

11 Log-lineare Modelle

11.1 Einleitung

- a Abhängigkeiten zwischen zwei kategoriellen Variablen zu untersuchen, bringt ähnlich viel und wenig wie die Berechnung von einfachen Korrelationen zwischen kontinuierlichen Variablen oder die Berechnung einer einfachen Regression. Wenn eine kontinuierliche Zielgrösse von mehreren erklärenden Variablen abhängt, braucht man die multiple Regression – sonst ist die Gefahr von falschen Schlüssen gross. Das ist für kategorielle Variablen nicht anders; wir brauchen Modelle, die es erlauben, mehrere erklärende Faktoren gleichzeitig zu erfassen. Dies soll an einem Beispiel gezeigt werden.
- b **Beispiel Zulassung zum Studium.** Im Zuge der Kritik an den Universitäten stellten Frauen im Jahre 1973 fest, dass sie anteilmässig seltener zum Graduierten-Studium zugelassen wurden als die Männer. Dies zeigt sich deutlich in Tabelle 11.1.b (i). (Quelle: Bickel et. al, 1975, Science 187, 398-404. Es handelt sich um Daten der 6 grössten Departemente.)

Geschl.	Anzahlen			Prozente		
	zugel.	abgew.	Σ	zugel.	abgew.	Σ
w	557	1278	1835	30.4	69.6	100
m	1198	1493	2691	44.5	55.5	100
Σ	1755	2771	4526	38.8	61.2	100

Tabelle 11.1.b (i): Zulassungen zum Graduierten-Studium

Da Männer die Daten schliesslich publizierten, ist dies wohl nicht die ganze Story. Man kann vermuten, dass jedermann den Eindruck hatte, dass die Diskriminierung nicht in seinem Departement stattfindet. Und jeder hatte damit mehr oder weniger Recht, wie Tabelle 11.1.b(ii) zeigt!

Wie kann das sein? In keinem Departement scheinen die Frauen ernsthaft diskriminiert worden zu sein, in vier von sechs sogar bevorzugt, und trotzdem ergibt sich insgesamt eine klare Benachteiligung! Es wäre leicht, die Zahlen so zu verändern, dass innerhalb aller Departemente sogar eine Bevorzugung der Frauen festgestellt würde, ohne den Gesamtbefund zu verändern.

- c Zusammenhänge, die innerhalb von verschiedenen Gruppen vorhanden sind, können also scheinbar ins Gegenteil verkehrt werden, wenn die Gruppierung nicht berücksichtigt wird. Dieses Phänomen ist unter dem Namen **Simpson's Paradox** bekannt. Der Effekt kann nur zu Stande kommen, wenn die beiden Variablen, zwischen denen ein Zusammenhang untersucht wird, innerhalb der verschiedenen Gruppen deutlich verschiedene Verteilungen zeigen.

Dept.	Geschl.	Anzahlen			Prozente		
		zugel.	abgew.	Σ	zugel.	abgew.	Σ
A	w	89	19	108	82.4	17.6	100
	m	512	313	825	62.1	37.9	100
B	w	17	8	25	68.0	32.0	100
	m	353	207	560	63.0	37.0	100
C	w	202	391	593	34.1	65.9	100
	m	120	205	325	36.9	63.1	100
D	w	131	244	375	34.9	65.1	100
	m	138	279	417	33.1	66.9	100
E	w	94	299	393	23.9	76.1	100
	m	53	138	191	27.7	72.3	100
F	w	24	317	341	7.0	93.0	100
	m	22	351	373	5.9	94.1	100
Σ		1755	2771	4526	38.8	61.2	100

Tabelle 11.1.b (ii): Zulassungen zum Studium, aufgeteilt nach Departementen

Bei quantitativen Variablen kann ein analoges Phänomen auftreten: Innerhalb der Gruppen kann die Korrelation ganz anders aussehen als bei der Betrachtung aller Gruppen zusammen. Man spricht dort von „**Inhomogenitäts-Korrelation**“ (Stahel (2002), 3.4.c). Das Problem macht den Einbezug möglichst aller erklärenden Variablen nötig, wenn man indirekte Effekte von erklärenden Variablen auf Zielgrößen vermeiden will (siehe Rg1, Abschnitt 3.3).

Analog zu jenen Überlegungen zeigt Simpson's Paradox die Notwendigkeit, auch Modelle für die Zusammenhänge zwischen mehreren Faktoren zu entwickeln. Zunächst befassen wir uns nochmals mit dem Fall von zwei Variablen, um die Begriffe aufzubauen.

11.2 Log-lineare Modelle für zwei Faktoren

- a Um log-lineare Modelle zu beschreiben, betrachten wir noch einmal die **Poisson-Regression** (9.1). Die Zielgröße Y_i war eine Anzahl, und das Modell lautete

$$Y_i \sim \mathcal{P}\langle \lambda_i \rangle, \quad \lambda_i = \mathcal{E}\langle Y_i \rangle, \quad \log \langle \lambda_i \rangle = \eta_i = \underline{x}_i^T \underline{\beta}.$$

Das Modell der Poisson-Regression lässt sich auf Kontingenztafeln anwenden – und damit steht die gesamte Methodik inklusive Schätzung, Residuenanalyse und Modelldiagnose zur Verfügung. Wir haben darauf hingewiesen (7.2.h), dass man die Anzahlen N_{hk} zunächst als Poisson-Variablen auffassen und später die Bedingungen über Randsummen oder Gesamtzahl der Beobachtungen berücksichtigen kann. Es gilt also zunächst

$$N_{hk} \sim \mathcal{P}\langle \lambda_{hk} \rangle.$$

Falls die beiden Faktoren A und B unabhängig sind, gilt (siehe BUCH 7.3.a)

$$\begin{aligned} \lambda_{hk} &= n\pi_{h+}\pi_{+k} \\ \eta_{hk} = \log \langle \lambda_{hk} \rangle &= \log \langle n \rangle + \log \langle \pi_{h+} \rangle + \log \langle \pi_{+k} \rangle. \end{aligned}$$

was man mit anderen Bezeichnungen als

$$\eta_{hk} = \mu + \alpha_h + \beta_k$$

schreiben kann. Das sieht aus wie das Modell einer **Zweiweg-Varianzanalyse ohne Wechselwirkungen** – bis auf den Fehlerterm. Dieser wurde, wie bei den verallgemeinerten linearen Modellen üblich, vermieden, indem man das Modell für den Erwartungswert der Zielgrösse statt für diese selber formuliert.

Die Analogie der hier zu besprechenden Modelle zu denen der faktoriellen Varianzanalyse wird sich auch für mehr als zwei Faktoren als sehr nützlich erweisen.

- b Die Summe aller Wahrscheinlichkeiten für jeden Faktor muss 1 sein. Das ergibt **Nebenbedingungen** für die Parameter. In der Varianzanalyse haben wir solche einführen müssen, damit die Haupteffekte α_h und β_k eindeutig definiert waren. Die gebräuchlichsten waren

$$(a) \quad \sum_h \alpha_h = 0$$

$$(b) \quad \alpha_1 = 0$$

(und entsprechend für die β_k). Im vorliegenden Modell führt $\alpha_h = \log \langle \pi_{h+} \rangle$ wegen $\sum_h \pi_{h+} = 1$ zu

$$(c) \quad \sum_h \exp \langle \alpha_h \rangle = 1.$$

Das ist zwar eine naheliegende Art, die α_h eindeutig zu machen, aber sie ist mathematisch komplizierter als die ersten beiden Bedingungen und wird deshalb oft durch eine von diesen ersetzt. Das log-lineare Modell, das die Unabhängigkeit zwischen den Faktoren A und B beschreibt, besitzt wegen diesen Nebenbedingungen $1 + (r - 1) + (s - 1)$ freie Parameter (wobei r und s die Anzahlen der Niveaus von A und B sind).

- c Die **Haupteffekte** charakterisieren die Wahrscheinlichkeiten π_{h+} (respektive π_{+k}), also die Randverteilungen. Wir haben gesehen (7.2.f, 7.2.h), dass die Randtotale n_{h+} , die ja die Randwahrscheinlichkeiten bestimmen, oft vorgegeben sind und, auch wenn sie sich zufällig ergeben, kaum je interessieren. Deshalb hat die in der Varianzanalyse so bedeutende Nullhypothese, dass keine Haupteffekte vorhanden seien, also $\alpha_h = 0$ oder $\beta_k = 0$, für die log-linearen Modelle kaum je eine Bedeutung. Sie würde bedeuten, dass alle Stufen des Faktors A gleiche Wahrscheinlichkeit $1/r$ hätten. (Bei Nebenbedingung (c) können die α_h nicht 0 sein. Die Nullhypothese müsste dann lauten, dass alle α_h gleich, nämlich $= -\log \langle r \rangle$ seien.)
- d Das Modell 11.2.a enthält also nur uninteressante Parameter. Wenn wir auch die Wechselwirkungen einführen,

$$\log \langle \lambda_{hk} \rangle = \mu + \alpha_h + \beta_k + (\alpha\beta)_{hk},$$

so erhalten wir das **maximale** oder **gesättigte Modell** (*saturated model*), das heisst, das Modell, das so viele freie Parameter wie Beobachtungen hat, so dass die Beobachtungen durch die geschätzten Parameter perfekt nachgebildet werden, wie wenn es keine zufälligen Abweichungen gäbe, vergleiche 9.3.g. (Die obige Gleichung hat zunächst sogar mehr Parameter als Beobachtungen; durch die erwähnten Nebenbedingungen für die Haupteffekte und ebensolche für die Wechselwirkungen wird das ausgeglichen.)

So ergibt sich aus der Theorie der verallgemeinerten linearen Modelle der folgende Test für die Unabhängigkeit der Faktoren A und B , also für das „Haupteffekt-Modell“: Die Testgrösse

$$D = 2 \cdot (\ell \langle \ell^{(M)} \rangle - \ell \langle \text{Haupteffektmodell} \rangle)$$

ist unter der Nullhypothese „ $(\alpha\beta)_{hk} = 0$ für alle h und k “ chiquadrat-verteilt mit $(r - 1)(s - 1)$ Freiheitsgraden.

- e Die **Testgrösse** lässt sich wie im allgemeinen Fall auch schreiben als doppelte Quadratsumme, deren Terme als quadrierte Devianz-Residuen identifiziert werden,

$$\begin{aligned} D &= 2 \sum_{h,k} d_{hk} = 2 \sum_{h,k} \left(R_{hk}^{(d)} \right)^2 \\ d_{hk} &= N_{hk} \log \left\langle N_{hk} / \hat{\lambda}_{hk} \right\rangle - N_{hk} + \hat{\lambda}_{hk} \\ R_{hk}^{(d)} &= \text{sign} \left\langle N_{hk} - \hat{\lambda}_{hk} \right\rangle \sqrt{d_{hk}} \end{aligned}$$

- f Die Testgrösse lässt sich auch vereinfachen zu

$$D = 2 \sum_{h,k} N_{hk} \log \left\langle N_{hk} / \hat{\lambda}_{hk} \right\rangle .$$

Der Test mit dieser Testgrösse ist nicht genau identisch mit dem Chiquadrat-Test aus 7.3.e, aber er liefert „fast immer“ die gleiche Antwort (und ist asymptotisch äquivalent). Er wurde lange vor der Entwicklung der Theorie der verallgemeinerten linearen Modelle im Zusammenhang mit der informationstheoretischen Sichtweise der Statistik erfunden und trägt auch den Namen **G-Test** (vergleiche Sachs (2004), Abschnitt 461).

- g Eine **Bemerkung**: In der Zweiweg-Varianzanalyse ohne wiederholte Beobachtungen mussten die Wechselwirkungen als Zufallsfehler interpretiert werden, und mit deren Hilfe konnten Hypothesen über die Haupteffekte getestet werden. (Allgemein gelangt man von gesättigten Modellen zu brauchbaren Zufallsmodellen, indem man geeignete Terme als zufällig behandelt.) Es war aber nicht (oder nur teilweise, mit Tricks) möglich, eine Hypothese über die Wechselwirkungen zu testen.

Bei den hier verwendeten **Häufigkeitsdaten** (vergleiche 7.1.g) sind wir hier, wie bereits bei der Poisson-Regression, besser dran: Da die Streuung einer Poisson-verteiltern Zufallsvariablen durch ihren Erwartungswert bereits gegeben ist (Varianz = Erwartungswert = λ), kann man testen, ob die Anzahlen N_{hk} zu stark von den Werten $\hat{\lambda}_{hk}$ abweichen, die man für das angepasste Modell erhält. Man kann also, wie oben besprochen, die Signifikanz der Wechselwirkungen testen. Das ist die Grundlage des gerade beschriebenen Devianz-Tests wie auch des Chiquadrat-Tests für Unabhängigkeit.

- h Tabelle 11.2.h zeigt eine Computer-Ausgabe für das **Beispiel der Umwelt-Umfrage** (7.1.c). Dabei wurden nur die Haupteffekte der beiden Faktoren A (Beeinträchtigung) und B (Schulbildung) ins Modell geschrieben (vergleiche 11.2.a). Die Residuen-Devianz wird benützt, um zu testen, ob die geschätzten Wechselwirkungen nur die unter dem Haupteffektmodell erwartete Streuung zeigen, also mit der Nullhypothese der Unabhängigkeit verträglich sind. Der Test zeigt Signifikanz. Die Testgrösse $T = 111.36$ stimmt übrigens annähernd mit der Testgrösse des Chiquadrat-Tests, $T = 110.26$, überein. Die Tabelle enthält Angaben zu den Haupteffekten, also lauter unnütze Information! Die einzige nützliche Zahl ist der Wert der Testgrösse „Residual Deviance“.

```
Call: glm(formula = Freq ~ Schule + Beeintr, family = poisson,
          data = t.dt)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	5.03154	0.06098	82.514	< 2e-16	***
SchuleLehre	0.84462	0.06674	12.656	< 2e-16	***
Schuleohne.Abi	0.17136	0.07576	2.262	0.0237	*
SchuleAbitur	-0.42433	0.08875	-4.781	1.74e-06	***
SchuleStudium	-0.70885	0.09718	-7.294	3.01e-13	***
Beeintretwas	-0.42292	0.05398	-7.835	4.71e-15	***
Beeintrziemlich	-1.17033	0.06979	-16.769	< 2e-16	***
Beeintrsehr	-2.03765	0.10001	-20.374	< 2e-16	***

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 1429.15 on 19 degrees of freedom
 Residual deviance: 111.36 on 12 degrees of freedom
 AIC: 246.50

Tabelle 11.2.h: Resultate des Haupteffekt-Modells im Beispiel der Umwelt-Umfrage

i **Signifikante Wechselwirkungen** lassen sich auf zwei Arten interpretieren:

- Die Erwartungswerte $\mathcal{E}\langle N_{hk} \rangle$ sind durch das Haupteffekt-Modell nicht richtig erfasst, das heisst, die Faktoren A und B sind nicht unabhängig. Im Beispiel ist dies die naheliegende und gewünschte Interpretation.
- Es könnte aber auch sein, dass die Erwartungswerte dem Modell der Unabhängigkeit von A und B genügen, aber die Streuung der N_{hk} grösser ist, als es die Poisson-Annahme ausdrückt. Es wäre beispielsweise möglich, dass die Einzelbeobachtungen abhängig sind. In einer Umfrage könnten oft mehrere Mitglieder der gleichen Familie befragt worden sein; diese würden wohl tendenziell ähnliche Meinungen und ähnliche Schulbildung haben. Dies würde zu vergrösserter Streuung (overdispersion) führen und könnte der Grund sein für den signifikanten Wert der Residual Deviance.

Die beiden Interpretationen lassen sich normalerweise nicht unterscheiden. Es ist also wichtig, dass die Unabhängigkeit der Beobachtungen durch die Datenaufnahme sichergestellt wird, damit man die zweite Möglichkeit ausschliessen kann.

- j Die Wechselwirkungsterme $(\alpha\beta)_{hk}$ zeigen einen engen Zusammenhang mit den **Doppelverhältnissen**: Es gilt gemäss 11.2.a

$$\log \langle \pi_{hk} \rangle = \log \langle \lambda_{hk} \rangle - \log \langle n \rangle = \mu + \alpha_h + \beta_k + (\alpha\beta)_{hk} - \log \langle n \rangle$$

und deshalb

$$\begin{aligned} \log \langle \theta_{hk,h'k'} \rangle &= \log \left\langle \frac{P \langle B = k \mid A = h \rangle}{P \langle B = k' \mid A = h \rangle} \middle/ \frac{P \langle B = k \mid A = h' \rangle}{P \langle B = k' \mid A = h' \rangle} \right\rangle \\ &= \log \langle \pi_{hk} \rangle - \log \langle \pi_{hk'} \rangle - \left(\log \langle \pi_{h'k} \rangle - \log \langle \pi_{h'k'} \rangle \right) \\ &= (\alpha\beta)_{hk} + (\alpha\beta)_{h'k'} - ((\alpha\beta)_{h'k} + (\alpha\beta)_{hk'}) , \end{aligned}$$

da sich alle Haupteffekte, μ und $\log \langle n \rangle$ wegheben. Die odds ratios werden allein durch die Wechselwirkungen bestimmt.

In einer 2×2 -Tafel gilt, wenn die Nebenbedingungen $\sum_h (\alpha\beta)_{hk} = 0$ und $\sum_k (\alpha\beta)_{hk} = 0$ festgelegt wurden, $(\alpha\beta)_{11} = -(\alpha\beta)_{12} = -(\alpha\beta)_{21} = (\alpha\beta)_{22}$ und damit wird $\log \langle \theta \rangle = 4(\alpha\beta)_{11}$.

- k Im Beispiel und in vielen Anwendungen spielt einer der Faktoren die Rolle einer Antwort- oder Zielgrösse auf der Ebene der Beobachtungs-Einheiten (vergleiche 7.1.h, 7.2.d). Die **Zielgrösse** N_{hk} der Poisson-Regression hat dagegen nur **technische Bedeutung**; sie ist auf der Ebene der Häufigkeitsdaten, also der zusammengefassten Beobachtungen, definiert.
- l **Daten-Eingabe.** Dem entsprechend müssen die Daten der Funktion `glm` in der Form eingegeben werden, die in 7.3.p angegeben wurde: In jeder Zeile von `t.dt` steht ein Wert h des Faktors A (Schule), ein Wert k des Faktors B (Beeinträchtigung) und die Anzahl Y_{hk} , die dieser Kombination entspricht, in der Spalte `Freq`. Tabelle 11.2.1 macht diese Anordnung für das Beispiel klar. Wenn die Daten in Form der ursprünglichen Beobachtungen vorhanden sind, gibt es normalerweise Funktionen, die die Zusammenfassung besorgen. In `S` geht das mit `data.frame(table(A,B))`.

	Schule	Beeintr	Freq
1	ungelernt	nicht	196
2	Lehre	nicht	410
3	ohne.Abi	nicht	152
	
6	ungelernt	etwas	73
7	Lehre	etwas	224
	
19	Abitur	sehr	16
20	Studium	sehr	17

Tabelle 11.2.1: Daten des Beispiels der Umwelt-Umfrage

Diese Form erlaubt es auch, die Anpassung eines multinomialen Regressionsmodells mit der Poisson-Regression vorzunehmen, wie es in 10.1.h angedeutet wurde.

- m Dieser Abschnitt hat wenig neue Einsichten gebracht, vielmehr haben wir verschiedene Aspekte bei der Analyse kategorialer Daten mit einem Modell verknüpft. Wir haben den Test auf Unabhängigkeit der beiden Faktoren im Jargon der Poisson-Regression besprochen – das ist komplizierter als die Diskussion im Rahmen der Kontingenztafeln. Es bildet aber die Basis für die Analyse höher-dimensionaler Kontingenztafeln, für die die „Maschinerie“ der log-linearen Modelle tiefere Einsichten erlaubt.

11.3 Log-lineare Modelle für mehr als zwei Faktoren

- a **Beispiel Umwelt-Umfrage.** In der Umfrage wurde, wie erwähnt, auch die Frage gestellt, ob der Einzelne viel zum Umweltschutz beitragen könne, ob die Lösung des Problems von staatlicher Seite kommen müsse, oder ob beides nötig sei. Tabelle 11.3.a zeigt den Zusammenhang dieses Faktors „Hauptverantwortung“ (C) mit der Beeinträchtigung B , aufgeschlüsselt nach der Schulbildung (A).

Beeintr.	Hauptverantwortung			
	Einzel.	Staat	beide	total
ungelernt				
p-Wert 0.00230				
nicht	81	110	19	210
etwas	38	36	9	83
ziemlich	23	9	6	38
sehr	12	3	5	20
total	154	158	39	351
Lehre				
p-Wert 4.57e-06				
nicht	206	193	33	432
etwas	150	67	28	245
ziemlich	58	19	8	85
sehr	22	7	6	35
total	436	286	75	797
ohne.Abi				
p-Wert 0.0696				
nicht	86	66	17	169
etwas	89	40	17	146
ziemlich	43	22	9	74
sehr	14	8	8	30
total	232	136	51	419

Beeintr.	Hauptverantwortung			
	Einzel.	Staat	beide	total
Abitur				
p-Wert 0.468				
nicht	41	24	14	79
etwas	51	17	24	92
ziemlich	25	16	13	54
sehr	12	6	3	21
total	129	63	54	246
Studium				
p-Wert 0.0668				
nicht	19	19	7	45
etwas	39	14	14	67
ziemlich	27	15	6	48
sehr	5	8	6	19
total	90	56	33	179

Tabelle 11.3.a: Ergebnisse für drei Fragen im Beispiel der Umwelt-Umfrage, mit p-Werten für den Test auf Unabhängigkeit der Beeinträchtigung und der Hauptverantwortung, gegeben die Schulstufe

- b Wenn, wie im Beispiel, drei Faktoren A , B und C vorliegen, so lautet das **gesättigte Modell**

$$\eta_{hkl} = \log \langle \lambda_{hkl} \rangle = \mu + \alpha_h + \beta_k + \gamma_l + (\alpha\beta)_{hk} + (\beta\gamma)_{kl} + (\alpha\gamma)_{hl} + (\alpha\beta\gamma)_{hkl}.$$

Schätzungen und Tests werden wie früher berechnet. Das Weglassen verschiedener Terme im gesättigten Modell führt zu sinnvollen und weniger sinnvollen Untermodellen, aus denen in der Anwendung ein geeignetes ausgewählt werden soll.

- c **Vollständige Unabhängigkeit** (A, B, C): $\eta_{hkl} = \mu + \alpha_h + \beta_k + \gamma_l$. Es bleiben wie im zweidimensionalen Fall nur die uninteressanten Haupteffekte im Modell. Dieses Modell bildet die einfachste Nullhypothese.
- d **Unabhängige Variablen-Gruppen** (AB, C): Das Modell $\eta_{hkl} = \mu + \alpha_h + \beta_k + \gamma_l + (\alpha\beta)_{hk}$ bedeutet, dass Faktor C unabhängig ist von (der gemeinsamen Verteilung von) A und B . Im Beispiel wäre dann die Ansicht über die Hauptverantwortung als unabhängig von der (Kombination von) Schulbildung und Beeinträchtigung vorausgesetzt, aber die Schulbildung und die Beeinträchtigung könnten zusammenhängen.

- e **Bedingte Unabhängigkeit** (AB, AC): Das Modell $\eta_{hkl} = \mu + \alpha_h + \beta_k + \gamma_\ell + (\alpha\beta)_{hk} + (\alpha\gamma)_{h\ell}$ bedeutet, dass die Faktoren B und C voneinander unabhängig sind, wenn A gegeben ist. Anders gesagt: Die bedingte gemeinsame Verteilung von B und C , gegeben A , zeigt Unabhängigkeit.

Im Beispiel wäre dann für jede Schulbildungsklasse die Meinung zur Hauptverantwortung unabhängig von der Beeinträchtigung durch Umwelteinflüsse.

- f **Partieller Zusammenhang** (AB, AC, BC) (*partial association*): Im Modell $\eta_{hkl} = \mu + \alpha_h + \beta_k + \gamma_\ell + (\alpha\beta)_{hk} + (\beta\gamma)_{k\ell} + (\alpha\gamma)_{h\ell}$ fehlt nur die dreifache Wechselwirkung.
- g Wie bereits erwähnt (7.1.h), kann man von der Problemstellung her meistens eine **Antwortgrösse** C angeben, deren Abhängigkeit von den erklärenden Faktoren man studieren will. In den Modellen werden dann die zweifachen Wechselwirkungen $(\alpha\gamma)_{h\ell}$ und $(\beta\gamma)_{k\ell}$ als Einfluss von A respektive B auf C interpretiert. Die Wechselwirkung $(\alpha\beta)_{hk}$ ist dann nicht von Interesse – sie entspricht der Korrelation von erklärenden Variablen in der gewöhnlichen Regression.

Im **Beispiel** betrachten wir zunächst die Zuteilung der Hauptverantwortung als Antwortfaktor und sowohl Schulbildung als auch Beeinträchtigung als erklärende. Den Zusammenhang zwischen Schulbildung und Beeinträchtigung haben wir bereits untersucht (7.2.d). Gehen wir für den Moment davon aus, dass die Beeinträchtigung die objektiven Gegebenheit am Wohnort der Befragten erfasst – was bei der verwendeten Erhebungsmethode, einer Befragung der Betroffenen, kaum gerechtfertigt ist. Dann erfasst ein Modell, das die Hauptverantwortung als Funktion der Schulbildung und der Beeinträchtigung beschreibt, die direkten Einflüsse dieser erklärenden Faktoren, unter Ausschaltung der jeweils anderen Grösse. (Eine subjektiv empfundene Beeinträchtigung ist etwas weniger gut als erklärende Variable geeignet, da sie mit der Antwortgrösse in Form eines wechselseitigen Einflusses zusammenhängen könnte.)

- h Im **Beispiel** wurde das Modell mit Haupteffekten und zweifachen Wechselwirkungen, aber ohne die dreifachen, angepasst. Es ergab sich eine Residuen-Devianz von 28.4 bei 24 Freiheitsgraden, was auf eine gute Anpassung schliessen lässt (P-Wert 0.24) und die Unterdrückung der dreifachen Wechselwirkungen rechtfertigt.

		RSS	Sum of Sq	Df	p.value
	„volles“ Modell	27.375	NA	NA	NA
ohne	Schule:Beeintr	130.011	102.636	12	0
ohne	Schule:Hauptv	64.781	37.406	8	0
ohne	Beeintr:Hauptv	83.846	56.471	6	0

Tabelle 11.3.h: Tests für das Weglassen der zweifachen Wechselwirkungen im Beispiel der Umwelt-Umfrage

Tabelle 11.3.h zeigt die Tests für das Weglassen der zweifachen Wechselwirkungen. Alle zweifachen Wechselwirkungen sind signifikant. Die erste davon ist die Wechselwirkung zwischen den erklärenden Faktoren. Die andern beiden bedeuten, dass beide erklärenden Grössen einen Einfluss auf den Antwortfaktor haben.

- i Das gesättigte Modell enthält die **dreifache Wechselwirkung**. Wenn sie sich als signifikant erweist, wirken die erklärenden Faktoren A und B nicht additiv im Sinne der log-linearen Modelle. Wie in der Varianzanalyse erschwert dies die Deutung der Einflüsse. Eine signifikante dreifache Wechselwirkung kann aber auch als Folge von übermässiger Streuung entstehen. Im Beispiel liegt dieser kompliziertere Fall, wie erwähnt, nicht vor.

j **Mehrere Antwortfaktoren.** Da die Variable „Beeinträchtigung“ nicht unbedingt eine „objektive“ Gegebenheit bedeutet, sondern eine subjektive Meinung erfasst, macht es Sinn, sie ebenfalls als Antwortgrösse zu studieren. Dann sind B und C Antwortgrößen und A die einzige erklärende Variable. Neben den Abhängigkeiten jeder der beiden Antwortgrößen von A kann jetzt auch von Interesse sein, Modell 11.3.e zu prüfen, das die bedingte Unabhängigkeit von B und C , gegeben A , formuliert. Würde es passen, während gleichzeitig eine einfache Kreuztabelle von B und C signifikante Abhängigkeit zeigt, dann könnte diese Abhängigkeit als „durch die Inhomogenität von A (der Schulbildungen) verursacht“ interpretiert werden. Im Beispiel ist das nicht der Fall.

Das Ergebnis lässt sich also so formulieren: Sowohl die Beeinträchtigung als auch die Zuweisung der Hauptverantwortung hängen von der Schulbildung ab, und diese beiden Faktoren hängen (innerhalb der Bildungsklassen) zusammen (keine bedingte Unabhängigkeit).

k Wie können die geschätzten Parameter **interpretiert** werden? Tabelle 11.3.k zeigt die geschätzten Effekte im Beispiel in der Anordnung von zweidimensionalen Kreuztabellen.

Schulbildung	Haupteff.	Beeinträchtigung			
		nicht	etwas	ziemlich	sehr
Haupteff.	$\hat{\mu} = 2.983$	$\hat{\beta}_1 = 0.682$	$\hat{\beta}_2 = 0.496$	$\hat{\beta}_3 = -0.205$	$\hat{\beta}_4 = -0.974$
ungelernt	$\hat{\alpha}_1 = -0.131$	0.483	-0.169	-0.274	-0.041
Lehre	$\hat{\alpha}_2 = 0.588$	0.450	0.106	-0.282	-0.273
ohne.Abi	$\hat{\alpha}_3 = 0.197$	-0.041	0.025	0.019	-0.004
Abitur	$\hat{\alpha}_4 = -0.190$	-0.280	0.036	0.187	0.057
Studium	$\hat{\alpha}_5 = -0.464$	-0.612	0.002	0.349	0.261

Schulbildung	Haupteff.	Hauptverantwortung		
		Einzelne	Staat	beide
Haupteff.	$\hat{\mu} = 2.983$	$\hat{\gamma}_1 = 0.593$	$\hat{\gamma}_2 = 0.038$	$\hat{\gamma}_3 = -0.631$
ungelernt	$\hat{\alpha}_1 = -0.131$	-0.096	0.246	-0.150
Lehre	$\hat{\alpha}_2 = 0.588$	0.171	0.094	-0.265
ohne.Abi	$\hat{\alpha}_3 = 0.197$	0.104	0.005	-0.109
Abitur	$\hat{\alpha}_4 = -0.190$	-0.064	-0.297	0.361
Studium	$\hat{\alpha}_5 = -0.464$	-0.115	-0.048	0.163

Beeinträchtigung	Haupteff.	Hauptverantwortung		
		Einzelne	Staat	beide
Haupteff.	$\hat{\mu} = 2.983$	$\hat{\gamma}_1 = 0.593$	$\hat{\gamma}_2 = 0.038$	$\hat{\gamma}_3 = -0.631$
nicht	$\hat{\beta}_1 = 0.682$	-0.109	0.389	-0.281
etwas	$\hat{\beta}_2 = 0.496$	0.075	-0.087	0.012
ziemlich	$\hat{\beta}_3 = -0.205$	0.127	-0.067	-0.060
sehr	$\hat{\beta}_4 = -0.974$	-0.093	-0.235	0.328

Tabelle 11.3.k: Geschätzte Haupteffekte und Wechselwirkungen im Beispiel der Umwelt-Umfrage. Restriktionen sind $\sum \alpha_k = \sum \beta_h = \sum \gamma_j = 0$.

Die Haupteffekte der erklärenden Variablen sind nicht von Bedeutung; sie charakterisieren lediglich deren Verteilungen, beispielsweise die Häufigkeiten der Personen mit verschiedenen Stufen der Schulbildung. **Haupteffekte der Antwortfaktoren** können von Interesse sein. Dabei führen, wie in der Varianzanalyse, Differenzen zu interpretierbaren Größen,

nämlich zu Odds. Es gilt $\log \langle \hat{\pi}_\ell / \hat{\pi}_{\ell'} \rangle = \hat{\gamma}_\ell - \hat{\gamma}_{\ell'}$. Für die Zuweisung der Hauptverantwortung sind die odds für die Einzelnen gegenüber dem Staat gleich $\exp \langle 0.593 - 0.038 \rangle = 1.742$.

Das Hauptinteresse gilt den Wechselwirkungen zwischen dem Antwortfaktor und den erklärenden Faktoren. Die Zweifach-Wechselwirkungen $(\alpha\gamma)_{h\ell}$ führen zu log odds ratios,

$$\log \left\langle \frac{\pi_{h\ell}}{\pi_{h\ell'}} \middle/ \frac{\pi_{h'\ell}}{\pi_{h'\ell'}} \right\rangle = (\alpha\gamma)_{h\ell} - (\alpha\gamma)_{h\ell'} - (\alpha\gamma)_{h'\ell} + (\alpha\gamma)_{h'\ell'}$$

Die odds für individuelle Verantwortung ($\ell = 1$) gegenüber staatlicher Verantwortung ($\ell' = 2$) verändern sich, wenn man Ungelernte ($h = 1$) mit Studierenden ($h' = 4$) vergleicht, um einen Faktor von $\exp \langle -0.096 - 0.246 - (-0.115) + (-0.048) \rangle = \exp \langle -0.275 \rangle = 0.760$; Studierende geben dem Einzelnen mehr Gewicht.

- 1* !!! multinomiale Regression: Die Parameter-Schätzung kann mit den Programmen für Verallgemeinerte Lineare Modelle erfolgen, wenn keine Funktion für multinomiale Regression zur Verfügung steht. Dabei müssen die Daten in einer zunächst unerwarteten Form eingegeben werden: Aus jeder Beobachtung Y_i machen wir k^* Beobachtungen Y_{ik}^* nach der Regel $Y_{ik}^* = 1$, falls $Y_i = k$ und $=0$ sonst – also wie dummy Variable für Faktoren, aber die Y_{ik}^* werden alle als Beobachtungen der Zielvariablen Y^* unter einander geschrieben. Gleichzeitig führt man als erklärende Variable einen Faktor $X^{(Y)}$ ein, dessen geschätzte Haupteffekte nicht von Interesse sein werden. Für die Zeile i, k der Datenmatrix (mit Y_{ik}^* als Wert der Zielgrösse) ist $X^{(Y)} = k$. Die Werte $X_i^{(j)}$ der übrigen erklärenden Variablen werden k^* Mal wiederholt. Mit diesen $n \cdot k^*$ „Beobachtungen“ führt man nun eine Poisson-Regression durch.

11.4 Vorgehen bei der Anpassung log-linearer Modelle

- a Aus allen vorangehenden Überlegungen ergibt sich ein **Vorgehen**, das auch auf mehr als drei Faktoren anwendbar ist:
Bei der Anwendung log-linearer Modelle ist es wichtig, sich vorher zu überlegen, was das Ziel der Analyse ist. Welches sind die Antwortfaktoren, welches die erklärenden? Sind die Antwortfaktoren ordinal? Dann ist wohl eine Analyse mit den dafür geeigneten Modellen (kumulative Logits) vorzuziehen.
- b Entwickeln eines Modells: Ausgehend vom gesättigten Modell lässt man schrittweise unwichtige Wechselwirkungsterme weg. Man fängt bei den höchsten Wechselwirkungen an. Man lässt aber alle Terme im Modell, die nur die erklärenden Variablen enthalten (um Inhomogenitäts-Korrelationen zu vermeiden!), ebenso alle Haupteffekte, auch diejenigen der Antwortfaktoren. Man versucht zwei Ziele zu erreichen:
- Das Modell soll komplex genug sein, um gute Anpassung zu erreichen, aber nicht komplexer als nötig, damit eine Überanpassung vermieden wird.
 - Das Modell soll einfach zu interpretieren sein.

- c Die **Interpretation** läuft analog zur Varianzanalyse, aber mit verschobener Bedeutung:
- Die Haupteffekte sind bedeutungslos.
 - Die zweifachen Wechselwirkungen zwischen Antwortfaktoren und erklärenden Faktoren haben die Bedeutung der Haupteffekte in der Varianzanalyse.
 - Wechselwirkungen zwischen erklärenden Grössen sind nicht sehr bedeutungsvoll. Wenn sie stark sind, können sie zu Interpretationsschwierigkeiten führen, ähnlich wie Unbalanciertheit in der Varianzanalyse oder Kollinearitäten in der Regression.
 - Wechselwirkungen zwischen Antwortfaktoren entsprechen Korrelationen zwischen mehreren Zielgrössen in der Varianzanalyse oder Regression – genauer: Korrelationen zwischen den Fehlertermen. Solche Abhängigkeiten sind ein Thema der multivariaten Varianzanalyse respektive Regression. Sie wurden im NDK nicht besprochen.
- Die dreifachen Wechselwirkungen, die einen Antwortfaktor und zwei erklärende Faktoren umfassen, entsprechen den zweifachen Wechselwirkungen in der Varianzanalyse.

11.5 Quantitative Variable

- a In Umfragen interessiert häufig die Abhängigkeit der Antworten vom Alter der Befragten. Man teilt das Alter üblicherweise in Kategorien ein und behandelt es wie eine kategorielle erklärende Grösse. Damit ist ein doppelter Informationsverlust verbunden: Die Klasseneinteilung führt zu einer ungenaueren Wiedergabe der Variablen; vor allem aber geht durch die Behandlung als ungeordneten Faktor die quantitative Interpretation, die „Intervallskala“ der Variablen, verloren.

Viele Antworten auf interessante Fragen zeigen eigentlich eine **Ordnung**, beispielsweise von „gar nicht einverstanden“ zu „ganz einverstanden“. Oft sind die Kategorien so gewählt, dass sogar eine quantitative Interpretation möglich ist. Für solche Variable lässt sich ebenfalls eine aussagekräftigere Analyse erhoffen, als sie durch die log-linearen Modelle mit ungeordneten Faktoren zu Stande kommt.

- b Im Beispiel der Umwelt-Umfrage wurde das Alter ebenfalls erhoben. Wir fragen, ob die Zuweisung der Hauptverantwortung von der quantitativen Grösse Alter abhängt. Wie soll ein entsprechendes Modell lauten?

Wir könnten postulieren, dass im log-linearen Modell für die Variablen A (Alter) und B (Hauptverantwortung) die η_{hk} linear vom Alter x_h abhängen sollen. Statt für jedes Niveau h des Alters einen eigenen Koeffizienten α_h zuzulassen, schreiben wir den Effekt von A wie in der Regression als $\alpha \cdot x_h$. Für die Wechselwirkungen wird auf diese Weise aus $(\alpha\beta)_{hk}$ der Ausdruck $(\alpha\beta)_k x_h$, also für jedes Niveau von B eine lineare Abhängigkeit von x mit einer Steigung $(\alpha\beta)_k$. Das führt zu $\lambda_{hk} = \mu + \alpha x_h + \beta_k + (\alpha\beta)_k x_h$.

Dieses Modell macht auch für die Anzahlen N_{h+} der Personen mit Alter x_h Einschränkungen, die sich aus der speziellen Form des Modells ergeben. Diese Anzahlen wollen wir aber wie in den bisher betrachteten Modellen frei an die Daten anpassen können, da sie mit der Abhängigkeit zwischen A und B nichts zu tun haben. Deshalb machen wir für den Haupteffekt von A die lineare Form rückgängig und führen wieder die Haupteffekte α_h ein (vergleiche 11.2.c). Nun lautet das Modell

$$\lambda_{hk} = \mu + \alpha_h + \beta_k + (\alpha\beta)_k x_h .$$

Wie üblich muss einer der Koeffizienten $(\alpha\beta)_k$ durch eine Nebenbedingung festgelegt werden, damit die Parameter identifizierbar werden.

- c Die Bedeutung eines solchen Modells zeigt sich bei der Betrachtung von Doppelverhältnissen. Die Rechnung aus 11.2.j ergibt jetzt

$$\begin{aligned} \log\langle\theta_{hk,h'k'}\rangle &= \log\left\langle\frac{P\langle B = k|A = x_h\rangle}{P\langle B = k'|A = x_h\rangle} \Big/ \frac{P\langle B = k|A = x_{h'}\rangle}{P\langle B = k'|A = x_{h'}\rangle}\right\rangle \\ &= ((\alpha\beta)_k - (\alpha\beta)_{k'})(x_h - x_{h'}). \end{aligned}$$

Die logarithmierten Doppelverhältnisse für den Vergleich von $B = k$ mit $B = k'$ sind also proportional zur Differenz der x -Werte.

- d **Beispiel Umwelt-Umfrage.** Die Antworten auf die Frage nach der Hauptverantwortung als Funktion des Alters der Antwortenden ist in Abbildung 11.5.d dargestellt. Die eingezeichnete Glättung zeigt eine Tendenz, mit zunehmendem Alter dem Staat eine grössere Rolle in diesem Bereich zuzuweisen.

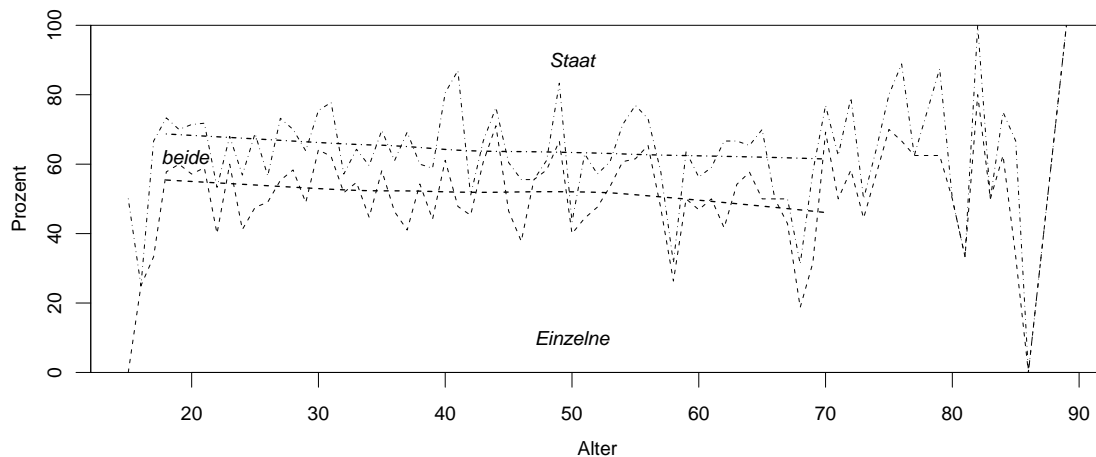


Abbildung 11.5.d: Zuweisung der Hauptverantwortung gegen Alter im Beispiel der Umwelt-Umfrage. Der untere Linienzug zeigt $N_{h1}/Nh+$ in Prozenten, der obere $1 - N_{h3}/Nh+$

Diese Tendenz lässt sich mit einem log-linearen Modell testen. Man bildet zunächst eine Kontingenztabelle mit der als Faktor behandelten quantitativen Variablen Alter (A) und dem Faktor Hauptverantwortung (B). Nun kann man das vorangehende Modell mit und ohne den Term $(\alpha\beta)_k x_h$ miteinander vergleichen. Man erhält eine Devianz-Differenz von 1.40 bei einem Unterschied der Freiheitsgrade von 2. Das ist klar nicht signifikant. (Dabei haben wir in der Figur für die Glättung wie in der Analyse nur die Personen zwischen 18 und 72 Jahren berücksichtigt; wenn alle Daten einbezogen werden, ist der Effekt noch schwächer.)

Konsequenterweise müssen wir auch weitere wichtige erklärende Variablen ins Modell einbeziehen, um einen allfälligen „reinen“ Alterseffekt nachweisen zu können. Wir wissen ja, dass die Schulbildung für die Zuweisung der Hauptverantwortung wesentlich ist. Nachdem wir Simpson's Paradox kennen, tun wir gut daran, diesen Faktor ins Modell aufzunehmen. Es zeigt sich aber auch dann keine Signifikanz.

- e Die Idee der Wechselwirkungen, die linear von einer quantitativen Variablen abhängen, lassen sich leicht auf mehrere quantitative Größen und allgemeinere quantitative Zusammenhänge als die Linearität erweitern. Christensen (1990) widmet dem Problem der "Factors with Quantitative Levels" ein Kapitel.

11.6 Ordinale Variable

- a Wie bereits festgestellt sind ordinale Variable in den Anwendungen häufiger als ungeordnete nominale. Die bisherigen Modelle berücksichtigen die Ordnung der Werte nicht. Das führt zu weniger gut bestimmten Zusammenhängen und, damit zusammenhängend, zu weniger mächtigen Tests.
- b Erste Idee: Für ordinale Zielgrößen haben wir bei den Verallgemeinerten Linearen Modellen die Idee der latenten Variablen eingeführt. Für den Zusammenhang zwischen zwei ordinalen Variablen könnten wir uns als Modell eine bivariate Normalverteilung von zwei latenten Variablen $Z^{(A)}$ und $Z^{(B)}$ denken, die durch Klassierung in die beiden beobachteten ordinalen Variablen A und B übergehen.
- c Zweite Idee: Falls die beiden ordinalen Variablen A und B einen „positiven“ Zusammenhang haben, sind die Abweichungen vom Modell der Unabhängigkeit in erhöhten Häufigkeiten der Fälle „ A klein und B klein“ und „ A gross und B gross“ und gleichzeitig zu kleinen Häufigkeiten der Fälle „ A klein und B gross“ und „ A gross und B klein“ zu finden. Ein entsprechendes Modell erhält man, wenn man setzt

$$\log \langle \lambda_{hk} \rangle = \mu + \alpha_h + \beta_k + \gamma \zeta_h \eta_k$$

mit geeigneten Zahlen ζ_h („zeta“ h) und η_k („eta“ k), die wir Scores nennen und die für kleine h (k) negativ und für grosse positiv sind.

Die Scores kann man festlegen, beispielsweise als fortlaufende ganze Zahlen, oder aus den Daten schätzen. Sie spielen eine ähnliche Rolle wie die Klassengrenzen für die latenten Variablen. Wenn man sie schätzen will, wird es technisch schwierig.

Da für feste h eine log-lineare Abhängigkeit der Häufigkeiten von (den Scores) der Variablen B vorliegt und umgekehrt, spricht man von „linear-by-linear association“.

- d Wenn die Scores nicht geschätzt, sondern als ganze Zahlen festgelegt werden, dann bedeutet das Modell, dass alle logarithmierten Doppelverhältnisse benachbarter Zellen gleich γ sind.
- e Ist nur eine Variable ordinal, so lautet ein Modell mit Scores so:

$$\log \langle \lambda_{hk} \rangle = \mu + \alpha_h + \beta_k + \gamma_h \eta_k$$

- f Zusammenhänge zwischen den beiden Ideen? Siehe Literatur.
Literatur: Agresti (2002, Kap. 8).