9 Verallgemeinerte Lineare Modelle

9.1 Das Modell der Poisson-Regression

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

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

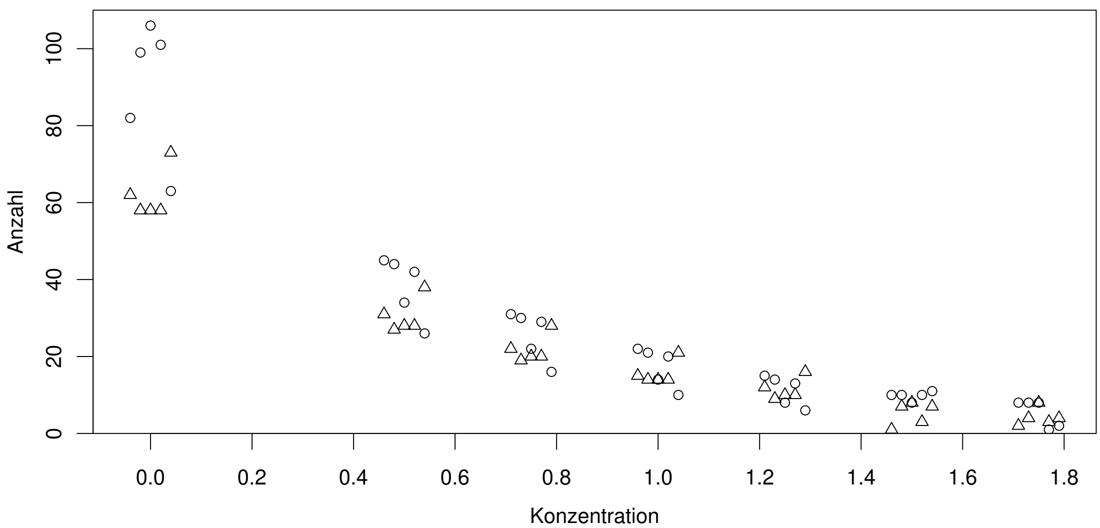


Abbildung 9.1.b: Anzahl reproduzierende Individuen im Beispiel der gehemmten Reproduktion. Die beiden Stämme sind mit verschiedenen Symbolen angegeben.

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

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

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

und die Y_i sollen stochastisch unabhängig sein.

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

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

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

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

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

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

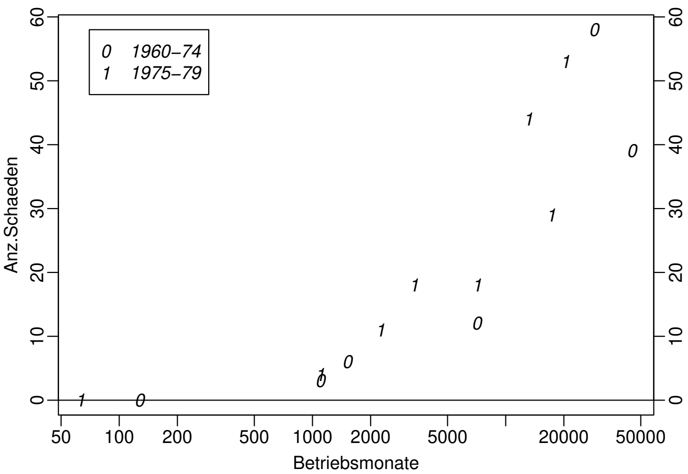
Die Zunahme von $x^{(j)}$ um eine Einheit bewirkt eine Multiplikation des Erwartungswertes λ um den Faktor $\widetilde{\beta}_j$, der auch als „Unit risk “ bezeichnet wird. Ist β_j positiv, so ist $\widetilde{\beta}_j > 1$, und der Erwartungswert wird mit zunehmendem $x^{(j)}$ grösser.

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

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

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

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



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

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

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

h Für dieses Beispiel ist das folgende Modell plausibel:

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

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

$$\mathtt{Y} \sim\; \mathtt{log10(M)\;+\;T\;+\;P\;+\;C}\;.$$

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 7.1.f erwähnt, ist diese Situation häufig. Es entstehen meistens Kreuztabellen. Wir werden in Kapitel 14.S.0.b sehen, dass die Poisson-Regression (oder besser -Varianzanalyse) für ihre Analyse eine entscheidende Rolle spielt.

9.2 Das Verallgemeinerte Lineare Modell

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

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

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

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

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

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

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

- Der Parameter θ heisst der **kanonische Parameter**. Die Eingangs-Variablen werden, wenn wir wieder zu den Verallgemeinerten Linearen Modellen zurückkehren, diesen kanonischen Parameter kontrollieren.

- ϕ ist ein weiterer Parameter, der mit der Varianz zu tun hat und **Dispersions-Parameter** genannt wird. Er ist normalerweise ein Störparameter und wird mit der Regression nichts zu tun haben. (Genau genommen ist die Familie nur eine Exponential-Familie, wenn ϕ als fest angenommen wird.)

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

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

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

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

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

Nun kann man auch die $b''\langle\theta\rangle$ direkt als Funktion von μ schreiben, $V\langle\mu\rangle = b''\left\langle (b')^{-1}\langle\mu\rangle \right\rangle$. 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

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

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

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

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

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

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

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

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 (8.2.j). Wenn es keine konkreten Gründe für eine spezielle Link-Funktion gibt, verwendet man aber in der Regel die kanonische. Zum einen besitzen „kanonische verallgemeinerte lineare Modelle" bessere theoretische Eigenschaften (Existenz und Eindeutigkeit des ML-Schätzers). Zum andern vereinfachen sich dadurch die Schätzgleichungen.

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

9.3 Schätzungen und Tests

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

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

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

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

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

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

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

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

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

Setzt man alle Komponenten gleich null,

$$s\langle\underline{\beta}\rangle = \sum_i \underline{s}_i\langle\underline{\beta}\rangle = \underline{0} \;,$$

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

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

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

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

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

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

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

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

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

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

Tabelle 9.3.f: Computer-Ausgabe von `regr` für das Beispiel der gehemmten Reproduktion g **Devianz.** Für die logistische Regression wurde die Likelihood, die mit der Anpassung der Modell-Parameter erreicht wird, mit einer maximalen Likelihood verglichen, und das lässt sich auch in den andern Verallgemeinerten Linearen Modellen tun. Die maximale Likelihood entsteht, indem ein maximales Modell angepasst wird, das für jede Beobachtung i den am besten passenden kanonischen Parameter $\widetilde{\theta}_i$ bestimmt. Die Devianz ist allgemein definiert als

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

wobei $\underline{y}$ der Vektor aller beobachteten Werte ist und $\underline{\widehat{\mu}}$ der Vektor der zugehörigen angepassten Erwartungswerte. Der Teil der Log-Likelihood-Funktion, der nicht von θ abhängt, fällt dabei weg. In der Formel ist $\widetilde{\theta}_i$ der Parameter, der am besten zu y_i passt. Er ist jeweils bestimmt durch $y_i = \mathcal{E}\langle Y_i \rangle = b'\left\langle \widetilde{\theta}_i \right\rangle$.

Ein Dispersions-Parameter ϕ lässt sich für das maximale Modell nicht mehr schätzen; man verwendet den geschätzten Wert des betrachteten Modells. Bei der Binomial- und der Poisson-Verteilung fällt dieses Problem weg, da $\phi = 1$ ist.

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

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

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

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

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

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

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

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

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

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

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

9.4 Übergrosse Streuung

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

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

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

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

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

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

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

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

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

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

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

9.5 Residuen-Analyse

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

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

  Wie erwähnt, haben diese Residuen verschiedene Varianzen.

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

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

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

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

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

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

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

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

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

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

b Die wichtigsten grafischen Darstellungen der Residuen-Analyse sind:

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

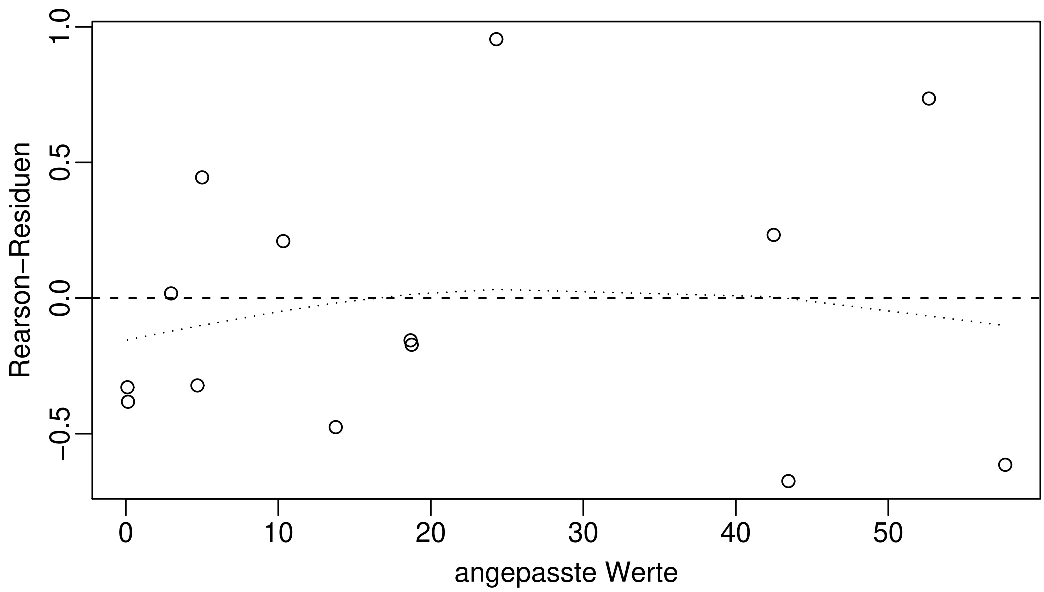


Abbildung 9.5.b: Tukey-Amscombe Plot zum Beispiel der Schiffs-Havarien

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 9.4.f.

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

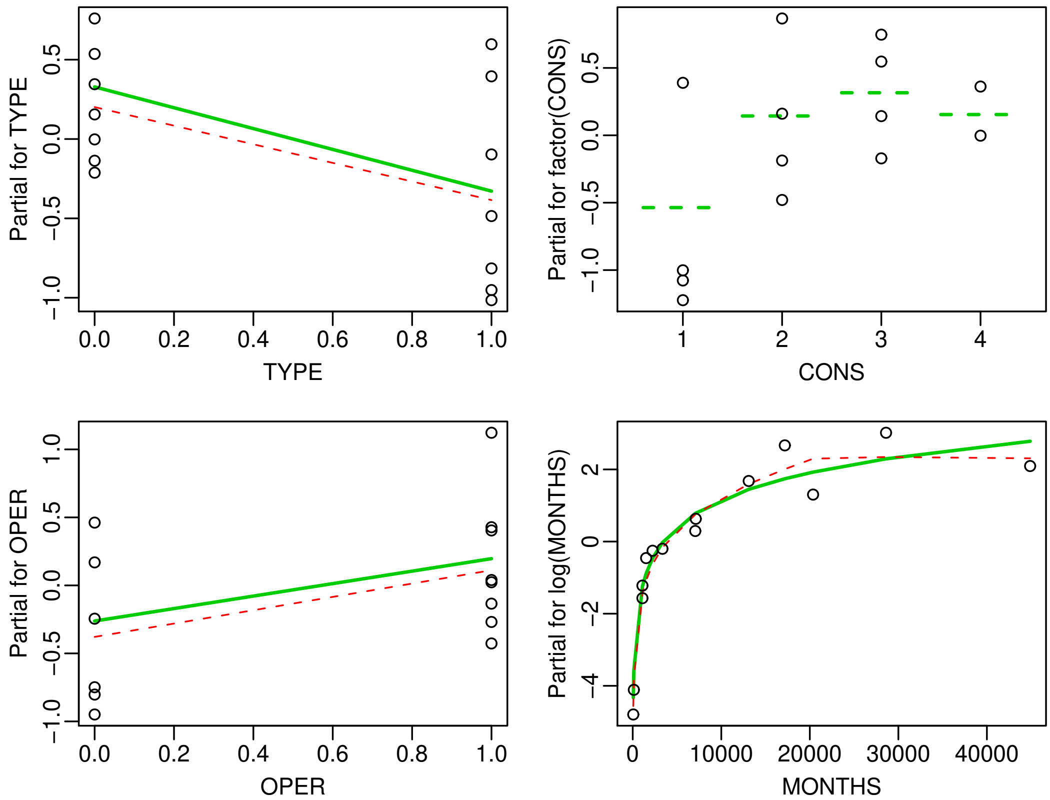


Abbildung 9.5.d: Partial residual Plots zu dem Havarie-Modell

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 8.4.k).

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

9.S S-Funktionen

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Damit ist der Scoring-Algorithmus festgelegt.

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

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

Die Lösung dieses Problems lautet

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Literaturverzeichnis

Agresti, A. (2002). *Categorical Data Analysis*, 2nd edn, Wiley, N.Y.

Agresti, A. (2007). *An Introduction to categorical data analysis*, Wiley Series in Probability & Math. Statistics, 2nd edn, Wiley, New York.

Bates, D. M. and Watts, D. G. (1988). *Nonlinear Regression Analysis and its Applications*, Wiley, N.Y.

Chambers, J. M. and Hastie, T. J. (1992). *Statistical Models in S*, Wadsworth & Brooks/Cole, Pacific Grove, Cal.

Chatterjee, S. and Price, B. (2000). *Regression Analysis By Example*, 3rd edn, Wiley, N.Y.

Christensen, R. (1990). *Log-linear models*, Springer, N.Y.

Cleveland, W. S. (1994). *The Elements of Graphing Data*, 2nd edn, Hobart Press, Summit, New Jersey.

Clogg, C. C. and Shihadeh, E. S. (1994). *Statistical models for ordinal variables*, Sage, Thousand Oaks, CA.

Cohen, A. (1980). On the graphical display of the significant components in a two-way contingency table, *Communications in Statistics – Theory and Methods* **A9**: 1025–1041.

Collet, D. (1991, 1999). *Modelling binary data*, Chapman & Hall/CRC Press LLC, Boca Raton, Florida.

Collet, D. (1994). *Modelling Survival Data in Medical Research*, Texts in Statistical Science, Chapman and Hall, London.

Cook, R. D. and Weisberg, S. (1999). *Applied regression including computing and graphics*, Wiley, N.Y.

Cox, D. R. (1989). *Analysis of Binary Data*, 2nd edn, Chapman and Hall, London.

Cox, D. R. and Snell, E. J. (1981). *Applied Statistics*, Chapman and Hall, London.

Crowder, M. J., Kimber, A. C., Smith, R. L. and Sweeting, T. J. (1991). *Statistical Analysis of Reliability Data*, Chapman and Hall.

Daniel, C. and Wood, F. S. (1980). *Fitting Equations to Data*, 2nd edn, Wiley, N.Y.

Davies, P. (1995). Data features, *Statistica Neerlandica* **49**: 185–245.

Devore, J. L. (2004). *Probability and Statistics for Engineering and the Sciences*, 6th edn, Duxbury Press, Belmont, California.

Dobson, A. J. (2002). *An Introduction to Generalized Linear Models*, 2nd edn, Chapman and Hall, London.

Draper, N. and Smith, H. (1998). *Applied Regression Analysis*, 3rd edn, Wiley, N.Y.

Fahrmeir, L. and Tutz, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models*, 2nd edn, Springer-Verlag, New York.

Fox, J. (2002). *An R and S-Plus companion to applied regression*, Sage, Thousand Oaks, CA.

Fox, J. and Monette, G. (1992). Generalized collinearity diagnostics, *Journal of the American Statistical Association* **87**: 178–183.

Fuller, W. A. (1987). *Measurement Error Models*, Wiley, N.Y.

Haaland, P. D. (1989). *Experimental Design in Biotechnology*, Marcel Dekker, N.Y.

Hampel, F. (1974). The influence curve and its role in robust estimation, *Journal of the American Statistical Association* **69**: 383–393.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*, Wiley, N.Y.

Harrell, F. E. J. (2002). *Regression Modeling Strategies. With Applications to Linear Models, Logistic Regression, and Survival Analysis*, Springer Series in Statistics, Springer, NY. *Corrected second printing*

Hartung, J., Elpelt, B. und Klösener, K. (2002). *Statistik. Lehr- und Handbuch der angewandten Statistik*, 13. Aufl., Oldenbourg, München.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*, number 43 in *Monographs on Statistics and Applied Probability*, Chapman and Hall, London.

Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Series in Statistics, Springer-Verlag, New York.

Hocking, R. R. (1996). *Methods and Applications of Linear Models; Regression and the Analysis of Variance*, Wiley Series in Probability and Statistics, Wiley, N.Y.

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

Huber, P. J. (1964). Robust estimation of a location parameter, **35**: 73–101.

Huber, P. J. and Ronchetti, E. M. (2009). *Robust Statistics*, 2nd edn, Wiley.

Kalbfleisch, J. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*, 2nd edn, Wiley, N.Y.

Lindsey, J. K. (1995). *Modelling Frequency and Count Data*, number 15 in *Oxford Statistical Science Series*, Clarendon Press, Oxford.

Maronna, R. A., Martin, R. D. and Yohai, V. J. (2006). *Robust Statistics, Theory and Methods*, Wiley Series in Probility and Statistics, Wiley, Chichester, England.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd edn, Chapman and Hall, London.

Mosteller, F. and Tukey, J. W. (1977). *Data Analysis and Regression: A Second Course in Statistics*, Addison-Wesley, Reading, Massachusetts.

Myers, R. H., Montgomery, D. C. and Vining, G. G. (2001). *Generalized Linear Models. With Applications in Engineering and the Sciences*, Wiley Series in Probability and Statistics, Wiley, NY.

Pokropp, F. (1994). *Lineare Regression und Varianzanalyse*, Oldenbourg.

Rice, J. A. (2007). *Mathematical Statistics and Data Analysis*, 3rd edn, Duxbury Press, Belmont, California.

Ripley, B. D. (1996). *Pattern Recognition and Neural Networks*, Cambridge Univ. Press, Cambridge, UK.

Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust Regression & Outlier Detection*, Wiley, N.Y.

Ryan, T. P. (1997). *Modern Regression Methods*, Series in Probability and Statistics, Wiley, N.Y. *includes disk*

Sachs, L. (2004). *Angewandte Statistik*, 11. Aufl., Springer, Berlin.

Schlittgen, R. (2003). *Einführung in die Statistik. Analyse und Modellierung von Daten*, 10. Aufl., Oldenbourg, München. *schoen, inkl. Sensitivity und breakdown, einfache regr mit resanal*

Sen, A. and Srivastava, M. (1990). *Regression Analysis; Theory, Methods, and Applications*, Springer-Verlag, N.Y.

Stahel, W. A. (2000). *Statistische Datenanalyse: Eine Einführung für Naturwissenschaftler*, 3. Aufl., Vieweg, Wiesbaden.

Stahel, W. A. (2007). *Statistische Datenanalyse: Eine Einführung für Naturwissenschaftler*, 5. Aufl., Vieweg, Wiesbaden.

van der Waerden, B. L. (1971). *Mathematische Statistik*, 3. Aufl., Springer, Berlin.

Venables, W. N. and Ripley, B. D. (1997). *Modern Applied Statistics with S-Plus*, Statistics and Computing, 2nd edn, Springer, Berlin.

Weisberg, S. (2005). *Applied Linear Regression*, 3rd edn, Wiley, N.Y.

Wetherill, G. (1986). *Regression Analysis with Applications*, number 27 in *Monographs on Statistics and Applied Probability*, Chapmann and Hall, London.