Statistische Regressionsmodelle

Teil II: Verallgemeinerte Lineare Modelle

Werner Stahel
Seminar für Statistik, ETH Zürich

März 2005 / Mai 2008

Zweiter Teil der Unterlagen zu einem Kurs über Regressionsmodelle, gehalten vom 4.-6. Juni 2008, veranstaltet von der Schweizerischen Gesellschaft für Statistik.

12 Zweiwertige Zielgrössen, logistische Regression

12.1 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 **Ausgangsgrössen** 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 Ausgangsgrössen hängt das Überleben von Frühgeburten ab? Hibbard (1986) stellte Daten von 247 Säuglingen zusammen. In Abbildung 12.1.b sind die beiden wichtigsten Ausgangsgrössen, 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 Ausgangsgrössen abhängt. Wir suchen also eine Funktion h mit

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

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

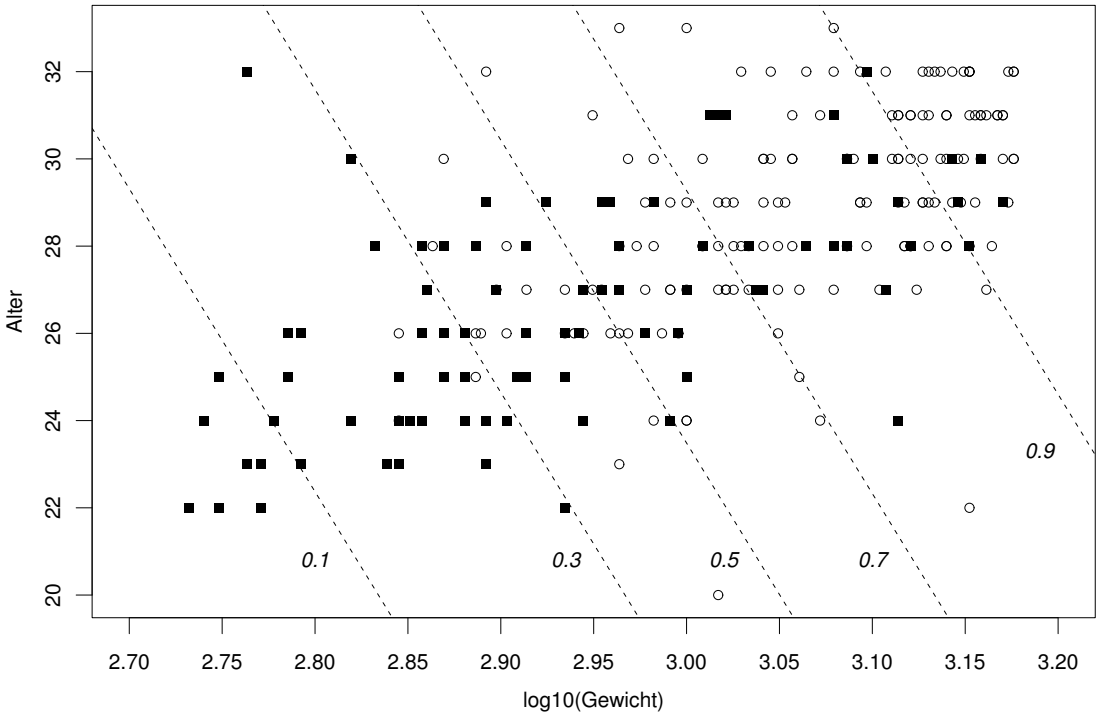


Abbildung 12.1.b: Logarithmiertes Gewicht und Alter im Beispiel der Frühgeburten. Die Überlebenden sind mit ∘, die anderen mit □ markiert. Die Geraden zeigen die Linien gleicher Überlebenswahrscheinlichkeiten (0.1, 0.3, 0.5, 0.7, 0.9) gemäss dem geschätzten logistischen Modell.

der Regression mit normalverteilten Fehlern

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

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,

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

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

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

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

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

Sie ordnet den Wahrscheinlichkeiten π das logarithmierte **Wettverhältnis** (die log odds) zu (11.4.e).

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

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

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

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

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

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

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

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

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

g Die Funktion g, die die Erwartungswerte $\mathcal{E}\langle Y_i \rangle$ in Werte des linearen Prädiktors verwandelt, nennt man die **Link-Funktion**. Die logistische Funktion ist zwar die üblichste, aber nicht die einzige geeignete Link-Funktion für binäre Zielgrössen. Im Prinzip eigenen sich alle strikt monotonen Funktionen, die den möglichen Werte zwischen 0 und 1 alle Zahlen zwischen $-\infty$ und $+\infty$ zuordnen – genauer, für die $g\langle 0\rangle = -\infty$ und $g\langle 1\rangle = \infty$ ist, vergleiche 12.2.j.

h Im **Beispiel der Frühgeburten** (12.1.b) wird die Wahrscheinlichkeit des Überlebens mit den weiter unten besprochenen Methoden geschätzt als

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

Die Linien gleicher geschätzter Wahrscheinlichkeit wurden in Abbildung 12.1.b bereits eingezeichnet. Abbildung 12.1.h zeigt die Beobachtungen und die geschätzte Wahrscheinlichkeit, aufgetragen gegen den linearen Prädiktor $\eta = -33.94 + 10.17 \cdot \texttt{log}_{10}(\texttt{Gewicht}) + 0.146 \cdot \texttt{Alter}$.

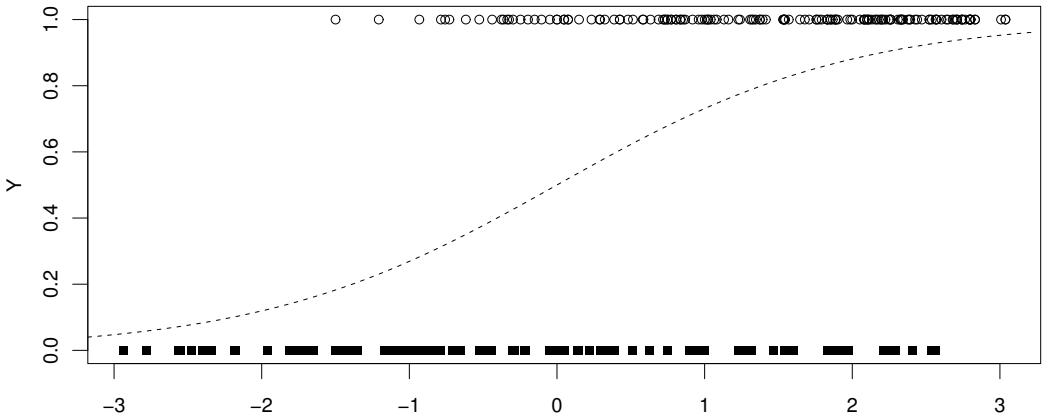


Abbildung 12.1.h: Die geschätzte Wahrscheinlichkeit $P\langle Y_i = 1\rangle$ als Funktion des linearen Prädiktors, zusammen mit den Beobachtungen, im Beispiel der Frühgeburten

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 Ausgangs-Variablen der Regression. Der lineare Prädiktor übernimmt die Rolle der Diskriminanzfunktion, die ja (in der Fisherschen Diskriminanzanalyse) ebenfalls linear in den $x_i^{(j)}$ war. Die Beobachtungen, für die $\widehat{\eta}_i > c$ mit $c = 0$ (oder allenfalls einer anderen geeigneten Grenze c) gilt, werden der einen, die übrigen der andern Gruppe zugeordnet.

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 Ausgangsgrössen 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 11.2.p ausführlicher behandeln.

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

l **Literatur.** Entsprechend dieser Einordnung gibt es umfassende und spezialisiertere Bücher:

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

12.2 Betrachtungen zum Modell

a Im Modell der logistischen Regression ist das logarithmierte Wettverhältnis gleich dem linearen Prädiktor η_i (12.1.f)

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

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

die so genannte **logistische Funktion**, die der logistischen Regression den Namen gegeben hat. Ihre Form ist durch die Linie in Abbildung 12.1.h gegeben.

b **Interpretation der Koeffizienten.** Koeffizienten!Interpretation 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

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

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

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

Solche Quotienten von Wettverhältnissen haben wir unter dem Namen **Doppelverhältnisse** oder **odds ratios** in 11.4.c eingeführt.

c Im **Beispiel** (12.1.b) lassen sich die Schätzungen (aus 12.3.h) folgendermassen interpretieren: Für ein Individum mit $\texttt{log}_{10}(\texttt{Gewicht}) = 3.1, \texttt{Alter} = 28$ erhält man als Schätzung für das logarithmierte Wettverhältnis $-33.94 + 10.17 \cdot 3.1 + 0.146 \cdot 28 = 1.68$ und damit ein Wettverhältnis für das Überleben von $\exp\langle 1.68 \rangle = 5.4$. Die geschätzte Wahrscheinlichkeit für das Überleben beträgt $g^{-1}\langle 5.4 \rangle = 0.84$. Vergleicht man nun dieses Wettverhältnis mit dem eines Individuums mit dem gleichen Alter und $\texttt{log}_{10}(\texttt{Gewicht}) = 2.9$, dann erhält man als odds ratio den Faktor $\exp\langle 10.17 \cdot (-0.2) \rangle = 0.13$, d.h. das Wettverhältnis im zweiten Fall ist auf 13% des vorherigen gesunken und wird $0.13 \cdot 5.4 = 0.70$, und die entsprechenden Wahrscheinlichkeit wird $0.70/1.70 = 0.41$.

d Im **Beispiel der Umweltumfrage** (11.1.c) sollte die Abhängigkeit der Zielgrösse „Beeinträchtigung“ von der Schulbildung erfasst werden. Die Zielgrösse hat hier vier mögliche geordnete Werte. Wir machen für die folgenden Betrachtungen daraus eine zweiwertige Variable, indem wir je zwei Kategorien zusammenfassen; später soll die feinere Unterteilung berücksichtigt werden.

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

e Die logistische Regression eignet sich also auch zur Analyse von **Kontingenztafeln**, sofern eine „Dimension“ der Tafel als Zielgrösse aufgefasst wird und nur 2 Stufen zeigt. Man kann von **logistischer Varianzanalyse** sprechen. Die Analyse von Kontingenztafeln wird im Kapitel über log-lineare Modelle (11.2.p) ausführlicher behandelt.

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

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

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

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

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

Es gilt

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

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

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

Für „Gruppen“ mit nur einer Beobachtung ($m_\ell = 1$) wird Y_ℓ wieder zweiwertig und die Binomialverteilung zur Bernoulli-Verteilung (12.1.c).

h **Beispiel Frühgeburten**. Um ein anschauliches Beispiel zu erhalten, untersuchen wir das Überleben von Frühgeburten nur als Funktion der Ausgangsgrösse Gewicht. Wenn wir Klassen von je 100 g Gewicht bilden, können wir die Daten zu den Häufigkeiten zusammenfassen, die in Tabelle 12.2.h gezeigt werden, zusammen mit einem Ausschnitt aus den ursprünglichen Beobachtungen. Abbildung 12.2.h zeigt sie mit dem angepassten Modell in dieser Form.

i **Transformierte Beobachtungen.** Laut dem Modell sind die logit-transformierten Erwartungswerte $\widetilde{\pi}_\ell$ der „Erfolgsraten“ $\widetilde{Y}_\ell / m_\ell$ gleich einer linearen Funktion der $\widetilde{x}_\ell^{(j)}$. Im Fall einer einzigen Ausgangsgrösse liegt es nahe, die beobachteten Werte $\widetilde{Y}_\ell / m_\ell$ selbst zu transformieren und gegen die Ausgangs-Variable aufzutragen; im Falle von mehreren Ausgangsgrössen kann man auf der horizontalen Achse stattdessen den linearen Prädiktor η verwenden. Es sollte sich dann statt des sigmoiden Zusammenhangs von Abbildung 12.2.h ein linearer ergeben.

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

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

(i)

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

(ii)

Tabelle 12.2.h: Beispiel Frühgeburten: Einige Einzel-Beobachtungen (i) und zusammengefasste Daten (ii). $m_\ell\widetilde{Y}_\ell$ ist die Anzahl Überlebende der insgesamt m_ℓ Kinder in der Gewichtsklasse ℓ mit mittlerem Gewicht $\widetilde{x}_\ell$

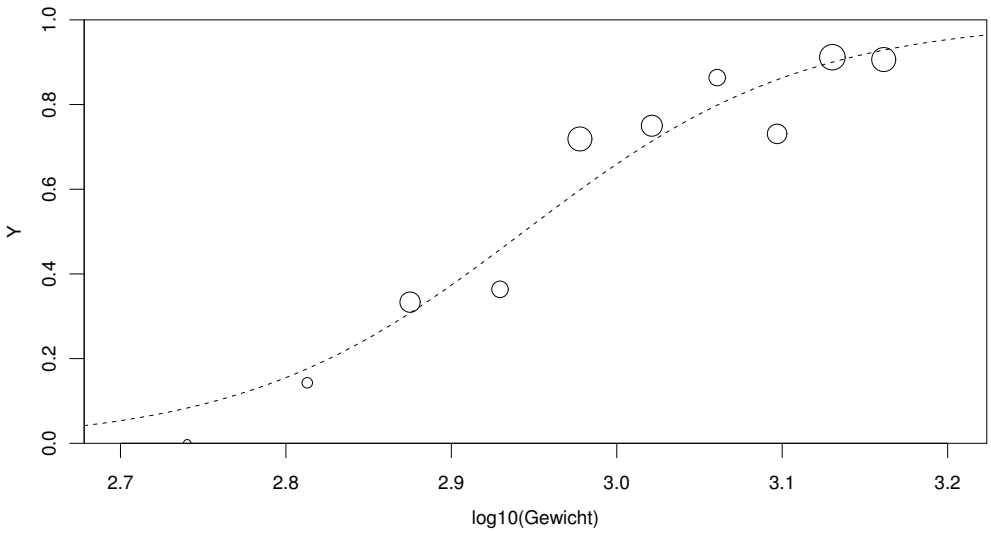


Abbildung 12.2.h: Überleben in Abhängigkeit vom Gewicht. Gruppierte Daten; die Fläche der Kreise ist proportional zur Anzahl Beobachtungen

Ausweg verwendet man die **empirischen Logits**

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

Abbildung 12.2.i zeigt die empirischen Logits für die Frühgeburtdaten und die angepasste lineare Funktion.

Wendet man auf die empirischen Logits eine gewöhnliche multiple Regression an, so erhält man eine alternative Schätzung der Koeffizienten. Sie bildet oft eine vernünftige Näherung für die optimalen Schätzwerte, die wir in 12.3.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.

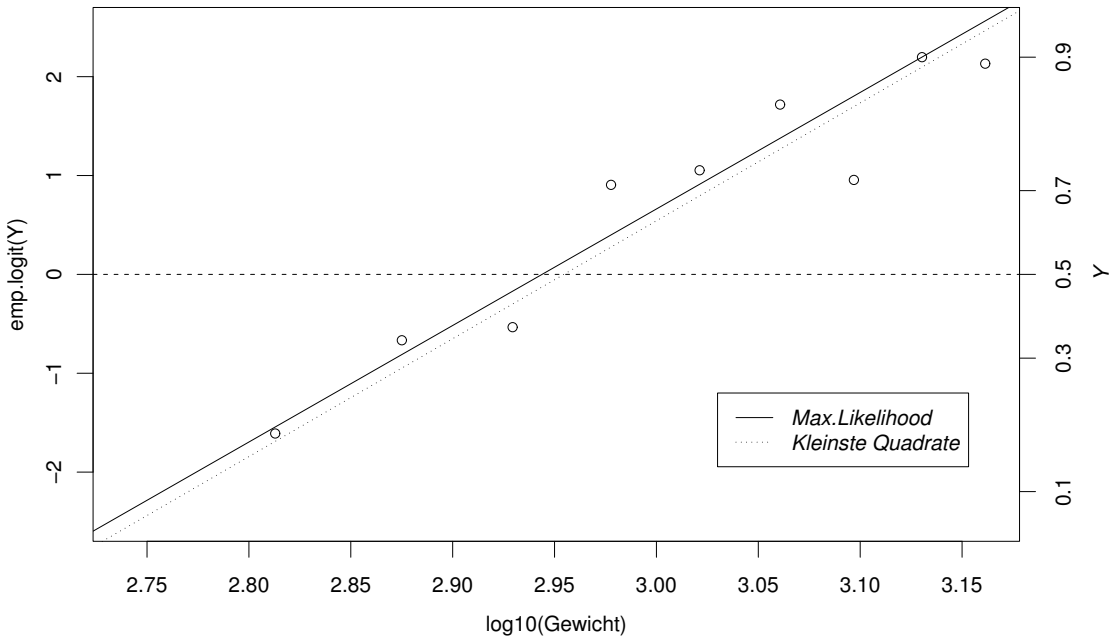


Abbildung 12.2.i: Modell in der logistischen Skala. In vertikaler Richtung sind die empirischen Logits der gruppierten Daten abgetragen. Die Geraden zeigen die geschätzten Werte $\widehat{\eta}_\ell$ des linearen Prädiktors. Die Markierungen auf der rechten Seite geben die untransformierten Werte der 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,

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

Die binäre Zielgrösse Y_i stellt fest, ob Z_i unterhalb oder oberhalb eines **Schwellenwertes** c liegt. Abbildung 12.2.j veranschaulicht diese Vorstellung.

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

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

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

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

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

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

Der Ausdruck $1 - F\langle -\eta \rangle$ ist selbst eine Verteilungsfunktion, nämlich diejenige von $-E$. Wenn wir diese Verteilungsfunktion gleich g^{-1} setzen, dann erhalten wir das Modell der lo-

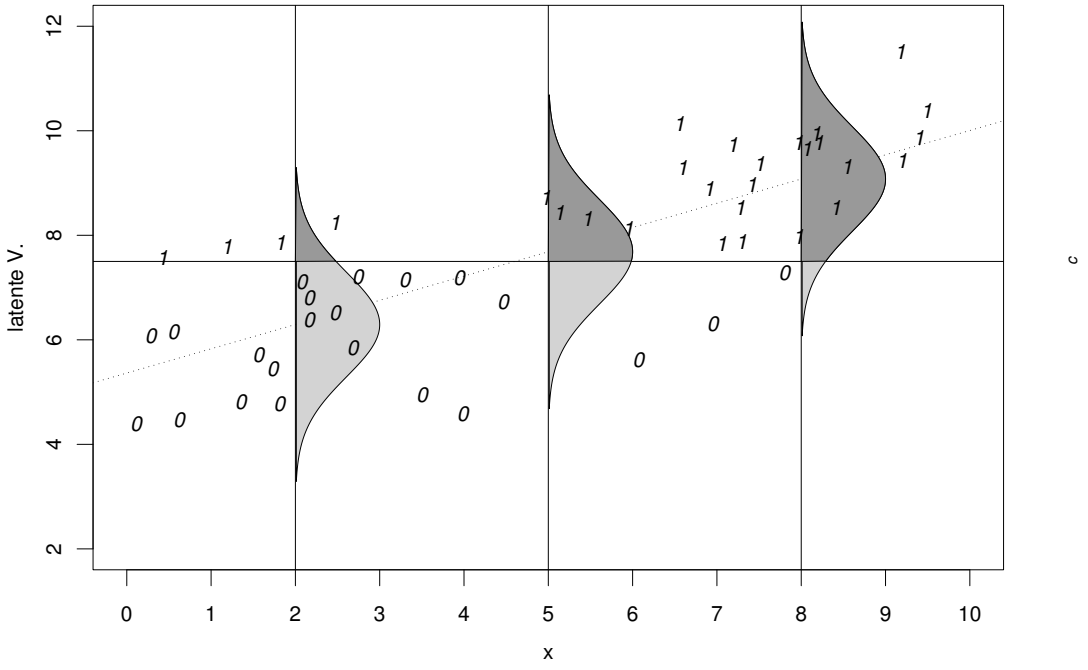


Abbildung 12.2.j: Zum Modell der latenten Variablen

gistischen Regression (12.1.f); die Funktion g ist die Umkehrfunktion der Verteilungsfunktion, also die entsprechende Quantil-Funktion. Wenn die E_i der **logistischen Verteilung** folgen, erhält man das logistische Regressionsmodell.

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

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

12.3 Schätzungen und Tests

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

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

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

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

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

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

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

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

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

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

Daraus erhält man

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

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

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

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

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

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

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

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

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

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

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

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

e **Berechnung.** Zur Lösung dieser Gleichungen braucht man ein iteratives Verfahren. Wie in der nichtlinearen Regression wird in jedem Schritt die Gleichung durch lineare Näherung so vereinfacht, dass sie zu einem linearen Regressionsproblem wird – hier zu einem mit Gewichten w_i. Wenn die Verbesserungsschritte schliesslich vernachlässigbar klein werden, ist die Lösung gefunden. Sie ist dann auch die exakte Lösung des genannten gewichteten linearen Regressionsproblems. Genaueres steht im Anhang 10.b.

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

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

hat eine genäherte Normalverteilung. Im Ausdruck für die Kovarianzmatrix müssen die geschätzten Koeffizienten eingesetzt werden, deshalb $\widehat{\boldsymbol{V}}^{(\beta)}$ statt $\boldsymbol{V}^{(\beta)}$. Da sie keine geschätzte Fehlervarianz $\widehat{\sigma}^2$ enthält, besteht keine theoretische Grundlage, 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 12.3.h) erscheint die Tabelle der geschätzten Koeffizienten, ihrer Standardfehler, der Werte der Teststatistiken und der P-Werte für die Hypothesen $\beta^{(j)} = 0$.

Auf den „Dispersion Parameter“ kommen wir päter zurück (13.2.f, 13.3.g, 13.4). Die „Null Deviance“ und die „Residual Deviance“ brauchen noch eine genauere Erklärung.

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

(Dispersion parameter for binomial family taken to be 1)

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

Number of Fisher Scoring iterations: 4
```

Tabelle 12.3.h: Computer-Ausgabe (leicht gekürzt) für das Beispiel Frühgeburten

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

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

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

* Diese erhält man, indem man in der Formel für $\ell\ell\langle \underline{\pi} \rangle$ (12.3.c) $\widetilde{\pi}_\ell$ durch $\widetilde{y}_\ell$ ersetzt. (Für $\widetilde{y}_\ell = 0$ und $\widetilde{y}_\ell = 1$ tritt $\log\langle 0 \rangle$ auf. Der Ausdruck wird aber in der Formel immer mit 0 multipliziert und die entsprechenden Terme können weggelassen werden.)
Setzt man dieses $\ell\ell^{(M)}$ und das erwähnte $\ell\ell\langle \underline{\pi} \rangle$ in die Definition der Devianz ein, so erhält man

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

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

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

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

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** (12.3.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 **??**

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** (12.2.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 12.3.l) sticht die Koeffizienten-Tabelle ins Auge, die wie in der multiplen linearen Regression Tests liefert, welche kaum interpretierbar sind. Die Ausgangsgrössen sind ja Faktoren, und es macht wieder wenig Sinn, einen einzelnen Koeffizienten auf Verschiedenheit von 0 zu testen – ausser für die zweiwertige Ausgangsgrösse 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 12.3.l: Computer-Ausgabe (gekürzt) für das Beispiel der Umweltumfrage

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

m Für **ungruppierte Daten** macht dieser Anpassungstest keinen Sinn. Für eine binäre Variable erhält man aus der Beobachtung nämlich keine Schätzung für ihre Varianz. (Es geht also nicht darum, dass die Näherung durch die Chiquadrat-Verteilung zu schlecht wäre.) In Anlehnung an den gerade erwähnten Test in der gewöhnlichen linearen Regression kann man aber die Daten auf der Basis des geschätzten Modells gruppieren. In SAS werden nach Hosmer and Lemeshow (2000) die angepassten Werte in 10 Gruppen mit (möglichst) gleich vielen Beobachtungen eingeteilt. Für jede Gruppe wird nun die Summe $\widetilde{Y}_\ell$ der „Erfolge“ Y_i gezählt und die Summe der geschätzten Wahrscheinlichkeiten $\widehat{\pi}_i$ gebildet. Auf diese Grössen wird dann der gewöhnliche Chiquadrat-Test angewandt. !!!???

n Der Vergleich zwischen einem grösseren und einem kleineren Modell wird gebraucht, um den **Einfluss einer nominalen Ausgangsgrösse** 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 12.3.n: Prüfung der Terme im Beispiel der Umweltumfrage

Tabelle 12.3.n zeigt, dass im **Beispiel der Umweltumfrage** für Geschlecht und Alter kein Einfluss auf die Beeinträchtugung nachgewiesen werden kann.

Für kontinuierliche und zweiwertige Ausgangs-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 12.3.l und in Tabelle 12.3.n erscheint eine Grösse **AIC**. Sie ist definiert als

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

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

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

Die Devianz ergibt sich wieder als Differenz zwischen

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

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

Im **Beispiel der Frühgeburten** liest man in der Computer-Ausgabe „Null Deviance: 318.42 on 245 degrees of freedom“ und „Residual Deviance: 235.89 on 243 degrees of freedom“. Die Teststatistik $318.42 - 235.89 = 82.53$ ergibt mit der Chiquadrat-Verteilung mit $245-243 = 2$ Freiheitsgraden einen P-Wert von 0. Die beiden Ausgangsgrössen 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 $2(\ell^{(G)} - \ell^{(K)}) = D\big\langle \underline{\widetilde{y}}; \underline{\widetilde{\pi}}^{(K)} \big\rangle - D\big\langle \underline{\widetilde{y}}; \underline{\widetilde{\pi}}^{(G)} \big\rangle$.
  Genäherte Verteilung unter H_0: $\chi^2_{p_G - p_K}$.
- Vergleich mit maximalem Modell, Anpassungstest: **Residuen-Devianz**.
  H_0: Angepasstes Modell mit p Parametern ist richtig.
  H_1: Maximales Modell M (mit einem Parameter für jede (Gruppen-) Beobachtung) ist richtig.
  Teststatistik $D\big\langle \underline{\widetilde{y}}; \underline{\widehat{\pi}} \big\rangle = 2(\ell^{(M)} - \ell\langle \underline{\widehat{\pi}} \rangle)$
  Genäherte Verteilung unter H_0, falls die m_ℓ genügend gross sind: $\chi^2_{\widetilde{n}-p}$ mit $\widetilde{n} =$ Anzahl (Gruppen-) Beobachtungen $\widetilde{Y}_\ell$. Dieser Test geht nur für gruppierte Beobachtungen!
- Gesamttest für die Regression: Vergleich von **Null-Devianz** $D\big\langle \underline{\widetilde{y}}; \underline{\widehat{\pi}}^0 \big\rangle$ und Residuen-Devianz.
  H_0: Null-Modell mit einem Parameter ist richtig.
  H_1: Angepasstes Modell mit p Parametern ist richtig.
  Teststatistik $D\big\langle \underline{\widetilde{y}}; \underline{\widehat{\pi}}^0 \big\rangle - D\big\langle \underline{\widetilde{y}}; \underline{\widehat{\pi}} \big\rangle = 2(\ell\langle \underline{\widehat{\pi}} \rangle - \ell\langle \underline{\widehat{\pi}}^0 \big\rangle)$.
  Genäherte Verteilung unter H_0: χ^2_{p-1}.

12.4 Residuen-Analyse

a Was **Residuen** sein sollen, ist nicht mehr eindeutig. Wir diskutieren hier die Definitionen für „zusammengefasste“ Daten, siehe 12.2.g. Für zweiwertige Zielgrössen ohne Gruppierung muss man $m_\ell = 1$ setzen.
Die Grössen

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

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

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

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

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

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

und haben genäherte Varianz 1.

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

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

war. Die Gewichte, die hier gebraucht werden, sind diejenigen, die im Algorithmus (12.3.e) für das angenäherte lineare Regressionsproblem verwendet werden. Sie sind gleich $w_\ell = m_\ell/(\widehat{\pi}_\ell(1-\widehat{\pi}_\ell))$ (vergleiche . 10.d). Die genauer standardisierten Residuen sind dann

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

e Ein weiterer, gut bekannter Typ von Residuen sind die **Devianz-Residuen**. Sie orientieren sich am Beitrag der $i$ten Beobachtung zur Devianz des Modells, der gemäss 12.3.i und 12.3.b gleich

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

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

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

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

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

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

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

verwenden (Abbildung 12.4.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.

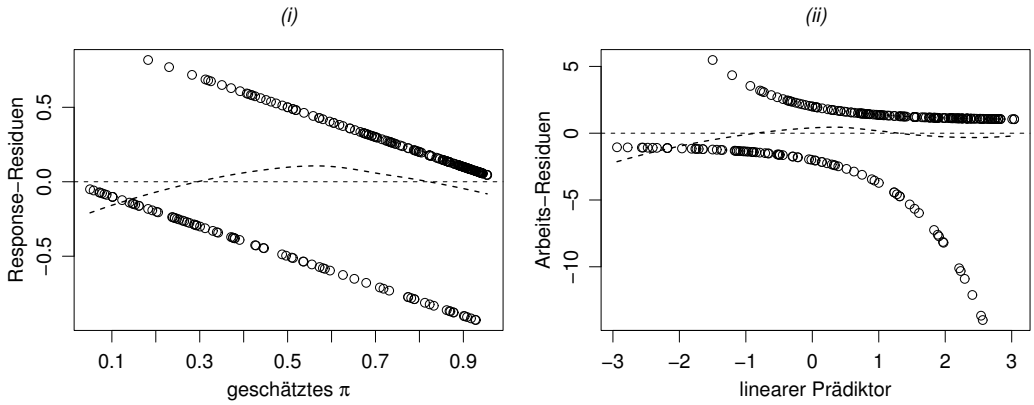


Abbildung 12.4.g: Tukey-Anscombe-Diagramme im Beispiel der Frühgeburten: (i) Response-Residuen und geschätzte π_i, (ii) Arbeits-Residuen und linearer Prädiktor

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

i Die Situation ist wesentlich besser, wenn **gruppierte Daten** vorliegen. Abbildung 12.4.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 erden die Überlebenschancen für leichte Neugeborene vom Modell überschätzt. Der Mangel sollte durch (weitere) Transformation dieser Ausgangsgrösse behoben werden.

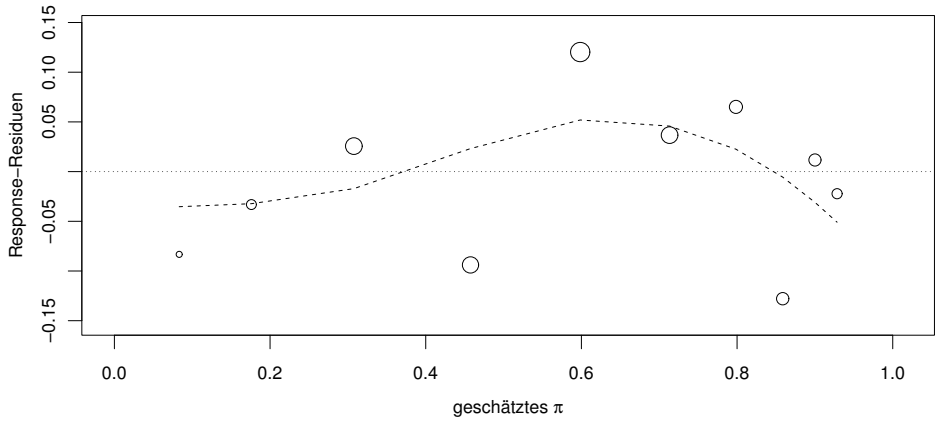


Abbildung 12.4.i: Tukey-Anscombe-Diagramm im Beispiel der Frühgeburten mit klassiertem Gewicht

j Um allfällige nicht-lineare Abhängigkeiten der Zielgrösse von den Ausgangsgrössen zu entdecken, kann man, wie in der multiplen linearen Regression, die **Residuen gegen die Ausgangs-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 Ausgangsgrösse addieren. So erhält man einen „**partial residual plot**“ oder „**term plot**“.

In Abbildung 12.4.j sieht man, dass für das Gewicht auch in dem erweiterten Modell der Effekt ungenügend modelliert ist (vergleiche 12.4.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 (10.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

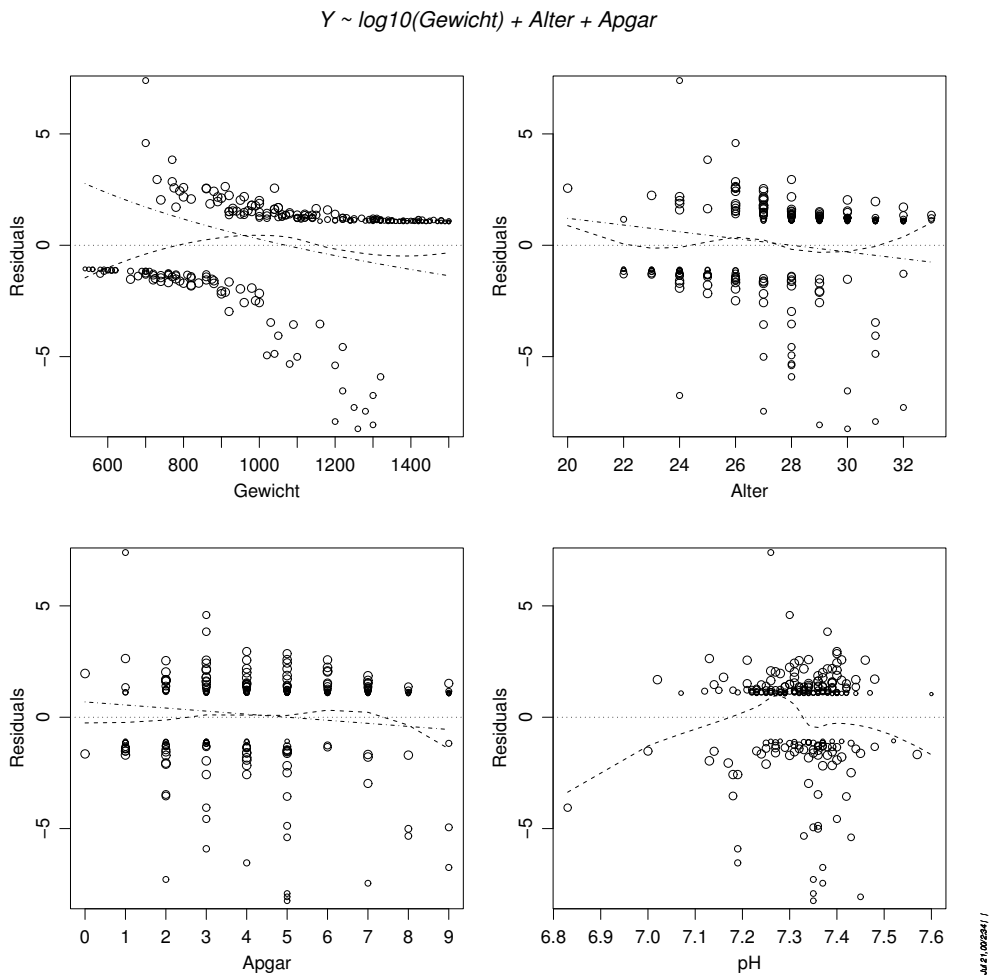


Abbildung 12.4.j: Residuen gegen Ausgangsgrössen im Beispiel der Frühgeburten. Die Radien der Kreise entsprechen den Gewichten. Einige extrem negative Residuen wurden weggeschnitten.

Analyse könnte das Modell versuchsweise ohne diese Beobachtungen angepasst werden.

12.S S-Funktionen

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

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

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

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 verwen-

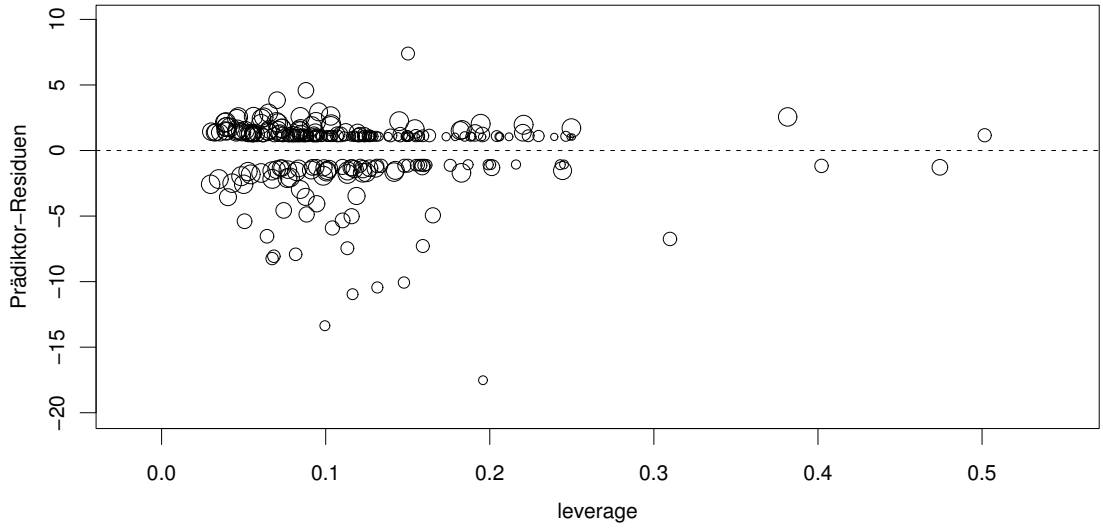


Abbildung 12.4.k: Residuen gegen Hebelarm-Werte $(\boldsymbol{H}_W)_{ii}$ für das Beispiel der Frühgeburten

det 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 Ausgangsvariablen 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.

13 Verallgemeinerte Lineare Modelle

13.1 Das Modell der Poisson-Regression

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

b **Beispiel gehemmte Reproduktion.** In einer Studie zur Schädlichkeit von Flugbenzin wurde die Reproduktion von Ceriodaphnia in Abhängigkeit von verschiedenen Konzentrationen des Schadstoffs für zwei Stämme von Organismen untersucht (Quelle: Myers, Montgomery and Vining (2001), example 4.5). Wie Abbildung 13.1.b zeigt, fällt die Anzahl der reproduzierenden Organismen stark ab; die Abnahme könnte etwa exponentielle Form haben.

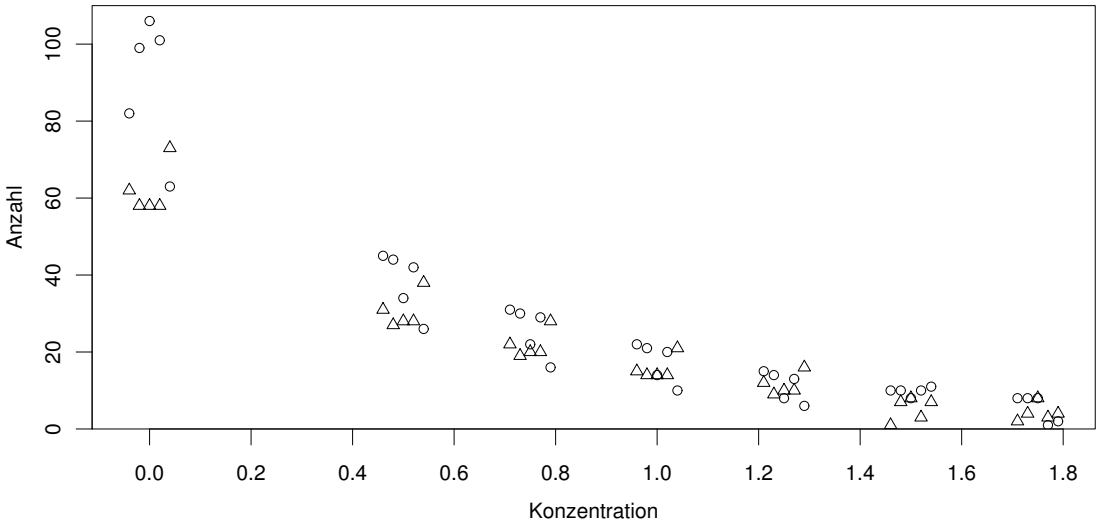


Abbildung 13.1.b: Anzahl reproduzierende Individuen im Beispiel der gehemmten Reproduktion. Die beiden Stämme sind mit verschiedenen Symbolen angegeben.

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

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

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

und die Y_i sollen stochastisch unabhängig sein.

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

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

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

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

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

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

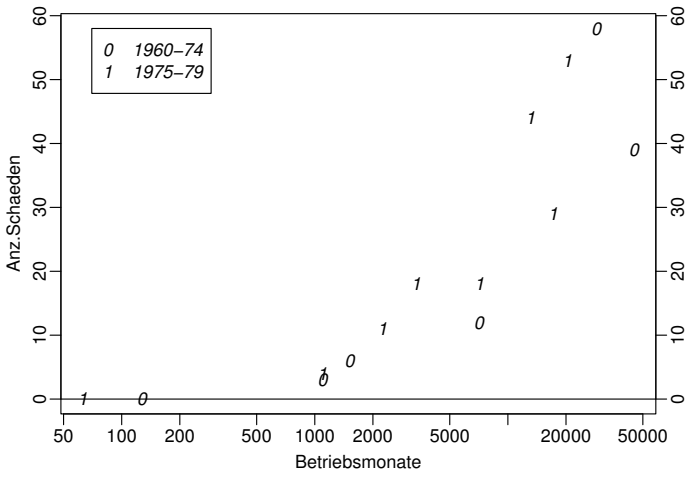
Die Zunahme von $x^{(j)}$ um eine Einheit bewirkt eine Multiplikation des Erwartungswertes λ um den Faktor $\widetilde{\beta}_j$, der auch als „Unit risk “ bezeichnet wird. Ist β_j positiv, so ist $\widetilde{\beta}_j > 1$, und der Erwartungswert wird mit zunehmendem $x^{(j)}$ grösser.

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

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

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

g **Beispiel Schiffs-Havarien**. Grosse Wellen können an Lastschiffen Schäden verursachen. Wovon hängen diese Havarien ab? Um diese Frage zu beantworten, wurden 7 „Flotten“ vergleichbarer Schiffe in je zwei Beobachtungsperioden untersucht (Quelle: McCullagh and Nelder (1989, p. 205), Teil der Daten). Für jede dieser 7×2 Beobachtungseinheiten wurde die Summe der Betriebsmonate über die Schiffe (M) erhoben und die Anzahl Y_i der Schadensereignisse eruiert. In der Tabelle in Abbildung 13.1.g sind ausserdem die Beobachtungsperiode (P), die Bauperiode (C) und Schiffstyp (T) notiert. Die Daten ergeben sich also aus einer Gruppierung von ursprünglichen Angaben über einzelne Schiffe, die entsprechend der Bauperiode, dem Schiffstyp und der Beobachungsperiode zusammengefasst wurden. Der wichtigste und offensichtlichste Zusammenhang – derjenige zwischen Anzahl Schadensereignisse und Anzahl Betriebsmonate – ist in der Abbildung grafisch festgehalten.



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

Abbildung 13.1.g: Daten zum Beispiel der Schiffs-Havarien. T: Schiffstyp, C: Bauperiode, P: Beobachtungsperiode, M: Betriebsmonate, Y: Anzahl Havarien

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

h Für dieses Beispiel ist das folgende Modell plausibel:

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

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

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

Weshalb wurde hier die Summe `M` der Betriebsmonate logarithmiert? Es ist plausibel, anzunehmen, dass die erwartete Anzahl Schadensfälle exakt proportional zu `M` ist, also, wenn man die anderen Einflussgrössen weglässt, $\mathcal{E}\langle Y_i \rangle = \alpha \mathtt{M}_i$, und deshalb $\log\langle \mathcal{E}\langle Y_i \rangle \rangle = \beta_0 + \beta_M \log\langle \mathtt{M}_i \rangle$ mit $\beta_0 = \log\langle \alpha \rangle$ und $\beta_M = 1$. Wir werden also erwarten, dass die Schätzung $\widehat{\beta}_M$ ungefähr 1 ergibt.

Dass sich eine allfällige Veränderung zwischen den Beobachtungsperioden P bzw. den Schiffstypen T ebenfalls multiplikativ auswirken sollte, ist sehr plausibel. Der Faktor $\exp\langle \beta_P \rangle$ beschreibt dann die Veränderung des Risikos, d.h. wie viel mal mehr Schäden in der zweiten Periode zu erwarten sind.

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

j Im Beispiel wurden die Schiffe, die eigentlich die natürlichen Beobachtungseinheiten wären, zu Gruppen zusammengefasst, und die Zielgrösse war dann die Summe der Zahlen der Havarien für die einzelnen Schiffe. Wie in 11.1.f erwähnt, ist diese Situation häufig. Es entstehen meistens Kreuztabellen. Wir werden in Kapitel 11.2.p sehen, dass die Poisson-Regression (oder besser -Varianzanalyse) für ihre Analyse eine entscheidende Rolle spielt.

13.2 Das Verallgemeinerte Lineare Modell

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

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

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

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

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

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

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

- Der Parameter θ heisst der **kanonische Parameter**. Parameter!kanonischer Die Ausgangs-Variablen werden, wenn wir wieder zu den Verallgemeinerten Linearen Modellen zurückkehren, diesen kanonischen Parameter kontrollieren.
- ϕ ist ein weiterer Parameter, der mit der Varianz zu tun hat und **Dispersions-Parameter** genannt wird. Er ist normalerweise ein Störparameter und wird mit

der Regression nichts zu tun haben. (Genau genommen ist die Familie nur eine Exponential-Familie, wenn ϕ als fest angenommen wird.)

- Die Grösse ω ist eine feste Zahl, die bekannt ist, aber von Beobachtung zu Beobachtung verschieden sein kann. Sie hat die Bedeutung eines **Gewichtes** der Beobachtung. Man könnte sie auch in die Grösse ϕ hineinnehmen. Bei mehreren Beobachtungen i wird ω von i abhängen, während ϕ für alle gleich ist. (Bei gruppierten Daten in der logistischen Regression wird $\omega_\ell = m_\ell$ sein, wie wir gleich feststellen werden.)
- Die Funktion $b\langle . \rangle$ legt fest, um welche Exponentialfamilie es sich handelt.
- Die Funktion $c\langle . \rangle$ wird benötigt, um die Dichte oder Wahrscheinlichkeitsfunktion auf eine Gesamt-Wahrscheinlichkeit von 1 zu normieren.

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

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

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

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

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

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

gilt.

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

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

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

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

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

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

f **Binomialverteilung**. In 12.2.g wurde der *Anteil* $\widetilde{Y}_\ell$ von „Erfolgen“ unter m_ℓ Versuchen als Zielgrösse verwendet und festgestellt, dass $m_\ell\widetilde{Y}_\ell$ binomial verteilt ist. Die Wahrscheinlichkeiten, ohne $\widetilde{\ }$ und Index ℓ geschrieben, sind dann $P\langle Y=y\rangle = \binom{m}{my}\pi^{my}(1-\pi)^{m-my}$ und ihre logarithmierten Werte kann man schreiben als

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

Hier ist

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

Für Erwartungswert und Varianz gilt $\mu = b'\langle\theta\rangle = \exp\langle\theta\rangle\,/(1+\exp\langle\theta\rangle) = \pi$ und $\mathrm{var}\langle Y\rangle = b''\langle\theta\rangle = \exp\langle\theta\rangle\,(1+\exp\langle\theta\rangle) - (\exp\langle\theta\rangle)^2/(1+\exp(\theta))^2 = \pi(1-\pi)$.
Für binäre Variable gilt die Formel natürlich auch, mit $m=1$.

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

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

Hier erhält man

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

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

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

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

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

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

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

k Prinzipiell kann man aber auch **andere Link-Funktionen** verwenden. Wenn beispielsweise $0 < \mathcal{E}\langle Y \rangle < 1$ gelten muss, lässt sich jede kumulative Verteilungsfunktion als inverse Link-Funktion einsetzen (12.2.j). Wenn es keine konkreten Gründe für eine spezielle Link-Funktion gibt, verwendet man aber in der Regel die kanonische. Zum einen besitzen „kanonische verallgemeinerte lineare Modelle“ bessere theoretische Eigenschaften (Existenz und Eindeutigkeit des ML-Schätzers). Zum andern vereinfachen sich dadurch die Schätzgleichungen.

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

13.3 Schätzungen und Tests

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

b **Likelihood**. Die Parameter, die uns interessieren, sind die Koeffizienten β_j. Sie bestimmen den Erwartungswert μ_i für jede Beobachtung, und dieser bestimmt schliesslich θ_i (siehe 13.2.d). Wir nehmen an, dass ϕ für alle Beobachtungen gleich ist. Der Beitrag einer Beobachtung i zur Log-Likelihood $\ell\ell$ ist gleich

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

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

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

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

c **Maximum-Likelihood-Schätzung**. Wir leiten hier die Schätzungen für den Spezialfall der Poisson-Regression mit „log-Link“ her. Die analoge, allgemeine Herleitung der Schätzgleichungen, eine Skizzierung des Schätzalgorithmus und einige Eigenschaften der Schätzer findet man im Anhang 10.A.

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

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

Setzt man alle Komponenten gleich null,

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

so entstehen die impliziten Gleichungen, die die Maximum-Likelihood-Schätzung $\widehat{\underline{\beta}}$ bestimmen; für den Poisson-Fall $\sum_i (y_i - \lambda_i) \cdot x_i^{(j)} = 0$.
Zur Lösung dieser Gleichungen geht man so vor, wie das für die logistische Regression in 12.3.e skizziert wurde und wie es in Anhang 10.b beschrieben ist.

d **Schätzung des Dispersions-Parameters.** Im allgemeinen Modell muss auch der Dispersions-Parameter ϕ geschätzt werden, und auch das erfolgt durch Maximieren der Likelihood. Für die spezifischen Modelle kommt dabei eine recht einfache Formel heraus. Für die Normalverteilung kommt, bis auf einen Faktor $(n-p)/n$, die übliche Schätzung der Varianz heraus. Für binomial- und Poisson-verteilte Zielgrössen muss kein Dispersions-Parameter geschätzt werden – wir werden in 13.4 diese gute Nachricht allerdings wieder einschränken.

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

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

wobei die Kovarianzmatrix $\boldsymbol{V}^{(\beta)}$ normalerweise von $\underline{\beta}$ abhängen wird. (Genaueres steht im Anhang, 10.e.) Damit lassen sich genäherte P-Werte für Tests und Vertrauensintervalle angeben. In der linearen Regression galt die Verteilung exakt, mit $\boldsymbol{V}^{(\beta)} = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$, und das ergab exakte P-Werte und Vertrauensintervalle.

f Für das **Beispiel der gehemmten Reproduktion** zeigt Tabelle 13.3.f den Aufruf der S-Funktion `regr` und die Computer-Ausgabe, die die bereits bekannte Form hat. Beide Ausgangsgrössen erweisen sich als hoch signifikant.

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

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

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

Tabelle 13.3.f: Computer-Ausgabe von `regr` für das Beispiel der gehemmten Reproduktion

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

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

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

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

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

Für binomial verteilte Zielgrössen wurde die Devianz in 12.3.i angegeben.

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

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

Die entsprechenden Devianz-Differenzen sind unter gewissen Bedingungen näherungsweise chiquadrat-verteilt. Für die Residuen-Devianz binärer Zielgrössen sind diese Bedingungen, wie erwähnt (12.3.k), nicht erfüllt.

* Die Bedingungen sind also für einmal nicht harmlos. Das liegt daran, dass im maximalen Modell M (13.3.g) für jede Beobachtung ein Parameter geschätzt wird; mit der Anzahl Beobachtungen geht also auch die Anzahl Parameter gegen unendlich, und das ist für asymptotische Betrachtungen gefährlich!

j Die Devianz wird für die Normalverteilung zur Summe der quadrierten Residuen, die ja bei der Schätzung nach dem Prinzip der Kleinsten Quadrate minimiert wird. Für andere Verteilungen haben die „rohen Residuen“ (12.4.a) verschiedene Varianz und sollten mit entsprechenden Gewichten summiert werden. Die Grösse

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

heisst **Pearson-Chiquadrat-Statistik**. Wenn $\widetilde{\phi}$ nicht aus den Daten geschätzt werden

muss, folgt sie in der Regel genähert einer Chiquadrat-Verteilung. Wenn T zu gross wird, müssen wir auf signifikante Abweichung vom Modell schliessen. Das legt einen **Anpassungstest** fest.

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

13.4 Übergrosse Streuung

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

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

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

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

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

c Der Parameter ϕ lässt sich analog zur Varianz der Normalverteilung schätzen $\widehat{\phi} = \frac{1}{n-p}\sum_i \frac{\omega_i(y_i-\widehat{\mu}_i)^2}{V\langle\mu_i\rangle}$.
Man teilt also die Pearson-Statistik durch ihre Freiheitsgrade. Üblicher ist es aber, statt der Pearson-Statistik die Devianz zu verwenden, die ja, wie gesagt (13.3.j), näherungsweise das Gleiche ist. Das ergibt $\widehat{\phi} = (1/(n-p))D\langle\underline{y};\underline{\widehat{\mu}}\rangle$. Im Beispiel der gehemmten Reproduktion erhält man mit den Angaben von 13.3.f $\widehat{\phi} = 88/67 = 1.3$.

d Im Anhang (10.e) kann man sehen, dass die Kovarianzmatrix der asymptotischen Verteilung der geschätzten Koeffizienten den Faktor ϕ enthält. (* $\widetilde{\boldsymbol{H}}$ enthält den Faktor $1/\phi$, siehe 10.c.) Durch die Einführung eines Dispersions-Parameters werden deshalb einfach Konfidenzintervalle um den Faktor $\sqrt{\hat{\phi}}$ breiter und die Werte der Teststatistiken um $1/\widehat{\phi}$ kleiner.

Die Funktion `regr` verwendet den geschätzten Streuungsparameter $\widehat{\phi}$ zur Berechnung der Tests von Koeffizienten und von Vertrauensintervallen, sofern der mittlere Wert der Zielgrösse gross genug ist (momentan wird als Grenze 3 verwendet) – ausser, dies werde mit dem Argument `calcdisp=FALSE` unterdrückt (wie es in 13.3.f getan wurde).

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

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

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

13.5 Residuen-Analyse

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

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

  Wie erwähnt, haben diese Residuen verschiedene Varianzen.

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

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

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

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

  Diese „unskalierten“ Pearson-Residuen dienen dazu, den Dispersions-Parameter zu schätzen oder zu prüfen, ob er gleich 1 sein kann, wie dies für das Binomial- und das Poisson-Modell gelten muss (vgl. 13.4). Die Grössen $R_i^{(P)}/\widehat{\phi}$ nennen wir skalierte Pearson-Residuen,

- **Devianz-Residuen**: Jede Beobachtung ergibt einen Beitrag d_i/ϕ zur Devianz (13.3.g), wobei

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

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

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

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

b Die wichtigsten grafischen Darstellungen der Residuen-Analyse sind:

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

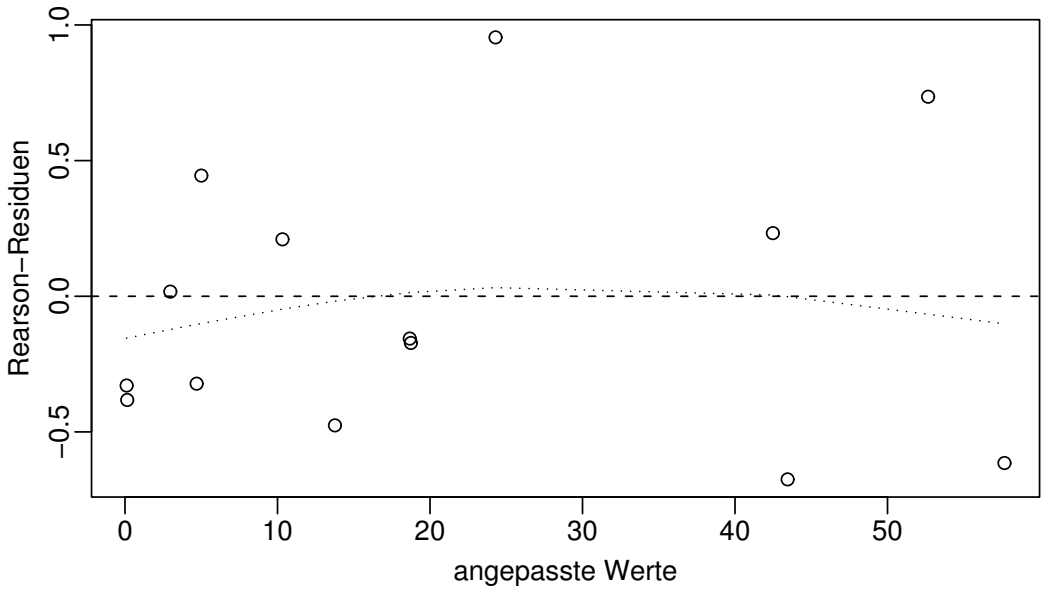


Abbildung 13.5.b: Tukey-Amscombe Plot zum Beispiel der Schiffs-Havarien

c • **Scale Plot**. Absolute (Pearson-) Residuen gegen angepasste Werte $\widehat{\mu}_i$ auftragen. Wenn eine Glättung einen Trend zeigt, ist die Varianzfunktion nicht passend. Man kann versuchen, sie direkt zu modellieren, siehe 13.4.f.

d • **Residuen gegen Ausgangs-Variable.** Prädiktor-Residuen $R_i^{(L)}$ werden gegen Ausgangs-Variable $x_i^{(j)}$ aufgetragen. Gekrümmte Glättungen deuten wie in der linearen Regression an, wie die Ausgangsgrössen transformiert werden sollten. Die Funktion `plresx` liefert wieder eine Referenzlinie für gleiche Werte des linearen Prädiktors. Da die Residuen mit verschiedenen Gewichten zur Regression beitragen, sollten sie dem entsprechend verschieden gross gezeichnet werden.
Wieder ist es üblicher, die **partiellen Residuen** zu verwenden und den Effekt der Ausgangs-Variablen mit einzuzeichnen, also einen **partial residual plot** oder **term plot** zu erstellen (vergleiche 12.4.j).

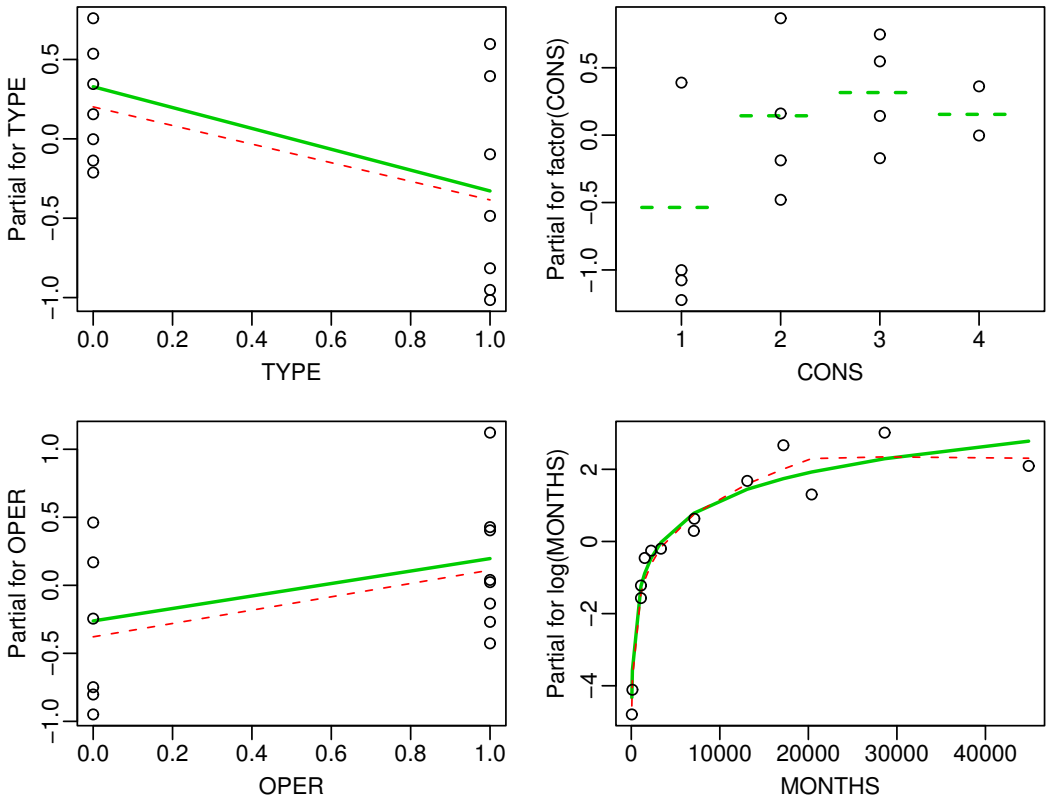


Abbildung 13.5.d: Partial residual Plots zu dem Havarie-Modell

e • **Leverage Plot**. Die Prädiktor-Residuen $R_i^{(L)}$ werden gegen die „fast ungewichteten" Hebelarm-Werte $\widetilde{h}_i$ aufgetragen und die Gewichte w_i durch verschieden grosse Kreis-Symbole dargestellt (vergleiche 12.4.k).

f Abbildungen 13.5.b und 13.5.d zeigen Residuenplots zum Modell im Beispiel Schiffs-Havarien. Bei so kleiner Beobachtungszahl sind Abweichungen kaum auszumachen.

13.S S-Funktionen

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

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

11 Kategorielle Zielgrössen

11.1 Multinomiale Zielgrössen

a In der logistischen Regression war die Zielgrösse zweiwertig. Im Beispiel der Umweltumfrage (12.2.d) hatte die Zielgrösse „Beeinträchtigung“ eigentlich vier mögliche Werte, die wir für das dortige Modell zu zwei Werten zusammengefasst haben. Die vier Werte zeigen eine Ordnung von „gar nicht“ bis „stark“. In der gleichen Umfrage wurde auch eine weitere Frage gestellt: „Wer trägt im Umweltschutz die Hauptverantwortung? – Einzelne, der Staat oder beide?“. Diese drei Auswahlantworten haben keine eindeutige Ordnung, denn vielleicht nehmen jene, die mit „beide“ antworten, den Umweltschutz besonders ernst, und deshalb liegt diese Antwort nicht unbedingt zwischen den beiden anderen.

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

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

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

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

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

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

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

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

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

d* **Gruppierte Daten**. Wie in der logistischen Regression (11.2.n) kann man die Beobachtungen mit gleichen Werten der Ausgangsgrössen zusammenfassen und zählen, wie viele von ihnen die verschiedenen Werte k der Zielgrösse zeigen. Es sei wieder m_ℓ die Anzahl der Beobachtungen mit $\underline{x}_i = \underline{\widetilde{x}}_\ell$, und $\widetilde{Y}_\ell^{(k)}$ der *Anteil* dieser Beobachtungen, für die $Y_i = k$ ist. Die *Anzahlen* $m_\ell \cdot \widetilde{Y}_\ell^{(k)}$ folgen dann der multinomialen Verteilung mit den Parametern $\widetilde{\pi}_\ell^{(1)}, ..., \widetilde{\pi}_\ell^{(k^*)}$, die durch das oben angegebene Modell bestimmt sind. Die Wahrscheinlichkeiten sind

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

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

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

f **S-Funktionen**. Im Statistik-System R steht im package `nnet` die Funktion `multinom` zur Verfügung, um solche Modelle anzupassen. Für das **Beispiel der Umweltumfrage** zeigt Tabelle 11.1.f ein `summary` des Modells, das die Frage nach der Hauptverantwortung in Abhängigkeit vom Alter und Geschlecht der Befragten beschreibt. Man kann die geschätzten Koeffizienten $\widehat{\beta}_{j\ell}$ und ihre Standardfehler ablesen.

Die Referenzkategorie ist „Einzelne“. Der Koeffizient von j = Alter für k = Staat ist $\widehat{\beta}_j^{(k)} = -0.00270$. In 50 Jahren nehmen also die log odds von „Staat“:„Einzelne“ um $0.0027 \cdot 50 = 0.135$ ab; als odds ratio ergibt sich $\exp\langle -0.135 \rangle = 0.874$. Allerdings ist der Koeffizient nicht signifikant, da $\widehat{\beta}_j^{(k)}/\text{standard error}_j^{(k)} = -0.0027/0.0034 = 0.79$ einen klar nicht signifikanten z-Wert ergibt. Zwischen den Geschlechtern besteht ein signifikantes Doppelverhältnis von $\exp\langle -0.244 \rangle = 0.78$. Frauen weisen die Verantwortung stärker den Einzelnen anstelle des Staates zu als Männer.

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

S-Funktionen. Im R-System sieht die Funtion `drop1` für multinomiale Modelle leider keinen Test vor. Man muss mit der Funktion `anova` die einzelnen Modelle vergleichen (oder `drop1` entsprechend ergänzen). Tabelle 11.1.g zeigt die Resultate einer erweiterten Funktion `drop1`, die den Test durchführt, für ein ausführlicheres Modell.

Erstaunlicherweise haben weder die politische Partei, noch das Alter oder die Wohnlage einen signifikanten Einfluss auf die Zuweisung der Hauptverantwortung. Das liegt nicht an einem starken Zusammenhang der Ausgangs-Variablen analog zum Kollinearitätspro-

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

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

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

Residual Deviance: 3385
AIC: 3425
```

Tabelle 11.1.f: Ergebnisse einer multinomialen Logit-Regression im Beispiel der Umweltumfrage

blem, das in der linearen Regression besprochen wurde, denn auch bei einer schrittweisen Elimination bleiben diese drei Variablen nicht-signifikant.

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

Tabelle 11.1.g: Signifikanzen von einzelnen Termen im Beispiel der Umweltumfrage

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

Eine Möglichkeit, die genauen Schätzungen zu erhalten, führt über eine andere Anordnung der Daten, die in 12.2.l besprochen wird.

i Die **Residuen-Devianz** ist wie in der logistischen Regression (Buch 12.3.i) sinnvoll bei Daten, die zu Anzahlen zusammengefasst werden können (mit $m_\ell > 3$ oder so). Hier wird die maximale Likelihood erreicht für $\widehat{\pi}_\ell^{(k)} = \widetilde{y}_\ell^{(k)}$ und man erhält

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

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

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

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

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

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

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

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

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

11.2 Geordnete Zielgrössen

a Wie früher erwähnt (11.1.a), haben Variable oft einen geordneten Wertebereich. Wie kann man diesen Aspekt ausnützen, wenn eine solche Grösse die Zielgrösse einer Regression ist?

Im **Beispiel der Umweltumfrage** (11.1.c) interessierte uns die Frage nach der Beeinträchtigung mit ihren geordneten Antwortmöglichkeiten von „überhaupt nicht“ bis „sehr“. Bei der Auswertung mit Kreuztabellen wurde diese Ordnung nicht berücksichtigt. Nun soll sie als Zielgrösse betrachtet und ihr Zusammenhang mit Ausgangsgrössen wie Schulbildung, Geschlecht und Alter untersucht werden.

b **Modell.** Zur Beschreibung eines Modells hilft, wie für die binäre Zielgrösse (12.2.j), die Annahme einer **latenten Variablen** Z, aus der sich die Kategorien der Zielgrösse durch Klassieren ergeben. Das frühere Modell wird erweitert, indem man mehrere **Schwellenwerte** α_k festlegt. Die Zielgrösse Y ist =0, wenn Z kleiner ist als die kleinste Schwelle α_1, sie ist =1, wenn Z zwischen α_1 und α_2 liegt, usw. Bei k^* Schwellenwerten nimmt Y die k^*+1 Werte $0, 1, ..., k^*$ an.

In Formeln:

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

Das bedeutet, dass

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

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

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

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

Abbildung 11.2.b veranschaulicht diese Vorstellung für eine einzige Ausgangs-Variable. Bei mehreren Ausgangsgrössen wäre auf der horizontalen Achse, wie üblich, der lineare Prädiktor $\eta_i = \underline{x}_i^T \underline{\beta}$ zu verwenden.

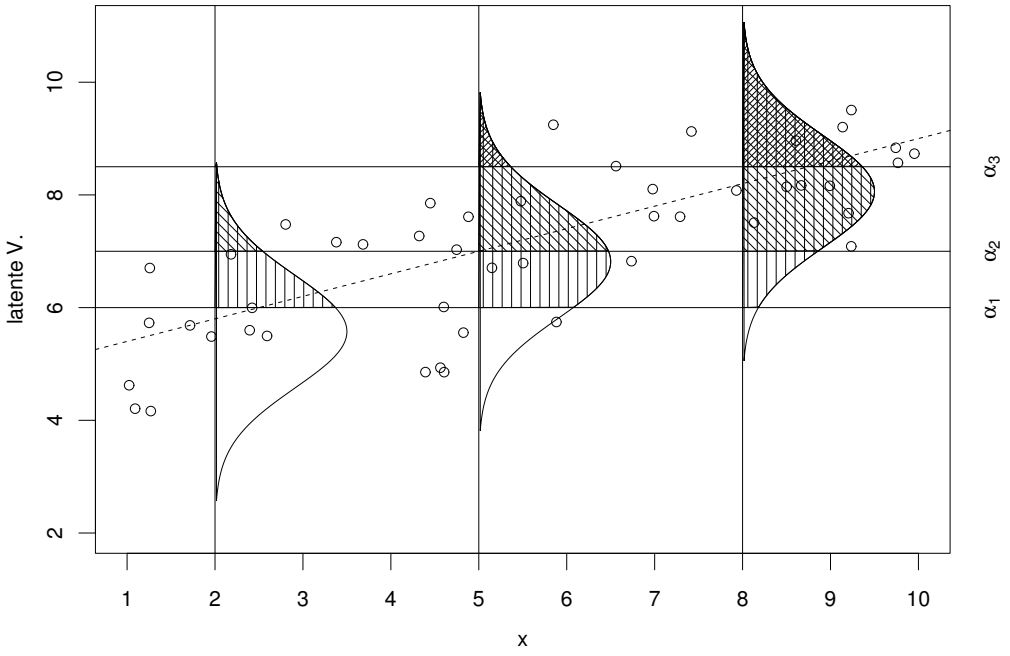


Abbildung 11.2.b: Zum Modell der latenten Variablen

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

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

wobei F die kumulative Verteilungsfunktion der Zufallsabweichungen E_i bezeichnet.

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

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

$$g\left\langle \gamma_k\langle \underline{x}_i \rangle \right\rangle = \sum\nolimits_j \beta_j x_i^{(j)} - \alpha_k$$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

l **S-Funktionen**. Im R findet man die Funktion `polr`, was für „Proportional Odds Logistic Regression" steht. Das `summary` (Tabelle 11.2.l (i)) liefert, wie üblich, die Tabelle der Koeffizienten mit Werten der t-Statistik für die Tests $\beta_j = 0$, die für Faktoren mit mehr als 2 Werten wenig Sinn machen. (Die P-Werte werden nicht mitgeliefert; man muss sie selbst ausrechnen.)

Wie in früheren Modellen zeigt die Funktion `drop1(t.r,test="Chisq")` die Signifikanz der Faktoren (Tabelle 11.2.l (ii)).

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

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

Residual Deviance: 4114.67
AIC: 4144.67
```

Tabelle 11.2.l (i): Resultate für die Regression der geordneten Zielgrösse Beeinträchtigung auf mehrere Ausgangsgrössen im Beispiel der Umweltumfrage

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

Tabelle 11.2.l (ii): Signifikanz der einzelnen Terme im Beispiel

Achtung! Eine kleine Simulationsstudie mit 500 Beobachtungen und 2-3 Variablen (davon ein 3-4-stufiger Faktor) und einer Zielgrösse mit 3 Werten hat alarmierende Resultate gebracht: Die ausgewiesenen Standardfehler waren um einen Faktor von 2 bis 3 zu klein. Die Resultate von `polr` stimmten zudem schlecht mit einer alternativen Berechnungsmethode überein, die gleich geschildert wird. Die Resultate sind also mit äusserster Vorsicht zu geniessen. Es ist bis auf Weiteres angezeigt, die Bootstrap-Methode zu benützen, um die Unsicherheiten zu erfassen. Für Vorhersagen der richtigen Klasse sind die Methoden vermutlich zuverlässiger.

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

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

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

oder, tabellarisch,

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

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

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

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

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

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

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

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

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

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

Die entsprechende S-Funktion soll demnächst entstehen und in die `f.reg` eingebaut werden.

11.S S-Funktionen

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

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

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

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

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

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

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. 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.

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.

Hosmer, D. W. and Lemeshow, S. (2000). *Applied Logistic Regression*, 2nd edn, Wiley, N.Y.

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.

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.

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.

Vincze, I. (1984). *Mathematische Statistik mit industriellen Anwendungen*, Band1, 2, 2. Aufl., Bibliograhisches Institut, Mannheim.

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*, Chapmann and Hall, London.