LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Schadenversicherungsmathematik

Teil 2: Tarifkalkulation

Dr. Ulrich Riegel

Mathematisches Institut
Ludwig-Maximilians-Universität München

Wintersemester 2023/24

Inhalt

Problemstellung

Ziel: Schätzung von $\mathsf{E}(R)$ für jedes Einzelrisiko R (da der erwartete Schaden die Basis für die Prämie ist).

Vorgehen: Bilden möglichst großer Gruppen von jeweils möglichst ähnlichen Risiken. Anwendung des Gesetzes der großen Zahlen.

Frage: Wann sind zwei Risiken ähnlich/gleich?

Annahme: Bei gleichen bzw. ähnlichen Ausprägungen ihrer Risikomerkmale.

Definition: Unter einem *Risikomerkmal* versteht man eine von außen a priori feststellbare Eigenschaft von Risiken, die mit dem künftigen Schadenverlauf korreliert ist.

Risikomerkmale

Beispiele für Risikomerkmale:

- Lebensalter
- Raucher/Nichtraucher
- Jahresfahrleistung in KH
- Anfänger/Nicht-Anfänger
- Bauart in Gebäude-Feuer
- Betriebsart
- Geographische Lage in Sturm
- Bauform

Relevante Risikomerkmale sind aus Intuition und Erfahrung weitgehend bekannt.

Inhalt

Problemstellung

Differenzierung des „Unverbindlichen Risikoprämientarif Kraftfahrzeug-Haftpflichtversicherung“ des GDV:

Risikomerkmal	Anzahl Klassen
Typklasse	16
Regionalklasse	12
Schadenfreiheits-Klasse	39
Kilometerleistung	8
Tarifgruppen	3
Wohneigentum	2
Nutzerkreis	2
Differenziertes Nutzeralter	16
Fahrzeugalter beim Erwerb	12

$\Rightarrow 16 \cdot 12 \cdot 39 \cdot 8 \cdot 3 \cdot 2 \cdot 2 \cdot 16 \cdot 12 = 138.018.816$ mögliche Tarifzellen bzw. Risikogruppen (und viele Gesellschaften unterteilen noch weiter)!

Kreuzklassifikation

Folgerung: Viele Zellen sind schwach (oder gar nicht) besetzt und haben einen instabilen (oder gar keinen) Schadenverlauf!

Idee: Heranziehen der Schadenerfahrung benachbarter Zellen mittels *Marginalfaktoren* oder *Marginalsummanden*.

Im Folgenden betrachten wir nur zwei Tarifmerkmale:

	y_1	$\cdots$	y_k			$\cdots$	y_K
x_1			$\vdots$				
$\vdots$			$\vdots$				
x_i		$\cdots$	$\mathsf{E}(Z_{ik}) = x_i y_k$	oder	$\mathsf{E}(Z_{ik}) = x_i + y_k$	$\cdots$	
$\vdots$			$\vdots$				
x_I			$\vdots$				

Kreuzklassifikation

Multiplikativer Fall: Parameter x_i und y_k sind nur bis auf einen Faktor bestimmt, denn $x_i y_k = (x_i c)\frac{y_k}{c}$.

Additiver Fall: Parameter x_i und y_k sind nur bis auf einen Summanden bestimmt, denn $x_i + y_k = (x_i + c) + (y_k - c)$.

Folgerung: Reduktion der Parameterzahl von $I \cdot K$ auf $1 + (I - 1) + (K - 1)$.

$\Rightarrow$ Im Fall des KH-Risikoprämien-Tarifs des GDV: Reduktion von 138.018.816 Parametern auf $102 = 1 + 15 + 11 + 38 + 7 + 2 + 1 + 1 + 15 + 11)$ Parameter.

Dieses Vorgehen erlaubt eine einfache Darstellung eines Tarifes mit intuitiver Tariforganik.

Beispiel: Ehemaliger Schwedischer Autotarif

Jahresfahrleistung (km)	
0-10.000	0,8
10.001-15.000	0,9
15.001-20.000	1,0
20.001-25.000	1,1
25.001-∞	1,2

Schadenfreiheit	
keine	1,0
1 Jahr	0,8
2 Jahre	0,7
3 Jahre	0,6
4 Jahre	0,5
5 Jahre	0,4
6 Jahre	0,2

Fahrzeugtyp
10 Modellklassen mit
Faktoren zwischen 1,0 und 2,0

Fahrgebiet
7 Gebiete mit Faktoren
zwischen 0,81 und 1,14

Aufgabenstellung

Schätze x_i, y_k aus Realisierungen $Z_{ik} = S_{ik}/v_{ik}$ mit $1 \leq i \leq I$ und $1 \leq k \leq K$ wobei die Volumina v_{ik} (in KH Jahreseinheiten) bekannt sind.

Falls Realisierungen aus mehreren Jahren verwendet werden sollen, kann man das Jahr als zusätzliches Risikomerkmal betrachten.

Im Folgenden betrachten wir verschiedene Verfahren zur Schätzung von x_i und y_k

Tarifierung mittels Marginaldurchschnitten

Mit $\overline{Z} = S_{++}/v_{++}$ sei

$$\widehat{x}_i := \frac{S_{i+}}{v_{i+}} \quad \text{und} \quad \widehat{y}_k := \frac{S_{+k}}{v_{+k}} \cdot \frac{1}{\overline{Z}},$$

wobei $S_{i+} := \sum_{k=1}^{K} S_{ik}$ und S_{+k} , v_{i+}, v_{+k}, S_{++}, v_{++} analog definiert sind.

Dann ist

$$\widehat{\mathrm{E}}(Z_{ik}) = \widehat{x}_i \cdot \widehat{y}_k = \overline{Z} \cdot \frac{S_{i+}/v_{i+}}{\overline{Z}} \cdot \frac{S_{+k}/v_{+k}}{\overline{Z}}.$$

Beachte: Falls das Volumen dimensionsfrei ist, so hat $\widehat{x}_i$ die Dimension EUR und $\widehat{y}_k$ ist dimensionsfrei.

Beispiel: 2-fach klassifizierte KH-Statistik

	privat	gewerblich	gesamt
leicht	1.800.000	69.000	1.869.000
	9.000	300	9.300
	200,0	230,0	201,0
mittel	1.320.000	177.100	1.497.100
	6.000	700	6.700
	220,0	253,0	223,4
schwer	720.000	276.000	996.000
	3.000	1.000	4.000
	240,0	276,0	249,0
gesamt	3.840.000	522.100	4.362.100
	18.000	2.000	20.000
	213,3	261,1	218,1

Jede Zelle enthält den Gesamtschaden S_{ik}, das Volumen v_{ik} und den Durchschnittsschaden $Z_{ik} = S_{ik}/v_{ik}$.

Beispiel: 2-fach klassifizierte KH-Statistik

Im Zahlenbeispiel erhalten wir mit den Maginaldurchschnitten die Schätzer:

$\widehat{E}(Z_{ik})$			$\widehat{x}_i$
	196,6	240,5	201,0
	218,6	267,5	223,4
	243,6	298,0	249,0
$\widehat{y}_k$	$\frac{213,3}{218,1}$	$\frac{261,1}{218,1}$	

Diese Lösung ist schlecht, da es hier eine exakte Lösung gibt:

$\widehat{E}(Z_{ik})$			$\widehat{x}_i$
	200,0	230,0	200,0
	220,0	253,0	220,0
	240,0	276,0	240,0
$\widehat{y}_k$	1,00	1,15	

Das Verfahren von Bailey und Simon

Bei dieser Methode bestimmt man $\widehat{x}_i$ und $\widehat{y}_k$ durch Minimierung von

$$\sum_{i=1}^{I}\sum_{k=1}^{K}\frac{(S_{ik}-v_{ik}x_iy_k)^2}{v_{ik}x_iy_k}=\sum_{i=1}^{I}\sum_{k=1}^{K}v_{ik}\frac{(Z_{ik}-x_iy_k)^2}{x_iy_k}$$

analog zum Minimum-χ^2-Schätzer. Nullsetzen der partiellen Ableitungen liefert:

$$\frac{\partial}{\partial x_i}\sum_{i=1}^{I}\sum_{k=1}^{K}v_{ik}\frac{(Z_{ik}-x_iy_k)^2}{x_iy_k}=0 \quad\Rightarrow\quad \widehat{x}_i=\sqrt{\frac{\sum_k v_{ik}Z_{ik}^2/\widehat{y}_k}{\sum_k v_{ik}\widehat{y}_k}},\quad 1\le i\le I,$$

$$\frac{\partial}{\partial y_k}\sum_{i=1}^{I}\sum_{k=1}^{K}v_{ik}\frac{(Z_{ik}-x_iy_k)^2}{x_iy_k}=0 \quad\Rightarrow\quad \widehat{y}_k=\sqrt{\frac{\sum_i v_{ik}Z_{ik}^2/\widehat{x}_i}{\sum_i v_{ik}\widehat{x}_i}},\quad 1\le k\le K.$$

Das Verfahren von Bailey und Simon

Rekursion:

Mit den Startwerten $\widehat{y}_k^{(0)} := 1$ iteriert man

$$\widehat{x}_i^{(\nu+1)} := \sqrt{\frac{\sum_k v_{ik} Z_{ik}^2 / \widehat{y}_k^{(\nu)}}{\sum_k v_{ik} \widehat{y}_k^{(\nu)}}}, \quad 1 \le i \le I,$$

$$\widehat{y}_k^{(\nu+1)} := \sqrt{\frac{\sum_i v_{ik} Z_{ik}^2 / \widehat{x}_i^{(\nu+1)}}{\sum_i v_{ik} \widehat{x}_i^{(\nu+1)}}}, \quad 1 \le k \le K.$$

Das Verfahren von Bailey und Simon

Bemerkungen:

- Konvergiert rasch!
- Ergibt im Beispiel von Folie 12 die exakte Lösung!
- Wurde 1962–1994 in (West-)Deutschland eingesetzt.
- Ausreißerempfindlich!
- Als gewichteter Kleinste-Quadrate-Schätzer akzeptabel, *falls* das Gewicht $1/v_{ik}x_iy_k$ indirekt proportional zu $\mathrm{Var}(S_{ik})$ ist, d.h. $\mathrm{Var}(S_{ik}) = c \cdot v_{ik}x_iy_k = c \cdot \mathrm{E}(S_{ik})$.

Das Marginalsummenverfahren

Bei diesem Verfahren werden $\widehat{x}_i$ und $\widehat{y}_k$ so geschätzt, dass

$$\sum_{k=1}^{K} v_{ik}\widehat{x}_i\widehat{y}_k = \sum_{k=1}^{K} S_{ik}, \quad 1 \le i \le I,$$

$$\sum_{i=1}^{I} v_{ik}\widehat{x}_i\widehat{y}_k = \sum_{i=1}^{I} S_{ik}, \quad 1 \le k \le K.$$

Beachte, dass die Schätzer nach Konstruktion die Schäden nicht überschätzen.

Diese Bedingungen führen zu

$$\widehat{x}_i = \frac{\sum_{k=1}^{K} S_{ik}}{\sum_{k=1}^{K} v_{ik}\widehat{y}_k} \quad \text{und} \quad \widehat{y}_k = \frac{\sum_{i=1}^{I} S_{ik}}{\sum_{i=1}^{I} v_{ik}\widehat{x}_i}.$$

Das Marginalsummenverfahren

Rekursion: Mit den Startwerten $\widehat{y}_k^{(0)} := 1$ iteriert man

$$\widehat{x}_i^{(\nu+1)} := \frac{\sum_{k=1}^{K} S_{ik}}{\sum_{k=1}^{K} v_{ik}\widehat{y}_k^{(\nu)}} \quad \text{und}$$

$$\widehat{y}_k^{(\nu+1)} := \frac{\sum_{i=1}^{I} S_{ik}}{\sum_{i=1}^{I} v_{ik}\widehat{x}_i^{(\nu+1)}}.$$

Bemerkungen:

- Konvergiert rasch
- Ergibt im Beispiel von Folie 12 die exakte Lösung
- Weniger ausreißerempfindlich als die Methode von Bailey-Simon
- Wird seit 1995 in Deutschland eingesetzt

Kritik an den vorgestellten Verfahren

Bemerkung: Die bisher vorgestellten Verfahren sind nicht stochastisch:

- Keine Angabe zur Genauigkeit der Schätzer
  - Sind die Parameterschätzer benachbarter Zellen so verschieden, dass verschiedene Prämien gerechtfertigt sind?
  - Muss eine etwas abweichende Prämie vom letzten Jahr angepasst werden?
- Keine Angabe zur Anpassungsgüte
  - Modellüberprüfung der Kreuzklassifikations-Annahme
  - Entscheidung $x_i \cdot y_k$ gegen $x_i + y_k$
- Keine Angabe zur Prognosegenauigkeit
  - Konfidenzintervall für den Schadenbedarf des nächsten Jahres

Ein stochastisches Ausgleichsverfahren für die Anzahl Schäden

Sei jetzt speziell S_{ik} die Anzahl Schäden in der Zelle (i,k). Dann heißt $Z_{ik} = S_{ik}/v_{ik}$ die *Schadenhäufigkeit*.

Tarifierung auf Basis der Schadenzahl ist besonders relevant in KH, da dort die *Schadendurchschnitt* (d.h. Schadenaufwand/Schadenzahl) von Zelle zu Zelle nur wenig variiert.

Beachte:

$$\text{Schadenbedarf} = \frac{\text{Gesamtschaden}}{\#\text{ Jahreseinheiten}} = \underbrace{\frac{\text{Gesamtschaden}}{\#\text{ Schäden}}}_{\text{Schadendurchschnitt}} \cdot \underbrace{\frac{\#\text{ Schäden}}{\#\text{ Jahreseinheiten}}}_{\text{Schadenhäufigkeit}}$$

Poisson-Verteilung

Definition:

Eine Zufallsgröße $N\colon \Omega \to \mathbb{N}_0$ heißt *Poisson-verteilt* mit Parameter $\lambda \geq 0$, wenn

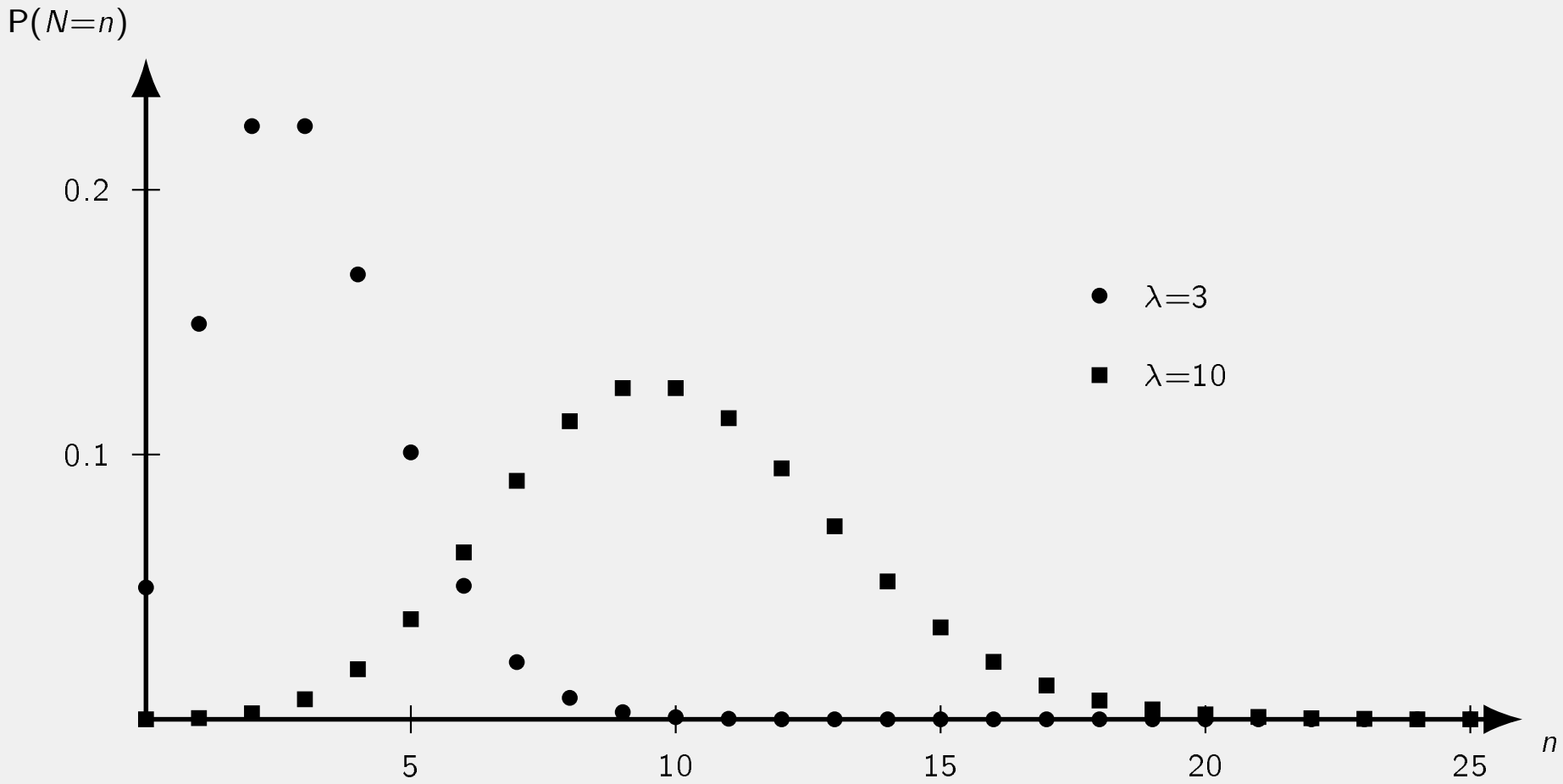
$$\mathrm{P}(N = n) = e^{-\lambda}\frac{\lambda^n}{n!}$$

für $n \in \mathbb{N}_0$. Wir schreiben dann $N \sim \mathrm{Poi}(\lambda)$.

Lemma: *Für* $N \sim \mathrm{Poi}(\lambda)$ *gilt*

$$\mathrm{E}(N) = \lambda = \mathrm{Var}(N).$$

Poisson-Verteilung



Poisson-Prozess

Bemerkung:

Entstehen die Schäden mit einem sog. *Poisson-Prozesses*, d.h.

(1) die Schadenanzahlen in disjunkten Zeitintervallen sind unabhängig,
(2) es treten nie zwei oder mehr Schäden im exakt gleichen Zeitpunkt ein und
(3) die erwartete Anzahl Schäden in einem Zeitintervall hängt nur von dessen Länge ab,

so ist die Anzahl der Schäden in einem festen Zeitraum automatisch Poisson-verteilt.

Die Poisson-Verteilung ist also eine sehr plausible Schadenzahlverteilung.

Poisson-Modell

Wir verwenden folgende Modell-Annahmen:

Poisson-Modell:

Für $1 \leq i \leq I$ und $1 \leq k \leq K$ nehmen wir an, dass die S_{ik} unabhängig sind mit

$$\text{Poi}(v_{ik} x_i y_k).$$

Es ist dann $\mathsf{E}(Z_{ik}) = x_i y_k$. Das Modell muss aber für $S_{ik} = v_{ik} Z_{ik}$ formuliert werden, da im Allgemeinen $Z_{ik} \notin \mathbb{N}_0$.

Poisson-Modell

Die Likelihoodfunktion ist

$$L(x_1, \ldots, x_I, y_1 \ldots, y_K) = \prod_{i=1}^{I} \prod_{k=1}^{K} \left(e^{-v_{ik} x_i y_k} \frac{(v_{ik} x_i y_k)^{S_{ik}}}{S_{ik}!} \right).$$

Hieraus erhalten wir die log-Likelihoodfunktion

$$\ln L = \sum_{i=1}^{I} \sum_{k=1}^{K} (S_{ik} \ln (v_{ik} x_i y_k) - v_{ik} x_i y_k - \ln(S_{ik}!)).$$

Es gilt

$$0 = \frac{\partial \ln L}{\partial x_i} = \sum_{k=1}^{K} \left(\frac{S_{ik}}{x_i} - v_{ik} y_k \right) = \frac{1}{x_i} \sum_{k=1}^{K} (S_{ik} - v_{ik} x_i y_k)$$

$$\iff \quad \sum_{k=1}^{K} v_{ik} x_i y_k = \sum_{k=1}^{K} S_{ik}, \quad 1 \leq i \leq I.$$

Poisson-Modell

Analog

$$0 = \frac{\partial \ln L}{\partial y_k} \quad \Longleftrightarrow \quad \sum_{i=1}^{I} v_{ik} x_i y_k = \sum_{i=1}^{I} S_{ik}, \quad 1 \leq k \leq K.$$

Folgerung:

Für das Poisson-Modell liefert das Marginalsummenverfahren also den ML-Schätzer!

Beispiel: KH-Statistik 2013 – Schadenzahl

Als Beispiel betrachten wir die Schadenzahlen aus der Kraftfahrt-Haftpflicht-Statistik von 2013 (nach Rücksprache mit dem GDV leicht verändert).

Wir beschränken uns auf PKW in der besten Schadenfreiheitsklasse und wenden das Marginalsummenverfahren an um nach den Merkmalen *Regionalklasse* und *Fahrleistung* zu differenzieren.

Als Anpassungstest verwenden wir folgenden χ^2-Test

$$\sum_{i=1}^{I}\sum_{k=1}^{K}\frac{(S_{ik}-v_{ik}\widehat{x}_i\widehat{y}_k)^2}{v_{ik}\widehat{x}_i\widehat{y}_k} \overset{\text{asymp.}}{\sim} \chi^2_{IK-(I+K-1)}.$$

(Heuristik: Wären $S_{ik}\sim\mathcal{N}(v_{ik}x_iy_k, v_{ik}x_iy_k)$, dann $\sum_{i,k}\frac{(S_{ik}-v_{ik}x_iy_k)^2}{v_{ik}x_iy_k} \overset{\text{exakt}}{\sim} \chi^2_{IK}$.)

Beispiel: KH-Statistik 2013 – Schadenzahl

Jahreseinheiten v_{ik}

Regionalklasse \ Fahrleistung	0 -	7 -	10 -	13 -	16 -	21 -	26 -	31 -	$\sum$
1	56.115	54.271	34.827	20.839	16.293	4.220	2.299	1.116	189.981
2	88.917	83.849	54.873	32.825	24.696	6.392	3.346	1.640	296.537
3	78.547	78.371	55.437	33.331	24.167	6.053	3.115	1.485	280.506
4	57.183	58.366	42.122	25.678	18.239	4.646	2.329	1.104	209.667
5	71.114	73.248	54.320	33.876	24.301	6.148	3.035	1.614	267.656
6	68.684	69.543	50.373	31.824	22.715	5.846	3.034	1.437	253.457
7	91.589	88.399	64.416	40.337	28.194	7.254	3.857	1.813	325.857
8	42.382	41.478	28.905	18.002	12.738	3.494	1.763	861	149.623
9	70.663	71.200	50.719	30.397	22.172	5.759	2.918	1.428	255.256
10	45.475	44.641	30.869	18.454	13.336	3.506	1.829	914	159.025
11	22.642	21.398	14.431	8.398	6.080	1.655	831	379	75.814
12	49.095	45.551	30.981	17.334	10.473	2.219	1.148	586	157.387
$\sum$	742.405	730.315	512.275	311.294	223.403	57.192	29.503	14.377	2.620.764

Beispiel: KH-Statistik 2013 – Schadenzahl

Schadenanzahl S_{ik}

Regionalklasse \ Fahrleistung	0 -	7 -	10 -	13 -	16 -	21 -	26 -	31 -	∑
1	2.265	2.393	1.521	1.024	759	226	106	65	8.359
2	3.685	3.698	2.593	1.577	1.286	322	180	106	13.447
3	3.278	3.387	2.555	1.584	1.277	373	174	97	12.725
4	2.291	2.675	1.915	1.213	952	250	145	91	9.532
5	2.970	3.231	2.558	1.619	1.297	338	197	92	12.302
6	2.830	3.050	2.367	1.550	1.163	355	186	88	11.589
7	3.941	4.084	3.130	2.042	1.495	417	222	103	15.434
8	1.832	1.936	1.464	933	688	190	102	55	7.200
9	3.098	3.278	2.562	1.576	1.203	333	187	86	12.323
10	1.964	2.021	1.553	938	759	220	110	76	7.641
11	1.049	1.057	722	446	328	90	42	36	3.770
12	2.387	2.492	1.705	1.038	699	168	84	39	8.612
∑	31.590	33.302	24.645	15.540	11.906	3.282	1.735	934	122.934

Beispiel: KH-Statistik 2013 – Schadenzahl

Schadenhäufigkeit Z_{ik}

Regionalklasse \ Fahrleistung	0 -	7 -	10 -	13 -	16 -	21 -	26 -	31 -	∅
1	4,04%	4,41%	4,37%	4,91%	4,66%	5,35%	4,61%	5,83%	4,40%
2	4,14%	4,41%	4,73%	4,80%	5,21%	5,04%	5,38%	6,46%	4,53%
3	4,17%	4,32%	4,61%	4,75%	5,28%	6,16%	5,59%	6,53%	4,54%
4	4,01%	4,58%	4,55%	4,72%	5,22%	5,38%	6,23%	8,24%	4,55%
5	4,18%	4,41%	4,71%	4,78%	5,34%	5,50%	6,49%	5,70%	4,60%
6	4,12%	4,39%	4,70%	4,87%	5,12%	6,07%	6,13%	6,12%	4,57%
7	4,30%	4,62%	4,86%	5,06%	5,30%	5,75%	5,76%	5,68%	4,74%
8	4,32%	4,67%	5,06%	5,18%	5,40%	5,44%	5,79%	6,39%	4,81%
9	4,38%	4,60%	5,05%	5,18%	5,43%	5,78%	6,41%	6,02%	4,83%
10	4,32%	4,53%	5,03%	5,08%	5,69%	6,27%	6,01%	8,31%	4,80%
11	4,63%	4,94%	5,00%	5,31%	5,40%	5,44%	5,06%	9,49%	4,97%
12	4,86%	5,47%	5,50%	5,99%	6,67%	7,57%	7,32%	6,65%	5,47%
∅	4,26%	4,56%	4,81%	4,99%	5,33%	5,74%	5,88%	6,50%	4,69%

Beispiel: KH-Statistik 2013 – Schadenzahl

ausgeglichene Schadenhäufigkeit $\hat{x}_i\hat{y}_k$ (Marginalsummenverfahren)

Regionalklasse \ Fahrleistung	0 -	7 -	10 -	13 -	16 -	21 -	26 -	31 -	$\hat{x}_i$
1	3,99%	4,28%	4,52%	4,69%	5,02%	5,41%	5,55%	6,13%	4,41%
2	4,12%	4,42%	4,66%	4,84%	5,18%	5,58%	5,72%	6,32%	4,55%
3	4,11%	4,41%	4,65%	4,83%	5,17%	5,57%	5,71%	6,30%	4,53%
4	4,11%	4,41%	4,65%	4,83%	5,17%	5,57%	5,71%	6,31%	4,54%
5	4,15%	4,45%	4,69%	4,88%	5,22%	5,62%	5,76%	6,37%	4,58%
6	4,13%	4,43%	4,67%	4,85%	5,19%	5,60%	5,74%	6,34%	4,56%
7	4,29%	4,60%	4,85%	5,04%	5,39%	5,81%	5,95%	6,58%	4,73%
8	4,36%	4,67%	4,93%	5,12%	5,48%	5,91%	6,05%	6,68%	4,81%
9	4,37%	4,69%	4,94%	5,13%	5,49%	5,92%	6,07%	6,70%	4,82%
10	4,36%	4,67%	4,93%	5,12%	5,48%	5,90%	6,05%	6,68%	4,81%
11	4,52%	4,85%	5,11%	5,31%	5,68%	6,13%	6,28%	6,93%	4,99%
12	5,01%	5,37%	5,66%	5,88%	6,29%	6,78%	6,95%	7,68%	5,52%
$\hat{y}_k$	90,61%	97,18%	102,53%	106,50%	113,93%	122,81%	125,88%	139,02%	

Beispiel: KH-Statistik 2013 – Schadenzahl

Residuenquadrate $(S_{ik}-v_{ik}\hat{x}_i\hat{y}_k)^2/(v_{ik}\hat{x}_i\hat{y}_k)$

Regionalklasse \ Fahrleistung	0 -	7 -	10 -	13 -	16 -	21 -	26 -	31 -	∑
1	0,26	2,04	1,76	2,16	4,26	0,03	3,64	0,17	14,31
2	0,13	0,01	0,47	0,10	0,04	3,42	0,69	0,05	4,91
3	0,80	1,28	0,20	0,41	0,65	3,82	0,08	0,12	7,36
4	1,56	3,94	1,04	0,64	0,08	0,31	1,07	6,53	15,17
5	0,13	0,24	0,02	0,65	0,68	0,17	2,79	1,12	5,81
6	0,02	0,31	0,07	0,02	0,24	2,35	0,81	0,10	3,91
7	0,06	0,11	0,01	0,05	0,39	0,05	0,25	2,20	3,12
8	0,11	0,00	1,06	0,14	0,14	1,30	0,21	0,12	3,07
9	0,04	1,01	1,19	0,15	0,18	0,19	0,56	0,99	4,30
10	0,14	1,98	0,66	0,05	1,13	0,82	0,00	3,63	8,41
11	0,65	0,38	0,35	0,00	0,88	1,27	1,98	3,58	9,08
12	2,02	0,89	1,42	0,32	2,41	2,03	0,22	0,80	10,10
∑	5,92	12,18	8,25	4,68	11,08	15,74	12,29	19,41	89,55

Beispiel: KH-Statistik 2013 – Schadenzahl

Für unser Beispiel der KH-Statistik 2013 liefert die Teststatistik für das Marginalsummenverfahren einen Wert von $89{,}6 < \chi^2_{77;95\%} = 98{,}5$. Man muss also bei einem Konfidenzniveau von 95% nicht ablehnen.

In Fällen, in denen der Test ablehnt, passen eine oder mehrere der Modellannahmen nicht:

- Poisson-Verteilung
- Multiple Kreuzklassifikation
- Unabhängigkeit der S_{ik}

Die Schätzer der Methode von Bailey und Simon minimieren die Teststatistik. Der Wert liegt mit 89,4 leicht unter dem Wert für das Marginalsummenverfahren.

Das auf der Gamma-Verteilung beruhende Ausgleichsverfahren

Wir verwenden folgende Modell-Annahmen:

Gamma-Modell: Die Schadensätze/Schadenbedarfe Z_{ik} sind unabhängig mit

$$Z_{ik} = \frac{S_{ik}}{v_{ik}} \sim \Gamma(x_i y_k, v_{ik}\alpha)$$

für $1 \leq i \leq I$, $1 \leq k \leq K$ (α für alle i, j gleich).

Das Modell entspricht unseren Überlegungen zum Individuelle Modell, erweitert um die Annahme der Kreuzklassifikation $\mathsf{E}(Z_{ik}) = x_i y_k$.

Die Annahme, dass α in allen Zellen gleich ist, d.h. dass die individuellen Risiken in allen Zellen den gleichen Formparameter haben, sollte überprüft werden.

ML-Schätzung der x_i, y_k

Maximum-Likelihood-Schätzung der x_i, y_k: Es gilt

$$L(\{x_i, y_k\}, \alpha) = \prod_{i=1}^{I} \prod_{k=1}^{K} \left(\left(\frac{Z_{ik} v_{ik} \alpha}{x_i y_k} \right)^{v_{ik}\alpha} \exp\left(-\frac{Z_{ik} v_{ik} \alpha}{x_i y_k} \right) (Z_{ik} \Gamma(v_{ik}\alpha))^{-1} \right)$$

und somit

$$\ln L(\{x_i, y_k\}, \alpha) = \sum_{i=1}^{I} \sum_{k=1}^{K} \left(v_{ik}\alpha \ln\left(\frac{S_{ik}\alpha}{x_i y_k} \right) - \frac{S_{ik}\alpha}{x_i y_k} - \ln(Z_{ik}\Gamma(v_{ik}\alpha)) \right).$$

Nullsetzen der partielle Ableitung liefert

$$0 = \frac{\partial \ln L}{\partial x_i} = \sum_{k=1}^{K} \left(-\frac{v_{ik}\alpha}{x_i} + \frac{S_{ik}\alpha}{x_i^2 y_k} \right) = \frac{\alpha}{x_i^2} \sum_{k=1}^{K} \left(\frac{S_{ik}}{y_k} - v_{ik} x_i \right).$$

ML-Schätzung der x_i, y_k

Somit erhalten wir

$$\widehat{x}_i = \frac{1}{v_{i+}} \left(\sum_{k=1}^{K} \frac{S_{ik}}{\widehat{y}_k} \right) = \sum_{k=1}^{K} \frac{v_{ik}}{v_{i+}} \frac{Z_{ik}}{\widehat{y}_k}, \quad 1 \leq i \leq I,$$

und analog aufgrund der Symmetrie

$$\widehat{y}_k = \sum_{i=1}^{I} \frac{v_{ik}}{v_{+k}} \frac{Z_{ik}}{\widehat{x}_i}, \quad 1 \leq k \leq K.$$

ML-Schätzung der x_i, y_k

Bemerkungen:

- Beide Gleichungen können gemeinsam alternierend iterativ gelöst werden (wie beim Marginalsummenverfahren). Zum Start werden alle $\widehat{y}_k = 1$ gesetzt.
- Konvergiert rasch (fünf Iterationen).
- Liefert im Beispiel von Folie 12 die exakte Lösung.
- Der Schätzer $\widehat{\alpha}$ wird nicht benötigt.
- Die Höhe der Likelihood-Funktion L ermöglicht die Entscheidung zwischen dem additiven und multiplikativen Modell (aber hierzu wird $\widehat{\alpha}$ benötigt).

Beispiel: KH-Statistik 2013 – Schadenaufwand

Als Beispiel betrachten wir den Schadenaufwand aus der Kraftfahrt-Haftpflicht-Statistik von 2013 (nach Rücksprache mit dem GDV leicht verändert).

Wir beschränken uns wieder auf PKW in der besten Schadenfreiheitsklasse und wenden die ML-Schätzer des Gamma-Modells an um nach den Merkmalen *Regionalklasse* und *Fahrleistung* zu differenzieren.

Beispiel: KH-Statistik 2013 – Schadenaufwand

Jahreseinheiten v_{ik}

Regionalklasse \ Fahrleistung	0 -	7 -	10 -	13 -	16 -	21 -	26 -	31 -	$\sum$
1	56.115	54.271	34.827	20.839	16.293	4.220	2.299	1.116	189.981
2	88.917	83.849	54.873	32.825	24.696	6.392	3.346	1.640	296.537
3	78.547	78.371	55.437	33.331	24.167	6.053	3.115	1.485	280.506
4	57.183	58.366	42.122	25.678	18.239	4.646	2.329	1.104	209.667
5	71.114	73.248	54.320	33.876	24.301	6.148	3.035	1.614	267.656
6	68.684	69.543	50.373	31.824	22.715	5.846	3.034	1.437	253.457
7	91.589	88.399	64.416	40.337	28.194	7.254	3.857	1.813	325.857
8	42.382	41.478	28.905	18.002	12.738	3.494	1.763	861	149.623
9	70.663	71.200	50.719	30.397	22.172	5.759	2.918	1.428	255.256
10	45.475	44.641	30.869	18.454	13.336	3.506	1.829	914	159.025
11	22.642	21.398	14.431	8.398	6.080	1.655	831	379	75.814
12	49.095	45.551	30.981	17.334	10.473	2.219	1.148	586	157.387
$\sum$	742.405	730.315	512.275	311.294	223.403	57.192	29.503	14.377	2.620.764

Beispiel: KH-Statistik 2013 – Schadenaufwand

Schadenaufwand S_{ik}

Regionalklasse \ Fahrleistung	0 -	7 -	10 -	13 -	16 -	21 -	26 -	31 -	∑
1	7.023.290	6.218.655	3.983.868	3.161.259	2.161.277	928.689	262.259	150.989	23.890.286
2	9.375.689	10.734.040	7.599.432	3.808.610	3.190.079	834.130	383.924	276.049	36.201.953
3	8.193.664	8.876.313	6.846.592	8.259.134	3.829.077	1.172.159	464.056	240.595	37.881.590
4	7.019.032	7.980.952	5.495.448	4.089.568	2.372.583	813.518	611.486	210.673	28.593.260
5	7.987.158	9.529.338	7.295.884	4.908.361	5.382.239	872.635	886.138	362.372	37.224.125
6	8.975.545	8.665.627	6.721.325	6.167.412	3.194.202	1.089.993	444.365	208.617	35.467.086
7	9.963.360	12.029.761	12.559.903	5.446.003	4.043.389	1.232.733	623.844	574.594	46.473.587
8	5.199.963	4.986.751	4.030.205	3.255.314	1.898.046	453.804	374.692	129.674	20.328.449
9	8.907.966	12.048.585	12.903.667	4.393.695	3.998.992	998.566	530.992	261.454	44.043.917
10	5.768.359	5.335.671	4.205.425	2.443.276	2.330.993	580.119	253.944	236.318	21.154.105
11	2.528.342	2.680.101	2.352.112	1.113.933	1.424.014	336.952	99.165	72.567	10.607.186
12	8.877.132	7.435.399	4.514.678	2.469.732	2.251.598	342.870	231.534	390.453	26.513.396
∑	89.819.500	96.521.193	78.508.539	49.516.297	36.076.489	9.656.168	5.166.399	3.114.355	368.378.940

Beispiel: KH-Statistik 2013 – Schadenaufwand

Schadenbedarf Z_{ik}

Regionalklasse \ Fahrleistung	0 -	7 -	10 -	13 -	16 -	21 -	26 -	31 -	∅
1	125,2	114,6	114,4	151,7	132,6	220,0	114,1	135,3	125,8
2	105,4	128,0	138,5	116,0	129,2	130,5	114,7	168,3	122,1
3	104,3	113,3	123,5	247,8	158,4	193,6	149,0	162,0	135,0
4	122,7	136,7	130,5	159,3	130,1	175,1	262,5	190,8	136,4
5	112,3	130,1	134,3	144,9	221,5	141,9	292,0	224,6	139,1
6	130,7	124,6	133,4	193,8	140,6	186,4	146,5	145,2	139,9
7	108,8	136,1	195,0	135,0	143,4	169,9	161,7	317,0	142,6
8	122,7	120,2	139,4	180,8	149,0	129,9	212,6	150,6	135,9
9	126,1	169,2	254,4	144,5	180,4	173,4	182,0	183,1	172,5
10	126,8	119,5	136,2	132,4	174,8	165,5	138,9	258,5	133,0
11	111,7	125,2	163,0	132,6	234,2	203,7	119,4	191,3	139,9
12	180,8	163,2	145,7	142,5	215,0	154,5	201,6	666,2	168,5
∅	121,0	132,2	153,3	159,1	161,5	168,8	175,1	216,6	140,6

Beispiel: KH-Statistik 2013 – Schadenaufwand

ausgeglichene Schadenbedarfe $\hat{x}_i\hat{y}_k$ (Gamma-Modell)

Regionalklasse \ Fahrleistung	0 -	7 -	10 -	13 -	16 -	21 -	26 -	31 -	$\hat{x}_i$
1	109,6	119,5	137,5	145,5	146,6	154,7	159,4	194,5	1,000
2	106,4	116,0	133,4	141,2	142,3	150,2	154,7	188,8	0,970
3	114,6	125,0	143,8	152,2	153,4	161,9	166,7	203,5	1,046
4	117,7	128,4	147,7	156,3	157,5	166,2	171,2	209,0	1,074
5	118,4	129,1	148,5	157,2	158,4	167,2	172,2	210,1	1,080
6	120,4	131,3	151,0	159,8	161,0	169,9	175,1	213,6	1,098
7	122,0	133,1	153,1	162,0	163,2	172,3	177,4	216,5	1,113
8	117,0	127,6	146,8	155,4	156,5	165,2	170,2	207,7	1,068
9	147,9	161,3	185,6	196,4	197,9	208,9	215,1	262,5	1,350
10	115,2	125,6	144,5	152,9	154,1	162,6	167,5	204,4	1,051
11	120,2	131,1	150,8	159,6	160,8	169,7	174,8	213,3	1,097
12	149,0	162,6	187,0	197,9	199,4	210,5	216,8	264,5	1,360
$\hat{y}_k$	109,6	119,5	137,5	145,5	146,6	154,7	159,4	194,5	

ML-Schätzung von α

Definition: Die Funktion $\Psi := \Gamma'/\Gamma$ heißt *Digamma-Funkton*. Die Funktion Ψ' heißt *Trigamma-Funkton*.

Zum Schätzen von α setzen wir

$$0 = \frac{\partial \ln(L)}{\partial \alpha} = \sum_{i=1}^{I}\sum_{k=1}^{K}\left(v_{ik} \ln\left(\frac{S_{ik}\alpha}{\widehat{x}_i \widehat{y}_k}\right) \underbrace{+ v_{ik} - \frac{S_{ik}}{\widehat{x}_i \widehat{y}_k}}_{\sum_{i,k}=0} - v_{ik} \underbrace{\frac{\Gamma'(v_{ik}\alpha)}{\Gamma(v_{ik}\alpha)}}_{\Psi(v_{ik}\alpha)} \right)$$

$$= \sum_{i=1}^{I}\sum_{k=1}^{K}\left(\ln\left(\frac{S_{ik}\alpha}{\widehat{x}_i \widehat{y}_k}\right) - \Psi(v_{ik}\alpha) \right) =: f(\alpha)$$

Lösung mit Newton-Raphson

$$\widehat{\alpha}^{(\nu+1)} := \widehat{\alpha}^{(\nu)} - \frac{f(\widehat{\alpha}^{(\nu)})}{f'(\widehat{\alpha}^{(\nu)})}.$$

ML-Schätzung von α

Als Startwert $\widehat{\alpha}^{(0)}$ verwenden wir den Momentenschätzer $\widehat{\alpha}^{M}$.

Zur Berechnung von f und f' benötigt man die Digamma- und die Trigamma-Funktion.

In Programmiersprachen wie `R`, `MATLAB` und `python` sind diese verfügbar.

Ansonsten kann man für $x > 1$ folgende Approximationen verwenden:

$$\Psi(x) \approx \ln(x) - \frac{1}{2x} - \frac{1}{12x^2} + \frac{1}{120x^4} - \frac{1}{260x^6} \quad \text{und}$$

$$\Psi'(x) \approx \frac{1}{x} + \frac{1}{2x^2} + \frac{1}{6x^3} - \frac{1}{30x^5} + \frac{6}{260x^7}$$

(für $x > 1$ relative Fehler $< 1\%$, für $x > 2$ relative Fehler $< 0{,}01\%$)

Momentenschätzer für α

Mit $\mathrm{Var}(Z_{ik}) = \frac{(x_i y_k)^2}{v_{ik}\alpha}$ erhalten wir $\frac{v_{ik}\,\mathrm{Var}(Z_{ik})}{(x_i y_k)^2} = \frac{1}{\alpha}$. Somit wäre $\frac{v_{ik}(Z_{ik}-x_i y_k)^2}{(x_i y_k)^2}$ ein erwartungstreuer Schätzer für $\frac{1}{\alpha}$. Da die Erwartungswerte $x_i y_k$ nicht bekannt sind sondern geschätzt werden ($I + K - 1$ Parameter), verwenden wir

$$\frac{1}{IK - (I + K - 1)} \sum_{i=1}^{I} \sum_{k=1}^{K} \frac{v_{ik}(Z_{ik} - \widehat{x}_i \widehat{y}_k)^2}{(\widehat{x}_i \widehat{y}_k)^2}$$

als Schätzer für $\frac{1}{\alpha}$. Wir erhalten den gewünschten Momentenschätzer

$$\widehat{\alpha}^M = \frac{IK}{\sum_{i=1}^{I} \sum_{k=1}^{K} \frac{v_{ik}(Z_{ik}-\widehat{x}_i\widehat{y}_k)^2}{(\widehat{x}_i\widehat{y}_k)^2}}.$$

Im Fall der KH-Statistik 2013:

- Momentenschätzer: $1{,}197 \cdot 10^{-3}$
- ML-Schätzer: $1{,}699 \cdot 10^{-3}$

Genauigkeit der Parameter

Wir legen oBdA $x_1 := 1$ fest, damit die Lösung eindeutig ist.

Als Approximation der Kovarianzmatrix der Parameter benötigen wir die Fisher-Informationsmatix:

$$\mathrm{Cov}(\widehat{\vartheta}) \approx I(\vartheta)^{-1} = \mathrm{E}\left(\left(-\frac{\partial^2 \ln L}{\partial\vartheta_i \partial\vartheta_j}\right)_{i,j}\right)^{-1}$$

mit $\widehat{\vartheta} = (\widehat{\vartheta}_1, \ldots, \widehat{\vartheta}_{I+K}) = (\widehat{x}_2, \ldots, \widehat{x}_I, \widehat{y}_1, \ldots, \widehat{y}_K, \widehat{\alpha})$.

Genauigkeit der Parameter

Partielle Ableitungen nach x_i und x_j:

$$-\frac{\partial^2 \ln(L)}{\partial x_i \partial x_j} = -\delta_{ij}\alpha \sum_{k=1}^{K} \left(\frac{v_{ik}}{x_i^2} - \frac{2S_{ik}}{x_i^3 y_k} \right)$$

$$\Rightarrow \mathrm{E}\left(-\frac{\partial^2 \ln(L)}{\partial x_i \partial x_j} \right) = -\delta_{ij}\alpha \sum_{k=1}^{K} \left(\frac{v_{ik}}{x_i^2} - \frac{2v_{ik}x_i y_k}{x_i^3 y_k} \right) = \frac{\delta_{ij}\alpha}{x_i^2} \sum_{k=1}^{K} v_{ik}$$

Partielle Ableitungen nach x_i und y_k:

$$-\frac{\partial^2 \ln(L)}{\partial x_i \partial y_k} = \frac{S_{ik}\alpha}{x_i^2 y_k^2} \qquad \Rightarrow \qquad \mathrm{E}\left(-\frac{\partial^2 \ln(L)}{\partial x_i \partial y_k} \right) = \frac{v_{ik}\alpha}{x_i y_k}$$

Genauigkeit der Parameter

Partielle Ableitungen nach y_k und y_l:

$$\mathsf{E}\left(-\frac{\partial^2 \ln(L)}{\partial y_k \partial y_l}\right) = \frac{\delta_{kl}\alpha}{y_k^2}\sum_{i=1}^{I} v_{ik} \qquad \text{(Symmetrie!)}$$

Partielle Ableitungen nach x_i und α:

$$-\frac{\partial^2 \ln(L)}{\partial x_i \partial \alpha} = -\frac{1}{x_i^2}\sum_{k=1}^{K}\left(\frac{S_{ik}}{y_k} - v_{ik}x_i\right)$$

$$\Rightarrow \mathsf{E}\left(-\frac{\partial^2 \ln(L)}{\partial x_i \partial \alpha}\right) = -\frac{1}{x_i^2}\sum_{k=1}^{K} 0 = 0$$

Partielle Ableitungen nach y_k und α:

$$\mathsf{E}\left(-\frac{\partial^2 \ln(L)}{\partial y_k \partial \alpha}\right) = 0 \qquad \text{(Symmetrie!)}$$

Genauigkeit der Parameter

Partielle Ableitungen nach α:

$$-\frac{\partial^2 \ln(L)}{\partial \alpha^2} = -\sum_{i=1}^{I}\sum_{k=1}^{K} v_{ik}\left(\frac{1}{\alpha} - v_{ik}\Psi'(v_{ik}\alpha)\right)$$

mit der Trigamma-Funktion $\Psi'(x) = \frac{1}{x} + \frac{1}{2x^2} + \frac{1}{6x^3} + \cdots$. Somit

$$\begin{aligned}
\mathrm{E}\left(-\frac{\partial^2 \ln(L)}{\partial \alpha^2}\right) &= \sum_{i=1}^{I}\sum_{k=1}^{K} v_{ik}^2\left(\frac{1}{2v_{ik}^2\alpha^2} + \frac{1}{6v_{ik}^3\alpha^3} + \cdots\right) \\
&= \frac{IK}{2\alpha^2}\left(1 + \frac{1}{IK}\sum_{i=1}^{I}\sum_{k=1}^{K}\frac{1}{3v_{ik}\alpha} + \cdots\right) \\
&\gtrsim \frac{IK}{2\alpha^2} \qquad \text{wenn für die meisten Zellen } v_{ik}\alpha \gg 1 \text{ gilt.}
\end{aligned}$$

Genauigkeit der Parameter

Insgesamt erhalten wir die folgende Informationsmatrix $I_{IK}(\{x_2, \ldots, x_I, y_1, \ldots, y_K, \alpha)$:

$$
\begin{array}{c|c|cc|c}
 & & \overbrace{x_2 \ \ldots \ x_I}^{I-1} & \overbrace{y_1 \ \ldots \ y_K}^{K} & \begin{matrix}1\\ \alpha\end{matrix} \\
\hline
I-1 & \begin{matrix}x_2\\ \vdots \\ x_I\end{matrix} & \left(\frac{\delta_{ij} v_{i+}}{x_i^2}\alpha\right)_{i,j} & \left(\frac{v_{ik}}{x_i y_k}\alpha\right)_{i,k} & 0 \\
K & \begin{matrix}y_1\\ \vdots \\ y_K\end{matrix} & \left(\frac{v_{ik}}{x_i y_k}\alpha\right)^t_{i,k} & \left(\frac{\delta_{kl} v_{+k}}{y_k^2}\alpha\right)_{k,l} & 0 \\
1 & \alpha & 0 & 0 & \approx \frac{IK}{2\alpha^2}
\end{array}
=: \begin{pmatrix} \alpha M & 0 \\ 0 & \frac{IK}{2\alpha^2} \end{pmatrix}
$$

Asymptotische Kovarianzmatrix

Da die asymptotische Kovarianzmatrix die Inverse der Informationsmatrix ist, erhalten wir schließlich

$$\mathrm{Cov}(\widehat{\vartheta}) \approx \begin{pmatrix} (\alpha M)^{-1} & 0 \\ 0 & \frac{2\alpha^2}{IK} \end{pmatrix}.$$

Insbesondere erhalten wir $\mathrm{Var}(\widehat{\alpha}) \approx \frac{2\alpha^2}{IK} \approx \frac{2\widehat{\alpha}^2}{IK}$, d.h.

$$\mathrm{Sd}(\widehat{\alpha}) \approx \frac{\widehat{\alpha}}{\sqrt{IK/2}}.$$

Im Fall der KH-Statistik 2013 ist $\sqrt{IK/2} = \sqrt{12 \cdot 8/2} \approx 6.9$, d.h. wir schätzen die Standardabweichung von $\widehat{\alpha}$ auf ungefähr 14% von $\widehat{\alpha}$.

Durch Invertieren der kompletten Informationsmatrix erhält man auch Schätzer für die Varianzen und Kovarianzen von $\widehat{x}_i$ und $\widehat{y}_k$.

Genauigkeit (Schätzfehler) der Schadenschätzer

Letztlich interessiert aber die Genauigkeit (Schätzfehler) der Prämienschätzer $\widehat{x}_i \cdot \widehat{y}_k$, da diese die Basis der Prämienkalkulation sind.

Wir betrachten die Abbildungen

$$T_{ik}(x_2, \ldots, x_I, y_1, \ldots, y_K, \alpha) := x_i y_k.$$

Nach dem Transformationssatz für ML-Schätzer erhalten wir

$$\begin{aligned}
\operatorname{Var}(\widehat{x}_i \widehat{y}_k) &\approx \left(\frac{\partial T_{ik}}{\partial \vartheta}\right) \operatorname{Cov}(\widehat{\vartheta}) \left(\frac{\partial T_{ik}}{\partial \vartheta}\right)^t \\
&= (y_k, x_i) \begin{pmatrix} \operatorname{Var}(\widehat{x}_i) & \operatorname{Cov}(\widehat{x}_i, \widehat{y}_k) \\ \operatorname{Cov}(\widehat{y}_k, \widehat{x}_i) & \operatorname{Var}(\widehat{y}_k) \end{pmatrix} \begin{pmatrix} y_k \\ x_i \end{pmatrix} \\
&= \operatorname{Var}(\widehat{x}_i)\, y_k^2 + 2 x_i \operatorname{Cov}(\widehat{x}_i, \widehat{y}_k) y_k + x_i^2 \operatorname{Var}(\widehat{y}_k).
\end{aligned}$$

Beispiel: KH-Statistik 2013 – Schadenaufwand

$\widehat{Vco}(\hat{x}_i\hat{y}_k)$

Regionalklasse	Fahrleistung 0 -	7 -	10 -	13 -	16 -	21 -	26 -	31 -
1	0.060	0.061	0.064	0.069	0.074	0.115	0.151	0.209
2	0.050	0.051	0.054	0.061	0.066	0.110	0.147	0.207
3	0.052	0.052	0.055	0.061	0.067	0.110	0.148	0.207
4	0.058	0.058	0.061	0.067	0.072	0.113	0.150	0.209
5	0.053	0.053	0.056	0.062	0.068	0.111	0.148	0.207
6	0.054	0.054	0.057	0.063	0.069	0.111	0.148	0.207
7	0.049	0.049	0.052	0.059	0.065	0.109	0.147	0.206
8	0.067	0.067	0.070	0.075	0.080	0.118	0.154	0.211
9	0.054	0.054	0.057	0.063	0.069	0.111	0.148	0.207
10	0.065	0.065	0.068	0.073	0.078	0.117	0.153	0.211
11	0.091	0.091	0.093	0.097	0.101	0.134	0.166	0.220
12	0.065	0.066	0.068	0.074	0.079	0.118	0.154	0.211

Beispiel: KH-Statistik 2013 – Schadenaufwand

Für die KH-Statistik 2013 sehen wir

$$\widehat{\mathrm{Vco}}(\widehat{x}_i\widehat{y}_k) = \frac{\widehat{\mathrm{Sd}}(\widehat{x}_i\widehat{y}_k)}{\widehat{x}_i\widehat{y}_k} \in [0{,}049, 0{,}220].$$

Mit $\pm$ zwei Standardabweichungen als Konfidenzintervall, erhalten wir beispielsweise

- für die Zelle (Regionalklasse 3 / Fahrleistung 10–): $\widehat{x}_i\widehat{y}_k = 144 \pm 16$,
- für die Zelle (Regionalklasse 12 / Fahrleistung 31–): $\widehat{x}_i\widehat{y}_k = 265 \pm 112$.

Diese Schätzung ist relevant für die Fragestellung

- ob zwei Zellen signifikant verschiedene Erwartungswert-Schätzer haben und
- ob man eine etwas abweichende „alte“ Prämie beibehalten kann.

Beispiel: KH-Statistik 2013 – Schadenaufwand

ausgeglichene Schadenbedarfe $\hat{x}_i\hat{y}_k$ (Gamma-Modell)

Regionalklasse	Fahrleistung 0 -	7 -	10 -	13 -	16 -	21 -	26 -	31 -	$\hat{x}_i$
1	109,6	119,5	137,5	145,5	146,6	154,7	159,4	194,5	1,000
2	106,4	116,0	133,4	141,2	142,3	150,2	154,7	188,8	0,970
3	114,6	125,0	143,8	152,2	153,4	161,9	166,7	203,5	1,046
4	117,7	128,4	147,7	156,3	157,5	166,2	171,2	209,0	1,074
5	118,4	129,1	148,5	157,2	158,4	167,2	172,2	210,1	1,080
6	120,4	131,3	151,0	159,8	161,0	169,9	175,1	213,6	1,098
7	122,0	133,1	153,1	162,0	163,2	172,3	177,4	216,5	1,113
8	117,0	127,6	146,8	155,4	156,5	165,2	170,2	207,7	1,068
9	147,9	161,3	185,6	196,4	197,9	208,9	215,1	262,5	1,350
10	115,2	125,6	144,5	152,9	154,1	162,6	167,5	204,4	1,051
11	120,2	131,1	150,8	159,6	160,8	169,7	174,8	213,3	1,097
12	149,0	162,6	187,0	197,9	199,4	210,5	216,8	264,5	1,360
$\hat{y}_k$	109,6	119,5	137,5	145,5	146,6	154,7	159,4	194,5	

Beispiel: KH-Statistik 2013 – Schadenaufwand

$\widehat{Vco}(\hat{x}_i\hat{y}_k)$

Regionalklasse	Fahrleistung 0 -	7 -	10 -	13 -	16 -	21 -	26 -	31 -
1	0.060	0.061	0.064	0.069	0.074	0.115	0.151	0.209
2	0.050	0.051	0.054	0.061	0.066	0.110	0.147	0.207
3	0.052	0.052	0.055	0.061	0.067	0.110	0.148	0.207
4	0.058	0.058	0.061	0.067	0.072	0.113	0.150	0.209
5	0.053	0.053	0.056	0.062	0.068	0.111	0.148	0.207
6	0.054	0.054	0.057	0.063	0.069	0.111	0.148	0.207
7	0.049	0.049	0.052	0.059	0.065	0.109	0.147	0.206
8	0.067	0.067	0.070	0.075	0.080	0.118	0.154	0.211
9	0.054	0.054	0.057	0.063	0.069	0.111	0.148	0.207
10	0.065	0.065	0.068	0.073	0.078	0.117	0.153	0.211
11	0.091	0.091	0.093	0.097	0.101	0.134	0.166	0.220
12	0.065	0.066	0.068	0.074	0.079	0.118	0.154	0.211

Beispiel: KH-Statistik 2013 – Schadenaufwand

ausgeglichene Schadenbedarfe $\hat{x}_i\hat{y}_k$ (Gamma-Modell)

Regionalklasse \ Fahrleistung	0 -	7 -	10 -	13 -	16 -	21 -	26 -	31 -	$\hat{x}_i$
1	109,6	119,5	137,5	145,5	146,6	154,7	159,4	194,5	1,000
2	106,4	116,0	133,4	141,2	142,3	150,2	154,7	188,8	0,970
3	114,6	125,0	143,8	152,2	153,4	161,9	166,7	203,5	1,046
4	117,7	128,4	147,7	156,3	157,5	166,2	171,2	209,0	1,074
5	118,4	129,1	148,5	157,2	158,4	167,2	172,2	210,1	1,080
6	120,4	131,3	151,0	159,8	161,0	169,9	175,1	213,6	1,098
7	122,0	133,1	153,1	162,0	163,2	172,3	177,4	216,5	1,113
8	117,0	127,6	146,8	155,4	156,5	165,2	170,2	207,7	1,068
9	147,9	161,3	185,6	196,4	197,9	208,9	215,1	262,5	1,350
10	115,2	125,6	144,5	152,9	154,1	162,6	167,5	204,4	1,051
11	120,2	131,1	150,8	159,6	160,8	169,7	174,8	213,3	1,097
12	149,0	162,6	187,0	197,9	199,4	210,5	216,8	264,5	1,360
$\hat{y}_k$	109,6	119,5	137,5	145,5	146,6	154,7	159,4	194,5	

Beispiel: KH-Statistik 2013 – Schadenaufwand

$\widehat{Vco}(\hat{x}_i \hat{y}_k)$

Regionalklasse \ Fahrleistung	0 -	7 -	10 -	13 -	16 -	21 -	26 -	31 -
1	0.060	0.061	0.064	0.069	0.074	0.115	0.151	0.209
2	0.050	0.051	0.054	0.061	0.066	0.110	0.147	0.207
3	0.052	0.052	0.055	0.061	0.067	0.110	0.148	0.207
4	0.058	0.058	0.061	0.067	0.072	0.113	0.150	0.209
5	0.053	0.053	0.056	0.062	0.068	0.111	0.148	0.207
6	0.054	0.054	0.057	0.063	0.069	0.111	0.148	0.207
7	0.049	0.049	0.052	0.059	0.065	0.109	0.147	0.206
8	0.067	0.067	0.070	0.075	0.080	0.118	0.154	0.211
9	0.054	0.054	0.057	0.063	0.069	0.111	0.148	0.207
10	0.065	0.065	0.068	0.073	0.078	0.117	0.153	0.211
11	0.091	0.091	0.093	0.097	0.101	0.134	0.166	0.220
12	0.065	0.066	0.068	0.074	0.079	0.118	0.154	0.211

Mittlerer Prognose-Fehler

Ebenso wichtig ist die Abweichung der künftigen (nächstjährigen) Realisierung Z_{ik}^* vom Schätzer $\widehat{x}_i\widehat{y}_k$, d.h. der sog. *mittlere quadratische (Prognose-)Fehler*

$$\begin{aligned}
\mathrm{mse}(\widehat{x}_i\widehat{y}_k) &:= \mathrm{E}\left((Z_{ik}^* - \widehat{x}_i\widehat{y}_k)^2\right) \\
&= \mathrm{E}\left(\Big(\big(Z_{ik}^* - \mathrm{E}(Z_{ik}^*)\big) + \big(\mathrm{E}(Z_{ik}^*) - \mathrm{E}(\widehat{x}_i\widehat{y}_k)\big) + \big(\mathrm{E}(\widehat{x}_i\widehat{y}_k) - \widehat{x}_i\widehat{y}_k\big)\Big)^2\right) \\
&\approx \underbrace{\mathrm{Var}(Z_{ik}^*)}_{\textit{Zufallsfehler}} + \underbrace{\left[\mathrm{E}(Z_{ik}^*) - \mathrm{E}(\widehat{x}_i\widehat{y}_k)\right]^2}_{\textit{Bias}} + \underbrace{\mathrm{Var}(\widehat{x}_i\widehat{y}_k)}_{\textit{Schaetzfehler}}.
\end{aligned}$$

Hierbei verwendet: $\mathrm{Cov}(Z_{ik}^*, \widehat{x}_i\widehat{y}_k) \approx 0$, da die Zufallsvariablen aus verschiedenen (und daher nahezu) unabhängigen Jahren sind.

Der Bias sollte wegen der Konsistenz der ML-Schätzer nahezu null sein.

Mittlerer Prognose-Fehler

In unserem Modell ist $Z_{ik}^* \sim \Gamma(x_i y_k, v_{ik}^* \alpha)$, wobei v_{ik}^* die Volumina des nächsten Jahres sind. Daher erhalten wir den Zufallsfehler

$$\mathrm{Var}(Z_{ik}^*) = \frac{(x_i y_k)^2}{v_{ik}^* \alpha}.$$

In dem Beispiel der KH-Statistik 2013 ist für $v_{ik}^* = v_{ik}$ der Zufallsfehler für alle Zellen größer als der Schätzfehler, und es ist

$$\widehat{\mathrm{Vco}}(Z_{ik}^*) = \frac{1}{\sqrt{v_{ik} \widehat{\alpha}}} \in [0{,}080, 1{,}246].$$

Beispiel: KH-Statistik 2013 – Schadenaufwand

$\widehat{Vco}(Z_{ik})$

Regionalklasse \ Fahrleistung	0 -	7 -	10 -	13 -	16 -	21 -	26 -	31 -
1	0,102	0,104	0,130	0,168	0,190	0,373	0,506	0,726
2	0,081	0,084	0,104	0,134	0,154	0,303	0,419	0,599
3	0,087	0,087	0,103	0,133	0,156	0,312	0,435	0,629
4	0,101	0,100	0,118	0,151	0,180	0,356	0,503	0,730
5	0,091	0,090	0,104	0,132	0,156	0,309	0,440	0,604
6	0,093	0,092	0,108	0,136	0,161	0,317	0,440	0,640
7	0,080	0,082	0,096	0,121	0,144	0,285	0,391	0,570
8	0,118	0,119	0,143	0,181	0,215	0,410	0,578	0,827
9	0,091	0,091	0,108	0,139	0,163	0,320	0,449	0,642
10	0,114	0,115	0,138	0,179	0,210	0,410	0,567	0,802
11	0,161	0,166	0,202	0,265	0,311	0,596	0,842	1,246
12	0,109	0,114	0,138	0,184	0,237	0,515	0,716	1,002

Anpassungstest (Goodness of Fit-Test)

Anpassungstest durch Vergleich der Likelihood L mit der Likelihood L_1 des vollen Modells mit lauter individuellen $\mathsf{E}(Z_{ik}) =: \mu_{ik}$ mittels Likelihood-Quotiententest. Hierbei muss α bekannt sein (z.B. aus Einzelrisikodaten), sonst ist L_1 überparametrisiert. Wie oben berechnen wir

$$\ln L_1(\{\mu_{ik}\}) = \sum_{i=1}^{I}\sum_{k=1}^{K}\left(v_{ik}\alpha \ln\left(\frac{S_{ik}\alpha}{\mu_{ik}}\right) - \frac{S_{ik}\alpha}{\mu_{ik}} - \ln\left(Z_{ik}\Gamma(v_{ik}\alpha)\right)\right)$$

Partielle Ableitung liefert

$$0 = \frac{\partial \ln L_1(\{\mu_{ik}\})}{\partial \mu_{ik}} = -\frac{v_{ik}\alpha}{\mu_{ik}} + \frac{S_{ik}\alpha}{\mu_{ik}^2} \qquad \Rightarrow \qquad \widehat{\mu}_{ik} = \frac{S_{ik}}{v_{ik}} = Z_{ik}$$

(keine Überraschung!).

Anpassungstest (Goodness of Fit-Test)

Daher erhalten wir

$$\ln L_1(\{\widehat{\mu}_{ik}\}) = \sum_{i=1}^{I}\sum_{k=1}^{K}\left(v_{ik}\alpha\ln(v_{ik}\alpha) - v_{ik}\alpha - \ln\left(Z_{ik}\Gamma(v_{ik}\alpha)\right)\right)$$

$$\begin{aligned}\ln L(\{\widehat{x}_i,\widehat{y}_k\}) &= \sum_{i=1}^{I}\sum_{k=1}^{K}\left(v_{ik}\alpha\ln\left(\frac{S_{ik}\alpha}{\widehat{x}_i\widehat{y}_k}\right) - \frac{S_{ik}\alpha}{\widehat{x}_i\widehat{y}_k} - \ln\left(Z_{ik}\Gamma(v_{ik}\alpha)\right)\right)\\ &= \sum_{i=1}^{I}\sum_{k=1}^{K}\left(v_{ik}\alpha\ln\left(\frac{S_{ik}\alpha}{\widehat{x}_i\widehat{y}_k}\right) - v_{ik}\alpha - \ln\left(Z_{ik}\Gamma(v_{ik}\alpha)\right)\right).\end{aligned}$$

Anpassungstest (Goodness of Fit-Test)

Mittels Likelihood-Quotiententest erhalten wir

$$2 \ln \frac{L_1}{L} = 2(\ln L_1 - \ln L) = 2 \sum_{i=1}^{I} \sum_{k=1}^{K} \left(v_{ik} \alpha \ln \left(\frac{v_{ik} \widehat{x}_i \widehat{y}_k \alpha}{S_{ik} \alpha} \right) \right)$$

$$= 2\alpha \sum_{i=1}^{I} \sum_{k=1}^{K} v_{ik} \ln \left(\frac{\widehat{x}_i \widehat{y}_k}{Z_{ik}} \right) \sim \chi^2_{IK-(I+K-1)},$$

wobei $IK - (I + K - 1)$ die Anzahl der eingesparten Parameter ist.

Anpassungstest (Goodness of Fit-Test)

In unserem Beispiel der KH-Statistik 2013 erhalten wir $2\ln\frac{L_1}{L} = 57.792 \cdot \alpha$ mit $12 \cdot 8 - 19 = 77$ Freiheitsgraden, d.h. keine Ablehnung der Kreuzklassifikation für

$$\alpha \leq \frac{\chi^2_{77;95\%}}{57.792} = \frac{98{,}5}{57.792} \approx 0{,}001704$$

bei einem Konfidenzniveau von 95%.

Beachte, dass $\widehat{\alpha} = 0{,}001699$ in unserem Beispiel nicht herangezogen werden sollte um die Kreuzklassifikation anzunehmen bzw. abzulehnen, da α im Modell geschätzt wurde. Wie oben bemerkt muss α bekannt sein.

Großschadenproblematik und Kupierung

Im Beispiel KH 2013 ist Zelle (Regionalklasse 9 / Fahrleistung 10–) ein Ausreißer:

- $Z_{ik} = 254{,}4$
- $\widehat{x}_i \cdot \widehat{y}_k = 185{,}6$

Wegen $v_{ik} = 50.719$ ist dies ein realisierter Schaden von 12,9 Mio. statt der Schätzung von 9,4 Mio. Diese Abweichung kann durch einen einzelnen Großschaden bedingt sein.

Noch extremer ist die Situation in der Feuer-Versicherung, wo oft über 50% des Gesamtschadens von den 1% größten Schäden stammen.

Beispiel: KH-Statistik 2013 – Schadenaufwand

Schadenbedarf Z_{ik}

Regionalklasse \ Fahrleistung	0 -	7 -	10 -	13 -	16 -	21 -	26 -	31 -	∅
1	125,2	114,6	114,4	151,7	132,6	220,0	114,1	135,3	125,8
2	105,4	128,0	138,5	116,0	129,2	130,5	114,7	168,3	122,1
3	104,3	113,3	123,5	247,8	158,4	193,6	149,0	162,0	135,0
4	122,7	136,7	130,5	159,3	130,1	175,1	262,5	190,8	136,4
5	112,3	130,1	134,3	144,9	221,5	141,9	292,0	224,6	139,1
6	130,7	124,6	133,4	193,8	140,6	186,4	146,5	145,2	139,9
7	108,8	136,1	195,0	135,0	143,4	169,9	161,7	317,0	142,6
8	122,7	120,2	139,4	180,8	149,0	129,9	212,6	150,6	135,9
9	126,1	169,2	254,4	144,5	180,4	173,4	182,0	183,1	172,5
10	126,8	119,5	136,2	132,4	174,8	165,5	138,9	258,5	133,0
11	111,7	125,2	163,0	132,6	234,2	203,7	119,4	191,3	139,9
12	180,8	163,2	145,7	142,5	215,0	154,5	201,6	666,2	168,5
∅	121,0	132,2	153,3	159,1	161,5	168,8	175,1	216,6	140,6

Beispiel: KH-Statistik 2013 – Schadenaufwand

ausgeglichene Schadenbedarfe $\hat{x}_i\hat{y}_k$ (Gamma-Modell)

Regionalklasse	Fahrleistung 0 -	7 -	10 -	13 -	16 -	21 -	26 -	31 -	$\hat{x}_i$
1	109,6	119,5	137,5	145,5	146,6	154,7	159,4	194,5	1,000
2	106,4	116,0	133,4	141,2	142,3	150,2	154,7	188,8	0,970
3	114,6	125,0	143,8	152,2	153,4	161,9	166,7	203,5	1,046
4	117,7	128,4	147,7	156,3	157,5	166,2	171,2	209,0	1,074
5	118,4	129,1	148,5	157,2	158,4	167,2	172,2	210,1	1,080
6	120,4	131,3	151,0	159,8	161,0	169,9	175,1	213,6	1,098
7	122,0	133,1	153,1	162,0	163,2	172,3	177,4	216,5	1,113
8	117,0	127,6	146,8	155,4	156,5	165,2	170,2	207,7	1,068
9	147,9	161,3	185,6	196,4	197,9	208,9	215,1	262,5	1,350
10	115,2	125,6	144,5	152,9	154,1	162,6	167,5	204,4	1,051
11	120,2	131,1	150,8	159,6	160,8	169,7	174,8	213,3	1,097
12	149,0	162,6	187,0	197,9	199,4	210,5	216,8	264,5	1,360
$\hat{y}_k$	109,6	119,5	137,5	145,5	146,6	154,7	159,4	194,5	

Großschadenproblematik und Kupierung

- Wenn Großschäden zufällig über die Risikogruppen verteilt sind, sollte man sie kupieren (= abschneiden), aber nicht ganz eliminieren.
- Die Kupierungsgrenze sollte nicht für alle Risikogruppen gleich gewählt werden sondern z.B. das gleiche Perzentil (z.B. 99.9%) der Verteilung der Einzelschadenhöhen.
- Dies ist bei kleiner Schadenzahl nicht praktikabel. Dann kann die Grenze mit Tschebyscheff oder Cantelli ermittelt werden (mit einem kleineren Perzentil, z.B. 99%):

$$\frac{\mathrm{Var}(X)}{\mathrm{Var}(X)+a^2} = 1\% \quad \Rightarrow \quad a^2 = 99 \cdot \mathrm{Var}(X) \quad \Rightarrow \quad a \approx 10 \cdot \mathrm{Sd}(X).$$

Großschadenproblematik und Kupierung

- Dieser Ansatz führt zu Kupierung bei einer Grenze von 10 Standardabweichungen über dem Erwartungswert.
- Der Gesamtbetrag der durch Kupierung weggefallenen Schadenteile muss durch eine einheitliche prozentuale Erhöhung der Erwartungswertschätzer nach Kupierung wieder eingerechnet werden.

Inhalt

Verallgemeinerte Lineare Modelle (GLMs)

Ein *lineares Modell* besteht aus unabhängigen Beobachtungen $Z = (Z_1, \ldots, Z_N)$ mit gleichen Varianzen $\mathsf{Var}(Z_n) = \sigma^2$, $1 \leq n \leq N$, und dem linearen Prediktor

$$\mathsf{E}(Z_n) = \sum_{m=1}^{M} \xi_{nm}\beta_m, \quad \text{d.h. } \mathsf{E}(Z) = X\beta$$

mit bekannter Designmatrix $X = (\xi_{nm})$ mit vollem Rang $M < N$ und unbekannten Parametern $\beta = (\beta_1, \ldots, \beta_M)$.

Ziel: Viele Beobachtungen Z_n durch wenige Parameter β_m erklären, d.h. $M \ll N$.

Lineare Modelle

Dann ist $\widehat{\beta} := (X^t X)^{-1} X^t Z$ der *Gauß-Markov-Schätzer*, d.h. der (eindeutig bestimmte) lineare erwartungstreue Schätzer, der unter allen linearen erwartungstreuen Schätzern $\widetilde{\beta} = \widetilde{\beta}(Z_1, \ldots, Z_N)$ die minimale erwartete quadratische Abweichung

$$\mathrm{E}\left(\|\mathrm{E}(Z) - X\widetilde{\beta}\|^2\right)$$

hat (BLUE = Best Linear Unbiased Estimator).

$\widehat{\beta}$ minimiert unter den linearen Schätzern $\widetilde{\beta}$ die Summe der Fehlerquadrate

$$\sum_{n=1}^{N} \left(Z_n - \sum_{m=1}^{M} \xi_{nm} \widetilde{\beta}_m \right)^2 = \|Z - X\widetilde{\beta}\|^2.$$

Sind die Z_n normalverteilt, so sind auch die Verteilungen von $\widehat{\beta}$ und den Teststatistiken für diverse lineare Hypothesen angebbar. Außerdem ist $\widehat{\beta}$ dann der ML-Schätzer für β.

Verallgemeinerte Lineare Modelle

Problematik bei Linearen Modellen:

- In der Praxis sind die Z_n oft nicht normalverteilt.
- Die Linearität ist oft erst nach gewisser Transformation von $E(Z_n)$ gegeben (z.B. $\ln(x_i y_k) = \ln x_i + \ln y_k$).
- Die Varianzen sind oft nicht für alle Z_n gleich.
- Außerdem hätte man gerne eine Software für die komplexen Kalkulationen wie im Abschnitt über das kreuzklassifizierte Gamma-Modell.

Dies alles leisten die GLMs!

Exponentialfamilien

Definition: Eine *einparametrige Exponentialfamilie mit Volumenmaßen* ist eine Schar von Verteilungen P_ϑ, wobei ϑ ein offenes Intervall $\Theta \subset \mathbb{R}$ durchläuft, so dass für jedes ϑ die Verteilung P_ϑ eine Dichte bezüglich eines gemeinsamen dominierenden Maßes ν von der Form

$$f(x \mid \vartheta, \phi, v) = \exp\left(\frac{\vartheta x - b(\vartheta)}{\phi / v} + c(x, \phi, v)\right), \quad x \in \mathbb{R}$$

besitzt. Hierbei ist

- $b\colon \Theta \to \mathbb{R}$ zweimal stetig differenzierbar mit $b' > 0$ und $b'' > 0$,
- $v > 0$ ein bekanntes Volumenmaß,
- $\phi > 0$ ein Skalenparameter und
- $c\colon \mathbb{R} \times (0, \infty)^2 \to \mathbb{R}$ messbar.

Wir bezeichnen die Verteilung von P_ϑ dann mit $\mathrm{EDF}_\nu(\vartheta, \phi, v, b, c)$.

Exponentialfamilien

Lemma: *Für* $X_\vartheta \sim \mathrm{EDF}_\nu(\vartheta, \phi, v, b, c)$ *gilt*

$$\mathrm{E}(X_\vartheta) = b'(\vartheta) \quad \textit{und} \quad \mathrm{Var}(X_\vartheta) = \frac{\phi}{v} b''(\vartheta).$$

Definition: Die Funktion $V : b'(\Theta) \to \mathbb{R}$ mit

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

heißt *Varianzfunktion* der Exponentialfamilie.

Beispiele für Exponentialfamilien

Normalverteilung:

Wir wählen $\Theta := \mathbb{R}$, $\vartheta := \mu$, $\phi > 0$, $v > 0$, $\sigma^2 := \phi/v$, $b(\vartheta) := \vartheta^2/2$ und

$$c(x, \phi, v) := -\frac{x^2}{2\phi/v} - \frac{1}{2}\ln(2\pi\phi/v) = -\frac{x^2}{2\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2).$$

Dann gilt

$$\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) = \exp\left(\frac{\vartheta x - b(\vartheta)}{\phi/v} + c(x, \phi, v)\right),$$

d.h. $\mathrm{EDF}_\nu(\vartheta, \phi, v, b, c) = \mathcal{N}(\mu, \sigma^2)$.

Es ist $\mathrm{E}(X_\vartheta) = b'(\vartheta) = \vartheta = \mu$, $\mathrm{Var}(X_\vartheta) = \frac{\phi}{v} b''(\vartheta) = \phi/v = \sigma^2$ und $V(\mu) = 1$.

Beispiele für Exponentialfamilien

Overdispersed Poisson:

Wir wählen $\Theta := \mathbb{R}$, $\phi > 0$, $\vartheta := \ln(\lambda)$, $b(\vartheta) := \exp(\vartheta)$, $v > 0$ und $c(x, \phi, v) := -\ln(\Gamma(\frac{xv}{\phi} + 1)) - \frac{xv}{\phi} \ln(\phi/v)$. Als dominierendes Maß wählen wir das Zählmaß auf $\frac{\phi}{v} \cdot \mathbb{N}_0$. Dann gilt für $x \in \frac{\phi}{v} \cdot \mathbb{N}_0$

$$\begin{aligned} e^{-\lambda v/\phi} \frac{(\lambda v/\phi)^{xv/\phi}}{(xv/\phi)!} &= \exp\left(\frac{xv \ln(\lambda) - \lambda v}{\phi} + c(x, \phi, v) \right) \\ &= \exp\left(\frac{\vartheta x - b(\vartheta)}{\phi/v} + c(x, \phi, v) \right), \end{aligned}$$

d.h. für eine Zufallsgröße N mit dieser Verteilung gilt $Nv/\phi \sim \text{Poi}(\lambda v/\phi)$.

Es ist $\mathsf{E}(X_\vartheta) = b'(\vartheta) = e^\vartheta = \lambda$, $\text{Var}(X_\vartheta) = \frac{\phi}{v} b''(\vartheta) = \frac{\phi}{v} e^\vartheta = \frac{\phi}{v}\lambda$ und $V(\mu) = \mu$. Für $\phi = v$ ergibt sich die normale Poisson-Verteilung.

Beispiele für Exponentialfamilien

Gamma-Verteilung:

Wir wählen $\Theta := (-\infty, 0)$, $\vartheta := -1/\mu$, $\phi := 1/\alpha$, $v > 0$, $b(\vartheta) := -\ln(-\vartheta)$ und

$$c(x, \phi, v) := (v/\phi - 1)\ln(x) - \ln(\Gamma(v/\phi)) + \frac{v}{\phi}\ln(v/\phi).$$

Dann gilt

$$\frac{\left(\frac{\alpha v}{\mu}\right)^{\alpha v}}{\Gamma(\alpha v)} x^{\alpha v - 1} \exp(-\frac{\alpha v x}{\mu}) = \exp\left(\frac{\vartheta x - b(\vartheta)}{\phi/v} + c(x, \phi, v)\right),$$

d.h. $\mathrm{EDF}_\nu(\vartheta, \phi, v, b, c) = \Gamma(\mu, v\alpha)$.

Es ist $\mathrm{E}(X_\vartheta) = b'(\vartheta) = -1/\vartheta = \mu$, $\mathrm{Var}(X_\vartheta) = \frac{\phi}{v} b''(\vartheta) = \frac{\phi}{v\vartheta^2} = \frac{\mu^2}{v\alpha}$ und $V(\mu) = \mu^2$.

Verallgemeinerte Lineare Modelle

Definition: Ein *verallgemeinertes lineares Modell (GLM)* besteht aus

- unabhängigen Beobachtungen $Z_1, \ldots, Z_N$ mit bekannten Volumenmaßen $v_1, \ldots, v_N$, so dass
$$Z_n \sim \mathrm{EDF}_{\nu_n}(\vartheta_n, \phi, v_n, b, c),$$
mit für alle n gleichen b, c (bekannt) und ϕ (unbekannt),
- einer bekannten Designmatrix $X = (\xi_{ij}) \in \mathbb{R}^{N \times M}$,
- unbekannten Parametern $\beta = (\beta_1, \ldots, \beta_M)$ und
- einer zweimal stetig differenzierbaren *Linkfunktion* $g\colon (a, b) \to \mathbb{R}$ mit $g'(x) \neq 0$ für alle $x \in (a, b)$,

so dass für die Erwartungswerte $\mu_n := \mathrm{E}(Z_n)$ gilt

$$g(\mu_n) = \sum_{m=1}^{M} \xi_{nm} \beta_m.$$

ML-Schätzung und Fisher-Informationsmatrix bei GLMs

Lineare Modelle sind spezielle GLMs mit $v_n = 1$, $\phi = \sigma^2$, $V(\mu_n) = 1$ und $g(\mu_n) = \mu_n$.

Lemma: *In einem GLM gilt*

$$\frac{\partial}{\partial \beta_m} \ln L(\beta, \phi) = \frac{1}{\phi} \sum_{n=1}^{N} \frac{v_n (Z_n - \mu_n)}{V(\mu_n)} \frac{\xi_{nm}}{g'(\mu_n)}.$$

Bezeichnet $I(\beta, \phi) = \Big(I_{mk}(\beta, \phi) \Big)_{mk}$ *die Fisher-Informationsmatrix bezüglich der Parameter* β*, so gilt*

$$I_{mk}(\beta, \phi) = \frac{1}{\phi} \sum_{n=1}^{N} \frac{v_n}{V(\mu_n)} \frac{\xi_{nm} \xi_{nk}}{g'(\mu_n)^2}.$$

ML-Schätzung und Fisher-Informationsmatrix bei GLMs

Bemerkungen:

- Die Lösung der ML-Gleichungen für β hängt offenbar weder von den dominierenden Maßen ν_n noch von der Funktion $c(x, \phi, v)$ ab.
- Von der Verteilungsfamilie geht nur die Varianzfunktion $V(\mu)$ in die Likelihoodgleichnungen ein.
- Ferner kann man β schätzen, ohne ϕ zu kennen bzw. zu schätzen (wie α im Gamma-Modell). Zur Berechnung der Fisher-Informationsmatrix wird aber ein Schätzer von ϕ benötigt.
- Es gibt Softwarepakete für die GLMs. Dort werden diese ML-Gleichungen üblicherweise mit dem Fisher-Scoring-Algorithmus gelöst.

Fisher Scoring

Fisher Scoring-Verfahren:

Ersetzt man im Newton-Raphson-Verfahren zur Lösung von

$$\nabla_\beta \phi \ell(\beta, \phi) = \nabla_\beta \phi \ln L(\beta, \phi) = 0$$

die Matrix $\nabla_\beta \nabla_\beta^t \phi \ell(\beta, \phi)$ durch $-\phi I(\beta, \phi) = \phi \, \mathrm{E}_\beta(\nabla_\beta \nabla_\beta^t \ell(\beta, \phi))$, so spricht man vom *Fisher-Scoring-Verfahren*.

Die Iteration lautet dann

$$\widehat{\beta}^{(\nu+1)} := \widehat{\beta}^{(\nu)} + [\phi I(\widehat{\beta}^{(\nu)}, \phi)]^{-1} \cdot \nabla_\beta \phi \ell(\widehat{\beta}^{(\nu)}, \phi).$$

Schätzung des Dispersionsparameters

Schätzung von ϕ:

Ist $\widehat{\beta}$ der ML-Schätzer für β und sind die $\widehat{\mu}_n$ die zugehörigen Schätzer für μ_n, so lautet ein Schätzer für den Skalenparameter ϕ

$$\widehat{\phi} := \frac{1}{N-M} \sum_{n=1}^{N} \frac{v_n (Z_n - \widehat{\mu}_n)^2}{V(\widehat{\mu}_n)}.$$

Dieser wird benötigt um einen Schätzer für die Kovarianzmatrix von $\widehat{\beta}$ zu berechnen.

Devianz

Wir fixieren nun Exponentialfamilie, Dispersion ϕ und Linkfunktion g und stellen ein GLM mit Designmatrix $X \in \mathbb{R}^{N \times M}$ als M-dimensionale Fläche in $\mathbb{R}^N$ dar:

$$\mathcal{M} := \left\{ (\mu_1, \ldots, \mu_N) \,\middle|\, g(\mu_n) = \sum_{m=1}^{M} \xi_{nm} \beta_m \right\}.$$

Extremfall ist das volle Modell $\mathcal{M}_N = \mathbb{R}^N$ mit der N-dimensionalen Einheitsmatrix als Designmatrix.

Devianz

Definition: Es sei

$$L(\mathcal{M}) := 2 \sup_{b'(\vartheta)\in\mathcal{M}} \sum_{n=1}^{N} v_i(\vartheta_i Z_i - b(\vartheta_i)).$$

Dann heißt

$$D(\mathcal{M}) := L(\mathcal{M}_N) - L(\mathcal{M})$$

die *Devianz* des Modells $\mathcal{M}$.

Devianz

Bemerkungen:

- Im linearen Modell ist die Devianz gleich der Summe der Fehlerquadrate.
- Die Devianz $D(\mathcal{M})$ ist die Differenz der log-Likelihoods des vollen Modells und $\mathcal{M}$ multipliziert mit 2ϕ. Die Devianz ist unabhängig vom Dispersionsparameter ϕ.
- Sind $\mathcal{M}_a \subset \mathcal{M}_b$ geschachtelte Modelle mit $N_a < N_b$ Freiheitsgraden und ist das kleinere Modell $\mathcal{M}_a$ richtig, so ist

$$\frac{D(\mathcal{M}_a) - D(\mathcal{M}_b)}{\phi}$$

χ^2-verteilt mit $N_b - N_a$ Freiheitsgraden (Likelihood-Quotiententest).

Anwendung von GLMs auf das kreuzklassifizierte Ausgleichsproblem

Linearisierung durch log-Link, d.h. $g(\mu) = \ln \mu$. Wir verwenden die Parameter

$$(\beta_1, \ldots, \beta_I, \beta_{I+1}, \ldots, \beta_{I+K}) = (\ln x_1, \ldots, \ln x_I, \ln y_1, \ldots, \ln y_K)$$

und die Design-Matrix $X = (\xi_{ik,m})$ mit

$$\xi_{ik,m} = \begin{cases} 1 & \text{für } m = i \text{ oder } m = I + k \\ 0 & \text{sonst,} \end{cases}$$

d.h. die Design-Matrix besteht aus Dummy-Variablen, die nur das Zutreffen einer bestimmten Merkmalsausprägungskombination anzeigen. Hiermit $\mathrm{E}(Z_{ik}) = x_i y_k$ und

$$g(x_i y_k) = \ln x_i + \ln y_k = \beta_i + \beta_{I+k} = \sum_{m=1}^{M} \xi_{ik,m} \beta_m.$$

ML-Gleichungen

Damit erhalten wir die ML-Gleichungen

$$0 = \frac{1}{\phi} \sum_{i=1}^{I} \sum_{k=1}^{K} \frac{v_{ik}(Z_{ik} - x_i y_k)}{V(x_i y_k)} \cdot x_i y_k \cdot \xi_{ik,m} = \sum_{k=1}^{K} \frac{Z_{mk} - x_m y_k}{\mathrm{Var}(Z_{mk})} \cdot x_m y_k$$

für $1 \leq m \leq I$ und

$$0 = \frac{1}{\phi} \sum_{i=1}^{I} \sum_{k=1}^{K} \frac{v_{ik}(Z_{ik} - x_i y_k)}{V(x_i y_k)} \cdot x_i y_k \cdot \xi_{ik,m} = \sum_{i=1}^{I} \frac{Z_{i,m-I} - x_i y_{m-I}}{\mathrm{Var}(Z_{i,m-I})} \cdot x_i y_{m-I}$$

für $I + 1 \leq m \leq I + K$.

Im Poisson-Fall $\mathrm{Var}(Z_{ik}) = x_i y_k / v_{ik}$ ergeben sich die Marginalsummen-Gleichungen. Auch im Gamma-Fall $\mathrm{Var}(Z_{ik}) = (x_i y_k)^2 / (v_{ik} \alpha)$ ergeben sich die dieselben Gleichungen wie oben.

Bemerkungen

- Um eine invertierbare Informationsmatrix zu bekommen muss man wieder einen der Parameter x_i oder y_k fest wählen (z.B. $x_1 := 1$). Das wurde oben nicht berücksichtigt, wirkt sich aber nicht auf die Gleichungen aus.
- Bei GLMs vom Typ $v_n \operatorname{Var}(Z_n) = \phi V(\mu_n) = \phi \mu_n^\zeta$ mit $\zeta \in \{0, 1, 2, 3\}$ kann die Wahl von ζ, d.h. die Wahl der Verteilung der Z_n, mittels eines Plots von $\ln\left(v_n(Z_n - \widehat{\mu}_n)^2\right)$ gegen $\ln \widehat{\mu}_n$ plausibilisiert werden: Die Steigung der resultierenden Regressionsgeraden sollte gleich dem ζ der zugrunde gelegten Verteilungsannahme sein.
- Wenn man $\mu_n = \mathsf{E}(Z_n)$ und $\operatorname{Var}(Z_n)$ innerhalb jeder Risikogruppe n schätzen kann (z.B. aus Einzelrisikodaten oder aus mehreren Jahren), dann sollte man die Geeignetheit eines GLM und die Schätzwerte der Parameter ϕ and ζ aus der Regression $\ln\left(v_n \widehat{\operatorname{Var}}(Z_n)\right)$ versus $\ln \phi + \zeta \cdot \ln \widehat{\mu}_n$ entnehmen.

Inhalt

Problemstellung

Gegeben sei ein nominal skaliertes Risikomerkmal mit vielen Ausprägungen, z.B. geographische Regionen (PLZ, Zulassungsbezirke), Betriebsarten (Apotheke bis Zeitschriftenhandel), Autotypen.

Ziel:	Zusammenfassen von Ausprägungen zu Ausprägungsklassen mit ähnlichem Schadenbedarf/-satz.
Beachte:	Dies ist bei metrisch oder ordinal skalierten Merkmalen einfacher bzw. evtl. sogar überflüssig, wenn ein funktionaler Zusammenhang angenommen werden kann, z.B. zwischen Jahresfahrleistung und Schadenbedarf.
Problem:	Manche Ausprägungen können extrem schwach besetzt und daher z.B. zufällig schadenfrei sein (in allen Jahren). Diese müssen von fachmännischer Hand zugeordnet werden.

Problemstellung

Cluster-Analyse:	Generell ist die Aufgabe einer Cluster-Analyse die Zusammenfassung von Objekten zu Klassen/Clusters, wobei Objekte derselben Klasse möglichst ähnlich, Objekte verschiedener Klassen möglichst verschieden sein sollen.
Entscheidend:	Quantifizierung des Ähnlichkeitsbegriffs!
Problem:	Ein Durchrechnen aller Möglichkeiten von Klasseneinteilungen ist nicht möglich, da deren Anzahl viel zu groß ist (auch wenn man ein Optimalitätskriterium hätte). Daher induktives Vorgehen und heuristische Verfahren.

Bei uns sind die Objekte gleich den Ausprägungen (Gruppe der Risiken mit gleicher Merkmalsausprägung) und das Ähnlichkeitskriterium ist der Erwartungswert des Schadenbedarfs/-satzes.

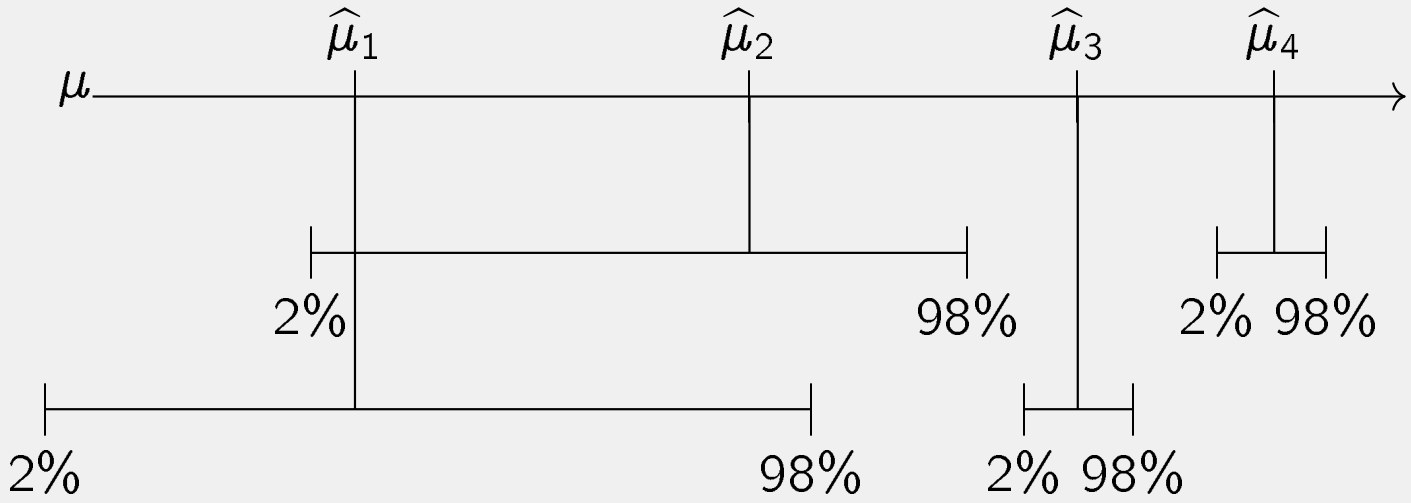
Problemstellung

Besonderheiten:

- Die Objekte sind verschieden groß und der Wert des Ähnlichkeitskriteriums ist nicht exakt bekannt, d.h. muss geschätzt werden und ist daher zufallsabhängig.
- Daher ist die Quantifizierung nicht durch die euklidische Distanz $|\widehat{\mu}_i - \widehat{\mu}_k|$ möglich, sondern durch Test auf Gleichheit der Erwartungswerte μ_i und μ_k.

Problemstellung

Warum nicht $|\widehat{\mu}_k - \widehat{\mu}_l|$ als Distanzmaß?



Zwar ist $|\widehat{\mu}_3 - \widehat{\mu}_4| < |\widehat{\mu}_1 - \widehat{\mu}_2|$ aber dennoch sind offenbar μ_3 und μ_4 signifikant verschieden, μ_1 und μ_2 dagegen nicht. Eine vernünftige Teststatistik berücksichtigt die unterschiedlichen Varianzen.

Agglomerative Klassenbildung

Es seinen I Ausprägungen eines Risikomerkmals gegeben (d.h. I Risikogruppen) jeweils mit J Beobachtungen Z_{ij} des Schadensatzes mit bekannten Volumina v_{ij}.

Gesucht sind K disjunkte Klassen $C_1, \cdots, C_K$ mit $\bigcup_{k=1}^{K} C_k = \{1, 2, \cdots, I\}$, $K \ll I$, die „besser“ sind als andere K Klassen.

Gamma-Modell

Wir verwenden das Gamma-Modell $Z_{ij} \sim \text{Gamma}\,(\mu_i, v_{ij}\alpha_i)$ und verwenden das Ähnlichkeits- bzw. Distanzmaß $d(i,k)$ zwischen zwei Ausprägungen i und k, das durch den Likelihood-Quotiententest

$$d(i,k) = 2 \cdot \ln \frac{L_1\left(\widehat{\mu}_i, \widehat{\alpha}_i, \widehat{\mu}_k, \widehat{\alpha}_k\right)}{L_0\left(\widehat{\mu}, \widehat{\alpha}\right)}$$

auf Gleichheit der Parameter $(\mu_i, \alpha_i) = (\mu_k, \alpha_k) = (\mu, \alpha)$ definiert wird. Genauer gilt

$$L_1 = \prod_{j=1}^{J} \left(g_{\widehat{\mu}_i, v_{ij}\widehat{\alpha}_i}(Z_{ij}) \cdot g_{\widehat{\mu}_k, v_{kj}\widehat{\alpha}_k}(Z_{kj}) \right) \quad \text{mit} \quad \widehat{\mu}_i = \frac{\sum_j v_{ij} Z_{ij}}{\sum_j v_{ij}}$$

und $\widehat{\alpha}_i$ so dass $\sum_{j=1}^{J} v_{ij} \left(\ln\left(\frac{v_{ij} Z_{ij} \widehat{\alpha}_i}{\widehat{\mu}_i} \right) - \Psi\left(v_{ij}\widehat{\alpha}_i\right) \right) = 0$, siehe oben.

Gamma-Modell

Analog für $\widehat{\mu}_k$ und $\widehat{\alpha}_k$. Ferner

$$L_0 = \prod_{j=1}^{J} \left(g_{\widehat{\mu}, v_{ij}\widehat{\alpha}}(Z_{ij}) \cdot g_{\widehat{\mu}, v_{kj}\widehat{\alpha}}(Z_{kj}) \right) \quad \text{mit} \quad \widehat{\mu} = \frac{\sum_j \left(v_{ij} Z_{ij} + v_{kj} Z_{kj} \right)}{\sum_j \left(v_{ij} + v_{kj} \right)}$$

und $\widehat{\alpha}$ so dass

$$\sum_{j=1}^{J} \left(v_{ij} \ln \left(\frac{v_{ij} Z_{ij} \widehat{\alpha}}{\widehat{\mu}} \right) - v_{ij} \Psi \left(v_{ij} \widehat{\alpha} \right) + v_{kj} \ln \left(\frac{v_{kj} Z_{kj} \widehat{\alpha}}{\widehat{\mu}} \right) - v_{kj} \Psi \left(v_{kj} \widehat{\alpha} \right) \right) = 0.$$

Gamma-Modell

Damit geht man agglomerativ vor, d.h.

- schrittweise Reduktion der Anzahl Ausprägungsklassen durch Fusion der Ausprägungen/Klassen mit der jeweils kleinsten Distanz $d(i,k)$
- Stopp, sobald $d(i,k) \sim \chi^2_{4-2}$ das zuvor festgelegte Signifikanzniveau übersteigt (d.h. der Test die Gleichheit der Verteilungsparameter ablehnt)

Beachte: Gilt $(\mu_i, \alpha_i) = (\mu_k, \alpha_k) = (\mu, \alpha)$, so ist

$$Z_{(ik),j} = \frac{v_{ij} Z_{ij} + v_{kj} Z_{kj}}{v_{ij} + v_{kj}} \sim \Gamma\left(\mu, \left(v_{ij} + v_{kj}\right)\alpha\right),$$

d.h. der Schadensatz der durch Fusion entstandenen Klasse ist wieder Gamma-verteilt.

Poisson-Modell

Betrachten wir nur die Schadenhäufigkeiten und modellieren

$$S_{ij} \sim \text{Poi}(v_{ij}\mu_i),$$

so vereinfachen sich die Rechnungen erheblich und die Testgröße (Distanzmaß) $d(i,k)$ kann explizit angegeben werden:

$$d(i,k) = 2\ln\frac{L_1(\widehat{\mu}_i,\widehat{\mu}_k)}{L_0(\widehat{\mu})} = 2S_{i+}\ln\frac{\widehat{\mu}_i}{\widehat{\mu}} + 2S_{k+}\ln\frac{\widehat{\mu}_k}{\widehat{\mu}} \sim \chi^2_1.$$

Normalverteilungs-Modell

Im Modell $Z_{ij} \sim \mathcal{N}(\mu_i, \sigma^2/v_{ij})$ mit einem bei allen Ausprägungen gleichen und *bekannten* σ^2 erhält man als LQ-Test

$$d(i,k) = \frac{(\widehat{\mu}_i - \widehat{\mu}_k)^2}{\frac{\sigma^2}{v_{i+}} + \frac{\sigma^2}{v_{k+}}}$$

wobei der überall gleiche Faktor σ^{-2} auch weggelassen werden kann, d.h. doch nicht bekannt sein muss.

Die so entstehende sog. *Ward-Distanz*

$$\sigma^2 d(i,k) = \frac{v_{i+} \cdot v_{k+}}{v_{i+} + v_{k+}} (\widehat{\mu}_i - \widehat{\mu}_k)^2$$

kann auch bei Daten aus nur einem Jahr ($J = 1$) angewandt werden und ist daher recht populär. Allerdings hat man dann kein Stopp-Kriterium.

Agglomerative Klassenbildung

Bemerkungen:

- Man kann beim erstmaligen Überschreiten des Signifikanzniveaus meist durch einen Umordnungsversuch noch etwas weiter kommen. Dass solche Umordnungen in der Regel möglich sind, zeigt den heuristischen Charakter des Vorgehens.
- Problematisch bei diesem Vorgehen ist die isolierte/separate Betrachtung einzelner Risikomerkmale. Werden zwei Ausprägungsklassen als signifikant unterschiedlich erkannt, so kann das durch andere Risikomerkmale bedingt sein.
- Bei der Bestimmung der optimalen Clusterzahl sind meist nicht vorrangig formale, sondern inhaltliche Kriterien von Bedeutung.

Inhalt

Problemstellung

In den meisten Versicherungsbranchen gibt es viele Risikomerkmale. Z.B. in KH:

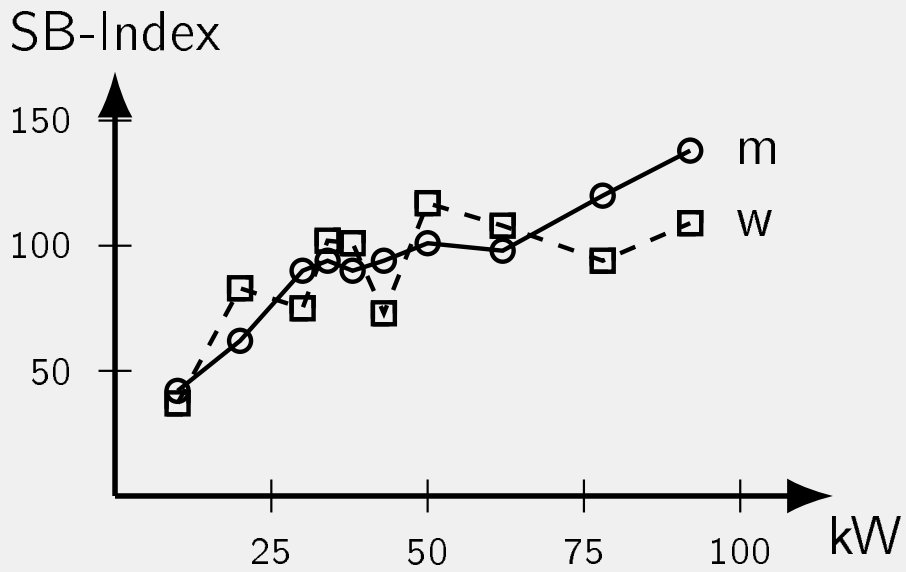
- Fahrzeug: kW, Typ, Gewicht, Alter
- Fahrgebiet: Zulassungsbezirk, Stadt/Land
- Fahrer: Alter, Beruf, Geschlecht, Familienstand, Anzahl Fahrer, Dauer des Führerscheinbesitzes, Garagenbesitzer, Nationalität
- Fahrintensität: Kilometerleistung, Nutzungsart (privat/beruflich)

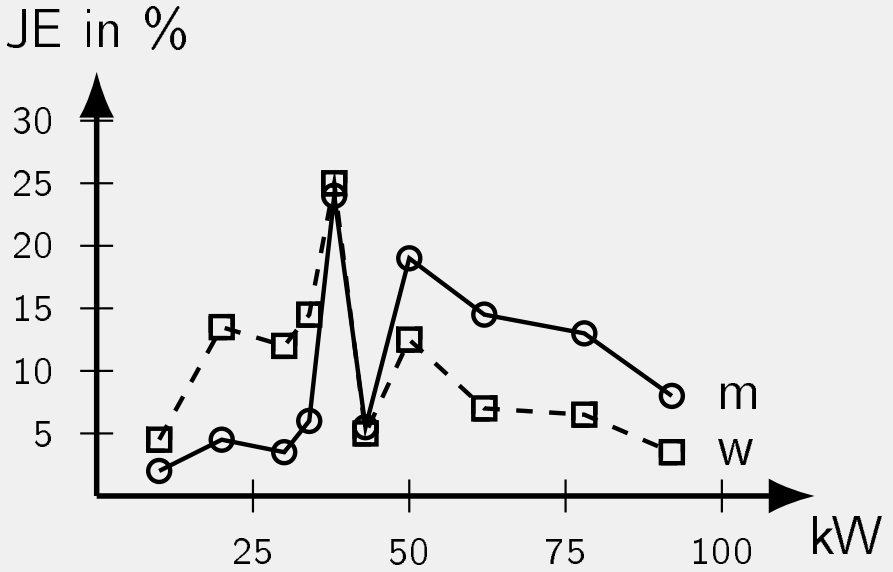
Aber es bestehen gegenseitige Abhängigkeiten, z.B. eine (alte) Stichprobe in KH:

Geschlecht	Anteil	relativer Schadenbedarf	mittlere Leistung
weiblich	25%	90%	40 kW
männlich	75%	103%	53 kW

Abhängigkeit des Schadenbedarfs vom Geschlecht

Tatsächlich ist für gleiche kW der Schadenbedarf für männlich und weiblich nicht signifikant unterschiedlich:





Analysen eines einzelnen Merkmals können irreführend sein. kW und Geschlecht sollten beispielsweise nicht gleichzeitig Tarifmerkmal sein. Ziel ist die Auswahl der effizientesten Risikomerkmale für einen Tarif.

Schrittweise Merkmalsauswahl

Wir nehmen an, dass bereits $t > 0$ Tarifmerkmale ausgewählt und ein GLM $\mathcal{M}_a$ angepasst wurde.

Nimmt man ein weiteres Risikomerkmal mit K Ausprägungsklassen hinzu und passt ein GLM $\mathcal{M}_b$ an, so ist bei bekannter Dispersion ϕ

$$\frac{D(\mathcal{M}_a) - D(\mathcal{M}_b)}{\phi}$$

χ^2-verteilt mit $K - 1$ Freiheitsgraden.

Hiermit kann man testen, ob das weitere Risikomerkmal signifikant ist.

Schrittweise Merkmalsauswahl

S Auswahl des signifikantesten Risikomerkmals. $\xrightarrow{\text{keines}}$ **Ende**

A Hinzunahme des gegenüber den derzeitigen TM signifikantesten RM (egal, ob über Signifikanz-Schranke oder nicht).

B Ist jedes TM (immer noch) signifikant gegenüber den jeweils anderen TM?

- ja → Ist diese Menge der TM erstmals entstanden?
  - ja → **A**
  - nein → **Ende**
- nein → Elimination des am wenigsten signifikanten TM. Ist dies das TM, das beim letzten **A**-Schritt hinzukam?
  - ja → **Ende**
  - nein → **B**

Schrittweise Merkmalsauswahl

Bemerkungen:

- Die Dispersion ϕ ist nicht bekannt und muss geschätzt werden. Der Schätzer hängt jedoch davon ab, welche Tarifmerkmale man in dem GLM verwendet. Für die Merkmalsauswahl sollte man den Parameter ϕ mit einem der Modelle schätzen und dann für die Analyse konstant lassen.
- Alternativ kann man das maximale Modell (d.h. das Modell mit allen Merkmalen) als Ausgangspunkt nehmen und dann sukzessive die am wenigsten signifikanten Merkmale aus dem Modell entfernen.
- Neben der formalen statistischen Analyse ist bei der Auswahl der Merkmale eine intensive Auseinandersetzung mit inhaltlichen Aspekten der Modelle bzw. einzelner Merkmale unabdingbar. Bei der Auswahl ist immer auch der gesunde Menschenverstand gefragt.