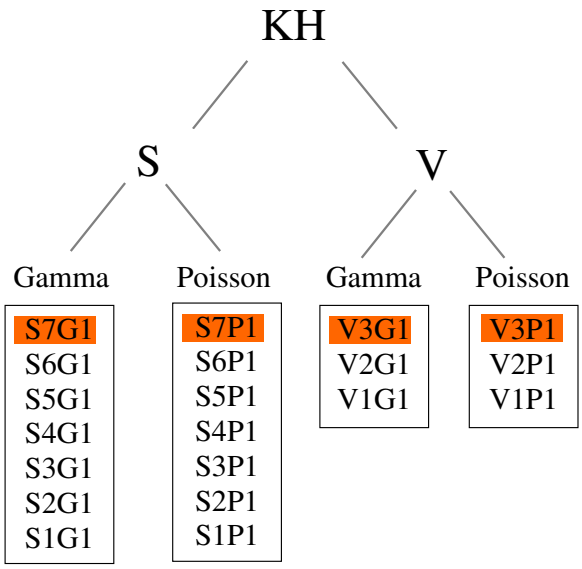
Insgesamt gilt hier, dass das Modell mit allen Variablen am besten an die Daten angepasst ist.

4.8 Zusammenfassung der Ergebnisse des Goodness of Fit, des Residual Devianz Tests und des Partial Devianz Tests

In den vorherigen Abschnitten wurden vier verschiedene Gruppen an Modellen untersucht. Pro Gruppe wird entschieden welches Modell am besten an die Daten angepasst ist.

In Abbildung 4.1 ist die Entstehung dieser Gruppen aufgezeigt. Aus dem KH Datensatz sind durch Aggregation der S Datensatz und der V Datensatz entstanden. An die aggregierten Datensätze werden jeweils Gamma-verteilte und Poisson-verteilte Modelle angepasst.

Abbildung 4.1: Ergebnisse der Modellwahl basierend auf der Anpassungsgüte, dem Residual Devianz Test und dem Partial Devianz Test



Betrachtet man die Gruppe der Gamma-verteilten Modelle des S Datensatzes so ist das S7G1 Modell das beste Modell, siehe Abbildung 4.1. Für die Poisson-verteilten Modelle des S Datensatzes ist das S7P1 Modell am besten an die Daten angepasst. Das beste Modell des V Datensatzes für die Gamma-verteilten Modelle ist das Modell V3G1 und für die Poisson-verteilten Modellen das V3P1 Modell.

5 Residuenanalyse und Interpretation für ausgewählte Modelle

In diesem Kapitel werden zuerst die Standardisierten Pearson Residuen und die Standardisierten Devianz Residuen der im letzten Abschnitt ausgewählten vier Modelle betrachtet. Im Anschluss wird der Einfluss der Kovariablen dieser Modelle illustriert.

5.1 Standardisierte Pearson Residuen und Standardisierte Devianz Residuen

Da bei der Modellierung mit Gewichten gearbeitet wird, werden die standardisierten Residuen betrachtet.

In den beiden vorhergehenden Abschnitten wurde mit Hilfe der Anpassungsgüte und durch Hypothesentests entschieden, welches Modell pro Gruppe am besten an die Daten angepasst ist.

Die vier besten Modelle sind das S7G1 Modell, das S7P1 Modell, das V3G1 Modell und das V3P1 Modell, siehe Tabelle 4.2 und Tabelle 4.3. Für diese Modelle werden die Standardisierten Pearson Residuen und die Standardisierten Devianz Residuen untersucht, siehe Abschnitt 2.3.5.

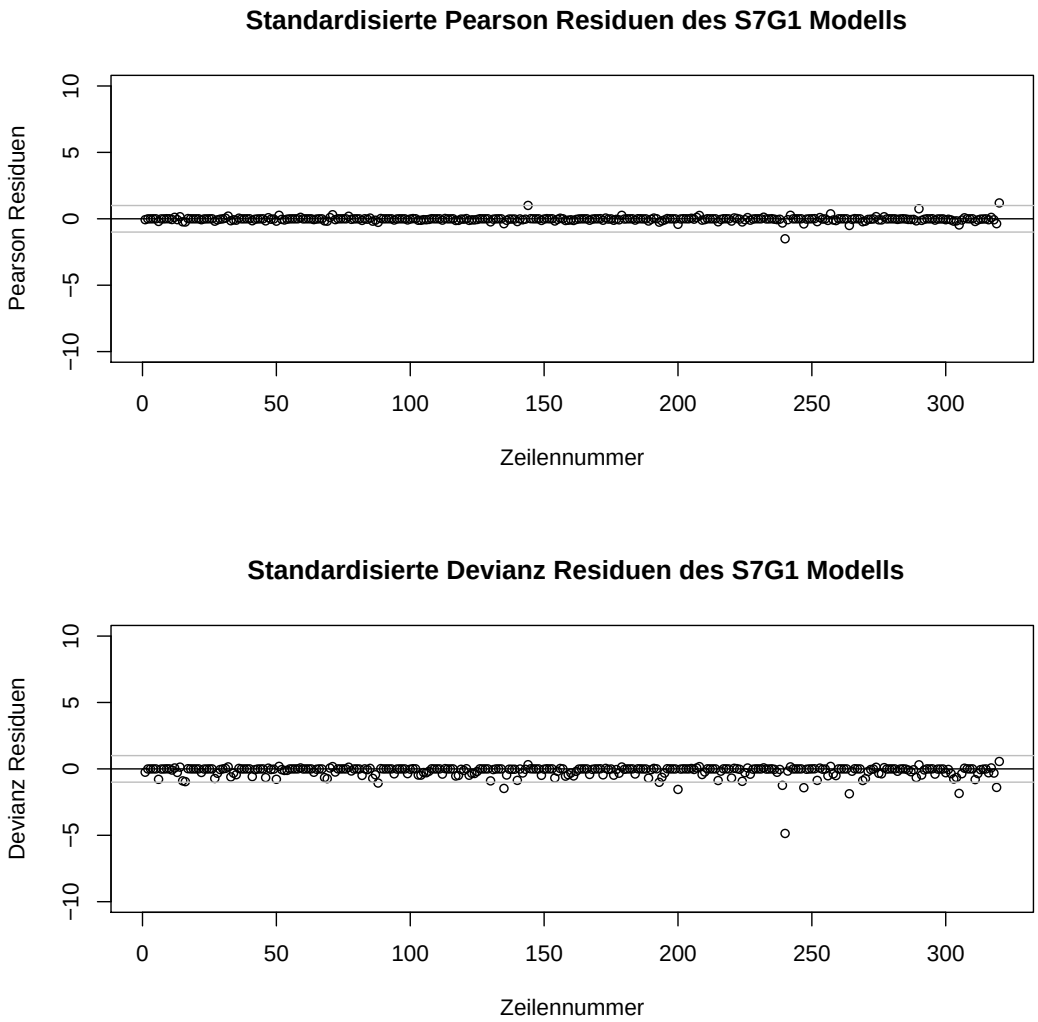
5.1.1 Standardisierte Pearson Residuen und Standardisierte Devianz Residuen des S7G1 Modells

Zuerst wird das S7G1 Modell betrachtet. Dieses Modell ist Gamma-verteilt und wurde an den S Datensatz mit dem Merkmalen a, b und c angepasst.

In Abbildung 5.1 sind die Standardisierten Pearson Residuen und die Standardisierten Devianz Residuen dargestellt. Die Residuen sind gegen die entsprechende Zeilennummer des S Datensatzes angetragen. Wie in Abschnitt 2.3.7 beschrieben berechnet man die Standardisierten Pearson Residuen und die Standardisierten Devianz Residuen.

Betrachtet man die beiden Plots, dann erkennt man, dass es bei diesem Modell kaum Residuen gibt, die den Wert ±1 übersteigen.

Abbildung 5.1: Standardisierte Pearson und Devianz Residuen des S7G1 Modells



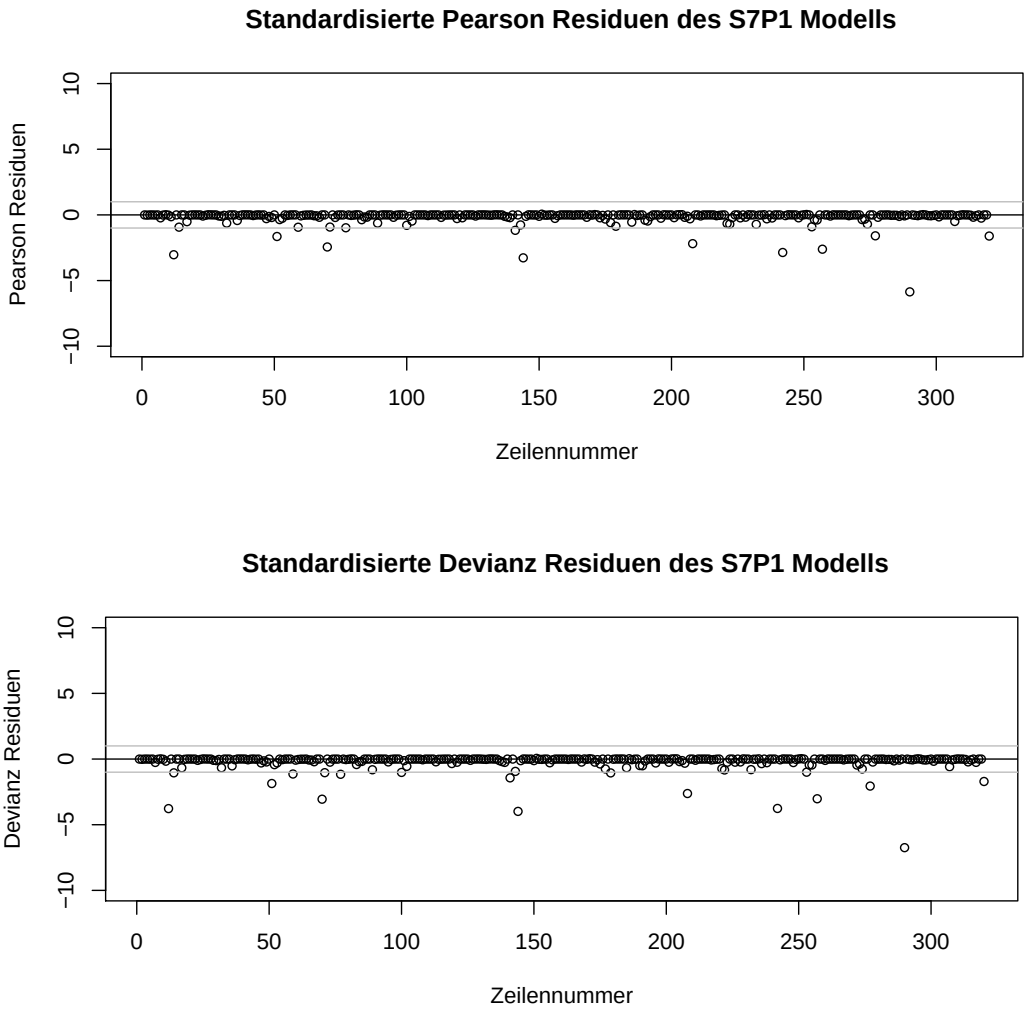
Bei den Standardisierten Devianz Residuen kann man wiederum erkennen, dass im Vergleich zu den Standardisierten Pearson Residuen mehr Residuen einen Wert größer als ±1 haben. Alle Standardisierten Devianz Residuen, die die Grenze ±1 überschreiten, sind negativ.

5.1.2 Standardisierte Pearson Residuen und Standardisierte Devianz Residuen des S7P1 Modells

Das nächste Modell, das untersucht wird, ist das S7P1 Modell. Wie das S7G1 Modell wurde es mit den Kovariablen a, b und c an den S Datensatz angepasst, aber die zugrunde liegende Verteilung ist die Poisson-Verteilung.

Eine Illustration der Standardisierten Pearson Residuen und der Standardisierten Devianz Residuen findet sich in Abbildung 5.2. Die Residuen sind wiederum gegen die entsprechende Zeilennummer des S Datensatzes angetragen. Die Berechnung der Standardisierten Pearson Residuen und der Standardisierten Devianz Residuen erfolgt wie in Abschnitt 2.3.6 angegeben.

Abbildung 5.2: Standardisierte Pearson und Devianz Residuen des S7P1 Modells



Es fällt auf, dass die Residuen für das S7P1 Modell weiter um Null streuen als beim vorherigen Modell. Es fällt auf, dass alle Residuen, die einen Wert größer als ± 1 haben negativ sind.

Allgemein ist zu sehen, dass die Beobachtungen, die große Standardisierte Pearson Residuen haben, auch große Standardisierte Devianz Residuen aufweisen. Die Streuung der Residuen um die Regressionsgerade erscheint ansonsten regelmäßig.

Vergleicht man die Residuen des S7G1 Modells mit den Residuen des S7P1 Modells, dann scheint das S7G1 Modell die Daten insgesamt besser anzupassen.

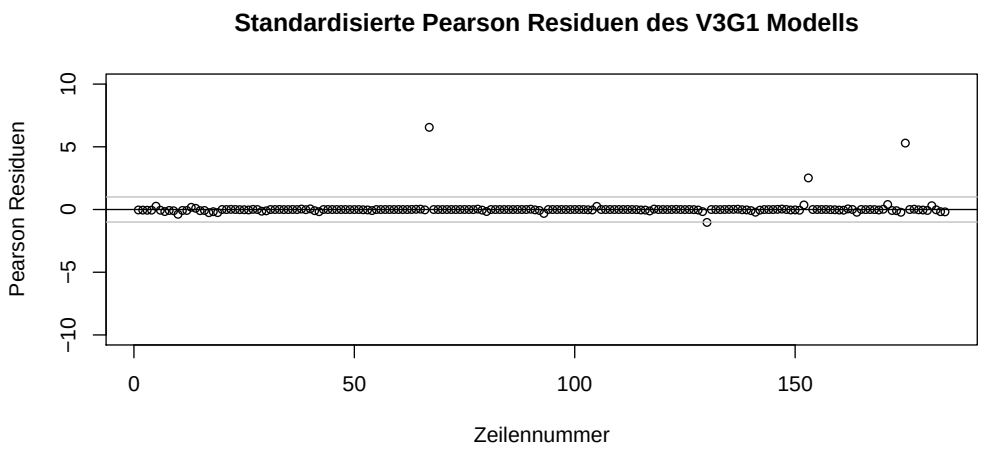
5.1.3 Standardisierte Pearson Residuen und Standardisierte Devianz Residuen des V3G1 Modells

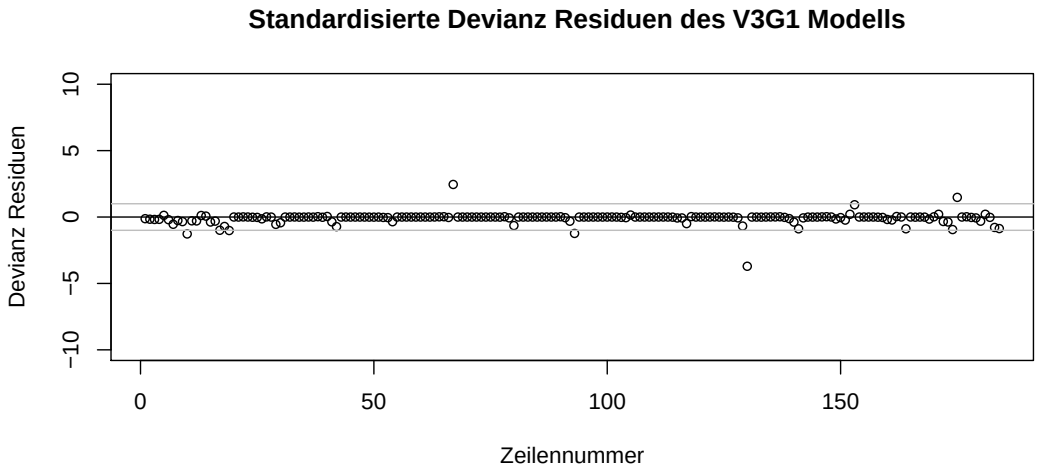
Es wird nun das V3G1 Modell untersucht. Dieses Modell verwendet die Gamma-Verteilung und wurde mit den Kovariablen A und B an den V Datensatz angepasst.

In Abbildung 5.3 sind die Standardisierten Pearson Residuen und die Standardisierten Devianz Residuen dargestellt. Die Residuen sind gegen die entsprechende Nummer des V Datensatzes angetragen und wurden wie in Abschnitt 2.3.7 beschrieben berechnet.

Betrachtet man die beiden Plots, dann erkennt man, dass sowohl für die Standardisierten Devianz Residuen als auch für die Standardisierten Pearson Residuen kaum Residuen vorhanden sind, die den Wert ± 1 übersteigen. Die Residuen streuen regelmäßig um die Regressionsgerade.

Abbildung 5.3: Standardisierte Pearson und Devianz Residuen des V3G1 Modells



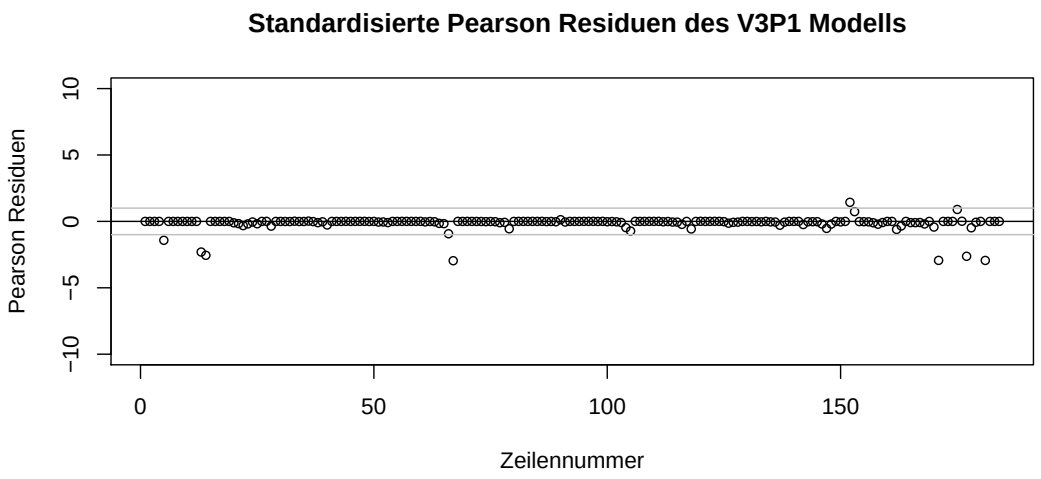


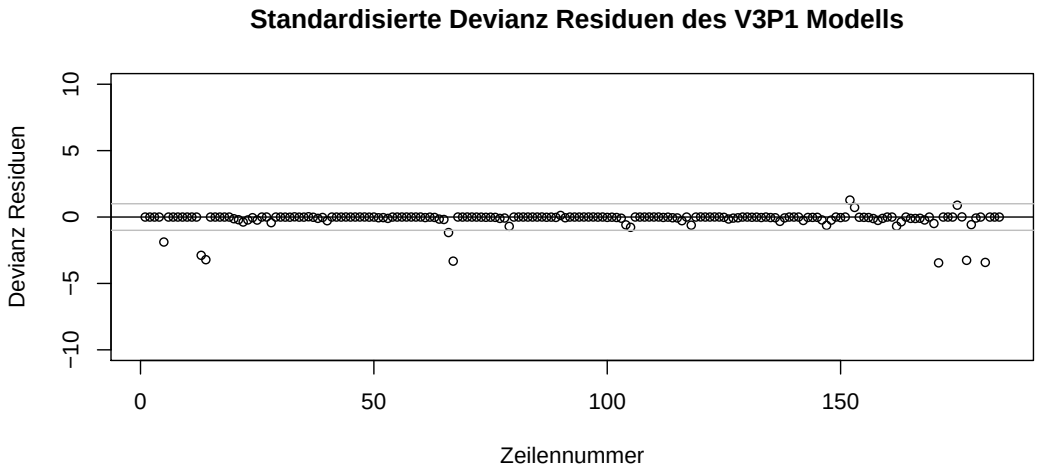
5.1.4 Standardisierte Pearson Residuen und Standardisierte Devianz Residuen des V3P1 Modells

Das letzte Modell, das betrachtet wird, ist das V3P1 Modell. Auch für dieses Modell wurden die Merkmale A und B an den V Datensatz angepasst. In diesem Fall wurde die Poisson-Verteilung verwendet.

Eine Illustration der Standardisierten Pearson Residuen und der Standardisierten Devianz Residuen findet man in Abbildung 5.4. Die Residuen sind gegen die entsprechende Zeilennummer des V Datensatzes angetragen.

Abbildung 5.4: Standardisierte Pearson und Devianz Residuen des V3P1 Modells





Die Standardisierten Pearson Residuen und die Standardisierten Devianz Residuen werden wie in Anschnitt 2.3.6 beschrieben berechnet.

Betrachtet man die Residuenplots, dann kann man erkennen, dass auch dieses Modell mehr Residuen als das vorherige hat, die einen größeren Wert als ± 1 haben. Auch hier ist zu sehen, dass die Beobachtungen, die große Standardisierte Pearson Residuen haben, auch große Standardisierte Devianz Residuen aufweisen.

Somit scheint insgesamt das Modell V3G1 besser an die Daten angepasst zu sein als das Modell V3P1.

5.2 Illustration der Schätzer und der Fitted Values

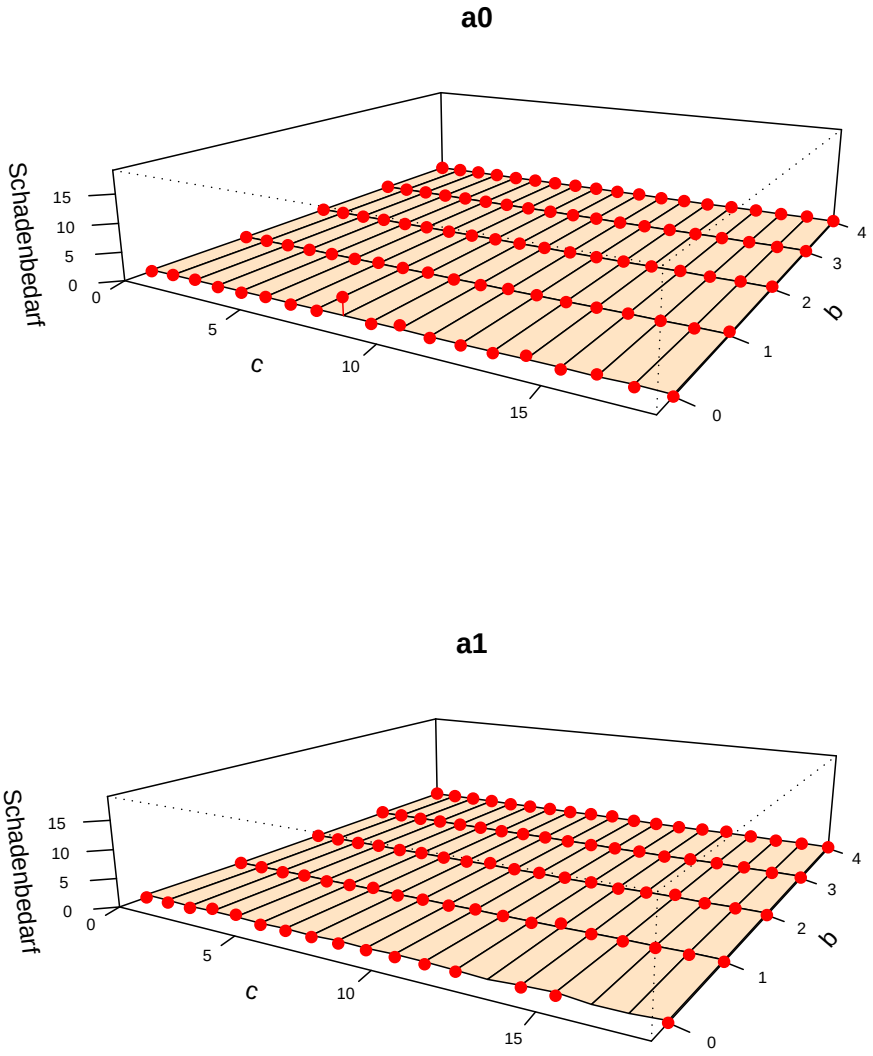
Dieser Abschnitt enthält die Perspektiven Plots der vier Modelle S7G1, S7P1, V3G1 und V3P1, siehe Tabelle 4.2 und Tabelle 4.3. Dieselben Modelle wurden bereits im vorherigen Abschnitt im Bezug auf ihre Residuen untersucht.

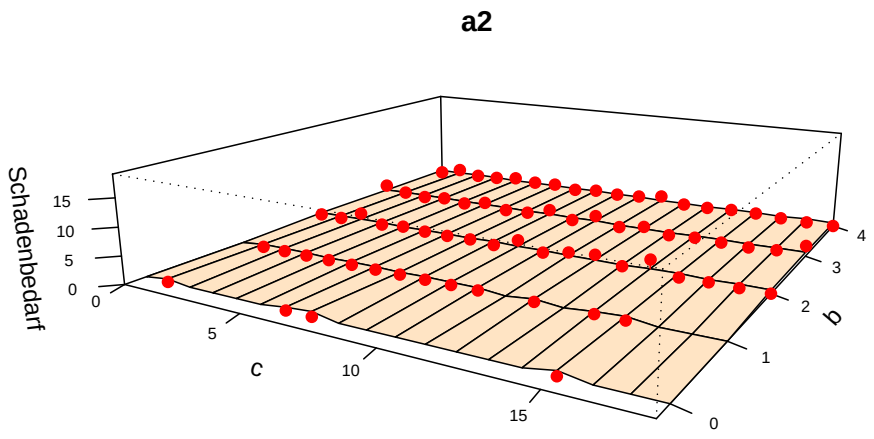
Die Perspektiven Plots enthalten die Fitted Values, die als Datenpunkte dargestellt sind. Die geschätzten Regressionsparameter werden als Ebene in den Plots illustriert.

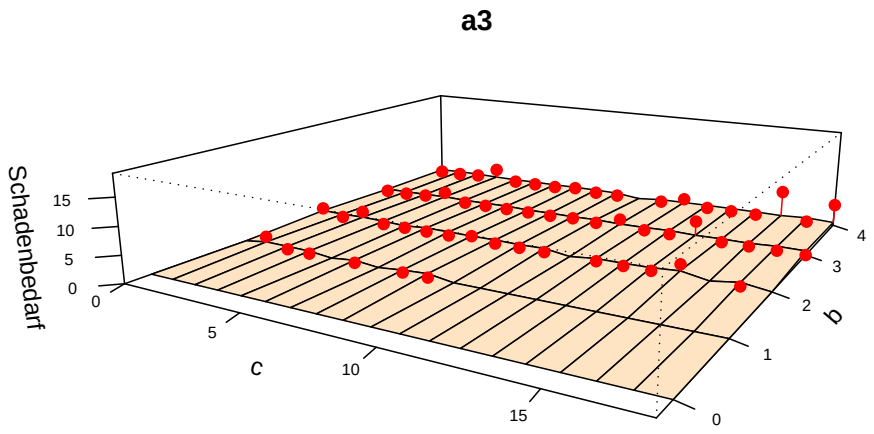
5.2.1 Illustration des gemeinsamen Einflusses von c und b bei unterschiedlicher Ausprägung des Merkmals a des S7G1 Modells

Abbildung 5.5 illustriert den gemeinsamen Einfluss von c und b bei unterschiedlicher Ausprägung des Merkmals a des S7G1 Modells. Dieses Modell ist Gamma-verteilt und wurde mit den Merkmalen a, b und c an den S Datensatz angepasst.

Abbildung 5.5: Illustration des gemeinsamen Einflusses von c und b bei unterschiedlicher Ausprägung des Merkmals a des S7G1 Modells





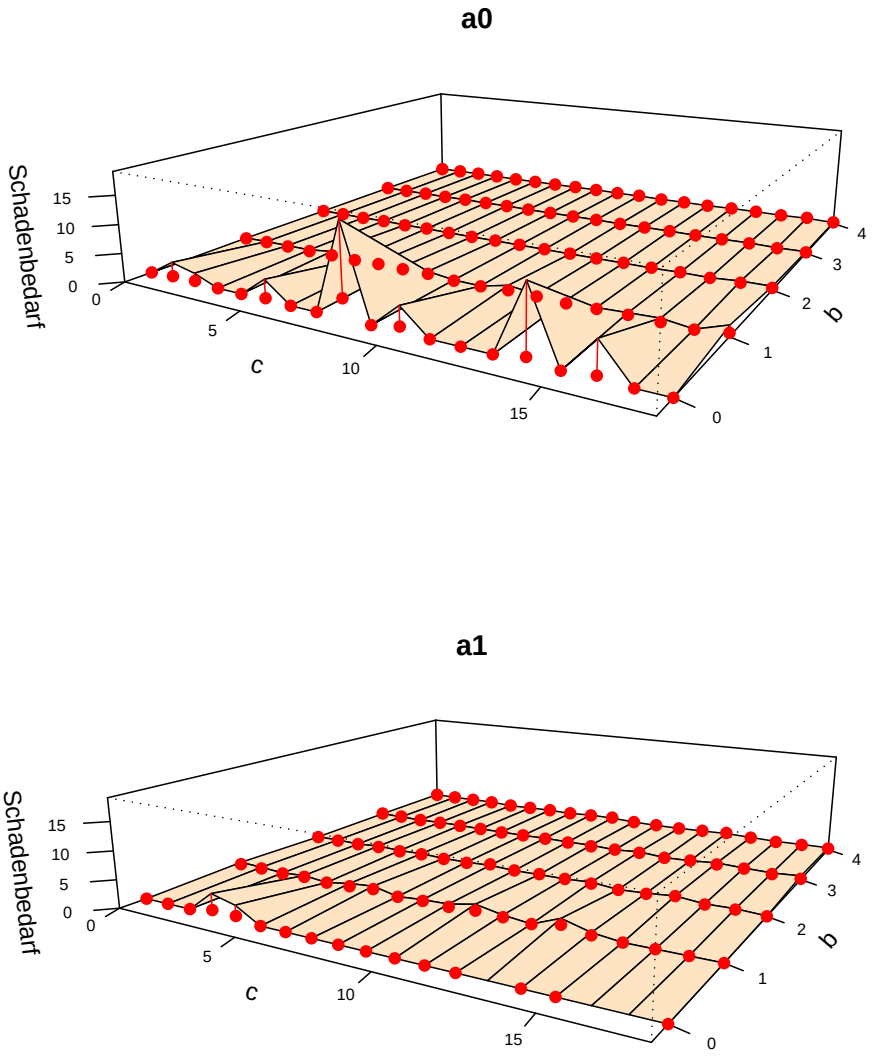


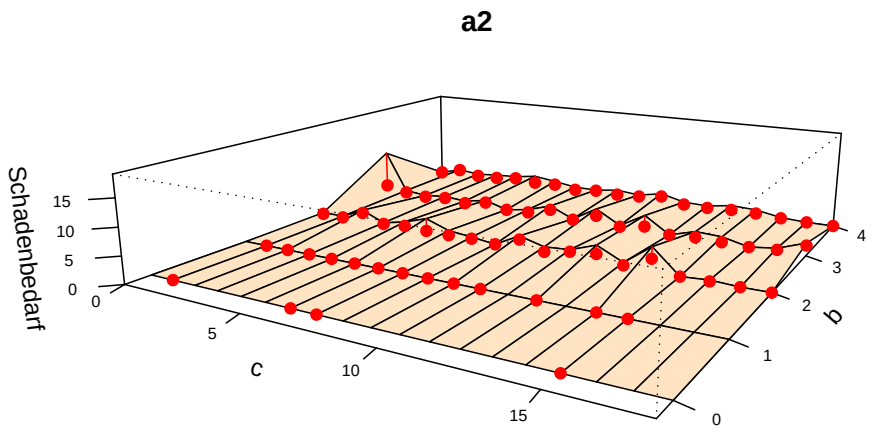
Der Perspektiven Plot ist in die verschiedenen Ausprägungen des Merkmals a unterteilt. Man erkennt erneut, dass für die Ausprägungen a2 und a3 nicht alle Beobachtungen vorhanden sind.

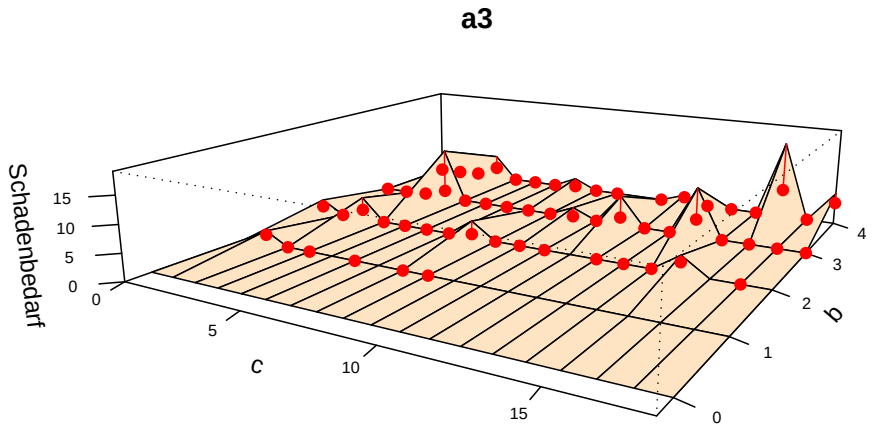
5.2.2 Illustration des gemeinsamen Einflusses von c und b bei unterschiedlicher Ausprägung des Merkmals a des S7P1 Modells

Als nächstes wird der gemeinsam Einfluss von c und b bei unterschiedlicher Ausprägung des Merkmals a des S7P1 Modells in Abbildung 5.6 illustriert. Das Modell S7P1 wurde mit allen Variablen und der Poisson-Verteilung an den S Datensatz angepasst.

Abbildung 5.6: Illustration des gemeinsamen Einflusses von c und b bei unterschiedlicher Ausprägung des Merkmals a des S7P1 Modells







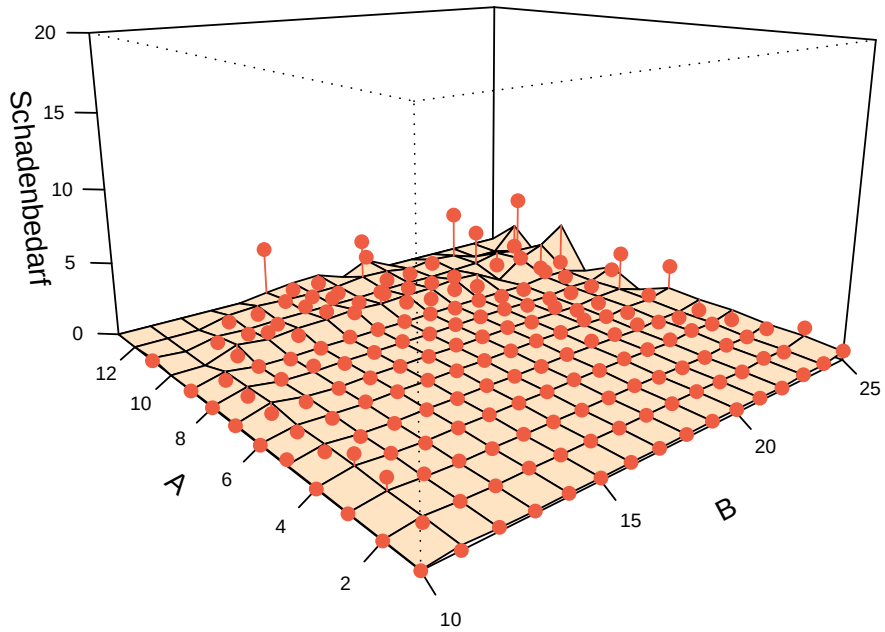
Im Vergleich zum vorherigen Modell ist zu erkennen, dass die geschätzten Regressionsparameter in fast allen Fällen über den Fitted Values liegen.

Dies deutet darauf hin, dass die Gamma-Verteilung die Daten besser anpasst als die Poisson-Verteilung und somit das Modell S7G1 besser an die Daten angepasst ist.

5.2.3 Illustration des gemeinsamen Einflusses von A und B des V3G1 Modells

Die Illustration des gemeinsamen Einflusses von A und B in Abbildung 5.7 betrachtet das volle Gamma-verteilte Modell des V Datensatzes.

Abbildung 5.7: Illustration des gemeinsamen Einflusses von A und B des V3G1 Modells

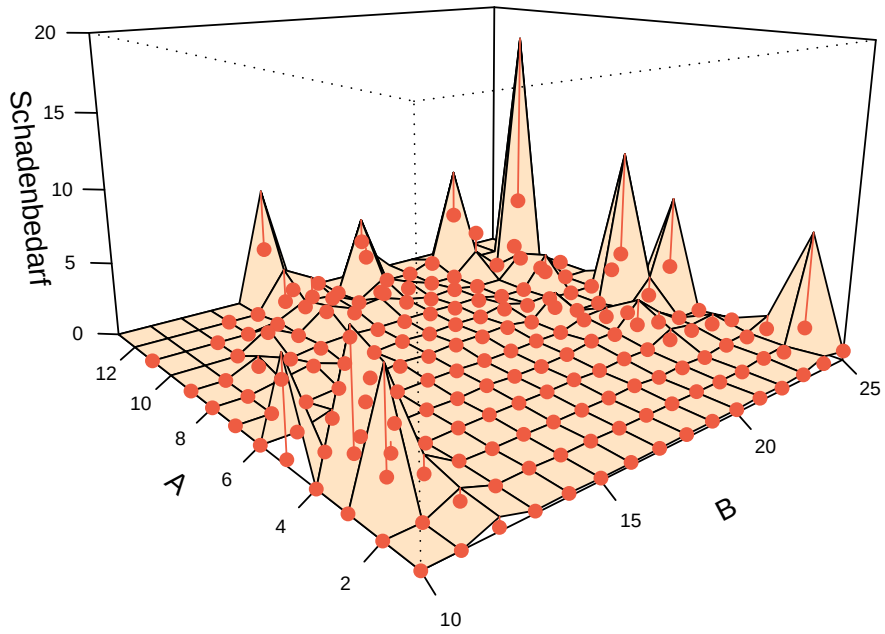


Die Regressionsebene erscheint flach. Nur im Bereich der Ausprägungen über A10 sind Erhebungen zu erkennen. Die Fitted Values liegen sowohl oberhalb als auch unterhalb der Regressionsebene.

5.2.4 Illustration des gemeinsamen Einflusses von A und B des V3P1 Modells

Die letzte Illustration, Abbildung 5.8, stellt den gemeinsamen Einfluss von A und B für volle Poisson-verteilte Modelle des V Datensatzes dar.

Abbildung 5.8: Illustration des gemeinsamen Einflusses von A und B des V3P1 Modells



Die Regressionsebene hat mehr Erhebungen als die des vorherigen Modells. Hier scheinen auch fast alle Regressionsparameter über den Fitted Values zu liegen. Somit kann vermutet werden, dass das Gamma-verteilte Modell besser an die Daten angepasst ist.

6 Nicht genestete und genestete Modellvergleiche

In den nächsten Abschnitten wird Schritt für Schritt untersucht, welches der Modelle am besten an die Daten angepasst ist. Um die Modelle miteinander zu vergleichen, werden der Vuong Test und der Distribution-Free Test verwendet. Es werden die Modelle S6G1, S6P1, S7G1, S7P1, V3G1 und V3P1 betrachtet, siehe Tabelle 4.2 und Tabelle 4.3.

Die Schätzung der Regressionsparameter dieser Modelle erfolgt mit dem S und dem V Datensatz. Um die Teststatistiken des Vuong Tests und des Distribution-Free Tests zu berechnen ist es jedoch notwendig auf die Einzeldatenebene, den KH Datensatz, zurückzukehren.

6.1 Wahl der Verteilung der Zielvariable Schadenbedarf

Als erstes wird die Verteilung der Zielvariable Schadenbedarf mithilft des Vuong Tests und des Distribution-Free Tests untersucht.

Betrachtet man die Gamma-Regression, dann gilt für den Schadenbedarf SB_i auf der aggregierten Datenebene:

$$SB_i \sim \Gamma\left(\frac{\mu_i}{B_i}, \nu J_i\right)$$

Kehrt man zur Einzeldatenebene zurück dann gilt für SB_{ij}:

$$SB_{ij} \sim \Gamma\left(\frac{\mu_i u_{ij}}{B_i J_i}, \nu J_i\right) = \Gamma\left(\mu_i^E, \nu J_i\right)$$

für $\mu_i^E = \frac{\mu_i u_{ij}}{B_i J_i}$, siehe Abschnitt 4.2.

Für die Gamma-verteilten Modelle wird somit folgende Dichte verwendet:

$$F_{\boldsymbol{\mu}^E,\nu} = f\left(SB_{ij}|X_{ij};\boldsymbol{\mu}^E,\nu\right) = \frac{\left[\frac{\nu^\nu}{\Gamma(\nu)\ \mu_i^{E\ \nu}}\ SB_{ij}{}^{(\nu-1)}\exp(-\frac{\nu}{\mu_i^E}\ SB_{ij})\right]^{u_{ij}}}{c_{ij}},$$

wobei für c_{ij} gilt, dass $\int_0^\infty f(SB_{ij})^{u_{ij}} = c_{ij}$ für $i = 1, ..., m;\ j = 1, ..., k_i$ und f die Dichte der Gamma-Verteilung ist.

Der Schadenbedarf SB_i der Poisson-Regression auf den aggregierten Datensätzen hat den Erwartungswert $E(SB_i) = \frac{D_i}{B_iJ_i}\lambda_i$ und die Varianz $Var(SB_i) = \left(\frac{D_i}{B_iJ_i}\right)^2\lambda_i$ mit $\lambda_i^E = \frac{D_i}{B_iJ_i}\lambda_i$. Auf dem KH Datensatz hat der Schadenbedarf SB_{ij} den Erwartungswert $E(SB_{ij}) = \frac{D_{ij}}{B_iJ_i}\lambda_i$ und die Varianz $Var(SB_{ij}) = \left(\frac{D_{ij}}{B_iJ_i}\right)^2\lambda_i$, siehe Abschnitt 4.3.

Die Wahrscheinlichkeitsfunktion der Poisson-verteilten Modelle ist gegeben durch:

$$F_{\boldsymbol{\lambda}^E} = g(SB_{ij}|X_{ij};\boldsymbol{\lambda}^E) = \frac{\left[\exp\left(\lambda_i^E\right)\ \frac{\lambda_i^{E\ SB_{ij}}}{SB_{ij}!}\right]^{\frac{u_{ij}B_i}{D_i}}}{d_{ij}},$$

wobei für d_{ij} gilt, dass $\sum_{i=1}^\infty g(SB_{ij})^{u_{ij}} = d_{ij}$ für $i = 1, ..., m;\ j = 1, ..., k_i$ und g die Wahrscheinlichkeitsfunktion der Poisson-Verteilung ist.

Zuerst wird die Verteilung ausgewählt, die das Modell besser an die Daten anpasst. Dazu werden der Vuong Test und der Distribution-Free Test verwendet. Die Modelle, die verglichen werden, haben dieselben Variablen aber unterschiedliche Verteilungen.

6.1.1 Entwicklung der Vuong Teststatistik für die Wahl zwischen Poisson- und Gamma-Verteilung der Zielvariable

Zunächst wird betrachtet, wie man mit dem Vuong Test, Abschnitt 2.4.1, zwischen der Gamma-Regression und der Poisson-Regression auswählt. Die Teststatistik des Vuong Tests für die Wahl einer Verteilung sieht in diesem Fall wie folgt aus:

$$\boldsymbol{V} = \frac{\boldsymbol{LR}_k(\hat{\boldsymbol{\mu}}_k, \hat{\nu}_k, \hat{\boldsymbol{\lambda}}_k)}{\sqrt{k}\hat{w}_k}$$

mit

$$
\begin{aligned}
& \boldsymbol{LR}_k\left(\hat{\boldsymbol{\mu}}_k^E, \hat{\nu}_k, \hat{\boldsymbol{\lambda}}_k^E\right) = L_k^f\left(\hat{\boldsymbol{\mu}}_k^E, \hat{\nu}_k\right) - L_k^g\left(\hat{\boldsymbol{\lambda}}_k^E\right) \\
= & \sum_{i=1}^{m}\sum_{j=1}^{k_i} \log f(SB_{ij}|X_{ij}; \hat{\boldsymbol{\mu}}^E, \hat{\nu}) - \sum_{i=1}^{m}\sum_{j=1}^{k_i} \log g(SB_{ij}|Z_{ij}; \hat{\boldsymbol{\lambda}}^E) \\
= & \sum_{i=1}^{m}\sum_{j=1}^{k_i} u_{ij}\left[\hat{\nu}\log\frac{\hat{\nu}}{\hat{\mu}_i^E} - \log\Gamma(\hat{\nu}) + (\hat{\nu}-1)\log SB_{ij} - \frac{\hat{\nu}}{\hat{\mu}_i^E}SB_{ij}\right] - \log(c_{ij}) \\
& -\frac{u_{ij}B_i}{D_i}\left[\hat{\lambda}_i^E - SB_{ij}\log(\hat{\lambda}_i^E) + \log(SB_{ij}!)\right] + \log(d_{ij})
\end{aligned}
$$

und

$$
\hat{w}_k^2 = \frac{1}{k}\sum_{i=1}^{m}\sum_{j=1}^{k_i}\left[\log\left(\frac{f(SB_{ij}|X_{ij}; \hat{\boldsymbol{\mu}}^E, \hat{\nu})}{g(Y_{ij}|Z_{ij}; \hat{\boldsymbol{\lambda}}^E)}\right)\right]^2 - \left[\frac{1}{k}\sum_{i=1}^{m}\sum_{j=1}^{k_i}\log\left(\frac{f(SB_{ij}|X_{ij}; \hat{\boldsymbol{\mu}}^E, \hat{\nu})}{g(Y_{ij}|Z_{ij}; \hat{\boldsymbol{\lambda}}^E)}\right)\right]^2,
$$

wobei $k = \sum_{i=1}^{m} k_i$ mit

$$
\begin{aligned}
& \log\left(\frac{f(SB_{ij}|X_{ij}; \hat{\boldsymbol{\mu}}^E, \hat{\nu})}{g(SB_{ij}|Z_{ij}; \hat{\boldsymbol{\lambda}}^E)}\right) \\
= & \; u_{ij}\left[\hat{\nu}\log\left(\hat{\nu}\frac{SB_{ij}}{\hat{\mu}_i^E}\right) + \log\left(\frac{SB_{ij}}{\Gamma(\hat{\nu})}\right) - \hat{\nu}\frac{SB_{ij}}{\hat{\mu}_i^E}\right] - \log(c_{ij}) \\
& -\frac{u_{ij}B_i}{D_i}\left[\hat{\lambda}_i^E - SB_{ij}\log(\hat{\lambda}_i^E) + \log(SB_{ij}!)\right] + \log(d_{ij}).
\end{aligned}
$$

Es ist kein Korrekturfaktor notwendig, da die beiden Modelle, die gegenübergestellt werden, die gleichen Variablen und somit auch dieselbe Anzahl an Regressionsparametern haben. Man erhält folgende Hypothese und Alternative:

$$
\boldsymbol{H} : \text{Gamma GLM} \quad \textit{versus} \quad \boldsymbol{K} : \text{Poisson GLM}
$$

Verwerfe $\boldsymbol{H}$ zum Niveau α, falls $\boldsymbol{V} < z_{\frac{\alpha}{2}}$. Für $z_{\frac{\alpha}{2}} < \boldsymbol{V} < z_{1-\frac{\alpha}{2}}$ trifft der Vuong Test keine Entscheidung, $\boldsymbol{H}$ wird weder verworfen noch nicht verworfen.

6.1.2 Entwicklung der Distribution-Free Teststatistik für die Wahl zwischen Poisson- und Gamma-Verteilung der Zielvariable

Der Distribution-Free Test aus Abschnitt 2.4.2 kann ebenfalls verwendet werden um zu entscheiden, ob die Gamma-Verteilung oder die Poisson-Verteilung die Modelle besser an die Daten anpasst.

Die Teststatistik des Distribution-Free Test sieht für die Wahl zwischen zwei Verteilungen wie folgt aus:

$$\boldsymbol{B} = \sum_{i=1}^{m} \sum_{j=1}^{k_i} \boldsymbol{I}_{(0,\infty)}(t_{ij})$$

mit

$$\begin{aligned} t_{ij} &= \log\left(\frac{f(SB_{ij}|X_{ij};\hat{\boldsymbol{\mu}}_i^E,\hat{\nu})}{g(SB_{ij}|Z_{ij};\hat{\boldsymbol{\lambda}}_i^E)}\right) \\ &= u_{ij}\left[\hat{\nu}\log\left(\hat{\nu}\frac{SB_{ij}}{\hat{\mu}_i^E}\right) + \log\left(\frac{SB_{ij}}{\Gamma(\hat{\nu})}\right) - \hat{\nu}\frac{SB_{ij}}{\hat{\mu}_i^E}\right] - \log(c_{ij}) \\ &\quad -\frac{u_{ij}B_i}{D_i}\left[\hat{\lambda}_i^E - SB_{ij}\log(\hat{\lambda}_i^E) + \log(SB_{ij}!)\right] + \log(d_{ij}) \end{aligned}$$

für $i = 1, ..., m;\ j = 1, ..., k_i$.

Man erhält die folgende Hypothese und die Alternative:

$$\boldsymbol{H} : \text{Gamma GLM} \quad versus \quad \boldsymbol{K} : \text{Poisson GLM}$$

Verwerfe $\boldsymbol{H}$, falls $\boldsymbol{B} < \frac{k}{2}$ mit $k = \sum_{i=1}^{m} k_i$.

6.1.3 Berechnung der Teststatistiken des Vuong Tests und des Distribution-Free Tests für den KH Datensatz zur Verteilungswahl

Es werden nun die Werte der Teststatistiken des Vuong Tests und des Distribution-Free Tests berechnet. Es werden die Modelle S6 und S7 des S Datensatzes und die V3 Modelle des V Datensatzes verglichen.

In Tabelle 6.1 sind die Ergebnisse des Vuong Tests und des Distribution-Free Tests für die Wahl einer geeigneten Verteilung zusammengefasst.

Tabelle 6.1: Werte der Teststatistiken des Vuong Tests und des Distribution-Free Tests für die Verteilungswahl

Modelle	**Testent-scheidung**	**Distribution-Free Test**	$\frac{k}{2}$	**Vuong Test**	95 % Quantil der Standard-normalverteilung
$\boldsymbol{H}$: S6G1 vs $\boldsymbol{K}$: S6P1	accept $\boldsymbol{H}$	234904	117453	1.00	1.64
$\boldsymbol{H}$: S7G1 vs $\boldsymbol{K}$: S7P1	accept $\boldsymbol{H}$	234902	117453	1.00	1.64
$\boldsymbol{H}$: V3G1 vs $\boldsymbol{K}$: V3P1	accept $\boldsymbol{H}$	234098	117453	1.00	1.64

Der Wert $\frac{k}{2}$ aus der obigen Tabelle ist der Wert der Binomialverteilung mit den Parametern k und 0,5 (Bin(k; 0,5)). Der Distribution-Free Test verwirft $\boldsymbol{H}$ nicht genau dann, wenn der Wert der Teststatistik größer ist als $\frac{k}{2}$. Somit entscheidet der Distribution-Free Test, dass in allen drei Fällen das Gamma-verteilte Modell besser an die Daten angepasst ist, siehe Tabelle 6.1. Dies bestätigt die Vermutung aus dem Abschnitt 5.1.

Die Hypothese $\boldsymbol{H}$ des Vuong Tests wird verworfen für Werte kleiner als -1,64, für Werte größer als 1,64 wird $\boldsymbol{H}$ nicht verworfen. Liegt der Wert der Teststatistik dazwischen, dann trifft der Vuong Test keine Entscheidung. Betrachtet man die Teststatistiken des Vuong Tests in der obigen Tabelle, dann entscheidet dieser in keinem der drei Fälle, ob die Gamma-Verteilung oder die Poisson-Verteilung besser ist.

6.2 Modellwahl für genestete Modelle des S Datensatzes

Hier werden nun zwei genestete Gamma-verteilte Modelle mit unterschiedlicher Anzahl an geschätzten Parametern betrachtet. Es wird mit dem Vuong Test und dem Distribution-Free Test untersucht, welches der Modelle besser an den S Datensatz angepasst ist. In Abschnitt 6.1 ist die Dichte der Modelle beschrieben. Man arbeitet erneut auf der unaggregierten Datenebene, dem KH Datensatz.

Da die Modelle, die verglichen werden, eine unterschiedliche Anzahl an geschätzten Parametern haben, muss ein Korrekturfaktor verwendet werden.

Die Teststatistik $\boldsymbol{V}$ des Vuong Tests ist dieselbe wie die Teststatistik (6.1) im folgenden Abschnitt 6.3. Dasselbe gilt für die Teststatistik $\boldsymbol{B}$ des Distribution-Free Tests. Auch die Teststatistik $\boldsymbol{B}$ stimmt mit der Teststatistik (6.2) im folgenden Abschnitt überein.

Tabelle 6.2 enthält die Ergebnisse der Hypothesentests für die Gamma-verteilten genesteten Modelle. Der Vollständigkeit wegen sind auch die Werte der Teststatistik für den Vergleich der zwei genesteten Poisson-verteilten Modelle angegeben.

Tabelle 6.2: Werte der Teststatistiken des Vuong Tests und des Distribution-Free Tests für genestete Modelle

Modelle	**Testent-scheidung**	**Distribution-Free Test**	$\frac{k}{2}$	**Vuong Test**	95 % Quantil der Standard-normalverteilung
$\boldsymbol{H}$: S7P1 vs $\boldsymbol{K}$: S6P1	reject $\boldsymbol{H}$	26893	117453	-1.00	1.64
$\boldsymbol{H}$: S7G1 vs $\boldsymbol{K}$: S6G1	reject $\boldsymbol{H}$	22783	117453	-1.00	1.64

Der Distribution-Free Test verwirft die Hypothese $\boldsymbol{H}$, da der Wert der Teststatistik kleiner als $\frac{k}{2}$ ist. Somit entscheidet dieser in beiden Fällen, dass das um das Merkmal b reduzierte Modell S6G1 bzw. S6P1 besser an die Daten angepasst ist. Dies bestätigt die Vermutung aus Abschnitt 4.6. Dort wurde zwar das volle Modell S7G1 bzw. S7P1 als das bessere Modell ausgewählt, wobei jedoch der Unterschied zwischen den beiden Modellen gering war.

Der Vuong Test trifft erneut keine Entscheidung, erst für Werte der Teststatistik unter -1,64 oder über 1,64 würde der Vuong Test $\boldsymbol{H}$ verwerfen oder nicht verwerfen.

6.3 Nicht genestete Modellwahl zwischen den S Modellen und den V Modellen

Im nächsten Schritt werden nun Modelle des S Datensatzes mit Modellen des V Datensatzes verglichen. Dies bedeutet, man untersucht nicht genestete Modelle. Auch hier werden der Vuong Test und des Distribution-Free Test verwendet.

Wie in den beiden vorhergehenden Abschnitten arbeitet man auf dem unaggregierten KH Datensatz. Die Dichte der Modelle findet man in Anschnitt 6.1.

6.3.1 Entwicklung der Vuong Teststatistik für die Wahl nicht genesteter Gamma-verteilter Modelle

Die Teststatistik zum Vergleich nicht genesteter Gamma-verteilter Modelle hat die Form:

$$\boldsymbol{V} = \frac{\boldsymbol{LR}_k(\hat{\boldsymbol{\mu}}_k^E, \hat{\nu}_k, \hat{\boldsymbol{\alpha}}_k^E, \hat{\beta}_k)}{\sqrt{k}\hat{w}_k}$$

mit $k = \sum_{i=1}^m k_i$ und

$$\begin{aligned}
& \boldsymbol{LR}_k(\hat{\boldsymbol{\mu}}_k^E, \hat{\nu}_k, \hat{\boldsymbol{\alpha}}_k^E, \hat{\beta}_k) = L_k^f(\hat{\boldsymbol{\mu}}_k^E, \hat{\nu}_k) - L_k^g(\hat{\boldsymbol{\alpha}}_k^E, \hat{\beta}_k) \\
= & \sum_{i=1}^m \sum_{j=1}^{k_i} \log f(SB_{ij}|X_{ij}; \hat{\boldsymbol{\mu}}_i^E, \hat{\nu}) - \sum_{i=1}^m \sum_{j=1}^{k_i} \log g(SB_{ij}|Z_{ij}; \hat{\boldsymbol{\alpha}}_i^E, \hat{\beta}) \\
= & \sum_{i=1}^m \sum_{j=1}^{k_i} u_{ij} \left[\hat{\nu}\log(\hat{\nu}) - \hat{\nu}\log(\hat{\mu}_i^E) - \log(\Gamma(\hat{\nu})) + (\hat{\nu}-1)\log(SB_{ij}) - \frac{\hat{\nu}}{\hat{\mu}_i^E} SB_{ij} \right] - \log(c_{ij}) \\
- & u_{ij} \left[\hat{\beta}\log(\hat{\beta}) + \hat{\beta}\log(\hat{\alpha}_i^E) + \log(\Gamma(\hat{\beta})) - (\hat{\beta}-1)\log(SB_{ij}) + \frac{\hat{\beta}}{\hat{\alpha}_i^E} SB_{ij} \right] + \log(d_{ij})
\end{aligned}$$

und

$$\hat{w}_k^2 = \frac{1}{k} \sum_{i=1}^m \sum_{j=1}^{k_i} \left[\log \left(\frac{f(SB_{ij}|X_{ij}; \hat{\boldsymbol{\mu}}^E, \hat{\nu})}{g(SB_{ij}|Z_{ij}; \hat{\boldsymbol{\alpha}}^E, \hat{\beta})} \right) \right]^2 - \left[\frac{1}{k} \sum_{i=1}^m \sum_{j=1}^{k_i} \log \left(\frac{f(SB_{ij}|X_{ij}; \hat{\boldsymbol{\mu}}^E, \hat{\nu})}{g(SB_{ij}|Z_{ij}; \hat{\boldsymbol{\alpha}}^E, \hat{\beta})} \right) \right]^2$$

wobei $\sum_{i=1}^m k_i = k$ mit

$$\begin{aligned}
& \log \left(\frac{f(SB_{ij}|X_{ij}; \hat{\boldsymbol{\mu}}^E, \hat{\nu})}{g(SB_{ij}|Z_{ij}; \hat{\boldsymbol{\alpha}}^E, \hat{\beta})} \right) \\
= & u_{ij} \left[\hat{\nu}\log(\hat{\nu}) - \hat{\nu}\log(\hat{\mu}_i^E) - \log(\Gamma(\hat{\nu})) + (\hat{\nu}-1)\log(SB_{ij}) - \frac{\hat{\nu}}{\hat{\mu}_i^E} SB_{ij} \right] - \log(c_{ij}) \\
- & u_{ij} \left[\hat{\beta}\log(\hat{\beta}) + \hat{\beta}\log(\hat{\alpha}_i^E) + \log(\Gamma(\hat{\beta})) - (\hat{\beta}-1)\log(SB_{ij}) + \frac{\hat{\beta}}{\hat{\alpha}_i^E} SB_{ij} \right] + \log(d_{ij}).
\end{aligned}$$

Da in diesem Fall die beiden Modelle eine unterschiedliche Anzahl an geschätzten Parameter haben ist ein Korrekturfaktor notwendig. Schließlich erhält man folgende Teststatistik:

$$\boldsymbol{V} = \frac{\widehat{\boldsymbol{LR}}_k(\hat{\boldsymbol{\mu}}_k^E, \hat{\nu}_k, \hat{\boldsymbol{\alpha}}_k^E, \hat{\beta}_k)}{\sqrt{k}\hat{w}_k} = \frac{\boldsymbol{LR}_k(\hat{\boldsymbol{\mu}}_k^E, \hat{\nu}_k, \hat{\boldsymbol{\alpha}}_k^E, \hat{\beta}_k) - K(p,q)}{\sqrt{k}\hat{w}_k} \tag{6.1}$$

mit

$$K(p,q) = \frac{p}{2}\log(k) - \frac{q}{2}\log(k),$$

wobei p die Anzahl der geschätzten Parameter des Modells f und q die Anzahl der geschätzten Parameter des Modells g ist.

Für Hypothese und Alternative ergibt sich:

$$\boldsymbol{H} : \text{Modelle des S Datensatzes} \quad versus \quad \boldsymbol{K} : \text{Modelle des V Datensatzes}$$

Verwerfe $\boldsymbol{H}$ zum Niveau α, falls $\boldsymbol{V} < z_{\frac{\alpha}{2}}$. Für $z_{\frac{\alpha}{2}} < \boldsymbol{V} < z_{1-\frac{\alpha}{2}}$ trifft der Vuong Test keine Entscheidung, $\boldsymbol{H}$ wird weder verworfen noch nicht verworfen.

Um die Teststatistiken für die Poisson-Verteilung zu erhalten wird auf dieselbe Weise vorgegangen, die Teststatistiken findet man im Anhang.

6.3.2 Entwicklung der Distribution-Free Teststatistik für die Wahl nicht genesteter Gamma-verteilter Modelle

Hier werden nicht genestete Modelle mit dem Distribution-Free Test verglichen. Die Teststatistik für die Gamma-verteilten Modelle sieht wie folgt aus:

$$\boldsymbol{B} = \sum_{i=1}^{m}\sum_{j=1}^{k_i} \boldsymbol{I}_{(0,+\infty)}(t_{ij})$$

mit

$$
\begin{aligned}
t_{ij} &= log\left(f\left(SB_{ij}|X_{ij};\hat{\boldsymbol{\mu}}^E,\hat{\nu}\right)\right) - log\left(g\left(SB_{ij}|Z_{ij};\hat{\boldsymbol{\alpha}}^E,\hat{\beta}\right)\right) \\
&= u_{ij}\left[\hat{\nu}\log(\hat{\nu}) - \hat{\nu}\log(\hat{\mu}_i^E) - \log(\Gamma(\hat{\nu})) + (\hat{\nu}-1)\log(SB_{ij}) - \frac{\hat{\nu}}{\hat{\mu}_i^E}SB_{ij}\right] - \log(c_{ij}) \\
&- u_{ij}\left[\hat{\beta}\log(\hat{\beta}) + \hat{\beta}\log(\hat{\alpha}_i^E) + \log(\Gamma(\hat{\beta})) - (\hat{\beta}-1)\log(SB_{ij}) + \frac{\hat{\beta}}{\hat{\alpha}_i^E}SB_{ij}\right] + \log(d_{ij})
\end{aligned}
$$

für $i = 1, ..., m;\ j = 1, ..., k_i$. Auch hier ist durch die unterschiedliche Anzahl der geschätzten Parameter in den Modellen ein Korrekturfaktor notwendig.

$$K(p,q) = \frac{p}{2k}\log(k) - \frac{q}{2k}\log(k)$$

Schließlich erhält man

$$\boldsymbol{B} = \sum_{i=1}^{m}\sum_{j=1}^{k_i}\boldsymbol{I}_{(0,+\infty)}\left(t_{ij}\right) \tag{6.2}$$

mit

$$t_{ij} = log\left(f\left(SB_{ij}|X_{ij};\hat{\boldsymbol{\mu}}^E,\hat{\nu}\right)\right) - log\left(g\left(SB_{ij}|Z_{ij};\hat{\boldsymbol{\alpha}}^E,\hat{\beta}\right)\right) - K(p,q)$$

für $i = 1, ..., m;\ j = 1, ..., k_i$, wobei p die Anzahl der geschätzten Parameter des Modells f und q die Anzahl der geschätzten Parameter des Modells g ist.

Die Hypothese und die Alternative sind gegeben durch:

$$\boldsymbol{H} : \text{Modelle des S Datensatzes} \quad versus \quad \boldsymbol{K} : \text{Modelle des V Datensatzes}$$

Verwerfe $\boldsymbol{H}$, falls $\boldsymbol{B} < \frac{k}{2}$ mit $k = \sum_{i=1}^{m} k_i$.

Auch hier wird um die Teststatistik für Poisson-verteilte Modelle zu erhalten auf dieselbe Weise vorgegangen. Diese Teststatistik findet man ebenfalls im Anhang.

6.3.3 Berechnung der Teststatistiken des Vuong Tests und des Distribution-Free Tests für den KH Datensatz zur Wahl nicht genesteter Modelle

Es werden die Werte der Teststatistiken des Vuong Tests und des Distribution-Free Tests berechnet. Das volle Modell des V Datensatzes, das volle Modell des S Datensatzes und das um das Merkmal b reduzierte Modell des S Datensatzes werden verglichen.

Die Werte der Teststatistiken sind in Tabelle 6.3 zusammengefasst. Der Vollständigkeit wegen sind auch die Werte der Teststatistiken bei Poisson-Verteilungsannahme angefügt.

Tabelle 6.3: Werte der Teststatistiken des Vuong Tests und des Distribution-Free Tests für nicht genestete Modelle

Modelle	**Testent-scheidung**	**Distribution-Free Test**	$\frac{k}{2}$	**Vuong Test**	95 % Quantil der Standard-normalverteilung
$\boldsymbol{H}$: S7G1 vs $\boldsymbol{K}$: V3G1	accept $\boldsymbol{H}$	180203	117453	1.00	1.64
$\boldsymbol{H}$: S6G1 vs $\boldsymbol{K}$: V3G1	accept $\boldsymbol{H}$	186788	117453	1.00	1.64
$\boldsymbol{H}$: S7P1 vs $\boldsymbol{K}$: V3P1	accept $\boldsymbol{H}$	194266	117453	1.00	1.64
$\boldsymbol{H}$: S6P1 vs $\boldsymbol{K}$: V3P1	accept $\boldsymbol{H}$	221646	117453	1.00	1.64

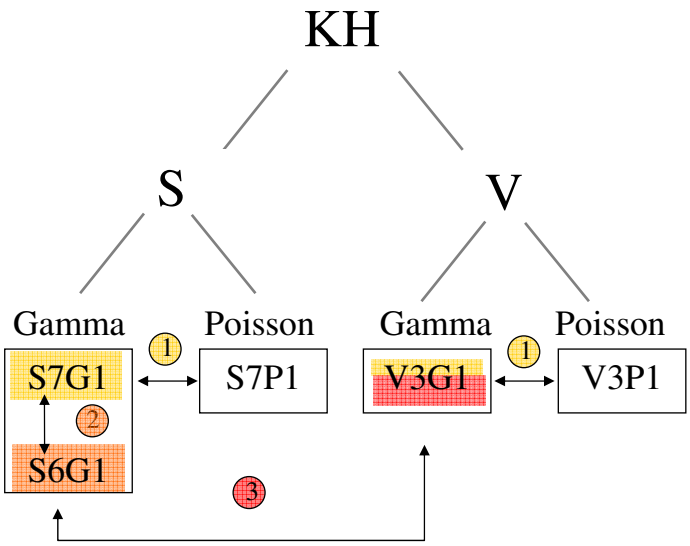
Wie man aus der obigen Tabelle erkennt, entscheidet sich der Distribution-Free Test in allen Fällen für die Modelle des S Datensatzes. Da für alle Modelle des S Datensatzes der Wert der Teststatistik größer als $\frac{k}{2}$ ist.

Der Vuong Test trifft auch für nicht genestete Modelle keine Entscheidung, welches Modell besser an die Daten angepasst ist. Er würde $\boldsymbol{H}$ verwerfen oder nicht verwerfen, wenn die Teststatistiken Werte unter -1,64 oder über 1,64 annehmen würden.

6.4 Zusammenfassung der Ergebnisse der Modellwahl mit dem Distribution-Free Test und dem Vuong Test

Die Ergebnisse der Modellwahl mit dem Distribution-Free Test und dem Vuong Test werden nun in Abbildung 6.1 zusammengefasst.

Abbildung 6.1: Zusammenfassung der Modellwahl mit Hilfe des Distribution-Free Tests und des Vuong Tests



Im 1. Schritt, siehe Abbildung 6.1, wird betrachtet, welche Verteilung die Daten besser anpasst. Wie man in der Abbildung sieht, entscheidet sich der Distribution-Free Test für die Gamma-Verteilung.

Im 2. Schritt, siehe Abbildung 6.1, werden genestete Modelle des S Datensatzes mit Hilfe des Distribution-Free Tests verglichen. Hier entscheidet der Distribution-Free Test, dass das um das Merkmal b reduzierte Modell besser an die Daten angepasst ist.

Schließlich werden im 3. Schritt, siehe Abbildung 6.1, nicht genestete Modelle verglichen. Hier wählt der Distribution-Free Test das Modell S6G1 als bestes Modell.

In Tabelle 6.4 sind die Werte der Teststatistiken des Distribution-Free Tests für ausgewählte Hypothesen und Alternativen des KH Datensatzes zusammengefasst.

Tabelle 6.4: Werte der Teststatistiken des Distribution-Free Tests für ausgewählte Hypothesen und Alternativen des KH Datensatzes

	K: S6G1	***K***: S6P1	***K***: S7G1	***K***: S7P1	***K***: V3G1	***K***: V3P1
H: S6G1		234904 accept ***H***			186788 accept ***H***	
H: S6P1						221646 accept ***H***
H: S7G1	22783 reject ***H***			234902 accept ***H***	180203 accept ***H***	
H: S7P1		26893 reject ***H***				194266 accept ***H***
H: V3G1						234089 accept ***H***

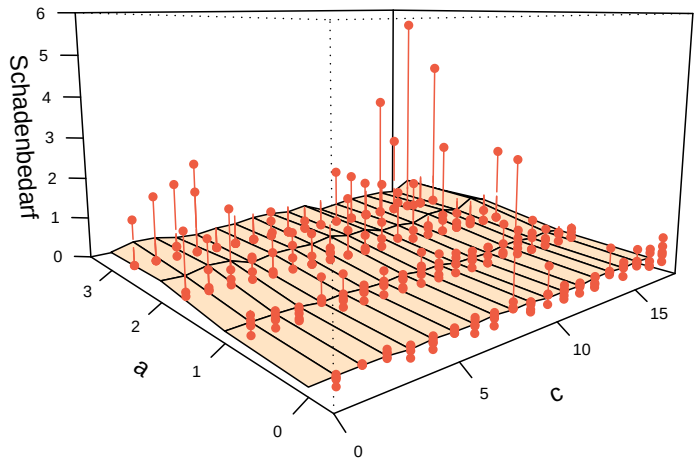
Betrachtet man die Werte der Teststatistiken des Vuong Tests für ausgewählte Hypothesen und Alternativen in Tabelle 6.5, dann erkennt man, dass der Vuong Test in keinem der Fälle eine Entscheidung trifft. Somit scheint der Vuong Test ungeeignet für diese Fragestellung.

Tabelle 6.5: Werte der Teststatistiken des Vuong Tests für ausgewählte Hypothesen und Alternativen des KH Datensatzes

	K: S6G1	***K***: S6P1	***K***: S7G1	***K***: S7P1	***K***: V3G1	***K***: V3P1
H: S6G1		1.00 no decision			1.00 no decision	
H: S6P1						1.00 no decision
H: S7G1	-1.00 no decision			1.00 no decision	1.00 no decision	
H: S7P1		-1.00 no decision				1.00 no decision
H: V3G1						1.00 no decision

Abbildung 6.2 illustriert den gemeinsamen Einfluss der Merkmale a und c des S6G1 Modells.

Abbildung 6.2: Illustration des gemeinsamen Einflusses der Schätzer der Merkmale a und c des S6G1 Modells



Das S6G1 Modell ist Gamma-verteilt und mit den Variablen a und c an den S Datensatz angepasst

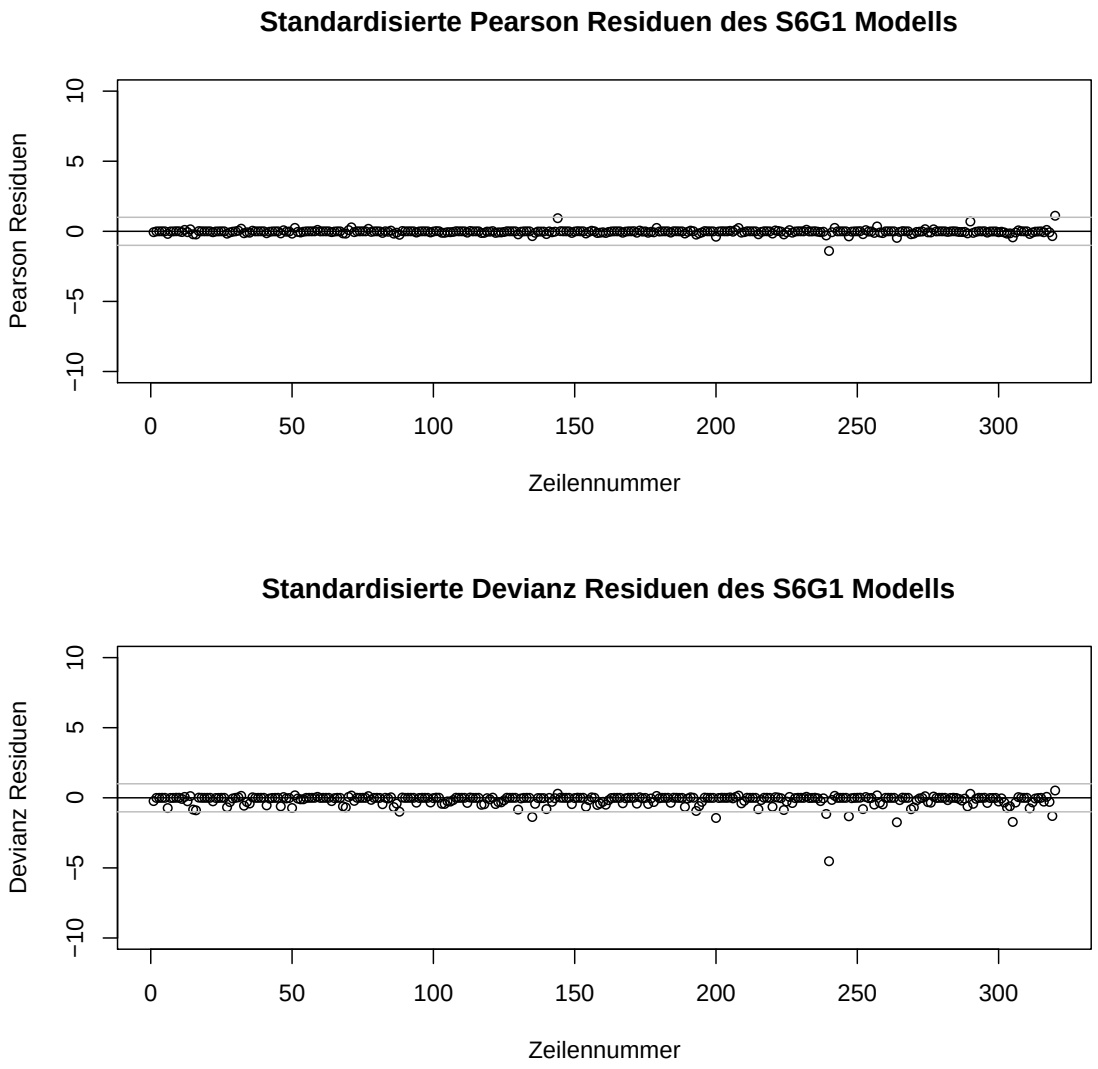
Die Regressionsebene erscheint im Allgemeinen flach, ab a2 steigt sie leicht an. Für die Klassen des Merkmals c ist eine Steigung bei niedrigen und sehr hohen Klassen zu erkennen.

Die Standardisierten Pearson Residuen und die Standardisierten Devianz Residuen des Modells S6G1 sind in Abbildung 6.3 dargestellt. Die dargestellten Residuen sind gegen die entsprechende Zeilennummer des S Datensatzes angetragen.

Die Berechnung der Standardisierten Pearson Residuen und der Standardisierten Devianz Residuen erfolgt wie in Abschnitt 2.3.7 angegeben.

Es ist zu erkennen, dass die Residuen kleine Werte annehmen. Sowohl die Standardisierten Pearson Residuen als auch die Standardisierten Devianz Residuen haben kaum Residuen, die einen Wert größer als ± 1 haben. Außerdem streuen die Residuen regelmäßig um die Regressionsgerade.

Abbildung 6.3: Standardisierte Pearson Residuen und Standardisierte Devianz Residuen des Modells S6G1



7 Simulation der Gütefunktion für den Distribution-Free Test und den Vuong Test

Im letzten Kapitel dieser Arbeit wird untersucht in wieweit man sich auf die Ergebnisse des Vuong Tests und des Distribution-Free Tests aus den Abschnitten 6.1, 6.2 und 6.3 verlassen kann.

Hierzu werden Datensätze simuliert und die Werte der Teststatistiken berechnet. Der Ablauf kann wie folgt beschrieben werden:

(i) Zwei Modelle werden ausgewählt. Es wird festgelegt, welches Modell als *Wahres Modell* und welches als *Alternative* betrachtet wird.

(ii) Es werden R Datensätze der Länge *Beob* simuliert, basierend auf den Schätzern des *Wahren Modells* und dessen Verteilung.

(iii) Die Schätzer für das *Wahre Modell* und die *Alternative* werden berechnet.

(iv) Berechnung der Werte der Teststatistiken für den Vuong Test und den Distribution-Free Test.

7.1 Simulation der Gütefunktion ohne Gewicht mit Wahrem Modell S6G1

Es werden zuerst Simulationen ohne Gewicht betrachtet. Simulationen ohne Gewicht bedeutet, dass nur die Datensätze aus dem KH 2007 Datensatz ausgewählt werden, bei denen das Merkmal Jahreseinheit den Wert eins hat.

Für die Simulation wird S6G1 als das Wahre Modell gewählt, die Alternative ist das Modell V3G1. Es wird nun das Wahre Modell, die Alternative und das Vorgehen für diese Simulation genau beschrieben.

Wie bereits erwähnt wird das Modell S6G1 als das Wahre Modell gewählt. Dieses Modell enthält neben dem Intercept auch die Variablen a und c. Sei nun $i \in I^H$ mit $I^H = \{i = (a, c)$, wobei die Kombination von a und c im betrachteten S Datensatz auftritt$\}$.

Für die Verteilung des Schadenbedarfs S_SB_i pro Zelle gilt in diesem Fall:

$$S\_SB_i \sim \Gamma(\frac{\mu_i^H}{B_i}, \nu^H J_i)$$

und für den Schadenbedarf auf Einzeldatenebene gilt somit:

$$SB_{ij} \sim \Gamma(\frac{\mu_i^H u_{ij}}{B_i J_i}, \nu^H u_{ij}),$$

mit $j = 1, ..., k_i$ und $SB_i = \sum_{j=1}^{k_i} SB_{ij}$.

Der Wert μ_i^H setzt sich wie folgt zusammen:

$$\mu_i^H = \exp\{\boldsymbol{x}^T \boldsymbol{\beta}^H\},$$

mit $\boldsymbol{\beta}^H = (\lambda^H, \lambda_1^{a,H}, ..., \lambda_4^{a,H}, \lambda_1^{c,H}, ..., \lambda_{18}^{c,H})^T$.

Das Modell V3G1 wird als das alternative Modell gewählt. Dieses Modell enthält neben dem Intercept die Variablen A und B. Sei nun $i \in I^K$ mit $I^K = \{i = (B, A)$, wobei die Kombination von B und A im betrachteten V Datensatz auftritt$\}$.

Für die Verteilung des Schadenbedarfs V_SB_i pro Zelle gilt hier:

$$V\_SB_i \sim \Gamma(\frac{\mu_i^K}{B_i}, \nu^K J_i)$$

und für den Schadenbedarf auf Einzeldatenebene erhält man:

$$SB_{ij} \sim \Gamma(\frac{\mu_i^K u_{ij}}{B_i J_i}, \nu^K u_{ij})$$

mit $j = 1, ..., k_i$ und $V\_SB_i = \sum_{j=1}^{k_i} V\_SB_{ij}$.

Der Wert μ_i^K setzt sich wie folgt zusammen:

$$\mu_i^K = \exp\{\boldsymbol{x}^T \boldsymbol{\beta}^K\}$$

mit $\boldsymbol{\beta}^K = (\lambda^K, \lambda_1^{B,K}, ..., \lambda_{16}^{B,K}, \lambda_1^{A,K}, ..., \lambda_{13}^{A,K})^T$.

Die Parameter μ_i^H und ν^H, $i \in I^H$, des *Wahren Modells* werden nun als Basis für die Simulation der Datensätze gewählt. Mit Hilfe dieser Parameter werden die Daten SB_{ij}^r simuliert mit $r = 1, .., R$. Somit gibt R die Anzahl der insgesamt simulierten Datensätze an.

Die Offsets sind bekannte Parameter und werden daher nicht simuliert. Bei dieser Simulation gilt für alle u_{ij}: $u_{ij} = 1$. Es werden also nur Datensätze mit Gewicht eins betrachtet. Durch diese Einschränkung reduziert sich die Anzahl der Datensätze auf 24.128 Beobachtungen (Beob = 24128). Die simulierten SB_{ij}^r sind wie folgt verteilt:

$$SB_{ij}^r \sim \Gamma(\frac{\mu_i^H}{B_i J_i}, \nu^H)$$

mit $i \in I^H, j = 1, ..., k_i$. Es wird somit das Gewicht eins und der Offset $\frac{1}{B_i J_i}$ berücksichtigt.

Für jeden der R Datensätze werden die Parameter des Gamma-verteilten Modells S6G1 mit Offset und Gewicht, $\mu_i^{r,H}$ und $\nu^{r,H}$, und die Parameter des Gamma-verteilten Modells V3G1 mit Offset und Gewicht, $\mu_i^{r,K}$ und $\nu^{r,K}$, geschätzt. Anschließend werden die Werte der Teststatistiken des Vuong Tests und der Distribution-Free Tests berechnet.

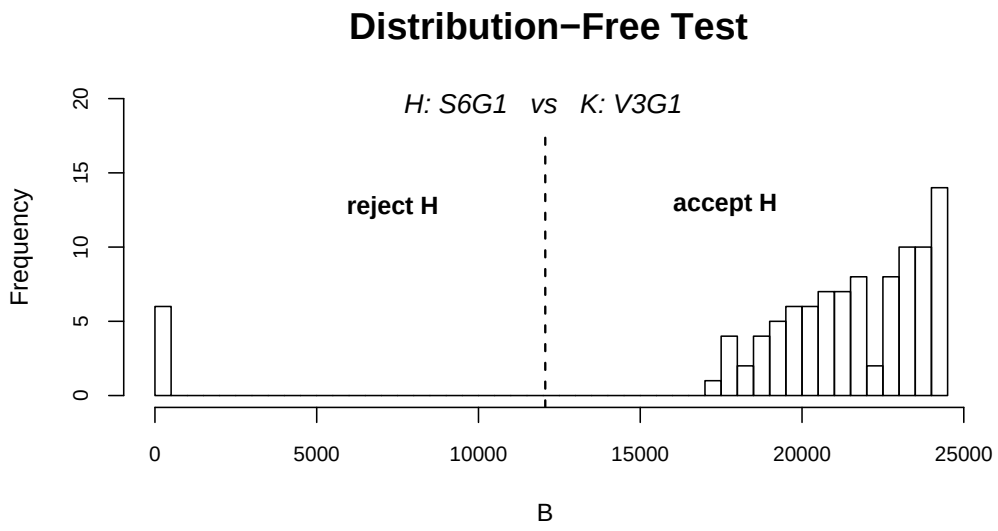
Fasst man die wichtigen Informationen dieses Abschnitts in tabellarischer Form zusammen, dann erhält man:

$$\begin{aligned}
\textit{Wahres Modell} &= S6G1 \\
\textit{Alternative} &= V3G1 \\
R &= 100 \\
\textit{Beob} &= 24.128 \\
\textit{Verteilung} &= \text{Gamma}
\end{aligned}$$

Im Folgenden wird auf eine ausführliche Beschreibung des *Wahren Modells*, der *Alternative*und des Vorgehens bei der Simulation der Datensätze verzichtet. Die wichtigen Informationen werden wie oben als tabellarische Übersicht angegeben.

Stellt man die Werte der Teststatistiken des Distribution-Free Tests dieser Simulation als Histogramm dar, dann erhält man folgende Abbildung:

Abbildung 7.1: Histogramm der Ergebnisse des Distribution-Free Tests für ohne Gewicht simulierte Datensätze mit Wahrem Modell S6G1



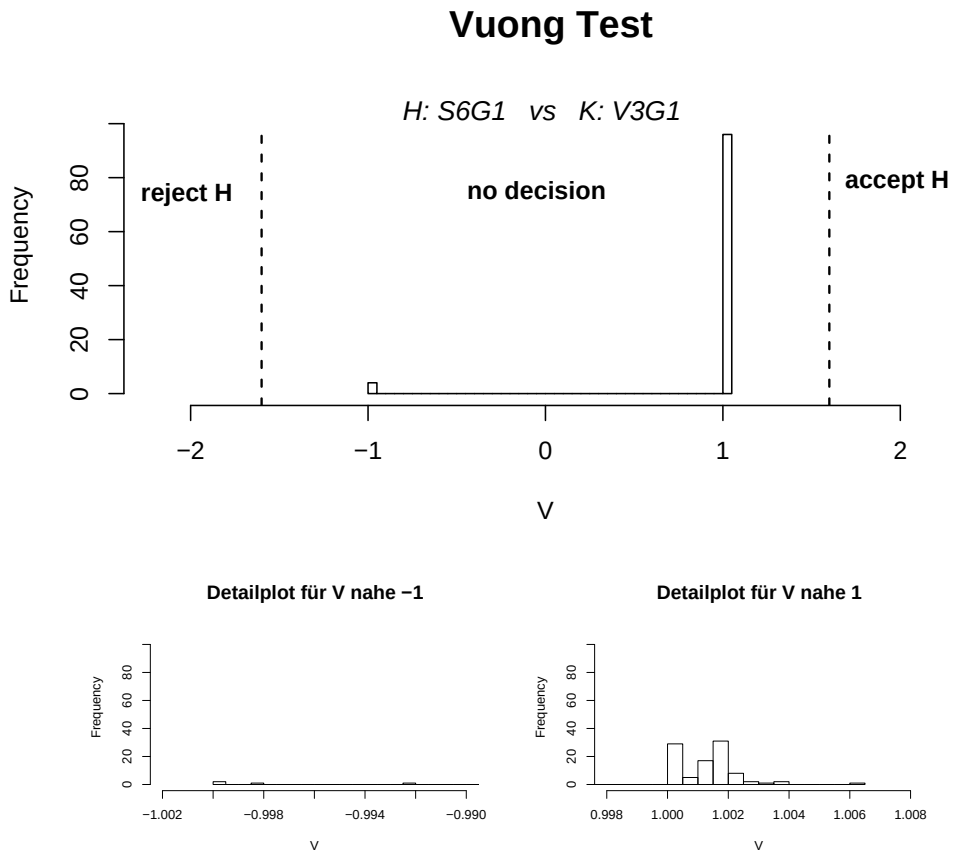
Die vertikale Linie in Abbildung 7.1 beim Wert 12.064 unterteilt Annahme- und Verwerfungsbereich des Distribution-Free Tests. Liegt der Wert der Teststatistik über $\frac{Beob}{2}$ wird $\boldsymbol{H}$ akzeptiert. Von den 100 simulierten Datensätzen entscheidet der Distribution-Free Test in 94 Fällen, dass das *Wahre Modell* S6G1 besser an die Daten angepasst ist. Somit erhält man eine geschätzte Güte von $\frac{94}{100}$.

Die nächste Grafik, Abbildung 7.2, zeigt das Histogramm des Vuong Tests für ohne Gewicht simulierte Datensätze mit S6G1 als *Wahres Modell*.

Die beiden vertikalen, gestrichelten Linien bei Wert -1,64 und 1,64 entsprechen dem 95% bzw. dem 5% Quantil der Standardnormalverteilung. Ist der Wert der Teststatistik kleiner als -1,64, dann wird das *Wahre Modell* als besser an die Daten angepasstes Modell abgelehnt. Für Werte über 1,64 wird die Hypothese $\boldsymbol{H}$ nicht verworfen. Liegt der Wert dazwischen, dann trifft der Vuong Test keine Entscheidung.

Für die simulierten Datensätze trifft der Vuong Test in 100 von 100 Fällen keine Entscheidung. Die geschätzte Güte beim Vuong Test liegt bei $\frac{0}{100}$. Hiermit bestätigt sich die

Abbildung 7.2: Histogramm der Ergebnisse des Vuong Tests für ohne Gewicht simulierte Datensätze mit Wahrem Modell S6G1



Vermutung der vorherigen Abschnitte, dass der Vuong Test zur Modellwahl bei dieser Problemstellung nicht geeignet ist.

In Tabelle 7.1 sind die Lagemaße der Werte der Teststatistiken des Distribution-Free Tests und des Vuong Tests für diese Simulation aufgeführt. Auch hier ist zu sehen, dass der Vuong Test keine Entscheidung trifft, der Distribution-Free Test jedoch das *Wahre Modell* als das bessere Modell identifiziert.

Tabelle 7.1: Zusammenfassung der Ergebnisse des Distribution-Free Tests und des Vuong Tests für ohne Gewicht simulierte Datensätze mit Wahrem Modell S6G1

Test	**Min**	**1st Quantil**	**Median**	**Mean**	**3rd Quantil**	**Max**
Distribution-Free	8877	18200	22240	20820	24120	24128
Vuong	-0.9997	1.0000	1.0000	0.8407	1.0000	1.0060

7.2 Simulation der Gütefunktion ohne Gewicht mit Wahrem Modell V3G1

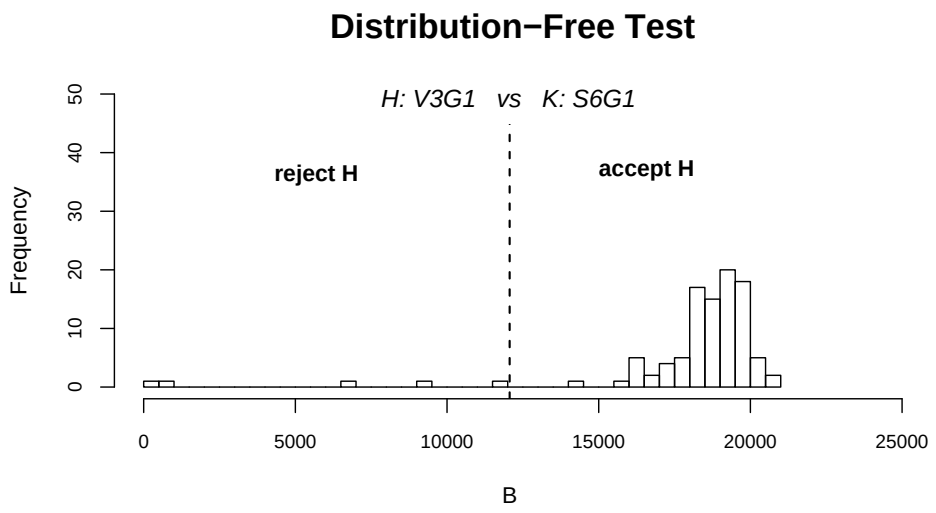
Bei der nächsten Simulation von Datensätzen wird das *Wahre Modell* und die *Alternative* vertauscht. Es wird untersucht, ob sich die Hypothesentests systematisch für ein Modell entscheiden.

Das Modell V3G1 wird nun als das *Wahre Modell* gewählt. Die Datensätze werden erneut ohne Gewicht simuliert, somit bleiben auch hier nur 24.128 Datensätze vom KH Datensatz übrig. Wie im vorherigen Abschnitt werden 100 Datensätze simuliert. Fasst man diese Informationen als tabellarische Übersicht zusammen, dann erhält man:

$$\begin{aligned} \textit{Wahres Modell} &= V3G1 \\ \textit{Alternative} &= S6G1 \\ R &= 100 \\ \textit{Beob} &= 24.128 \\ \textit{Verteilung} &= \text{Gamma} \end{aligned}$$

Abbildung 7.3 zeigt die Werte der Teststatistiken des Distribution-Free Tests als Histogramm dargestellt.

Abbildung 7.3: Histogramm der Ergebnisse des Distribution-Free Tests für ohne Gewicht simulierte Datensätze mit Wahrem Modell V3G1

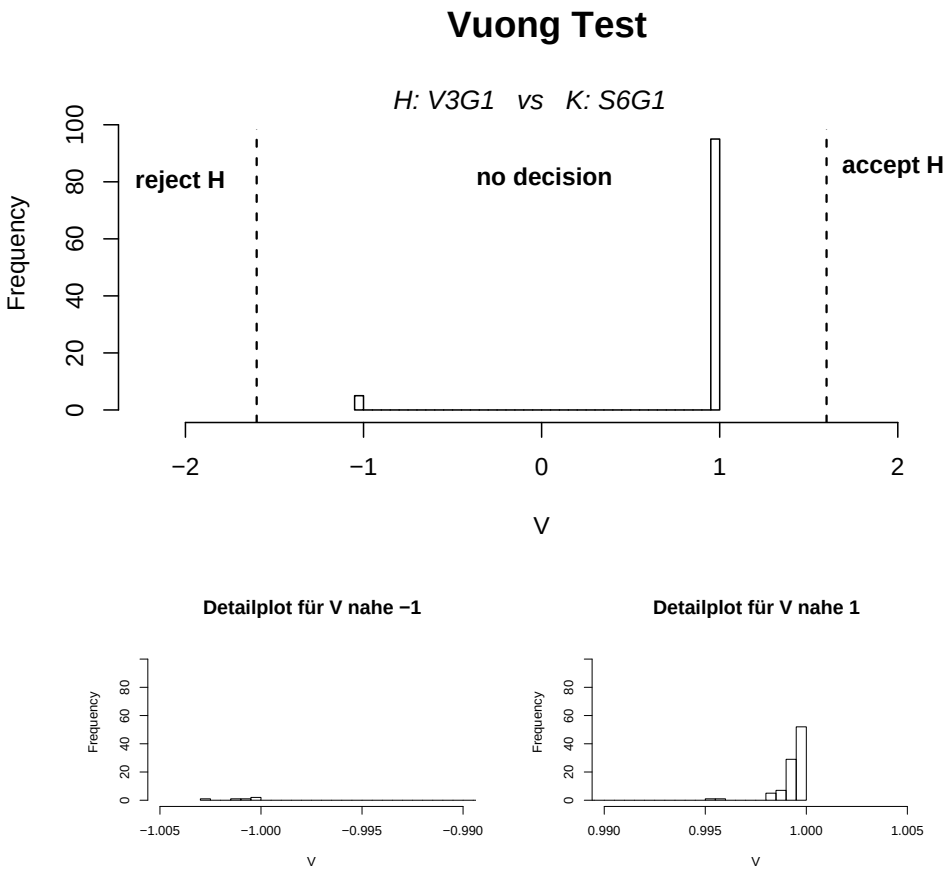


Die vertikale, gestrichelte Linie separiert auch hier den Annahme- und den Verwerfungsbereich des Distribution-Free Test. Die Linie ist am Wert $\frac{Beob}{2} = 12.064$ angetragen. Die Hypothese $\boldsymbol{H}$ wird bei 95 von 100 simulierten Datensätzen akzeptiert. Es ergibt sich eine geschätzte Güte von $\frac{95}{100}$.

Insgesamt kann man sagen, dass das Modell, für das sich der Distribution-Free Test bei Gamma-verteilten Modellen und ungewichteten Beobachtungen entscheidet, besser an die Daten angepasst ist.

Das Histogramm der Ergebnisse der Teststatistiken des Vuong Tests für diese Simulationen ist in Abbildung 7.4 illustriert.

Abbildung 7.4: Histogramm der Ergebnisse des Vuong Tests für ohne Gewicht simulierte Datensätze mit Wahrem Modell V3G1



Die beiden gestrichelten, vertikalen Linien unterteilen den Bereich für den der Vuong Test keine Entscheidung trifft, den Verwerfungsbereich und den Annahmebereich. Die Linien sind an den Werten des 5% und des 95% Quantils der Standardnormalverteilung ange-

tragen. Der Vuong Test trifft in keinem der 100 Fälle eine Entscheidung. Die geschätzte Güte ist $\frac{0}{100}$.

Somit ist der Vuong Test als Modellwahlkriterium für Gamma-verteilte nicht genestete Modelle mit ungewichteten Beobachtungen ungeeignet.

Abschließend sind in Tabelle 7.2 die Lagemaße der Werte der Teststatistiken dieser Simulation des Distribution-Free Tests und des Vuong Tests angegeben.

Tabelle 7.2: Zusammenfassung der Ergebnisse des Distribution-Free Tests und des Vuong Tests für ohne Gewicht simulierte Datensätze mit Wahrem Modell V3G1

Test	**Min**	**1st Quantil**	**Median**	**Mean**	**3rd Quantil**	**Max**
Distribution-Free	0	18050	18880	18050	19490	20700
Vuong	-1.0030	0.9993	0.9995	0.8993	0.9997	0.9999

Die Werte der Teststatistiken des Vuong Tests variieren um ± 1. Man erkennt, dass der Distribution-Free Test die Hypothese $\boldsymbol{H}$ in fast allen Fällen deutlich akzeptiert.

7.3 Simulation der Gütefunktion mit Gewicht und Wahrem Modell S6G1

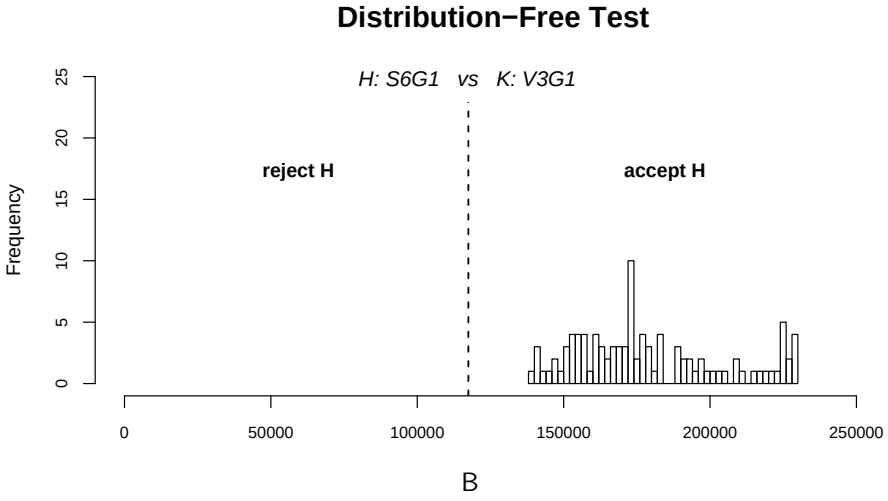
Bei den nächsten beiden Simulationen werden auch die gewichteten Beobachtungen mit einbezogen. Somit werden für alle 234.906 Beobachtungen des KH Datensatzes Werte simuliert.

Das Modell S6G1 wird als *Wahres Modell* gewählt. Es werden 100 Datensätze simuliert. Die für diese Simulation wichtigen Informationen können wie folgt zusammengefasst werden:

$$\begin{aligned} \textit{Wahres Modell} &= S6G1 \\ \textit{Alternative} &= V3G1 \\ R &= 100 \\ \textit{Beob} &= 234.906 \\ \textit{Verteilung} &= \text{Gamma} \end{aligned}$$

Berechnet man die Werte der Teststatistiken des Distribution-Free Tests und stellt diese als Histogramm dar, dann entsteht folgende Abbildung.

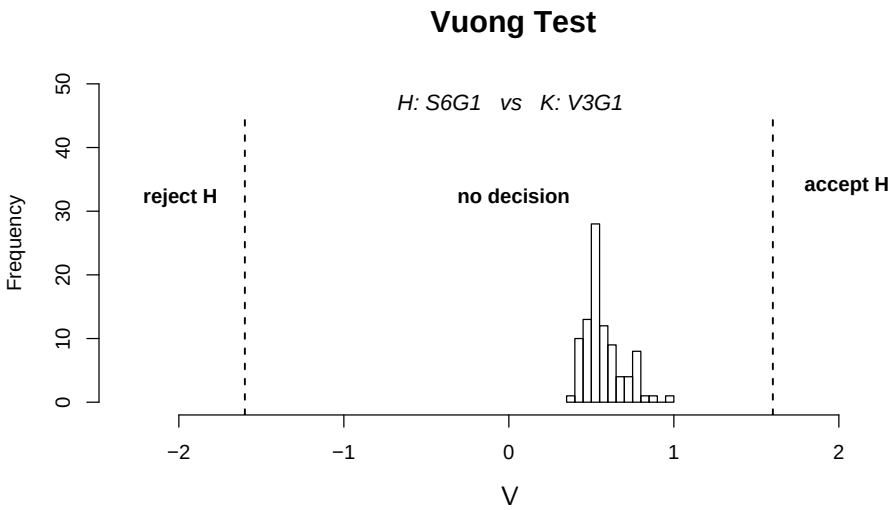
Abbildung 7.5: Histogramm der Ergebnisse des Distribution-Free Tests für mit Gewicht simulierte Datensätze mit Wahrem Modell S6G1



Durch die vertikale, gestrichelte Linie sind der Annahmebereich und der Verwerfungsbereich in Abbildung 7.5 separiert. Die Linie ist beim Wert 117.453 angetragen. In 100 von 100 Fällen wird die Hypothese $\boldsymbol{H}$ akzeptiert. Die geschätzte Güte ist $\frac{100}{100}$.

Das Histogramm der Werte der Teststatistiken des Vuong Tests ist in Abbildung 7.6 dargestellt.

Abbildung 7.6: Histogramm der Ergebnisse des Vuong Tests für mit Gewicht simulierte Datensätze mit Wahrem Modell S6G1



Der Bereich in dem der Vuong Test keine Entscheidung trifft, der Verwerfungsbereich und der Annahmebereich sind durch vertikale, gestrichelte Linien an den Werten des 5% bzw. des 95% Quantils der Standardnormalverteilung unterteilt. Der Vuong Test trifft für alle 100 Simulationen keine Entscheidung, die geschätzte Güte ist $\frac{0}{100}$.

Die Lagemaße der Werte des Distribution-Free Tests und des Vuong Tests für diese Simulation sind in Tabelle 7.3 zusammengefasst. Die Werte des Vuong Tests variierten zwischen 0,421 und 0,990.

Tabelle 7.3: Zusammenfassung der Ergebnisse des Distribution-Free Tests und des Vuong Tests für mit Gewicht simulierte Datensätze mit Wahrem Modell S6G1

Test	**Min**	**1st Quantil**	**Median**	**Mean**	**3rd Quantil**	**Max**
Distribution-Free	140000	160000	173000	179000	195000	229000
Vuong	0.421	0.470	0.528	0.582	0.607	0.990

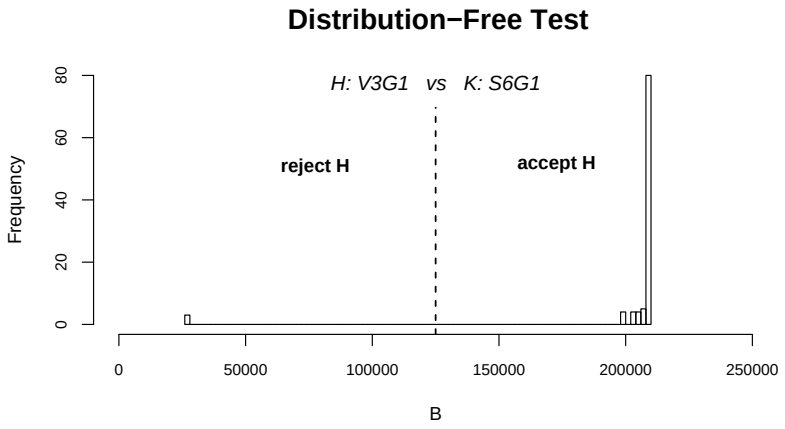
7.4 Simulation der Gütefunktion mit Gewicht und Wahrem Modell V3G1

Diese Simulation von Datensätzen hat denselben Aufbau wie die Simulation im vorherigen Abschnitt. Es werden nur das *Wahre Modell* und die *Alternative* vertauscht. Auch hier soll vermieden werden, dass sich die Hypothesentests systematisch für ein Modell entscheiden.

Man wählt das Modell V3G1 als *Wahres Modell*. Es wird der volle KH Datensatz simuliert, 234.906 Beobachtungen. Wie im vorherigen Abschnitt werden 100 Datensätze simuliert. Man kann diese Informationen in folgender Form tabellarisch zusammenfassen.

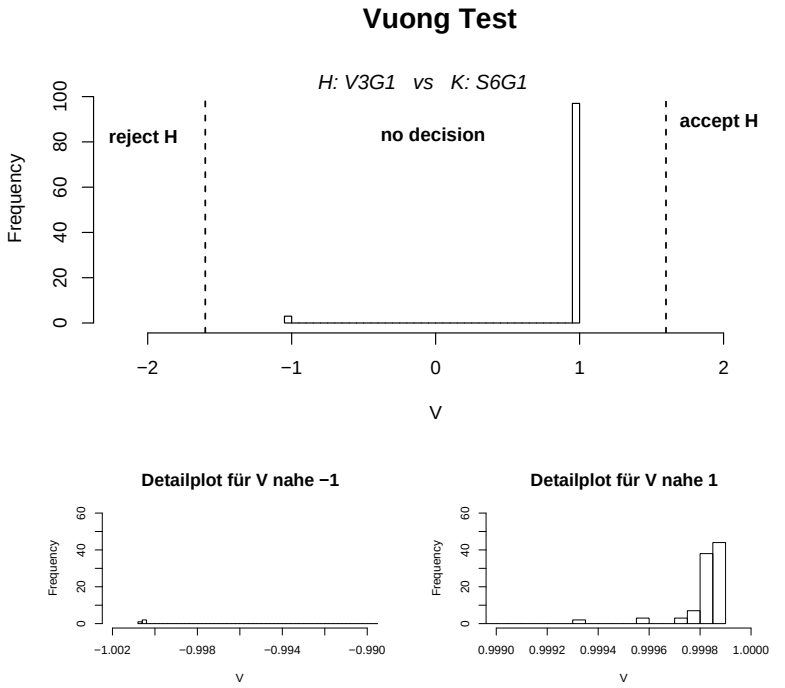
$$\begin{aligned} \textit{Wahres Modell} &= V3G1 \\ \textit{Alternative} &= S6G1 \\ R &= 100 \\ \textit{Beob} &= 234.906 \\ \textit{Verteilung} &= \text{Gamma} \end{aligned}$$

Abbildung 7.7: Histogramm der Ergebnisse des Distribution-Free Tests für mit Gewicht simulierte Datensätze mit Wahrem Modell V3G1



In Abbildung 7.7 sind die Werte der Teststatistiken des Distribution-Free Tests als Histogramm illustriert. Die vertikale, gestrichelte Linie am Wert 117.453 unterteilt den Annahmebereich und den Verwerfungsbereich des Distribution Free Tests. Für die simulierten Datensätze wird die Hypothese $\boldsymbol{H}$ in 98 von 100 Fällen akzeptiert. Die geschätzte Güte ist $\frac{98}{100}$.

Abbildung 7.8: Histogramm der Ergebnisse des Vuong Tests für mit Gewicht simulierte Datensätze mit Wahrem Modell V3G1



Somit eignet sich der Distribution-Free Test bei dieser Problemstellung auch für die Modellwahl von Gamma-verteilten Modellen mit gewichteten Beobachtungen.

Die Abbildung 7.8 beinhaltet die Werte der Teststatistiken des Vuong Tests dieser Simulation. Die verschiedenen Bereiche sind durch gestrichelte, vertikale Linien unterteilt. Diese sind am 5% bzw. 95% Quantil der Standardnormalverteilung angetragen.

Fällt der Wert der Teststatistik in den linken Bereich, dann wird die Hypothese $\boldsymbol{H}$ verworfen, im rechten Bereich wird die Hypothese $\boldsymbol{H}$ akzeptiert. Fallen die Werte in die Mitte, dann trifft der Vuong Test keine Entscheidung. Dies ist in der Abbildung 7.8 der Fall. Somit ergibt sich eine geschätzte Güte von $\frac{0}{100}$.

Daher ist der Vuong Test ebenfalls ungeeignet für die Modellwahl nicht genesteter Gamma-verteilter Modelle mit gewichteten Beobachtungen.

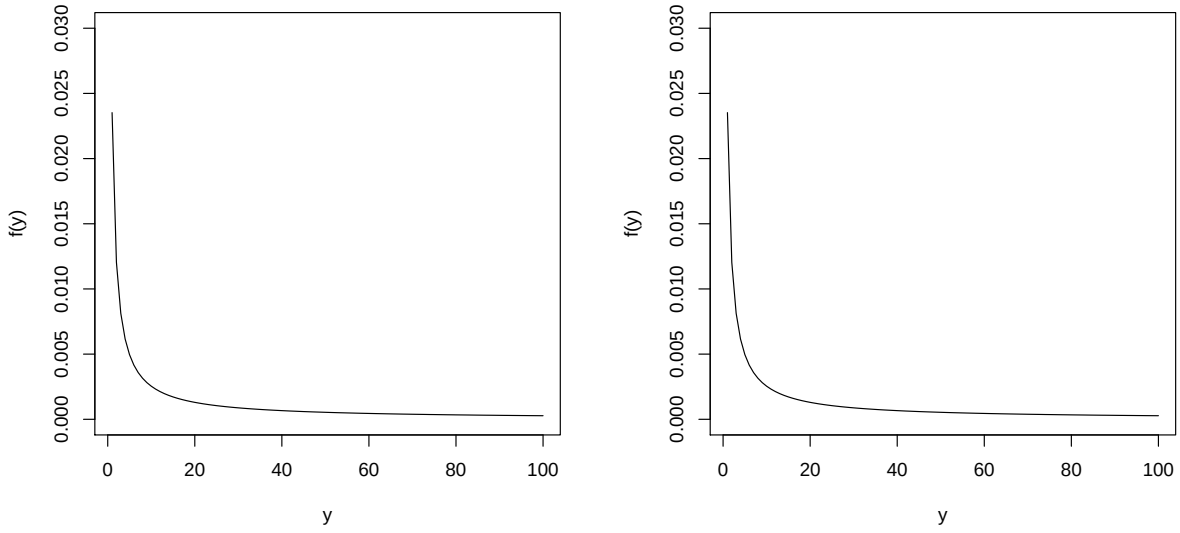
Clarke stellt fest, siehe Clarke (2007), dass der Vuong Test für sehr spitze Verteilungen ungeeignet ist. Betrachtet man dem Mittelwert für die Erwartungswerte der simulierten Datensätze mit und ohne Gewicht (Wahres Modell S6G1) in Tabelle 7.4, dann erkennt man, dass diese sehr klein sind.

Tabelle 7.4: Mittelwert für die Erwartungswerte der simulierten Datensätze mit und ohne Gewicht bei Wahrem Modell S6G1

Abschnitt	*Wahres Modell*	**Gewicht**	**geschätztes Modell**	**Mittelwert der μ_i**
7.1	S6G1	ohne	S6G1	0.0016
7.1	S6G1	ohne	V3G1	0.0027
7.3	S6G1	mit	S6G1	0.0005
7.3	S6G1	mit	V3G1	0.0003

Liegt der Wert des Dispersionsparameters bei 34,1, dann hat die Dichte der Gamma-Verteilung für die Werte $\mu_i = 0,0005$ und $\mu_i = 0,001$ die folgende Gestalt, siehe Abbildung 7.9.

Abbildung 7.9: Darstellung der Dichte der Gamma-Verteilung



(a) $\mu = 0,001$ und Dispersion= $34,1$

(b) $\mu = 0,0005$ und Dispersion= $34,1$

Tabelle 7.5 enthält die Lagemaße des Distribution-Free Tests und des Vuong Tests für diese Simulation.

Tabelle 7.5: Zusammenfassung der Ergebnisse des Distribution-Free Tests und des Vuong Tests für mit Gewicht simulierte Datensätze mit Wahrem Modell V3G1

Test	**Min**	**1st Quantil**	**Median**	**Mean**	**3rd Quantil**	**Max**
Distribution-Free	26400	208501	203000	179000	208944	209000
Vuong	-1.001	0.998	0.998	0.940	0.999	0.999

7.5 Simulation der Gütefunktion mit Gewicht und Wahrem Modell S6P1

Die nächsten beiden Abschnitte betrachten Simulationen der Gütefunktion von Poisson-verteilten Modellen. Es werden nur gewichtete Beobachtungen simuliert, da bei den Poisson Modellen das Gewicht eins im KH Datensatz nicht vorkommt. Somit werden Werte für alle 234.906 Beobachtungen des KH Datensatzes simuliert.

In diesem Fall wird das Modell S6P1 als *Wahres Modell* gewählt und 100 Datensätze simuliert. Man kann die für diese Simulation wichtigen Informationen wie folgt tabellarisch zusammenfassen:

$$\begin{aligned} \textit{Wahres Modell} &= S6P1 \\ \textit{Alternative} &= V3P1 \\ R &= 100 \\ \textit{Beob} &= 234.906 \\ \textit{Verteilung} &= \text{Poisson} \end{aligned}$$

Abbildung 7.10: Histogramm der Ergebnisse des Distribution-Free Tests für mit Gewicht simulierte Datensätze mit Wahrem Modell S6P1

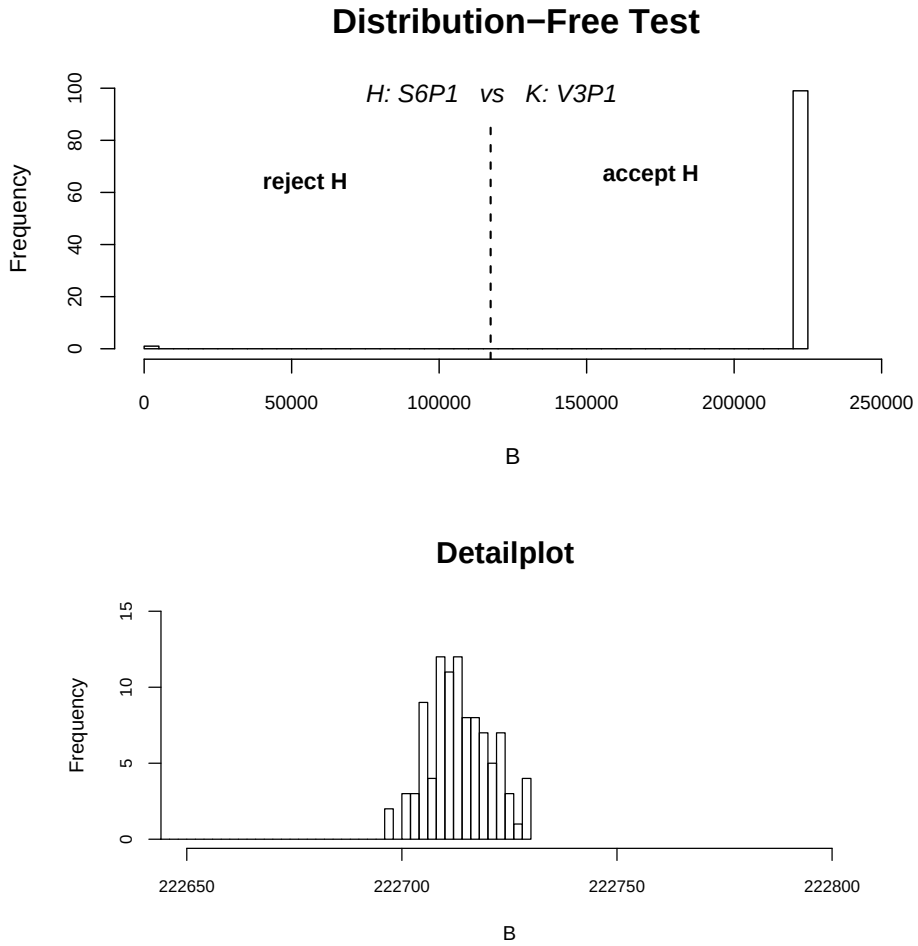
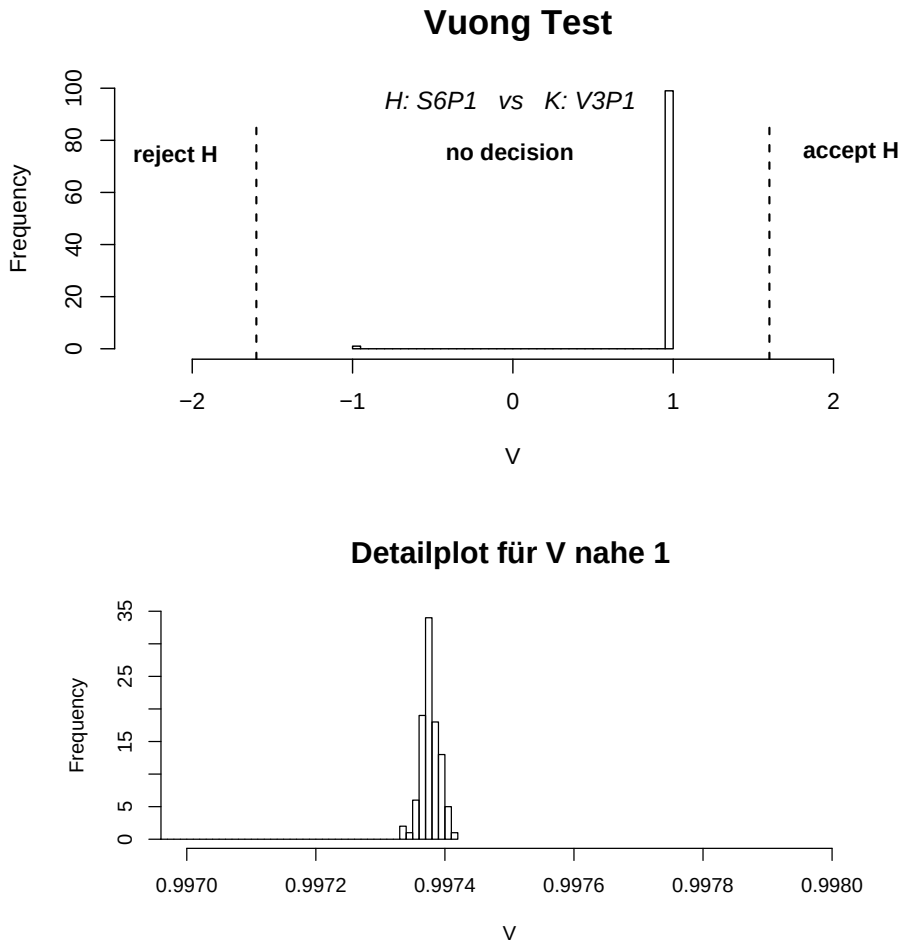


Abbildung 7.10 illustriert die Werte der Teststatistiken des Distribution-Free Tests als Histogramm.

Die vertikale, gestrichelte Linie beim Wert 117.453 separiert Annahmebereich und Verwerfungsbereich des Distribution-Free Tests. Wie man in Abbildung 7.10 sieht, ist die geschätzte Güte dieser Simulation $\frac{99}{100}$. Das heißt die Hypothese $\boldsymbol{H}$ wird in 99 von 100 Fällen akzeptiert.

Eine Darstellung der Werte der Teststatistiken des Vuong Tests findet sich in Abbildung 7.11.

Abbildung 7.11: Histogramm der Ergebnisse des Vuong Tests für mit Gewicht simulierte Datensätze mit Wahrem Modell S6P1



Der Bereich in dem der Vuong Test keine Entscheidung trifft, der Annahmebereich und der Verwerfungsbereich sind durch vertikale, gestrichelte Linien an den Werten des 5% und des 95% Quantils der Standardnormalverteilung getrennt.

Die geschätzte Güte des Vuong Tests ist $\frac{0}{100}$, siehe Abbildung 7.11. Also wird die Hypothese $\boldsymbol{H}$ in keinem der Fälle akzeptiert.

In Tabelle 7.6 werden die Lagemaße der Werte des Distribution-Free Tests und des Vuong Tests zusammengefasst.

Tabelle 7.6: Zusammenfassung der Ergebnisse des Distribution-Free Tests und des Vuong Tests für mit Gewicht simulierte Datensätze mit Wahrem Modell S6P1

Test	**Min**	**1st Quantil**	**Median**	**Mean**	**3rd Quantil**	**Max**
Distribution-Free	7	222709	222713	220486	222719	222730
Vuong	-0.997	0.997	0.997	0.987	0.997	0.997

7.6 Simulation der Gütefunktion mit Gewicht und Wahrem Modell V3P1

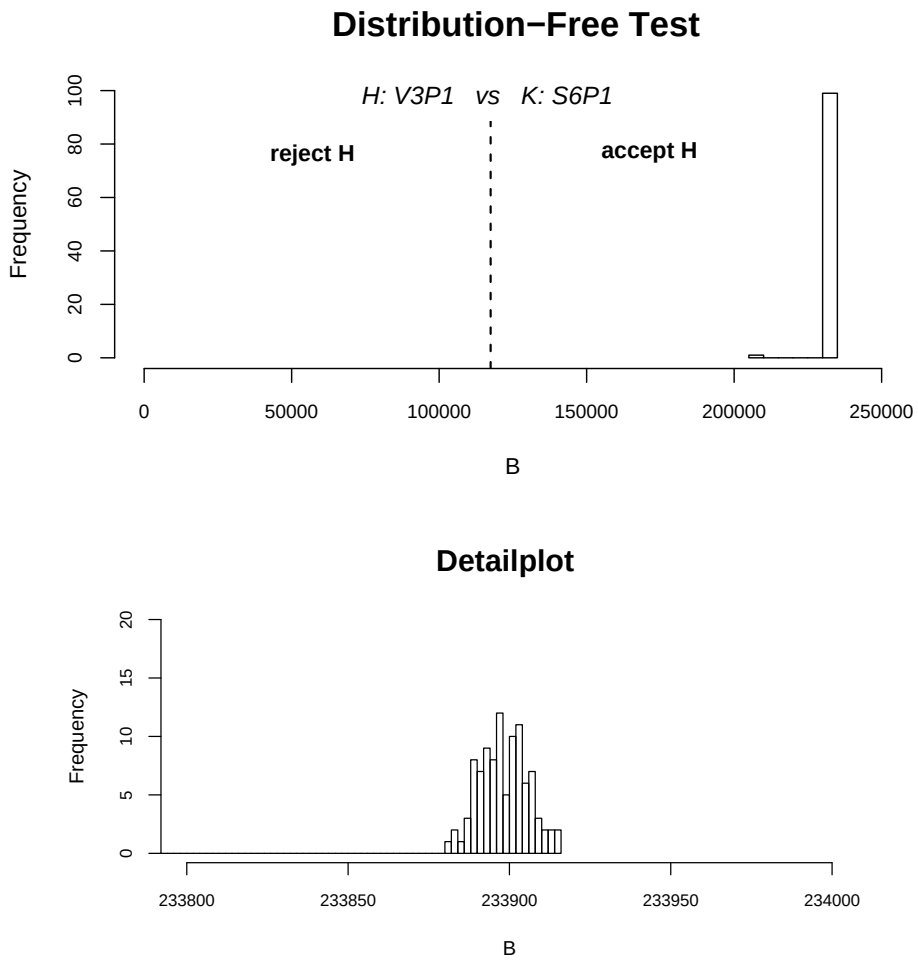
Die Alternative und die Hypothese werden in dieser Simulation erneut vertauscht. Damit wird untersucht, ob sich die Hypothesentests systematisch für oder gegen eines der Modelle entscheiden. Sonst hat die Simulation denselben Aufbau wie die Simulation im vorherigen Abschnitt.

Es werden 100 Datensätze simuliert. Für alle 234.906 Beobachtungen des KH Datensatzes werden Werte simuliert. Die untersuchten Modelle sind Poisson-verteilt. Das Modell V3P1 wird als *Wahres Modell* gewählt. Die tabellarische Übersicht der für diese Simulation wichtigen Informationen sieht wie folgt aus:

$$\begin{aligned} \textit{Wahres Modell} &= V3P1 \\ \textit{Alternative} &= S6P1 \\ R &= 100 \\ \textit{Beob} &= 234.906 \\ \textit{Verteilung} &= \text{Poisson} \end{aligned}$$

Die Werte der Teststatistiken des Distribution-Free Tests sind in Abbildung 7.12 als Histogramm dargestellt.

Abbildung 7.12: Histogramm der Ergebnisse des Distribution-Free Tests für mit Gewicht simulierte Datensätze mit Wahrem Modell V3P1



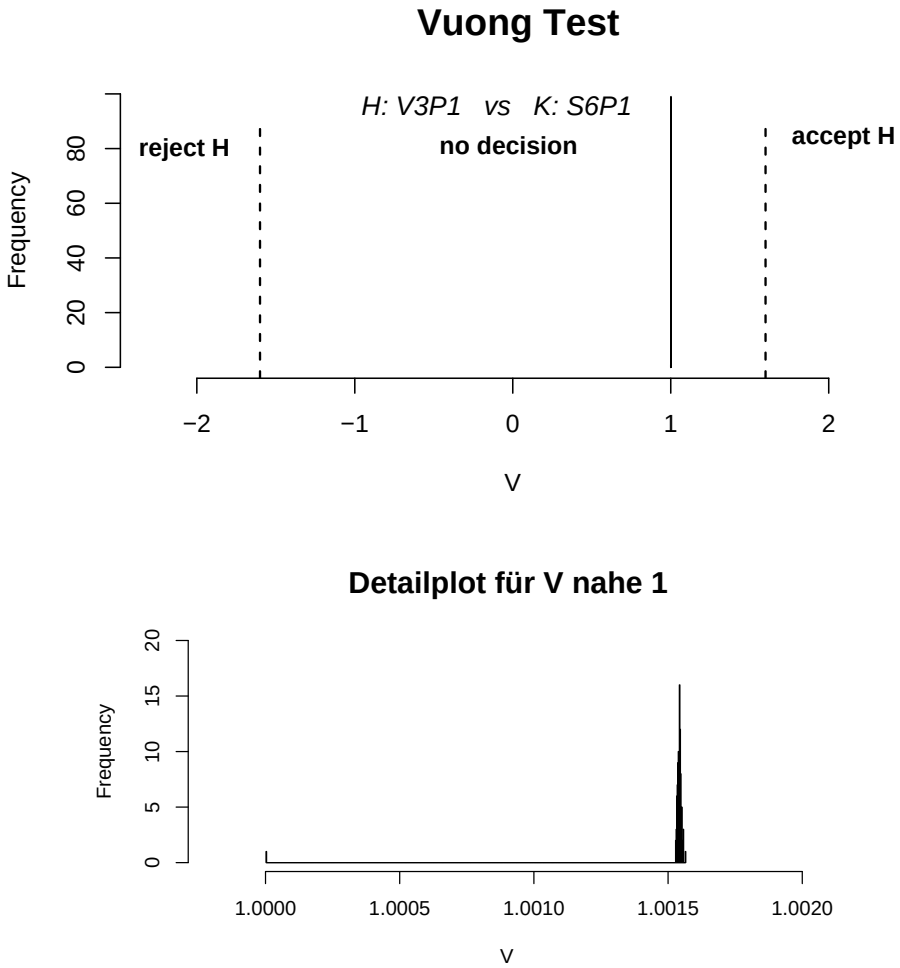
Der Annahmebereich und der Verwerfungsbereich des Distribution-Free Tests werden durch vertikale, gestrichelte Linien am Wert 117.453 unterteilt. Die Hypothese $\boldsymbol{H}$ des Distribution-Free Tests wird in 100 von 100 Fällen akzeptiert und man erhält somit eine geschätzte Güte von $\frac{100}{100}$.

Der Distribution-Free Test kann somit zum Vergleich nicht genesteter Poisson-verteilter Modelle verwendet werden.

In Abbildung 7.13 sind die Werte der Teststatistiken des Vuong Tests illustriert. Die beiden vertikalen, gestrichelten Linien unterteilen den Bereich in dem der Vuong Test keine

Entscheidung trifft, den Annahmebereich und den Verwerfungsbereich. Die Linien sind an den Werten des 5% und des 95% Quantils der Standardnormalverteilung angetragen.

Abbildung 7.13: Histogramm der Ergebnisse des Vuong Tests für mit Gewicht simulierte Datensätze mit Wahrem Modell V3P1



Auch hier trifft der Vuong Test keine Entscheidung, wie man in Abbildung 7.13 erkennt. Man erhält eine geschätzte Güte von $\frac{0}{100}$.

Der Vuong Test ist somit ungeeignet, um nicht genestete Poisson-verteilte Modelle mit Gewicht zu vergleichen.

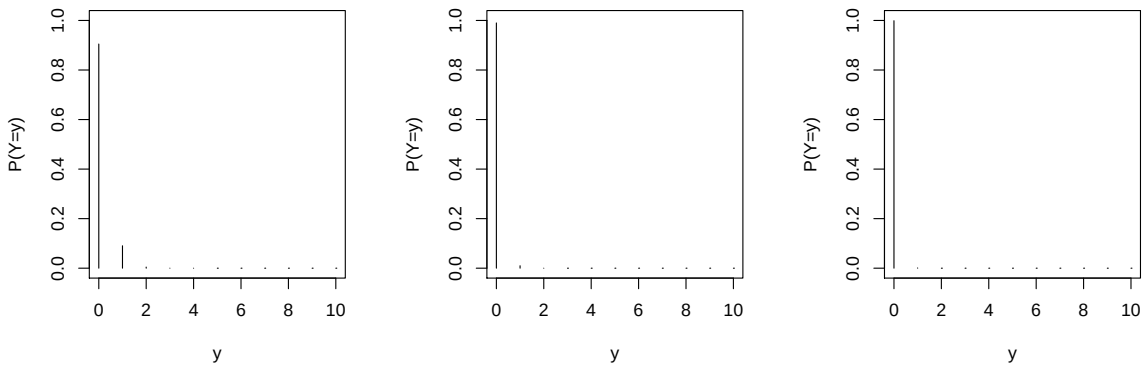
In Abschnitt 7.4 wurde bereits erwähnt, dass der Vuong Test für spitze Verteilungen nicht geeignet ist, siehe Clarke (2007). Betrachtet man diese Simulation, dann erkennt man, dass das Mittel der Erwartungswerte für die Hypothese V3P1 bei 0.0177 und das Mittel der Erwartungswerte für die Alternative bei 0.0176 liegt, siehe Tabelle 7.7.

Tabelle 7.7: Mittelwert für die Erwartungswerte der simulierten Datensätze mit und ohne Gewicht bei Wahrem Modell V3P1

Wahres Modell	**Gewicht**	**geschätztes Modell**	**Mittelwert der μ_i**
V3P1	mit	V3P1	0.0177
V3P1	mit	S6P1	0.0176

Die Wahrscheinlichkeitsfunktion für verschiedene Werte von λ werden in Abbildung 7.14 betrachtet. Man erkennt, dass die Wahrscheinlichkeitsfunktion umso spitzer wird je kleiner der Wert von λ ist.

Abbildung 7.14: Darstellung der Wahrscheinlichkeitsfunktion der Poisson-Verteilung für verschiedene Werte von λ



(a) $\boldsymbol{\lambda} = 0,1$ (b) $\boldsymbol{\lambda} = 0,01$ (c) $\boldsymbol{\lambda} = 0,001$

Die Lagemaße des Distribution-Free Tests und des Vuong Tests werden abschließend in Tabelle 7.8 angegeben.

Tabelle 7.8: Zusammenfassung der Ergebnisse des Distribution-Free Tests und des Vuong Tests für mit Gewicht simulierte Datensätze mit Wahrem Modell V3P1

Test	**Min**	**1st Quantil**	**Median**	**Mean**	**3rd Quantil**	**Max**
Distribution-Free	206208	233890	233898	233622	233904	233916
Vuong	1.000	1.002	1.002	1.002	1.002	1.002

8 Zusammenfassung

Ziel dieser Diplomarbeit ist es nicht genestete Kraftfahrt Haftpflicht Tarife zu vergleichen und den besten Tarif auszuwählen.

Es werden Kundendaten der Versicherungskammer Bayern aus den Jahren 2006 und 2007 verwendet. Es ist nötig den ursprünglichen KH Datensatz auf zwei Weisen zu aggregieren, so entstehen der S und der V Datensatz. Die unterschiedliche Verdichtung der Daten ist notwendig, da die beiden Tarife, die modelliert werden, nicht genestet sind. Unterschiedliche Verteilungen werden als Verteilung der Zielvariable diskutiert und schließlich die Gamma-Verteilung und die Poisson-Verteilung ausgewählt.

Durch die Wahl der Gamma-Verteilung wird eine Manipulation der Zielvariable Schadenbedarf nötig. Die geschätzten Regressionsparameter der ursprünglichen und der manipulierten Zielvariable werden verglichen. Es wird festgestellt, dass die Manipulation die geschätzten Regressionsparameter kaum verändert.

Durch die beiden Verteilungen und die zwei aggregierten Datensätze entstehen vier Gruppen genesteter Modelle, die zuerst getrennt voneinander betrachtet werden. Der Goodness of Fit, der Partial Devianz Test und der Residual Devianz Test dieser Modelle werden untersucht und das beste Modell aus jeder Gruppe ausgewählt.

Es folgt eine Modellwahl für nicht genestete Modelle. Zuerst wird die Verteilung der Zielvariablen verglichen. Die Modellwahl erfolgt durch den Distribution-Free Test und den Vuong Test. Mit diesen beiden Hypothesentests können sowohl genestete als auch nicht genestete Modelle betrachtet werden. Man kommt zu dem Ergebnis, dass die Gamma-verteilten Modelle besser an die Daten angepasst sind. Mit denselben beiden Tests werden anschließend die nicht genesteten Modelle des S und des V Datensatzes verglichen. Das Ergebnis ist, dass das Modell S6G1, siehe Tabelle 4.3, am besten an die Daten angepasst ist.

Es schließt ein Kapitel mit Simulationen der Gütefunktion des Distribution-Free Tests und des Vuong Tests an. Sowohl beim Vergleich der nicht genesteten Modelle als auch bei

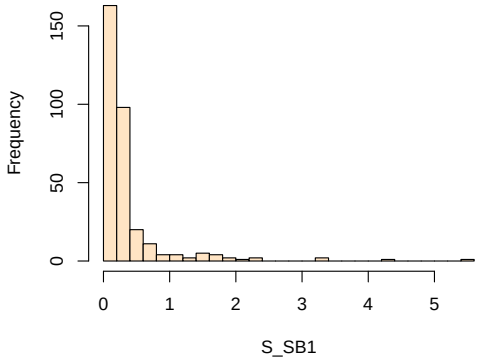
der Simulation der Gütefunktion wird festgestellt, dass der Vuong Test in keinem Fall eine Entscheidung zugunsten oder gegen ein Modell trifft. Clarke (2007) stellt fest, dass der Vuong Test für sehr spitze Verteilungen ungeeignet ist. Der Vuong Tests kann somit zum Vergleich von Kraftfahrt Haftpflicht Tarifen nicht verwendet werden, weder zum Vergleich genesteter noch zum Vergleich nicht genesteter Modelle.

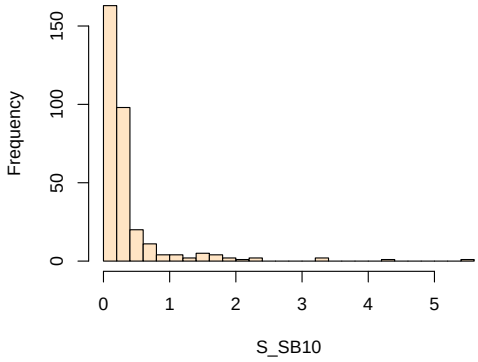
Der Distribution-Free Test hingegen eignet sich zur Wahl nicht genesteter Modelle, wie die Simulationen im letzten Kapitel zeigen. Er kann auch für Datensätze mit Gewicht verwendet werden.

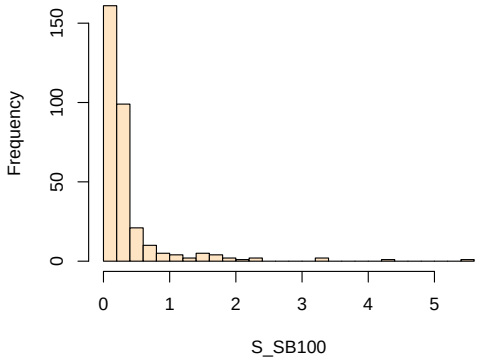
Mit dem Distribution-Free Test hat man nun ein Instrument um nicht genestete Modelle vergleichen zu können. Dies stellt eine deutliche Erweiterung der Hypothesentests dar, da man mit den üblichen Hypothesentests nur genestete Modelle vergleichen kann.

Anhang

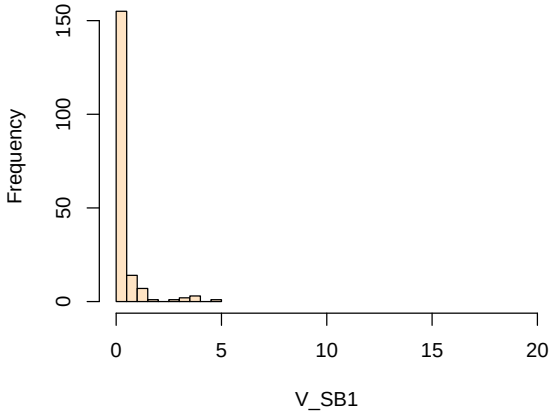
Histogramme von S_SB1, S_SB10 und S_SB100 (Abschnitt 3.5.1)

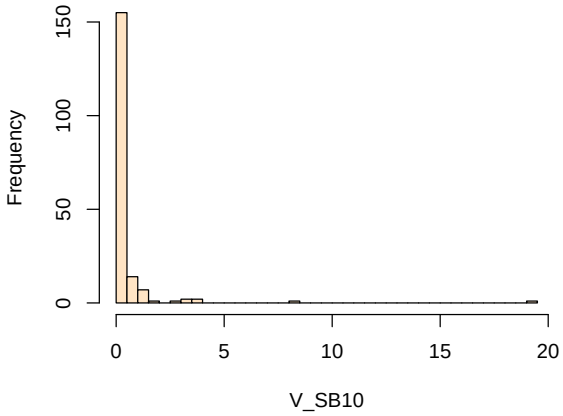


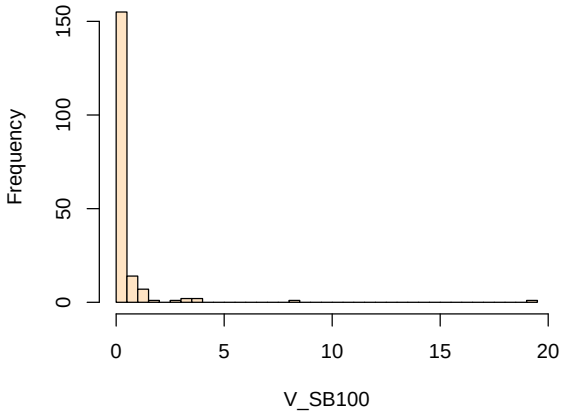




Histogramme von V_SB1, V_SB10 und V_SB100 (Abschnitt 3.5.2)







Vergleich der Regressionsparameter der manipulierten S6G Modelle (Abschnitt 4.4)

	S6G1	**S6G10**	**Differenz**	**S6G100**	**Differenz**
Intercept	4.727	4.729	−0.002	4.747	−0.019
a0	0.000	0.000	0.000	0.000	0.000
a1	0.425	0.425	0.001	0.420	0.006
a2	1.081	1.080	0.001	1.070	0.011
a3	1.219	1.218	0.001	1.208	0.011
c0	0.000	0.000	0.000	0.000	0.000
c1	0.065	0.065	0.000	0.063	0.002
c2	0.108	0.108	0.000	0.105	0.003
c3	0.118	0.117	0.000	0.114	0.004
c4	−0.078	−0.078	0.000	−0.078	0.000
c5	0.131	0.130	0.000	0.127	0.003
c6	0.051	0.051	0.000	0.048	0.003
c7	0.046	0.046	0.000	0.044	0.003
c8	0.231	0.231	0.000	0.226	0.005
c9	0.057	0.057	0.000	0.054	0.003
c10	0.176	0.176	0.000	0.171	0.005
c11	−0.118	−0.118	0.000	−0.118	0.000
c12	−0.032	−0.032	0.000	−0.033	0.001
c13	−0.074	−0.074	0.000	−0.076	0.001
c14	0.205	0.205	0.000	0.201	0.004
c15	0.274	0.273	0.001	0.268	0.006
c16	0.151	0.150	0.000	0.147	0.004
c17	0.333	0.333	0.001	0.327	0.006
c18	0.197	0.196	0.000	0.192	0.005

Vergleich der Regressionsparameter der manipulierten S7G Modelle (Abschnitt 4.4)

	S7G1	**S7G10**	**Differenz**	**S7G100**	**Differenz**
Intercept	5.174	5.176	−0.002	5.191	−0.017
a0	0.000	0.000	0.000	0.000	0.000
a1	0.425	0.424	0.001	0.419	0.006
a2	1.082	1.081	0.001	1.071	0.011
a3	1.213	1.212	0.001	1.203	0.011
c0	0.000	0.000	0.000	0.000	0.000
c1	0.067	0.067	0.000	0.065	0.002
c2	0.106	0.106	0.000	0.103	0.003
c3	0.125	0.124	0.000	0.121	0.004
c4	−0.083	−0.083	0.000	−0.083	0.000
c5	0.137	0.136	0.000	0.133	0.004
c6	0.056	0.055	0.000	0.052	0.003
c7	0.052	0.052	0.000	0.049	0.003
c8	0.233	0.232	0.000	0.228	0.005
c9	0.066	0.066	0.000	0.063	0.003
c10	0.182	0.181	0.000	0.177	0.005
c11	−0.103	−0.103	0.000	−0.103	0.000
c12	−0.025	−0.025	0.000	−0.026	0.001
c13	−0.068	−0.068	0.000	−0.069	0.001
c14	0.214	0.213	0.000	0.209	0.005
c15	0.289	0.288	0.001	0.283	0.006
c16	0.152	0.152	0.000	0.148	0.004
c17	0.350	0.349	0.001	0.343	0.007
c18	0.215	0.214	0.000	0.210	0.005
b0	0.000	0.000	0.000	0.000	0.000
b1	−0.587	−0.586	−0.001	−0.580	−0.006
b2	−0.500	−0.500	0.000	−0.497	−0.004
b3	−0.411	−0.411	0.000	−0.409	−0.002
b4	−0.464	−0.464	0.000	−0.461	−0.002

Vergleich der Regressionsparameter der manipulierten S7P Modelle (Abschnitt 4.4)

	S7P	S7P1	Differenz	S7P10	Differenz	S7P100	Differenz
Intercept	−0.910	2.255	−3.165	4.553	−5.462	6.760	−7.670
a0	0.000	0.000	0.000	0.000	0.000	0.000	0.000
a1	−1.487	−1.327	−0.160	−0.693	−0.794	−0.103	−1.384
a2	−3.355	−2.293	−1.062	−0.507	−2.848	0.615	−3.970
a3	−5.154	−2.802	−2.353	−0.537	−4.617	0.905	−6.059
c0	0.000	0.000	0.000	0.000	0.000	0.000	0.000
c1	1.014	0.876	0.138	0.314	0.700	−0.266	1.280
c2	0.278	0.067	0.211	−0.828	1.106	−1.835	2.113
c3	1.380	1.295	0.084	0.985	0.394	0.681	0.698
c4	0.274	0.334	−0.060	0.428	−0.154	0.404	−0.130
c5	1.130	0.998	0.132	0.516	0.614	0.084	1.045
c6	1.039	1.088	−0.050	1.093	−0.054	0.894	0.145
c7	0.984	0.791	0.193	−0.041	1.025	−0.933	1.918
c8	0.524	0.418	0.106	0.033	0.490	−0.321	0.845
c9	1.094	0.967	0.127	0.538	0.556	0.220	0.874
c10	0.581	0.413	0.168	−0.225	0.807	−0.788	1.370
c11	0.245	0.084	0.161	−0.488	0.733	−0.942	1.186
c12	0.507	0.527	−0.020	0.537	−0.030	0.481	0.026
c13	−0.109	−0.183	0.074	−0.360	0.252	−0.396	0.287
c14	0.666	0.444	0.223	−0.504	1.170	−1.673	2.339
c15	0.358	0.361	−0.004	0.266	0.091	0.027	0.331
c16	−0.016	0.125	−0.140	0.459	−0.475	0.662	−0.678
c17	−0.648	−0.424	−0.224	0.024	−0.671	0.256	−0.903
c18	−0.788	−0.505	−0.283	−0.023	−0.766	0.139	−0.927
b0	0.000	0.000	0.000	0.000	0.000	0.000	0.000
b1	2.242	0.054	2.187	−0.534	2.776	−0.846	3.088
b2	4.506	1.348	3.158	−0.957	5.463	−2.395	6.901
b3	4.964	1.854	3.110	−0.207	5.171	−1.329	6.293
b4	4.577	1.431	3.146	−0.811	5.388	−2.156	6.732

Vergleich der Regressionsparameter der manipulierten V3P Modelle (Abschnitt 4.4)

	V3P	V3P1	Differenz	V3P10	Differenz	V3P100	Differenz
Intercept	−15.519	−10.272	−5.247	−7.919	−7.600	−5.551	−9.967
A1	0.000	0.000	0.000	0.000	0.000	0.000	0.000
A2	0.184	0.175	0.008	0.119	0.065	−0.008	0.192
A3	0.134	0.137	−0.003	0.160	−0.026	0.210	−0.075
A4	0.420	0.410	0.011	0.341	0.080	0.198	0.222
A5	0.246	0.223	0.023	0.144	0.102	0.339	−0.093
A6	0.564	0.519	0.045	0.273	0.290	−0.044	0.608
A7	−0.266	−0.329	0.063	−0.329	0.064	−0.136	−0.130
A8	0.081	0.001	0.080	−0.182	0.263	−0.116	0.197
A9	−2.661	−1.353	−1.308	−0.327	−2.334	0.093	−2.754
A10	−1.721	−0.790	−0.931	−0.259	−1.461	0.039	−1.760
A11	−7.738	−6.793	−0.945	−4.834	−2.904	−3.314	−4.424
A12	−9.978	−8.279	−1.699	−6.183	−3.795	−3.931	−6.047
A13	−9.365	−7.868	−1.497	−5.820	−3.545	−3.528	−5.837
B10	0.000	0.000	0.000	0.000	0.000	0.000	0.000
B11	1.849	−0.025	1.874	0.003	1.845	0.016	1.832
B12	4.854	0.928	3.925	0.558	4.296	0.254	4.600
B13	13.402	8.107	5.295	5.527	7.875	3.645	9.756
B14	13.785	8.550	5.235	6.240	7.545	3.889	9.896
B15	13.773	8.536	5.237	6.226	7.547	3.897	9.876
B16	13.872	8.632	5.240	6.307	7.565	3.914	9.958
B17	13.835	8.598	5.237	6.290	7.545	3.959	9.876
B18	13.949	8.710	5.239	6.400	7.549	4.074	9.875
B19	14.037	8.793	5.244	6.457	7.579	4.023	10.014
B20	13.992	8.724	5.268	6.184	7.808	3.940	10.052
B21	8.424	2.374	6.051	0.477	7.947	−0.150	8.574
B22	5.676	1.189	4.487	0.416	5.260	0.017	5.659
B23	5.577	0.965	4.612	0.279	5.298	−0.038	5.615
B24	4.725	0.552	4.172	0.120	4.605	−0.320	5.045
B25	5.160	0.593	4.567	−0.024	5.183	−0.382	5.542

R Output des Modells S6G1 (Abschnitt 4.5)

```
Call:
glm(formula = S_SB1 ~ offset(-log(S_B)) + a + c, family = Gamma(link = "log"),
weights = S_J)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-27.33   -6.91   -3.01    1.84   15.64

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
Intercept    4.7271     0.1048   45.11  < 2e-16 ***
a1           0.4254     0.0538    7.91  5.2e-14 ***
a2           1.0814     0.1595    6.78  6.4e-11 ***
a3           1.2188     0.4603    2.65   0.0085 **
c1           0.0648     0.1213    0.53   0.5935
c2           0.1085     0.1369    0.79   0.4288
c3           0.1177     0.1180    1.00   0.3191
c4          -0.0776     0.1377   -0.56   0.5735
c5           0.1306     0.1217    1.07   0.2840
c6           0.0511     0.1224    0.42   0.6764
c7           0.0464     0.1239    0.37   0.7082
c8           0.2314     0.1345    1.72   0.0864 .
c9           0.0570     0.1211    0.47   0.6381
c10          0.1761     0.1347    1.31   0.1921
c11         -0.1183     0.1387   -0.85   0.3943
c12         -0.0322     0.1334   -0.24   0.8093
c13         -0.0744     0.1496   -0.50   0.6194
c14          0.2050     0.1305    1.57   0.1171
c15          0.2737     0.1375    1.99   0.0474 *
c16          0.1508     0.1500    1.01   0.3154
c17          0.3333     0.1853    1.80   0.0730 .
c18          0.1968     0.1920    1.02   0.3062
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 36.8)

    Null deviance: 22474  on 319  degrees of freedom
Residual deviance: 16416  on 298  degrees of freedom
AIC: -241004

Number of Fisher Scoring iterations: 4
```

R Output des Modells S7G1 (Abschnitt 4.5)

```
Call:
glm(formula = S_SB1 ~ offset(-log(S_B)) + a + c + b, family = Gamma(link = "log"),
weights = S_J)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-27.39   -6.87   -2.81    1.40   14.88

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
Intercept    5.1743     0.3606   14.35  < 2e-16 ***
a1           0.4247     0.0499    8.50  9.5e-16 ***
a2           1.0821     0.1480    7.31  2.5e-12 ***
a3           1.2132     0.4272    2.84   0.0048 **
c1           0.0675     0.1126    0.60   0.5495
c2           0.1065     0.1271    0.84   0.4029
c3           0.1247     0.1095    1.14   0.2559
c4          -0.0828     0.1280   -0.65   0.5179
c5           0.1367     0.1130    1.21   0.2274
c6           0.0556     0.1137    0.49   0.6251
c7           0.0518     0.1151    0.45   0.6532
c8           0.2329     0.1249    1.87   0.0631 .
c9           0.0659     0.1125    0.59   0.5584
c10          0.1815     0.1251    1.45   0.1479
c11         -0.1028     0.1288   -0.80   0.4255
c12         -0.0250     0.1239   -0.20   0.8401
c13         -0.0676     0.1389   -0.49   0.6271
c14          0.2135     0.1211    1.76   0.0790 .
c15          0.2890     0.1277    2.26   0.0243 *
c16          0.1520     0.1392    1.09   0.2758
c17          0.3499     0.1721    2.03   0.0429 *
c18          0.2148     0.1784    1.20   0.2294
b1          -0.5866     0.3630   -1.62   0.1072
b2          -0.5002     0.3491   -1.43   0.1530
b3          -0.4112     0.3486   -1.18   0.2391
b4          -0.4638     0.3492   -1.33   0.1851
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 31.7)

    Null deviance: 22474  on 319  degrees of freedom
Residual deviance: 16151  on 294  degrees of freedom
AIC: -242858

Number of Fisher Scoring iterations: 2
```

R Output des Modells S6P1 (Abschnitt 4.5)

```
Call:
glm(formula = S_SB1 ~ offset(offs) + a + c, family = poisson,
     weights = (S_B * S_J)/S_D)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-12.27   -2.24   -0.54    1.57   15.67

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
Intercept    3.4618     0.0677   51.16  < 2e-16 ***
a1          -1.3144     0.0308  -42.63  < 2e-16 ***
a2          -2.1518     0.0478  -45.05  < 2e-16 ***
a3          -2.5197     0.0630  -40.02  < 2e-16 ***
c1           0.8437     0.0799   10.55  < 2e-16 ***
c2           0.0461     0.0934    0.49  0.62149
c3           1.2746     0.0759   16.80  < 2e-16 ***
c4           0.3218     0.0880    3.65  0.00026 ***
c5           0.9774     0.0786   12.43  < 2e-16 ***
c6           1.0641     0.0777   13.69  < 2e-16 ***
c7           0.7582     0.0810    9.36  < 2e-16 ***
c8           0.3969     0.0865    4.59  4.5e-06 ***
c9           0.9551     0.0790   12.09  < 2e-16 ***
c10          0.4018     0.0868    4.63  3.7e-06 ***
c11          0.0858     0.0934    0.92  0.35847
c12          0.5148     0.0848    6.07  1.3e-09 ***
c13         -0.1546     0.1001   -1.54  0.12239
c14          0.4316     0.0861    5.01  5.4e-07 ***
c15          0.3372     0.0875    3.85  0.00012 ***
c16          0.1548     0.0928    1.67  0.09531 .
c17         -0.3958     0.1073   -3.69  0.00023 ***
c18         -0.5037     0.1100   -4.58  4.7e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 12004.9  on 319  degrees of freedom
Residual deviance:  4881.6  on 298  degrees of freedom
AIC: Inf

Number of Fisher Scoring iterations: 13
```

R Output des Modells S7P1 (Abschnitt 4.5)

```
Call:
glm(formula = S_SB1 ~ offset(offs) + a + c + b, family = poisson,
    weights = (S_B * S_J)/S_D)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-6.124  -1.731  -0.653   1.001  12.644

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
Intercept    2.2552     0.0836   26.97  < 2e-16 ***
a1          -1.3267     0.0308  -43.03  < 2e-16 ***
a2          -2.2929     0.0478  -47.99  < 2e-16 ***
a3          -2.8015     0.0630  -44.44  < 2e-16 ***
c1           0.8762     0.0799   10.96  < 2e-16 ***
c2           0.0668     0.0934    0.71  0.47477
c3           1.2952     0.0759   17.07  < 2e-16 ***
c4           0.3338     0.0880    3.79  0.00015 ***
c5           0.9981     0.0786   12.70  < 2e-16 ***
c6           1.0882     0.0777   14.01  < 2e-16 ***
c7           0.7907     0.0810    9.77  < 2e-16 ***
c8           0.4175     0.0865    4.83  1.4e-06 ***
c9           0.9671     0.0790   12.25  < 2e-16 ***
c10          0.4133     0.0868    4.76  1.9e-06 ***
c11          0.0842     0.0934    0.90  0.36718
c12          0.5268     0.0848    6.22  5.1e-10 ***
c13         -0.1826     0.1001   -1.82  0.06806 .
c14          0.4437     0.0861    5.15  2.6e-07 ***
c15          0.3612     0.0875    4.13  3.6e-05 ***
c16          0.1249     0.0928    1.35  0.17834
c17         -0.4238     0.1073   -3.95  7.9e-05 ***
c18         -0.5052     0.1100   -4.59  4.4e-06 ***
b1           0.0542     0.0695    0.78  0.43527
b2           1.3476     0.0564   23.89  < 2e-16 ***
b3           1.8542     0.0543   34.18  < 2e-16 ***
b4           1.4306     0.0560   25.56  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 12004.9  on 319  degrees of freedom
Residual deviance:  2056.3  on 294  degrees of freedom
AIC: Inf

Number of Fisher Scoring iterations: 14
```

R Output des Modells V3P1 (Abschnitt 4.5)

```
Call:
glm(formula = V_SB1 ~ offset(offs) + A + B, family = poisson,
    weights = 1/offs)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-10.302   -0.563    2.174    5.372   20.725

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
Intercept -1.03e+01   7.64e-02 -134.48  < 2e-16 ***
A2         1.75e-01   3.76e-02    4.66  3.1e-06 ***
A3         1.37e-01   4.38e-02    3.13  0.00172 **
A4         4.10e-01   4.24e-02    9.66  < 2e-16 ***
A5         2.23e-01   5.12e-02    4.36  1.3e-05 ***
A6         5.19e-01   4.78e-02   10.86  < 2e-16 ***
A7        -3.29e-01   8.51e-02   -3.87  0.00011 ***
A8         7.59e-04   8.27e-02    0.01  0.99267
A9        -1.35e+00   1.40e-01   -9.68  < 2e-16 ***
A10       -7.90e-01   1.68e-01   -4.72  2.4e-06 ***
A11       -6.79e+00   1.51e-01  -45.06  < 2e-16 ***
A12       -8.28e+00   1.53e-01  -53.93  < 2e-16 ***
A13       -7.87e+00   3.35e-01  -23.49  < 2e-16 ***
B11       -2.53e-02   1.52e-01   -0.17  0.86782
B12        9.28e-01   1.47e-01    6.30  3.0e-10 ***
B13        8.11e+00   1.20e-01   67.42  < 2e-16 ***
B14        8.55e+00   8.37e-02  102.13  < 2e-16 ***
B15        8.54e+00   8.02e-02  106.44  < 2e-16 ***
B16        8.63e+00   7.96e-02  108.42  < 2e-16 ***
B17        8.60e+00   7.88e-02  109.10  < 2e-16 ***
B18        8.71e+00   8.10e-02  107.50  < 2e-16 ***
B19        8.79e+00   8.24e-02  106.68  < 2e-16 ***
B20        8.72e+00   9.59e-02   90.93  < 2e-16 ***
B21        2.37e+00   1.63e-01   14.58  < 2e-16 ***
B22        1.19e+00   1.38e-01    8.63  < 2e-16 ***
B23        9.65e-01   1.29e-01    7.49  6.9e-14 ***
B24        5.52e-01   1.25e-01    4.41  1.0e-05 ***
B25        5.93e-01   2.02e-01    2.94  0.00329 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 81336.1  on 183  degrees of freedom
Residual deviance:  6418.1  on 156  degrees of freedom
AIC: Inf

Number of Fisher Scoring iterations: 15
```

Entwicklung der Vuong Teststatistik für die Wahl nicht genesteter Poisson-verteilter Modelle (Abschnitt 6.3.1)

Die Teststatistik zum Vergleich nicht genesteter Poisson-verteilter Modelle hat die Form:

$$\boldsymbol{V} = \frac{\boldsymbol{LR}_k(\hat{\boldsymbol{\lambda}}_k, \hat{\boldsymbol{\alpha}}_k^E,)}{\sqrt{k}\hat{w}_k}$$

mit $k = \sum_{i=1}^m k_i$ und

$$\begin{aligned}
&\boldsymbol{LR}_k(\hat{\boldsymbol{\lambda}}_k^E, \hat{\boldsymbol{\alpha}}_k^E) = L_k^f(\hat{\boldsymbol{\lambda}}_k^E) - L_k^g(\hat{\boldsymbol{\alpha}}_k^E) \\
= &\sum_{i=1}^m \sum_{j=1}^{k_i} \log f(SB_{ij}|X_{ij}; \hat{\boldsymbol{\lambda}}_i^E) - \sum_{i=1}^m \sum_{j=1}^{k_i} \log g(SB_{ij}|Z_{ij}; \hat{\boldsymbol{\alpha}}_i^E) \\
= &\sum_{i=1}^m \sum_{j=1}^{k_i} \frac{u_{ij}B_i}{D_i} \left[\hat{\lambda}_i^E - SB_{ij} \log\left(\hat{\lambda}_i^E\right) + \log\left(SB_{ij}!\right)\right] - \log\left(c_{ij}\right) \\
&-\frac{u_{ij}B_i}{D_i} \left[\hat{\alpha}_i^E - SB_{ij} \log\left(\hat{\alpha}_i^E\right) + \log\left(SB_{ij}!\right)\right] + \log(d_{ij})
\end{aligned}$$

und

$$\hat{w}_k^2 = \frac{1}{k} \sum_{i=1}^m \sum_{j=1}^{k_i} \left[\log\left(\frac{f(SB_{ij}|X_{ij}; \hat{\boldsymbol{\lambda}}^E)}{g(SB_{ij}|Z_{ij}; \hat{\boldsymbol{\alpha}}^E)}\right)\right]^2 - \left[\frac{1}{k} \sum_{i=1}^m \sum_{j=1}^{k_i} \log\left(\frac{f(SB_{ij}|X_{ij}; \hat{\boldsymbol{\lambda}}^E)}{g(SB_{ij}|Z_{ij}; \hat{\boldsymbol{\alpha}}^E}\right)\right]^2$$

wobei $\sum_{i=1}^m k_i = k$ mit

$$\begin{aligned}
&\log\left(\frac{f(SB_{ij}|X_{ij}; \hat{\boldsymbol{\lambda}}^E)}{g(SB_{ij}|Z_{ij}; \hat{\boldsymbol{\alpha}}^E)}\right) \\
= &\frac{u_{ij}B_i}{D_i} \left[\hat{\lambda}_i^E - SB_{ij} \log\left(\hat{\lambda}_i^E + \log\left(SB_{ij}!\right)\right)\right] - \log(c_{ij}) \\
- &\frac{u_{ij}B_i}{D_i} \left[\hat{\alpha}_i^E - SB_{ij} \log\left(\hat{\alpha}_i^E + \log\left(SB_{ij}!\right)\right)\right] + \log(d_{ij}).
\end{aligned}$$

Da in diesem Fall die beiden Modelle eine unterschiedliche Anzahl an geschätzten Parameter haben ist ein Korrekturfaktor notwendig. Schließlich erhält man folgende Teststatistik:

$$\boldsymbol{V} = \frac{\widehat{\boldsymbol{LR}}_k(\hat{\boldsymbol{\lambda}}_k^E, \hat{\boldsymbol{\alpha}}_k^E)}{\sqrt{k}\hat{w}_k} = \frac{\boldsymbol{LR}_k(\hat{\boldsymbol{\lambda}}_k^E, \hat{\boldsymbol{\alpha}}_k^E) - K(p,q)}{\sqrt{k}\hat{w}_k} \tag{8.1}$$

mit

$$K(p,q) = \frac{p}{2}\log(k) - \frac{q}{2}\log(k),$$

wobei p die Anzahl der geschätzten Parameter des Modells f und q die Anzahl der geschätzten Parameter des Modells g ist.

Für Hypothese und Alternative ergibt sich:

$$\boldsymbol{H} : \text{Modelle des S Datensatzes} \quad versus \quad \boldsymbol{K} : \text{Modelle des V Datensatzes}$$

Verwerfe $\boldsymbol{H}$ zum Niveau α, falls $\boldsymbol{V} < z_{\frac{\alpha}{2}}$. Für $z_{\frac{\alpha}{2}} < \boldsymbol{V} < z_{1-\frac{\alpha}{2}}$ trifft der Vuong Test keine Entscheidung, $\boldsymbol{H}$ wird weder verworfen noch nicht verworfen.

Entwicklung der Distribution-Free Teststatistik für die Wahl nicht genesteter Poisson-verteilter Modelle (Abschnitt 6.3.2)

Hier werden nicht genestete Modelle mit dem Distribution-Free Test verglichen. Die Teststatistik für die Poisson-verteilten Modelle sieht wie folgt aus:

$$\boldsymbol{B} = \sum_{i=1}^{m} \sum_{j=1}^{k_i} \boldsymbol{I}_{(0,+\infty)} \left(t_{ij}\right)$$

mit

$$\begin{aligned} t_{ij} &= log\left(f\left(SB_{ij}|X_{ij};\hat{\boldsymbol{\lambda}}^E\right)\right) - \log\left(g\left(SB_{ij}|Z_{ij};\hat{\boldsymbol{\alpha}}^E\right)\right) \\ &= \frac{u_{ij}B_i}{D_i}\left[\hat{\lambda}_i^E - SB_{ij}\log\left(\hat{\lambda}_i^E + \log\left(SB_{ij}!\right)\right)\right] - \log(c_{ij}) \\ &- \frac{u_{ij}B_i}{D_i}\left[\hat{\alpha}_i^E - SB_{ij}\log\left(\hat{\alpha}_i^E + \log\left(SB_{ij}!\right)\right)\right] + \log(d_{ij}). \end{aligned}$$

für $i = 1, ..., m;\ j = 1, ..., k_i$. Auch hier ist durch die unterschiedliche Anzahl der geschätzten Parameter in den Modellen ein Korrekturfaktor notwendig.

$$K(p,q) = \frac{p}{2k}\log(k) - \frac{q}{2k}\log(k)$$

Schließlich erhält man:

$$\boldsymbol{B} = \sum_{i=1}^{m} \sum_{j=1}^{k_i} \boldsymbol{I}_{(0,+\infty)} \left(t_{ij}\right) \tag{8.2}$$

mit

$$t_{ij} = log\left(f\left(SB_{ij}|X_{ij};\hat{\boldsymbol{\mu}}^E,\hat{\nu}\right)\right) - log\left(g\left(SB_{ij}|Z_{ij};\hat{\boldsymbol{\alpha}}^E,\hat{\beta}\right)\right) - K(p,q)$$

für $i = 1, ..., m;\ j = 1, ..., k_i$, wobei p die Anzahl der geschätzten Parameter des Modells f und q die Anzahl der geschätzten Parameter des Modells g ist.

Die Hypothese und die Alternative sind gegeben durch:

$$\boldsymbol{H} : \text{Modelle des S Datensatzes} \quad versus \quad \boldsymbol{K} : \text{Modelle des V Datensatzes}$$

Verwerfe $\boldsymbol{H}$, falls $\boldsymbol{B} < \frac{k}{2}$ mit $k = \sum_{i=1}^{m} k_i$.

R Packet

`dist.vuong`	*Distribution Free and Vuong Test for Gamma or Poisson distribution*

Description

`dist.vuong` compares two nested or nonnested models. The Distribution-Free Test and the Vuong Test for weighted or non weighted Gamma or Poisson distributed models is calculated.

Usage

```
dist.vuong( x , mu.f , mu.g , nu.f , nu.g ,
weight, penalty , distribution = " G ")
```

Arguments

`X`	a numeric vector, the observed data
`mu.f, mu.g`	a numeric vector of the same length than `X`. One wants to compare the expectation for the two models f and g.
`nu.f, nu.g`	1/dispersion for models f and g
`weights`	a numeric vector of length x, if no `weights` are needed: `weights` =1
`penalty`	-(p-q) /2 with p and q number of parameters of the statistical models f and g

Details

`dist.vuong` returns an vector of length 2. The first value is the value of the Distribution-Free Test the second one is the value of the Vuong Test.

References

Vuong, Q. (1989) Likelihood Ratio Tests for Model Selection and nonnested Hypotheses. Ecometrica.

Clarke K. A. (2003) A simple Distribution-Free Test for nonnested Hypotheses. Political Analysis.

Abbildungsverzeichnis

Tabellenverzeichnis

Literaturverzeichnis

Belasco, E. (2002). *Modelling Risk in Fed Cattle Production.* Dissertation, North Carolina State University.

Bickel, P. and K. Doksum (1977). *Mathematical Statistics.* Prentice Hall.

Clarke, K. A. (2001). Testing nonnested models of international relations. *American Journal of Political Science 45*, 724–744.

Clarke, K. A. (2003). Nonparametric model discrimination in international relations. *Political Analysis 47.*

Clarke, K. A. (2007). A simple distribution-free test for nonnested model selection. *Political Analysis 15*, 347–364.

de Jong, P. and G. Heller (2008). *Generalized Linear Models for Insurance Data.* Cambridge University Press.

Fahrmeier, L. and T. Kneib (2007). *Regression.* Springer.

Fahrmeir, L. and R. Künstler (2002). *Statistik.* Springer.

Fahrmeir, L. and G. Tutz (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models.* Springer.

Hu, F. and J. Zidek (2002). The weighted likelihood. *The Canadian Journal of Statistics 30*, 347–371.

Lambert, D. (1992). Zero-inflated-poisson regression with application to defects in manufacturing. *Technometrics, 34(1):1-14.*

Lindsey, J. (1997). *Applying Generalized Linear Models.* Springer.

Mack, T. (2002). *Schadenversicherungsmathematik.* DGVM.

Myers, R. and D. Montgomery (2001). *Generalized Linear Models.* John Wiley and Sons.

Nelder, J. and A. McCullagh (1991). *Generalized Linear Models*. Chapman and Hall.

Vuong, Q. (1989). Likelihood ratio tests for model selection and nonnested hypotheses. *Econometrica 57*, 307–333.