7 Eine und zwei kategorielle Variable

7.1 Einleitung

a In Umfragen wird für jede Frage vorzugsweise eine Liste von Auswahlantworten angeboten. Es wird beispielsweise gefragt, welches von 5 Produkten man bevorzugt. In der Medizin wird eine Diagnose bestimmt, die den Patienten einer Gruppe von Kranken zuweist. In der Botanik kann man die Blütenfarbe oder die Blattform festhalten. In der Technik kann bei Geräte-Ausfällen eine Ursache, der Hersteller, die Produktions-Schicht u.a.m. notiert werden.

b In all diesen Beispielen entstehen **kategorielle** Daten. Eine kategorielle Variable hält fest, zu welcher **Kategorie** oder **Klasse** jede Beobachtungseinheit (Person, Objekt, Zeitperiode, ...) bezüglich eines Merkmals gehört. In der Regression haben wir solche Variable bisher nur als Eingangsvariable benützt und sie dann als **Faktoren** bezeichnet.
Manchmal entstehen solche Daten auch durch **Klassierung** von kontinuierlichen Merkmalen: Man teilt beispielsweise Personen in die Altersklassen „unter 26“, „26-45“, „46-65“, „über 65“ ein. Dabei geht Information verloren, aber manchmal wird die Auswertung einfacher verständlich.

c ▷ **Beispiel**. In einer **Umfrage** zum Umweltschutz wurde unter anderem gefragt, ob man sich durch **Umweltschadstoffe** beeinträchtigt fühle (Quelle: „Umweltschutz im Privatbereich“. Erhebung des EMNID, Zentralarchiv für empirische Sozialforschung der Universität Köln, vergleiche Stahel (2002), 10.3.a). Die möglichen Antworten waren: (1) „überhaupt nicht beeinträchtigt“, (2) „etwas beeinträchtigt“, (3) „ziemlich beeinträchtigt“ und (4) „sehr beeinträchtigt“.

Man interessiert sich u.a. dafür, ob die Beeinträchtigung etwas mit der Schulbildung zu tun hat. Man wird also dieses soziologische Merkmal ebenfalls erfragen und dazu die Schulbildung beispielsweise in die fünf Kategorien (1) Volks-, Hauptschule ohne Lehrabschluss; (2) mit Lehrabschluss; (3) weiterbildende Schule ohne Abitur; (4) Abitur, Hochschulreife, Fachhochschulreife; (5) Studium (Universität, Akademie, Fachhochschule) einteilen.

In der Umfrage wurde natürlich auch das Alter und das Geschlecht erfasst. Wir werden das Beispiel in den folgenden Kapiteln immer wieder aufgreifen und dabei auch Verbindungen mit Antworten auf die Frage nach der Hauptverantwortung untersuchen, die die Befragten (1) dem Staat, (2) den Einzelnen oder (3) beiden zusammen zuweisen konnten. ◁

d Die Auswertung solcher Daten muss berücksichtigen,

- dass **Differenzen** zwischen den Kategorien nicht sinnvoll als Unterschiede zwischen Beobachtungseinheiten interpretiert werden können, auch wenn man sie oft mit numerischen **Codes** 1,2,..., bezeichnet;
- dass die möglichen Werte oft keine natürliche **Ordnung** aufweisen; ist eine solche doch vorhanden (Gefährlichkeit einer Krankheit, Antworten von „gar nicht einverstanden“ bis „vollkommen einverstanden“, klassierte quantitative Variable usw.), so spricht man von **ordinalen** Daten, andernfalls von **nominalen** Daten;
- dass für die meisten solchen Variablen nur **wenige, vorgegebene Werte** möglich sind.

Eine Normalverteilung oder eine andere stetige Verteilung kommt für solche Daten nicht in Frage – ausser allenfalls als grobes erstes Modell, wenn wenigstens eine ordinale Skala vorliegt.

e Den ersten Schritt der Auswertung solcher Daten bildet ihre **Zusammenfassung**: Man **zählt**, wie viele Beobachtungseinheiten in die möglichen Kategorien oder Kombinationen von Kategorien fallen.

Die (absoluten oder relativen) Häufigkeiten werden in einem **Stabdiagramm** (Abbildung 7.1.e), einem Histogramm oder einem **Kuchendiagramm** *(pie chart)* dargestellt. Wir zeichnen hier kein Kuchendiagramm, weil empirische Untersuchungen gezeigt haben, dass diese weniger genau erfasst werden als Stabdiagramme (Cleveland, 1994).

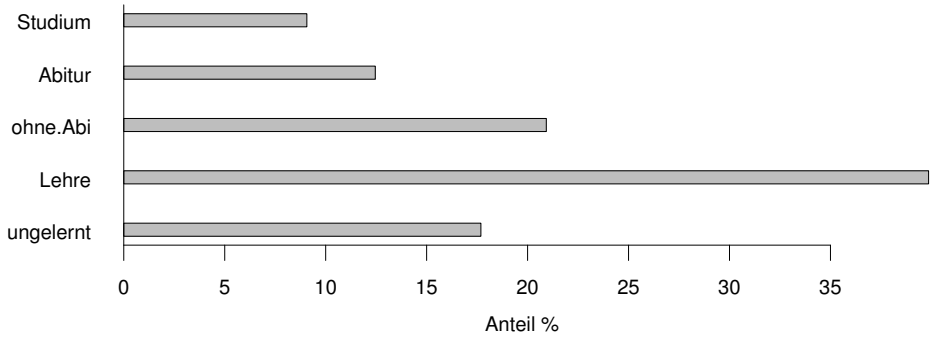


Abbildung 7.1.e: Stabdiagramm der Schulbildung im Beispiel der Umweltumfrage

f Mit zwei kategoriellen Variablen entsteht eine (zweidimensionale) **Kreuztabelle** oder **Kontingenztafel**.

Im **Beispiel** der Umweltumfrage zeigt Tabelle 7.1.f die Ergebnisse für die zwei Variablen „Schulbildung“ und „Beeinträchtigung“.

		Beeinträchtigung (B)				
		nicht	etwas	zieml.	sehr	Summe
Schulbildung (A)	ungelernt	196	73	35	17	321
	Lehre	410	224	78	35	747
	ohne.Abi	152	131	70	28	381
	Abitur	67	81	46	16	210
	Studium	42	59	40	17	158
	Summe	867	568	269	113	1817

Tabelle 7.1.f: Schulbildung und Beeinträchtigung durch Umweltschadstoffe

Man kann natürlich auch die Anzahlen für alle Kombinationen von drei und mehr Variablen festhalten und spricht dann von höher-dimensionalen Kontingenztafeln.

g Durch die Zusammenfassung entstehen **Häufigkeitsdaten**, oft auch **Zähldaten** genannt. Modelle, die die Grundlage für die schliessende Statistik bilden, legen dann fest, mit welchen Wahrscheinlichkeiten welche Anzahlen auftreten werden.

Lindsey (1995) legt Wert auf eine nützliche Unterscheidung: Zähldaten, die auf die geschilderte Weise durch Auszählen der Beobachtungseinheiten, die in bestimmte Kategorien fallen, zu Stande kommen, nennt er „**frequency data**“ (also Häufigkeitsdaten).

Wenn für jede Beobachtungseinheit eine Anzahl angegeben wird, beispielsweise die Zahl der aufgetretenen Fehler in jeder Woche oder die Zahl der beobachteten Hirsche pro Begehung, so spricht er von „**count data**“, was wir zur Unterscheidung vom zweideutigen Wort Zähldaten mit **Anzahldaten** bezeichnen wollen. Ein solcher count kann irgendwelche Objekte oder Ereig-

nisse zählen. Der wesentliche Unterschied ist der, dass für Häufigkeitsdaten die unabhängigen Beobachtungen zuerst zusammengefasst werden müssen. Die Variablen für die ursprünglichen Beobachtungen sind dann keine Anzahlen, sondern kategorielle Variable.

h Häufig kann man bei statistischen Studien von der Problemstellung her eine Variable als **Zielgrösse** oder **Antwortfaktor** erkennen, deren Zusammenhänge mit anderen, den **erklärenden Variablen** oder Faktoren durch ein Modell beschrieben werden sollen. Im Beispiel der Umweltumfrage wird man die Beeinträchtigung oder auch die Benennung der Hauptverantwortung als Antwortfaktor ansehen und die Einflüsse der Schulbildung oder anderer soziologischer Merkmale auf diese Grösse erfassen wollen.

Es geht also darum, ein Regressionsmodell zu entwickeln, bei dem die Zielgrösse kategoriell ist. Wenn die Zielgrösse nur zwei mögliche Werte hat, also binär ist, bietet die **logistische Regression** das brauchbarste und einfachste Modell an. Die Verallgemeinerung auf mehr als zwei mögliche Werte heisst multinomiale Regression. Für geordnete Zielgrössen gibt es ebenfalls Erweiterungen; die wichtigste läuft unter dem Namen „kumulative Logits“.

Diese Modelle gehören zum allgemeineren Gebiet der **Verallgemeinerten Linearen Modelle** *(Generalized Linear Models)*, die bereits behandelt wurden.

i Wenn die Variablen „gleichberechtigt“ behandelt werden sollen, könnte man von einer Fragestellung der **multivariaten Statistik kategorieller Daten** sprechen. Die Analyse von Zusammenhängen entspricht dann der Korrelations-Analyse von stetigen Daten.

Hierfür bieten sich Methoden für Kontingenztafeln, vor allem die **loglinearen Modelle** an, die wir in Kapitel 10.S.0.c behandeln werden. Loglineare Modelle eignen sich auch dazu, Fragestellungen mit mehreren Antwortgrössen zu behandeln. Sie gehören ebenfalls zu den Verallgemeinerten Linearen Modellen.

7.2 Modelle für Kreuztabellen

a Zunächst wollen wir uns mit Zusammenhängen zwischen zwei Variablen befassen. Die Daten aus einer Umfrage, Beobachtungsstudie oder einem Versuch kann man, wie in 7.1.f gesagt, in einer Kreuztabelle zusammenfassen. Wir führen Bezeichnungen ein:

		Variable B							
		1	2	3		k		s	$\sum$
	1	n_{11}	n_{12}	n_{13}	$\dots$	n_{1k}	$\dots$	n_{1s}	n_{1+}
	2	n_{21}	n_{22}	n_{23}	$\dots$	n_{2k}	$\dots$	n_{2s}	n_{2+}
Variable A		$\vdots$	$\vdots$			$\vdots$		$\vdots$	$\vdots$
	h	n_{h1}	n_{h2}	$\dots$		n_{hk}	$\dots$	n_{hs}	n_{h+}
		$\vdots$	$\vdots$			$\vdots$		$\vdots$	$\vdots$
	r	n_{r1}	n_{r2}	$\dots$		n_{rk}	$\dots$	n_{rs}	n_{r+}
	$\sum$	n_{+1}	n_{+2}	$\dots$		n_{+k}	$\dots$	n_{+s}	n

Die Tabelle enthält die absoluten Häufigkeiten n_{hk} von Beobachtungen für zwei Variable A und B, mit r resp. s Kategorien. Insgesamt gibt es rs Kombinationen. Die **Randhäufigkeiten** für die einzelnen Variablen werden mit n_{h+} und n_{+k} bezeichnet.

b Die Tabelle macht klar, welche Art von Daten wir erwarten. Damit wir irgendwelche Fragen statistisch beantworten können, brauchen wir ein **Modell**, das beschreibt, **welche Wahrscheinlichkeit jede mögliche Kombination von Werten** für *eine einzelne Beobachtung* hat. Wir bezeichnen die Wahrscheinlichkeit, dass Variable A Ausprägung h und Variable B Ausprägung k erhält, mit π_{hk}. Die Wahrscheinlichkeiten π_{hk} legen die gemeinsame Verteilung von A und B fest. Es muss $\sum_{h,k} \pi_{hk} = 1$ gelten.

Die **Randverteilungen** der Variablen sind durch die Randsummen $\pi_{h+} = \sum_k \pi_{hk}$ und $\pi_{+k} = \sum_h \pi_{hk}$ bestimmt. Interessante Modelle werden dadurch entstehen, dass man für die π_{hk} Einschränkungen einführt.

c Das einfachste Modell macht keine Einschränkungen. Die Wahrscheinlichkeiten werden dann durch die **relativen Häufigkeiten** geschätzt,

$$\widehat{\pi}_{hk} = N_{hk}/n$$

Hier wurden die N_{hk} gross geschrieben, da sie jetzt Zufallsvariable sind. Die gesamte Anzahl Beobachtungen n wird dagegen üblicherweise als feste Zahl angenommen.

▷ Im Beispiel der Umweltumfrage (7.1.c) ergibt sich Tabelle 7.2.c. ◁

		Beeinträchtigung (B)				
		nicht	etwas	zieml.	sehr	Summe
Schulbildung (A)	ungelernt	10.8	4.0	1.9	0.9	17.7
	Lehre	22.6	12.3	4.3	1.9	41.1
	ohne.Abi	8.4	7.2	3.9	1.5	21.0
	Abitur	3.7	4.5	2.5	0.9	11.6
	Studium	2.3	3.2	2.2	0.9	8.7
	Summe	47.7	31.3	14.8	6.2	100.0

Tabelle 7.2.c: Relative Häufigkeiten in Prozenten im Beispiel der Umweltumfrage

d Wenn der Faktor A eine erklärende Variable für die Zielgrösse oder den Antwortfaktor B ist, dann ist es informativ, die Wahrscheinlichkeitsverteilung von B auf jeder Stufe von A zu bilden, also die **bedingten Wahrscheinlichkeiten**

$$\pi_{k|h} = P\langle B = k \mid A = h\rangle = \frac{\pi_{hk}}{\pi_{h+}}$$

zu betrachten. Eine Schätzung für diese Grössen erhält man, indem man die N_{hk} durch die Randsummen N_{h+} teilt, $\widehat{\pi}_{k|h} = N_{hk}/N_{h+}$.

▷ Für das Beispiel zeigt Tabelle 7.2.d, dass die Beeinträchtigung mit höherer Schulstufe zunimmt. Dies sieht man noch besser in einer grafischen Darstellung, in der die Verteilungen der Beeinträchtigung für die verschiedenen Schulbildungsklassen mit Histogrammen verglichen werden (Abbildung 7.2.d). ◁

e Die π_{hk} legen die Wahrscheinlichkeiten fest, mit denen die *einzelnen Beobachtungen* in die Zellen $[h, k]$ der Tabelle fallen. Wenn wir nun n Beobachtungen machen, stellt sich die Frage, welcher Verteilung die *Häufigkeiten* N_{hk} *der Beobachtungen* folgen.

Die Antwort liefert die **Multinomiale Verteilung**, die genau für solche Fälle eingeführt wurde (Stahel (2002), 5.5). Dass die Einzelwahrscheinlichkeiten π_{hk} hier zwei Indizes tragen, ändert an der Situation nichts. Es gilt also

$$P\langle N_{11} = n_{11}, N_{12} = n_{12}, ..., N_{rs} = n_{rs}\rangle = \frac{n!}{n_{11}!n_{12}!...n_{rs}!}\, \pi_{11}^{n_{11}}\pi_{12}^{n_{12}}, ..., \pi_{rs}^{n_{rs}} \ .$$

		Beeinträchtigung (B)				
		nicht	etwas	zieml.	sehr	Summe
Schulbildung (A)	ungelernt	61.1	22.7	10.9	5.3	100
	Lehre	54.9	30.0	10.4	4.7	100
	ohne.Abi	39.9	34.4	18.4	7.3	100
	Abitur	31.9	38.6	21.9	7.6	100
	Studium	26.6	37.3	25.3	10.8	100
	Summe	47.7	31.3	14.8	6.2	100

Tabelle 7.2.d: Beeinträchtigung der Gruppen in Prozentzahlen im Beispiel der Umweltumfrage

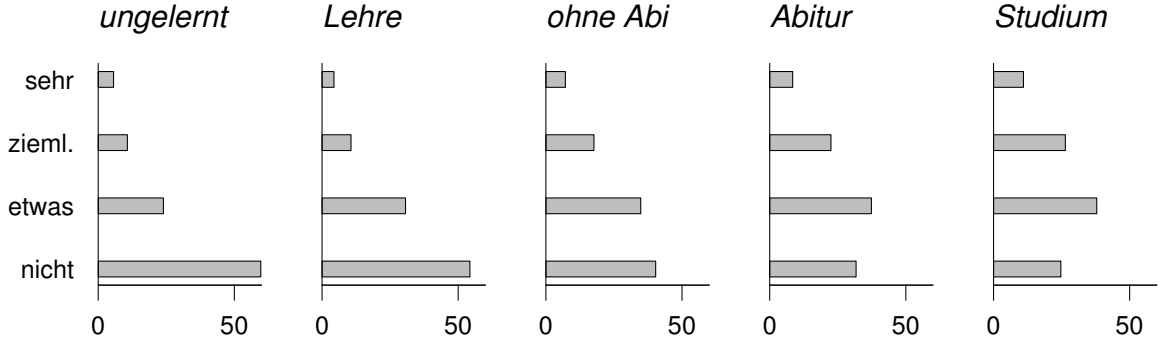


Abbildung 7.2.d: Histogramme zum Vergleich der Beeinträchtigung für die Schulbildungsklassen im Beispiel der Umweltumfrage

Wir schreiben

$$[N_{11}, N_{12}, ..., N_{rs}] \sim \mathcal{M}\langle n; \pi_{11}, \pi_{12}, ..., \pi_{rs}\rangle \ .$$

In englischen Büchern spricht man von *multinomial sampling*. Die Erwartungswerte der Anzahlen N_{hk} sind $\mathcal{E}\langle N_{hk}\rangle = n\pi_{hk}$.

f In manchen Studien sind die einen Randtotale im Voraus festgelegt: Man befragt beispielsweise gleich viele Frauen und Männer oder eine vorbestimmte Anzahl Mitarbeitende aus jeder Hierarchiestufe. Im Sinne der Stichproben-Erhebungen zieht man eine **geschichtete Stichprobe**. Die N_{h+} sind also vorgegeben, $N_{h+} = n_{h+}$. Man erhält r unabhängige Stichproben, und jede folgt einer Multinomialen Verteilung,

$$[N_{h1}, N_{h2}, ..., N_{hs}] \sim \mathcal{M}\langle n_{h+}; \pi_{h1}, \pi_{h2}, ..., \pi_{hs}\rangle \ , \quad \text{unabhängig,} \quad \text{für } h = 1, ..., r \ .$$

Man spricht von *independent multinomial sampling*.

g Rechnungen und Überlegungen können einfacher werden, wenn man das folgende Modell verwendet, das nicht nur die Randtotale frei lässt, sondern sogar die Gesamtzahl N der Beobachtungen als zufällig annimmt:

Zur Herleitung der Poisson-Verteilung wurden in Stahel (2002), 5.2.a, Regentropfen betrachtet, die auf Platten fallen. Hier stellen wir uns $r \cdot s$ Platten mit den Flächen π_{hk} vor. Zählt man die Regentropfen, die in einem festen Zeitabschnitt auf die Platten fallen, dann wird ihre Gesamtzahl gemäss der erwähnten Herleitung eine Poisson-Verteilung $\mathcal{P}\langle\lambda\rangle$ haben, wobei λ die erwartete Anzahl misst. Die Überlegung gilt aber auch für jede einzelne Platte: N_{hk} ist Poisson-verteilt, und die erwartete Anzahl λ_{hk} ist proportional zur Fläche, nämlich $\lambda_{hk} = \pi_{hk} \cdot \lambda$, da π_{hk} der

Anteil der Platte $[h,k]$ an der Gesamtfläche ist. Die Zahlen der Tropfen, die im betrachteten Zeitraum auf die einzelnen Platten fallen, sind stochastisch unabhängig.

Es ergibt sich das **Modell der unabhängigen Poisson-Verteilungen** *(Poisson sampling),*

$$N_{hk} \sim \mathcal{P}\langle \pi_{hk} \cdot \lambda \rangle \ , \quad \text{unabhängig für } h = 1,...,r \text{ und } k = 1,...,s \ .$$

Die Wahrscheinlichkeiten werden

$$P\langle N_{11} = n_{11}, N_{12} = n_{12}, ..., N_{rs} = n_{rs} \rangle = \prod_{h,k} \frac{\lambda_{hk}^{n_{hk}}}{n_{hk}!}\, e^{-\lambda_{hk}}$$

mit $\lambda_{hk} = \pi_{hk}\lambda$.

h Man kann im letzten Modell die Gesamtzahl N festhalten und die bedingte Verteilung der N_{hk}, gegeben $N = n$, betrachten. Das ergibt exakt das Modell der Multinomialen Verteilung (7.2.e), ...

* ... denn es gilt $\lambda = \sum_{h,k} \lambda_{hk}$, $\pi_{hk} = \lambda_{hk}/\lambda$ und deshalb

$$P\langle N_{11} = n_{11}, N_{12} = n_{12}, ..., N_{rs} = n_{rs} \mid N = n \rangle = \prod_{h,k} \frac{\lambda_{hk}^{n_{hk}}}{n_{hk}!}\, e^{-\lambda_{hk}} \Big/ \frac{\lambda^n}{n!}\, e^{-\lambda}$$

$$= \frac{n!}{\prod_{h,k} n_{hk}!} \cdot \frac{\prod_{h,k} \lambda_{hk}^{n_{hk}}}{\lambda^{\sum_{h,k} n_{hk}}} \cdot \frac{e^{-\sum_{h,k} \lambda_{hk}}}{e^{-\lambda}} = \frac{n!}{\prod_{h,k} n_{hk}!} \cdot \prod_{h,k} \pi_{hk}^{n_{hk}} \ .$$

Hält man zudem die Randtotale $N_{h+} = n_{h+}$ fest, dann erhält man die unabhängigen Multinomialen Verteilungen von 7.2.f. (Später werden wir auch noch die anderen Randsummen festhalten, siehe 7.3.d, 7.3.l.)

Diese Zusammenhänge werden wir bei Wahrscheinlichkeitsrechnungen im Zusammenhang mit kategoriellen Daten immer wieder ausnützen. Ein grundlegender Trick wird darin bestehen, mit dem sehr einfachen Modell der unabhängigen Poisson-Variablen N_{hk} zu arbeiten und nachher für die „Bedingtheit“ Korrekturen vorzunehmen.

7.3 Unabhängigkeit von zwei Variablen und Vergleich von Stichproben

a Die Frage, ob zwei Variable mit einander in einem Zusammenhang stehen, ist eine grundlegende Frage der Wissenschaft. Sie verlangt nach einem Test für die stochastische Unabhängigkeit – in unserem Zusammenhang die Unabhängigkeit von zwei kategoriellen Grössen.

Eine **Nullhypothese**, die statistisch getestet werden soll, muss durch ein Wahrscheinlichkeitsmodell beschrieben sein. Hier geht es darum, die der Nullhypothese entsprechenden Einschränkungen an die π_{hk} zu formulieren. Wenn die Variablen A und B **unabhängig** sind, dann heisst das, dass

$$\pi_{hk} = P\langle A = h, B = k \rangle = P\langle A = h \rangle \cdot P\langle B = k \rangle = \pi_{h+}\pi_{+k}$$

gilt. Für die Anzahlen N_{hk} erhalten wir gemäss 7.2.e die Erwartungswerte $\mathcal{E}\langle N_{hk} \rangle = n\pi_{h+}\pi_{+k}$.

b Um die Nullhypothese zu prüfen, schätzen wir die πs und bilden die Differenzen

$$\widehat{\pi}_{hk} - \widehat{\pi}_{h+}\widehat{\pi}_{+k} = \frac{N_{hk}}{n} - \frac{N_{h+}}{n} \cdot \frac{N_{+k}}{n} \ .$$

Multipliziert man diese Ausdrücke mit n, so werden sie zu Differenzen zwischen den Anzahlen N_{hk} und

$$\widehat{\lambda}^{(0)}_{hk} = N_{h+}N_{+k}/n = n\widehat{\pi}_{h+}\widehat{\pi}_{+k} \ ,$$

welche man gemäss dem vorhergehenden Absatz als die geschätzten Erwartungswerte dieser Anzahlen unter der Nullhypothese erkennt.

Wenn diese Differenzen zu stark von null verschieden sind, ist die Nullhypothese zu verwerfen. Wie stark „zu stark“ ist, können wir beurteilen, da gemäss 7.2.h (näherungsweise) $N_{hk} \sim \mathcal{P}\left\langle \lambda^{(0)}_{hk} \right\rangle$ und deshalb $\operatorname{var}\langle N_{hk}\rangle \approx \lambda_{hk}$ ist. Es ist also

$$R^{(P)}_{hk} = \frac{N_{hk} - \widehat{\lambda}^{(0)}_{hk}}{\sqrt{\widehat{\lambda}^{(0)}_{hk}}}$$

näherungsweise eine Grösse mit Erwartungswert 0 und Varianz 1. Für nicht allzu kleine $\widehat{\lambda}^{(0)}_{hk}$ ist die Poisson-Verteilung näherungsweise eine Normalverteilung, und $R^{(P)}_{hk}$ ist standard-normalverteilt.

c Um aus den standardisierten Differenzen eine einzige Teststatistik zu erhalten, bilden wir wie beim Kriterium der Kleinsten Quadrate in der Regression ihre Quadratsumme

$$T = \sum_{h,k} (R^{(P)}_{hk})^2 = \sum_{h,k} \frac{(N_{hk} - \widehat{\lambda}^{(0)}_{hk})^2}{\widehat{\lambda}^{(0)}_{hk}} = \sum_{h,k} \frac{(N_{hk} - N_{h+}N_{+k}/n)^2}{N_{h+}N_{+k}/n} \ .$$

Diese Summe entspricht der allgemeinen „Merkform“ einer Chi-Quadrat-Teststatistik

$$T = \sum_{h,k} \frac{(\ \text{beobachtet}_{hk} - \ \text{erwartet}_{hk})^2}{\text{erwartet}_{hk}}$$

Eine Quadratsumme von unabhängigen, standard-normalverteilten Grössen ist chiquadrat-verteilt; die Anzahl Freiheitsgrade ist gleich der Zahl der Summanden. Die „kleine Korrektur“, die durch das „Bedingen“ auf die geschätzten $\lambda^{(0)}_{h+}$ und $\lambda^{(0)}_{+k}$ (oder die Randsummen der Kreuztabelle) nötig werden, besteht (wie in der linearen Regression mit normalverteilten Fehlern) darin, dass die Zahl der Freiheitsgrade um die Anzahl solcher Bedingungen reduziert wird. Es gibt r Bedingungen für die Zeilen und danach noch $s-1$ unabhängige Bedingungen für die Spalten (da die Summen der Randsummen gleich sein müssen). So erhält man $rs-r-(s-1) = (r-1)(s-1)$ Freiheitsgrade.

d* In 7.2.f wurden die Randsummen n_{h+} als fest betrachtet. Das entspricht dem Verlust der Freiheitsgrade durch die Schätzung der $\lambda^{(0)}_{h+}$. Mit diesem Modell kann man also die bedingte Verteilung der Teststatistik, gegeben die $\widehat{\lambda}^{(0)}_{h+}$ oder die n_{h+}, untersuchen. Da auch die $\lambda^{(0)}_{+k}$ geschätzt werden, muss auch auf die n_{+k} bedingt werden. Man kann zeigen, dass die Chiquadrat-Verteilung mit der angegebenen Zahl von Freiheitsgraden eine gute Näherung für diese doppelt bedingte Verteilung ist, vergleiche auch 7.3.l.

e Zusammengefasst erhalten wir den **Chiquadrat-Test für Kontingenztafeln**:
Es sei zu testen
$H_0 : \pi_{hk} = \pi_{h+} \cdot \pi_{+k}$ – Unabhängigkeit von A und B oder
$H_0 : \pi_{k|h} = \pi_{k|h'}$ – (bedingte) Verteilung von B gegeben $A = h$ ist gleich für alle h.
Teststatistik:

$$T = \sum_{h,k} \frac{(N_{hk} - N_{h+}N_{+k}/n)^2}{N_{h+}N_{+k}/n} \ .$$

Verteilung unter der Nullhypothese: $T \sim \chi^2_{(r-1)(s-1)}$

Damit die genäherte Verteilung brauchbar ist, dürfen die *geschätzten erwarteten Anzahlen* $\widehat{\lambda}_{hk}^{(0)} = N_{h+}N_{+k}/n$ nicht zu klein sein. Nach van der Waerden (1971) und F. Hampel (persönliche Mitteilung aufgrund eigener Untersuchungen) kann folgende Regel aufgestellt werden: Etwa 4/5 der $\widehat{\lambda}_{hk}^{(0)}$ müssen ≥ 4 sein, die übrigen ≥ 1. Bei vielen Klassen (rs gross) können einzelne $\widehat{\lambda}_{hk}^{(0)}$ sogar noch kleiner sein (aus Stahel, 2002, Abschnitt 10.1.n).

f ▷ Im **Beispiel der Umweltumfrage** (7.1.c) fragten wir, ob die empfundene Beeinträchtigung etwas mit der Schulbildung zu tun hat. Tabelle 7.3.f enthält die erwarteten Anzahlen und die $R_{hk}^{(P)}$. Deren Quadratsumme $T = 110.26$ ist deutlich zu gross für eine chiquadrat-verteilte Grösse mit $(5-1)(4-1) = 12$ Freiheitsgraden; der kritische Wert beträgt 21.03. Dem entsprechend gibt R als P-Wert eine blanke Null an. Die Nullhypothese der Unabhängigkeit wird also klar verworfen. ◁

$\widehat{\lambda}_{hk}^{(0)}$	k				$R_{hk}^{(P)}$	k			
h	1	2	3	4	h	1	2	3	4
1	153.2	100.3	47.5	20.0	3.5	-2.7	-1.8	-0.7	
2	356.4	233.5	110.6	46.5	2.8	-0.6	-3.1	-1.7	
3	181.8	119.1	56.4	23.7	-2.2	1.1	1.8	0.9	
4	100.2	65.6	31.1	13.1	-3.3	1.9	2.7	0.8	
5	75.4	49.4	23.4	9.8	-3.8	1.4	3.4	2.3	

Tabelle 7.3.f: Geschätzte erwartete Anzahlen $\widehat{\lambda}_{hk}^{(0)}$ und Pearson-Residuen $R_{hk}^{(P)}$ im Beispiel der Umweltumfrage

g Die standardisierten Differenzen $R_{hk}^{(P)}$ werden **Pearson-Residuen** genannt. Sie können anzeigen, wie die Abweichung von der Nullhypothese zu Stande kommt.

Abbildung 7.3.g zeigt sie grafisch in Form eines **association plots** (Cohen (1980)). Die gezeichneten Rechtecke richten sich in ihrer Höhe nach den Pearson-Residuen und in ihrer Breite nach ihrem Nenner $\sqrt{\widehat{\lambda}_{hk}^{(0)}}$, so dass die Flächen proportional zu den (Absolutwerten der) Differenzen der N_{hk} von ihren geschätzten Erwartungswerten $\widehat{\lambda}_{hk}^{(0)}$ werden.

h Man kann die vorherige Frage auch anders formulieren: Antworten die Personen mit verschiedener Schulbildung auf die Frage nach der Belästigung gleich oder verschieden? Das ist dann eine Frage des Vergleichs von Stichproben – den Stichproben aus den verschiedenen Schulstufen. Diese Formulierung läge vor allem dann nahe, wenn die Stichprobe entsprechend der Schulbildung geschichtet erhoben worden wäre, wenn man also aus den verschiedenen Stufen jeweils eine vorgegebene Anzahl Personen befragt hätte. Sie wäre auch dann noch sinnvoll, wenn die Stichprobenumfänge in den verschiedenen Schichten keinen Bezug zu ihrem Anteil in der Grundgesamtheit hätten.

Die Stichproben in den Schichten werden unabhängig gezogen. Es geht also um den **Vergleich von unabhängigen Stichproben**. Im Falle von kontinuierlichen Zufallsvariablen war bei einem Vergleich unabhängiger Stichproben meistens der „Lageparameter“ (Erwartungswert oder Median) von Interesse. Für kategorielle Variable macht diese Frage keinen Sinn; man will hier testen, ob die ganzen Verteilungen der Variablen in den Schichten übereinstimmen. – Für geordnete Grössen ist die Gleichheit der Mediane oft wieder von besonderer Bedeutung, und man kann die Rangtests (U-Test oder Kruskal-Wallis) verwenden.

i Es zeigt sich, dass die erwarteten Anzahlen für die einzelnen Zellen der Tabelle unter der Nullhypothese, dass alle Stichproben der gleichen Verteilung folgen, genau nach der Formel in 7.3.c zu berechnen sind – auch wenn jetzt die Randtotale n_{h+} nicht mehr zufällig sind. Die Teststatistik T, die dort angeführt wurde, zeigt auch die Abweichungen von der neuen Nullhypothese an.

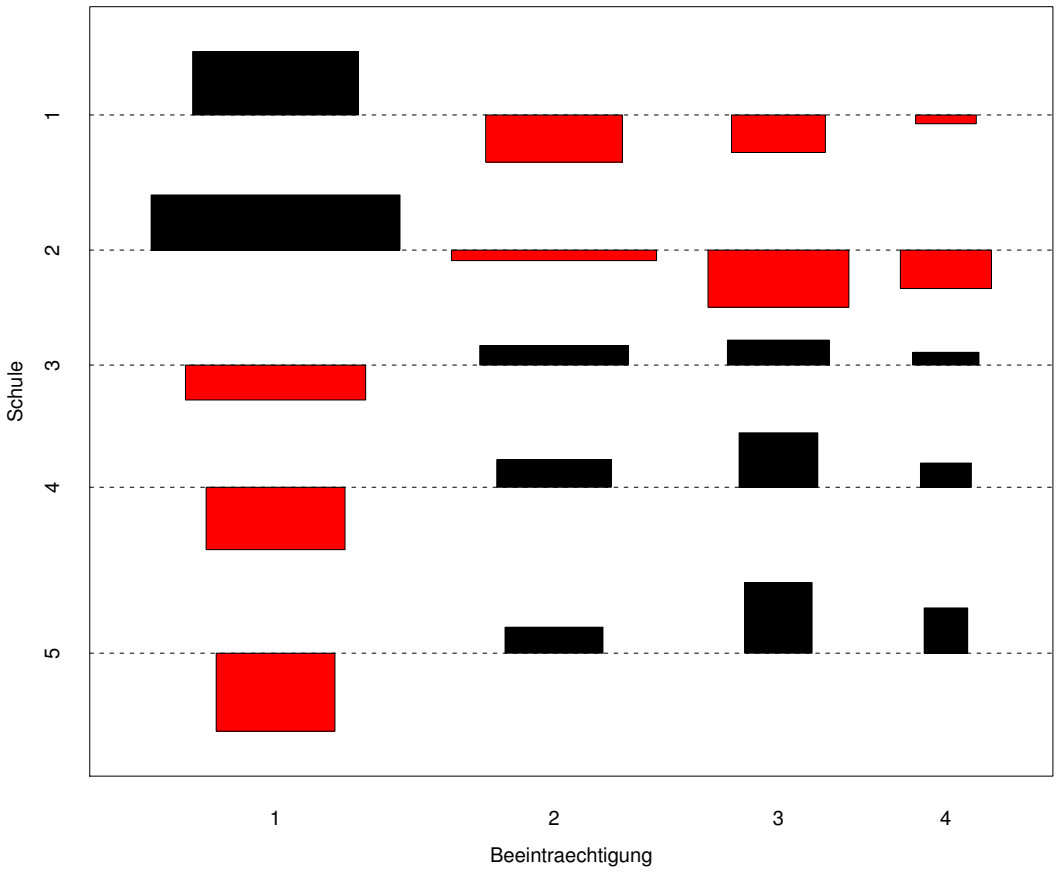


Abbildung 7.3.g: Association Plot für das Beispiel der Umweltumfrage

Ihre Verteilung müsste jetzt, genau genommen, unter dem Modell des independent multinomial sampling bestimmt werden. Das macht aber keinen Unterschied, da bereits für den Test der Unabhängigkeit die bedingte Verteilung, gegeben die Randtotale, verwendet wurde.

Der Test zum Vergleich von unabhängigen Stichproben ist deshalb mit dem Test für die Unabhängigkeit zweier Variablen identisch.

j Eine Kreuztabelle mit nur zwei Zeilen und zwei Spalten wird **Vierfeldertafel** genannt.

▷ **Beispiel Herzinfarkt und Verhütungsmittel** (Agresti, 2002, 2.1.3). Die 58 verheirateten Patientinnen unter 45 Jahren, die in zwei englischen Spitalregionen wegen Herzinfarkt behandelt wurden, und etwa drei Mal mehr Patientinnen, die aus anderen Gründen ins Spital kamen, wurden befragt, ob sie je Verhütungspillen verwendet hätten. Die Ergebnisse zeigt Tabelle 7.3.j. Die Frage ist, ob Verhütungspillen einen Einfluss auf Herzinfarkte haben.

		Herzinfarkt (B)		
		ja	nein	Summe
Verhütungspille (A)	ja	23	34	57
	nein	35	132	167
	Summe	58	166	224

Tabelle 7.3.j: Kreuztabelle der Verwendung von Verhütungspillen und Herzinfarkt.

Zur Beantwortung der Frage vergleichen wir in den beiden Gruppen die Anteile derer, die Pil-

len benützt hatten. Ist $N_{11}/n_{+1} = 23/58 = 40\%$ signifikant von $N_{12}/n_{+2} = 34/166 = 20\%$ verschieden? ◁

k Wir vergleichen also zwei Stichproben in Bezug auf eine binäre Zielgrösse, oder anders gesagt: Wir fragen, ob die Wahrscheinlichkeit für ein Ereignis (die Pillenverwendung) in zwei Gruppen (Herzinfarkt ja oder nein) gleich sei, was oft auch als **Vergleich zweier Wahrscheinlichkeiten** bezeichnet wird.

Wie im allgemeinen Fall eignet sich der gleiche Test, um die **Unabhängigkeit** von zwei Variablen zu testen – in diesem Fall **von zwei binären Variablen**.

Die Teststatistik aus 7.3.c kann umgeformt werden zu

$$T = \frac{n(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1+}n_{2+}n_{+1}n_{+2}} .$$

Sie ist wieder genähert chiquadrat-verteilt, mit gerade mal $(2-1)(2-1) = 1$ Freiheitsgrad. Die Näherung wird noch etwas besser, wenn man die so genannte „continuity correction“ von Yates verwendet (Hartung, Elpelt und Klösener, 2002, VII.1.2.1).

▷ Im Beispiel erhält man

```
Pearson's Chi-squared test with Yates' continuity correction
X-squared = 7.3488, df = 1, p-value = 0.00671
```
◁

l* Die Verteilung der Teststatistik unter der Nullhypothese lässt sich in diesem Fall exakt bestimmen. Wenn die Randtotale wieder als fest betrachtet werden, dann ist die ganze Tabelle bestimmt, wenn noch eine der vier Zahlen aus dem Inneren der Vierfeldertafel bekannt ist – beispielsweise N_{11}. Die Teststatistik hat ja einen einzigen Freiheitsgrad!

Die Verteilung ist durch die Wahrscheinlichkeiten

$$P\langle N_{11} = n_{11}\rangle = \frac{\binom{n_{1+}}{n_{11}}\binom{n_{2+}}{n_{21}}}{\binom{n}{n_{+1}}} = \frac{n_{1+}!}{n_{11}!n_{12}!} \cdot \frac{n_{2+}!}{n_{21}!n_{22}!} \Big/ \frac{n!}{n_{+1}!n_{+2}!} = \frac{n_{1+}!n_{2+}!n_{+1}!n_{+2}!}{n!n_{11}!n_{12}!n_{21}!n_{22}!}$$

gegeben. Sie wird **hypergeometrische Verteilung** genannt. Wenn diese Verteilung benützt wird, spricht man vom **exakten Test von Fisher**.

Hier werden die Randsummen nicht nur für einen Faktor festgehalten wie in 7.2.f, sondern für beide. Die hypergeometrische Verteilung entsteht also aus dem Modell der unabhängigen Multinomialen Verteilungen, indem man in diesem Modell die bedingte Verteilung von N_{11}, gegeben N_{+1} und N_{+2}, bestimmt, vergleiche 7.3.d.

m Für kontinuierliche Variable werden in Statistik-Einführungsbüchern nicht nur unabhängige, sondern auch **verbundene Stichproben** verglichen. Für jede Beobachtungseinheit werden also zwei Variable $Y^{(1)}$ und $Y^{(2)}$ ermittelt, beispielsweise das gleiche Merkmal vor und nach einer Behandlung. Man fragt meistens, ob sich der Erwartungswert (oder ein anderer Lageparameter) verändert hat. Dazu bildet man Differenzen $Y^{(2)} - Y^{(1)}$ und prüft, ob sie zufällig um 0 herum streuen.

Für kategorielle Variable machen Lageparameter und Differenzen keinen Sinn. Wir fragen wieder allgemeiner, ob sich die Verteilungen der beiden Variablen unterscheiden. Damit die Frage Sinn macht, müssen zunächst beide gleich viele mögliche Werte haben ($r = s$), und diese müssen einander in natürlicher Weise entsprechen. Die Verteilungen sind nun nicht nur dann gleich, wenn alle $Y_i^{(1)} = Y_i^{(2)}$ sind, sondern auch dann, wenn die „Übergangs-Wahrscheinlichkeiten“ π_{hk} paarweise übereinstimmen, also $\pi_{hk} = \pi_{kh}$ gilt. Das lässt sich recht einfach testen.

n In einer Vierfeldertafel verwendet man dazu den **McNemar-Test**. Die Nullhypothese heisst $H_0 : \pi_{1+} = \pi_{+1}$ oder, äquivalent dazu, $\pi_{12} = \pi_{21}$.
Teststatistik und Verteilung: $N_{12} \sim \mathcal{B}\langle N_{12} + N_{21}, 1/2\rangle$. Man betrachtet also die bedingte Verteilung der Anzahl der Wechsel von 1 nach 2 (oder von 2 nach 1), gegeben die Anzahl aller Wechsel. Die Beobachtungen, für die beide Variablen den gleichen Wert haben, gehen nicht direkt in den Test ein. Sie verringern nur die „Anzahl Versuche“ für die Binomialverteilung.

o* Wenn die Kreuztabelle mehr als zwei Zeilen und Spalten hat, lässt sich die Nullhypothese $\pi_{hk} = \pi_{kh}$ für alle $h < k$ mit einer Erweiterung dieses Tests prüfen: Es ist

$$T = \sum_{h<k} \frac{(N_{hk} - N_{kh})^2}{N_{hk} + N_{kh}}$$

genähert chiquadrat-verteilt; die Anzahl Freiheitsgrade stimmt mit der Anzahl Summanden überein. Es ist aber wichtig, zu bemerken, dass ein solcher Test nicht eigentlich das prüft, was am Anfang gefragt wurde; die Verteilungen von $Y^{(1)}$ und $Y^{(2)}$ können nämlich auch gleich sein, wenn nicht alle $\pi_{hk} = \pi_{kh}$ sind! Wie man es richtig macht, ist dem Autor im Moment nicht bekannt.

p Die **Statistik-Programme** setzen normalerweise voraus, dass die Daten in der Form der ursprünglichen Daten-Matrix eingegeben werden, dass also für jede Beobachtung i der Wert der Faktoren, A_i, B_i, in einer Zeile eingegeben wird. Im Beispiel der Herzinfarkte sind das 224 Zeilen, für jede Patientin eine. Die Kreuztabelle mit den N_{hk} erstellt das Programm dann selbst.

Wenn man die Kreuztabelle direkt zeilenweise eingibt, können die meisten Programme nichts damit anfangen. Immerhin kann man jeweils die Beobachtungen, die in beiden (später: allen) Variablen übereinstimmen, zusammenfassen. In einer Zeile der Eingabe stehen dann die Werte der beiden Variablen und die Anzahl der entsprechenden Beobachtungen. Für das Beispiel 7.3.j schreibt man die Daten in der folgenden Form auf:

A	B	N
1	1	23
1	2	35
2	1	34
2	2	132

Die Spalte mit den Anzahlen muss dann oft als „Gewicht“ angesprochen werden.

7.4 Abhängigkeit von zwei Variablen

a Wenn zwei Variable nicht unabhängig sind, möchte man ihre Abhängigkeit durch eine Zahl charakterisieren, die die Stärke des Zusammenhangs misst. Für quantitative Variable gibt es dafür die verschiedenen Korrelationen (Pearson- und Rangkorrelationen), die eng miteinander verwandt sind (Stahel (2002) 3.2). Für kategorielle Merkmale gibt es verschiedene Vorschläge.

Besonders bedeutungsvoll und gleichzeitig einfach zu interpretieren sind solche Masse im Fall eines binären Antwortfaktors B, weshalb dieser Fall ausführlicher diskutiert werden soll. Die Wortwahl der Begriffe stammt teilweise aus der Medizin, in der das Vorhandensein einer Krankheit ($B=1$) in Zusammenhang gebracht mit einer Gruppierung (Faktor A), die die „Exposition“ oder „Risikogruppe“ erfasst.

b Wir bezeichnen die bedingte Wahrscheinlichkeit des betrachteten Ereignisses $B = 1$, gegeben die Gruppe $A = h$, als das **Risiko** $\pi_{1|h} = P\langle B=1|A=h\rangle = \pi_{h1}/\pi_{h+}$ für die Gruppe h.

Zum Vergleich des Risikos zwischen zwei Gruppen dienen

- die Risiko-Differenz, $\pi_{1|1} - \pi_{1|2}$. Dieses Mass ist wenig bedeutungsvoll; es kann allenfalls sinnvoll interpretiert werden, wenn man die einzelnen $\pi_{1|h}$ ungefähr kennt.
- das **relative Risiko**, $\pi_{1|1}/\pi_{1|2}$. Für kleine Risiken ist dies brauchbarer als die Risiko-Differenz. Ein relatives Risiko von 4 bedeutet, dass die Wahrscheinlichkeit für das Ereignis in Gruppe eins 4 mal grösser ist als in Gruppe zwei.

c Das nützlichste Mass für den Vergleich von Risiken bildet das **Doppelverhältnis**, englisch präziser **odds ratio** genannt.
Zunächst brauchen wir den Begriff des **Wettverhältnisses** , englisch **odds**. Zu einer Wahrscheinlichkeit, hier $P\langle B=1\rangle$, gehört ein Wettverhältnis $P\langle B=1\rangle/(1-P\langle B=1\rangle) = P\langle B=1\rangle/P\langle B=0\rangle$. Es drückt aus, wie eine Wette abgeschlossen werden müsste, wenn die Wahrscheinlichkeit eines Ereignisses bekannt wäre und die Wette keinem Partner einen positiven Erwartungswert des Gewinns/Verlusts bringen sollte. Eine Wahrscheinlichkeit von 0.75 entspricht einem Wettverhältnis von $3:1=3$.
Wir vergleichen nun die Wettverhältnisse für die beiden Gruppen $h=1$ und $h=2$, indem wir ihren Quotienten bilden,

$$\theta = \frac{P\langle B=1 \mid A=1\rangle}{P\langle B=2 \mid A=1\rangle} \Bigg/ \frac{P\langle B=1 \mid A=2\rangle}{P\langle B=2 \mid A=2\rangle} = \frac{\pi_{1|1}}{\pi_{2|1}} \Bigg/ \frac{\pi_{1|2}}{\pi_{2|2}} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} \,.$$

So entsteht ein Verhältnis von Verhältnissen; deshalb der Name Doppelverhältnis. Es fällt auf, dass im Falle von zwei Gruppen, also einer binären Variablen A, die Rollen von A und B vertauschbar sind. Das Doppelverhältnis ist also ein symmetrisches Mass für die Abhängigkeit von zwei binären Variablen – wie die Korrelation für kontinuierliche Variable es ist.

d Ein odds ratio von 1 bedeutet, dass die odds und damit die (bedingten) Wahrscheinlichkeiten in beiden Gruppen gleich sind. Wenn nur zwei Gruppen vorhanden sind, ist dies gleichbedeutend mit der Unabhängigkeit von A und B. Ein Doppelverhältnis, das >1 ist, bedeutet in diesem Fall, dass die Wahrscheinlichkeit, für beide Variablen den gleichen Wert zu erhalten, gegenüber der Unabhängigkeit erhöht ist – also eine „positive Abhängigkeit“.

e Noch einfacher zu handhaben ist das **logarithmierte Doppelverhältnis** *(log odds ratio)* $\ell\theta = \log\langle\theta\rangle$.
Wir betrachten zunächst den Logarithmus der Wettverhältnisse, die „**log odds**“ $\log\langle P\langle B=1 \mid A=h\rangle/(1-P\langle B=1 \mid A=h\rangle\rangle$ für die beiden Gruppen $A=h=1$ und $h=0$.
Das logarithmierte Doppelverhältnis ist gleich der Differenz der log odds für die beiden Gruppen,

$$\begin{aligned} \ell\theta &= \log\left\langle \frac{P\langle B=1 \mid A=1\rangle}{(1-P\langle B=1 \mid A=1\rangle)} \right\rangle - \log\left\langle \frac{P\langle B=1 \mid A=0\rangle}{(1-P\langle B=1 \mid A=0\rangle)} \right\rangle \\ &= \log\langle\pi_{11}/\pi_{10}\rangle - \log\langle\pi_{01}/\pi_{00}\rangle = \log\langle\pi_{11}\rangle - \log\langle/\pi_{10}\rangle - \log\langle\pi_{01}\rangle + \log\langle\pi_{00}\rangle \,. \end{aligned}$$

Diese Grösse hat folgende Eigenschaften:

- $\ell\theta = 0$ bei Unabhängigkeit,
- $\ell\theta > 0$ bei positiver Abhängigkeit,
- $\ell\theta < 0$ bei negativer Abhängigkeit.
- Vertauscht man die Kategorien (1 und 2) der einen Variablen, so wechselt nur das Vorzeichen von $\ell\theta$.

Im Unterschied zu einer „gewöhnlichen“ (Pearson-) Korrelation ist $\ell\theta$ aber nicht auf das Intervall $[-1,1]$ begrenzt.

f Zurück zum Begriff des Risikos. Für kleine Risiken ist $\pi_{1+} \approx \pi_{12}$ und ebenso $\pi_{2+} \approx \pi_{22}$. Deshalb wird das relative Risiko näherungsweise gleich

$$\frac{\pi_{1|1}}{\pi_{1|2}} = \frac{\pi_{11}\pi_{2+}}{\pi_{1+}\pi_{21}} \approx \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} \,,$$

also gleich dem Doppelverhältnis.

g Wenn man die Randverteilung der Variablen A ändert, die bedingten Wahrscheinlichkeiten von B gegeben A aber unverändert lässt, so ändert sich das Doppelverhältnis nicht. Das erweist sich als sehr nützlich, wenn man an geschichtete Stichproben denkt: Wenn man die Schichten verschieden intensiv untersucht, ändert man dadurch zwar die π_{h+}, aber nicht die $\pi_{k|h}$, und die Doppelverhältnisse bleiben gleich!

h Wenn **mehr als zwei Klassen** für die Faktoren vorliegen, ist die sinnvolle Definition von odds ratios nicht mehr eindeutig. Man kann für jede Kombination von Klassen $[h, k]$ das Doppelverhältnis θ_{hk} für die Ergebnisse $B = k$ und $B \neq k$ für $A = h$ gegenüber $A \neq h$ bilden und erhält

$$\theta_{hk} = \frac{\pi_{hk} \sum_{h' \neq h, k' \neq k} \pi_{h'k'}}{(\pi_{h+} - \pi_{hk})(\pi_{+k} - \pi_{hk})} .$$

Die Doppelverhältnisse hängen dann wieder nicht von den Randsummen ab.

Eine andere sinnvolle Definition lautet

$$\theta_{hk,h'k'} = \frac{P\langle B = k \mid A = h\rangle}{P\langle B = k'|A = h\rangle} \Big/ \frac{P\langle B = k|A = h'\rangle}{P\langle B = k'|A = h'\rangle} = \frac{\pi_{k|h}}{\pi_{k'|h}} \Big/ \frac{\pi_{k|h'}}{\pi_{k'|h'}} = \frac{\pi_{hk}\pi_{h'k'}}{\pi_{h'k}\pi_{hk'}}$$

– das heisst, man vergleicht nur die Populationen von 2 Gruppen mit einander und lässt alle übrigen Beobachtungen unberücksichtigt.

Unabhängigkeit der beiden Faktoren bedeutet, dass alle Doppelverhältnisse gleich 1 sind.

Es gibt Vorschläge für Gesamt-Masse der Abhängigkeit zwischen kategoriellen Variablen. Wir verweisen auf Agresti, 2002, 2.3.

i Die Doppelverhältnisse müssen in den Anwendungen ja **geschätzt** werden. Es ist zunächst naheliegend, statt der Wahrscheinlichkeiten π_{hk} jeweils relative Häufigkeiten N_{hk}/n in die Definition einzusetzen. Da $N_{hk} = 0$ werden kann, muss man diesen Vorschlag abändern: Man schätzt

$$\widehat{\theta}_{hk} = \frac{(N_{hh} + 0.5)\,(N_{kk} + 0.5)}{(N_{hk} + 0.5)\,(N_{kh} + 0.5)} .$$

Diese Schätzungen weichen natürlich um eine zufällige Grösse von ihrem Modellwert ab. Die Streuung der Abweichung hängt von den Randsummen ab, im Gegensatz zum zu schätzenden Parameter selbst!

j* Weitere Abhängigkeitsmasse siehe Clogg and Shihadeh (1994).

7.5 Anmerkungen zu medizinischen Anwendungen

a In der Studie zum Herzinfarkt-Risiko (7.3.j) wurde eine Gruppe von Patientinnen, die einen Infarkt erlitten hatten, verglichen mit einer Gruppe von Frauen, die davon nicht betroffen waren. Eine solche Untersuchung wird **retrospektive Studie** (oder nach dem englischen *case control study* auch Fall-Kontroll-Studie) genannt; man versucht nach der Manifestation der Krankheit rückblickend zu ergründen, welche Faktoren sie begünstigt haben.

Aus der genannten Studie lässt sich das Risiko für einen Herzinfarkt nicht abschätzen, denn der Anteil der Frauen mit Herzinfarkt wurde ja durch den Rahmen der Untersuchung auf 58/224=26% festgelegt. Das ist glücklicherweise nicht das Risiko für einen Herzinfarkt! Was sich aus einer retrospektiven Studie korrekt schätzen lässt, sind Doppelverhältnisse, die die Erhöhung des Risikos durch die untersuchten „Risikofaktoren“ messen.

Wie für die meisten Krankheiten ist auch für den Herzinfarkt bei Frauen das absolute Risiko in der Bevölkerung bekannt. Aus einer entsprechenden Angabe und einem Doppelverhältnis kann man die Risiken für die untersuchten Gruppen bestimmen (siehe Übungen).

b Ein absolutes Risiko kann man auch schätzen, wenn man eine Zufallsstichprobe aus der Bevölkerung zieht. Eine solche Vorgehensweise nennt man auch **Querschnittstudie** *(cross sectional study).* Sie eignet sich allerdings nur für verbreitete Krankheiten, da sonst eine riesige Stichprobe gezogen werden muss, um wenigstens einige Betroffene drin zu haben. Wenn man untersuchen will, wie die Lebensgewohnheiten mit einer Krankheit zusammenhängen, muss man ausserdem mit der Schwierigkeit rechnen, dass sich die Leute nur schlecht an ihre früheren Gewohnheiten erinnern und dass diese Erinnerung ausserdem durch die Krankheit beeinflusst sein könnte.

c Zu präziseren Daten gelangt man – allerdings mit viel grösserem Aufwand – mit einer so genannten **Kohorten-Studie**: Eine (grosse) Gruppe von Menschen (Kohorte) wird ausgewählt aufgrund von Merkmalen, die mit der Krankheit nichts zu tun haben und bevor die Krankheit bei jemandem von ihnen ausgebrochen ist. Im Idealfall zieht man eine einfache Stichprobe aus einer Grundgesamtheit, über die man etwas aussagen möchte. Die Ausgangslage wird durch die Erfassung der Lebensgewohnheiten oder -umstände u.a. festgehalten. Nach oft recht langer Zeit untersucht man, welche Personen bestimmte Krankheitssymptome entwickelt haben, und prüft, ob sich Gruppen mit verschiedenen Ausgangssituationen diesbezüglich unterscheiden. Ein allfälliger Unterschied hängt mit der Ausgangssituation direkt oder indirekt zusammen.

d Die präzisesten Schlussfolgerungen erlauben die **klinischen Studien** *(clinical trials):* Ein Kollektiv von Patienten wird festgelegt, beispielsweise alle Patienten, die mit bestimmten Symptomen in eine Klinik eintreten. Sie werden mit einem Zufallsmechanismus (Zufallszahlen) einer Behandlungsgruppe zugeteilt. Wenn sich Krankheitsmerkmale nach erfolgter Behandlung in den verschiedenen Gruppen unterschiedlich zeigen, kommt wegen der zufälligen Zuordnung nur die Behandlung als Ursache dafür in Frage. Diese Untersuchungen eignen sich deshalb, um die Wirksamkeit und die Nebenwirkungen von Medikamenten und anderen Behandlungen genau zu erfassen.

e Die Kohorten- und die klinischen Studien werden im Gegensatz zu den retrospektiven Studien als **prospektiv** bezeichnet, da man die Personen in die Untersuchung einbezieht, wenn die unterschiedlichen Behandlungen oder Bedingungen noch in der Zukunft liegen. Schlüsse auf **Wirkungszusammenhänge** sind nur für die klinischen Studien zulässig. Die andern drei Typen von Studien werden meist verwendet, um Fragestellungen der **Präventivmedizin** zu untersuchen; sie gehören zum Gebiet der **Epidemiologie**.

8 Zweiwertige Zielgrössen, logistische Regression

8.1 Einleitung

a Die **Regressionsrechnung** ist wohl die am meisten verwendete und am besten untersuchte Methodik in der Statistik. Es wird der Zusammenhang zwischen einer **Zielgrösse** (allenfalls auch mehrerer solcher Variablen) und einer oder mehreren **Eingangsgrössen** oder **erklärenden Grössen** untersucht.

Wir haben die multiple lineare Regression ausführlich behandelt und dabei vorausgesetzt, dass die Zielgrösse eine kontinuierliche Grösse sei. Nun wollen wir andere Fälle behandeln – zunächst den Fall einer **binären** (zweiwertigen) **Zielgrösse**. Viele Ideen der multiplen linearen Regression werden wieder auftauchen; einige müssen wir neu entwickeln. Wir werden uns wieder kümmern müssen um

- Modelle,
- Schätzungen, Tests, Vertrauensintervalle für die Parameter,
- Residuen-Analyse,
- Modellwahl.

b ▷ **Beispiel Frühgeburten**. Von welchen Eingangsgrössen hängt das Überleben von Frühgeburten ab? Hibbard (1986) stellte Daten von 247 Säuglingen zusammen. In Abbildung 8.1.b sind die beiden wichtigsten Eingangsgrössen, Gewicht und Alter, gegeneinander aufgetragen. Das Gewicht wurde logarithmiert. Die überlebenden Säuglinge sind durch einen offenen Kreis markiert. Man sieht, dass die Überlebenschancen mit dem Gewicht und dem Alter steigen – was zu erwarten war.

In der Abbildung wird auch das Ergebnis einer logistischen Regressions-Analyse gezeigt, und zwar mit „Höhenlinien“ der geschätzten Wahrscheinlichkeit des Überlebens. ◁

c Die Zielgrösse Y ist also eine zweiwertige (binäre) Zufallsvariable. Wir codieren die beiden Werte als 0 und 1. Im Beispiel soll $Y_i = 1$ sein, wenn das Baby überlebt, und andernfalls $= 0$. Die Verteilung einer binären Variablen ist die einfachste Verteilung, die es gibt. Sie ist durch die Wahrscheinlichkeit $P\langle Y = 1\rangle$ festgelegt, die wir kurz mit π bezeichnen. Es gilt $P\langle Y = 0\rangle = 1-\pi$. Diese einfachste Verteilung wird **Bernoulli-Verteilung** genannt; ihr Parameter ist π.

d Wir wollten untersuchen, wie die Wahrscheinlichkeit $P\langle Y_i = 1\rangle$ von den Eingangsgrössen abhängt. Wir suchen also eine Funktion h mit

$$P\langle Y_i = 1\rangle = h\left\langle x_i^{(1)}, x_i^{(2)}, ..., x_i^{(m)}\right\rangle \ .$$

Könnten wir die multiple lineare Regression anwenden? – Das ist schwierig, denn es gibt keine natürliche Aufteilung $Y_i = h\left\langle x_i^{(1)}, ..., x_i^{(m)}\right\rangle + E_i$ in Regressionsfunktion h und Zufallsabweichung E_i.

Man kann aber die Erwartungswerte betrachten. Es gilt gemäss der Regression mit normalverteilten Fehlern

$$\mathcal{E}\langle Y_i\rangle = h\left\langle x_i^{(1)}, x_i^{(2)}, ..., x_i^{(m)}\right\rangle \ .$$

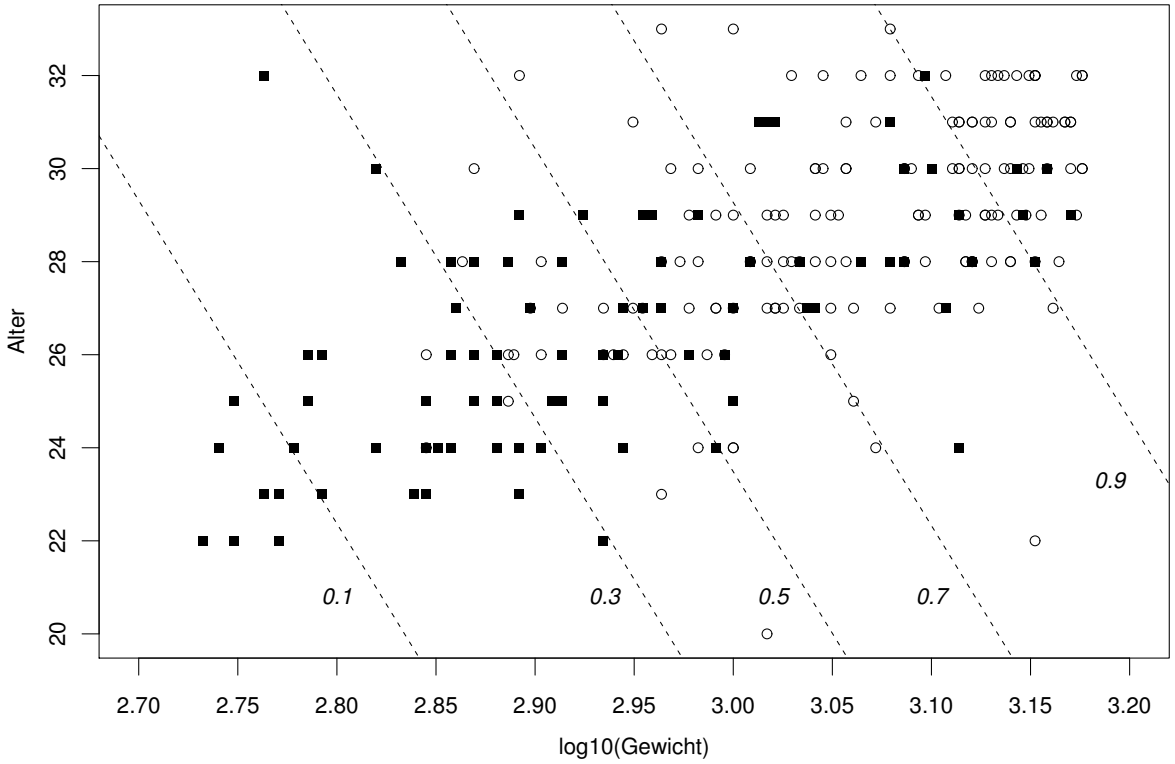


Abbildung 8.1.b: Logarithmiertes Gewicht und Alter im Beispiel der Frühgeburten. Die Überlebenden sind mit ∘, die anderen mit □ markiert. Die Geraden zeigen die Linien gleicher Überlebenswahrscheinlichkeiten (0.1, 0.3, 0.5, 0.7, 0.9) gemäss dem geschätzten logistischen Modell.

Für eine binäre Grösse Y_i gilt

$$\mathcal{E}\langle Y_i\rangle = 0 \cdot P\langle Y_i = 0\rangle + 1 \cdot P\langle Y_i = 1\rangle = P\langle Y_i = 1\rangle \ .$$

Also kann man in der ersten Gleichung $P\langle Y_i = 1\rangle$ durch $\mathcal{E}\langle Y_i\rangle$ ersetzen. In diesem Sinne sind die beiden Modelle gleich.

e In der multiplen linearen Regression wurde nun für h die lineare Form vorausgesetzt,

$$h\left\langle x^{(1)}, x^{(2)}, ..., x^{(m)}\right\rangle = \beta_0 + \beta_1 x^{(1)} + \beta_2 x^{(2)} + \ldots + \beta_m x^{(m)}$$

Können wir eine solche Funktion h für die Wahrscheinlichkeit $P\langle Y_i = 1\rangle$ brauchen? – Leider nein: Wenn ein $\beta_j \neq 0$ ist, werden für genügend extreme $x^{(j)}$-Werte die Grenzen 0 und 1, die für eine Wahrscheinlichkeit gelten müssen, überschritten.

In der linearen Regression wurden Transformationen der Zielgrösse in Betracht gezogen, um die Gültigkeit der Annahmen zu verbessern. Ebenso werden wir jetzt die Wahrscheinlichkeit $P\langle Y_i = 1\rangle$ so transformieren, dass ein lineares Modell sinnvoll erscheint.

f **Modell**. Eine übliche Transformation, die Wahrscheinlichkeiten (oder anderen Grössen, die zwischen 0 und 1 liegen) Zahlen mit unbegrenztem Wertebereich zuordnet, ist die so genannte **Logit-Funktion**

$$g\langle\pi\rangle = \log\left\langle \frac{\pi}{1-\pi} \right\rangle = \log\langle\pi\rangle - \log\langle 1-\pi\rangle \ .$$

Sie ordnet den Wahrscheinlichkeiten π das logarithmierte **Wettverhältnis** (die log odds) zu (7.4.e).

Für $g\langle P\langle Y_i=1\rangle\rangle$ können wir nun das einfache und doch so flexible Modell ansetzen, das sich bei der multiplen linearen Regression bewährt hat. Das Modell der logistischen Regression lautet

$$g\langle P\langle Y_i=1\rangle\rangle = \log\left\langle \frac{P\langle Y_i=1\rangle\rangle}{\langle 1-P\langle Y_i=1\rangle} \right\rangle = \beta_0+\beta_1 x_i^{(1)}+\beta_2 x_i^{(2)}+\ldots+\beta_m x_i^{(m)} \ .$$

Die rechte Seite heisst auch **linearer Prädiktor** und wird mit η_i (sprich „äta“) bezeichnet,

$$\eta_i = \beta_0+\beta_1 x_i^{(1)}+\beta_2 x_i^{(2)}+\ldots+\beta_m x_i^{(m)} \ .$$

Mit den Vektoren $\underline{x}_i = [1, x_i^{(1)}, x_i^{(2)}, ..., x_i^{(m)}]^T$ und $\underline{\beta} = [1, \beta_1, \beta_2, ..., \beta_m]^T$ kann man das abkürzen zu

$$\eta_i = \underline{x}_i^T \underline{\beta} \ .$$

Wie in der linearen Regression wird vorausgesetzt, dass die Beobachtungen Y_i stochastisch unabhängig sind.

An die X-Variablen werden ebenso wenige Anforderungen gestellt wie in der multiplen linearen Regression 3.2 Es können auch nominale Variable (Faktoren) (3.2.e) oder abgeleitete Terme (quadratische Terme, 3.2.v, Wechselwirkungen, 3.2.t) verwendet werden.

Es ist nützlich, wie in der linearen Regression zwischen den **Eingangsgrössen** und den daraus gebildeten X-Variablen oder **Regressoren** zu unterscheiden.

g Die Funktion g, die die Erwartungswerte $\mathcal{E}\langle Y_i\rangle$ in Werte des linearen Prädiktors verwandelt, nennt man die **Link-Funktion**. Die logistische Funktion ist zwar die üblichste, aber nicht die einzige geeignete Link-Funktion für binäre Zielgrössen. Im Prinzip eigenen sich alle strikt monotonen Funktionen, die den möglichen Werte zwischen 0 und 1 alle Zahlen zwischen $-\infty$ und $+\infty$ zuordnen – genauer, für die $g\langle 0\rangle = -\infty$ und $g\langle 1\rangle = \infty$ ist, vergleiche 8.2.j.

h ▷ Im **Beispiel der Frühgeburten** (8.1.b) wird die Wahrscheinlichkeit des Überlebens mit den weiter unten besprochenen Methoden geschätzt als

$$g\langle P\langle Y=1 \mid \mathtt{log}_{10}\langle \mathtt{Gewicht}\rangle, \mathtt{Alter}\rangle\rangle = -33.94+10.17\cdot \mathtt{log}_{10}\langle \mathtt{Gewicht}\rangle + 0.146\cdot \mathtt{Alter} \ .$$

Die Linien gleicher geschätzter Wahrscheinlichkeit wurden in Abbildung 8.1.b bereits eingezeichnet. Abbildung 8.1.h zeigt die Beobachtungen und die geschätzte Wahrscheinlichkeit, aufgetragen gegen den linearen Prädiktor $\eta = -33.94+10.17\cdot \mathtt{log}_{10}\langle \mathtt{Gewicht}\rangle + 0.146\cdot \mathtt{Alter}$. ◁

i In der Multivariaten Statistik wird die **Diskriminanzanalyse** für zwei Gruppen behandelt. Wenn man die Gruppen-Zugehörigkeit als (binäre) Zielgrösse Y_i betrachtet, kann man für solche Probleme auch die logistische Regression als Modell verwenden. Die multivariaten Beobachtungen $x_i^{(j)}$, aus denen die Gruppenzugehörigkeit ermittelt werden soll, sind jetzt die Eingangs-Variablen der Regression. Der lineare Prädiktor übernimmt die Rolle der Diskriminanzfunktion, die ja (in der Fisherschen Diskriminanzanalyse) ebenfalls linear in den $x_i^{(j)}$ war. Die Beobachtungen, für die $\widehat{\eta}_i > c$ mit $c=0$ (oder allenfalls einer anderen geeigneten Grenze c) gilt, werden der einen, die übrigen der andern Gruppe zugeordnet.

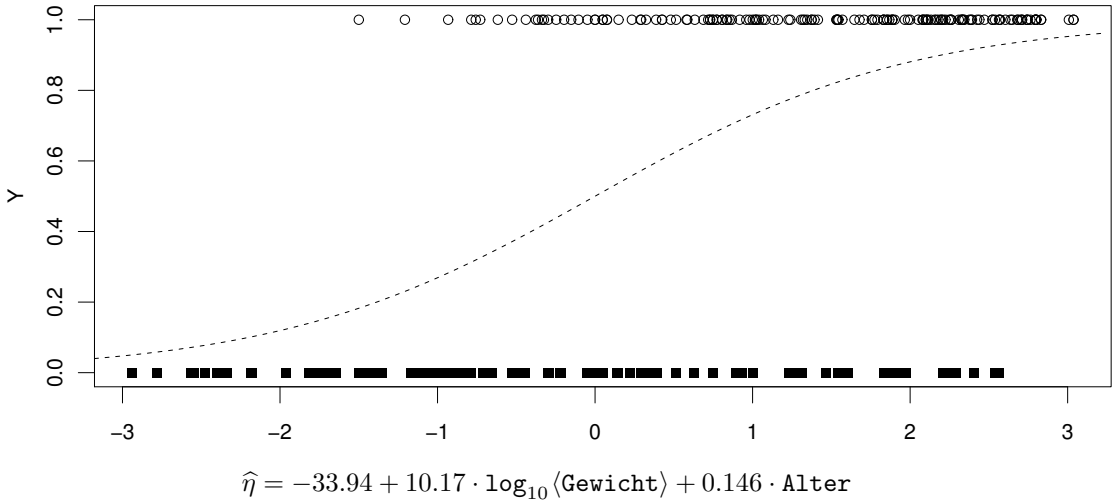


Abbildung 8.1.h: Die geschätzte Wahrscheinlichkeit $P\langle Y_i = 1\rangle$ als Funktion des linearen Prädiktors, zusammen mit den Beobachtungen, im Beispiel der Frühgeburten

j **Typische Anwendungen** für die logistische Regression sind:

- In toxikologischen Untersuchungen Toxikologie wird die Wahrscheinlichkeit festgestellt, mit der eine Maus bei einer bestimmten Giftkonzentration überlebt (oder stirbt). Stichwort **Dosis-Wirkungskurven** (dose-response curves).
- In der Medizin denken wir lieber an den entgegengesetzten Fall: Wird ein Patient bei einer bestimmten Konzentration eines Medikaments innerhalb einer vorgegebenen Zeit gesund oder nicht?
- Oft ist von Interesse, mit welcher Wahrscheinlichkeit Geräte in einer bestimmten Zeitperiode ausfallen, gegeben einflussreiche Grössen wie z.B. die Temperatur.
- In der **Qualitätskontrolle** wird das Auftreten eines Fehlers an einem Produkt untersucht, z.B. vergleichend für verschiedene Herstellungsverfahren.
- In der Biologie stellt sich häufig die Frage, ob ein bestimmtes Merkmal bei Lebewesen vorhanden ist und inwieweit ein Unterschied beispielsweise zwischen weiblichen und männlichen Lebewesen besteht.
- Im Kreditgeschäft oder im Customer relationship management sollen die „guten“ von den „schlechten“ Kunden getrennt werden.
- Wie gross ist die Wahrscheinlichkeit, dass es morgen regnet, wenn man berücksichtigt, wie das Wetter heute ist? Allgemein soll die Zugehörigkeit zu einer von zwei Gruppen erfasst und es soll untersucht werden, inwieweit sie durch gegebene Eingangsgrössen genauer bestimmt werden kann.

k **Ausblick.** In der logistischen Regression wird also eine binäre Zielgrösse untersucht.

In anderen Situationen *zählt* man Fälle (Individuen, Einheiten) mit bestimmten Eigenschaften. Das führt zu ähnlichen Schwierigkeiten bei Verwendung von Kleinsten Quadraten und zu Modellen, in denen die Zielgrösse Poisson-verteilt ist. Die für diese Situation geeignete Methodik heisst **Poisson-Regression**.

Solche Modelle dienen auch der Analyse von **Kontingenztafeln**, die in den Sozialwissenschaften eine wesentliche Rolle spielen. Sie heissen dann **log-lineare Modelle**. Wir werden sie in Kapitel 10.S.0.c ausführlicher behandeln.

Logistische Regression, Poisson-Regression und log-lineare Modelle bilden Spezialfälle des **Verallgemeinerten Linearen Modells**. Die statistische Methodik kann zum grossen Teil allgemein für alle diese Modelle formuliert werden. Wir behandeln hier zuerst den wichtigsten Spezialfall, die logistische Regression, werden aber teilweise auf Theorie verweisen, die allgemein für Verallgemeinerte Lineare Modelle gilt und deshalb dort behandelt wird.

l **Literatur.**

Entsprechend dieser Einordnung gibt es umfassende und spezialisiertere Bücher:

- Schwerpunktmässig mit logistischer Regression befassen sich Cox (1989) und Collet (1991, 1999). Beide Bücher sind gut zu lesen und enthalten auch wertvolle Tipps zur Datenanalyse. Umfassender ist das Buch von Agresti (2002). Es behandelt auch log-lineare Modelle. Die einfachere Variante Agresti (2007) ist sehr zu empfehlen.
- Bücher über Generalized Linear Models enthalten jeweils mindestens ein Kapitel über logistische Regression. Das klassische Buch von McCullagh and Nelder (1989) entwickelt die grundlegende Theorie und ist „trotzdem“ gut verständlich geschrieben. Das Kapitel über logistische Regression (“Binary Data”) behandelt dieses Thema in vorzüglicher Art. Eine elegante, kurze Abhandlung der Theorie bietet Dobson (2002).

8.2 Betrachtungen zum Modell

a Im Modell der logistischen Regression ist das logarithmierte Wettverhältnis gleich dem linearen Prädiktor η_i (8.1.f)

Umgekehrt kann man auch aus solchen η-Werten auf die Wahrscheinlichkeiten zurückschliessen. Dazu braucht man die „**inverse Link-Funktion**“, also die Umkehrfunktion

$$g^{-1}\langle\eta\rangle = \frac{\exp\langle\eta\rangle}{1+\exp\langle\eta\rangle} ,$$

die so genannte **logistische Funktion**, die der logistischen Regression den Namen gegeben hat. Ihre Form ist durch die Linie in Abbildung 8.1.h gegeben.

b **Interpretation der Koeffizienten.** Die logarithmierten Wettverhältnisse für $Y_i = 1$ sind, wie gesagt, eine lineare Funktion der Prädiktoren $x_i^{(j)}$. In Analogie zur linearen Regression können wir jetzt die Wirkung der einzelnen x-Variablen formulieren: Erhöht man $x^{(j)}$ um eine Einheit, dann erhöht sich das logarithmierte Wettverhältnis zu Gunsten von $Y = 1$ um β_j – wenn alle anderen $x^{(k)}$ dabei gleich bleiben. (Das Letztere ist nicht immer möglich. Beispielsweise ist ja in der quadratischen Regression $x^{(2)} = (x^{(1)})^2$.)

Für die unlogarithmierten Wettverhältnisse gilt

$$\begin{aligned} \operatorname{odds}\langle Y=1 \mid \underline{x}\rangle &= \frac{P\langle Y=1\rangle}{P\langle Y=0\rangle} = \exp\Big\langle \beta_0 + \sum\nolimits_j \beta_j x^{(j)} \Big\rangle = e^{\beta_0}\cdot e^{\beta_1 x^{(1)}}\cdot \ldots \cdot e^{\beta_m x^{(m)}} \\ &= e^{\beta_0}\cdot \exp\langle\beta_1\rangle^{x^{(1)}} \cdot \ldots \cdot \exp\langle\beta_m\rangle^{x^{(m)}} . \end{aligned}$$

Erhöht man $x^{(j)}$ um eine Einheit, dann erhöht sich deshalb das Wettverhältnis zu Gunsten von $Y = 1$ um den Faktor e^{β_j}. Anders ausgedrückt: Setzt man das Wettverhältnis für den erhöhten Wert $x^{(j)} = x + 1$ zum Wettverhältnis für den Ausgangswert $x^{(j)} = x$ ins Verhältnis, so erhält man

$$\frac{\operatorname{odds}\big\langle Y=1 \mid x^{(j)} = x+1\big\rangle}{\operatorname{odds}\big\langle Y=1 \mid x^{(j)} = x\big\rangle} = e^{\beta_j} .$$

Solche Quotienten von Wettverhältnissen haben wir unter dem Namen **Doppelverhältnisse** oder **odds ratios** in 7.4.c eingeführt.

c ▷ Im **Beispiel** (8.1.b) lassen sich die Schätzungen (aus 8.3.h) folgendermassen interpretieren: Für ein Individum mit $\texttt{log}_{10}\langle\texttt{Gewicht}\rangle = 3.1, \texttt{Alter} = 28$ erhält man als Schätzung für das logarithmierte Wettverhältnis $-33.94 + 10.17 \cdot 3.1 + 0.146 \cdot 28 = 1.68$ und damit ein Wettverhältnis für das Überleben von $\exp\langle 1.68\rangle = 5.4$. Die geschätzte Wahrscheinlichkeit für das Überleben beträgt $g^{-1}\langle 5.4\rangle = 0.84$. Vergleicht man nun dieses Wettverhältnis mit dem eines Individuums mit dem gleichen Alter und $\texttt{log}_{10}\langle\texttt{Gewicht}\rangle = 2.9$, dann erhält man als odds ratio den Faktor $\exp\langle 10.17 \cdot (-0.2)\rangle = 0.13$, d.h. das Wettverhältnis im zweiten Fall ist auf 13% des vorherigen gesunken und wird $0.13 \cdot 5.4 = 0.70$, und die entsprechenden Wahrscheinlichkeit wird $0.70/1.70 = 0.41$. ◁

d ▷ Im **Beispiel der Umweltumfrage** (7.1.c) sollte die Abhängigkeit der Zielgrösse „Beeinträchtigung“ von der Schulbildung erfasst werden. Die Zielgrösse hat hier vier mögliche geordnete Werte. Wir machen für die folgenden Betrachtungen daraus eine zweiwertige Variable, indem wir je zwei Kategorien zusammenfassen; später soll die feinere Unterteilung berücksichtigt werden.

Im logistischen Regressionsmodell bildet jede antwortende Person eine Beobachtung Y_i mit zugehörigen Werten $\underline{x}_i$ der Regressoren. ◁

e Die logistische Regression eignet sich also auch zur Analyse von **Kontingenztafeln**, sofern eine „Dimension“ der Tafel als Zielgrösse aufgefasst wird und nur 2 Stufen zeigt. Man kann von **logistischer Varianzanalyse** sprechen. Die Analyse von Kontingenztafeln wird im Kapitel über log-lineare Modelle (10.S.0.c) ausführlicher behandelt.

f **Gruppierte Beobachtungen**. Wenn mehrere (m_ℓ) Beobachtungen Y_i zu gleichen Bedingungen $\underline{x}_i = \underline{\widetilde{x}}_\ell$ gemacht werden, können wir sie zusammenfassen und die Anzahl der „Erfolge“, also die Zahl der i mit $Y_i = 1$, festhalten. Wir ziehen es vor, statt dieser Anzahl den Anteil der Erfolge als neue Grösse einzuführen; man kann diesen schreiben als

$$\widetilde{Y}_\ell = \tfrac{1}{m_\ell} \sum\nolimits_{i\,:\,\underline{x}_i = \underline{\widetilde{x}}_\ell} Y_i \ .$$

Das ist in der Kontingenztafel bereits geschehen: Alle Personen mit gleicher Schulbildung $\widetilde{x}_\ell$ wurden zusammengefasst, und die Zahlen in den Spalten liefern die Angaben für $\widetilde{Y}_\ell$: Wir haben für die gegenwärtige Betrachtung die letzten drei Spalten zusammengefasst. Die Summe über die drei Zahlen, dividiert durch die Randsumme, liefert den Anteil der mindestens „etwas“ beeinträchtigten Personen. Werden mehrere Eingangsgrössen betrachtet, so ist $\widetilde{Y}_\ell$ der Anteil der beeinträchtigten Personen i unter den m_ℓ Befragten, die gleiche Schulbildung $x_i^{(1)} = \widetilde{x}_\ell^{(1)}$, gleiches Geschlecht $x_i^{(2)} = \widetilde{x}_\ell^{(2)}$ und Alter $x_i^{(3)} = \widetilde{x}_\ell^{(3)}$ haben – allgemein, der Anteil der „Erfolge“ unter den m_ℓ „Versuchen“, die unter den Bedingungen $\underline{\widetilde{x}}_\ell$ durchgeführt wurden.

g Wenn für die einzelnen Beobachtungen Y_i das Modell der logistischen Regression vorausgesetzt wird, sind die Y_i mit $\underline{x}_i = \underline{\widetilde{x}}_\ell$ unabhängige Versuche mit gleicher Erfolgswahrscheinlichkeit $\widetilde{\pi}_\ell = h\langle\underline{\widetilde{x}}_\ell\rangle$. Die *Anzahl* der Erfolge $m_\ell\widetilde{Y}_\ell$ ist also **binomial verteilt**,

$$m_\ell\widetilde{Y}_\ell \sim \mathcal{B}\langle m_\ell, \widetilde{\pi}_\ell\rangle \ , \qquad g\langle\widetilde{\pi}_\ell\rangle = \underline{\widetilde{x}}_\ell^T\underline{\beta} \ .$$

Es gilt

$$\mathcal{E}\left\langle\widetilde{Y}_\ell\right\rangle = \tfrac{1}{m_\ell} \sum\nolimits_{i\,:\,\underline{x}_i = \underline{\widetilde{x}}_\ell} \mathcal{E}\langle Y_i\rangle = \widetilde{\pi}_\ell \ .$$

Ein Vorteil von gruppierten Daten besteht darin, dass man sie kompakter und informativer darstellen kann. Zudem sind manche Approximationen, die wir im Rahmen der Residuen-Analyse und unter dem Stichwort Anpassungsgüte besprechen werden, nur für gruppierte Daten aussagekräftig.

Es ist wichtig, anzumerken, dass das Modell sich durch die Gruppierung der Daten nicht geändert hat.

Für „Gruppen“ mit nur einer Beobachtung ($m_\ell = 1$) wird Y_ℓ wieder zweiwertig und die Binomialverteilung zur Bernoulli-Verteilung (8.1.c).

h ▷ **Beispiel Frühgeburten**. Um ein anschauliches Beispiel zu erhalten, untersuchen wir das Überleben von Frühgeburten nur als Funktion der Eingangsgrösse Gewicht. Wenn wir Klassen von je 100 g Gewicht bilden, können wir die Daten zu den Häufigkeiten zusammenfassen, die in Tabelle 8.2.h gezeigt werden, zusammen mit einem Ausschnitt aus den ursprünglichen Beobachtungen. Abbildung 8.2.h zeigt sie mit dem angepassten Modell in dieser Form. ◁

i	Y_i	weight	Age
1	1	1350	32
2	0	725	27
3	0	1090	27
4	0	1300	24
5	0	1200	31
...		...	
245	0	900	27
246	1	1150	27
247	0	790	27

(i)

ℓ	$\widetilde{x}_\ell$	m_ℓ	$m_\ell\widetilde{Y}_\ell$	$m_\ell(1-\widetilde{Y}_\ell)$
1	550	10	0	10
2	650	14	2	12
3	750	27	9	18
4	850	22	8	14
5	950	32	23	9
6	1050	28	21	7
7	1150	22	19	3
8	1250	26	19	7
9	1350	34	31	3
10	1450	32	29	3

(ii)

Tabelle 8.2.h: Beispiel Frühgeburten: Einige Einzel-Beobachtungen (i) und zusammengefasste Daten (ii). $m_\ell\widetilde{Y}_\ell$ ist die Anzahl Überlebende der insgesamt m_ℓ Kinder in der Gewichtsklasse ℓ mit mittlerem Gewicht $\widetilde{x}_\ell$

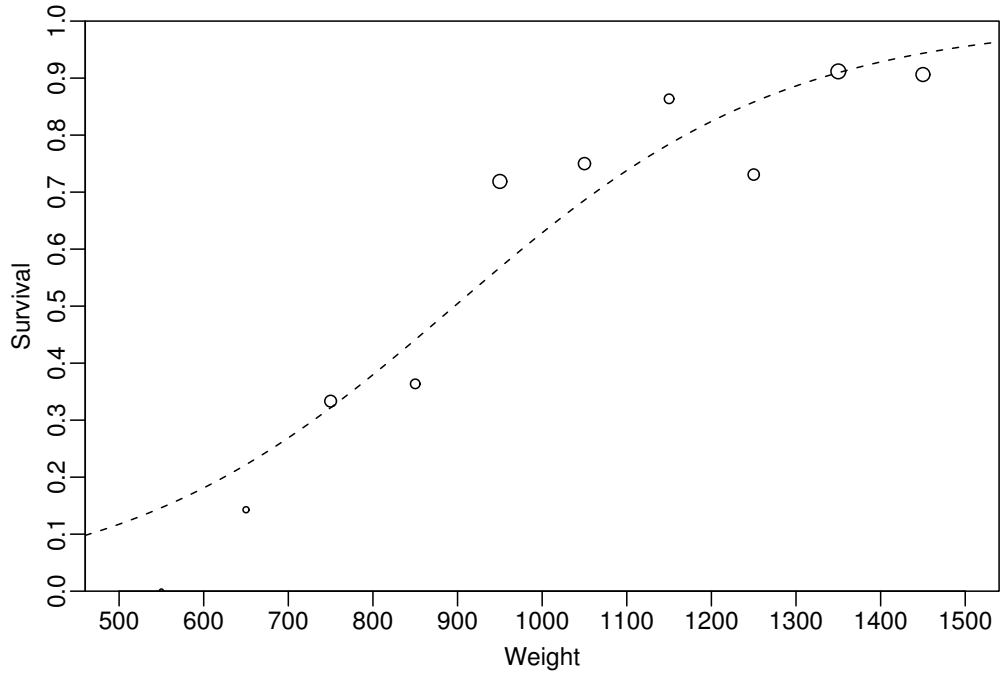


Abbildung 8.2.h: Überleben in Abhängigkeit vom Gewicht. Gruppierte Daten; die Fläche der Kreise ist proportional zur Anzahl Beobachtungen

i **Transformierte Beobachtungen.** Laut dem Modell sind die logit-transformierten Erwartungswerte $\widetilde{\pi}_\ell$ der „Erfolgsraten“ $\widetilde{Y}_\ell/m_\ell$ gleich einer linearen Funktion der $\widetilde{x}_\ell^{(j)}$. Im Fall einer einzigen Eingangsgrösse liegt es nahe, die beobachteten Werte $\widetilde{Y}_\ell/m_\ell$ selbst zu transformieren und gegen die Eingangs-Variable aufzutragen; im Falle von mehreren Eingangsgrössen kann man auf der horizontalen Achse stattdessen den linearen Prädiktor η verwenden. Es sollte sich dann statt des sigmoiden Zusammenhangs von Abbildung 8.2.h ein linearer ergeben.

Nun ist aber $g\langle 0\rangle$ und $g\langle 1\rangle$ für die Logit-Funktion nicht definiert, also erhält man für $\widetilde{Y}_\ell = 0$ und für $\widetilde{Y}_\ell = m_\ell$ keinen (endlichen) Wert der transformierten Grösse. Als pragmatischen Ausweg verwendet man die **empirischen Logits**

$$\log\left\langle \frac{\widetilde{Y}_\ell + 0.5}{m_\ell - \widetilde{Y}_\ell + 0.5} \right\rangle \ .$$

Abbildung 8.2.i zeigt die empirischen Logits für die Frühgeburtdaten und die angepasste lineare Funktion.

Wendet man auf die empirischen Logits eine gewöhnliche multiple Regression an, so erhält man eine alternative Schätzung der Koeffizienten. Sie bildet oft eine vernünftige Näherung für die optimalen Schätzwerte, die wir in 8.3.b besprechen werden. Für kleine m_ℓ ist die Übereinstimmung schlechter, und für ungruppierte, binäre Zielgrössen wird die Schätzung über Kleinste Quadrate von empirischen Logits unbrauchbar.

j **Modell der latenten Variablen**. Das logistische Regressionsmodell lässt sich noch von einer weiteren Überlegung her begründen: Man stellt sich vor, dass es eine nicht beobachtbare Variable Z_i gibt, die linear von den Regressoren abhängt,

$$Z_i = \widetilde{\beta}_0 + \sum\nolimits_j x_i^{(j)} \widetilde{\beta}_j + E_i = \widetilde{\eta}_i + E_i \ .$$

Die binäre Zielgrösse Y_i stellt fest, ob Z_i unterhalb oder oberhalb eines **Schwellenwertes** c liegt. Abbildung 8.2.j veranschaulicht diese Vorstellung.

Bei Pflanzen mag beispielsweise die Frosttoleranz eine kontinuierliche Grösse sein, die man in der Natur nicht messen kann. Man kann lediglich feststellen, ob die Pflenzen nach einem Frostereignis entsprechende Schäden zeigen, und gleichzeitig erklärende Variable aufnehmen, die die Pflanze selbst und ihre nähere Umgebung charakterisieren. Im Beispiel der Frühgeburten kann man sich eine Variable „Lebensenergie“ vorstellen, die einen Schwellenwert überschreiten muss, damit das Überleben gewährleistet ist.

Die Zielgrösse Y_i erfasst entsprechend dieser Idee, ob $Z_i \geq c$ gilt, und es wird

$$\pi_i = P\langle Y_i = 1\rangle = P\langle Z_i \geq c\rangle = P\langle E_i \geq c - \widetilde{\eta}_i\rangle = 1 - F_E\langle c - \widetilde{\eta}_i\rangle \ .$$

Dabei ist F_E die kumulative Verteilungsfunktion der E_i. Die Verteilung der binären Grösse Y und die der latenten Variablen Z hängen also direkt zusammen.

Setzt man $\beta_0 = \widetilde{\beta}_0 - c$ und $\beta_j = \widetilde{\beta}_j$, $j = 1, ..., m$, dann ergibt sich mit $\eta_i = \beta_0 + \sum_j \beta_j x_i^{(j)}$

$$P\langle Y_i = 1 \mid \underline{x}_i\rangle = 1 - F_E\langle -\eta_i\rangle \ .$$

Der Ausdruck $1 - F\langle -\eta\rangle$ ist selbst eine Verteilungsfunktion, nämlich diejenige von $-E$. Wenn wir diese Verteilungsfunktion gleich g^{-1} setzen, dann erhalten wir das Modell der logistischen Regression (8.1.f); die Funktion g ist die Umkehrfunktion der Verteilungsfunktion, also die entsprechende Quantil-Funktion. Wenn die E_i der **logistischen Verteilung** folgen, erhält man das logistische Regressionsmodell.

Je nach Annahme für die Verteilung der Zufallsfehler E_i ergibt sich ein anderes Regressionsmodell:

logistische Vert.	$\longrightarrow$ logistische Regression	$P\langle Y_i = 1\rangle = e^{\eta_i}/(1+e^{\eta_i})$
Normalvert.	$\longrightarrow$ Probitmodell	$P\langle Y_i = 1\rangle = \Phi\langle\eta_i\rangle$
Extremwertvert.	$\longrightarrow$ Komplementäres log-log Mod.	$P\langle Y_i = 1\rangle = 1 - \exp\langle -\exp\langle\eta_i\rangle\rangle$

8.3 Schätzungen und Tests

a Schätzungen und Tests beruhen auf der Methodik der Likelihood. Es existieren Programme (unterdessen in allen Statistik-Paketen, die diesen Namen verdienen), die es erlauben, Regressionen mit binären Variablen ebenso durchzuführen wie gewöhnliche lineare Regressionen.

b Die **Schätzung der Koeffizienten** erfolgt nach dem Prinzip der Maximalen Likelihood. Zur Erinnerung: Wir betrachten die Wahrscheinlichkeit für das beobachtete Ergebnis als Funktion der Parameter und suchen ihr Maximum. Die Wahrscheinlichkeit $P\langle Y_1 = y_1, Y_2 = y_2, ..., Y_n = y_n\rangle$ ist, da die Beobachtungen stochastisch unabhängig sind, gleich dem Produkt $\prod_i P\langle Y_i = y_i\rangle$. Logarithmiert man diesen Ausdruck, so verwandelt sich das Produkt in eine Summe. Deshalb ist es schlau, die logarithmierte Likelihood $\ell\ell = \sum_i \log\langle P\langle Y_i = y_i\rangle\rangle$ statt der unlogarithmierten zu maximieren.

Die Wahrscheinlichkeiten für die einzelnen Beobachtungen sind im logistischen Modell $P\langle Y_i = 1\rangle = \pi_i$ und $P\langle Y_i = 0\rangle = 1 - \pi_i$, wobei $\text{logit}\langle\pi_i\rangle = \underline{x}_i^T\underline{\beta}$ ist. Man kann dies auch ohne Fallunterscheidung hinschreiben als $P\langle Y_i = y_i\rangle = \pi_i^{y_i}(1-\pi_i)^{1-y_i}$. Die Beiträge der Beobachtungen zur logarithmierten Likelihood sind deshalb

$$\ell\ell_i\langle\underline{\pi}\rangle = \log\left\langle P\langle Y_i = y_i\rangle\right\rangle = y_i\log\langle\pi_i\rangle + (1-y_i)\log\langle 1-\pi_i\rangle \ ,$$

und die gesamte Log-Likelihood ist dann wie üblich die Summe aus diesen Beiträgen für die Einzelbeobachtungen,

$$\ell\ell\langle\underline{\pi}\rangle = \sum\nolimits_i \ell\ell_i\langle\underline{\beta}\rangle = \sum\nolimits_i (y_i\log\langle\pi_i\rangle + (1-y_i)\log\langle 1-\pi_i\rangle) \ .$$

Die Parameter $\underline{\beta}$ sind in den π_i „versteckt“, $\text{logit}\langle\pi_i\rangle = \underline{x}_i^T\underline{\beta}$.

Die Schätzung $\widehat{\underline{\beta}}$ ergibt sich durch Maximieren dieses Ausdrucks, also durch Ableiten und Null-Setzen.

c* Für gruppierte Daten waren die Grössen $m_\ell\widetilde{Y}_\ell$ binomial verteilt; die Wahrscheinlichkeiten sind deshalb

$$P\left\langle\widetilde{Y}_\ell = \widetilde{y}_\ell\right\rangle = \tbinom{m_\ell}{\widetilde{y}_\ell}\pi_\ell^{m_\ell\widetilde{y}_\ell}\cdot(1-\pi_\ell)^{m_\ell(1-\widetilde{y}_\ell)} \ .$$

Daraus erhält man

$$\ell\ell\langle\underline{\pi}\rangle = \sum\nolimits_\ell \Big(c_\ell + m_\ell\widetilde{y}_\ell\log\langle\widetilde{\pi}_\ell\rangle + m_\ell(1-\widetilde{y}_\ell)\log\langle 1-\widetilde{\pi}_\ell\rangle\Big)$$

mit $c_\ell = \log\left\langle\tbinom{m_\ell}{m_\ell\widetilde{y}_\ell}\right\rangle$.

d* Um den Ausdruck $\pi_i = g^{-1}\langle \underline{\widetilde{x}}_\ell^T \underline{\beta}\rangle$ nach β_j abzuleiten, benützt man die Kettenregel mit $dg^{-1}\langle\eta\rangle/d\eta = \exp\langle\eta\rangle/(1+\exp\langle\eta\rangle)^2 = \pi(1-\pi)$ und $\partial\eta_i/\partial\beta_j = x_i^{(j)}$. Man erhält

$$\frac{\partial \log\langle\pi_i\rangle}{\partial\beta_j} = \frac{1}{\pi_i}\cdot\pi_i(1-\pi_i)\frac{\partial\eta_i}{\partial\beta_j} = (1-\pi_i)\,x_i^{(j)}$$

und ebenso $\partial\log\langle 1-\pi_i\rangle/\partial\beta_j = -\pi_i\,x_i^{(j)}$. Deshalb ist

$$\frac{\partial\ell\ell\langle\underline{\pi}\rangle}{\partial\beta_j} = \sum\nolimits_{y_i=1}(1-\pi_i)\,x_i^{(j)} + \sum\nolimits_{y_i=0}(0-\pi_i)\,x_i^{(j)} = \sum\nolimits_i (y_i-\pi_i)\,x_i^{(j)}\ .$$

Die Maximum-Likelihood-Schätzung erhält man durch null setzen dieser Ausdrücke für alle j, was man zusammenfassen kann zu

$$\sum\nolimits_i (y_i - \widehat{\pi}_i)\,\underline{x}_i = \underline{0}\ .$$

Dies ist ein implizites Gleichungssystem für die in den $\widehat{\pi}_i$ versteckten Parameter $\beta^{(j)}$.

Geht man von gruppierten Beobachtungen aus, dann erhält man mit einer etwas komplizierteren Rechnung

$$\sum\nolimits_\ell m_\ell(\widetilde{y}_\ell - \widehat{\pi}_\ell)\,\underline{\widetilde{x}}_\ell = \underline{0}\ .$$

Es ist beruhigend, zu sehen, dass man das Gleiche erhält, wenn man die Summe in der vorhergehenden Gleichung zunächst über alle i bildet, für die $\underline{x}_i = \underline{\widetilde{x}}_\ell$ ist.

e **Berechnung.** Zur Lösung dieser Gleichungen braucht man ein iteratives Verfahren. Wie in der nichtlinearen Regression wird in jedem Schritt die Gleichung durch lineare Näherung so vereinfacht, dass sie zu einem linearen Regressionsproblem wird – hier zu einem mit Gewichten w_i. Wenn die Verbesserungsschritte schliesslich vernachlässigbar klein werden, ist die Lösung gefunden. Sie ist dann auch die exakte Lösung des genannten gewichteten linearen Regressionsproblems. Genaueres steht im Anhang 9.b.

f **Verteilung der geschätzten Koeffizienten.** In der multiplen linearen Regression konnte mit linearer Algebra recht einfach hergeleitet werden, dass der Vektor $\underline{\widehat{\beta}}$ der geschätzten Koeffizienten multivariat normalverteilt ist mit Erwartungswert $\underline{\beta}$ und Kovarianzmatrix $\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. Da die geschätzten Koeffizienten in der logistischen Regression die Lösung des näherungsweise äquivalenten gewichteten linearen Regressionsproblems sind, kann man daraus die Verteilung von $\underline{\widehat{\beta}}$ ableiten. Die geschätzten Koeffizienten sind also näherungsweise multivariat normalverteilt, haben genähert den Erwartungswert $\underline{\beta}$ und eine Kovarianzmatrix $\boldsymbol{V}^{(\beta)}$, die wir bei den Verallgemeinerten Linearen Modellen (9.3.e) angeben werden.

Die Näherung wird für grössere Stichproben immer genauer. Wie viele Beobachtungen es für eine genügende Näherung braucht, hängt von den Werten der Regressoren ab und ist deshalb nicht allgemein anzugeben.

g Genäherte **Tests** und **Vertrauensintervalle für die einzelnen Koeffizienten** erhält man aus diesen Angaben mit dem üblichen Rezept: Der Standardfehler von $\widehat{\beta}_j$ ist die Wurzel aus dem $j$ten Diagonalelement V_{jj} der angegebenen Kovarianzmatrix, und

$$T = \frac{\widehat{\beta}_j - \beta_j}{\sqrt{\widehat{V}_{jj}^{(\beta)}}}$$

hat eine genäherte Normalverteilung. Im Ausdruck für die Kovarianzmatrix müssen die geschätzten Koeffizienten eingesetzt werden, deshalb $\widehat{\boldsymbol{V}}^{(\beta)}$ statt $\boldsymbol{V}^{(\beta)}$. Da sie keine geschätzte Fehlervarianz $\widehat{\sigma}^2$ enthält, besteht kein theoretischer Grund, die Standard-Normalverteilung durch eine t-Verteilung zu ersetzen.

h Die **Computer-Ausgabe** enthält ähnliche Teile wie bei der gewöhnlichen linearen Regression. In `summary(r.babysurv)` (Tabelle 8.3.h) erscheint die Tabelle der geschätzten Koeffizienten, ihrer Standardfehler, der Werte der Teststatistiken und der P-Werte für die Hypothesen $\beta^{(j)} = 0$.

Auf den „Dispersion Parameter“ kommen wir päter zurück (9.2.f, 9.3.g, 9.4). Die „Null Deviance“ und die „Residual Deviance“ brauchen noch eine genauere Erklärung.

```
Call: glm(formula = Y ~ log10(Gewicht) + Alter, family = binomial,
          data = d.babysurv)
Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    -33.9449     4.9897   -6.80  1.0e-11 ***
log10(Gewicht)  10.1688     1.8812    5.41  6.5e-08 ***
Alter            0.1464     0.0745    1.96    0.049 *

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 318.42  on 245  degrees of freedom
Residual deviance: 235.89  on 243  degrees of freedom
AIC: 241.9

Number of Fisher Scoring iterations: 4
```

Tabelle 8.3.h: Computer-Ausgabe (leicht gekürzt) für das Beispiel Frühgeburten

i **Residuen-Devianz**. In der multiplen linearen Regression ist die Summe der Residuenquadrate ein Mass dafür, wie gut die Zielvariable durch die Einflussgrössen erklärt wird. In der logistischen Regression übernimmt die Residuen-Devianz diese Rolle. Sie ist für zusammengefasste, binomial verteilte $\widetilde{Y}_\ell$ definiert als 2 mal die Differenz zwischen der maximalen Log-Likelihood $\ell^{(M)}$ und dem Wert für das angepasste Modell,

$$D\left\langle \underline{\widetilde{y}}; \underline{\widehat{\pi}} \right\rangle := 2\left(\ell\ell^{(M)} - \ell\ell\left\langle \underline{\widehat{\beta}} \right\rangle \right) .$$

Was ist die maximale erreichbare Log-Likelihood? Es gilt ja $m_\ell \widetilde{Y}_\ell \sim \mathcal{B}\langle m_\ell, \widetilde{\pi}_\ell \rangle$. Wenn wir $\widetilde{\pi}_\ell$ für jede Gruppe frei wählen können, ist $\widetilde{\pi}_\ell = \widetilde{y}_\ell$ die Wahl, die die Likelihood maximiert.

* Diese erhält man, indem man in der Formel für $\ell\ell\langle \underline{\pi} \rangle$ (8.3.c) $\widetilde{\pi}_\ell$ durch $\widetilde{y}_\ell$ ersetzt. (Für $\widetilde{y}_\ell = 0$ und $\widetilde{y}_\ell = 1$ tritt $\log\langle 0 \rangle$ auf. Der Ausdruck wird aber in der Formel immer mit 0 multipliziert und die entsprechenden Terme können weggelassen werden.)

Setzt man dieses $\ell\ell^{(M)}$ und das erwähnte $\ell\ell\langle \underline{\pi} \rangle$ in die Definition der Devianz ein, so erhält man

$$D\left\langle \underline{\widetilde{y}}; \underline{\widehat{\pi}} \right\rangle = 2 \sum\nolimits_\ell \left(m_\ell \widetilde{y}_\ell \log\left\langle \frac{\widetilde{y}_\ell}{\widetilde{\pi}_\ell} \right\rangle + m_\ell (1 - \widetilde{y}_\ell) \log\left\langle \frac{1 - \widetilde{y}_\ell}{1 - \widetilde{\pi}_\ell} \right\rangle \right) .$$

Für ungruppierte, binäre Daten ergibt sich $\ell\ell^{(M)} = 0$ und somit
$D\left\langle \underline{\widetilde{y}}; \underline{\widehat{\pi}} \right\rangle = -2 \sum_i (y_i \log\langle \pi_i \rangle + (1 - y_i) \log\langle 1 - \pi_i \rangle)$.

j Die Devianz ist vor allem wertvoll beim Vergleich von geschachtelten Modellen. Für zwei Modelle, von denen das grössere (G) das kleinere (K) umfasst, kann man nach der allgemeinen Theorie des **Likelihood-Quotienten-Tests** prüfen, ob das grössere eine „echte“ Verbesserung bringt. Die Teststatistik ist

$$\begin{aligned} 2(\ell\ell^{(G)} - \ell\ell^{(K)}) &= 2(\ell\ell^{(M)} - \ell\ell^{(K)}) - 2(\ell\ell^{(M)} - \ell\ell^{(G)}) \\ &= D\left\langle \underline{\widetilde{y}}; \underline{\widehat{\pi}}^{(K)} \right\rangle - D\left\langle \underline{\widetilde{y}}; \underline{\widehat{\pi}}^{(G)} \right\rangle \end{aligned}$$

und wird als **Devianz-Differenz** bezeichnet. Sie ist asymptotisch chiquadrat-verteilt, wenn das kleine Modell stimmt; die Anzahl Freiheitsgrade ist, wie früher, gleich der Differenz der Anzahl Parameter in den beiden Modellen.

k Unter diesem Gesichtspunkt ist die **Residuen-Devianz** (8.3.i) die Teststatistik für den Likelihood-Quotienten-Test, der das angepasste Modell mit dem grösstmöglichen Modell vergleicht. Bei **gruppierten Daten** gibt dieses maximale Modell eine nicht zu unterbietende Streuung der Zielgrösse an, die sich aus der Binomialverteilung ergibt. Der Vergleich dieser minimalen Streuung mit der Streuung im angepassten Modell liefert eine Art „**Anpassungstest**“ (goodness of fit test), der sagt, ob die Streuung dem entspricht, was gemäss dem Modell der Binomialverteilung zu erwarten ist. Wenn die Streuung grösser ist, dann ist es sinnvoll, nach weiteren erklärenden Variablen zu suchen.

In der linearen Regression konnte man ebenfalls eine solche minimale Streuung erhalten, wenn mehrere Beobachtungen mit gleichen $\underline{x}_i$-Werten vorlagen, siehe 4.8.a

Eine genäherte Chiquadrat-Verteilung dieser Statistik ist nur gegeben, wenn die m_ℓ genügend gross sind. Es müssen also gruppierte Daten vorliegen mit genügend vielen Beobachtungen pro Gruppe (vergleiche 9.3.i). Deshalb muss ein hoher Wert für die Devianz nicht immer bedeuten, dass das Modell ungeeignet ist (vgl. McCullagh and Nelder (1989), Sect. 4.4.3 und 4.4.5).

l ▷ Im **Beispiel der Umweltumfrage** (8.2.d) kann man die Daten gruppieren. Damit die Gruppen nicht zu klein werden, soll das Alter in Klassen von 20 Jahren eingeteilt werden.

In der üblichen Computer-Ausgabe (Tabelle 8.3.l) sticht die Koeffizienten-Tabelle ins Auge, die wie in der multiplen linearen Regression Tests liefert, welche kaum interpretierbar sind. Die Eingangsgrössen sind ja Faktoren, und es macht wieder wenig Sinn, einen einzelnen Koeffizienten auf Verschiedenheit von 0 zu testen – ausser für die zweiwertige Eingangsgrösse Geschlecht.

```
Call:
glm(formula = cbind(Beeintr.gr, Beeintr.kl) ~ Schule +
    Geschlecht + Alter, family = binomial, data = d.umw1)

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     -1.6045     0.1656   -9.69  < 2e-16 ***
SchuleLehre     -0.1219     0.1799   -0.68  0.49803
Schuleohne.Abi   0.4691     0.1900    2.47  0.01355 *
SchuleAbitur     0.7443     0.2142    3.47  0.00051 ***
SchuleStudium    1.0389     0.2223    4.67  3.0e-06 ***
Geschlechtw      0.0088     0.1135    0.08  0.93818
Alter.L         -0.1175     0.1557   -0.75  0.45044
Alter.Q          0.1033     0.1304    0.79  0.42810
Alter.C          0.1436     0.1080    1.33  0.18364

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 105.95  on 38  degrees of freedom
Residual deviance:  36.71  on 30  degrees of freedom
AIC: 191.2
Number of Fisher Scoring iterations: 4
```

Tabelle 8.3.l: Computer-Ausgabe (gekürzt) für das Beispiel der Umweltumfrage

Die Residuen-Devianz ist mit 36.71 bei 30 Freiheitgraden im Bereich der zufälligen Streuung; der P-Wert ist 0.19. Das heisst, dass das Modell gut passt – aber nicht, dass keine weiteren erklärenden Variablen die Zielgrösse beeinflussen könnnten; wenn weitere Variable berücksichtigt werden, unterteilen sich die Anzahlen feiner, und das führt zu einem genaueren maximalen Modell. ◁

m **Für ungruppierte Daten macht dieser Anpassungstest keinen Sinn.** Für eine binäre Variable erhält man aus der Beobachtung nämlich keine Schätzung für ihre Varianz. (Es geht also nicht darum, dass die Näherung durch die Chiquadrat-Verteilung zu schlecht wäre.)

In Anlehnung an den gerade erwähnten Test in der gewöhnlichen linearen Regression kann man aber die Daten auf der Basis des geschätzten Modells gruppieren. In SAS werden nach Hosmer and Lemeshow (2000) die Beobachtungen aufgrund der angepassten Werte in 10 Gruppen mit (möglichst) gleich vielen Beobachtungen eingeteilt. Für jede Gruppe wird nun die Summe $\widetilde{Y}_\ell$ der „Erfolge“ Y_i gezählt und die Summe der geschätzten Wahrscheinlichkeiten $\widehat{\pi}_i$ gebildet. Auf diese Grössen wird dann der gewöhnliche Chiquadrat-Test angewandt, und zwar aufgrund von „ausgiebigen“ Simulationen mit 10-2=8 Freiheitsgraden.

Für die Einteilung in 10 gleich grosse Klassen habe ich keine Begründung gefunden. Ebenso wie für Anpassungstests für Verteilungen würde ich es vorziehen, an beiden Enden zwei kleine Klassen mit etwa 5 erwarteten „Erfolgen“ resp. „Misserfolgen“ zu machen und den Rest in 4 oder 5 gleich grosse Klassen einzuteilen.

n Der Vergleich zwischen einem grösseren und einem kleineren Modell wird gebraucht, um den **Einfluss einer nominalen Eingangsgrösse** auf die Zielgrösse zu prüfen – wie dies schon in der linearen Regression der Fall war. In der S-Sprache prüft die Funktion `drop1`, ob die einzelnen Terme einer Modell-Formel weggelassen werden können.

```
> drop1(r.umw,test="Chisq")

Single term deletions
Model:
cbind(Beeintr.gr, Beeintr.kl) ~ Schule + Geschlecht + Alter
           Df Deviance   AIC   LRT Pr(Chi)
<none>           36.7 191.2
Schule      4     89.4 235.9  52.7 9.7e-11 ***
Geschlecht  1     36.7 189.2 0.006    0.94
Alter       3     40.1 188.5   3.4    0.34
```

Tabelle 8.3.n: Prüfung der Terme im Beispiel der Umweltumfrage

Tabelle 8.3.n zeigt, dass im **Beispiel der Umweltumfrage** für Geschlecht und Alter kein Einfluss auf die Beeinträchtugung nachgewiesen werden kann.

Für kontinuierliche und zweiwertige Eingangs-Variable wird mit `drop1` die gleiche Nullhypothese geprüft wie mit dem Test, der in der Koeffizienten-Tabelle steht. Es wird aber nicht der genau gleiche Test angewandt. (* Der erste ist ein Likelihood-Ratio Test, der zweite ein „Wald“-Test.) Näherungsweise (asymptotisch) geben sie immerhin die gleichen Resultate.

o Der Vergleich eines kleineren mit einem grösseren Modell bildete in der linearen Regression den Grund-Baustein für die **Modellwahl**, vor allem für die schrittartigen automatisierten Verfahren. Am Ende von Tabelle 8.3.l und in Tabelle 8.3.n erscheint eine Grösse **AIC**. Sie ist definiert als

$$\mathrm{AIC} = D\left\langle \underline{y}; \underline{\widehat{\pi}} \right\rangle + 2p$$

und kann wie in der linearen Regression als Gütemass der Modelle verwendet und optimiert werden. (p ist die Anzahl geschätzter Koeffizienten.)

p Das **kleinste sinnvolle Modell** sagt, dass die Eingangsgrössen überhaupt keinen Einfluss haben, dass also die Wahrscheinlichkeiten $\widetilde{\pi}_\ell$ alle gleich seien. Der Schätzwert für diesen einzigen Parameter ist natürlich $\widetilde{\pi} = \sum_\ell \widetilde{y}_\ell / \sum_\ell m_\ell = \sum_\ell \widetilde{y}_\ell / n$. Die Log-Likelihood für dieses Modell ist gleich

$$\begin{aligned} \ell\ell^{(0)} &= \sum\nolimits_\ell c_\ell + \sum\nolimits_\ell \widetilde{y}_\ell \log\langle\widetilde{\pi}\rangle + \sum\nolimits_\ell (m_\ell - \widetilde{y}_\ell) \log\langle 1-\widetilde{\pi}\rangle \\ &= \sum\nolimits_\ell c_\ell + n\,(\widetilde{\pi} \log\langle\widetilde{\pi}\rangle + (1-\widetilde{\pi}) \log\langle 1-\widetilde{\pi}\rangle) \ . \end{aligned}$$

Die Devianz ergibt sich wieder als Differenz zwischen

$$D\left\langle \underline{\widetilde{y}}; \underline{\widetilde{\pi}} \right\rangle = 2\left(\ell\ell^{(M)} - \ell\ell^{(0)}\right)$$

und wird **Null-Devianz** genannt. Sie entspricht der „ totalen Quadratsumme“ $\sum_i (Y_i - \overline{Y})^2$ in der linearen Regression. Wieder ist es sinnvoll, jedes Modell mit diesem einfachsten zu vergleichen, um zu prüfen, ob es überhaupt einen erklärenden Wert hat.

▷ Im **Beispiel der Frühgeburten** liest man in der Computer-Ausgabe „Null Deviance: 318.42 on 245 degrees of freedom“ und „Residual Deviance: 235.89 on 243 degrees of freedom“. Die Teststatistik $318.42 - 235.89 = 82.53$ ergibt mit der Chiquadrat-Verteilung mit $245 - 243 = 2$ Freiheitsgraden einen P-Wert von 0. Die beiden Eingangsgrössen haben also gemeinsam (selbstverständlich) einen klar signifikanten Erklärungswert. ◁

q **Zusammenfassung der Likelihood-Quotienten-Tests.** Da diese Tests beim „Modellbauen“ wichtig sind, hier eine Übersicht:

- Vergleich zweier Modelle: **Devianz-Differenz.**
  H_0: Modell K mit p_K Parametern ist richtig (kleineres Modell).
  H_1: Modell G mit $p_G > p_K$ Parametern ist richtig (grösseres Modell).
  Teststatistik $2(\ell^{(G)} - \ell^{(K)}) = D\left\langle \underline{\widetilde{y}}; \underline{\widetilde{\pi}}^{(K)} \right\rangle - D\left\langle \underline{\widetilde{y}}; \underline{\widetilde{\pi}}^{(G)} \right\rangle$.
  Genäherte Verteilung unter H_0: $\chi^2_{p_G - p_K}$.
- Vergleich mit maximalem Modell, Anpassungstest: **Residuen-Devianz**.
  H_0: Angepasstes Modell mit p Parametern ist richtig.
  H_1: Maximales Modell M (mit einem Parameter für jede (Gruppen-) Beobachtung) ist richtig.
  Teststatistik $D\left\langle \underline{\widetilde{y}}; \underline{\widehat{\pi}} \right\rangle = 2(\ell^{(M)} - \ell\langle\underline{\widehat{\pi}}\rangle)$
  Genäherte Verteilung unter H_0, falls die m_ℓ genügend gross sind: $\chi^2_{\widetilde{n}-p}$ mit $\widetilde{n} =$ Anzahl (Gruppen-) Beobachtungen $\widetilde{Y}_\ell$. Dieser Test geht nur für gruppierte Beobachtungen!
- Gesamttest für die Regression: Vergleich von **Null-Devianz** $D\left\langle \underline{\widetilde{y}}; \underline{\widehat{\pi}}^0 \right\rangle$ und Residuen-Devianz.
  H_0: Null-Modell mit einem Parameter ist richtig.
  H_1: Angepasstes Modell mit p Parametern ist richtig.
  Teststatistik $D\left\langle \underline{\widetilde{y}}; \underline{\widehat{\pi}}^0 \right\rangle - D\left\langle \underline{\widetilde{y}}; \underline{\widehat{\pi}} \right\rangle = 2(\ell\langle\underline{\widehat{\pi}}\rangle - \ell\left\langle \underline{\widehat{\pi}}^0 \right\rangle)$.
  Genäherte Verteilung unter H_0: χ^2_{p-1}.

8.4 Residuen-Analyse

a Was **Residuen** sein sollen, ist nicht mehr eindeutig. Wir diskutieren hier die Definitionen für „zusammengefasste“ Daten, siehe 8.2.g. Für zweiwertige Zielgrössen ohne Gruppierung muss man $m_\ell = 1$ setzen.
Die Grössen

$$R_\ell = \widetilde{Y}_\ell - \widehat{\pi}_\ell \ , \qquad \widehat{\pi}_\ell = g^{-1}\langle\widehat{\eta}_\ell\rangle$$

werden **rohe Residuen** oder **response residuals** genannt.

b Residuen werden dazu gebraucht, die Form des Modells zu überprüfen. Die zentrale Rolle dabei spielt der lineare Prädiktor $\eta_i = \underline{x}_i^T \underline{\beta}$. Zwischen der Zielgrösse und dem linearen Prädiktor steht die Link-Funktion. Damit die Residuen direkt mit dem linearen Prädiktor in Beziehung gebracht werden können, ist es sinnvoll, die rohen Residuen „in den Raum des linearen Prädiktors zu transformieren". Die **Prädiktor-Residuen**, englisch meist *working residuals* oder, in S, *link residuals* genannt, sind gegeben durch

$$R_\ell^{(L)} = R_\ell \frac{d\eta}{d\pi}\langle\widehat{\pi}_\ell\rangle = R_\ell \left(\frac{1}{\pi_\ell} + \frac{1}{1-\pi_\ell}\right) .$$

c Beide Arten von Residuen haben eine Varianz, die von $\widehat{y}_\ell$ abhängt. Es ist deshalb naheliegend, diese Abhängigkeit durch eine Standardisierung zu vermeiden: Die **Pearson-Residuen** sind definiert als

$$R_\ell^{(P)} = R_\ell \Big/ \sqrt{\widehat{\pi}_\ell(1-\widehat{\pi}_\ell)/m_\ell}$$

und haben genäherte Varianz 1.

d* In der linearen Regression wurde gezeigt, dass die Varianz der Residuen R_i nicht ganz gleich ist, auch wenn die Fehler E_i gleiche Varianz haben. Es war für die gewichtete Regression (Reg 1, 4.7) $\mathrm{var}\langle R_i\rangle = \sigma^2\,(1/w_i - (H_W)_{ii})$, wobei $(H_W)_{ii}$ das $i$te Diagonalelement der Matrix

$$\boldsymbol{H}_W = \boldsymbol{X}\,(\,\boldsymbol{X}^T\boldsymbol{W}\,\boldsymbol{X}\,)^{-1}\,\boldsymbol{X}^T$$

war. Die Gewichte, die hier gebraucht werden, sind diejenigen, die im Algorithmus (8.3.e) für das angenäherte lineare Regressionsproblem verwendet werden. Sie sind gleich $w_\ell = m_\ell/(\widehat{\pi}_\ell(1-\widehat{\pi}_\ell))$ (vergleiche . 9.d). Die genauer standardisierten Residuen sind dann

$$\widetilde{R}_\ell^{(P)} = R_\ell \Big/ \sqrt{(1/w_\ell - (H_W)_{ii})} \ .$$

e Ein weiterer, gut bekannter Typ von Residuen sind die **Devianz-Residuen**. Sie orientieren sich am Beitrag der $i$ten Beobachtung zur Devianz des Modells, der gemäss 8.3.i und 8.3.b gleich

$$\begin{aligned} d_i &= m_\ell\,(y_\ell \log\langle y_\ell\rangle + (1-y_\ell)\log\langle 1-y_\ell\rangle - y_\ell\log\langle\pi_\ell\rangle + (1-y_\ell)\log\langle 1-\pi_\ell\rangle) \\ &= m_\ell\left(y_\ell\log\left\langle\frac{y_\ell}{\pi_\ell}\right\rangle + (1-y_\ell)\log\left\langle\frac{1-y_\ell}{1-\pi_\ell}\right\rangle\right) \end{aligned}$$

ist. Er entspricht dem quadrierten Residuum R_i^2 in der gewöhnlichen linearen Regression. Um aus ihm ein sinnvolles Residuum zu erhalten, ziehen wir die Wurzel und versehen sie mit dem Vorzeichen der Abweichung; so wird

$$R_i^{(D)} = \mathrm{sign}\langle Y_i - \widehat{\pi}_i\rangle\,\sqrt{d_i}\ .$$

f Residuen sind dazu da, grafisch dargestellt zu werden. Allerdings ergeben sich Schwierigkeiten, vor allem bei ungruppierten, zweiwertigen Daten (oder wenn die m_ℓ klein sind). Die R_ℓ haben verschiedene Verteilungen. Die $R_\ell^{(P)}$ haben zwar gleiche Varianzen, aber trotzdem nicht die gleichen Verteilungen: Wenn man mit den ursprünglichen binären Y_i arbeitet, sind für jede Beobachtung i nur zwei Werte von $R_i^{(P)}$ möglich. Welche zwei Werte das sind und mit welchen Wahrscheinlichkeiten sie angenommen werden, hängt vom (angepassten) Wert π_i der Regressionsfunktion (oder von η_i) ab. Diese wiederum sind durch die Werte der Regressoren bestimmt, und eine Verteilungsannahme für die Regressoren gibt es normalerweise nicht. Es hat also keinen Sinn, die Normalverteilung der Residuen mit einem QQ-Plot zu überprüfen – obwohl einige Programme eine solche Darstellung liefern!

Wenn gruppierte Daten vorliegen, dann kann man die Binomialverteilungen mit Normalverteilungen annähern, und die Pearson-Residuen sollten näherungsweise eine Standard-Normalverteilung zeigen. **Ein Normalverteilungs-Diagramm** macht also **nur** Sinn, wenn Pearson-Residuen für **gruppierte Daten mit nicht zu kleinen** m_ℓ vorliegen.

g Das **Tukey-Anscombe-Diagramm** bleibt ein wichtiges Instrument der Modell-Überprüfung. Für seine Festlegung bieten sich mehrere Möglichkeiten an: Man kann einerseits auf der vertikalen Achse prinzipiell alle Typen von Residuen auftragen und andererseits auf der horizontalen Achse die angepassten Werte $\widehat{\eta}_i$ für den linearen Prädiktor oder die entsprechenden geschätzten Wahrscheinlichkeiten $\widehat{\pi}_i$. Der Zweck soll wieder vor allem darin bestehen, Abweichungen von der Form der Regressionsfunktion zu zeigen. Man wird deshalb

- entweder Response-Residuen und geschätzte π_i
- oder Arbeits-Residuen und Werte des linearen Prädiktors

verwenden (Abbildung 8.4.g).

Die erste Variante verwendet die Begriffe, die einfach definiert sind, während die zweiten Variante der besten Näherung durch eine lineare Regression entspricht und deshalb die entsprechenden Beurteilungen von Nichtlineritäten zulässt.

h Das Diagramm ist schwieriger zu interpretieren als in der gewöhnlichen Regression, da Artefakte auftreten: Die Punkte liegen für ungruppierte Daten für die erste Variante auf zwei Geraden mit Abstand 1 – jedes Y_i kann ja nur zwei Werte annehmen! Bei anderen Residuen wird es nicht viel besser: Statt zwei Geraden zeigen sich zwei Kurven.

In einem solchen Diagramm kann man deshalb nur Abweichungen vom Modell sehen, wenn man eine **Glättung** einzeichnet, also eigentlich mit einem nichtparametrischen Modell für die π_i oder η_i vergleicht. Dabei sollte eine Glättungsdsmethode verwendet werden, die den verschiedenen Varianzen der Residuen mittels Gewichten Rechung trägt. Es ist wichtig, dass im Fall von ungruppierten Daten keine robuste Glättung verwendet wird. Sonst werden für tiefe und hohe geschätzte π_i die wenigen Beobachtungen mit $Y_i = 1$ resp. mit $Y_i = 0$ als Ausreisser heruntergewichtet, auch wenn sie genau dem Modell entsprechen.
Im Idealfall wird die glatte Funktion nahe an der Nulllinie verlaufen. Im Beispiel zeigt sich eine recht deutliche Abweichung. Auch wenn man berücksichtigt, dass die Glättung sich an den beiden Rändern eher unsinnig verhält, sieht man doch, dass für kleine vorhergesagt Werte die Überlebens-Wahrscheinlichkeit immer noch überschätzt wird, und dass auch in der Mitte die Anpassung besser sein könnte.

i Die Situation ist wesentlich besser, wenn **gruppierte Daten** vorliegen. Abbildung 8.4.i zeigt, was im Beispiel der Frühgeburten mit klassiertem Gewicht herauskommt. Es zeigt sich eine deutliche Abweichung vom angenommenen Modell. Da nur ein Regressor vorliegt, nämlich das logarithmierte Geburtsgewicht, wird klar, dass sein Zusammenhang mit dem Logit der Überlebenswahrscheinlichkeit nicht linear ist. Das ist durchaus plausibel: Sobald das Gewicht genügend hoch ist, wird das Kind wohl überleben, und höhere Werte erhöhen die Wahrscheinlichkeit für diesen günstigen Verlauf nicht mehr stark. Andererseits werden die Überlebenschancen für leichte Neugeborene vom Modell überschätzt. Der Mangel sollte durch (weitere) Transformation dieser Eingangsgrösse behoben werden.

j Um allfällige nicht-lineare Abhängigkeiten der Zielgrösse von den Eingangsgrössen zu entdecken, kann man, wie in der multiplen linearen Regression, die **Residuen gegen die Eingangs-Variablen** auftragen. Da die Regressoren einen Teil des linearen Prädiktors ausmachen, ist es sinnvoll, Prädiktor-Residuen zu verwenden.

Als Variante kann man, wieder wie in der gewöhnlichen linearen Regression, zu den Residuen den „Effekt“ der betrachteten Eingangsgrösse addieren. So erhält man einen „**partial residual plot**“ oder „**term plot**“.

In Abbildung 8.4.j sieht man, dass für das Gewicht auch in dem erweiterten Modell der Effekt ungenügend modelliert ist (vergleiche 8.4.i). Für die Variable `pH`, die im Modell nicht enthalten ist, sollte ein quadratischer Effekt geprüft werden; das ist ja auch durchaus plausibel, da der pH einen optimalen Bereich aufweist.

k **Einflussreiche Beobachtungen** können hier, wie in der gewöhnlichen linearen Regression, aus einem Diagramm geeigneter Residuen gegen die „**Hebelarm**-Werte“ *(leverages)* ersehen werden. Der Einfluss ist proportional zum Residuum und zum Gewicht im linearen Regressionsproblem, das bei der iterativen Berechnung die letzte Korrektur ergibt. Diese Residuen sind die Prädiktor-Residuen, und die Gewichte w_i sind im Anhang (9.d) festgelegt.

Die merkwürdige Struktur eines doppelten Bumerangs kommt dadurch zustande, dass für die zentralen Beobachtungen, die wenig leverage haben, die vorhergesagten Wahrscheinlichkeits-Werte in diesem Beispiel bei 0.5 liegen und deshalb die Prädiktor-Residuen nicht gross werden können.

Im Beispiel zeigen sich zwei bis vier Beobachtungen mit hohen Hebelwerten und eine mit etwas kleinerem Hebelwert, aber recht grossem (negativem) Residuum. In einer vertieften Analyse könnte das Modell versuchsweise ohne diese Beobachtungen angepasst werden.

8.S S-Funktionen

a **Funktion** `glm`. `glm` steht für *generalized linear model.* Man muss der Funktion über das Argument `family` deshalb angeben, dass die Zielgrösse binomial (oder Bernoulli-) verteilt ist. Der Aufruf lautet

```
> r.glm <- glm( Y~log10(Gewicht)+Alter, family=binomial,
                data=d.babysurv )
```

Die Modell-Formel `Y~log10(Gewicht)+Alter` gibt die Zielgrösse und die Terme des linearen Prädiktors an, vgl. 3.2.j.

b Die **Link-Funktion** muss nicht angegeben werden, wenn die übliche Wahl der logit-Funktion gewünscht wird; das Programm wählt sie auf Grund der Angabe der `family` selbst. Eine andere Link-Funktion kann über das Argument `family` auf etwas überraschende Art verlangt werden: `..., family=binomial(link="probit")`. (`binomial` ist nämlich selbst eine Funktion, die ihrerseits Funktionen erzeugt, die von `glm` dann verwendet werden. Wie diese Funktionen aussehen, hängt vom Argument `link` ab.)

c **Funktion** `summary` gibt wie üblich die Ergebnisse der Anpassung sinnvoll aus,

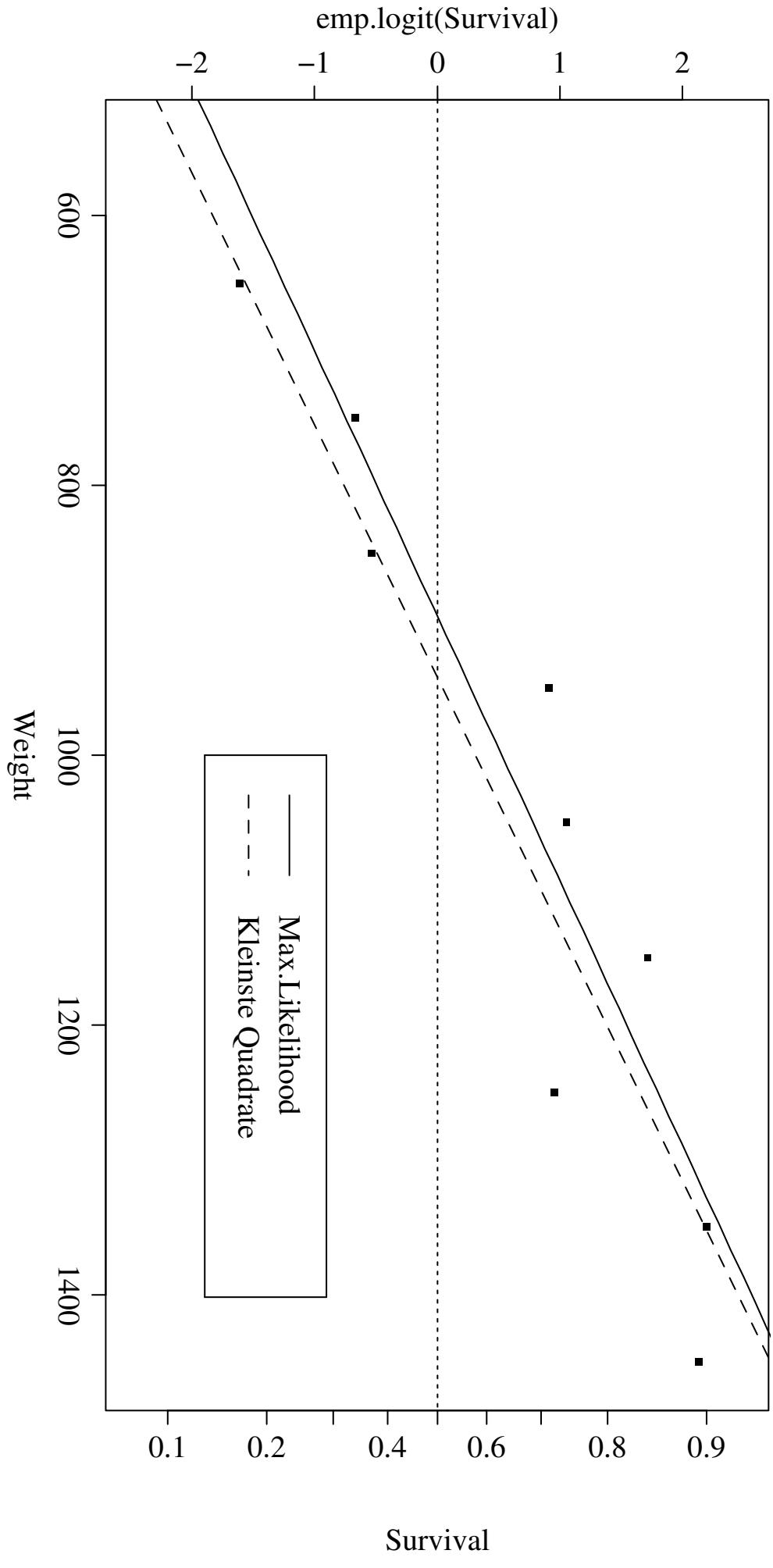
```
> summary(r.glm, corr=FALSE)
```

d **Funktion** `regr` funktioniert mit den gleichen Argumenten wie `glm`, liefert aber (ohne `summary`) vollständigere Resultate, wie im Fall der gewöhnlichen linearen Regression.

e **Funktion** `plot`. Wendet man `plot` auf das Ergebnis von `glm` an, dann werden bisher Darstellungen zur Residuen-Analyse gezeichnet, die nicht auf die logistische Regression passen.

Für das Resultat von `regr` wird kein Normalverteilungs-Diagramm gezeichnet (ausser man verlangt es ausdrücklich), und die Glättungen im Tukey-Anscombe plot und den Streudiagrammen der Residuen gegen die Eingangsvariablen ermöglichen eine sinnvolle Beurteilung dieser Darstellungen. Als Residuen werden die Arbeitsresiduen verwendet. Ihr Gewicht, das sie in der letzten Iteration des Algorithmus erhalten, wird durch die Symbolgrösse angezeigt.

f **Andere Verallgemeinerte Lineare Modelle**. Mit der entsprechenden Wahl des Arguments `family` können auch andere GLM angepasst werden, insbesondere die Poisson-Regression.


emp.logit(Survival)
−2
−1
0
1
2
600
800
1000
1200
1400
Weight
Max.Likelihood
Kleinste Quadrate
0.1
0.2
0.4
0.6
0.8
0.9
Survival

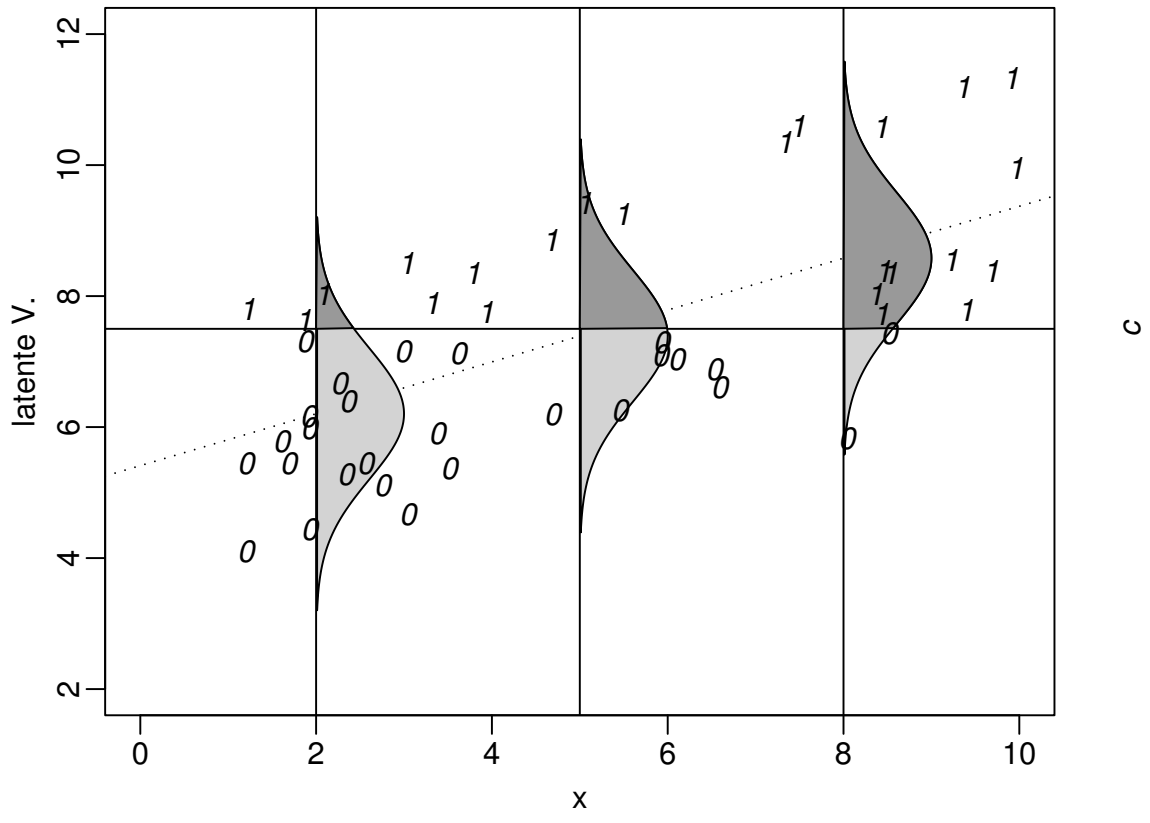


Abbildung 8.2.j: Zum Modell der latenten Variablen

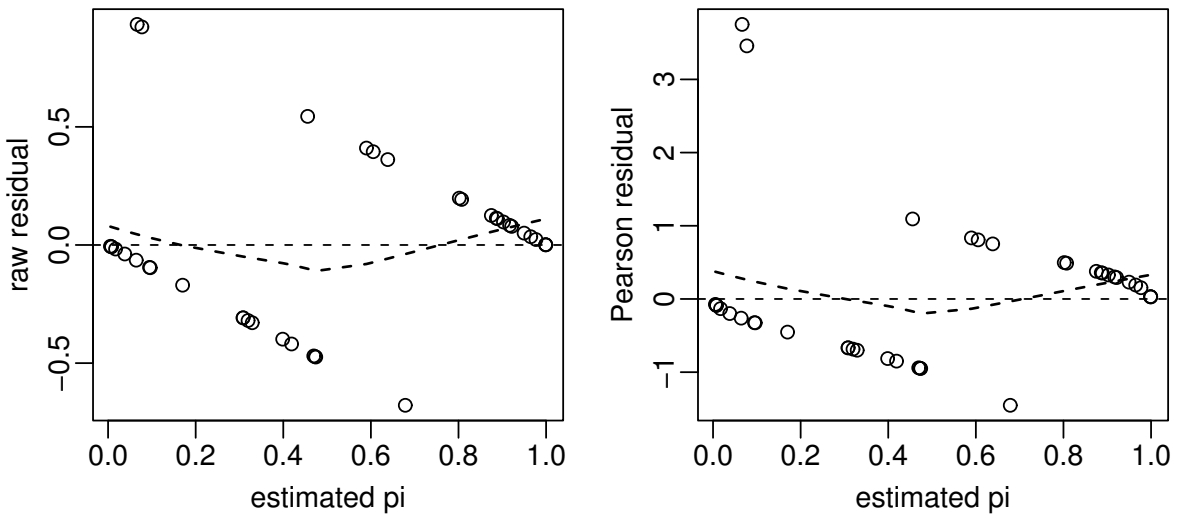


Abbildung 8.4.g: Tukey-Anscombe-Diagramme im Beispiel der Frühgeburten: (i) Response-Residuen und geschätzte π_i, (ii) Arbeits-Residuen und linearer Prädiktor

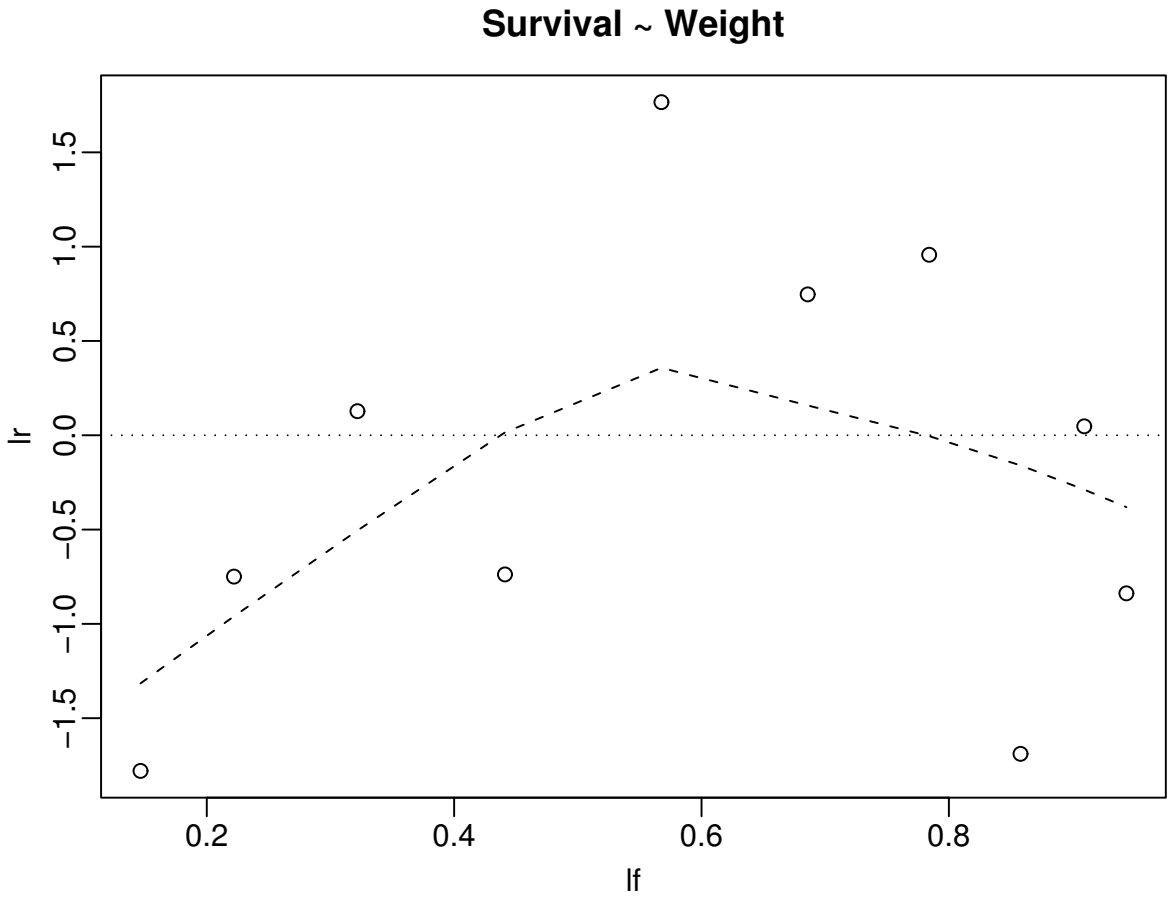


Abbildung 8.4.i: Tukey-Anscombe-Diagramm im Beispiel der Frühgeburten mit klassiertem Gewicht

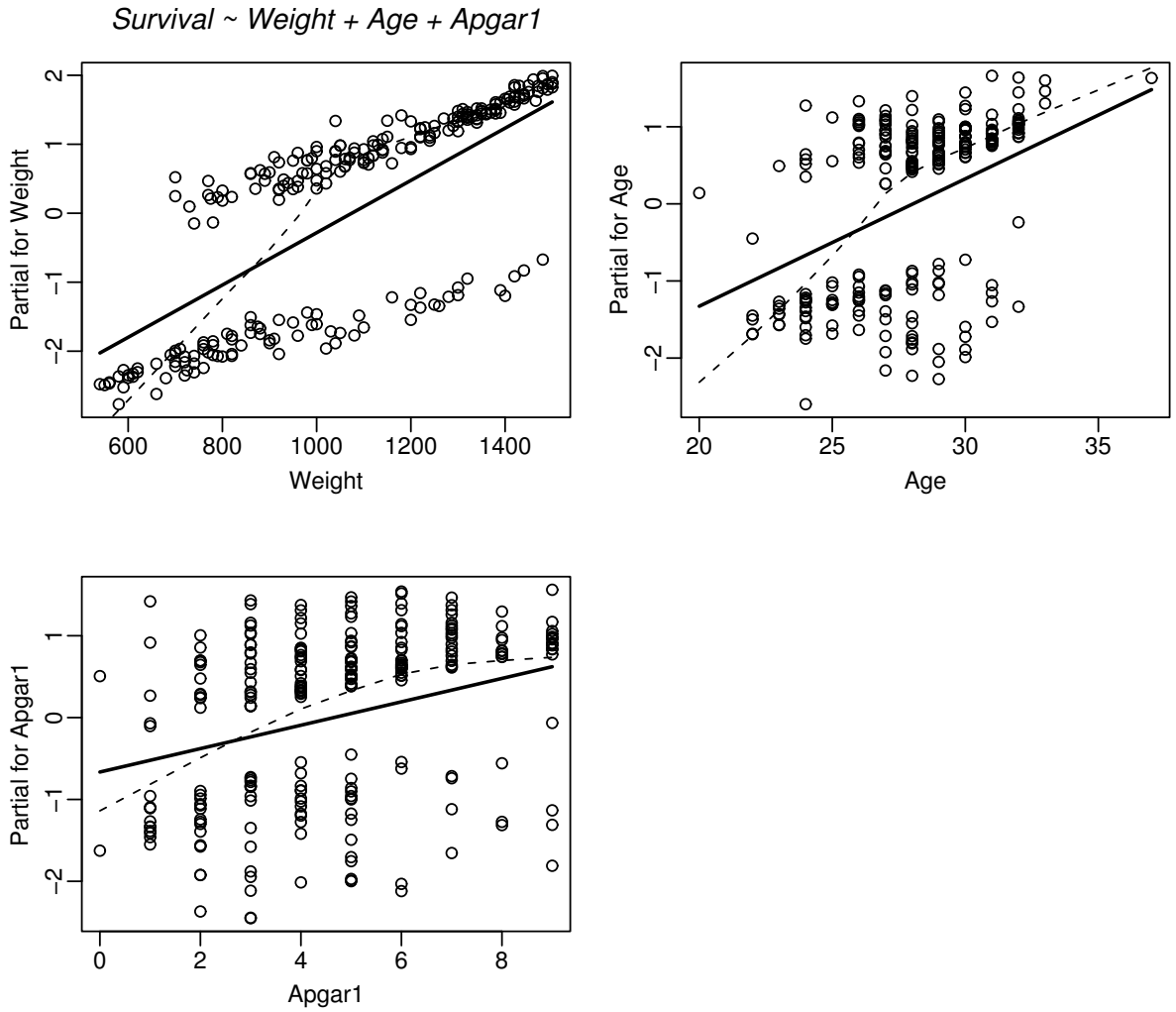


Abbildung 8.4.j: Residuen gegen Eingangsgrössen im Beispiel der Frühgeburten. Die Radien der Kreise entsprechen den Gewichten. Einige extrem negative Residuen wurden weggeschnitten.

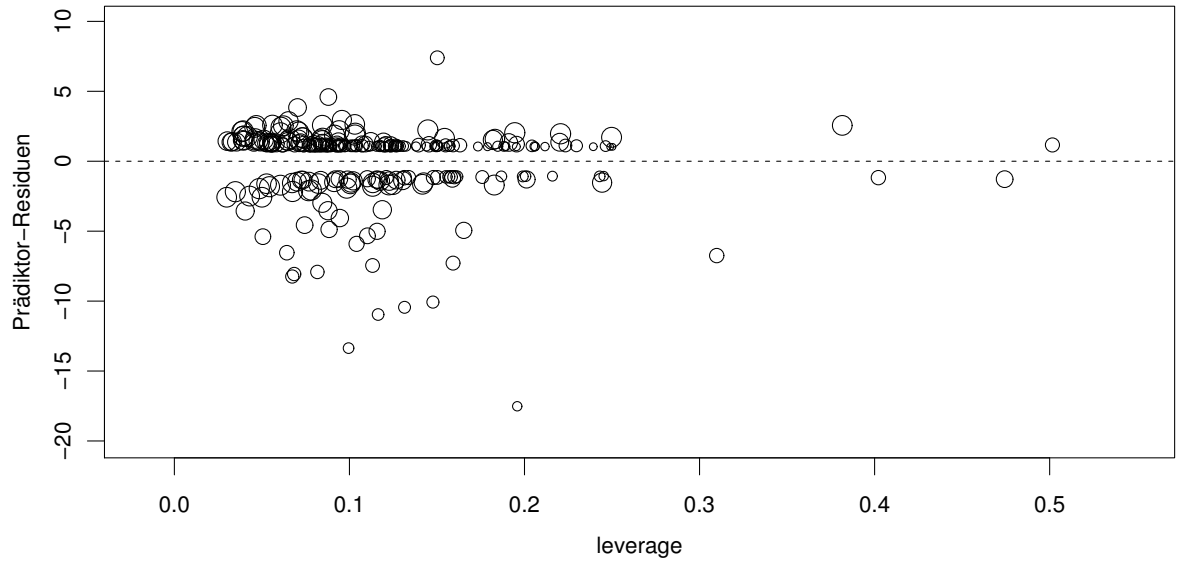


Abbildung 8.4.k: Residuen gegen Hebelarm-Werte $(\boldsymbol{H}_W)_{ii}$ für das Beispiel der Frühgeburten

9 Verallgemeinerte Lineare Modelle

9.1 Das Modell der Poisson-Regression

a Während sich die logistische Regression mit binären Zielgrössen befasst, liefert die Poisson-Regression Modelle für andere Zähldaten. Wir wollen diesen Fall nicht mehr ausführlich behandeln, sondern ihn benützen, um auf eine allgemeinere Klasse von Modellen vorzubereiten.

b **Beispiel gehemmte Reproduktion.** In einer Studie zur Schädlichkeit von Flugbenzin wurde die Reproduktion von Ceriodaphnia in Abhängigkeit von verschiedenen Konzentrationen des Schadstoffs für zwei Stämme von Organismen untersucht (Quelle: Myers, Montgomery and Vining (2001), example 4.5). Wie Abbildung 9.1.b zeigt, fällt die Anzahl der reproduzierenden Organismen stark ab; die Abnahme könnte etwa exponentielle Form haben.

c **Verteilung.** Die Zielgrösse Y_i ist eine Anzahl von Individuen. Deswegen liegt es nahe, ihre Verteilung, gegeben die Eingangsgrössen, als Poisson-verteilt anzunehmen, $Y_i \sim \mathcal{P}\langle\lambda_i\rangle$. Der Parameter λ_i wird von den Regressoren $\underline{x}_i$ abhängen.

Erinnern wir uns, dass der Parameter λ der Poisson-Verteilung gleich ihrem Erwartungswert ist. Für diesen Erwartungswert nehmen wir nun, wie in der multiplen linearen und der logistischen Regression, an, dass er eine Funktion der Regressoren ist, zusammen also

$$Y_i \sim \mathcal{P}\langle\lambda_i\rangle \ , \qquad \mathcal{E}\langle Y_i\rangle = \lambda_i = h\langle\underline{x}_i\rangle \ ,$$

und die Y_i sollen stochastisch unabhängig sein.

d **Link-Funktion.** Da der Erwartungswert nicht negativ sein kann, ist eine lineare Funktion $\beta_0 + \sum_j \beta_j x_i^{(j)}$ wieder nicht geeignet als Funktion h. Für binäre Zielgrössen verwendeten wir diesen „linearen Prädiktor“ trotzdem und setzten ihn gleich einer Transformation des Erwartungswertes,

$$g\langle\mathcal{E}\langle Y_i\rangle\rangle = \eta_i = \underline{x}_i^T\underline{\beta} \ .$$

(Wir schreiben, wie früher, der Kürze halber $\underline{x}_i^T\underline{\beta}$ statt $\beta_0+\sum_j \beta_j x_i^{(j)}$ oder statt $\sum_j \beta_j x_i^{(j)}$, wenn kein Achsenabschnitt β_0 im Modell vorkommen soll.) Als Transformations-Funktion eignet sich der **Logarithmus**, denn er macht aus den positiven Erwartungswerten transformierte Werte, die keine Begrenzung haben. Der *Logarithmus* des Erwartungswertes der Zielgrösse Y_i ist also gemäss dem Modell eine *lineare* Funktion der Regressoren $\underline{x}_i$. Man nennt solche Modelle **log-linear**.

Die **Poisson-Regression** kombiniert nun die logarithmische Link-Funktion mit der Annahme der Poisson-Verteilung für die Zielgrösse.

e Der Logarithmus verwandelt, wie wir bereits in der linearen und der logistischen Regression erörtert haben, **multiplikative Effekte** in additive Terme im Bereich des linearen Prädiktors, oder umgekehrt: Wenn $g\langle\lambda\rangle = \log\langle\lambda\rangle$ ist, gilt

$$\begin{aligned}\mathcal{E}\langle Y_i\rangle &= \lambda = \exp\langle\underline{x}_i^T\underline{\beta}\rangle = e^{\beta_0}\cdot e^{\beta_1 x_i^{(1)}}\cdot\ldots\cdot e^{\beta_m x_i^{(m)}} \\ &= e^{\beta_0}\cdot\exp\langle\beta_1\rangle^{x^{(1)}}\cdot\ldots\cdot\exp\langle\beta_m\rangle^{x^{(m)}} \ .\end{aligned}$$

Die Zunahme von $x^{(j)}$ um eine Einheit bewirkt eine Multiplikation des Erwartungswertes λ um den Faktor $\widetilde{\beta}_j$, der auch als „Unit risk “ bezeichnet wird. Ist β_j positiv, so ist $\widetilde{\beta}_j > 1$, und der Erwartungswert wird mit zunehmendem $x^{(j)}$ grösser.

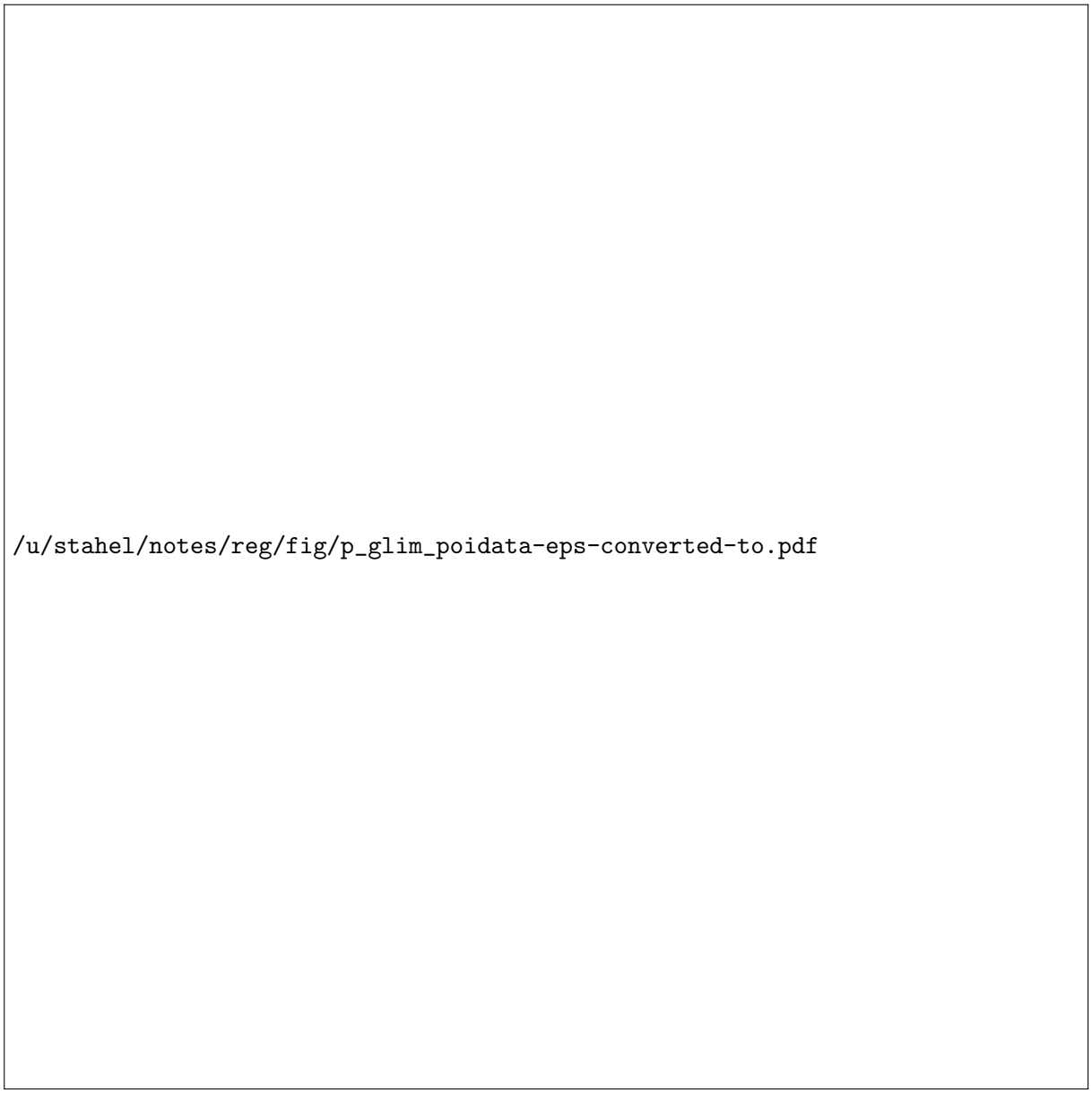


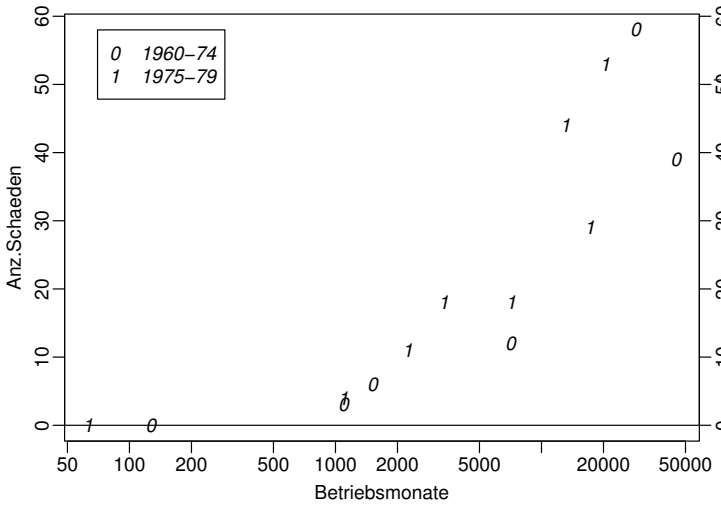
Abbildung 9.1.b: Anzahl reproduzierende Individuen im Beispiel der gehemmten Reproduktion. Die beiden Stämme sind mit verschiedenen Symbolen angegeben.

f Im **Beispiel der gehemmten Reproduktion** sind die Konzentration `C` des Benzins und der verwendete Stamm `S` die Eingangsgrössen. Die erwartete Anzahl nimmt mit der Erhöhung der Konzentration um eine Einheit gemäss einem Haupteffekt-Modell

$$\log\langle\mathcal{E}\,\langle Y_i\rangle\rangle = \eta_i = \beta_0 + \beta_C\ \mathtt{C}_i + \beta_S\ \mathtt{S}_i$$

um einen Faktor $\exp\langle\beta_C\rangle$ ab, was einer exponentiellen Abnahme gleich kommt, deren „Geschwindigkeit“ für beide Stämme gleich ist. Die beiden Stämme unterscheiden sich durch einen konstanten Faktor $\exp\langle\beta_S\rangle$. Wenn die „Geschwindigkeiten“ für die beiden Stämme unterschiedlich sein sollen oder, anders gesagt, der Unterschied zwischen den Stämmen für die verschiedenen Konzentrationen nicht den gleichen Faktor ergeben soll, dann braucht das Modell einen Wechselwirkungs-Term $\beta_{CS}\ \mathtt{C}\cdot\mathtt{S}$.

g **Beispiel Schiffs-Havarien**. Grosse Wellen können an Lastschiffen Schäden verursachen. Wovon hängen diese Havarien ab? Um diese Frage zu beantworten, wurden 7 „Flotten“ vergleichbarer Schiffe in je zwei Beobachtungsperioden untersucht (Quelle: McCullagh and Nelder (1989, p. 205), Teil der Daten). Für jede dieser 7×2 Beobachtungseinheiten wurde die Summe der Betriebsmonate über die Schiffe (M) erhoben und die Anzahl Y_i der Schadensereignisse eruiert. In der Tabelle in Abbildung 9.1.g sind ausserdem die Beobachtungsperiode (P), die Bauperiode (C) und Schiffstyp (T) notiert. Die Daten ergeben sich also aus einer Gruppierung von ursprünglichen Angaben über einzelne Schiffe, die entsprechend der Bauperiode, dem Schiffstyp und der Beobachungsperiode zusammengefasst wurden. Der wichtigste und offensichtlichste Zusammenhang – derjenige zwischen Anzahl Schadensereignisse und Anzahl Betriebsmonate – ist in der Abbildung grafisch festgehalten.



	T	C	P	M	Y
1	0	60	0	127	0
2	0	60	1	63	0
3	0	65	0	1095	3
4	0	65	1	1095	4
5	0	70	0	1512	6
6	0	70	1	3353	18
7	0	75	1	2244	11
8	1	60	0	44882	39
9	1	60	1	17176	29
10	1	65	0	28609	58
11	1	65	1	20370	53
12	1	70	0	7064	12
13	1	70	1	13099	44
14	1	75	1	7117	18

Abbildung 9.1.g: Daten zum Beispiel der Schiffs-Havarien. T: Schiffstyp, C: Bauperiode, P: Beobachtungsperiode, M: Betriebsmonate, Y: Anzahl Havarien

Es interessiert uns, welchen Einfluss die Eingangsgrössen auf die Schadensfälle haben. Welcher Schiffstyp ist anfälliger? Gibt es Unterschiede zwischen den beiden Beobachtungsperioden?

h Für dieses Beispiel ist das folgende Modell plausibel:

$$\log\langle\mathcal{E}\langle Y_i\rangle\rangle \;=\; \beta_0 + \beta_M \log\langle \mathtt{M}_i\rangle + \beta_T \mathtt{T}_i + \beta_P \mathtt{P}_i + \gamma_1 \cdot \mathtt{(C1)}_i + \gamma_2 \cdot \mathtt{(C2)}_i + \gamma_3 \cdot \mathtt{(C3)}_i$$

wobei `C1, C2` und `C3` dummy Variable sind, die der Variablen `C` (Bauperiode) entsprechen, welche hier als Faktor einbezogen wird. In der Sprache der Modell-Formeln wird das vereinfacht zu

```
Y ~ log10(M) + T + P + C .
```

Weshalb wurde hier die Summe `M` der Betriebsmonate logarithmiert? Es ist plausibel, anzunehmen, dass die erwartete Anzahl Schadensfälle exakt proportional zu `M` ist, also, wenn man die anderen Einflussgrössen weglässt, $\mathcal{E}\langle Y_i\rangle = \alpha \mathtt{M}_i$, und deshalb $\log\langle\mathcal{E}\langle Y_i\rangle\rangle = \beta_0 + \beta_M \log\langle \mathtt{M}_i\rangle$ mit $\beta_0 = \log\langle\alpha\rangle$ und $\beta_M = 1$. Wir werden also erwarten, dass die Schätzung $\widehat{\beta}_M$ ungefähr 1 ergibt. Dass sich eine allfällige Veränderung zwischen den Beobachtungsperioden P bzw. den Schiffstypen T ebenfalls multiplikativ auswirken sollte, ist sehr plausibel. Der Faktor $\exp\langle\beta_P\rangle$ beschreibt dann die Veränderung des Risikos, d.h. wie viel mal mehr Schäden in der zweiten Periode zu erwarten sind.

i **Term ohne Koeffizient.** Nochmals zum Einfluss der Betriebsmonate: Da wir für β_M aus guten Gründen den Wert 1 erwarten, muss dieser Koeffizient eigentlich nicht aus den Daten geschätzt werden. In der gewöhnlichen linearen Regression liesse sich eine solche Idee einfach umsetzen: Wir würden statt der Anzahl der Schäden Y_i die „Rate“ $Y_i/\texttt{M}_i$ der Zielgrösse verwenden (und `M` für eine Gewichtung verwenden). Hier geht das schief, weil $Y_i/\texttt{M}_i$ keine Poisson-Verteilung hat. Deshalb muss das Programm die Option einer „Vorgabe“ für jede Beobachtung vorsehen. In der S-Funktion `glm` gibt es dafür ein Argument `offset`.

j Im Beispiel wurden die Schiffe, die eigentlich die natürlichen Beobachtungseinheiten wären, zu Gruppen zusammengefasst, und die Zielgrösse war dann die Summe der Zahlen der Havarien für die einzelnen Schiffe. Wie in 7.1.f erwähnt, ist diese Situation häufig. Es entstehen meistens Kreuztabellen. Wir werden in Kapitel 10.S.0.c sehen, dass die Poisson-Regression (oder besser -Varianzanalyse) für ihre Analyse eine entscheidende Rolle spielt.

9.2 Das Verallgemeinerte Lineare Modell

a Logistische und Poisson-Regression bilden zwei Spezialfälle der **Verallgemeinerten Linearen Modelle** *(generalized linear models)*, und auch die gewöhnliche lineare Regression gehört dazu. Wir haben bereits die wichtigste Annahme, die allen gemeinsam ist, formuliert: **Der Erwartungswert der Zielgrösse, geeignet transformiert, ist gleich einer linearen Funktion der Parameter** β_j, genannt der **lineare Prädiktor**,

$$g\langle\mathcal{E}\langle Y_i\rangle\rangle = \eta_i = \underline{x}_i^T\,\underline{\beta}\ .$$

Die Funktion g, die Erwartungswerte von Y in Werte für den linearen Prädiktor η verwandelt, wird **Link-Funktion** genannt.
In der gewöhnlichen linearen Regression ist g die Identität, in der logistischen die Logit-Funktion und in der Poisson-Regression der Logarithmus.

b Damit ist noch nichts über die Form der **Verteilung** von Y_i gesagt. In der gewöhnlichen Regression wurde eine Normalverteilung angenommen, mit einer Varianz, die nicht vom Erwartungswert abhängt. Es war sinnvoll, die additive Zufallsabweichung E_i einzuführen und für sie im üblichen Fall eine (Normal-) Verteilung anzunehmen, die für alle i gleich war. Das wäre für die logistische und die Poisson-Regression falsch. Hier ist die Verteilung von Y_i jeweils durch den Erwartungswert (und m_ℓ im Fall von gruppierten Daten in der logistischen Regression) bereits festgelegt.
Die Verallgemeinerten Linearen Modelle lassen hier einen grossen Spielraum offen. Die Verteilung von Y_i, gegeben ihr Erwartungswert, soll zu einer parametrischen Familie gehören, die ihrerseits der grossen Klasse der **Exponentialfamilien** angehört. Diese ist so weit gefasst, dass möglichst viele übliche Modelle dazugehören, dass aber trotzdem nützliche mathematische Theorie gemacht werden kann, die zum Beispiel sagt, wie Parameter geschätzt und getestet werden können.

c **Exkurs: Exponentialfamilien**. Eine Verteilung gehört einer so genannten einfachen Exponentialfamilie an, wenn sich ihre Dichte $f\langle y\rangle$ oder Wahrscheinlichkeitsfunktion $P\langle Y=y\rangle$ schreiben lässt als

$$\exp\left\langle \frac{y\theta - b\langle\theta\rangle}{\phi}\,\omega + c\langle y;\phi,\omega\rangle \right\rangle\ .$$

Das sieht kompliziert aus! Es ist, wie beabsichtigt, allgemein genug, um nützliche und bekannte Spezialfälle zu umfassen. Was bedeuten die einzelnen Grössen?

- Der Parameter θ heisst der **kanonische Parameter**. Die Eingangs-Variablen werden,

wenn wir wieder zu den Verallgemeinerten Linearen Modellen zurückkehren, diesen kanonischen Parameter kontrollieren.

- ϕ ist ein weiterer Parameter, der mit der Varianz zu tun hat und **Dispersions-Parameter** genannt wird. Er ist normalerweise ein Störparameter und wird mit der Regression nichts zu tun haben. (Genau genommen ist die Familie nur eine Exponential-Familie, wenn ϕ als fest angenommen wird.)
- Die Grösse ω ist eine feste Zahl, die bekannt ist, aber von Beobachtung zu Beobachtung verschieden sein kann. Sie hat die Bedeutung eines **Gewichtes** der Beobachtung. Man könnte sie auch in die Grösse ϕ hineinnehmen. Bei mehreren Beobachtungen i wird ω von i abhängen, während ϕ für alle gleich ist. (Bei gruppierten Daten in der logistischen Regression wird $\omega_\ell = m_\ell$ sein, wie wir gleich feststellen werden.)
- Die Funktion $b\langle . \rangle$ legt fest, um welche Exponentialfamilie es sich handelt.
- Die Funktion $c\langle . \rangle$ wird benötigt, um die Dichte oder Wahrscheinlichkeitsfunktion auf eine Gesamt-Wahrscheinlichkeit von 1 zu normieren.

d **Erwartungswert und Varianz** können allgemein ausgerechnet werden,

$$\mu = \mathcal{E}\langle Y \rangle = b'\langle \theta \rangle \ , \qquad \mathrm{var}\langle Y \rangle = b''\langle \theta \rangle \cdot \phi/\omega \ .$$

Da die Ableitung $b'\langle . \rangle$ der Funktion b jeweils umkehrbar ist, kann man auch θ aus dem Erwartungswert μ ausrechnen,

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

Nun kann man auch die $b''\langle \theta \rangle$ direkt als Funktion von μ schreiben, $V\langle \mu \rangle = b''\big\langle (b')^{-1}\langle \mu \rangle \big\rangle$. Man nennt diese Funktion die **Varianzfunktion**, da gemäss der vorhergehenden Gleichung

$$\mathrm{var}\langle Y \rangle = V\langle \mu \rangle \cdot \phi/\omega$$

gilt.

e Wir wollen nun einige Verteilungen betrachten, die sich in dieser Form darstellen lassen. Zunächst zur **Normalverteilung**! Ihre logarithmierte Dichte ist

$$\begin{aligned} \log\big\langle\, f\big\langle y; \mu, \sigma^2 \big\rangle\,\big\rangle &= -\log\Big\langle \sqrt{2\pi^o}\sigma \Big\rangle - \frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2 \\ &= \frac{\mu y - \frac{1}{2}\mu^2}{\sigma^2} \ - \ y^2/(2\sigma^2) - \frac{1}{2}\log\big\langle 2\pi^o\sigma^2 \big\rangle \end{aligned}$$

(wobei wir $\pi^o = 3.14159...$ schreiben zur Unterscheidung vom Parameter π). Sie entspricht mit

$$\begin{aligned} \theta &= \mu \ , \qquad b\langle \theta \rangle = \theta^2/2 \ , \qquad \phi = \sigma^2 \ , \qquad \omega = 1 \\ c\langle y; \phi, \omega \rangle &= -y^2/(2\phi) - (1/2)\log\langle 2\pi^o\phi \rangle \end{aligned}$$

der vorhergehenden Form – auch wenn man sich zum Seufzer: „Wieso auch einfach, wenn es kompliziert auch geht!“ veranlasst sieht.

Die obigen Formeln für Erwartungswert und Varianz sind rasch nachgeprüft: $b'\langle \theta \rangle = \theta = \mu$ und $b''\langle \theta \rangle = 1$ und damit $\mathrm{var}\langle Y \rangle = \phi/\omega = \sigma^2$.

f **Binomialverteilung**. In 8.2.g wurde der *Anteil* $\widetilde{Y}_\ell$ von „Erfolgen“ unter m_ℓ Versuchen als Zielgrösse verwendet und festgestellt, dass $m_\ell \widetilde{Y}_\ell$ binomial verteilt ist. Die Wahrscheinlichkeiten, ohne $\widetilde{\ }$ und Index ℓ geschrieben, sind dann $P\langle Y=y\rangle = \binom{m}{my}\pi^{my}(1-\pi)^{m-my}$ und ihre logarithmierten Werte kann man schreiben als

$$\begin{aligned}\log\big\langle\, P\langle Y=y\rangle\,\big\rangle &= \log\Big\langle\binom{m}{my}\Big\rangle + (my)\log\langle\pi\rangle + m\log\langle 1-\pi\rangle - (my)\log\langle 1-\pi\rangle \\ &= my\log\langle\pi/(1-\pi)\rangle + m\log\langle 1-\pi\rangle + \log\Big\langle\binom{m}{my}\Big\rangle\ .\end{aligned}$$

Hier ist

$$\begin{aligned}\theta &= \log\langle\pi/(1-\pi)\rangle \implies \pi = e^\theta/(1+e^\theta) \\ b\langle\theta\rangle &= \log\langle 1+\exp\langle\theta\rangle\rangle\ , \qquad \omega = m\ , \qquad \phi = 1 \\ c\langle y;\phi;\omega\rangle &= \log\Big\langle\binom{m}{my}\Big\rangle\end{aligned}$$

Für Erwartungswert und Varianz gilt $\mu = b'\langle\theta\rangle = \exp\langle\theta\rangle/(1+\exp\langle\theta\rangle) = \pi$ und $\operatorname{var}\langle Y\rangle = b''\langle\theta\rangle = \exp\langle\theta\rangle\,(1+\exp\langle\theta\rangle) - (\exp\langle\theta\rangle)^2/(1+\exp(\theta))^2 = \pi(1-\pi)$.

Für binäre Variable gilt die Formel natürlich auch, mit $m=1$.

g **Poisson-Verteilung**. Die Wahrscheinlichkeiten sind

$$P\langle Y=y\rangle = \frac{1}{y!}\,\lambda^y\,e^{-\lambda}\ , \qquad \log\langle P\langle Y=y\rangle\rangle = -\log\langle y\,!\rangle + y\log\langle\lambda\rangle - \lambda\ .$$

Hier erhält man

$$\begin{aligned}\theta &= \log(\lambda)\ , \qquad b\langle\theta\rangle = \exp(\theta) = \lambda \\ \phi &= 1\ , \qquad \omega = 1\ , \qquad c\langle y;\phi;\omega\rangle = -\log\langle y\,!\rangle \\ \mu &= b'\langle\theta\rangle = \exp(\theta)\ , \qquad \operatorname{var}\langle Y\rangle = b''\langle\theta\rangle = \exp\langle\theta\rangle\end{aligned}$$

h Weitere wichtige Verteilungen, die in die gewünschte Form gebracht werden können, sind die **Exponentialverteilung** und allgemeiner die **Gamma-Verteilung** und die **Weibull-Verteilung**, die für kontinuierliche positive Grössen wie Überlebenszeiten geeignet sind und deshalb unter anderem in der Zuverlässigkeits-Theorie eine wichtige Rolle spielen.

i Zurück zum **Regressionsmodell:** Bei logistischer und Poisson-Regression haben wir den Zusammenhang zwischen Ziel- und Einflussgrössen mit Hilfe der **Link-Funktion** g modelliert. Sie hat zunächst den Zweck, die möglichen Erwartungswerte auf den Bereich der möglichen Werte des linearen Prädiktors – also alle (reellen) Zahlen – auszudehnen. Die naheliegenden Link-Funktionen sind

$g\langle\mu\rangle = \log\langle\mu\rangle$,	wenn $\mathcal{E}\langle Y\rangle > 0$ sein muss, aber sonst beliebig ist,
$g\langle\mu\rangle = \operatorname{logit}\langle\mu\rangle = \log\langle\mu/(1-\mu)\rangle$,	wenn $\mathcal{E}\langle Y\rangle$ zwischen 0 und 1 liegen muss,
$g\langle\mu\rangle = \mu$,	wenn $\mathcal{E}\langle Y\rangle$ keinen Einschränkungen unterliegt,

Die Link-Funktion verknüpft den Erwartungswert μ mit dem linearen Prädiktor η, und μ ist seinerseits eine Funktion des kanonischen Parameters θ. Dies kann man zusammen schreiben als

$$\eta = g\big\langle\, b'\langle\theta\rangle\,\big\rangle = \widetilde{g}\langle\theta\rangle\ .$$

j Die bisher betrachteten verallgemeinerten linearen Modelle haben noch eine spezielle Eigenschaft: Die gewählte Link-Funktion führt den Erwartungswert μ in den kanonischen Parameter θ über. Damit wird $\theta = \eta$ oder $\widetilde{g}$ gleich der Identität. Es wird also angenommen, dass die Kovariablen-Effekte linear auf den kanonischen Parameter wirken. Diese Funktionen nennt man **kanonische Link-Funktionen**.

k Prinzipiell kann man aber auch **andere Link-Funktionen** verwenden. Wenn beispielsweise $0 < \mathcal{E}\langle Y\rangle < 1$ gelten muss, lässt sich jede kumulative Verteilungsfunktion als inverse Link-Funktion einsetzen (8.2.j). Wenn es keine konkreten Gründe für eine spezielle Link-Funktion gibt, verwendet man aber in der Regel die kanonische. Zum einen besitzen „kanonische verallgemeinerte lineare Modelle“ bessere theoretische Eigenschaften (Existenz und Eindeutigkeit des ML-Schätzers). Zum andern vereinfachen sich dadurch die Schätzgleichungen.

Wenn sich in der Praxis auf Grund der Residuenanalyse ein Hinweis auf ein schlecht passendes Modell zeigt, ist es oft sinnvoll, wie in der multiplen linearen Regression, zunächst durch Transformationen der Eingangsgrössen zu versuchen, die Anpassung des Modells zu verbessern. Wenn das nichts hilft, wird man die Link-Funktion ändern.

9.3 Schätzungen und Tests

a Der Vorteil einer Zusammenfassung der betrachteten Modelle zu einem allgemeinen Modell besteht darin, dass theoretische Überlegungen und sogar Berechnungsmethoden für alle gemeinsam hergeleitet werden können. Die Schätzung der Parameter erfolgt nach der Methode der Maximalen Likelihood, und die Tests und Vertrauensintervalle beruhen auf genäherten Verteilungen, die für Maximum-Likelihood-Schätzungen allgemein hergeleitet werden können.

b **Likelihood**. Die Parameter, die uns interessieren, sind die Koeffizienten β_j. Sie bestimmen den Erwartungswert μ_i für jede Beobachtung, und dieser bestimmt schliesslich θ_i (siehe 9.2.d). Wir nehmen an, dass ϕ für alle Beobachtungen gleich ist. Der Beitrag einer Beobachtung i zur Log-Likelihood $\ell\ell$ ist gleich

$$\ell\ell_i\big\langle y_i;\underline{\beta}\big\rangle = \log\big\langle\, P\big\langle Y_i = y_i \mid \underline{x}_i,\underline{\beta}\big\rangle\,\big\rangle = \big(y_i\theta_i - b\big\langle\,\theta_i\,\big\rangle\big)\,\omega_i/\phi + c\langle y_i;\phi,\omega_i\rangle \;\;,\quad \theta_i = \widetilde{g}\big\langle \underline{x}_i^T\underline{\beta}\big\rangle \;.$$

Für Poisson-verteilte Zielgrössen mit der kanonischen Link-Funktion erhält man

$$\ell\ell_i\big\langle y_i;\underline{\beta}\big\rangle = y_i\cdot\log\langle\lambda_i\rangle - \lambda_i - \log(y_i!) = y_i\eta_i - e^{\eta_i} - \log(y_i!)\;,\qquad \eta_i = \underline{x}_i^T\underline{\beta}\;.$$

Da es sich um unabhängige Beobachtungen handelt, erhält man die Log-Likelihood als Summe $\ell\ell\big\langle\underline{y};\underline{\beta}\big\rangle = \sum_i \ell\ell_i\big\langle y_i;\underline{\beta}\big\rangle$.

c **Maximum-Likelihood-Schätzung**. Wir leiten hier die Schätzungen für den Spezialfall der Poisson-Regression mit „log-Link“ her. Die analoge, allgemeine Herleitung der Schätzgleichungen, eine Skizzierung des Schätzalgorithmus und einige Eigenschaften der Schätzer findet man im Anhang 9.A.

Die Ableitung der Log-Likelihood nach den Parametern setzt sich, wie die Log-Likelihood, aus Beiträgen der einzelnen Beobachtungen zusammen, die **Scores** genannt werden,

$$s_i^{(j)}\big\langle\underline{\beta}\big\rangle = \frac{\partial\ell\ell_i\big\langle\underline{\beta}\big\rangle}{\partial\beta_j} = \frac{\partial\widetilde{\ell\ell}}{\partial\eta}\langle\eta_i\rangle\cdot\frac{\partial\eta_i}{\partial\beta_j} = (y_i-\lambda_i)\cdot x_i^{(j)}\;.$$

Setzt man alle Komponenten gleich null,

$$s\big\langle\underline{\beta}\big\rangle = \sum\nolimits_i \underline{s}_i\big\langle\underline{\beta}\big\rangle = \underline{0}\;,$$

so entstehen die impliziten Gleichungen, die die Maximum-Likelihood-Schätzung $\widehat{\underline{\beta}}$ bestimmen; für den Poisson-Fall $\sum_i(y_i-\lambda_i)\cdot x_i^{(j)} = 0$.

Zur Lösung dieser Gleichungen geht man so vor, wie das für die logistische Regression in 8.3.e skizziert wurde und wie es in Anhang 9.b beschrieben ist.

d **Schätzung des Dispersions-Parameters.** Im allgemeinen Modell muss auch der Dispersions-Parameter ϕ geschätzt werden, und auch das erfolgt durch Maximieren der Likelihood. Für die spezifischen Modelle kommt dabei eine recht einfache Formel heraus. Für die Normalverteilung kommt, bis auf einen Faktor $(n-p)/n$, die übliche Schätzung der Varianz heraus. Für binomial- und Poisson-verteilte Zielgrössen muss kein Dispersions-Parameter geschätzt werden – wir werden in 9.4 diese gute Nachricht allerdings wieder einschränken.

e Um **Tests und Vertrauensbereiche** festzulegen, braucht man die Verteilung der Schätzungen. Es lässt sich zeigen, dass als „asymptotische Näherung“ eine multivariate Normalverteilung gilt,

$$\underline{\widehat{\beta}} \overset{a}{\sim} \mathcal{N}\left\langle \underline{\beta}, \boldsymbol{V}^{(\beta)} \right\rangle \ ,$$

wobei die Kovarianzmatrix $\boldsymbol{V}^{(\beta)}$ normalerweise von $\underline{\beta}$ abhängen wird. (Genaueres steht im Anhang, 9.e.) Damit lassen sich genäherte P-Werte für Tests und Vertrauensintervalle angeben. In der linearen Regression galt die Verteilung exakt, mit $\boldsymbol{V}^{(\beta)} = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$, und das ergab exakte P-Werte und Vertrauensintervalle.

f Für das **Beispiel der gehemmten Reproduktion** zeigt Tabelle 9.3.f den Aufruf der S-Funktion `regr` und die Computer-Ausgabe, die die bereits bekannte Form hat. Beide Eingangsgrössen erweisen sich als hoch signifikant.

```
Call: regr(formula = count ~ ., data = d.ceriofuel, family = poisson,
           calcdisp = F)
Terms:
              coef stcoef signif df p.value
(Intercept)  4.455  0.000  57.02  1       0
fuel        -1.546 -0.869 -16.61  1       0
strain      -0.274 -0.138  -2.84  1       0

         deviance df p.value
Model        1276  2  0.0000
Residual       88 67  0.0433
Null         1364 69      NA

Family is poisson.  Dispersion parameter taken to be 1.
AIC: 417.3
```

Tabelle 9.3.f: Computer-Ausgabe von `regr` für das Beispiel der gehemmten Reproduktion

g **Devianz.** Für die logistische Regression wurde die Likelihood, die mit der Anpassung der Modell-Parameter erreicht wird, mit einer maximalen Likelihood verglichen, und das lässt sich auch in den andern Verallgemeinerten Linearen Modellen tun. Die maximale Likelihood entsteht, indem ein maximales Modell angepasst wird, das für jede Beobachtung i den am besten passenden kanonischen Parameter $\widetilde{\theta}_i$ bestimmt. Die Devianz ist allgemein definiert als

$$\begin{aligned} D\langle \underline{y}; \underline{\widehat{\mu}} \rangle &= 2(\ell\ell^{(M)} - \ell\ell\left\langle \underline{\widehat{\beta}} \right\rangle) = \frac{2}{\phi} \sum\nolimits_i \omega_i \left(y_i(\widetilde{\theta}_i - \widehat{\theta}_i) - b\left\langle \widetilde{\theta}_i \right\rangle + b\left\langle \widehat{\theta}_i \right\rangle \right) \\ \widehat{\theta}_i &= \widetilde{g}\left\langle \underline{x}_i^T \underline{\widehat{\beta}} \right\rangle \end{aligned}$$

wobei $\underline{y}$ der Vektor aller beobachteten Werte ist und $\underline{\widehat{\mu}}$ der Vektor der zugehörigen angepassten Erwartungswerte. Der Teil der Log-Likelihood-Funktion, der nicht von θ abhängt, fällt dabei weg. In der Formel ist $\widetilde{\theta}_i$ der Parameter, der am besten zu y_i passt. Er ist jeweils bestimmt durch $y_i = \mathcal{E}\langle Y_i \rangle = b'\left\langle \widetilde{\theta}_i \right\rangle$.
Ein Dispersions-Parameter ϕ lässt sich für das maximale Modell nicht mehr schätzen; man verwendet den geschätzten Wert des betrachteten Modells. Bei der Binomial- und der Poisson-Verteilung fällt dieses Problem weg, da $\phi = 1$ ist.

h Im Poisson-Modell sind die geschätzten Parameter im maximalen Modell gleich $\widetilde{\theta}_i = \log\langle y_i \rangle$ und man erhält

$$\begin{aligned} D\langle \underline{y}; \underline{\widehat{\mu}} \rangle &= 2 \sum\nolimits_i \left(y_i(\log\langle y_i \rangle - \log\langle \widehat{\mu}_i \rangle) - e^{\log\langle y_i \rangle} + e^{\log\langle \widehat{\mu}_i \rangle} \right) \\ &= 2 \sum\nolimits_i y_i \log\langle y_i / \widehat{\mu}_i \rangle - (y_i - \widehat{\mu}_i) \end{aligned}$$

Für binomial verteilte Zielgrössen wurde die Devianz in 8.3.i angegeben.

i Mit Hilfe der Devianz lassen sich auch allgemein die Fragen beantworten, die für die logistische Regression bereits angesprochen wurden:

- Vergleich von Modellen.
- Überprüfung des Gesamt-Modells.
- Anpassungstest.

Die entsprechenden Devianz-Differenzen sind unter gewissen Bedingungen näherungsweise chiquadratverteilt. Für die Residuen-Devianz binärer Zielgrössen sind diese Bedingungen, wie erwähnt (8.3.k), nicht erfüllt.

* Die Bedingungen sind also für einmal nicht harmlos. Das liegt daran, dass im maximalen Modell M (9.3.g) für jede Beobachtung ein Parameter geschätzt wird; mit der Anzahl Beobachtungen geht also auch die Anzahl Parameter gegen unendlich, und das ist für asymptotische Betrachtungen gefährlich!

j Die Devianz wird für die Normalverteilung zur Summe der quadrierten Residuen, die ja bei der Schätzung nach dem Prinzip der Kleinsten Quadrate minimiert wird. Für andere Verteilungen haben die „rohen Residuen“ (8.4.a) verschiedene Varianz und sollten mit entsprechenden Gewichten summiert werden. Die Grösse

$$T = \sum\nolimits_i \frac{\omega_i (y_i - \widehat{\mu}_i)^2}{\widetilde{\phi}\, V\langle \widehat{\mu}_i \rangle}$$

heisst **Pearson-Chiquadrat-Statistik**. Wenn $\widetilde{\phi}$ nicht aus den Daten geschätzt werden muss, folgt sie in der Regel genähert einer Chiquadrat-Verteilung. Wenn T zu gross wird, müssen wir auf signifikante Abweichung vom Modell schliessen. Das legt einen **Anpassungstest** fest.

Vorher haben wir die Residuen-Devianz als Teststatistik für genau den gleichen Zweck verwendet. Sie hatte näherungsweise ebenfalls die gleiche Chiquadrat-Verteilung. Die beiden Teststatistiken sind „asymptotisch äquivalent“.

9.4 Übergrosse Streuung

a Die Residuen-Devianz des angepassten Modells kann man für einen Anpassungstest verwenden, falls der Dispersions-Parameter *nicht* aus den Daten geschätzt werden *muss.* Im Fall von binomial und Poisson-verteilten Zielgrössen ist die Varianz ja durch das Modell festgelegt, und der Anpassungstest kann zur Ablehnung des Modells führen. Die Devianz misst in gewissem Sinne die Streuung der Daten und der Test vergleicht diese geschätzte Streuung mit der Varianz, die unter dem Modell zu erwarten wäre. Ein statistisch signifikanter, erhöhter Wert bedeutet also, dass die Daten – genauer die Residuen – eine **übergrosse Streuung** zeigen. Man spricht von **over-dispersion**.

Im Beispiel der gehemmten Reproduktion war die Residuen-Devianz knapp signifkant; es ist also eine übergrosse Streuung angezeigt.

b Damit wir dennoch Statistik treiben können, brauchen wir ein neues Modell. Statt einer Poisson-Verteilung könnten wir beispielsweise eine so genannte **Negative Binomialverteilung** postulieren. Es zeigt sich aber, dass es gar nicht nötig ist, sich auf eine bestimmte Verteilungsfamilie festzulegen. Wesentlich ist nur, wie die Varianz $V\langle\mu\rangle\,\phi/\omega$ der Verteilung von Y von ihrem Erwartungwert μ abhängt. Dies bestimmt die asymptotischen Verteilungen der geschätzten Parameter.

Die einfachste Art, eine grössere Streuung als im Poisson- oder Binomialmodell zuzulassen, besteht darin, die jeweilige Varianzfunktion beizubehalten und den Dispersions-Parameter ϕ nicht mehr auf 1 festzulegen. Dieser wird dann zu einem Störparameter.

Da damit kein Wahrscheinlichkeits-Modell eindeutig festgelegt ist, spricht man von Quasi-Modellen und von **Quasi-Likelihood**.

c Der Parameter ϕ lässt sich analog zur Varianz der Normalverteilung schätzen $\widehat{\phi} = \frac{1}{n-p}\sum_i \frac{\omega_i(y_i-\widehat{\mu}_i)^2}{V\langle\mu_i\rangle}$. Man teilt also die Pearson-Statistik durch ihre Freiheitsgrade. Üblicher ist es aber, statt der Pearson-Statistik die Devianz zu verwenden, die ja, wie gesagt (9.3.j), näherungsweise das Gleiche ist. Das ergibt $\widehat{\phi} = (1/(n-p))D\langle\underline{y};\underline{\widehat{\mu}}\rangle$. Im Beispiel der gehemmten Reproduktion erhält man mit den Angaben von 9.3.f $\widehat{\phi} = 88/67 = 1.3$.

d Im Anhang (9.e) kann man sehen, dass die Kovarianzmatrix der asymptotischen Verteilung der geschätzten Koeffizienten den Faktor ϕ enthält. (* $\widetilde{\boldsymbol{H}}$ enthält den Faktor $1/\phi$, siehe 9.c.) Durch die Einführung eines Dispersions-Parameters werden deshalb einfach Konfidenzintervalle um den Faktor $\sqrt{\hat{\phi}}$ breiter und die Werte der Teststatistiken um $1/\widehat{\phi}$ kleiner.

Die Funktion `regr` verwendet den geschätzten Streuungsparameter $\widehat{\phi}$ zur Berechnung der Tests von Koeffizienten und von Vertrauensintervallen, sofern der mittlere Wert der Zielgrösse gross genug ist (momentan wird als Grenze 3 verwendet) – ausser, dies werde mit dem Argument `calcdisp=FALSE` unterdrückt (wie es in 9.3.f getan wurde).

e Beachte: Der Schluss gilt nicht in umgekehrter Richtung. Wenn der Dispersions-Parameter kleiner als 1 ist, verkleinern sich nicht die Konfidenzintervalle. Häufig ist ein kleiner Dispersions-Parameter ein Hinweis darauf, dass in einem Modell für gruppierte Beobachtungen die Unabhängigkeitsannahme zwischen den Einzel-Beobachtungen nicht erfüllt ist.

Diese Erscheinung tritt in der Ökologie immer wieder auf, wenn die **Anzahl Arten** auf einer Untersuchungsfläche als Zielgrösse benützt wird. Die Poisson-Verteilung ist hier nicht adäquat, da „Ereignisse“ mit ganz verschiedenen Wahrscheinlichkeiten gezählt werden. Eine häufige Art ist vielleicht auf allen Untersuchungsflächen anzutreffen, und wenn es vorwiegend solche Arten hätte, wäre die Variation der Artenzahl sicher wesentlich kleiner, als das von einer Poisson-Verteilung festgelegt wird. Eine Poisson-verteilte Variable zählt unabhängige „Ereignisse“, die gleichartig und deshalb gleich wahrscheinlich sind.

f **Quasi-Modelle**. Die Idee, einen Dispersions-Parameter einzuführen, ohne ein genaues Modell festzulegen, lässt sich verallgemeinern: Das Wesentliche am Modell sind die Link- und die Varianzfunktion. Man legt also nur fest, wie der Erwartungswert und die Varianz von Y vom linearen Prädiktor η abhängt.

9.5 Residuen-Analyse

a Für die Definition von **Residuen** gibt es die vier für die logistische Regression eingeführten Vorschläge:

- Rohe Residuen oder **response residuals**: $R_i = Y_i - \widehat{\mu}_i$.

  Wie erwähnt, haben diese Residuen verschiedene Varianzen.

- Die **Prädiktor-Residuen** *(working residuals* oder *link residuals)* erhält man, indem man die Response-Residuen „in der Skala des Prädiktors ausdrückt:

$$R_i^{(L)} = R_i \cdot g'\langle\widehat{\mu}_i\rangle \ ,$$

- **Pearson-Residuen**: Die rohen Residuen werden durch ihre Standardabweichung, ohne Dispersions-Parameter ϕ, dividiert,

$$R_i^{(P)} = R_i / \sqrt{V\langle\widehat{\mu}_i\rangle / \omega_i} \ .$$

  Diese „unskalierten" Pearson-Residuen dienen dazu, den Dispersions-Parameter zu schätzen oder zu prüfen, ob er gleich 1 sein kann, wie dies für das Binomial- und das Poisson-Modell gelten muss (vgl. 9.4). Die Grössen $R_i^{(P)}/\widehat{\phi}$ nennen wir skalierte Pearson-Residuen,

- **Devianz-Residuen**: Jede Beobachtung ergibt einen Beitrag d_i/ϕ zur Devianz (9.3.g), wobei

$$d_i = 2\omega_i \left(Y_i(\widetilde{\theta}_i - \widehat{\theta}_i) - b\left\langle\widetilde{\theta}_i\right\rangle + b\left\langle\widehat{\theta}_i\right\rangle\right) \ .$$

  Für die Normalverteilung sind dies die quadrierten Residuen. Um sinnvolle Residuen zu erhalten, zieht man daraus die Wurzel und setzt als Vorzeichen diejenigen der rohen Residuen, also

$$R_i^{(D)} = \operatorname{sign}\langle Y_i - \widehat{\mu}_i\rangle \sqrt{d_i} \ .$$

  Sie werden unskalierte „Devianz-Residuen" genannt – unskaliert, weil wieder der Faktor ϕ weggelassen wurde. Wenn man ihn einbezieht, erhält man die skalierten Devianz-Residuen.

b Die wichtigsten grafischen Darstellungen der Residuen-Analyse sind:

- **Tukey-Anscombe-Plot**: Prädiktor-Residuen $R_i^{(L)}$ werden gegen den linearen Prädiktor $\widehat{\eta}_i$ aufgetragen. Die Residuen sollten über den ganzen Bereich um 0 herum streuen. Wenn eine Glättung (von Auge oder berechnet) eine Abweichung zeigt, soll man eine Transformation von Eingangs-Variablen (siehe term plot, unten) oder allenfalls eine andere Link-Funktion prüfen.

c • **Scale Plot**. Absolute (Pearson-) Residuen gegen angepasste Werte $\widehat{\mu}_i$ auftragen. Wenn eine Glättung einen Trend zeigt, ist die Varianzfunktion nicht passend. Man kann versuchen, sie direkt zu modellieren, siehe 9.4.f.

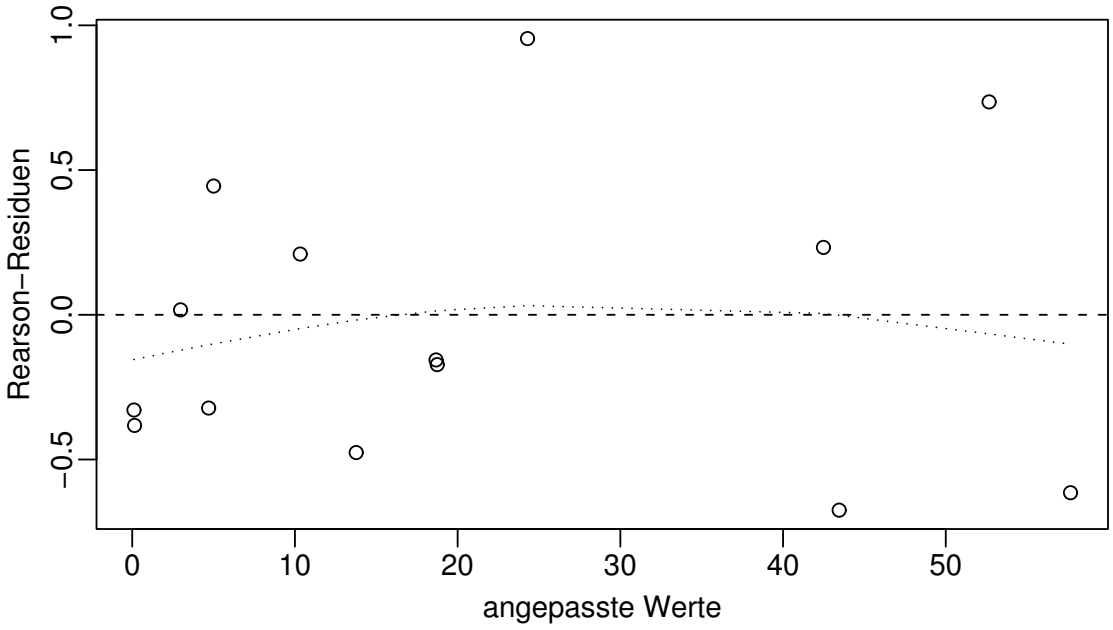


Abbildung 9.5.b: Tukey-Amscombe Plot zum Beispiel der Schiffs-Havarien

d • **Residuen gegen Eingangs-Variable.** Prädiktor-Residuen $R_i^{(L)}$ werden gegen Eingangs-Variable $x_i^{(j)}$ aufgetragen. Gekrümmte Glättungen deuten wie in der linearen Regression an, wie die Eingangsgrössen transformiert werden sollten. Die Funktion `plresx` liefert wieder eine Referenzlinie für gleiche Werte des linearen Prädiktors. Da die Residuen mit verschiedenen Gewichten zur Regression beitragen, sollten sie dem entsprechend verschieden gross gezeichnet werden.
Wieder ist es üblicher, die **partiellen Residuen** zu verwenden und den Effekt der Eingangs-Variablen mit einzuzeichnen, also einen **partial residual plot** oder **term plot** zu erstellen (vergleiche 8.4.j).

e • **Leverage Plot**. Die Prädiktor-Residuen $R_i^{(L)}$ werden gegen die „fast ungewichteten“ Hebelarm-Werte $\widetilde{h}_i$ aufgetragen und die Gewichte w_i durch verschieden grosse Kreis-Symbole dargestellt (vergleiche 8.4.k).

f Abbildungen 9.5.b und 9.5.d zeigen Residuenplots zum Modell im Beispiel Schiffs-Havarien. Bei so kleiner Beobachtungszahl sind Abweichungen kaum auszumachen.

9.S S-Funktionen

a Zur von Verallgemeinerten Linearen Modellen dienen die S-Funktionen `glm` oder `regr`, die wir schon für die logistische Regression verwendet haben. Die Angabe `family=poisson` legt die gewählte Verteilungsfamilie fest.

`summary`, `plot`, `drop1`, ...

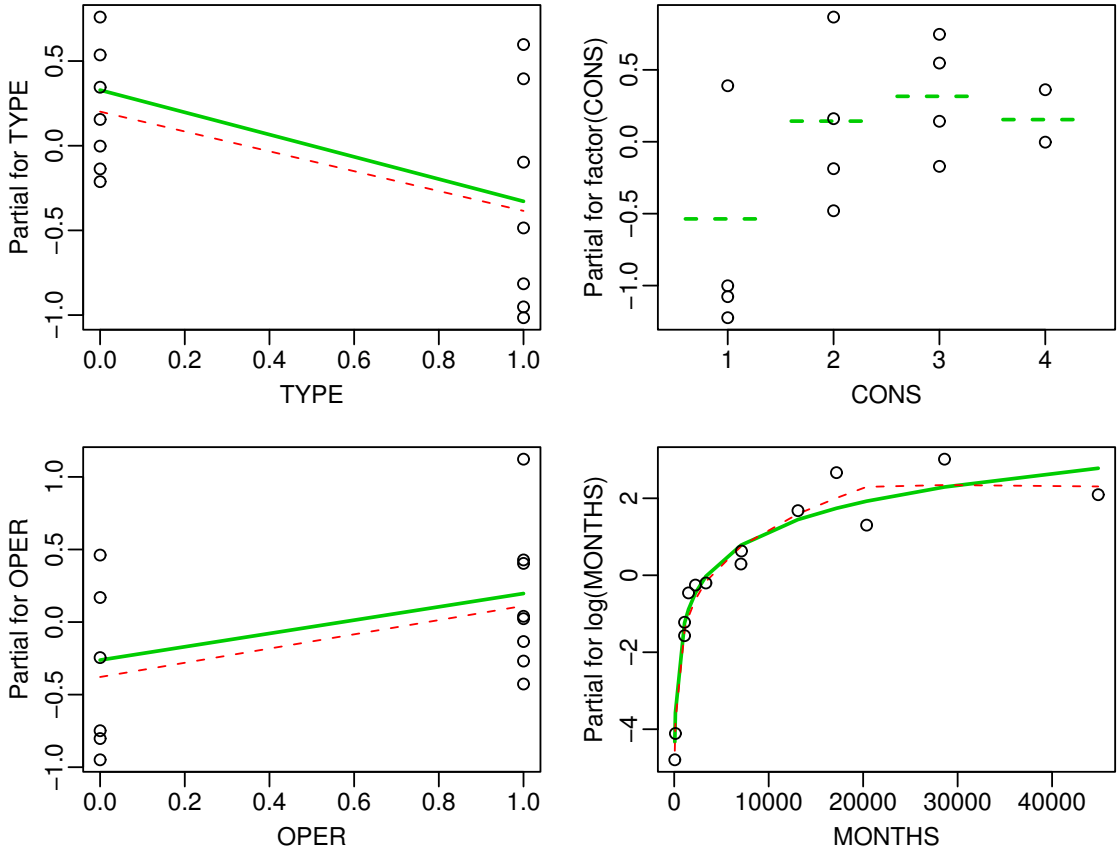


Abbildung 9.5.d: Partial residual Plots zu dem Havarie-Modell

9.A Anhang: Genaueres zur Schätzung der Parameter und zur asymptotischen Verteilung

a **Maximum Likelihood.** Der Beitrag $\ell\ell_i$ einer Beobachtung zur Log-Likelihood ist in 9.3.b angegeben. Um die Maximum-Likelihood-Schätzung zu bestimmen, wird man wie üblich die Ableitungen der Summe dieser Beiträge nach den Parametern gleich null setzen. Die Ableitung von $\ell\ell_i$ nach den Parametern hat hier und auch später eine fundamentale Bedeutung. Sie wird **Score-Funktion** genannt. Wir erhalten wie in 9.3.c

$$s^{(j)}\left\langle y_i, \underline{x}_i; \underline{\beta}\right\rangle = \frac{\partial \ell\ell_i\left\langle \underline{\beta}\right\rangle}{\partial \beta_j} = \frac{\partial \ell}{\partial \theta}\left\langle \theta_i\right\rangle \cdot \frac{d\theta}{d\mu}\left\langle \mu_i\right\rangle \cdot \frac{d\mu}{d\eta}\left\langle \eta_i\right\rangle \cdot \frac{\partial \eta_i}{\partial \beta_j} \ .$$

(Für Funktionen $f\langle x\rangle$ eines einzigen Argumentes schreiben wir die (gewöhnliche) Ableitung als df/dx.) Da $\mu(\theta) = b'\langle\theta\rangle$ und $\eta_i = \underline{x}_i^T\underline{\beta}$ ist, werden die Ableitungen zu

$$\begin{aligned}
\frac{\partial \ell\ell}{\partial \theta}\left\langle \theta_i\right\rangle &= (y_i - b'\left\langle \theta_i\right\rangle)\cdot \omega_i/\phi = (y_i - \mu_i)\cdot \omega_i/\phi \\
\frac{d\mu}{d\theta}\left\langle \theta_i\right\rangle &= b''\left\langle \theta_i\right\rangle = V\left\langle \mu_i\right\rangle \quad \Longrightarrow \quad \frac{d\theta}{d\mu}\left\langle \mu_i\right\rangle = 1/V\left\langle \mu_i\right\rangle \\
\frac{d\mu}{d\eta}\left\langle \eta_i\right\rangle &= (g^{-1})'\left\langle \eta_i\right\rangle \ , \qquad \frac{\partial \eta_i}{\partial \beta_j} = x_i^{(j)} \ .
\end{aligned}$$

(In der mittleren Zeile wurde die Regel für die Ableitung einer Umkehrfunktion verwendet: $(f^{-1})'\langle y\rangle = 1/f'\langle x\rangle$ mit $y = f\langle x\rangle$.) Zusammen erhält man

$$s^{(j)}\left\langle y_i, \underline{x}_i; \underline{\beta}\right\rangle = (y_i - \mu_i)\cdot \frac{\omega_i}{\phi\, V\left\langle \mu_i\right\rangle}\cdot (g^{-1})'\left\langle \eta_i\right\rangle \cdot x_i^{(j)} \ .$$

Setzt man alle Komponenten der Scores-Summe gleich null, $\sum_i \underline{s}\left\langle y_i, \underline{x}_i; \underline{\beta}\right\rangle = \underline{0}$, so entstehen die impliziten Gleichungen, die die Maximum-Likelihood-Schätzungen $\widehat{\beta}_j$ bestimmen.

b **Algorithmus.** Für die Lösung dieser impliziten Gleichungen wird ein Algorithmus angewandt, der allgemein für Maximum-Likelihood-Schätzungen geeignet ist und **Scoring-Algorithmus** heisst. Er ist mit dem allgemein bekannten **Newton-Raphson-Algorithmus** für numerische Optimierung verwandt. Dieser ist ein iteratives Rechenschema: Ausgehend von einem Startwert $\underline{\beta}^{(0)}$ wird eine Verbesserung $\Delta\underline{\beta}$ berechnet, die zu einer Verbesserung der Zielfunktion – in unserem Fall zu einer Erhöhung der Log-Likelihood – führt. Solche Schritte werden wiederholt, bis die Verbesserungen sehr klein werden.

Der Verbesserungsschritt des Newton-Raphson-Algorithmus verlangt die Berechnung von Ableitungen der Funktionen $s^{(j)}\left\langle \underline{\beta}\right\rangle$, die null werden sollen, also von zweiten Ableitungen der Zielfunktion. Das ergibt eine ganze Matrix $\boldsymbol{H}\left\langle \underline{\beta}\right\rangle = \partial \underline{s}\left\langle \underline{\beta}\right\rangle/\partial\underline{\beta} = [\partial s^{(j)}\left\langle \underline{\beta}\right\rangle/\partial\beta_k]_{jk}$, die Hessesche Matrix genannt wird. Die Funktion $\underline{s}\left\langle \underline{\beta}\right\rangle$ ist in der Nähe eines Vektors $\underline{\beta}^{(s)}$ gemäss linearer Näherung gleich

$$\underline{s}\left\langle \underline{\beta}\right\rangle \approx \underline{s}\left\langle \underline{\beta}^{(s)}\right\rangle + \boldsymbol{H}\left\langle \underline{\beta}^{(s)}\right\rangle\ (\underline{\beta} - \underline{\beta}^{(s)}) \ .$$

Wenn man die rechte Seite gleich null setzt, erhält man die Korrektur

$$\Delta\underline{\beta} = \underline{\beta}^{(s+1)} - \underline{\beta}^{(s)} = -(\boldsymbol{H}\left\langle \underline{\beta}^{(s)}\right\rangle)^{-1}\cdot \underline{s}\left\langle \underline{\beta}^{(s)}\right\rangle \ .$$

So weit die allgemeine Idee des Newton-Raphson-Algorithmus.

c Bei der Maximum-Likelihood-Schätzung ist die Funktion $\underline{s}$ die Summe $\sum_i s\langle y_i, \underline{x}_i; \underline{\beta}\rangle$, also

$$\Delta\underline{\beta} = -\boldsymbol{H}\left\langle \underline{\beta}^{(s)}\right\rangle^{-1} \cdot \sum\nolimits_i \underline{s}\left\langle y_i, \underline{x}_i; \underline{\beta}^{(s)}\right\rangle \quad \text{mit} \quad \boldsymbol{H}\langle\underline{\beta}\rangle = \sum\nolimits_i \partial\underline{s}\langle y_i, \underline{x}_i; \underline{\beta}\rangle / \partial\underline{\beta} \; .$$

Die Idee des Scoring-Algorithmus besteht darin, die Summanden in $\boldsymbol{H}$ durch ihren Erwartungswert $\widetilde{\boldsymbol{H}}$ unter der (vorläufig) geschätzten Verteilung zu ersetzen. Man erhält, da die Verteilung der Beobachtungen von den $\underline{x}_i$ abhängt, weiterhin eine Summe,

$$\widetilde{\boldsymbol{H}}\left\langle \underline{\beta}^{(s)}\right\rangle = \sum\nolimits_i \mathcal{E}\left\langle \partial\underline{s}\langle Y, \underline{x}_i; \underline{\beta}\rangle / \partial\underline{\beta}\right\rangle \; .$$

Die Ableitungen $\partial s^{(j)}/\partial\beta^{(k)}$ schreiben wir als

$$\frac{\partial s^{(j)}\langle Y, \underline{x}_i; \underline{\beta}\rangle}{\partial\beta^{(k)}} = -\frac{\partial\mu_i}{\partial\beta^{(k)}} \cdot \frac{\omega_i}{\phi V\langle\mu_i\rangle}\,(g^{-1})'\langle\eta_i\rangle\; x_i^{(j)} + (Y-\mu_i)\,\frac{\omega_i}{\phi}\,\frac{\partial}{\partial\beta^{(k)}}\left\langle \frac{(g^{-1})'\langle\eta_i\rangle}{V\langle\mu_i\rangle}\right\rangle \cdot x_i^{(j)} \; .$$

Den komplizierteren zweiten Teil müssen wir glücklicherweise nicht ausrechnen, da sein Erwartungswert null ist – es ist ja nur Y zufällig, und $\mathcal{E}\langle Y-\mu_i\rangle = 0$. Der erste Teil hängt nicht von Y ab; man muss also gar keinen Erwartungswert bilden. Es ist $\partial\mu_i/\partial\beta^{(k)} = (g^{-1})'\langle\eta_i\rangle\,\underline{x}_i^{(k)}$. Deshalb wird

$$-\widetilde{\boldsymbol{H}}\langle\underline{\beta}\rangle = \sum\nolimits_i \underline{x}_i\underline{x}_i^T \cdot \left((g^{-1})'\langle\eta_i\rangle\right)^2 \cdot \frac{1}{V\langle\mu_i\rangle} \cdot \frac{\omega_i}{\phi} \; .$$

Damit ist der Scoring-Algorithmus festgelegt.

Die Matrix $-\widetilde{\boldsymbol{H}}$ hat auch eine zentrale Bedeutung bei der asymptotischen Verteilung der Schätzung und trägt deshalb einen Namen: Sie heisst **Fisher-Information** und wird als $\boldsymbol{J}_n\langle\underline{\beta}\rangle$ notiert. Der Index n soll daran erinnern, dass es sich um die Summe der „Fisher-Informationen" aller Beobachtungen handelt.

d Wir wollen eine Überlegung anführen, die uns zu Vertrauterem führt: Man kann unschwer sehen, dass die Korrektur-Schätzung $\Delta\underline{\beta}$ im Scoring-Algorithmus als Lösung eines gewichteten Kleinste-Quadrate-Problems geschrieben werden kann. Ein solches Problem besteht in der Minimierung des Ausdrucks $\sum_i w_i(\widetilde{y}_i - \underline{x}_i^T\underline{\beta})^2$ mit vorgegebenen Gewichten w_i. (Die w_i sind nicht die ω_i des verallgemeinerten linearen Modells! Wir schreiben $\widetilde{y}_i$ statt einfach y_i, um eine Verwechslung mit den bisher verwendeten y_i zu vermeiden.)

Die Lösung dieses Problems lautet

$$\widehat{\underline{\beta}} = \left(\sum\nolimits_i w_i\,\underline{x}_i\underline{x}_i^T\right)^{-1} \sum\nolimits_i w_i\,\underline{x}_i\widetilde{y}_i \; .$$

Diese Schätzung besteht also auch aus einer Matrix, die eine Summe darstellt und invertiert wird, multipliziert mit einer Summe von Vektoren. Wenn wir Gewichte w_i einführen als

$$w_i = \left((g^{-1})'\langle\widehat{\eta}_i\rangle\right)^2 \frac{1}{V\langle\mu_i\rangle} \cdot \frac{\omega_i}{\phi} \; ,$$

dann stimmt die zu invertierende Matrix in beiden Fällen überein. Nun setzen wir $\widetilde{y}_i = r_i^{(L)}$, wobei

$$r_i^{(L)} = (y_i - \widehat{\mu}_i)\,\frac{d\eta}{d\mu}\langle\mu_i\rangle = r_i \cdot g'\langle\widehat{\mu}_i\rangle$$

die Prädiktor-Residuen sind, die in 9.5.a erwähnt wurden. Jetzt stimmt auch $\underline{s}_i\langle\beta\rangle$ mit $\underline{x}_i w_i\widetilde{y}_i$ überein, und die Lösung $\widehat{\underline{\beta}}$ des gewichteten Kleinste-Quadrate-Problems liefert die Korrektur $\Delta\beta$.

Es ist üblich, auf beiden Seiten noch die vorhergehende Schätzung $\underline{\beta}^{(s)}$ dazu zu zählen – rechts in der Form $(\widetilde{\boldsymbol{H}}\langle\underline{\beta}\rangle)^{-1}\widetilde{\boldsymbol{H}}\langle\underline{\beta}\rangle\,\underline{\beta}^{(s)}$. Man erhält

$$\begin{aligned}\underline{\beta}^{(s+1)} &= \underline{\beta}^{(s)} + \Delta\underline{\beta} = (\boldsymbol{H}\left\langle\underline{\beta}^{(s)}\right\rangle)^{-1}\boldsymbol{H}\left\langle\underline{\beta}^{(s)}\right\rangle\,\underline{\beta}^{(s)} + (\boldsymbol{H}\left\langle\underline{\beta}^{(s)}\right\rangle)^{-1} \cdot \sum\nolimits_i \underline{s}\left\langle y_i, \underline{x}_i; \underline{\beta}^{(s)}\right\rangle \\ &= (\boldsymbol{H}\left\langle\underline{\beta}^{(s)}\right\rangle)^{-1} \cdot \sum\nolimits_i w_i\,\underline{x}_i\,(\underline{x}_i^T\underline{\beta}^{(s)} + r_i^{(L)}) \; .\end{aligned}$$

Man kann also die korrigierte Schätzung $\underline{\beta}^{(s+1)}$ direkt als gewichtete Kleinste-Quadrate-Lösung erhalten, indem man $\widetilde{y}_i = \underline{x}_i^T \underline{\beta}^{(s)} + r_i^{(L)}$ statt $\widetilde{y}_i = r_i^{(L)}$ setzt.

e **Asymptotische Verteilung.** Die „Einkleidung“ des Verbesserungsschrittes des Scoring-Algorithmus als gewichtetes Kleinste-Quadrate-Problem ist nützlich, um die Verteilung der Schätzfunktion $\widehat{\underline{\beta}}$ zu studieren. Man kann zeigen, dass die asymptotische Verteilung gerade die ist, die die gewichtete Kleinste-Quadrate-Schätzung hat, wenn man „vergisst“, dass die Beobachtungen $\widetilde{y}_i$ und die Gewichte w_i von den Schätzwerten selber abhängen (und die Lösungswerte $\widehat{\underline{\beta}}$ einsetzt).

Das gleiche Ergebnis liefert auch die allgemeine Theorie der Maximum-Likelihood-Schätzung: Der geschätzte Parametervektor ist asymptotisch normalverteilt und erwartungstreu mit der inversen Fisher-Information als Kovarianzmatrix,

$$\widehat{\underline{\beta}} \overset{a}{\sim} \mathcal{N}_p\left\langle \underline{\beta}, (\widetilde{\boldsymbol{H}}\left\langle \underline{\beta} \right\rangle)^{-1} \right\rangle \ .$$

(* Der Zusammenhang zwischen dem Scoring-Algorithmus und der asymptotischen Verteilung gilt allgemein für Maximum-Likelihood- und M-Schätzungen. Interessierte können versuchen, dies mit Hilfe der Einflussfunktion, die in der robusten Statistik eingeführt wurde, nachzuvollziehen.)

Mit diesem Ergebnis lassen sich in der üblichen Weise Tests und Vertrauensintervalle angeben, die asymptotisch den richtigen Fehler erster Art respektive den richtigen Vertrauenskoeffizienten haben. Tests, die auf der genäherten asymptotischen Normalverteilung der Schätzungen beruhen, heissen **Wald-Tests**.

10 Kategorielle Zielgrössen

10.1 Multinomiale Zielgrössen

a In der logistischen Regression war die Zielgrösse zweiwertig. Im Beispiel der Umweltumfrage (8.2.d) hatte die Zielgrösse „Beeinträchtigung“ eigentlich vier mögliche Werte, die wir für das dortige Modell zu zwei Werten zusammengefasst haben. Die vier Werte zeigen eine Ordnung von „gar nicht“ bis „stark“. In der gleichen Umfrage wurde auch eine weitere Frage gestellt: „Wer trägt im Umweltschutz die Hauptverantwortung? – Einzelne, der Staat oder beide?“. Diese drei Auswahlantworten haben keine eindeutige Ordnung, denn vielleicht nehmen jene, die mit „beide“ antworten, den Umweltschutz besonders ernst, und deshalb liegt diese Antwort nicht unbedingt zwischen den beiden anderen.

Hier soll zunächst ein **Modell für eine ungeordnete, kategorielle Zielgrösse** behandelt werden. Im nächsten Abschnitt wird der Fall einer geordneten Zielgrösse untersucht.

b **Modell.** Für eine einzelne Beobachtung bildet das Modell eine einfache Erweiterung des Falles der zweiwertigen Zielgrösse. Wir müssen festlegen, wie die Wahrscheinlichkeiten $P\langle Y_i = k\rangle$ der möglichen Werte k von den Werten $\underline{x}_i$ der Regressoren abhängen.

Die möglichen Werte der Zielgrösse wollen wir mit 0 beginnend durchnummerieren, damit die zweiwertige Zielgrösse ein Spezialfall der allgemeineren Formulierung wird. Zunächst zeichnen wir eine Kategorie als „**Referenzkategorie**“ aus. Wir wollen annehmen, dass es die Kategorie $k=0$ sei.

Eine einfache Erweiterung des logistischen Modells besteht nun darin, dass wir für jedes $k \geq 1$ für das logarithmierte Wettverhältnis gegenüber der Referenzkategorie ein separates lineares Modell ansetzen,

$$\log\left\langle \frac{P\langle Y_i = k\rangle}{P\langle Y_i = 0\rangle} \right\rangle = \log\left\langle \frac{\pi_i^{(k)}}{\pi_i^{(0)}} \right\rangle = \eta_i^{(k)} = \beta_0^{(k)} + \sum\nolimits_j \beta_j^{(k)} x_i^{(j)} \quad k = 1, 2, ..., k^* \ .$$

Zunächst scheint es, dass je nach Wahl der Referenzkategorie ein anderes Modell herauskommt. Es zeigt sich aber, dass sich diese Modelle nicht wirklich unterscheiden (ähnlich wie es in der Varianzanalyse keine wesentliche Rolle spielt, welche Kategorie, welches Niveau eines Faktors, im formalen Modell weggelassen wird, um die Lösung eindeutig zu machen).

c* Wählen wir beispielsweise $k=1$ statt $k=0$ als Referenz. Für $k \geq 2$ ergibt sich

$$\begin{aligned} \log\left\langle \frac{P\langle Y_i = k \mid \underline{x}_i\rangle}{P\langle Y_i = 1 \mid \underline{x}_i\rangle} \right\rangle &= \log\left\langle \frac{P\langle Y_i = k \mid \underline{x}_i\rangle}{P\langle Y_i = 0 \mid \underline{x}_i\rangle} \right\rangle - \log\left\langle \frac{P\langle Y_i = 1 \mid \underline{x}_i\rangle}{P\langle Y_i = 0 \mid \underline{x}_i\rangle} \right\rangle \\ &= \beta_0^{(k)} + \sum\nolimits_j \beta_j^{(k)} x_i^{(j)} - \beta_0^{(1)} + \sum\nolimits_j \beta_j^{(1)} x_i^{(j)} \\ &= (\beta_0^{(k)} - \beta_0^{(1)}) + \sum\nolimits_j \left(\beta_j^{(k)} - \beta_j^{(1)}\right) x_i^{(j)} \ . \end{aligned}$$

Das hat genau die selbe Form wie das Ausgangsmodell, wenn man die Differenzen $(\beta_j^{(k)} - \beta_j^{(1)})$ als neue Koeffizienten $\widetilde{\beta}_j^{(k)}$ einsetzt. Für $k=0$ muss man $\widetilde{\beta}_j^{(0)} = -\beta_j^{(1)}$ setzen.

d* **Gruppierte Daten**. Wie in der logistischen Regression (10.2.n) kann man die Beobachtungen mit gleichen Werten der Eingangsgrössen zusammenfassen und zählen, wie viele von ihnen die verschiedenen Werte k der Zielgrösse zeigen. Es sei wieder m_ℓ die Anzahl der Beobachtungen mit $\underline{x}_i = \underline{\widetilde{x}}_\ell$, und $\widetilde{Y}_\ell^{(k)}$ der *Anteil* dieser Beobachtungen, für die $Y_i = k$ ist. Die *Anzahlen* $m_\ell \cdot \widetilde{Y}_\ell^{(k)}$ folgen dann der multinomialen Verteilung mit den Parametern $\widetilde{\pi}_\ell^{(1)}, ..., \widetilde{\pi}_\ell^{(k^*)}$, die durch das oben angegebene Modell bestimmt sind. Die Wahrscheinlichkeiten sind

$$\begin{aligned} P\left\langle \underline{\widetilde{Y}}_\ell = \underline{\widetilde{y}}_\ell \right\rangle &= P\langle m_\ell \widetilde{Y}_0 = m_\ell \widetilde{y}_0,\, m_\ell \widetilde{Y}_1 = m_\ell \widetilde{y}_1,\, \dots,\, m_\ell \widetilde{Y}_{k^*} = m_\ell \widetilde{y}_{k^*} \rangle \\ &= \frac{m_\ell !}{(m_\ell \widetilde{y}_\ell^{(0)})! \cdot \ldots \cdot (m_\ell \widetilde{y}_\ell^{(k^*)})!} \, (\widetilde{\pi}_\ell^{(0)})^{m_\ell \widetilde{y}_\ell^{(0)}} (\widetilde{\pi}_\ell^{(2)})^{m_\ell \widetilde{y}_\ell^{(2)}} \cdot \ldots \cdot (\widetilde{\pi}_\ell^{(k^*)})^{m_\ell \widetilde{y}_\ell^{(k^*)}} \, . \end{aligned}$$

Die multinomiale Verteilung bildet eine multivariate Exponentialfamilie. Mit einer geeigneten Link-Funktion versehen, legt die multinomiale Verteilung ein multivariates verallgemeinertes lineares Modell fest. Die kanonische Link-Funktion ist diejenige, die durch das angegebene Modell beschrieben wird.

e Die Tatsache, dass für zusammengefasste Beobachtungen eine multinomiale Verteilung entsteht, erklärt den Namen **multinomiales Logit-Modell** für das oben formulierte Modell. Es ist recht flexibel, denn es erlaubt für jeden möglichen Wert k der Zielgrösse eine eigene Form der Abhängigkeit ihrer Wahrscheinlichkeit von den Regressoren. Ein positiver Koeffizient $\beta_j^{(k)} > 0$ bedeutet für zunehmendes $x^{(j)}$ eine steigende Neigung zur Kategorie k im Verhältnis zur Neigung zur Referenzkategorie 0.
Die Flexibilität bedingt, dass recht viele Parameter zu schätzen sind; die Anzahl ist das Produkt aus k^* und der Anzahl Prädiktoren (plus 1 für die Achsenabschnitte $\beta_0^{(k)}$). Mit kleinen Datensätzen sind diese Parameter schlecht bestimmt.

f **S-Funktionen**. Im Statistik-System R steht im package `nnet` die Funktion `multinom` zur Verfügung, um solche Modelle anzupassen. Für das **Beispiel der Umweltumfrage** zeigt Tabelle 10.1.f ein `summary` des Modells, das die Frage nach der Hauptverantwortung in Abhängigkeit vom Alter und Geschlecht der Befragten beschreibt. Man kann die geschätzten Koeffizienten $\widehat{\beta}_{j\ell}$ und ihre Standardfehler ablesen.

```
Call:
multinom(formula = Hauptv ~ Alter + Schulbildung + Beeintr + Geschlecht,
    data = t.d)

Coefficients:
     (Intercept)    Alter Sch.Lehre Sch.ohne.Abi Sch.Abitur Sch.Studium
Staat       0.599 -0.00270    -0.518       -0.500      -0.66      -0.366
beide      -1.421  0.00262    -0.562       -0.257       0.34       0.220
      Beeintretwas Beeintrziemlich Beeintrsehr Geschlechtw
Staat       -0.722          -0.719      -0.685      -0.244
beide        0.135           0.106       0.716      -0.179

Std. Errors:
      (Intercept)   Alter Sch.Lehre Sch.ohne.Abi Sch.Abitur Sch.Studium
Staat       0.228 0.00340     0.149        0.174      0.221       0.231
beide       0.349 0.00495     0.234        0.257      0.284       0.307
      Beeintretwas Beeintrziemlich Beeintrsehr Geschlechtw
Staat        0.123           0.163       0.243       0.107
beide        0.179           0.224       0.271       0.154

Residual Deviance: 3385
AIC: 3425
```

Tabelle 10.1.f: Ergebnisse einer multinomialen Logit-Regression im Beispiel der Umweltumfrage

Die Referenzkategorie ist „Einzelne“. Der Koeffizient von j = Alter für k = Staat ist $\widehat{\beta}_j^{(k)} = -0.00270$. In 50 Jahren nehmen also die log odds von „Staat“:„Einzelne“ um $0.0027 \cdot 50 = 0.135$

ab; als odds ratio ergibt sich $\exp\langle -0.135\rangle = 0.874$. Allerdings ist der Koeffizient nicht signifikant, da $\widehat{\beta}_j^{(k)}/\text{standard error}_j^{(k)} = -0.0027/0.0034 = 0.79$ einen klar nicht signifikanten z-Wert ergibt. Zwischen den Geschlechtern besteht ein signifikantes Doppelverhältnis von $\exp\langle -0.244\rangle = 0.78$. Frauen weisen die Verantwortung stärker den Einzelnen anstelle des Staates zu als Männer.

g Ob eine **Eingangsgrösse** einen **Einfluss** auf die Zielgrösse hat, sollte man nicht an den einzelnen Koeffizienten festmachen, da ja k^* Koeffizienten null sein müssen, wenn kein Einfluss da ist. Es muss also ein grösseres mit einem kleineren Modell verglichen werden, und das geschieht wie üblich mit den log-likelihoods oder den Devianzen.

S-Funktionen. Im R-System sieht die Funtion `drop1` für multinomiale Modelle leider keinen Test vor. Man muss mit der Funktion `anova` die einzelnen Modelle vergleichen (oder `drop1` entsprechend ergänzen). Tabelle 10.1.g zeigt die Resultate einer erweiterten Funktion `drop1`, die den Test durchführt, für ein ausführlicheres Modell.

Erstaunlicherweise haben weder die politische Partei, noch das Alter oder die Wohnlage einen signifikanten Einfluss auf die Zuweisung der Hauptverantwortung. Das liegt nicht an einem starken Zusammenhang der Eingangs-Variablen analog zum Kollinearitätsproblem, das in der linearen Regression besprochen wurde, denn auch bei einer schrittweisen Elimination bleiben diese drei Variablen nicht-signifikant.

	Df	AIC	Chisq	p.value
<none>	58	3436	NA	NA
Alter	56	3433	1.35	0.508
Schulbildung	50	3454	34.00	0.000
Beeintr	52	3488	64.34	0.000
Geschlecht	56	3437	5.56	0.062
Ortsgroesse	46	3455	43.10	0.000
Wohnlage	46	3422	9.82	0.632
Partei	44	3418	10.56	0.720

Tabelle 10.1.g: Signifikanzen von einzelnen Termen im Beispiel der Umweltumfrage

h* Wenn man kein geeignetes Programm zur Verfügung hat, kann man die $\beta_j^{(k)}$ für die verschiedenen k getrennt schätzen, indem man k^* logistische Regressionen rechnet, jeweils mit den Daten der Kategorie k und der Referenzkategorie. Das gibt zwar leicht andere Resultate, aber die Unterschiede sind nicht allzu gross, wenn die Referenzkategorie einen genügenden Anteil der Beobachtungen umfasst.

Eine Möglichkeit, die genauen Schätzungen zu erhalten, führt über eine andere Anordnung der Daten, die in 11.2.l besprochen wird.

i Die **Residuen-Devianz** ist wie in der logistischen Regression (8.3.i) sinnvoll bei Daten, die zu Anzahlen zusammengefasst werden können (mit $m_\ell > 3$ oder so). Hier wird die maximale Likelihood erreicht für $\widehat{\pi}_\ell^{(k)} = \widetilde{y}_\ell^{(k)}$ und man erhält

$$D\langle \underline{\widetilde{y}}; \underline{\widehat{\pi}}\rangle = 2(\ell\ell^{(M)} - \ell\ell\langle \underline{\widetilde{y}}; \underline{\widehat{\pi}}\rangle) = 2\sum\nolimits_{\ell,k} m_\ell \widetilde{y}_\ell^{(k)} \log\left\langle \frac{\widetilde{y}_\ell^{(k)}}{\widehat{\pi}_\ell^{(k)}} \right\rangle .$$

Dies gilt für alle möglichen Links zwischen den Wahrscheinlichkeiten $\underline{\pi}$ und den Koeffizienten $\beta_j^{(k)}$ der linearen Prädiktoren.

j Eine weitere Anwendung des multinomialen Logitmodells ist die **Diskriminanzanalyse mit mehr als 2 Kategorien**. Ähnlich wie beim binären logistischen Modell schätzt man einen Score aus der Modellgleichung für jede Kategorie. Dann ordnet man die Beobachtung derjenigen Kategorie zu, für die der lineare Prädiktor maximal ist.

k* Ein noch allgemeineres Modell erlaubt es, die Eingangs-Variablen von den möglichen Werten der Zielgrösse abhängig zu machen.

$$\log\left\langle \frac{P\langle Y_i = k \mid \underline{x}_i\rangle}{P\langle Y_i = 0 \mid \underline{x}_i\rangle} \right\rangle = \beta_0^{(k)} + \sum\nolimits_j \beta_j^{(k)} x_i^{(jk)} .$$

Es werden also jeweils 2 Wahlmöglichkeiten miteinander verglichen. Man erlaubt für jedes Verhältnis eine andere Wirkung der Eingangsgrössen.

Diese Form wird auch „Discrete Choice Models“ genannt, da sie bei Studien des Wahlverhaltens von Konsumenten verwendet wird.

Literatur: Agresti (2002), Kap. 9, Fahrmeir and Tutz (2001), Kap. 3.2.

l **Residuen-Analyse.** Was Residuen sein sollen, ist im Zusammenhang mit der multinomialen Regression nicht klar. Zunächst gibt es für jede der logistischen Regressionen, auf denen sie beruht, die entsprechenden Residuen, und diese hängen von von der Referenzkategorie ab. Man könnte also für jedes Paar von Werten der Zielgrösse für jede Beobachtung ein Residuum definieren. Wie diese in geeigneter Form gemeinsam dargestellt werden können, ist dem Autor zurzeit noch zu wenig klar. Hinweise werden gerne entgegen genommen.

10.2 Geordnete Zielgrössen

a Wie früher erwähnt (7.1.a), haben Variable oft einen geordneten Wertebereich. Wie kann man diesen Aspekt ausnützen, wenn eine solche Grösse die Zielgrösse einer Regression ist?

Im **Beispiel der Umweltumfrage** (7.1.c) interessierte uns die Frage nach der Beeinträchtigung mit ihren geordneten Antwortmöglichkeiten von „überhaupt nicht“ bis „sehr“. Bei der Auswertung mit Kreuztabellen wurde diese Ordnung nicht berücksichtigt. Nun soll sie als Zielgrösse betrachtet und ihr Zusammenhang mit Eingangsgrössen wie Schulbildung, Geschlecht und Alter untersucht werden.

b **Modell.** Zur Beschreibung eines Modells hilft, wie für die binäre Zielgrösse (8.2.j), die Annahme einer **latenten Variablen** Z, aus der sich die Kategorien der Zielgrösse durch Klassieren ergeben. Das frühere Modell wird erweitert, indem man mehrere **Schwellenwerte** α_k festlegt. Die Zielgrösse Y ist =0, wenn Z kleiner ist als die kleinste Schwelle α_1, sie ist =1, wenn Z zwischen α_1 und α_2 liegt, usw. Bei k^* Schwellenwerten nimmt Y die k^*+1 Werte $0, 1, ..., k^*$ an.

In Formeln:

$$\begin{aligned} Y = 0 &\iff Z < \alpha_1 \\ Y = k &\iff \alpha_k \le Z < \alpha_{k+1} \qquad k = 1, ..., k^* - 1 \\ Y = k^* &\iff \alpha_{k^*} \le Z . \end{aligned}$$

Das bedeutet, dass

$$P\langle Y \ge k\rangle = P\langle Z \ge \alpha_k\rangle \qquad k = 1, \ldots, k^* .$$

Für die latente Variable Z soll der Einfluss der Eingangsgrössen durch eine multiple lineare Regression gegeben sein, also

$$Z_i = \beta_0 + \sum\nolimits_j x_i^{(j)} \beta_j + E_i .$$

Der Fehlerterm in dieser Regression hat einen bestimmten Verteilungstyp F, z. B. eine logistische oder eine Normalverteilung.

Abbildung 10.2.b veranschaulicht diese Vorstellung für eine einzige Eingangs-Variable. Bei mehreren Eingangsgrössen wäre auf der horizontalen Achse, wie üblich, der lineare Prädiktor $\eta_i = \underline{x}_i^T \underline{\beta}$ zu verwenden.

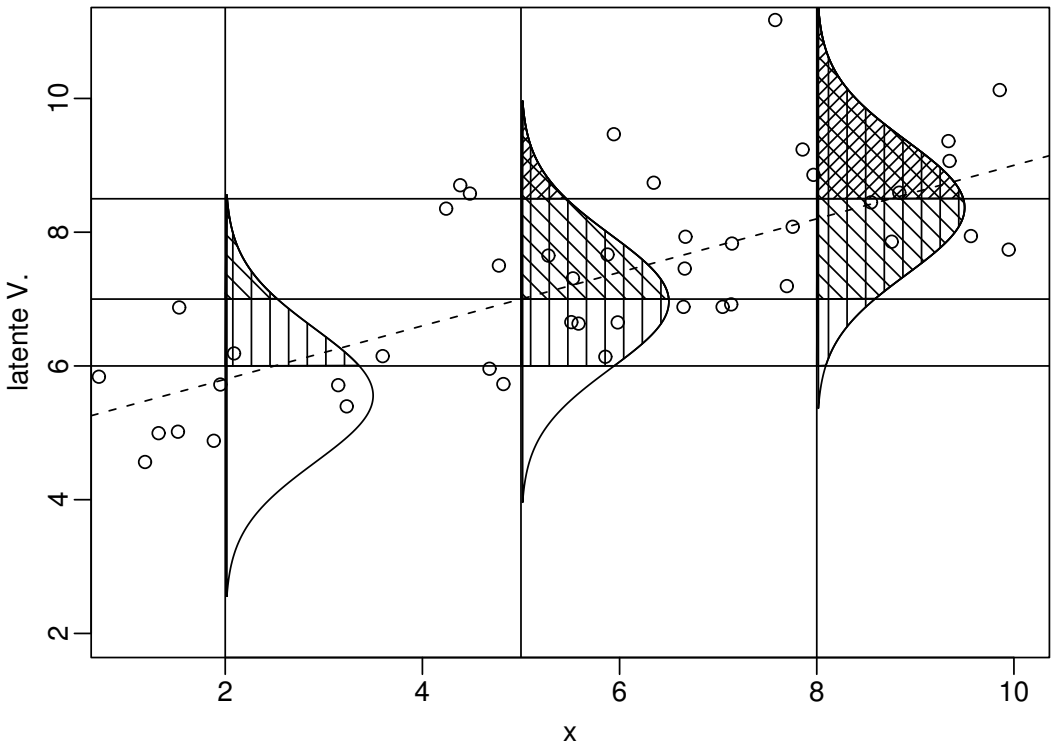


Abbildung 10.2.b: Zum Modell der latenten Variablen

c Wir betrachten die Ereignisse $\{Y_i \geq k\} = \{Z_i \geq \alpha_k\}$ und erhalten für ihre Wahrscheinlichkeiten

$$\begin{aligned} \gamma_k\langle \underline{x}_i\rangle := P\langle Y_i \geq k\rangle &= P\langle Z_i > \alpha_k\rangle = P\Big\langle E_i > \alpha_k - \beta_0 - \textstyle\sum_j \beta_j x_i^{(j)}\Big\rangle \\ &= 1 - F\Big\langle \alpha_k - \Big(\beta_0 + \textstyle\sum_j \beta_j x_i^{(j)}\Big)\Big\rangle \ , \end{aligned}$$

wobei F die kumulative Verteilungsfunktion der Zufallsabweichungen E_i bezeichnet.

d Man sieht leicht, dass β_0 unbestimmt ist, da wir zu jedem Schwellenwert α_k eine Konstante hinzuzählen und diese von β_0 abzählen können, ohne dass sich die Y_i ändern. Wir setzen daher $\beta_0 = 0$. – Die Streuung der latenten Variablen ist ebenfalls nicht bestimmt. Wir können Z und alle Schwellenwerte mit einer Konstanten multiplizieren, ohne Y_i zu ändern. Für die kumulative Verteilungsfunktion F der Zufallsfehler kann man daher eine feste Verteilung, ohne den in der multiplen Regression üblichen Streuungsparameter σ, annehmen.

Wenn wir jetzt, wie bei der Regression mit binärer Zielgrösse, $1 - F\langle -\eta\rangle =: g^{-1}\langle\eta\rangle$ setzen, wird

$$g\langle\gamma_k\langle\underline{x}_i\rangle\rangle = \textstyle\sum_j \beta_j x_i^{(j)} - \alpha_k$$

Für jeden Schwellenwert α_k ergibt sich also ein Regressions-Modell mit der binären Zielgrösse, die 1 ist, wenn $Y \geq k$ ist. Diese Modelle sind miteinander verknüpft, da für alle die gleichen Koeffizienten β_j der Regressoren vorausgesetzt werden.

Die üblichste Wahl der Link-Funktion ist wieder die Logit-Funktion. Man spricht dann vom Modell der **kumulativen Logits**. Die inverse Link-Funktion g^{-1} ist dann die logistische Funktion, und die Verteilung der $-E_i$ ist damit die logistische Verteilung.

e Die **Schwellenwerte** α_k müssen nicht etwa gleich-abständig sein. Sie sind unbekannt, und man wird versuchen, sie gleichzeitig mit den Haupt-Parametern β_j zu schätzen. In der Regel sind sie Hilfsparameter, die nicht weiter interessieren.

f Der Name **kumulatives Modell** bezeichnet die Tatsache, dass das Modell die Wahrscheinlichkeiten $P\langle Y \geq k\rangle$, also für die „von oben her kumulierten" Wahrscheinlichkeiten der möglichen Werte k von Y, festlegt.

In Büchern und Programmen wird üblicherweise umgekehrt ein Modell für die „von unten her kumulierten" Wahrscheinlichkeiten formuliert. Das hat den Nachteil, dass diese Wahrscheinlichkeiten mit zunehmendem $\underline{x}^T\underline{\beta}$ abnehmen, so dass positive Koeffizienten β_j einen negativen Zusammenhang der betreffenden Eingangs-Variablen mit der Zielgrösse bedeuten. Wenn so vorgegangen wird, wie wir es hier getan haben, dann bedeutet dagegen ein positiver Koeffizient β_j, dass eine Zunahme von $x^{(j)}$ zu einer Zunahme von Y (oder der latenten Variablen Z) führt. Zudem wird der Fall der Regression mit einer binären Zielgrösse, insbesondere die logistische Regression, ein Spezialfall des neuen Modells, nämlich der Fall von $k^* = 1$.

g Die Wahrscheinlichkeiten für die einzelnen Kategorien erhält man aus sukzessiven Differenzen,

$$P\langle Y_i = k\rangle = \gamma_k\langle \underline{x}_i\rangle - \gamma_{k+1}\langle \underline{x}_i\rangle$$

h Bei einer logistischen Verteilung hat man den Vorteil, dass das Ergebnis mit Hilfe der **Wettverhältnisse** (odds) interpretiert werden kann. Dazu wird jeweils das Wettverhältnis bezüglich eines Schwellenwerts gebildet („cumulative odds"): Wahrscheinlichkeit für niedrigere Kategorien vs. Wahrscheinlichkeit für höhere Kategorien

$$\text{odds}\langle Y_i \geq k \mid \underline{x}_i\rangle = \frac{P\langle Y_i \geq k\rangle}{P\langle Y_i < k\rangle} = \frac{\gamma_k}{1-\gamma_k} = \exp\langle -\alpha_k\rangle \cdot \exp\langle \beta_1\rangle^{x^{(1)}} \cdot \ldots \cdot \exp\langle \beta_m\rangle^{x^{(m)}} \ .$$

Die Eingangsgrössen wirken auf alle Unterteilungen $Y_i < k$ vs. $Y_i \geq k$ gleich. Die einzelnen Regressoren wirken multiplikativ auf die Wettverhältnisse. Ein solches Modell heisst deshalb Modell der proportionalen Verhältnisse, **proportional-odds model**.

Die Formel vereinfacht sich noch, wenn man die logarithmierten **Doppelverhältnisse** (log odds ratios) für verschiedene Werte $\underline{x}_i$ der Regressoren betrachtet,

$$\log\left\langle \frac{\text{odds}\langle Y_1 \geq k \mid \underline{x}_1\rangle}{\text{odds}\langle Y_2 \geq k \mid \underline{x}_2\rangle} \right\rangle = \beta_1 \cdot (x_1^{(1)} - x_2^{(1)}) + \cdots + \beta_m \cdot (x_1^{(m)} - x_2^{(m)}) \ .$$

In dieser Gleichung kommt α_k nicht vor. Die Doppelverhältnisse sind also für alle Kategorien k der Zielgrösse gleich!

Wenn $x^{(j)}$ nur eine Indikatorvariable ist, die Behandlung B_0 von Behandlung B_1 unterscheidet, so ist der Koeffizient $\beta^{(j)}$ ein Mass für den Behandlungs-Effekt („unit risk"), der gemäss dem Modell für alle Schwellenwerte gleich ist.

i* Für die logistische Regression wurden neben der Verwendung der logit-Funktion als **Link** noch zwei weitere vorgestellt. Zunächst wurde erwähnt, dass die Annahme einer Normalverteilung für die latente Variable zur Probit-Funktion führt, dass aber die Unterschiede höchstens in riesigen Datensätzen spürbar werden könnten; die beiden Verteilungen unterscheiden sich nur in den Schwänzen, und diese werden mit den hier betrachteten Beobachtungen nur ungenau erfasst. Die Verwendung der Probit-Funktion hat den Nachteil, dass die Interpretation der Koeffizienten über ihre Veränderung der log odds nicht mehr (genau) gilt.

j* Die dritte gebräuchliche Link-Funktion war die „**komplementäre Log-Log-Funktion**“

$$g\langle\mu\rangle = \log\left\langle -\log\langle 1-\mu\rangle \right\rangle , \quad 0<\mu<1$$

Die entsprechende inverse Link-Funktion ist $g^{-1}(\eta) = 1 - \exp\langle -\exp\langle\eta\rangle\rangle$, und das ist die Verteilungsfunktion der Gumbel-Verteilung.

Für Überlebens- oder Ausfallzeiten bewährt sich die Weibull-Verteilung. Logarithmiert man solche Variable, dann erhält man die Gumbel-Verteilung. Hinter einer Gumbel-verteilten Zielgrösse mit additiven Wirkungen der Regressoren steht oft die Vorstellung einer Weibull-verteilten Grösse und multiplikativen Wirkungen.

k* In der Literatur gibt es neben dem kumulativen Logit-Modell für geordnete Zielgrössen auch das Modell, das für aufeinanderfolgende Kategorien proportionale Wettverhältnisse postuliert. Clogg and Shihadeh (1994) zeigt, dass die Normalverteilung der latenten Variablen dieses Modell der **adjacent classes logits** näherungsweise rechtfertigt.

l **S-Funktionen**. Im R findet man die Funktion `polr`, was für „Proportional Odds Logistic Regression“ steht. Das `summary` (Tabelle 10.2.l (i)) liefert, wie üblich, die Tabelle der Koeffizienten mit Werten der t-Statistik für die Tests $\beta_j = 0$, die für Faktoren mit mehr als 2 Werten wenig Sinn machen. (Die P-Werte werden nicht mitgeliefert; man muss sie selbst ausrechnen.)

```
Call: polr(formula = Beeintr ~ Alter + Schule + Geschlecht
           + Ortsgroesse, data = t.d)
Coefficients:
                            Value Std. Error t value p.value
Alter                     -0.00268    0.00299 -0.8992   0.369
SchuleLehre                0.08594    0.13937  0.6166   0.538
Schuleohne.Abi             0.63084    0.15546  4.0578   0.000
SchuleAbitur               0.81874    0.18502  4.4251   0.000
SchuleStudium              1.07522    0.19596  5.4869   0.000
Geschlechtw                0.00699    0.09110  0.0768   0.939
Ortsgroesse2000-4999       0.57879    0.27104  2.1354   0.033
Ortsgroesse5000-19999      0.58225    0.23455  2.4825   0.013
Ortsgroesse20000-49999     0.85579    0.27155  3.1515   0.002
Ortsgroesse50000-99999     0.60140    0.29400  2.0456   0.041
Ortsgroesse100000-499999   0.87548    0.23167  3.7790   0.000
Ortsgroesse>500000         1.10828    0.21568  5.1386   0.000

Intercepts:
               Value  Std. Error t value
nicht|etwas     0.995  0.273      3.644
etwas|ziemlich  2.503  0.278      9.007
ziemlich|sehr   3.936  0.290     13.592

Residual Deviance: 4114.67
AIC: 4144.67
```

Tabelle 10.2.l (i): Resultate für die Regression der geordneten Zielgrösse Beeinträchtigung auf mehrere Eingangsgrössen im Beispiel der Umweltumfrage

Wie in früheren Modellen zeigt die Funktion `drop1(t.r,test="Chisq")` die Signifikanz der Faktoren (Tabelle 10.2.l (ii)).

Achtung! Eine kleine Simulationsstudie mit 500 Beobachtungen und 2-3 Variablen (davon ein 3-4-stufiger Faktor) und einer Zielgrösse mit 3 Werten hat alarmierende Resultate gebracht: Die ausgewiesenen Standardfehler waren um einen Faktor von 2 bis 3 zu klein. Die Resultate von

```
Model:
Beeintr ~ Alter + Schule + Geschlecht + Ortsgroesse
             Df  AIC    LRT Pr(Chi)
<none>          4145
Alter         1 4143      1    0.369
Schule        4 4196     59    0.000 ***
Geschlecht    1 4143 0.0059    0.939
Ortsgroesse   6 4174     42    0.000 ***
```

Tabelle 10.2.l (ii): Signifikanz der einzelnen Terme im Beispiel

`polr` stimmten zudem schlecht mit einer alternativen Berechnungsmethode überein, die gleich geschildert wird. Die Resultate sind also mit äusserster Vorsicht zu geniessen. Es ist bis auf Weiteres angezeigt, die Bootstrap-Methode zu benützen, um die Unsicherheiten zu erfassen. Für Vorhersagen der richtigen Klasse sind die Methoden vermutlich zuverlässiger.

Die Resultate für das **Beispiel der Umweltumfrage** zeigen auch hier, dass Schulbildung und Ortsgrösse einen klaren Einfluss auf die Beurteilung der Beeinträchtigung haben, während Alter und Geschlecht keinen Einfluss zeigen. (Die P-Werte für die beiden letzteren konnten schon in der ersten Tabelle abgelesen werden, da beide nur einen Freiheitsgrad haben.)

m* Man kann das Modell auch **mit Hilfe einer Funktion für die logistische Regression** anpassen. Dazu muss man allerdings die Daten speziell arrangieren. Aus jeder Beobachtung Y_i machen wir k^* Beobachtungen Y_{ik}^* nach der Regel

$$\widetilde{Y}_{ik}^* = \begin{cases} 1 & \text{falls} \quad Y_i \geq k \\ 0 & \text{falls} \quad Y_i < k \end{cases}$$

oder, tabellarisch,

	Y_{i1}^*	Y_{i2}^*	Y_{i3}^*
$Y_i = 0$	0	0	0
1	1	0	0
2	1	1	0
3	1	1	1

Gleichzeitig führt man als Eingangs-Variable einen Faktor $X^{(Y)}$ ein, dessen geschätzte Haupteffekte die Schwellenwerte α_k sein werden. Die neue Datenmatrix besteht jetzt aus n Gruppen von k^* Zeilen. Die k-te Zeile der Gruppe i enthält Y_{ik}^* als Wert der Zielgrösse, k als Wert von $X^{(Y)}$ und die $x_i^{(j)}$ als Werte der anderen Regressoren. Mit diesen $n \cdot k^*$ „Beobachtungen“ führt man nun eine logistische Regression durch.

n* Wie bei der binären und der multinomialen Regression kann man Beobachtungen mit gleichen Werten $\underline{x}_i$ der Regressoren zusammenfassen. Die Zielgrössen sind dann

$$\widetilde{Y}_\ell^{(k)} = \text{Anzahl}\{i \mid Y_i = k \text{ und } \underline{x}_i = \underline{\widetilde{x}}_\ell\} / m_\ell \,,$$

also die Anteile der Personen mit Regressor-Werten $\underline{\widetilde{x}}_\ell$, die die $k$te Antwort geben.

Die Funktion `polr` erlaubt die Eingabe der Daten in aggregierter Form mittels dem Argument `weights`.

o Im Vergleich mit dem **multinomialen Logit-Modell** muss man im kumulativen Logit-Modell deutlich weniger Parameter schätzen: Anstelle von $k^* \cdot p$ sind es hier $k^* + p$. Deswegen wird man bei ordinalen Kategorien das kumulative Modell vorziehen. Wenn die Annahme der gleichen Steigungen verletzt ist, ist es jedoch sinnvoll, auch ordinale Daten mit einem multinomialen Regressions-Modell auszuwerten. Diese Überlegung zeigt auch, wie man diese Annahme überprüfen kann: Man passt ein multinomiales Logit-Modell an und prüft mit einem Modellvergleichs-Test, ob die Anpassung signifikant besser ist.

(Wenn man es genau nimmt, sind die beiden Modelle allerdings nicht geschachtelt, weshalb die Voraussetzungen für den Test nicht exakt erfüllt sind.)

p **Residuen-Analyse**. Wie für die ungeordneten Zielgrössen sind dem Autor keine dem Modell angepassten Definitionen für Residuen bekannt. Eine sinnvolle Definition erscheint mir die Differenz zwischen dem bedingten Erwartungswert der latenten Variablen Z, gegeben die beobachtete Kategorie und der lineare Prädiktor, und dem Wert des linearen Prädiktors,

$$R_i = \mathcal{E}\langle Z \mid Y_i, \widehat{\eta}_i\rangle - \widehat{\eta}_i \ .$$

Die entsprechende S-Funktion ist im Package `regr0` eingebaut.

10.S S-Funktionen

a **Funktion** `polr`. Die S-Funktion `polr` (proportional odds linear regression) aus dem Package `MASS` passt Modelle mit geordneter Zielgrösse an.

```
> t.r  <- polr(y~x1+x2+..., data=t.d, weigths, ...)
```

Die linke Seite der Formel, `y`, muss ein Faktor sein. Die Niveaus werden in der Reihenfolge geordnet, wie sie unter `levels(t.y)` erscheinen. Damit man keine Überraschungen erlebt, sollte man einen Faktor vom Typ `ordered` verwenden.

```
> t.y  <- ordered(t.d$groups, levels=c("low","medium","high"))
```

Gruppierte Daten können nicht als Matrix eingegeben werden. (Man muss die Anzahlen untereinander schreiben und als `weights` angeben. ...)

b **Funktion** `multinom`. Für multinomiale Regression gibt es die Funktion `multinom`. Sie ist im Package `nnet` versorgt, weil die Berechnung Methoden braucht, die auch für „neural networks“ Anwendung finden. Die linke Seite der Formel kann ein Faktor sein oder für gruppierte Daten, analog zur logistischen Regression, eine Matrix mit k^* Spalten, in denen die Anzahlen mit $Y_i = k$ stehen.

c **Funktion** `regr`. Beide Funktionen sind auch über die Funktion `regr` des Packages `regr0` verfügbar und werden automatisch gewählt, wenn die Zielgrösse ein `ordered`-Faktor resp. ein gewöhnlicher `factor` ist.

Im ersten Fall wird mit `plot` eine spezielle Residuen-Darstellung gewählt, s. oben und Dokumentation zu `regr0`.

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 anteilsmässig 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.)

	Anzahlen			Prozente		
Geschl.	zugel.	abgew.	$\sum$	zugel.	abgew.	$\sum$
w	557	1278	1835	30.4	69.6	100
m	1198	1493	2691	44.5	55.5	100
$\sum$	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 stattfinde. Und jeder hatte damit mehr oder weniger Recht, wie Tabelle 11.1.b(ii) zeigt!

		Anzahlen			Prozente		
Dept.	Geschl.	zugel.	abgew.	$\sum$	zugel.	abgew.	$\sum$
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
	$\sum$	1755	2771	4526	38.8	61.2	100

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

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 Benach-

teiligung! 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.

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

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

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-Variable 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

$$\begin{aligned} &\text{(a)} \qquad \sum\nolimits_h \alpha_h = 0 \\ &\text{(b)} \qquad \alpha_1 = 0 \end{aligned}$$

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

$$\text{(c)} \qquad \sum\nolimits_h \exp\langle\alpha_h\rangle = 1 \; .$$

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

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

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,

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

f Die Testgrösse lässt sich auch vereinfachen zu

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

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-verteilten Zufallsvariablen durch ihren Erwartungswert bereits gegeben ist (Varianz = Erwartungswert = λ), kann man testen, ob die Anzahlen N_{hk} zu stark von den Werten $\widehat{\lambda}_{hk}$ 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“.

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.

```
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

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

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

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.l 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))`.

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 kategorieller Daten mit einem Modell verknüpft. Wir haben den Test auf Unabhängigkeit der beiden Faktoren im Jargon der Poisson-Regression besprochen – das ist

	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.l: Daten des Beispiels der Umwelt-Umfrage

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			
	Einz.	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			
	Einz.	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**

$$\eta_{hk\ell} = \log\langle\lambda_{hk\ell}\rangle = \mu + \alpha_h + \beta_k + \gamma_\ell + (\alpha\beta)_{hk} + (\beta\gamma)_{k\ell} + (\alpha\gamma)_{h\ell} + (\alpha\beta\gamma)_{hk\ell} \; .$$

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): $\eta_{hk\ell} = \mu + \alpha_h + \beta_k + \gamma_\ell$. 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 $\eta_{hk\ell} = \mu + \alpha_h + \beta_k + \gamma_\ell + (\alpha\beta)_{hk}$ 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 $\eta_{hk\ell} = \mu + \alpha_h + \beta_k + \gamma_\ell + (\alpha\beta)_{hk} + (\alpha\gamma)_{h\ell}$ 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 $\eta_{hk\ell} = \mu + \alpha_h + \beta_k + \gamma_\ell + (\alpha\beta)_{hk} + (\beta\gamma)_{k\ell} + (\alpha\gamma)_{h\ell}$ 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.

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 $\log\langle\widehat{\pi}_\ell/\widehat{\pi}_{\ell'}\rangle = \widehat{\gamma}_\ell - \widehat{\gamma}_{\ell'}$ 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,

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

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

l* !!! 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 $X^{(j)}_i$ der übrigen erklärenden Variablen werden k^* Mal wiederholt. Mit diesen $n \cdot k^*$ „Beobachtungen“ führt man nun eine Poisson-Regression durch.

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

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

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

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

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 abhänge. 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.

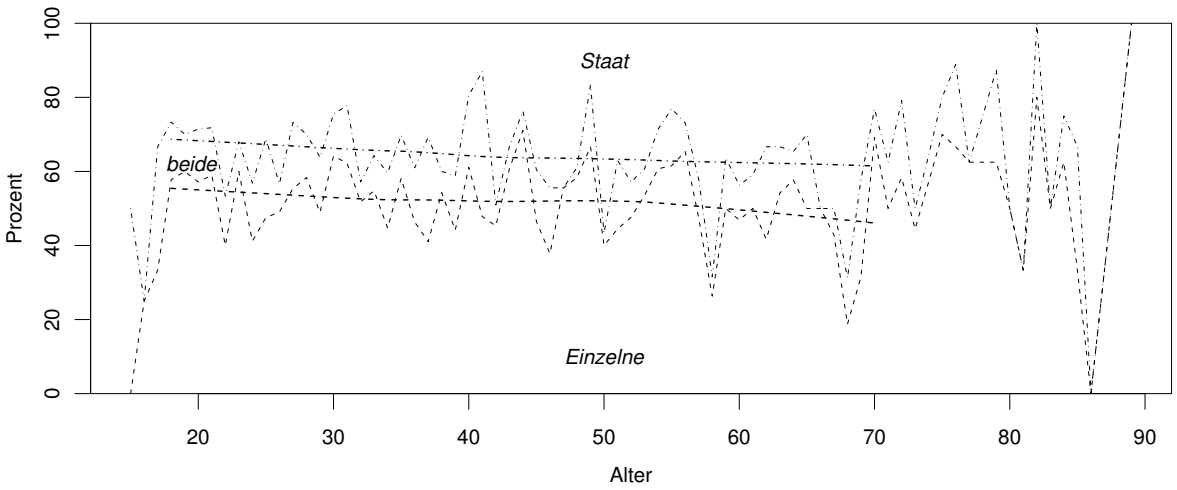


Abbildung 11.5.d: Zuweisung der Hauptverantwortung gegen Alter im Beispiel der Umwelt-Umfrage. Der untere Linienzug zeigt $N_{h1}/Nh+$ in Prozenten, der obere $1 - N_{h3}/Nh+$

Diese Tendenz lässt sich mit einem log-linearen Modell testen. Man bildet zunächst eine Kontingenztafel mit der als Faktor behandelten quantiativen 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 quantiative 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).

12 Ergänzungen zu kategoriellen Regressionsmodellen

Notizen und Stichworte.

12.1 Korrespondenz-Analyse

a Die Korrespondenz-Analyse ist ein grafisches Mittel zur Veranschaulichung von Datenmatrizen. Sie ist mit dem Biplot und damit mit der Hauptkomonenten-Analyse eng verwandt. Sie wurde ursprünglich zur Darstellung von Kreuztabellen entwickelt, ist seither aber erweitert worden für andere Arten von Datenmatrizen.

Hier soll nur die Interpretation des Resultats diskutiert werden. Auf den Biplot werden wir im Block Mu-2a zurückkommen.

b Tabelle 12.1.b zeigt für eine Umfrage bei 193 Angestellten einer grossen Firma den Zusammenhang zwischen Angestellten-Kategorien und Rauch- resp. Trinkgewohnheiten.

	smoking				drinking		
Anzahlen	no.sm	light.sm	medium.sm	heavy.sm	no.alc	yes.alc	total
sen.man	4	2	3	2	0	11	11
jun.man	4	3	7	4	1	17	18
sen.emp	25	10	12	4	5	46	51
jun.emp	18	24	33	13	10	78	88
secr	10	6	7	2	7	18	25
total	61	45	62	25	23	170	193

Zeilen-%	no.sm	light.sm	medium.sm	heavy.sm	no.alc	yes.alc	total
sen.man	36	18	27	18	0	100	100
jun.man	22	17	39	22	6	94	100
sen.emp	49	20	24	8	10	90	100
jun.emp	20	27	38	15	11	89	100
secr	40	24	28	8	28	72	100
total	32	23	32	13	12	88	100

Tabelle 12.1.b: Rauch- und Trink-Gewohnheiten von Angestellten in einer grossen Firma: Anzahlen und Zeilenprozentwerte

c Wir führen „Distanzen“ zwischen standardisierten Zeilen $\widetilde{\underline{x}}_h = [x_{hk}/x_{h+}]$ ($x_{h+} = \sum_k x_{hk}$) ein:

$$d\langle h, h'\rangle = \sum_k (\widetilde{x}_{hk} - \widetilde{x}_{h'k})^2 / x_{+k} \; .$$

Sie sollen möglichst gut dargestellt werden in zwei Dimensionen. Wie bei der Hauptkomponenten-Analyse müssen diese als Linearkombinationen der Zeilen von X zustandekommen. – Analog für Kolonnen. Beide Darstellungen werden kombiniert. So bedeuten grafische Beziehungen zwischen den Punkten der Darstellung näherungsweise folgendes:

Abbildung 12.1.b: Korrespondenz-Analyse-Diagramm für das Beispiel der Rauch- und Trink-Gewohnheiten

- Zwei Zeilen- (Spalten-) Punkte, die nahe beieinander liegen, bedeuten ähnliche Proportionen der Zeilen (Spalten).
- Zeilen- (Spalten-) Punkte, die nahe beim Nullpunkt liegen, bedeuten „durchschnittliche“ Proportionen.
- Liegen ein Zeilenpunkt und ein Spaltenpunkt in ähnlicher Richtung vom Nullpunkt weg, so ist die entsprechende Kombination übermässig häufig in den Daten.

Die gesamte Darstellung zeigt also die Abweichungen der Kreuztabelle von der Unabhängigkeit.

d Im Beispiel sind eigentlich zwei Kreuztabellen dargestellt. Die Rauch-Gewohnheiten wurden zur Berechnung der Darstellungs-Achsen verwendet. Die Trink-Gewohnheiten sind Spalten, die auf die gleiche Art wie die anderen Spalten dargestellt werden. Ebenso werden als zusätzliche Zeile die durchschnittlichen Rauch-Gewohnheiten über die ganzen USA dargestellt.

12.2 Kombination unabhängiger Tests, Meta-Analyse

a Beispiel Herzinfarkte und Verhütungsmittel: Studie an mehreren Spitälern = Multicenter-Studie. Ergibt mehrere Vierfeldertafeln mit verschiedenen Randsummen.
Kombinierte Evidenz?

„Beispiel“:

Teststatistik	1.50	2.30	5.10	0.90	3.20
P.Wert	0.22	0.13	0.02	0.34	0.07

Ist damit der Zusammenhang bewiesen?

b Möglichkeiten für einen Gesamttest:

- Teststatistiken zusammenzählen: $T = \sum T_\ell \sim \chi^2_m$.
- „Mantel-Haenszel-Statistik“: $U = \sum_\ell N_{11,\ell}$. Erwartungswert und Varianz bestimmen. $(U - \mathcal{E}\langle U\rangle)/\sqrt{\mathrm{var}\langle U\rangle}$ mit $\mathcal{N}\langle 0,1\rangle$ oder das Quadrat mit χ^2_1 vergleichen.
- P-Werte mitteln (arithmetisch, geometrisch, ...). Verteilung bestimmen ist nicht so schwierig.

Die dritte Möglichkeit und Varianten davon können auch eingesetzt werden, wenn die P-Werte von ganz verschiedenartigen Tests stammen.

12.3 Discrete Choice Models

a Antwortgrösse: Wahl einer Kategorie
aus einer Auswahl von Möglichkeiten.
Beispiel: Verkehrsmittel.

Erklärende Variable I: charakterisieren die Möglichkeiten.
Beispiel: Fahrzeiten, Umsteigen, ...

Erklärende Variable II: Gewöhnliche erklärende Variable,
unabhängig von der Kategorie der Antwortgrösse.
Beispiel: Alter, Geschlecht, Einkommen, ...

führt zu Multinomialer Regression (multinomial logit models).

b Literatur: Agresti, Kap. 9, Fahrmeir and Tutz (2001), Kap. 3.2.