

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).

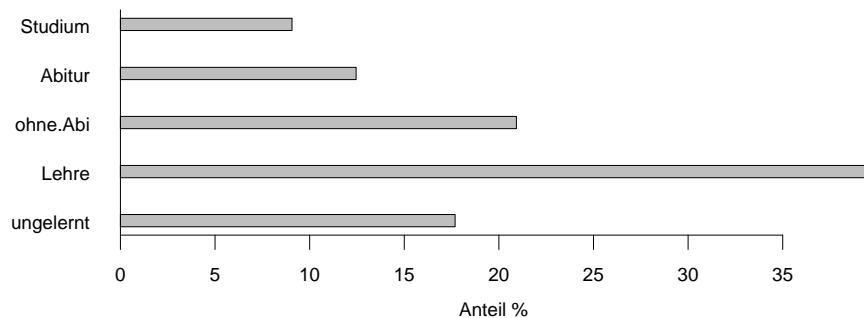


Abbildung 7.1.e: Stabdiagramm der Schulbildung im Beispiel der Umweltumfrage

- 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 (<i>B</i>)				Summe
		nicht	etwas	zieml.	sehr	
Schulbildung (<i>A</i>)	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 **Zähl**daten 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: Zähl

daten, 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 Zähl

daten mit **Anzahldaten** bezeichnen wollen. Ein solcher count kann irgendwelche Objekte oder Ereignisse 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 14.S.0.b 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					Σ		
		1	2	3	k	s			
Variable A	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+}
	\vdots	\vdots	\vdots			\vdots		\vdots	\vdots
	h	n_{h1}	n_{h2}	\dots		n_{hk}	\dots	n_{hs}	n_{h+}
	\vdots	\vdots	\vdots			\vdots		\vdots	\vdots
	r	n_{r1}	n_{r2}	\dots		n_{rk}	\dots	n_{rs}	n_{r+}
Σ	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,

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

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. ◁

- 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, $\hat{\pi}_{k|h} = N_{hk}/N_{h+}$.

		Beeinträchtigung (B)				Summe
		nicht	etwas	zieml.	sehr	
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

▷ 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). ◀

		Beeinträchtigung (B)				Summe
		nicht	etwas	zieml.	sehr	
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

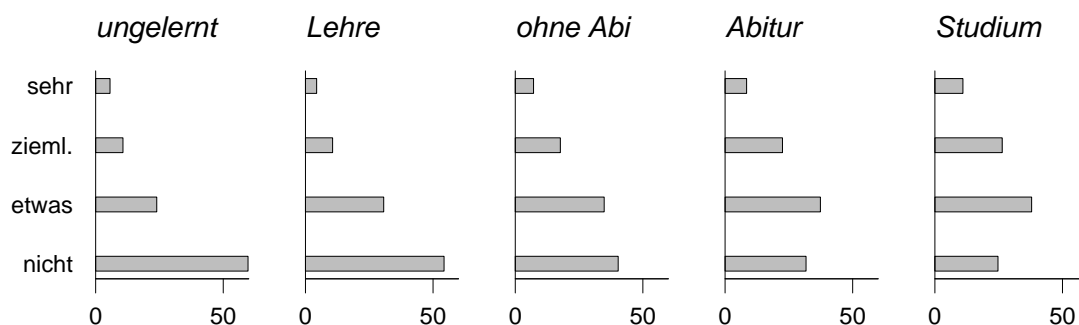


Abbildung 7.2.d: Histogramme zum Vergleich der Beeinträchtigung für die Schulbildungsklassen im Beispiel der Umweltumfrage

- 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

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

Wir schreiben

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

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,

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

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*),

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

Die Wahrscheinlichkeiten werden

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

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

$$\begin{aligned}
 P\langle N_{11} = n_{11}, N_{12} = n_{12}, \dots, N_{rs} = n_{rs} \mid N = n \rangle &= \prod_{h,k} \frac{\lambda_{hk}^{n_{hk}}}{n_{hk}!} e^{-\lambda_{hk}} \bigg/ \frac{\lambda^n}{n!} e^{-\lambda} \\
 &= \frac{n!}{\prod_{h,k} n_{hk}!} \cdot \frac{\prod_{h,k} \lambda_{hk}^{n_{hk}}}{\lambda^{\sum_{h,k} n_{hk}}} \cdot \frac{e^{-\sum_{h,k} \lambda_{hk}}}{e^{-\lambda}} = \frac{n!}{\prod_{h,k} n_{hk}!} \cdot \prod_{h,k} \pi_{hk}^{n_{hk}}.
 \end{aligned}$$

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 π s und bilden die Differenzen

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

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

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

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) $N_{hk} \sim \mathcal{P}\langle \lambda_{hk}^{(0)} \rangle$ und deshalb $\text{var}\langle N_{hk} \rangle \approx \lambda_{hk}$ ist. Es ist also

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

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

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

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 $\lambda_{h+}^{(0)}$ und $\lambda_{+k}^{(0)}$ (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 $\lambda_{h+}^{(0)}$. Mit diesem Modell kann man also die bedingte Verteilung der Teststatistik, gegeben die $\hat{\lambda}_{h+}^{(0)}$ oder die n_{h+} , untersuchen. Da auch die $\lambda_{+k}^{(0)}$ 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 7.3.1.

- 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: $T \sim \chi_{(r-1)(s-1)}^2$

Damit die genäherte Verteilung brauchbar ist, dürfen die *geschätzten erwarteten Anzahlen* $\hat{\lambda}_{hk}^{(0)} = N_{h+}N_{+k}/n$ 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 $\hat{\lambda}_{hk}^{(0)}$ müssen ≥ 4 sein, die übrigen ≥ 1 . Bei vielen Klassen (*rs* gross) können einzelne $\hat{\lambda}_{hk}^{(0)}$ 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. ◀

$\hat{\lambda}_{hk}^{(0)}$ h	k				$R_{hk}^{(P)}$ h	k			
	1	2	3	4		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 $\hat{\lambda}_{hk}^{(0)}$ 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 $\sqrt{\hat{\lambda}_{hk}^{(0)}}$, so dass die Flächen proportional zu den (Absolutwerten der) Differenzen der N_{hk} von ihren geschätzten Erwartungswerten $\hat{\lambda}_{hk}^{(0)}$ 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.

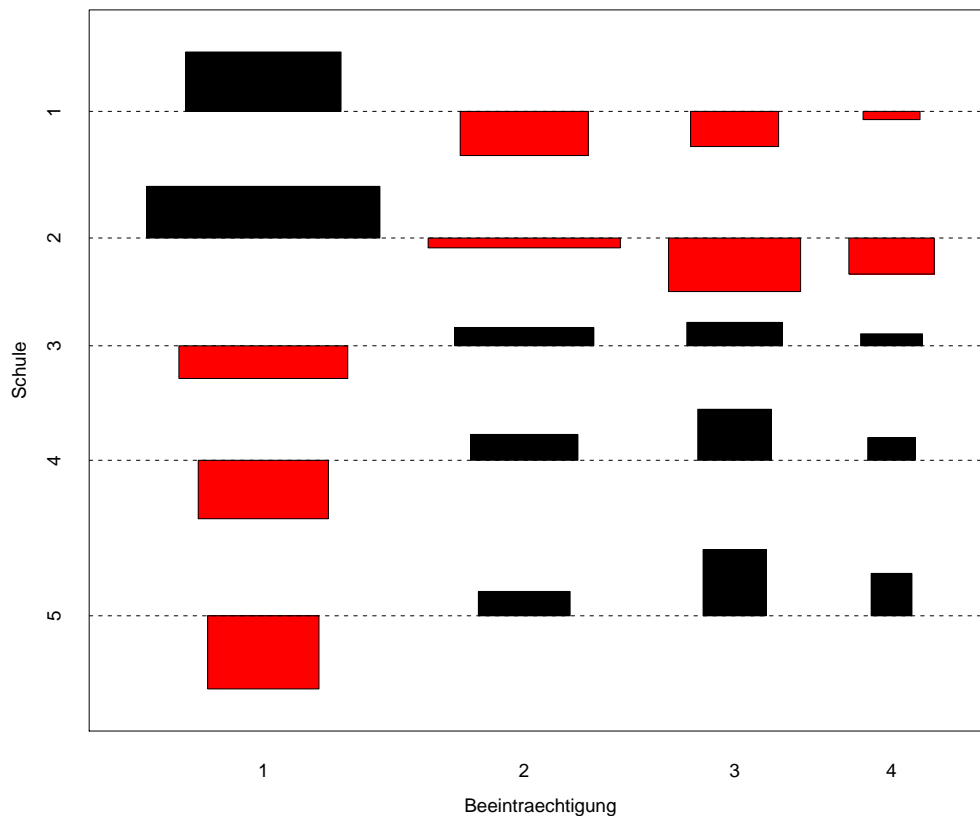


Abbildung 7.3.g: Association Plot für das Beispiel der Umweltumfrage

- 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. 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.

Zur Beantwortung der Frage vergleichen wir in den beiden Gruppen die Anteile derer, die Pillen benützt hatten. Ist $N_{11}/n_{+1} = 23/58 = 40\%$ signifikant von $N_{12}/n_{+2} = 34/166 = 20\%$ verschieden? ◀

		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.

- 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 „Übergangswahrscheinlichkeiten“ 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: $N_{12} \sim \mathcal{B}(N_{12} + N_{21}, 1/2)$. 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 kategoriale 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,

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

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*) $\theta = \log\langle\theta\rangle$.

Wir betrachten zunächst den Logarithmus der Wettverhältnisse, die „**log odds**“ $\log\langle P\langle B=1 | A=h\rangle / (1 - P\langle B=1 | A=h\rangle)\rangle$ für die beiden Gruppen $A=h=1$ und $h=0$.

Das logarithmierte Doppelverhältnis ist gleich der Differenz der log odds für die beiden Gruppen,

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

Diese Grösse hat folgende Eigenschaften:

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

Im Unterschied zu einer „gewöhnlichen“ (Pearson-) Korrelation ist θ 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

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

– 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

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

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

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

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

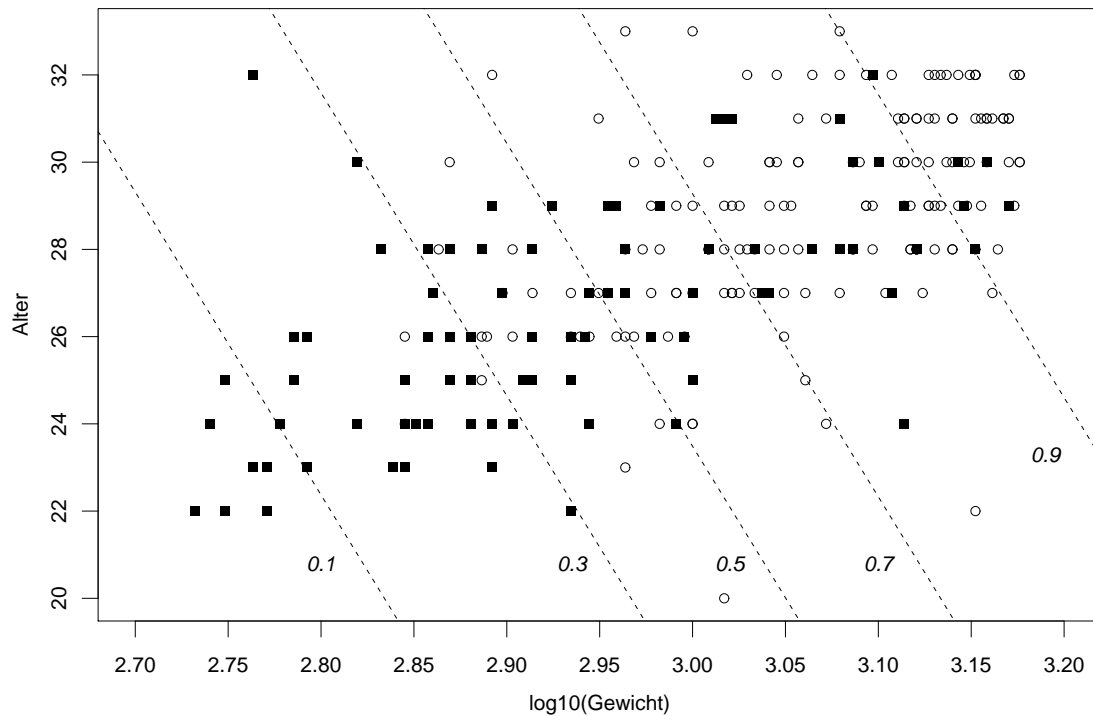


Abbildung 8.1.b: Logarithmiertes Gewicht und Alter im Beispiel der Frühgeburten. Die Überlebenden sind mit \circ , die anderen mit \square markiert. Die Geraden zeigen die Linien gleicher Überlebenswahrscheinlichkeiten (0.1, 0.3, 0.5, 0.7, 0.9) gemäss dem geschätzten logistischen Modell.

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

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

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,

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

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

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

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

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

Mit den Vektoren $\underline{x}_i = [1, x_i^{(1)}, x_i^{(2)}, \dots, x_i^{(m)}]^T$ und $\underline{\beta} = [1, \beta_1, \beta_2, \dots, \beta_m]^T$ 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 eignen 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 \triangleright Im **Beispiel der Frühgeburten** (8.1.b) wird die Wahrscheinlichkeit des Überlebens mit den weiter unten besprochenen Methoden geschätzt als

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

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 $\eta = -33.94 + 10.17 \cdot \log_{10}\langle \text{Gewicht} \rangle + 0.146 \cdot \text{Alter}$. \triangleleft

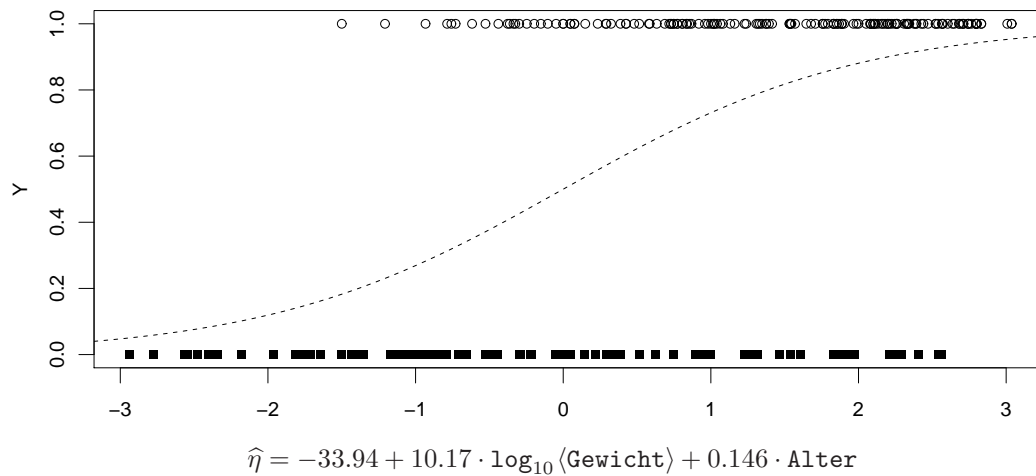


Abbildung 8.1.h: Die geschätzte Wahrscheinlichkeit $P\langle Y_i = 1 \rangle$ als Funktion des linearen Prädiktors, zusammen mit den Beobachtungen, im Beispiel der Frühgeburten

- 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 $\hat{\eta}_i > c$ mit $c=0$ (oder allenfalls einer anderen geeigneten Grenze c) gilt, werden der einen, die übrigen der andern Gruppe zugeordnet.
- 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 14.S.0.b 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.

1 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

$$g^{-1}(\eta) = \frac{\exp(\eta)}{1 + \exp(\eta)},$$

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

$$\begin{aligned} \text{odds}\langle Y = 1 \mid \underline{x} \rangle &= \frac{P\langle Y = 1 \rangle}{P\langle Y = 0 \rangle} = \exp\left\langle \beta_0 + \sum_j \beta_j x^{(j)} \right\rangle = e^{\beta_0} \cdot e^{\beta_1 x^{(1)}} \cdot \dots \cdot e^{\beta_m x^{(m)}} \\ &= e^{\beta_0} \cdot \exp\langle \beta_1 \rangle^{x^{(1)}} \cdot \dots \cdot \exp\langle \beta_m \rangle^{x^{(m)}} . \end{aligned}$$

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

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

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

- c \triangleright Im **Beispiel** (8.1.b) lassen sich die Schätzungen (aus 8.3.h) folgendermassen interpretieren: Für ein Individuum mit $\log_{10}\langle \text{Gewicht} \rangle = 3.1$, $\text{Alter} = 28$ 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 $\log_{10}\langle \text{Gewicht} \rangle = 2.9$, 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 $0.70/1.70 = 0.41$. \triangleleft
- d \triangleright Im **Beispiel der Umweltumfrage** (7.1.c) sollte die Abhängigkeit der Zielgrösse „Beinträchtigung“ 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. \triangleleft
- 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 (14.S.0.b) ausführlicher behandelt.

- f **Gruppierte Beobachtungen.** Wenn mehrere (m_ℓ) Beobachtungen Y_i zu gleichen Bedingungen $\underline{x}_i = \tilde{\underline{x}}_\ell$ 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

$$\tilde{Y}_\ell = \frac{1}{m_\ell} \sum_{i: \underline{x}_i = \tilde{\underline{x}}_\ell} Y_i .$$

Das ist in der Kontingenztafel bereits geschehen: Alle Personen mit gleicher Schulbildung $\tilde{\underline{x}}_\ell$ wurden zusammengefasst, und die Zahlen in den Spalten liefern die Angaben für \tilde{Y}_ℓ : 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 \tilde{Y}_ℓ der Anteil der beeinträchtigten Personen i unter den m_ℓ Befragten, die gleiche Schulbildung $x_i^{(1)} = \tilde{x}_\ell^{(1)}$, gleiches Geschlecht $x_i^{(2)} = \tilde{x}_\ell^{(2)}$ und Alter $x_i^{(3)} = \tilde{x}_\ell^{(3)}$ haben – allgemein, der Anteil der „Erfolge“ unter den m_ℓ „Versuchen“, die unter den Bedingungen $\tilde{\underline{x}}_\ell$ durchgeführt wurden.

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

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

Es gilt

$$\mathcal{E}\langle \tilde{Y}_\ell \rangle = \frac{1}{m_\ell} \sum_{i: \underline{x}_i = \tilde{\underline{x}}_\ell} \mathcal{E}\langle Y_i \rangle = \tilde{\pi}_\ell .$$

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 \triangleright **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. \triangleleft
- i **Transformierte Beobachtungen.** Laut dem Modell sind die logit-transformierten Erwartungswerte $\tilde{\pi}_\ell$ der „Erfolgsraten“ \tilde{Y}_ℓ/m_ℓ gleich einer linearen Funktion der $\tilde{x}_\ell^{(j)}$. Im Fall einer einzigen Eingangsgrösse liegt es nahe, die beobachteten Werte \tilde{Y}_ℓ/m_ℓ 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.

i	Y_i	weight	Age	ℓ	\tilde{x}_ℓ	m_ℓ	$m_\ell \tilde{Y}_\ell$	$m_\ell(1 - \tilde{Y}_\ell)$
1	1	1350	32	1	550	10	0	10
2	0	725	27	2	650	14	2	12
3	0	1090	27	3	750	27	9	18
4	0	1300	24	4	850	22	8	14
5	0	1200	31	5	950	32	23	9
...		...		6	1050	28	21	7
245	0	900	27	7	1150	22	19	3
246	1	1150	27	8	1250	26	19	7
247	0	790	27	9	1350	34	31	3
				10	1450	32	29	3

(i)

(ii)

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

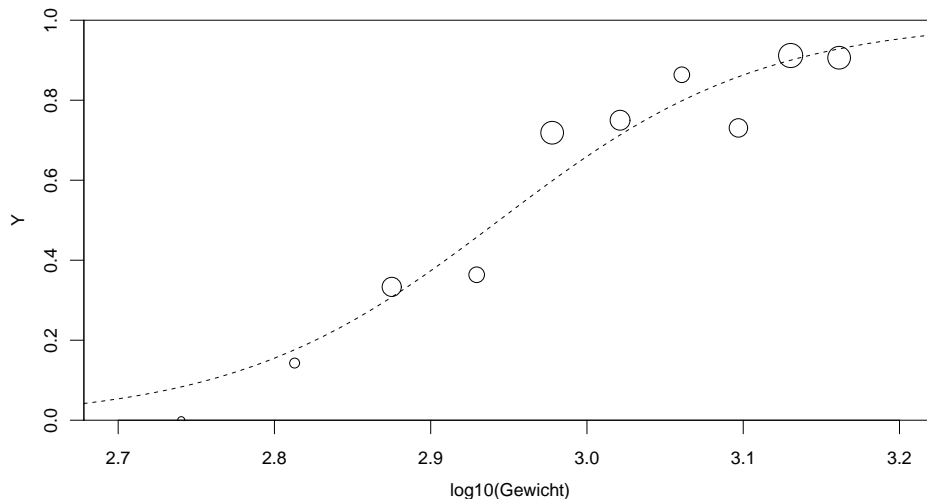


Abbildung 8.2.h: Überleben in Abhängigkeit vom Gewicht. Gruppierte Daten; die Fläche der Kreise ist proportional zur Anzahl Beobachtungen

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

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

Abbildung 8.2.i zeigt die empirischen Logits für die Frühgeburtendaten 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.

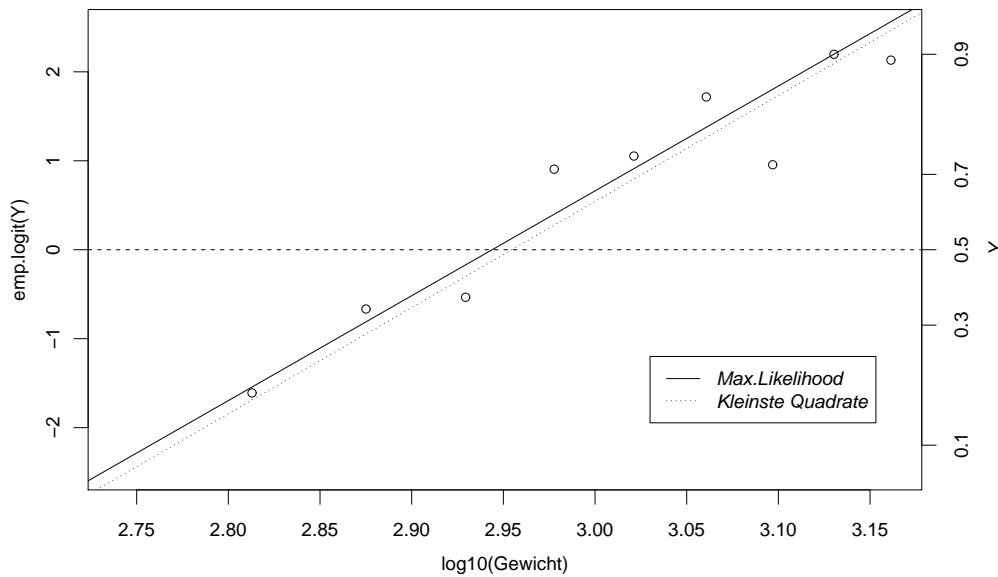


Abbildung 8.2.i: Modell in der logistischen Skala. In vertikaler Richtung sind die empirischen Logits der gruppierten Daten abgetragen. Die Geraden zeigen die geschätzten Werte $\hat{\eta}_\ell$ des linearen Prädiktors. Die Markierungen auf der rechten Seite geben die untransformierten Werte der Wahrscheinlichkeit des Überlebens an.

- 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,

$$Z_i = \tilde{\beta}_0 + \sum_j x_i^{(j)} \tilde{\beta}_j + E_i = \tilde{\eta}_i + E_i .$$

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 Pflanzen 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

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

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 $\beta_0 = \tilde{\beta}_0 - c$ und $\beta_j = \tilde{\beta}_j$, $j = 1, \dots, m$, 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

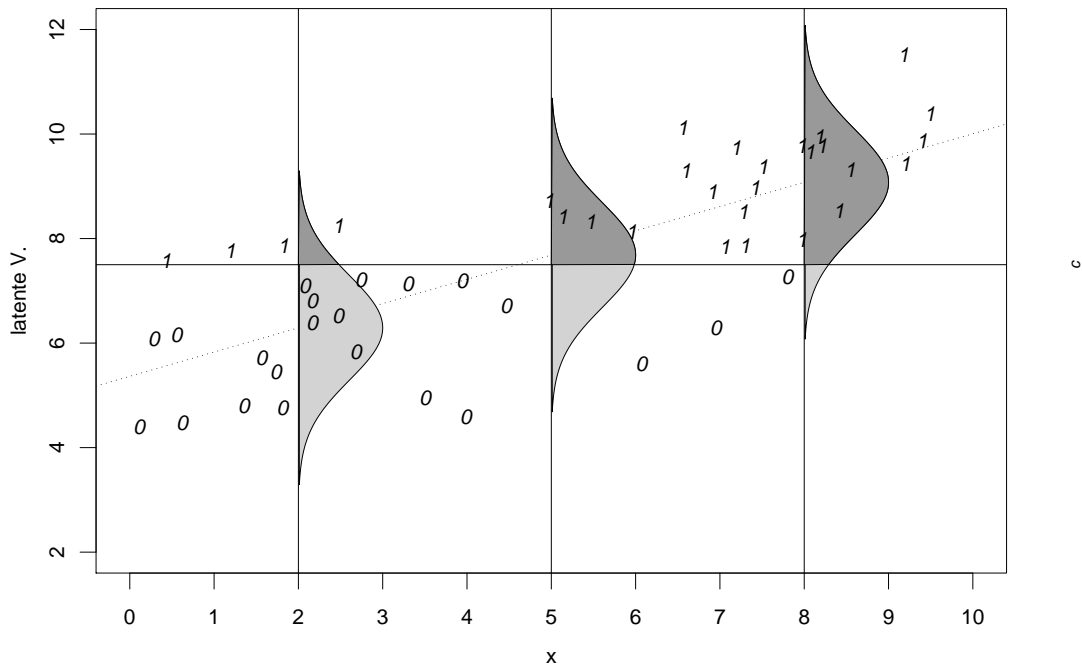


Abbildung 8.2.j: Zum Modell der latenten Variablen

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

8.3 Schätzungen und Tests

- 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.
- 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 $P\langle Y_1 = y_1, Y_2 = y_2, \dots, Y_n = y_n \rangle$ 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

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

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

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

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 $m_\ell \tilde{Y}_\ell$ binomial verteilt; die Wahrscheinlichkeiten sind deshalb

$$P \langle \tilde{Y}_\ell = \tilde{y}_\ell \rangle = \binom{m_\ell}{\tilde{y}_\ell} \pi_\ell^{m_\ell \tilde{y}_\ell} \cdot (1 - \pi_\ell)^{m_\ell (1 - \tilde{y}_\ell)} .$$

Daraus erhält man

$$\ell\ell \langle \underline{\pi} \rangle = \sum_\ell \left(c_\ell + m_\ell \tilde{y}_\ell \log \langle \tilde{\pi}_\ell \rangle + m_\ell (1 - \tilde{y}_\ell) \log \langle 1 - \tilde{\pi}_\ell \rangle \right)$$

mit $c_\ell = \log \langle \binom{m_\ell}{m_\ell \tilde{y}_\ell} \rangle$.

d* Um den Ausdruck $\pi_i = g^{-1} \langle \tilde{\underline{x}}_i^T \underline{\beta} \rangle$ 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

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

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

$$\sum_i (y_i - \hat{\pi}_i) \underline{x}_i = \underline{0} .$$

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

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

$$\sum_\ell m_\ell (\tilde{y}_\ell - \hat{\pi}_\ell) \tilde{\underline{x}}_\ell = \underline{0} .$$

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 $\underline{x}_i = \tilde{\underline{x}}_\ell$ 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 13.b.

f **Verteilung der geschätzten Koeffizienten.** In der multiplen linearen Regression konnte mit linearer Algebra recht einfach hergeleitet werden, dass der Vektor $\hat{\underline{\beta}}$ der geschätzten Koeffizienten multivariat normalverteilt ist mit Erwartungswert $\underline{\beta}$ und Kovarianzmatrix $\sigma^2(\mathbf{X}^T \mathbf{X})^{-1}$. Da die geschätzten Koeffizienten in der logistischen Regression die Lösung des näherungsweise äquivalenten gewichteten linearen Regressionsproblems sind, kann man daraus die Verteilung von $\hat{\underline{\beta}}$ ableiten. Die geschätzten Koeffizienten sind also näherungsweise multivariat normalverteilt, haben genähert den Erwartungswert $\underline{\beta}$ und eine Kovarianzmatrix $\mathbf{V}^{(\beta)}$, die wir bei den Verallgemeinerten Linearen Modellen (13.3.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 $\hat{\beta}_j$ ist die Wurzel aus dem j ten Diagonalelement V_{jj} der angegebenen Kovarianzmatrix, und

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

hat eine genäherte Normalverteilung. Im Ausdruck für die Kovarianzmatrix müssen die geschätzten Koeffizienten eingesetzt werden, deshalb $\hat{\mathbf{V}}^{(\beta)}$ statt $\mathbf{V}^{(\beta)}$. Da sie keine geschätzte Fehlervarianz $\hat{\sigma}^2$ 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 später zurück (13.2.f, 13.3.g, 13.4). Die „Null Deviance“ und die „Residual Deviance“ brauchen noch eine genauere Erklärung.

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 \tilde{Y}_ℓ definiert als 2 mal die Differenz zwischen der maximalen Log-Likelihood $\ell^{(M)}$ und dem Wert für das angepasste Modell,

$$D(\tilde{\mathbf{y}}; \hat{\underline{\pi}}) := 2 \left(\ell^{(M)} - \ell(\hat{\underline{\beta}}) \right).$$

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

```
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

* Diese erhält man, indem man in der Formel für $\ell(\underline{\pi})$ (8.3.c) $\tilde{\pi}_\ell$ durch \tilde{y}_ℓ ersetzt. (Für $\tilde{y}_\ell = 0$ und $\tilde{y}_\ell = 1$ 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 $\ell^{(M)}$ und das erwähnte $\ell(\underline{\pi})$ in die Definition der Devianz ein, so erhält man

$$D\langle \tilde{\underline{y}}; \hat{\underline{\pi}} \rangle = 2 \sum_{\ell} \left(m_{\ell} \tilde{y}_{\ell} \log \left\langle \frac{\tilde{y}_{\ell}}{\tilde{\pi}_{\ell}} \right\rangle + m_{\ell} (1 - \tilde{y}_{\ell}) \log \left\langle \frac{1 - \tilde{y}_{\ell}}{1 - \tilde{\pi}_{\ell}} \right\rangle \right).$$

Für ungruppierte, binäre Daten ergibt sich $\ell^{(M)} = 0$ und somit

$$D\langle \tilde{\underline{y}}; \hat{\underline{\pi}} \rangle = -2 \sum_i (y_i \log \langle \pi_i \rangle + (1 - y_i) \log \langle 1 - \pi_i \rangle).$$

- 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

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

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 13.3.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 Freiheitsgraden 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 könnten; 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 \tilde{Y}_ℓ der „Erfolge“ Y_i gezählt und die Summe der geschätzten Wahrscheinlichkeiten $\hat{\pi}_i$ gebildet. Auf diese Grössen wird dann der gewöhnliche Chi-Quadrat-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 Beeinträchtigung 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

$$AIC = D\langle y; \hat{\pi} \rangle + 2p$$

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 $\tilde{\pi}_\ell$ alle gleich seien. Der Schätzwert für diesen einzigen Parameter ist natürlich $\tilde{\pi} = \sum_\ell \tilde{y}_\ell / \sum_\ell m_\ell = \sum_\ell \tilde{y}_\ell / n$. Die Log-Likelihood für dieses Modell ist gleich

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

Die Devianz ergibt sich wieder als Differenz zwischen

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

und wird **Null-Devianz** genannt. Sie entspricht der „totalen Quadratsumme“ $\sum_i (Y_i - \bar{Y})^2$ 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 Chi-Quadrat-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 $2(\ell^{(G)} - \ell^{(K)}) = D \langle \tilde{y}; \tilde{\pi}^{(K)} \rangle - D \langle \tilde{y}; \tilde{\pi}^{(G)} \rangle$.
 Genäherte Verteilung unter H_0 : $\chi_{p_G - p_K}^2$.
- 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 $D \langle \tilde{y}; \hat{\pi} \rangle = 2(\ell^{(M)} - \ell \langle \hat{\pi} \rangle)$
 Genäherte Verteilung unter H_0 , falls die m_ℓ genügend gross sind: $\chi_{\tilde{n} - p}^2$ mit $\tilde{n} =$ Anzahl (Gruppen-) Beobachtungen \tilde{Y}_ℓ . Dieser Test geht nur für gruppierte Beobachtungen!
- Gesamttest für die Regression: Vergleich von **Null-Devianz** $D \langle \tilde{y}; \hat{\pi}^0 \rangle$ und Residuen-Devianz.
 H_0 : Null-Modell mit einem Parameter ist richtig.
 H_1 : Angepasstes Modell mit p Parametern ist richtig.
 Teststatistik $D \langle \tilde{y}; \hat{\pi}^0 \rangle - D \langle \tilde{y}; \hat{\pi} \rangle = 2(\ell \langle \hat{\pi} \rangle - \ell \langle \hat{\pi}^0 \rangle)$.
 Genäherte Verteilung unter H_0 : χ_{p-1}^2 .

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

$$R_\ell = \tilde{Y}_\ell - \hat{\pi}_\ell, \quad \hat{\pi}_\ell = g^{-1}(\hat{\eta}_\ell)$$

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

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

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

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

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) $\text{var}\langle R_i \rangle = \sigma^2 (1/w_i - (H_W)_{ii})$, wobei $(H_W)_{ii}$ das i te Diagonalelement der Matrix

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

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 $w_\ell = m_\ell / (\hat{\pi}_\ell(1 - \hat{\pi}_\ell))$ (vergleiche . 13.d). Die genauer standardisierten Residuen sind dann

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

- 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

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

- 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 $\hat{\eta}_i$ für den linearen Prädiktor oder die entsprechenden geschätzten Wahrscheinlichkeiten $\hat{\pi}_i$. 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 Nichtlinearitäten zulässt.

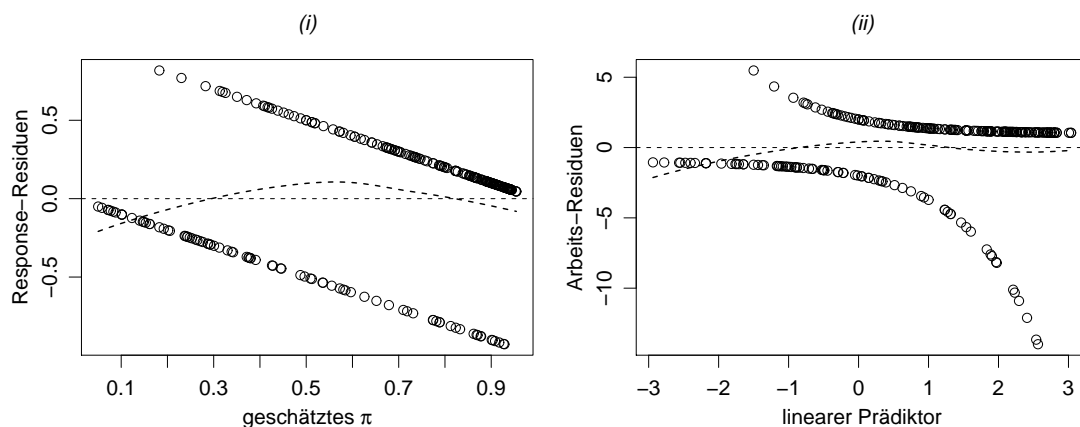


Abbildung 8.4.g: Tukey-Anscombe-Diagramme im Beispiel der Frühgeburten: (i) Response-Residuen und geschätzte π_i , (ii) Arbeits-Residuen und linearer Prädiktor

- 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 Glättungsmethode verwendet werden, die den verschiedenen Varianzen der Residuen mittels Gewichten Rechnung 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 vorhergesagte 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.

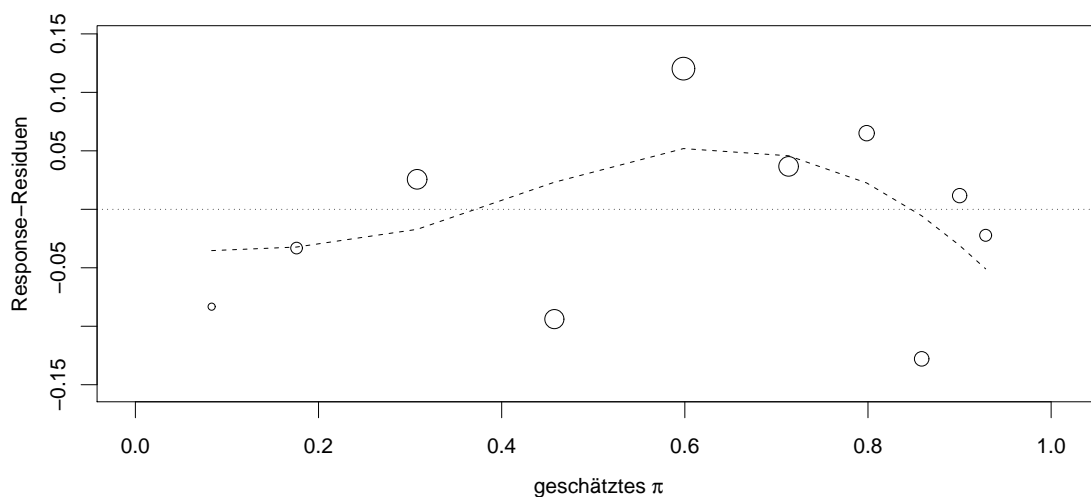


Abbildung 8.4.i: Tukey-Anscombe-Diagramm im Beispiel der Frühgeburten mit klassiertem Gewicht

- 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**“.

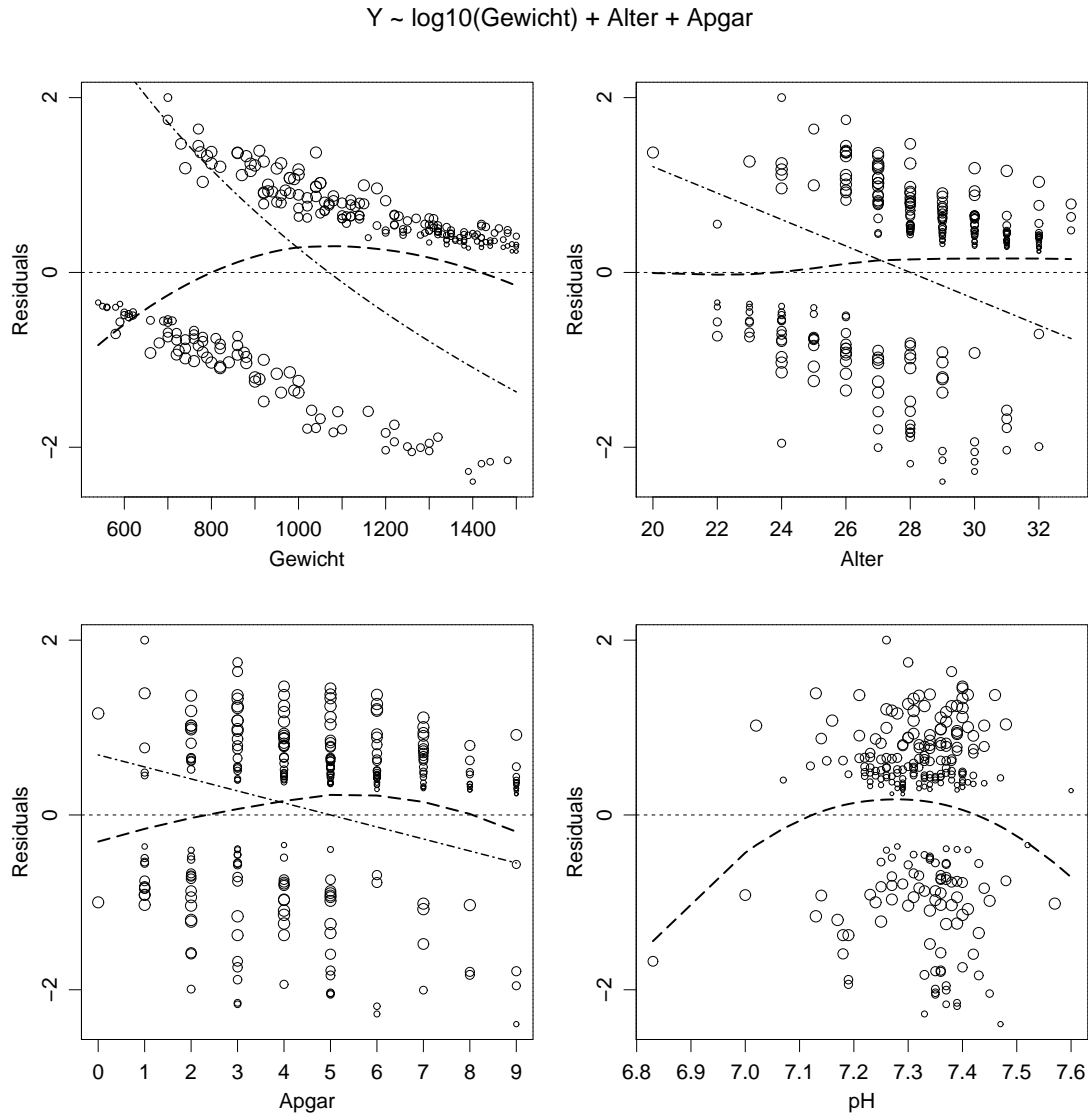


Abbildung 8.4.j: Residuen gegen Eingangsgrößen im Beispiel der Frühgeburten. Die Radien der Kreise entsprechen den Gewichten. Einige extrem negative Residuen wurden weggeschnitten.

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 (13.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.

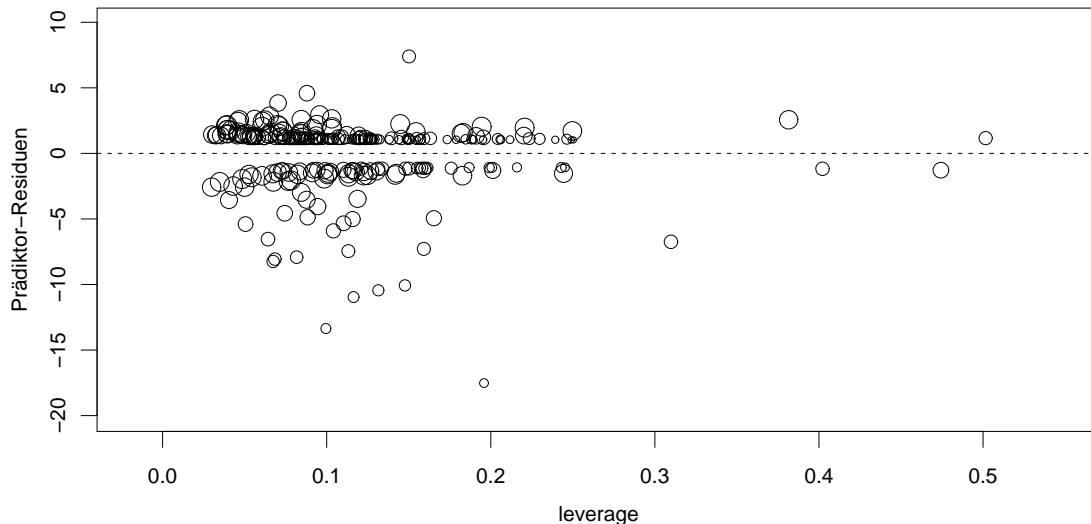


Abbildung 8.4.k: Residuen gegen Hebelarm-Werte $(H_W)_{ii}$ für das Beispiel der Frühgeburten

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 $Y \sim \log_{10}(\text{Gewicht}) + \text{Alter}$ 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.

Literaturverzeichnis

- Agresti, A. (2002). *Categorical Data Analysis*, 2nd edn, Wiley, N.Y.
- Agresti, A. (2007). *An Introduction to categorical data analysis*, Wiley Series in Probability & Math. Statistics, 2nd edn, Wiley, New York.
- Bates, D. M. and Watts, D. G. (1988). *Nonlinear Regression Analysis and its Applications*, Wiley, N.Y.
- Chambers, J. M. and Hastie, T. J. (1992). *Statistical Models in S*, Wadsworth & Brooks/Cole, Pacific Grove, Cal.
- Chatterjee, S. and Price, B. (2000). *Regression Analysis By Example*, 3rd edn, Wiley, N.Y.
- Christensen, R. (1990). *Log-linear models*, Springer, N.Y.
- Cleveland, W. S. (1994). *The Elements of Graphing Data*, 2nd edn, Hobart Press, Summit, New Jersey.
- Clogg, C. C. and Shihadeh, E. S. (1994). *Statistical models for ordinal variables*, Sage, Thousand Oaks, CA.
- Cohen, A. (1980). On the graphical display of the significant components in a two-way contingency table, *Communications in Statistics – Theory and Methods* **A9**: 1025–1041.
- Collet, D. (1991, 1999). *Modelling binary data*, Chapman & Hall/CRC Press LLC, Boca Raton, Florida.
- Collet, D. (1994). *Modelling Survival Data in Medical Research*, Texts in Statistical Science, Chapman and Hall, London.
- Cook, R. D. and Weisberg, S. (1999). *Applied regression including computing and graphics*, Wiley, N.Y.
- Cox, D. R. (1989). *Analysis of Binary Data*, 2nd edn, Chapman and Hall, London.
- Cox, D. R. and Snell, E. J. (1981). *Applied Statistics*, Chapman and Hall, London.
- Crowder, M. J., Kimber, A. C., Smith, R. L. and Sweeting, T. J. (1991). *Statistical Analysis of Reliability Data*, Chapman and Hall.
- Daniel, C. and Wood, F. S. (1980). *Fitting Equations to Data*, 2nd edn, Wiley, N.Y.
- Davies, P. (1995). Data features, *Statistica Neerlandica* **49**: 185–245.
- Devore, J. L. (2004). *Probability and Statistics for Engineering and the Sciences*, 6th edn, Duxbury Press, Belmont, California.
- Dobson, A. J. (2002). *An Introduction to Generalized Linear Models*, 2nd edn, Chapman and Hall, London.
- Draper, N. and Smith, H. (1998). *Applied Regression Analysis*, 3rd edn, Wiley, N.Y.
- Fahrmeir, L. and Tutz, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models*, 2nd edn, Springer-Verlag, New York.

- Fox, J. (2002). *An R and S-Plus companion to applied regression*, Sage, Thousand Oaks, CA.
- Fox, J. and Monette, G. (1992). Generalized collinearity diagnostics, *Journal of the American Statistical Association* **87**: 178–183.
- Fuller, W. A. (1987). *Measurement Error Models*, Wiley, N.Y.
- Haaland, P. D. (1989). *Experimental Design in Biotechnology*, Marcel Dekker, N.Y.
- Hampel, F. (1974). The influence curve and its role in robust estimation, *Journal of the American Statistical Association* **69**: 383–393.
- Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*, Wiley, N.Y.
- Harrell, F. E. J. (2002). *Regression Modeling Strategies. With Applications to Linear Models, Logistic Regression, and Survival Analysis*, Springer Series in Statistics, Springer, NY. Corrected second printing
- Hartung, J., Elpelt, B. und Klösener, K. (2002). *Statistik. Lehr- und Handbuch der angewandten Statistik*, 13. Aufl., Oldenbourg, München.
- Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*, number 43 in *Monographs on Statistics and Applied Probability*, Chapman and Hall, London.
- Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Series in Statistics, Springer-Verlag, New York.
- Hocking, R. R. (1996). *Methods and Applications of Linear Models; Regression and the Analysis of Variance*, Wiley Series in Probability and Statistics, Wiley, N.Y.
- Hosmer, D. W. and Lemeshow, S. (2000). *Applied Logistic Regression*, 2nd edn, Wiley, N.Y.
- Huber, P. J. (1964). Robust estimation of a location parameter, **35**: 73–101.
- Huber, P. J. and Ronchetti, E. M. (2009). *Robust Statistics*, 2nd edn, Wiley.
- Kalbfleisch, J. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*, 2nd edn, Wiley, N.Y.
- Lindsey, J. K. (1995). *Modelling Frequency and Count Data*, number 15 in *Oxford Statistical Science Series*, Clarendon Press, Oxford.
- Maronna, R. A., Martin, R. D. and Yohai, V. J. (2006). *Robust Statistics, Theory and Methods*, Wiley Series in Probability and Statistics, Wiley, Chichester, England.
- McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd edn, Chapman and Hall, London.
- Mosteller, F. and Tukey, J. W. (1977). *Data Analysis and Regression: A Second Course in Statistics*, Addison-Wesley, Reading, Massachusetts.
- Myers, R. H., Montgomery, D. C. and Vining, G. G. (2001). *Generalized Linear Models. With Applications in Engineering and the Sciences*, Wiley Series in Probability and Statistics, Wiley, NY.
- Pokropp, F. (1994). *Lineare Regression und Varianzanalyse*, Oldenbourg.
- Rice, J. A. (2007). *Mathematical Statistics and Data Analysis*, 3rd edn, Duxbury Press, Belmont, California.
- Ripley, B. D. (1996). *Pattern Recognition and Neural Networks*, Cambridge Univ. Press, Cambridge, UK.
- Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust Regression & Outlier Detection*, Wiley, N.Y.
- Ryan, T. P. (1997). *Modern Regression Methods*, Series in Probability and Statistics, Wiley, N.Y. includes disk

- Sachs, L. (2004). *Angewandte Statistik*, 11. Aufl., Springer, Berlin.
- Schlittgen, R. (2003). *Einführung in die Statistik. Analyse und Modellierung von Daten*, 10. Aufl., Oldenbourg, München. *schoen, inkl. Sensitivity und breakdown, einfache regr mit resanal*
- Sen, A. and Srivastava, M. (1990). *Regression Analysis; Theory, Methods, and Applications*, Springer-Verlag, N.Y.
- Stahel, W. A. (2000). *Statistische Datenanalyse: Eine Einführung für Naturwissenschaftler*, 3. Aufl., Vieweg, Wiesbaden.
- Stahel, W. A. (2007). *Statistische Datenanalyse: Eine Einführung für Naturwissenschaftler*, 5. Aufl., Vieweg, Wiesbaden.
- van der Waerden, B. L. (1971). *Mathematische Statistik*, 3. Aufl., Springer, Berlin.
- Venables, W. N. and Ripley, B. D. (1997). *Modern Applied Statistics with S-Plus*, Statistics and Computing, 2nd edn, Springer, Berlin.
- Weisberg, S. (2005). *Applied Linear Regression*, 3rd edn, Wiley, N.Y.
- Wetherill, G. (1986). *Regression Analysis with Applications*, number 27 in *Monographs on Statistics and Applied Probability*, Chapman and Hall, London.