EIDESSTATTLICHE ERKLÄRUNG

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommenen Stellen als solche kenntlich gemacht habe. Das in TUGRAZonline hochgeladene Textdokument ist mit der vorliegenden Masterarbeit identisch.

Datum

Unterschrift

Vorwort

Im Zuge meines Studiums für Finanz- und Versicherungsmathematik an der Technischen Universität Graz ermöglichte mir die Kärntner Landesversicherung ein zweimonatiges beziehungsweise einmonatiges Praktikum im Sommer 2016 und 2017. Dabei durfte ich versicherungsmathematische Daten aufbereiten und auswerten. Da nach jeder Analyse neue Interessen und Fragen aufkamen und viele davon am Ende meines ersten Praktikums noch offen blieben, darf ich diese in Form meiner Diplomarbeit ausarbeiten.

Um von zu Hause aus arbeiten zu können und gleichzeitig den Datenschutz zu gewährleisten, wurden bei den zur Verfügung gestellten Daten sämtliche Polizzennummern und Namen der Versicherungsnehmer verändert beziehungsweise gelöscht. Ziel dieser Arbeit ist die genaue Analyse von Kfz-Daten mit Hilfe von generalisierten linearen Modellen. Anhand der Auswertungen können Prämienkalkulationen sowohl für die Kfz-Haftpflichtversicherung als auch die Kfz-Kaskoversicherung durchgeführt werden.

In dieser Arbeit wird die Vorgehensweise der Berechnung lediglich für die Kfz-Haftpflichtversicherung angeführt, da diese für alle drei Versicherungssparten dieselbe ist. Die detailierten Ergebnisse der Vollkasko- und Teilkaskoversicherung werden aber selbstverständlich ebenfalls angeführt.

Abstract

Es werden die Kfz-Versicherungsdaten der Kärntner Landesversicherung betrachtet. Mit Hilfe von generalisierten linearen Modellen werden jene Merkmale einer Polizze (sowohl fahrzeugspezifische als auch versicherungsnehmerspezifische) bestimmt, welche signifikanten Einfluss auf den Schadenbedarf und somit auf die Versicherungsprämie haben. Dafür werden 3 verschiedene Methoden verwendet. Bei einer Methode wird der Schadenbedarf in die Schadenfrequenz und die durchschnittliche Schadenshöhe aufteilt und jeweils getrennt modelliert. Die anderen zwei Methoden modellieren den Schadenbedarf direkt. Anschließend werden die Ergebnisse miteinander verglichen und analysiert.

The used insurance data is from Kärntner Landesversicherung. Generalised linear models are used to determine those attributes (vehicle specific as well as insuranceholder specific) of an insurance policy which have a significant influence on the claims requirement and therefore on the insurance premium. For that 3 different methods are applied. One of them seperates the claims requirement into the claim frequency and the claim severity and models them segregated. The other two methods determine the claims requirement directly. Afterwards the results of the different methods are being compared and analysed.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Problemstellung und Erklärung des Datensatzes

Das Ziel dieser Arbeit richtet sich auf die Analysierung der vorhandenen Daten einer Versicherung in Österreich um anschließend die Prämienkalkulation genauer gestalten zu können. Dabei wird im Speziellen die Kfz-Versicherung mit deren Versicherungsnehmern und Fahrzeugen betrachtet. Dies inkludiert den Anteil der männlichen und weiblichen Versicherungsnehmer, sowie deren Alter, Wohnort, Ausbildungsabschluss und Ähnliches. Des Weiteren sind die Eigenschaften der Fahrzeuge, welche sich im Bestand befinden, das heißt, Fahrzeugart, Konzern, Antriebsart, Leistung etc. von Interesse. Um die Prämie eines Versicherungsnehmers möglichst genau bestimmen zu können, müssen jene Merkmale identifiziert werden, welche die Wahrscheinlichkeit der Inanspruchnahme einer Versicherungsleistung erhöhen.

Der gegebene Datensatz ist in Abb.1 zu sehen, wobei hier nur die ersten 10 Zeilen grafisch dargestellt werden.

BetreuerNr	yearvec	POL._S	gueltigAb	gueltigBis	VerbandSparte	KundeSex	KundeGebDatum	KundePLZ	AkadTitel
101882	2011	[redacted]	01.03.2010	01.03.2011	71	M	12.02.1961	1130	B/>=1
101882	2012	[redacted]	01.03.2012	01.06.2012	71	M	12.02.1961	1130	B/>=1
101882	2012	[redacted]	01.03.2011	01.03.2012	71	M	12.02.1961	1130	B/>=1
101882	2011	[redacted]	01.03.2011	01.03.2012	71	M	12.02.1961	1130	B/>=1
101924	2012	[redacted]	01.12.2011	18.07.2012	71	M	25.07.1925	9065	A/0
101924	2014	[redacted]	12.05.2014	01.02.2015	71	W	01.01.1900	9081	A/0
101924	2011	[redacted]	01.06.2011	01.06.2012	71	M	15.09.1983	1090	A/0
101924	2013	[redacted]	01.10.2012	01.10.2013	71	W	07.09.1974	1180	B/>=1
101924	2015	[redacted]	23.09.2014	04.05.2015	71	M	17.05.1944	9020	A/0
101924	2015	[redacted]	01.07.2015	03.03.2016	71	M	19.08.1943	9300	A/0

SonstTitel	VNfamilienstand	VNnatio	VNnatjur	vmerk_I3SFR	vmerk_bmsvb	vmerk_STAT	vmerk_LEAS
A/0	unbekannt	Austria	N	A/-6 bis -3	A/-6 bis -3	PKW ohne besond. Bestimmung	N
A/0	unbekannt	Austria	N	A/-6 bis -3	A/-6 bis -3	PKW ohne besond. Bestimmung	N
A/0	unbekannt	Austria	N	A/-6 bis -3	A/-6 bis -3	PKW ohne besond. Bestimmung	N
A/0	unbekannt	Austria	N	A/-6 bis -3	A/-6 bis -3	PKW ohne besond. Bestimmung	N
A/0	unbekannt	Austria	N	A/-6 bis -3	A/-6 bis -3	PKW ohne besond. Bestimmung	N
A/0	unbekannt	Austria	N	C/2 bis 5	C/2 bis 5	PKW ohne besond. Bestimmung	N
A/0	unbekannt	Austria	N	A/-6 bis -3	A/-6 bis -3	PKW ohne besond. Bestimmung	N
A/0	unbekannt	Austria	N	C/2 bis 5	C/2 bis 5	PKW ohne besond. Bestimmung	N
A/0	unbekannt	Austria	N	C/2 bis 5	C/2 bis 5	PKW ohne besond. Bestimmung	N
A/0	unbekannt	Austria	N	A/-6 bis -3	A/-6 bis -3	PKW ohne besond. Bestimmung	N

vmerk_WECHS	vmerk_KHVAR	ERSTZUlassung	HUBRAU	LEISTU	KFZKEY_S	VKBEZL	OVKC1
N	A	23.07.1999	C/ <2000	C/>70<=90kW	PKW	Audi A3	Diesel
N	A	23.07.1999	C/ <2000	C/>70<=90kW	PKW	Audi A3	Diesel
N	A	23.07.1999	C/ <2000	C/>70<=90kW	PKW	Audi A3	Diesel
N	A	23.07.1999	C/ <2000	C/>70<=90kW	PKW	Audi A3	Diesel
N	A	01.01.1900	C/ <2000	B/>50<=70	PKW	Ford Sierra	Diesel
N	A	23.01.2003	C/ <2000	C/>70<=90kW	PKW	Toyota	Diesel
N	A	22.08.2002	B/ <1700	B/>50<=70	PKW	Opel Astra	Diesel
N	A	09.01.2004	A/ <1400	B/>50<=70	PKW	VW GOLF	Benzin
N	A	06.07.2000	D/ >2000	E/>110kW	PKW	Jaguar XKR Coupe	Benzin
N	A	24.06.2005	A/ <1400	B/>50<=70	PKW	Suzuki	Benzin

Konzern	KONZNR_S	JNPR_EUR	vers_jahre	anteilige_JNPR	Schadenanzahl	Schadenanzahl_o_H_S	Wirkschaden	Zahlungen
AUDI	2	232.68	0.16164384	37.61129	0	0	0.00000	0
AUDI	2	250.27	0.25205479	63.08175	0	0	0.00000	0
AUDI	2	237.92	0.16438356	39.11014	0	0	0.00000	0
AUDI	2	237.92	0.83835616	199.46170	0	0	0.00000	0
FORD	23	237.70	0.54520548	129.59534	0	0	0.00000	0
TOYOTA	78	506.20	0.64109589	324.52274	0	0	0.00000	0
OPEL	57	294.96	0.58630137	172.93545	0	0	0.00000	0
VW	84	378.74	0.74794521	283.27677	0	0	0.00000	0
JAGUAR	33	756.58	0.33698630	254.95710	0	0	0.00000	0
SUZUKI	77	279.83	0.50410959	141.06499	0	0	0.00000	0

Kosten	alter_years	jahre_seit_erstzulassung	Zugehoerigkeit
0	50-64	12	unbekannt
0	50-64	13	unbekannt
0	50-64	13	unbekannt
0	50-64	12	unbekannt
0	75-100	113	Makler
0	>100/Firma	11	Makler
0	26-30	9	Makler
0	31-49	9	Makler
0	65-74	15	Makler
0	65-74	10	Makler

Abbildung 1: Struktur des Datensatzes

1.1 Merkmale und deren Ausprägungen

Da nicht alle Merkmalsnamen selbsterklärend sind, werden hier die genauen Bezeichnungen angegeben:

Merkmalsname	Erklärung
VerbandSparte	71 steht für Haftpflicht und 73 für Kasko.
KundeSex	Die Geschlechter sind männlich, weiblich, Firma sowie unbekannt.
AkadTitel	Steht für die Anzahl der akademischen Titel, nicht für die Titel selbst.
SonstTitel	Wie AkadTitel.
VNnatjur	Ja steht für juristische Person, Nein für natürliche.
vmerk_I3SFR	Gibt die Bonus-Malus Stufe an.
vmerk_bmsvb	Gibt die interne Bonus-Malus Stufe an.
vmerk_STAT	Gibt den Verwendungszweck an.
vmerk_LEAS	Steht für geleastes Fahrzeug Ja oder Nein.
vmerk_KHVAR	Steht für Kfz-Haftpflicht Varianten (A,B oder keine Variante).
KFZKEY_S	Steht für den Fahrzeugtyp (PKW, LKW, einspuriges Kfz, ...).
OVKC1	Gibt die Antriebsart an (Diesel, Benzin, ...).
vers_jahre	Steht für die anteiligen vers. Jahre dieser Polizze im gewählten Jahr.
anteilige_JNPR	Gibt die anteilige Jahresnettoprämie an.
Wirkschaden	Ist der bereits ausregulierte Wirkschaden für das jeweilige Jahr.
Zugehoerigkeit	Steht für die Betreuergruppe (hauptberuflicher Mitarbeiter, Makler,..).

Tabelle 1: Merkmalsnamen

Hierbei sei noch zu erwähnen, wie der Bestand zustande kommt. Angenommen man betrachtet eine Polizze mit der Laufzeit von 01.02.2013 bis 01.02.2014 und wählt das Jahr 2014. Dann ist natürlich klar, dass hier nicht die Polizze mit 1 Jahr gewichtet werden darf, sondern nur mit dem Anteil des Jahres, in dem sie auch gültig ist. In unserem Fall wären dies 31 Tage im Jänner. Somit erhält man für die versichterten Jahre $31/365 = 0.0849$ Jahre.

Ein weiterer wichtiger Aspekt ist, dass Verischerungen durch das Gesetz der großen Zahlen gerechtfertigt sind, Quijano und Garrido [13]. Daher ist es von Bedeutung, stetige Merkmalsausprägungen zu diskretisieren, um besser Gruppieren zu können und nicht nur total unterschiedliche Risiken zu haben. Betrachtet man zum Beispiel die Leistung eines Fahrzeuges, so gibt es sehr viele verschiedene Leistungsklassen. Dabei stellt sich die Frage ob es einen gravierenden Unterschied macht ob ein Fahrzeug 60,61 oder 62 kW hat. Da dies nicht der Fall ist und man möglichst viele vergleichbare Fahrzeuge in den Klassen haben möchte, werden diese geclustert. Hier zum Beispiel < 50kW, $\geq 50 < 70$kW, $\geq 70 < 90$kW, $\geq 90 < 110$kW und > 110kW.

1.2 Key Ratios

Bei der Analyse von Versicherungsdaten ist man an bestimmten Key Ratios interessiert, siehe Ohlsson, Johansson [12, Kapitel 1]. Key Ratios sind Verhältnisse zwischen dem Resultat einer Zufallsvariable und einem Volumensmaß (Exposure), $Y = Z/\omega$, wobei Z als die Response und ω als die Exposure bezeichnet wird. Die in unserem Fall wichtigsten Faktoren sind die anteiligen versicherten Jahre in einem gewählten Jahr (wird auch als Bestand bezeichnet), die Anzahl der Schäden in demselben Jahr, sowie die Schadenhöhe. Aus diesen Faktoren kann man jene Key Ratios berechnen, welche von größter Bedeutung sind, nämlich die Schadenfrequenz, der Durchschnittsschaden sowie der Schadenbedarf.

$$\text{Schadenfrequenz} = \frac{\text{Anzahl der Schäden}}{\text{Bestand}}, \tag{1}$$

$$\text{durchschnittliche Schadenhöhe} = \frac{\text{Summe Schadenhöhe}}{\text{Anzahl der Schäden}}, \tag{2}$$

$$\text{Schadenbedarf} = \text{Schadenfrequenz} \cdot \text{durchschnittliche Schadenhöhe}. \tag{3}$$

1.3 Allgemeine Modellannahmen

Es werden hier die Annahmen wie in Ohlsson, Johansson [12, Kapitel 1] gewählt.

1.3.1 Unabhängigkeit der Polizzen

Man betrachte n verschiedene Polizzen. Des Weiteren sei Z_i die Response für die Polizze i. Dann sind $Z_1, ..., Z_n$ unabhängig.

Wenn man diese Annahme genauer überdenkt, fallen einem sofort Beispiele in der Kfz-Versicherung ein, welche diese nicht erfüllen. Ein Beispiel dafür wäre ein Unfall, wobei alle Autos bei derselben Versicherung versichert sind, wodurch die Annahme verletzt ist, allerdings ist der Einfluss nur sehr gering und kann somit vernachlässigt werden.

1.3.2 Unabhängigkeit der Zeit

Man betrachte n disjunkte Zeitintervalle. Des Weiteren sei Z_i die Response für das Zeitintervall i. Dann sind $Z_1, ..., Z_n$ unabhängig.

Auch hier findet man Beispiele in der Praxis, welche diese Annahme nicht vollständig erfüllen, wie zum Beispiel ein Autofahrer, welcher erst kürzlich einen Unfall hatte. In diesem Fall kann man davon ausgehen, dass dieser anschließend einen vorsichtigeren Fahrstil aufweist und somit eine geringere Schadenfrequenz hat. Allerdings zahlt es sich auch hier aus, das Modell durch die Unabhängigkeit zu vereinfachen.

Der Vorteil dieser beiden Annahme liegt in der Unabhängigkeit aller Schäden. Diese treten entweder in unterschiedlichen Polizzen auf oder in unterschiedlichen Zeitintervallen.

Das Ziel eines Versicherers ist es, sein Portfolio in homogene Gruppen zu unterteilen und allen Personen innerhalb derselben Gruppe, dieselbe Prämie zu verrechnen.

1.3.3 Homogenität

Man wähle zwei Polizzen aus der selben Tarifzelle mit dem selben Exposure. Sei weiters Z_i die Response von Polizze i (Schadenanzahl oder Schadenhöhe). Dann folgt, dass Z_1 und Z_2 die selbe Wahrscheinlichkeitsverteilung haben.

Diese Annahme ist in der Realität leider auch nicht exakt erfüllt, aber man möchte homogene Gruppen bilden, innerhalb derer dieselbe Nettoprämie verlangt wird, welche einigermaßen fair ist. Dies kann auch dazu führen, dass es sich hierbei auch um ein und dieselbe Polizze handelt, welche in unterschiedlichen Jahren aufrecht bleibt.

Zum Beispiel hat ein Versicherungsnehmer sein Auto seit Oktober 2014 bei der Versicherung gemeldet. Nun ist diese Polizze unterteilt in jene Polizze von Oktober 2014 bis Oktober 2015, in eine von Oktober 2015 bis Oktober 2016 und eine von Oktober 2016 bis heute, da eine Kfz-Versicherung immer 1 Jahr lang läuft. Befindet sich der Versicherungsnehmer zum Beispiel in der Mitte einer Altersklasse so ändert sich für diesen nichts im Oktober 2015. Dies führt dazu, dass alle Polizzen dieselben Eigenschaften – insbesondere dieselbe Laufzeit – haben, mit Ausnahme des Jahres, in dem sie aufrecht sind. Somit ist es nicht von Bedeutung, wann die jeweilige Polizze startet und endet, sondern nur wie lange diese aufrecht ist.

Somit können durch Homogenität wiederholte Beobachtungen in die statistische Analyse miteinbezogen werden. Dies führt allerdings dazu, dass einer Polizze im Jahr 2010 zugewiesener Schaden der Höhe 500 € gleich behandelt wird wie ein einer Polizze im Jahr 2015 zugewiesener Schaden derselben Höhe. Um dies nun zu vermeiden, werden Schäden mit dem Kraftfahrzeughaftpflicht-Versicherungsleistungspreisindex (KVLPI) beziehungsweise dem Verbraucherpreisindex in der Kaskoversicherung skaliert.

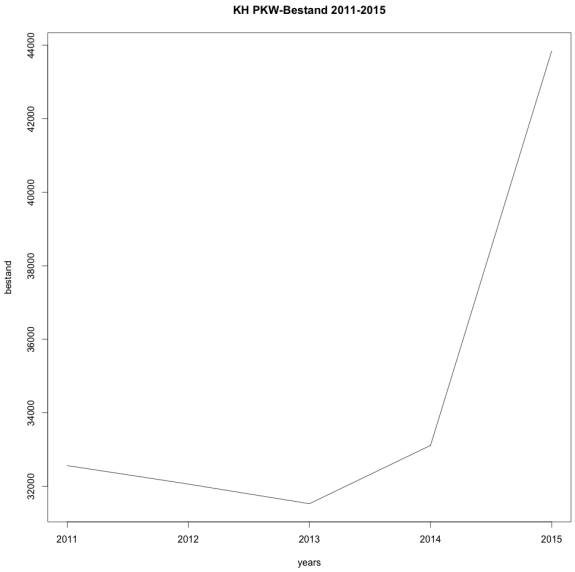
1.4 Erwartungswerte und Varianzen

Da man bei der Prämienkalkulation Erwartungswerte und Varianzen von Key Ratios $Y = Z/\omega$ betrachten muss, werden diese hier, wie auch in Ohlsson, Johansson [12, Kapitel 1] kurz berechnet. Sei zum Beispiel ω die Schadensanzahl, sodass man Z als die Summe von ω Respones $W_1, ...W_\omega$ schreiben kann. Die Annahmen bezüglich der Unabhängigkeit der Polizzen und der Zeit implizieren, dass die W_k's unabhängig sind, da die Schäden von unterschiedlichen Polizzen oder Zeitpunkten stammen. Die dritte Annahme der Homogenität liefert eine gleiche Verteilung, sodass $E[W_k] = \mu$ und $Var[W_k] = \sigma^2$. Somit gilt:

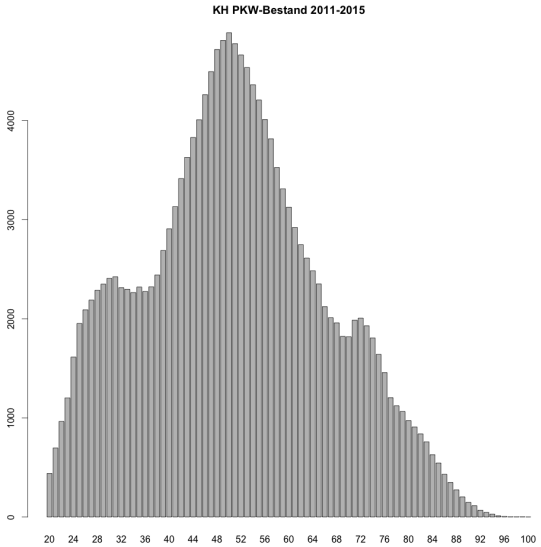
$$E[Z] = E\left[\sum_{i=1}^{\omega} W_k\right] = \omega\mu, \qquad Var[Z] = Var\left[\sum_{i=1}^{\omega} W_k\right] = \omega\sigma^2, \tag{4}$$

$$E[Y] = E[Z/\omega] = \mu, \qquad Var[Y] = Var[Z/\omega] = \sigma^2/\omega. \tag{5}$$

Dieses Ergebnis gilt auch für andere Key Ratios mit Bestand als Exposure, siehe Ohlsson, Johansson [12, Lemma 1.1].

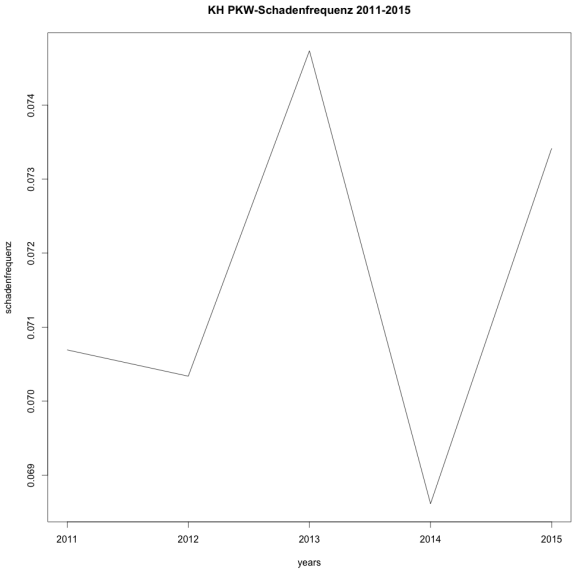


(a) Bestandsverlauf

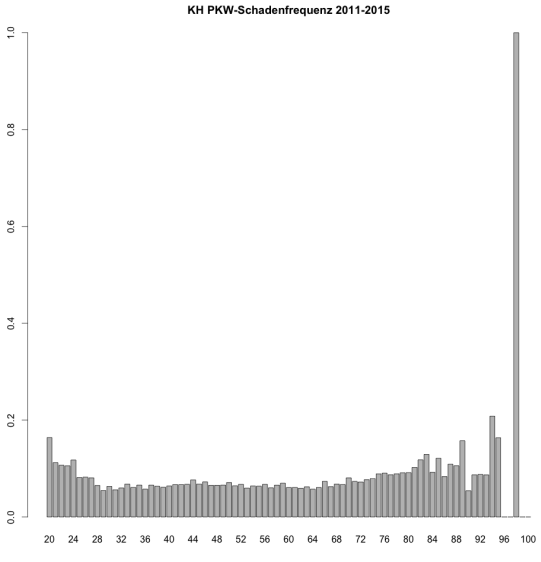


(b) Gesamtbestand je Altersgruppe

Abbildung 2: Bestand 2011-2015

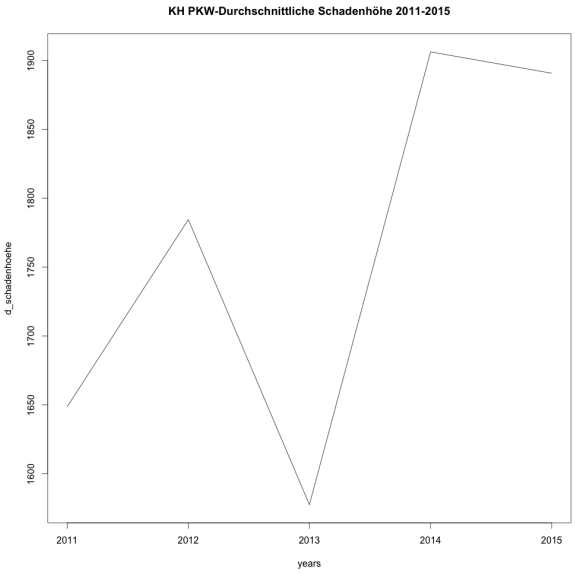


(a) Verlauf Schadenfrequenz

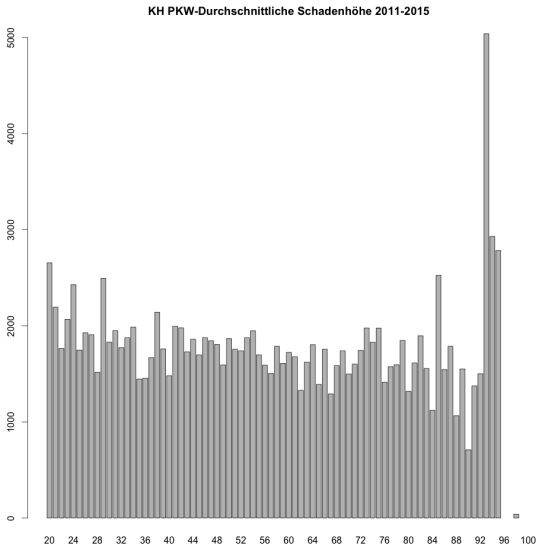


(b) Gesamtschadenfrequenz je Altersgruppe

Abbildung 3: Schadenfrequenz 2011-2015

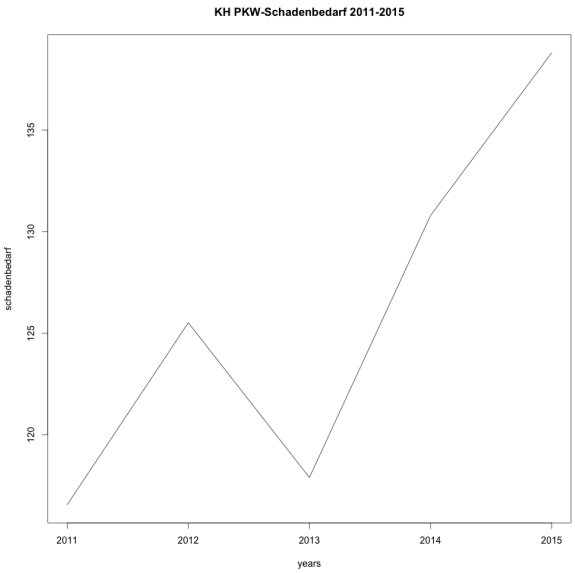


(a) Verlauf durchschnittliche Schadenhöhe

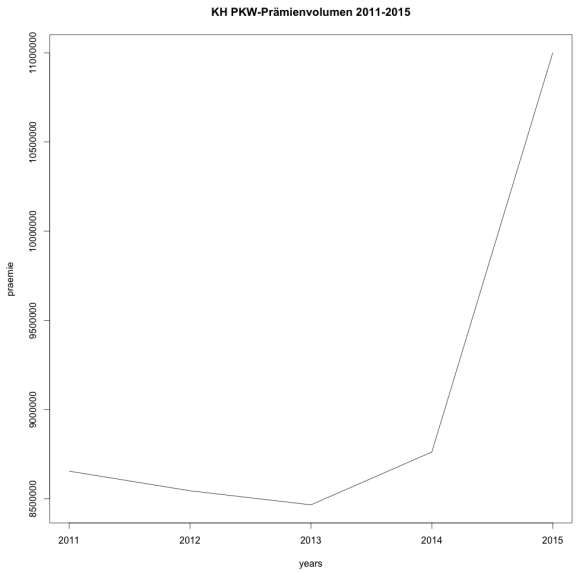


(b) Durchschnittliche Schadenhöhe je Altersgruppe

Abbildung 4: Durchschnittliche Schadenhöhe 2011-2015



(a) Verlauf Schadenbedarf 2011-2015



(b) Verlauf Prämienvolumen 2011-2015

1.5 Clusterverfahren

Bei der Analyse von Kfz-Versicherungsdaten weisen einige zu betrachtende Merkmale eine sehr große Anzahl an Merkmalsausprägungen auf, wie zum Beispiel das Merkmal Fahrzeugkonzern. Hier gibt es weit über hundert verschiedene, welche zu Klassen zusammengefasst werden sollten. Ein anderes Beispiel sind die Postleitzahlen der Kunden, welche man ebenso in Gruppen zusammenfassen sollte.

Um eine solche Klasseneinteilung objektiv gestalten zu können, bedient man sich mathematischer Methoden, den sogenannten Clusterverfahren, siehe Heep-Altiner, Klemmstein [14]. Der Vorteil gegenüber künstlichen Klassen liegt hierbei auch darin, dass Klassen aus einem Clustering robuster gegenüber Schwankungen in den Folgejahren sind und nicht leer sein können.

Allgemein versteht man unter dem Begriff „Clusteranalyse“ ein mathematisches Verfahren, das aus einer heterogenen Menge von Objekten (z.B.: verschiedene PLZ) homogene Teilmengen von Objekten (Regionalklassen) aus der Objektgesamtheit identifiziert, Heep-Altiner, Klemmstein [14, 7.3.1 Vorgehensweise bei einer Clusteranalyse].

Es gibt hierbei zwei unterschiedliche Arten von hierarchischen Clusterverfahren, nämlich die agglomerativen und die divisiblen Verfahren. Die Vorgehensweise bei agglomerativen Verfahren ist Folgende: Zunächst befindet sich jedes Objekt in einer eigenen Klasse und nach und nach werden immer mehr Objekte zu Klassen zusammengefasst.

Bei den divisiblen Verfahren wird umgekehrt vorgegangen. Hier startet man bei einer großen Gesamtklasse, welche alle Objekte enthält und unterteilt diese sukzessive in einzelne Unterklassen.

In weiterer Folge wird nur noch das Ward-Verfahren beschrieben, da dieses in der Kfz-Versicherung das am häufigsten verwendete ist. Dieses gehört zur Gruppe der agglomerativen Clusterverfahren, welche iterativ vorgehen. Das bedeutet, dass man wie zuvor erwähnt, zu Beginn alle Objekte in eigene Klassen gliedert und anschließend jene zwei Klassen zusammenfasst, welche sich am „ähnlichsten“ sind. Diese Vorgehensweise wird so lange wiederholt, bis alle ursprünglichen Objekte einer Klasse zugeteilt wurden.

Diese Vorgehensweise ist bei allen agglomerativen Clusterverfahren gleich. Die Unterscheidung zwischen dem Ward-Verfahren und anderen, wie zum Beispiel „Single Linkage“ oder „Complete Linkage“ findet in der unterschiedlichen Definition der „Ähnlichkeit“ statt. Dieses Maß der Ähnlichkeit, auch Abstandsmaß oder Approximationsmaß genannt kann zum Beispiel die Anzahl von Werten oder die Streuung innerhalb einer Klasse sein.

Zwei Klassen werden immer dann zusammengefasst, wenn der Zuwachs der Streuung bei der Fusion der beiden Klassen minimal ist. Der Grund liegt darin, dass man Merkmalsausprägungen zusammenfassen möchte, welche ein „ähnliches“ Risiko darstellen.

Vorteile des Ward-Verfahrens:

- Tendenz zur Bildung homogener Klassen (ca. gleiche Klassengröße)
- geringe Reaktion auf Ausreißer
- geringe Streuung innerhalb der Klassen
- Optimales Verfahren hinsichtlich der Auswahl einer guten Klassenanzahl

1.5.1 Genaue Funktionsweise des Ward-Verfahrens

Zunächst werden die benötigten Variablen wie in Heep-Altiner, Klemmstein [14, Kapitel 7] definiert, um anschließend das genaue Vorgehen des Ward-Verfahrens zu beschreiben.

Z_{ij}	j-te Beobachtung in der i-ten Klasse,
N_{ij}	Gewichtung der j-ten Beobachtung in der i-ten Klasse,
m_i	Anzahl der Beobachtungen in der i-ten Klasse,
$N_i := \sum_{j=1}^{m_i} N_{ij}$	Summe aller Gewichte in der i-ten Klasse,
$\bar{Z}_i := \sum_{j=1}^{m_i} \frac{N_{ij} Z_{ij}}{N_i}$	Gewichtetes Mittel der Beobachtungen in der i-ten Klasse,
$N := \sum_{i=1}^{k} N_i$	Summe der Gewichte über alle Werte.

Die Fehlerquadratsumme innerhalb einer Klasse ist gegeben durch:

$$d_i^2 := \sum_{j=1}^{m_i} \frac{N_{ij}}{N_i} \times (Z_{ij} - \bar{Z}_i)^2. \tag{6}$$

Die Gesamtvarianz über alle noch vorhandenen k Klassen ist somit die Summe der Varianzen aller Klassen.

$$D_k := \sum_{i=1}^{k} d_i^2. \tag{7}$$

Wie in der Vorgehensweise beschrieben, werden schrittweise jene 2 Klassen zusammengefasst, welche einen minimalen Varianzzuwachs aufweisen (Minimum von $D_{k-l} - D_k$). Das Ward-Verfahren liefert somit für jede gewünschte Klassenanzahl k, jene k Klassen, welche die kleinste Gesamtvarianz aufweisen.

Die Bestimmung einer guten Klassenanzahl ist mit Hilfe der Analyse des Informationsverlustes möglich, wurde aber hier nicht durchgeführt. Stattdessen wurde zu Beginn eine größere Klassenanzahl angenommen. Anschließend wurden mit der in R implementierten Funktion „fit.contrast“ jene Klassen zusammengefasst, welche einen ähnlichen Erwartungswert haben. Die Klassen waren hier im konkreten Fall Fahrzeugkonzerne oder Postleitzahlen der Versicherungsnehmer, wobei die Gewichtung N_{ij} mit 1 festgelegt wurde.

1.6 Multiplikatives Modell

M sei die Anzahl der Merkmale, wobei jedes Merkmal weitere Ausprägungen/Klassen hat. Die Anzahl der Ausprägungen des Merkmals i werden mit m_i bezeichnet. Zunächst werden, der Einfachheit halber, wie in Ohlsson, Johansson [12, Kapitel 1], nur 2 Merkmale betrachtet. In einer Tarifzelle (i, j) wobei i und j die jeweilige Ausprägung des jeweiligen Merkmals definieren, sei der Exposure w_{ij} (zum Beispiel: Bestand) und die Response Z_{ij} (zum Beispiel: Schadenanzahl). Dies führt zur Key Ratio $Y_{ij} = Z_{ij}/w_{ij}$ (in diesem Fall: Schadenfrequenz).

Der Erwartungswert von Y_{ij} sei μ_{ij}, falls $w_{ij} = 1$. Das multiplikative Modell hat somit folgende Darstellung:

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

Die Menge aller γ_{1i} für $i = 1, ..., m_1$ seien die zum Merkmal 1 gehörigen Parameter und die Menge aller γ_{2j} für $j = 1, ..., m_2$ seien die zum Merkmal 2 gehörigen Parameter. γ_0 sei hierbei ein Basiswert.

Das hier dargestellte Modell an sich ist überparametrisiert, da der Erwartungswert μ_{ij} gleich bleibt, wenn man zum Beispiel γ_{1i} mit 10 multipliziert und dafür γ_{2j} wiederum durch 10 dividiert. Um Eindeutigkeit zu gewährleisten wird eine Ausprägung jedes Merkmals als Referenzklasse gewählt. Bevorzugt wird hierbei jene mit dem höchsten Exposure/Bestand. Der Einfachheit halber seien das hier die jeweils ersten Ausprägungen der beiden Merkmale. Somit erhalten wir $\gamma_{11} = \gamma_{21} = 1$. Nun kann γ_0 als der Basiswert interpretiert werden. Für alle anderen Ausprägungen erhält man somit den relativen Unterschied zur Basiszelle. Zum Beispiel sei $\gamma_{12} = 1.3$. Dann ist er Erwartungswert der Zelle $(2, 1)$ um 30 % höher als in Zelle $(1, 1)$.

1.6.1 Warum sollte man überhaupt ein multiplikatives Modell verwenden?

Der große Vorteil des multiplikativen Modells ist, dass es keine Interaktionen zwischen 2 Merkmalen gibt. Das Alter sei Merkmal 1, der Bezirk sei Merkmal 2 und die Key Ratio die Schadenfrequenz. Das Verhältnis der Schadenfrequenz zwischen 2 Altersklassen ist somit für jeden Bezirk dasselbe.

Zum Beispiel sei die Schadenfrequenz in Bezirk 1 der 17-20-Jährigen um 30 % höher als die der 30-50-Jährigen. Dann gilt dies auch für jeden anderen Bezirk. Seien die Schadenfrequenzen der Altersklasse 30-50 Jahre für Bezirk 1 10 % und für Bezirk 2 30 %.

Im multiplikativen Modell erhält man bei einem 20 %-igen Anstieg der Schadenfrequenz für die 30-50-Jährigen somit die Schadenfrequenzen von 12 % und 36 %, was eine durchaus plausible Annahme darstellt.

Wählt man hingegen ein additives Modell $\mu_{ij} = \gamma_0 + \gamma_{1i} + \gamma_{2j}$, so würde man für den 1. Bezirk dieselbe Schadenfrequenz von 12 % erhalten, allerdings für den 2. Bezirk nur 32 %. Somit wäre die Veränderung im Vergleich zum 1. Bezirk unverhältnismäßig klein. Daher ist die Annahme des multiplikativen Modells durchaus gerechtfertigt.

Das allgemeine Modell hat somit folgende Darstellung:

$$\mu_{i_1, i_2, ..., i_M} = \gamma_0 \gamma_{1 i_1} \gamma_{2 i_2} \cdots \gamma_{M i_M}.$$

2 Generalisierte lineare Modelle

Ziel der Analyse der Daten ist es, bestimmen zu können, welche Merkmale einen signifikanten Einfluss auf eine der Key Ratios Y haben, siehe Ohlsson, Johansson [12, Kapitel 2]. Das bedeutet, wie verändert sich die abhängige Variable Y in einem multiplen linearen Regressionsmodell bei Änderung der beschreibenden Variablen x (explanatory variables). Das Problem der linearen Regression liegt in den Voraussetzungen, welche die Kfz-Daten im Allgemeinen nicht erfüllen.

Lineare Regressionsmodelle setzen normalverteilte Zufallsfehler voraus, allerdings folgt die Anzahl der Schäden einer diskreten Verteilung, welche keine negativen Werte annehmen kann. Die zweite Voraussetzung ist, dass bei linearen Modellen angenommen wird, dass der Erwartungswert eine lineare Funktion der beschreibenden Variablen ist, hingegen für die Kfz-Analyse ein muliplikatives Modell geeigneter ist.

2.1 Exponentialfamilie

Generalisierte lineare Modelle folgen einer allgemeinen Klasse von Verteilungen, welche sowohl bekannte diskrete als auch stetige Verteilungen inkludiert, wie zum Beispiel die Normal- oder Poissonverteilung. Aufgrund der Modellannahmen sind die Variablen $Y_1, ..., Y_n = \mathbf{Y}$ unabhängig und folgen einer Exponentialfamilie mit Dichtefunktion, McCullagh, Nelder [10, Kapitel 2]

$$f_Y(y;\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}, \tag{8}$$

für spezielle Funktionen $a(\cdot), b(\cdot)$ und $c(\cdot)$. Falls ϕ bekannt ist, so handelt es sich um ein Modell der Exponentialfamilie mit Paramter θ.

Daher gilt zum Beispiel für die Normalverteilung, Ohlsson, Johansson [12, Kapitel 2]

$$f_y(y;\theta,\phi) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{-(y-\mu)^2}{2\sigma^2}\right\} \tag{9}$$

$$= \exp\left\{\frac{(y\mu - \mu^2/2)}{\sigma^2} - \frac{1}{2}\left(\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2)\right)\right\}, \tag{10}$$

sodass $\theta = \mu$, $\phi = \sigma^2$ und $a(\phi) = \phi$, $b(\theta) = \frac{\theta^2}{2}$ und $c(y,\phi) = -\frac{1}{2}\left\{\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2)\right\}$.

Sei nun $l(\theta,\phi;y) = \log f_Y(y;\theta,\phi)$ die log-Likelihood Funktion wie in McCullagh, Nelder [10, Kapitel 2] als Funktion von θ und ϕ, wobei y gegeben ist. Für die log-Likelihood Funktion von (8) folgt

$$l(\theta;y) = \frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi). \tag{11}$$

Der Erwartungswert und die Varianz können mit folgenden, unter den Regularitätsbedingungen gültigen (sind für die Exponentialfamilie erfüllt) Gleichungen für die Scorefunktion und die Informationszahl ermittelt werden

$$E\left[\frac{\partial l(\theta;Y)}{\partial\theta}\right] = 0, \qquad E\left[-\frac{\partial^2 l(\theta;Y)}{\partial\theta^2}\right] = E\left[\left(\frac{\partial l(\theta;Y)}{\partial\theta}\right)^2\right]. \tag{12}$$

Durch Bildung der ersten und zweiten Ableitung von $l(\theta; y)$ nach θ erhält man

$$\frac{\partial l(\theta; y)}{\partial \theta} = \frac{y - b'(\theta)}{a(\phi)} \quad \text{und} \tag{13}$$

$$\frac{\partial^2 l(\theta; y)}{\partial \theta^2} = \frac{-b''(\theta)}{a(\phi)}, \tag{14}$$

wobei die Ableitungen nach θ hier mit b', b'' dargestellt werden.

Bildung des Erwatungswertes der ersten Gleichung liefert unter Verwendung von (12)

$$E\left[\frac{\partial l(\theta; Y)}{\partial \theta}\right] = E\left[\frac{Y - b'(\theta)}{a(\phi)}\right] = \frac{\mu - b'(\theta)}{a(\phi)} = 0 \Leftrightarrow \mu = b'(\theta), \tag{15}$$

sodass $E[Y] = \mu = b'(\theta)$.

Des Weiteren gilt mit (12)

$$0 = -\frac{b''(\theta)}{a(\phi)} + E\left[\left(\frac{Y - b'(\theta)}{a(\phi)}\right)^2\right] = -\frac{b''(\theta)}{a(\phi)} + \frac{Var(Y)}{a^2(\phi)} \Leftrightarrow b''(\theta) = \frac{Var(Y)}{a(\phi)}, \tag{16}$$

sodass $Var(Y) = b''(\theta)a(\phi)$, das Produkt zweier Funktionen, wobei $b''(\theta)$ nur vom kanonischen Parameter und daher vom Mittelwert abhängt und als Varianzfunktion bezeichnet wird und $a(\phi)$ nur von ϕ abhängt und daher unabhängig von θ ist.

Die Varianzfunktion wird standardmäßig mit $V(\mu)$ bezeichnet und $a(\phi)$ ist üblicherweise von der Form $a(\phi) = \frac{\phi}{\omega}$. Der Dispersionsparameter ϕ, häufig auch σ^2 ist den Beobachtungen gegenüber konstant, während ω je nach Beobachtung variiert und sozusagen ein Gewicht darstellt.

Die Parameter und Funktionen der wichtigsten Verteilungen sind hier in einer Tabelle zusammengefasst.

	Normal	Poisson	Gamma
Notation	$N(\mu, \sigma^2)$	$Poi(\mu)$	$\Gamma(\alpha, \tau)$
Bildbereich von y	$(-\infty, \infty)$	$\mathbb{N}$	$(0, \infty)$
Dispersionsparameter: ϕ	$\phi = \sigma^2$	1	$\phi = \frac{1}{\tau}$
Kumulantenfunktion: $b(\theta)$	$\frac{\theta^2}{2}$	$\exp(\theta)$	$-\log(-\theta)$
$c(y; \phi)$	$-\frac{1}{2}\left(\frac{y^2}{\phi} + \log(2\pi\phi)\right)$	$-\log(y!)$	$\tau\log(\tau y) - \log(y) - \log(\Gamma(\tau))$
$\mu(\theta) = E[Y; \theta]$	θ	$\exp(\theta)$	$\frac{-1}{\theta}$
Kanonische Linkfkt.: $\theta(\mu)$	Identität	Logarithmus	Reziproke
Varianzfunktion: $V(\mu)$	1	μ	μ^2

Tabelle 2: Verteilungen der Exponentialfamilie

2.2 Linkfunktion

Wie bereits zu Beginn erwähnt, gibt es bei generalisierten linearen Modellen zwei wesentliche Unterschiede zum linearen Regressionsmodell. Hier wird der zweite Unterschied, nämlich die Funktion des Erwartungswertes, die sogenannte Linkfunktion betrachtet.

Man beginnt, der Einfachheit halber, wie in Ohlsson, Johansson [12, Kapitel 2] mit 2 Merkmalen, wobei eines 2 und das andere 3 Ausprägungen hat. Sei μ_{ij} der Erwartungswert der Key Ratio in Zelle (i, j). Lineare Modelle setzen eine additive Struktur des Erwartungswertes voraus,

$$\mu_{ij} = \gamma_0 + \gamma_{1i} + \gamma_{2j}. \tag{17}$$

Dieses Modell ist überparametrisiert, daher wird eine Restriktion hinzugefügt, welche erfordert, dass die Parameter der Basiszelle 0 sind. Man nehme an, $(1, 1)$ sei die Basiszelle, dann sei $\gamma_{11} = \gamma_{21} = 0$, sodass gilt $\mu_{11} = \gamma_0$. Die anderen Parameter messen somit den Abstand zur Basiszelle.

Als nächstes wird das Modell in Listenform dargestellt. Dazu werden die Parameter neu benannt,

$$\beta_1 = \gamma_0; \quad \beta_2 = \gamma_{12}; \quad \beta_3 = \gamma_{22}; \quad \beta_4 = \gamma_{23}. \tag{18}$$

Die Erwartungswerte der einzelnen Zellen befinden sich in der linken Tabelle. Anschließend verwendet man Dummy Variablen x_{ij}, wobei $x_{ij} = 1$, falls β_j in μ_i enthalten ist, ansonsten gilt $x_{ij} = 0$. Man erhält für die linke Tabelle eine neue Darstellung (rechte Tabelle).

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

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

Tabelle 3: Listendarstellung additives Modell

Mit Hilfe dieser Dummy Variablen erhält man für den Erwartungswert des linearen Modells

$$\mu_i = \sum_{j=1}^{4} x_{ij}\beta_j \quad i = 1, 2, 3, 4, 5, 6. \tag{19}$$

Dasselbe Ergebnis in der Matrixdarstellung $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}$ mit

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \\ \mu_5 \\ \mu_6 \end{pmatrix} \qquad \boldsymbol{X} = \begin{pmatrix} x_{11} & x_{12} & x_{13} & x_{14} \\ x_{21} & x_{22} & x_{23} & x_{24} \\ x_{31} & x_{32} & x_{33} & x_{34} \\ x_{41} & x_{42} & x_{43} & x_{44} \\ x_{51} & x_{52} & x_{53} & x_{54} \\ x_{61} & x_{62} & x_{63} & x_{64} \end{pmatrix} \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix}, \tag{20}$$

wobei $\boldsymbol{X}$ die sogenannte Designmatrix/Modellmatrix ist.

Wir haben aber bereits gesehen, Ohlsson, Johansson [12, Kapitel 1] dass es für Kfz-Daten wesentlich sinnvoller ist, ein multiplikatives Modell zu verwenden. Mit einfacher Logarithmus-Bildung erhält man wiederum ein additives Modell

$$\log(\mu_{ij}) = \log(\gamma_0) + \log(\gamma_{1i}) + \log(\gamma_{2j}). \tag{21}$$

In diesem Fall seien $\gamma_{11} = \gamma_{21} = 1$. Ein analoges Vorgehen liefert somit für jede Zelle i

$$\eta_i = \log(\mu_i) = \sum_{j=1}^{4} x_{ij}\beta_j \quad i = 1, 2, 3, 4, 5, 6. \tag{22}$$

Diese Gleichung hat dieselbe Struktur wie zuvor, bis auf den Logarithmus auf der linken Seite. Für GLMs ist es erlaubt eine monotone Funktion des Erwartungswertes auf der linken Seite zu haben.

Eine Verallgemeinerung für die Response Y_i, welche von r Merkmalen $x_1, x_2, ..., x_r$ beeinflusst ist und für jedes $i = 1, ..., n$ einer Zelle zugeordnet wird, liefert

$$\eta_i = \sum_{j=1}^{r} x_{ij}\beta_j \quad i = 1, 2, ..., n, \tag{23}$$

wobei x_{ij} den Wert des Merkmals x_j an der i-ten Stelle beschreibt. In einem linearen Modell gilt $\mu_i = \eta_i$, wie wir zuvor gesehen haben. Allgemein gilt $g(\mu_i) = \eta_i$, wobei $g(\cdot)$ eine monotone, differenzierbare Funktion sein muss. Diese Funktion $g(\cdot)$ heißt Linkfunktion, da sie den Erwartungswert mit der linearen Struktur verbindet. Für unser multiplikatives Modell gilt somit, dass die Linkfunktion der Logarithmus ist

$$g(\mu_i) = \log(\mu_i).$$

2.2.1 Kanonische Linkfunktion

Wie bereits bei der Exponentialfamilie gesehen, werden unterschiedliche Parametrisierungen von GLMs verwendet. Diese Parameter sind eindeutige Funktionen voneinander, sodass (siehe Ohlsson, Johansson [12, Kapitel 2])

$$\theta \xrightarrow{b'(\cdot)} \mu \xrightarrow{g(\cdot)} \nu. \tag{24}$$

Durch die Wahl von $g(\cdot)$ als die Inverse von $b'(\cdot)$, sodass $\theta = \eta$, wird $g(\cdot)$ die kanonische Linkfunktion genannt. $b(\cdot)$ und damit auch $b'(\cdot)$ sind durch die Struktur der Zufallskomponenten und die eindeutige Wahl der Varianzfunktion bestimmt und $g(\cdot)$ durch die Modellierung des Erwartungswertes.

Betrachtet man, wie Ohlsson, Johansson [12, Kapitel 2] zum Beispiel die Normalverteilung so gilt $\mu = b'(\theta) = \theta$, sodass die kanonische Linkfunktion der Identität entspricht. Bei der Poisson-Verteilung gilt $\mu = b'(\theta) = e^{\theta}$, sodass die kanonische Linkfunktion der Logarithmus ist. Als letztes Beispiel wähle man die Gamma-Verteilung mit $\mu = b'(\theta) = -1/\theta$, so folgt für die kanonische Linkfunktion die Inverse Funktion.

2.3 Offset

In manchen Fällen ist ein Teil des Erwartungswertes μ bereits im Vorhinein bekannt. In unserem Fall ist dies der Kfz-Bestand, welcher als Offset in die GLM-Analyse miteinbezogen wird. Bei der Verwendung eines multiplikativen Modells und der Verwendung der Variable u für den bekannten Faktor, sowie $z_i = \log(u_i)$, erhält man

$$\eta_i = z_i + \sum_{j=1}^{r} x_{ij}\beta_j \quad i = 1, 2, ..., n. \tag{25}$$

2.4 Fisher Scoring Algorithmus

2.4.1 Maximum Likelihood Schätzung

Die allgemeine Vorgehensweise bei der Suche des Maximum Likelihood Schätzers der $\boldsymbol{\beta}$-Paramter in einem GLM funktioniert wie in Ohlsson, Johansson [12, Kapitel 2]. Die n Beobachtungen $Y_1, ..., Y_n$ folgen der Verteilung einer Exponentialfamilie mit $Y_i \sim Exp(\theta_i, \omega_i, \phi)$ mit $a(\phi) = \frac{\phi}{\omega}$ und auf Grund der Unabhängigkeit folgt für die log-Likelihood Funktion von $\theta = (\theta_1, ..., \theta_n)^T$

$$l(\boldsymbol{\theta}; \phi, \boldsymbol{y}) = \frac{1}{\phi}\sum_i \omega_i(y_i\theta_i - b(\theta_i)) + \sum_i c(y_i, \phi, \omega_i). \tag{26}$$

Da $\frac{1}{\phi}$ für die Maximierung nach $\boldsymbol{\theta}$ irrelevant ist, wird es hier weggelassen. Um die log-Likelihood Funktion als eine Funktion von $\boldsymbol{\beta}$ zu erhalten, werden die Relation $\mu_i = b'(\theta_i)$ sowie die Linkfunktion $g(\mu_i) = \eta_i = \sum_j x_{ij}\beta_j$ herangezogen. Die Ableitung von l nach β_j ergibt wegen der Kettenregel folgende Gleichung:

$$\frac{\partial l}{\partial \beta_j} = \sum_i \frac{\partial l}{\partial \theta_i}\frac{\partial \theta_i}{\partial \beta_j} = \frac{1}{\phi}\sum_i (\omega_i y_i - \omega_i b'(\theta_i))\frac{\partial \theta_i}{\partial \beta_j} \tag{27}$$

$$= \frac{1}{\phi}\sum_i (\omega_i y_i - \omega_i b'(\theta_i))\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j} \tag{28}$$

Da $\mu_i = b'(\theta_i)$, gilt $\frac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i)$ und $\frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{b''(\theta_i)}$, wobei $b''(\theta_i) = V(\mu_i)$. Des Weiteren sei $\frac{\partial \mu_i}{\partial \eta_i} = \left[\frac{\partial \eta_i}{\partial \mu_i}\right]^{-1} = \frac{1}{g'(\mu_i)}$ und mit $\eta_i = \sum_j x_{ij}\beta_j$ erhält man $\frac{\partial \eta_i}{\partial \beta_j} = x_{ij}$.

Fügt man nun alle Ergebnisse in die obige Gleichung ein so erhält man die sogenannte Score-Funktion

$$\frac{\partial l}{\partial \beta_j} = \frac{1}{\phi}\sum_{i=1}^{n} \omega_i \frac{y_i - \mu_i}{V(\mu_i)g'(\mu_i)}x_{ij}. \tag{29}$$

Um das Maximum bestimmen zu können werden alle r partiellen Ableitungen 0 gesetzt, was zu anschließenden ML Gleichungen führt

$$\frac{1}{\phi}\sum_{i=1}^{n} \omega_i \frac{y_i - \mu_i}{V(\mu_i)g'(\mu_i)}x_{ij} = 0, \qquad j = 1, 2, 3, ..., r. \tag{30}$$

Sofort lässt sich vermuten, dass die Lösung $\mu_i = y_i$ ist. Dies ist allerdings nur im saturierten/vollen Modell der Fall, bei dem die Anzahl der Parameter gleich der Anzahl der

Beobachtungen ist. Da dieses Modell eher uninteressant ist, muss man die zusätzlichen Bedingungen an die μ_i's heranziehen, nämlich $\mu_i = \mu_i(\boldsymbol{\beta})$, sowie

$$\mu_i = g^{-1}(\eta_i) = g^{-1}\left(\sum_j x_{ij}\beta_j\right). \tag{31}$$

Da eine Matrixdarstellung oft einfacher zu verwenden ist, siehe McCulloch, Searle, Neuhaus [1, Kapitel 5], führt man die Diagonalmatrizen $\boldsymbol{W}$ und $\boldsymbol{\Delta}$ ein,

$$\tilde{w_i} = \frac{w_i}{V(\mu_i)g'(\mu_i)^2}, \qquad d_i = g'(\mu_i),$$

sodass $\boldsymbol{W} = \text{diag}(\tilde{w_i}; i = 1, ..., n)$, $\quad \boldsymbol{\Delta} = \text{diag}(d_i; i = 1, ..., n)$,.

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = \frac{1}{\phi}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{\Delta}(\boldsymbol{y} - \boldsymbol{\mu}). \tag{32}$$

Die ML-Gleichungen (30) können somit in Matrixform dargestellt werden

$$\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{\Delta}\boldsymbol{y} = \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{\Delta}\boldsymbol{\mu}, \tag{33}$$

wobei $\boldsymbol{X}$ die Designmatrix ist und $\boldsymbol{W}$, $\boldsymbol{\Delta}$ und $\boldsymbol{\mu}$ das unbekannte $\boldsymbol{\beta}$ enthalten und meist nichtlineare Funktionen in $\boldsymbol{\beta}$ und somit nicht analystisch lösbar sind.

2.4.2 Fisher Scoring Algorithmus

Um eine Lösung der ML Gleichungen (30) für $\boldsymbol{\beta}$ zu erhalten wird eine iterative „Weighted Least Square Methode“ wie in McCullogh, Searle, Neuhaus [1, Kapitel 5] verwendet. Ein numerisches Verfahren ist der Fisher-Scoring-Algorithmus. Die Iterationsschritte sind wie folgt:

$$\boldsymbol{\theta}^{[n+1]} = \boldsymbol{\theta}^{[n]} + \boldsymbol{I}(\boldsymbol{\theta}^{[n]})^{-1}\frac{\partial l}{\partial \boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{[n]}}, \tag{34}$$

wobei $[n]$ die Iterationszahl angibt, $\boldsymbol{I}(\boldsymbol{\theta}^{[n]})$ für die Fisher-Information und $\boldsymbol{\theta}$ für den gesamten Paramtervektor steht.

Da die Fisher-Information dem negativen Erwartungswert der zweiten Ableitung der log-Likelihood Gleichungen entspricht, wird dieser im nächsten Schritt, mit Hilfe der Produktregel, berechnet,

$$\frac{\partial^2 l}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T} = -\frac{1}{\phi}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{\Delta}\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}^T} + \frac{1}{\phi}\boldsymbol{X}^T\frac{\partial \boldsymbol{W}\boldsymbol{\Delta}}{\partial \boldsymbol{\beta}^T}(\boldsymbol{y} - \boldsymbol{\mu}), \tag{35}$$

$$\boldsymbol{I} := E\left[-\frac{\partial^2 l}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T}\right] = \frac{1}{\phi}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{\Delta}\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}^T} \tag{36}$$

$$= \frac{1}{\phi}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{\Delta}\boldsymbol{\Delta}^{-1}\boldsymbol{X} \tag{37}$$

$$= \frac{1}{\phi}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}, \tag{38}$$

da

$$\frac{\partial \mu_i}{\partial \boldsymbol{\beta}} = \frac{\partial \mu_i}{\partial g(\mu_i)} \frac{\partial g(\mu_i)}{\partial \boldsymbol{\beta}} = \left(\frac{\partial g(\mu_i)}{\partial \mu_i} \right)^{-1} \frac{\partial g(g^{-1}(\boldsymbol{x_i}^T \boldsymbol{\beta}))}{\partial \boldsymbol{\beta}} = \left(\frac{\partial g(\mu_i)}{\partial \mu_i} \right)^{-1} \boldsymbol{x_i}^T. \tag{39}$$

Setzt man diese Ergebnisse in den Fisher Scoring Algorithmus ein erhält man

$$\boldsymbol{\beta}^{[n+1]} = \boldsymbol{\beta}^{[n]} + (\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{\Delta} (\boldsymbol{y} - \boldsymbol{\mu}). \tag{40}$$

Wichtig ist hierbei, dass $\boldsymbol{W}$, $\boldsymbol{\Delta}$ und $\boldsymbol{\mu}$ im Iterationsschritt $[n]$ mit $\boldsymbol{\beta}^{[n]}$ berechnet werden. Mit Hilfe der Definition von $\boldsymbol{z} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\Delta}(\boldsymbol{y} - \boldsymbol{\mu})$ kann obige Gleichung in die Form einer Iterativen Weighted Least Squares Notation gebracht werden.

Ein alternativer Ansatz, siehe Ohlsson, Johansson [12, Kapitel 3], um die Fisher-Information zu bestimmen, verwendet dieselben Relationen die bereits bei der Exponentialfamilie benutzt wurden. $\frac{\partial \mu_i}{\partial \eta_i} = \frac{1}{g'(\mu_i)}$ und $\frac{\partial \eta_i}{\partial \beta_j} = x_{ij}$. Für $j = 1, 2, ..., r$; $k = 1, 2, ..., r$ folgt somit

$$\frac{\partial^2 l}{\partial \beta_j \beta_k} = \sum_i \frac{w_i}{\phi} \frac{\partial}{\partial \mu_i} \left[\frac{y_i - \mu_i}{v(\mu_i) g'(\mu_i)} \right] x_{ij} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_k} \tag{41}$$

$$= \sum_i \frac{w_i}{\phi} \frac{\partial}{\partial \mu_i} \left[\frac{y_i - \mu_i}{v(\mu_i) g'(\mu_i)} \right] x_{ij} \frac{1}{g'(\mu_i)} x_{ik} \tag{42}$$

$$= \sum_i \frac{w_i}{\phi} \frac{\partial}{\partial \mu_i} \left[\frac{y_i - \mu_i}{v(\mu_i)(g'(\mu_i))^2} \right] x_{ij} x_{ik} \tag{43}$$

$$= \sum_i \frac{w_i}{\phi} \frac{-v(\mu_i)(g'(\mu_i))^2 - (y_i - \mu_i)[v'(\mu_i)(g'(\mu_i))^2 + 2g'(\mu_i) g''(\mu_i) v(\mu_i)]}{[v(\mu_i)(g'(\mu_i))^2]^2} x_{ij} x_{ik} \tag{44}$$

$$= \sum_i \frac{w_i}{\phi} \frac{-v(\mu_i) g'(\mu_i) - (y_i - \mu_i)[v'(\mu_i) g'(\mu_i) + 2g''(\mu_i) v(\mu_i)]}{(v(\mu_i))^2 (g'(\mu_i))^3} x_{ij} x_{ik} \tag{45}$$

$$= \sum_i \frac{1}{\phi} \underbrace{\frac{w_i}{v(\mu_i)(g'(\mu_i))^2} \left[\frac{-v(\mu_i) g'(\mu_i)}{v(\mu_i) g'(\mu_i)} - (y_i - \mu_i) \frac{[v'(\mu_i) g'(\mu_i) + 2g''(\mu_i) v(\mu_i)]}{v(\mu_i) g'(\mu_i)} \right]}_{-a_i} x_{ij} x_{ik} \tag{46}$$

$$= -\frac{1}{\phi} \sum_i x_{ij} a_i x_{ik}. \tag{47}$$

Sei nun $\boldsymbol{A} = \text{diag}(a_i; i = 1, ..., n)$, so kann (47) in Matrixform als $\boldsymbol{H} = -\frac{1}{\phi} \boldsymbol{X}^T \boldsymbol{A} \boldsymbol{X}$ dargestellt werden. Da der Erwartungswert von Y_i gleich μ_i ist folgt für den Erwartungswert von $\boldsymbol{A}$, dass der hintere Teil von a_i wegfällt und

$$E[\boldsymbol{A}] = \boldsymbol{W} = \text{diag}(\tilde{w}_i; i = 1, ..., n), \quad \text{mit } \tilde{w}_i = \frac{w_i}{v(\mu_i)(g'(\mu_i))^2}. \tag{48}$$

Für die Fisher-Information $\boldsymbol{I}$ ergibt sich nun

$$\boldsymbol{I} = -E[\boldsymbol{H}] = \frac{1}{\phi} \boldsymbol{X}^T E[\boldsymbol{A}] \boldsymbol{X} = \frac{1}{\phi} \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X}. \tag{49}$$

Der Vorteil dieser Darstellung ist, dass sie einerseits besser nachzurechnen ist und andererseits eine Aussage über das Maximum getroffen werden kann, Ohlsson, Johansson [12,

Kapitel 3]. Es könnte sich hierbei auch um ein lokales Maximum handeln. Für kanonische Linkfunktionen ist dies aber nicht der Fall und man erhält sehr wohl ein globales Maximum, da $\boldsymbol{A} = \boldsymbol{W}$. Dies liegt daran, dass für kanonische Linkfunktionen die Gleichungen $g'(\mu) = \frac{1}{v(\mu)}$ sowie $g''(\mu) = -\frac{v'(\mu)}{(v(\mu))^2}$ gelten.

Setzt man diese beiden Gleichungen in a_i ein, so erhält man

$$a_i = \frac{w_i}{v(\mu_i)(g'(\mu_i))^2}\left[1 + (y_i - \mu_i)\frac{[v'(\mu_i)\frac{1}{v(\mu_i)} - \frac{v'(\mu)}{(v(\mu))^2}v(\mu_i)]}{v(\mu_i)g'(\mu_i)}\right] = \frac{w_i}{v(\mu_i)(g'(\mu_i))^2}. \tag{50}$$

Daher folgt, dass $\boldsymbol{A}$ positiv definit ist, und bei passender Wahl der Merkmale, kann verhindert werden, dass $\boldsymbol{X}$ linear abhängige Spalten hat. Folglich ist $\boldsymbol{H} = -\frac{1}{\phi}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}$ negativ definit und die log-Likelihood Funktion l ist konkav. Somit geben die ML-Gleichungen tatsächlich ein Maximum.

2.4.3 Asymptotische Eigenschaften der MLs

Die ML-Schätzer sind asymptotisch normalverteilt und erwartungstreu mit Kovarianzmatrix gleich der Inversen der Fisher-Information $\boldsymbol{I}$, siehe Ohlsson, Johansson [12, Kapitel 3],

$$\hat{\boldsymbol{\beta}} \overset{d}{\approx} N(\boldsymbol{\beta}; \boldsymbol{I}^{-1}). \tag{51}$$

Um dies zu beweisen, wird ein multivariater Zentraler Grenzwertsatz verwendet. Seien $\boldsymbol{S}_n = (S_{n1}, ..., S_{nr})^T$ Zufallsvektoren mit $S_{nj} = \sum_{i=1}^n X_{ij}$, wobei die Vektoren $\boldsymbol{X}_i = (X_{i1}, ..., X_{ir}); \quad i = 1, 2, ...$ unabhängig sind. Die einzelnen Komponenten von $(X_{i1}, ..., X_{ir})$ müssen hingegen nicht unabhängig sein. Man nehme an, dass der Erwartungswert der X_{ij} gleich 0 ist für alle i, j.

Des Weiteren werden symmetrische, nicht negativ definite Kovarianzmatrizen $\boldsymbol{V}_n$ mit $v_{jk} = Cov(S_{nj}, S_{nk}) = \sum_{i=1}^n Cov(X_{ij}, X_{jk})$ definiert. Daher kann für solche Matrizen auch die Wurzel berechnet werden, sodass $\boldsymbol{V}_n = \boldsymbol{V}_n^{\frac{1}{2}}\boldsymbol{V}_n^{\frac{1}{2}}$. Falls $\boldsymbol{V}_n$ positiv definit ist, so existiert auch die Inverse von $\boldsymbol{V}_n^{\frac{1}{2}}$, nämlich $\boldsymbol{V}_n^{-\frac{1}{2}}$.

Bezeichne mit $N(\mathbf{0}; \mathbb{1})$ eine multivariate Normalverteilung mit Mittelwertvektor $\mathbf{0}$, welcher nur Nullen beinhaltet und Einheitsmatrix $\mathbb{1}$, welche in der Hauptdiagonale Einsen und sonst nur Nullen enthält. Der multivariate zentrale Grenzwertsatz besagt, dass

$$\boldsymbol{V}_n^{-\frac{1}{2}}\boldsymbol{S}_n \overset{d}{\approx} N(\mathbf{0}; \mathbb{1}), \tag{52}$$

für n sehr groß und $\boldsymbol{S}_n$ als Spaltenvektor.

Nun gilt es diesen Satz auf unseren Fall anzuwenden. Definiere dazu, wie Ohlsson, Johansson [12, Kapitel 3] die Funktionen $s_j(\boldsymbol{y}, \boldsymbol{\beta})$ wie folgt:

$$s_j(\boldsymbol{y}, \boldsymbol{\beta}) = \frac{1}{\phi}\sum_{i=1}^n \omega_i \frac{y_i - \mu_i}{v(\mu_i)g'(\mu_i)}x_{ij}; \quad j = 1, 2, ..., r, \tag{53}$$

wobei n die Anzahl der Beobachtungen ist. Den ML-Schätzer $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, ..., \hat{\beta}_r)$ erhält man mit (30) und daher ergibt sich

$$s_j(\boldsymbol{y}, \hat{\boldsymbol{\beta}}) = 0; \quad j = 1, ..., r. \tag{54}$$

Sei $\boldsymbol{\beta}^0 = (\beta_1^0, ..., \beta_r^0)$ der wahre Wert von $\boldsymbol{\beta}$ und es gilt, dass jedes $s_j(\boldsymbol{Y}, \boldsymbol{\beta}^0)$ eine Summe unabhängiger Zufallsvariablen mit Mittelwert 0 ist, da die Y_i's unabhängig sind. Mit der Likelihood Funktion L, der log-Likelihood Funktion $l = \log(L)$ und der Beziehung $\frac{\partial L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} / L(\boldsymbol{\beta}) = \frac{\partial \log(L(\boldsymbol{\beta}))}{\partial \boldsymbol{\beta}} = \frac{\partial l}{\partial \boldsymbol{\beta}}$ erhält man schließlich

$$\boldsymbol{I}(\boldsymbol{\beta}) = -\int \frac{\partial^2 \log(L(\boldsymbol{\beta}))}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} L(\boldsymbol{\beta}) dy = -\int \frac{\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} L(\boldsymbol{\beta}) - \frac{\partial L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \frac{\partial L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}}{(L(\boldsymbol{\beta}))^2} L(\boldsymbol{\beta}) dy \tag{55}$$

$$= \underbrace{\int \frac{\partial^2 L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} dy}_{\frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \int L(\boldsymbol{\beta}) = 0} + \int \underbrace{\frac{\partial l}{\partial \boldsymbol{\beta}}}_{s(\boldsymbol{\beta})} \underbrace{\frac{\partial l}{\partial \boldsymbol{\beta}^T}}_{s(\boldsymbol{\beta}^T)} L(\boldsymbol{\beta}) dy \tag{56}$$

$$= E[s(\boldsymbol{\beta}) s(\boldsymbol{\beta}^T)] = Cov(s(\boldsymbol{\beta})), \quad \text{da } E[s(\boldsymbol{\beta})] = 0. \tag{57}$$

Da die Matrix $\boldsymbol{V}_n$ mit den Einträgen $v_{jk} = Cov(s_j(\boldsymbol{Y}, \boldsymbol{\beta}^0), s_k(\boldsymbol{Y}, \boldsymbol{\beta}^0))$, wegen vorheriger Gleichung, der Fisher-Information $\boldsymbol{I}$ entspricht, folgt mit dem zentralen Grenzwertsatz

$$\boldsymbol{I}^{-\frac{1}{2}} \boldsymbol{s}(\boldsymbol{y}, \boldsymbol{\beta}) \overset{d}{\approx} N(\boldsymbol{0}; \mathbb{1}), \tag{58}$$

wobei $\boldsymbol{s}(\boldsymbol{y}, \boldsymbol{\beta})$ den Spaltenvektor mit Elementen $s_j(\boldsymbol{y}, \boldsymbol{\beta}); j = 1, ..., r$ darstellt.

Da man aber an der asymptotischen Verteilung von $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, ..., \hat{\beta}_r)$ interessiert ist, geht man, wie Ohlsson, Johansson [12, Kapitel 3] folgendermaßen vor:

Man bildet eine Approximation erster Ordnung von $s_j(\boldsymbol{y}, \boldsymbol{\beta})$ als eine Funktion von $\boldsymbol{\beta} = (\beta_1, ..., \beta_r)$ um $\boldsymbol{\beta}^0 = (\beta_1^0, ..., \beta_r^0)$.

$$s_j(\boldsymbol{y}, \boldsymbol{\beta}) \approx s_j(\boldsymbol{y}, \boldsymbol{\beta}^0) + \sum_{k=1}^{r} \left(\frac{\partial}{\partial \beta_k} s_j(\boldsymbol{y}, \boldsymbol{\beta}^0) \right) (\beta_k - \beta_k^0) \tag{59}$$

Mit (41) ergibt sich, dass

$$\frac{\partial}{\partial \beta_k} s_j(\boldsymbol{y}, \boldsymbol{\beta^0}) = -\sum_{i=1}^{n} x_{ij} a_i x_{ik} \tag{60}$$

Wie bereits im alternativen Ansatz zur Bestimmung der Fisher-Information, erhält man, bei Verwendung einer kanonischen Linkfunktion (log-Link für Poisson), dass $a_i = \tilde{\omega}_i$, sodass

$$\frac{\partial}{\partial \beta_k} s_j(\boldsymbol{y}, \boldsymbol{\beta^0}) = -\sum_{i=1}^{n} x_{ij} d_i x_{ik} = -\boldsymbol{I}_{jk}. \tag{61}$$

$\boldsymbol{I}_{jk}$ bezeichnet hier die j-te Zeile und k-te Spalte der Fisher-Informationsmatrix $\boldsymbol{I}$. Setzt man diese Abschätzung nun in (59) ein so folgt

$$s_j(\boldsymbol{y}, \boldsymbol{\beta}) \approx s_j(\boldsymbol{y}, \boldsymbol{\beta}^0) - \sum_{k=1}^{r} \boldsymbol{I}_{jk} (\beta_k - \beta_k^0). \tag{62}$$

Im nächsten Schritt wird $\boldsymbol{\beta}$ durch dessen Schätzer $\hat{\boldsymbol{\beta}}$ ersetzt und zusammen mit (54) ergibt sich

$$\sum_{k=1}^{r} \boldsymbol{I}_{jk} (\hat{\beta}_k - \beta_k^0) \approx s_j(\boldsymbol{y}, \boldsymbol{\beta}^0). \tag{63}$$

Dies ist äquivalent zu $\boldsymbol{I}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) \approx \boldsymbol{s}(\boldsymbol{y}, \boldsymbol{\beta}^0)$ und es gilt mit (58), dass $\boldsymbol{I}^{\frac{1}{2}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) \approx \boldsymbol{I}^{-\frac{1}{2}}\boldsymbol{s}(\boldsymbol{y}, \boldsymbol{\beta}^0) \approx N(\mathbf{0}; \mathbb{1})$ für n groß. Mit Hilfe von Umformungen und der Wahl von $\boldsymbol{\beta} = \boldsymbol{\beta}^0$ erhält man schlussendlich

$$\boldsymbol{I}^{\frac{1}{2}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \approx N(\mathbf{0}; \mathbb{1}) \Leftrightarrow \frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{\frac{1}{\boldsymbol{I}^{\frac{1}{2}}}} \approx N(\mathbf{0}; \mathbb{1}) \Rightarrow \hat{\boldsymbol{\beta}} \approx N(\boldsymbol{\beta}, \boldsymbol{I}^{-1}). \tag{64}$$

3 Nettoprämienkalkulation

Die Nettoprämie entspricht den gesamten erwarteten Kosten aller Schäden je Versicherungsnehmer, welche innerhalb einer Versicherungsperiode anfallen, siehe Mihaela [11]. Die Nettoprämie ist somit der durchschnittliche Schaden pro versichertem Jahr,

$$S = \sum_{i=1}^{N} Z_i, \tag{65}$$

wobei N die Schadensanzahl und Z_i die Schadenshöhen beschreiben. Das Ziel dieser Arbeit ist es nun, diese Nettoprämie von Versicherungsnehmern, für welche die zu Beginn angesprochenen Merkmale und deren Ausprägungen zutreffen, zu bestimmen.

Es werden nun 3 verschiedene Methoden vorgestellt, anhand derer die Nettoprämie bestimmt werden kann.

3.1 Getrennte Modellierung

Der Erwartungswert des Gesamtschadens eines Versicherungsnehmers wird durch Schadenfrequenz und durchschnittliche Schadenhöhe seiner Polizze beschrieben. Folglich ist die Nettoprämie dieser Polizze das Produkt aus der Schadenfrequenz und der durchschnittlichen Schadenhöhe, unter der Annahme, dass die beiden unabhängig sind, siehe Ohlsson, Johansson [12, Kapitel 1]. Hier entsprechen $E[N]$, $E[Z_i]$ und $E[S]$ der Schadenfrequenz, der durchschnittlichen Schadenhöhe und der Nettoprämie,

$$E\left[\sum_{i=1}^{N} Z_i\right] = E[N] \cdot E[Z_i]. \tag{66}$$

Somit werden nur die Schadenfrequenz, welche, wie im nächsten Unterpunkt erklärt, der relativen Poisson Verteilung folgt und die durchschnittliche Schadenhöhe jeweils getrennt modelliert und anschließend miteinander multipliziert.

Da die Schadensanzahl beziehungsweise Schadenfrequenz zuerst modelliert wird kann anschließend die durchschnittliche Schadenshöhe geschätzt werden, da auf die Schadensanzahl bedingt werden kann.

Bei den zwei weiteren Methoden, welche anschließend vorgestellt werden, wird der Erwartungswert des Gesamtschadens je versichertem Jahr direkt modelliert. Die Vorteile der getrennten Modellierung sind laut Ohlsson, Johansson [12, Kapitel 2] folgende:

1.) Die Modellierung der Schadenfrequenz ist wesentlich stabiler als jene der durchschnittlichen Schadenhöhe. Grund dafür ist der wesentlich größere Datensatz, da bei der Modellierung der durchschnittlichen Schadenhöhe die Gamma Verteilung verwendet wird und nur Daten verwendet werden dürfen, welche einen echt positiven Schaden aufweisen.

2.) Eine differenzierte Analyse gibt Auskunft über die Schadenfrequenz und die durchschnittliche Schadenhöhe eines Merkmals. So kann es zum Beispiel sein, dass bei einem Merkmal alle Ausprägungen dieselbe Schadenfrequenz haben, aber unterschiedliche durchschnittliche Schadenhöhen, was sich klarerweise auf die Prämie auswirkt, allerdings von entscheidender Bedeutung für den Versicherer sein kann. Diese Information geht bei der direkten Analyse natürlich verloren.

3.1.1 Spezialfall des multiplikativen Poisson-Modells für die Schadenfrequenz

Man beginnt wieder, wie Ohlsson, Johansson [12, Kapitel 2] mit dem vereinfachten multiplikativen Modell wie zuvor:

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

Wir wissen bereits, dass die passende Linkfunktion der Logarithmus ist.

Seien X_i die Schadensanzahl, ω_i die Laufzeit und μ_i der Erwartungswert bei $\omega_i = 1$. Dann gilt $E(X_i) = \omega_i \mu_i$ und X_i folgt einer Poisson Verteilung mit Dichtefunktion

$$f_{X_i}(x_i; \mu_i) = e^{-\omega_i \mu_i} \frac{(\omega_i \mu_i)^{x_i}}{x_i!}, \quad x_i = 0, 1, 2, ... \tag{67}$$

Die Dichtefunktion für y_i, sodass $\omega_i y_i$ eine nichtnegative Zahl ist, ist

$$f_{Y_i}(y_i; \mu_i) = P(Y_i = y_i) = P(X_i = \omega_i y_i) = e^{-\omega_i \mu_i} \frac{(\omega_i \mu_i)^{\omega_i y_i}}{(\omega_i y_i)!} \tag{68}$$

$$= \exp\{\omega_i [y_i \log(\mu_i) - \mu_i] + c(y_i, \omega_i\}, \tag{69}$$

wobei $c(y_i, \omega_i) = \omega_i y_i \log(\omega_i) - \log(\omega_i y_i!)$. Die Verteilung der Schadenfrequenz $Y_i = X_i/\omega_i$ nennt man relative Poisson Verteilung. Durch Reparametrisierung von $\theta_i = \log(\mu_i)$ sieht man, dass es sich hier um ein Modell der Exponentialfamile handelt. Mit $\phi = 1$ und Kumulantenfunktion $b(\theta_i) = e^{\theta_i}$ gilt

$$f_{Y_i}(y_i; \theta_i) = \exp\{\omega_i (y_i \theta_i - e^{\theta_i}) + c(y_i, \omega_i)\}, \tag{70}$$

wobei wegen $\mu_i > 0$, $-\infty < \theta_i < \infty$ gilt.

Anschließend wird die log-Likelihood Funktion gebildet und nach den γ-Parametern abgeleitet, was zu einem Gleichungssystem führt. Durch Einsetzen und aufgrund der Unabhängigkeit der Polizzen folgt für die log-Likelihood Funktion

$$l = \sum_i \sum_j \omega_{ij} \{y_{ij} [\log(\gamma_0) + \log(\gamma_{1i}) + \log(\gamma_{2j})] - \gamma_0 \gamma_{1i} \gamma_{2j}\} + c, \tag{71}$$

wobei c unabhängig von den γ-Paramtern ist. Ableiten nach den γ-Parametern liefert ein Gleichungssystem, welches die gewünschten Schätzer liefert und mit Hilfe von numerischen Methoden, welche in R bereits implementiert sind, gelöst werden kann.

$$\gamma_0 = \frac{\sum_i \sum_j \omega_{ij} y_{ij}}{\sum_i \sum_j \omega_{ij} \gamma_{1i} \gamma_{2j}}, \tag{72}$$

$$\gamma_{1i} = \frac{\sum_j \omega_{ij} y_{ij}}{\gamma_0 \sum_j \omega_{ij} \gamma_{2j}}; \quad i = 2, ..., m_1, \tag{73}$$

$$\gamma_{2j} = \frac{\sum_i \omega_{ij} y_{ij}}{\gamma_0 \sum_i \omega_{ij} \gamma_{1i}} \quad j = 2, ..., m_2. \tag{74}$$

3.1.2 Speziallfall des multiplikativen Gamma Modells für die Duchschnittsschadenhöhe

Da wir nun die Schadensanzahl beziehungsweise Schadenfrequenz bereits modelliert haben können wir die durchschnittliche Schadenshöhe schätzen. Daher betrachte man, Ohlsson, Johansson [12, Kapitel 2] die Key Ratio $Y = X/\omega$, wobei X den Gesamtschaden bezeichne und ω die Schadensanzahl. Die Wahl der Verteilung von X ist nun nicht mehr so klar wie zuvor. Man möchte aber auf jeden Fall eine Verteilung, welche nur positive Werte annimmt und rechtsschief ist, wodurch die Normalverteilung nicht mehr gewählt werden kann.

Standardmäßig Ohlsson, Johansson [12, Kapitel 2] wird die Gamma Verteilung gewählt, da auch diese eine Verteilung der Exponentialfamilie ist. Es gelte nun, dass die Kosten jedes Schadens Gamma verteilt sind mit $\omega = 1$.

Die Dichtefunktion einer $\Gamma(\alpha, \tau)$ verteilten Zufallsvariable habe folgende Darstellung

$$f(x) = \frac{\tau^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\tau x}; \quad x > 0, \tag{75}$$

mit Indexparameter $\alpha > 0$ und Scale Parameter $\tau > 0$. Der Erwartungswert einer Gamma verteilten Zufallsvariable $Z \sim \Gamma(\alpha, \tau)$, ist $\frac{\alpha}{\tau}$ und die Varianz ist $\frac{\alpha}{\tau^2}$. Des Weiteren gilt für ω unabhängige gammaverteilte Zufallsvariablen Z_i: $X = Z_1 + ... + Z_\omega \sim \Gamma(\omega\alpha, \tau)$. Somit folgt für $Y = X/\omega$

$$f_Y(y) = \omega f_X(\omega y). \tag{76}$$

und daher $Y \sim \Gamma(\omega\alpha, \omega\tau)$ mit $E[Y] = \frac{\alpha}{\tau}$. Durch Reparametrisieung von $\mu = \frac{\alpha}{\tau}$ und $\phi = \frac{1}{\alpha}$ auf dem Paramterraum $\mu > 0$ und $\phi > 0$ erhält man die Darstellung

$$f_Y(y) = f_Y(y; \mu, \phi) = \frac{1}{\Gamma(\omega/\phi)} \left(\frac{\omega}{\mu\phi}\right)^{\omega/\phi} y^{(\omega/\phi)-1} e^{-\omega y/(\mu\phi)} \tag{77}$$

$$= \exp\left\{\frac{-y/\mu - \log(\mu)}{\phi/\omega} + c(y, \phi, \omega)\right\}, \quad y > 0, \tag{78}$$

wobei $c(y, \phi, \omega) = \log(\omega y/\phi)\omega/\phi - \log(y) - \log(\Gamma(\omega/\phi))$. Es folgt $E[Y] = \frac{\omega\alpha}{\omega\tau} = \mu$ und $Var[Y] = \frac{\omega\alpha}{\omega^2\tau^2} = \frac{\phi\mu^2}{\omega}$.

Um nun zu Zeigen, dass die Gamma-Verteilung der Exponentialfamilie angehört, wähle man $\theta = -\frac{1}{\mu}$, wodurch der neue Parameter Werte in der offenen Menge $\theta < 0$ annimmt. Mit $b(\theta_i) = -\log(-\theta_i)$ und der nachfolgenden Darstellung der Dichtefunktion erhält man die gewünschte Darstellung der Dichte der Exponentialfamilie

$$f_{Y_i}(y_i; \theta_i, \phi) = \exp\left\{\frac{y_i\theta_i + \log(-\theta_i)}{\phi/\omega_i} + c(y_i, \phi, \omega_i)\right\}. \tag{79}$$

Da sowohl Erwartungswert als auch Varianz bekannt sind, kann der Variationskoeffizient betrachtet werden.

$$CV = \frac{Var(Y_i)}{(E[Y_i])^2} = \frac{\phi}{\omega_i}. \tag{80}$$

Nun ist ersichtlich, dass für Zellen mit gleichem Exposure ω_i der Variationskoeffizient konstant bleibt, was so viel bedeutet, wie dass die Standardabweichung proportional zum Mittelwert ist. Wie bereits bei der Erklärung „Warum sollte man überhaupt ein multiplikatives Modell verwenden?“ macht das durchaus Sinn.

Man kann sich hierfür ein kurzes Beispiel überlgen. Sei der Mittelwert einer Tarifzelle 50 und die Standardabweichung 10 und der Mittelwert einer anderen Tarifzelle 200. Hierbei würde man sich dann eine Standardabweichung von 40 anstatt von 10 erwarten, was durch die Proportionalität gegeben ist.

Die ML-Gleichungen können mit Hilfe von (30) und der Annahme eines multiplikativen Modells berechnet werden, da $g(\mu_i) = \log(\mu_i)$ und somit $g'(\mu_i) = \frac{1}{\mu_i}$ gilt. Zusätzlich benötigt man, dass die Gamma-Verteilung ein Spezialfall der Tweedie-Familie ist (wird später in Kapitel 3.2 genauer beschrieben) und deshalb für die Varianzfunktion $v(\mu_i) = \mu_i^2$ gilt

$$\frac{1}{\phi}\sum_i \omega_i \frac{y_i - \mu_i}{v(\mu_i)g'(\mu_i)} x_{ij} = \frac{1}{\phi}\sum_i \omega_i \frac{y_i - \mu_i}{\mu_i^2 \frac{1}{\mu_i}} x_{ij} = 0 \Leftrightarrow \sum_i \omega_i \frac{y_i}{\mu_i} x_{ij} = \sum_i \omega_i x_{ij}. \tag{81}$$

Verwendet man wieder den vereinfachten Fall mit nur 2 Merkmalen und $\mu_{ij} = \gamma_0\gamma_{1i}\gamma_{2j}$ so erhält man das Gleichungssystem

$$\sum_i \sum_j \frac{\omega_{ij} y_{ij}}{\gamma_0\gamma_{1i}\gamma_{2j}} = \sum_i \sum_j \omega_{ij}, \tag{82}$$

$$\sum_j \frac{\omega_{ij} y_{ij}}{\gamma_0\gamma_{1i}\gamma_{2j}} = \sum_j \omega_{ij} \quad i = 2, ..., m_1, \tag{83}$$

$$\sum_i \frac{\omega_{ij} y_{ij}}{\gamma_0\gamma_{1i}\gamma_{2j}} = \sum_i \omega_{ij} \quad j = 2, ..., m_2. \tag{84}$$

$$\tag{85}$$

3.1.3 Schätzer Nettoprämie

Wie im Paper von Quijano, Garrido [13] nehme man an, dass in der Schadenfrequenzanalyse k_N signifikante Ausprägungen übrig bleiben. Sei dazu $\boldsymbol{\beta^N}$ der dazugehörige Vektor der Regressionskoeffizienten. Des Weiteren nehme man an, dass das Modell der Durchschnittsschadenhöhen k_Y signifikante Ausprägungen beinhaltet mit dem Koeffizientenvektor $\boldsymbol{\beta_Y}$.

Da beide Modelle nicht zwangsweise dieselben Merkmale beinhalten müssen und somit auch nicht dieselben Ausprägungen, müssen zunächst Adaptierungen vorgenommen werden, um die Schätzer anschließend miteinander multiplizieren zu können.

Man nehme an, dass das Modell für die Schadenfrequenz sowie das Modell für die Durchschnittsschadenhöhen k gemeinsame Ausprägungen haben. Definiere deren gemeinsame Designmatrix $\boldsymbol{X_*}$, wobei die ersten k Spalten zu den gemeinsamen Ausprägungen gehören, die nächsten $(k_N - k)$ Spalten entsprechen den Ausprägungen des Schadenfrequenz Modells und die letzten $(k_Y - k)$ Spalten entsprechen denen des Durschnittsschadenhöhen Modells.

Anschließend definiere den neuen Koeffizientenvektor des Schadenfrequenz Modells $\boldsymbol{\beta^N_*} := (\boldsymbol{\beta^N_1}, \boldsymbol{\beta^N_2}, \mathbf{0})$, wobei $\boldsymbol{\beta^N_1}$ ein Vektor der Dimension k mit den gemeinsamen Koeffizienten des Vektors $\boldsymbol{\beta^N}$ ist. $\boldsymbol{\beta^N_2}$ ist jener $(k_N - k)$-dimensionale Vektor mit Werten aus $\boldsymbol{\beta^N}$, welche nur im Schadenfrequenz Modell vorkommen und $\mathbf{0}$ ist ein Nullvektor der Dimension $(k_Y - k)$.

Analoges Vorgehen für das Durchschnittsschadenhöhen Modell liefert den Vektor $\boldsymbol{\beta^Y_*} := (\boldsymbol{\beta^Y_1}, \mathbf{0}, \boldsymbol{\beta^Y_2})$ mit dem k-dimensionalen Vektor $\boldsymbol{\beta^Y_1}$, dem $(k_N - k)$-dimensionalen Nullvektor und dem Vektor $\boldsymbol{\beta^Y_2}$ der Dimension $(k_Y - k)$.

Sei S_i^* die Responsevariable der i-ten Klasse mit $\boldsymbol{X_i^*}$ die i-te Zeile von $\boldsymbol{X_*}$ und dem Logarithmus als Linkfunktion so gilt

$$E[S_i^*] := \mu_i^* = \exp\{\boldsymbol{X_i^*}(\boldsymbol{\beta_*^N} + \boldsymbol{\beta_*^Y})\}. \tag{86}$$

S_i^* ist zusammengesetzt Poisson-Gamma verteilt mit den Paramtern N_i^*, α und τ_i, wobei der Erwartungswert der Schadenfrequenz sowie der Erwartungswert der Durchschnittsschadenhöhe der i-ten Klasse, wie in Quijano, Garrido [13] gegeben sind durch

$$E[N_i] := \mu_{N_i^*} = \exp\{\boldsymbol{X_i^*}\boldsymbol{\beta_*^N}\} \tag{87}$$

$$E[Y_i] := \mu_{Y_i^*} = \frac{\alpha}{\tau_i} = \exp\{\boldsymbol{X_i^*}\boldsymbol{\beta_*^{Y_i}}\}. \tag{88}$$

3.2 Direkte Modellierung mit Hilfe der Tweedie-Familie

Verteilungen der Tweedie-Familie sind Verteilungen der Exponentialfamilie welche durch die Varianzfunktion definiert sind, siehe Quijano, Garrido [13]. Eine Verteilung aus der Exponentialfamilie nennt man Verteilung der Tweedie-Familie, wenn der Bereich der Varianzfunktion V auf $(0, \infty)$ gegeben ist und für ein $p \in \mathbb{R}$

$$V(\mu) = \mu^p. \tag{89}$$

Wird $p < 0$ gewählt, so werden durch die Tweedie-Familie Verteilungen auf ganz $\mathbb{R}$ charakterisiert. Es existiert keine Exponentialfamilie mit $p \in (0, 1)$ in der Varianzfunktion. Wird $p > 1$ gewählt so werden Verteilungen auf $(0, \infty)$ charakterisiert. Somit sind diese Verteilungen für uns interessant. Die nachfolgende Tabelle, siehe Ohlsson, Johansson [12, Kapitel 2] zeigt ein paar bekannte Verteilungen, welche der Tweedie-Familie angehören. Hier betrachten wir im Speziellen $p \in (1, 2)$.

p	Typ	Name	Key Ratio
0	stetig	Normal	-
1	diskret	Poisson	Schadenfrequenz
(1,2)	mixed, nicht negativ	mixed Poisson-Gamma	Nettoprämie
2	stetig, positiv	Gamma	durchschnittliche Schadenshöhe
3	stetig, positiv	Inverse Normal	durchschnittliche Schadenshöhe

Tabelle 4: Verteilungen der Tweedie-Familie

Eine Exponentialfamilie kann laut Quijano, Garrido [13] nicht nur durch dessen kanonische Parameter definiert werden, sondern auch in Form von dessen Erwartungswert. Sei $Tw(p, \mu, \lambda)$ eine Tweedie-Verteilung mit Varianzfunktionsexponent p, Erwartungswert μ und Indexparameter λ. Sei S wie zuvor

$$S = \sum_{i=1}^{N} Z_i \tag{90}$$

mit $N \sim Poi(\mu_N)$ und $Z_i \overset{iid}{\sim} \Gamma(\alpha, \tau)$, welche von N unabhängig sind. Definiere mit $CPG(\mu_N, \alpha, \tau)$ eine zusammengesetzte Poisson-Gamma-Verteilung mit Poissonrate μ_N

und Sprunghöhenverteilung $\Gamma(\alpha, \tau)$, welche die Verteilung von S ist. Für $p \in (1,2)$ stimmen die Tweedie-Familie und die zusammengesetzte Poisson-Gamma-Verteilung überein.

Für die nachstehenden Definitionen und die Vorgehensweise bei der Umwandlung zwischen Tweedie- und zusammengesetzter Poisson-Gamma-Verteilung siehe Jørgensen [6, Kapitel 4]. Sei $V_p(\mu) = \mu^p$, für $\mu \in \Omega_p$, die Tweedie Varianzfunktion und seien κ_p die Kumulantenfunktion und τ_p die Mittelwertsfunktion auf dem Definitionsbereich Θ_p. Diese seien wie folgt definiert

$$\tau_p(\theta) = \left(\frac{\theta}{\alpha - 1}\right)^{\alpha - 1}, \quad \text{für } p \neq 1, \tag{91}$$

$$\kappa_p(\theta) = \frac{\alpha - 1}{\alpha}\left(\frac{\theta}{\alpha - 1}\right)^{\alpha}, \quad \text{für } p \neq 1, 2. \tag{92}$$

Für Tweedie Modelle gilt $Tw^*(p, \theta, \lambda) = Tw(p, \lambda\tau_p(\theta), \lambda^{1-p})$, wobei Tw^* eine reproduktive/additive Form von Tw ist. Die Kumulantenerzeugende Funktion von Tw^* für $s \in \Theta_p - \theta$ ist

$$K_p^*(s; \theta, \lambda) = \lambda\kappa_p(\theta)[\Big(1 + \frac{s}{\theta}\Big)^{\alpha} - 1], \quad \text{für } p \neq 1, 2. \tag{93}$$

Sei nun $N, Z_1, Z_2....$ eine Folge unabhängiger Zufallsvariablen, sodass N Poisson verteilt mit $Poi(m)$ und die Z_i seien identisch verteilt, sodass

$$S = \sum_{i=1}^{N} Z_i. \tag{94}$$

Für $N = 0$ sei auch $S = 0$ und außerdem nehme man an, dass $m = \lambda\kappa_p(\theta)$ und $Z_i \sim \Gamma(-\alpha, -\theta), \quad i = 1, 2, 3, ...,$ wobei $\alpha, \theta < 0$ und α und p über die Beziehung $\alpha = \frac{p-2}{p-1}$ verbunden sind. Für die Summe gammaverteilter Zufallsvariablen folgt somit

$$S|N = n \sim \Gamma(-n\alpha, -\theta), \quad \text{für } n \geq 1. \tag{95}$$

Sei $M(s, p, b) = (1 + \frac{-s}{b})^{-p}, \quad p, b > 0$ die Momenterzeugende Funktion einer $\Gamma(p, b)$ verteilten Zufallsvariable. Die Momentenerzeugende Funktion von S hat daher folgende Form

$$\begin{aligned} E[e^{sS}] &= E[E(e^{sS}|N)] && (96)\\ &= E[(M(s; -\alpha, -\theta))^N] && (97)\\ &= \exp\{m[M(s; -\alpha, -\theta) - 1]\} && (98)\\ &= \exp\left\{\lambda\kappa_p(\theta)\left[(1 + \frac{s}{\theta})^{\alpha} - 1\right]\right\}, \quad \alpha, \theta < 0. && (99) \end{aligned}$$

Das zeigt, dass $S \sim Tw^*(p, \theta, \lambda)$ und somit, wegen $Tw^*(p, \theta, \lambda) = Tw(p, \lambda\left(\frac{\theta}{\alpha-1}\right)^{\alpha-1}, \lambda^{1-p})$,

$$Tw(p, \lambda\left(\frac{\theta}{\alpha - 1}\right)^{\alpha - 1}, \lambda^{1-p}) = CPG(\lambda\kappa_p(\theta), -\alpha, -\theta). \tag{100}$$

Setzt man $\mu = \lambda\left(\frac{\theta}{\alpha-1}\right)^{\alpha-1}$ und $\tilde{\lambda} = \frac{1}{\lambda^{1-p}}$, so ist es möglich eine Parametrisierung durch die andere darzustellen. Sei dazu $p \in (1,2)$, $\mu > 0$ und $\tilde{\lambda} > 0$, dann gilt

$$Tw(p, \mu, \tilde{\lambda}) = CPG\left(\frac{\tilde{\lambda}\mu^{2-p}}{2-p}, -\frac{p-2}{p-1}, \frac{\tilde{\lambda}\mu^{1-p}}{p-1}\right). \tag{101}$$

Mit $m, \alpha, \tau > 0$ und durch einfaches Nachrechnen erhält man ebenso

$$CPG(\mu_N, \alpha, \tau) = Tw\left(\frac{\alpha+2}{\alpha+1}, \frac{\mu_N \alpha}{\tau}, \frac{(\mu_N\alpha)^{\frac{\alpha+2}{\alpha+1}-1}\tau^{2-\frac{\alpha+2}{\alpha+1}}}{\alpha+1}\right). \tag{102}$$

Um nun mit Hilfe von GLMs den Schätzer für den Erwartungswert von S zu bestimmen, muss zunächst $p \in (1,2)$ geschätzt werden, siehe Quijano, Garrido [13]. Eine numerische Methode für den Maximum Likelihood Schätzer von p wird in Gilchrist, Drinkwater [4] vorgestellt, welche in R als tweedie.profile() implementiert ist (siehe spätere Tweedie Analyse).

3.2.1 Schätzer Nettoprämie

Man nehme wie Quijano, Garrido in [13] an, dass das Modell k_S signifikante Ausprägungen beinhalte und sei dazu $\boldsymbol{\beta^S}$ der dazugehörige Koeffizientenvektor. Des Weiteren seien $\boldsymbol{X}$ die Designmatrix und S_i die Responsevariable der i-ten Klasse, wobei S_i einer $Tw(p, \mu_i, \lambda)$ Verteilung mit $p \in (1,2)$ und $\lambda > 0$ folgt. Somit gilt

$$E[S_i] := \mu_i = \exp\{\boldsymbol{X_i}\boldsymbol{\beta^S}\}. \tag{103}$$

3.3 Direkte Modellierung mit Hilfe der quasi-Poisson-Familie

Bei dieser Methode beginnt man damit, dass man den Schaden einer Polizze als zusammengesetzt Poisson-verteilte Zufallsvariable betrachtet. Sei $S = Z_1 + Z_2 + ... + Z_N$, wobei $N \sim Poi(\lambda)$ mit $E[N] = \lambda$ und $Z_i \sim \Gamma(\alpha, \beta)$. Für eine zusammengesetzt Poisson-verteilte Zufallsvariable gilt, siehe Kass et al. [7, Theorem 3.7.1 (CLT for compound Poisson distributions)]:

$$Var[S] = \lambda E[Z^2] = \lambda((E[Z]^2 + Var[Z]) = \lambda\left(\left(\frac{\alpha}{\beta}\right)^2 + \frac{\alpha}{\beta^2}\right) \tag{104}$$

$$= \lambda\frac{\alpha(\alpha+1)}{\beta^2} = E[S]\frac{\alpha+1}{\beta}, \tag{105}$$

wobei $E[S] = E[N] \cdot E[Z] = \lambda\frac{\alpha}{\beta}$. Man nimmt nun an, siehe Kaas et al. [7, Kapitel 9], dass λ viel variabler ist als der Shapeparameter α und der Scaleparameter β. Somit geht man davon aus, dass $Var[S]$ approximativ proportional zu $E[S]$ ist. Aufgrund dieser Annahme kann man Modelle mit der quasi-Poisson-Familie fitten, welche für die Varianz einen Parameter c frei lässt, $Var[S] = c \cdot E[S]$, mit c nicht notwendigerweise gleich $\frac{\alpha+1}{\beta}$.

Die Theorie dieser Vorgangsweise stammt aus McCullagh, Nelder [10, Kapitel 9]. Die Komponenten eines Zufallsvektors Y seien unabhängig mit Erwartungswert μ und bekannter Varianzfunktion $Var(Y) = \phi V(\mu)$ (hier: $V(\mu) = \mu$). Dann ist die quasi-Scorefunktion über die quasi-Likelihood Funktion $Q(\mu; y)$ definiert durch

$$U = u(\mu; y) = \frac{\partial Q(\mu; y)}{\partial \mu} := \frac{y-\mu}{\phi V(\mu)}. \tag{106}$$

Dabei hat die Funktion U folgende Eigenschaften mit einer log-Likelihood Funktion gemeinsam

$$E[U] = 0 \text{ und} \tag{107}$$

$$Var(U) = Var\left(\frac{\partial Q(\mu;y)}{\partial \mu}\right) = \frac{Var(y)}{\phi^2 V^2(\mu)} = \frac{1}{\phi V(\mu)} = -E\left(\frac{\partial^2 Q(\mu;y)}{\partial \mu^2}\right) = -E\left(\frac{\partial U}{\partial \mu}\right). \tag{108}$$

Somit kann die quasi-Likelihood Funktion $Q(\mu;y)$ (genauer gesagt: log-quasi-Likelihood Funktion) über das Integral (falls es existiert)

$$Q(\mu;y) = \int_y^\mu \frac{y-t}{\phi V(t)} dt, \tag{109}$$

bestimmt werden und verhält sich wie eine log-Likelihood Funktion für μ.

Seien nun y_i unabhängige Responses, für welche jeweils $E[y_i] = \mu_i$ und $Var(y_i) = \phi V(\mu_i)$ gilt, so ist die quasi-Likelihood der gesamten Daten die Summe der einzelnen Beiträge

$$Q(\boldsymbol{\mu};\boldsymbol{y}) = \sum Q_i(\mu_i;y_i). \tag{110}$$

Es gelte wieder für μ_i, dass $g(\mu_i) = x_i^T\boldsymbol{\beta}$ und um nun den Maximum quasi-Likelihood Schätzer $\hat{\boldsymbol{\beta}}$ bestimmen zu können, muss lediglich die Gleichung (110) nach $\boldsymbol{\beta}$ abgeleitet und 0 gesetzt werden. Dies ergibt

$$U(\boldsymbol{\beta};y) = \sum_i^n \frac{y_i - \mu_i}{\phi V(\mu_i)} \frac{\partial \mu_i}{\partial \boldsymbol{\beta}}. \tag{111}$$

Verwendet man in weiterer Folge die Matrixschreibweise $\boldsymbol{V} = \text{diag}(V(\mu_1), ..., V(\mu_n))$ und $\boldsymbol{D} = \frac{\partial \mu}{\partial \boldsymbol{\beta}^T}$. Mit $\boldsymbol{D}$ wird die Ableitungsmatrix von $\mu(\boldsymbol{\beta})$ nach $\boldsymbol{\beta}$ bezeichnet, welche von der Ordnung $(n \times p)$ ist. Die quasi-Scorefunktion hat dann folgende Darstellung

$$\boldsymbol{U}(\boldsymbol{\beta};y) = \frac{1}{\phi}\boldsymbol{D}^T\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}). \tag{112}$$

Die Kovarianzmatrix von $\boldsymbol{U}(\boldsymbol{\beta};y)$, welche wegen (108) gleich dem negativen Erwartungswert der Ableitung von U nach $\boldsymbol{\beta}$ ist, ist

$$\boldsymbol{i}_\beta = \frac{1}{\phi}\boldsymbol{D}^T\boldsymbol{V}^{-1}\boldsymbol{D}. \tag{113}$$

Diese Matrix spielt für quasi-Likelihood Funktionen dieselbe Rolle wie die Fisher-Information bei gewöhnlichen Likelihood Funktionen. Für die asymptotische Kovarianzmatrix von $\hat{\boldsymbol{\beta}}$ gilt

$$Cov(\hat{\boldsymbol{\beta}}) \backsimeq \boldsymbol{i}_\beta = \frac{1}{\phi}(\boldsymbol{D}^T\boldsymbol{V}^{-1}\boldsymbol{D})^{-1}. \tag{114}$$

Nun wählt man einen beliebigen Startwert $\hat{\boldsymbol{\beta}}_0$, welcher idealerweise nahe bei $\hat{\boldsymbol{\beta}}$ liegt und führt die Newton-Raphson Methode mit Fisher-Scoring durch.

Der erste Schritt ist folgender:

$$\hat{\boldsymbol{\beta}}_1 = \hat{\boldsymbol{\beta}}_0 + (\hat{\boldsymbol{D}}^T \hat{\boldsymbol{V}}^{-1} \hat{\boldsymbol{D}}_0)^{-1} \hat{\boldsymbol{D}}^T \hat{\boldsymbol{V}}^{-1} (\boldsymbol{y} - \boldsymbol{\mu_0}), \tag{115}$$

wobei $\boldsymbol{\mu}_0 = \boldsymbol{\mu}(\boldsymbol{\beta}_0)$. Die Iterationen werden solange fortgesetzt bis Konvergenz eintritt um den quasi-Likelihood Schätzer $\hat{\boldsymbol{\beta}}$ zu erhalten. Interessant hierbei ist die Unabhängigkeit von ϕ und, dass sich die quasi-Likelihoods wie gewöhnliche log-Likelihoods verhalten. Nur für die Schätzung von ϕ ist Vorsicht geboten, da sich $Q(\cdot; y)$ hier nicht wie ein log-Likelihood verhält. Deshalb wird hierfür die mittlere Pearson Statistik verwendet.

$$\hat{\phi} = \frac{1}{n-p} \sum_{i=1}^{n} \frac{(y_i - \hat{\mu_i})^2}{V(\hat{\mu_i})}. \tag{116}$$

3.3.1 Schätzer Nettoprämie

Man nehme wieder wie Quijano, Garrido in [13] an, dass das Modell k_S signifikante Ausprägungen beinhalte und sei dazu $\boldsymbol{\beta^S}$ der dazugehörige Koeffizientenvektor. Des Weiteren seien $\boldsymbol{X}$ die Designmatrix und S_i die Responsevariable der i-ten Klasse. Somit gilt wieder

$$E[S_i] := \mu_i = \exp\{\boldsymbol{X_i} \boldsymbol{\beta^S}\}. \tag{117}$$

4 Grundlagen und Erklärungen wichtiger R Befehle

4.1 Deviance/LRT/Pearson's χ^2

Bei der Suche nach einem passenden Modell für einen gegebenen Datensatz ist man mit dem Ziel konfrontiert, die tatsächlichen Beobachtungen $\boldsymbol{y}$ durch gefittete Werte $\hat{\boldsymbol{\mu}}$ zu ersetzen, siehe McCullagh, Nelder [10, Kapitel 2]. Dabei sollte die Anzahl der Parameter in diesem Modell relativ klein sein, da es sich um eine Vereinfachung der Daten handeln soll. Somit werden die μ's nicht exakt den y's entsprechen. Da eine geringe Abweichung sehr wohl in Ordnung ist, ist eine große Abweichung ein Kriterium dafür, dass das Modell nicht passt.

Hat man zum Beispiel n Beobachtungen, so kann man ein Modell mit bis zu n Parametern fitten. Das einfachste Modell ist das sogenannte Nullmodell mit nur einem gemeinsamen Parameter μ für alle y's. Das genaue Gegenteil dazu ist das volle/saturierte Modell mit n Parametern mit jeweils einem μ für jede Beobachtung y. Die Daten werden in diesem Modell exakt gematcht. In der Praxis sind beide Modelle nicht interessant, da das Nullmodell zu ungenau und das volle Modell zu keiner Vereinfachung der tatsächlichen Daten führt. Das saturierte Modell liefert allerdings einen Vergleichswert für ein Modell mit $p < n$ Parametern.

Es wird die log-Likelihood Funktion als Funktion der Mittelwertsparamter $\boldsymbol{\mu}$ anstatt der kanonischen Paramter $\boldsymbol{\theta}$ ausgedrückt. Sei $l(\hat{\boldsymbol{\mu}}, \phi; \boldsymbol{y})$ die log-Likelihood Funktion, welche über $\boldsymbol{\beta}$ maximiert wird, während ϕ fix ist.

Die Messgröße, wie gut der Fit eines Modells ist, ist proportional zur doppelten Differenz zwischen dem erreichbaren log-Likelihood Wert des saturierten Modells und dem log-Likelihood Wert des betrachteten Modells. Die skalierte Deviance $D^*(\boldsymbol{y}, \hat{\boldsymbol{\mu}})$ ist wie folgt definiert, siehe Ohlsson, Johansson [12, Kapitel 3]:

$$D^* = D^*(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = 2[l(\boldsymbol{y}) - l(\hat{\boldsymbol{\mu}})], \tag{118}$$

wobei angenommen wird, dass ϕ sowohl in $l(\boldsymbol{y})$ als auch in $l(\hat{\boldsymbol{\mu}})$ gleich ist. Durch Einsetzen der log-Likelihood Gleichungen und der neu definierten inversen Funktion h von b' ($\mu_i = b'(\theta_i) \Leftrightarrow \theta_i = h(\mu_i)$) erhält man

$$D^* = \frac{2}{\phi} \sum_i \omega_i (y_i h(y_i) - b(h(y_i))) - y_i h(\hat{\mu}_i) + b(h(\hat{\mu}_i)). \tag{119}$$

Die nicht skalierte Deviance sei mit $D = \phi D^*$ definert.

Um zwei konkrete Beispiele für die Deviance anzugeben, wurden die Poisson- und die Gamma-Verteilung gewählt. Die Deviances der beiden Verteilungen sehen wie folgt aus

$$D^*(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = 2 \sum_i \omega_i (y_i \log(y_i) - y_i \log(\hat{\mu}_i) - y_i + \hat{\mu}_i), \tag{120}$$

$$D^*(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = 2 \sum_i \omega_i \left(\frac{y_i}{\hat{\mu}_i} - 1 - \log\left(\frac{y_i}{\hat{\mu}_i} \right) \right). \tag{121}$$

Die quasi-Deviance wird wie in McCullagh, Nelder [10, Kapitel 2] definiert durch

$$D(\boldsymbol{y}; \hat{\boldsymbol{\mu}}) = -2\phi(q(\hat{\boldsymbol{\mu}}; \boldsymbol{y}) - q(\boldsymbol{y}; \boldsymbol{y})) = -2 \sum_{i=1}^{n} \int_{y_i}^{\hat{\mu}_i} \frac{y_i - t}{V(t)} dt. \tag{122}$$

Eine weiteres wichtiges Maß für die Güte des Fits eines statistischen Modells ist Pearson's Chi-Quadrat X^2,

$$X^2 = \sum_i \frac{(y_i - \hat{\mu}_i)^2}{Var(Y_i)} = \frac{1}{\phi} \sum_i \omega_i \frac{(y_i - \hat{\mu}_i)^2}{v(\hat{\mu}_i)}. \tag{123}$$

Es existiert auch hier eine unskalierte Variante ϕX^2. Das Problem dabei ist, dass ϕ im Allgemeinen, wie zum Beispiel für die Gamma-Verteilung, nicht immer bekannt ist. Für den Poisson Fall ist, auf Grund der Definition der Exponentialfamilie, $\phi = 1$. Kennt man ϕ nicht, so kann man sich der nachfolgenden approximativen Lösung bedienen, welche einen erwartungstreuen Schätzer liefert.

Pearson's X^2 ist approximativ $\chi^2(n-r)$ verteilt, wobei r die Anzahl der zu schätzenden β-Parameter ist. Da für eine χ^2_k-verteilte Zufallsvariable gilt, dass deren Erwartungswert k ist, gilt hier, dass $E[X^2] \approx (n-r)$. Daher ist $\hat{\phi}$ ein approximativer erwartungstreuer Schätzer von ϕ,

$$\hat{\phi} = \frac{\phi X^2}{n-r} = \frac{1}{n-r} \sum_i \omega_i \frac{(y_i - \hat{\mu}_i)^2}{v(\hat{\mu}_i)}. \tag{124}$$

Laut McCullagh, Nelder [10, Kapitel 2] ist dieser approximative Schätzer $\hat{\phi}$ von ϕ für die Gamma-Verteilung der beste (Voraussetzung: nicht gehäufter Datensatz).

Die obigen Definitionen sind nun für die Durchführung einer GLM-Analyse von großer Bedeutung. Speziell die Deviance wird im nächsten Kapitel benötigt, da man, um herauszufinden, ob ein Merkmal bei der GLM-Analyse von Relevanz ist, Hypothesentests durchführt. Dies wird hier, wie auch in Ohlsson, Johansson [12, Kapitel 3] mittels eines LRT (Likelihood Ratio Tests) gemacht, bei welchem zwei Modelle gegenübergestellt werden und die Deviances für das Ergebnis ausschlaggebend sind.

Eines der beiden Modelle beinhaltet ein zu untersuchendes Merkmal, das andere nicht. Für beide Modelle muss gelten, dass eines eine Teilmenge des anderen Modells ist und man in der Nullhypothese annimmt, dass das Modell mit weniger Merkmalen besser ist als jenes mit einem zusätzlichen Merkmal.

Satz 4.1. (*Likelihood Ratio Test, siehe Ohlsson, Johansson [12, Theorem 3.1]*)
Man wähle 2 Modelle H_r und H_s, wobei $H_s \subset H_r$. Seien $\hat{\boldsymbol{\mu}}^{(r)}$ die MLEs unter H_r, sowie gleichermaßen für H_s. Dann ist die LRT Statistik für das Testen von H_s gegen H_r gegeben durch

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

wobei $D^*(y, \hat{\mu}) = 2 \cdot (l(y) - l(\hat{\mu}))$.

Beweis: Da ein Modell eine Teilmenge des anderen ist, müssen beide der selben Exponentialfamilie (Poisson, Gamma, ...) angehören und daher auch dasselbe ϕ haben. Unter allgemeinen Annahmen sind LRT's approximativ χ^2-verteilt, siehe dazu Lindgren [9]. Falls die Modelle je f_r beziehungsweise f_s nicht redundante β-Parameter haben so gilt für den Likelihood Ratio Test mit $f_r - f_s$ Freiheitsgraden

$$LRT = -2 \log \left(\frac{L_s(\hat{\boldsymbol{\mu}}^{(r)})}{L_s(\hat{\boldsymbol{\mu}}^{(s)})} \right) \tag{125}$$

$$= -2 \log(L_s(\hat{\boldsymbol{\mu}}^{(r)})) + 2 \log(L_s(\hat{\boldsymbol{\mu}}^{(s)})) \tag{126}$$

$$= D^*(\boldsymbol{y}, \hat{\boldsymbol{\mu}}^{(s)}) - D^*(\boldsymbol{y}, \hat{\boldsymbol{\mu}}^{(r)}). \tag{127}$$

Verwendet man den Schätzer $\hat{\phi}$ für ϕ, so kann der LRT umgeformt werden zu

$$\frac{\phi}{\hat{\phi}}(D^*(\boldsymbol{y}, \hat{\boldsymbol{\mu}}^{(s)}) - D^*(\boldsymbol{y}, \hat{\boldsymbol{\mu}}^{(r)})) = \frac{D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}^{(s)}) - D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}^{(r)})}{\hat{\phi}}, \tag{128}$$

wobei $\hat{\phi}$ anhand des größeren Modells geschätzt werden muss. Würde man stattdessen das kleinere Modell verwenden so beinhaltet der Schätzer auch die Variation der im größeren Modell zusätzlichen Merkmale. Falls dann die Nullhypothese abgelehnt werden würde, so können diese Merkmale nicht als zufällig angesehen werden.

Das Ergebnis eines Likelihood Ratio Tests setzt sich meist aus der Teststatstik, welche χ^2-verteilt ist mit $f_r - f_s$ Freiheitsgraden, sowie einem p-Wert zusammen. Der p-Wert der Stichprobe gibt dabei an, wie groß die Wahrscheinlichkeit ist, dass bei bestehender Nullhypothese ein gleiches oder ein extremeres Ergebnis eintritt. Ist ein p-Wert kleiner als ein vorgegebenes Signifikanzniveau (hier: 5 %), so wird die Nullhypothese verworfen. Ist dies nicht der Fall, kann keine Aussage getroffen werden. Dabei ist es irrelevant ob der p-Wert 0.1 oder 0.8 ist.

4.2 fit.contrast()

Die Funktion fit.contrast(model, varname, coeff, ...), siehe Warners et al. [5], berechnet die gewünschten Kontraste durch erneutes Anpassen des Modells mit den entsprechenden Argumenten. Dafür wird ein Modell benötigt. In unserem Fall handelt es sich hier um das optimale generalisierte lineare Modell.

Des Weiteren muss mit „varname“ eine Merkmalsausprägung definiert werden und mit Hilfe von „coeff“ (Vektor: z.B.: c(-1,0,0,1)) werden 2 konkrete Merkmalsausprägungen gewählt. In diesem Fall würde der Erwartungswert der ersten Merkmalsausprägung mit dem Erwartungswert der vierten verglichen werden. Mit Hilfe eines Hypothesentests für Kontraste kann berechnet werden, ob 2 Merkmalsausprägungen weiter zusammengefasst werden sollten. Die Nullhypothese setzt hierbei eine Gleichheit der Erwartungswerte voraus, welche zum Niveau $\alpha = 5$ % verworfen werden soll. Dies geschieht über die Berechnung der Teststatistik, sowie der dazugehörigen zweiseitigen p-Werte.

4.3 drop1() (Backward Selection)

Die Vorgehensweise bei der Backward Selection ist folgende, siehe Kang [8]:

Man beginnt mit dem vollen Modell, das bedeutet, dass man zunächst mit allen verfügbaren Merkmalen startet. Im ersten Schritt wird jedes Merkmal einmal aus dem Modell entfernt und mit dem vollen Modell „verglichen“. Dieser Vergleich wird mit Hilfe des Likelihood Ratio Tests durchgeführt. Man erhält somit so viele p-Werte wie Merkmale im vollen Modell vorhanden sind.

Anschließend wird der größte p-Wert betrachtet. Ist dieser größer als ein vorgegebenes Signifikanzniveau (hier: $\alpha = 5$ %), so wird dieses Merkmal aus dem Modell entfernt und man beginnt von vorne, mit dem Unterschied, dass das volle Modell nun 1 Merkmal weniger hat. Diese Vorgehensweise wird solange wiederholt, bis kein p-Wert mehr größer als $\alpha = 5$ % ist. Das dadurch gefundene Modell ist jetzt das in diesem Sinn optimale Modell.

In R gibt es hierfür eine vorimplementierte Funktion, die sogenannte drop1() Funktion, welche in dieser Arbeit verwendet wurde.

5 Analyse der Versicherungsdaten

In diesem Kapitel wird die genaue Vorgehensweise zur Bestimmung der gesuchten Parameter des multiplikativen Modells beschrieben. Am Ende erhält man eine genaue Übersicht der Ergebnisse aller drei Versicherungssparten (Haftpflicht-, Vollkasko-, Teilkaskoversicherung). Da aber das Verfahren für alle Versicherungssparten dasselbe ist, werden die einzelnen Schritte nur für die Haftpflichtversicherung angeführt. Der Vorteil dieser Versicherungssparte ist, dass deutlich mehr Daten zur Verfügung stehen, als bei den anderen beiden. Zu Beginn steht die Datenaufbereitung an, welche meist sehr viel Zeit in Anspruch nimmt. Ein wesentlicher Punkt dabei ist das Clustern von Merkmalsausprägungen.

5.1 Clusterdiagramme

Zunächst wird jeder Postleitzahl der dazugehörige Bezirk zugeordnet und alle Bezirke, die einen geforderten minimalen Bestand nicht erfüllen, werden zu einer Klasse „Sonstige“ zusammengefasst. Anschließend wird der erwartete Schaden (Schadenbedarf) aller Bezirke berechnet. Dies erfolgt durch die Multiplikation der Schadenfrequenz mit der durchschnittlichen Schadenhöhe. Anhand dieser Werte wird das Ward-Clustering angewendet. Dies führt zu dem auf der nächsten Seite abgebildetetn, Cluster-Dendogramm der Bezirke. In weiterer Folge ist ein Clustering der Automarken/Konzerne notwendig. Die Vorgehensweise ist analog zu der des Bezirks-Clusterings und liefert ebenso ein Cluster-Dendogramm.

Der Konzern „Manuell“ steht dabei für selbst gebaute Fahrzeuge, welche eine Straßenzulassung erhalten haben.

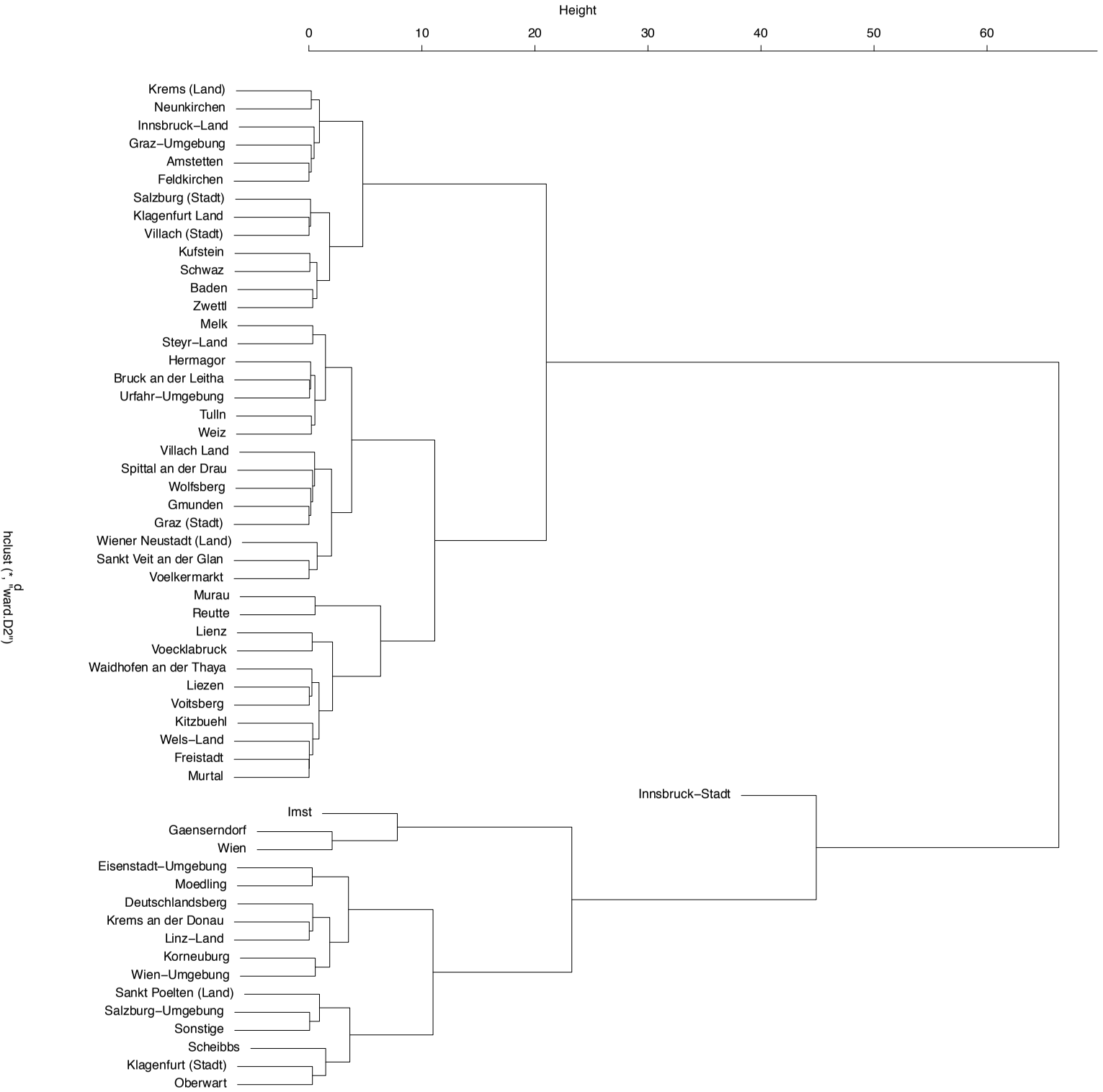
Auf den folgenden 3 Seiten sind zunächst das Cluster-Dendogramm der Bezirke, danach eine Österreichkarte, mit den gewählten Bezirksklassen sowie das Cluster-Dendogramm der Bezirke zu sehen.

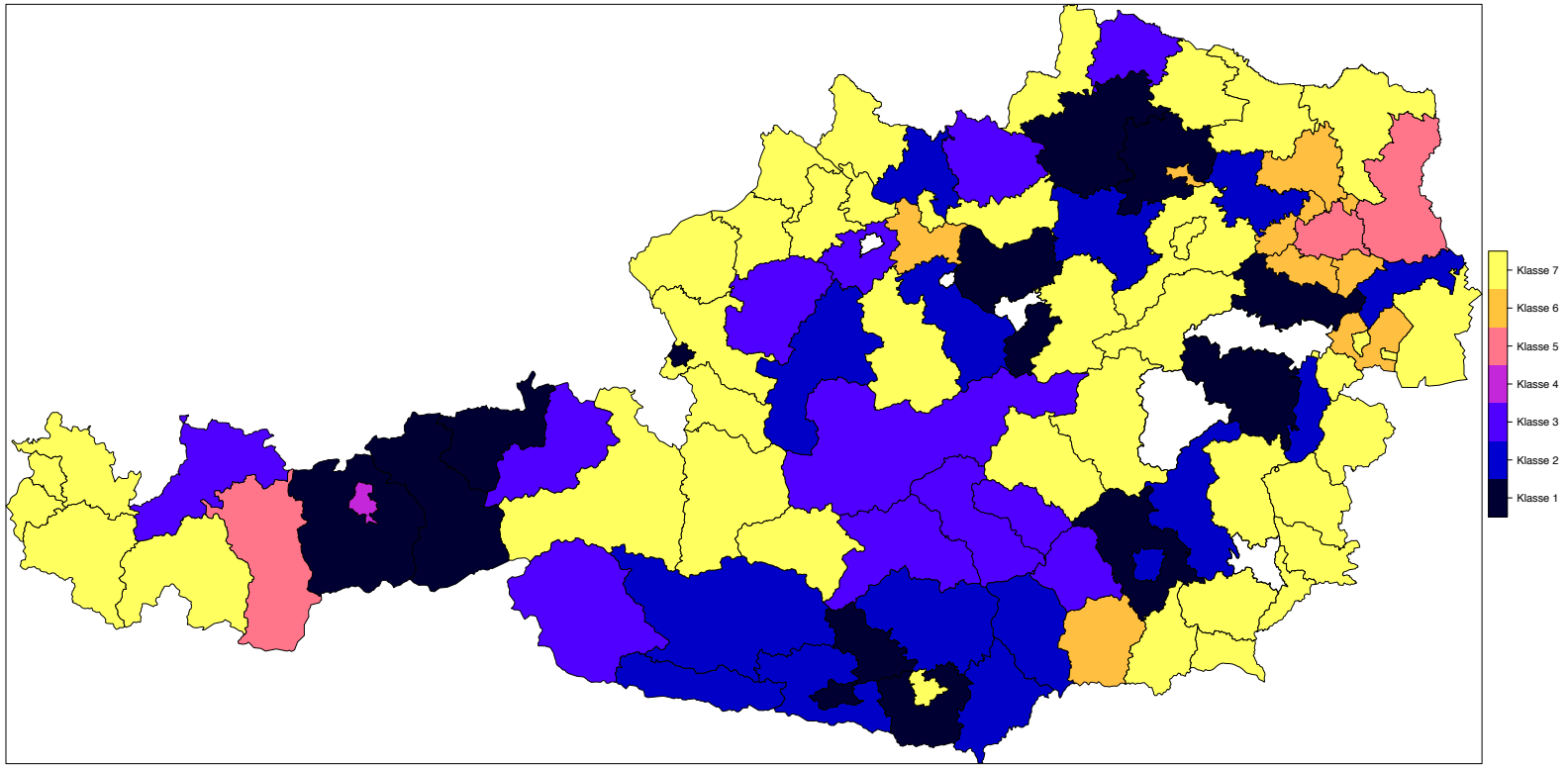
Klasse	Bezirk
1	Krems (Land), Neunkirchen, Innsbruck-Land, Graz-Umgebung, Amstetten, Feldkirchen, Salzburg (Stadt), Klagenfurt-Land, Villach (Stadt), Kufstein, Schwaz, Baden, Zwettl
2	Melk, Steyr-Land, Hermagor, Bruck an der Leitha, Urfahr-Umgebung, Tulln, Weiz, Villach-Land, Spittal an der Drau, Wolfsberg, Gmunden, Graz (Stadt, Wiener Neustadt (Land), Sankt Veit an der Glan, Völkermarkt
3	Murau, Reutte, Lienz, Vöcklabruck, Waidhofen an der Thaya, Liezen, Voitsberg, Kitzbühel, Wels-Land, Freistadt, Murtal
4	Innsbruck-Stadt
5	Imst, Gänserndorf, Wien
6	Eisenstadt-Umgebung, Mödling, Deutschlandsberg, Krems an der Donau, Linz-Land, Korneuburg, Wien-Umgebung
7	Sankt Pölten (Land), Salzburg-Umgebung, Sonstige, Scheibbs, Klagenfurt (Stadt), Oberwart

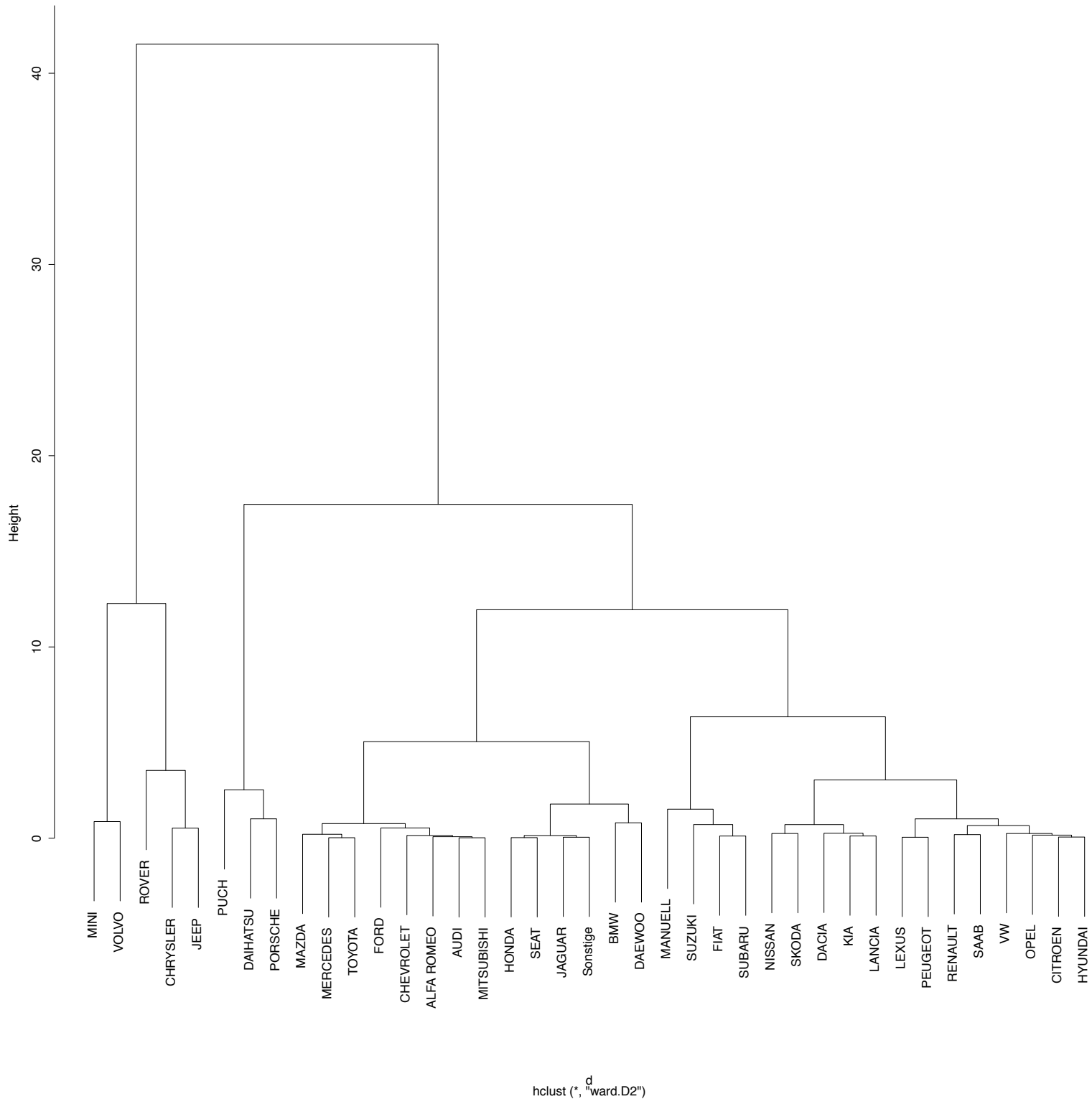
Tabelle 5: Clustering der Bezirke

Klasse	Konzern
1	Mini, Volvo, Rover, Chrysler, Jeep
2	Puch, Daihatsu, Porsche
3	Mazda, Mercedes, Toyota, Ford, Chevrolet, Alfa-Romeo, Audi, Mitsubishi
4	Honda, Seat, Jaguar, Sonstige, BMW, Daewoo
5	Manuell, Suzuki, Fiat, Subaru
6	Nissan, Skoda, Dacia, Kia, Lancia, Lexus, Peugeot, Renault, Saab, VW, Opel Citroen, Hyundai

Tabelle 6: Clustering der Konzerne


Cluster-Dendogramm Bezirk (KH, PKW)
Height
0
10
20
30
40
50
60
Krems (Land)
Neunkirchen
Innsbruck–Land
Graz–Umgebung
Amstetten
Feldkirchen
Salzburg (Stadt)
Klagenfurt Land
Villach (Stadt)
Kufstein
Schwaz
Baden
Zwettl
Melk
Steyr–Land
Hermagor
Bruck an der Leitha
Urfahr–Umgebung
Tulln
Weiz
Villach Land
Spittal an der Drau
Wolfsberg
Gmunden
Graz (Stadt)
Wiener Neustadt (Land)
Sankt Veit an der Glan
Voelkermarkt
Murau
Reutte
Lienz
Voecklabruck
Waidhofen an der Thaya
Liezen
Voitsberg
Kitzbuehl
Wels–Land
Freistadt
Murtal
Innsbruck–Stadt
Imst
Gaenserndorf
Wien
Eisenstadt–Umgebung
Moedling
Deutschlandsberg
Krems an der Donau
Linz–Land
Korneuburg
Wien–Umgebung
Sankt Poelten (Land)
Salzburg–Umgebung
Sonstige
Scheibbs
Klagenfurt (Stadt)
Oberwart
hclust (*, "ward.D2")
d


Klasse 7
Klasse 6
Klasse 5
Klasse 4
Klasse 3
Klasse 2
Klasse 1


Cluster−Dendogramm Konzern (KH, PKW)
Height
0
10
20
30
40
MINI
VOLVO
ROVER
CHRYSLER
JEEP
PUCH
DAIHATSU
PORSCHE
MAZDA
MERCEDES
TOYOTA
FORD
CHEVROLET
ALFA ROMEO
AUDI
MITSUBISHI
HONDA
SEAT
JAGUAR
Sonstige
BMW
DAEWOO
MANUELL
SUZUKI
FIAT
SUBARU
NISSAN
SKODA
DACIA
KIA
LANCIA
LEXUS
PEUGEOT
RENAULT
SAAB
VW
OPEL
CITROEN
HYUNDAI
d
hclust (*, "ward.D2")

Nach erfolgreicher Klasseneinteilung wird als nächstes für jedes Merkmal jene Ausprägung als Referenz gewählt, welche den höchsten Bestand aufweist, da die Anwendung der GLM-Analyse auf diesen Klassen sehr stabil ist. Würde man zum Beispiel eine Ausreißerklasse als Referenzklasse wählen, würden sehr starke Schwankungen bei den Schätzern auftreten. Auf diese Art und Weise kann man anschließend die weiteren Ausprägungen jedes Merkmals mit der jeweiligen Referenz vergleichen und somit interpretieren.

5.2 GLM-Analyse für die Schadenfrequenz

Dabei wählt man zunächst das vollständige Modell, bei dem man alle verfügbaren Merkmale hinzufügt. Anschließend werden mit Hilfe des Likelihood Ratio Tests alle nicht signifikanten Merkmale aus dem Modell entfernt (drop1() Funktion). Dabei wird für jedes Merkmal das Gesamtmodell mit einem Modell ohne dieses Merkmal verglichen. Man entfernt jenes Merkmal, welches am wenigsten signifikant ist. Diese Vorgangsweise (das Gesamtmodell ist aber bereits ein Teilmodell des ursprünglichen Gesamtmodells) wiederholt man so lange, bis alle Merkmale signifikant sind und keine weitere Verbesserung des Modells mehr möglich ist.

Zusätzlich betrachtet man für jedes Modell dessen AIC (Akaike Infomation Criterion), welcher durch die Summe des negativen zweifachen Likelihood Werts und der zweifachen Anzahl der zu schätzenden Parameter definiert ist. Dieser ermöglicht es uns ebenfalls, Modelle miteinander zu vergleichen. Je niedriger dieser Wert, umso besser ist das Modell. Man beginnt mit dem vollständigen Modell, bei welchem alle Merkmale hinzugefügt werden.

Nun werden der Reihe nach alle nicht signifikanten Merkmale entfernt bis nur noch das in diesem Sinn optimale Modell übrig bleibt:

```
Model:
glm_data$Schadenanzahl ~ glm_data$AkadTitel + glm_data$SonstTitel +
    glm_data$HUBRAU + glm_data$LEISTU + glm_data$vmerk_LEAS +
    glm_data$VNfamilienstand + glm_data$KundeSex + glm_data$vmerk_WECHS +
    glm_data$OVKC1 + glm_data$Konzern + glm_data$Bezirk + glm_data$alter_years +
    glm_data$vmerk_I3SFR + glm_data$Zugehoerigkeit + offset(log(glm_data$vers_jahre))
                        Df Deviance    AIC    LRT  Pr(>Chi)
<none>                       79690 103683
glm_data$AkadTitel       1    79697 103689   7.38 0.0066095 **
glm_data$SonstTitel      1    79702 103694  12.82 0.0003435 ***
glm_data$HUBRAU          3    79700 103688  10.56 0.0143503 *
glm_data$LEISTU          4    79713 103699  23.48 0.0001015 ***
glm_data$vmerk_LEAS      1    79718 103710  28.58 8.991e-08 ***
glm_data$VNfamilienstand 6    79723 103704  33.13 9.880e-06 ***
glm_data$KundeSex        3    79706 103694  16.36 0.0009569 ***
glm_data$vmerk_WECHS     1    79695 103687   5.89 0.0152233 *
glm_data$OVKC1           3    79697 103685   7.78 0.0508697 .
glm_data$Konzern         5    79752 103735  62.00 4.680e-12 ***
glm_data$Bezirk          6    79980 103962 290.62 < 2.2e-16 ***
glm_data$alter_years     7    80118 104098 428.94 < 2.2e-16 ***
glm_data$vmerk_I3SFR     5    80133 104116 443.08 < 2.2e-16 ***
glm_data$Zugehoerigkeit  3    79698 103686   8.46 0.0374819 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Abbildung 5: Optimales Poisson-Modell

Zu Beginn wurden Annahmen über unser Modell aufgestellt, welche leider nicht immer der Realität entsprechen. Eine dieser Annahmen war die Homogenität zwischen den Tarifzellen.

In der Praxis stellt sich hingegen heraus, dass die Poisson-Verteilung die tatsächlichen Daten nicht optimal widerspiegelt, da die Homogenität innerhalb der einzelnen Zellen nicht gegeben ist, siehe Ohlsson, Johansson [12, Kapitel 3]. Grund dafür sind meist nicht ausreichend vorhandene Informationen über die Versicherungsnehmer, was dazu führt, dass die Varianz der Beobachtungen größer als jene der Poisson-Verteilung ist. Dieses Phänomen ist auch in diesem Datensatz vorhanden. Zunächst wurden der Erwartungswert und die Varianz gegenübergestellt, welche sich bei gleicher Verteilung, $E[N] = Var[N]$, um die Gerade häufen sollten.

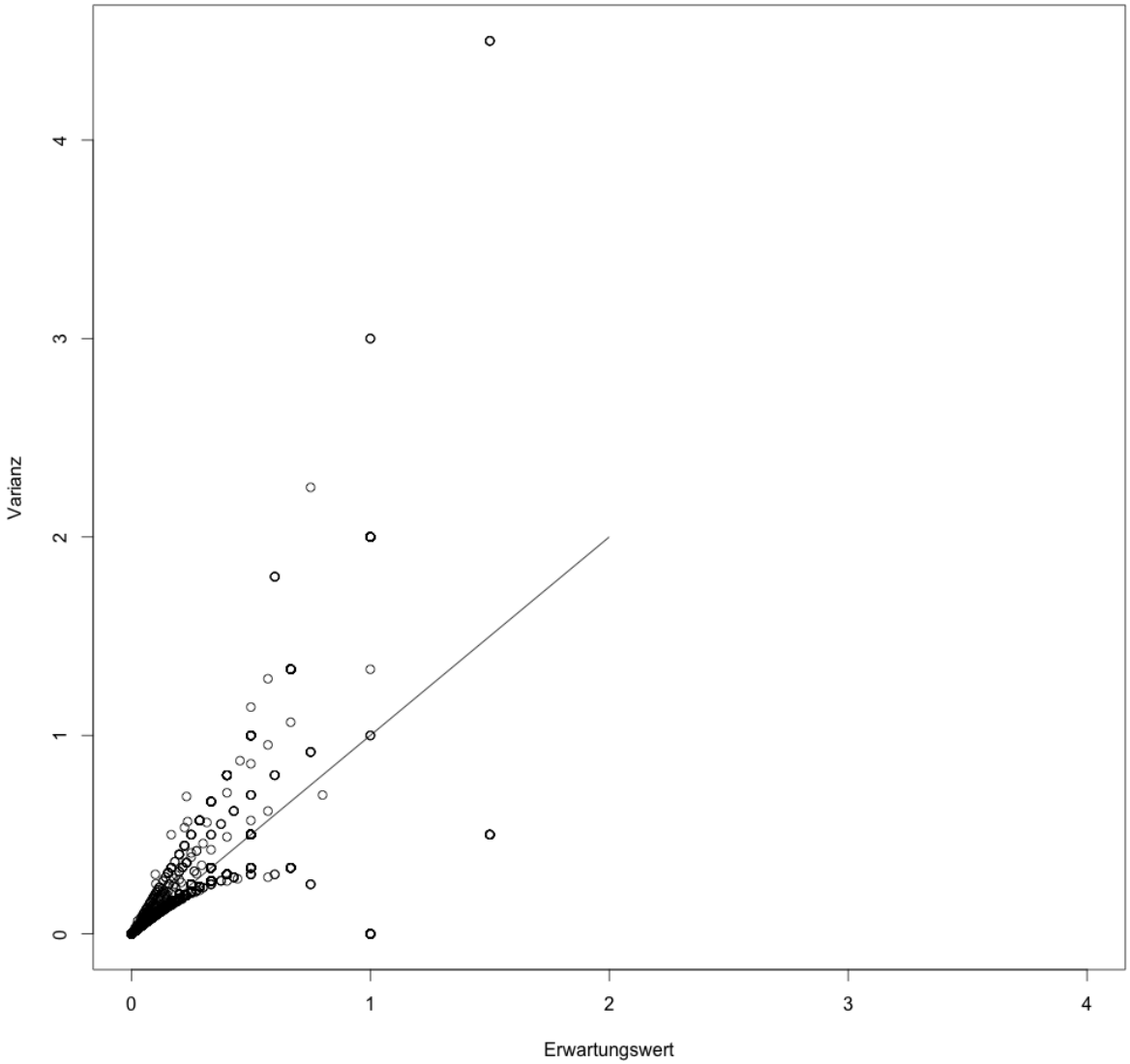


Abbildung 6: Overdispersion-Plot Poisson-Modell

Im obigen Plot zwischen Erwartungswert und Varianz von N ist bereits eine Tendenz zur Overdispersion erkennbar. Daher wurden zwei Tests auf Overdispersion durchgeführt.

```
> dispersiontest(opt_model, trafo=1)

        Overdispersion test

data:  opt_model
z = 7.182, p-value = 3.436e-13
alternative hypothesis: true alpha is greater than 0
sample estimates:
     alpha
0.03665674
```

Abbildung 7: Overdispersion-Test Poisson-Modell

Der erste Test (dispersiontest()), siehe Zeiler, Kleiber [15] deutet ebenfalls auf eine Overdispersion hin. Bei einem standardmäßigen Poisson-Modell ist, wie zuvor erwähnt, der Erwartungswert $E[y] = \mu$ gleich der Varianz $Var[y] = \mu$. Sei nun $Var[y] = \mu + \alpha \cdot trafo(\mu)$, wobei α bei einer vorliegenden gleichen Verteilung 0 sein müsste. Die Funktion dispersiontest() überprüft nun die Nullhypothese: $\alpha = 0$ gegen die Alternativhypothese $\alpha > 0$ für Overdispersion beziehungsweise $\alpha < 0$ für Underdispersion. Da hier $\alpha = 0.036$, liegt Overdispersion vor.

Die Notation bei der Erklärung des zweiten Tests wurde nun bewusst anders gewählt, um Missverständnisse zu vermeiden. DieserTest wird in Denuit et al. [2, Kapitel 2] vorgestellt. Hier sei die Varianzfunktion eines heterogenen Modells gegeben durch $Var[N_i] = \lambda_i + \tau\lambda_i^2$ mit $\tau = Var[\theta_i]$ der Varianz eines Zufallseffekts. Folglich ist die Nullhypothese $H_0 : \tau = 0$ gegen die Alternativhypothese $H_1 : \tau > 0$. Mit Hilfe der folgenden Teststatistiken kann man nun die Poisson-Verteilung gegen ein heterogenes Modell mit zuvor beschriebenes Varianzfunktion testen.

$$T_1 = \frac{\sum_{i=1}^{n}\left((k_i - \hat{\lambda}_i)^2 - k_i\right)}{\sqrt{2\sum_{i=1}^{n}\hat{\lambda}_i^2}}, \tag{129}$$

$$T_2 = \frac{\sum_{i=1}^{n}\left((k_i - \hat{\lambda}_i)^2 - k_i\right)}{\sqrt{\sum_{i=1}^{n}\left((k_i - \hat{\lambda}_i)^2 - k_i\right)^2}}, \tag{130}$$

$$T_3 = \frac{\sum_{i=1}^{n}\left((k_i - \hat{\lambda}_i)^2 - k_i\right)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\hat{\lambda}_i^2}\left((k_i - \hat{\lambda}_i)^2 - k_i\right)^2}\sqrt{\sum_{i=1}^{n}\hat{\lambda}_i^2}}, \tag{131}$$

wobei k_i in unserem Fall für die Schadenanzahl steht. Alle drei Teststatistiken sind standard-normalverteilt. Die Ergebnisse für unsere Daten lauten: $T_1 = 13,36$, $T_2 = 9,76$ und $T_3 = 6,23$. Die p-Werte sind somit alle kleiner als 10^{-4}, was zu einer Ablehnung der Nullhypothese (Erwartungswert gleich Varianz) führt.

Eine Möglichkeit Overdispersion zu vermeiden ist die Negativ-Binomialverteilung anstatt der Poisson-Verteilung für die Schadenfrequenz zu verwenden, siehe Ohlsson, Johansson [12, Kapitel 3].

5.2.1 Übergang von Poisson Verteilung zu Negativ Binomialverteilung

Diese Möglichkeit wird hier angewandt. Dafür wird der Erwartungswert der Poisson Verteilung als Zufallsvariable betrachtet. Sei $\Lambda_1, \Lambda_2, ...$ eine Folge von Zufallsvariablen, welche auf $(0, \infty)$ verteilt sind. Weiters seien $X_1, X_2, ...$ unabhängige Zufallsvariablen, welche bedingt darauf, dass $\Lambda_i = \lambda_i$, Poisson verteilt sind mit Erwartungswert λ_i. Es gelten somit $E[X_i|\Lambda_i] = \Lambda_i$ und $Var[X_i|\Lambda_i] = \Lambda_i$. Folglich gelten

$$E[X_i] = E[E(X_i|\Lambda_i)] = E[\Lambda_i] \tag{132}$$
$$Var[X_i] = E[Var(X_i|\Lambda_i)] + Var(E[X_i|\Lambda_i]) = E[\Lambda_i] + Var[\Lambda_i]. \tag{133}$$

Die Annahme eines Zusammenhangs von Erwartungswert und Varianz für $\{\Lambda_i\}$ impliziert somit einen für $\{X_i\}$. Eine gängige Möglichkeit ist $Var[\Lambda_i] = \nu E[\Lambda_i]$, für ein positives ν. Sei nun X_i die Schadensanzahl mit Erwartungswert $\omega_i\mu_i$ so folgt

$$Var[X_i] = E[\Lambda_i] + Var[\Lambda_i] = \omega_i\mu_i + \nu\omega_i\mu_i = (1+\nu)\omega_i\mu_i, \tag{134}$$
$$Var[Y_i] = Var\left[\frac{X_i}{\omega_i}\right] = \frac{Var[X_i]}{\omega_i^2} = \frac{(1+\nu)\omega_i\mu_i}{\omega_i^2} = \frac{(1+\nu)\mu_i}{\omega_i}. \tag{135}$$

Setzt man $(1+\nu) = \phi$ gilt $Var[Y_i] = \phi\mu_i/\omega_i$, was von der Form der Varianz einer Poisson Verteilung ähnelt, mit zusätzlichem Dispersion Parameter $\phi > 1$.

Nun muss nur noch eine Verteilung für die Λ_i gewählt werden. Diese wird mit der Gamma-Verteilung festgelegt, welche den Zusammenhang von Erwartungswert und Varianz aufrecht erhalten soll. Mit dem Lemma von der Totalen Wahrscheinlichkeit, siehe Ohlsson, Johansson [12, Lemma A.4] kann man nun die Wahrscheinlichkeitsdichtefunktion für X_i bestimmen.

$$P(X_i = x_i) = \int_0^\infty f_{Poi(\lambda)}(x_i) f_{\Gamma(\frac{\omega_i\mu_i}{\nu}, \frac{1}{\nu})}(\lambda) d\lambda \tag{136}$$
$$= \int_0^\infty \frac{\lambda^{x_i}}{x_i!} e^{-\lambda} \frac{(\frac{1}{\nu})^{\omega_i\mu_i/\nu}}{\Gamma(\omega_i\mu_i/\nu)} \lambda^{\omega_i\mu_i/\nu - 1} e^{-\lambda/\nu} d\lambda \tag{137}$$
$$= \frac{(\frac{1}{\nu})^{\omega_i\mu_i/\nu}}{\Gamma(\omega_i\mu_i/\nu)x_i!} \int_0^\infty \frac{\lambda^{x_i + \omega_i\mu_i/\nu - 1}}{x_i!} e^{-\lambda - \lambda/\nu} d\lambda \tag{138}$$

Nun nützt man $1 + \frac{1}{\nu} = \frac{1+\nu}{\nu}$ und folgende Eigenschaft aus:

$$1 = \int_0^\infty f_{\Gamma(\omega_i\mu_i/\nu + x_i, \frac{1+\nu}{\nu})}(\lambda) d\lambda \tag{139}$$
$$= \int_0^\infty \left(\frac{1+\nu}{\nu}\right)^{\omega_i\mu_i/\nu + x_i} \frac{1}{\Gamma(\omega_i\mu_i/\nu + x_i)} \lambda^{x_i + \omega_i\mu_i/\nu - 1} e^{-\lambda(\frac{1+\nu}{\nu})} d\lambda \tag{140}$$
$$\Leftrightarrow \tag{141}$$
$$\frac{\Gamma(\omega_i\mu_i/\nu + x_i)}{\left(\frac{1+\nu}{\nu}\right)^{\omega_i\mu_i/\nu + x_i}} = \int_0^\infty \lambda^{x_i + \omega_i\mu_i/\nu - 1} e^{-\lambda(\frac{1+\nu}{\nu})} d\lambda \tag{142}$$
$$\tag{143}$$

Setzt man dieses Ergebnis nun oben ein erhält man

$$P(X_i = x_i) = \frac{(\frac{1}{\nu})^{\omega_i\mu_i/\nu}}{\Gamma(\omega_i\mu_i/\nu)x_i!}\frac{\Gamma(\omega_i\mu_i/\nu + x_i)}{\left(\frac{1+\nu}{\nu}\right)^{\omega_i\mu_i/\nu+x_i}} \tag{144}$$

$$= \frac{\Gamma(\omega_i\mu_i/\nu + x_i)}{\Gamma(\omega_i\mu_i/\nu)x_i!}\frac{(\frac{1}{\nu})^{\omega_i\mu_i/\nu}}{\left(\frac{1+\nu}{\nu}\right)^{\omega_i\mu_i/\nu+x_i}} \tag{145}$$

$$= \frac{\Gamma(\omega_i\mu_i/\nu + x_i)}{\Gamma(\omega_i\mu_i/\nu)x_i!}\left(\frac{\frac{1}{\nu}}{\frac{\nu+1}{\nu}}\right)^{\omega_i\mu_i/\nu}\left(\frac{1}{\frac{\nu+1}{\nu}}\right)^{x_i} \tag{146}$$

$$= \frac{\Gamma(\omega_i\mu_i/\nu + x_i)}{\Gamma(\omega_i\mu_i/\nu)x_i!}\left(\frac{1}{\nu+1}\right)^{\omega_i\mu_i/\nu}\left(\frac{\nu}{\nu+1}\right)^{x_i}. \tag{147}$$

Vergleicht man diese Funktion nun mit der Wahrscheinlichkeitsfunktion der Negativ-Binomial-Verteilung

$$f(k; r, p) = \binom{k+r-1}{k}p^r(1-p)^k \tag{148}$$

$$= \binom{(r-1)+k}{r-1}p^r(1-p)^k = \frac{(r-1+k)!}{(r-1)!(r-1+k-(r-1))!}p^r(1-p)^k \tag{149}$$

$$= \frac{(r-1+k)!}{(r-1)!k!}p^r(1-p)^k = \frac{\Gamma(r+k)}{\Gamma(r)k!}p^r(1-p)^k, \tag{150}$$

so erkennt man sofort, dass die Form der beiden Gleichungen dieselbe ist.

Anschließend bestimmt man die log-Likelihood Funktion und leitet diese nach $\{\beta_j\}$ und ν ab um zu den ML-Gleichungen zu gelangen, welche wieder numerisch gelöst werden können.

Daher werden diese im nächsten Schritt bestimmt

$$P(Y_i = y_i) = P(X_i/\omega_i = y_i) = P(X_i = \omega_i y_i) = \frac{\Gamma(\omega_i\mu_i/\nu + \omega_i y_i)}{\Gamma(\omega_i\mu_i/\nu)(\omega_i y_i)!}\left(\frac{1}{\nu+1}\right)^{\omega_i\mu_i/\nu}\left(\frac{\nu}{\nu+1}\right)^{\omega_i y_i}. \tag{151}$$

Betrachte zuerst nur $\frac{\Gamma(\omega_i\mu_i/\nu+\omega_i y_i)}{\Gamma(\omega_i\mu_i/\nu)}$ und vereinfache, wobei man annimmt, dass $\omega_i\mu_i/\nu+\omega_i y_i$ aus den natürlichen Zahlen stammen,

$$\frac{\Gamma(\omega_i\mu_i/\nu + \omega_i y_i)}{\Gamma(\omega_i\mu_i/\nu)} = \frac{(\omega_i\mu_i/\nu + \omega_i y_i - 1)!}{(\omega_i\mu_i/\nu - 1)!} \tag{152}$$

$$= \frac{(\omega_i\mu_i/\nu + \omega_i y_i - 1)(\omega_i\mu_i/\nu + \omega_i y_i - 2)\cdots(\omega_i\mu_i/\nu + \omega_i y_i - \omega_i y_i)}{1} \tag{153}$$

$$= \prod_{k=1}^{\omega_i y_i}\left[\omega_i\mu_i/\nu + \omega_i y_i - k\right]. \tag{154}$$

$$P(Y_i = y_i) = \nu^{\omega_i y_i} \left(\prod_{k=1}^{\omega_i y_i} \omega_i \mu_i / \nu + \omega_i y_i - k \right) \left(\frac{1}{\nu+1} \right)^{\omega_i \mu_i / \nu} \left(\frac{1}{\nu+1} \right)^{\omega_i y_i} \frac{1}{(\omega_i y_i)!} \tag{155}$$

$$= \left(\prod_{k=1}^{\omega_i y_i} \omega_i \mu_i + \nu(\omega_i y_i - k) \right) \left(\frac{1}{\nu+1} \right)^{\omega_i \mu_i / \nu} \left(\frac{1}{\nu+1} \right)^{\omega_i y_i} \frac{1}{(\omega_i y_i)!}. \tag{156}$$

Für die log-Likelihood Gleichung gilt nun

$$l(\boldsymbol{y}; \boldsymbol{\mu}, \nu) = \sum_i \left\{ \sum_{k=1}^{\omega_i y_i} \log(\omega_i \mu_i + (k-1)\nu) - \left(\frac{\omega_i \mu_i}{\nu} + \omega_i y_i \right) \log(1+\nu) - \log((\omega_i y_i)!) \right\}, \tag{157}$$

da

$$\sum_{k=1}^{\omega_i y_i} \log(\omega_i \mu_i + (k-1)\nu) = \sum_{k=1}^{\omega_i y_i} \log(\omega_i \mu_i + \nu(\omega_i y_i - k) \tag{158}$$

$$= \sum_{i=0}^{\omega_i y_i - 1} \log(\omega_i \mu_i + i\nu) \tag{159}$$

$$= \sum_{j=1}^{\omega_i y_i} \log(\omega_i \mu_i + (j-1)\nu). \tag{160}$$

Setzt man, auf Grund des multiplikativen Modells, $\mu_i = e^{\eta_i}$ mit $\eta_i = \beta_1 x_{i1} + ... + \beta_r x_{ir}$ und leitet nach β_j beziehungsweise ν ab, so erhält man durch 0 setzen die ML-Gleichungen.

$$\frac{\partial l}{\partial \beta_j} = \sum_i \left(\sum_{k=1}^{\omega_i y_i} \frac{1}{\omega_i e^{\beta_1 x_{i1} + ... + \beta_r x_{ir}} + (k-1)\nu} \omega_i e^{\beta_1 x_{i1} + ... + \beta_r x_{ir}} x_{ij} \right) - \log(1+\nu) \frac{\omega_i e^{\beta_1 x_{i1} + ... + \beta_r x_{ir}}}{\nu} x_{ij} \tag{161}$$

$$= \sum_i \frac{\omega_i}{\nu} \left[\sum_{k=1}^{\omega_i y_i} \frac{\mu_i \nu}{\omega_i \mu_i + (k-1)\nu} x_{ij} - \log(1+\nu) \mu_i x_{ij} \right] \tag{162}$$

$$= \sum_i \omega_i \left[\sum_{k=1}^{\omega_i y_i} \frac{\mu_i \nu}{\omega_i \mu_i + (k-1)\nu} x_{ij} - \log(1+\nu) \mu_i x_{ij} \right] = 0. \tag{163}$$

$$\frac{\partial l}{\partial \nu} = \sum_i \sum_{k=1}^{\omega_i y_i} \frac{1}{\omega_i \mu_i + (k-1)\nu} (k-1) + \frac{\omega_i \mu_i}{\nu^2} \log(1+\nu) - \frac{1}{1+\nu} \left(\frac{\omega_i \mu_i}{\nu} + \omega_i y_i \right) = 0 \tag{164}$$

Die ML-Gleichung für ν sieht hier anders aus als in Ohlsson, Johansson [12, Seite 68].

5.2.2 Merkmalsauswahl

Da man nun ein passendes Modell zur Modellierung der Schadenfrequenz gefunden hat, ist es wieder an der Zeit die wesentlichen Merkmale zu bestimmen. Die Vorgehensweise ist analog zum Poisson-Modell.

Anschließend werden die einzelnen Ausprägungen aller relevanten Merkmale mit Hilfe der Funktion fit.contrast() auf Signifikanz überprüft und – falls erforderlich – zu gemeinsamen Ausprägungen zusammengefasst. Schlussendlich erhält man ein Modell, welches nur noch hoch-signifikante Merkmale und Merkmalsausprägungen aufweist. Zu beachten ist stets, dass bei Hinzunahme eines Merkmales, immer alle dazugehörigen Ausprägungen zum Modell hinzugefügt werden müssen. Es dürfen keine Ausprägungen eines Merkmals, welches als signifikant angesehen wird aus dem Modell entfernt werden.

```
glm.nb(formula = glm_data$Schadenanzahl ~ offset(log(glm_data$vers_jahre)) +
    glm_data$AkadTitel + glm_data$SonstTitel + glm_data$HUBRAU +
    glm_data$LEISTU + glm_data$vmerk_LEAS + glm_data$VNfamilienstand +
    glm_data$KundeSex + glm_data$vmerk_WECHS + glm_data$Konzern +
    glm_data$Bezirk + glm_data$alter_years + glm_data$vmerk_I3SFR +
    glm_data$Zugehoerigkeit, link = log, init.theta = 1.241004502)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0535  -0.3054  -0.2372  -0.1622   4.9331

Coefficients:
                                                      Estimate Std. Error  z value Pr(>|z|)
(Intercept)                                          -3.07424    0.02422 -126.936  < 2e-16 ***
glm_data$AkadTitelB/>=1                               0.08060    0.03153    2.556 0.010585 *
glm_data$SonstTitelB/>=1                              0.34957    0.09576    3.651 0.000262 ***
glm_data$HUBRAUA/ <1400                              -0.10586    0.02465   -4.294 1.75e-05 ***
glm_data$HUBRAUD/ >2000                               0.09505    0.03265    2.911 0.003601 **
glm_data$LEISTUA/<=50kW                              -0.09967    0.03002   -3.320 0.000901 ***
glm_data$LEISTUD/E >90kW                              0.08854    0.02639    3.355 0.000793 ***
glm_data$vmerk_LEASJ                                  0.19322    0.03425    5.641 1.69e-08 ***
glm_data$VNfamilienstandgeschieden/verwitwet/verheiratet  0.12399    0.02364    5.245 1.56e-07 ***
glm_data$KundeSexF/U/W                                0.08657    0.01967    4.400 1.08e-05 ***
glm_data$vmerk_WECHSJ                                 0.07641    0.03447    2.216 0.026672 *
glm_data$KonzernKlasse 1                              0.23701    0.05650    4.195 2.73e-05 ***
glm_data$KonzernKlasse 2                             -0.74269    0.18870   -3.936 8.29e-05 ***
glm_data$KonzernKlasse 3 und 4                        0.05429    0.02084    2.605 0.009196 **
glm_data$KonzernKlasse 5                             -0.13061    0.03655   -3.574 0.000352 ***
glm_data$BezirkKlasse 1                               0.14017    0.02461    5.696 1.23e-08 ***
glm_data$BezirkKlasse 3                              -0.29530    0.09894   -2.985 0.002839 **
glm_data$BezirkKlasse 4 und 5                         0.51535    0.04770   10.803  < 2e-16 ***
glm_data$BezirkKlasse 6 und 7                         0.34642    0.02401   14.428  < 2e-16 ***
glm_data$alter_years17-20                             0.94499    0.07908   11.949  < 2e-16 ***
glm_data$alter_years21-25                             0.35305    0.04246    8.316  < 2e-16 ***
glm_data$alter_years65-74                             0.16809    0.02987    5.627 1.83e-08 ***
glm_data$alter_years75-100                            0.54778    0.03200   17.118  < 2e-16 ***
glm_data$vmerk_I3SFRB/-2 bis 1                        0.17752    0.02336    7.599 2.98e-14 ***
glm_data$vmerk_I3SFRC/2 bis 5                         0.33155    0.02855   11.613  < 2e-16 ***
glm_data$vmerk_I3SFRD/E 6 bis 13                      0.61751    0.03157   19.561  < 2e-16 ***
glm_data$vmerk_I3SFRF/14 bis 17                       1.50564    0.21160    7.115 1.12e-12 ***
glm_data$ZugehoerigkeitMakler und unbekannt          -0.05159    0.02327   -2.217 0.026619 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(1.241) family taken to be 1)

    Null deviance: 73213  on 352910  degrees of freedom
Residual deviance: 71713  on 352883  degrees of freedom
AIC: 103516

Number of Fisher Scoring iterations: 1


              Theta:  1.241
          Std. Err.:  0.123

 2 x log-likelihood:  -103457.856
```

Abbildung 8: Summary optimales Negativ-Binomial-Modell

In der Summary sind nicht alle Schätzer aufgelistet, da alle Schätzer der Referenzklasse 0 sind und somit hier nicht ersichtlich. Die anderen Schätzer werden im Verhältnis zur Referenzklasse angegeben. Zur Erklärung wähle man das Merkmal Bonus-Malus-Stufe. Hier ist die Referenzklasse -6 bis -3, da diese Merkmalsausprägung in der Summary nicht ersichtlich ist. Die Schätzer der anderen Klassen sind alle positiv. Betrachtet man zum Beispiel die nächsthöhere Stufe mit -2 bis 1, so erkennt man, dass die erwartete Schadenanzahl bei steigender Bonus-Malus-Stufe ebenfalls steigt. Und zwar muss man exp(Estimate) berechnen um den prozentuellen Zuwachs (die prozentuelle Abnahme) bestimmen zu können. Das heißt für die Referenzklasse -6 bis -3 gilt 100 % ($\exp(0) = 1$). Vergleicht man diese dann mit der Klasse -2 bis 1 so erhält man $\exp(0{,}17752) = 1{,}1943$. Das bedeutet, dass die Schadenwahrscheinlichkeit um 19,43 % steigt, wenn man sich in dieser Klasse befindet. Daraus folgend steigt auch die Schadenfrequenz um 19,43 %, da der Bestand konstant ist.

Dass es sich bei diesem optimalen Modell tatsächlich um eine signifikante Verbesserung im Vergleich zum vollen Modell handelt, sieht man auch bei den Deviances. Die Null-Deviance steht hierbei für die Deviance im vollen Modell mit der Anzahl der dazugehörigen Freiheitsgrade. Die Residual-Deviance dementsprechend für das optimale Modell.

Hier ist die Null-Deviance 73213 bei 352910 Freiheitsgraden und die Residual Deviance 71713 bei 352883 Freiheitsgraden. Die Deviance sinkt somit um 73213 - 71713 = 1500 bei einem Verlust von 27 Freiheitsgraden. Das χ^2-Quantil mit $\alpha = 5$ % und 27 Freiheitsgraden lautet 40,11. Da 1500 deutlich größer als 40,11 ist, liefert auch der χ^2-Test einen p-Wert von 0, was für eine wesentliche Verbesserung spricht.

5.3 GLM-Analyse für die Schadenhöhe

Hierbei wird anstelle der Schadenanzahl, wie es bei der negativen Binomialverteilung der Fall war, der Durchschnittsschaden modelliert. Dabei werden nur jene Polizzen herangezogen, welche einen positiven Schaden aufweisen. Klarerweise verringert sich somit die Datenmenge beträchtlich, da in der Haftpflichtversicherung nur ungefähr jeder 14. Versicherungsnehmer einen Schaden herbeiführt.

Nach erfolgreicher Durchführung der Likelihood Ratio Tests, erhält man das folgende in diesem Sinn optimale Modell:

```
Single term deletions

Model:
mean_schaden ~ glm_data$vmerk_WECHS + glm_data$Konzern + glm_data$alter_years
                     Df Deviance    AIC scaled dev. Pr(>Chi)
<none>                    15796 208402
glm_data$vmerk_WECHS  1    15807 208405      5.1344 0.023457 *
glm_data$Konzern      5    15820 208403     10.8250 0.054962 .
glm_data$alter_years  7    15841 208409     20.5191 0.004551 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Abbildung 9: Optimales Gamma-Modell

Man erkennt hier sofort, dass die Schadenhöhe nur durch 3 Merkmale dominiert wird, in diesem Fall ob es sich um einen PKW mit Wechselkennzeichen handelt, von welchem Konzern das Fahrzeug stammt und wie alt der Versicherungsnehmer ist. Die Analyse gibt deutlich weniger Auskunft als jene für die Schadenfrequenz.

Nach weiterem Zusammenfassen der Ausprägungen einzelner Merkmale erhält man somit die endgültige Summary.

```
Call:
glm(formula = mean_schaden ~ +glm_data$vmerk_WECHS + glm_data$Konzern +
    glm_data$alter_years, family = Gamma(link = log), weights = glm_data$Schadenanzahl)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.9987  -0.9315  -0.4412   0.1781   6.2254

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                    7.47192    0.01736 430.468  < 2e-16 ***
glm_data$vmerk_WECHSJ          0.11235    0.04871   2.307 0.021090 *
glm_data$KonzernKlasse 2 und 3 -0.06955    0.02936  -2.369 0.017851 *
glm_data$alter_years17-30      0.14989    0.03971   3.775 0.000161 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 2.213398)

    Null deviance: 15881  on 11762  degrees of freedom
Residual deviance: 15825  on 11759  degrees of freedom
AIC: 208410

Number of Fisher Scoring iterations: 6
```

Abbildung 10: Summary optimales Gamma-Modell

Nun ist es möglich, die Nettoprämie genau zu bestimmen, da man nur die Schätzer der Schadenfrequenz mit den Schätzern der durchschnittlichen Schadenhöhe multiplizieren muss. Aufpassen muss man jedoch trotzdem noch, da die Schätzer aus der Summary zuerst noch exponenziert werden müssen. Eine detailierte Auflistung der Schätzer erfolgt am Ende, nachdem alle 3 Methoden zur Nettoprämienkalkulation vorgestellt wurden.

5.4 Direkte GLM-Analyse quasi-Poisson

An der Vorgehensweise ändert sich nichts im Vergleich zur Analyse für die Schadenfrequenz, nur dass man hier direkt zu den Schätzern für die Nettoprämie kommt. Das optimale Modell unterscheidet sich ein wenig von jenem der Schadenfrequenz, da hier weniger signifikante Merkmale ausfindig gemacht werden.

```
Model:
glm_data$Wirkschaden/glm_data$vers_jahre ~ glm_data$HUBRAU +
    glm_data$vmerk_LEAS + glm_data$vmerk_WECHS + glm_data$Konzern +
    glm_data$Bezirk + glm_data$alter_years + glm_data$vmerk_I3SFR
                    Df  Deviance scaled dev.  Pr(>Chi)
<none>                 169367621
glm_data$HUBRAU      3 169623538      26.045 9.333e-06 ***
glm_data$vmerk_LEAS  1 169449179       8.300  0.003964 **
glm_data$vmerk_WECHS 1 169417706       5.097  0.023964 *
glm_data$Konzern     5 169578283      21.439  0.000669 ***
glm_data$Bezirk      6 170097578      74.289 5.377e-14 ***
glm_data$alter_years 7 170294666      94.346 < 2.2e-16 ***
glm_data$vmerk_I3SFR 5 170326980      97.635 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Abbildung 11: Optimales quasi-Poisson-Modell

Nach weiterem Zusammenfassen der Ausprägungen einzelner Merkmale, erhält man wieder die endgültige Summary.

```
> summary(opt_model_direct)

Call:
glm(formula = glm_data$Wirkschaden/glm_data$vers_jahre ~ +glm_data$HUBRAU +
    glm_data$vmerk_LEAS + glm_data$vmerk_WECHS + glm_data$Konzern +
    glm_data$Bezirk + glm_data$alter_years + glm_data$vmerk_I3SFR,
    family = quasipoisson, weights = glm_data$vers_jahre)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-41.98  -12.89   -9.96   -6.76  755.96

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                      4.50833    0.04522  99.701  < 2e-16 ***
glm_data$HUBRAUA/B <1700        -0.17731    0.04691  -3.779 0.000157 ***
glm_data$HUBRAUD/ >2000          0.15875    0.06752   2.351 0.018716 *
glm_data$vmerk_LEASJ             0.24539    0.07728   3.175 0.001497 **
glm_data$vmerk_WECHSJ            0.17822    0.07616   2.340 0.019278 *
glm_data$KonzernKlasse 1         0.32623    0.11888   2.744 0.006064 **
glm_data$KonzernKlasse 2        -1.04190    0.49725  -2.095 0.036141 *
glm_data$KonzernKlasse 5        -0.21459    0.08459  -2.537 0.011183 *
glm_data$BezirkKlasse 1          0.18250    0.05598   3.260 0.001114 **
glm_data$BezirkKlasse 3         -0.51336    0.25513  -2.012 0.044202 *
glm_data$BezirkKlasse 4,5 und 6  0.59142    0.09151   6.463 1.03e-10 ***
glm_data$BezirkKlasse 7          0.35881    0.05451   6.582 4.65e-11 ***
glm_data$alter_years17-20        1.09212    0.15832   6.898 5.27e-12 ***
glm_data$alter_years21-25        0.44242    0.09005   4.913 8.96e-07 ***
glm_data$alter_years75-100       0.49212    0.07374   6.674 2.50e-11 ***
glm_data$vmerk_I3SFRB/-2 bis 1   0.17707    0.05423   3.265 0.001095 **
glm_data$vmerk_I3SFRC/2 bis 5    0.38182    0.06434   5.934 2.96e-09 ***
glm_data$vmerk_I3SFRD/E/F 6 bis 17  0.66753    0.06948   9.607  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasipoisson family taken to be 9851.86)

    Null deviance: 173075162  on 352910  degrees of freedom
Residual deviance: 169462018  on 352893  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 9
```

Abbildung 12: Summary optimales quasi-Poisson-Modell

5.5 Direkte GLM-Analyse Tweedie

Ein Tweedie-Modell für zusammengesetzt Poisson-verteilte Zufallsvariablen muss ein $p \in (1, 2)$ haben. Daher wird hier mit der Bestimmung des Parameters p begonnen. Dies erfolgt nach Dunn [3] mit der Funktion tweedie.profile().

Dabei wird jener Wert von p gewählt, welcher den Maximum Likelihood Wert maximiert. Dies ist im nachstehendem Plot schön ersichtlich. Es werden dabei alle Punkte zwischen 1,1 und 1,9 betrachtet. Der Grund, dass die Funktion nur bis 1,7 eingezeichnet ist liegt in der Konvergenz der Funktion tweedie.profile(), da die Funktion in diesem Fall für p-Werte größer 1,7, auf Grund numerischer Instabilität nicht mehr konvergiert und der Maximum Likelihood Wert somit $-\infty$ ist.

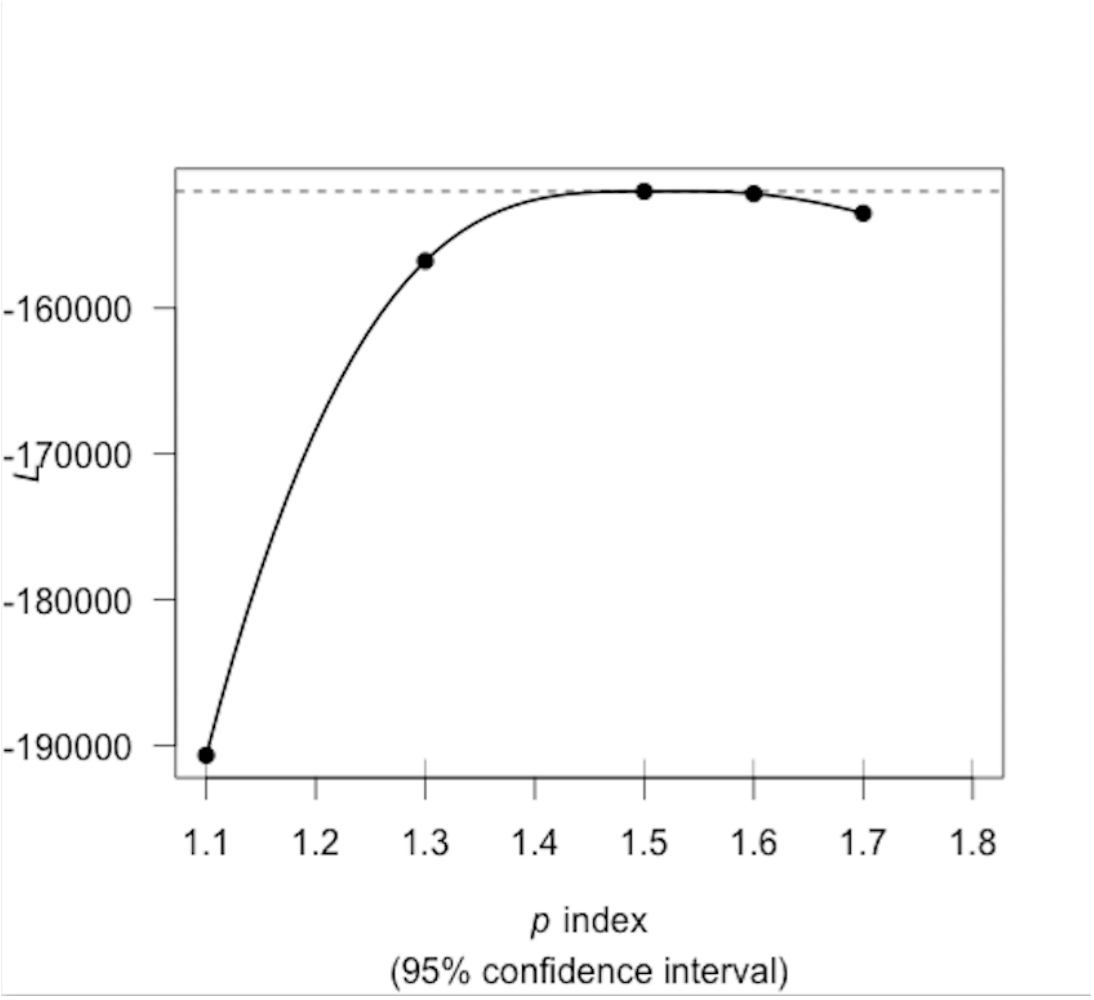


Abbildung 13: Parameter p Tweedie-Modell

Da nun ein Modell mit dem gefundenen p-Wert aufgestellt werden kann, ist es an der Zeit, nicht-signifikante Merkmale aus dem Modell zu entfernen. Dies geschieht wieder mit dem drop1() Befehl in R. Es folgt somit das in diesem Sinne optimale Tweedie-Modell.

```
Model:
glm_data$Wirkschaden/glm_data$vers_jahre ~ glm_data$HUBRAU +
    glm_data$vmerk_LEAS + glm_data$vmerk_WECHS + glm_data$Konzern +
    glm_data$Bezirk + glm_data$alter_years + glm_data$vmerk_I3SFR
                     Df Deviance scaled dev.  Pr(>Chi)
<none>                  15098267
glm_data$HUBRAU       3 15124735      23.595 3.034e-05 ***
glm_data$vmerk_LEAS   1 15106622       7.448  0.006349 **
glm_data$vmerk_WECHS  1 15103431       4.604  0.031907 *
glm_data$Konzern      5 15117832      17.442  0.003734 **
glm_data$Bezirk       6 15172772      66.420 2.212e-12 ***
glm_data$alter_years  7 15189693      81.505 6.792e-15 ***
glm_data$vmerk_I3SFR  5 15196142      87.254 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Abbildung 14: Optimales Tweedie-Modell

Nach weiterem Zusammenfassen der Ausprägungen einzelner Merkmale, erhält man wieder die endgültige Summary.

```
> summary(opt_model_tweedie)

Call:
glm(formula = glm_data$Wirkschaden/glm_data$vers_jahre ~ +glm_data$HUBRAU +
    glm_data$vmerk_LEAS + glm_data$vmerk_WECHS + glm_data$Konzern +
    glm_data$Bezirk + glm_data$alter_years + glm_data$vmerk_I3SFR,
    family = tweedie(var.power = tweedie.mle$p.max, link.power = 0),
    weights = glm_data$vers_jahre)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-11.992   -5.923   -4.607   -3.145  161.712

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                       4.50742    0.04500 100.169  < 2e-16 ***
glm_data$HUBRAUA/B <1700         -0.17160    0.04799  -3.575  0.00035 ***
glm_data$HUBRAUD/ >2000           0.16511    0.07254   2.276  0.02284 *
glm_data$vmerk_LEASJ              0.24328    0.08296   2.933  0.00336 **
glm_data$vmerk_WECHSJ             0.17563    0.08047   2.183  0.02907 *
glm_data$KonzernKlasse 1          0.30556    0.13666   2.236  0.02536 *
glm_data$KonzernKlasse 2         -0.98902    0.40950  -2.415  0.01573 *
glm_data$KonzernKlasse 5         -0.18783    0.08219  -2.285  0.02230 *
glm_data$BezirkKlasse 1           0.18664    0.05763   3.239  0.00120 **
glm_data$BezirkKlasse 3          -0.49994    0.22661  -2.206  0.02737 *
glm_data$BezirkKlasse 4,5 und 6   0.57799    0.10529   5.489 4.04e-08 ***
glm_data$BezirkKlasse 7           0.35758    0.05767   6.201 5.63e-10 ***
glm_data$alter_years17-20         1.09440    0.21604   5.066 4.07e-07 ***
glm_data$alter_years21-25         0.45326    0.10280   4.409 1.04e-05 ***
glm_data$alter_years75-100        0.49300    0.07947   6.204 5.52e-10 ***
glm_data$vmerk_I3SFRB/-2 bis 1    0.16402    0.05535   2.963  0.00304 **
glm_data$vmerk_I3SFRC/2 bis 5     0.37815    0.06754   5.599 2.16e-08 ***
glm_data$vmerk_I3SFRD/E/F 6 bis 17 0.66278   0.07778   8.521  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Tweedie family taken to be 1124.871)

    Null deviance: 15474465  on 352910  degrees of freedom
Residual deviance: 15106878  on 352893  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 7
```

Abbildung 15: Summary optimales Tweedie-Modell

6 Clusterings und Ergebnisse der einzelnen Versicherungssparten

6.1 Haftpflichtversicherung

Ward-Clustering der Bezirke sowie der Autokonzerne:

Klasse	Bezirk
1	Krems (Land), Neunkirchen, Innsbruck-Land, Graz-Umgebung, Amstetten, Feldkirchen, Salzburg (Stadt), Klagenfurt-Land, Villach (Stadt), Kufstein, Schwaz, Baden, Zwettl
2	Melk, Steyr-Land, Hermagor, Bruck an der Leitha, Urfahr-Umgebung, Tulln, Weiz, Villach-Land, Spittal an der Drau, Wolfsberg, Gmunden, Graz (Stadt, Wiener Neustadt (Land), Sankt Veit an der Glan, Völkermarkt
3	Murau, Reutte, Lienz, Vöcklabruck, Waidhofen an der Thaya, Liezen, Voitsberg, Kitzbühel, Wels-Land, Freistadt, Murtal
4	Innsbruck-Stadt
5	Imst, Gänserndorf, Wien
6	Eisenstadt-Umgebung, Mödling, Deutschlandsberg, Krems an der Donau, Linz-Land, Korneuburg, Wien-Umgebung
7	Sankt Pölten (Land), Salzburg-Umgebung, Sonstige, Scheibbs, Klagenfurt (Stadt), Oberwart

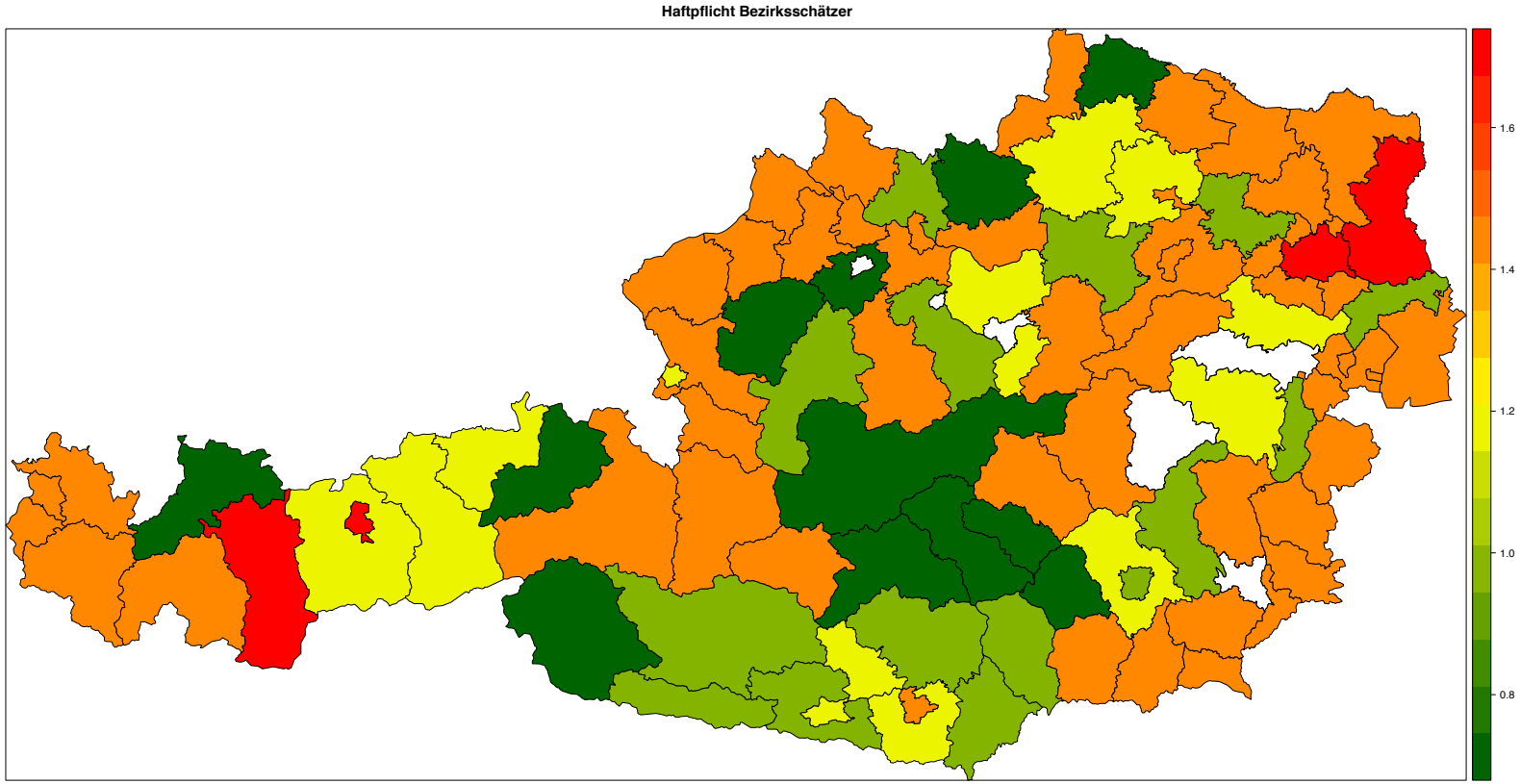
Tabelle 7: Clustering der Bezirke

Klasse	Konzern
1	Mini, Volvo, Rover, Chrysler, Jeep
2	Puch, Daihatsu, Porsche
3	Mazda, Mercedes, Toyota, Ford, Chevrolet, Alfa-Romeo, Audi, Mitsubishi
4	Honda, Seat, Jaguar, Sonstige, BMW, Daewoo
5	Manuell, Suzuki, Fiat, Subaru
6	Nissan, Skoda, Dacia, Kia, Lancia, Lexus, Peugeot, Renault, Saab, VW, Opel Citroen, Hyundai

Tabelle 8: Clustering der Konzerne

Haftpflicht	**Schätzer getrennt**	*untere Schranke (getrennt)*	*obere Schranke (getrennt)*	**Schätzer quasi-Poisson**	**Schätzer Tweedie**
(Intercept)	**81,26**	*76,65*	*86,15*	**90,77**	**90,69**
Anzahl Akademischer Titel					
keine	**1,00**	*1,00*	*1,00*	**1,00**	**1,00**
mindestens 1	**1,08**	*1,02*	*1,15*	**1,00**	**1,00**
Anzahl Sonstiger Titel					
keine	**1,00**	*1,00*	*1,00*	**1,00**	**1,00**
mindestens 1	**1,42**	*1,18*	*1,71*	**1,00**	**1,00**
Hubraum					
<1400ccm	**0,90**	*0,86*	*0,94*	**0,84**	**0,84**
>=1400 <1700ccm	**1,00**	*1,00*	*1,00*	**0,84**	**0,84**
>=1700 <2000ccm	**1,00**	*1,00*	*1,00*	**1,00**	**1,00**
>2000cmm	**1,10**	*1,03*	*1,17*	**1,17**	**1,18**
Leistung					
<=50kW	**0,91**	*0,85*	*0,96*	**1,00**	**1,00**
>50 <= 70kW	**1,00**	*1,00*	*1,00*	**1,00**	**1,00**
>70 <=90kW	**1,00**	*1,00*	*1,00*	**1,00**	**1,00**
>90 <= 110kW	**1,09**	*1,04*	*1,15*	**1,00**	**1,00**
>110kW	**1,09**	*1,04*	*1,15*	**1,00**	**1,00**
Leasingfahrzeug					
Nein	**1,00**	*1,00*	*1,00*	**1,00**	**1,00**
Ja	**1,21**	*1,13*	*1,30*	**1,28**	**1,28**
Familienstand					
geschieden	**1,13**	*1,08*	*1,19*	**1,00**	**1,00**
verwitwet	**1,13**	*1,08*	*1,19*	**1,00**	**1,00**
verheiratet	**1,13**	*1,08*	*1,19*	**1,00**	**1,00**
unbekannt	**1,00**	*1,00*	*1,00*	**1,00**	**1,00**
verstorben	**1,00**	*1,00*	*1,00*	**1,00**	**1,00**
Lebensgemeinschaft	**1,00**	*1,00*	*1,00*	**1,00**	**1,00**
ledig	**1,00**	*1,00*	*1,00*	**1,00**	**1,00**
Geschlecht					
Firma	**1,09**	*1,05*	*1,13*	**1,00**	**1,00**
unbekannt	**1,09**	*1,05*	*1,13*	**1,00**	**1,00**
männlich	**1,00**	*1,00*	*1,00*	**1,00**	**1,00**
weiblich	**1,09**	*1,05*	*1,13*	**1,00**	**1,00**
Wechselkennzeichen					
Nein	**1,00**	*1,00*	*1,00*	**1,00**	**1,00**
Ja	**1,21**	*1,07*	*1,36*	**1,20**	**1,19**
Konzernklassen					
1 (Mini, Volvo, Rover, Chrysler, Jeep)	**1,27**	*1,13*	*1,42*	**1,39**	**1,36**
2 (Puch, Daihatsu, Porsche)	**0,44**	*0,31*	*0,65*	**0,35**	**0,37**
3 (Mazda, Mercedes, Toyota, Ford, Chevrolet, Alfa Romea, Audi, Mitsubishi)	**0,98**	*0,92*	*1,06*	**1,00**	**1,00**
4 (Honda, Seat, Jaguar, Sonstige, BMW, Daewoo)	**1,06**	*1,01*	*1,10*	**1,00**	**1,00**
5 (Manuell, Suzuki, Fiat, Subaru)	**0,88**	*0,82*	*0,94*	**0,81**	**0,83**
6 (Nissan, Skoda, Dacia, Kia, Lancia, Lexus, Peugeot, Renault, VW, Opel, Citroen, Hyundai)	**1,00**	*1,00*	*1,00*	**1,00**	**1,00**

Bezirksklassen					
1 (Krems (Land), Neunkirchen, Innsbruck-Land, Graz-Umgebung, Amstetten, Feldkirchen, Salzburg (Stadt), Klagenfurt Land, Villach) (Stadt), Kufstein, Schwaz, Baden, Zwettl)	**1,15**	*1,10*	*1,21*	**1,20**	**1,21**
2 (Melk, Steyr-Land, Hermagor, Bruck an der Leitha, Urfahr-Umgebung, Tulln, Weiz, Villach Land , Spittal an der Drau, Wolfsberg, Gmunden, Graz (Stadt), Wiener Neustadt (Land), Sankt Veit an der Glan, Voelkermarkt)	**1,00**	*1,00*	*1,00*	**1,00**	**1,00**
3 (Murau, Reutte, Lienz, Voecklabruck, Waidhofen an der Thaya, Liezen, Voitsberg, Kitzbuehl, Wels-Land, Freistadt, Murtal)	**0,74**	*0,61*	*0,90*	**0,60**	**0,61**
4 (Innsbruck-Stadt)	**1,67**	*1,52*	*1,84*	**1,81**	**1,78**
5 (Imst, Gaenserndorf, Wien)	**1,67**	*1,52*	*1,84*	**1,81**	**1,78**
6 (Eisenstadt-Umgebung, Moedling, Deutschlandsberg, Krems an der Donau, Linz-Land, Korneuburg, Wien-Umgebung)	**1,41**	*1,35*	*1,48*	**1,81**	**1,78**
7 (Sankt Poelten (Land), Salzburg-Umgebung, Sonstige, Scheibbs, Klagenfurt (Stadt), Oberwart)	**1,41**	*1,35*	*1,48*	**1,43**	**1,43**
Altersklassen					
>100 und Firma	**1,00**	*1,00*	*1,00*	**1,00**	**1,00**
17-20	**2,99**	*2,51*	*3,55*	**2,98**	**2,99**
21-25	**1,65**	*1,48*	*1,85*	**1,56**	**1,57**
26-30	**1,16**	*1,07*	*1,26*	**1,00**	**1,00**
31-49	**1,00**	*1,00*	*1,00*	**1,00**	**1,00**
50-64	**1,00**	*1,00*	*1,00*	**1,00**	**1,00**
65-74	**1,18**	*1,12*	*1,25*	**1,00**	**1,00**
75-100	**1,73**	*1,62*	*1,84*	**1,64**	**1,64**
Bonus Malus Stufen					
-6 bis -3	**1,00**	*1,00*	*1,00*	**1,00**	**1,00**
-2 bis 1	**1,19**	*1,14*	*1,25*	**1,19**	**1,18**
2 bis 5	**1,39**	*1,32*	*1,47*	**1,46**	**1,46**
6 bis 9	**1,85**	*1,74*	*1,97*	**1,95**	**1,94**
10 bis 13	**1,85**	*1,74*	*1,97*	**1,95**	**1,94**
14 bis 17	**4,51**	*2,98*	*6,82*	**1,95**	**1,94**
Zugehörigkeit					
HBM	**1,00**	*1,00*	*1,00*	**1,00**	**1,00**
Makler	**0,95**	*0,91*	*0,99*	**1,00**	**1,00**
IGV	**1,00**	*1,00*	*1,00*	**1,00**	**1,00**
unbekannt	**0,95**	*0,91*	*0,99*	**1,00**	**1,00**


Haftpflicht Bezirksschätzer
1.6
1.4
1.2
1.0
0.8

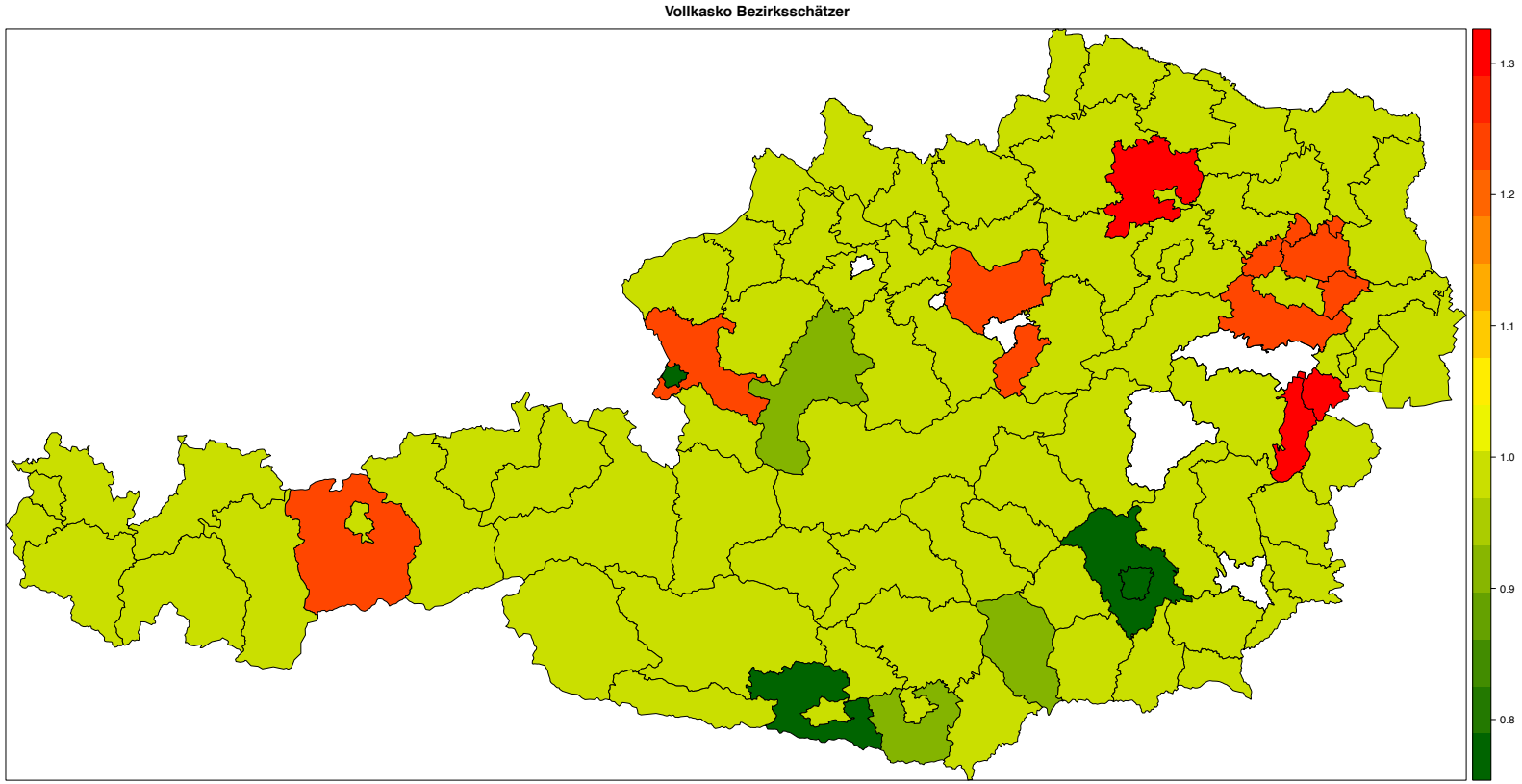
6.2 Vollkaskoversicherung

Ward-Clustering der Bezirke sowie der Autokonzerne:

Klasse	Bezirk
1	Wiener Neustadt (Land), Krems (Land), Mattersburg
2	Wien
3	Salzburg-Umgebung, Amstetten, Innsbruck-Land, Baden, Wien-Umgebung
4	Salzburg (Stadt), Graz (Stadt), Graz-Umgebung, Villach Land
5	Klagenfurt Land, Gmunden, Wolfsberg
6	Klagenfurt (Stadt), Mödling, Hermagor, Völkermarkt
7	Spittal an der Drau, Lienz, Sonstige, Feldkirchen, Sankt Veit an der Glan, Villach (Stadt)

Klasse	Konzern
1	Fiat
2	Manuell, Chevrolet, Skoda, Nissan, Renault, Kia, Dacia, Peugeot
3	Toyota, Mazda, Mitsubishi
4	Hyundai, Suzuki, Seat, VW, Ford, Opel, Alfa-Romeo, Honda
5	Volvo, Citroen, Subaru
6	Jeep, Mini, Mercedes, Audi, Lancia
7	Sonstige, BMW, Rover

Vollkasko	**Schätzer getrennt**	*untere Schranke (getrennt)*	*obere Schranke (getrennt)*	**Schätzer quasi-Poisson**	**Schätzer Tweedie**
(Intercept)	**381,86**	*357,79*	*407,55*	**363,94**	**370,69**
Anzahl Akademischer Titel					
keine	**1,00**	*1,00*	*1,00*	**1,00**	**1,00**
mindestens 1	**0,98**	*0,91*	*1,06*	**1,00**	**1,00**
Anzahl Sonstiger Titel					
keine	**1,00**	*1,00*	*1,00*	**1,00**	**1,00**
mindestens 1	**1,23**	*1,05*	*1,44*	**1,40**	**1,00**
Leistung					
<=50kW	**0,80**	*0,71*	*0,89*	**1,00**	**0,84**
>50 <= 70kW	**1,00**	*1,00*	*1,00*	**1,00**	**1,00**
>70 <=90kW	**1,14**	*1,14*	*1,14*	**1,24**	**1,21**
>90 <= 110kW	**1,32**	*1,25*	*1,39*	**1,43**	**1,41**
>110kW	**1,60**	*1,47*	*1,74*	**1,80**	**1,78**
Geschlecht					
Firma	**1,00**	*1,00*	*1,00*	**1,00**	**1,00**
unbekannt	**1,18**	*1,06*	*1,31*	**1,00**	**1,00**
männlich	**1,00**	*1,00*	*1,00*	**1,00**	**1,00**
weiblich	**1,00**	*1,00*	*1,00*	**1,00**	**1,00**
Antriebsart					
Diesel	**1,00**	*1,00*	*1,00*	**1,00**	**1,00**
Benzin	**0,84**	*0,81*	*0,88*	**0,82**	**0,83**
Elektro	**0,41**	*0,23*	*0,73*	**0,82**	**0,83**
Sonstige	**0,41**	*0,23*	*0,73*	**0,82**	**0,83**
Konzernklassen					
1 (Fiat)	**0,89**	*0,85*	*0,93*	**0,85**	**0,86**
2 (Manuell, Chevrolet, Skoda, Nissan, Renault, Kia, Dacia, Peugeot)	**0,89**	*0,85*	*0,93*	**0,85**	**0,86**
3 (Toyota, Mazda, Mitsubishi)	**1,00**	*1,00*	*1,00*	**1,00**	**1,00**
4 (Hyundai, Suzuki, Seat, VW, Ford, Opel, Alfa Romeo, Honda)	**1,00**	*1,00*	*1,00*	**1,00**	**1,00**
5 (Volvo, Citroen, Subaru)	**1,06**	*1,01*	*1,11*	**1,00**	**1,00**
6 (Jeep, Mini, Mercedes, Audi, Lancia)	**1,06**	*1,01*	*1,11*	**1,00**	**1,00**
7 (Sonstige, BMW, Rover)	**1,14**	*1,07*	*1,23*	**1,00**	**1,00**
Bezirksklassen					
1 (Wiener Neustadt (Land), Krems (Land), Mattersburg)	**1,29**	*1,13*	*1,47*	**1,31**	**1,31**
2 (Wien)	**1,23**	*1,13*	*1,34*	**1,31**	**1,31**
3 (Salzburg-Umgebung, Amstetten, Innsbruck-Land, Baden, Wien-Umgebung)	**1,23**	*1,13*	*1,34*	**1,31**	**1,31**
4 (Salzburg (Stadt), Graz (Stadt), Graz-Umgebung, Villach Land)	**0,79**	*0,74*	*0,85*	**0,86**	**0,87**
5 (Klagenfurt Land, Gmunden, Wolfsberg)	**0,90**	*0,86*	*0,94*	**0,86**	**0,87**
6 (Klagenfurt (Stadt), Moedling, Hermagor, Voelkermarkt)	**1,00**	*1,00*	*1,00*	**1,00**	**1,00**
7 (Spittal an der Drau, Lienz, Sonstige, Feldkirchen, Sankt Veit an der Glan, Villach (Stadt))	**0,99**	*0,93*	*1,05*	**1,00**	**1,00**
Altersklassen					
>100 und Firma	**1,21**	*1,12*	*1,32*	**1,42**	**1,41**
17-20	**3,72**	*2,45*	*5,66*	**3,42**	**3,63**
21-25	**1,37**	*1,25*	*1,50*	**1,69**	**1,69**
26-30	**1,09**	*1,01*	*1,17*	**1,00**	**1,00**
31-49	**1,00**	*1,00*	*1,00*	**1,00**	**1,00**
50-64	**1,00**	*1,00*	*1,00*	**1,00**	**1,00**
65-74	**1,11**	*1,06*	*1,17*	**1,00**	**1,00**
75-100	**1,29**	*1,18*	*1,41*	**1,36**	**1,40**
Zugehörigkeit					
HBM	**1,00**	*1,00*	*1,00*	**1,00**	**1,00**
Makler	**1,00**	*1,00*	*1,00*	**1,00**	**1,00**
IGV	**1,08**	*1,01*	*1,15*	**1,00**	**1,00**
unbekannt	**1,14**	*1,08*	*1,20*	**1,00**	**1,00**


Vollkasko Bezirksschätzer
1.3
1.2
1.1
1.0
0.9
0.8

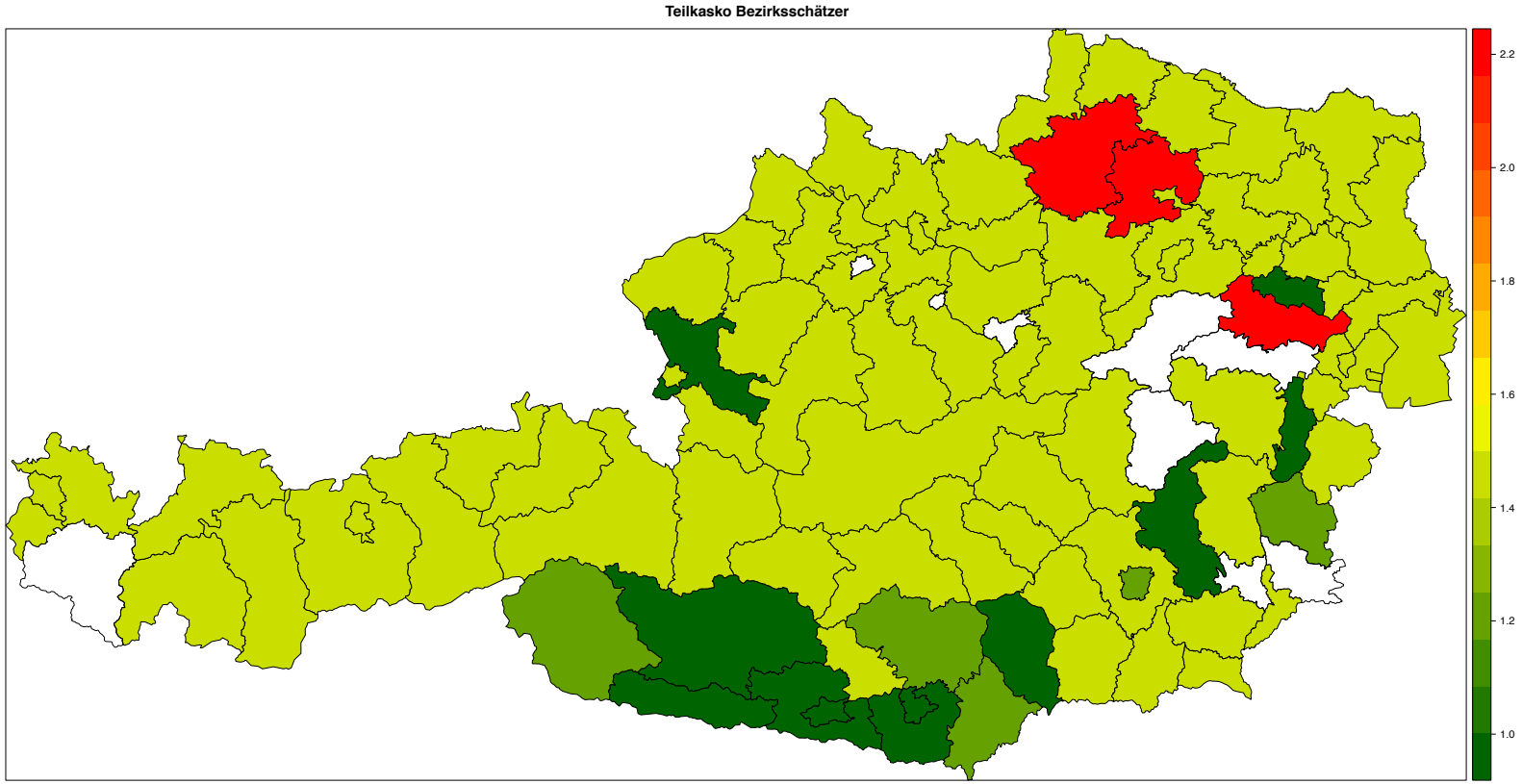
6.3 Teilkaskoversicherung

Ward-Clustering der Bezirke sowie der Autokonzerne:

Klasse	Bezirk
1	Graz (Stadt), Oberwart, Lienz, Sankt Veit an der Glan, Völkermarkt
2	Sonstige, Amstetten, Feldkirchen
3	Salzburg-Umgebung, Weiz
4	Spittal an der Drau, Hermagor, Wolfsberg, Klagenfurt Land, Villach Land, Villach (Stadt), Wiener Neustadt (Land), Klagenfurt (Stadt), Mödling
5	Zwettl
6	Baden, Krems (Land)
7	Wien-Umgebung, Graz-Umgebung, Wien

Klasse	Konzern
1	Rover, Volvo, Audi, BMW
2	Subaru, Dacia, Lancia
3	Toyota, Fiat, Kia, Mini, Renault
4	Hyundai, Chevrolet, Citroen, Suzuki
5	Sonstige, Alfa-Romeo, Honda
6	Chrysler, Mazda, VW, Mitsubishi, Opel
7	Jeep, Seat, Mercedes, Skoda
8	Ford, Manuell, Nissan, Peugeot

Teilkasko	**Schätzer getrennt**	*untere Schranke (getrennt)*	*obere Schranke (getrennt)*	**Schätzer quasi-Poisson**	**Schätzer Tweedie**
(Intercept)	**134,21**	*119,72*	*150,46*	**119,76**	**121,02**
Hubraum					
<1400ccm	**1,24**	*1,12*	*1,38*	**1,00**	**1,00**
>=1400 <1700ccm	**1,24**	*1,12*	*1,38*	**1,00**	**1,00**
>=1700 <2000ccm	**1,00**	*1,00*	*1,00*	**1,00**	**1,00**
>2000cmm	**1,00**	*1,00*	*1,00*	**1,00**	**1,00**
Leistung					
<=50kW	**0,64**	*0,53*	*0,77*	**0,75**	**0,76**
>50 <= 70kW	**0,91**	*0,83*	*1,00*	**1,00**	**1,00**
>70 <=90kW	**1,00**	*1,00*	*1,00*	**1,00**	**1,00**
>90 <= 110kW	**1,12**	*1,01*	*1,24*	**1,19**	**1,18**
>110kW	**1,75**	*1,48*	*2,07*	**1,69**	**1,74**
Geschlecht					
Firma	**1,00**	*1,00*	*1,00*	**1,00**	**1,00**
unbekannt	**1,00**	*1,00*	*1,00*	**1,00**	**1,00**
männlich	**1,00**	*1,00*	*1,00*	**1,00**	**1,00**
weiblich	**0,78**	*0,69*	*0,88*	**0,79**	**0,79**
Antriebsart					
Diesel	**1,00**	*1,00*	*1,00*	**1,00**	**1,00**
Benzin	**0,75**	*0,68*	*0,82*	**0,78**	**0,78**
Elektro	**0,75**	*0,68*	*0,82*	**0,78**	**0,78**
Sonstige	**0,75**	*0,68*	*0,82*	**0,78**	**0,78**
Konzernklassen					
1 (Rover, Volvo, Audi, BMW)	**1,10**	*0,99*	*1,22*	**1,19**	**1,17**
2 (Subaru, Dacia, Lancia)	**0,73**	*0,66*	*0,81*	**0,71**	**0,37**
3 (Toyota, Fiat, Kia, Mini, Renault)	**0,73**	*0,66*	*0,81*	**0,71**	**0,72**
4 (Hyundai, Chevrolet, Citroen, Suzuki)	**0,73**	*0,66*	*0,81*	**0,71**	**0,72**
5 (Sonstige, Alfa Romeo, Honda)	**1,00**	*1,00*	*1,00*	**1,00**	**1,00**
6 (Chrysler, Mazda, VW, Mitsubishi, Opel)	**1,00**	*1,00*	*1,00*	**1,00**	**1,00**
7 (Jeep, Seat, Mercedes, Skoda)	**0,82**	*0,74*	*0,90*	**0,86**	**0,83**
8 (Ford, Manuell, Nissan, Peugeot)	**0,82**	*0,74*	*0,90*	**0,86**	**0,83**
Bezirksklassen					
1 (Graz (Stadt), Oberwart, Lienz, Sankt Veit an der Glan, Voelkermarkt)	**1,21**	*1,12*	*1,31*	**1,38**	**1,39**
2 (Sonstige, Amstetten, Feldkirchen)	**1,42**	*1,28*	*1,58*	**1,38**	**1,39**
3 (Salzburg-Umgebung, Weiz)	**1,00**	*1,00*	*1,00*	**1,00**	**1,00**
4 (Spittal an der Drau, Hermagor, Wolfsberg, Klagenfurt Land, Villach Land, Villach (Stadt), Wiener Neustadt (Land), Klagenfurt (Stadt), Moedling)	**1,00**	*1,00*	*1,00*	**1,00**	**1,00**
5 (Zwettl)	**2,16**	*1,67*	*2,79*	**2,57**	**2,85**
6 (Baden, Krems (Land))	**2,16**	*1,67*	*2,79*	**2,57**	**2,85**
7 (Wien-Umgebung,Graz-Umgebung, Wien)	**1,42**	*1,28*	*1,58*	**1,38**	**1,39**
Altersklassen					
>100 und Firma	**1,00**	*1,00*	*1,00*	**1,53**	**1,00**
17-20	**1,43**	*1,22*	*1,68*	**1,53**	**1,66**
21-25	**1,43**	*1,22*	*1,68*	**1,53**	**1,66**
26-30	**1,00**	*1,00*	*1,00*	**1,00**	**1,00**
31-49	**1,00**	*1,00*	*1,00*	**1,00**	**1,00**
50-64	**0,72**	*0,64*	*0,81*	**0,74**	**0,75**
65-74	**0,64**	*0,56*	*0,74*	**0,58**	**0,59**
75-100	**0,50**	*0,40*	*0,62*	**0,58**	**0,59**


Teilkasko Bezirksschätzer
2.2
2.0
1.8
1.6
1.4
1.2
1.0

7 Interpretation der Ergebnisse

Für die Analyse wurden Großschäden ausgenommen, das heißt, dass alle Schäden größer als 50.000 € aus dem Datensatz entfernt wurden. Für diese müsste man spezielle Verfahren mit Extremwertverteilungen anwenden, um das sogenannte Großschadenloading berechnen zu können. Dabei wäre es mit der hier verwendeten Methode vermutlich problematisch geworden, da zu wenige Großschäden vorhanden waren.

Bei der ersten Methode ist deutlich erkennbar, dass es sich hierbei um die detailierteste Analyse handelt, da es am meisten Schätzer ungleich 1 gibt. Dies bestätigen auch die Schätzer für die Nettoprämie aus Kapitel 3. Der größere Arbeitsaufwand durch die getrennte Modellierung von Schadenfrequenz und Durchschnittsschadenhöhe wird somit mit mehr Information belohnt. Es können nämlich auch Aussagen rein über die Schadenfrequenz und die Durchschnittsschadenhöhe getroffen werden. Dies ist bei den direkten Methoden nicht möglich.

Die Schätzer selbst unterscheiden sich meist nur geringfügig von jenen der beiden anderen Methoden (Schätzer der direkten Methoden sind meist innerhalb des Konfidenzintervalls beziehungsweise nahe dran). Dort, wo sich die Schätzer gravierend unterscheiden, liegt nur ein kleiner Bestand einer Ausprägung eines bestimmten Merkmals vor, was zu größeren Schwankungen der Schätzer führt. Daher ist es – wie zu Beginn erwähnt – von großer Bedeutung jene Ausprägungen als Referenzklassen zu definieren, welche den größten Bestand aufweisen. Die Aussagekraft der Schätzer von Ausprägungen mit geringem Bestand ist somit sehr gering.

Bei den direkten Methoden werden diese Ausprägungsklassen daher meist überhaupt nicht separiert behandelt, was in manchen Fällen dafür Vorteile bringt. Ein Beispiel aus der Haftpflichtversicherung sind die Bonus-Malus Stufen 14-17. In dieser Klasse befinden sich kaum Versicherungsnehmer. Die erste Methode liefert hier einen Schätzer von 4,51, was der viereinhalbfachen Prämie entsprechen würde. Dies ist aber total unrealistisch. Die Schätzer der beiden direkten Methoden sind hingegen 1,95 und 1,94, welche aus Sicht der Versicherung durchaus erklärbar sind. Die doppelte Prämie für eine Person in dieser Bonus-Malus-Stufe wäre vertretbar und nicht so unrealistisch.

Interessant ist, welche Merkmale bei welcher Methode als signifikant auftreten. Die beiden direkten Methoden liefern hier, bis auf die Vollkaskoversicherung, bei der die Anzahl der sonstigen Titel im Tweedie-Modell als nicht signifikant angesehen wird, immer diesselben Merkmale, auch wenn die Schätzer minimal verschieden sind. Es sind aber bei allen drei Versicherungssparten weniger signifikante Merkmale, als bei der getrennten Methode.

Bei den direkten Methoden kommt es auch häufiger vor, dass unterschiedliche Merkmalsausprägungen exakt den gleichen Schätzer haben. Dies ist kein Zufall, sondern liegt daran, dass diese Merkmalsausprägungen zusammengefasst wurden, da sie in der jeweiligen Methode als nicht signifikant unterschiedlich erkannt wurden.

7.1 Detailierte Betrachtung der einzelnen Merkmale

Bei den nachfolgenden Betrachtungen und Erklärungen beziehe ich mich immer auf die Ergebnisse der getrennten Modellierung, da diese am detailiertesten sind. Die Schätzer der anderen beiden Methoden befinden sich meist noch im Konfidenzintervall dieser, beziehungsweise nahe daran. Der Grund, trotzdem eine direkte Analyse durchzuführen liegt in erster Linie daran, Sicherheit zu erhalten und wie bereits zuvor erwähnt in der geringeren Sensibilität gegenüber Merkmalsausprägungen mit geringem Bestand. Liefern alle Methoden, bei Merkmalen mit ausreichend großem Bestand, ähnliche Ergebnisse, so kann man davon ausgehen, dass die Ergebnisse passen.

Interessant sind die Ergebnisse der Versicherungssparten bezüglich des Geschlechts. Bei der Haftpflichtversicherung ist erkenntlich, dass Frauen eine um 9 % höhere Prämie als die Männer zahlen sollten, wo hingegen bei der Teilkaskoversicherung die Frauen um 22 % weniger zahlen sollten als Männer. Aufgrund einer Verordnung darf bei der Prämie von Versicherungsverträgen allerdings nicht bezüglich des Geschlechts unterschieden werden

Für die Teilkasko- und Vollkaskoversicherung ist die Antriebsart ein entscheidendes Merkmal zur Festlegung der Prämie. In beiden Fällen erfordert die Antriebsart Diesel eine wesentlich höhere Prämie. Dasselbe gilt für Fahrzeuge mit mehr Leistung, im Vergleich zu Fahrzeugen mit weniger Leistung. Durchaus interessant ist hierbei die Betrachtung der Haftpflichtversicherung. Bei dieser Sparte hat die Leistung so gut wie keinen Einfluss auf die Prämienhöhe.

Bei der Haftpflichversicherung ist ein treibendes Merkmal die Bonus-Malus Stufe, da die Prämie bei steigender Bonus-Malus Stufe rapide ansteigen müsste. Eine versicherte Person in der Stufe 6 müsste statistisch gesehen bereits um 85 % mehr Prämie bezahlen, als eine sich in den Stufen -6 bis -3 befindende versicherte Person. Ein weiteres Merkmal ist der Vermerk über ein Wechselkennzeichen. Fahrzeuge mit Wechselkennzeichen weisen deutlich schlechtere Werte als standardmäßig angemeldete auf. Diese Beobachtung lässt sich auch auf Leasing Fahrzeuge übertragen.

Durchaus erwähnenswert ist auch die Anzahl akademischer und sonstiger Titel. Entgegen den Erwartungen sollten nämlich die Haftpflichtversicherungsprämien versicherter Personen mit mindestens einem Titel um 8 % (akademische Titel) beziehungsweise 42 % (sonstige Titel) höher sein, als für Personen ohne Titel. Dieses Phänomen tritt auch bei der Vollkaskoversicherung auf, allerdings nicht ganz so stark.

Die drei wichtigsten Merkmale, welche bei allen Versicherungssparten von signifikanter Bedeutung sind, sind die Konzernklasse des Fahrzeuges, der Wohnbezirk des Versicherten, sowie das Alter der versicherten Person. Zu beachten ist hierbei, dass lediglich das Merkmal Alter einer versicherten Person über die drei verschiedenen Versicherungssparten betrachtet werden kann. Der Grund liegt in den unterschiedlichen Clusterings der Konzerne und Bezirke innerhalb der Versicherungssparten.

Bei den Altersklassen ist deutlich erkennbar, dass sehr junge Autofahrer (17-25 Jahre) eine höhere Prämie leisten müssten, als Personen im Alter zwischen 30 und 64 Jahren. Diese Altersklasse stellt auch meist das Referenzalter dar, da sich die meisten versicherten Personen in diesem Alter befinden. In der Haftpflicht- und Vollkaskoversicherung ist eine höhere Prämie auch für ältere Personen (ab 65 Jahren) erforderlich. Im krassen Gegenzug dazu ist dies bei der Teilkaskoversicherung jedoch nicht der Fall. Hier ist die erforderliche Prämie für Personen ab 50 Jahren um 28 % und für Personen ab 75 Jahren sogar um

bis zu 50 % niedriger. Der Grund hierfür ist vermutlich jener, dass Personen in diesem Alter weniger Wert auf deren Fahrzeuge legen und somit die Meldefrequenz von Schäden geringer ist.

Vergleiche von Konzernklassen und Bezirksklassen dürfen nur innerhalb der jeweiligen Versicherungssparte durchgeführt werden, da diese im Unterschied zu den anderen Merkmalen nicht einheitlich geclustert wurden.

Betrachtet man zunächst die Haftpflichtversicherung so erkennt man, dass die Referenzklasse jene ist, bei welcher der Konzern Volkswagen vorkommt, da mit Abstand am meisten Fahrzeuge dieses Konzerns im Bestand sind. Krasse Unterschiede sind hier zu den Konzernklassen 1,2,4 und 5 erkennbar, wobei die erste Klasse eine deutlich höhere Prämie aufbringen müsste. Der Schätzer von 0,44 für die Klasse 2 ist nicht ernst zu nehmen, da der Bestand dieser Klasse nur sehr gering ist. Fahrzeuge der Konzerne Suzuki, Fiat und Subaru weisen einen deutlich niedrigeren Schadenbedarf als die Refernzklasse auf, wo hingegen Fahrzeuge der Konzerne Honda, Seat, Jaguar, BMW und Daewoo einen höheren Schadenbedarf haben.

Bei den Bezirken ist ersichtlich, dass Kärntner Bezirke wesentlich besser abschneiden als jene im Nordosten von Österreich. Ein Ausreißer ist die Stadt Innsbruck, welche in einer eigenen Klasse ist, da der Bestand dort groß genug ist und gleichzeitig der Schadenbedarf enorm hoch ist.

Als Nächstes betrachte man die Vollkaskoversicherung. Bei den Konzernen wird wieder die Klasse mit VW als Referenzklasse gewählt. Hier stimmen die theoretischen Ergebnisse mit den Vermutungen/Vorurteilen aus der Realität überein. Im Detail bedeutet das, dass Fahrzeuge der teureren Marken wie zum Beispiel Mini, Mercedes, Audi, BMW und Rover einen höheren Schadenbedarf aufweisen als jene der Referenzklasse.

Bei den Bezirken sieht man innerhalb Kärntens interessanterweise keine starken Unterschiede zwischen Stadt und Land. Höhere Prämien sollten in Wien und den Umgebungen um Wien, Innsbruck und Salzburg eingehoben werden, da dort der Schadenbedarf deutlich höher ist.

Zum Schluss wird noch die Teilkaskoversicherung begutachtet. Bei den Konzernen ist hier erkennbar, dass Fahrzeuge der Konzerne BMW, Volvo, Audi, Alfa-Romeo, Honda, Chrysler, Mazda, VW, Mitsubishi und Opel jene Konzerne mit dem höchsten Schadenbedarf sind. Die Unterschiede der erforderlichen Prämie zu den anderen Marken sind hier sogar recht hoch und befinden sich im Bereich zwischen 20 % und 30 %. Bei den Bezirksklassen befinden sich die meisten Bezirke von Kärnten in der hier guten Referenzklasse. Lediglich Sankt Veit an der Glan und Völkermarkt verhalten sich ähnlich schlecht wie zum Beispiel Graz (Stadt) und Feldkirchen weist bereits einen noch höheren Schadenbedarf auf. Dieser ist bereits gleich hoch wie in Wien, Wien-Umgebung und Graz-Umgebung. Am deutlich schlechtesten sind die Bezirke Zwettl, Baden und Krems (Land), welche laut Statistik eine doppelt so hohe Prämie wie die Bezirke aus der Referenzklasse zahlen sollten.

Gesamtheitlich gesehen fällt auf, dass die Bezirke nordöstlich von Österreich wie zum Beispiel Wien, Wien-Umgebung wesentlich schlechter abschneiden als die Bezirke mit den Hauptbeständen in Kärnten. Des Weiteren bestätigt sich, dass teurere Automarken auch einen höheren Schadenbedarf haben und somit für diese Konzerne entsprechend höhere Prämien eingehoben werden müssten.

7.2 Wofür benötigt eine Versicherung diese Ergebnisse?

Da man die erhaltenen Informationen in der Praxis nicht direkt umsetzen kann, da Diskriminierungen auf Grund des Geschlechts oder Alters nicht erlaubt sind und kein Versicherungsnehmer bei einer Versicherung bleiben würde, wenn diese wegen des Fahrzeugtyps des Versicherungsnehmers plötzlich 20 % mehr Prämie verlangt, sind alternative Ideen gefragt.

Eine Möglichkeit wäre die versicherungsinternen Kundenbetreuer zu informieren, wo das Geschäft statistisch gesehen eher schlecht beziehungsweise gut ist. Somit könnte man das Anwerben neuer Kunden in gewissen Gebieten einschränken oder fördern.

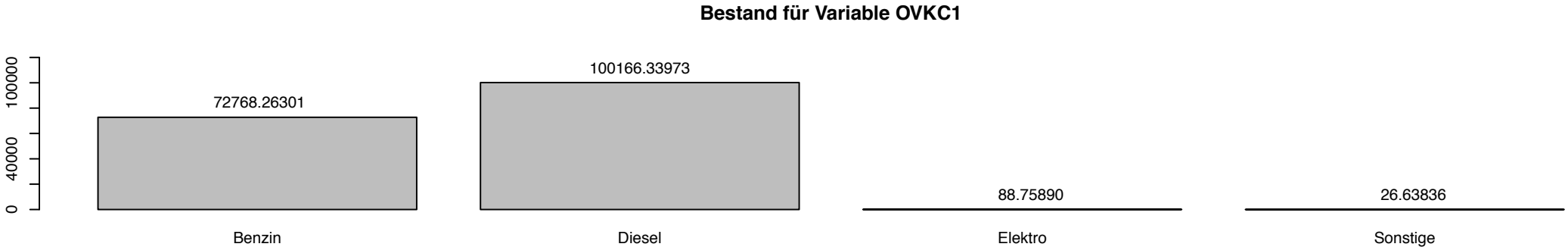
Ein weiterer Gedanke wären unterschiedliche Tarife oder Boni für Kunden mit gewissen Automarken anzubieten und damit statistisch gesehen gute Kunden anzulocken.

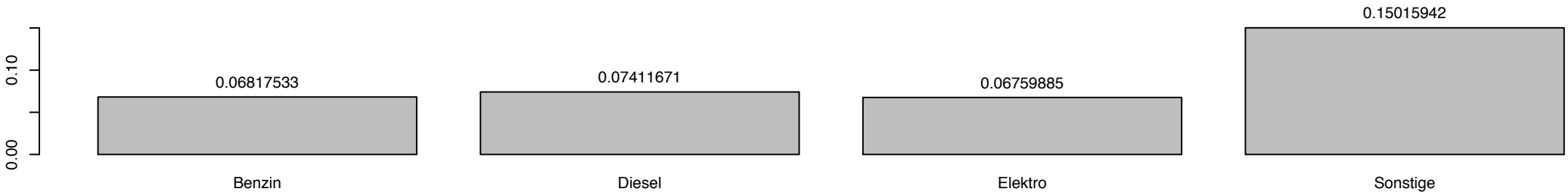
Man könnte versuchen die Ergebnisse fahrzeugspezifischer Merkmale wie zum Beispiel den Konzern, die Antriebsart, die Leistung und ähnliche so zu nützen, dass die Tarife nur geringfügig voneinander abweichen, sodass der Versicherugsnehmer dies zunächst überhaupt nicht wahrnimmt. Bei der jährlichen Tarifanpassung könnte man dann zum Beispiel die Tarife der „guten“ Versicherungsnehmer gleich lassen und nur die Prämien jener Polizzen erhöhen, welche statistisch gesehen „schlechter“ sind. Somit würde man eine schleichende Differenzierung zwischen „besseren“ und „schlechteren“ Kunden erhalten.

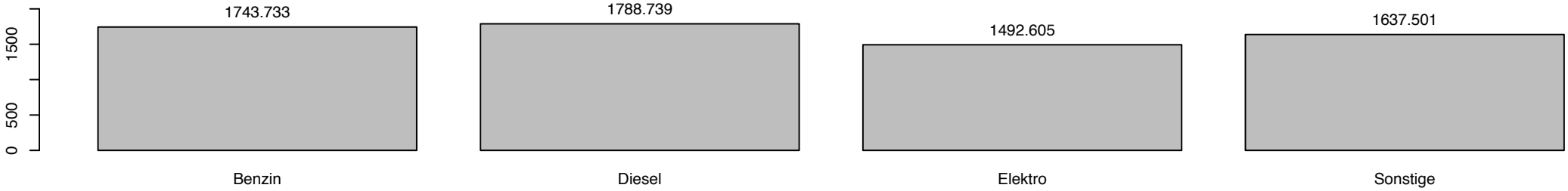
8 Anhang

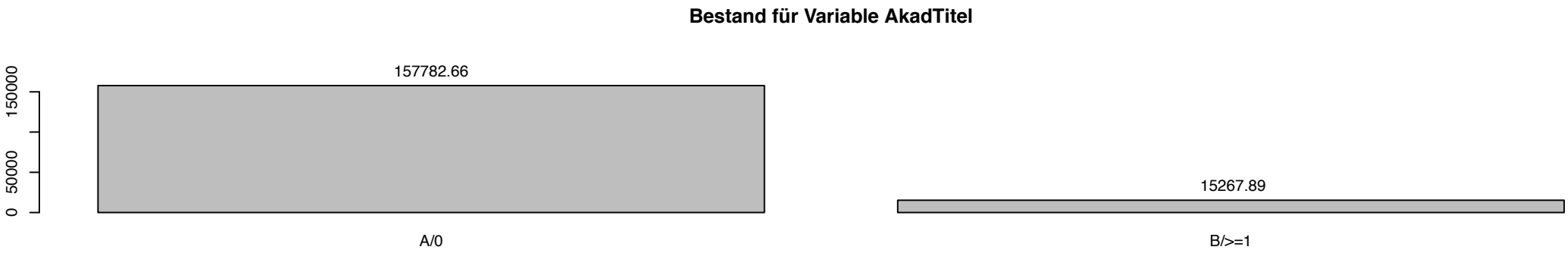
8.1 Deskriptive Analyse für die Haftpflichtversicherung

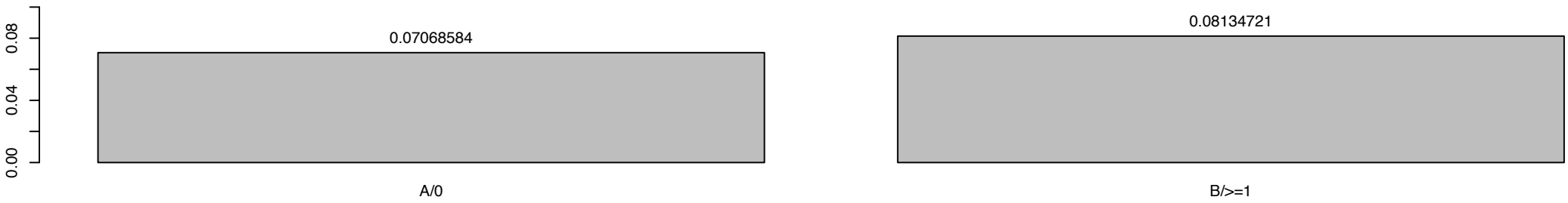
- Antriebsart
- Anzahl akademischer Titel
- Anzahl sonstiger Titel
- Altersklassen
- Bezirksklassen
- Bonus-Malus-Stufen
- Familienstand
- Geschlecht
- Haftpflicht-Variante
- Hubraum
- Interne-Bonus-Malus-Stufen
- Konzern
- Leasing
- Leistung
- Nation
- Natürliche/Juristische Person
- Verwendungszweck
- Wechselkennzeichen
- Zugehörigkeit

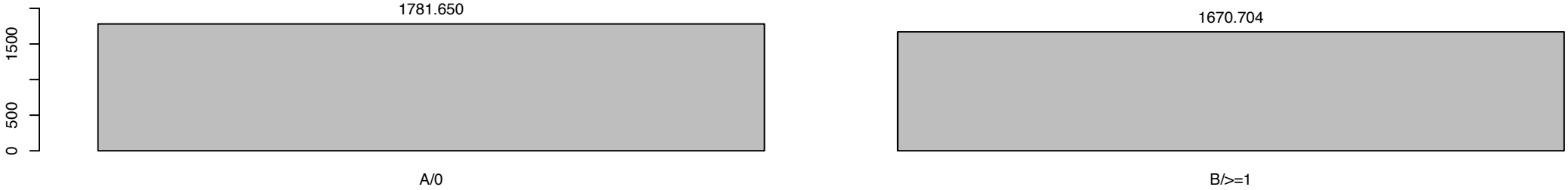

Bestand für Variable OVKC1
0
40000
100000
72768.26301
100166.33973
88.75890
26.63836
Benzin
Diesel
Elektro
Sonstige

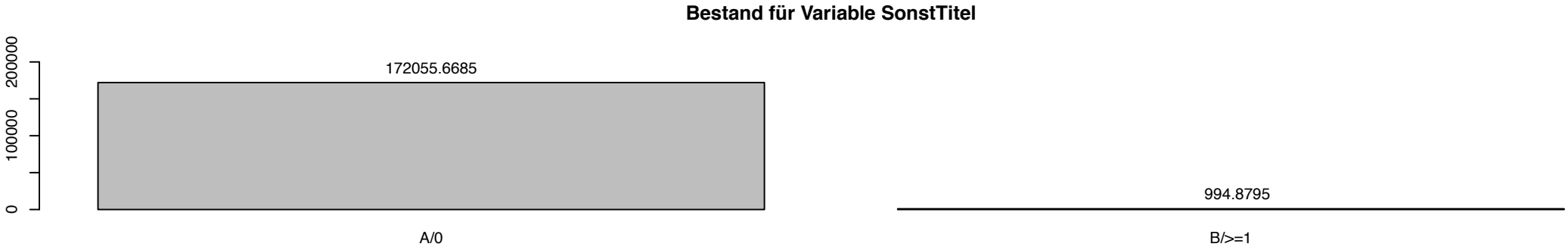

Schadenfrequenz für Variable OVKC1
0.00
0.10
0.06817533
0.07411671
0.06759885
0.15015942
Benzin
Diesel
Elektro
Sonstige

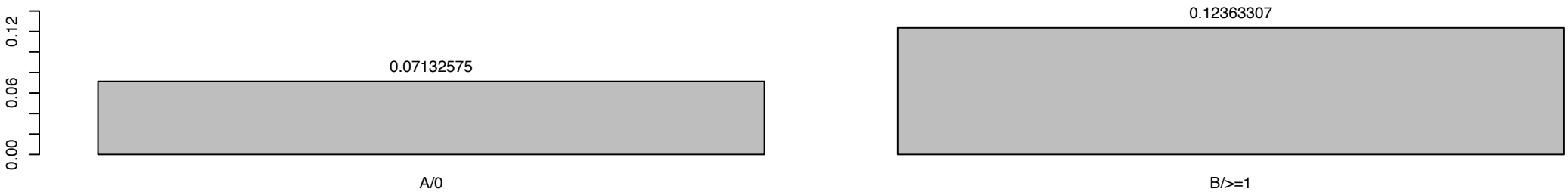

Durchschnittliche Schadenhöhe für Variable OVKC1
0
500
1500
1743.733
1788.739
1492.605
1637.501
Benzin
Diesel
Elektro
Sonstige

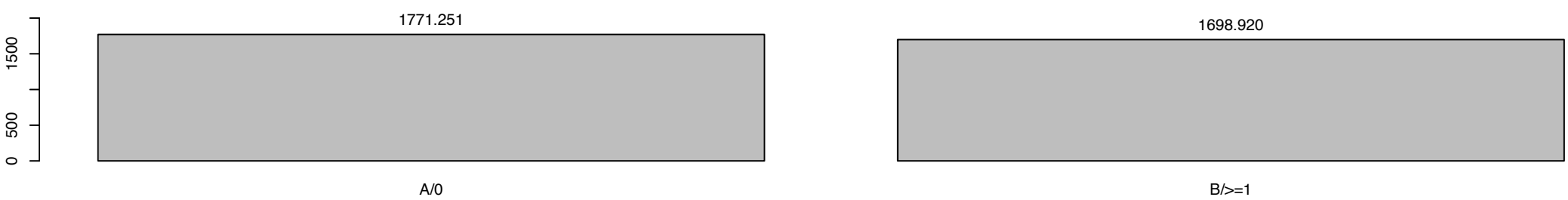

Bestand für Variable AkadTitel
157782.66
15267.89
0
50000
150000
A/0
B/>=1

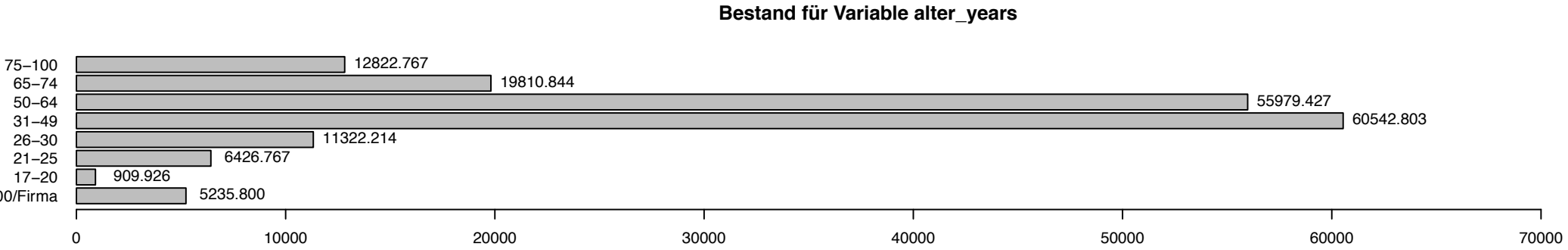

Schadenfrequenz für Variable AkadTitel
0.07068584
0.08134721
0.00
0.04
0.08
A/0
B/>=1

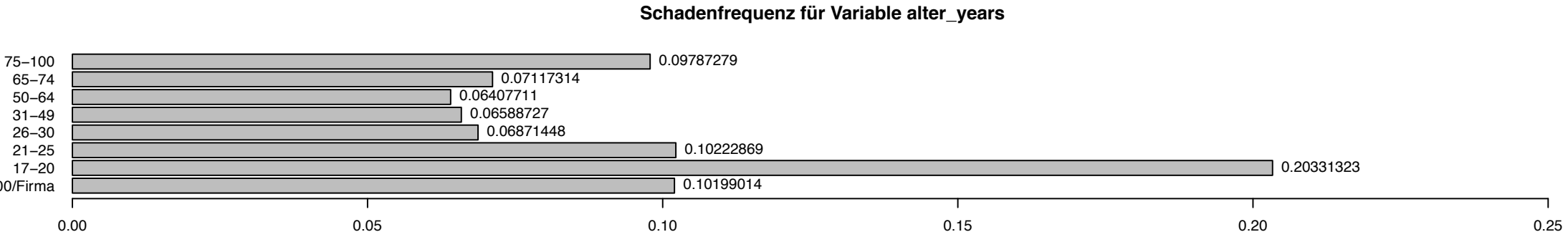

Durchschnittliche Schadenhöhe für Variable AkadTitel
1781.650
1670.704
0
500
1500
A/0
B/>=1

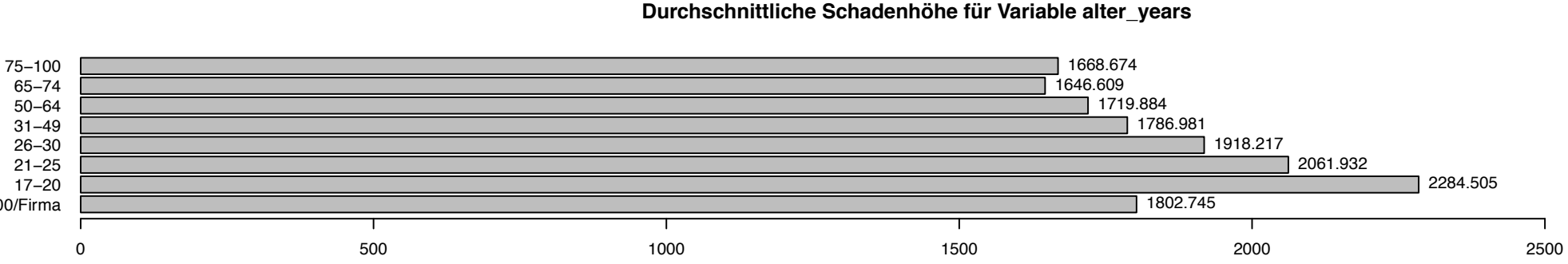

Bestand für Variable SonstTitel
0
100000
200000
172055.6685
994.8795
A/0
B/>=1

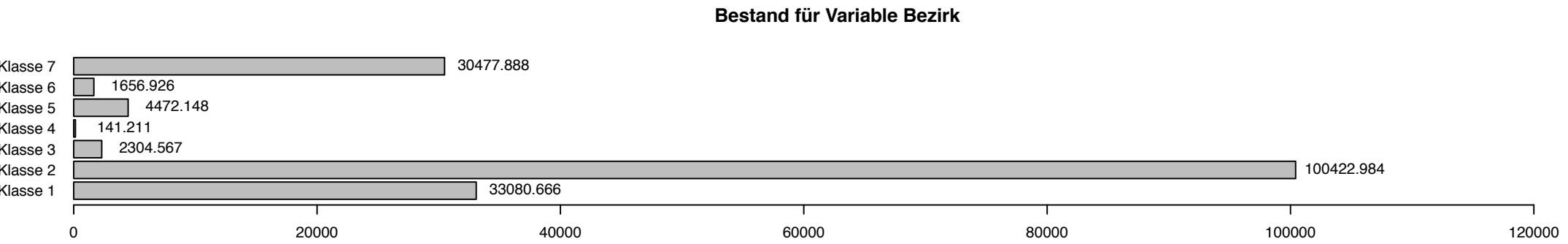

Schadenfrequenz für Variable SonstTitel
0.00
0.06
0.12
0.07132575
0.12363307
A/0
B/>=1

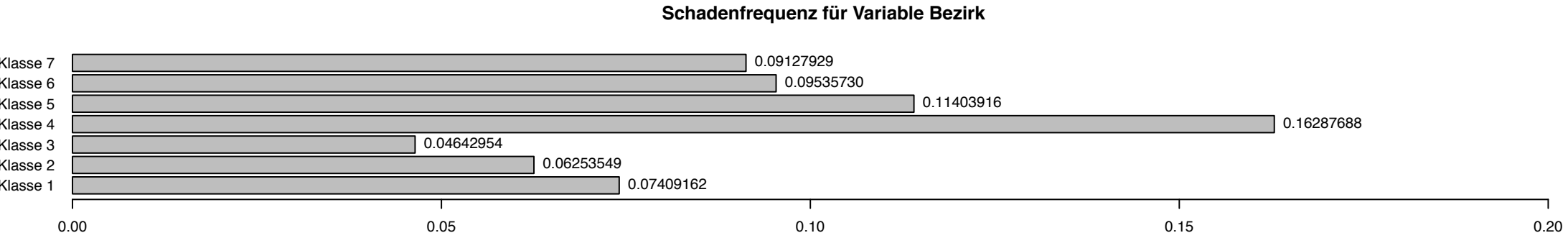

Durchschnittliche Schadenhöhe für Variable SonstTitel
0
500
1500
1771.251
1698.920
A/0
B/>=1

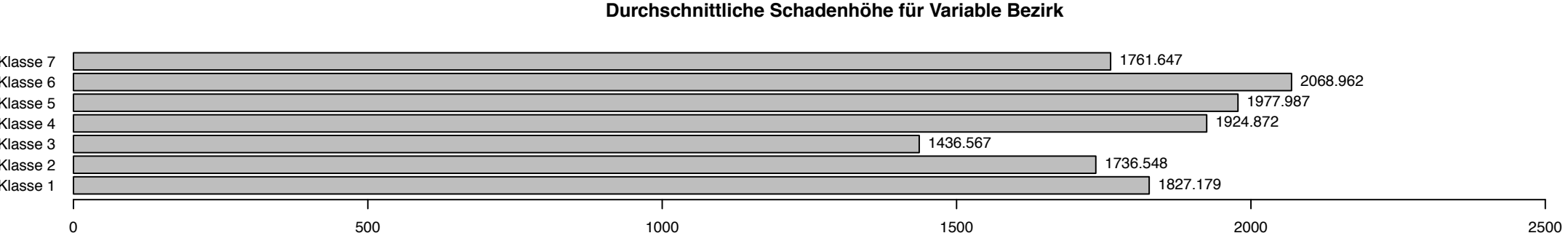

Bestand für Variable alter_years
75–100
65–74
50–64
31–49
26–30
21–25
17–20
00/Firma
12822.767
19810.844
55979.427
60542.803
11322.214
6426.767
909.926
5235.800
0
10000
20000
30000
40000
50000
60000
70000

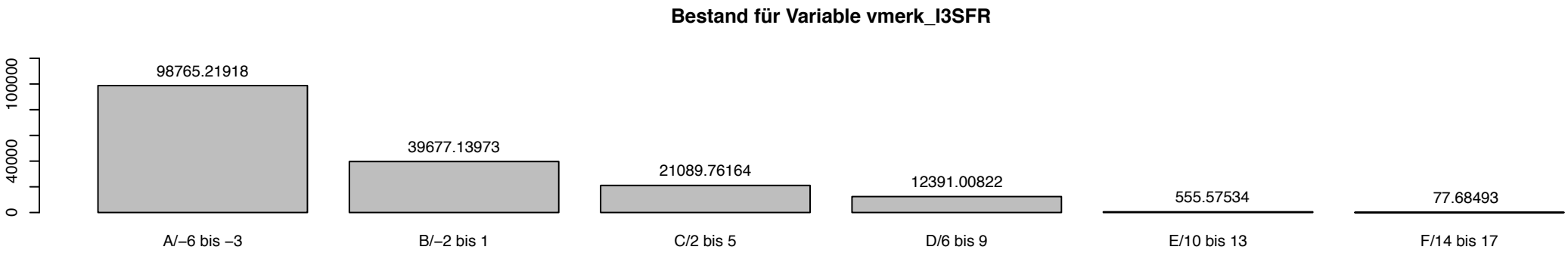

Schadenfrequenz für Variable alter_years
75–100
65–74
50–64
31–49
26–30
21–25
17–20
00/Firma
0.09787279
0.07117314
0.06407711
0.06588727
0.06871448
0.10222869
0.20331323
0.10199014
0.00
0.05
0.10
0.15
0.20
0.25

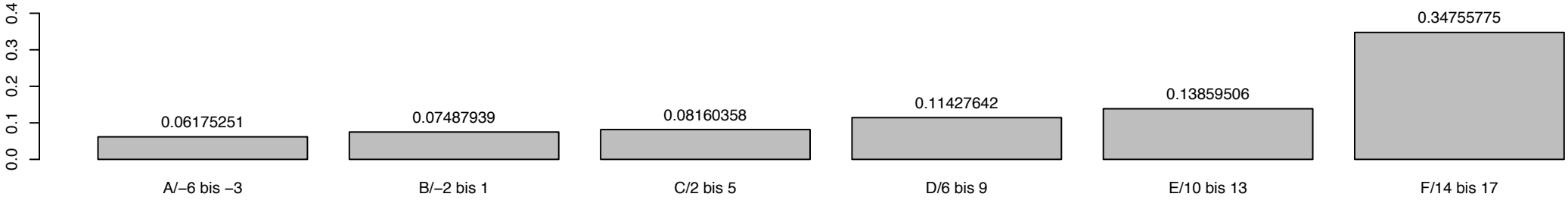

Durchschnittliche Schadenhöhe für Variable alter_years
75–100
65–74
50–64
31–49
26–30
21–25
17–20
00/Firma
1668.674
1646.609
1719.884
1786.981
1918.217
2061.932
2284.505
1802.745
0
500
1000
1500
2000
2500

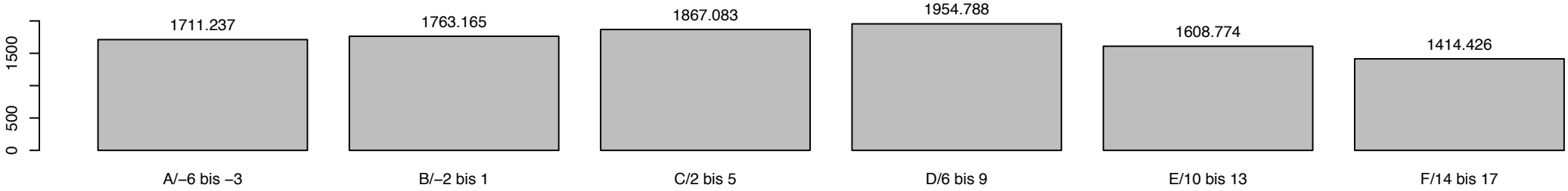

Bestand für Variable Bezirk
Klasse 7
Klasse 6
Klasse 5
Klasse 4
Klasse 3
Klasse 2
Klasse 1
30477.888
1656.926
4472.148
141.211
2304.567
100422.984
33080.666
0
20000
40000
60000
80000
100000
120000

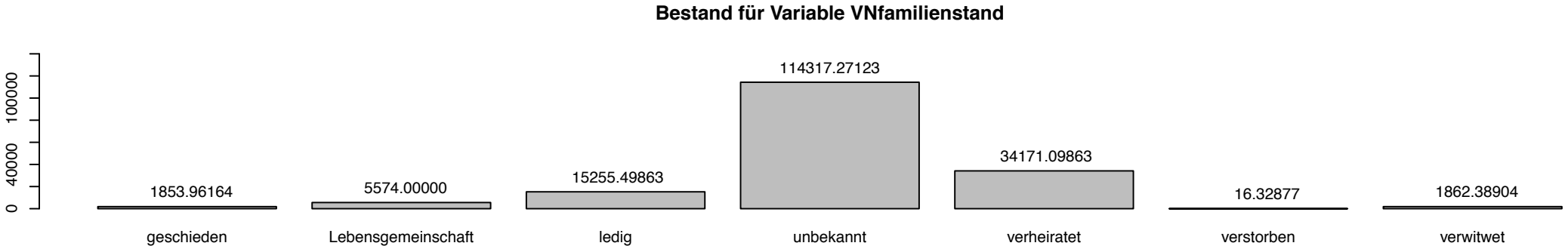

Schadenfrequenz für Variable Bezirk
Klasse 7
Klasse 6
Klasse 5
Klasse 4
Klasse 3
Klasse 2
Klasse 1
0.09127929
0.09535730
0.11403916
0.16287688
0.04642954
0.06253549
0.07409162
0.00
0.05
0.10
0.15
0.20

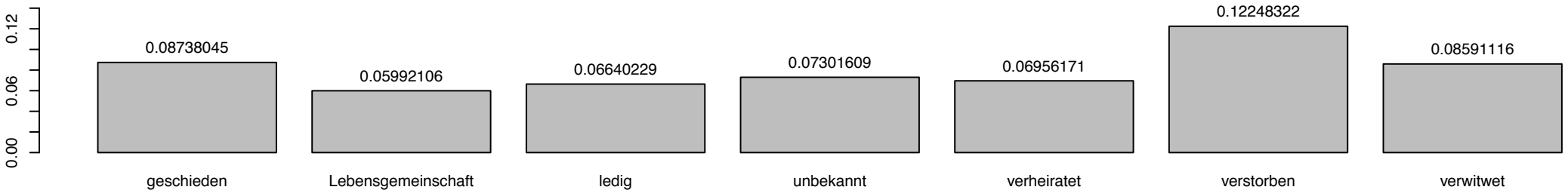

Durchschnittliche Schadenhöhe für Variable Bezirk
Klasse 7
Klasse 6
Klasse 5
Klasse 4
Klasse 3
Klasse 2
Klasse 1
1761.647
2068.962
1977.987
1924.872
1436.567
1736.548
1827.179
0
500
1000
1500
2000
2500

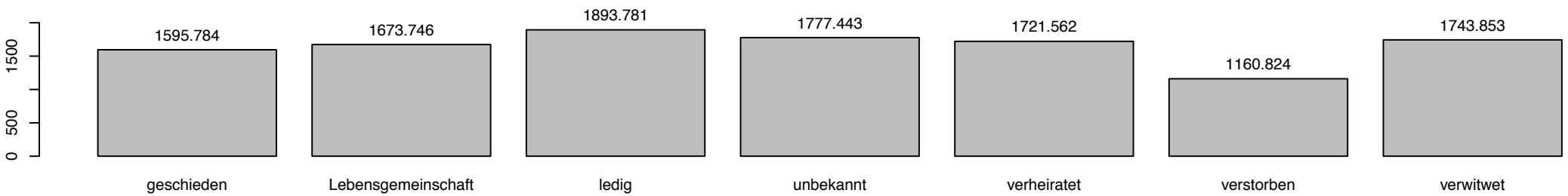

Bestand für Variable vmerk_l3SFR
0
40000
100000
98765.21918
39677.13973
21089.76164
12391.00822
555.57534
77.68493
A/−6 bis −3
B/−2 bis 1
C/2 bis 5
D/6 bis 9
E/10 bis 13
F/14 bis 17

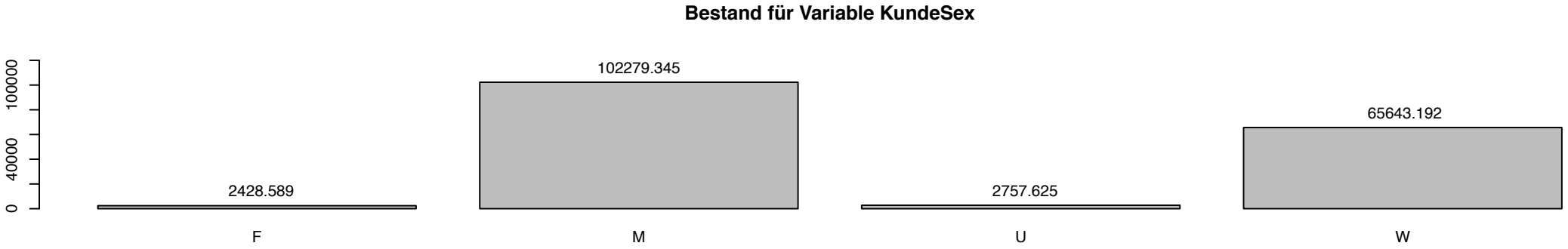

Schadenfrequenz für Variable vmerk_l3SFR
0.0
0.1
0.2
0.3
0.4
0.06175251
0.07487939
0.08160358
0.11427642
0.13859506
0.34755775
A/−6 bis −3
B/−2 bis 1
C/2 bis 5
D/6 bis 9
E/10 bis 13
F/14 bis 17

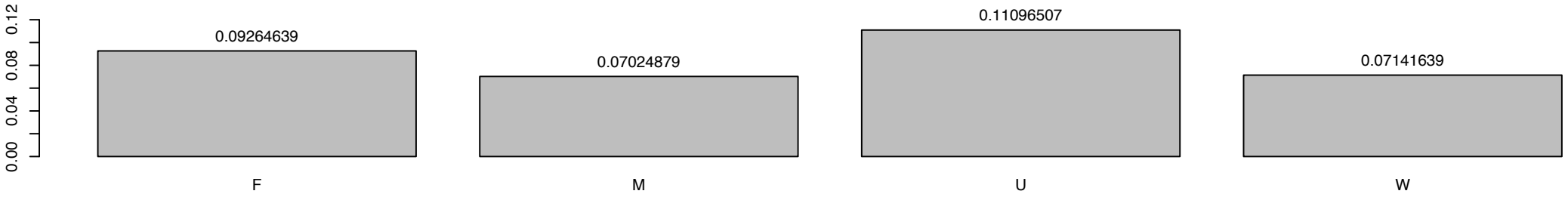

Durchschnittliche Schadenhöhe für Variable vmerk_l3SFR
0
500
1500
1711.237
1763.165
1867.083
1954.788
1608.774
1414.426
A/−6 bis −3
B/−2 bis 1
C/2 bis 5
D/6 bis 9
E/10 bis 13
F/14 bis 17

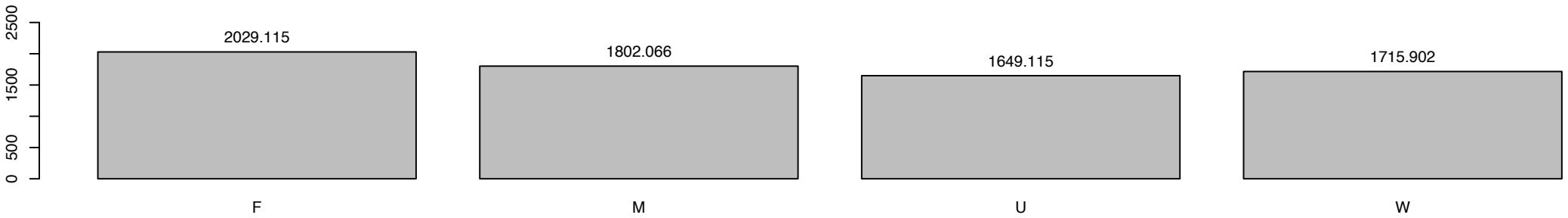

Bestand für Variable VNfamilienstand
0
40000
100000
1853.96164
5574.00000
15255.49863
114317.27123
34171.09863
16.32877
1862.38904
geschieden
Lebensgemeinschaft
ledig
unbekannt
verheiratet
verstorben
verwitwet

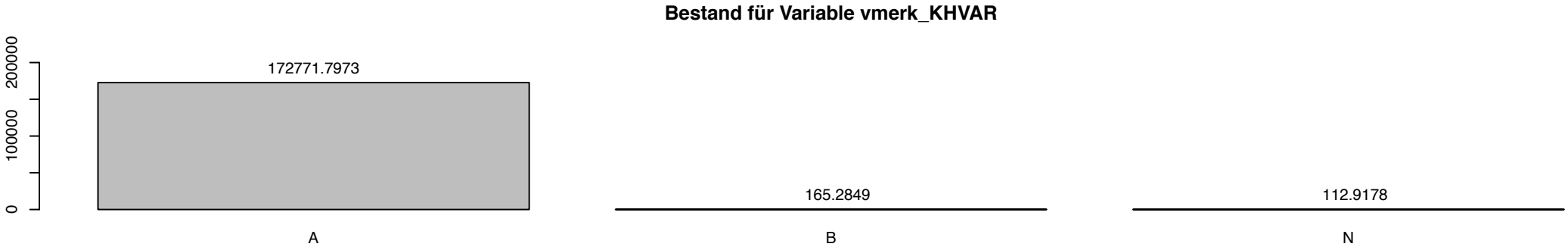

Schadenfrequenz für Variable VNfamilienstand
0.00
0.06
0.12
0.08738045
0.05992106
0.06640229
0.07301609
0.06956171
0.12248322
0.08591116
geschieden
Lebensgemeinschaft
ledig
unbekannt
verheiratet
verstorben
verwitwet

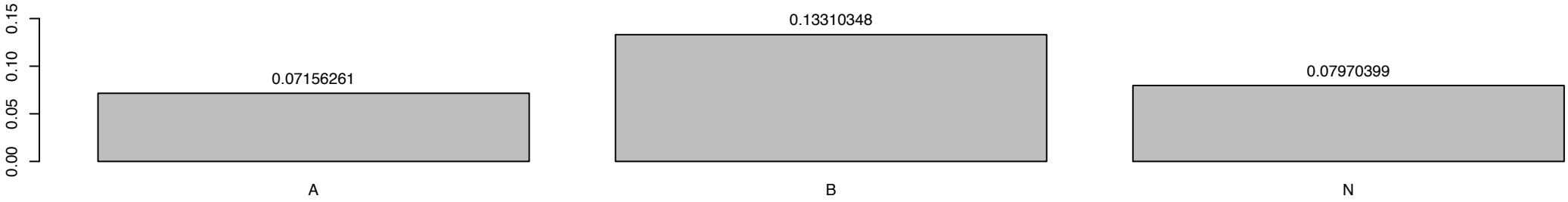

Durchschnittliche Schadenhöhe für Variable VNfamilienstand
0
500
1500
1595.784
1673.746
1893.781
1777.443
1721.562
1160.824
1743.853
geschieden
Lebensgemeinschaft
ledig
unbekannt
verheiratet
verstorben
verwitwet


Bestand für Variable KundeSex
0
40000
100000
2428.589
102279.345
2757.625
65643.192
F
M
U
W

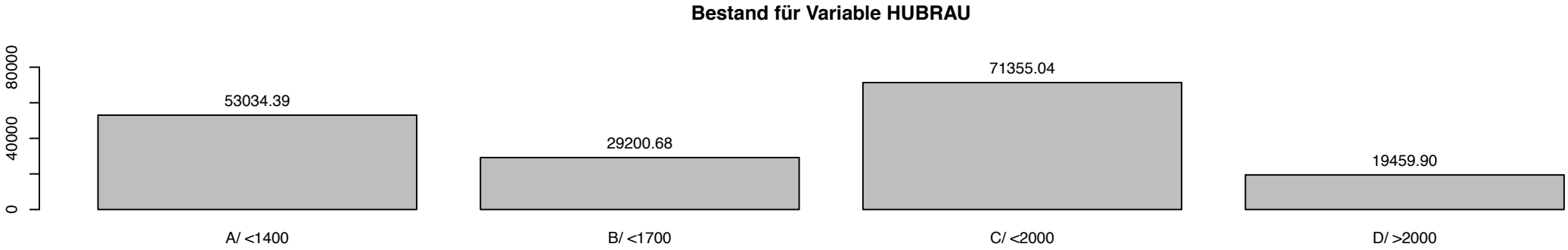

Schadenfrequenz für Variable KundeSex
0.00
0.04
0.08
0.12
0.09264639
0.07024879
0.11096507
0.07141639
F
M
U
W

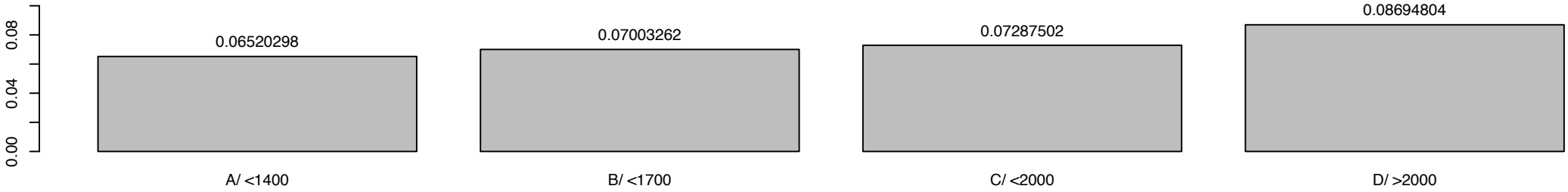

Durchschnittliche Schadenhöhe für Variable KundeSex
0
500
1500
2500
2029.115
1802.066
1649.115
1715.902
F
M
U
W

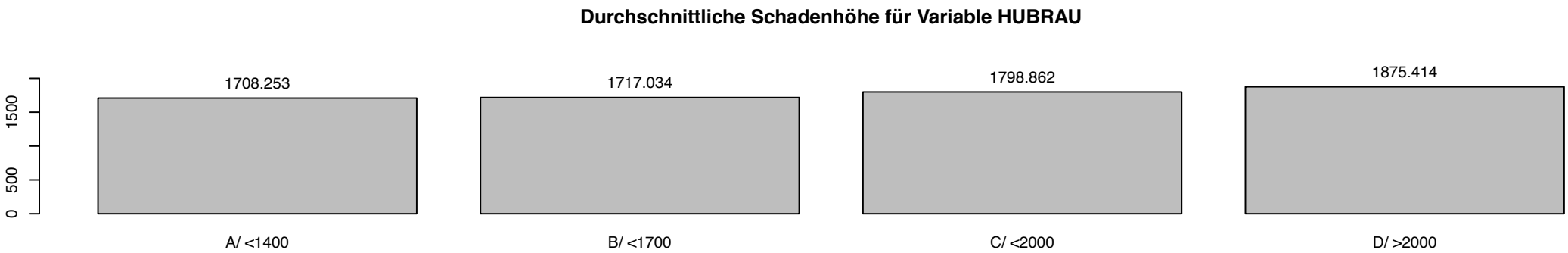

Bestand für Variable vmerk_KHVAR
0
100000
200000
172771.7973
165.2849
112.9178
A
B
N

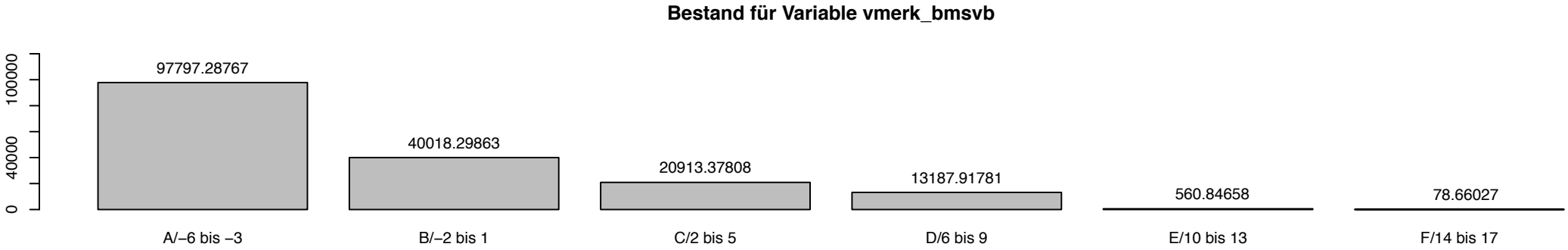

Schadenfrequenz für Variable vmerk_KHVAR
0.00
0.05
0.10
0.15
0.07156261
0.13310348
0.07970399
A
B
N

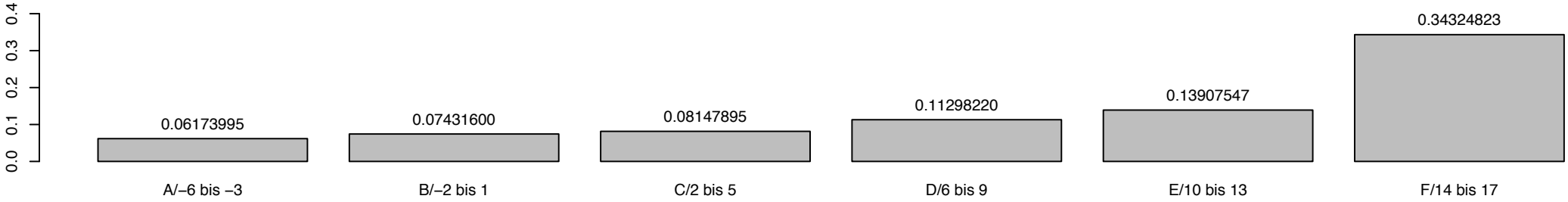

Durchschnittliche Schadenhöhe für Variable vmerk_KHVAR
0
500
1500
1771.395
1240.759
1881.913
A
B
N

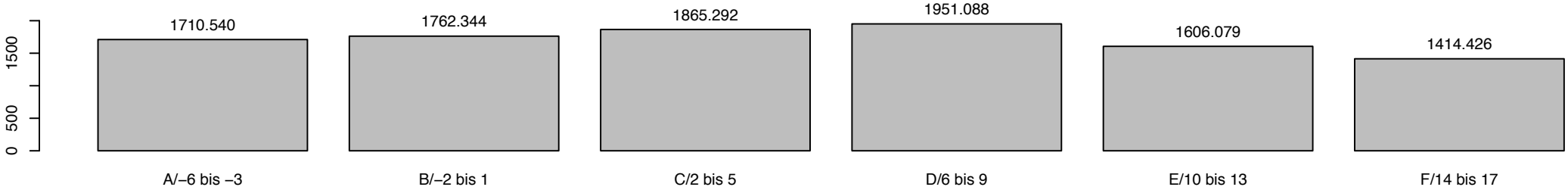

Bestand für Variable HUBRAU
0
40000
80000
53034.39
29200.68
71355.04
19459.90
A/ <1400
B/ <1700
C/ <2000
D/ >2000

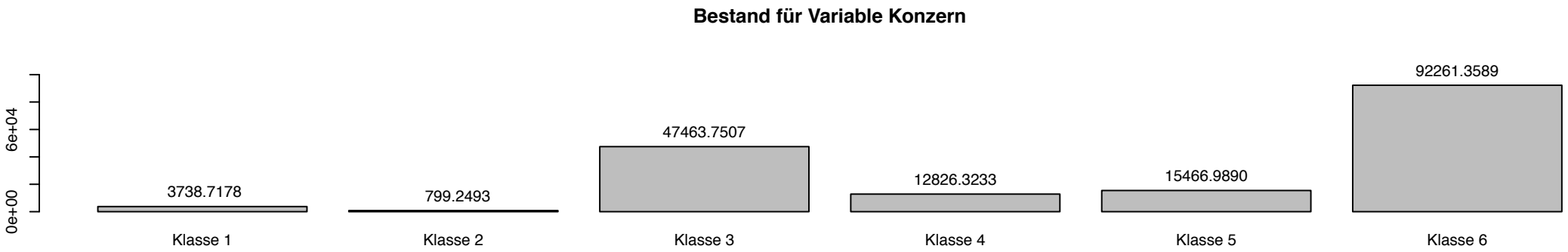

Schadenfrequenz für Variable HUBRAU
0.00
0.04
0.08
0.06520298
0.07003262
0.07287502
0.08694804
A/ <1400
B/ <1700
C/ <2000
D/ >2000

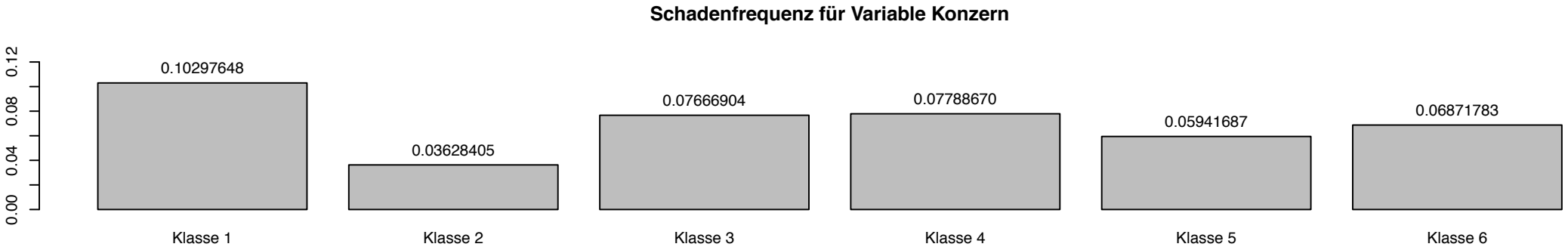

Durchschnittliche Schadenhöhe für Variable HUBRAU
0
500
1500
1708.253
1717.034
1798.862
1875.414
A/ <1400
B/ <1700
C/ <2000
D/ >2000

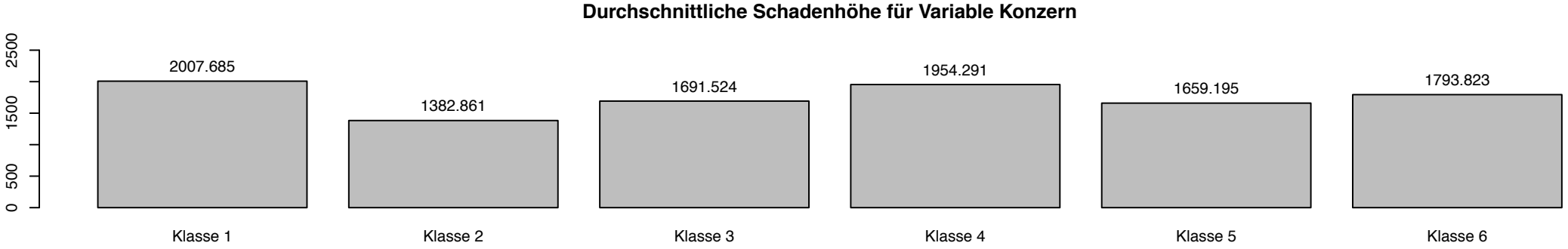

Bestand für Variable vmerk_bmsvb
0
40000
100000
97797.28767
40018.29863
20913.37808
13187.91781
560.84658
78.66027
A/−6 bis −3
B/−2 bis 1
C/2 bis 5
D/6 bis 9
E/10 bis 13
F/14 bis 17

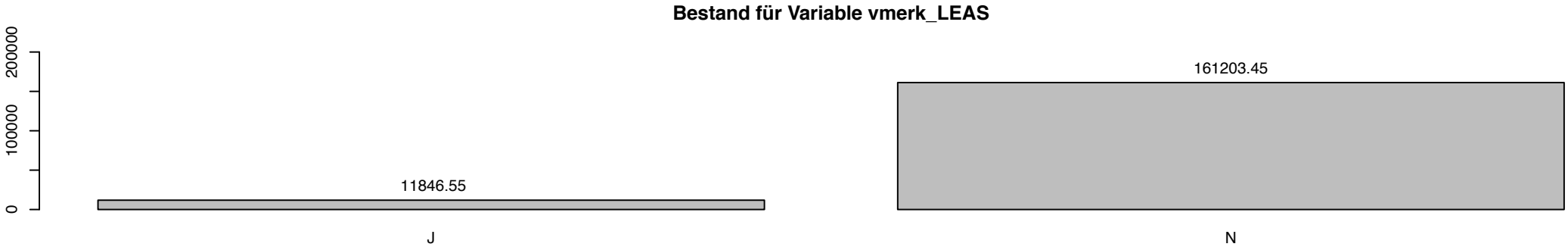

Schadenfrequenz für Variable vmerk_bmsvb
0.0
0.1
0.2
0.3
0.4
0.06173995
0.07431600
0.08147895
0.11298220
0.13907547
0.34324823
A/−6 bis −3
B/−2 bis 1
C/2 bis 5
D/6 bis 9
E/10 bis 13
F/14 bis 17

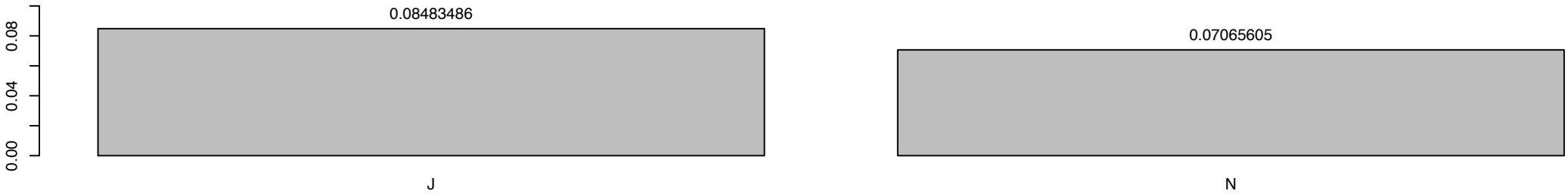

Durchschnittliche Schadenhöhe für Variable vmerk_bmsvb
0
500
1500
1710.540
1762.344
1865.292
1951.088
1606.079
1414.426
A/−6 bis −3
B/−2 bis 1
C/2 bis 5
D/6 bis 9
E/10 bis 13
F/14 bis 17

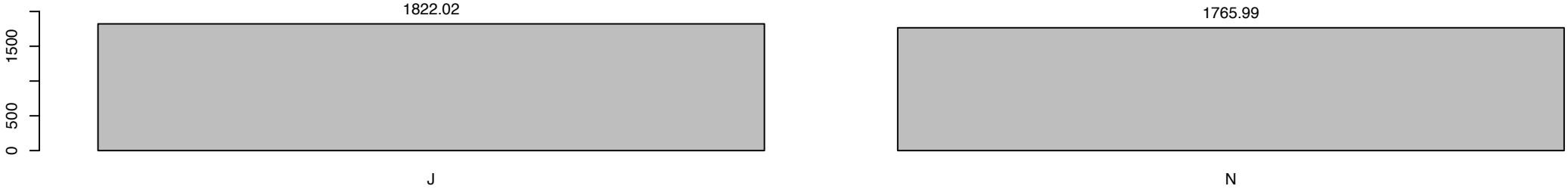

Bestand für Variable Konzern
0e+00
6e+04
3738.7178
799.2493
47463.7507
12826.3233
15466.9890
92261.3589
Klasse 1
Klasse 2
Klasse 3
Klasse 4
Klasse 5
Klasse 6

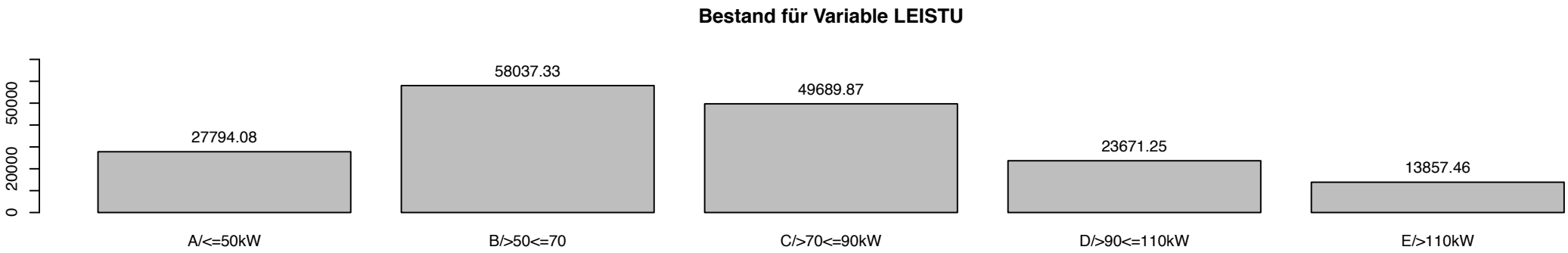

Schadenfrequenz für Variable Konzern
0.00
0.04
0.08
0.12
0.10297648
0.03628405
0.07666904
0.07788670
0.05941687
0.06871783
Klasse 1
Klasse 2
Klasse 3
Klasse 4
Klasse 5
Klasse 6

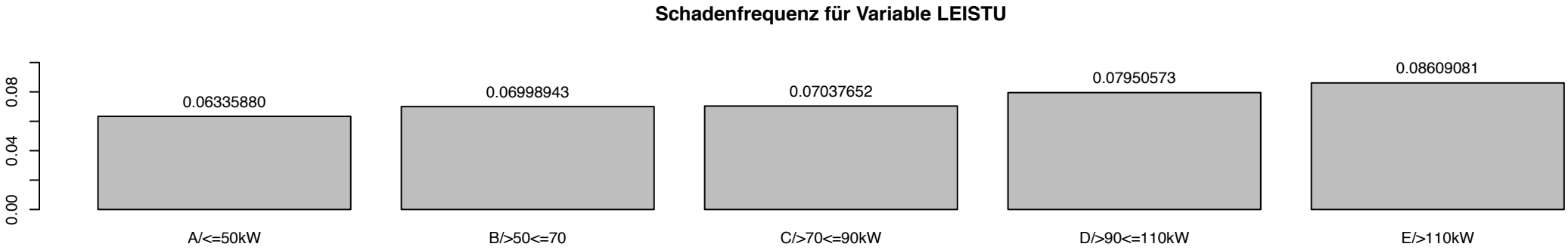

Durchschnittliche Schadenhöhe für Variable Konzern
0
500
1500
2500
2007.685
1382.861
1691.524
1954.291
1659.195
1793.823
Klasse 1
Klasse 2
Klasse 3
Klasse 4
Klasse 5
Klasse 6

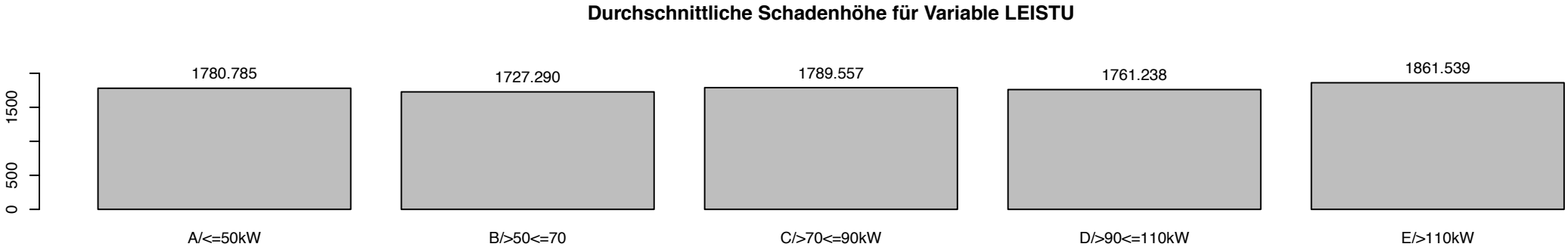

Bestand für Variable vmerk_LEAS
200000
100000
0
11846.55
161203.45
J
N

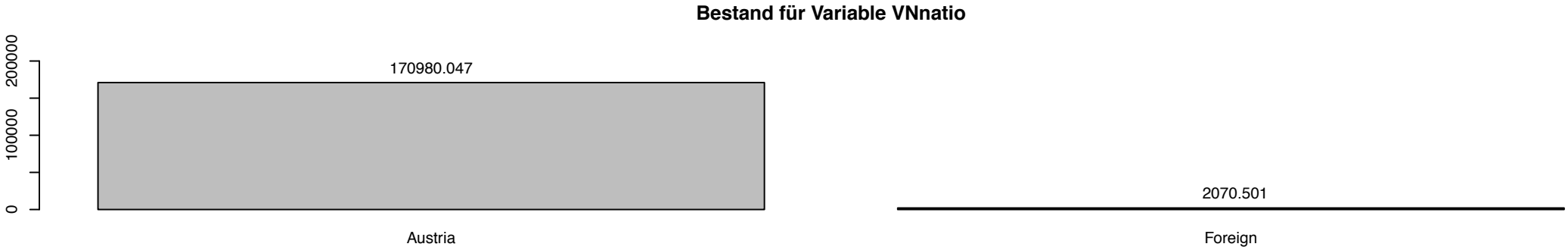

Schadenfrequenz für Variable vmerk_LEAS
0.08
0.04
0.00
0.08483486
0.07065605
J
N

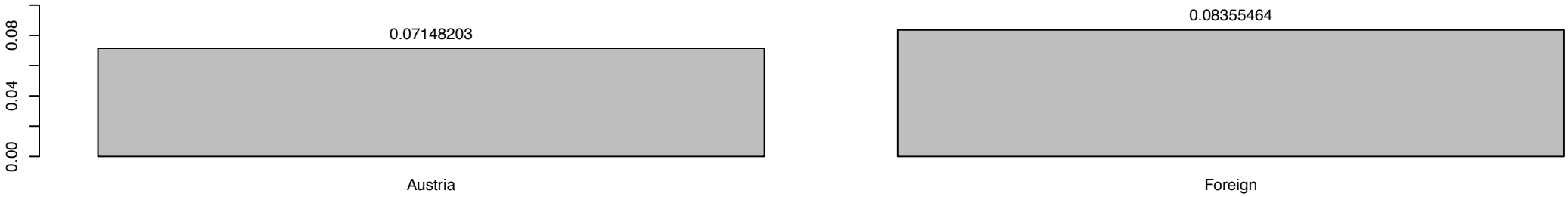

Durchschnittliche Schadenhöhe für Variable vmerk_LEAS
1500
500
0
1822.02
1765.99
J
N

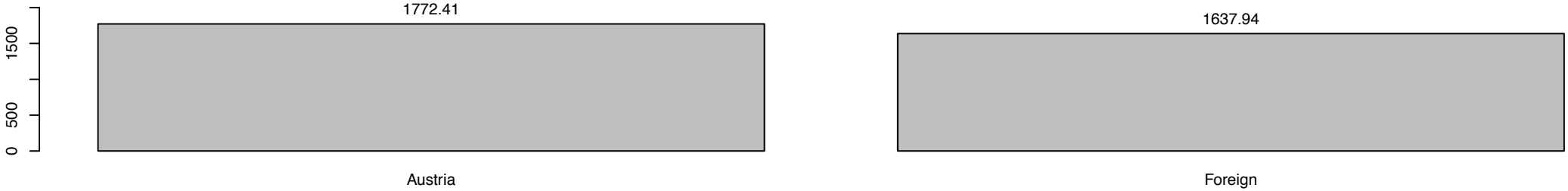

Bestand für Variable LEISTU
0
20000
50000
27794.08
58037.33
49689.87
23671.25
13857.46
A/<=50kW
B/>50<=70
C/>70<=90kW
D/>90<=110kW
E/>110kW

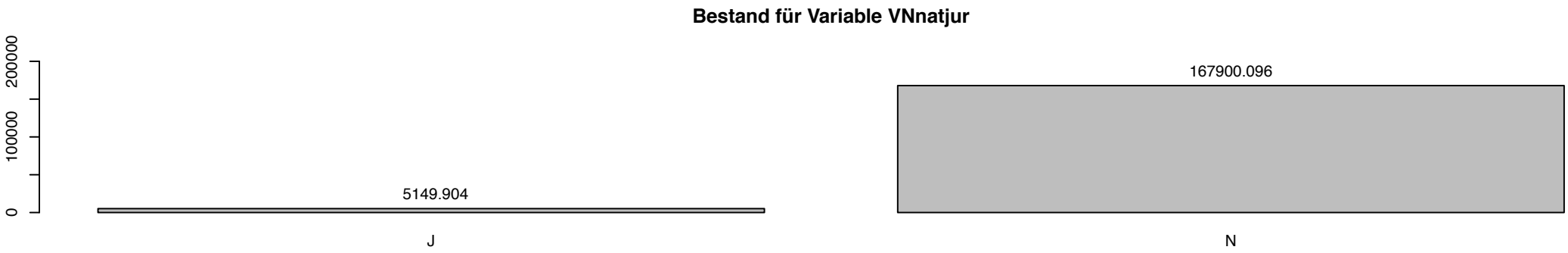

Schadenfrequenz für Variable LEISTU
0.00
0.04
0.08
0.06335880
0.06998943
0.07037652
0.07950573
0.08609081
A/<=50kW
B/>50<=70
C/>70<=90kW
D/>90<=110kW
E/>110kW

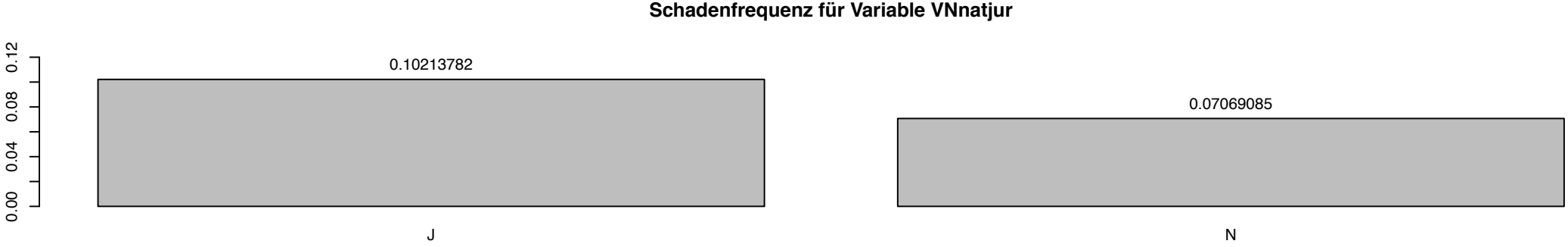

Durchschnittliche Schadenhöhe für Variable LEISTU
0
500
1500
1780.785
1727.290
1789.557
1761.238
1861.539
A/<=50kW
B/>50<=70
C/>70<=90kW
D/>90<=110kW
E/>110kW

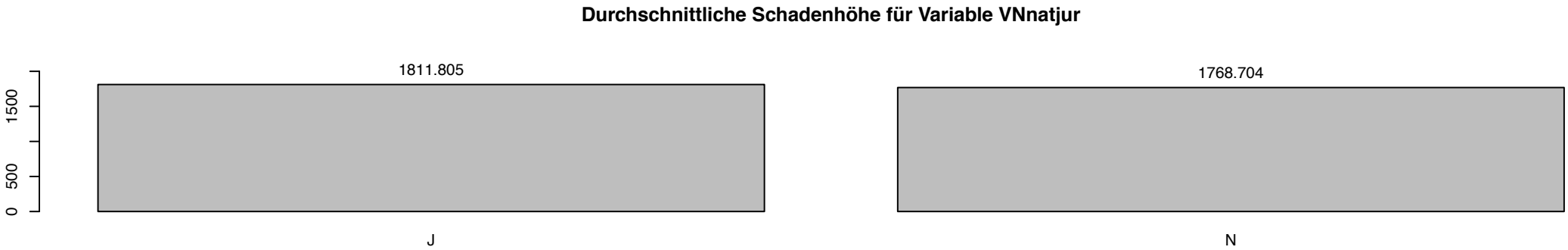

Bestand für Variable VNnatio
0
100000
200000
170980.047
2070.501
Austria
Foreign

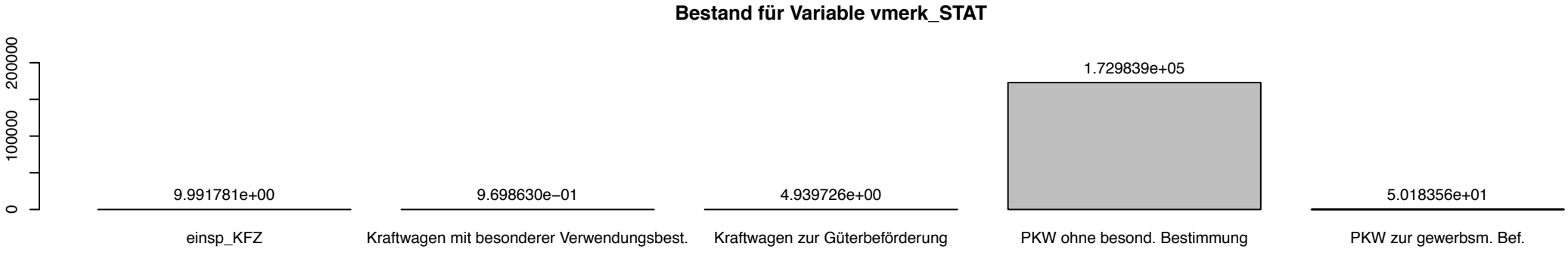

Schadenfrequenz für Variable VNnatio
0.00
0.04
0.08
0.07148203
0.08355464
Austria
Foreign

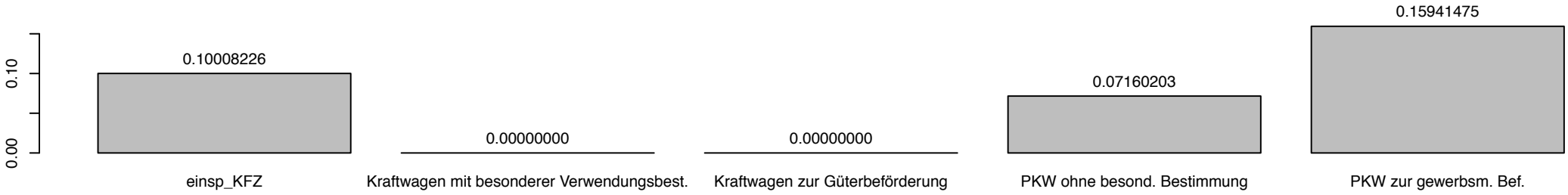

Durchschnittliche Schadenhöhe für Variable VNnatio
0
500
1500
1772.41
1637.94
Austria
Foreign

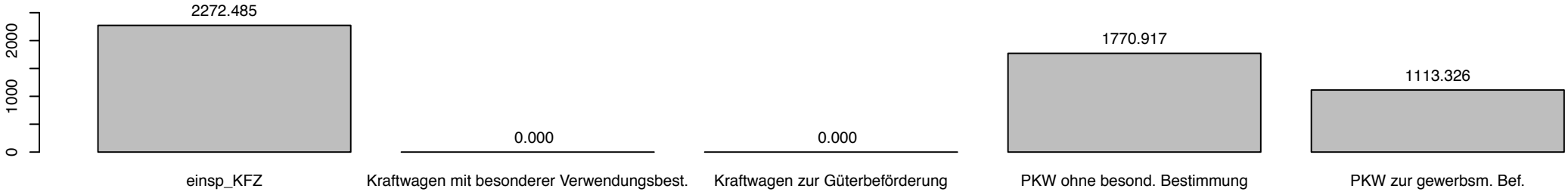

Bestand für Variable VNnatjur
0
100000
200000
5149.904
167900.096
J
N

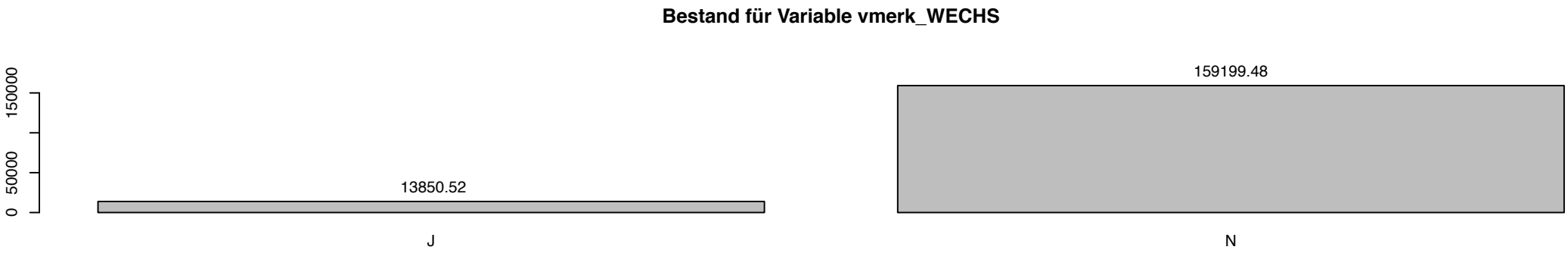

Schadenfrequenz für Variable VNnatjur
0.00
0.04
0.08
0.12
0.10213782
0.07069085
J
N

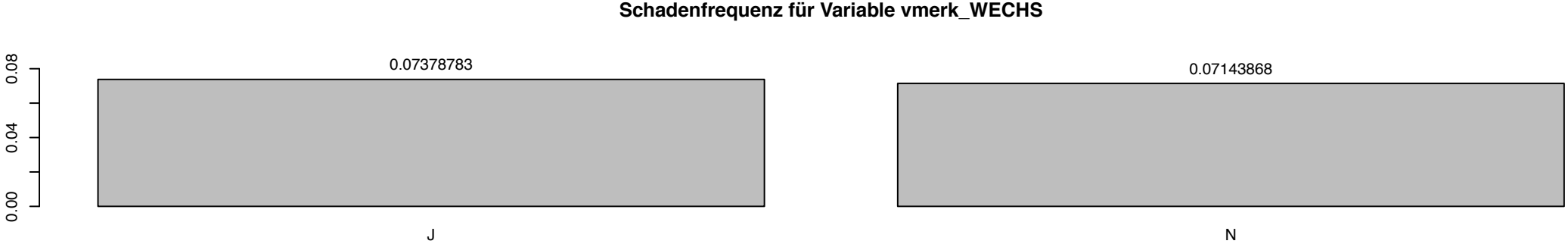

Durchschnittliche Schadenhöhe für Variable VNnatjur
0
500
1500
1811.805
1768.704
J
N

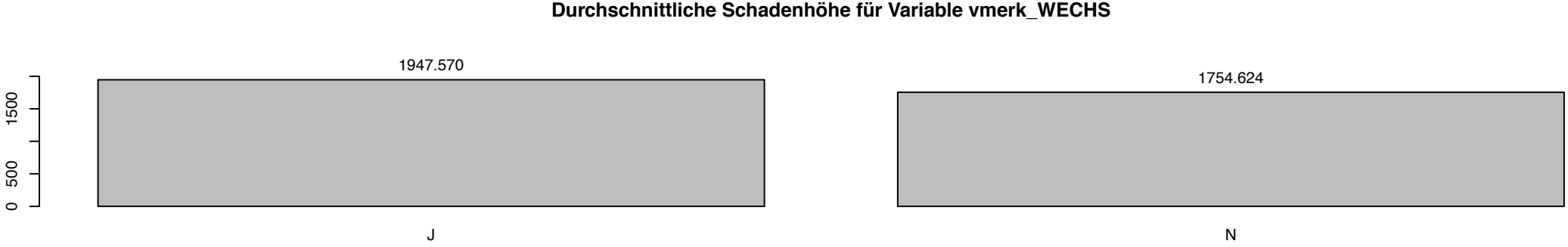

Bestand für Variable vmerk_STAT
0
100000
200000
9.991781e+00
9.698630e−01
4.939726e+00
1.729839e+05
5.018356e+01
einsp_KFZ
Kraftwagen mit besonderer Verwendungsbest.
Kraftwagen zur Güterbeförderung
PKW ohne besond. Bestimmung
PKW zur gewerbsm. Bef.

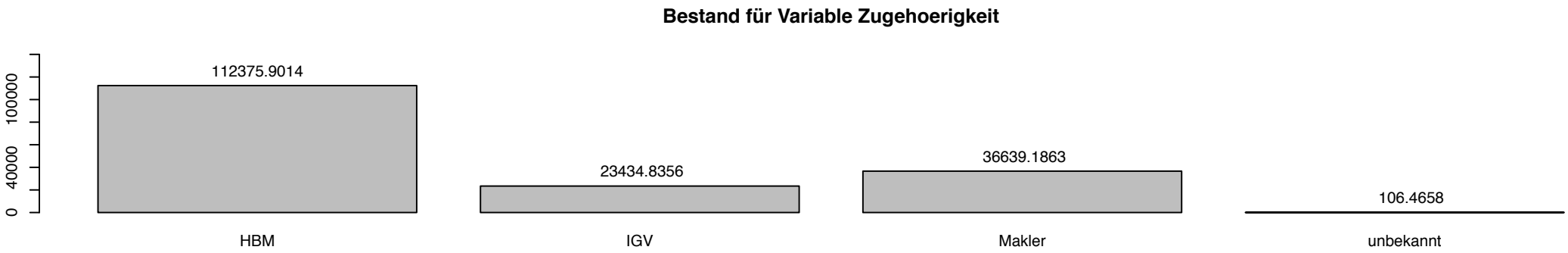

Schadenfrequenz für Variable vmerk_STAT
0.00
0.10
0.10008226
0.00000000
0.00000000
0.07160203
0.15941475
einsp_KFZ
Kraftwagen mit besonderer Verwendungsbest.
Kraftwagen zur Güterbeförderung
PKW ohne besond. Bestimmung
PKW zur gewerbsm. Bef.

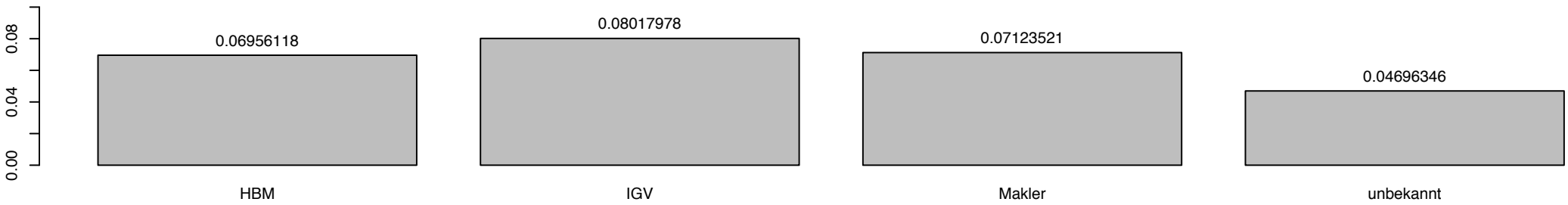

Durchschnittliche Schadenhöhe für Variable vmerk_STAT
0
1000
2000
2272.485
0.000
0.000
1770.917
1113.326
einsp_KFZ
Kraftwagen mit besonderer Verwendungsbest.
Kraftwagen zur Güterbeförderung
PKW ohne besond. Bestimmung
PKW zur gewerbsm. Bef.

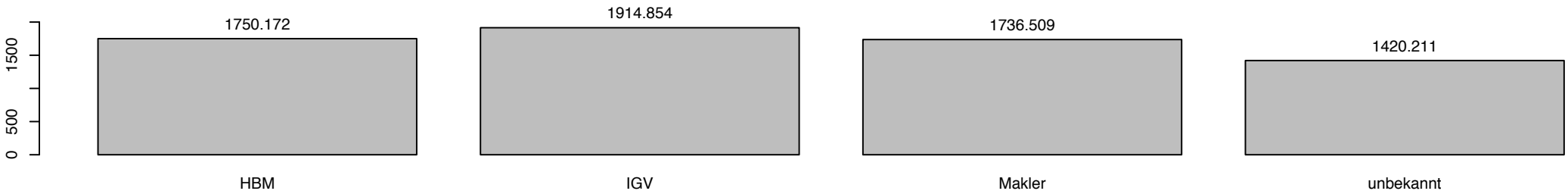

Bestand für Variable vmerk_WECHS
0
50000
150000
13850.52
159199.48
J
N


Schadenfrequenz für Variable vmerk_WECHS
0.00
0.04
0.08
0.07378783
0.07143868
J
N


Durchschnittliche Schadenhöhe für Variable vmerk_WECHS
0
500
1500
1947.570
1754.624
J
N

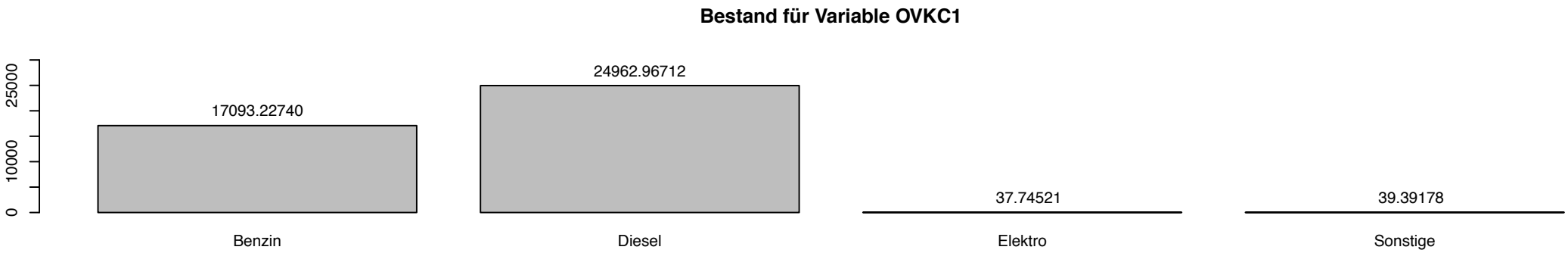

Bestand für Variable Zugehoerigkeit
112375.9014
23434.8356
36639.1863
106.4658
0
40000
100000
HBM
IGV
Makler
unbekannt

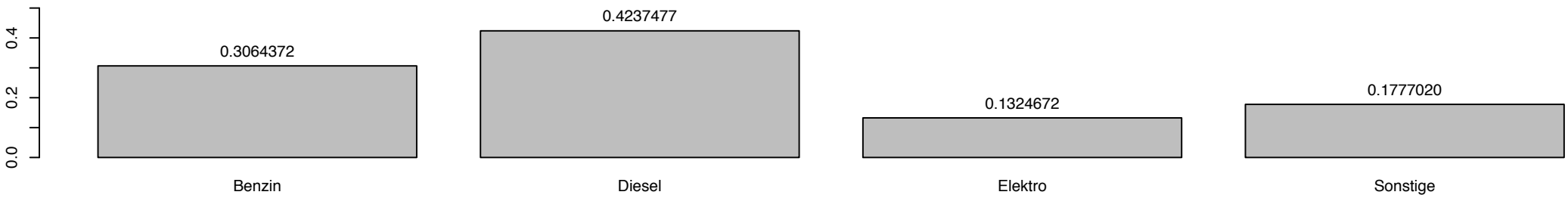

Schadenfrequenz für Variable Zugehoerigkeit
0.06956118
0.08017978
0.07123521
0.04696346
0.00
0.04
0.08
HBM
IGV
Makler
unbekannt

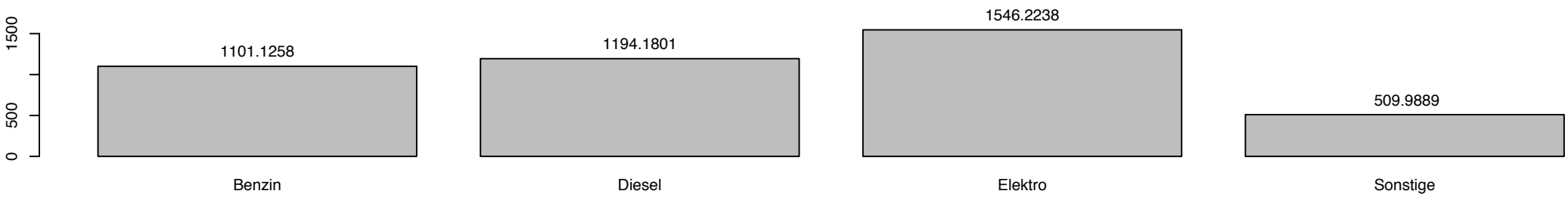

Durchschnittliche Schadenhöhe für Variable Zugehoerigkeit
1750.172
1914.854
1736.509
1420.211
0
500
1500
HBM
IGV
Makler
unbekannt

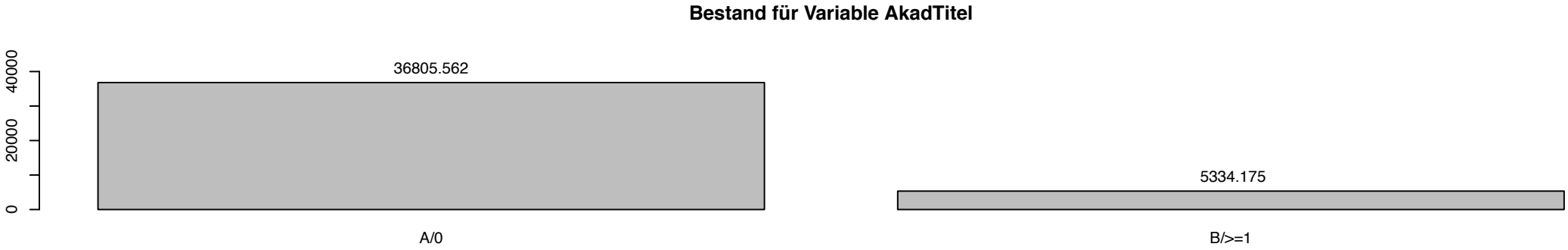
8.2 Deskriptive Analyse für die Vollkaskoversicherung

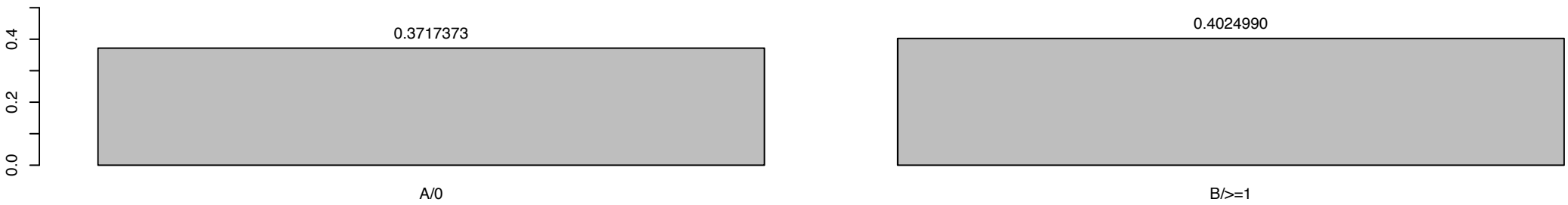
- Antriebsart
- Anzahl akademischer Titel
- Anzahl sonstiger Titel
- Altersklassen
- Bezirksklassen
- Familienstand
- Geschlecht
- Hubraum
- Konzern
- Leistung
- Nation
- Natürliche/Juristische Person
- Wechselkennzeichen
- Zugehörigkeit

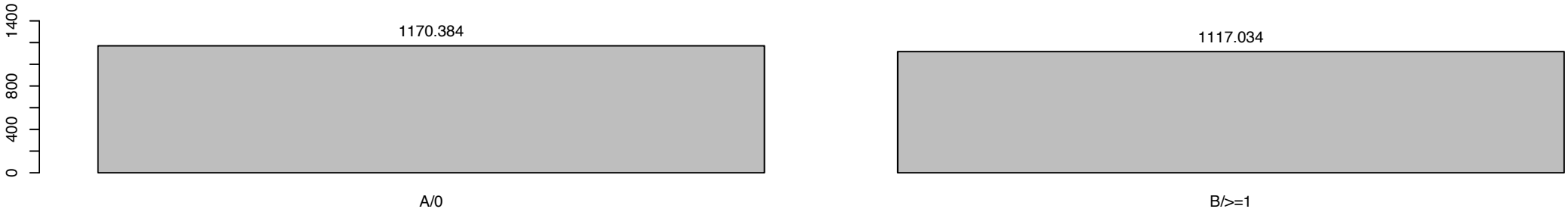

Bestand für Variable OVKC1
0
10000
25000
17093.22740
24962.96712
37.74521
39.39178
Benzin
Diesel
Elektro
Sonstige

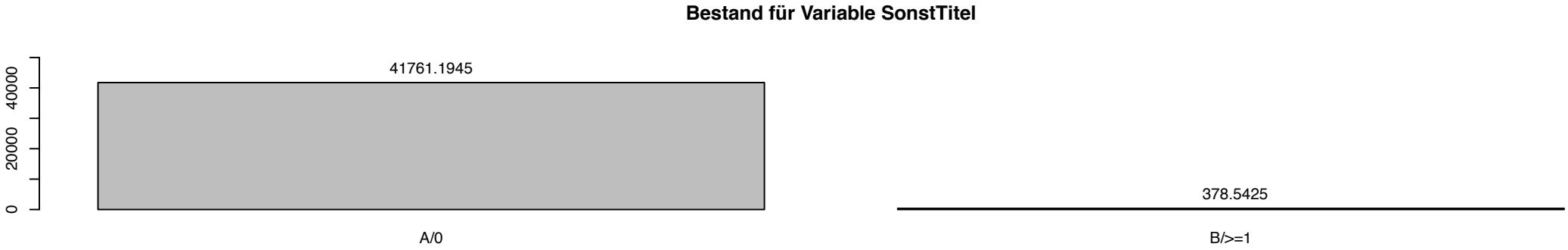

Schadenfrequenz für Variable OVKC1
0.0
0.2
0.4
0.3064372
0.4237477
0.1324672
0.1777020
Benzin
Diesel
Elektro
Sonstige

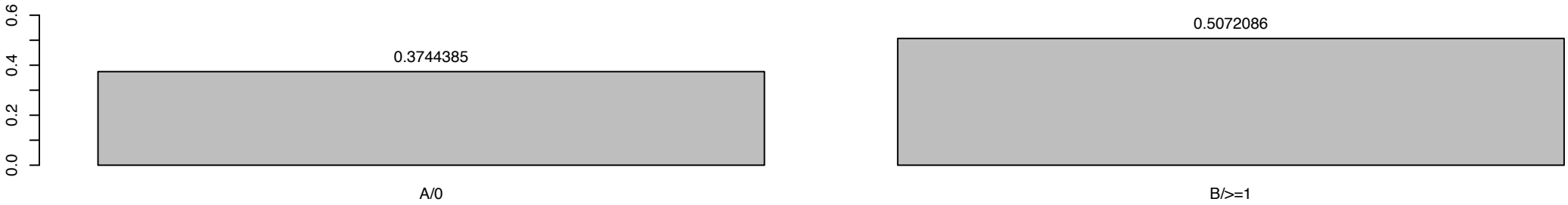

Durchschnittliche Schadenhöhe für Variable OVKC1
0
500
1500
1101.1258
1194.1801
1546.2238
509.9889
Benzin
Diesel
Elektro
Sonstige

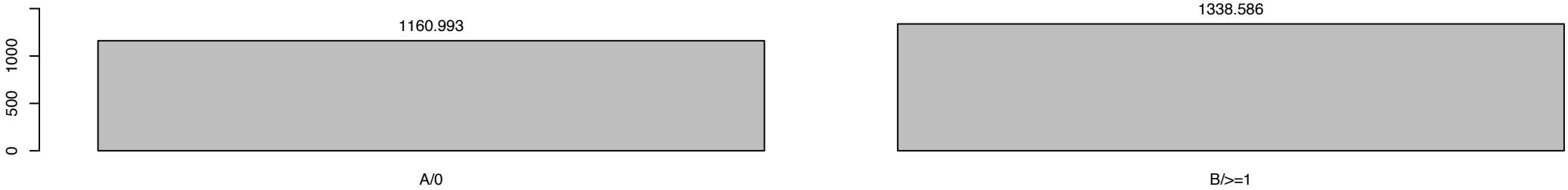

Bestand für Variable AkadTitel
0
20000
40000
36805.562
5334.175
A/0
B/>=1

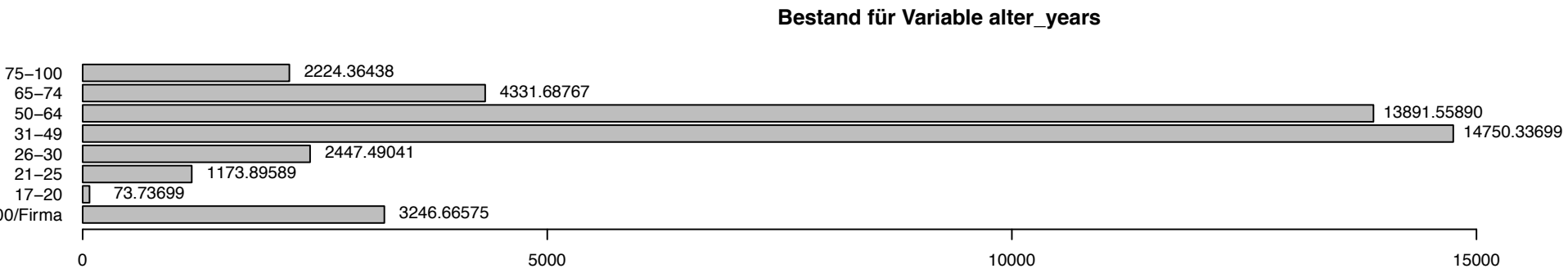

Schadenfrequenz für Variable AkadTitel
0.0
0.2
0.4
0.3717373
0.4024990
A/0
B/>=1

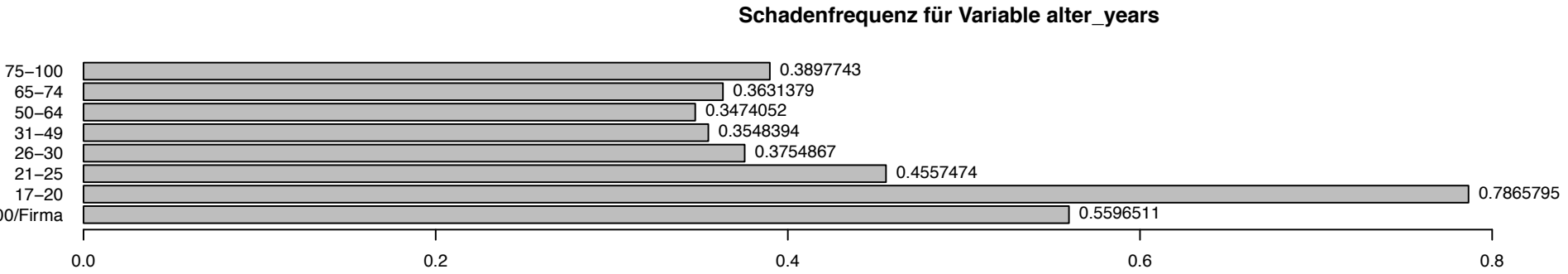

Durchschnittliche Schadenhöhe für Variable AkadTitel
0
400
800
1400
1170.384
1117.034
A/0
B/>=1

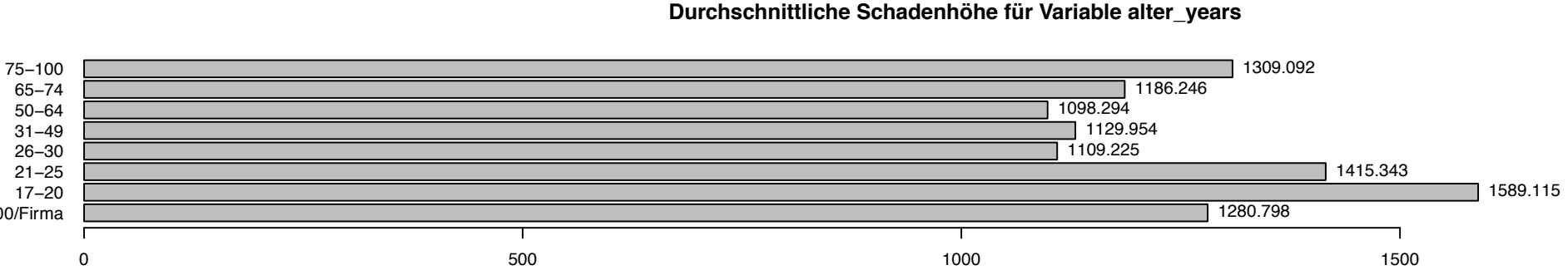

Bestand für Variable SonstTitel
41761.1945
378.5425
0
20000
40000
A/0
B/>=1

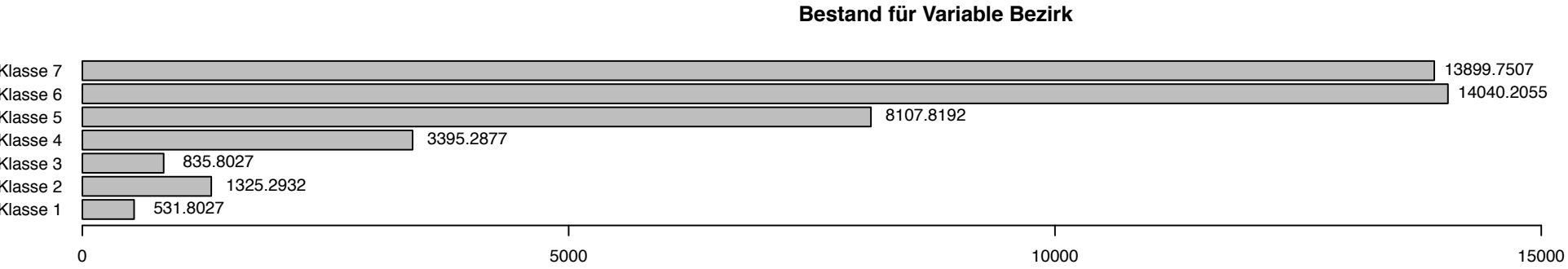

Schadenfrequenz für Variable SonstTitel
0.3744385
0.5072086
0.0
0.2
0.4
0.6
A/0
B/>=1

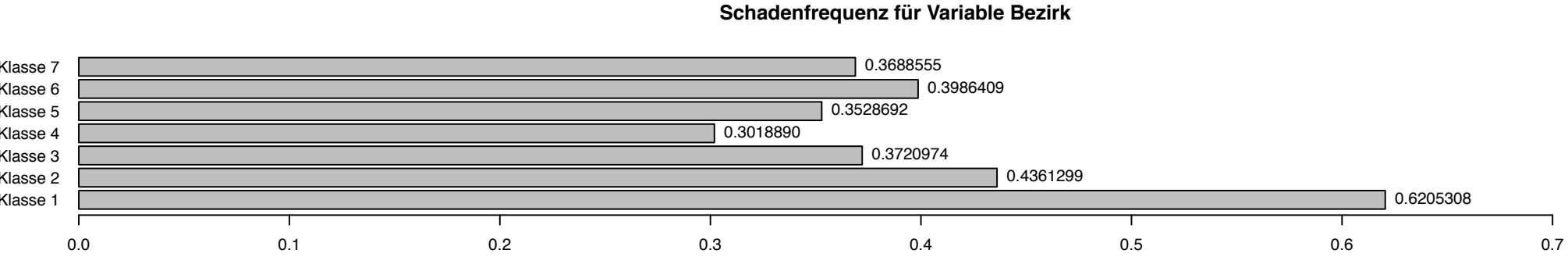

Durchschnittliche Schadenhöhe für Variable SonstTitel
1160.993
1338.586
0
500
1000
A/0
B/>=1

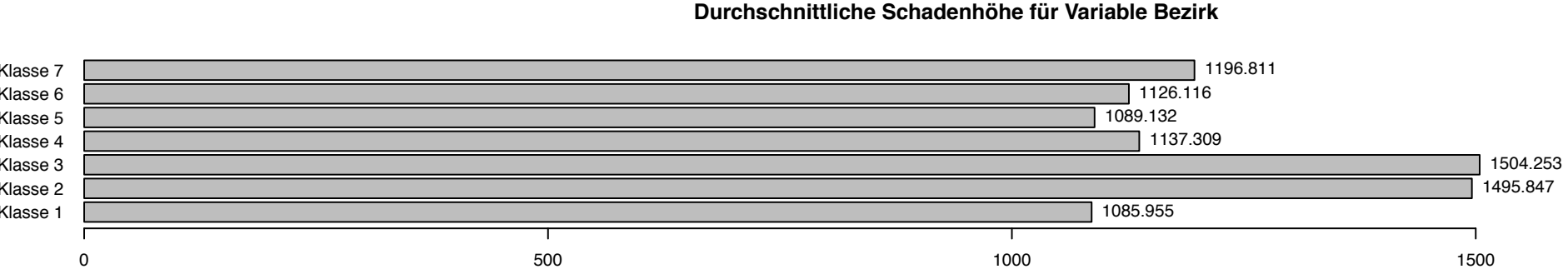

Bestand für Variable alter_years
75–100
65–74
50–64
31–49
26–30
21–25
17–20
00/Firma
2224.36438
4331.68767
13891.55890
14750.33699
2447.49041
1173.89589
73.73699
3246.66575
0
5000
10000
15000

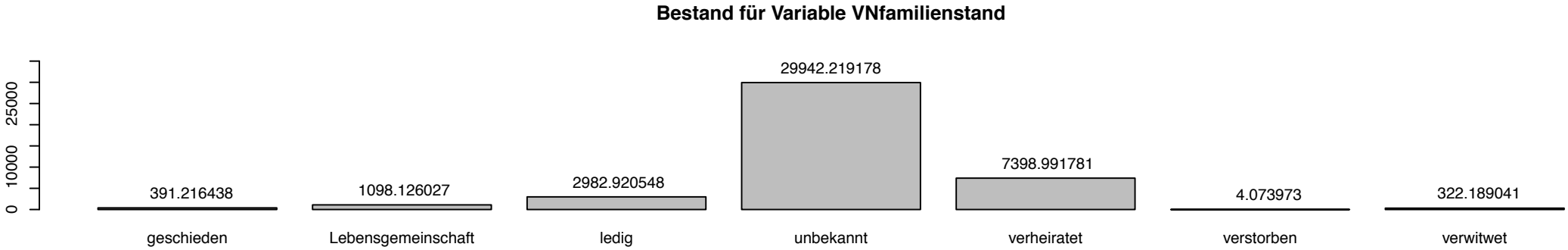

Schadenfrequenz für Variable alter_years
75–100
65–74
50–64
31–49
26–30
21–25
17–20
00/Firma
0.3897743
0.3631379
0.3474052
0.3548394
0.3754867
0.4557474
0.7865795
0.5596511
0.0
0.2
0.4
0.6
0.8

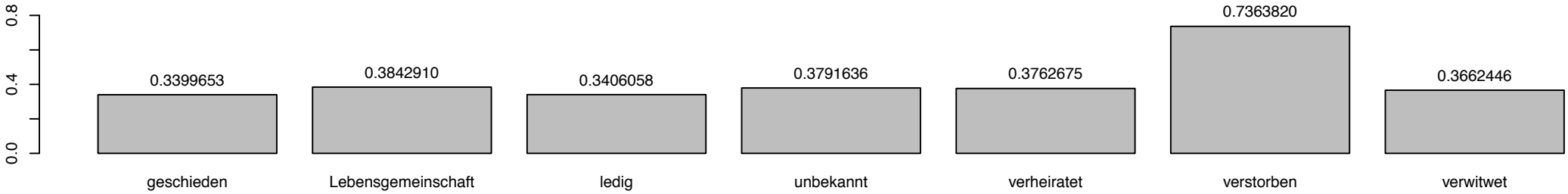

Durchschnittliche Schadenhöhe für Variable alter_years
75–100
65–74
50–64
31–49
26–30
21–25
17–20
00/Firma
1309.092
1186.246
1098.294
1129.954
1109.225
1415.343
1589.115
1280.798
0
500
1000
1500

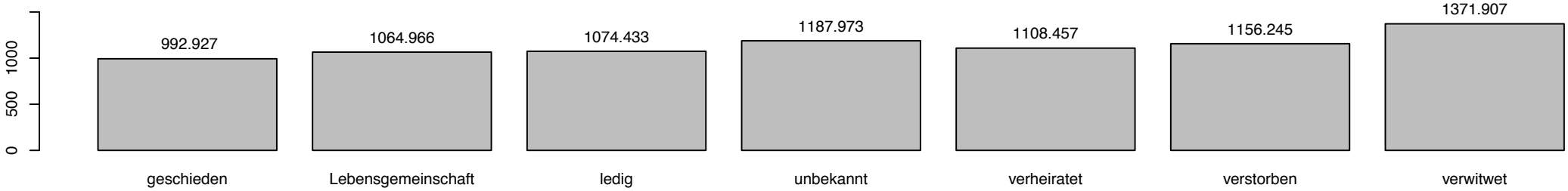

Bestand für Variable Bezirk
Klasse 7
Klasse 6
Klasse 5
Klasse 4
Klasse 3
Klasse 2
Klasse 1
13899.7507
14040.2055
8107.8192
3395.2877
835.8027
1325.2932
531.8027
0
5000
10000
15000

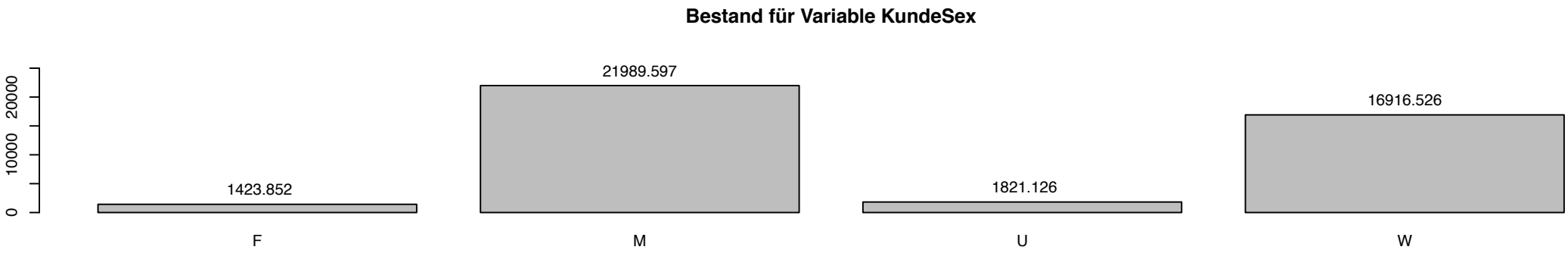

Schadenfrequenz für Variable Bezirk
Klasse 7
Klasse 6
Klasse 5
Klasse 4
Klasse 3
Klasse 2
Klasse 1
0.3688555
0.3986409
0.3528692
0.3018890
0.3720974
0.4361299
0.6205308
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7

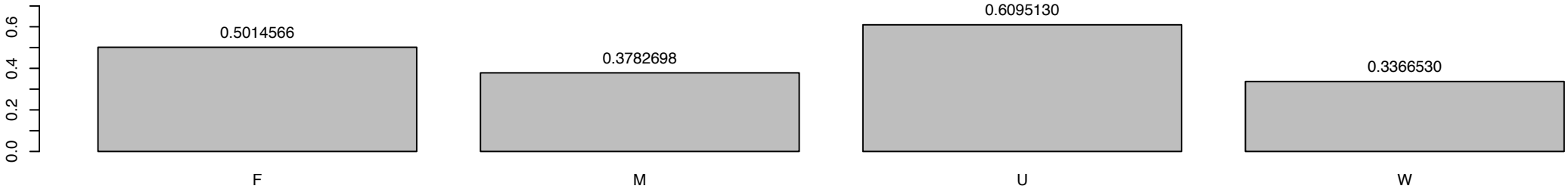

Durchschnittliche Schadenhöhe für Variable Bezirk
Klasse 7
Klasse 6
Klasse 5
Klasse 4
Klasse 3
Klasse 2
Klasse 1
1196.811
1126.116
1089.132
1137.309
1504.253
1495.847
1085.955
0
500
1000
1500

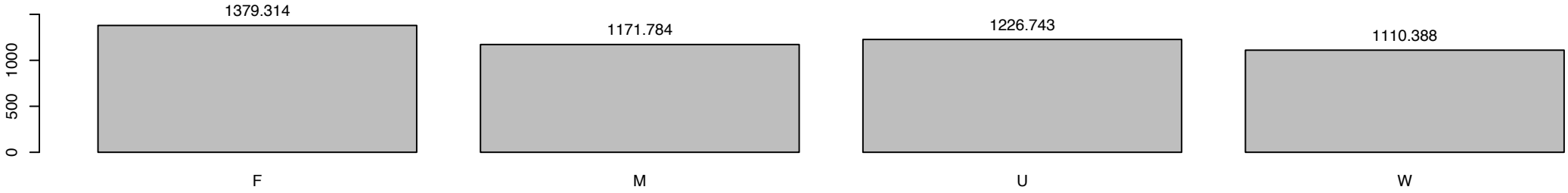

Bestand für Variable VNfamilienstand
0
10000
25000
391.216438
1098.126027
2982.920548
29942.219178
7398.991781
4.073973
322.189041
geschieden
Lebensgemeinschaft
ledig
unbekannt
verheiratet
verstorben
verwitwet

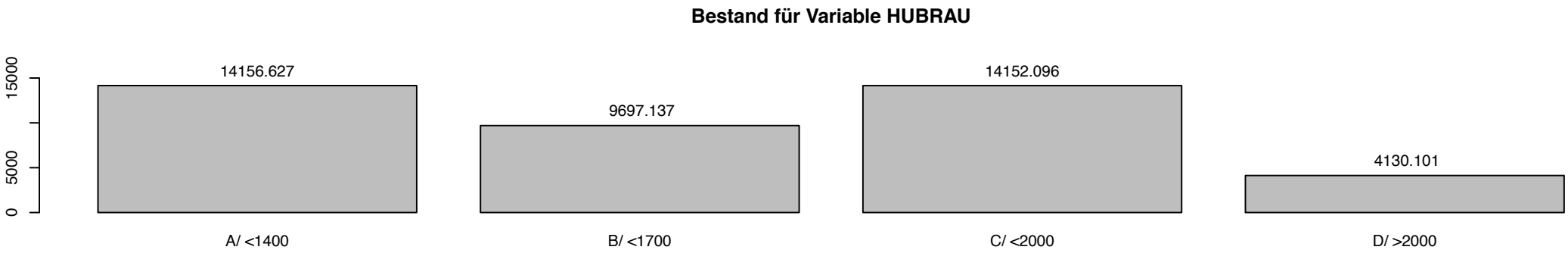

Schadenfrequenz für Variable VNfamilienstand
0.0
0.4
0.8
0.3399653
0.3842910
0.3406058
0.3791636
0.3762675
0.7363820
0.3662446
geschieden
Lebensgemeinschaft
ledig
unbekannt
verheiratet
verstorben
verwitwet

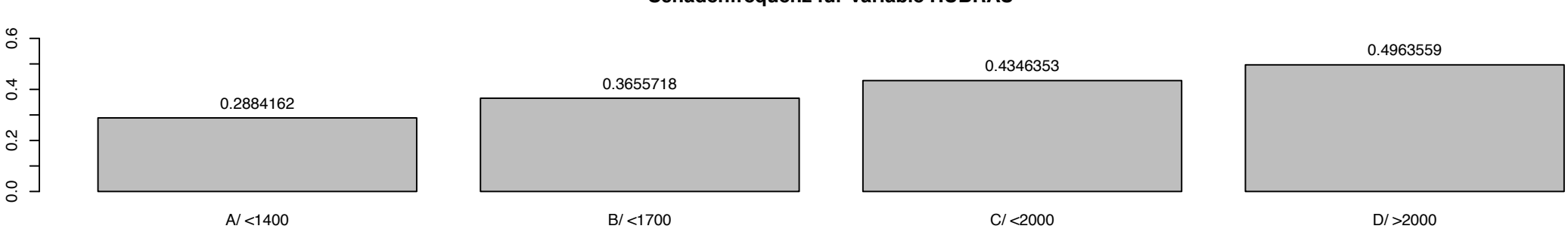

Durchschnittliche Schadenhöhe für Variable VNfamilienstand
0
500
1000
992.927
1064.966
1074.433
1187.973
1108.457
1156.245
1371.907
geschieden
Lebensgemeinschaft
ledig
unbekannt
verheiratet
verstorben
verwitwet

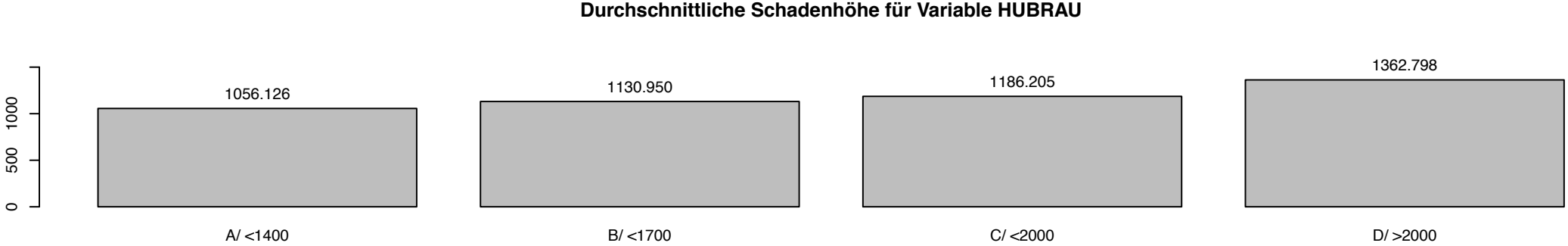

Bestand für Variable KundeSex
0
10000
20000
1423.852
21989.597
1821.126
16916.526
F
M
U
W

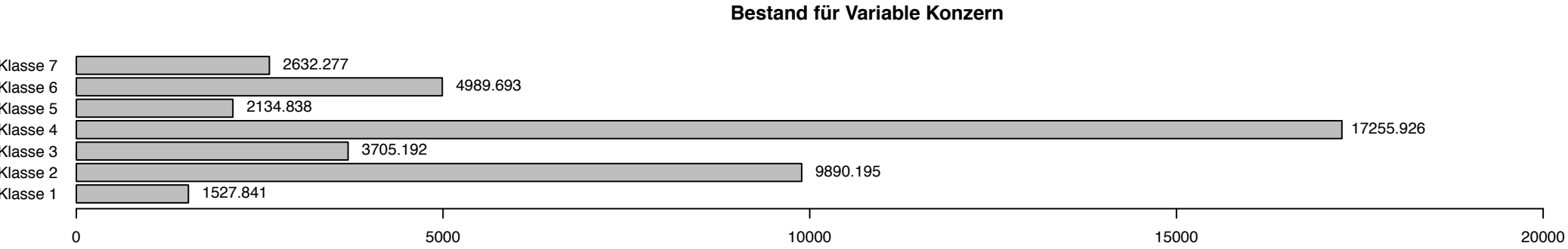

Schadenfrequenz für Variable KundeSex
0.0
0.2
0.4
0.6
0.5014566
0.3782698
0.6095130
0.3366530
F
M
U
W

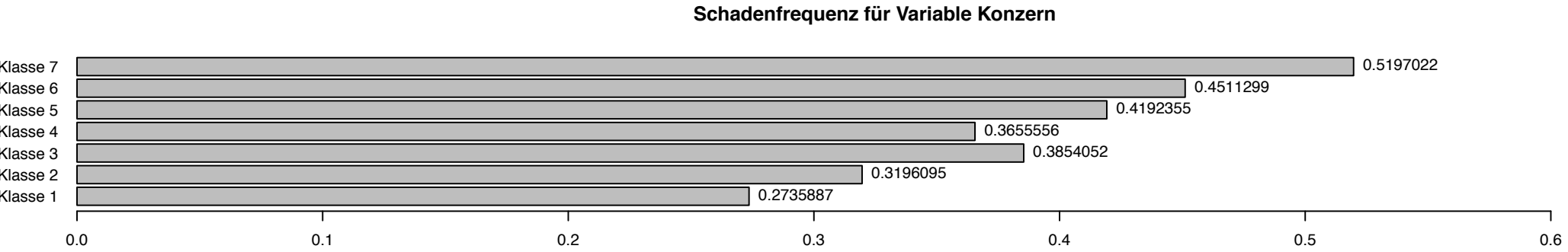

Durchschnittliche Schadenhöhe für Variable KundeSex
0
500
1000
1379.314
1171.784
1226.743
1110.388
F
M
U
W

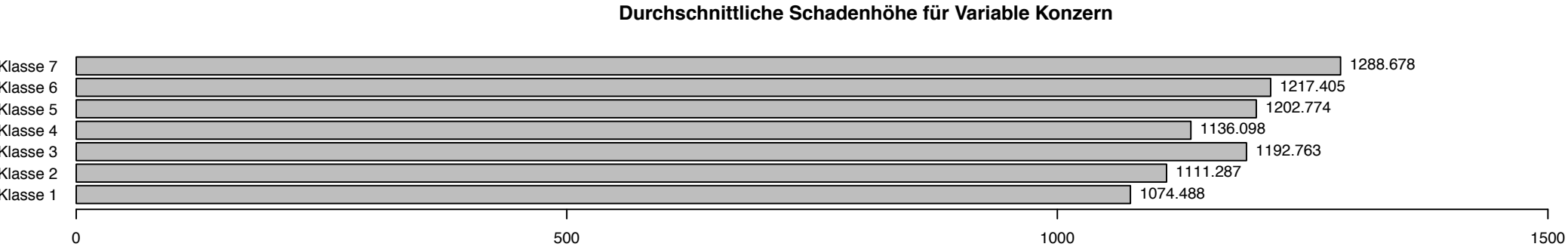

Bestand für Variable HUBRAU
14156.627
9697.137
14152.096
4130.101
0
5000
15000
A/ <1400
B/ <1700
C/ <2000
D/ >2000

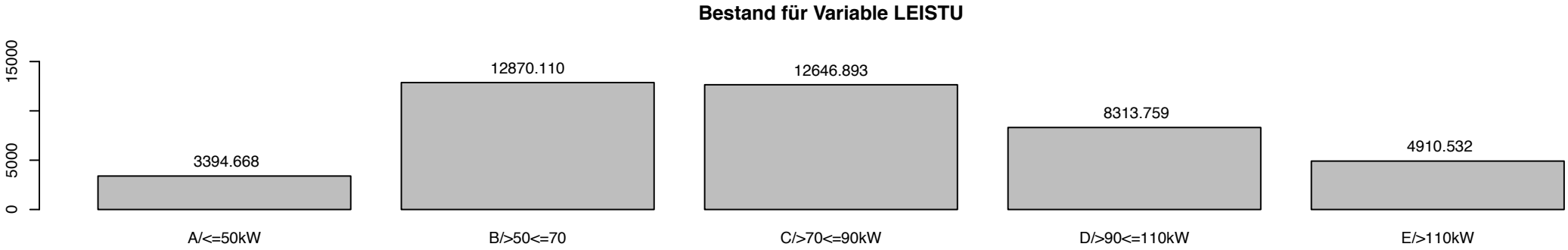

Schadenfrequenz für Variable HUBRAU
0.2884162
0.3655718
0.4346353
0.4963559
0.0
0.2
0.4
0.6
A/ <1400
B/ <1700
C/ <2000
D/ >2000

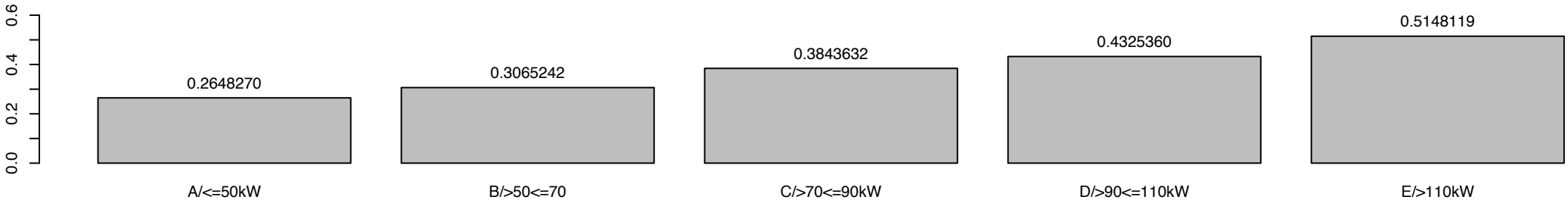

Durchschnittliche Schadenhöhe für Variable HUBRAU
1056.126
1130.950
1186.205
1362.798
0
500
1000
A/ <1400
B/ <1700
C/ <2000
D/ >2000

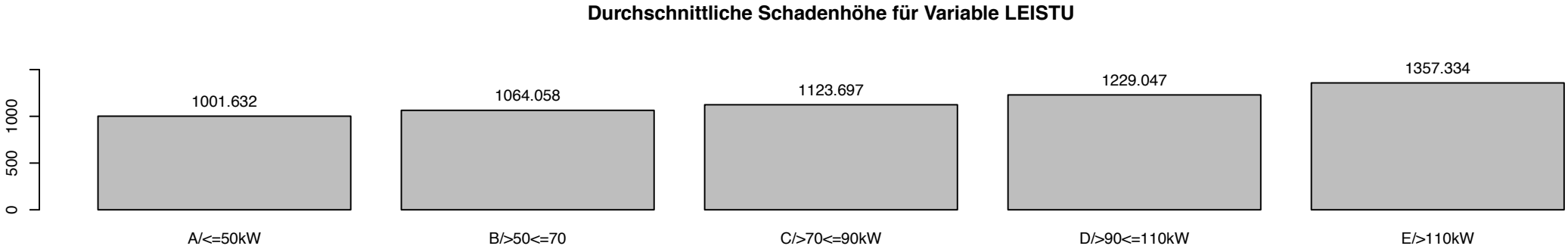

Bestand für Variable Konzern
Klasse 7
Klasse 6
Klasse 5
Klasse 4
Klasse 3
Klasse 2
Klasse 1
2632.277
4989.693
2134.838
17255.926
3705.192
9890.195
1527.841
0
5000
10000
15000
20000

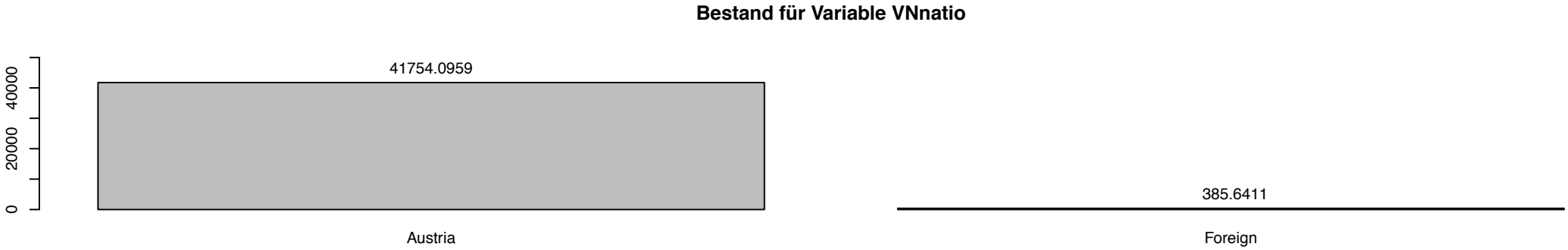

Schadenfrequenz für Variable Konzern
Klasse 7
Klasse 6
Klasse 5
Klasse 4
Klasse 3
Klasse 2
Klasse 1
0.5197022
0.4511299
0.4192355
0.3655556
0.3854052
0.3196095
0.2735887
0.0
0.1
0.2
0.3
0.4
0.5
0.6

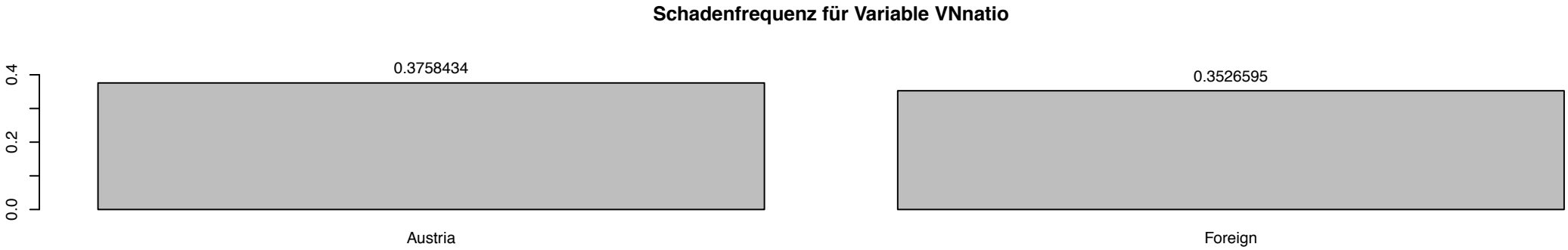

Durchschnittliche Schadenhöhe für Variable Konzern
Klasse 7
Klasse 6
Klasse 5
Klasse 4
Klasse 3
Klasse 2
Klasse 1
1288.678
1217.405
1202.774
1136.098
1192.763
1111.287
1074.488
0
500
1000
1500

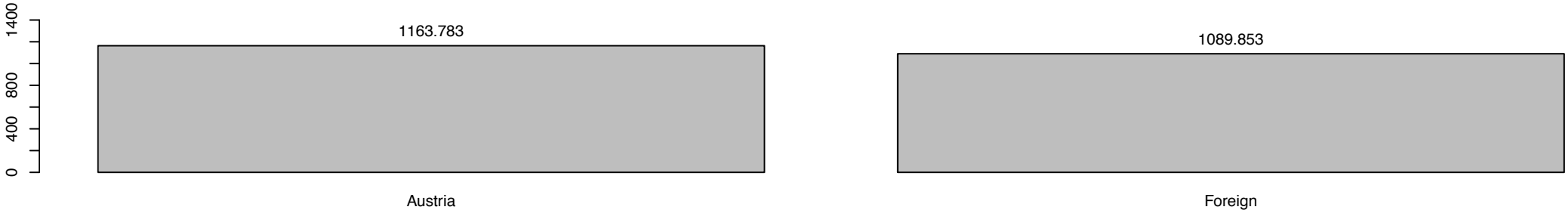

Bestand für Variable LEISTU
15000
5000
0
3394.668
12870.110
12646.893
8313.759
4910.532
A/<=50kW
B/>50<=70
C/>70<=90kW
D/>90<=110kW
E/>110kW

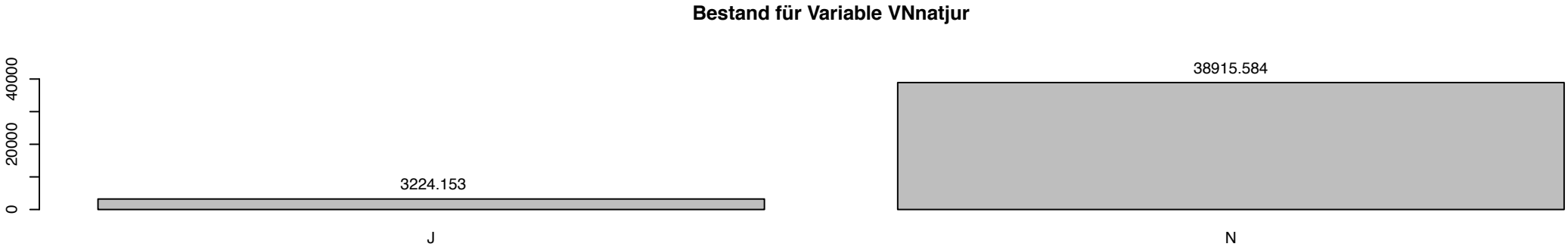

Schadenfrequenz für Variable LEISTU
0.6
0.4
0.2
0.0
0.2648270
0.3065242
0.3843632
0.4325360
0.5148119
A/<=50kW
B/>50<=70
C/>70<=90kW
D/>90<=110kW
E/>110kW

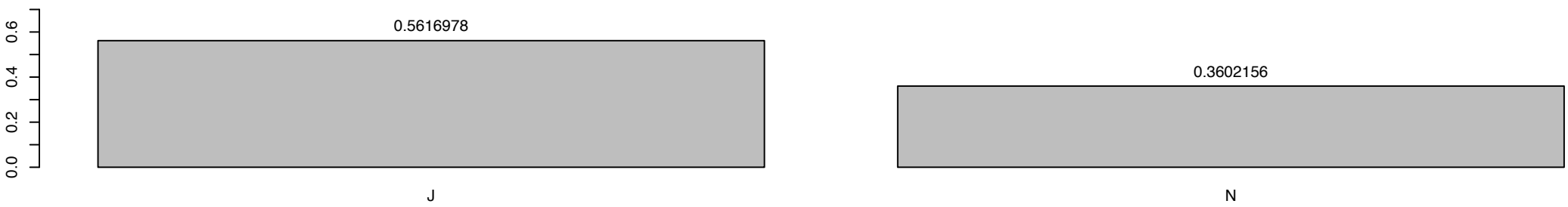

Durchschnittliche Schadenhöhe für Variable LEISTU
1000
500
0
1001.632
1064.058
1123.697
1229.047
1357.334
A/<=50kW
B/>50<=70
C/>70<=90kW
D/>90<=110kW
E/>110kW

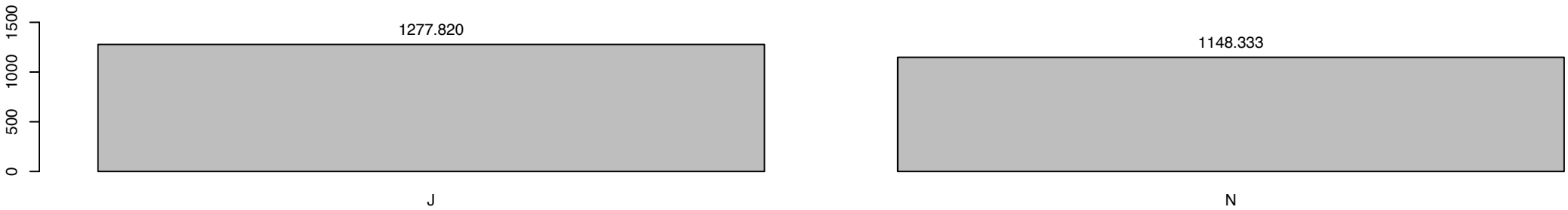

Bestand für Variable VNnatio
41754.0959
385.6411
0
20000
40000
Austria
Foreign

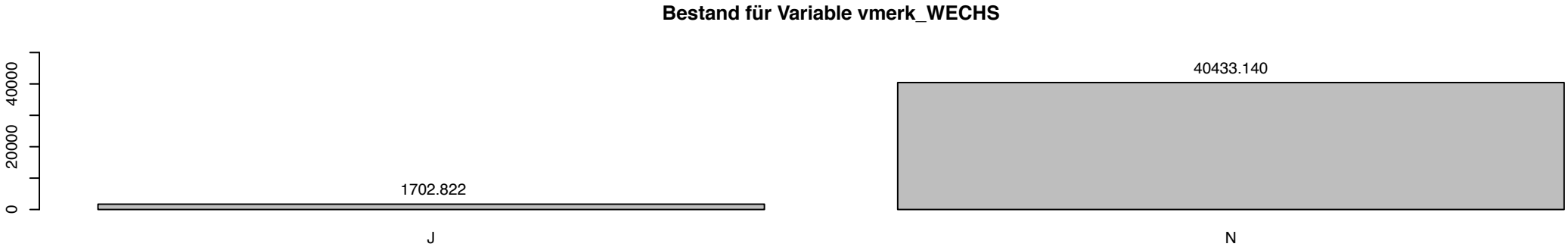

Schadenfrequenz für Variable VNnatio
0.3758434
0.3526595
0.0
0.2
0.4
Austria
Foreign

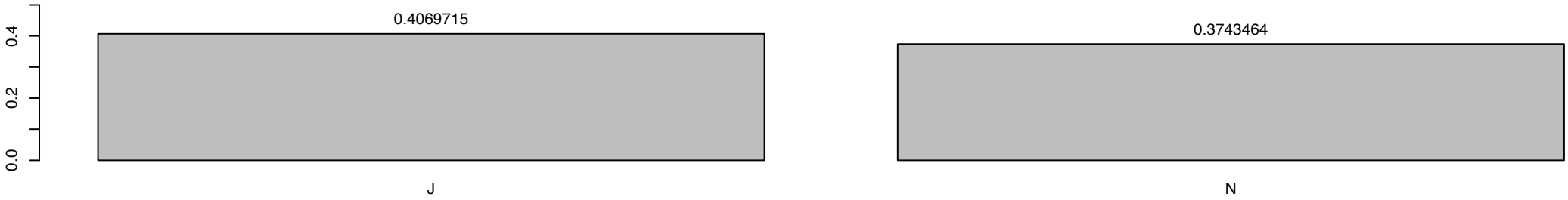

Durchschnittliche Schadenhöhe für Variable VNnatio
1163.783
1089.853
0
400
800
1400
Austria
Foreign

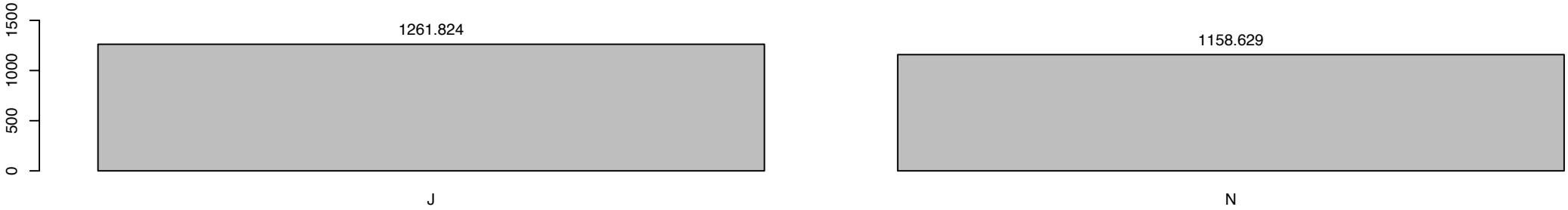

Bestand für Variable VNnatjur
0
20000
40000
3224.153
38915.584
J
N

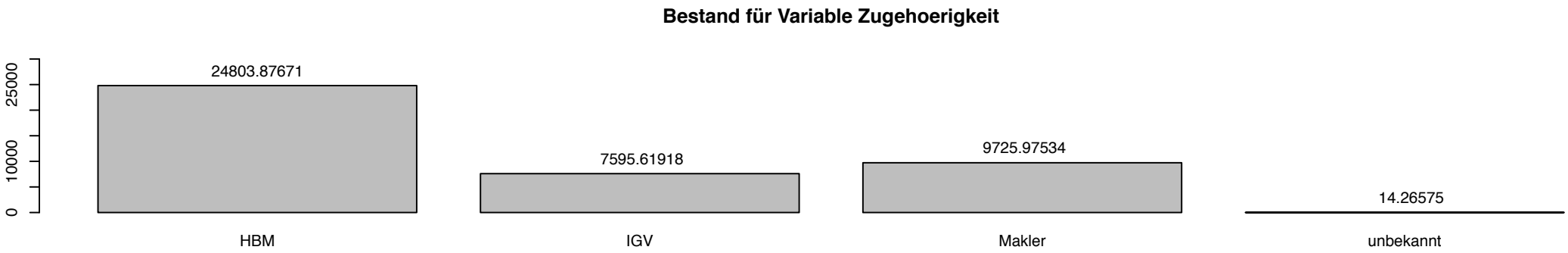

Schadenfrequenz für Variable VNnatjur
0.0
0.2
0.4
0.6
0.5616978
0.3602156
J
N

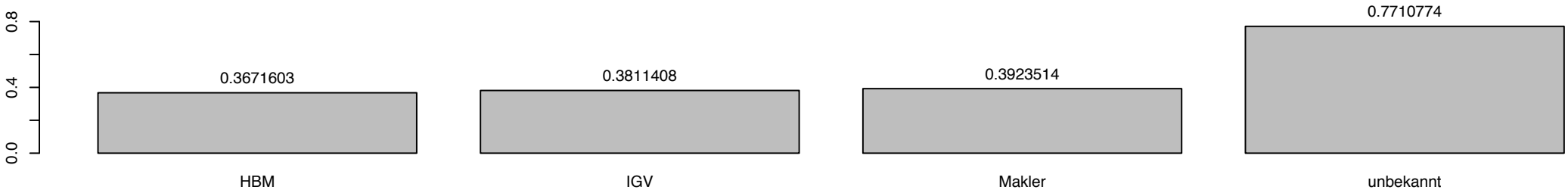

Durchschnittliche Schadenhöhe für Variable VNnatjur
0
500
1000
1500
1277.820
1148.333
J
N

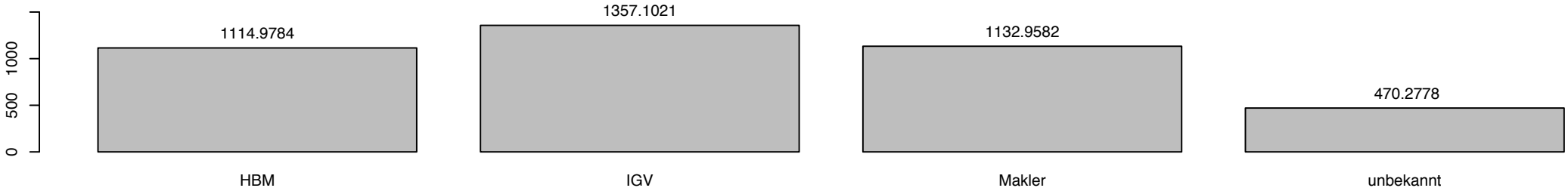

Bestand für Variable vmerk_WECHS
0
20000
40000
1702.822
40433.140
J
N


Schadenfrequenz für Variable vmerk_WECHS
0.0
0.2
0.4
0.4069715
0.3743464
J
N


Durchschnittliche Schadenhöhe für Variable vmerk_WECHS
0
500
1000
1500
1261.824
1158.629
J
N

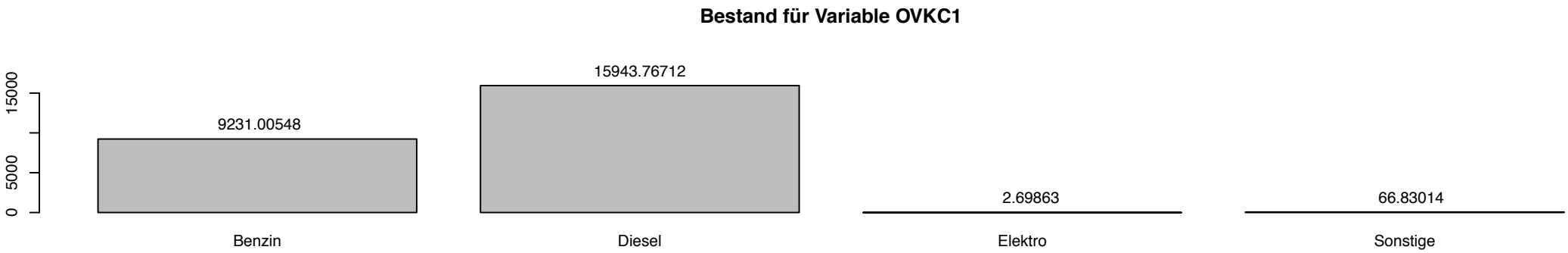

Bestand für Variable Zugehoerigkeit
0
10000
25000
24803.87671
7595.61918
9725.97534
14.26575
HBM
IGV
Makler
unbekannt

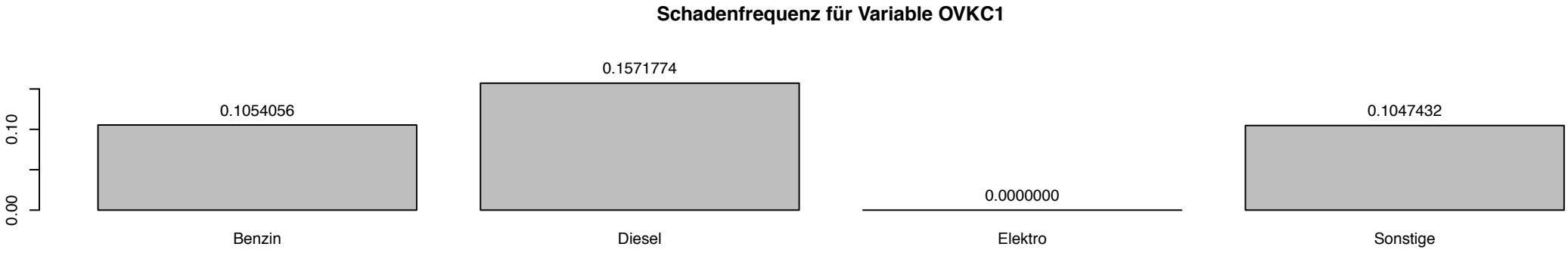

Schadenfrequenz für Variable Zugehoerigkeit
0.0
0.4
0.8
0.3671603
0.3811408
0.3923514
0.7710774
HBM
IGV
Makler
unbekannt

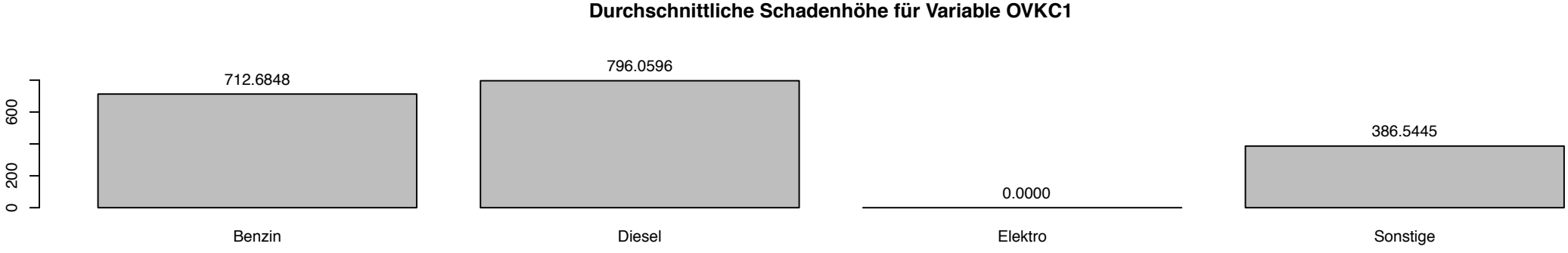

Durchschnittliche Schadenhöhe für Variable Zugehoerigkeit
0
500
1000
1114.9784
1357.1021
1132.9582
470.2778
HBM
IGV
Makler
unbekannt

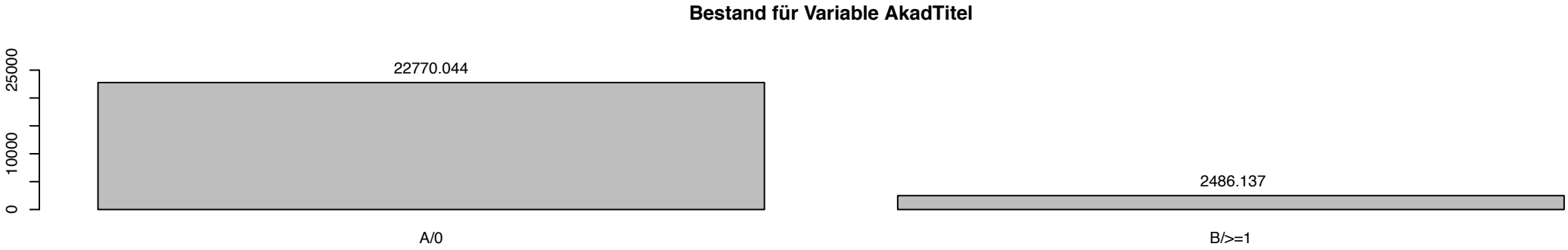
8.3 Deskriptive Analyse für die Teilkaskoversicherung

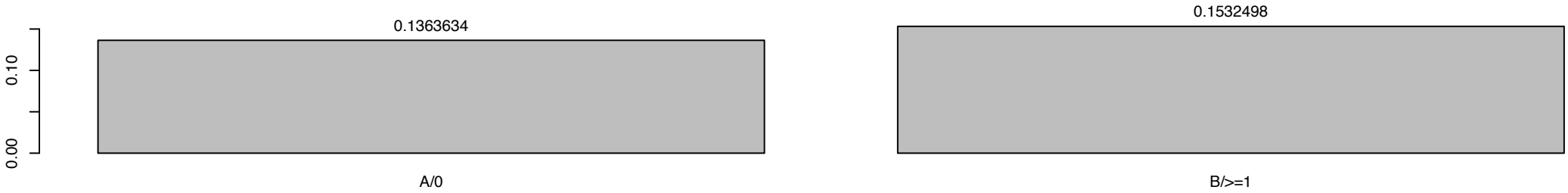
- Antriebsart
- Anzahl akademischer Titel
- Anzahl sonstiger Titel
- Altersklassen
- Bezirksklassen
- Familienstand
- Geschlecht
- Hubraum
- Konzern
- Leistung
- Nation
- Natürliche/Juristische Person
- Wechselkennzeichen
- Zugehörigkeit

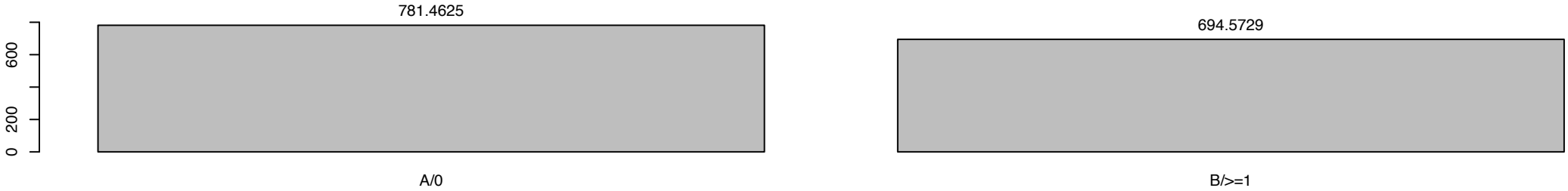

Bestand für Variable OVKC1
0
5000
15000
9231.00548
15943.76712
2.69863
66.83014
Benzin
Diesel
Elektro
Sonstige

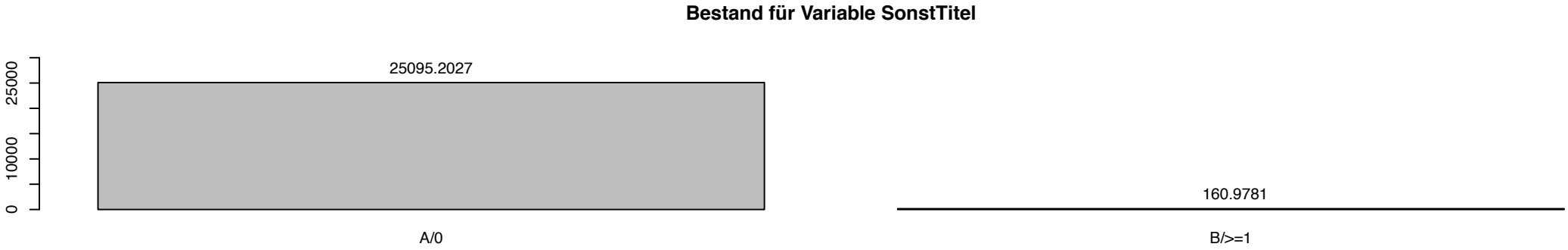

Schadenfrequenz für Variable OVKC1
0.00
0.10
0.1054056
0.1571774
0.0000000
0.1047432
Benzin
Diesel
Elektro
Sonstige

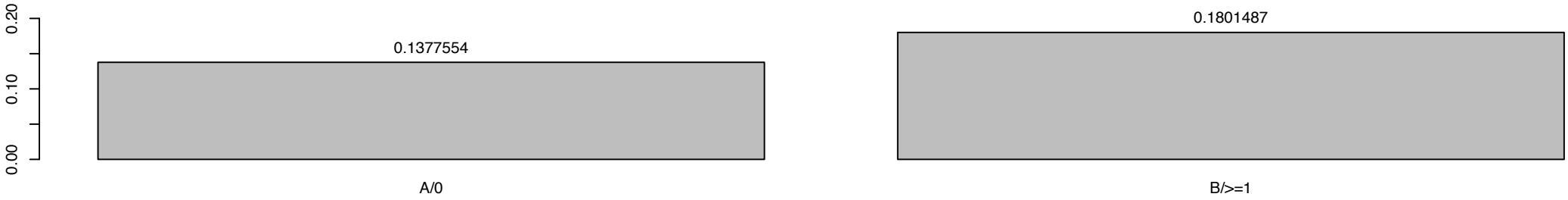

Durchschnittliche Schadenhöhe für Variable OVKC1
0
200
600
712.6848
796.0596
0.0000
386.5445
Benzin
Diesel
Elektro
Sonstige

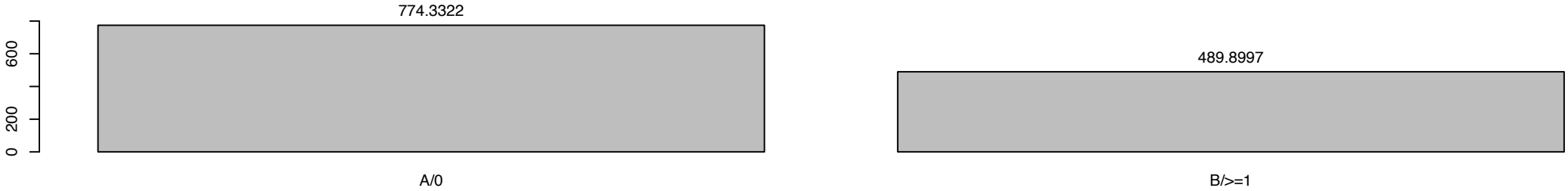

Bestand für Variable AkadTitel
25000
10000
0
22770.044
2486.137
A/0
B/>=1

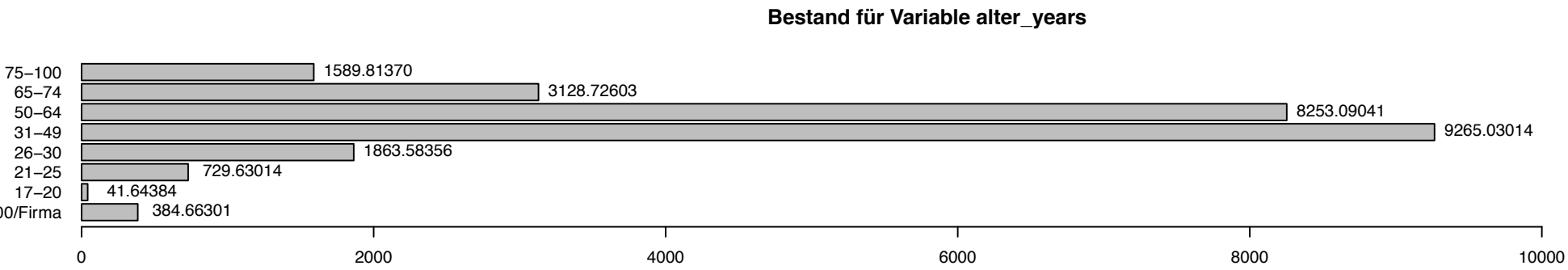

Schadenfrequenz für Variable AkadTitel
0.10
0.00
0.1363634
0.1532498
A/0
B/>=1

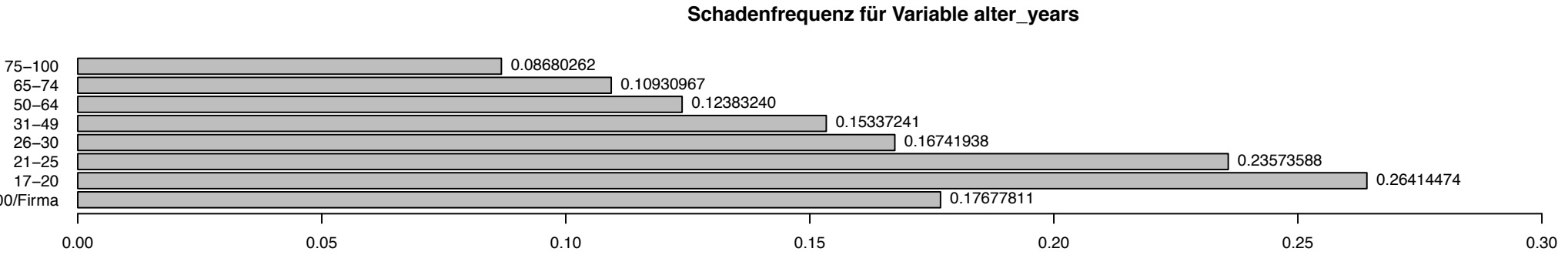

Durchschnittliche Schadenhöhe für Variable AkadTitel
600
200
0
781.4625
694.5729
A/0
B/>=1

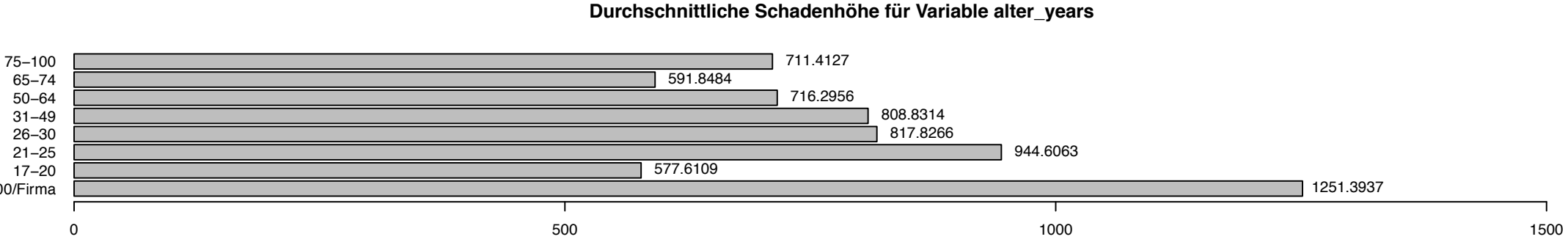

Bestand für Variable SonstTitel
25095.2027
160.9781
0
10000
25000
A/0
B/>=1

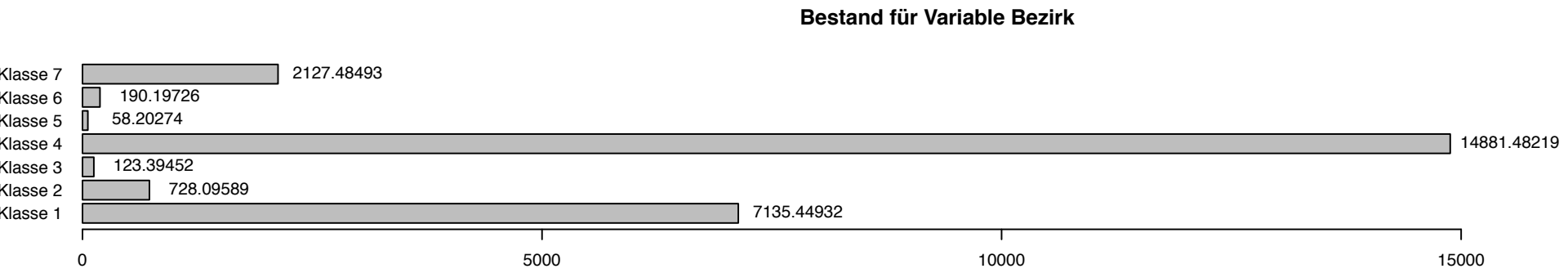

Schadenfrequenz für Variable SonstTitel
0.1377554
0.1801487
0.00
0.10
0.20
A/0
B/>=1

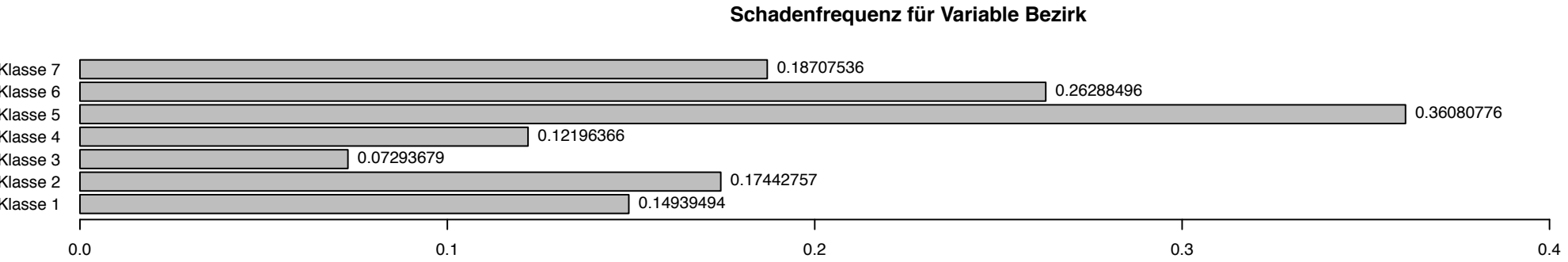

Durchschnittliche Schadenhöhe für Variable SonstTitel
774.3322
489.8997
0
200
600
A/0
B/>=1

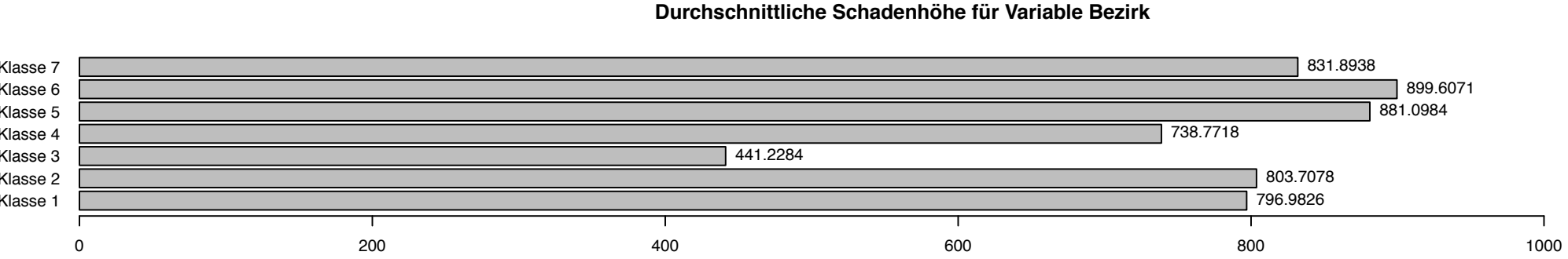

Bestand für Variable alter_years
75–100
65–74
50–64
31–49
26–30
21–25
17–20
00/Firma
1589.81370
3128.72603
8253.09041
9265.03014
1863.58356
729.63014
41.64384
384.66301
0
2000
4000
6000
8000
10000

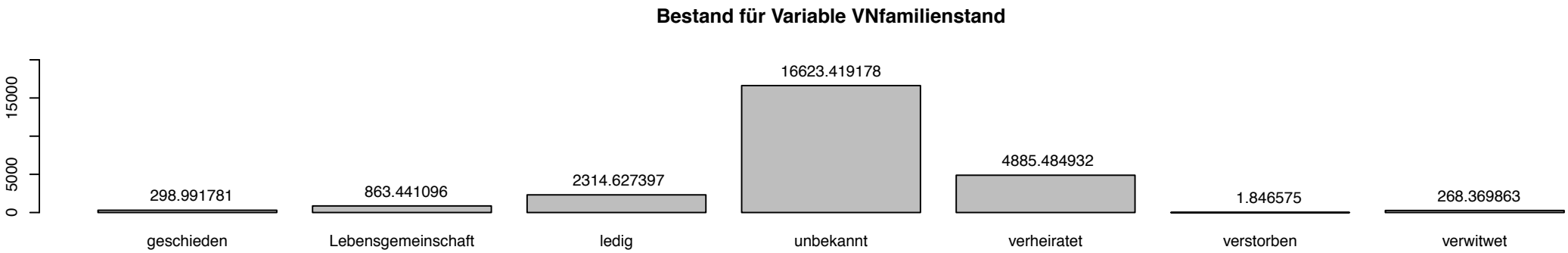

Schadenfrequenz für Variable alter_years
75–100
65–74
50–64
31–49
26–30
21–25
17–20
00/Firma
0.08680262
0.10930967
0.12383240
0.15337241
0.16741938
0.23573588
0.26414474
0.17677811
0.00
0.05
0.10
0.15
0.20
0.25
0.30

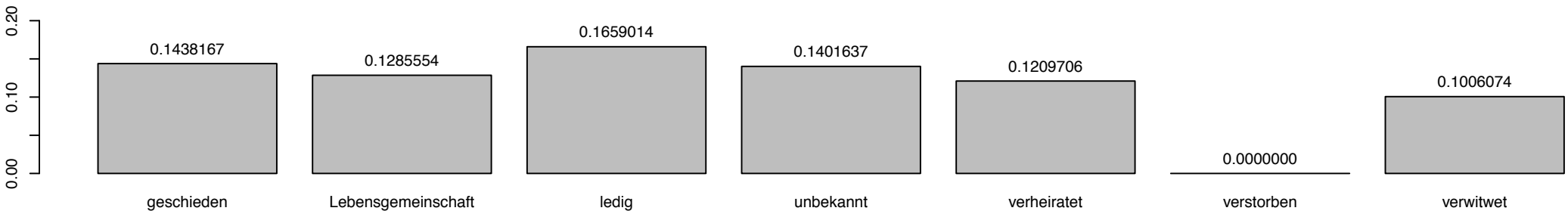

Durchschnittliche Schadenhöhe für Variable alter_years
75–100
65–74
50–64
31–49
26–30
21–25
17–20
00/Firma
711.4127
591.8484
716.2956
808.8314
817.8266
944.6063
577.6109
1251.3937
0
500
1000
1500

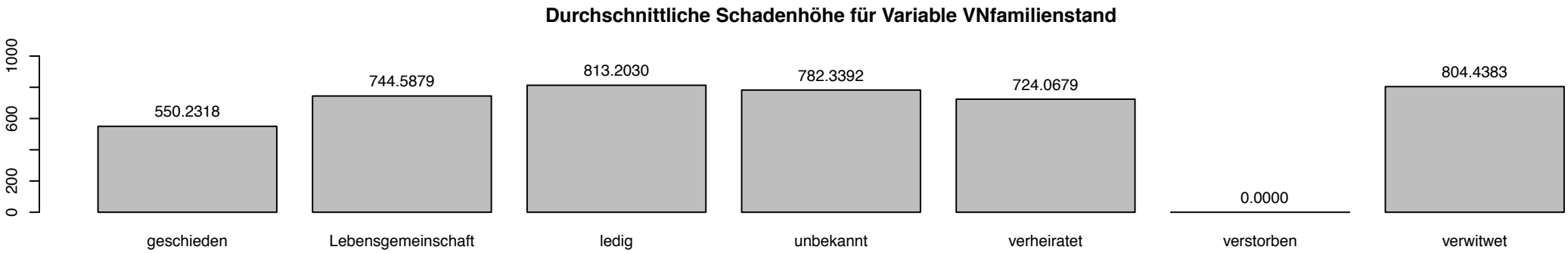

Bestand für Variable Bezirk
Klasse 7
Klasse 6
Klasse 5
Klasse 4
Klasse 3
Klasse 2
Klasse 1
2127.48493
190.19726
58.20274
14881.48219
123.39452
728.09589
7135.44932
0
5000
10000
15000

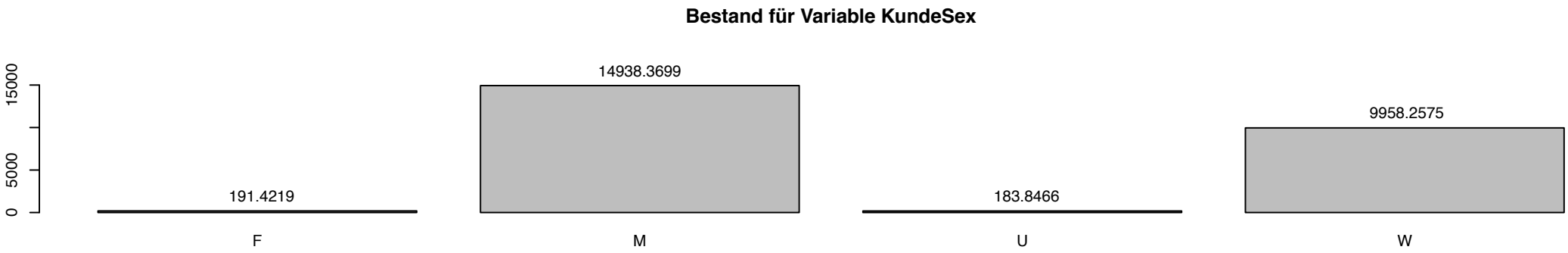

Schadenfrequenz für Variable Bezirk
Klasse 7
Klasse 6
Klasse 5
Klasse 4
Klasse 3
Klasse 2
Klasse 1
0.18707536
0.26288496
0.36080776
0.12196366
0.07293679
0.17442757
0.14939494
0.0
0.1
0.2
0.3
0.4

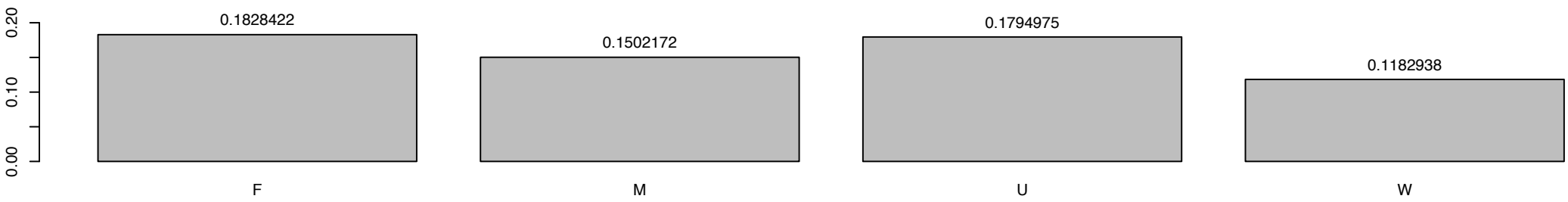

Durchschnittliche Schadenhöhe für Variable Bezirk
Klasse 7
Klasse 6
Klasse 5
Klasse 4
Klasse 3
Klasse 2
Klasse 1
831.8938
899.6071
881.0984
738.7718
441.2284
803.7078
796.9826
0
200
400
600
800
1000

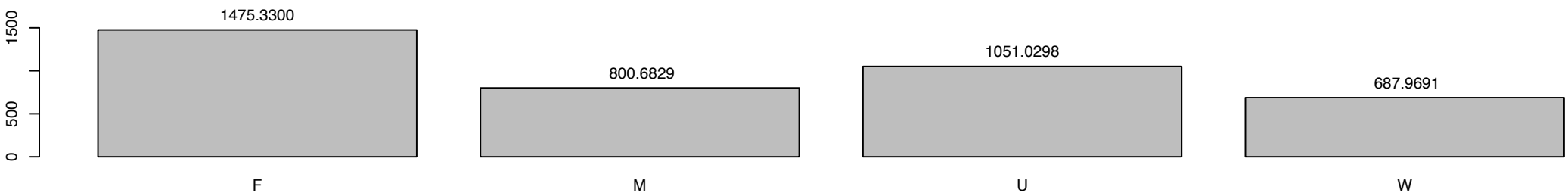

Bestand für Variable VNfamilienstand
0
5000
15000
298.991781
863.441096
2314.627397
16623.419178
4885.484932
1.846575
268.369863
geschieden
Lebensgemeinschaft
ledig
unbekannt
verheiratet
verstorben
verwitwet

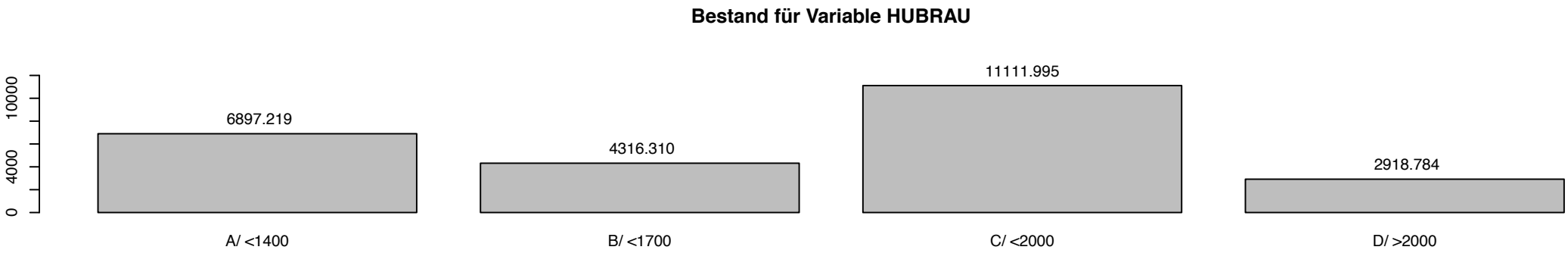

Schadenfrequenz für Variable VNfamilienstand
0.00
0.10
0.20
0.1438167
0.1285554
0.1659014
0.1401637
0.1209706
0.0000000
0.1006074
geschieden
Lebensgemeinschaft
ledig
unbekannt
verheiratet
verstorben
verwitwet

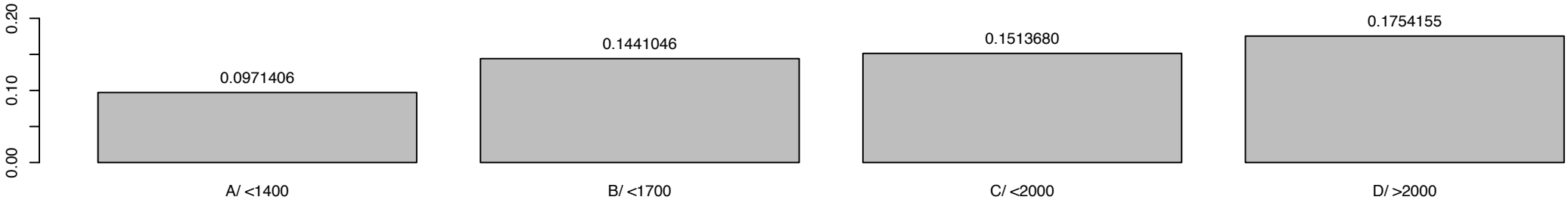

Durchschnittliche Schadenhöhe für Variable VNfamilienstand
0
200
600
1000
550.2318
744.5879
813.2030
782.3392
724.0679
0.0000
804.4383
geschieden
Lebensgemeinschaft
ledig
unbekannt
verheiratet
verstorben
verwitwet

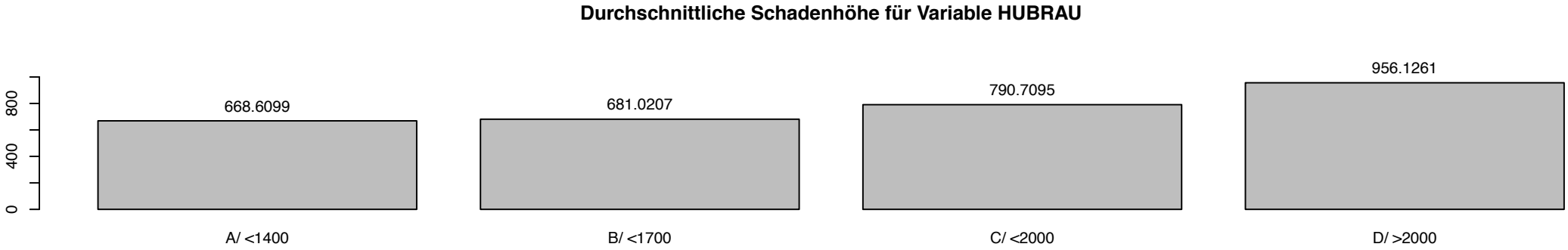

Bestand für Variable KundeSex
0
5000
15000
191.4219
14938.3699
183.8466
9958.2575
F
M
U
W

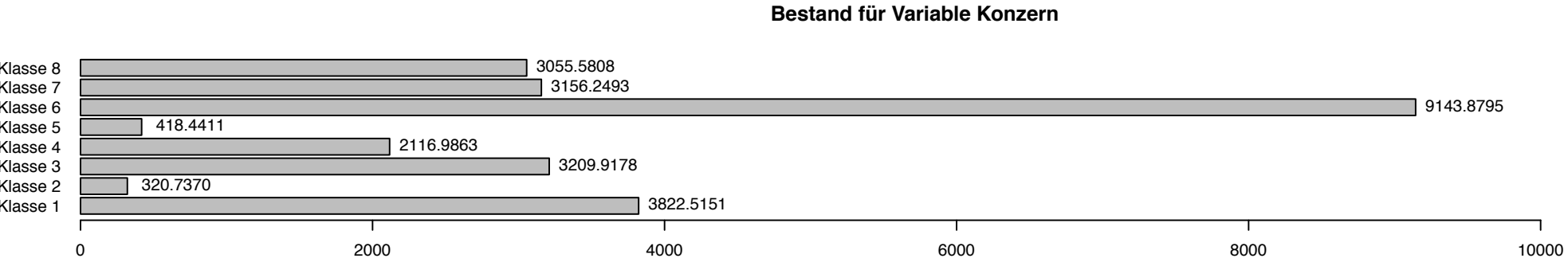

Schadenfrequenz für Variable KundeSex
0.00
0.10
0.20
0.1828422
0.1502172
0.1794975
0.1182938
F
M
U
W

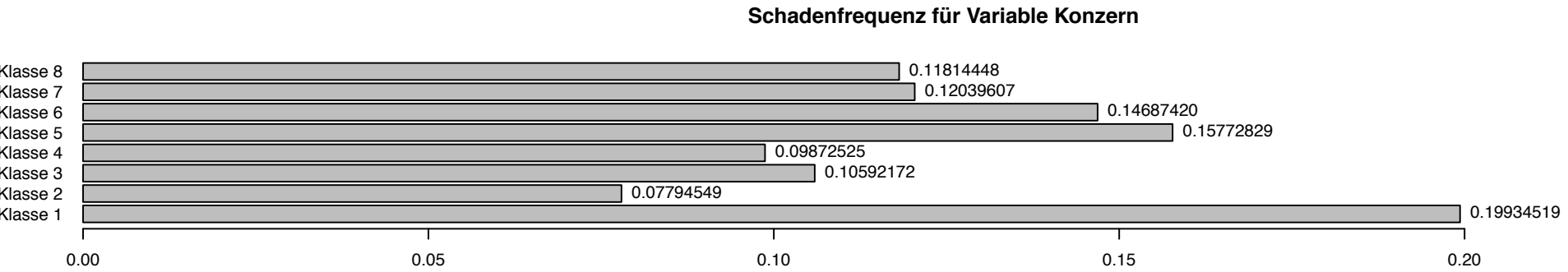

Durchschnittliche Schadenhöhe für Variable KundeSex
0
500
1500
1475.3300
800.6829
1051.0298
687.9691
F
M
U
W

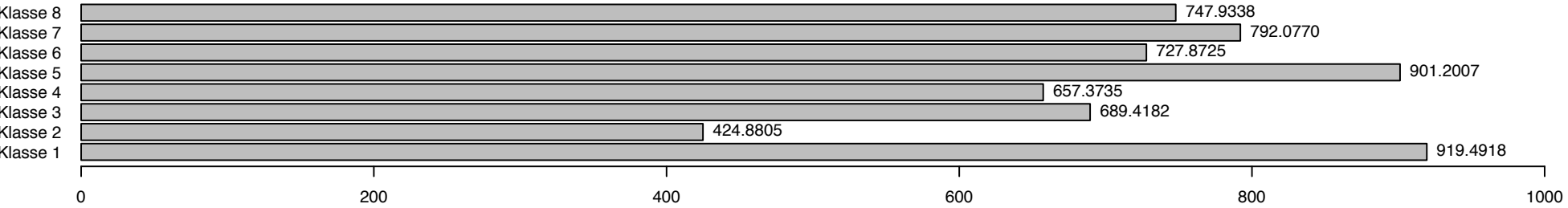

Bestand für Variable HUBRAU
0
4000
10000
6897.219
4316.310
11111.995
2918.784
A/ <1400
B/ <1700
C/ <2000
D/ >2000

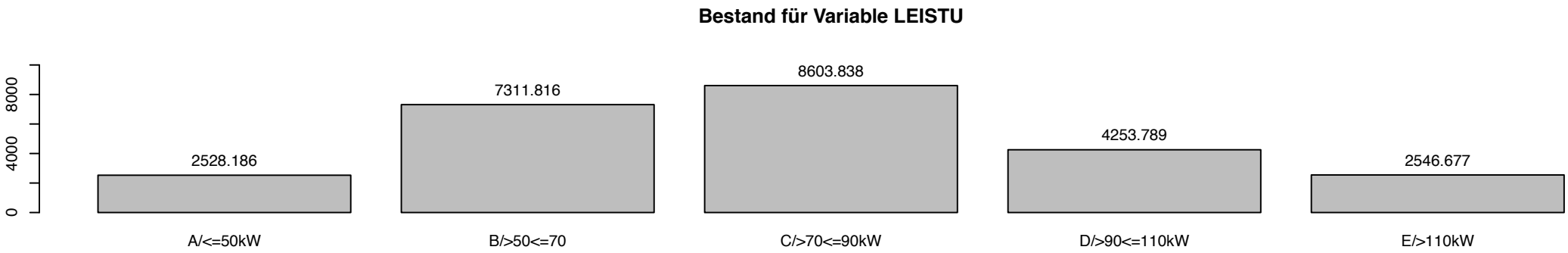

Schadenfrequenz für Variable HUBRAU
0.00
0.10
0.20
0.0971406
0.1441046
0.1513680
0.1754155
A/ <1400
B/ <1700
C/ <2000
D/ >2000

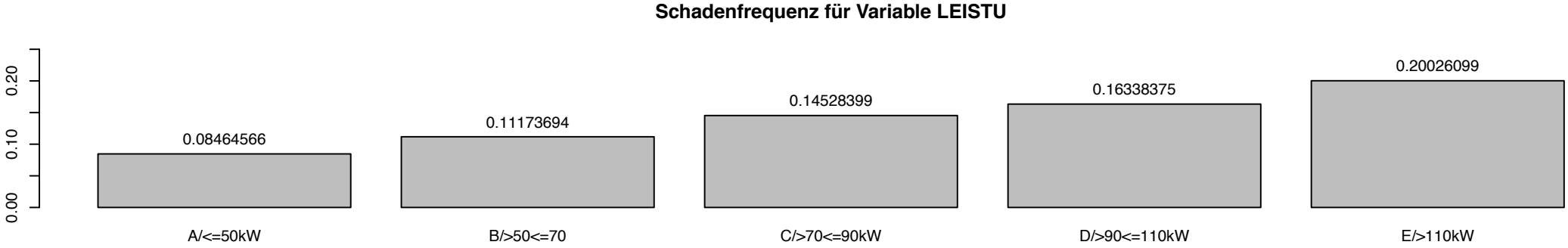

Durchschnittliche Schadenhöhe für Variable HUBRAU
0
400
800
668.6099
681.0207
790.7095
956.1261
A/ <1400
B/ <1700
C/ <2000
D/ >2000

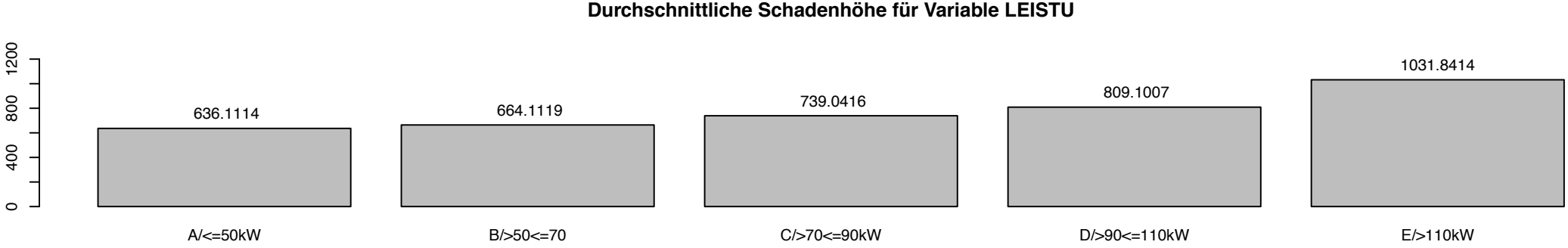

Bestand für Variable Konzern
Klasse 8
Klasse 7
Klasse 6
Klasse 5
Klasse 4
Klasse 3
Klasse 2
Klasse 1
3055.5808
3156.2493
9143.8795
418.4411
2116.9863
3209.9178
320.7370
3822.5151
0
2000
4000
6000
8000
10000

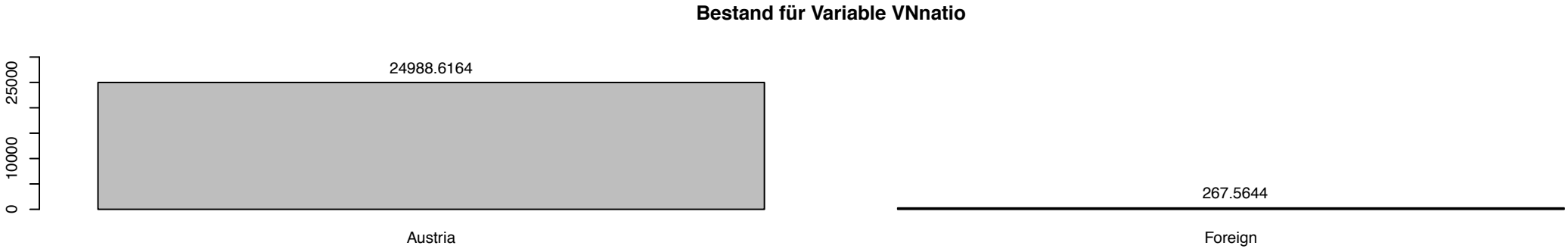

Schadenfrequenz für Variable Konzern
Klasse 8
Klasse 7
Klasse 6
Klasse 5
Klasse 4
Klasse 3
Klasse 2
Klasse 1
0.11814448
0.12039607
0.14687420
0.15772829
0.09872525
0.10592172
0.07794549
0.19934519
0.00
0.05
0.10
0.15
0.20

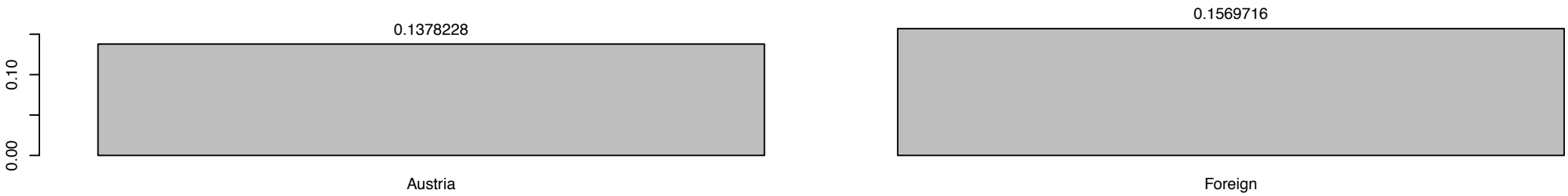

Durchschnittliche Schadenhöhe für Variable Konzern
Klasse 8
Klasse 7
Klasse 6
Klasse 5
Klasse 4
Klasse 3
Klasse 2
Klasse 1
747.9338
792.0770
727.8725
901.2007
657.3735
689.4182
424.8805
919.4918
0
200
400
600
800
1000

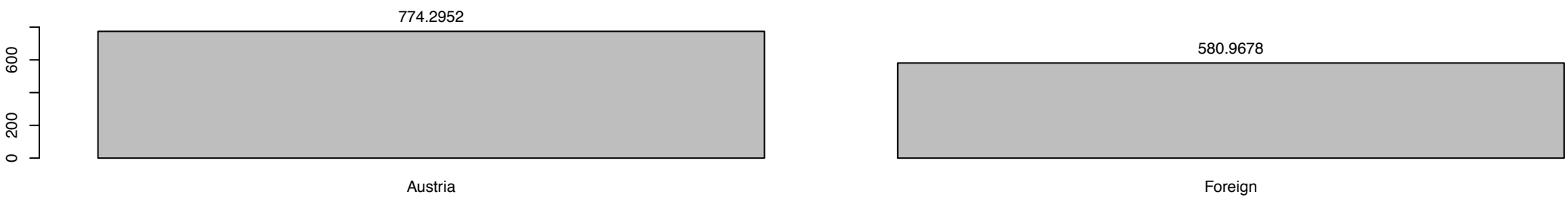

Bestand für Variable LEISTU
0
4000
8000
2528.186
7311.816
8603.838
4253.789
2546.677
A/<=50kW
B/>50<=70
C/>70<=90kW
D/>90<=110kW
E/>110kW

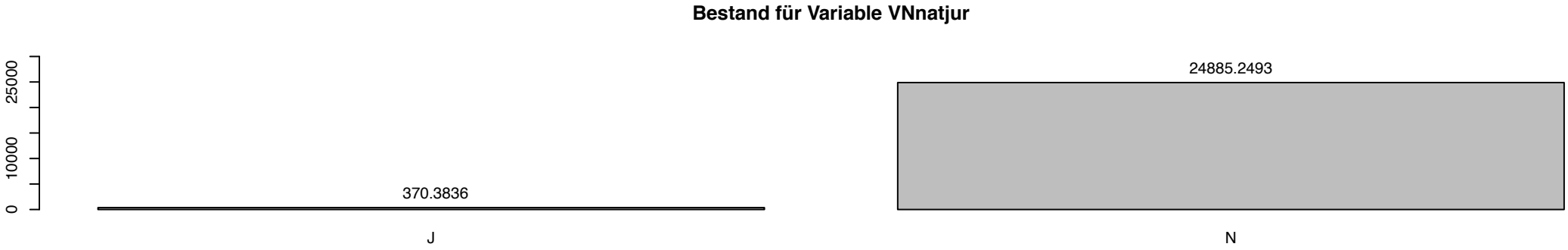

Schadenfrequenz für Variable LEISTU
0.00
0.10
0.20
0.08464566
0.11173694
0.14528399
0.16338375
0.20026099
A/<=50kW
B/>50<=70
C/>70<=90kW
D/>90<=110kW
E/>110kW

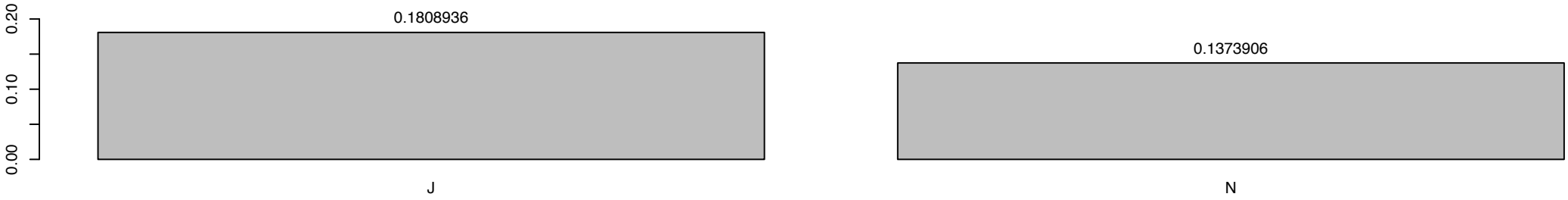

Durchschnittliche Schadenhöhe für Variable LEISTU
0
400
800
1200
636.1114
664.1119
739.0416
809.1007
1031.8414
A/<=50kW
B/>50<=70
C/>70<=90kW
D/>90<=110kW
E/>110kW

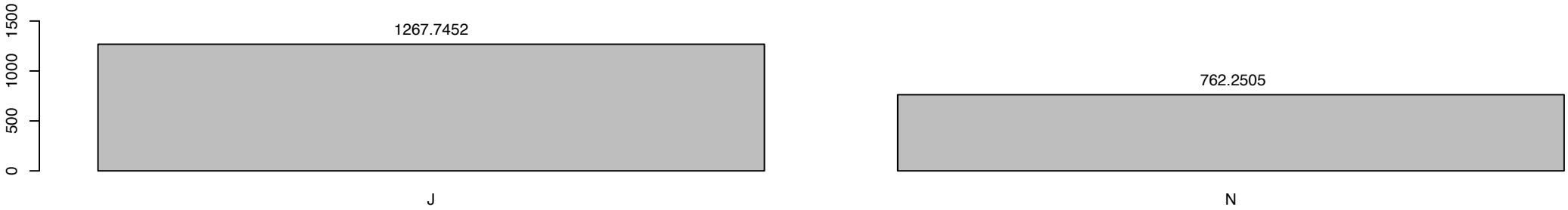

Bestand für Variable VNnatio
24988.6164
267.5644
0
10000
25000
Austria
Foreign

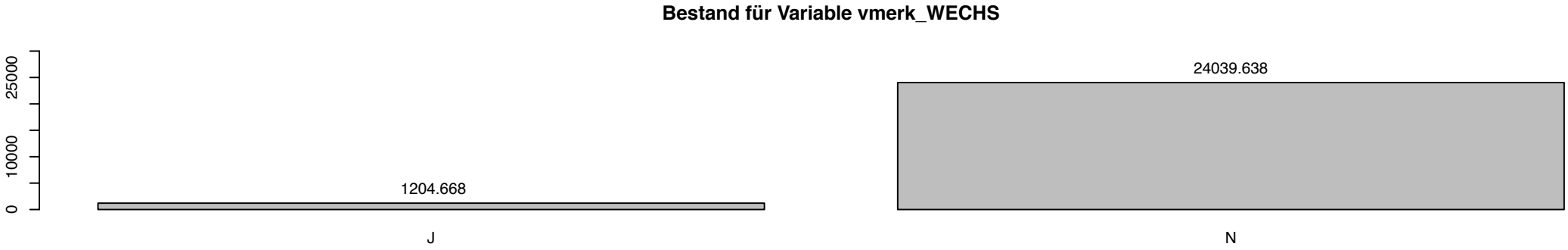

Schadenfrequenz für Variable VNnatio
0.1378228
0.1569716
0.00
0.10
Austria
Foreign

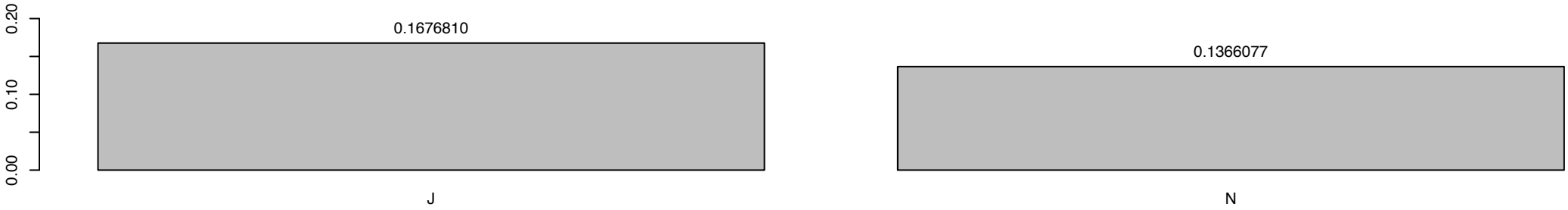

Durchschnittliche Schadenhöhe für Variable VNnatio
774.2952
580.9678
0
200
600
Austria
Foreign

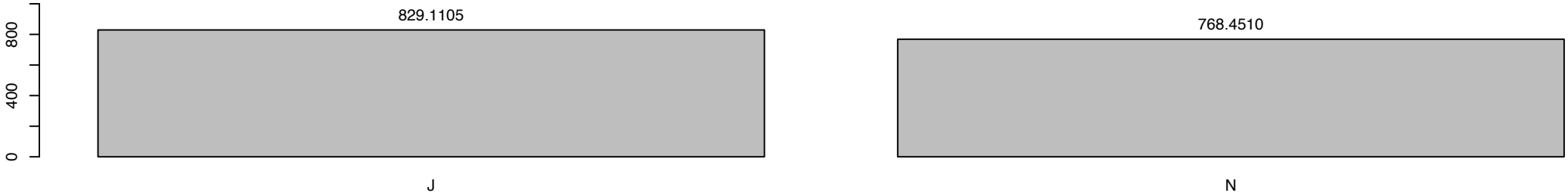

Bestand für Variable VNnatjur
0
10000
25000
370.3836
24885.2493
J
N

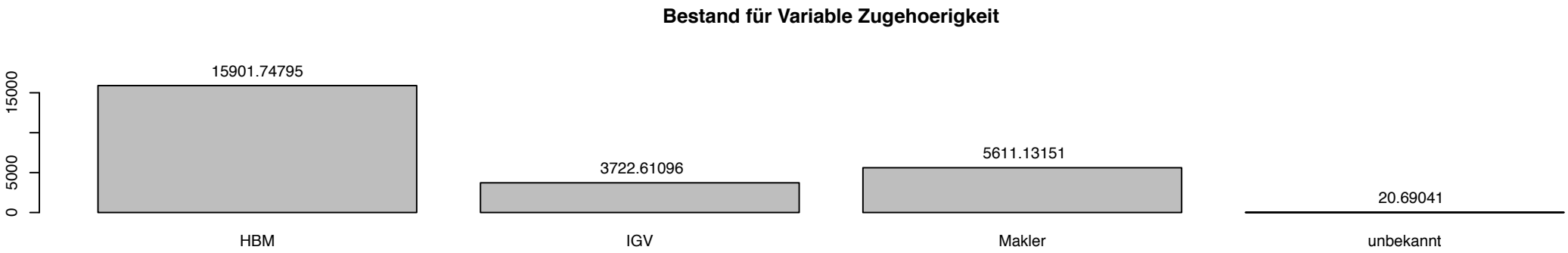

Schadenfrequenz für Variable VNnatjur
0.00
0.10
0.20
0.1808936
0.1373906
J
N

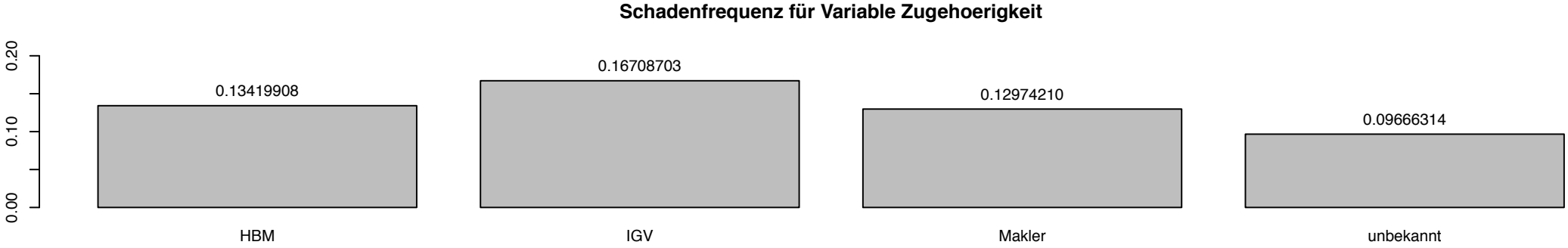

Durchschnittliche Schadenhöhe für Variable VNnatjur
0
500
1000
1500
1267.7452
762.2505
J
N

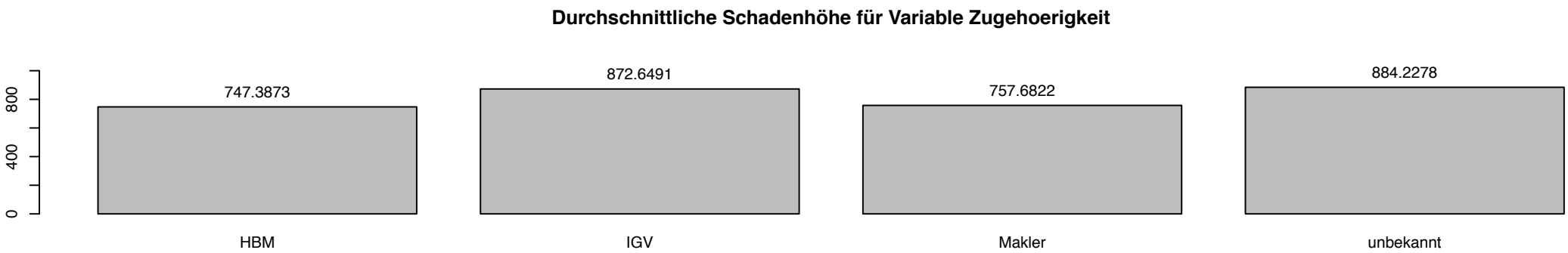

Bestand für Variable vmerk_WECHS
0
10000
25000
1204.668
24039.638
J
N


Schadenfrequenz für Variable vmerk_WECHS
0.00
0.10
0.20
0.1676810
0.1366077
J
N


Durchschnittliche Schadenhöhe für Variable vmerk_WECHS
0
400
800
829.1105
768.4510
J
N


Bestand für Variable Zugehoerigkeit
15901.74795
3722.61096
5611.13151
20.69041
0
5000
15000
HBM
IGV
Makler
unbekannt


Schadenfrequenz für Variable Zugehoerigkeit
0.13419908
0.16708703
0.12974210
0.09666314
0.00
0.10
0.20
HBM
IGV
Makler
unbekannt


Durchschnittliche Schadenhöhe für Variable Zugehoerigkeit
747.3873
872.6491
757.6822
884.2278
0
400
800
HBM
IGV
Makler
unbekannt

Literatur

[1] Shayle R. Searle Charles E. McCulloch and John M. Neuhaus. *Generalized, Linear, and Mixed Models*. John Wiley and Sons, Inc., Hoboken, New Jersey, 2008. Second Edition.

[2] Michel Denuit, Xavier Maréchal, Sandra Pitrebois, and Jean-François Walhin. *Actuarial modelling of claim counts: Risk classification, credibility and bonus-malus systems*. John Wiley & Sons, 2007.

[3] PK Dunn. tweedie: Tweedie exponential family models. *R package version, URL https://cran.r-project.org/web/packages/tweedie/tweedie.pdf*, 2016.

[4] R. Gilchrist and D. Drinkwater. Fitting Tweedie models to data with probability of zero responses. *Proceedings of the 14th International Workshop on Statistical Modelling*, pages 207–214, 1999.

[5] Thomas Lumley Gregory R. Warnes, Ben Bolker and Randall C Johnson. Various R programming tools for model fitting. *R package version, URL https://cran.r-project.org/web/packages/gmodels/gmodels.pdf*, 2015.

[6] Bent Jorgensen. *The theory of dispersion models*, volume 76 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London, 1997.

[7] Rob Kaas, Marc Goovaerts, Jan Dhaene, and Michel Denuit. *Modern actuarial risk theory: using R*. Springer-Verlag, Berlin Heidelberg, 2008. Volume 128.

[8] Hyunseung Kang. Lecture on model selection, stat 431. *University of Pennsylvania, URL http://pages.stat.wisc.edu/ hyunseung/stat431/ModelSelection.pdf*, pages 1–9, 2012.

[9] Bernard Lindgren. *Statistical theory*. CRC Press, 1993. Volume 22.

[10] P. McCullagh and J. A. Nelder. *Generalized linear models*. Monographs on Statistics and Applied Probability. Chapman & Hall, London, 1989. Second Edition [of MR0727836].

[11] David Mihaela. Auto insurance premium calculation using generalized linear models. *Procedia Economics and Finance*, 20(1):147–156, 2015.

[12] Esbjörn Ohlsson and Björn Johansson. *Non-Life Insurance Pricing with Generalized Linear Models*. Springer-Verlag, Berlin Heidelberg, 2010.

[13] Oscar Alberto Quijano Xacur and José Garrido. Generalised linear models for aggregate claims: to Tweedie or not? *Eur. Actuar. J.*, 5(1):181–202, 2015.

[14] Maria Heep-Altiner und Monika Klemmstein. *Versicherungsmathematische Anwendungen in der Praxis mit Schwerpunkt Kraftfahrt und Allgemeine Haftpflicht*. Verlag Versicherungswirtschaft, Karlsruhe, 2001.

[15] Achim Zeileis and Christian Kleiber. AER: applied econometrics with R. *R package version 0.9-0, URL http://CRAN. R-project. org/package= AER*, (1), 2017.