Johannes Gutenberg-Universität Mainz

Fachbereich Mathematik

Multivariate Klassifikation in der Kraftfahrzeughaftpflichtversicherung

Diplomarbeit

von

Alexandra Franzmann

Themenstellerin:	Prof. Dr. Czado
Betreuerin:	Prof. Dr. Czado
Abgabetermin:	27. März 2001

Inhaltsverzeichnis

Kapitel 1

Einleitung

Wir beginnen die Einleitung mit einer Motivation aus Sicht eines Versicherungsunternehmens. Die Deregulierung des Versicherungsmarktes in Deutschland führte vor allem in der Kraftfahrzeugversicherung zu neuen Tarifen mit neuen Kriterien wie gefahrene Jahreskilometer, Garage, Geschlecht etc. Folgende Probleme stellen sich seitdem einem Versicherungsunternehmen: Ist die Einführung eines neuen Merkmals wie Garage sinnvoll? Wenn bei zwei Kriterien eine Ermäßigung der Prämie gerechtfertigt ist, sollen dann auch beide Merkmale zugleich eingesetzt werden? Als Beispiel dazu führen wir die unterschiedliche Handhabung der Rabatte aufgrund des Geschlechts (Frau) und aufgrund der Fahrleistung (geringe Jahreskilometer) an. Woran erkennt man einen Versicherungsnehmer mit geringem oder hohem Unfallrisiko? Im folgenden erörtern wir, warum Versicherungsunternehmen (VU) ihre Tarife differenzieren müssen, wenn andere das tun. Wir nehmen einen sehr vereinfachten Markt an mit zwei Typen von Versicherungsnehmern, Typ A mit geringem Unfallrisiko und Typ B mit hohem Unfallrisiko, sowie zwei VU I und II. VU I legt für alle Versicherungsnehmer eine gleiche Prämie π fest, während VU II von A die Prämie π_A und von B die Prämie π_B verlangt. Wenn jeder Versicherungsnehmer sich das VU auswählen darf, so besitzen die VU I und II nur dann positive Marktanteile, wenn $\pi_A < \pi < \pi_B$ gilt. Die guten Risiken werden eher von VU I zu II wechseln, so daß der Anteil von A in VU II steigt (Gegenselektion für VU I mangels Differenzierung). VU I muß die Prämie π erhöhen. Schließlich sind alle Versicherungsnehmer der Gruppe A in VU II, und für die letztliche Prämie gilt: $\pi > \pi_A$. Wenn außerdem $\pi < \pi_B$, so werden alle unfallträchtigen Versicherungsnehmer von VU II zu I wechseln. (Der Fall $\pi > \pi_B$ ist uninteressant.) Am Ende dieser Entwicklung gibt es zwei Möglichkeiten:
1) $\pi = \pi_B$: Alle Versicherungsnehmer vom Typ A sind bei VU II, was II einen Vorteil verschafft.
2) $\pi < \pi_B$: Alle Versicherungsnehmer vom Typ B sind bei VU I, alle vom Typ A bei VU II.
Diese Motivation zeigt die Notwendigkeit, ein Verfahren zu finden, das die wichtigen Merkmale zur Tarifierung aus einer Vielzahl von möglichen Merkmalen herausfiltert, potentielle Wechselwirkungen mit anderen Merkmalen erkennt und das vorhandene Datenmaterial möglichst gut anpaßt.
Da die Tarifbildung in der Kraftfahrzeugversicherung auf der Anzahl der Schäden fußt, bietet

sich als Ausgangsmodell für die Analyse dieser Zähldaten ein Regressionmodell mit Poissonverteilungsannahme an, ähnlich wie die Normalverteilungsannahme Grundlage für stetige Daten ist. Bewährt haben sich in den vergangenen drei Jahrzehnten die Analysen auf der Basis von verallgemeinerten linearen Modellen, zu denen u. a. die Regressionsmodelle mit Poissonverteilung gehören. Nun besitzt die Poissonverteilung die recht strenge Eigenschaft, daß die Varianz genauso groß ist wie der Erwartungswert ist. In der Praxis der Datenanalyse stellt sich jedoch oft heraus, daß die Varianz in den Daten (bei gegebenen Regressoren) größer als der Erwartungswert ist. Zwar werden die Parameter des Modells in dieser Situation noch korrekt geschätzt, wenn die Spezifizierung des Erwartungswerts richtig ist, doch die Varianz der Parameterschätzer wird unterschätzt. Das führt bei Hypothesentests auf die Signifikanz eines Regressors zu einem zu optimistischen Erkennen der Signifikanzen. Darum benötigen wir Tests, die die Gültigkeit eines Poissonmodells bei dem Verdacht auf Verletzung der Gleichheit von Varianz und Erwartungswert feststellen können. Falls ein derartiger Test zur Ablehnung der Poissonverteilungsannahme führt, ist es üblich, die Poissonverteilung in eine allgemeinere Familie von Verteilungen einzubetten und die Regressionsanalyse mit der allgemeineren Verteilungsfamilie fortzusetzen.
Diese Ideen stellen das Programm der Diplomarbeit dar.
Der erste Teil von Kapitel 2 führt in die Theorie der verallgemeinerten linearen Modelle mit Schwerpunkt auf dem Poissonmodell ein. Sodann beschreiben wir verschiedene Mechanismen, die eine für die Poissonverteilung zu große Varianz erzeugen, sowohl innerhalb der verallgemeinerten linearen Modelle als auch mit Hilfe der stochastischen Prozesse und stellen zwei Modellierungsansätze vor. Der folgende Teil des zweiten Kapitels ist ein theoretischer Schwerpunkt der Arbeit. In ihm behandeln wir die asymptotischen Verteilungen des Likelihood-Quotienten- und des Lagrangeschen Multiplikatoren-Tests unter verschiedenen Voraussetzungen. Es folgt Kapitel 3 mit einer Anwendung des Lagrangeschen Multiplikatoren-Tests, um den Zusammenhang zwischen Varianz und Erwartungswert zu überprüfen. Als Beispiele werden drei Teststatistiken entwickelt, mit denen wir die Gültigkeit eines Poissonmodells beurteilen können. Schließlich untersuchen wir in Kapitel 4 einen Datensatz mittels der zuvor entwickelten Verfahren und Diagnosekennzahlen.

Kapitel 2

Theoretische Grundlagen

2.1 Verallgemeinerte lineare Modelle

In diesem Abschnitt betrachten wir eine Klasse von statistischen Modellen, die eine Verallgemeinerung des klassischen linearen Modells sind. Diese sogenannten verallgemeinerten linearen Modelle (GLMe) beinhalten u.a. lineare und Poisson-Regression als Spezialfälle. Wir beschreiben im folgenden die Komponenten eines GLMs, eine Methode zum Schätzen der Parameter sowie kurz ihr asymptotisches Verhalten und ein Maß für die Anpassungsgüte.

2.1.1 Datenstruktur

Wir betrachten die übliche Regressionsstruktur, bei der eine univariate Variable y, *die Zielvariable*, durch einen Vektor $\mathbf{x} = (x_1, \ldots, x_p)^T$ von *Regressoren* x_j, $j = 1, \ldots, p$, erklärt werden soll. Andere, geläufige Bezeichnungen für y sind abhängige Variable oder Response, während x_j auch erklärende Variable, Kovariable oder unabhängige Variable heißt. Die Zielvariable kann eine stetige, reelle Variable (wie im linearen Modell), nichtnegativ, eine Zählvariable oder binär sein. Die Regressoren können sowohl als metrische wie auch als qualitative (geordnete oder ungeordnete kategorische) Variablen oder gemischt vorliegen. Wir lassen deterministische und stochastische Regressoren zu, solange in den einzelnen Abschnitten keine Angaben dazu gemacht werden. Desweiteren seien n Beobachtungen $(y_i, \mathbf{x_i})$ von $(y, \mathbf{x})$ gegeben. Wir fassen die y_i als eine Realisierung der Zufallsvariablen Y_i auf, die gegeben $\mathbf{x_i}$ unabhängig verteilt sind mit Erwartungswert μ_i.

2.1.2 Komponenten des GLMs

(i) **zufällige Komponente**
Die Verteilung von Y_i gehört zur exponentiellen Familie, die die allgemeine Gestalt hat

$$f_{Y_i}(y_i; \theta_i, \phi) = \exp\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\} \tag{2.1}$$

mit spezifischen Funktionen $a(.)$, $b(.)$ und $c(.)$. Dabei heißen θ_i kanonischer Parameter und ϕ Skalenparameter.

Definition und Beispiel 2.1 (Poissonverteilung als Mitglied der exponentiellen Familie) *Y sei eine Zufallsvariable mit Werten in $\mathbb{N}_0$. Y hat genau dann eine Poissonverteilung mit Parameter $\lambda, \lambda > 0$, wenn die Wahrscheinlichkeitsfunktion lautet*

$$f_Y(y;\lambda) = \frac{\lambda^y}{y!}e^{-\lambda}.$$

Wir schreiben: $Y \sim Poi(\lambda)$.
Durch Umformen der Wahrscheinlichkeitsfunktion $f_Y(y;\lambda) = \exp\{y\ln\lambda - \lambda - \ln y!\}$ und mittels der Festlegung $\theta := \ln\lambda$, $\phi := 1$ sowie $a(\phi) = 1$, $b(\theta) = \exp\theta$, $c(y,\phi) = -\ln y!$ ist gezeigt, daß die Poissonverteilung zur exponentiellen Familie gehört.

Während die Funktion $c(.)$ nur der Normalisierung dient, ist vor allem die Funktion $b(.)$ zur Charakterisierung der Verteilungseigenschaften wichtig. Wir schreiben $\ell(\theta_i, \phi; y_i) := \ln f_{Y_i}(y_i;\theta_i,\phi)$ für die log-Likelihoodfunktion, die wir als Funktion von θ_i und ϕ betrachten, wenn y_i gegeben ist. Der Erwartungswert μ_i und die Varianz von Y_i können leicht aus der bekannten Formel (für einen Beweis siehe z. B. Casella/Berger [1990, S. 309])

$$E\left(\frac{\partial\ell}{\partial\theta_i}\right) = 0 \tag{2.2}$$

und der Informationsgleichung (s. Casella/Berger [1990, S. 312])

$$E\left(\frac{\partial^2\ell}{\partial\theta_i^2}\right) + E\left(\frac{\partial\ell}{\partial\theta_i}\right)^2 = 0 \tag{2.3}$$

hergeleitet werden.
Wir erhalten aus (2.1)

$$\ell(\theta_i,\phi;y_i) = \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)$$

sowie

$$\frac{\partial\ell}{\partial\theta_i} = \left(y_i - b'(\theta_i)\right)/a(\phi) \tag{2.4}$$

und

$$\frac{\partial^2\ell}{\partial\theta_i^2} = -b''(\theta_i)/a(\phi). \tag{2.5}$$

Aus (2.2) und (2.4) ergibt sich

$$0 = E\left(\frac{\partial\ell}{\partial\theta_i}\right) = \left(\mu_i - b'(\theta_i)\right)/a(\phi)$$

und damit der Erwartungswert von Y_i gegeben $\mathbf{x_i}$

$$E(Y_i) = \mu_i = b'(\theta_i). \tag{2.6}$$

Analog ermitteln wir mit (2.3) – (2.5) zunächst

$$0 = E\left(-\frac{b''(\theta_i)}{a(\phi)}\right) + E\left[\left(\frac{Y_i - b'(\theta_i)}{a(\phi)}\right)^2\right] = -\frac{b''(\theta_i)}{a(\phi)} + \frac{E(Y_i - E(Y_i))^2}{a^2(\phi)} = -\frac{b''(\theta_i)}{a(\phi)} + \frac{VarY_i}{a^2(\phi)}$$

und dann die Varianz von Y_i

$$VarY_i = b''(\theta_i)a(\phi). \tag{2.7}$$

Der zweite Term auf der linken Seite von (2.3) ist gerade die Fisher-Information, so daß wir die Varianz von Y_i auch als Vielfaches der Fisher-Information auffassen können.
Wir sehen in (2.7), daß die Varianz das Produkt aus zwei Funktionen ist. Die erste Funktion, $b''(\theta_i)$, hängt ausschließlich vom kanonischen Parameter und folglich vom Erwartungswert ab. Wir nennen sie die *Varianzfunktion* und schreiben, indem wir sie als Funktion von μ_i betrachten, $V(\mu_i)$.

Beispiel 2.2 (Erwartungswert und Varianz der Poissonverteilung als Funktion des kanonischen Parameters) *Wir bestimmen den Erwartungswert und die Varianz einer Poisson-verteilten Zufallsvariable Y mit Hilfe der Formeln (2.6) und (2.7). Es gilt:* $b(\theta) = \exp\theta$ *und* $a(\phi) = 1$.
Daraus folgt $E(Y) = b'(\theta) = \exp\theta$ *und* $VarY = b''(\theta)a(\phi) = \exp\theta$.

Die Funktion $a(\phi)$ besitzt häufig die Gestalt $a(\phi) = \phi/\omega_i$, wobei ω_i ein bekanntes Gewicht ist. Bei ungruppierten Daten, d.h. jede Beobachtung entspricht genau einer Einheit (z. B. Individuum), wählen wir $\omega_i = 1$. Der Fall $\omega_i \neq 1$ tritt bei gruppierten Daten ein: Wenn bei mehreren Beobachtungen $(y_i, \mathbf{x_i})$ die Werte der Regressoren identisch sind, fassen wir die Beobachtungen so zusammen, daß nur noch verschiedene Regressorkombinationen verbleiben. Zusätzlich ermitteln wir pro Regressorkombination die Anzahl der beobachteten Wiederholungen m_k sowie das arithmetische Mittel der ursprünglich individuellen Zielvariablen. In einem neuen Regressionsmodell mit dem arithmetischen Mittel als Zielvariable sind die Gewichte $\omega_k = m_k$. Anstelle des arithmetischen Mittels können wir als neue Zielvariable auch die Summe der einzelnen Zielvariablen in einer Regressorkombination betrachten. Dann ergeben sich die Gewichte $\omega_k = 1/m_k$.

(ii) **systematische Komponente**
Die Regressoren werden durch einen *linearen Prädiktor* $\eta_i = \mathbf{x}_i^T\beta = \sum_{j=1}^p x_{ij}\beta_j$ in das Modell eingeführt, wobei $\beta = (\beta_1, \ldots, \beta_p)^T$ ein Vektor mit unbekannten Parametern ist, die aus den Daten geschätzt werden sollen. In obiger Formel bedeutet x_{ij}, daß es sich um den Wert des j-ten Regressors bei der i-ten Beobachtung handelt.
Zu dem Vektor $\mathbf{x}_i$ ist das gleiche zu sagen, was bereits im linearen Modell gilt: Häufig wird eine Konstante, der sogenannte *Intercept*, hinzugefügt, so daß der Regressorvektor die Gestalt $(1, \mathbf{x}_i)^T$ annimmt. Metrische Regressoren können auch nichtlineare Transformationen $f(z_l)$ der „zugrundeliegenden" Variablen z_l sein. Ein kategorieller Regressor muß zuvor als Dummyvektor umkodiert werden. Wichtig bei dem Ansatz der GLMe ist die Linearität in β.

(iii) **Verbindung zwischen zufälliger und systematischer Komponente**
Zuletzt müssen wir in unserem Modell die zufällige Komponente, d.h. die Verteilung, mit der systematischen, dem Prädiktor, verbinden. Das geschieht mit Hilfe einer monotonen, differenzierbaren Funktion g, die den Erwartungswert als Funktion des Prädiktors auffaßt:

$$g(\mu_i) = \eta_i \qquad \text{bzw.} \qquad \mu_i = g^{-1}(\eta_i).$$

g heißt *Linkfunktion*.
Für jede Verteilung der exponentiellen Familie gibt es eine spezielle Linkfunktion, die die Existenz einer suffizienten Statistik garantiert. Diese *kanonische Linkfunktion*, wie sie genannt wird, ist durch die Bedingung $\eta_i = \theta_i$ definiert.

Beispiel 2.3 (kanonische Linkfunktion der Poissonverteilung) *Für das Regressionsmodell mit Poissonverteilung ist die log-Linkfunktion $g(\mu) = \ln\mu$ die kanonische Linkfunktion, weil aus $b(\theta) = \exp\theta$ folgt $\mu = b'(\theta) = \exp\theta$, so daß die Bedingung $\eta = \theta$ erfüllt ist, wenn $\eta = \ln\mu(= \ln(\exp\theta) = \theta)$.*

Zusammenfassend halten wir fest, daß ein GLM vollständig charakterisiert ist durch die drei Komponenten

- den Typ der exponentiellen Familie
- den Vektor mit den Regressoren (Designvektor)
- die Linkfunktion.

2.1.3 ML-Schätzung der Parameter

Nach Einführung des GLMs besteht die Regressionsanalyse darin, daß wir die Parameter β_j, $j = 1, \ldots, p$, schätzen müssen. Da wir annehmen, die zugrundeliegende Verteilung vollständig und korrekt zu kennen, liegt es nahe, die Maximum Likelihood-Methode (ML-Methode) zu benutzen. Zur Vollständigkeit wiederholen wir kurz diese Methode.

Definition 2.4 (Maximum Likelihood) *Wir nehmen an, daß eine Zufallsvariable Y, gegeben den Designvektor $\mathbf{x}$ und den Parametervektor ϑ, die (stetige oder diskrete) Dichte $f(y|\mathbf{x},\vartheta)$ besitzt. Diese Dichte, wenn wir sie als Funktion der Parameter bei festem y auffassen, nennen wir* Likelihoodfunktion *und schreiben*

$$L(\vartheta) = f(y|\mathbf{x},\vartheta) \ .$$

Liegt nun eine Stichprobe mit n unabhängigen Realisierungen vor, so hat die Likelihoodfunktion die Gestalt eines Produkts

$$L(\vartheta) = \prod_{i=1}^{n} f(y_i|\mathbf{x},\vartheta) \ .$$

Das ML-Prinzip zur Konstruktion eines Schätzers $\hat{\vartheta}$ beruht auf der Maximierung dieser Funktion

$$L(\hat{\vartheta}) = \max_{\vartheta} L(\vartheta).$$

Die Idee der Likelihood-Methode läßt sich anschaulich für diskrete Dichten erklären. Die Likelihood-Methode wählt denjenigen Wert von ϑ als Schätzer aus, für den die Wahrscheinlichkeit, daß die gegebenen Werte $y_1, \ldots, y_n$ angenommen werden, am größten ist (s. Fahrmeir/Künstler/Pigeot/ Tutz [1997, S. 379]).

Um ML-Schätzer für die β_j zu erhalten, setzen wir ferner voraus, daß der Skalenparameter ϕ bekannt ist und daß die Matrix $X = (\mathbf{x}_1^T, \ldots, \mathbf{x}_n^T)^T$ mit $\mathbf{x}_i^T = (x_{i1}, \ldots, x_{ip})$ vollen Spaltenrang p besitzt. Maximierung der Likelihoodfunktion einer Beobachtung i ist gleichbedeutend mit der Maximierung ihres oft leichter handhabbaren Logarithmus ℓ_i, da das Logarithmieren eine streng monotone Transformation ist:

$$\ell_i(\theta_i) = \ln f_Y(y_i; \theta_i, \phi) = \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi).$$

Für die gesamte Stichprobe ergibt sich die Summe $\ell = \sum_{i=1}^n \ell_i$.
Wir bestimmen die ML-Schätzer $\hat{\beta}_j$ von β_j, indem wir die partiellen Ableitungen von ℓ_i nach β_j mit Hilfe der Kettenregel bilden und gleich Null setzen:

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^n \frac{\partial \ell_i}{\partial \beta_j} = \sum_{i=1}^n \frac{\partial \ell_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \beta_j} = \sum_{i=1}^n \frac{y_i - \mu_i}{a(\phi)}\frac{1}{V(\mu_i)}\frac{\partial \mu_i}{\partial \beta_j} = 0 \qquad j = 1, \ldots, p \tag{2.8}$$

Beispiel 2.5 (ML-Schätzung bei der Poissonverteilung) *Wir nehmen an, daß $y_1, \ldots, y_n$ Realisierungen Poisson-verteilter Zufallsvariablen gegeben die zugehörigen Regressoren $x_{i1}, \ldots, x_{ip}$, $i = 1, \ldots, n$, sind. Die Matrix $X = (\mathbf{x}_1^T, \ldots, \mathbf{x}_n^T)^T$ besitze Rang p. Desweiteren wählen wir den kanonischen Link als Linkfunktion. Nun bestimmen wir den ML-Schätzer $\hat{\beta}$ von $\beta = (\beta_1, \ldots, \beta_p)^T$. Die log-Likelihoodfunktion lautet*

$$\ell(\beta) = \ln L(\beta) = \sum_{i=1}^n y_i \mathbf{x}_i^T\beta - \exp(\mathbf{x}_i^T\beta) - \ln y_i! \qquad \textit{mit } \mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T \ .$$

Somit löst $\hat{\beta}_j$ nach (2.8) die Gleichung

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^n (y_i - \exp\{\sum_{k=1}^p x_{ik}\beta_k\}) \left(\exp\{\sum_{k=1}^p x_{ik}\beta_k\}\right)^{-1} \exp\{\sum_{k=1}^p x_{ik}\beta_k\} x_{ij} = 0 \qquad j = 1, \ldots, p \tag{2.9}$$

und $\hat{\beta}$ das Gleichungssystem

$$\frac{\partial \ell}{\partial \beta} = \sum_{i=1}^n (y_i - \exp\{\mathbf{x}_i^T\beta\})\mathbf{x_i} = \mathbf{0} \ .$$

Nun ist aber das Kriterium, Nullstelle der ersten partiellen Ableitung zu sein, nicht hinreichend die Bestimmung eines Maximums. Aufschluß verschaffen uns die zweiten partiellen Ableitungen

$$\frac{\partial^2 \ell}{\partial \beta_j \partial \beta_r} = -\sum_{i=1}^n \exp\{\sum_{k=1}^p x_{ik}\beta_k\} x_{ir} x_{ij} \qquad j, r = 1, \ldots, p \tag{2.10}$$

Daraus ergibt sich die Hesse-Matrix H

$$H(\beta) = \frac{\partial^2 \ell}{\partial\beta\partial\beta'} = -\sum_{i=1}^{n} \exp(\mathbf{x}_i^T\beta)\mathbf{x}_i\mathbf{x}_i^T.$$

Da H negativ definit ist, ist ℓ global konkav, so daß der ML-Schätzer $\hat{\beta}$ als lokales Maximum von ℓ mit dem globalen Maximum übereinstimmt.

Wie im obigen Beispiel sind auch im allgemeinen die Likelihoodgleichungen (2.8) nichtlinear in β, weswegen es gewöhnlich keine analytische Lösung für $\hat{\beta}$ gibt, sondern die Gleichungen numerisch mit Hilfe eines Iterationsalgorithmus (z. B. der Newton-Raphson-Methode, s. Fahrmeir/Tutz [1994, S. 40 f] und für weitere Referenzen McCullagh/Nelder [1989, S. 43]) gelöst werden müssen.

2.1.4 Asymptotische Eigenschaften des ML-Schätzers

Eine weitere Konsequenz, die sich aus dem Fehlen einer analytischen Lösung für $\hat{\beta}$ ergibt, ist, daß wir exakte Ergebnisse über die Verteilung von $\hat{\beta}$ nur schwer erhalten. Darum basieren die Folgerungen auf asymptotischen Ergebnissen.
Zunächst geben wir die „Regularitätsbedingungen“ an, unter denen Konsistenz und asymptotische Normalverteilung des ML-Schätzers gelten (s. Cameron/Trivedi [1998, S. 23 f.]):

(i) Die Wahrscheinlichkeitsfunktion oder Dichte $f(y, \mathbf{x}, \beta)$ ist vollständig bekannt und injektiv in β.

(ii) Der Parameterraum Θ mit $\beta \in \Theta$ ist endlichdimensional, abgeschlossen und kompakt.

(iii) Die ersten drei Ableitungen von ℓ nach β existieren und sind stetig und beschränkt.

(iv) Die Reihenfolge von Differentiation und Integration der Likelihoodfunktion kann vertauscht werden.

(v) Die Regresorvektoren $\mathbf{x}_i, i = 1, \ldots, n$, genügen den Bedingungen

(a) $\mathbf{x}_i^T\mathbf{x}_i < \infty$

(b) $\frac{E(w_i^2)}{\sum_{i=1}^n E(w_i^2)} = 0 \qquad$ mit $w_i = \mathbf{x}_i^T \frac{\partial\ell}{\partial\beta}$

(c) $\lim_{n\to\infty} \frac{\sum_{i=1}^n E(w_i^2|\Omega_{i-1})}{\sum_{i=1}^n E(w_i^2)} = 1 \qquad$, wobei $\Omega_{i-1} = (\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_{i-1}})^T$

Die erste Bedingung sichert, daß ℓ ein eindeutiges Maximum hat. Die zweite Bedingung schließt mögliche Probleme an den Rändern von Θ aus und kann gelockert werden, falls z. B. ℓ global konkav ist. Bei der dritten Bedingung reicht oft schon die Existenz der Ableitungen bis zur zweiten Ordnung. Die vierte Bedingung ist eine Schlüsselbedingung, die solche Wahrscheinlichkeitsfunktionen und Dichten ausschließt, bei denen der Wertebereich der y_i von β abhängt. Die letzte Bedingung sorgt dafür, daß nur Beobachtungen eingehen, deren Anteil an der Likelihood

nicht zu groß ist.
Im folgenden skizzieren wir die Beweisidee zur Konsistenz und asymptotischen Normalverteilung des ML-Schätzers, wie sie sich in Winkelmann [1996, S. 61] befindet. Der ausführliche Beweis steht z. B. in Cox/Hinkley [1973, S. 288 f, S. 294]. Wir nehmen an, daß die Dichte im Regressionsmodell korrekt spezifiziert ist, d. h. es gibt einen wahren Parameterwert β_0, für den der datenerzeugende Prozeß der y_i die Dichte $f(y_i|\mathbf{x}_i, \beta_0)$ hat. Durch eine multivariate Taylorentwicklung des Gradienten $\frac{\partial \ell}{\partial \beta}$, den wir mit $g(.)$ bezeichnen, im Punkt β_0 erhalten wir

$$g(\hat{\beta}) \approx g(\beta_0) + H(\beta_0)(\hat{\beta} - \beta_0) \qquad \text{mit} \quad H(\beta) = \frac{\partial g}{\partial \beta'} = \frac{\partial^2 \ell}{\partial \beta \partial \beta'} \ .$$

Diese Gleichung formen wir wegen $g(\hat{\beta}) = \mathbf{0}$ nach (2.8) um zu

$$\sqrt{n}(\hat{\beta} - \beta_0) \approx \left(-\frac{1}{n} H(\beta_0)\right)^{-1} \frac{1}{\sqrt{n}} g(\beta_0) \ . \tag{2.11}$$

Wir betrachten nun die beiden Terme auf der rechten Seite. Bei dem Term in der Klammer handelt es sich um ein Mittel, so daß wir das Gesetz der großen Zahlen darauf anwenden und die Unabhängigkeit der Stichprobe ausnutzen:

$$-\frac{1}{n} H(\beta_0) = -\frac{1}{n} \sum_{i=1}^{n} H_i(\beta_0) \quad \xrightarrow{p} \quad \mathcal{I} \ ,$$

wobei H_i die Hesse-Matrix der i-ten Beobachtung und $\mathcal{I} = -E(\frac{\partial^2 \ell}{\partial \beta \partial \beta'})$ die Fisher-Informationsmatrix (näheres s. Anhang A) ist. Auf den zweiten Term der rechten Seite von (2.11) wenden wir Liapunovs Zentralen Grenzwertsatz (s. Rao [1973, S. 127]) an und bekommen

$$\frac{1}{\sqrt{n}} g(\beta_0) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathcal{I}) \ .$$

Bei den beiden Grenzübergängen haben wir stillschweigend die Informationsgleichung

$$E(\frac{\partial^2 \ell}{\partial \beta_i \partial \beta_j}) = -E(\frac{\partial \ell}{\partial \beta_i} \frac{\partial \ell}{\partial \beta_j}) \qquad i, j = 1, \ldots, p \tag{2.12}$$

ausgenutzt, die aus der dritten und vierten Regularitätsbedingung folgt. Schließlich ergibt sich

$$\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathcal{I}^{-1}) \ .$$

Damit haben wir gezeigt, daß $\hat{\beta}$ ein konsistenter Schätzer für β_0 ist und daß $\hat{\beta}$ asymptotisch normalverteilt ist. Außerdem besitzt der ML-Schätzer die wünschenswerte Eigenschaft, daß seine asymptotische Kovarianzmatrix die untere Cramer-Rao-Schranke erreicht (s. Cox/Hinkley [1974, S. 304]) und somit asymptotisch effizient ist.

2.1.5 Goodness of fit-Maße

Devianz

Im vorangegangenen Abschnitt haben wir eine Methode zur Schätzung der Parameter kennengelernt sowie einige wichtige Eigenschaften dieser Schätzer. Wir benötigen jetzt ein Maß, das uns

anzeigt, wie gut die gefundenen Schätzer (und damit das Modell) an die Datenstruktur angepaßt sind. Für ein solches Maß verwenden wir erneut die log-Likelihoods.
Um die grundlegende Idee darzustellen, lösen wir uns von unseren gegebenen Regressoren. Wenn uns n Beobachtungen vorliegen, können wir beliebige Modelle mit bis zu n Parametern daran anpassen. Da einfachste Modell, das *Nullmodell*, besitzt einen einzigen Parameter μ, der zu einem gemeinsamen geschätzten Wert $\hat{\mu}$ für alle Beobachtungen führt. Damit ordnet das Nullmodell die gesamte Variation zwischen den Beobachtungen der zufälligen Komponente zu. Das andere Extrem ist das *volle Modell* mit n Parametern. Wir können im vollen Modell die Parameter durch Lösen eines linearen Gleichungssystems mit n Gleichungen und n Parametern bestimmen. Auf diese Weise ordnet das volle Modell die gesamte Variation in den Beobachtungen der systematischen Komponente zu und überläßt keine der zufälligen Komponente. In der Praxis ist das Nullmodell zu einfach und das volle Modell nicht aussagekräftig genug, denn es faßt die Daten nicht in ihrer wesentlichen Struktur zusammen, sondern gibt sie vollständig wieder. Trotzdem dient uns das volle Modell als Ausgangspunkt, um die Diskrepanz zu einem Zwischenmodell mit p Parametern zu messen.
Wir setzen im folgenden voraus, daß der Skalenparameter ϕ bekannt ist oder zumindest fest gewählt. Mit $\ell(\hat{\mu})$ bezeichnen wir die log-Likelihood des zu untersuchenden Modells mit den p Parametern, während $\ell(\mathbf{y})$ die log-Likelihood des vollen Modells bezeichnet, denn $\mathbf{y}$ ist im vollen Modell derjenige Schätzer, der die log-Likelihood maximiert. Nun können wir ein Maß für die Anpassungsgüte definieren.

Definition 2.6 (Devianz) *Die skalierte Devianz D^* ist definiert durch*

$$D^*(\mathbf{y}, \hat{\mu}, \phi) = -2[\ell(\hat{\mu}) - \ell(\mathbf{y})] \ .$$

Somit ist die skalierte Devianz eine Funktion von $\mathbf{y}$ und $\hat{\mu}$, die durch $a(\phi)$ geteilt wird. Üblicherweise multiplizieren wir D^* mit dem Faktor $a(\phi)$ und erhalten die Devianz D:

$$D(\mathbf{y}, \hat{\mu}) = -2a(\phi)[\ell(\hat{\mu}) - \ell(\mathbf{y})] \ .$$

Beispiel 2.7 (Devianz im Poissonmodell) *In Beispiel 2.5 haben wir die log-Likelihood im Poissonmodell bestimmt. Da wir aus Beispiel 2.2 wissen, daß für die Poissonverteilung $a(\phi) = 1$ gilt, stimmen die Devianz D und die skalierte Devianz D^* überein. Wir erhalten*

$$\begin{aligned} D(\mathbf{y}, \hat{\mu}) &= D^*(\mathbf{y}, \hat{\mu}) = -2[\ell(\hat{\mu}) - \ell(\mathbf{y})] \\ &= -2\{\sum_{i=1}^{n}(y_i \ln \hat{\mu}_i - \hat{\mu}_i - \ln y_i!) - (y_i \ln y_i - y_i - \ln y_i!)\} \\ &= 2\{\sum_{i=1}^{n} y_i \ln(y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i)\}, \end{aligned} \tag{2.13}$$

wobei $y_i \ln y_i = 0$ gesetzt wird, wenn $y_i = 0$.
Bei der kanonischen Linkfunktion $\mu_i = \exp(\mathbf{x}_i^T \beta)$ mit Regressoren, die einen Intercept enthalten, $\mathbf{x}_i^T = (1, x_{i2}, \ldots, x_{ip})$ vereinfacht sich die Devianz. Denn für $i = 1, \ldots, n$ gilt $\frac{\partial \mu_i}{\partial \beta_1} = \mu_i$ und für

den ML-Schätzer $\hat{\mu}_i = \exp(\mathbf{x}_i^T \hat{\beta})$ *nach (2.9):* $\frac{\partial \ell}{\partial \beta_1} = \sum_{i=1}^n (y_i - \hat{\mu}_i) = 0$. *Wir können also den letzten Term in (2.13) wegfallen lassen und erhalten*

$$D(\mathbf{y}, \hat{\mu}) = 2 \sum_{i=1}^n y_i \ln(y_i/\hat{\mu}_i).$$

Wir bemerken, daß die Konstante 2 in der Definition der Devianz lediglich der Normalisierung dient, damit im Falle der Normalverteilung die Devianz mit der Residuenquadratsumme übereinstimmt. Dadurch ist sofort klar, daß die Devianz eine χ^2-Verteilung mit $n-p$ Freiheitsgraden bei Modellen mit Normalverteilung hat. Bei anderen Verteilungen läßt sich zeigen (s. Satz 2.24), daß die Devianz asymptotisch χ^2_{n-p}-verteilt ist (selbstverständlich nur unter der Annahme, daß das Modell gilt).
Die Devianz sollte nicht als absolutes Maß für die Anpassung eines einzelnen Modells dienen, sondern zum Vergleich zwischen zwei geschachtelten Modellen herangezogen werden. Wir möchten beispielsweise testen, ob die Hinzunahme eines weiteren Regressors die Anpassung signifikant verbessert. Bezeichne der Vektor $\hat{\mu}_0$ die angepaßten Werte des zu testenden Modells und $\hat{\mu}_A$ die angepaßten Werte des erweiterten Modells mit dem zusätzlichen Regressor. Wir messen die Reduktion in der Devianz durch $D(\mathbf{y}, \hat{\mu}_0) - D(\mathbf{y}, \hat{\mu}_A) = 2\,[\ell(\hat{\mu}_A) - \ell(\hat{\mu}_0)]$. Diese Statistik ist asymptotisch χ^2_1-verteilt, wobei die χ^2-Approximation der Devianzdifferenz meist erheblich genauer ist, selbst wenn sie für $D(\mathbf{y}, \hat{\mu}_0)$ oder $D(\mathbf{y}, \hat{\mu}_A)$ ungenau ist.

Pearson χ^2

Ein weiteres Maß für die Gesamtanpassung eines Regressionsmodells ist die intuitive Pearson-Statistik.

Definition 2.8 (Pearson-Statistik) *Die Pearson-Statistik lautet*

$$P = \sum_{i=1}^n \frac{(y_i - \hat{\mu}_i)^2}{\hat{\omega}_i},$$

wobei $\hat{\mu}_i$ *Schätzer von* μ_i *und* $\hat{\omega}_i$ *Schätzer von* $Var\, Y_i$ *sind.*

Anhand dieser allgemeinen Darstellung erkennen wir, daß dieses Maß für beliebige Regressionsmodelle geeignet ist, solange nur der Erwartungswert und die Varianz bekannt sind. Im Gegensatz dazu beschränkt sich die Devianz auf parametrische Regressionsmodelle, denn die Likelihood muß definiert sein. Wenn der Erwartungswert und die Varianz korrekt spezifiziert sind, gilt $E(\sum_{i=1}^n \frac{(Y_i-\mu_i)^2}{Var\, Y_i}) = n$ wegen $E\frac{(Y_i-\mu_i)^2}{Var\, Y_i} = 1$. Da wir μ_i und $Var\, Y_i$ aber schätzen müssen, nehmen wir eine Korrektur der Freiheitsgrade vor und vergleichen P mit $(n-p)$. Offensichtlich besitzt P bei Modellen mit Normalverteilung eine exakte χ^2_{n-p}-Verteilung und ist wiederum die Residuenquadratsumme. Bei anderen Verteilungen müssen wir die Daten zunächst so gruppieren, daß alle verbleibenden μ_i verschieden sind, um mit Hilfe des Zentralen Grenzwertsatzes die asymptotische χ^2_{n-p}-Verteilung zu erhalten, solange die relative Gruppengröße für wachsendes n konstant ist (s. Fahrmeir/Tutz [1994, S. 48]).

Beispiel 2.9 (Pearson χ^2 im Poissonmodell) *Aus Beispiel 2.2 wissen wir, daß im Poissonmodell gilt* $\operatorname{Var} Y_i = \mu_i$*, so daß die Pearson-Statistik die Gestalt annimmt*

$$P_{Poi} = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}.$$

Weil der Quotient auf der rechten Gleichungsseite die empirische Varianz mit dem empirischen Erwartungswert vergleicht, wird P_{Poi} *oft als Indikator für die theoretisch geforderte Gleichheit der beiden benutzt. Dabei interpretieren wir* $P_{Poi} > n - p$ *als Hinweis darauf, daß die Varianz im Modell größer als der Erwartungswert ist, und* $P_{Poi} < n - p$ *als Hinweis, daß die Varianz kleiner als der Erwartungswert ist. Daß die Pearson-Statistik nur eingeschränkt als ein solches Diagnosemittel taugt, sehen wir an folgender korrigierter Modellierung: Wir verallgemeinern die Poissonverteilung, indem wir den Skalenparameter* ϕ *aus (2.1) frei wählen, statt ihn wie in Beispiel 2.1 gleich* 1 *zu setzen, und wählen* $a(\phi) = \phi$*. Dann ist wegen (2.7) die Varianz ein Vielfachen vom Erwartungswert:* $\operatorname{Var} Y_i = \phi\mu_i$*. Verwenden wir nun als Schätzer* $\hat{\omega}_i = \hat{\phi}\hat{\mu}_i$ *mit* $\hat{\phi} = \frac{1}{n-p}\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2/\hat{\mu}_i$*, so ist die Pearson-Statistik immer gleich* $n - p$ *unabhängig von* ϕ*, womit sie als Indiz für die Gleichheit von Varianz und Erwartungswert untauglich geworden ist. Das Verhältnis von empirischer Varianz und empirischen Erwartungswert in Poissonmodellen wird im folgenden Abschnitt näher untersucht.*

2.2 Überdispersion

Bisher betrachteten wir GLMe ganz allgemein. In diesem Abschnitt wollen wir uns ausführlicher mit den Eigenschaften eines speziellen GLMs, und zwar des Poissonmodells, beschäftigen. Das Poissonmodell ist das einfachste Regressionsmodell für Zähldaten.
Die in Abschnitt 2.1 beschriebene Gleichheit von Erwartungswert und Varianz ist charakteristisch für die Poissonverteilung. Sie spielt eine entscheidende Rolle in der folgenden Diskussion und wird als *Äquidispersion* bezeichnet. Abweichungen von der Äquidispersion sind entweder *Überdispersion*, bei der die Varianz größer als der Erwartungswert ist, oder *Unterdispersion*, bei der die Varianz kleiner als der Erwartungswert ist. Im Gegensatz zu anderen Verteilungen ist die Verletzung der Äquidispersion eine hinreichende Bedingung, um die Poissonverteilungsannahme zu verletzen. Übertragen auf die Poissonregression bedeutet Überdispersion, daß bei gegebenen Regressoren die Varianz der Zielvariable ihren Erwartungswert übersteigt.
In vielen Untersuchungen erweisen sich die Daten variabler als durch das Poissonmodell vorhergesagt; es liegt Überdispersion vor. Im folgenden Abschnitt geben wir mögliche Ursachen der Überdispersion an und beschreiben sodann, wie wir die auftretende Überdispersion modellieren können.

2.2.1 Gründe für Überdispersion

Wir nehmen jetzt an, daß unsere Daten Überdispersion aufweisen. Dieses Phänomen tritt in unterschiedlicher Weise auf. Mögliche Mißspezifizierungen sind:

(i) Weitere Regressoren fehlen.

(ii) Die vorhandenen Regressoren gehen durch eine bislang unbekannte Transformation in den Prädiktor ein.

(iii) Der Prädiktor im GLM sieht nur die Linearität in den Parametern vor, während zusätzlich ein nichtlinearer Zusammenhang zwischen den Parametern und den Regressoren besteht. So geht z. B. im untersuchten Modell der Regressor x nur in der Form βx in den Prädiktor ein. Tatsächlich aber gibt es außerdem den unberücksichtigten Zusammenhang $e^{-\beta x}$.

(iv) Die Linkfunktion ist falsch gewählt. Beispielsweise können wir den Logarithmus als Linkfunktion in unserer Analyse gewählt haben, während die wahre Linkfunktion die Wurzel aus dem Erwartungswert ist.

(v) Die zugrundeliegende Beobachtungseinheit (Zeit, Volumen, Fläche, etc.) ist nicht fest, sondern zufällig.

Die ersten vier aufgelisteten Mißspezifizierungen können in beliebigen Regressionsmodellen auftreten, während die zuletzt genannte spezifisch für Poissonmodelle ist.
Um zu beweisen, daß die oben erwähnten Punkte zu Überdispersion führen, verdeutlichen wir, daß unser gewähltes Poissonmodell nicht zur vollständigen Erklärung der beobachteten Heterogenität ausreicht. Im folgenden Beweis modellieren wir die unbeobachtete Heterogenität als Zufallsvariable, von der die Zielvariable neben den Regressoren ebenfalls abhängt, und nennen sie U.

Lemma 2.10 (unbeobachtete Heterogenität) *Seien Y und U zwei Zufallsvariablen. $Y|U = u$ habe eine Poissonverteilung mit Erwartungswert $E(Y|u)$ und Varianz $Var(Y|u)$. Dann gilt:*

$$Var(Y) > E(Y).$$

Beweis: *Wir verwenden die bekannte Formel für die bedingte Varianz (s. Casella/Berger [1990, S. 158])*

$$Var(Y) = E(Var(Y|u)) + Var(E(Y|u)). \tag{2.14}$$

Dann nutzen wir die Gleichheit von bedingter Varianz und bedingtem Erwartungswert der Poissonverteilung, um (2.14) umzuformen zu

$$Var(Y) = E(E(Y|u)) + Var(E(Y|u)).$$

Den zweiten Term auf der rechten Seite in obiger Gleichung schätzen wir nach unten durch 0 *ab, und auf den ersten Term wenden wir die Formel für den bedingten Erwartungswert $E(Y) = E(E(Y|u))$ an, womit wir die behauptete Überdispersion gezeigt haben.*

Wir können eine Poissonverteilung auch mittels der Terminologie der stochastischen Prozesse beschreiben. Alle dazu notwendigen Begriffe werden im nachstehenden Exkurs eingeführt, die auf der Darstellung in Cox [1966, Kap. 1-3] beruht. Eine Einführung jüngeren Datums in stochastische Prozesse befindet sich in Resnick [1992].

Knappe Einführung in die Theorie der stochastischen Prozesse

Definition 2.11 (stochastischer Prozeß) *Ein stochastischer Prozeß $\{X(t), t \geq 0\}$ ist eine zeitindizierte Familie von reellwertigen Zufallsvariablen, die auf einem beliebigen Wahrscheinlichkeitsraum definiert sind.*

Definition 2.12 (Zählprozeß) *Ein stochastischer Prozeß $\{N(t), t \geq 0\}$ heißt Zählprozeß, wenn $N(t)$ die Anzahl aller Ereignisse, die bis zum Zeitpunkt t eintreten, bezeichnet.*

$N(t)$ ist nichtnegativ, ganzzahlig und besitzt die Eigenschaft, daß $N(s) \leq N(t)$ für $s < t$. Offensichtlich gibt $N(t) - N(s)$ die Anzahl der Ereignisse an, die im Intervall $]s, t]$ eintreten. Wir stellen uns den eindimensionalen Zählprozeß als eine Punktmenge auf der Zeitachse vor, die eine zufällige Folge von zu gewissen Zeitpunkten eintretenden Ereignissen darstellt.
Statt die Anzahl der Ereignisse in einem Intervall zu analysieren, kann man die Zeitspanne zwischen zwei aufeinanderfolgenden Ereignissen untersuchen. Wir transformieren somit den Zählprozeß $\{N(t), t \geq 0\}$ in eine Folge $\tau_k, k \in \mathbb{N}$, von Wartezeiten zwischen dem $(k-1)$-ten und dem k-ten Ereignis.

Definition 2.13 (Ankunftszeit und Wartezeit) *Die Ankuntszeit S_r bis zum r-ten Ereignis ist definiert durch $S_r = \inf\{t \geq 0 | N(t) = r\}$.*
Mit Hilfe der Ankunftszeit definieren wir die Wartezeit $\tau_k, k \in \mathbb{N}$: $\tau_k = S_k - S_{k-1}$ mit $S_0 := 0$.
Die zufällige, nichtnegative Folge $\{\tau_k, k \geq 1\}$ heißt Folge der Wartezeiten.
Damit können wir S_r auch schreiben als $S_r = \sum_{k=1}^{r} \tau_k, r \in \mathbb{N}$.

Aus der Definition von $N(t)$ und S_r ist sofort klar, daß der Zusammenhang

$$N(t) < r \qquad \Longleftrightarrow \qquad S_r > t$$

besteht. Bezeichne F_r die Verteilungsfunktion von S_r, so folgt

$$P(N(t) < r) = P(S_r > t) = 1 - F_r(t)$$

und weiter

$$P(N(t) = r) = P(N(t) < r + 1) - P(N(t) < r) = F_r(t) - F_{r+1}(t). \tag{2.15}$$

Schließlich führen wir noch eine zur Charakterisierung der Wartezeitverteilung wichtige Funktion ein.

Definition 2.14 (Ausfallrate) *Die Ausfallrate oder Hazardrate λ ist definiert als*

$$\lambda(t) = \lim_{dt \to 0+} \frac{P(t \leq \tau < t + dt | \tau > t)}{dt}. \tag{2.16}$$

Eine Anwendung des Satz von Bayes liefert die Darstellung $\lambda(t) = \frac{f(t)}{1-F(t)}$, wobei f die Dichte und F die Verteilungsfunktion der Wartezeit τ sind. Wir interpretieren $\lambda(t)dt$ als Wahrscheinlichkeit, daß bei bereits verstrichener Wartezeit t ein Ereignis im Intervall $[t, t+dt[$ eintritt.
Die Ausfallrate beinhaltet damit die zugrundeliegende Zeitabhängigkeit des Prozesses. Ist $\lambda(t)$ eine monoton wachsende Funktion von t, so sprechen wir von *positiver Alterung*. Sie bedeutet, daß das nächste Ereignis umso wahrscheinlicher im nächsten Augenblick eintritt, je länger die Wartezeit ist. Analog sprechen wir von *negativer Alterung*, falls die Ausfallrate eine monoton fallende Funktion ist. Dann tritt das nächste Ereignis umso unwahrscheinlicher gleich ein, je größer die Wartezeit ist. Bei einer konstanten Ausfallrate gibt es keine Alterung.

Beispiel 2.15 (Poisson-Prozeß) *Wir betrachten eine Folge von Wartezeiten $\{\tau_i\}$, wobei das Ende einer Wartezeit sofort zum Beginn einer neuen führt. Die τ_i seien unabhängig und identisch exponentiell verteilt mit Dichte $f_\tau(t) = \mu e^{-\mu t}$. $N(T)$ bezeichne die Anzahl der Ereignisse in $]0,T]$. Wir wollen die Verteilung der Ereignisse $N(t)$ bestimmen. Durch (2.15) ist die Verteilung mittels der Verteilungsfunktionen der Ankunftszeiten S_r gegeben. Andererseits ist S_r die Summe von iid Zufallsvariablen. Indem wir die Laplace-Transformierten verwenden, können wir zeigen, daß S_r eine Gammaverteilung besitzt mit Dichte $f_r(t) = \frac{\mu}{(r-1)!}(\mu t)^{r-1}e^{-\mu t}$ (s. Cox [1966, S. 17-21]). Die entsprechende Verteilungsfunktion lautet:*

$$F_r(t) = \int_0^t \frac{\mu}{(r-1)!}(\mu x)^{r-1}e^{-\mu x}\,dx \overset{u=\mu x}{=} \frac{1}{(r-1)!}\int_0^{\mu t} u^{r-1}e^{-u}\,du$$

Wegen $r \in \mathbb{N}$ können wir $(r-1)$-mal partiell integrieren, so daß sich F_r auch schreiben läßt als

$$F_r(t) = 1 - e^{-\mu t}\sum_{k=0}^{r-1}\frac{(\mu t)^k}{k!}.$$

Daraus erhalten wir die Verteilung von $N(t)$:

$$P(N(t) = r) = F_r(t) - F_{r+1}(t) = \frac{(\mu t)^r}{r!}e^{-\mu t},$$

die sich als Poisson-Verteilung mit Parameter μt entpuppt.
Wir ermitteln noch die Ausfallrate

$$\lambda(t) = \frac{f_\tau(t)}{1-F_\tau(t)} = \frac{\mu e^{-\mu t}}{1-(1-e^{-\mu t})} = \mu,$$

die sich als konstant herausstellt.

Demnach ist die Poisson-Verteilung angemessen, wenn die aufeinanderfolgenden Ereignisse während eines festen Zeitintervalls unabhängig voneinander und mit konstanter Ausfallrate eintreten. Die Verletzung der Unabhängigkeitsannahme kann ebenfalls zu Überdispersion führen, wie wir jetzt zeigen wollen. Die zu beobachtenden Ereignisse treten in Gruppen auf, deren Größe zufällig ist.

Satz 2.16 *Gegeben sei* $Y = Z_1 + Z_2 + \cdots + Z_N$ *mit* $Z_i \in \mathbb{N}_0$ *und iid. Die Anzahl der Summanden* N *sei Poisson-verteilt und unabhängig von* $Z_i \, \forall \, i = 1, \ldots, N$.
Wenn gilt $E({Z_i}^2) > E(Z_i)$ *und* $Var Z_i < \infty$, *so liegt Überdispersion vor.*
Beweis: *Wir bestimmen zunächst den Erwartungswert von* Y.
Mit $\mathcal{P}^{(N)}(s)$, $\mathcal{P}^{(Y)}(s)$ *bzw. mit* $\mathcal{P}^{(Z_i)}(s)$ *bezeichnen wir die wahrscheinlichkeitserzeugenden Funktionen (s. Anhang A) von* N, Y *bzw.* Z_i. *Laut Feller [1968, S.286f.] gilt*

$$\mathcal{P}^{(Y)}(s) = \mathcal{P}^{(N)}(\mathcal{P}^{(Z_i)}(s)) \tag{2.17}$$

Differenzieren obiger Formel nach s *mit der Kettenregel liefert:*

$$\mathcal{P}^{(Y)'}(s) = \mathcal{P}^{(N)'}(\mathcal{P}^{(Z_i)}(s)) \cdot \mathcal{P}^{(Z_i)'}(s).$$

Indem wir den Zusammenhang $(A.3)$ *ausnutzen, erhalten wir den Erwartungswert*

$$E(Y) = \mathcal{P}^{(Y)'}(1) = \mathcal{P}^{(N)'} \underbrace{(\mathcal{P}^{(Z_i)}(1))}_{=1} \cdot \mathcal{P}^{(Z_i)'}(1) = E(N) \cdot E(Z_i)$$

Als nächsten Schritt berechnen wir die Varianz von Y.
Durch zweimaliges Ableiten von (2.17) nach s *unter Hinzunahme der Produktregel ermitteln wir*

$$\mathcal{P}^{(Y)''}(s) = \mathcal{P}^{(N)''}(\mathcal{P}^{(Z_i)}(s)) \cdot [\mathcal{P}^{(Z_i)'}(s)]^2 + \mathcal{P}^{(N)'}(\mathcal{P}^{(Z_i)}(s)) \cdot \mathcal{P}^{(Z_i)''}(s).$$

Nun wenden wir die Gleichung $(A.4)$ *auf* Y *an:*

$$\begin{aligned}
Var Y &= \mathcal{P}^{(Y)''}(1) + \mathcal{P}^{(Y)'}(1) - [\mathcal{P}^{(Y)'}(1)]^2 \\
&= \mathcal{P}^{(N)''} \underbrace{(\mathcal{P}^{(Z_i)}(1))}_{=1} \left[\mathcal{P}^{(Z_i)'}(1)\right]^2 + \mathcal{P}^{(N)'} \underbrace{(\mathcal{P}^{(Z_i)}(1))}_{=1} \cdot \mathcal{P}^{(Z_i)''}(1) + E(Y) - [E(Y)]^2 \\
&= \mathcal{P}^{(N)''}(1)[E(Z_i)]^2 + E(N) \cdot \mathcal{P}^{(Z_i)''}(1) + E(N)E(Z_i) - [E(N)E(Z_i)]^2 \\
&= \left\{\mathcal{P}^{(N)''}(1) + E(N) - [E(N)]^2 - E(N)\right\} [E(Z_i)]^2 + \\
&\quad + \left\{\mathcal{P}^{(Z_i)''}(1) + E(Z_i) - [E(Z_i)]^2 + [E(Z_i)]^2\right\} E(N) \\
&= Var N \, [E(Z_i)]^2 + Var Z_i \, E(N) \\
&\overset{Var N = E(N)}{=} E(N)\{Var Z_i + [E(Z_i)]^2\} \\
&= E(N) \, E({Z_i}^2) \\
&> E(N) \, E(Z_i) = E(Y)
\end{aligned}$$

Wegen der Voraussetzung $E({Z_i}^2) > E(Z_i)$ *ist die behauptete Überdispersion bewiesen.*

Als nächstes zeigen wir, daß Überdispersion auch auftreten kann, wenn die Annahme der konstanten Ausfallrate verletzt ist.

Satz 2.17 *(vgl. Winkelmann [1996, S. 46]) Sei* $\{\tau_i\}$ *eine Folge von iid Zufallsvariablen und* $N(t)$ *die Anzahl der Ereignisse in* $]0, t[$. *Desweiteren nehmen wir an, daß die Dichten der Wartezeiten* τ_i *eine monotone Ausfallrate besitzen. Dann erzeugt negative (positive) Alterung der Dichten*

von τ_i Überdispersion (Unterdispersion) der Verteilung von $N(t)$ für $t \to \infty$.
Beweis: *Wir führen folgende Bezeichnungen ein:*

$$E(\tau_i) = \mu \qquad Var\ \tau_i = \sigma^2 \qquad S_r = \tau_1 + \cdots + \tau_r$$

Da S_r eine Summe von iid Zufallsvariablen ist, folgt nach dem Zentralen Grenzwertsatz, daß die Zufallsvariable S_r für $r \to \infty$ asymptotisch normalverteilt mit Erwartungswert μr und Varianz $\sigma^2 r$ ist.
Wir setzen $r_t := \frac{t}{\mu} + y_t\sigma\sqrt{\frac{t}{\mu^3}}$, $y_t \in \mathbb{R}^+$, was sich umformen läßt zu $\mu r_t - t = y_t\sigma\sqrt{\frac{t}{\mu}}$.
In der Definition tritt eine kleine analytische Schwierigkeit auf, da r_t eine ganze Zahl sein muß. Für $t \to \infty$ müssen wir eigentlich einen Grenzprozeß betrachten, in dem $y_t = y + \varepsilon_t$ ist, wobei ε_t der kleinste Wert größer als Null ist, so daß r_t ganzzahlig ist. Es bleibt also zu zeigen, daß ε_t für $t \to \infty$ verschwindet, um die Schreibweise zu rechtfertigen. Dazu folgern wir, daß wegen der Bedingung $r_t \in \mathbb{N}_0$ mit $r_t = \frac{t}{\mu} + y_t\sigma\sqrt{\frac{t}{\mu^3}} \overset{y_t = y+\varepsilon_t}{=} \frac{t}{\mu} + y\sigma\sqrt{\frac{t}{\mu^3}} + \varepsilon_t\sigma\sqrt{\frac{t}{\mu^3}} =: A + \varepsilon_t\sigma\sqrt{\frac{t}{\mu^3}}$ gelten muß

$$0 \le \varepsilon_t\sigma\sqrt{\frac{t}{\mu^3}} < 1 \qquad da\ [A] \le A < [A] + 1.$$

Diese Ungleichung für ε_t ist äquivalent zu $0 \le \varepsilon_t\sqrt{t} < \frac{\sqrt{\mu^3}}{\sigma}$. Da μ und σ fest sind, ist $\varepsilon_t\sqrt{t}$ beschränkt. Lassen wir nun $t \to \infty$ laufen, so konvergiert ε_t aufgrund der Beschränktheit von $\varepsilon_t\sqrt{t}$ gegen Null.
Dann gilt für die Verteilung von $N(t)$:

$$\begin{aligned} P\left(N(t) < r_t\right) &= P\left(S_{r_t} > t\right) = P\left(\frac{S_{r_t} - r_t\mu}{\sigma\sqrt{r_t}} > \frac{t - \mu r_t}{\sigma\sqrt{r_t}}\right) \\ &= P\left(\frac{S_{r_t} - r_t\mu}{\sigma\sqrt{r_t}} > \frac{-y_t\sigma\sqrt{t}}{\sigma\sqrt{r_t\mu}}\right) = P\left(\frac{S_{r_t} - r_t\mu}{\sigma\sqrt{r_t}} > \frac{-y_t\sqrt{t}}{(t + y_t\sigma\sqrt{t/\mu})^{1/2}}\right) \\ &= P\left(\frac{S_{r_t} - r_t\mu}{\sigma\sqrt{r_t}} > -y_t\left(1 + \frac{y_t\sigma}{\sqrt{t\mu}}\right)^{-1/2}\right) \end{aligned}$$

Wegen der asymptotischen Normalverteilung von S_r und insbesondere deren Symmetrie gilt:

$$P\left(\frac{S_{r_t} - r_t\mu}{\sigma\sqrt{r_t}} > -y_t\left(1 + \frac{y_t\sigma}{\sqrt{t\mu}}\right)^{-1/2}\right) \overset{as}{=} P\left(\frac{S_{r_t} - r_t\mu}{\sigma\sqrt{r_t}} < y_t\left(1 + \frac{y_t\sigma}{\sqrt{t\mu}}\right)^{-1/2}\right)$$

Wir halten nun $y_t = y$ fest und lassen $t \to \infty$ gehen. Dann folgt aus obiger Gleichung und aus dem Zentralen Grenzwertsatz für S_r, daß

$$\lim_{t\to\infty} P\left(N(t) < r_t\right) = \lim_{t\to\infty} P\left(\frac{S_{r_t} - r_t\mu}{\sigma\sqrt{r_t}} < y\right) = \Phi(y)\ , \tag{2.18}$$

wobei Φ die Standardnormalverteilungsfunktion bezeichnet. Damit haben wir bewiesen, daß $N(t)$ asymptotisch normalverteilt ist mit dem Erwartungswert t/μ und der Varianz $\sigma^2 t/\mu^3$.
Um jetzt die behauptete Überdispersion (Unterdispersion) zu beweisen, verwenden wir ein Resultat von Barlow und Proschan [1965, S. 33], nach dem aus negativer (positiver) Alterung folgt,

daß die relative Streuung $v = \sigma/\mu$ der Wartezeit größer (kleiner) als 1 ist.
Für die Grenzverteilung von $N(T)$ gilt:

$$\frac{Varianz}{Erwartungswert} \sim \frac{\sigma^2 t}{\mu^3}\frac{\mu}{t} = \frac{\sigma^2}{\mu^2} \ ,$$

weshalb $v = \sigma/\mu$ genau dann größer (kleiner) als 1 ist, wenn obiger Quotient aus Varianz und Erwartungswert größer (kleiner) als 1 ist.

Wir bemerken abschließend, daß bei einem Poisson-Prozeß mit exponentiell verteilten Wartezeiten stets $v = 1$ exakt gilt, während der eben bewiesene Satz nur ein asymptotisches Ergebnis liefert.

2.2.2 Modellierung der Überdispersion

Nachdem wir gezeigt haben, daß es aus mehreren Gründen zu Abweichungen von der Äquidispersion kommt, stellen wir in diesem Abschnitt zwei Ansätze zur Modellierung der Überdispersion vor. Dabei parametrisieren wir die Überdispersion mit einer negativen Binomialverteilung, die wir mit ihren wichtigsten Eigenschaften zunächst einführen.

Definition 2.18 (Negative Binomialverteilung) *Eine Zufallsvariable X mit Werten in $\mathbb{N}_0$ ist negativ binomialverteilt mit Parametern a und b, wenn X die Wahrscheinlichkeitsfunktion*

$$P(X = k) = \frac{\Gamma(a+k)}{\Gamma(a)\Gamma(k+1)}\left(\frac{1}{1+b}\right)^a\left(\frac{b}{1+b}\right)^k \qquad k \in \mathbb{N}_0, \quad a, b \in \mathbb{R}^+$$

besitzt. Wir schreiben kurz: $X \sim NB(a, b)$.
Setzen wir $a = 1$, erhalten wir die geometrische Verteilung als Spezialfall.

Es gibt eine bekannte Interpretation der negativen Binomialverteilung für $a \in \mathbb{N}$. Dann beschreibt $P(X = k)$ die Wahrscheinlichkeit, daß in $(a + k)$ unabhängigen Bernoulliversuchen genau k Mißerfolge dem a-ten Erfolg vorausgehen.

Lemma 2.19 (Eigenschaften der negativen Binomialverteilung)
Sei $X \sim NB(a, b)$. Dann gilt:

(i) Die wahrscheinlichkeitserzeugende Funktion lautet (für einen Beweis s. Feller [1957, S. 164]):
$\mathcal{P}(s) = [1 + b(1 - s)]^{-a}$

(ii) Erwartungswert und Varianz von X sind gegeben durch
$E(X) = \mathcal{P}'(1) = ab$
$VarX = \mathcal{P}''(1) + \mathcal{P}'(1) - [\mathcal{P}'(1)]^2 = ab(1 + b)$
Wir bemerken, daß wegen $b > 0$ die Varianz von X größer als der Erwartungswert ist, also immer Überdispersion auftritt.

(iii) Wenn $a \to \infty$ und $b \to 0$, so daß $ab = \lambda$ konstant ist, dann konvergiert die negative Binomialverteilung gegen die Poissonverteilung mit Parameter λ.

Beweis: *Wir betrachten die wahrscheinlichkeitserzeugende Funktion von X und ersetzen b durch λ/a:*

$$\lim_{\substack{a\to\infty \\ b\to 0}} [1 + b(1-s)]^{-a} = \lim_{a\to\infty} [1 + \frac{\lambda(1-s)}{a}]^{-a} = e^{-\lambda(1-s)} \tag{2.19}$$

Aber das ist gemäß Beispiel A.1 gerade die wahrscheinlichkeitserzeugende Funktion einer Zufallsvariablen, die Poisson-verteilt ist mit Parameter λ.

Wir kehren jetzt zur Modellierung der Überdispersion zurück.

(i) zufällige Effekte

Bei einer Poissonregression ohne unbeobachtete Heterogenität ist die Verteilung von $(Y_i|\mathbf{x}_i)$ unter den gegebenen Regressoren $\mathbf{x}_i$ spezifiziert. Dies ist gleichbedeutend mit der Spezifizierung des Erwartungswerts als nichtstochastische Funktion von $\mathbf{x}_i$. Dagegen spezifizieren wir in gemischten Modellen die Verteilung von $(Y_i|\mathbf{x}_i, u_i)$, wobei u_i den Ausdruck für die unbeobachtete zufällige Heterogenität in der i-ten Beobachtung bezeichnet. Der genaue spezielle Zusammenhang zwischen Y_i und $(\mathbf{x}_i, u_i)$ muß bekannt sein. Eine übliche funktionelle Gestalt ist der exponentielle Erwartungswert mit multiplikativem Fehler: $E(Y_i|\mathbf{x}_i, U_i = u_i) = \exp(\mathbf{x}_i^T\beta)u_i$, wobei die U_i nichtnegative, iid Zufallsvariablen und unabhängig von den Regressoren $\mathbf{x}_i$ sind. Die multiplikative Heterogenitätsannahme ist recht speziell, aber mathematisch gängig und attraktiver als ein additiver Fehler, der zu einer Verletzung der Nichtnegativitätsvoraussetzung der Y_i führen könnte. Durch diese Modellierung ist $(Y_i|\mathbf{x}_i, u_i) \sim Poi(\theta_i)$ mit $\theta_i := \exp(\mathbf{x}_i\beta)u_i$. Der Erwartungswert θ_i ist nicht mehr fest, sondern selbst eine Zufallsvariable. Obwohl wir θ_i und U_i nicht kennen, sind wir in der Lage, Aussagen über die Verteilung von $(Y_i|\mathbf{x}_i)$ zu machen. Dazu normalisieren wir $E(U_i) = 1$, wenn es einen Intercept gibt, um Identifizierbarkeit zu garantieren. Weiter gelten die Bezeichnungen ${\sigma_u}^2 := Var\, U_i$ und $\mu_i := \exp(\mathbf{x}_i^T\beta)$. Mit den bekannten Formeln für den bedingten Erwartungswert und die bedingte Varianz (s. Casella/Berger [1990, S. 156 und S. 158]) bestimmen wir die ersten beiden Momente von $(Y_i|\mathbf{x}_i)$:

$$E(Y_i|\mathbf{x}_i) = E_{U_i}(E[Y_i|\mathbf{x}_i, U_i]) = E_{U_i}(\mu_i U_i) = \mu_i$$

$$\begin{aligned} Var\,(Y_i|\mathbf{x}_i) &= E_{U_i}(Var[Y_i|\mathbf{x}_i, U_i]) + Var_{U_i}(E[Y_i|\mathbf{x}_i, U_i]) \\ &= E_{U_i}(\mu_i U_i) + Var_{U_i}(\mu_i U_i) \\ &= \mu_i E_{U_i}(U_i) + {\mu_i}^2 Var_{U_i}(U_i) = \mu_i + {\sigma_u}^2 {\mu_i}^2 \end{aligned}$$

Wir sehen daraus, daß unbeobachtete, zufällige Heterogenität Überdispersion erzeugt, sofern U_i nicht degeneriert ist, d. h. $Var\, U_i = 0$.

Ist die Dichte $g(u_i)$ von U_i bekannt, so erhalten wir die Randverteilung von $(Y_i|\mathbf{x}_i)$ durch Integration der gemeinsamen Dichte über U_i:

$$h(y_i|\mu_i) = \int f(y_i|\mu_i, u_i)\, g(u_i)\, du_i,$$

wobei $f(y_i|\mu_i, u_i)$ die Poisson-Wahrscheinlichkeitsfunktion von $Y_i|\mathbf{x}_i, u_i$ bezeichnet.

Beispiel 2.20 *(s. Cameron/Trivedi [1998, S. 100 f.]) Wir nehmen an, daß die U_i gammaverteilt sind, und zeigen, daß die gemischte Randverteilung von $(Y_i|\mathbf{x}_i)$ negativ binomial ist. Für den Rest des Beispiels lassen wir den Index i weg, da Doppeldeutigkeit auszuschließen ist. Sei also $U \sim G(\alpha, \beta)$, d. h. U besitzt die Dichte $g(u) = \frac{\beta^\alpha}{\Gamma(\alpha)} u^{\alpha-1} e^{-u\beta}$ mit $E(U) = \alpha/\beta$ und $Var\, U = \alpha/\beta^2$. Um die Identifizierbarkeitsbedingung $E(U) = 1$ zu erfüllen, müssen wir $\beta = \alpha$ setzen. Dadurch verlieren wir einen Freiheitsgrad bei der Gammaverteilung, und die Varianz von U lautet $1/\alpha$. Dann ist die Randverteilung von Y gegeben durch*

$$
\begin{aligned}
h(y|\mu,\alpha) &= \int_0^\infty f(y|\mu,u)g(u)\,du \\
&\overset{\theta=\mu u}{=} \int_0^\infty f(y|\mu,\theta)g(\theta)\frac{1}{\mu}\,d\theta \\
&= \int_0^\infty \frac{\theta^y}{y!}e^{-\theta}\,\frac{\alpha^\alpha}{\Gamma(\alpha)}\left(\frac{\theta}{\mu}\right)^{\alpha-1} e^{-\frac{\theta\alpha}{\mu}}\frac{1}{\mu}\,d\theta \\
&\overset{\Gamma(y+1)=y!}{=} \frac{1}{\Gamma(y+1)\Gamma(\alpha)}\int_0^\infty \left(\frac{\alpha}{\mu}\right)^\alpha \theta^{\alpha+y-1}e^{-\theta(1+\alpha/\mu)}\,d\theta \\
&= \frac{\Gamma(\alpha+y)}{\Gamma(y+1)\Gamma(\alpha)}\left(\frac{\alpha}{\mu}\right)^\alpha\left(\frac{\mu}{\alpha+\mu}\right)^{\alpha+y} \\
&\quad \underbrace{\int_0^\infty \frac{1}{\Gamma(\alpha+y)}\left(\frac{\alpha+\mu}{\mu}\right)^{\alpha+y}\theta^{\alpha+y-1}e^{-\theta\frac{\alpha+\mu}{\mu}}\,d\theta}_{=1 \text{ da Integrand } G(\alpha+y,\frac{\alpha+\mu}{\mu})\text{-Dichte}} \\
&= \frac{\Gamma(\alpha+y)}{\Gamma(y+1)\Gamma(\alpha)}\left(\frac{\alpha}{\alpha+\mu}\right)^\alpha\left(\frac{\mu}{\alpha+\mu}\right)^{y} \\
&= \frac{\Gamma(\alpha+y)}{\Gamma(y+1)\Gamma(\alpha)}\left(\frac{1}{1+\mu/\alpha}\right)^\alpha\left(\frac{\mu/\alpha}{1+\mu/\alpha}\right)^{y}
\end{aligned}
$$

Damit ist die Randverteilung $NB(\alpha, \mu/\alpha)$ mit den ersten beiden Momenten $E(Y|\mu,\alpha) = \mu$ und $Var(Y|\mu,\alpha) = \mu(1 + \frac{1}{\alpha}\mu)$.
Um jetzt den Schritt zu einer Regression zu vollziehen, müssen wir den Parameter α mittels den Regressoren spezifizieren. Am einfachsten wählen wir α konstant, also $\alpha = 1/{\sigma_u}^2$. Dadurch ist ein quadratischer Zusammenhang zwischen Varianz und Erwartungswert gegeben. Negative Binomialmodelle mit quadratischer Varianz bezeichnen wir kurz mit NB2. Lassen wir α von den Beobachtungen abhängen, indem wir $\alpha = 1/{\sigma_u}^2\mu$ setzen, so ist die Varianz eine lineare Funktion des Erwartungswerts. Negative Binomialmodelle mit linearer Varianzfunktion bezeichnen wir kurz mit NB1.
Bei der Berechnung der Randverteilung substituierten wir u durch θ/μ. Schauen wir uns die daraus entstandene Dichte für θ, $\frac{1}{\mu}g(\theta)$, an, so stellen wir fest, daß dies die Dichte

einer Gammaverteilung mit den Parametern α und α/μ ist. Da das Argument θ als Erwartungswert im Poissonmodell mit unbeobachteter Heterogenität eine Zufallsvariable ist, können wir allgemeiner formulieren: Ist $\theta \sim G(\alpha, \alpha/\mu)$ und $Y|\theta \sim Poi(\theta)$, dann führt die Mischung zu einer $NB(\alpha, \mu/\alpha)$-Verteilung für Y.

(ii) zufällige Summen von Zufallsvariablen
Wir haben gezeigt, daß die Poissonverteilung einen Zählprozeß während eines festen Zeitintervalls angemessen beschreibt, wenn die Ereignisse voneinander unabhängig eintreten und ihre Wartezeiten keine Alterung aufweisen. Überdispersion tritt auf, wenn die Unabhängigkeitsannahme verletzt ist. Jetzt beweisen wir, daß auch diese Situation mit der negativen Binomialverteilung modelliert werden kann.

Satz 2.21 *Sei $Y = Z_1 + \cdots + Z_N$ mit $Z_i \in I\!N$ und iid und $N \sim Poi(\lambda)$. Z_i besitze eine logarithmische Verteilung mit Parameter θ, $\theta \in]0,1[$, d.h. $P(Z_i = k) = \alpha\theta^k/k$ mit $\alpha = -[\ln(1-\theta)]^{-1}$. Dann: $Y \sim NB(\alpha\lambda, \frac{\theta}{1-\theta})$.*
Beweis: *Wir bestimmen die wahrscheinlichkeitserzeugende Funktion von Y.*
Wegen Beispiel (A.1) und Beispiel (A.2) lauten die wahrscheinlichkeitserzeugenden Funktionen von Z_i und N
$\mathcal{P}^{(Z_i)}(s) = -\alpha\ln(1-\theta s)$ und $\mathcal{P}^{(N)}(s) = e^{-\lambda+\lambda s}$.
Wir wenden erneut den Satz $\mathcal{P}^{(Y)}(s) = \mathcal{P}^{(N)}(\mathcal{P}^{(Z_i)}(s))$ aus Feller [1968, S 286 f] an:

$$\begin{aligned}
\mathcal{P}^{(Y)}(s) &= \exp[-\lambda - \lambda\alpha\ln(1-\theta s)] \\
&= (1-\theta s)^{-\alpha\lambda}\exp(-\lambda) = (1-\theta s)^{-\alpha\lambda}\exp(1/\alpha)^{-\alpha\lambda} \\
&\overset{1/\alpha=\ln(1-\theta)}{=} \left[\frac{1-\theta s}{1-\theta}\right]^{-\alpha\lambda} = \left(1 - \frac{\theta}{1-\theta}(1-s)\right)^{-\alpha\lambda}
\end{aligned}$$

Das ist genau die wahrscheinlichkeitserzeugende Funktion einer negativen Binomialverteilung mit den Parametern $\alpha\lambda$ und $\frac{\theta}{1-\theta}$.

Liegen die Regressionsdaten in Form einer Kreuzklassifikation vor, dann können wir bei einer negativen Binomialverteilung nicht unterscheiden, welcher der beiden hier erörterten Mechanismen, nämlich zufällige Effekte wie in Beispiel 2.20 oder zufällige Summen wie in Satz 2.21, der Überdispersion zugrunde liegt.

2.3 Beurteilung der Anpassung

Methoden zur Beurteilung der Modellanpassung können entweder formell oder nicht-formell sein. Nicht-formelle Methoden stützen sich auf die Meinung und das Auge des Statistikers, um Muster zu erkennen. Ein Modell wird anhand solcher Methoden als geeignet beurteilt, wenn neben anderen Kriterien die Residuen kein Muster aufweisen. Die Argumentation lautet, daß wir beim Erkennen eines Musters in den Residuen ein angemesseneres Modell finden können. Das praktische Problem dabei besteht darin, daß wir uns vor einer Überinterpretation schützen

müssen. Denn jede endliche Residuenmenge kann derart gemacht werden, daß sie zu einem Muster führt, wenn wir nur genau genug hinschauen.
Formelle Methoden stützen sich darauf, das aktuelle Modell in eine umfassendere Familie mit einem zusätzlichen Parameter einzubetten. Wenn θ ein solcher Parameter ist und den Wert θ_0 in dem zu untersuchenden Modell annimmt, dann bestimmt die formelle Methode einen Schätzer $\hat{\theta}$, der die beste Anpassung innerhalb der größeren Familie darstellt. $\hat{\theta}$ wird mit θ_0 verglichen, und wenn die Hinzunahme des zusätzlichen Parameters die Anpassung nicht deutlich verbessert, behalten wir das bisherige Modell bei.
Wir stellen in diesem Abschnitt beide Methoden allgemein vor. Im folgenden Kapitel konkretisieren wir den formellen Ansatz, indem wir die Varianzfunktion des Poissonmodells in eine größere Familie von Varianzfunktionen einbetten.

2.3.1 Testen von Hypothesen

Es gibt drei „klassische“ Tests: den Likelihood-Quotienten-, den Wald- und den Lagrangeschen Multiplikatoren-Test (auch Score-Test genannt). In diesem Abschnitt gehen wir ausführlich auf die asymptotischen Eigenschaften des Likelihood-Quotienten-Test (LQ-Test) und des Lagrangeschen Multiplikatoren-Test (LM-Test) ein, während wir nur kurz bemerken, daß alle Ergebnisse auch für den Wald-Test gelten.
Für diesen Abschnitt seien stets $Y_1, \ldots, Y_n$ unabhängige, nicht notwendig identisch verteilte Zufallsvariablen mit (stetigen oder diskreten) Dichten $f_i(y;\theta), i = 1, \ldots, n, \theta \in \mathbb{R}^q$, wenn nicht explizit eine andere Voraussetzung angegeben ist. Es gelten die Regularitätsbedingungen i) – iv) aus Abschnitt 2.1.4 für den ML-Schätzer. $\Omega \subseteq \mathbb{R}^q$ bezeichne den gesamten Parameterraum, Ω_0 den Parameterraum unter H_0 und Ω_A mit $\Omega = \Omega_0 \dot{\cup} \Omega_A$ den Parameterraum unter H_1. Desweiteren setzen wir voraus, daß der wahre Parameter im Inneren von Ω liegt. Wir führen noch folgende Bezeichnungen ein:
$\hat{\theta}$ sei der uneingeschränkte ML-Schätzer von θ. Für alle $i = 1, \ldots, n$ und alle $j, r, s = 1, \ldots, q$ gelte:

$$\ell_i(\theta) := \ln f_i(y;\theta) \qquad \ell(\theta) := \sum_{i=1}^{n} \ell_i(\theta)$$

$$\frac{\partial}{\partial\theta_j}\ell_i(\theta_j{}^*) := \frac{\partial}{\partial\theta_j}\ell_i(\theta_j)|_{\theta_j=\theta_j{}^*} \qquad \frac{\partial}{\partial\theta_j}\ell(\theta_j{}^*) := \frac{\partial}{\partial\theta_j}\ell(\theta_j)|_{\theta_j=\theta_j{}^*}$$

$$\frac{\partial}{\partial\theta}\ell_i(\theta) := \left(\frac{\partial}{\partial\theta_1}\ell_i(\theta), \ldots, \frac{\partial}{\partial\theta_q}\ell_i(\theta)\right)^T \qquad \frac{\partial}{\partial\theta}\ell(\theta) := \left(\frac{\partial}{\partial\theta_1}\ell(\theta), \ldots, \frac{\partial}{\partial\theta_q}\ell(\theta)\right)^T$$

$$\mathcal{I}_i(\theta) := E\left[\frac{\partial}{\partial\theta}\ell_i(\theta)\frac{\partial}{\partial\theta^T}\ell_i(\theta)\right] = (\mathcal{I}^i_{rs}) \quad \text{mit} \quad \mathcal{I}^i_{rs} = E\left[\frac{\partial}{\partial\theta_r}\ell_i(\theta)\frac{\partial}{\partial\theta_s}\ell_i(\theta)\right]$$

$$\mathcal{I}_+(\theta) := E\left[\frac{\partial}{\partial\theta}\ell(\theta)\frac{\partial}{\partial\theta^T}\ell(\theta)\right] = (\mathcal{I}^+_{rs}) \quad \text{mit} \quad \mathcal{I}^+_{rs}(\theta) = E\left[\frac{\partial}{\partial\theta_r}\ell(\theta)\frac{\partial}{\partial\theta_s}\ell(\theta)\right]$$

Wir setzen voraus, daß die Fisher-Information $\mathcal{I}_+(\theta)$ der gesamten Stichprobe in einer Umgebung des wahren Parameters positiv definit und endlich ist. Schließlich fordern wir, daß die Voraussetzungen des schwachen Gesetz der großen Zahlen für $\frac{\partial^2}{\partial\theta_r\,\partial\theta_s}\ell(\theta), r, s = 1, \ldots, q$, erfüllt sind. In diesem Abschnitt bezeichne $\|.\|$ eine beliebige Norm im $\mathbb{R}^q$, die für einen Satz jeweils

festgehalten wird. Mit den folgenden Bemerkungen stellen wir einige Zusammenhänge vor, die wir in den kommenden Beweisen immer wieder verwenden werden.

a) Wegen der Unabhängigkeit der $Y_i, i = 1, \ldots, n$, gilt $\mathcal{I}_+(\theta) = \sum_{i=1}^n \mathcal{I}_i(\theta)$.

b) Die Regularitätsbedingungen iii) und iv) aus Abschnitt 2.1.4 sichern die Informationsgleichung (2.12), weswegen für alle $r, s = 1, \ldots, q$ mit dem schwachen Gesetz der großen Zahlen gilt:

$$\frac{1}{n}\frac{\partial^2}{\partial\theta_r\,\partial\theta_s}\ell(\theta) - \frac{1}{n}E\left(\frac{\partial^2}{\partial\theta_r\,\partial\theta_s}\ell(\theta)\right) = \frac{1}{n}\frac{\partial^2}{\partial\theta_r\,\partial\theta_s}\ell(\theta) + \frac{1}{n}E\left(\frac{\partial}{\partial\theta_r}\ell(\theta)\,\frac{\partial}{\partial\theta_s}\ell(\theta)\right)$$

$$= \frac{1}{n}\sum_{i=1}^n \frac{\partial^2}{\partial\theta_r\partial\theta_s}\ell_i(\theta) + \frac{1}{n}\sum_{i=1}^n E\left(\frac{\partial}{\partial\theta_r}\ell_i(\theta)\,\frac{\partial}{\partial\theta_s}\ell_i(\theta)\right) \xrightarrow{p} 0$$

Dann gilt für die Matrix

$$-\frac{1}{n}\frac{\partial^2}{\partial\theta\,\partial\theta^T}\ell(\theta) \overset{\text{as}}{=} \frac{1}{n}\mathcal{I}_+(\theta) \tag{2.20}$$

c) Aufgrund der Definition des ML-Schätzers $\hat\theta$ ist $\frac{\partial}{\partial\theta}\ell(\hat\theta) = \mathbf{0}$. Mit einer multivariaten Taylorentwicklung im wahren Parameter θ^* erhalten wir

$$\mathbf{0} = \frac{\partial}{\partial\theta}\ell(\hat\theta) = \frac{\partial}{\partial\theta}\ell(\theta^*) + \frac{\partial^2}{\partial\theta\,\partial\theta^T}\ell(\tilde\theta)(\hat\theta - \theta^*) \qquad \text{mit } \|\tilde\theta - \theta^*\| \le \|\hat\theta - \theta^*\| \tag{2.21}$$

und somit

$$\frac{\partial}{\partial\theta}\ell(\theta^*) = -\frac{\partial^2}{\partial\theta\,\partial\theta^T}\ell(\tilde\theta)(\hat\theta - \theta^*).$$

$\tilde\theta$ ist wegen $\|\tilde\theta - \theta^*\| \le \|\hat\theta - \theta^*\|$ ein konsistenter Schätzer für θ^*. Außerdem ist $\frac{\partial^2}{\partial\theta\,\partial\theta^T}\ell(\theta)$ stetig nach Regularitätsbedingung iii), so daß $\frac{\partial^2}{\partial\theta\,\partial\theta^T}\ell(\tilde\theta)$ ein konsistenter Schätzer für $\frac{\partial^2}{\partial\theta\,\partial\theta^T}\ell(\theta^*)$ ist. Wegen Bemerkung b) folgt insgesamt

$$\frac{\partial}{\partial\theta}\ell(\theta^*) \overset{\text{as}}{=} \mathcal{I}_+(\theta^*)(\hat\theta - \theta^*) \quad \text{bzw.} \quad \hat\theta - \theta^* \overset{\text{as}}{=} \mathcal{I}_+^{-1}(\theta^*)\frac{\partial}{\partial\theta}\ell(\theta^*) \tag{2.22}$$

Definition 2.22 (Likelihood-Quotienten-Test) *Der Likelihood-Quotienten-Test (LQ-Test) besitzt für $H_0 : \theta \in \Omega_0$ gegen $H_1 : \theta \in \Omega_A$ die Teststatistik*

$$e^{\frac{1}{2}T_{LQ}} = \frac{\sup_{\theta\in\Omega}\prod_{i=1}^n f_i(y;\theta)}{\sup_{\theta\in\Omega_0}\prod_{i=1}^n f_i(y;\theta)},$$

die äquivalent ist zu

$$T_{LQ} = 2\,[\sup_{\theta\in\Omega}\ell(\theta) - \sup_{\theta\in\Omega_0}\ell(\theta)] = 2\,[\ell(\hat\theta) - \sup_{\theta\in\Omega_0}\ell(\theta)]. \tag{2.23}$$

Wir lehnen H_0 zu einem vorgegeben Signifikanzniveau α ab, wenn $T_{LQ} \ge c(\alpha)$ mit $P(T_{LQ} \ge c(\alpha)) = \alpha$ ist.

Der LQ-Test benötigt den uneingeschränkten ML-Schätzer $\hat{\theta}$ von θ und den ML-Schätzer $\hat{\theta}_0$ unter H_0. Ist der uneingeschränkte ML-Schätzer schwer zu bestimmen, verwenden wir lieber den Lagrangeschen Multiplikatoren-Test (LM-Test).

Definition 2.23 (Lagrangescher Multiplikatoren-Test) *Der LM-Test besitzt für $H_0 : \theta \in \Omega_0$ gegen $H_1 : \theta \in \Omega_A$ die Teststatistik*

$$T_{LM} = \frac{\partial}{\partial \theta^T} \ell(\hat{\theta}_0) \left[\mathcal{I}_+(\hat{\theta}_0)\right]^{-1} \frac{\partial}{\partial \theta} \ell(\hat{\theta}_0). \tag{2.24}$$

Wie bei dem LQ-Test lehnen wir H_0 zum Signifikanzniveau α ab, wenn der Wert der Teststatistik T_{LM} den kritischen Wert $c(\alpha)$ annimmt oder übersteigt: $T_{LM} \geq c(\alpha)$ mit $P(T_{LM} \geq c(\alpha)) = \alpha$.

Wir bestimmen nun die asymptotischen Verteilungen des LQ-Tests und des LM-Tests sowohl unter einer einfachen als auch unter einer zusammengesetzten Nullhypothese.

Einfache Hypothesen

Bei der einfachen Hypothese $H_0 : \theta = \theta_0$ gegen $H_1 : \theta \neq \theta_0$ vereinfachen sich T_{LQ} und T_{LM} wegen $\Omega_0 = \{\theta_0\}$ zu $T_{LQ} = 2[\ell(\hat{\theta}) - \ell(\theta_0)]$ und $T_{LM} = \frac{\partial}{\partial \theta^T} \ell(\theta_0) [\mathcal{I}_+(\theta_0)]^{-1} \frac{\partial}{\partial \theta} \ell(\theta_0)$.

Satz 2.24 (Asymptotische Verteilung des LQ-Tests unter H_0 bei einfacher Hypothese) *T_{LQ} besitzt unter $H_0 : \theta = \theta_0$ gegen $H_1 : \theta \neq \theta_0$ eine asymptotische χ^2_q-Verteilung, falls $\theta \in \mathbb{R}^q$.*

Beweis: *H_0 ist wahr, d. h. θ_0 ist der wahre Parameterwert. Eine multivariate Taylorentwicklung von $\ell(\theta_0)$ in $\hat{\theta}$ analog (2.21) liefert:*

$$\ell(\theta_0) = \ell(\hat{\theta}) + \frac{1}{2}(\hat{\theta} - \theta_0)^T \frac{\partial^2}{\partial \theta \, \partial \theta^T} \ell(\tilde{\theta})(\hat{\theta} - \theta_0) \qquad \text{mit } \|\tilde{\theta} - \theta_0\| \leq \|\hat{\theta} - \theta_0\|. \tag{2.25}$$

Wir setzen diese Formel in T_{LQ} ein:

$$\begin{aligned} T_{LQ} = 2[\ell(\hat{\theta}) - \ell(\theta_0)] &= -(\hat{\theta} - \theta_0)^T \frac{\partial^2}{\partial \theta \, \partial \theta^T} \ell(\tilde{\theta})(\hat{\theta} - \theta_0) \\ &= n(\hat{\theta} - \theta_0)^T \left[-\frac{1}{n} \frac{\partial^2}{\partial \theta \, \partial \theta^T} \ell(\tilde{\theta}) \right] (\hat{\theta} - \theta_0) \end{aligned}$$

Nun konvergiert der Term in eckiger Klammer nach Bemerkungen b) und c) gegen $\frac{1}{n}\mathcal{I}_+(\theta_0)$ und $\sqrt{n}(\hat{\theta} - \theta_0)$ gegen $\mathcal{N}_q(\mathbf{0}, \mathcal{I}_+^{-1}(\theta_0))$ wegen der asymptotischen Normalität des ML-Schätzers, so daß T_{LQ} asymptotisch χ^2-verteilt ist mit Freiheitsgraden, die gleich der Dimension von θ sind.

Im folgenden wird der Wechsel von LQ-Test zu LM-Test rechtfertigt.

Satz 2.25 *T_{LM} besitzt unter H_0 eine asymptotische χ^2_q-Verteilung.*

Beweis: *Es reicht zu zeigen, daß T_{LM} asymptotisch äquivalent zu T_{LQ} unter H_0 ist.*

Erneut benutzen wir eine Taylorentwicklung und die Formeln (2.22) und (2.20) aus den Bemerkungen mit $\theta^* = \theta_0$, um T_{LQ} wie in Satz 2.24 umzuformen:

$$\begin{aligned} T_{LQ} &= 2[\ell(\hat{\theta}) - \ell(\theta_0)] \\ &\overset{(2.25)}{=} n(\hat{\theta} - \theta_0)^T \left[-\frac{1}{n} \frac{\partial^2}{\partial\theta\, \partial\theta^T} \ell(\tilde{\theta}) \right] (\hat{\theta} - \theta_0) \\ &\overset{\text{as}}{=} \left(\frac{\partial}{\partial\theta} \ell(\theta_0) \right)^T \mathcal{I}_+^{-1}(\theta_0) \mathcal{I}_+(\theta_0) \mathcal{I}_+^{-1}(\theta_0) \frac{\partial}{\partial\theta} \ell(\theta_0) \\ &= \frac{\partial}{\partial\theta^T} \ell(\theta_0) \mathcal{I}_+^{-1}(\theta_0) \frac{\partial}{\partial\theta} \ell(\theta_0) \\ &= T_{LM} \end{aligned}$$

Wir nehmen jetzt an, daß der wahre Parameter in Ω_A liegt. Bei unserer Untersuchung unterscheiden wir zwischen einer festen Alternative und lokalen Alternativen. Dazu benötigen wir den Begriff des *konsistenten Tests*, den wir zuerst einführen.

Definition 2.26 (Konsistenter Test) *In Analogie zur Konsistenz von Schätzern nennen wir eine Folge von Tests (T_n) oder kurz einen Test T_n für zwei Hypothesen H_0 gegen H_1 konsistent, wenn die Wahrscheinlichkeit für einen Fehler 2. Art gegen 0 strebt, wenn also mit der Bezeichnung B für den Annahmebereich gilt:*

$$P_\theta(T_n \in B) \overset{n\to\infty}{\longrightarrow} 0 \qquad \text{bzw.} \quad P_\theta(T_n \in B^C) = 1 - P_\theta(T_n \in B) \overset{n\to\infty}{\longrightarrow} 1 \qquad \forall\, \theta \in \Omega_A$$

Satz 2.27 *Ist (die feste Alternative) $H_1 : \theta = \theta_A$ wahr, dann ist der LQ-Test konsistent.*
Beweis: *Wir entwickeln $\ell(\theta_A)$ in $\hat{\theta}$: $\ell(\theta_A) = \ell(\hat{\theta}) + \frac{1}{2}(\theta_A - \hat{\theta})^T \frac{\partial^2}{\partial\theta\, \partial\theta^T} \ell(\tilde{\theta})(\theta_A - \hat{\theta})$ mit $\|(\tilde{\theta} - \hat{\theta})\| \leq \|\theta_A - \hat{\theta}\|$. Diese Taylorentwicklung setzen wir in T_{LQ} ein:*

$$\begin{aligned} T_{LQ} &= 2\left[\ell(\hat{\theta}) - \ell(\theta_0)\right] = 2\left[\ell(\hat{\theta}) - \ell(\theta_A) + \ell(\theta_A) - \ell(\theta_0)\right] \\ &= 2\left[\ell(\theta_A) - \ell(\theta_0)\right] - (\theta_A - \hat{\theta})^T \frac{\partial^2}{\partial\theta\, \partial\theta^T} \ell(\hat{\theta})\, (\theta_A - \hat{\theta}) \end{aligned}$$

Wie im Beweis von Satz 2.24 gilt im Grenzübergang $n \to \infty$, daß $-(\theta_A - \hat{\theta})^T \frac{\partial^2}{\partial\theta\, \partial\theta^T} \ell(\hat{\theta})\, (\theta_A - \hat{\theta})$ eine $\chi^2(\dim\theta_A)$-Verteilung besitzt. Können wir jetzt noch zeigen, daß $\ell(\theta_A) - \ell(\theta_0)$ unbeschränkt ist, dann gilt für jedes $d \in \mathbb{R}$

$$P_{\theta_A}(T_{LQ} > d) \overset{n\to\infty}{\longrightarrow} 1$$

mit P_{θ_A} als Wahrscheinlichkeit bzgl. der Dichte $\prod_{i=1}^n f_i(y; \theta_A)$, womit die Konsistenz von dem LQ-Test bewiesen ist.
Wir führen den Beweis der Unbeschränktheit von $\ell(\theta_A) - \ell(\theta_0)$ durch Widerspruch.
Angenommen: $\exists\, K > 0 : \quad \ell(\theta_A) - \ell(\theta_0) \leq K,$
dann gilt $\frac{1}{n}[\ell(\theta_A) - \ell(\theta_0)] \leq \frac{K}{n}.$
und

$$\frac{1}{n}[\ell(\theta_A) - \ell(\theta_0)] \leq 0 \qquad \text{für } n \to \infty \tag{2.26}$$

Andererseits haben wir aufgrund der Jensenschen Ungleichung (und $\ln(\cdot)$ *konvex) folgende Ungleichung:*

$$-E_{\theta_A}[\ell(\theta_A) - \ell(\theta_0)] = E_{\theta_A}\left[\ln \frac{f(\mathbf{x};\theta_0)}{f(\mathbf{x};\theta_A)}\right] \leq \ln E_{\theta_A}\left[\frac{f(\mathbf{x};\theta_0)}{f(\mathbf{x};\theta_A)}\right],$$

wobei E_{θ_A} *den Erwartungswert bzgl. der gemeinsamen Dichte* $\prod_{i=1}^{n} f_i(y;\theta_A)$ *bezeichnet. Gleichheit gilt genau dann, wenn der Quotient* $\frac{f(\mathbf{x};\theta_0)}{f(\mathbf{x};\theta_A)}$ *f. s. eine Konstante ist. Wir folgern die echte Ungleichheit aus der Regularitätsbedingung, daß die Dichte injektiv in* θ *ist. Außerdem gilt:*

$$E_{\theta_A}\left[\frac{f(\mathbf{x};\theta_0)}{f(\mathbf{x};\theta_A)}\right] = \int\limits_S \frac{f(\mathbf{x};\theta_0)}{f(\mathbf{x};\theta_A)} f(\mathbf{x};\theta_A) d\mathbf{x} = \int\limits_S f(\mathbf{x};\theta_0) d\mathbf{x} \leq 1$$

mit $S = \{x \,|\, f(\mathbf{x};\theta_A) > 0\}$*, wobei Gleichheit genau dann eintritt, wenn* $S = \{x \,|\, f(\mathbf{x};\theta_0) > 0\}$ *f. s. Beide Ungleichungen zusammen liefern*

$$E_{\theta_A}[\ell(\theta_A) - \ell(\theta_0)] > \ln E_{\theta_A}\left[\frac{f(\mathbf{x};\theta_0)}{f(\mathbf{x};\theta_A)}\right] \geq \ln 1 = 0$$

Diese Ungleichung beinhaltet aufgrund einer Anwendung des starken Gesetz der großen Zahlen, daß für $n \to \infty$ *gilt:*

$$\frac{1}{n}[\ell(\theta_A) - \ell(\theta_0)] > 0 \quad f.s.$$

Damit ist ein Widerspruch zu (2.26) hergestellt und die Unbeschränktheit von $\ell(\theta_A) - \ell(\theta_0)$ *bewiesen.*

Im folgenden nehmen wir an, daß der wahre Parameter in H_1 liegt und in Abhängigkeit von n die Werte θ_n mit $\theta_n \overset{n\to\infty}{\longrightarrow} \theta_0$ annimmt (lokale Alternativen). Dabei unterscheiden wir zwei Konvergenzgeschwindigkeiten. Zunächst betrachten wir den Fall, daß $\sqrt{n}(\theta_n - \theta_0)$ konvergiert, und darauf den Fall, daß $\sqrt{n}(\theta_n - \theta_0)$ divergiert. In beiden Fällen gilt aufgrund der asymptotischen Normalverteilung des ML-Schätzers $\hat{\theta}$ und aufgrund der asymptotischen Äquivalenz von $\mathcal{I}_+(\theta_n)$ und $\mathcal{I}_+(\theta_0)$

$$\sqrt{n}(\hat{\theta} - \theta_n) \overset{as}{\sim} \mathcal{N}_q(\mathbf{0}, \mathcal{I}_+(\theta_0)^{-1}) \tag{2.27}$$

Satz 2.28 *Wenn die lokale Alternative* $H_1 : \theta = \theta_n$ *mit* $\theta_n \overset{n\to\infty}{\longrightarrow} \theta_0$ *wahr ist und* $\sqrt{n}(\theta_n - \theta_0)$ *konvergiert, besitzt der LQ-Test asymptotisch eine nichtzentrale* χ^2_q*-Verteilung mit Nichtzentralitätsparameter* $(\theta_n - \theta_0)^T \mathcal{I}_+(\theta_0)(\theta_n - \theta_0)$*.*

Beweis*: Wir benutzen die Taylorentwicklungen von*

$$\begin{aligned}
\ell(\theta_n) &= \ell(\hat{\theta}) + \frac{1}{2}(\hat{\theta} - \theta_n)^T \frac{\partial^2}{\partial\theta\,\partial\theta^T}\ell(\theta_1)(\hat{\theta} - \theta_n) \quad mit \quad \|(\theta_1 - \theta_n)\| \leq \|(\hat{\theta} - \theta_n)\| \quad und \\
\ell(\theta_0) &= \ell(\theta_n) + (\theta_n - \theta_0)^T \frac{\partial}{\partial\theta}\ell(\theta_n) + \frac{1}{2}(\theta_n - \theta_0)^T \frac{\partial^2}{\partial\theta\,\partial\theta^T}\ell(\theta_2)(\theta_n - \theta_0) \\
&\quad mit \quad \|(\theta_2 - \theta_n)\| \leq \|(\hat{\theta} - \theta_n)\|,
\end{aligned}$$

um T_{LQ} *umzuformen:*

$$\begin{aligned}
T_{LQ} &= 2\,[\ell(\hat{\theta}) - \ell(\theta_0)] = 2\,[\ell(\hat{\theta}) - \ell(\theta_n) + \ell(\theta_n) - \ell(\theta_0)] \\
&= 2\,[\ell(\theta_n) - \ell(\theta_0)] - (\hat{\theta} - \theta_n)^T \frac{\partial^2}{\partial\theta\,\partial\theta^T}\ell(\theta_1)(\hat{\theta} - \theta_n) \\
&= 2\,(\theta_n - \theta_0)^T \frac{\partial}{\partial\theta}\ell(\theta_n) + (\theta_n - \theta_0)^T \frac{\partial^2}{\partial\theta\,\partial\theta^T}\ell(\theta_2)(\theta_n - \theta_0) - (\hat{\theta} - \theta_n)^T \frac{\partial^2}{\partial\theta\,\partial\theta^T}\ell(\theta_1)(\hat{\theta} - \theta_n)
\end{aligned}$$

Weil θ_1 und θ_2 konsistente Schätzer von θ_n sind und $\theta_n \overset{n\to\infty}{\longrightarrow} \theta_0$, konvergieren $-\frac{1}{n}\frac{\partial^2}{\partial\theta\,\partial\theta^T}\ell(\theta_1)$ und $-\frac{1}{n}\frac{\partial^2}{\partial\theta\,\partial\theta^T}\ell(\theta_2)$ aufgrund des schwachen Gesetz der großen Zahlen gegen $\frac{1}{n}\mathcal{I}_+(\theta_0)$. Wir schreiben damit für T_{LQ}:

$$T_{LQ} \overset{\text{as}}{=} 2\,(\theta_n-\theta_0)^T\frac{\partial}{\partial\theta}\ell(\theta_n) + (\theta_n-\theta_0)^T\mathcal{I}_+(\theta_0)(\theta_n-\theta_0) + (\hat\theta-\theta_n)^T\mathcal{I}_+(\theta_0)(\hat\theta-\theta_n) \tag{2.28}$$

Nun wenden wir (2.27) und (2.22) auf $\frac{\partial}{\partial\theta}\ell(\theta_n)$ an, um $\frac{\partial}{\partial\theta}\ell(\theta_n) = \mathcal{I}_+(\theta_0)(\hat\theta-\theta_n)$ zu erhalten, und setzen $\delta := \lim_{n\to\infty}\sqrt{n}(\theta_n-\theta_0)$.

$$\begin{aligned}
T_{LQ} &\overset{\text{as}}{=} 2\,(\theta_n-\theta_0)^T\mathcal{I}_+(\theta_0)(\hat\theta-\theta_n) + (\theta_n-\theta_0)^T\mathcal{I}_+(\theta_0)(\theta_n-\theta_0) + (\hat\theta-\theta_n)^T\mathcal{I}_+(\theta_0)(\hat\theta-\theta_n)\\
&= (\hat\theta-\theta_n+\theta_n-\theta_0)^T\mathcal{I}_+(\theta_0)(\hat\theta-\theta_n+\theta_n-\theta_0) \qquad (2.29)\\
&\overset{\text{as}}{=} (\hat\theta-\theta_n+\frac{1}{\sqrt{n}}\delta)^T\mathcal{I}_+(\theta_0)^{1/2}\mathcal{I}_+(\theta_0)^{1/2}(\hat\theta-\theta_n+\frac{1}{\sqrt{n}}\delta) \qquad (2.30)\\
&= \left[\mathcal{I}_+(\theta_0)^{1/2}(\hat\theta-\theta_n)+\mathcal{I}_+(\theta_0)^{1/2}\frac{1}{\sqrt{n}}\delta\right]^T\left[\mathcal{I}_+(\theta_0)^{1/2}(\hat\theta-\theta_n)+\mathcal{I}_+(\theta_0)^{1/2}\frac{1}{\sqrt{n}}\delta\right] \qquad (2.31)
\end{aligned}$$

Bei der Umforumg (2.30) nutzten wir aus, daß die Fisher-Information $\mathcal{I}_+(\theta_0)$ positiv definit nach Voraussetzung ist und sich deshalb als Produkt zweier identischer, symmetrischer Matritzen, $\mathcal{I}_+(\theta_0)^{1/2}$, schreiben läßt. Da der Term $\mathcal{I}_+(\theta_0)^{1/2}(\hat\theta-\theta_n)$ in (2.31) eine asymptotische Standardnormalverteilung besitzt, folgt sofort, daß T_{LQ} asymptotisch $\chi^2(\dim\theta_0)$-verteilt ist mit Nichtzentralitätsparameter $\frac{1}{\sqrt{n}}\delta^T\mathcal{I}_+(\theta_0)\frac{1}{\sqrt{n}}\delta$.

Es ist klar, daß der Nichtzentralitätsparameter für $n\to\infty$ verschwindet, weil auch im asymptotischen Grenzfall der wahre Parameter θ_0 in Ω_0, dem Parameterraum unter H_0, liegt.
Wir sehen nun leicht, daß der LM-Test unter den Voraussetzungen des obigen Satzes dieselbe asymptotische Verteilung hat. Denn Formel (2.29) vereinfacht sich zu $(\hat\theta-\theta_0)^T\mathcal{I}_+(\theta_0)(\hat\theta-\theta_0)$, was nach Satz 2.25 asymptotisch gleich T_{LM} ist.

Satz 2.29 *Seien $Y_1,\ldots,Y_n$ iid. Unter der lokalen Alternative $H_1:\theta=\theta_n$ mit $\theta_n \overset{n\to\infty}{\longrightarrow}\theta_0$ und $\sqrt{n}(\theta_n-\theta_0)$ divergent ist der LQ-Test konsistent.*

Beweis: *Der Beweis vom vorangegangegen Satz 2.28 gilt bis Gleichung (2.28) auch hier. Wegen der iid-Annahme ist $\mathcal{I}_+(\theta_0) = n\mathcal{I}_1(\theta_0)$, so daß (2.28) wird zu*

$$T_{LQ} \overset{\text{as}}{=} 2\sqrt{n}(\theta_n-\theta_0)^T\frac{1}{\sqrt{n}}\frac{\partial}{\partial\theta}\ell(\theta_n) + n(\theta_n-\theta_0)^T\mathcal{I}_1(\theta_0)(\theta_n-\theta_0) + n(\hat\theta-\theta_n)^T\mathcal{I}_1(\theta_0)(\hat\theta-\theta_n)$$

Die zweite quadratische Form, $n(\hat\theta-\theta_n)^T\mathcal{I}_1(\theta_0)(\hat\theta-\theta_n)$, auf der rechten Seite ist wegen (2.27) asymptotisch χ^2_q-verteilt und somit in Wahrscheinlichkeit beschränkt. Ein ähnliches Argument gilt für den Ausdruck $\frac{1}{\sqrt{n}}\frac{\partial}{\partial\theta}\ell(\theta_n)$. Denn mit (2.22) haben wir $\frac{\partial}{\partial\theta}\ell(\theta_n) \overset{\text{as}}{=} n\mathcal{I}_1(\theta_n)(\hat\theta-\theta_n)$ und mit (2.27) schließen wir $\frac{1}{\sqrt{n}}\frac{\partial}{\partial\theta}\ell(\theta_n) \overset{as}{\sim} \mathcal{N}_q(\mathbf{0},\mathcal{I}_1(\theta_0))$, so daß der gesamte Term $2\sqrt{n}(\theta_n-\theta_0)^T\frac{1}{\sqrt{n}}\frac{\partial}{\partial(}\ell(\theta)_n)$ in Wahrscheinlichkeit divergiert. Da $\mathcal{I}_1(\theta_0)$ positiv definit ist, haben wir $(\theta_n-\theta_0)^T\mathcal{I}_1(\theta_0)(\theta_n-\theta_0)>0$. Mit der Divergenz von $\sqrt{n}(\theta_n-\theta_0)$ folgern wir $\sqrt{n}(\theta_n-\theta_0)^T\mathcal{I}_1(\theta_0)\sqrt{n}(\theta_n-\theta_0) \overset{n\to\infty}{\longrightarrow}\infty$. Somit haben wir gezeigt, daß $\forall\, d\in\mathbb{R}$ gilt

$$P_{\theta_n}(T_{LQ}>d) \overset{n\to\infty}{\longrightarrow} 1,$$

was ja gerade die behauptete Konsistenz ist.

Zusammengesetzte Hypothesen

Nachdem wir das asymptotische Verhalten des LQ-Tests und des LM-Tests bei einfachen Hypothesen untersucht haben, betrachten wir jetzt ihr asymptotisches Verhalten bei zusammengesetzten Hypothesen. Wir beschränken uns auf solche Hypothesen, bei denen sich der q-dimensionale Parameter θ unter H_0 als Funktion eines niederdimensionalen Parameters β ausdrücken läßt:

$$\theta_j = g_j(\beta_1, \ldots, \beta_k) \quad j = 1, \ldots, q, \quad k < q \quad \text{für } \theta \in \Omega_0 \tag{2.32}$$

Der einfachste Fall liegt vor, wenn wir θ in $\theta = (\psi, \lambda)^T$ aufspalten können, so daß λ ein d-dimensionaler, $1 \leq d < q$, getrennter Nebenparameter ist, der nicht durch Datenreduktion eliminierbar ist, und $\Omega_0 = \{\theta | \psi = \psi_0\}$. Da der ML-Quotient unter Parametertransformation invariant ist, kann die allgemeine Einschränkung (2.32) auf die Gestalt $\theta = (\psi_0, \lambda)^T$ unter H_0 reduziert werden. Die Partitionierung von θ bringt einige neue Schreibweisen mit sich. So lautet die zu testende Hypothese $H_0 : \psi = \psi_0$ gegen $H_1 : \psi \neq \psi_0$. Wir schreiben entsprechend der Partitionierung für die log-Likelihood $\ell(\theta) = \ell(\psi, \lambda)$, ihre partiellen Ableitungen

$$\frac{\partial}{\partial\theta}\ell(\theta) = \left(\frac{\partial}{\partial\psi}\ell(\psi,\lambda), \frac{\partial}{\partial\lambda}\ell(\psi,\lambda)\right)^T \quad \text{und} \quad \frac{\partial^2}{\partial\theta\,\partial\theta^T}\ell(\theta) = \begin{pmatrix} \frac{\partial^2}{\partial\psi\,\partial\psi^T}\ell(\psi,\lambda) & \frac{\partial^2}{\partial\lambda\,\partial\psi^T}\ell(\psi,\lambda) \\ \frac{\partial^2}{\partial\psi\,\partial\lambda}\ell(\psi,\lambda) & \frac{\partial^2}{\partial\lambda\,\partial\lambda^T}\ell(\psi,\lambda) \end{pmatrix}$$

Außerdem wird die Fisher-Information der gesamten Stichprobe zu

$$\mathcal{I}_+(\theta) = \mathcal{I}_+(\psi,\lambda) = \begin{pmatrix} \mathcal{I}_{\psi\psi}(\psi,\lambda) & \mathcal{I}_{\psi\lambda}(\psi,\lambda) \\ \mathcal{I}_{\lambda\psi}(\psi,\lambda) & \mathcal{I}_{\lambda\lambda}(\psi,\lambda) \end{pmatrix}$$

$$\text{und ihre Inverse zu} \quad \mathcal{I}_+^{-1}(\theta) = \begin{pmatrix} i^{\psi\psi}(\psi,\lambda) & i^{\psi\lambda}(\psi,\lambda) \\ i^{\lambda\psi}(\psi,\lambda) & i^{\lambda\lambda}(\psi,\lambda) \end{pmatrix}.$$

An dieser Stelle erinnern wir an den nützlichen Satz zur Invertierung von Blockmatrizen (s. Witting [1995, S. 368]). Sei $J = (J_{il})_{1\leq i,l\leq 2} \in \mathbb{R}^{n\times n}$ eine symmetrische Blockmatrix mit positiv definiten Matrizen J_{22} und $J_{11.2} := J_{11} - J_{12}J_{22}^{-1}J_{21}$. Dann gilt

$$J = \begin{pmatrix} J_{11} & J_{12} \\ J_{21} & J_{22} \end{pmatrix} \qquad J^{-1} = \begin{pmatrix} J_{11.2}^{-1} & -J_{11.2}^{-1}J_{12}J_{22}^{-1} \\ J_{22}^{-1}J_{21}J_{11.2}^{-1} & J_{22}^{-1} + J_{22}^{-1}J_{21}J_{11.2}^{-1}J_{12}J_{22}^{-1} \end{pmatrix} \tag{2.33}$$

wie wir sofort durch Nachrechnen von $JJ^{-1} = E$ bestätigen können.
Unter der Nullhypothese wird λ durch die ML mit $\psi = \psi_0$ fest geschätzt. Wir schreiben $\hat{\lambda}_0$ für den ML-Schätzer von λ bzgl. $\psi = \psi_0$. Wir bemerken, daß der eingeschränkte ML-Schätzer $\hat{\lambda}_0$ die Gleichung $(\frac{\partial}{\partial\psi}\ell(\psi_0, \hat{\lambda}_0), \frac{\partial}{\partial\lambda}\ell(\psi_0, \hat{\lambda}_0))^T = \mathbf{0}$ erfüllt, während für die uneingeschränkten ML-Schätzer $\hat{\psi}$ und $\hat{\lambda}$ gilt $(\frac{\partial}{\partial\psi}\ell(\hat{\psi}, \hat{\lambda}), \frac{\partial}{\partial\lambda}\ell(\hat{\psi}, \hat{\lambda}))^T = \mathbf{0}$. Deshalb können wir (2.21) unter der Nullhypothese $\theta^* = (\psi_0, \lambda)^T$ zum einen anwenden mit dem eingeschränkten ML-Schätzer $\hat{\theta} = (\hat{\psi}, \hat{\lambda})^T$ und zum anderen anwenden mit $\hat{\theta} = (\psi_0, \hat{\lambda}_0)^T$. Dann folgt wegen (2.22) durch Gleichsetzen der beiden Formeln mit der partitionierten Schreibweise

$$\begin{pmatrix} \mathcal{I}_{\psi\psi}(\psi_0,\lambda) & \mathcal{I}_{\psi\lambda}(\psi_0,\lambda) \\ \mathcal{I}_{\lambda\psi}(\psi_0,\lambda) & \mathcal{I}_{\lambda\lambda}(\psi_0,\lambda) \end{pmatrix} \begin{pmatrix} \hat{\psi} - \psi_0 \\ \hat{\lambda} - \lambda \end{pmatrix} \overset{\text{as}}{=} \begin{pmatrix} \mathcal{I}_{\psi\psi}(\psi_0,\lambda) & \mathcal{I}_{\psi\lambda}(\psi_0,\lambda) \\ \mathcal{I}_{\lambda\psi}(\psi_0,\lambda) & \mathcal{I}_{\lambda\lambda}(\psi_0,\lambda) \end{pmatrix} \begin{pmatrix} \mathbf{0} \\ \hat{\lambda}_0 - \lambda \end{pmatrix}$$

$$\Longleftrightarrow \begin{pmatrix} \mathcal{I}_{\psi\psi}(\psi_0,\lambda)(\hat{\psi}-\psi_0)+\mathcal{I}_{\psi\lambda}(\psi_0,\lambda)(\hat{\lambda}-\lambda) \\ \mathcal{I}_{\lambda\psi}(\psi_0,\lambda)(\hat{\psi}-\psi_0)+\mathcal{I}_{\lambda\lambda}(\psi_0,\lambda)(\hat{\lambda}-\lambda) \end{pmatrix} \stackrel{\text{as}}{=} \begin{pmatrix} \mathcal{I}_{\psi\lambda}(\psi_0,\lambda)(\hat{\lambda}_0-\lambda) \\ \mathcal{I}_{\lambda\lambda}(\psi_0,\lambda)(\hat{\lambda}_0-\lambda) \end{pmatrix}$$

Die zweite Zeile des partitionierten Vektors liefert

$$\hat{\lambda}_0-\lambda \stackrel{\text{as}}{=} \hat{\lambda}-\lambda+\mathcal{I}_{\lambda\lambda}^{-1}(\psi_0,\lambda)\mathcal{I}_{\lambda\psi}(\psi_0,\lambda)(\hat{\psi}-\psi_0),$$

so daß für den uneingeschränkten ML-Schätzer $\hat{\lambda}_0$ gilt

$$\hat{\lambda}_0 \stackrel{\text{as}}{=} \hat{\lambda}+\mathcal{I}_{\lambda\lambda}^{-1}(\psi_0,\lambda)\mathcal{I}_{\lambda\psi}(\psi_0,\lambda)(\hat{\psi}-\psi_0). \tag{2.34}$$

Mit der neuen Schreibweise ergibt sich insbesondere aus $\sqrt{n}(\hat{\theta}-\theta^*) \stackrel{as}{\sim} \mathcal{N}_q(\mathbf{0},\mathcal{I}_+^{-1}(\theta^*))$,
$\theta^*=(\psi^*,\lambda^*)^T$ wahrer Parameterwert,
$\sqrt{n}(\hat{\psi}-\psi^*) \stackrel{as}{\sim} \mathcal{N}_{q-d}(\mathbf{0},i^{\psi\psi}(\psi^*,\lambda^*)) = \mathcal{N}_{q-d}(\mathbf{0},[\mathcal{I}_{\psi\psi}(\psi^*,\lambda^*)-\mathcal{I}_{\psi\lambda}(\psi^*,\lambda^*)\mathcal{I}_{\lambda\lambda}^{-1}(\psi^*,\lambda^*)\mathcal{I}_{\lambda\psi}(\psi^*,\lambda^*)]^{-1})$
$\sqrt{n}(\hat{\lambda}-\lambda^*) \stackrel{as}{\sim} \mathcal{N}_d(\mathbf{0},i^{\lambda\lambda}(\psi^*,\lambda^*))$.

Wir zeigen im folgenden, daß sich alle asymptotischen Verteilungseigenschaften des LQ-Tests von den einfachen auf die zusammengesetzten Hypothesen übertragen. Bei den Beweisen benutzen wir die gleichen Argumentationen wie in den entsprechenden Sätzen für einfache Hypothesen.

Satz 2.30 *Ist $H_0:\psi=\psi_0$ wahr, so besitzt der LQ-Test eine asymptotische $\chi^2(\dim\Omega-\dim\Omega_0)$-Verteilung.*

Beweis: *Mit den Taylorentwicklungen*

$$\ell(\psi_0,\lambda)=\ell(\hat{\psi},\hat{\lambda})+\frac{1}{2}\begin{pmatrix}\hat{\psi}-\psi_0\\ \hat{\lambda}-\lambda\end{pmatrix}^T\begin{pmatrix}\frac{\partial^2}{\partial\psi\,\partial\psi^T}\ell(\psi_1,\lambda_1) & \frac{\partial^2}{\partial\lambda\,\partial\psi^T}\ell(\psi_1,\lambda_1)\\ \frac{\partial^2}{\partial\psi\,\partial\lambda^T}\ell(\psi_1,\lambda_1) & \frac{\partial^2}{\partial\lambda\,\partial\lambda^T}\ell(\psi_1,\lambda_1)\end{pmatrix}\begin{pmatrix}\hat{\psi}-\psi_0\\ \hat{\lambda}-\lambda\end{pmatrix}$$

mit $\|(\psi_1-\psi_0,\lambda_1-\lambda)^T\| \le \|(\hat{\psi}-\psi_0,\hat{\lambda}-\lambda)^T\|$ *und*

$$\ell(\psi_0,\lambda)=\ell(\psi_0,\hat{\lambda}_0)+\frac{1}{2}\begin{pmatrix}\mathbf{0}\\ \hat{\lambda}_0-\lambda\end{pmatrix}^T\begin{pmatrix}\frac{\partial^2}{\partial\psi\,\partial\psi^T}\ell(\psi_2,\lambda_2) & \frac{\partial^2}{\partial\lambda\,\partial\psi^T}\ell(\psi_2,\lambda_2)\\ \frac{\partial^2}{\partial\psi\,\partial\lambda^T}\ell(\psi_2,\lambda_2) & \frac{\partial^2}{\partial\lambda\,\partial\lambda^T}\ell(\psi_2,\lambda_2)\end{pmatrix}\begin{pmatrix}\mathbf{0}\\ \hat{\lambda}_0-\lambda\end{pmatrix}$$

mit $\|(\psi_2-\psi_0,\lambda_2-\lambda)^T\| \le \|(\mathbf{0},\hat{\lambda}_0-\lambda)^T\|$ *erhalten wir*

$$\begin{aligned}
T_{LQ} &= 2\,[\ell(\hat{\psi},\hat{\lambda})-\ell(\psi_0,\hat{\lambda}_0)]\\
&= 2\,[\ell(\hat{\psi},\hat{\lambda})-\ell(\psi_0,\lambda)]-2\,[\ell(\psi_0,\hat{\lambda}_0)-\ell(\psi_0,\lambda)]\\
&= -\begin{pmatrix}\hat{\psi}-\psi_0\\ \hat{\lambda}-\lambda\end{pmatrix}^T\begin{pmatrix}\frac{\partial^2}{\partial\psi\,\partial\psi^T}\ell(\psi_1,\lambda_1) & \frac{\partial^2}{\partial\lambda\,\partial\psi^T}\ell(\psi_1,\lambda_1)\\ \frac{\partial^2}{\partial\psi\,\partial\lambda^T}\ell(\psi_1,\lambda_1) & \frac{\partial^2}{\partial\lambda\,\partial\lambda^T}\ell(\psi_1,\lambda_1)\end{pmatrix}\begin{pmatrix}\hat{\psi}-\psi_0\\ \hat{\lambda}-\lambda\end{pmatrix}+\\
&\quad+\begin{pmatrix}\mathbf{0}\\ \hat{\lambda}_0-\lambda\end{pmatrix}^T\begin{pmatrix}\frac{\partial^2}{\partial\psi\,\partial\psi^T}\ell(\psi_2,\lambda_2) & \frac{\partial^2}{\partial\lambda\,\partial\psi^T}\ell(\psi_2,\lambda_2)\\ \frac{\partial^2}{\partial\psi\,\partial\lambda^T}\ell(\psi_2,\lambda_2) & \frac{\partial^2}{\partial\lambda\,\partial\lambda^T}\ell(\psi_2,\lambda_2)\end{pmatrix}\begin{pmatrix}\mathbf{0}\\ \hat{\lambda}_0-\lambda\end{pmatrix}\\
&\stackrel{\text{as}}{=} \begin{pmatrix}\hat{\psi}-\psi_0\\ \hat{\lambda}-\lambda\end{pmatrix}^T\begin{pmatrix}\mathcal{I}_{\psi\psi}(\psi_0,\lambda) & \mathcal{I}_{\psi\lambda}(\psi_0,\lambda)\\ \mathcal{I}_{\lambda\psi}(\psi_0,\lambda) & \mathcal{I}_{\lambda\lambda}(\psi_0,\lambda)\end{pmatrix}\begin{pmatrix}\hat{\psi}-\psi_0\\ \hat{\lambda}-\lambda\end{pmatrix}-\\
&\quad-\begin{pmatrix}\mathbf{0}\\ \hat{\lambda}_0-\lambda\end{pmatrix}^T\begin{pmatrix}\mathcal{I}_{\psi\psi}(\psi_0,\lambda) & \mathcal{I}_{\psi\lambda}(\psi_0,\lambda)\\ \mathcal{I}_{\lambda\psi}(\psi_0,\lambda) & \mathcal{I}_{\lambda\lambda}(\psi_0,\lambda)\end{pmatrix}\begin{pmatrix}\mathbf{0}\\ \hat{\lambda}_0-\lambda\end{pmatrix}
\end{aligned}$$

Bei der letzten Umformung haben wir Formel (2.20) aus Bemerkung b) für die Blockmatrizen angewendet. Nun setzen wir Formel (2.34) in $\begin{pmatrix} \mathbf{0} \\ \hat{\lambda}_0 - \lambda \end{pmatrix}^T \begin{pmatrix} \mathcal{I}_{\psi\psi}(\psi_0,\lambda) & \mathcal{I}_{\psi\lambda}(\psi_0,\lambda) \\ \mathcal{I}_{\lambda\psi}(\psi_0,\lambda) & \mathcal{I}_{\lambda\lambda}(\psi_0,\lambda) \end{pmatrix} \begin{pmatrix} \mathbf{0} \\ \hat{\lambda}_0 - \lambda \end{pmatrix}$ *ein:*

$$\begin{pmatrix} \mathbf{0} \\ \hat{\lambda} - \lambda + \mathcal{I}_{\lambda\lambda}^{-1}(\psi_0,\lambda)\mathcal{I}_{\lambda\psi}(\psi_0,\lambda)(\hat{\psi}-\psi_0) \end{pmatrix}^T \begin{pmatrix} \mathcal{I}_{\psi\psi}(\psi_0,\lambda) & \mathcal{I}_{\psi\lambda}(\psi_0,\lambda) \\ \mathcal{I}_{\lambda\psi}(\psi_0,\lambda) & \mathcal{I}_{\lambda\lambda}(\psi_0,\lambda) \end{pmatrix} \begin{pmatrix} \mathbf{0} \\ \hat{\lambda} - \lambda + \mathcal{I}_{\lambda\lambda}^{-1}(\psi_0,\lambda)\mathcal{I}_{\lambda\psi}(\psi_0,\lambda)(\hat{\psi}-\psi_0) \end{pmatrix}$$

$$\begin{aligned} = \; & (\hat{\lambda}-\lambda)^T\mathcal{I}_{\lambda\lambda}(\psi_0,\lambda)(\hat{\lambda}-\lambda) + 2(\hat{\lambda}-\lambda)^T\mathcal{I}_{\lambda\psi}(\psi_0,\lambda)(\hat{\psi}-\psi_0) + \\ & (\hat{\psi}-\psi_0)^T\mathcal{I}_{\psi\lambda}(\psi_0,\lambda)\mathcal{I}_{\lambda\lambda}^{-1}(\psi_0,\lambda)\mathcal{I}_{\psi\lambda}(\psi_0,\lambda)(\hat{\psi}-\psi_0) \end{aligned}$$

Einsetzen dieser Gleichung in T_{LQ} *liefert*

$$\begin{aligned} T_{LQ} \overset{\text{as}}{=} & \begin{pmatrix} \hat{\psi}-\psi_0 \\ \hat{\lambda}-\lambda \end{pmatrix}^T \begin{pmatrix} \mathcal{I}_{\psi\psi}(\psi_0,\lambda) & \mathcal{I}_{\psi\lambda}(\psi_0,\lambda) \\ \mathcal{I}_{\lambda\psi}(\psi_0,\lambda) & \mathcal{I}_{\lambda\lambda}(\psi_0,\lambda) \end{pmatrix} \begin{pmatrix} \hat{\psi}-\psi_0 \\ \hat{\lambda}-\lambda \end{pmatrix} - \\ & - \begin{pmatrix} \mathbf{0} \\ \hat{\lambda}_0-\lambda \end{pmatrix}^T \begin{pmatrix} \mathcal{I}_{\psi\psi}(\psi_0,\lambda) & \mathcal{I}_{\psi\lambda}(\psi_0,\lambda) \\ \mathcal{I}_{\lambda\psi}(\psi_0,\lambda) & \mathcal{I}_{\lambda\lambda}(\psi_0,\lambda) \end{pmatrix} \begin{pmatrix} \mathbf{0} \\ \hat{\lambda}_0-\lambda \end{pmatrix} \\ \overset{\text{as}}{=} & (\hat{\psi}-\psi_0)^T\mathcal{I}_{\psi\psi}(\psi_0,\lambda)(\hat{\psi}-\psi_0) + 2(\hat{\lambda}-\lambda)^T\mathcal{I}_{\lambda\psi}(\psi_0,\lambda)(\hat{\psi}-\psi_0) \\ & +(\hat{\lambda}-\lambda)^T\mathcal{I}_{\lambda\lambda}(\psi_0,\lambda)(\hat{\lambda}-\lambda) \\ & -(\hat{\lambda}-\lambda)^T\mathcal{I}_{\lambda\lambda}(\psi_0,\lambda)(\hat{\lambda}-\lambda) \\ & -2\,(\hat{\lambda}-\lambda)^T\mathcal{I}_{\lambda\psi}(\psi_0,\lambda)(\hat{\psi}-\psi_0) \\ & -(\hat{\psi}-\psi_0)^T\mathcal{I}_{\psi\lambda}(\psi_0,\lambda)\mathcal{I}_{\lambda\lambda}^{-1}(\psi_0,\lambda)\mathcal{I}_{\psi\lambda}(\psi_0,\lambda)(\hat{\psi}-\psi_0) \\ = & (\hat{\psi}-\psi_0)^T\left[\mathcal{I}_{\psi\psi}(\psi_0,\lambda) - \mathcal{I}_{\psi\lambda}(\psi_0,\lambda)\mathcal{I}_{\lambda\lambda}^{-1}(\psi_0,\lambda)\mathcal{I}_{\lambda\psi}(\psi_0,\lambda)\right](\hat{\psi}-\psi_0) \\ = & (\hat{\psi}-\psi_0)^T[i^{\psi\psi}(\psi_0,\lambda)]^{-1}(\hat{\psi}-\psi_0) \end{aligned} \tag{2.35}$$

Da $i^{\psi\psi}(\psi_0,\lambda)$ *die Varianz-Kovarianzmatrix der asymptotischen Randverteilung von* $\hat{\psi}$ *unter* H_0 *ist und* $\hat{\psi}$ *als ML-Schätzer asymptotisch normalverteilt ist, folgt wie in Satz 2.24, daß der LQ-Test unter* H_0 *eine asymptotische* χ^2*-Verteilung mit* $\dim\psi_0 = \dim\Omega - \dim\Omega_0$ *Freiheitsgraden hat.*

Wir bemerken, daß bei der letzten Umformung zu (2.35) der unbekannte, wahre Parameter λ herausgefallen ist, so daß die asymptotische Verteilung von T_{LQ} unter H_0 unabhängig von λ ist. Diese Eigenschaft des LQ-Tests trifft auch bei lokalen Alternativen $H_1 : \psi = \psi_n$ mit $\psi_n \overset{n\to\infty}{\longrightarrow} \psi_0$ und $\sqrt{n}(\psi_n - \psi_0)$ konvergent zu, wie wir als nächstes beweisen.
Zunächst stellen wir fest, daß $\sqrt{n}(\hat{\psi}-\psi_n, \hat{\lambda}-\lambda)^T \overset{as}{\sim} \mathcal{N}_q(\mathbf{0}, \mathcal{I}_+^{-1}(\psi_0,\lambda))$, weil $\hat{\theta} = (\hat{\psi},\hat{\lambda})^T$ asymptotisch normalverteilt ist und weil mit der Stetigkeit der Fisher-Information sowie $\psi_n \overset{n\to\infty}{\longrightarrow} \psi_0$ gilt $\mathcal{I}_+(\psi_n,\lambda) \overset{\text{as}}{=} \mathcal{I}_+(\psi_0,\lambda)$. Außerdem haben wir damit für jeden konsistenten Schätzer $(\tilde{\psi},\tilde{\lambda})^T$ von $(\psi_0,\lambda)^T$:
$-\frac{1}{n}\begin{pmatrix} \frac{\partial^2}{\partial\psi\,\partial\psi^T}\ell(\tilde{\psi},\tilde{\lambda}) & \frac{\partial^2}{\partial\lambda\,\partial\psi^T}\ell(\tilde{\psi},\tilde{\lambda}) \\ \frac{\partial^2}{\partial\psi\,\partial\lambda^T}\ell(\tilde{\psi},\tilde{\lambda}) & \frac{\partial^2}{\partial\lambda\,\partial\lambda^T}\ell(\tilde{\psi},\tilde{\lambda}) \end{pmatrix} \overset{\text{as}}{=} \frac{1}{n}\begin{pmatrix} \mathcal{I}_{\psi\psi}(\tilde{\psi},\tilde{\lambda}) & \mathcal{I}_{\psi\lambda}(\tilde{\psi},\tilde{\lambda}) \\ \mathcal{I}_{\lambda\psi}(\tilde{\psi},\tilde{\lambda}) & \mathcal{I}_{\lambda\lambda}(\tilde{\psi},\tilde{\lambda}) \end{pmatrix}$
Durch diese Eigenschaften bleiben alle Umformungen von Satz 2.30 bis Gleichung (2.35) gültig.

Satz 2.31 *Wenn $H_1 : \psi = \psi_n$ mit $\psi_n \overset{n\to\infty}{\longrightarrow} \psi_0$ wahr ist und $\sqrt{n}(\psi_n - \psi_0)$ konvergiert, dann ist der LQ-Test asymptotisch nichtzentral-$\chi^2(\dim\Omega - \dim\Omega_0)$-verteilt.*
Beweis: *Wegen der positiven Definitheit können wir $[i^{\psi\psi}(\psi_0,\lambda)]^{-1}$ in ein Produkt aus zwei identischen, symmetrischen Matrizen $[i^{\psi\psi}(\psi_0,\lambda)]^{-1/2}$ zerlegen und formen (2.35) mit $\delta := \lim_{n\to\infty}\sqrt{n}(\psi_n - \psi_0)$ weiter um:*

$$\begin{aligned} T_{LQ} &\overset{\text{as}}{=} (\hat\psi - \psi_0)^T [i^{\psi\psi}(\psi_0,\lambda)]^{-1}(\hat\psi - \psi_0) \\ &= (\hat\psi - \psi_n + \psi_n - \psi_0)^T [i^{\psi\psi}(\psi_0,\lambda)]^{-1/2}[i^{\psi\psi}(\psi_0,\lambda)]^{-1/2}(\hat\psi - \psi_n + \psi_n - \psi_0) \\ &\overset{\text{as}}{=} \{[i^{\psi\psi}(\psi_0,\lambda)]^{-1/2}(\hat\psi - \psi_n) + [i^{\psi\psi}(\psi_0,\lambda)]^{-1/2}\frac{1}{\sqrt{n}}\delta\}^T \\ &\quad \{[i^{\psi\psi}(\psi_0,\lambda)]^{-1/2}(\hat\psi - \psi_n) + [i^{\psi\psi}(\psi_0,\lambda)]^{-1/2}\frac{1}{\sqrt{n}}\delta\} \end{aligned}$$

Aus der asymptotischen Standardnormalverteilung von $[i^{\psi\psi}(\psi_0,\lambda)]^{-1/2}(\hat\psi - \psi_n)$ ergibt sich, daß T_{LQ} eine asymptotische $\chi^2(\dim\Omega - \dim\Omega_0)$-Verteilung mit Nichtzentralitätsparameter $\frac{1}{n}\delta^T i^{\psi\psi}(\psi_0,\lambda)\delta$ besitzt.

Satz 2.32 *Seien $Y_1,\ldots,Y_n$ iid. Der LQ-Test ist konsistent, wenn $H_1 : \psi = \psi_n$ mit $\psi_n \overset{n\to\infty}{\longrightarrow} \psi_0$ wahr ist und $\sqrt{n}(\psi_n - \psi_0)$ divergiert.*
Beweis: *Die iid-Annahme läßt uns die Inverse der Fisher-Information $\mathcal{I}_+^{-1}(\theta)$ der gesamten Stichprobe schreiben als $\mathcal{I}_+^{-1}(\theta) = [n\mathcal{I}_1(\theta)]^{-1} = \frac{1}{n}\mathcal{I}_1^{-1}(\theta) = \frac{1}{n}\mathcal{I}_1^{-1}(\psi,\lambda)$, wobei $\mathcal{I}_1(\theta)$ die Fisher-Information einer einzelnen Beobachtung bezeichnet. Damit erhalten wir $i^{\psi\psi}(\psi,\lambda) = \frac{1}{n}i_1^{\psi\psi}(\psi,\lambda)$ mit $i_1^{\psi\psi}(\psi,\lambda)$ als inverser Blockmatrixanteil von $\mathcal{I}_1^{-1}(\psi,\lambda)$ und weiter aus (2.35)*

$$\begin{aligned} T_{LQ} &\overset{\text{as}}{=} (\hat\psi - \psi_0)^T [i^{\psi\psi}(\psi_0,\lambda)]^{-1}(\hat\psi - \psi_0) \\ &= (\hat\psi - \psi_n + \psi_n - \psi_0)^T n[i_1^{\psi\psi}(\psi_0,\lambda)]^{-1}(\hat\psi - \psi_n + \psi_n - \psi_0) \\ &= \underbrace{n(\hat\psi - \psi_n)^T [i_1^{\psi\psi}(\psi_0,\lambda)]^{-1}(\hat\psi - \psi_n)}_{\overset{as}{\sim}\chi^2_{q-d}} + 2\sqrt{n}(\psi_n - \psi_0)^T \underbrace{[i_1^{\psi\psi}(\psi_0,\lambda)]^{-1}\sqrt{n}(\hat\psi - \psi_n)}_{\overset{as}{\sim}\mathcal{N}_{q-d}(\mathbf{0},[i_1^{\psi\psi}(\psi_0,\lambda)]^{-1})} + \\ &\quad + \underbrace{(\psi_n - \psi_0)^T [i_1^{\psi\psi}(\psi_0,\lambda)]^{-1}(\psi_n - \psi_0)}_{\overset{n\to\infty}{\longrightarrow}\infty} \end{aligned}$$

Indem wir die gleiche Argumentation wie in Satz 2.29 anwenden, haben wir gezeigt, daß T_{LQ} in Wahrscheinlichkeit unbeschränkt ist, und der LQ-Test somit ein konsistenter Test ist.

Schließlich müssen wir noch die asymptotische Äquivalenz von LQ-Test, Wald-Test und LM-Test beweisen. Für den Wald-Test ergibt sie sich sofort aus $i^{\psi\psi}(\psi_0,\lambda) \overset{\text{as}}{=} i^{\psi\psi}(\hat\psi,\hat\lambda)$ und (2.35).

Satz 2.33 *Der LM-Test ist unter $H_0 : \psi = \psi_0$ und unter lokalen Alternativen $H_1 : \psi = \psi_n$ mit $\psi_n \overset{n\to\infty}{\longrightarrow} \psi_0$ und $\sqrt{n}(\psi_n - \psi_0)$ konvergent asymptotisch äquivalent zum LQ-Test.*
Beweis: *Wir benutzen (2.22) mit der Partitionierung von θ:*

$$\begin{pmatrix} \hat\psi - \psi_0 \\ \hat\lambda - \lambda \end{pmatrix} \overset{\text{as}}{=} \begin{pmatrix} i^{\psi\psi}(\psi_0,\lambda) & i^{\psi\lambda}(\psi_0,\lambda) \\ i^{\lambda\psi}(\psi_0,\lambda) & i^{\lambda\lambda}(\psi_0,\lambda) \end{pmatrix} \begin{pmatrix} \frac{\partial}{\partial\psi}\ell(\psi_0,\lambda) \\ \frac{\partial}{\partial\lambda}\ell(\psi_0,\lambda) \end{pmatrix}$$

$$\Longrightarrow \quad \hat{\psi} - \psi_0 \stackrel{\text{as}}{=} i^{\psi\psi}(\psi_0, \lambda) \frac{\partial}{\partial \psi} \ell(\psi_0, \lambda) + i^{\psi\lambda}(\psi_0, \lambda) \frac{\partial}{\partial \lambda} \ell(\psi_0, \lambda)$$

Einsetzen in (2.35) liefert:

$$\begin{aligned}
T_{LQ} &\stackrel{\text{as}}{=} (\hat{\psi} - \psi_0)^T [i^{\psi\psi}(\psi_0, \lambda)]^{-1} (\hat{\psi} - \psi_0) \\
&\stackrel{\text{as}}{=} \left[i^{\psi\psi}(\psi_0, \lambda) \frac{\partial}{\partial \psi} \ell(\psi_0, \lambda) + i^{\psi\lambda}(\psi_0, \lambda) \frac{\partial}{\partial \lambda} \ell(\psi_0, \lambda) \right]^T [i^{\psi\psi}(\psi_0, \lambda)]^{-1} \\
&\quad \left[i^{\psi\psi}(\psi_0, \lambda) \frac{\partial}{\partial \psi} \ell(\psi_0, \lambda) + i^{\psi\lambda}(\psi_0, \lambda) \frac{\partial}{\partial \lambda} \ell(\psi_0, \lambda) \right] \\
&= \left[\frac{\partial}{\partial \psi} \ell(\psi_0, \lambda) + \{i^{\psi\psi}(\psi_0, \lambda)\}^{-1} i^{\psi\lambda}(\psi_0, \lambda) \frac{\partial}{\partial \lambda} \ell(\psi_0, \lambda) \right]^T i^{\psi\psi}(\psi_0, \lambda) \\
&\quad \left[\frac{\partial}{\partial \psi} \ell(\psi_0, \lambda) + \{i^{\psi\psi}(\psi_0, \lambda)\}^{-1} i^{\psi\lambda}(\psi_0, \lambda) \frac{\partial}{\partial \lambda} \ell(\psi_0, \lambda) \right] \\
&= \left[\frac{\partial}{\partial \psi} \ell(\psi_0, \lambda) - \mathcal{I}_{\psi\lambda}(\psi_0, \lambda) \mathcal{I}_{\lambda\lambda}^{-1}(\psi_0, \lambda) \frac{\partial}{\partial \lambda} \ell(\psi_0, \lambda) \right]^T i^{\psi\psi}(\psi_0, \lambda) \\
&\quad \left[\frac{\partial}{\partial \psi} \ell(\psi_0, \lambda) - \mathcal{I}_{\psi\lambda}(\psi_0, \lambda) \mathcal{I}_{\lambda\lambda}^{-1}(\psi_0, \lambda) \frac{\partial}{\partial \lambda} \ell(\psi_0, \lambda) \right] \\
&=: T(\lambda)
\end{aligned}$$

Bei der vorletzten Umformung nutzten wir (2.33) aus, weswegen wir schreiben können

$$\{i^{\psi\psi}(\psi_0, \lambda)\}^{-1} i^{\psi\lambda}(\psi_0, \lambda) = -\{i^{\psi\psi}(\psi_0, \lambda)\}^{-1} i^{\psi\psi}(\psi_0, \lambda) \mathcal{I}_{\psi\lambda}(\psi_0, \lambda) \mathcal{I}_{\lambda\lambda}^{-1}(\psi_0, \lambda) = -\mathcal{I}_{\psi\lambda}(\psi_0, \lambda) \mathcal{I}_{\lambda\lambda}^{-1}(\psi_0, \lambda).$$

Selbstverständlich hängt T von dem unbekannten Wert λ des Nebenparameters ab, aber wegen der Konsistenz des ML-Schätzers $\hat{\lambda}_0$ können wir λ durch $\hat{\lambda}_0$ ersetzen, ohne die asymptotische Gleichheit zu verändern. Mit der Definition von $\hat{\lambda}_0$ gilt $\frac{\partial}{\partial \lambda} \ell(\psi_0, \hat{\lambda}_0) = 0$, so daß wir erhalten

$$T_{LQ} \stackrel{\text{as}}{=} T(\lambda) \stackrel{\text{as}}{=} T(\hat{\lambda}_0) = \left[\frac{\partial}{\partial \psi} \ell(\psi_0, \hat{\lambda}_0) \right]^T i^{\psi\psi}(\psi_0, \hat{\lambda}_0) \left[\frac{\partial}{\partial \psi} \ell(\psi_0, \hat{\lambda}_0) \right] = T_{LM} \tag{2.36}$$

Zusammengesetzte Hypothesen am Rand des Parameterraumes

In den vorangegangenen Abschnitten nahmen wir an, daß der wahre Parameter im Inneren des Parameterraums liegt. Wir betrachten nun den Fall, daß bei dem wahren Parameter $\theta^* \in \mathbb{R}^q$ mit $\theta_i^* \geq 0, i = 1, \ldots, q$, eine Komponente gleich 0 ist. Im folgenden bestimmen wir die asymptotische Verteilung des ML-Schätzers von θ^* unter der Nullhypothese und lokalen Alternativen und zeigen, daß die LM-Tests auch dann asymptotisch χ_1^2-verteilt sind.
Zuvor führen wir die benötigten Annahmen und Bezeichnungen ein, denn die Regularitätsbedingungen aus Abschnitt 2.1.4 gelten nicht länger. Wir setzen o.B.d.A. die erste Komponente $\theta_1^* = 0$. Statt des gesamten Parameterraums Ω beschränken wir uns auf $\Omega_1 \subseteq \Omega$, wobei Ω_1 abgeschlossen und beschränkt ist und den wahren Parameter θ^* enthält. O.B.d.A. nehmen wir an, daß Ω_1 ein Quader ist: $\Omega_1 = \{\theta \in \mathbb{R}^q \mid 0 \leq \theta_i \leq b_i, b_i > 0, i = 1, \ldots, q\}$. Desweiteren ist die unabhängige Stichprobe $Y_1, \ldots, Y_n$ mit den Dichten oder Wahrscheinlichkeitsfunktionen $f_i(Y_i; \theta)$, $i = 1, \ldots, n$, die stetig in θ sind, gegeben. Wir fassen hier jede Funktion $\hat{\theta}(Y_1, \ldots, Y_n)$,

die Werte in Ω_1 annimmt und für die $\ell(\theta) = \sum_{i=1}^{n} \ln f_i(y_i;\theta)$ in Ω_1 ein globales Maximum besitzt, als ML-Schätzer auf. Wegen der Stetigkeit aller f_i in θ gibt es eine solche Funktion. Falls mehrere derartige Funktionen existieren, wählen wir eine aus. Außerdem seien $f_i(Y_i;\theta)$ für fast alle Werte von Y_i zweimal stetig differenzierbar in θ. D_n bezeichne die Menge aller Stichprobenwerte $y_1, \ldots, y_n$, für die die zweiten partiellen Ableitungen von $f(Y;\theta)$ nach θ existieren. Wir setzen voraus, daß $\hat{\theta}$ ein konsistenter Schätzer für θ auf Ω ist. Aus der Abgeschlossenheit und Beschränktheit von Ω_1 folgt sofort, daß $\hat{\theta}$ auf Ω_1 ein gleichmäßig konsistenter Schätzer ist. Es gelte die Gleichheit

$$\int \frac{\partial}{\partial\theta_r} f_i(y;\theta)\,dy = \int \frac{\partial^2}{\partial\theta_r\,\partial\theta_s} f_i(y;\theta)\,dy = 0 \qquad \forall \quad i = 1,\ldots,n \text{ und } \forall\, r,s = 1,\ldots,q. \tag{2.37}$$

Wir nutzen diese Gleichheit und die stetige Differenzierbarkeit von f_i, um zu zeigen, daß die Erwartungswerte der ersten partiellen Ableitungen von $\ell_i(\theta) = \ln f_i(Y_i;\theta)$ existieren:

$$0 = \int \frac{\partial}{\partial\theta_r} f_i(y;\theta)\,dy = \int \frac{\frac{\partial}{\partial\theta_r} f_i(y;\theta)}{f_i(y;\theta)} f_i(y;\theta)\,dy = \int \frac{\partial}{\partial\theta_r} \ln f_i(y;\theta)\, f_i(y;\theta)\,dy = E\left(\frac{\partial}{\partial\theta_r}\ell_i(\theta)\right).$$

Schließlich fordern wir die Existenz von $E(\frac{\partial^2}{\partial\theta\partial\theta^T}\ell_i(\theta)) < \infty$, und von einem $K > 0$, so daß

$$\inf_{\theta\in\Omega_1}\left\{\det\left(-\frac{1}{n}E\left[\frac{\partial^2}{\partial\theta\,\partial\theta^T}\ell_i(\theta)\right]\right)\right\} = K > 0.$$

Damit sichern wir uns die Existenz der Fisher-Information $\mathcal{I}_+(\theta)$. Mit Hilfe von (2.37) zeigen wir die Informationsungleichung:

$$\text{Wegen} \quad \frac{\partial^2}{\partial\theta_r\,\partial\theta_s}\ln f_i(Y;\theta) = \frac{1}{f_i(Y;\theta)}\frac{\partial^2}{\partial\theta_r\,\partial\theta_s} f_i(Y;\theta) - \frac{\partial}{\partial\theta_r}\ln f_i(Y;\theta)\frac{\partial}{\partial\theta_s}\ln f_i(Y;\theta) \tag{2.38}$$

gilt

$$\begin{aligned} E\left[\frac{\partial^2}{\partial\theta_r\,\partial\theta_s}\ln f_i(Y;\theta)\right] &= E\left[\frac{1}{f_i(Y;\theta)}\frac{\partial^2}{\partial\theta_r\,\partial\theta_s} f_i(Y;\theta) - \frac{\partial}{\partial\theta_r}\ln f_i(Y;\theta)\,\frac{\partial}{\partial\theta_s}\ln f_i(Y;\theta)\right] \\ &= \int \frac{1}{f_i(Y;\theta)}\left[\frac{\partial^2}{\partial\theta_r\,\partial\theta_s} f_i(Y;\theta)\right] f_i(Y;\theta)\,dy - E\left[\frac{\partial}{\partial\theta_r}\ln f_i(Y;\theta)\,\frac{\partial}{\partial\theta_s}\ln f_i(Y;\theta)\right] \\ &\overset{(2.37)}{=} -E\left[\frac{\partial}{\partial\theta_r}\ln f_i(Y;\theta)\,\frac{\partial}{\partial\theta_s}\ln f_i(Y;\theta)\right] \end{aligned}$$

Die reelle Matrix $\mathcal{I}_i(\theta) = (\mathcal{I}^i_{rs}(\theta))$ ist als Varianz-Kovarianzmatrix des Zufallsvektors $\frac{\partial}{\partial\theta}\ell_i(\theta)$ symmetrisch, also hermitesch, und positiv semidefinit. Die positive Definitheit von $\mathcal{I}_i(\theta)$ folgern wir mit der Äquivalenz, daß eine beliebige quadratische Matrix genau dann positiv definit (semidefinit) ist, wenn sie hermitesch ist und alle ihre Eigenwerte positiv (nichtnegativ) sind. Wir müssen also zeigen, daß 0 nicht zum Spektrum von $\mathcal{I}_i(\theta)$ gehört. Dazu bringen wir $\mathcal{I}_i(\theta)$ auf Jordan-Normalform. Es ist bekannt, daß die Determinante der Jordan-Normalform gerade das Produkt aller Eigenwerte von $\mathcal{I}_i(\theta)$, gezählt mit den algebraischen Vielfachheiten, ist und mit der Determinante von $\mathcal{I}_i(\theta)$ übereinstimmt. Nun haben wir vorausgesetzt, daß gilt $\det\mathcal{I}_i(\theta) > 0$, weshalb $\mathcal{I}_i(\theta)$ nur positive Eigenwerte besitzt und somit positiv definit ist.

Wir folgern weiter, daß die Fisher-Information der gesamten Stichprobe $\mathcal{I}_+(\theta)$ als Summe positiv definiter Matrizen positiv definit ist und ihre Inverse $\mathcal{I}_+^{-1}(\theta)$ und ihre Wurzel $\mathcal{I}_+^{1/2}(\theta)$ existieren. Außerdem ist jede Hauptuntermatrix $(\mathcal{I}_{rs}(\theta))_{r,s\in J}$, $J \subset \{1,\ldots,q\}$, von $\mathcal{I}_+(\theta)$ positiv definit. Wir stellen ferner an die Fisher-Information $\mathcal{I}_+(\theta)$ die Forderung, daß es ein $\eta > 0$ gibt, so daß $E|\frac{\partial}{\partial\theta_r}\ell(\theta)|^{2+\eta}$ für $r = 1,\ldots,q$ eine beschränkte Funktion von θ in Ω_1 ist. Zuletzt verlangen wir, daß die Voraussetzungen des schwachen Gesetz der großen Zahlen für $\frac{\partial^2}{\partial\theta_r\,\partial\theta_s}\ell(\theta)$, $r,s = 1,\ldots,q$, erfüllt sind.

Satz 2.34 *Sei $\theta^* = (\theta_1^*,\ldots,\theta_q^*)$ der wahre Parameter mit $\theta_1^* = 0$, $0 < \theta_i^* < b_i, \forall\, i = 2,\ldots,q$. Dann konvergiert die Verteilungsfunktion $P((\hat\theta - \theta^*) < \mathbf{t})$ gleichmäßig in $\mathbf{t}$ und in θ^* gegen die gemischte Verteilungsfunktion $\frac{1}{2}F_1(\mathbf{t},\theta^*) + \frac{1}{2}F_2(\mathbf{t},\theta^*)$, wobei F_1 eine q-dimensionale Verteilungsfunktion auf dem Raum $t_1 > 0, t_i \in \mathbb{R}\ \forall i = 2,\ldots,q$ ist. F_1 besitzt dort eine Verteilung, die gleich dem Zweifachen einer $\mathcal{N}_q(\mathbf{0},\mathcal{I}_+^{-1}(\theta^*))$-Verteilung ist. F_2 ist eine $(q-1)$-dimensionale Verteilungsfunktion auf dem Unterraum $t_1 = 0, t_i \in \mathbb{R}\ \forall i = 2,\ldots,q$, so daß die gemeinsame Verteilung von $(\hat\theta_2 - \theta_2^*),\ldots,(\hat\theta_q - \theta_q^*)$ diejenige der Größen*

$$(\hat\theta_i - \theta_i^*) \stackrel{\text{as}}{=} \sum_{s=2}^{q} \sigma_{is}^{(1)} w_s \quad i = 2,\ldots,q$$

ist, wobei $(w_1,\ldots,w_q)^T \sim \mathcal{N}_q(\mathbf{0},\mathcal{I}_+(\theta^*))$ und die Verteilung von $w_2,\ldots,w_q$ gebildet wird unter der Bedingung

$$w_1 - \sum_{r=2}^{q} \mathcal{I}_{1r}(\theta^*) \sum_{s=2}^{q} \sigma_{is}^{(1)} w_s \le 0 \tag{2.39}$$

mit $\sigma_{is}^{(1)}$ als Elemente der Hauptuntermatrix von $\mathcal{I}_+^{-1}(\theta^*)$, die wir durch Streichen der ersten Zeile und ersten Spalte erhalten:

$$(\sigma_{is}^{(1)}) = \begin{pmatrix} \mathcal{I}_{22}^+(\theta^*) & \ldots & \mathcal{I}_{2q}^+(\theta^*) \\ \vdots & & \vdots \\ \mathcal{I}_{q2}^+(\theta^*) & \ldots & \mathcal{I}_{qq}^+(\theta^*) \end{pmatrix}^{-1}.$$

Beweis: *Wegen der Konvergenz des ML-Schätzers $\hat\theta$ liegt $\hat\theta$ f.s. in einer Umgebung von θ^*, die offen bzgl. Ω_1 ist. Da $\theta_i \in]\,0, b_i[$ für $i = 2,\ldots,q$, existieren die partiellen Ableitungen und es gilt*

$$\frac{\partial}{\partial\theta_i}\ell(\hat\theta) = 0 \quad i = 2,\ldots,q. \tag{2.40}$$

Für $\theta_1^* = 0$ ist die rechtsseitige partielle Ableitung definiert, die wir gleich $\frac{\partial}{\partial\theta_1}\ell(\theta^*)$ setzen.
Für alle Stichprobenwerte aus D_n gilt die Taylorentwicklung

$$\frac{\partial}{\partial\theta}\ell(\hat\theta) = \frac{\partial}{\partial\theta}\ell(\theta^*) + \frac{\partial^2}{\partial\theta\,\partial\theta^T}\ell(\tilde\theta)(\hat\theta - \theta^*)$$

mit $\|\tilde\theta - \theta^*\| \le \|\hat\theta - \theta^*\|$, wobei $\|.\|$ eine beliebige Norm auf $\mathbb{R}^q$ ist.

$$\frac{\partial}{\partial\theta}\ell(\hat\theta) \quad = \quad \frac{\partial}{\partial\theta}\ell(\theta^*) + \frac{\partial^2}{\partial\theta\,\partial\theta^T}\ell(\tilde\theta)(\hat\theta - \theta^*)$$

$$
\begin{aligned}
&= \frac{\partial}{\partial\theta}\ell(\theta^*) + \left[\frac{\partial^2}{\partial\theta\,\partial\theta^T}\ell(\tilde{\theta}) + \mathcal{I}_+(\theta^*) - \mathcal{I}_+(\theta^*)\right](\hat{\theta}-\theta^*) \\
&= \frac{\partial}{\partial\theta}\ell(\theta^*) + \left[\frac{\partial^2}{\partial\theta\,\partial\theta^T}\ell(\tilde{\theta}) + \mathcal{I}_+(\theta^*)\right](\hat{\theta}-\theta^*) - \mathcal{I}_+(\theta^*)(\hat{\theta}-\theta^*) \qquad (2.41)
\end{aligned}
$$

$\tilde{\theta}$ ist wegen $\|\tilde{\theta}-\theta^*\| \le \|\hat{\theta}-\theta^*\|$ ein konsistenter Schätzer für θ^*. Außerdem ist $\frac{\partial^2}{\partial\theta\,\partial\theta^T}\ell(\theta)$ stetig wegen (2.38) und der vorausgesetzten zweimaligen stetigen Differenzierbarkeit von f, so daß $\frac{\partial^2}{\partial\theta\,\partial\theta^T}\ell(\tilde{\theta})$ ein konsistenter Schätzer für $\frac{\partial^2}{\partial\theta\,\partial\theta^T}\ell(\theta^*)$ ist. Mit dem schwachen Gesetz der großen Zahlen konvergiert

$$
\begin{aligned}
\frac{1}{n}\frac{\partial^2}{\partial\theta\,\partial\theta^T}\ell(\theta^*) + \frac{1}{n}\mathcal{I}_+(\theta^*) &= \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta\partial\theta^T}\ell_i(\theta^*) - \frac{1}{n}E\left[\frac{\partial^2}{\partial\theta\,\partial\theta^T}\ell(\theta^*)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta\partial\theta^T}\ell_i(\theta^*) - \frac{1}{n}\sum_{i=1}^{n}E\left(\frac{\partial^2}{\partial\theta\partial\theta^T}\ell_i(\theta^*)\right) \xrightarrow{p} 0,
\end{aligned}
$$

weshalb die asymptotische Gleichheit $\frac{\partial^2}{\partial\theta\,\partial\theta^T}\ell(\theta^*) \stackrel{\text{as}}{=} -\mathcal{I}_+(\theta^*)$ gilt. Somit fällt in (2.39) der mittlere Term auf der rechen Seite asymptotisch heraus.

1. Fall: $\hat{\theta}_1 > 0$

Dann existiert die partielle Ableitung in $\hat{\theta}_1$ mit $\frac{\partial}{\partial\theta_1}\ell(\hat{\theta}) = 0$.

Es gilt $\mathbf{0} = \frac{\partial}{\partial\theta}\ell(\hat{\theta})$, und wir haben nach (2.39)

$$
\begin{aligned}
\frac{\partial}{\partial\theta}\ell(\theta^*) &\stackrel{\text{as}}{=} \mathcal{I}_+(\theta^*)(\hat{\theta}-\theta^*) \\
\iff \quad (\hat{\theta}-\theta^*) &\stackrel{\text{as}}{=} \mathcal{I}_+^{-1}(\theta^*)\frac{\partial}{\partial\theta}\ell(\theta^*). \qquad (2.42)
\end{aligned}
$$

Nun gilt $\frac{\partial}{\partial\theta}\ell(\theta^*) \overset{as}{\sim} \mathcal{N}_q(\mathbf{0}, \mathcal{I}_+(\theta^*))$ (für einen Beweis s. Rao [1974, S. 416]), weshalb wir für die Bedingung $\hat{\theta}_1 > 0$ folgern

$$
P(\hat{\theta}_1 > 0) = P((\hat{\theta}_1 - \theta_1^*) > 0) \stackrel{\text{as}}{=} \frac{1}{2}
$$

und weiter für die q-dimensionale Verteilung von $(\hat{\theta}_1 - \theta_1^*)$ unter $\hat{\theta}_1 > 0$, daß ihre Dichte 0 ist für $\hat{\theta}_1 \le 0$ und die Verteilung für $\hat{\theta}_1 > 0$ gegen $2\mathcal{N}_q(\mathbf{0}, \mathcal{I}_+^{-1}(\theta^*))$ konvergiert. Diese asymptotische Verteilung ist gerade das behauptete F_1, und das Mischgewicht $\frac{1}{2}$ die Wahrscheinlichkeit von der Bedingung $\hat{\theta}_1 > 0$.

Die gleichmäßige Konvergenz von θ^* auf dem Unterraum $\theta_1 > 0$ von Ω_1 folgt aus der vorausgesetzten gleichmäßigen Konvergenz von $\hat{\theta}$ zusammen mit der Stetigkeit von $\mathcal{I}_+^{-1}(\theta)\frac{\partial}{\partial\theta}\ell(\theta)$ und der Beschränktheit von $E(|\frac{\partial}{\partial\theta}\ell(\theta)|^{2+\eta})$ in θ.

2. Fall: $\hat{\theta}_1 = 0$

Die ML-Funktion $\hat{\theta}$ bestimmt das globale Maximum von ℓ auf Ω_1. Deshalb gelten weiterhin die Gleichungen (2.40), während wir für die erste Komponente erhalten

$$
\frac{\partial}{\partial\theta_1}\ell(\hat{\theta}) \le 0,
$$

denn das Maximum befindet sich am Rand von Ω_1. Mit Hilfe der Taylorentwicklung (2.39) und $\hat{\theta}_1 = \theta_1^*$ ermitteln wir für die erste Komponente von $\frac{\partial}{\partial\theta}\ell(\hat{\theta})$

$$0 \geq \frac{\partial}{\partial\theta_1}\ell(\hat{\theta}) \stackrel{\text{as}}{=} \frac{\partial}{\partial\theta_1}\ell(\theta^*) - \sum_{r=2}^{q} \mathcal{I}_{1r}^{+}(\theta^*)(\hat{\theta}_r - \theta_r^*). \tag{2.43}$$

Für die Betrachtung der restlichen Komponenten $\frac{\partial}{\partial\theta_i}\ell(\hat{\theta})$, $i = 2, \ldots, q$, ergibt sich aus dem Entfernen der ersten Komponente, daß wir in der Taylorentwicklung (2.39) die Fisher-Information $\mathcal{I}_+(\theta^*)$ durch ihre Hauptuntermatrix, die die erste Zeile und die erste Spalte von $\mathcal{I}_+(\theta^*)$ nicht enthält, ersetzen müssen, damit alle Umformungen bis (2.40) gültig bleiben. Bezeichne

$$\left(\sigma_{rs}^{(1)}\right) := \begin{pmatrix} \mathcal{I}_{22}^{+}(\theta^*) & \ldots & \mathcal{I}_{2q}^{+}(\theta^*) \\ \vdots & & \vdots \\ \mathcal{I}_{q2}^{+}(\theta^*) & \ldots & \mathcal{I}_{qq}^{+}(\theta^*) \end{pmatrix}^{-1}$$

die Hauptuntermatrix von $\mathcal{I}_+^{-1}(\theta^*)$, die durch Streichung der ersten Zeile und ersten Spalte von $\mathcal{I}_+^{-1}(\theta^*)$ entsteht, so schreibt sich (2.40) in der Form

$$\begin{pmatrix} (\hat{\theta}_2 - \theta_2^*) \\ \vdots \\ (\hat{\theta}_q - \theta_q^*) \end{pmatrix} \stackrel{\text{as}}{=} (\sigma_{rs}^{(1)}) \begin{pmatrix} \frac{\partial}{\partial\theta_2}\ell(\theta^*) \\ \vdots \\ \frac{\partial}{\partial\theta_q}\ell(\theta^*) \end{pmatrix}$$

bzw. komponentenweise

$$(\hat{\theta}_r - \theta_r^*) \stackrel{\text{as}}{=} \sum_{s=2}^{q} \frac{\partial}{\partial\theta_s}\ell(\theta^*)\sigma_{rs}^{(1)} \quad \text{für } r = 2, \ldots, q.$$

Setzen wir obige Formeln in (2.41) ein und identifizieren wir w_s mit $\frac{\partial}{\partial\theta_s}\ell(\theta^*), s = 1, \ldots, q$, so folgt sofort die behauptete Bedingung (2.39). Wegen $(w_1, \ldots, w_q)^T = \frac{\partial}{\partial\theta^*}\ell(\,) \stackrel{as}{\sim} \mathcal{N}_q(\mathbf{0}, \mathcal{I}_+(\theta^*))$ unter der Voraussetzung (2.39) besitzen die Komponenten $(\hat{\theta}_r - \theta_r^*)$, $r = 2, \ldots, q$, die behauptete gemeinsame Verteilungsfunktion F_2. Das Mischgewicht für F_2 erhalten wir aus $P(\hat{\theta}_1 = 0) = P((\hat{\theta}_1 - \theta_1^*) = 0) = 1 - P((\hat{\theta}_1 - \theta_1^*) > 0) = 1/2$, denn der betrachtete Parameterraum Ω_1 liegt in dem nichtnegativen Quadranten von $\mathbb{R}^q$.
Die gleichmäßige Konvergenz für θ schließen wir mit der gleichen Begründung wie im 1. Fall.

Satz 2.35 *Seien $Y_1, \ldots, Y_n$ iid. Es gelten die in diesem Unterabschnitt vor Satz 2.34 getroffenen Annahmen außer den Voraussetzungen für das schwache Gesetz der großen Zahlen. Die Komponenten $\theta_2^*, \ldots, \theta_q^*$ haben feste Werte in den offenen Intervallen $0 < \theta_i^* < b_i$, $i = 2, \ldots, q$, während $\theta_1^* = an^{-1/2}$ mit $0 \leq a < a_0$ und $n \in \mathbb{N}$. Dann konvergiert auf diesem Intervall die gemeinsame Verteilung von $\sqrt{n}(\hat{\theta}_1 - \theta_1^*), \ldots, \sqrt{n}(\hat{\theta}_q - \theta_q^*)$ gleichmäßig gegen die gemischte Verteilung von*

$$\alpha F_1(\mathbf{t}) + (1 - \alpha)F_2(\mathbf{t}),$$

wobei $F_1(\mathbf{t})$ eine q-variate Verteilungsfunktion auf dem Raum $t_1 > -a$, $t_i \in \mathbb{R}$, $i = 2, \ldots, q$, ist, deren Verteilung unter der Voraussetzung $\sqrt{n}(\hat{\theta}_1 - \theta_1^*) > -a$ gerade $2\,\mathcal{N}_q(\mathbf{0}, \mathcal{I}^{-1}(\theta^*))$ ist. Dabei

bezeichnet $\mathcal{I}(\theta^*) := \mathcal{I}_i(\theta^*)$ die Fisher-Information einer Beobachtung i, deren Index wir aufgrund der iid-Annahme weglassen. $F_2(\mathbf{t})$ ist eine $(q-1)$-variate Verteilungsfunktion auf dem Raum $t_1 = -a$, $t_i \in \mathbb{R}$, $i = 2, \ldots, q$, so daß wir die gemeinsame Verteilung von $\sqrt{n}(\hat{\theta}_2 - \theta_2^*), \ldots, \sqrt{n}(\hat{\theta}_q - \theta_q^*)$ wie folgt erhalten: Die $\sqrt{n}(\hat{\theta}_i - \theta_i^*)$ sind Lösungen der Gleichungen

$$\begin{aligned} \sqrt{n}(\hat{\theta}_1 - \theta_1^*) &= -a & (2.44)\\ w_i &\overset{\text{as}}{=} -a\mathcal{I}_{i1}(\theta^*) + \sum_{r=2}^{q} \sqrt{n}(\hat{\theta}_r - \theta_r^*)\mathcal{I}_{ir}(\theta^*) \quad i = 2, \ldots, q, & (2.45) \end{aligned}$$

wobei die w_i, $i = 1, \ldots, q$, eine $\mathcal{N}_q(\mathbf{0}, \mathcal{I}(\theta^*))$-Verteilung unter der Bedingung

$$w_1 + a\mathcal{I}_{11}(\theta^*) - \sum_{r=2}^{q} \sqrt{n}(\hat{\theta}_r - \theta_r^*)\mathcal{I}_{ir}(\theta^*) \leq 0 \qquad (2.46)$$

besitzen.

Beweis: *Der Beweis läuft analog zum vorangegangenen Satz. Wir haben wiederum die Taylorentwicklung (2.39) und begründen hier die Asymptotik des Terms $[\frac{\partial^2}{\partial\theta\,\partial\theta^T}\ell(\tilde{\theta}) + \mathcal{I}_+(\theta^*)](\hat{\theta} - \theta^*) \overset{\text{as}}{=} 0$ mit der iid-Annahme der Y_i, dem starken Gesetz der großen Zahlen und der Konsistenz von $\tilde{\theta}$. Aufgrund der Definition von $\hat{\theta}$ bekommen wir die Formeln $\frac{\partial}{\partial\theta_1}\ell(\hat{\theta}) \leq 0$ und $\frac{\partial}{\partial\theta_i}\ell(\hat{\theta}) = 0$ für $i = 2, \ldots, q$. Wir unterscheiden wieder die Fälle, daß $\hat{\theta}_1$ am Rand oder im Inneren von $[0, a_0[$ liegt. Ist $\hat{\theta}_1 > 0$, gilt $\frac{\partial}{\partial\theta_1}\ell(\hat{\theta}) = 0$ und wir folgern wie im 1. Fall von Satz 2.34, daß $(\hat{\theta} - \theta^*) \overset{as}{\sim} \mathcal{N}_q(\mathbf{0}, \mathcal{I}_+^{-1}(\theta^*))$. Wegen der iid-Annahme gilt $\mathcal{I}_+(\theta^*) = n\mathcal{I}(\theta^*)$, so daß wir $\sqrt{n}(\hat{\theta} - \theta^*) \overset{as}{\sim} \mathcal{N}_q(\mathbf{0}, \mathcal{I}^{-1}(\theta^*))$ schreiben können. Wir erhalten α, indem wir die Wahrscheinlichkeit von $\hat{\theta}_1 > 0$ bestimmen:*

$$P(\hat{\theta}_1 > 0) = P(\sqrt{n}(\hat{\theta}_1 - \theta_1^*) > -a) \overset{\text{as}}{=} 1 - \Phi(-a\mathcal{I}_{11}^{1/2}(\theta^*)) =: \alpha$$

Im anderen Fall ist $\hat{\theta}_1 = 0$, so daß gilt $\sqrt{n}(\hat{\theta}_1 - \theta_1^*) = \sqrt{n}(0 - an^{-1/2}) = -a$, womit wir Gleichung (2.42) gezeigt haben. Mit einer Taylorentwicklung von $\frac{\partial}{\partial\theta}\ell(\hat{\theta})$ erhalten wir abermals (2.39) und berechnen den Term

$\mathcal{I}_+(\theta^*)(\hat{\theta} - \theta^*) = n\mathcal{I}(\theta^*)(\hat{\theta} - \theta^*)$

$$= n \begin{pmatrix} \mathcal{I}_{11}(\theta^*) & \ldots & \mathcal{I}_{1q}(\theta^*) \\ \mathcal{I}_{21}(\theta^*) & \ldots & \mathcal{I}_{2q}(\theta^*) \\ \vdots & & \vdots \\ \mathcal{I}_{q1}(\theta^*) & \ldots & \mathcal{I}_{qq}(\theta^*) \end{pmatrix} \begin{pmatrix} -an^{-1/2} \\ \hat{\theta}_2 - \theta_2^* \\ \vdots \\ \hat{\theta}_q - \theta_q^* \end{pmatrix} = n \begin{pmatrix} -an^{-1/2}\mathcal{I}_{11}(\theta^*) + \sum_{r=2}^{q}(\hat{\theta}_r - \theta_r^*)\mathcal{I}_{1r}(\theta^*) \\ \vdots \\ -an^{-1/2}\mathcal{I}_{q1}(\theta^*) + \sum_{r=2}^{q}(\hat{\theta}_r - \theta_r^*)\mathcal{I}_{qr}(\theta^*) \end{pmatrix}$$

Daraus folgern wir für $i = 2, \ldots, q$

$$\frac{1}{\sqrt{n}}\frac{\partial}{\partial\theta_i}\ell(\theta^*) \overset{\text{as}}{=} -a\mathcal{I}_{i1}(\theta^*) + \sum_{r=2}^{q} \sqrt{n}(\hat{\theta}_r - \theta_r^*)\mathcal{I}_{ir}(\theta^*),$$

so daß durch die Festlegung $w_i := \frac{1}{\sqrt{n}}\frac{\partial}{\partial\theta_i}\ell(\theta^*)$, $i = 1, \ldots, q$, die Gleichungen (2.43) bewiesen sind. Für $i = 1$ schließen wir aus der Taylorentwicklung und aus $\frac{\partial}{\partial\theta_1}\ell(\hat{\theta}) \leq 0$ die Ungleichung

(2.46):

$$0 \geq \frac{1}{\sqrt{n}}\frac{\partial}{\partial\theta_1}\ell(\theta^*) + a\mathcal{I}_{11}(\theta^*) - \sum_{r=2}^{q}\mathcal{I}_{1r}(\theta^*)\sqrt{n}(\hat{\theta}_r - \theta_r^*) =: w_1 + a\mathcal{I}_{11}(\theta^*) - \sum_{r=2}^{q}\mathcal{I}_{1r}(\theta^*)\sqrt{n}(\hat{\theta}_r - \theta_r^*)$$

Somit ist die Verteilungsfunktion F_2 für $\hat{\theta}_1 = 0$ gezeigt. Die gleichmäßige Konvergenz sehen wir mit derselben Argumentation wie im vorangegangen Satz.

Nachdem wir die asymptotischen Verteilung des ML-Schätzers $\hat{\theta}$ für den wahren Parameter θ^*, bei dem eine Komponente θ_1^* am Rand des Parameterraums liegt, kennen, müssen wir als nächstes überprüfen, ob und wie sich die asymptotische Verteilung der Tests für die Hypothese $H_0 : \theta_1 = 0$ gegen $H_0 : \theta_1 > 0$ ändert. Es handelt sich also um eine zusammengesetzte Hypothese mit den Bezeichnungen $\psi := \theta_1$ und $\lambda := (\theta_2, \ldots, \theta_q)^T$ aus dem vorherigen Unterabschnitt. Zur Bestimmung der asymptotischen Testverteilungen unterscheiden wir erneut zwischen den Fällen $\hat{\theta}_1 > 0$ und $\hat{\theta}_1 = 0$.
Für $\hat{\theta}_1 > 0$ hat $(\hat{\theta} - \theta^*)$ eine asymptotische $\mathcal{N}_q(\mathbf{0}, \mathcal{I}_+^{-1}(\theta^*))$-Verteilung und erfüllt mit den Annahmen dieses Unterabschnitts sämtliche Voraussetzungen für die Sätze aus dem Unterabschnitt „Zusammengesetzte Hypothesen“. Damit übertragen sich alle Aussagen der Sätze 2.30 bis 2.33 auch für Parameter mit einer Komponente auf dem Rand des Parameterraums, solange deren ML-Schätzer im Inneren liegt.
Falls $\hat{\theta}_1 = 0$, dann stimmen der eingeschränkte ML-Schätzer $\hat{\lambda}_0$, der die Gleichung $\left(\frac{\partial}{\partial\psi}\ell(\psi_0, \hat{\lambda}_0), \frac{\partial}{\partial\lambda}\ell(\psi_0, \hat{\lambda}_0)\right)^T = \mathbf{0}$ mit ψ_0 als Parameterwert unter H_0 erfüllt, und der uneingeschränkte ML-Schätzer $\hat{\lambda}$, der die Gleichung $\left(\frac{\partial}{\partial\psi}\ell(\hat{\psi}, \hat{\lambda}), \frac{\partial}{\partial\lambda}\ell(\hat{\psi}, \hat{\lambda})\right)^T = \mathbf{0}$ erfüllt, wegen $\psi_0 = \theta_1 = 0 = \hat{\theta}_1 = \hat{\psi}$ überein. Deshalb liefert die Teststatistik $T_{LQ} = 2[\ell(\hat{\psi}, \hat{\lambda}) - \ell(\psi_0, \hat{\lambda}_0)]$ des LQ-Tests auch stets den Wert 0 und lehnt die Nullhypothese niemals ab. Mit Satz 2.34 haben wir folglich gezeigt, daß die asymptotisch Verteilung des LQ-Tests unter der Nullhypothese im Punkt 0 Wahrscheinlichkeitsmaß $0,5$ und für positive Werte eine $0,5 - \chi_1^2$-Verteilung besitzt.
Für den LM-Test mit der Teststatistik

$$\begin{aligned} T_{LM} &= \left(\frac{\partial}{\partial\psi}\ell(\psi_0, \hat{\lambda}_0), \frac{\partial}{\partial\lambda}\ell(\psi_0, \hat{\lambda}_0)\right)^T (\mathcal{I}_+(\psi_0, \hat{\lambda}_0))^{-1} \left(\frac{\partial}{\partial\psi}\ell(\psi_0, \hat{\lambda}_0), \frac{\partial}{\partial\lambda}\ell(\psi_0, \hat{\lambda}_0)\right) \\ &= \left(\frac{\partial}{\partial\psi}\ell(\hat{\psi}, \hat{\lambda}), \frac{\partial}{\partial\lambda}\ell(\hat{\psi}, \hat{\lambda})\right)^T (\mathcal{I}_+(\hat{\psi}, \hat{\lambda}))^{-1} \left(\frac{\partial}{\partial\psi}\ell(\hat{\psi}, \hat{\lambda}), \frac{\partial}{\partial\lambda}\ell(\hat{\psi}, \hat{\lambda})\right) \end{aligned}$$

ergibt sich jedoch wegen der Sätze 2.34 und 2.35 sowohl unter der Annahme, daß $H_0 : \theta_1 = \psi_0 = 0$ wahr ist, als auch unter der Annahme, daß die lokale Alternative $H_1 : \theta_1 = \psi_0 = an^{-1/2}$ mit a wie in Satz 2.35 wahr ist, daß $\left(\frac{\partial}{\partial\psi}\ell(\hat{\psi}, \hat{\lambda}), \frac{\partial}{\partial\lambda}\ell(\hat{\psi}, \hat{\lambda})\right)^T \overset{\text{as}}{\sim} \mathcal{N}(0, (\mathcal{I}_+(\hat{\psi}, \hat{\lambda}))^{-1})$, wenn $\hat{\theta}_1 = 0$ ist. Deshalb besitzt der LM-Test unter diesen Annahmen eine asymptotische χ_1^2-Verteilung. Wir weisen darauf hin, daß der LQ-Test und der LM-Test in dieser Situation *nicht* asymptotisch äquivalent sind. Nur für den LM-Test gilt die asymptotische χ^2-Verteilung, unabhängig von dem Wert des ML-Schätzers $\hat{\theta}_1$, falls die Nullhypothese oder die lokale Alternative wahr ist.

2.3.2 Residuenanalyse

Residuen messen die Abweichungen der angepaßten Werte von den gegebenen Werten der Zielvariable. Sie können benutzt werden, um Modellmißspezifizierungen, Ausreißer oder Beobachtungen mit schlechter Anpassung aufzuspüren. Die Residuenanalyse, insbesondere die visuelle Analyse, kann möglicherweise die Art der Mißspezifizierung und Wege, sie zu korrigieren, aufzeigen sowie ein Gefühl für die Größe der Mißspefizierungsauswirkung vermitteln.

Definition 2.36 (Rohresiduum) *Das natürliche Residuum ist das* Rohresiduum

$$r_i = y_i - \hat{\mu}_i \qquad i = 1, \ldots, n,$$

wobei der angepaßte Erwartungswert $\hat{\mu}_i$ der bedingte Erwartungswert $\mu_i = \mu(\mathbf{x}_i^T\beta)$ ausgewertet an der Stelle $\beta = \hat{\beta}$ ist.

Asymptotisch verhält sich dieses Residuum wie $y_i - \mu_i$, denn aus $\hat{\beta} \xrightarrow{p} \beta$ folgt $\hat{\mu}_i \xrightarrow{p} \mu_i$. Im klassischen linearen Regressionsmodell mit normalverteilten Fehlern und konstanter Varianz σ^2 gilt $(Y_i - \mu_i) \sim N(0, \sigma^2)$, so daß das Rohresiduum in großen Stichproben die erwünschte Eigenschaft besitzt, symmetrisch um den Nullpunkt mit konstanter Varianz verteilt zu sein. Jedoch ist $Y_i - \mu_i$ im allgemeinen heteroskedastisch und asymmetrisch. Wenn z. B. $Y_i \sim Poi(\mu)$, dann hat $Y_i - \mu_i$ Varianz μ_i und drittes zentriertes Moment μ_i. Für Zähldaten gibt es kein Residuum mit Erwartungswert 0, konstanter Varianz und symmetrischer Verteilung. Dies führt zu verschiedenen Residuen, die entsprechend den am meisten erwünschten Eigenschaften definiert sind.

Definition 2.37 (Pearson-Residuum) *Eine offensichtliche Korrektur der Heteroskedastizität ist das* Pearson-Residuum

$$p_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\omega}_i}}$$

mit $\hat{\omega}_i$ als Schätzer für die Varianz ω_i von Y_i.

Die Summe der quadrierten Residuen ergibt die in Definition 2.8 definierte Pearson-Statistik. In großen Stichproben bei korrekt gewählter Link- und Varianzfunktion besitzt dieses Residuum den Erwartungswert 0 und die Varianz 1, ist aber asymmetrisch verteilt. Gilt beispielsweise $Y \sim Poi(\mu)$, dann erhalten wir mit $\omega = \mu$: $\quad E[(Y - \mu)/\sqrt{\mu}]^3 = 1/\sqrt{\mu}$.

Definition 2.38 (Devianzresiduum) *Gehört die Verteilung von Y der exponentiellen Familie an, können wir das* Devianzresiduum *verwenden, das durch*

$$d_i = sign(y_i - \hat{\mu}_i)\sqrt{2\{\ell(y_i) - \ell(\hat{\mu}_i)\}}$$

gegeben ist, wobei $\ell(\hat{\mu}_i)$ die logarithmierte Dichte von Y_i ausgewertet bei $\mu_i = \hat{\mu}_i$ und $\ell(y_i)$ die logarithmierte Dichte an der Stelle $\mu_i = y_i$ bezeichnet.

Motiviert wird das Devianzresiduum durch die Tatsache, daß die Summe dieser quadrierten Residuen gerade die in Definition 2.6 definierte Devianz darstellt. Im Falle der Poissonverteilung von Y_i nimmt d_i folgende Gestalt an: $d_i = sign(y_i - \hat{\mu}_i)\sqrt{2\{y_i \ln(y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i)\}}$, wobei $y_i \ln y_i = 0$ gesetzt wird für $y_i = 0$.
McCullagh/Nelder [1989, S. 39] empfehlen den Gebrauch von Devianzresiduen in der Residuenanalyse, weil sie viel besser als die Pearson-Residuen an standardnormalverteilte Residuen herankommen (s. auch Cameron/Trivedi [1998, S. 142]). Die folgende Darstellung behandelt dennoch beliebige Residuen R_i, wenn nicht explizit ein bestimmter Residuentyp angegeben ist.

Die vielleicht erfolgreichste Art, Residuen einzusetzen, ist, sie gegen andere, interessierende Variablen graphisch darzustellen. Solche Graphiken schließen Residuen ein, die gegen geschätzte Werte der Zielvariable aufgetragen werden, um z. B. zu erkennen, ob die Anpassung bei kleinen oder großen Werten der Zielvariable schlecht ist, oder die gegen ausgelassene Regressoren aufgetragen werden, um einen möglichen Zusammenhang festzustellen, oder gegen vorhandene Regressoren, um zu sehen, ob sie in das Modell durch eine andere als die gewählte Transformation eingehen.
Es ist zunächst verlockend, die Residuen gegen die vorhandenen Werte der Zielvariable zu zeichnen, doch eine solche Graphik ist nicht aussagekräftig. Um dies zu verdeutlichen, betrachten wir diese Graphik mit Rohresiduen. Da gilt $Cov(Y_i - \mu_i, Y_i) = Var\, Y_i > 0$, gibt es einen positiven Zusammenhang zwischen $Y_i - \mu_i$ und Y_i. Stattdessen tragen wir die Rohresiduen gegen die angepaßten Erwartungswerte ab und beachten, daß $Cov(Y_i - \mu_i, \mu_i) = 0$ gilt. Alternativ können wir auch die gegebenen Werte von Y_i gegen die angepaßten $\hat{\mu}_i$ auftragen. Wenn wir den Erwartungswert μ_i richtig spezifiziert haben, streuen die Residuen R_i in einem zufälligen Muster um den Wert 0 und mit konstanter Varianz. Typische systematische Abweichungen davon sind zum einen das Auftreten einer Krümmung oder eines (linearen) Trends im Erwartungswert und zum anderen eine systematische Veränderung des Streuungsbereichs mit wachsenden, angepaßten Erwartungswerten. Ein Trend kann mehrere Gründe haben wie beispielsweise eine falsch gewählte Linkfunktion, die falsche Skalierung von einem oder mehreren metrischen Regressoren oder das Fehlen eines quadratischen Terms bei einem Regressor. Wir gehen noch auf die Unterscheidung zwischen diesen Ursachen ein.
Eine andere Art, die allgemeine Anpassung eines Modells graphisch zu beurteilen, besteht in der Untersuchung, wie stark die Residuen normalverteilt sind. Dazu ordnen wir die Devianzresiduen d_i der Größe nach und tragen sie gegen die Ordnungsstatistik einer $N(0,1)$-Stichprobe auf, d. h. d_i gegen

$$rnorm_i = \bar{r} + s_r \Phi^{-1}(\frac{i - 0,5}{n}) \qquad i = 1, \ldots, n$$

mit $\bar{r} = \sum_{i=1}^{n} d_i$ und $s_r^2 = \frac{1}{n-1}\sum_{i=1}^{n} d_i^2$ und Φ^{-1} als Inverse der Standardnormalverteilungsfunktion. Liegen keine Modellabweichungen vor, zeigt diese Graphik angenähert eine Gerade.
Um die Anpassung eines metrischen Regressors x_k aus dem linearen Prädiktor graphisch auszuwerten, tragen wir die Regressorwerte x_{ik} gegen die Residuen R_i auf. Wie bei der Graphik Residuen R_i gegen angepaßte Erwartungswerte $\hat{\mu}_i$ streuen die Residuen R_i zufällig mit konstan-

ter Varianz um 0, falls wir das Modell korrekt spezifiziert haben. Wiederum deutet ein Trend im Residuenmuster auf die falsche Wahl der Linkfunktion oder eine nichtbeachtete Transformation des Regressors x_k hin. Ein Trend kann auch durch eine falsche Transformation eines anderen Regressors $x_j, j \neq k$, der eng mit dem untersuchten x_k korrelliert ist, hervorgerufen werden.
Außerdem haben wir die Möglichkeit, die Varianzfunktion $V(\mu_i)$ eines Modells graphisch zu überprüfen. Dazu zeichnen wir die absoluten Residuen $|R_i|$ gegen die geschätzten Werte $\hat{\mu}_i$. Ist die richtige Varianzfunktion $V(\mu_i)$ gewählt, weist die Graphik keinen Trend auf, während sich eine falsch gewählte Varianzfunktion $V(\mu_i)$ in einem Trend niederschlägt. Ein positiver Trend deutet darauf hin, daß die Varianzfunktion $V(\mu_i)$ zu langsam mit dem Erwartungswert wächst, so daß z. B. die ursprüngliche Wahl $V(\mu_i) = \mu_i$ durch $V(\mu_i) = \mu_i^2$ ersetzt werden sollte. Ein negativer Trend deutet im Gegensatz auf eine zu langsam fallende Varianzfunktion $V(\mu_i)$ bezogen auf den Erwartungswert μ hin. Alternativ können wir die quadrierten Rohresiduen r_i^2 als Schätzer für die Varianz gegen die angepaßten Erwartungswerte $\hat{\mu}_i$ darstellen. Die Punkte in dieser Graphik sollten um die gewählte Varianzfunktion $V(\mu_i)$ streuen. Für das Poissonmodell mit $V(\mu_i) = \mu_i$ bedeutet das eine zufällige Streuung um die Gerade durch den Nullpunkt mit Steigung 1.
In analoger Weise verifizieren wir graphisch die Linkfunktion $g(\mu_i)$. Dabei stellen wir die geschätzten Prädiktoren $\hat{\eta}_i = \mathbf{x}_i^T \beta$ gegen die Werte y_i der Zielvariablen Y_i dar. Bei der richtig gewählten Linkfunktion $g(\mu_i)$ stimmt der funktionelle Zusammenhang in der Graphik mit dieser überein. Da der Prädiktor η_i von der Skalierung der metrischen Regressoren $x_{ik}, k = 1, \ldots, p$, abhängt, können Abweichungen von dem erwarteten Muster neben der falschen Linkfunktion auch auf Fehler in der Skalierung eines Regressors x_k oder mehrerer metrischer Regressoren zurückzuführen sein.

Kapitel 3

Testen auf Überdispersion

Da viele der in Abschnitt 2.2.1 genannten Mißspezifizierungen zu einer Verletzung der bedingten Äquidispersionsannahme führen, stellt die Varianzfunktion einen natürlichen Startpunkt für einen Mißspezifizierungstest dar. Gewöhnlich denken wir an eine spezielle Alternative mit einer allgemeinen Varianzfunktion, die gleichzeitig die Poisson-Varianzfunktion durch eine Parametereinschränkung einschließt. In dieser Situation kann einer der drei üblichen Tests – LQ-Test, Wald-Test, LM-Test – angewendet werden. Ein häufiges Problem bei diesen Tests besteht darin, daß unter der Nullhypothese der wahre Parameter am Rand des Parameterraums liegen kann. Wir haben dieses Problem bereits in Abschnitt 2.2.2(i) kennengelernt, als wir die Auswirkungen eines zufälligen Erwartungswerts im Poissonmodell untersuchten und zeigten, daß der Varianzfunktion ein quadratischer Term $\tau\mu^2$ hinzugefügt wird bezogen auf das gleiche Poissonmodell mit festem Erwartungswert, wobei $\tau > 0$ die Varianz des zufälligen Erwartungswerts bezeichnet. Wollen wir testen, ob ein zufälliger oder ein fester Erwartungswert im Poissonmodell die Daten in angemessener Weise beschreibt, so lautet unsere Hypothese $H_0 : \tau = 0$ gegen die einseitige Alternative $H_1 : \tau > 0$. In Unterabschnitt 2.3.1 „Zusammengesetzte Hypothesen am Rand des Parameterraums“ bewiesen wir, daß die asymptotische Normalverteilung des ML-Schätzers bei diesem Testproblem nicht mehr gilt und daß der LQ-Test keine asymptotische χ^2-Verteilung unter der Nullhypothese oder lokalen Alternativen besitzt. Als praktikable Alternative zu den LQ-Tests erwiesen sich die LM-Tests, deren asymptotische Verteilungseigenschaften erhalten bleiben und die die oft mühselige Berechnung der Likelihoods unter der Alternativhypothese vermeiden.
In diesem Kapitel leiten wir eine LM-Teststatistik zum Testen auf Überdispersion her und geben drei oft verwendete Beispiele für Poissonregressionsmodelle an. Die Teststatistik beruht auf der Darstellung (2.36) in Satz 2.3.3.

3.1 Herleitung der Teststatistik

Es seien die Beobachtungen $y_1, \ldots, y_n$ der unabhängigen Stichprobe $Y_1, \ldots, Y_n$ als Zielvariablen gegeben. Die Verteilung der Y_i gehöre zur exponentiellen Familie (vgl. Formel (2.1)), die wir hier

leicht modifizieren. Wir setzen die Funktion des Skalenparameters $a(\phi) = 1$ und reparametrisieren den Zusammenhang zwischen der Zielvariable Y_i und dem Parameter θ_i, indem wir den Zusammenhang zwischen einer Funktion $d(\theta_i)$ von θ_i und der Zielvariable Y_i einführen, so daß die Dichte von Y_i die Gestalt

$$f(Y_i;\theta_i) = \exp\{d_i(\theta_i)Y_i - b_i(\theta_i) + c_i(Y_i)\} \tag{3.1}$$

hat. Mit den Formeln (2.2) und (2.3) leiten wir wie in Abschitt 2.1.2(i) den Erwartungswert und die Varianz von Y_i her:

$$\begin{aligned} E(Y_i) &= \mu_i = \{d_i^{'}(\theta_i)\}^{-1} b_i^{'}(\theta_i) && (3.2)\\ Var\, Y_i &= \sigma_i^2 = \{d_i^{'}(\theta_i)\}^{-2}[b_i^{''}(\theta_i) - d_i^{''}(\theta_i)E(Y_i)] && (3.3) \end{aligned}$$

wobei $'$ Differentiation nach θ_i bezeichnet. θ_i ist eine Funktion des Regressorvektors $\mathbf{x}_i \in I\!R^p$ und des Parameters β: $\theta_i = \theta_i(\mathbf{x}_i;\beta)$. Für die Alternativhypothese H_1 konstruieren wir nun eine größere Familie, die überdispersierte Modelle enthält und für die $Var\, Y_i \geq \sigma_i^2$ gilt, wobei Gleichheit nur auftritt, wenn H_0 gilt. Dazu seien die θ_i nicht fest, sondern stetige, unabhängige Zufallsvariablen, die wir θ_i^* nennen, mit den folgenden Momenten:

$$\begin{aligned} E(\theta_i^*) &= \theta_i(\mathbf{x}_i;\beta) && (3.4)\\ Var\, \theta_i^* &= \tau k_i(\theta_i) > 0, \quad \tau \in I\!R_0^+,\ k_i \text{ diffenzierbare Funktion von } \theta_i \text{ in } I\!R^+ && (3.5)\\ E[(\theta_i^* - \theta_i)^r] &= \alpha_r \quad \text{mit } \alpha_r = o(\tau) \text{ für } r \geq 3. && (3.6) \end{aligned}$$

Bei gegebenem θ_i^* gehört die bedingte Dichte $f(Y_i;\theta_i)$ von Y_i zur exponentiellen Familie aus (3.1). Folglich ist θ_i^* so konstruiert, daß sich für $\tau \to 0$ das überdispersierte Modell auf (3.1) reduziert.

Als Beispiel für eine solche Modellierung erinnern wir an Abschnitt 2.2.2(i). Dort betrachteten wir ein Poissonmodell mit $\theta_i^* = \theta_i u_i$ und $\theta_i = \exp(\mathbf{x}_i^T\beta)$ als feste Parameter und U_i als Zufallsvariablen mit $E(U_i) = 1$. Im darauffolgenden Beispiel 2.20 wählten wir zum einen $Var\, U_i = \tau$, also unabhängig von $\mathbf{x}_i$, und zum anderen $Var\, U_i = \tau/\theta_i$, also abhängig von $\mathbf{x}_i$ und β, so daß wir Überdispersion auf verschiedene Arten modellierten (s. dazu die späteren Beispiele).

Wie in Abschnitt 2.2.2(i) erhalten wir die Randverteilung von Y_i durch Integration der gemeinsamen Dichte über θ_i^*:

$$f_M(Y_i;\theta_i) = E_{\theta_i^*}[f(Y_i;\theta_i^*)]. \tag{3.7}$$

Um nun einen LM-Test zu der Hypothese

$$H_0 : \tau = 0 \quad \text{gegen} \quad H_1 : \tau > 0$$

zu konstruieren, fordern wir die Gültigkeit der Annahmen aus Unterabschnitt 2.3.1 „Zusammengesetzte Hypothesen am Rand des Parameterraums“ sowie hinreichende Glattheit von $\mu_i, i = 1,\ldots,n$, als Funktionen von β, um die positive Definitheit der Fisher-Information in einer Umgebung von $\tau = 0$ zu garantieren. Außerdem sei der Erwartungswert unter H_0 korrekt spezifiziert.

Wir berechnen zuerst den Zähler $\sum_{i=1}^{n} \frac{\partial}{\partial\tau}\ell_i(\hat{\theta}_i)|_{\tau=0}$ mit $\ell_i(\theta_i) = \ln f(Y_i;\theta_i)$ und mit $\hat{\theta}_i$ als ML-Schätzer von θ_i unter der Bedingung $\tau = 0$ des LM-Tests (Definition 2.23) und dann die Fisher-Information $\mathcal{I}_+(\hat{\beta}, \tau = 0) = \sum_{i=1}^{n} \mathcal{I}(\hat{\beta}, \tau = 0)$ für den Nenner.
Dazu entwickeln wir die Dichten $f(Y_i;\theta_i^*), i = 1,\ldots,n$, in eine Taylorreihe:

$$f(Y_i;\theta_i^*) = f(Y_i;\theta_i)+\frac{\partial}{\partial\theta_i^*}f(Y_i;\theta_i)(\theta_i^*-\theta_i)+\frac{1}{2}\frac{\partial^2}{\partial\theta_i^{*\,2}}f(Y_i;\theta_i)(\theta_i^*-\theta_i)^2+\sum_{r=3}^{\infty}\frac{1}{r!}\frac{\partial^r}{\partial\theta_i^{*\,r}}f(Y_i;\theta_i)(\theta_i^*-\theta_i)^r.$$

Wir erhalten die Dichte f_M der Randverteilung mit (3.7) und benutzen (3.4) bis (3.6) dabei:

$$\begin{aligned}
f_M(Y_i;\theta_i) &= E_{\theta_i^*}[f(Y_i;\theta_i^*)] \\
&= f(Y_i;\theta_i) + 0 + \frac{1}{2}\tau k_i(\theta_i)\frac{\partial^2}{\partial\theta_i^{*\,2}}f(Y_i;\theta_i) + \sum_{r=3}^{\infty}\frac{\alpha_r}{r!}\frac{\partial^r}{\partial\theta_i^{*\,r}}f(Y_i;\theta_i) \\
&= f(Y_i;\theta_i)\left\{1 + \frac{1}{2}\tau k_i(\theta_i)\left[\frac{\partial^2}{\partial\theta_i^{*\,2}}f(Y_i;\theta_i)\right]f^{-1}(Y_i;\theta_i) + \sum_{r=3}^{\infty}\frac{\alpha_r}{r!}\left[\frac{\partial^r}{\partial\theta_i^{*\,r}}f(Y_i;\theta_i)\right]f^{-1}(Y_i;\theta_i)\right\}
\end{aligned}$$

Daraus folgen die log-Likelihoods:

$$\ell_i(\theta_i) = \ln f(Y_i;\theta_i)+\ln\{1+\frac{1}{2}\tau k_i(\theta_i)\left[\frac{\partial^2}{\partial\theta_i^{*\,2}}f(Y_i;\theta_i)\right]f^{-1}(Y_i;\theta_i)+\sum_{r=3}^{\infty}\frac{\alpha_r}{r!}\left[\frac{\partial^r}{\partial\theta_i^{*\,r}}f(Y_i;\theta_i)\right]f^{-1}(Y_i;\theta_i)\}$$

Wir bilden jetzt die partiellen Ableitungen von ℓ_i nach τ an der Stelle $\tau = 0$ und beachten, daß wegen (3.6) die Reihe verschwindet.

$$\begin{aligned}
\frac{\partial}{\partial\tau}\ell_i(\hat{\theta}_i)|_{\tau=0} &= \frac{1}{2}k_i(\hat{\theta}_i)\frac{\partial^2}{\partial\theta_i^{*\,2}}f(Y_i;\hat{\theta}_i)\,f^{-1}(Y_i;\hat{\theta}_i) \\
&= \frac{1}{2}k_i(\hat{\theta}_i)f(Y_i;\hat{\theta}_i)\left\{[d_i'(\hat{\theta}_i)Y_i - b_i'(\hat{\theta}_i)]^2 + d_i''(\hat{\theta}_i)Y_i - b_i''(\hat{\theta}_i)\right\}f^{-1}(Y_i;\hat{\theta}_i) \\
&= \frac{1}{2}k_i(\hat{\theta}_i)\{d_i'(\hat{\theta}_i)\}^2\left\{[Y_i - \{d_i'(\hat{\theta}_i)\}^{-1}b_i'(\hat{\theta}_i)]^2 - \{d_i'(\hat{\theta}_i)\}^{-2}[b_i''(\hat{\theta}_i) - d_i''(\hat{\theta}_i)Y_i]\right\} \qquad (3.8)
\end{aligned}$$

Für den weiteren Konstruktionsverlauf der LM-Teststatistik unterlassen wir es zur besseren Lesbarkeit, die Abhängigkeit der Funktionen b_i, d_i, k_i, ℓ_i von θ_i anzugeben. Somit erhalten wir für den Zähler der Teststatistik:

$$\frac{\partial}{\partial\tau}\ell(\hat{\theta}_i)|_{\tau=0} = \sum_{i=1}^{n}\frac{\partial}{\partial\tau}\ell_i(\hat{\theta}_i)|_{\tau=0} = \sum_{i=1}^{n}\frac{1}{2}k_i\{d_i'\}^2\left\{[Y_i - \{d_i'\}^{-1}b_i']^2 - \{d_i'\}^{-2}[b_i'' - d_i''Y_i]\right\} =: \sum_{i=1}^{n}T_i(\hat{\theta}_i).$$

Zur Berechnung des Nenners partitionieren wir die Fisher-Information

$$\mathcal{I}_+(\beta,\tau) = \begin{pmatrix} \mathcal{I}_{\beta\beta} & \mathcal{I}_{\beta\tau} \\ \mathcal{I}_{\tau\beta} & \mathcal{I}_{\tau\tau} \end{pmatrix}$$

und benutzen Formel (2.33). Die Bestimmung der einzelnen Untermatrizen von $\mathcal{I}_+(\beta,\tau)$ mit Ausnahme von $\mathcal{I}_{\tau\tau}$ erfolgt durch Verwendung der Kettenregel für Ableitungen:

$$\frac{\partial\ell_i}{\partial\beta_j} = \sum_{s=1}^{n}\frac{\partial\ell_i}{\partial\theta_s}\frac{\partial\theta_s}{\partial\beta_j}.$$

Dazu haben wir zunächst die Erwartungswerte $E(-\frac{\partial^2 \ell_i}{\partial \theta_i^2})|_{\tau=0}$ und $E(-\frac{\partial^2 \ell_i}{\partial \theta_i \tau})|_{\tau=0}$ zu ermitteln und führen die Abkürzung $h_i := h_i(\theta_i) = \ln d_i'(\theta_i)$ ein, woraus $h_i' = d_i'' (d_i')^{-1}$, $h_i'' = d_i''' (d_i')^{-1} - (d_i'')^2 (d_i')^{-2}$ und $h_i''' = d_i'''' (d_i')^{-1} - 3 d_i'' d_i''' (d_i')^{-2} + 2(d_i'')^3 (d_i')^{-3}$ folgt. Für $\tau = 0$ hat die Dichte $f_M(Y_i; \theta_i)$ die Gestalt (3.1), weshalb gilt

$$\frac{\partial^2 \ell_i}{\partial \theta_i^2}|_{\tau=0} = d_i'' Y_i - b_i''$$

$$\Longrightarrow \quad E\left(-\frac{\partial^2 \ell_i}{\partial \theta_i^2}\right)|_{\tau=0} = b_i'' - d_i'' E(Y_i)|_{\tau=0} \overset{(3.2)}{=} b_i'' - d_i'' (d_i')^{-1} b_i' = b_i'' - h_i' b_i' \tag{3.9}$$

Die Gleichung (3.8) liefert die erste partielle Ableitung $\frac{\partial}{\partial \tau} \ell_i|_{\tau=0}$, so daß sich für $\frac{\partial^2 \ell_i}{\partial \theta_i \, \partial \tau}|_{\tau=0}$ mit der Produktregel ergibt:

$$\begin{aligned}
\frac{\partial^2 \ell_i}{\partial \theta_i \, \partial \tau}|_{\tau=0} &= \frac{1}{2}[k_i' (d_i')^2 + 2 k_i d_i' d_i''] \left\{ [Y_i - (d_i')^{-1} b_i']^2 + (d_i')^{-2} [d_i'' Y_i - b_i''] \right\} \\
&\quad + \frac{1}{2} k_i (d_i')^2 \left\{ 2[Y_i - (d_i')^{-1} b_i'][-(d_i')^{-2} d_i'' b_i' - (d_i')^{-1} b_i''] \right. \\
&\quad \left. - (d_i')^{-2} [-2(d_i')^{-1} d_i'' (b_i'' - d_i'' Y_i) + b_i''' - d_i''' Y_i] \right\}
\end{aligned}$$

Unter Berücksichtigung von (3.2) und (3.3) folgt

$$\begin{aligned}
E\left(-\frac{\partial^2 \ell_i}{\partial \theta_i \, \partial \tau}\right)|_{\tau=0} &= \frac{1}{2} k_i \left[-2(d_i')^{-1} d_i'' (b_i'' - d_i'' (d_i')^{-1} b_i') + b_i''' - d_i''' (d_i')^{-1} b_i' \right] \\
&= \frac{1}{2} k_i \left[b_i' \left(2(d_i')^{-2} (d_i'')^2 - d_i''' (d_i')^{-1} \right) - 2 b_i'' (d_i')^{-1} d_i'' + b_i''' \right] \\
&= \frac{1}{2} k_i \left[b_i' \{ (h_i')^2 - h_i'' \} - 2 b_i'' h_i' + b_i''' \right]
\end{aligned} \tag{3.10}$$

Zur Bestimmung von $\mathcal{I}_{\tau\tau}$ verwenden wir (3.8):

$$\begin{aligned}
\mathcal{I}_{\tau\tau} &= E\left[\left(\frac{\partial \ell_i}{\partial \tau} \right)^2 \right]|_{\tau=0} = \frac{1}{4} k_i^2 E\left[(d_i')^4 \{ [Y_i - (d_i')^{-1} b_i']^2 - (d_i')^{-2} [b_i'' - d_i'' Y_i] \}^2 \right] \\
&= \frac{1}{4} k_i^2 E\left[(d_i')^4 [Y_i - (d_i')^{-1} b_i']^4 - 2(d_i')^2 [Y_i - (d_i')^{-1} b_i']^2 [b_i'' - d_i'' Y_i] + [b_i'' - d_i'' Y_i]^2 \} \right]
\end{aligned}$$

Wir benötigen hier das dritte und vierte Moment von Y_i. Dazu nehmen wir an, daß die Umkehrfunktion von d_i existiert, und reparametrisieren die Dichte (3.1) zurück auf die ursprüngliche Gestalt (2.1), von der wir die momentenerzeugende Funktion mit ihren ersten vier Ableitungen an der Stelle 0 bestimmen (s. Bickel/Doksum [1977, S. 70 f]). Einsetzen der Momente in obige Gleichung für $\mathcal{I}_{\tau\tau}$ liefert nach einiger Rechnung:

$$\mathcal{I}_{\tau\tau} = \sum_{i=1}^{n} \frac{1}{4} k_i^2 \left[b_i' \{ 5 h_i' h_i'' - 3(h_i')^3 - h_i''' \} + 2\{ b_i' h_i' - b_i'' \}^2 + b_i'' \{ 6(h_i')^2 - 4 h_i'' \} - 4 b_i''' h_i' + b_i'''' \right]$$

Setzen wir nun

$$\begin{aligned}
W_1 &= diag\{ E\left(-\frac{\partial^2 \ell_i}{\partial \theta_i^2} \right)|_{\tau=0},\ i = 1, \ldots, n \} \\
W_2 &= diag\{ E\left(-\frac{\partial^2 \ell_i}{\partial \theta_i \, \partial \tau} \right)|_{\tau=0},\ i = 1, \ldots, n \},
\end{aligned}$$

$U \in \mathbb{R}^{n \times p}$ mit dem (ir)-ten Eintrag $\frac{\partial \theta_i}{\partial \beta_r}$ und bezeichne $\mathbf{1} \in \mathbb{R}^n$ den Einheitsvektor, so erhalten wir die Untermatrizen

$$\mathcal{I}_{\beta\beta} = U^T W_1 U \quad \text{und} \quad \mathcal{I}_{\beta\tau} = U^T W_2 \mathbf{1}.$$

Daraus ergibt sich wegen (2.33) die asymptotische Varianz V^2 des Zählers $\sum_{i=1}^n T_i(\hat{\theta}_i)$

$$\begin{aligned} V^2 &= \mathcal{I}_{\tau\tau} - \mathcal{I}_{\tau\beta}\mathcal{I}_{\beta\beta}^{-1}\mathcal{I}_{\beta\tau} \\ &= \mathcal{I}_{\tau\tau} - \mathbf{1}^T W_2 U (U^T W_1 U) U^T W_2 \mathbf{1} \end{aligned}$$

und der LM-Test

$$S^2 = \frac{[\sum_{i=1}^n T_i(\hat{\theta}_i)]^2}{V^2}$$

mit den in Abschnitt 2.3.1 gezeigten Eigenschaften.
Da der LM-Test S^2 unter H_0 asymptotisch χ_1^2-verteilt ist, besitzt seine Wurzel $S = \frac{\sum_{i=1}^n T_i(\hat{\theta}_i)}{V}$ unter H_0 asymptotisch die Standardnormalverteilung. Die Teststatistik S soll in den folgenden Beispielen betrachtet werden.

3.2 Beispiele

Für die drei folgenden Beispiele gelte, daß $Y_i \sim Poi(\mu_i)$ unter H_0.

3.2.1 Log-lineares Poissonmodell mit additiven zufälligen Effekten

Das erste Modell stellt eine einfache Eweiterung des üblichen log-linearen Poissonmodells mit Erwartungswert μ_i dar, indem wir einen additiven, zufälligen Fehler Z_i einführen. Wir setzen in den Gleichungen (3.4) und (3.5)

$$\theta_i = \ln \mu_i = \mathbf{x}_i^T \beta \quad \theta_i^* = \mathbf{x}_i^T \beta + Z_i \quad k_i(\theta_i) = 1,$$

wobei die Z_i iid Zufallsvariablen mit $E(Z_i) = 0$ und $Var\, Z_i = \tau < \infty, i = 1, \ldots, n$, sind. Wir können uns die Z_i als unbeobachtete, stochastische Regressoren vorstellen. Nach Beispiel 2.1 identifizieren wir in (3.1) die Funktionen $d_i(\theta_i)$ mit θ_i und $b_i(\theta_i)$ mit e^{θ_i}. Durch eine Approximation der Exponentialfunktion und des Logarithmus mittels Taylorpolynomen ersten Grades, also $e^x \approx 1 + x$ bzw. $\ln x \approx x - 1$, bestimmen wir den Erwartungswert und die Varianz im gemischten Modell:

$$\begin{aligned} E_{\theta_i^*}(Y_i) &= E_{\theta_i^*}[E_{Y_i}(Y_i|\theta_i^*)] \overset{(3.2)}{=} E_{\theta_i^*}[(d_i')^{-1} b_i'] = E_{\theta_i^*}[e^{\theta_i^*}] \\ &\approx E_{\theta_i^*}(1 + \theta_i^*) = 1 + \theta_i = 1 + \ln \mu_i \approx 1 + \mu_i - 1 \\ &= \mu_i \\ Var_{\theta_i^*}(Y_i) &= E_{\theta_i^*}[Var_{Y_i}(Y_i|\theta_i^*)] + Var_{\theta_i^*}[E_{Y_i}(Y_i|\theta_i^*)] \\ &\overset{Var_{Y_i} = E_{Y_i}}{=} E_{\theta_i^*}[E_{Y_i}(Y_i|\theta_i^*)] + Var_{\theta_i^*}[e^{\theta_i^*}] \\ &\approx \mu_i + Var_{\theta_i^*}[e^{\theta_i + Z_i}] \end{aligned}$$

$$\begin{aligned} &= \mu_i + (e^{\theta_i})^2 Var_{\theta_i^*}[e^{Z_i}] \\ &\approx \mu_i + \mu_i^2 Var_{\theta_i^*}[1 + Z_i] = \mu_i + \mu_i^2 Var_{\theta_i^*}[Z_i] \\ &= \mu_i + \tau\mu_i^2 \end{aligned}$$

Wir erkennen aus diesen Formeln, daß die Modellierung eines additiven Fehlers angenähert dieselben ersten beiden Momente liefert wie die Modellierung der Überdispersion durch einen multiplikativen Fehler in 2.2.2(i). Da die verwendeten Taylorpolynome den Entwicklungspunkt 0 besitzen, ist die Approximation für kleine Werte von τ gut.
Um die Teststatistik $S = \frac{\sum_{i=1}^n T_i(\hat{\theta}_i)}{V}$ zu berechnen, beachten wir, daß wegen $d_i(\theta_i) = \theta_i$ die erste Ableitung $d_i'(\theta_i) = 1$ ist und alle höheren Ableitungen von d_i sowie sämtliche Ableitungen von h_i gleich 0 sind. Außerdem sind alle Ableitungen von $b_i(\theta_i) = e^{\theta_i}$ gerade wieder $b_i(\theta_i)$, und mit $\theta_i = \ln \mu_i$ ist $b_i(\hat{\theta}_i) = \hat{\mu}_i$. Deshalb erhalten wir für den Zähler nach (3.8)

$$T_i(\hat{\theta}_i) = \frac{1}{2}[(Y_i - \hat{\mu}_i)^2 - \hat{\mu}_i].$$

Für den Nenner ermitteln wir zuerst $E(-\frac{\partial^2 \ell_i}{\partial \theta_i^2})|_{\tau=0} = b_i''(\hat{\theta}_i) = \hat{\mu}_i$ und $E(-\frac{\partial^2 \ell_i}{\partial \theta_i \, \partial \tau})|_{\tau=0} = \frac{1}{2} b_i'''(\hat{\theta}_i) = \frac{1}{2}\hat{\mu}_i$, woraus folgt

$$\begin{aligned} W_1 &= diag\{\hat{\mu}_i, i = 1, \ldots, n\} \quad \text{und} \\ W_2 &= diag\{\frac{1}{2}\hat{\mu}_i, i = 1, \ldots, n\} = \frac{1}{2} W_1. \end{aligned}$$

Weiter gilt

$$\mathcal{I}_{\tau\tau} = \sum_{i=1}^n \frac{1}{4}\left[2\{b_i''(\hat{\theta}_i)\}^2 + b_i''''(\hat{\theta}_i)\right] = \sum_{i=1}^n \frac{1}{4}\left[2\hat{\mu}_i^2 + \hat{\mu}_i\right] = \frac{1}{2}\sum_{i=1}^n \hat{\mu}_i^2 \; + \; \frac{1}{4}\sum_{i=1}^n \hat{\mu}_i.$$

Um nun Formel (2.33) benutzen zu können, muß die Invertierbarkeit von $\mathcal{I}_{\beta\beta} = U^T W_1 U$ in unserem Poissonmodell garantiert sein. Eine hinreichende Bedingung ist die Existenz eines Intercepts. Zur Berechnung von $\mathcal{I}_{\tau\beta}\mathcal{I}_{\beta\beta}^{-1}\mathcal{I}_{\beta\tau}$ bemerken wir vorab, daß W_1 als Diagonalmatrix im Poissonmodell stets positiv semidefinit ist und darum eine Matrix $R \in \mathbb{R}^{n\times n}$ existiert mit $W_1 = R^T R$. Wir schreiben

$$\begin{aligned} \mathcal{I}_{\tau\beta}\mathcal{I}_{\beta\beta}^{-1}\mathcal{I}_{\beta\tau} &= \mathbf{1} W_2 U (U^T W_1 U)^{-1} U^T W_2 \mathbf{1} \\ &= \frac{1}{4}\mathbf{1}^T W_1 U (U^T W_1 U)^{-1} U^T W_1 \mathbf{1} \\ &= \frac{1}{4}\mathbf{1}^T R^T R U [(RU)^T RU]^{-1} (RU)^T R\, \mathbf{1} \end{aligned}$$

Können wir beweisen, daß der Term $RU[(RU)^T RU]^{-1}(RU)^T$ gerade die Einheitsmatrix E_n in $\mathbb{R}^{n\times n}$ darstellt, so ergibt sich

$$\mathcal{I}_{\tau\beta}\mathcal{I}_{\beta\beta}^{-1}\mathcal{I}_{\beta\tau} = \frac{1}{4}\mathbf{1}^T R^T R\, \mathbf{1} = \frac{1}{4}\mathbf{1}^T W_1 \mathbf{1} = \frac{1}{4}\sum_{i=1}^n \hat{\mu}_i.$$

Insgesamt erhalten wir für die Teststatistik S, die wir für dieses Modell P_A nennen:

$$P_A = \frac{\sum_{i=1}^n T_i(\hat{\theta}_i)}{V} = \frac{\frac{1}{2}\sum_{i=1}^n (Y_i - \hat{\mu}_i)^2 - \hat{\mu}_i}{(\frac{1}{2}\sum_{i=1}^n \hat{\mu}_i^2)^{1/2}} = \frac{\sum_{i=1}^n (Y_i - \hat{\mu}_i)^2 - \hat{\mu}_i}{(2\sum_{i=1}^n \hat{\mu}_i^2)^{1/2}}. \tag{3.11}$$

Es steht noch der Beweis der Gleichheit $RU[(RU)^T RU]^{-1}(RU)^T = E_n$ aus:
Es gilt $E_n = RU(RU)^T[(RU)^T RU]^{-1} = [(RU)^T RU]^{-1} RU(RU)^T$
und $E_p = (RU)^T RU[RU(RU)^T]^{-1}$,
wobei E_n bzw. E_p die Einheitsmatrizen im $\mathbb{R}^{n\times n}$ bzw. im $\mathbb{R}^{p\times p}$ bezeichnen. Wir multiplizieren die letzte Gleichung von links mit RU und von rechts mit $(RU)^T$:

$$RU(RU)^T = RU(RU)^T RU[RU(RU)^T]^{-1}(RU)^T = RU[RU(RU)^T]^{-1}(RU)^T RU(RU)^T$$

Nun multiplizieren wir $[RU(RU)^T]^{-1}$ von rechts:

$$E_n = RU(RU)^T[RU(RU)^T]^{-1} = RU[(RU)^T RU]^{-1}(RU)^T.$$

3.2.2 Poissonmodell mit multiplikativen zufälligen Effekten

Wenn wir die unbeobachtete Heterogenität lieber multiplikativ in unser Regressionsmodell einbauen wollen, dann wählen wir

$$\theta_i = \mu_i = \mathbf{x}_i^T\beta \quad \theta_i^* = U_i\mu_i \quad \text{und} \quad E(U_i) = 1,\ Var\, U_i = \tau < \infty$$

mit U_i als positive, iid Zufallsvariablen. Nach Abschnitt 2.2.2(i) gilt mit dieser Modellierung in dem gemischten Modell $E_{\theta_i^*}(Y_i) = \mu_i$ und $Var_{\theta_i^*}(Y_i) = \mu_i + \tau\mu_i^2$ exakt. Obwohl wir in 2.2.2(i) speziell die kanonische Linkfunktion betrachteten, ist der Beweis der beiden Formeln unabhängig von der gewählten Linkfunktion. Wir sehen also, daß die Wahl einer konstanten Varianz für die zufälligen Effekte U_i zu einer quadratischen Varianzfunktion im gemischten Modell führt (vgl. Beispiel 2.20).
Setzen wir $d_i(\theta_i) = \ln\theta_i$, dann ergibt sich wegen $\theta_i = \mu_i \overset{(3.2)}{=} (d_i'(\theta_i))^{-1} b_i'(\theta_i) = \theta_i\, b_i'(\theta_i)$, daß $b_i(\theta_i) = \theta_i$ gilt. Weiter erhalten wir $h_i(\theta_i) = -\ln\theta_i$, $h_i'(\theta_i) = -\frac{1}{\theta_i}$, $h_i''(\theta_i) = \frac{1}{\theta_i^2}$ und $h_i'''(\theta_i) = -2\frac{1}{\theta_i^3}$ sowie $k_i(\theta_i) = \theta_i^2$ mit der quadratischen Varianzfunktion. Damit berechnen wir für $T_i(\hat{\theta}_i)$ nach (3.8)

$$T_i(\hat{\theta}_i) = \frac{1}{2}\left[(Y_i - \hat{\mu}_i)^2 - Y_i\right].$$

Mit (3.10) bekommen wir außerdem für $E(-\frac{\partial^2\ell_i}{\partial\theta_i\,\partial\tau})|_{\tau=0}$:

$$E(-\frac{\partial^2\ell_i}{\partial\theta_i\,\partial\tau})|_{\tau=0} = \frac{1}{2}\hat{\theta}_i^2\left[\{h_i'(\theta_i)\}^2 - h_i''(\theta_i)\right] = \frac{1}{2}\hat{\theta}_i^2\left[\frac{1}{\theta_i^2} - \frac{1}{\theta_i^2}\right] = 0,$$

so daß $W_2 = 0$ und weiter $\mathcal{I}_{\tau\beta}\mathcal{I}_{\beta\beta}^{-1}\mathcal{I}_{\beta\tau} = 0$ ist. Somit vereinfacht sich V^2 wegen (2.33) zu $V^2 = \mathcal{I}_{\tau\tau}$ mit

$$\mathcal{I}_{\tau\tau} = \sum_{i=1}^n \frac{1}{4}\hat{\theta}_i^4\left[5h_i'(\hat{\theta}_i)h_i''(\hat{\theta}_i) - 3\{h_i'(\hat{\theta}_i)\}^3 - h_i'''(\hat{\theta}_i) + 2\{h_i'(\hat{\theta}_i)\}^2\right]$$

$$
\begin{aligned}
&= \sum_{i=1}^{n} \frac{1}{4}\hat{\theta}_i^4 \left[-\frac{5}{\hat{\theta}_i^3} + \frac{3}{\hat{\theta}_i^3} + \frac{2}{\hat{\theta}_i^3} + \frac{2}{\hat{\theta}_i^2} \right] \\
&= \sum_{i=1}^{n} \frac{1}{2}\hat{\theta}_i^2 = \frac{1}{2}\sum_{i=1}^{n} \hat{\mu}_i^2.
\end{aligned}
$$

Es ergibt sich daraus die Teststatistik

$$
P_B = \frac{\sum_{i=1}^n T_i(\hat{\theta}_i)}{V} = \frac{\frac{1}{2}\sum_{i=1}^n (Y_i - \hat{\mu}_i)^2 - Y_i}{(\frac{1}{2}\sum_{i=1}^n \hat{\mu}_i^2)^{1/2}} = \frac{\sum_{i=1}^n (Y_i - \hat{\mu}_i)^2 - Y_i}{(2\sum_{i=1}^n \hat{\mu}_i^2)^{1/2}}. \tag{3.12}
$$

Die Ähnlichkeit der Teststatistiken P_A und P_B fällt sofort auf. Tatsächlich stimmen beide überein, wenn wir ein Poissonmodell mit Intercept und der kanonischen Linkfunktion wählen, denn in Beispiel 2.7 haben wir $\sum_{i=1}^n Y_i = \sum_{i=1}^n \hat{\mu}_i$ gezeigt.

3.2.3 Poissonmodell mit multiplikativen zufälligen Effekten und einer alternativen Varianzfunkion

Eine dritte Art, zusätzliche Variation im Poissonmodell mit stochastischen Fehlern, die multiplikativ eingehen, darzustellen, ist der lineare Varianzansatz $Var_{\theta_i^*}(Y_i) = \mu_i(1+\tau)$. Die Gleichungen (3.4) und (3.5) spezifizieren wir deshalb wir in dem vorangegangenen Beispiel, während wir die Varianz der zufälligen Effekte in (3.5) von den Werten der Regressoren abhängig machen:

$$
\theta_i = \mu_i \quad \theta_i^* = U_i\mu_i \quad \text{mit } E(U_i) = 1,\ Var\, U_I = \tau/\mu_i \text{ mit } \tau < \infty
$$

Da wir im Vergleich zum vorherigen Beispiel nur die Varianzfunktion verändern, setzen wir erneut $d_i(\theta_i) = \ln\theta_i$. Dadurch bleibt die Spezifizierung für $b_i(\theta_i) = \theta_i$ erhalten, so daß wir zur Bestimmung der Teststatistik den Term $W_2 = 0$ unverändert übernehmen. Mit $k_i(\theta_i) = \theta_i$ ergeben sich auch die neuen Terme für $T_i(\hat{\theta}_i)$ und $V^2 = \mathcal{I}_{\tau\tau}$ ohne große Neuberechnung aus dem vorherigen Beispiel, denn dort war $k_i(\theta_i) = \theta_i^2$. Somit ermitteln wir jetzt

$$
\begin{aligned}
T_i(\hat{\theta}_i) &= \frac{1}{2}\frac{1}{\hat{\mu}_i}[(Y_i - \hat{\mu}_i)^2 - Y_i] \\
\mathcal{I}_{\tau\tau} &= \sum_{i=1}^{n} \frac{1}{4}\hat{\mu}_i^2 \frac{2}{\hat{\mu}_i^2} = \frac{n}{2}.
\end{aligned}
$$

Damit bilden wir die Teststatistik

$$
P_C = \frac{\sum_{i=1}^n T_i(\hat{\theta}_i)}{V} = \frac{\frac{1}{2}\sum_{i=1}^n \frac{1}{\hat{\mu}_i}[(Y_i - \hat{\mu}_i)^2 - Y_i]}{\sqrt{n/2}} = \frac{1}{\sqrt{2n}}\sum_{i=1}^{n} \frac{(Y_i - \hat{\mu}_i)^2 - Y_i}{\hat{\mu}_i}. \tag{3.13}
$$

Wenn die μ_i groß werden, konvergiert $\frac{Y_i}{\hat{\mu}_i}$ für $i = 1, \ldots, n$, n fest, in Wahrscheinlichkeit gegen 1, so daß ein auf P_C basierender Test äquivalent zu einem, der auf der Pearson-Statistik $P = \sum_{i=1}^n \frac{(Y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}$ beruht, ist. Diese Äquivalenz rechtfertigt den Gebrauch der Pearson-Statistik als ein Maß für Überdispersion (s. Beispiel 2.9). Sie kann aber auch zur Beurteilung, ob der Erwartungswert $\mu_i = \mu_i(\mathbf{x}_i; \beta)$ richtig spezifiziert wurde, dienen.

Die lineare Parametrisierung der Varianz ist nicht für alle Untersuchungen geeignet, da die unbeobachteten zufälligen Effekte U_i von den bekannten Regressoren abhängen. Wenn es für diese Varianzstruktur keine Rechtfertigung gibt, sollte laut Dean [1992, S. 456] besser das Modell des zweiten Beispiels mit der quadratischen Varianzfunktion verwendet werden.
Dean und Lawless [1989] untersuchten die Güteeigenschaften der drei Teststatistiken mittels der Monte Carlo-Methode. Diese belegen, daß P_B insbesondere bei mittleren Werten von τ eine größere Güte als die Devianz- und die Pearson-Statistik besitzt. Wenn τ wächst, nähert sich die Güte aller drei Tests 1 an. Die Unterschiede zwischen P_B einerseits und der Devianz- bzw. der Pearson-Statistik andererseits sind umso geringer, je kleiner die Auswirkungen der Regressoren sind und je weniger die μ_i's variieren. Da die Devianz- und Pearson-Statistiken nicht speziell als Test für Überdispersion konstruiert worden sind, sondern um die Angemessenheit der Spezifizierung der μ_i innerhalb des Poissonmodells zu beurteilen, verwundert es nicht, daß sie eine geringere Güte besitzen.

Kapitel 4

Datenanalyse

In diesem Kapitel werden mit Hilfe der zuvor beschriebenen Poissonregression und den Tests auf Überdispersion die Anzahl der Schäden für eine bestimmte, eng umgrenzte Fahrzeugart in der Kraftfahrzeughaftpflichtversicherung analysiert. Alle graphischen Darstellungen wurden mit dem Programm S-Plus entworfen, während die Berechnungen mit dem Programm SAS durchgeführt wurden. Für die Poissonregression verwendeten wir die von SAS bereitgestellte Prozedur GENMOD. Die Tests auf Überdispersion wurden selbst in SAS geschrieben, und der Quelltext befindet sich im Anhang.

4.1 Beschreibung der Daten

Als Datensatz stehen die Angaben eines deutschen Versicherungsunternehmens von vier aufeinanderfolgenden Jahren zur Verfügung. Jährlich erhoben wurden dabei die Merkmale Schadenanzahl, Schadenfreiheitsklasse, eine Regionalstruktur, die sich aus dem Zulassungsbezirk des Fahrzeugs ableitet, Alter und Stärke des Fahrzeugs sowie Tarifgruppe. Wir wollen die Schadenanzahl als Zielvariable mittels den anderen Merkmalen als Regressoren erklären. Die Tarifgruppe ist der einfachste Regressor, denn er gibt nur an, ob es sich bei dem Versicherungsnehmer um einen Beamten handelt oder nicht. Die Stärke des Fahrzeugs ist in kW (und nicht in PS) gemessen und beträgt maximal 99 kW. Das Alter der hier betrachteten Fahrzeuge liegt zwischen einem und zwanzig Jahren. Das Merkmal Regionalstruktur gliedert sich in die zwölf Klassen $0, \ldots, 11$. Mit 0 bezeichnen wir die Klasse mit den wenigsten zu erwartenden Schäden, während in Klasse 11 die meisten Schäden zu erwarten sind. Das Bonus-Malus-System besitzt die fünf Schadenfreiheitsklassen $0, 1/2, 1, 2$ und 3. Diese Einteilung basiert auf der Schadenanzahl des Vorjahres mit 0 als Klasse für Versicherungsnehmer mit besonders vielen Vorjahresschäden und als Einstiegsklasse für Neuversicherte und mit 3 als Klasse für unfallfreie, langjährige Versicherungsnehmer. Auf die Umstufungsregeln soll hier nicht weiter eingegangen werden, sondern dafür auf die Rabatte der Beitragssätze für die verschiedenen Schadenfreiheitsklassen. In Schadenfreiheitsklasse 0 wird der volle Beitragssatz (100%) erhoben, während ein Versicherungsnehmer in Schadenfreiheitsklasse 1/2 und 1 noch 70% des anfänglichen Beitragsatz bezahlt. Die Schaden-

freiheitsklasse 2 wird mit 55% des Beitragsatzes veranschlagt und die Schadenfreiheitsklasse 3 noch mit 40%. Die Daten liegen nun so gruppiert vor, daß jede Merkmalskombination der Regressoren höchstens einmal vorkommt. Die beobachtete Schadenanzahl bezieht sich also auf eine Gruppe von Versicherungsnehmern, die bezüglich der erhobenen Regressoren identisch sind, und nicht auf einzelne Individuen. Demnach benötigen wir ein Maß für die Größe jeder Gruppe: die Jahreseinheiten. Während der beobachteten vier Jahre gibt es Versicherungsnehmer, die nur für einen Teil der Zeit einer Regressorkombination angehören (z. B. wegen Umstufung in eine andere Schadenfreiheitsklasse oder wegen Kündigung der Police). Diese Versicherungsnehmer werden bezüglich des Volumenmaßes „Anzahl Risiken“ (Risiko = kleinste versicherbare Einheit) zeitanteilig gezählt. Dies bedeutet, daß z. B. ein halbjähriger Versicherungsnehmer als halbe Jahreseinheit gezählt wird, zwei solche Versicherungsnehmer werden also zusammen wie ein ganzjähriger Versicherungsnehmer behandelt. Wir beachten, daß hierbei für den Schadensprozeß die Homogenität und Unabhängigkeit in der Zeit unterstellt wird, d. h. daß Erwartungswert und Varianz eines halbjährigen Versicherungsnehmers gerade die Hälfte von Erwartungswert und Varianz eines ganzjährigen Versicherungsnehmers sind. Im vorliegenden Datensatz gibt es über eine halbe Millionen Jahreseinheiten, die sich auf 50 842 Regressorkombinationen verteilen.

4.2 Explorative Datenanalyse

Um uns einen ersten Überblick zu verschaffen, wie wir mit einer Poissonregression die beobachteten Schadenanzahlen am besten modellieren können, betrachten wir Graphiken, die einen Regressor auf der x-Achse gegen die beobachteten Schadenanzahlen auf der y-Achse darstellen. Nun beschrieben wir im vorangegangenen Abschnitt, daß uns die gesamte Schadenanzahl S_g einer Risikogruppe bzw. Regressorkombination g vorliegt, so daß wir für die Graphiken eine Größe benötigen, die uns das Vergleichen der Schadenanzahlen aus unterschiedlichen Gruppenvolumina ermöglicht. Darum ziehen wir statt der volumenabhängigen Gesamtschadenanzahl S_g die anschaulichere, volumenbezogene Größe Schadenhäufigkeit $SH_g = \frac{S_g}{v_g}$ heran, wobei v_g die Jahreseinheiten als Volumen einer Gruppe bezeichnet. Die Schadenhäufigkeit SH_g gibt die mittlere Schadenanzahl eines Versicherungsnehmers aus Gruppe g für ein Jahr an. Ausgangspunkt unserer Analyse ist das Poissonmodell mit dem Logarithmus als kanonische Linkfunktion, so daß wir unser Modell wie folgt spezifizieren:

Annahme: $S_g \sim Poi(\mu_g) \qquad g = 1, \ldots, 50\,842$

Spezifizierung des Erwartungswerts μ_g: $\ln \mu_g = \ln v_g + \mathbf{x}_g^T \beta$

mit $\mathbf{x}_g$ als Regressorvektor der Merkmale Schadenfreiheitsklasse, Regionalklasse, Tarifgruppe, Fahrzeugalter und Fahrzeugstärke, deren funktionelle Gestalt noch zu bestimmen ist, und mit $\ln v_g$ als Offset-Variable. Unter einer Offset-Variable verstehen wir einen Regressor mit konstantem Koeffizienten 1. Wir formen die letzte Gleichung um zu

$$\ln \frac{\mu_g}{v_g} = \mathbf{x}_g^T \beta.$$

Aus dieser Formel wird deutlich, daß die logarithmierten Schadenhäufigkeiten SH_g als Beobachtungen der unbekannten Größen $\ln \frac{\mu_g}{v_g}$ in den Graphiken auf der y-Achse gegen die einzelnen Regressoren aufzutragen sind.

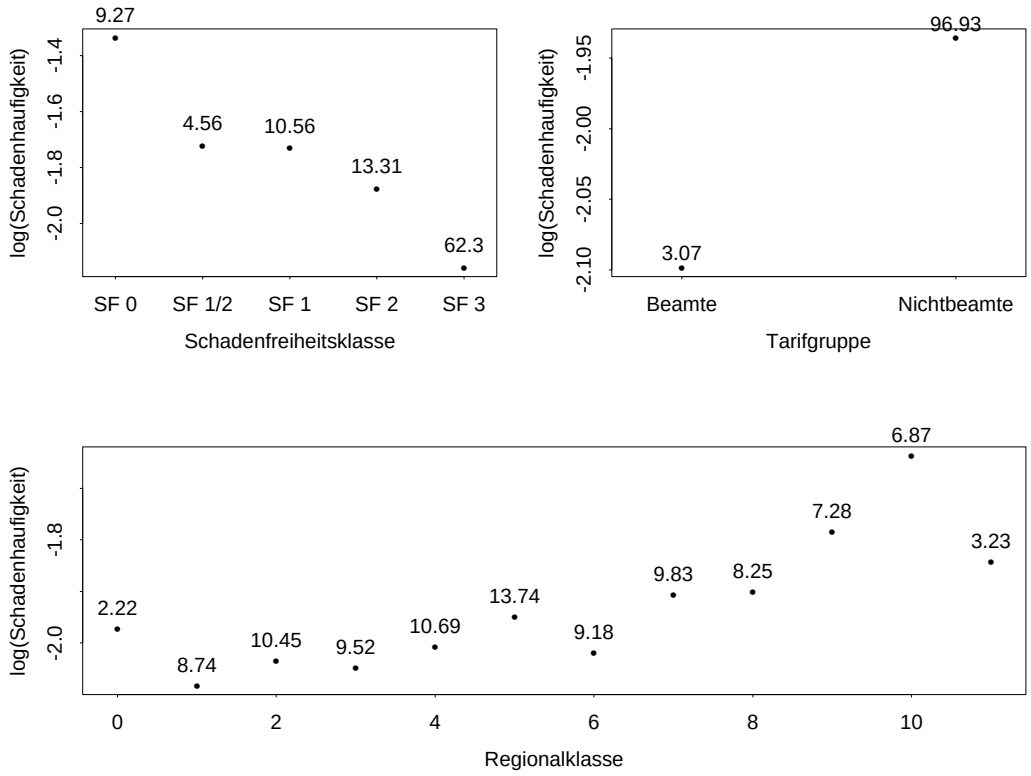


Abbildung 4.1: Explorative Datenanalyse für die kategorielle Regressoren Schadenfreiheitsklasse, Tarifgruppe, Regionalstruktur (Zahlen über den Punkten bezeichnen die Jahreseinheiten in Prozent)

Anhand der explorativen Datenanlyse sehen wir zunächst, daß Versicherungsnehmer, die als Beamte eingestuft sind, eine niedrigere Schadenhäufigkeit als Nichtbeamte aufweisen. Diesen Unterschied (beachte Maßstab) als signifikant zu klassifizieren, ist aufgrund der wenigen Jahreseinheiten für Beamte mit Vorsicht zu bewerten. Die Graphik für die Schadenfreiheitsklassen gibt gut sichtbar die Struktur der Beitragsermäßigung in den Schadenfreiheitklassen wider. Bestimmen wir die relativen Schadenhäufigkeiten SH(Klasse i)/SH(Klasse 0), $i = 0, 1/2, 1, 2, 3$, so decken sich die empirischen Werte 100%, 68%, 67%, 58% und 44% mit den tatsächlich tarifierten Werten 100%, 70%, 70%, 55% und 40% im Rahmen der statistischen Ungenauigkeit. Insbesondere belegt die Graphik die Gleichheit der Rabatte für die Schadenfreiheitsklassen 1/2 und 1. Auch die aufsteigenden Punkte in der Graphik für die Regionalklassen spiegelt deren Einteilung wider. Der zu niedrige Wert für die Schadenanzahl in Regionalklasse 11 und der zu hohe Wert in Regionalklasse 0 können mit den wenigen Jahreseinheiten erklärt werden, die dahinter stecken, so daß sich ungewöhnlich gut bzw. schlecht verlaufende, einzelne Schadensentwicklungen besonders deutlich in Form einer Schadenhäufigkeitssenkung niederschlagen.

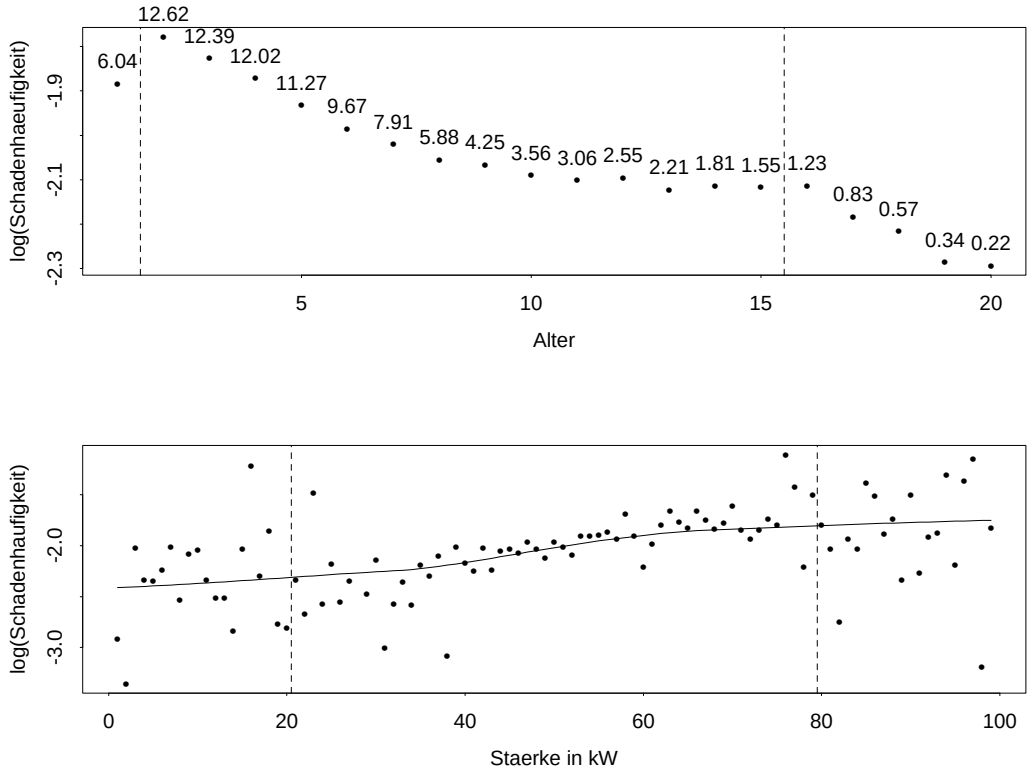


Abbildung 4.2: Explorative Datenanalyse für die stetige Regressoren Fahrzeugalter und Fahrzeugstärke
(Zahlen über den Punkten bezeichnen die Jahreseinheiten in Prozent)

Obwohl die Ausprägungen der Regressoren Fahrzeugalter und -stärke in diskreten, äquidistanten Werten vorliegen, wollen wir sie in der Analyse aufgrund ihrer Bedeutung als stetig auffassen. Die Graphik für den Regressor Fahrzeugalter weist eine Gliederung des Regressors in drei Abschnitte auf, die durch die senkrechten, gestrichelten Linien verdeutlicht wird. Demnach sollten wir das Fahrzeugalter in drei neue Variablen zerlegen: die erste ist eine Indikatorfunktion für Alter=1; die zweite beschreibt den Altersbereich zwischen 2 und 15 Jahren und deutet auf eine Transformation, die zur Potenzfamilie gehört. Die dritte neue Altersvariable weist auf eine andere Potenztransformation des Fahrzeugalters über 15 Jahren hin. Für den Regressor Stärke treten die Verhältnisse nicht derart offen zutage. In den Randbereichen bis 20 kW und über 80 kW streuen die logarithmierten Schadenhäufigkeiten wesentlich stärker als im mittleren Bereich von 20 kW bis 80 kW, was durch die gestrichelten Linien in der unteren Graphik von Abb. 4.2 hervorgehoben wird. Dieses Verhalten kann erneut mit den wenigen Jahreseinheiten, auf denen diese Werte basieren, erklärt werden: 1% der Jahreseinheiten gehören zum Stärkebereich unter 20 kW und 3% der Jahreseinheiten zum Stärkebereich über 80 kW. Für den mittleren Bereich verhält sich die Fahrzeugstärke linear zu den logarithmierten Schadenhäufigkeiten. Ob für den gesamten Wertebereich der Fahrzeugstärke eine Transformation des Regressors nötig ist, soll eine glättende Funktion, die auf einer Folge von lokalen Regressionsmodellen beruht, verdeutlichen.

Dazu wurde die Funktion *lowess*() verwendet. Ihr Graph ist in der Abbildung als Linie eingezeichnet und läßt eher einen nichtlinearen Zusammenhang zwischen Stärke und logarithmieren Schadenhäufigkeiten vermuten.

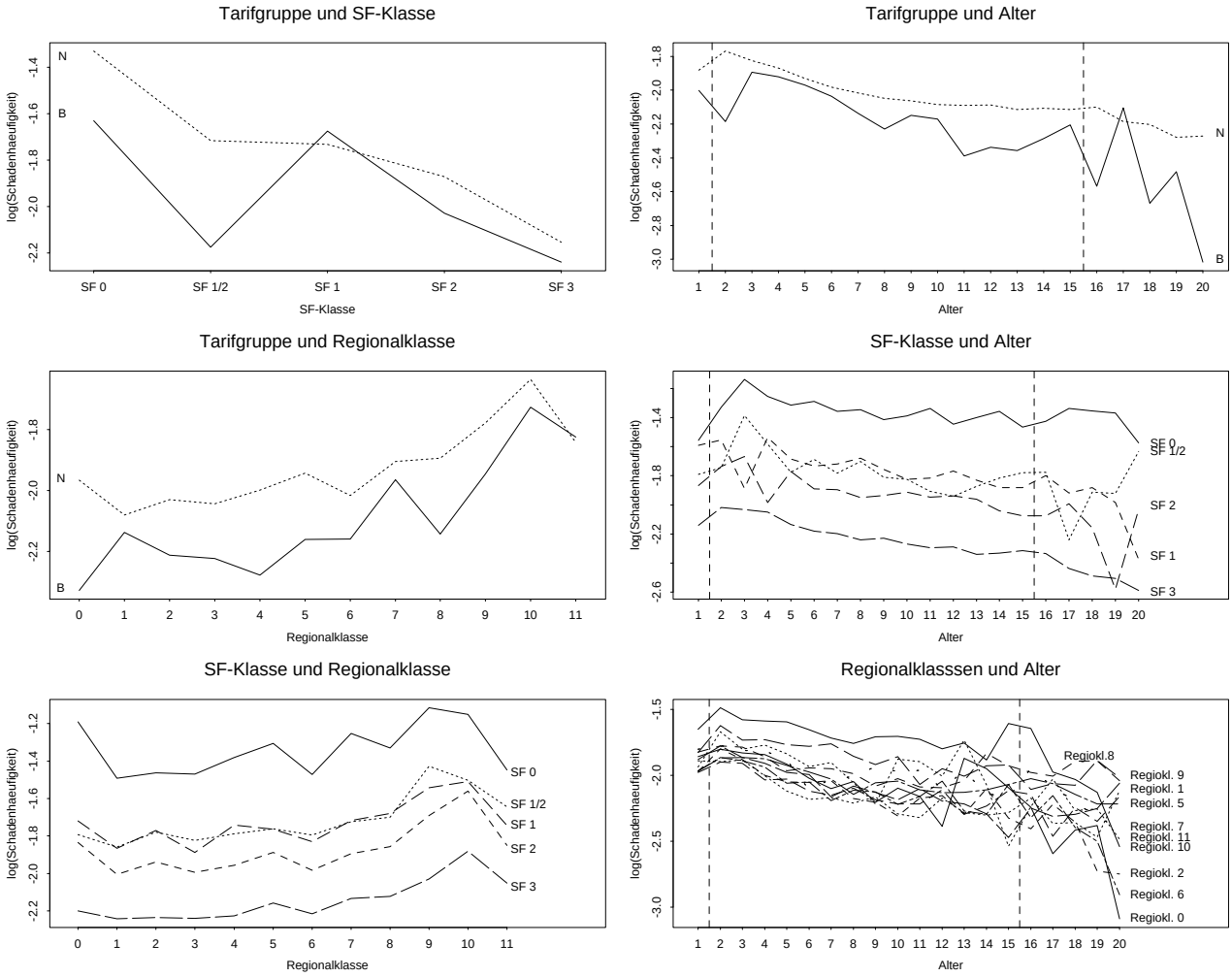


Abbildung 4.3: Interaktionen
N = Nichtbeamte, B = Beamte

Betrachten wir nun die Graphiken aller möglichen Zweifach-Interaktionen. Wir erklären zunächst die Interpretation der Interaktionsgraphiken an einem einfachen Modell. Dazu betrachten wir die Regressionsgleichung mit Zielvariable Y und den stetigen Regressoren x_1 und x_2 sowie der Identität als Linkfunktion:

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

Tragen wir in diesem Modell die Werte des Regressors x_2 gegen $E(Y)$ für feste Werte von x_1 auf, so erhalten wir eine Gerade mit y-Achsenabschnitt $\beta_0 + \beta_1 x_1$ und Steigung β_2. Verändern wir nun den Wert des Regressors x_1, dann behält der Graph von x_2 gegen $E(Y)$ seine Steigung β_2, nur der y-Achsenabschnitt wird von den Änderungen in x_1 beeinflußt. Bei den Graphen für

verschiedene Werte von x_1 handelt es sich also um parallele Linien. Betrachten wir jetzt das gleiche Regressionsmodell mit dem zusätzlichen Interaktionsterm x_1x_2:
$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$.
Die Gleichung formen wir um zu

$$E(Y) = \beta_0 + \beta_1 x_1 + (\beta_2 + \beta_3 x_1) x_2.$$

Wenn wir hier für einen festen Wert von x_1 die Werte des Regressors x_2 gegen $E(Y)$ auftragen, haben wir wiederum den y-Achsenabschnitt $\beta_0 + \beta_1 x_1$, aber die Steigung $\beta_2 + \beta_3 x_1$ hängt jetzt von dem Wert des anderen Regressors x_1 ab. Dementsprechend verändert sich bei verschiedenen Werten von x_1 sowohl der y-Achsenabschnitt als auch die Steigung des Graphen von x_2 gegen $E(Y)$. Somit sind die Graphen für unterschiedliche Werte von x_1 nicht mehr parallel.
Wir nutzen das unterschiedliche Steigungsverhalten in den beschriebenen Modellen mit und ohne Interaktionsterm in der explorativen Datenanalyse, um aus der Graphik für beobachtete Werte der Zielvariable gegen einen bestimmten Regressor bei vorgegebenen Werten eines anderen Regressors auf die Existenz von Interaktionstermen zu schließen: Weisen die Graphen im Rahmen der statistischen Ungenauigkeit Parallelität auf, so liegt keine Interaktion vor. Zeigen die Graphen ein nichtparalleles Muster, dann werden wir die Interaktion im Modell als weiteren Regressor berücksichtigen. Die Betrachtung solcher Zweifach-Interaktionsgraphiken für stetige Regressoren kann nicht für alle Werte der Regressoren erfolgen. Um das Problem der repräsentativen Auswahl einiger Werte zu umgehen, bündeln wir die Ausprägungen in wenige Gruppen, so daß wir den stetigen Regressor wie einen kategoriellen behandeln.
In unserem Datensatz gibt es zwei stetige Regressoren. Indem wir das Fahrzeugalter mit seinen 20 Ausprägungen auf die x-Achse auftragen, vermeiden wir eine Gruppierung. Lediglich die Werte der Fahrzeugstärke müssen wir in größere Gruppen zusammenfassen. Eine naheliegende Einteilung besteht aus einer Gruppierung mit etwa gleich vielen Jahreseinheiten in einer Gruppe, um die zufälligen Streuungseffekte in allen Gruppen gleich groß zu halten. Die Gliederung in sechs Gruppen (bis 42 kW, bis 45 kW, bis 51 kW, bis 56 kW, bis 58 kW, bis 99 kW) scheint ein guter Kompromiß zu sein, um einerseits nicht zuviel Stetigkeit zu aufzugeben und um andererseits nicht zu wenig Volumen in den Gruppen zu haben.

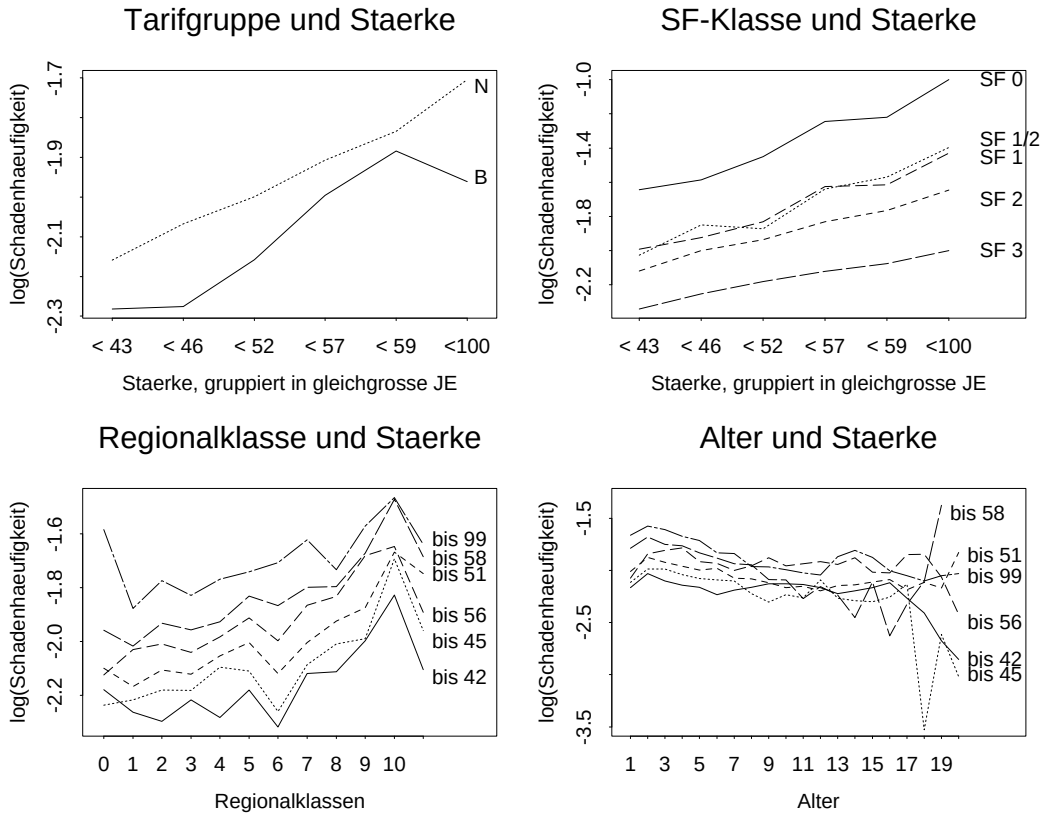


Abbildung 4.4: Interaktionen
N = Nichtbeamte, B = Beamte

An den Graphiken der Interaktionen lassen sich mehrere Eigenschaften erkennen. Zum einen bestätigen sie die Ähnlichkeit der Schadenfreiheitsklassen 1/2 und 1, weshalb wir in den folgenden Poissonregressionsmodellen anstelle der Schadenfreiheitsklassen die Beitragssätze in Prozent betrachten. Dadurch haben wir einen Parameter weniger zu schätzen. Desweiteren rechtfertigen die Graphiken für die Interaktionen mit dem Fahrzeugalter die Aufspaltung in drei neue, daraus hergeleitete Regressoren für die Poissonregression. Unter Berücksichtigung der unterschiedlichen Volumina in den einzelnen Gruppen stufen wir sämtliche Interaktionen mit dem Regressor Tarifgruppe als nichtsignifikant ein. Die Beurteilung der Signifikanz der restlichen Interakionen mit den Schadenfreiheitsklassen wird durch das Zusammenfassen von Klasse 1/2 und 1 etwas erschwert, wenn aber Interaktionen im Poissonmodell als signifikant erkannt werden, muß dies anhand der explorativen Datenanalyse auf die Klasse 0 zurückzuführen sein. Ebenso bei der Interaktion zwischen Fahrzeugalter und Regionalklasse liefert die explorative Datenanalyse keine sichere Aussage zur Signifikanz, insbesondere weil zuviele Gruppen nur wenige Jahreseinheiten besitzen und sich darum zufällige Schwankungen ausgeprägt bemerkbar machen. Anders verhält es sich mit den Interaktionen zwischen Alter und Stärke sowie zwischen Regionalklasse und Stärke, die wir aufgrund ihrer Graphiken als signifikant bewerten.
Bevor wir die Daten mit einer Poissonregression anpassen, müssen wir eine geeignete Transformation für die zwei neuen Altersvariablen und die Fahrzeugstärke finden. Dazu benutzen wir

das Box-Tidwell-Verfahren (s. Myers [1990, S. 307 f]), das uns Schätzer für die Exponenten einer Potenztransformation liefert. Box und Tidwell schlagen ein Iterationsverfahren vor, das die Exponenten $\alpha_1, \ldots, \alpha_k$ in einem Modell mit linearem Prädiktor

$$\eta = \beta_0 + \beta_1 w_1 + \ldots + \beta_k w_k, \text{ wobei } \quad w_j = \begin{cases} x_j^{\alpha_j} & falls \quad \alpha_j \neq 0 \\ \ln x_j & falls \quad \alpha_j = 0 \end{cases}$$

schätzt. Bei dieser Schreibweise und der folgenden Darstellung unterschlagen wir bei η, x und Y die Abhängigkeit von der i-ten Beobachtung zugunsten einer besseren Lesbarkeit. Fassen wir den linearen Prädiktor als Funktion $f(\alpha_1, \ldots, \alpha_k) = f(\alpha)$ von α auf und bezeichnen wir mit g die in Abschnitt 2.1.2 definierte Linkfunktion, so ergibt eine Taylorreihenentwicklung ersten Grades im Punkt $\alpha_0 = (\alpha_{1,0}, \ldots, \alpha_{k,0})$, angewendet auf den linearen Prädiktor:

$$E(Y) = g^{-1}(\eta) = g^{-1}(f(\alpha)) \approx g^{-1}\left(f(\alpha)|_{\alpha=\alpha_0} + (\alpha - \alpha_0)\frac{\partial}{\partial \alpha} f(\alpha)|_{\alpha=\alpha_0}\right)$$

Ein offensichtlicher Startwert für α_0 ist $\mathbf{1}$. Mit $f(\alpha)|_{\alpha=\mathbf{1}} = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$ und mit $\frac{\partial}{\partial \alpha_j}(x_j^{\alpha_j}) = \frac{\partial}{\partial \alpha_j}(\exp\{\alpha_j \ln x_j\}) = x_j^{\alpha_j} \ln x_j$ erhalten wir für die Taylorreihenentwicklung

$$\begin{aligned} E(Y) &= g^{-1}(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + (\alpha_1 - 1)\beta_1 x_1 \ln x_1 + \ldots + (\alpha_k - 1)\beta_k x_k \ln x_k) \\ &=: g^{-1}(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \gamma_1 z_1 + \ldots + \gamma_k z_k), \end{aligned}$$

wobei $\gamma_j := (\alpha_j - 1)\beta_j$ und $z_j := x_j \ln x_j$ für $j = 1, \ldots, k$ definiert sind.
Für ein Iterationsverfahren zur Schätzung von $\alpha_1, \ldots, \alpha_k$ sind folgende Schritte notwendig:

(i) Führe eine Regression mit dem Modell $E(Y) = g^{-1}(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k)$ durch und bezeichne die Parameterschätzer mit $\hat{\beta}_0, \ldots \hat{\beta}_k$.

(ii) Führe eine Regression mit dem Modell $E(Y) = g^{-1}(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \gamma_1 z_1 + \ldots + \gamma_k z_k)$ durch, um die Parameter $\gamma_1, \ldots, \gamma_k$ zu schätzen. Bezeichne die Schätzer mit $\hat{\gamma}_1, \ldots, \hat{\gamma}_k$.

(iii) Für $j = 1, \ldots, k$ schätze α_j durch

$$\hat{\alpha}_j = \frac{\hat{\gamma}_j}{\hat{\beta}_j} + 1$$

(iv) Die Iteration ist beendet, wenn $\hat{\gamma}_j$ nahe bei 0 liegt, denn dann gilt $\hat{\alpha}_j - 1 \approx 0$. Sind die Schätzer $\hat{\gamma}_j$ von γ groß, setze $w_j^* := x_j^{\alpha_j}$ und $z_j^* := w_j^* \ln w_j^*, j = 1, \ldots, k$, und wiederhole die Schritte (i) bis (iii) mit w_j^* bzw. z_j^* anstelle von x_j und z_j. Der neue Schätzer $\tilde{\alpha}_j$ berechnet sich aus $\tilde{\alpha}_j = (\frac{\hat{\gamma}_j}{\hat{\beta}_j} + 1)\hat{\alpha}_j$.

Wir beachten, daß sich allgemein nach n Iterationsdurchläufen der Schätzer $\tilde{\alpha}_j^{(n)}$ ergibt aus

$$\tilde{\alpha}_j^{(n)} = (\frac{\hat{\gamma}_j^{(n)}}{\hat{\beta}_j^{(n)}} + 1)\hat{\alpha}_j^{(n-1)},$$

wobei $\hat{\beta}_j^{(n)}$ und $\hat{\gamma}_j^{(n)}$ die Schätzer von β_j und γ_j in der n-ten Iteration und $\hat{\alpha}_j^{(n-1)}$ den Schätzer des Exponenten von x_j aus der $(n-1)$-ten Iteration bezeichnen.
Angewendet auf unseren Datensatz liefert uns das Box-Tidwell-Verfahren für den Regressor Fahrzeugstärke den Wert 1.075 für den Exponenten. Weil dieser Wert nahe bei 1 liegt, können wir die Fahrzeugstärke untransformiert lassen und einen linearen Zusammenhang zwischen ihr und der logarithmierten Schadenhäufigkeit annehmen. Für die neuen Altersvariablen in den Bereichen 2 bis 15 Jahre und über 15 Jahre bestimmt das Box-Tidwell-Verfahren die Werte 0.1029 und -3.3017 der Exponenten, die wir auf 0.1 und -3.3 runden. Somit wird das Fahrzeugalter zwischen 2 und 15 Jahren in Form der zehnten Wurzel in die Poissonregression eingehen, während wir das Fahrzeugalter über 15 Jahren zu $1/(Alter)^{3.3}$ transformieren.
Es sei darauf hingewiesen, daß wir durch die Schätzung der Exponenten zusätzliche Varianz in ein Regressionsmodell einbringen. Da wir in unseren Analysen die Exponenten als fest betrachten, unterschlagen wir diesen Einfluß auf die Varianz der Zielvariablen.

4.3 Poissonregression

Wir beginnen unsere Regressionsanalyse mit einem Modell, das alle Haupteffekte und alle sinnvollen Zweifach-Interaktionen als Regressoren umfaßt. Sinnvolle Zweifach-Interaktion heißt hier, daß wir nicht die Interaktionen zwischen den drei Altersvariablen in das Modell aufnehmen. Wir benutzen zur Spezifizierung des Modells die Bezeichnungen aus dem vorangegangen Abschnitt: S_g für die Schadenanzahl der Gruppe von Versicherungsnehmern in Regressorkombination g und v_g für die Jahreseinheiten der Regressorkombination g. Für die Regressoren verwenden wir die Bezeichnungen aus dem Quelltext des SAS-Programms, damit die SAS-Ausgabe in den späteren Abbildungen leichter zu lesen ist:

$$tg = \begin{cases} 1 & \text{für Tarifgruppe = Beamte} \\ 0 & \text{für Tarifgruppe = Nichtbeamte} \end{cases}$$

beitrag Beitragssatz einer Schadenfreiheitsklasse in Prozent
aufgeteilt in die drei Indikatorvariablen $beitrag = 40$, $beitrag = 55$, $beitrag = 100$
mit der Ausprägung 70 als Referenzkategorie

regiokl Regionalklasse $\in \{0, \ldots, 11\}$
aufgeteilt in die elf Indikatorvariablen $regiokl = 0$, $regiokl = 1, \ldots, regiokl = 10$
mit der Ausprägung 11 als Referenzkategorie

staerke Fahrzeugstärke in kW

$$age1 = \begin{cases} 1 & \text{für Fahrzeugalter = 1 Jahr} \\ 0 & \text{sonst} \end{cases}$$

$$age2to15 = \begin{cases} (\text{Fahrzeugalter})^{0,1} & \text{für Fahrzeugalter} \in \{2 \text{ Jahre}, \ldots, 15 \text{ Jahre}\} \\ 0 & \text{sonst} \end{cases}$$

$$age16min \quad = \quad \begin{cases} (\text{Fahrzeugalter})^{-3,3} & \text{für Fahrzeugalter} \in \{16 \text{ Jahre}, \ldots, 20 \text{ Jahre}\} \\ 0 & \text{sonst} \end{cases}$$

Mit diesen Bezeichnungen und dem Logarithmus als Linkfunktion spezifizieren wir den Erwartungswert μ_g, $g = 1, \ldots, 50\,842$, wie folgt

$$\begin{aligned}
\ln \mu_g \quad = \quad & \ln v_g + \beta_0 + \beta_{tg} tg_g + \beta_{40}(beitrag = 40)_g + \beta_{55}(beitrag = 55)_g + \beta_{100}(beitrag = 100)_g \\
& + \beta_{staerke} staerke_g + \beta_{age1} age1_g + \beta_{age2to15} age2to15_g + \beta_{age16min} age16min_g \\
& + \beta_{regiokl=0}(regiokl = 0)_g + \ldots + \beta_{regiokl=10}(regiokl = 10)_g + \beta_{staerke*tg} staerke_g * tg_g \\
& + \beta_{staerke*40} staerke_g * (beitrag = 40)_g + \beta_{staerke*55} staerke_g * (beitrag = 55)_g \\
& + \beta_{staerke*100} staerke_g * (beitrag = 100)_g \\
& + \beta_{staerke*regiokl=0} staerke_g * (regiokl = 0)_g + \ldots + \beta_{staerke*regiokl=10} staerke_g * (regiokl = 10)_g \\
& + \beta_{staerke*age1} staerke_g * age1_g + \beta_{staerke*age2to15} staerke_g * age2to15_g \\
& + \beta_{staerke*age16min} staerke_g * age16min_g \\
& + \beta_{regiokl=0*tg}(regiokl = 0)_g * tg_g + \ldots + \beta_{regiokl=10*tg}(regiokl = 10)_g * tg_g \\
& + \beta_{regiokl=0*age1}(regiokl = 0)_g * age1_g + \ldots + \beta_{regiokl=10*age1}(regiokl = 10)_g * age1_g \\
& + \beta_{regiokl=0*age2to15}(regiokl = 0)_g * age2to15_g + \\
& \quad \ldots + \beta_{regiokl=10*age2to15}(regiokl = 10)_g * age2to15_g \\
& + \beta_{regiokl=0*age16min}(regiokl = 0)_g * age16min_g + \\
& \quad \ldots + \beta_{regiokl=10*age16min}(regiokl = 10)_g * age16min_g \\
& + \beta_{regiokl=0*40}(regiokl = 0)_g * (beitrag = 40)_g + \\
& \quad \ldots + \beta_{regiokl=10*40}(regiokl = 10)_g * (beitrag = 40)_g \\
& + \beta_{regiokl=0*55}(regiokl = 0)_g * (beitrag = 55)_g + \\
& \quad \ldots + \beta_{regiokl=10*55}(regiokl = 10)_g * (beitrag = 55)_g \\
& + \beta_{regiokl=0*100}(regiokl = 0)_g * (beitrag = 100)_g + \\
& \quad \ldots + \beta_{regiokl=10*100}(regiokl = 10)_g * (beitrag = 100)_g \\
& + \beta_{age1*tg} age1_g * tg_g + \beta_{age2to15*tg} age2to15_g * tg_g + \beta_{age16min*tg} age16min_g * tg_g \\
& + \beta_{age1*40} age1_g * (beitrag = 40)_g + \beta_{age2to15*40} age2to15_g * (beitrag = 40)_g \\
& + \beta_{age16min*40} age16min_g * (beitrag = 40)_g \\
& + \beta_{age1*55} age1_g * (beitrag = 55)_g + \beta_{age2to15*55} age2to15_g * (beitrag = 55)_g \\
& + \beta_{age16min*55} age16min_g * (beitrag = 55)_g \\
& + \beta_{age1*100} age1_g * (beitrag = 100)_g + \beta_{age2to15*100} age2to15_g * (beitrag = 100)_g \\
& + \beta_{age16min*100} age16min_g * (beitrag = 100)_g \\
& + \beta_{40*tg}(beitrag = 40)_g * tg_g + \beta_{55*tg}(beitrag = 55)_g * tg_g + \beta_{100*tg}(beitrag = 100)_g * tg_g
\end{aligned}$$

Dieses Ausgangsmodell besitzt 129 zu schätzende Parameter, was auch in Hinblick auf eine Tarifbildung recht viel ist. Darum ist es unser Ziel, mit einer Rückwärtsanalyse Regressoren zu erkennen, die nicht signifikant zur Erklärung der Schadenanzahl beitragen und diese aus dem Modell zu entfernen, um die Veränderungen in der Schadenanzahl durch ein Modell mit wenigen Parametern erklären zu können. Dabei nehmen wir in jedem Schritt der Rückwärtsanalyse nur einen nichtsignifikanten Regressor aus dem Modell.
Kriterien für den Vergleich zwischen zwei Modellen sind die log-Likelihoods (s. Definition 2.4) und die skalierte Devianz (s. Definition 2.6) als Maße für den Verlust, wenn ein Regressor aus dem Modell entfernt wurde. Natürlich nimmt die log-Likelihood durch das Entfernen eines Regressors ab. Ist die Abnahme klein, besitzt das neue Modell mit einem Regressor weniger annähernd das gleiche Erklärungspotential wie das vorhergehende Modell. Ist die Abnahme im log-Likelihood groß, so verliert das neue Modell viel an Erklärungspotential. Da die skalierte Devianz als zweifache Differenz zwischen der maximal erreichbaren log-Likelihood und der erreichten log-Likelihood definiert ist, bedeutet eine große Abnahme der log-Likelihood dementsprechend einen großen Anstieg in der Devianz. In Abschnitt 2.1.5 erläuterten wir, daß die Devianzendifferenz von zwei aufeinander folgenden Modellen in der Rückwärtsanalyse asymptotisch χ^2-verteilt ist mit genau sovielen Freiheitsgraden, wie der entfernte Regressor besitzt. Kennen wir den asymptotischen p-Wert dieser Differenz, können wir die Entnahme des Regressors als signifikant oder nicht signifikant beurteilen. Dieses Kriterium für die Signifikanz eines Regressors stellt uns das Programm SAS in seiner Typ 3-Anova-Tabelle bereit. Die dritte Spalte der Anova-Tabelle enthält die Differenz zwischen der Devianz des untersuchten Modells und der Devianz des Modells, wenn der in der ersten Spalte der Anova-Tabelle bezeichnete Regressor aus dem untersuchten Modell entfernt wird. Mit anderen Worten: die dritte Spalte liefert die Teststatistik zu dem Test, ob der Parameter des Regressors aus der ersten Spalte im gewählten Modell gleich 0 ist. Die asymptotischen p-Werte dieser Tests stehen in der vierten Spalte, und die zugehörigen Freiheitsgrade befinden sich in der zweiten Spalte. Da unser Datensatz sehr groß ist und die Beobachtungen gruppiert vorliegen, verwenden wir diese p-Werte ohne Korrektur. Wir gehen bei der Rückwärtsanalyse so vor, daß wir jeweils denjenigen Regressor aus dem Modell entfernen, der den höchsten p-Wert $> 0,05$ hat. Einzige Einschränkung dabei ist, daß wir ein hierarchisches Modell beibehalten. Stellt sich also heraus, daß ein Haupteffekt in der Anova-Tabelle den größten p-Wert hat, während eine Interaktion zwischen ihm und einem anderen Haupteffekt als signifikant bewertet wird, behalten wir den nichtsignifikanten Haupteffekt im Modell und entfernen den Regressor mit dem zweithöchsten p-Wert $> 0,05$, falls möglich.
Mit Hilfe der graphischen Residuenanalyse beurteilen wir die Anpasung eines Modells bzgl. der Wahl der Varianz- und Linkfunktion, Anpassung an einen Haupteffekt und der Gesamtanpassung.
Die folgenden Abbildungen zeigen die Ergebnisse von der Analyse des Ausgangsmodells.

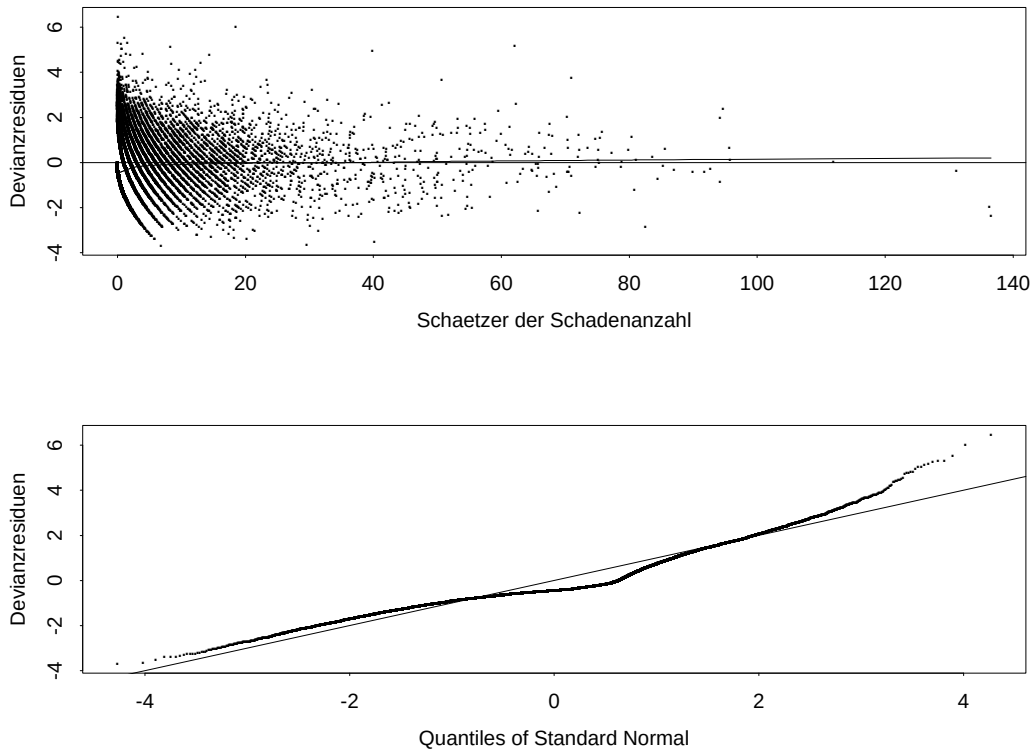


Abbildung 4.5: Analyse der Devianzresiduen des Ausgangsmodells
Die Linie nahe der x-Achse in der oberen Graphik ist der Graph der glättenden Funktion lowess

Zur Beurteilung der Gesamtanpassung haben wir zum einen die Devianzresiduen (s. Definition 2.38) gegen die Schätzer unserer Zielvariablen Schadenanzahl aufgetragen und zum anderen sie in einer Quantil-Quantil-Graphik dargestellt. In dem Streudiagramm erkennen wir sowohl an den Devianzresiduen als auch an der glättenden Funktion, daß vor allem die kleinen Schadenanzahlen schlecht von dem Modell geschätzt werden. Da aber auch bei höheren Schätzwerten der Schadenanzahl viele große Devianzresiduen auftreten, beurteilen wir die Gesamtanpassung als weniger gelungen. In der Graphik sind keine Konfidenzintervalle für die Schätzer eingezeichnet. Diese wurden aber für das 5%-Signifikanzniveau ermittelt und sind für den Großteil der geschätzten Schadenanzahlen so eng, daß bei dem Auftragen der unteren und oberen Konfidenzintervallgrenzen in die Graphik deren Werte mit dem Schätzer meist zusammenfallen. Oder die Konfidenzintervallgrenzen überdecken die Werte anderer Beobachtungen, so daß in beiden Fällen durch diese zusätzliche Information andere verloren geht. Aus diesen Gründen wird auf eine Abbildung verzichtet. Die Quantil-Quantil-Graphik zeigt, daß die Residuen für kleine und mittlere Werte annähernd normalverteilt sind, während die großen Residuen (> 4) doch deutlich davon abweichen. Die allgemeine leichte U-Form weist auf Schiefe hin, die sich durch die theoretisch zu erwartende Schiefe der Poissonverteilung erklären läßt (s. Erörterung der Residuenanalyse nach Definition 2.38).

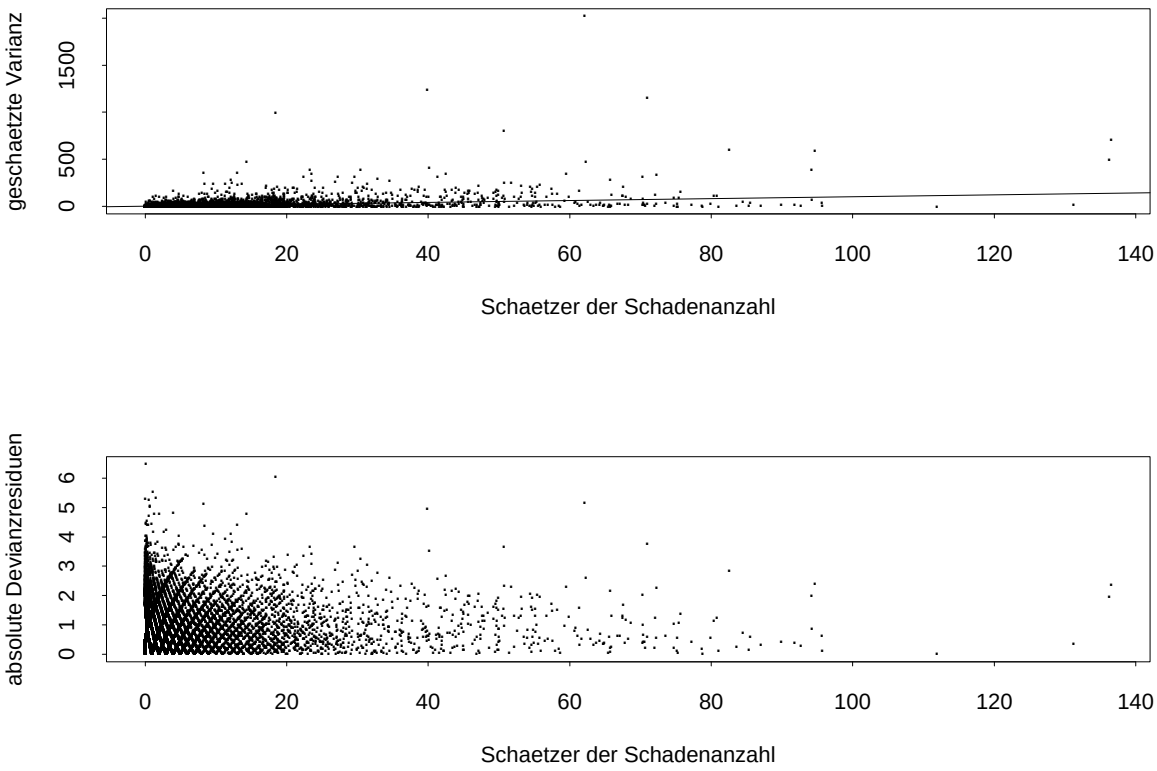


Abbildung 4.6a: Varianzfunktion im Ausgangsmodell
durchgezogene Linie : lowess-*Funktion*

Um die Ursachen der schlechten Gesamtanpassung herauszufinden, betrachten wir die Graphiken zur Überprüfung der Varianzfunktion. Vor allem die Graphik, in der die quadrierten Rohresiduen (s. Definition 2.36) als Schätzer für die Varianz auf der y-Achse gegen die geschätzte Schadenanzahl aufgetragen sind, weist auf eine massive Verletzung der Varianzannahme hin. Wir erinnern an Abschnitt 2.3.2, in dem dargelegt wurde, daß für das Poissonmodell diese Graphik ein zufälliges Streuungsmuster um die Gerade durch den Nullpunkt mit Steigung 1 darstellen sollte. Selbst wenn wir die besonders großen Ausreißer mit einer geschätzten Varianz von über 500 entfernen, um in einer neuen Varianzgraphik einen größeren Maßstab zu erhalten, ist die theoretisch geforderte Varianzfunktion $V(\mu_i) = \mu_i$ auch in einer solchen Darstellung nicht zu erkennen, wie Abb. 4.6b belegt. Die untere Graphik in Abbildung 4.6a zur Überprüfung der Varianzfunktion, in der die absoluten Devianzresiduen gegen die geschätzte Schadenanzahl aufgetragen sind, bestätigt die erste Graphik. Insbesondere werden die größten Ausreißer als solche auch hier widergegeben. Zusätzlich zeigt die zweite Graphik auch bei den kleinen geschätzten Schadenanzahlen große Ausreißer, die in der ersten Graphik nicht so deutlich hervortraten, was hauptsächlich an dem gewählten Maßstab und der Überlagerung vieler Werte liegt.

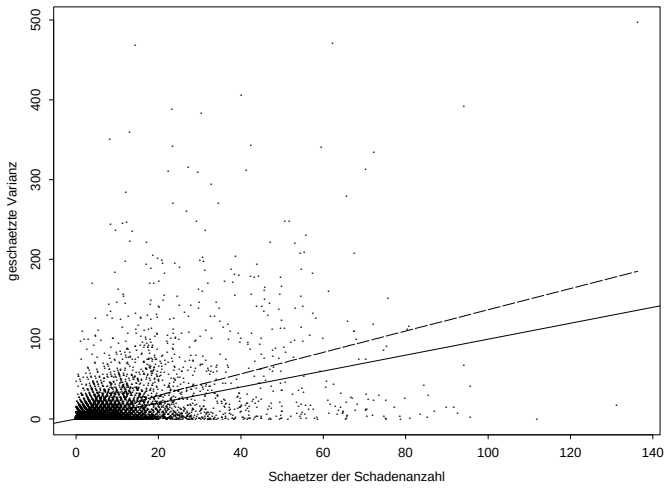


Abbildung 4.6b: Analyse der Varianzfunktion im Ausgangsmodell eingeschränkt auf Schätzwerte der Varianz < 500

Die durchgezogene Linie bezeichnet den Graphen der Varianzfunktion $V(\mu_i) = \mu_i$ *im Poissonmodell, während die gestrichelte Linie den Graphen der aus den Daten ermittelten glättenden Funktion* lowess *darstellt. Ihre Lage oberhalb der Varianzfunktion belegt die für eine Poissonverteilung zu große Streuung in den Daten.*

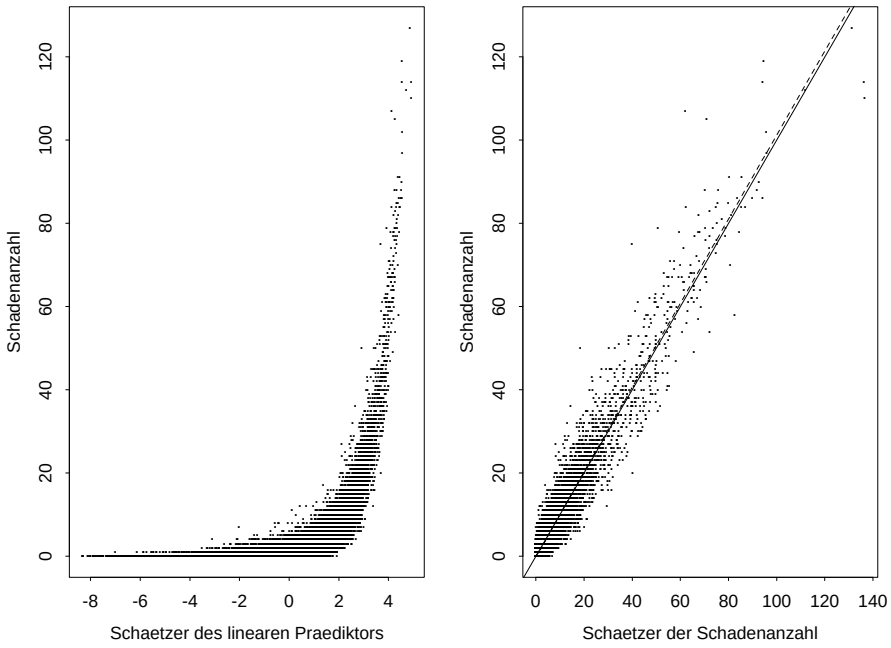
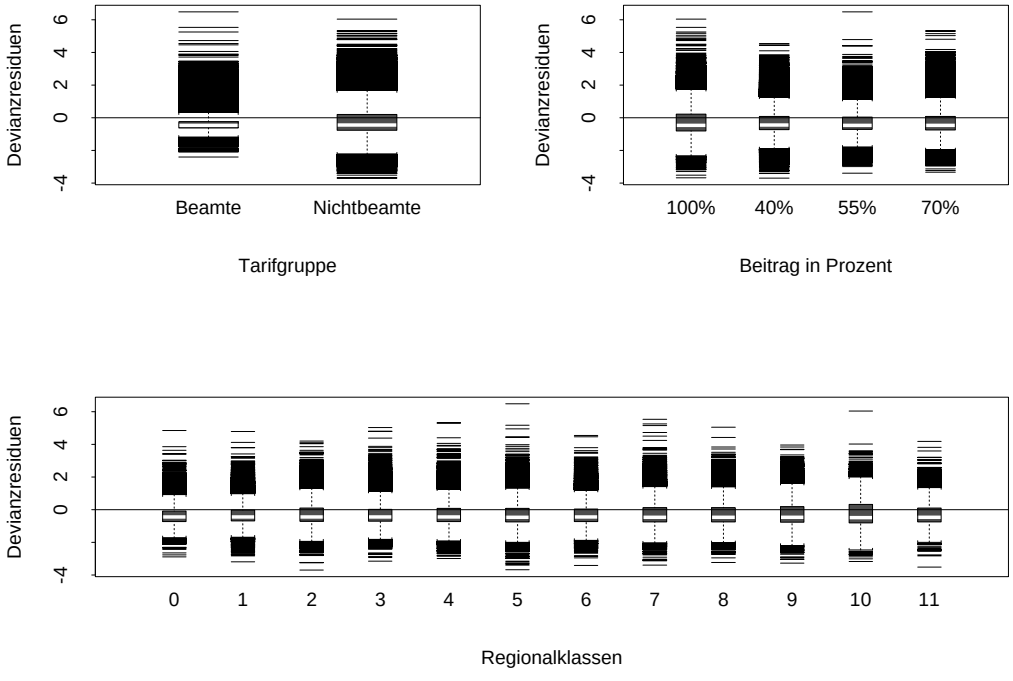


Abbildung 4.7: Linkfunktion im Ausgangsmodell

durchgezogene Linie: Graph der exponentiellen Linkfunktion $\exp(g(\mu_i)) = \exp(\ln \mu_i) = \mu_i$ *im Poissonmodell*

gestrichelte Linie: Graph der Funktion lowess

Als nächstes verifizieren wir graphisch die Wahl des Logarithmus als Linkfunktion. Bei geeigneter Wahl sollen die Punkte in der linken Graphik von Abbildung 4.7 zufällig um die Exponentialfunktion streuen, was augenscheinlich der Fall ist. Da visuell jedoch ein linearer Zusammenhang leichter zu erkennen ist, verwenden wir die exponentiellen Werte des geschätzten Prädiktors. Diese Transformation liefert uns gerade die Schätzer der Schadenanzahl, und ihre Darstellung gegen die beobachteten Schadenanzahlen befindet sich in der rechten Graphik von Abbildung 4.7. Sie bestätigt den Logarithmus als passende Linkfunktion, denn die Punkte streuen zufällig um die theoretisch geforderte Gerade durch den Nullpunkt mit Steigung 1, die zur Orientierung als durchgezogene Linie in die Graphik eingezeichnet ist. Aufgrund der mehrfachen Überlagerung von Punkten, insbesondere bei geschätzten Schadenanzahlen bis 30, ist es jedoch allein mittels der Streuung der Residuen nicht möglich, den exakten Verlauf der Linkfunktion zu erkennen. Deshalb verwenden wir hier die glättende Funktion *lowess*, die die Linkfunktion mit lokalen Regressionsmodellen aus den Daten schätzt und so den Informationsverlust durch die Überlagerungen ausgleicht. Ihr fast mit der Varianzfunktion identischer Graph liefert einen weiteren Beleg für die richtige Wahl der Linkfunktion.
Statt den geschätzten linearen Prädiktor zu transformieren, hätten wir auch die Schadenanzahl transformieren können, indem wir sie logarithmieren, um denselben linearen Zusammenhang zu überprüfen. Doch dabei gehen alle Beobachtungen verloren, bei denen keine Schäden aufgetreten sind, weil der Logarithmus für 0 nicht definiert ist. Dieser Informationsverlust wurde durch die hier gewählte Transformation vermieden.



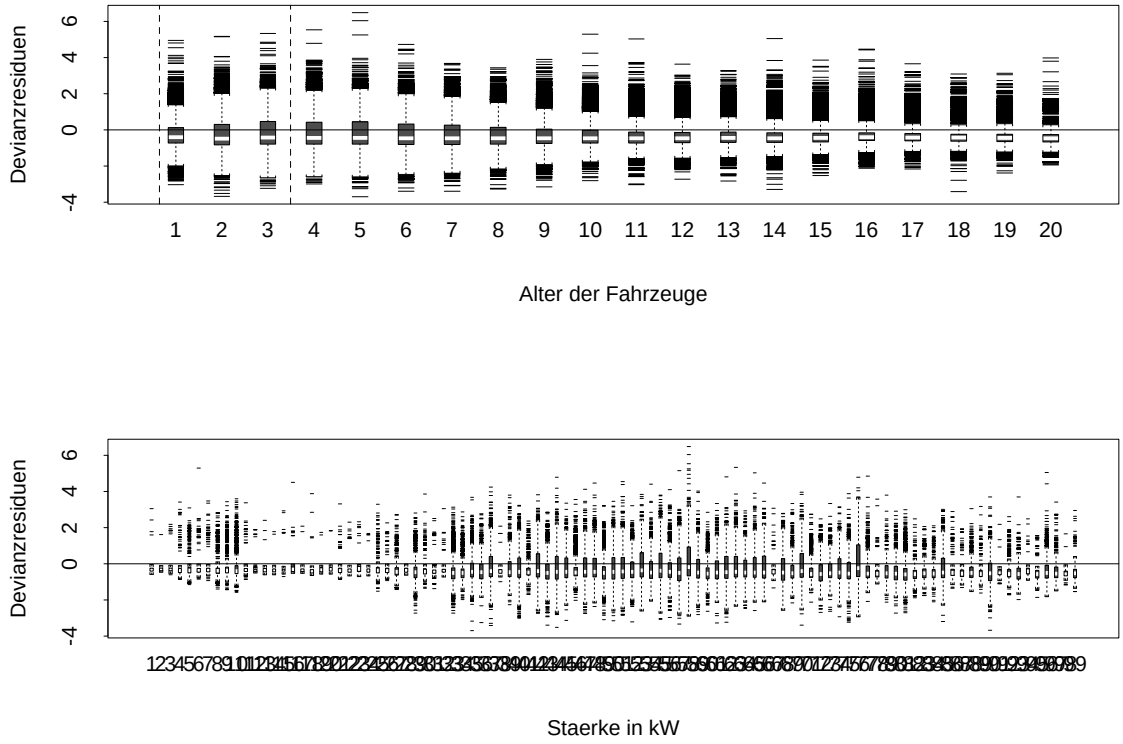


Abbildung 4.8: Residuenanalyse der Haupteffekte im Ausgangsmodell

Kommen wir nun zur graphischen Residuenanalyse der Haupteffekte. Um das Problem der vielfachen Überlagerung von Werten, das aus der großen Datenmenge herrührt, einzuschränken, haben wir sowohl für die kategoriellen als auch die stetigen Haupteffekte Boxplots erstellt. Zunächst fällt auf, daß alle Haupteffekte sehr viele Ausreißer in allen Ausprägungen haben. Das war wegen der großen Varianzen aus Abbildung 4.5 zu erwarten. Ebenso ist es plausibel, daß diejenigen Boxen und Nadeln länger sind, wo auch mehr Jahreseinheiten in den Ausprägungen stecken (vgl. dazu Abbildungen 4.1 und 4.2). Außerdem stellen wir fest, daß es deutlich mehr Ausreißer nach oben als nach unten gibt und daß der Median, der als weißer Balken in der Box sichtbar ist, stets kleiner als 0 ist. Mit der Schiefe der Poissonverteilung ist die nur geringe Abweichung aller Mediane von 0 zu erklären. Die negativen Mediane als systematischen Fehler und damit als eine Mißspezifizierung des Erwartungswerts zu deuten, scheint hier überinterpretiert.

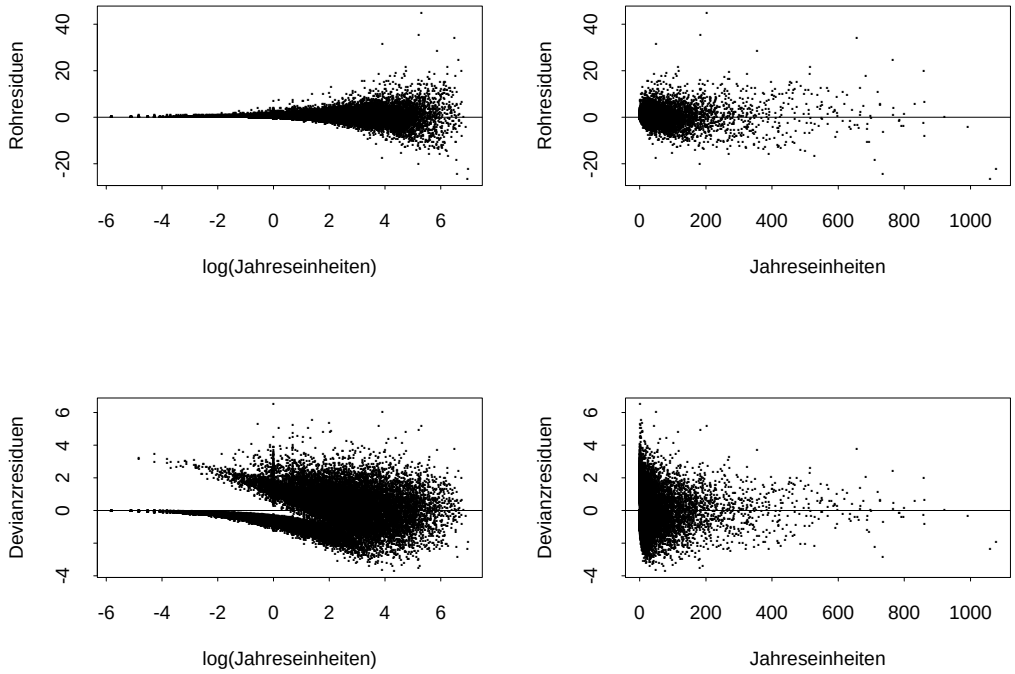


Abbildung 4.9: Offset im Ausgangsmodell

Mit Abbildung 4.9 beleuchten wir den Einfluß der Jahreseinheiten als Offset. Wir erkennen an der logarithmierten Darstellung der Jahreseinheiten, daß bei sehr schwach besetzen Beobachtungen mit weniger als einer Jahreseinheit generell die Schadenanzahl durch 0 geschätzt wird, was bei einer durchschnittlichen Schadenhäufigkeit von 0,14 für den gesamten Datensatz vernünftig ist. Ist nun doch ein Schaden (oder mehrere) eingetreten, sorgt dieser dafür, daß das Devianzresiduum dieser Beobachtung sprunghaft ansteigt, wodurch in der Graphik links unten von Abbildung 4.9 zwei Punktestränge zu sehen sind. Mit steigendem Jahreseinheitenvolumen in den Beobachtungen nehmen die Residuen die Gestalt einer Punktewolke an. Es fällt in der Grahik rechts unten in Abbildung 4.9 auf, daß bei den Beobachtungen, die ein großes Volumen von mindestens 400 Jahreseinheiten besitzen, nur noch ein Ausreißer, d. h. absolutes Devianzresiduum $|d_i| > 3$, bei den Devianzresiduen zu finden ist, wenn auch die Streuung um 0 groß bleibt. Dies und das Fehlen eines Trends lassen uns schließen, daß diese Beobachtungen, die etwa 13% der Jahreseinheiten ausmachen, von dem Poissonmodell vergleichsweise gut angepaßt werden, wenn auch die starke Streuung auf inhomogene Gruppen bzgl. der erhobenen Merkmale deutet. Wir können also das Auftreten von zufälligen Effekten bzw. unbeobachtete Heterogenität nicht ausschließen.

Nach der graphischen Beurteilung des Ausgangsmodells kommen wir kurz zu den formalen Anpassungskriterien, die von SAS ausgegeben werden.

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	51E3	45384.6887	0.8949
Scaled Deviance	51E3	45384.6887	0.8949
Pearson Chi-Square	51E3	65425.5584	1.2901
Scaled Pearson X2	51E3	65425.5584	1.2901
Log Likelihood	51E3	107990.9578	

Abbildung 4.10: Anpassungskriterien des Ausgangsmodells

Die hier ermittelten Werte für die Quotienten von Devianz- und Pearson-Statistik durch die Freiheitsgrade legen nahe, daß die Äquidispersionsannahme des Poissonmodells nur leicht verletzt ist, ja sogar Unterdispersion vorliegen könnte. Im Gegensatz dazu belegt die graphische Analyse vehemente Überdispersion. Somit rechtfertigt dieser Datensatz die in Abschnitt 2.1.5 empfohlene vorsichtige Anwendung der Devianz als Maß zur Beurteilung eines einzelnen Modells und den bevorzugten Gebrauch zum Vergleich zweier geschachtelter Modelle. Der Datensatz zeigt ebenso, daß die Pearson-Statistik aus Definition 2.8 als Teststatistik für Überdispersion im Poissonmodell mit linearer Varianzfunktion als Alternative geeignet ist (s. Beispiel 3.2.3).

LR-Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
staerke	1	139.75	<.0001
regiokl	11	47.05	<.0001
tg	1	0.43	0.5096
age1	1	11.73	0.0006
age2to15	1	15.18	<.0001
age16min	0	0.00	.
beitrag	3	43.76	<.0001
staerke*tg	1	0.94	0.3319
staerke*beitrag	3	280.02	<.0001
staerke*regiokl	11	73.36	<.0001
staerke*age1	1	27.73	<.0001
staerke*age2to15	1	27.80	<.0001
staerke*age16min	1	39.19	<.0001
tg*regiokl	11	13.29	0.2750
regiokl*age1	11	18.45	0.0728
regiokl*age2to15	11	17.43	0.0959
regiokl*age16min	0	0.00	.
regiokl*beitrag	33	79.39	<.0001
tg*age1	1	0.85	0.3578
tg*age2to15	1	0.53	0.4647
tg*age16min	0	0.00	.
tg*beitrag	3	8.29	0.0404
age1*beitrag	3	55.64	<.0001
age2to15*beitrag	3	77.23	<.0001
age16min*beitrag	3	94.29	<.0001

Abbildung 4.11: Anova-Tabelle des Ausgangsmodells
Zu den Definitionen der Abkürzungen s. Anfang dieses Abschnitts

Wir sind als nächstes daran interessiert, die Anzahl der zu schätzenden Parameter zu reduzieren. Dazu ziehen wir die Anova-Tabelle des Ausgangsmodells heran, um festzustellen, ob ein und welcher Regressor entfernt werden kann. Die Tabelle in Abbildung 4.11 zeigt, daß der Haupteffekt Tarifgruppe den höchsten nichtsignifikanten p-Wert besitzt. Da wir ein hierarchisches Modell beibehalten wollen, entfernen wir die Interaktion zwischen Tarifgruppe und dem Fahrzeugalter für den Altersbereich zwischen 2 und 15 Jahren, die den zweithöchsten nichtsignifikanten p-Wert hat. Die teilweise fehlenden Angaben für die dritte Altersvariable `age16min` werden von SAS nur dann erzeugt, wenn das Programm lineare Abhängigkeiten von den anderen Regressoren erkennt. Natürlich gibt es dafür inhaltlich überhaupt keine Grundlage und Rechtfertigung. Wir erklären uns dieses Phänomen durch Rundungsfehler, denn durch die gewählte Transformation $(\texttt{Alter})^{-3{,}3}$ entstehen sehr kleine Werte für `age16min`, die erst in der vierten Nachkommastelle ungleich 0 sind. Bei der Berechnung des Modells wird intern die Maximalstruktur an Zellen aufgebaut, d. h. alle möglichen Kombinationen der Regressorausprägungen, von denen im vorliegen Datensatz nur rund ein Viertel realisiert sind. Während der Iterationen erzeugen diese Werte für den Rechner mit seinen endlichen Maschinenzahlen praktisch Nullen, die lineare Abhängigkeit bewirken. Als Beleg für unseren Erklärungsansatz geben wir in Abbildung 4.12 die von SAS erstellte Typ 1-Anova-Tabelle an. In der vierten Spalte der Tabelle stehen die Devianzendifferenzen zwischen dem Modell, das aus allen in der ersten Spalte genannten Regressoren in den Zeilen oberhalb der betrachteten Zeile besteht, und dem Modell, das alle Regressoren von der ersten Zeile bis einschließlich der betrachteten Zeile enthält. Es werden somit auch in dieser Tabelle die Signifikanzen der Regressoren getestet, aber hier in einer Folge von geschachtelten Modellen, anfangend bei dem Nullmodell und bis zum vollständig im Prozeduraufruf spezifizierten Modell aufsteigend. Die asymptotischen p-Werte dieser Tests befinden sich in der letzten Spalte und die zugehörigen Freiheitsgrade in der dritten Spalte. Die zweite Spalte gibt in der n-ten Zeile die Devianz des Modells mit den Regressoren aus den ersten n Zeilen an.

Dieser hierarchische Aufbau der Tabelle bedeutet insbesondere, daß die Maximalstruktur der Modelle von der ersten Zeile bis zur letzten Zeile, wo die Maximalstruktur des Ausgangsmodells erreicht wird, anwächst. Wir sehen in Abbildung 4.12, daß es SAS noch bis zum Untermodell in der fünften Zeile von unten möglich ist, Signifikanztests ohne lineare Abhängigkeit zu berechnen. Erst dannach wird die Modellstruktur so groß, daß die Rechnerungenauigkeit zu falschen Testresultaten führt.

LR-Statistics For Type 3 Analysis

Source	Deviance	DF	Chi-Square	Pr > ChiSq
staerke	58393.0034	1	2638.37	<.0001
regiokl	55754.6297	11	1585.19	<.0001
tg	54169.4370	1	64.25	<.0001
age1	54105.1878	1	1.66	0.1975
age2to15	54103.5080	1	0.02	0.8897
age16min	53794.9709	1	308.54	<.0001
beitrag	46103.2585	3	7691.71	<.0001
staerke*tg	46100.0365	1	3.22	0.0727
staerke*beitrag	45864.1718	3	235.86	<.0001
staerke*regiokl	45787.0371	11	77.13	<.0001
staerke*age1	45786.9661	1	0.07	0.7898
staerke*age2to15	45785.3751	1	1.59	0.2072
staerke*age16min	45744.6389	1	40.74	<.0001
tg*regiokl	45730.2577	11	14.38	0.2126
regiokl*age1	45705.3167	11	24.94	0.0093
regiokl*age2to15	45691.9794	11	13.34	0.2719
regiokl*age16min	45671.0726	11	20.91	0.0343
regiokl*beitrag	45589.9056	33	81.17	<.0001
tg*age1	45589.5775	1	0.33	0.5668
tg*age2to15	45586.5297	1	3.05	0.0808
tg*age16min	45586.5285	1	0.00	0.9721
tg*beitrag	45578.7489	3	7.78	0.0404
age1*beitrag	45480.5487	3	98.20	0.0508
age2to15*beitrag	45478.9808	3	1.57	0.6667
age16min*beitrag	45384.6887	3	94.29	<.0001

Abbildung 4.12: Typ 1-Anova-Tabelle des Ausgangsmodells
Zu den Definitionen der Abkürzungen s. Anfang dieses Abschnitts

Im Laufe der Rückwärtsanalyse werden die Interaktionsterme `tg*Alter` mit allen drei Altersvariablen, `tg*staerke, tg*regiokl,` und `age2to15*regiokl` eliminiert. Nach sechs Schritten ist die Rückwärtsanalyse beendet, und das Endmodell der Poissonregression in Abbildung 4.16 dargestellt.

Poissonregression mit dem sechsten Modell
wie Ausgangsmodell ohne tg*Alter, tg*staerke, tg*regiokl und
ohne age2to15*regiokl

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	51E3	45419.8484	0.8952
Scaled Deviance	51E3	45419.8484	0.8952
Pearson Chi-Square	51E3	65338.2732	1.2878
Scaled Pearson X2	51E3	65338.2732	1.2878
Log Likelihood	51E3	107990.8440	

LR-Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
staerke	1	180.53	<.0001
regiokl	11	202.29	<.0001
tg	1	28.91	<.0001
age1	1	13.05	0.0003
age2to15	1	18.76	<.0001
age16min	0	0.00	.
beitrag	3	44.57	<.0001
staerke*beitrag	3	282.21	<.0001
staerke*regiokl	11	72.23	<.0001
staerke*age1	1	27.82	<.0001
staerke*age2to15	1	27.94	<.0001
staerke*age16min	1	39.35	<.0001
regiokl*age1	11	22.48	0.0209
regiokl*age16min	0	0.00	.
regiokl*beitrag	33	79.05	<.0001
tg*beitrag	3	8.68	0.0338
age1*beitrag	3	55.86	<.0001
age2to15*beitrag	3	78.44	<.0001
age16min*beitrag	3	95.84	<.0001

Abbildung 4.16: Rückwärtsanalyse: 6. Schritt

Es erstaunt, daß die Interaktion zwischen Alter und Regionalklasse für neue und ganz alte Fahrzeuge bestehen bleibt, während das Schadenverhalten der Fahrzeuge mittleren Alters von der Regionalstruktur unbeeinflußt bleibt. Das Verschwinden fast aller Interaktionen mit der Tarifgruppe bedeutet, daß Beamte kein spezifisches Schadenverhalten bezüglich der hier untersuchten Merkmale verglichen mit Nichtbeamten haben, was plausibel ist. Die Beitragsprämien der Beamten sollten jedoch nach der Schadenfreiheitsklasse gestaffelt sein, das besagt der Verbleib der Interaktion tg*beitrag im Modell.

Wir wollen die Anpassung des Endmodells an die Daten auch graphisch mittels der Residuenanalyse beurteilen. Da sich die Devianz schrittweise nur nichtsignifikant vom Ausgangs- zum Endmodell erhöht hat, erwarten wir, daß die Residuen im Endmodell nur unwesentlich stärker streuen als im Ausgangsmodell. Wie schon bei den Graphiken zum Ausgangsmodell wird die Analyse durch die große Datenmenge und die damit verbundene Überlagerung der Punkte erschwert. Ein graphischer Vergleich der beiden Modelle muß sich deshalb auf die Betrachtung

der Ausreißer beschränken, denn eine Änderung der Werte im tiefschwarzen Bereich, sozusagen dem Kern, der Punktewolke ist visuell nicht erkennbar. Wir können sehen, ob sich die Gestalt des Kerns der Punktewolke verändert hat, um Rückschlüsse auf einen Trend oder eine andere Varianz der Daten zu ziehen. Außerdem zeigen uns die Graphiken, ob extreme Ausreißer im Endmodell hinzugekommen sind, was bedeutet, daß die Streuung sich insgesamt vergrößert hat, oder ob die Ausreißer aus dem Anfangsmodell noch an Höhe gewonnen haben, was heißt, daß die Devianzzunahme auf einzelne schlecht angepaßte Beobachtungen zurückzuführen ist.

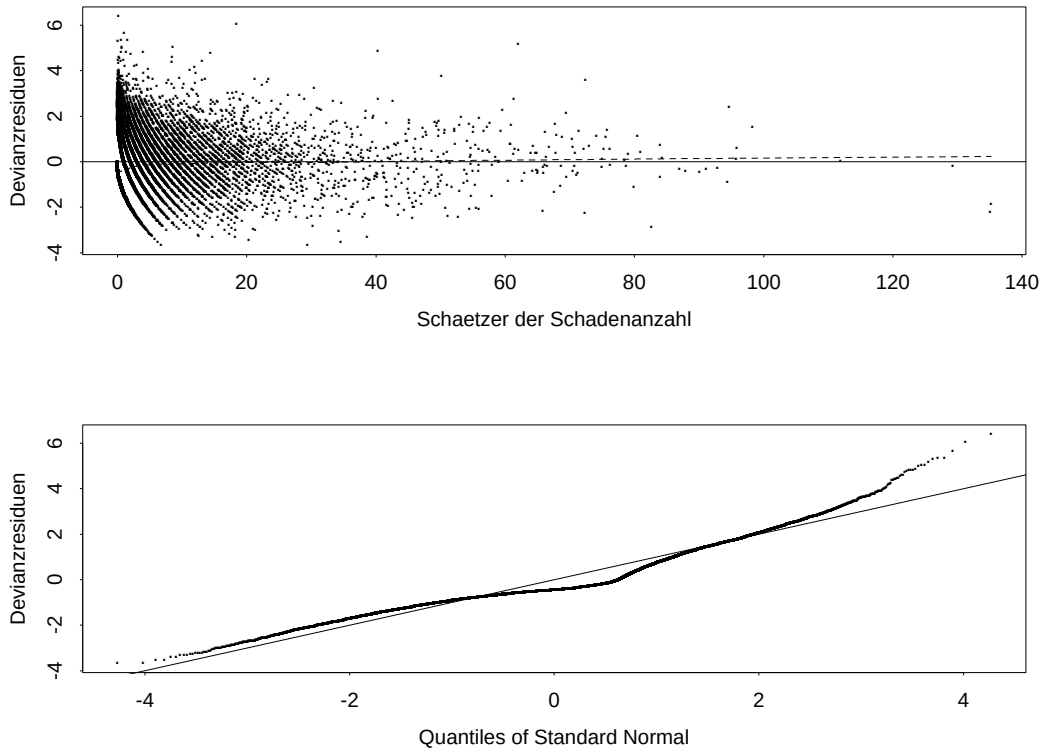


Abbildung 4.17: Gesamtanpassung des Endmodells
gestrichelte Linie: lowess-*Funktion*

Beide Graphiken zur Gesamtanpassung in Abbildung 4.17 belegen, daß sich die extremen Ausreißer vom Ausgangsmodell im Endmodell nicht weiter vergrößerten und daß, wie erwartet, die geringe Zunahme der Devianz graphisch auf den ersten Blick nicht festzustellen ist. Lediglich an der oberen Graphik in Abbildung 4.17 sind zwei neue Ausreißer nach unten bei einer geschätzten Schadenanzahl von etwa 35 zu sehen. Daß diese aber keinen nennenswerten Einfluß auf die Gesamtanpassung ausüben, drückt sich in dem unveränderten Graphen der *lowess*-Funktion aus.

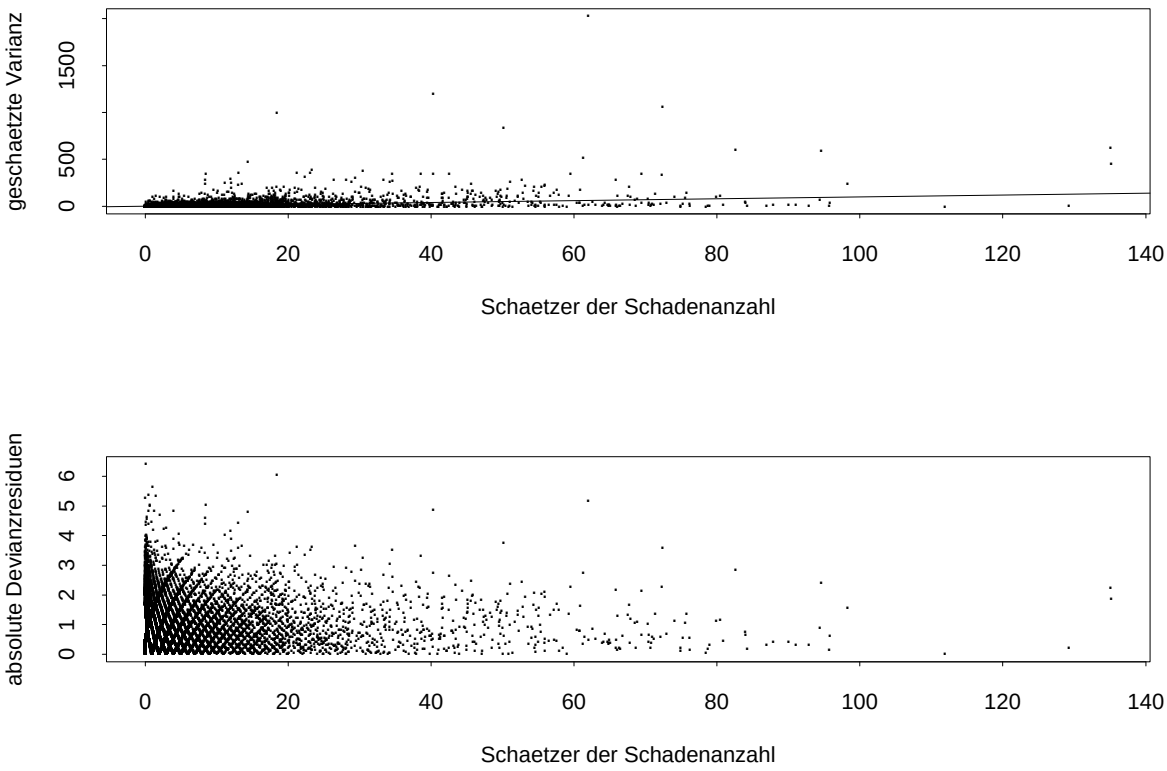


Abbildung 4.18: Varianzfunktion im Endmodell
Die durchgezogene Linie in der oberen Graphik stellt den Graphen der Varianzfunktion $V(\mu_i) = \mu_i$ *im Poissonmodell dar.*

Die Graphiken in Abbildung 4.18 zur Überprüfung der Äquidispersionsannahme des Poissonmodells bestätigen die kaum zu erkennende stärkere Streuung der Residuen. Nur die beiden neuen Ausreißer aus Abbildung 4.17 sind auch in den Graphiken aus Abbildung 4.18 als solche auszumachen. Doch ihr Auftreten bedeutet nicht, daß die Verletzung der Varianzfunktion noch erheblich gravierender als im Ausgangsmodell ist.

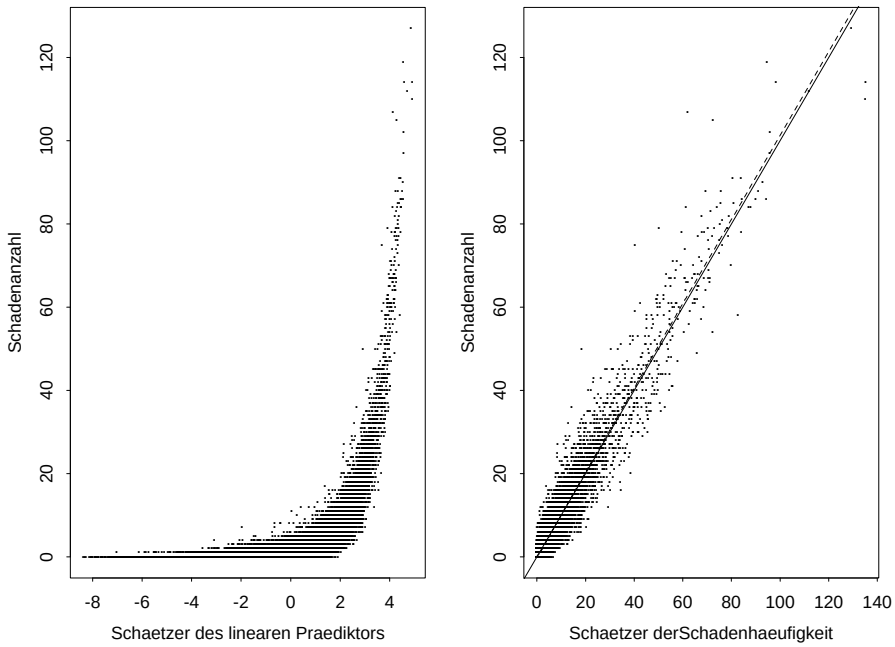


Abbildung 4.19: Linkfunktion im Endmodell
durchgezogene Linie: Graph der exponentiellen Linkfunktion $\exp(g(\mu_i)) = \exp(\ln \mu_i) = \mu_i$ *im Poissonmodell*
gestrichelte Linie: Graph der Funktion lowess

Für die graphische Verifikation der Linkfunktion gilt dasselbe wie für die bisherigen Graphiken: die Unterschiede zum Ausgangsmodell sind kaum zu entdecken, weshalb der Logarithmus als Linkfunktion im Endmodell genauso geeignet ist.

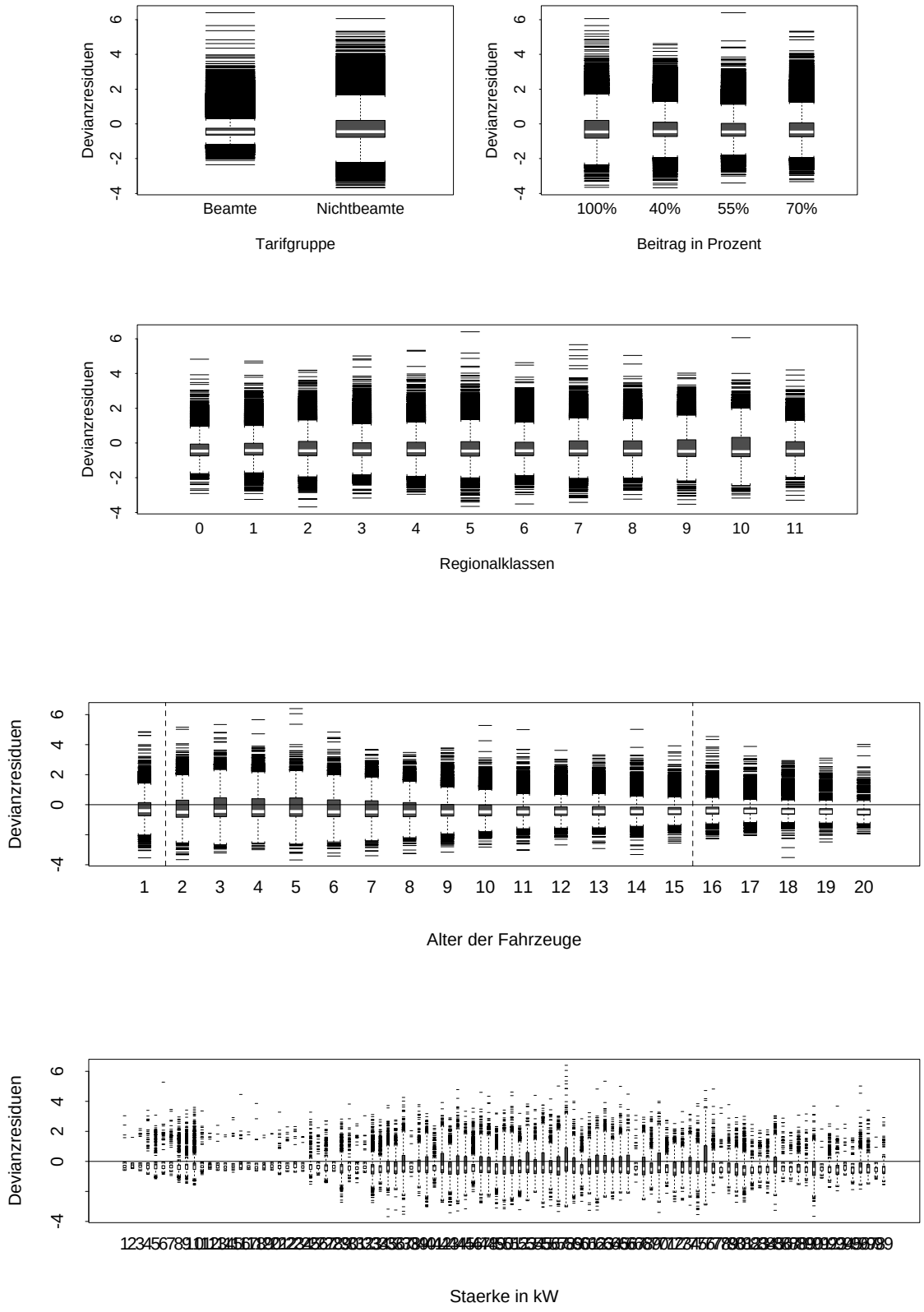


Abbildung 4.20: Residuenanalyse der Haupteffekte im Endmodell

Bei der Betrachtung der kategoriellen Regressoren in Abbildung 4.20 fällt die Zunahme der Devianz am deutlichsten in der Tarifgruppe auf. Dort sind insbesondere bei den Nichtbeamten neue Ausreißer nach oben zu erkennen. Das verwundert nicht weiter, denn dieses Merkmal wurde durch die Rückwärtsanalyse am stärksten beeinflußt. Im Gegensatz dazu wurde kein einziger Regressor, der mit dem Prämienbeitrag zu tun hat, aus dem Modell entfernt. Dies drückt sich auch in der fast unveränderten Residuengraphik dieses Haupteffekts rechts oben in Abbildung 4.20 aus. Ebenfalls bei den stetigen Haupteffekten wirkt sich die Reduktion der Parameter geringfügig auf die beiden Darstellungen in Abbildung 4.20 aus. Wir beobachten bei allen Regressoren auch die gleiche systematische negative Abweichung des Medians von 0 wie im Ausgangsmodell, so daß das Endmodell mit weniger Parametern gegenüber dem Ausgangsmodell zu bevorzugen ist.
Zusammenfassend halten wir fest, daß das Endmodell der Rückwärtsanalyse mit 104 Parametern die Daten in dem gleichen Maße anpaßt wie das Ausgangsmodell mit 129 Parametern. Jedoch beurteilen wir die Anpassung als nicht befriedigend, weil die Varianz durch die Poissonverteilung falsch spezifiziert wird. Eine Verbesserung des Modells ist nötig und muß bei der Varianzfunktion ansetzen.

4.4 Überdispersionstests

Nachdem die graphische Analyse einen klaren Hinweis auf Überdispersion ergab, führen wir nun auch zur Bestätigung die in Kapitel 3 beschriebenen formellen Tests auf diesen Sachverhalt durch. In unseren betrachteten Modellen wählten wir die kanonische Linkfunktion und benutzten einen Intercept, so daß die Teststatistiken P_A (s. Formel (3.11)) und P_B (s. Formel (3.12)) nach der in Beispiel 2.7 gezeigten Gleichheit $\sum y_i = \sum \hat{\mu}_i$ identische Werte liefern. In der Praxis werden die Berechnungen der beiden Statistiken mit dem Computer erfolgen, weshalb aufgrund von Rundungsfehlern kleine Unterschiede auftreten können. Die Ergebnisse der drei Tests mit ihren zugehörigen p-Werten zeigt Abbildung 4.21.

```
                    Testen auf Überdispersion
                   ===========================

                          Ausgangsmodell
                          --------------

      PA        PB        PC
   21.38     21.38     30.08

 pWertPA   pWertPB   pWertPC
  3.7E-6    3.7E-6    4.1E-8

 6. Modell: ohne TG*Alter, TG*Stärke, TG*Regionalkl, Alter2bis15*Regionalkl
 ------------------------------------------------------------------------

      PA        PB        PC
 21.4197   21.4197   30.4630

 pWertPA   pWertPB   pWertPC
  3.7E-6    3.7E-6    4.0E-8
```

Abbildung 4.21: Ergebnisse der Überdispersionstests

Bereits für das Ausgangsmodell der Poissonregression verwerfen alle drei Teststatistiken hochsignifikant die Äquidispersionsannahme zugunsten der Überdispersion. Das bedeutet, daß die Reduzierung der Regressoren und die damit einhergehende leichte Zunahme der Streuung in der Rückwärtsanalyse nicht die Ursache für die Ablehnung der Äquidispersion im Endmodell ist. Da die echte Varianz im Datensatz größer als die von dem Poissonmodell angenommene ist, liefern die Teststatistiken der Anova-Tabellen zur Signifikanz eines Regressors zu hohe Werte, womit die Signifikanz der im Endmodell verbliebenen Regressoren in Frage gestellt ist. Wir werden darum die Datenanalyse mit einer Verteilung fortsetzen, die auch Überdispersion zuläßt. Wir müssen dazu die Varianzfunktion neu wählen. Die Ablehnung der Äquidispersionsannahme bei beiden Teststatistiken P_B und P_C erlaubt uns sowohl die Verwendung einer linearen als auch einer quadratischen Varianzfunktion. Aufgrund der größeren Steigung der quadratischen Varianzfunktion und der auch bei kleinen Schäden groß geschätzten Varianz entscheiden wir uns für die quadratische Varianzfunktion.

4.5 Negative Binomialregression

Im folgenden modellieren wir die Daten, indem wir eine negative Binomialverteilung (s. Definition 2.18) mit quadratischer Varianzfunktion $V(\mu_i) = \mu_i + \tau\mu_i^2, \tau > 0$, zugrunde legen. Ist der Nichtlinearitätsparameter τ bekannt, so gehört die negative Binomialverteilung zur exponentiellen Familie. Wir verfahren hier so, daß wir den Nichtlinearitätsparameter zuerst mit der ML-Methode schätzen. Diesen Schätzer $\hat{\tau}$ behandeln wir in der anschließenden negativen Binomialregression als wahren Wert, wohlwissend, daß wir damit die aus der Schätzung von τ entstehende Varianz unterschlagen. Die Zugehörigkeit der negativen Binomialverteilung mit quadratischer Varianzfunktion und bekanntem Nichtlinearitätsparameter zur exponentiellen Familie nutzen

wir aus, um die Daten weiterhin mit der SAS-Prozedur GENMOD zu bearbeiten. Die logarithmische Linkfunktion und die Transformationen der stetigen Haupteffekte aus der Poissonregression behalten wir bei, damit die Vergleichbarkeit des neuen Modells mit dem Endmodell der Poissonregression gewährleistet ist. Wir wollen mit einer Rückwärtsanalyse herausfinden, ob die als signifikant eingestuften Regressoren aus dem Endmodell auch unter der flexibleren Varianzannahme ihre Signifikanz bewahren. Darum ist die Spezifizierung des Erwartungswerts im Startmodell der negativen Binomialregression identisch mit der Spezifizerung im Endmodell der Poissonregression. Wir beachten, daß die Devianz im negativen Binomialmodell von dem Nichtlinearitätsparameter abhängt, so daß wir den Schätzer von τ aus dem ersten Modell der negativen Binomialregression für die weiteren Modelle in der Rückwärtsanalyse konstant halten. Dies ermöglicht uns, die negativen Binomialmodelle der Rückwärtsanalyse miteinander zu vergleichen. Als Wert für den Schätzer von τ ermittelt die Prozedur GENMOD $\hat{\tau} = 0,0459$. Dies ist übrigens der gleiche Schätzwert wie derjenige, den wir erhalten, wenn wir das Ausgangsmodell der Poissonregression als Startmodell für die Rückwärtsanalyse der negativen Binomialregression benutzen. Das werten wir als ein Indiz, daß auch unter der neuen Verteilungsannahme die beiden Modelle recht ähnlich sind. Wir runden $\hat{\tau}$ auf $0,046$ und geben das Ergebnis der Regression in Abbildung 4.22 an.

NB2-Regression mit dem sechsten Modell und tau=0.046
wie 6. Modell der Poissonregression

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	51E3	41498.7857	0.8179
Scaled Deviance	51E3	41498.7857	0.8179
Pearson Chi-Square	51E3	61242.6406	1.2070
Scaled Pearson X2	51E3	61242.6406	1.2070
Log Likelihood	51E3	108209.2591	

LR-Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
staerke	1	138.63	<.0001
regiokl	11	117.58	<.0001
tg	1	27.43	<.0001
age1	1	8.47	0.0036
age2to15	1	13.16	0.0003
age16min	0	0.00	.
beitrag	3	21.87	<.0001
staerke*beitrag	3	207.34	<.0001
staerke*regiokl	11	55.00	<.0001
staerke*age1	1	19.68	<.0001
staerke*age2to15	1	20.40	<.0001
staerke*age16min	1	26.88	<.0001
regiokl*age1	11	13.04	0.2908
regiokl*age16min	11	12.74	0.2388
regiokl*beitrag	33	59.36	0.0033
tg*beitrag	3	9.03	0.0289
age1*beitrag	3	22.93	<.0001
age2to15*beitrag	3	39.61	<.0001
age16min*beitrag	3	52.11	<.0001

Abbildung 4.22: Ausgangsmodell der Rückwärtsanalyse mit negativer Binomialverteilung

Wir betonen hier nochmals, daß wir die sprunghafte Abnahme der Devianz bzw. den Anstieg der log-Likelihood bezogen auf die Poissonregression nur bedingt als Kriterien für eine verbesserte Anpassung heranziehen dürfen, denn je größer wir τ wählen, desto kleiner können wir die Devianz drücken bzw. desto größer die log-Likelihood wachsen lassen. Abbildung 4.22 zeigt, daß nun die Interaktion regiokl*age1 als nichtsignifikant eingestuft und daß die Interaktion regiokl*age16min als linear unabhängig behandelt wird (vgl. Abbildung 4.11). Auch für die Rückwärtsanalyse der negativen Binomialregression stellen wir nicht die einzelnen Schritte dar, sondern geben in Abbildung 4.24 gleich das Endmodell an, dem die Interaktionen `regiokl*age1` und `regiokl*age16min` nicht mehr angehören.

NB2-Regression mit dem achten Modell und tau=0.046
wie 6. Modell ohne regiokl*Alter

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	51E3	41524.9660	0.8181
Scaled Deviance	51E3	41524.9660	0.8181
Pearson Chi-Square	51E3	61273.9739	1.2071
Scaled Pearson X2	51E3	61273.9739	1.2071
Log Likelihood	51E3	108196.1689	

LR-Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
staerke	1	138.54	<.0001
regiokl	11	159.83	<.0001
tg	1	27.46	<.0001
age1	1	8.45	0.0037
age2to15	1	13.00	0.0003
age16min	0	0.00	.
beitrag	3	22.06	<.0001
staerke*beitrag	3	207.58	<.0001
staerke*regiokl	11	56.28	<.0001
staerke*age1	1	19.47	<.0001
staerke*age2to15	1	20.19	<.0001
staerke*age16min	1	26.56	<.0001
regiokl*beitrag	33	59.55	0.0031
tg*beitrag	3	8.92	0.0303
age1*beitrag	3	22.75	<.0001
age2to15*beitrag	3	39.61	<.0001
age16min*beitrag	3	52.33	<.0001

Abbildung 4.24: Endmodell der negativen Binomialregression

Es werden zwar nur drei Schritte der Rückwärtsanalyse benötigt, doch beinhalten diese eine weitere Parameterreduzierung auf schließlich 80 Parameter im letzten Modell. Auch inhaltlich ist das Entfernen der Interaktionen regiokl*age1 und regiokl*age16min bedeutsam, denn es leuchtet viel leichter ein, daß die Interaktion zwischen Regionalklasse und Fahrzeugalter komplett entfällt, als daß dies nur für bestimmte Altersbereiche der Fall ist, was ja das Ergebnis der Poissonregression ist.

Um die Graphiken der Residuenanalyse adäquat beurteilen zu können, verdeutlichen wir zuvor, was der Wechsel der Verteilungsannahme für die Parameterschätzung und damit für die Schätzer der Schadenanzahlen bedeutet.

Unter der Poissonverteilung mit kanonischer Linkfunktion lösen die ML-Schätzer $\hat{\beta}$ der Parameter β nach Beispiel 2.5 das Gleichungssystem $\frac{\partial \ell}{\partial \beta} = \sum_{i=1}^{n} (y_i - \mu(\mathbf{x}_i, \beta))\, \mathbf{x}_i = \mathbf{0}$ mit ℓ als log-Likelihood der gesamten Stichprobe und n als Anzahl der Beobachtungen. Benutzen wir nun die negative Binomialverteilung mit quadratischer Varianzfunktion $V(\mu_i) = \mu_i + \tau\mu_i^2$ und ebenfalls mit logarithmischer Linkfunktion, dann sind die ML-Schätzer $\hat{\beta}$ gemäß Formel (2.8) Lösungen des Gleichungssystems $\frac{\partial \ell}{\partial \beta} = \sum_{i=1}^{n} (y_i - \mu(\mathbf{x}_i, \beta)) \frac{1}{\mu(\mathbf{x}_i,\beta)+\tau\mu^2(\mathbf{x}_i,\beta)} \mu(\mathbf{x}_i, \beta)\mathbf{x}_i =$

$\sum_{i=1}^{n} (y_i - \mu(\mathbf{x}_i, \beta)) \frac{1}{1+\tau\mu(\mathbf{x}_i,\beta)} \mathbf{x}_i = \mathbf{0}$. Die Gleichungssysteme der beiden Verteilungen unterscheiden sich also lediglich durch den Term $\frac{1}{1+\tau\mu(\mathbf{x}_i,\beta)}$. In unserem Datensatz ist $\tau = 0,046$ recht klein gewählt, und $\mu(\mathbf{x}_i, \beta)$ liegt bei den meisten Beobachtungen zwischen 0 und 20, so daß der Faktor $\frac{1}{1+\tau\mu(\mathbf{x}_i,\beta)}$ sich für jede Beobachtung nur wenig ändert und als angenähert unabhängig von i betrachtet werden kann. Ein konstanter Faktor beeinflußt aber nicht die ML-Schätzung von β, weshalb wir im negativen Binomialmodell ähnliche Schätzwerte für die Schadenanzahl erwarten wie im Poissonmodell. Das bedeutet insbesondere, daß wir kaum mit Veränderungen bei der Überprüfung der Spezifizierung des Erwartungswerts rechnen dürfen. Wir haben mit der Änderung der Verteilungsannahme hauptsächlich eine größere Variation in den Daten zugelassen und erwarten darum vor allem bei den hohen Schadenanzahlen eine verbesserte Modellierung.

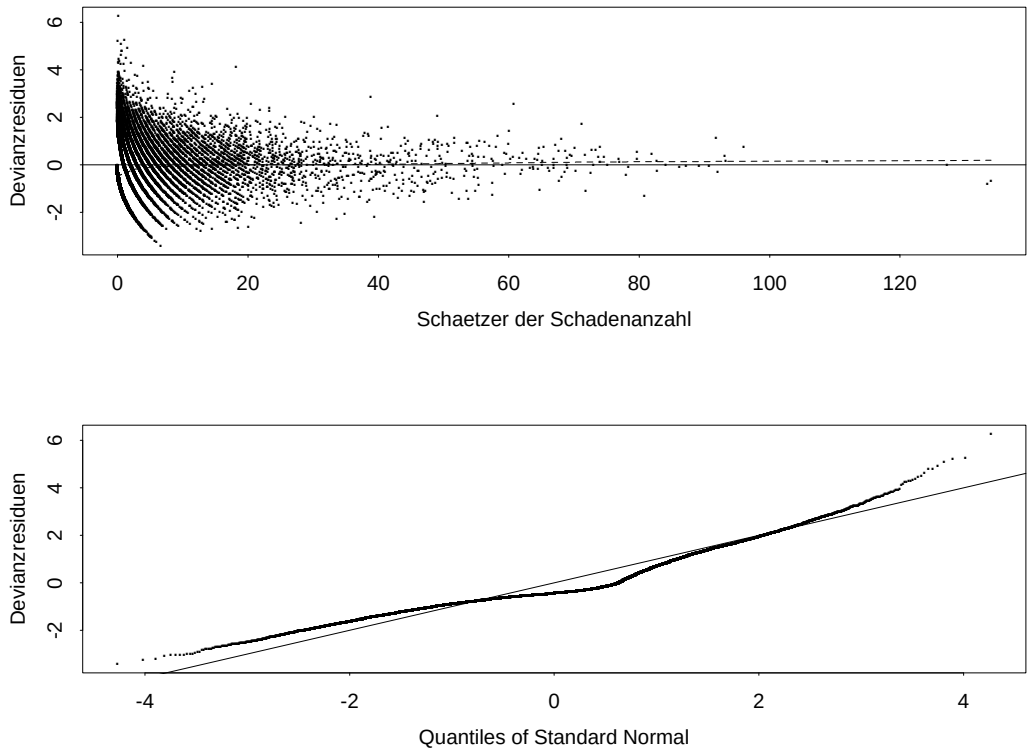


Abbildung 4.25: Gesamtanpassung im negativen Binomialmodell
gestrichelte Linie: lowess-*Funktion*

Abbildung 4.25 gibt die beschriebene erwartete Anpassung wider. Bereits ab einer geschätzten Anzahl von 30 Schäden streuen die Devianzresiduen in einem zufriedenstellend engen Intervall um 0, wie die obere Graphik aus Abbildung 4.25 zeigt. Wir erkennen an den Devianzresiduen und an dem Graphen der *lowess*-Funktion außerdem, daß vor allem die Beobachtungen mit einer klein geschätzten Schadenanzahl (< 10) schlecht modelliert werden.

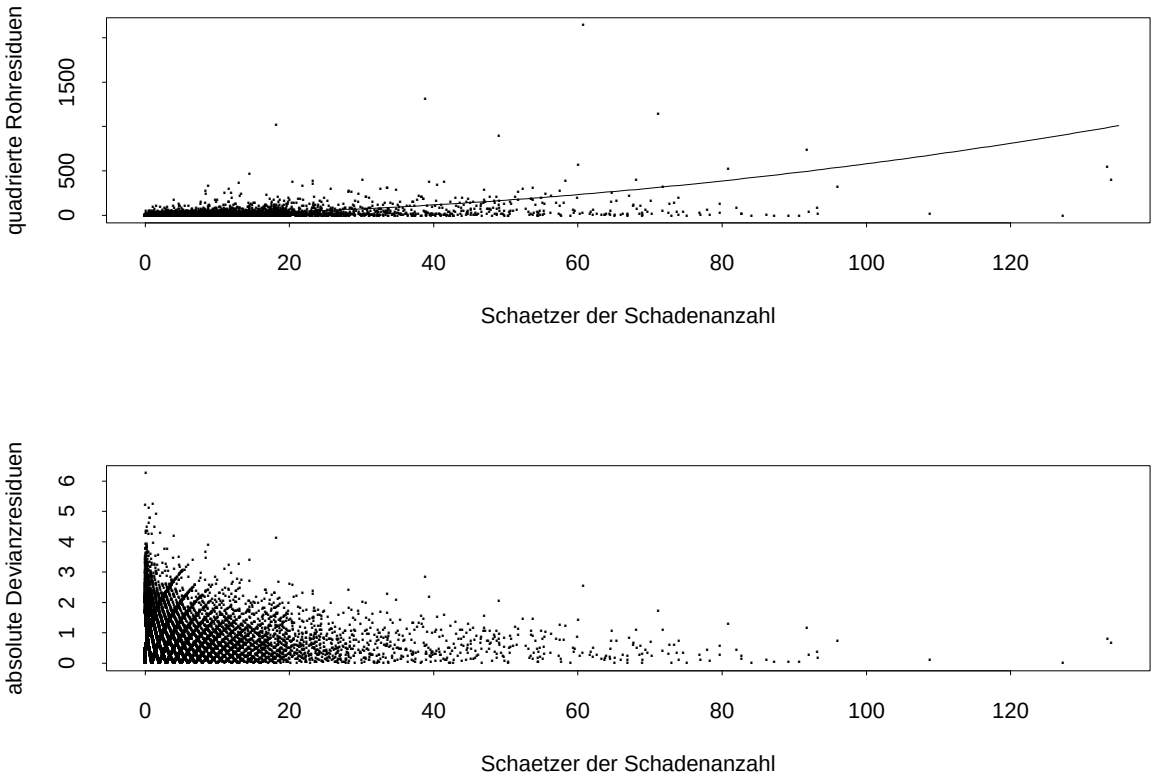


Abbildung 4.26a: Varianzfunktion im negativen Binomialmodell
durchgezogene Linie: Varianzfunktion $V(\mu_i) = \mu_i + 0,046\mu_i^2$ *im NB2-Modell*

Die Graphiken aus den Abbildungen 4.26a und 4.26b zur Überprüfung der Varianzfunktion erklären, warum das so ist. Vergleichen wir die Punktewolken aus der oberen Graphik von Abbildung 4.26a und aus Abbildung 4.26b mit den entsprechenden Graphiken in Abbildung 4.18 und Abbildung 4.6b für die Poissonverteilung, dann stellen wir fest, daß die Schadenanzahlen von dem negativen Binomialmodell ungefähr genauso geschätzt werden wir im Poissonmodell. Das heißt für die Beobachtungen mit maximal 10 geschätzten Schäden, daß ihre Varianz auch von der quadratischen Varianzfunktion des negativen Binomialmodells stark systematisch unterschätzt wird, während die Varianz für Beobachtungen mit mehr als 100 geschätzten Schäden schon überschätzt wird. Wir halten also fest, daß die Variabilität in den Daten von der quadratischen Varianzfunktion zwar deutlich besser modelliert wird als von der linearen Varianzfunktion der Poissonverteilung, doch der beobachtete funktionelle Zusammenhang zwischen Erwartungswert und Varianz ist auch damit nicht adäquat beschrieben.

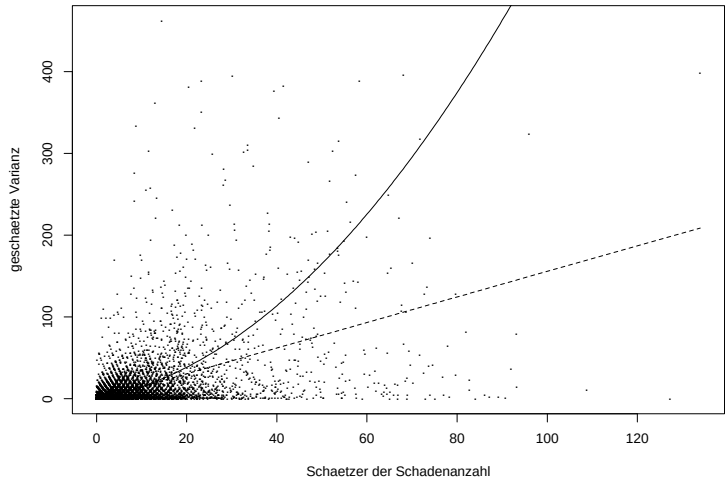


Abbildung 4.26b:
Varianzfunktion im NB2-Modell, eingeschränkt auf Schätzwerte der Varianz < 500
durchgezogene Linie: Varianzfunktion $V(\mu_i) = \mu_i + 0,046\mu_i^2$ *im NB2-Modell*
gestrichelte Linie: lowess-*Funktion*

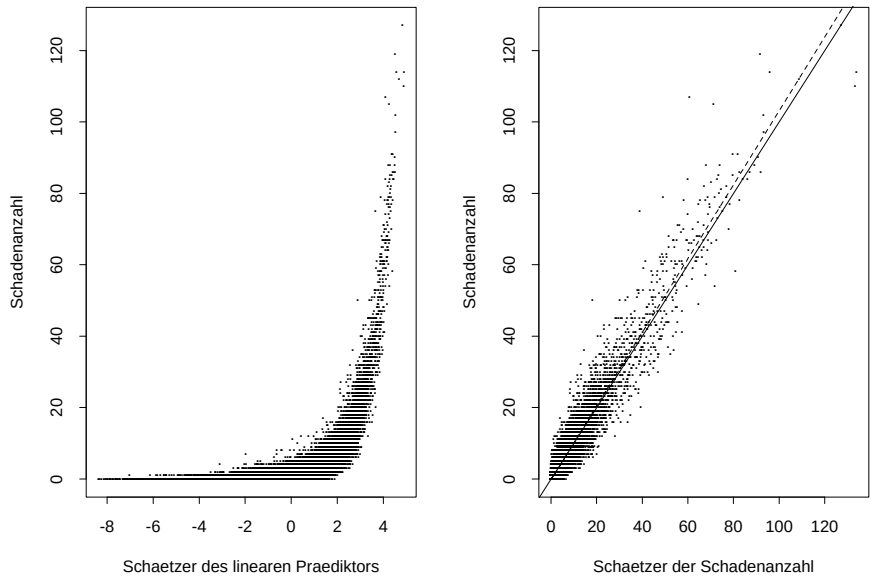
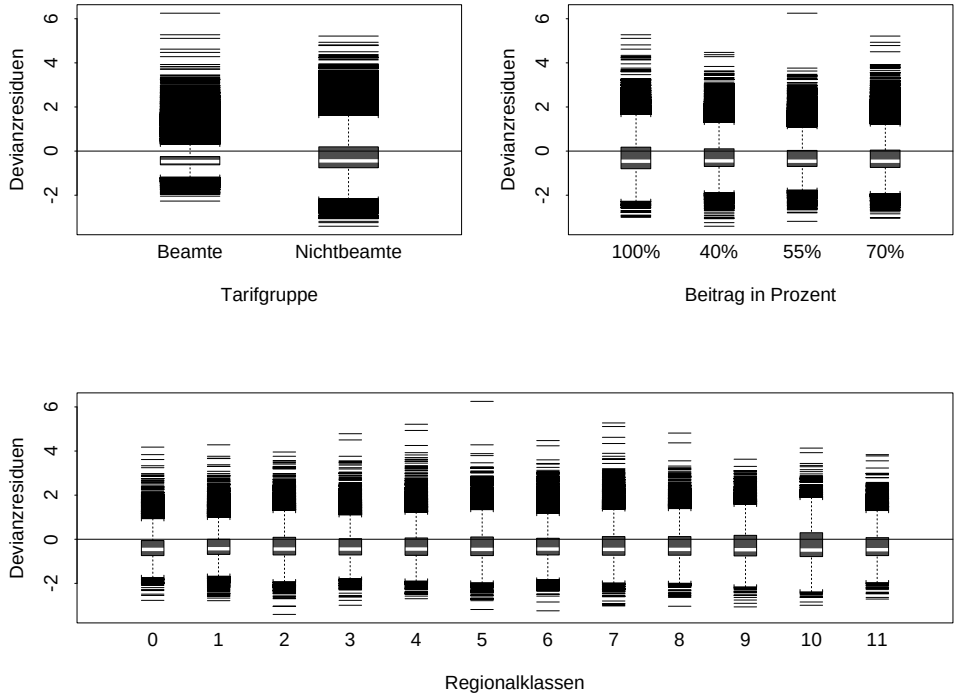


Abbildung 4.27: Linkfunktion im negativen Binomialmodell
durchgezogene Linie: exponentielle Linkfunktion μ_i
gestrichelte Linie: lowess-*Funktion*

Die Betrachtung der Linkfunktion in Abbildung 4.27 bestätigt die oben erläuterte Ähnlichkeit der Schätzer im negativen Binomialmodell und Poissonmodell, so daß die graphische Analyse

von Abbildung 4.27 zum gleichen Schluß kommt wie die von Abbildung 4.7: der Logarithmus ist die treffende Wahl der Linkfunktion.



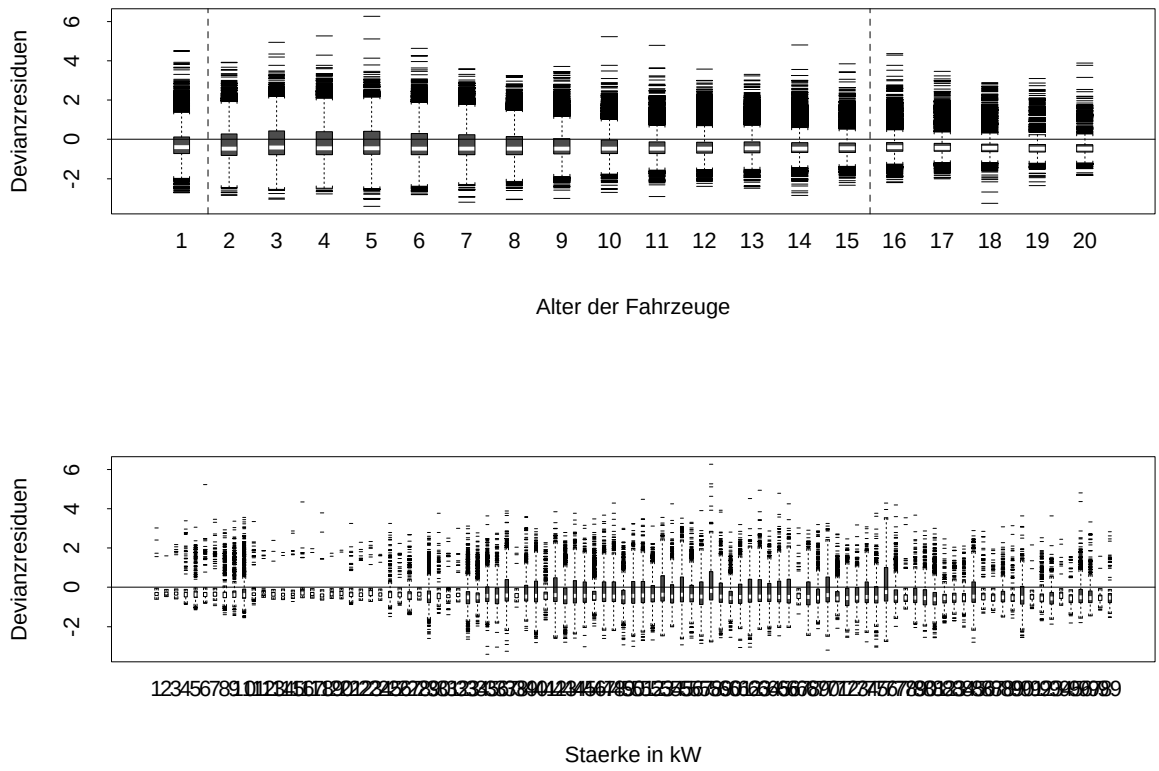


Abbildung 4.28: Residuenanalyse der Haupteffekte im negativen Binomialmodell

Die Residuenanalyse der Haupteffekte in Abbildung 4.28 zeigt, daß auch im negativen Binomialmodell der Median in allen Ausprägungen kleiner als 0 ist, was wiederum auf die Schiefe der Verteilung zurückzuführen ist. Die Boxplots zeigen andererseits auch, daß durch die größere Varianz des negativen Binomialmodells die Anzahl der Ausreißer abgenommen hat, zum Beispiel bei den Nichtbeamten oder der Regionalklasse 5, womit die Wahl einer flexibleren Varianzfunktion bestätigt wird. Damit stimmt die graphische Analyse mit unseren Erwartungen überein.

4.6 Diskussion und Fazit

Die hier untersuchten Regressionsmodelle setzen voraus, daß die beobachteten Schadenanzahlen bei gegebenen Regressoren unabhängig verteilt sind. Diese Annahme ist aus drei Gründen verletzt. Zum einen wurden die Daten über einen Zeitraum von vier Jahren erhoben, so daß ein Risiko vier Jahreseinheiten zählt, wenn es die ganze Zeit versichert war, während erst im Laufe des Beobachtungszeitraums dazugekommene Risiken mit weniger Jahreseinheiten gewichtet werden. Der zweite Grund, warum die Unabhängigkeitsvoraussetzung nicht gegeben ist, besteht darin, daß jedes betrachtete Risiko nicht automatisch einen anderen Versicherungsnehmer bzw. Fahrzeugfahrer darstellt. Wenn ein Versicherungsnehmer mehrere Fahrzeuge besitzt, muß er

für jedes Fahrzeug einen eigenen Versicherungsvertrag abschließen und wird deshalb mehrfach gezählt. Es ist naheliegend, daß das Schadenverhalten eines Versicherungsnehmers in verschiedenen Fahrzeugen nicht als unabhängig betrachtet werden kann. Weiterhin kann es vorkommen, daß ein Versicherungsnehmer während des Beobachtungszeitraums sein Fahrzeug aufgibt und sich ein neues kauft, wodurch er als neues, aber nicht unabhängiges Risisko in die Daten eingeht. Das Ausmaß der Unabhängigkeitsverletzung ist aus den Daten nicht zu erkennen, dazu liegen sie zu anonymisiert vor. Doch wir vermuten, daß der erstgenannte Grund mit Abstand den größten Einfluß auf die Datenanalyse hat. Eine Reduzierung dieses Fehlers kann nur bei der Planung der Untersuchung und der Datenerhebung erfolgen, indem wir uns beispielsweise nur auf ein Jahr als Beobachtungszeitraum beschränken und nur solche Risiken in die Untersuchung aufnehmen, die exakt eine Jahreseinheit an Volumen einbringen.

Aufgrund von Abbildung 4.2 in der explorativen Datenanalyse haben wir uns zu einer Aufteilung des Merkmals Fahrzeugalter in drei neue Regressoren, von denen wir zwei transformierten, entschieden. Nun könnte uns diese Graphik unter Vernachlässigung der neuen Fahrzeuge mit Alter = 1 Jahr dazu verleiten, die restlichen Punkte als annähernd linear aufzufassen. Damit ginge das Fahrzeugalter untransformiert in das Regressionsmodell ein. Wenn dieses untransformierte Regressionsmodell mit seiner einfachen Struktur die gleiche Anpassung wie die untersuchten Modelle besitzt, ist es zu bevorzugen. Wir überprüfen diese Möglichkeit, indem wir im Endmodell der Poissonregression die transformierten Altersvariablen gegen die untransformierten austauschen und erneut eine Poissonregression durchführen. Das untransformierte Regressionsmodell besitzt 32 Freiheitsgrade weniger als das Endmodell mit den transformierten Altersregressoren und eine Devianz von 45 566,38, was einer Differenz von 147 zum Endmodell entspricht. Der Test auf Gleichheit beider Modelle liefert einen asymptotischen p-Wert von $0,00$, womit die Nullhypothese der Gleichheit hochsignifikant verworfen wird. Mit diesen Überlegungen ist die vorgenommene Aufspaltung und Transformation des Fahrzeugalters gerechtfertigt. Da wir die Fahrzeugstärke untransformiert in den Modellen verwendeten und die anderen Regressoren als Indikatorvariablen, beurteilen wir die Modellierung des linearen Prädiktors als richtig. Mit der ausgezeichneten Wahl des Logarithmus als Linkfunktion halten wir insgesamt fest, daß der Erwartungswert korrekt spezifiziert wurde.

Damit ist nur eine der Voraussetzungen für die Überdispersionstests, nämlich unabhängige Beobachtungen, nicht vollständig erfüllt. Dennoch halten wir an der Ablehnung der Äquidispersionsannahme, die aus den Teststatistiken sowie der Residuenanalyse folgt, fest. Dies liegt zum einen an den sehr hohen Werten aller drei Teststatistiken und zum anderen an einem weiteren empirischen Kriterium, und zwar dem Vergleich von Mittelwert und Stichprobenvarianz der Schadenhäufigkeit aus dem Datensatz. Wir erhalten für den Mittelwert $\bar{SH} = 0,14$ und für die Stichprobenvarianz $\sigma^2_{SH} = 1,34$. Als Faustregel geben Cameron und Trivedi [1998, S. 77] an, daß die Daten auch nach der Hinzunahme der Regressoren wahrscheinlich Überdispersion aufweisen, wenn die Stichprobenvarianz mehr als dpooelt so groß wie der Mittelwert ist. Das ist hier offensichtlich der Fall.

Die Gründe für das Auftreten von Überdispersion in unseren Daten sind vielfältig. Fast alle in

Abschnitt 2.2.1 genannten Gründe kommen in Frage. Wir nehmen insbesondere an, daß weitere Regressoren wie das Wetter oder die Lichtverhältnisse zum Zeitpunkt des Schadeneintritts zur Beschreibung der Schadenanzahl fehlen. Desweiteren wird zwar das persönliche Schadenverhalten eines Versicherungsnehmers durch seine Klasse im Bonus-Malus-System bzw. seinen Beitragssatz charakterisiert, doch reicht dieser Haupteffekt allein nicht aus, um Aggresivität am Steuer, Kenntnis der Straßenverkehrsordnung und Trinkgewohnheiten vollständig zu beschreiben, wie die Signifikanz sämtlicher Interaktionen mit dem Beitragssatz in allen Modellen belegt. Auch unsere Entscheidung, die Beitragssätze anstelle der Schadenfreiheitsklassen als Regressor in das Modell zu nehmen, sorgt für einen Erklärungsverlust, da die Schadenfreiheitsklassen 1/2 und 1 mit dem gleichen Rabatt versehen sind und wir dadurch einen Freiheitsgrad verlieren. Aber auch die Art der Einstufung in eine Schadenfreiheitsklasse und das Verhalten der Versicherungsnehmer tragen zu einer verzerrten Widergabe der individuellen Schadenneigung bei. Wie in Abschnitt 4.1 erwähnt, dient ausschließlich die Schadenanzahl als Umstufungskriterium. Um Rückstufungen im Bonus-Malus-System zu vermeiden, bezahlen Versicherungsnehmer kleinere Schäden häufig selbst. Durch die Strategie, Schäden selbst zu regulieren, befinden sich Versicherungsnehmer oftmals in Klassen, die ihrem tatsächlichen persönlichen Schadenverlauf nicht entsprechen. Obwohl gleich hohe Prämien erhoben werden, setzen sich die Klassen dann zunehmend heterogen zusammen, so daß die korrekte individuelle Schadenneigung durch das praktizierte schadenhöhenunabhängige System nicht exakt modelliert werden kann. Auswege aus der mangelhaften Beschreibung des individuellen Schadenverhaltens kann die Hinzunahme von Stellvertretermerkmalen wie Alter des Fahrers bzw. des Versicherungsnehmers oder gefahrene Jahreskilometerleistung mit allen damit verbundenen Nachteilen sein oder die Einführung eines neuen Bonus-Malus-Systems, das auch die Schadenhöhe berücksichtigt und weitere Klassen enthält. Wie wir anhand von Abbildung 4.1 feststellen, sind die niedrigen Schadenfreiheitsklassen schwach besetzt, während die Schadenfreiheitsklasse 3 fast zwei Drittel aller Jahreseinheiten umfaßt. Die große Besetzung dieser Klasse wird wegen des oben beschriebenen Verhaltens der Versicherungsnehmer in den Jahren nach unserer Untersuchung noch zunehmen, so daß sich mit steigender Belegung auch die Heterogenität erhöht. Eine Neueinführung von Klassen mit niedrigen Beitragssätzen steuert dieser Entwicklung entgegen.

Neben der erläuterten inhaltlichen Erklärung für das Auftreten von Überdispersion gibt es auch mathematische Ursachen zu betrachten. So lassen GLMe nur einen linearen Zusammenhang zwischen Parameter und Regressor zu, wodurch ein tatsächlich vorhandener nichtlinearer Zusammenhang zu Überdispersion führt. Diese Art der Überdispersion kann überhaupt nicht durch die Veränderung der Verteilung modelliert werden. Ferner bewiesen wir in Abschnitt 2.2.1, daß auch die irrtümliche Annahme, daß die zugrundeliegende Beobachtungseinheit, hier Jahreseinheiten, fest ist statt zufällig, zu Überdispersion führt. Winkelmann [1994, S. 70] führt als ein Beispiel dafür an, daß häufig kurz nach einem eingetretenen Schaden kein neuer entstehen kann, da sich beispielsweise das Fahrzeug in Reparatur befindet oder gar durch ein neues ersetzt werden muß. Nun setzen sich die Jahreseinheiten definitionsgemäß aus der Anzahl der Versicherungsnehmer und ihrer Versicherungsdauer als zeitliche Komponente zusammen, so daß wir sie eigentlich als

zufällig betrachten müssen. Doch wir beurteilen den Effekt dieses Mechanismus zur Erzeugung von Überdispersion als nebensächlich, denn die Wahrscheinkeit für den Schadenseintritt ist wegen $\bar{SH} = 0,14$ eher klein und die Zeitspanne nach einem Schaden, in der das Fahrzeug nicht genutzt werden kann, ist im Vergleich zu dem Beobachtungszeitraum von vier Jahren recht kurz. Verlassen wir die Terminologie der GLMe und betrachten wir die Daten aus der Sicht der stochastischen Prozesse, so können wir die Überdispersion inhaltlich auch damit beschreiben. Wir erinnern, daß ein Poissonprozeß die Anzahl von Ereignissen, hier Schäden, während eines festen Zeitintervalls angemessen beschreibt, wenn die Ereignisse unabhänigig voneinander und mit konstanter Ausfallrate eintreten. In Satz 2.16 haben wir gezeigt, daß Überdispersion auftritt, sobald die Unabhängigkeitsannahme verletzt ist und die Schäden in Gruppen bestehend aus einer zufälligen Anzahl von Einzelschäden auftreten. Wir interpretieren diesen Sachverhalt so, daß ein unbeobachtetes Ereignis wie ein plötzlicher Schlechtwettereinbruch zu einer Reihe von Schäden führen, deren Anzahl zufällig ist. Abbildung 4.29 skizziert diese Art der Überdispersion.

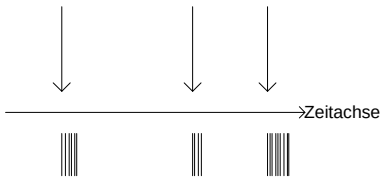


Abbildung 4.29: Die Pfeile oberhalb der Zeitachse markieren einen Schlechtwettereinbruch, aufgrund dessen sich eine zufällige Anzahl von Schäden ereignet, die durch Striche unterhalb der Zeitachse dargestellt sind.

In Satz 2.17 bewiesen wir, daß Überdispersion auch durch eine monoton fallende Ausfallrate hervorgerufen wird. Eine konstante Ausfallrate für unseren Datensatz anzunehmen, was wir mit der Poissonverteilung der Schadenanzahl tun, bedeutet, daß jeder Versicherungsnehmer über den gesamten Beobachtungszeitraum hinweg ohne Beachtung der von ihm verursachten Schäden eine gleich hohe individuelle Schadenneigung besitzt. Realistischer ist jedoch die Annahme, daß ein Versicherungsnehmer im Laufe der Zeit Fahrerfahrung sammelt und seine individuelle Schadenneigung abnimmt. Damit modellieren wir gerade die monoton fallende Ausfallrate für einen Versicherungsnehmer. Obwohl wir diskutierten, daß der Aufenthalt eines Versicherungsnehmers in einer Schadenfreiheitsklasse nicht immer seinem tatsächlichen Schadensverlauf entspricht, ziehen wir die hohe Besetzung der größten Schadenfreiheitsklasse als Stütze für die Annahme einer monoton fallenden Ausfallrate heran. Denn eine sinkende Schadenneigung drückt sich neben der Verringerung der Schadenanzahl auch in der Reduzierug der Schadenhöhe aus, die ja in dem praktizierten Bonus-Malus-System nicht berücksichtigt wird.
Nachdem wir das Auftreten der Überdispersion auf verschiedene Weisen erklären können, stellt sich die Frage nach einer angemessenen Modifizierung der Varianzfunktion. Die Graphen der

lowess-Funktion in den Abbildungen 4.6b und 4.26b legen eine lineare Varianzfunktion nahe. Aber die Graphiken zeigen auch, daß jede lineare Varianzfunktion die Variabilität in den Daten nicht in befriedigendem Maße modelliert, weil entweder die Steigung des Graphen für kleine Werte der geschätzen Schadenanzahl zu niedrig ist oder für groß geschätzte Schadenanzahlen zu hoch. Diese Schwierigkeit tritt ebenfalls bei der hier gewählten quadratischen Varianzfunktion auf. Aufgrund der Punktewolken in den Abbildungen 4.6a, 4.6b, 4.18, 4.26a und 4.26b scheint eine konvexe Varianzfunktion eine gute Wahl zu sein.
Zusammenfassend können wir festhalten, daß das Poissonmodell für den vorliegenden Datensatz wegen der zu strengen Äquidispersionsannahme keine gute Anpassung liefert. Dagegen modelliert die negative Binomialregeression mit quadratischer Varianzfunktion die Daten zufriedenstellend, auch wenn die Modellierung verbesserungsfähig ist.
Ein möglicher Ansatz ist der Einsatz einer Vorwärtsanalyse im Rahmen der NB2-Regression, die mit dem Nullmodell als Ausgangsmodell startet und dann schrittweise einen der möglichen Regressoren in das Modell hinzunimmt. Auswahlkriterium für den Regressor ist ein möglichst kleiner asymptotischer p-Wert zum Test, ob die Hinzunahme dieses Regressors zum bisherigen Modell die Anpassung signifikant verbessert, unterhalb eines zuvor gewählten Signifikanzniveaus α. Die Vorwärtsanalyse ist beendet, wenn alle Regressorkandidaten, die nicht dem Modell angehören, einen p-Wert$> \alpha$ besitzen. Liefert die Vorwärtsanalyse dasselbe Modell wie die hier durchgeführte Rückwärtsanalyse, ist dies eine gute Bestätigung unseres Endmodells der NB2-Regression. Beinhaltet das Endmodell der Vorwärtsanalyse weniger und andere Regressoren als das Endmodell der Rückwärtsanalyse, wird unsere Modellierung nur teilweise abgesichert. Der Einfluß derjenigen Regressoren, die nur in einem der beiden Endmodelle auftreten, ist dann schwierig zu beurteilen und bedarf weiterer Untersuchung.
Ein anderer Ansatz besteht wegen der kleinen Schadenhäufigkeit von 0,14 und der daraus resultierenden geringen Wahrscheinlichkeit für zwei oder mehr Schäden pro Versicherungsnehmer darin, die Zähldatenstruktur zu vernachlässigen und eine Binomialregression durchzuführen. Die Zielvariable in einem solchen Modell ist eine Indikatorvariable, die den Wert 1 annimmt, wenn ein Versicherungsnehmer mindestens einen Schaden im Beobachtungszeitraum gemeldet hat, und 0 bei keinem Schaden. Die Binomialregression läßt sich auch für gruppierte Daten, wie sie hier vorliegen, durchführen, indem einfach die Summe aller Indikatorvariablen in einer Gruppe als neue Zielvariable betrachtet wird. Wieder können wir Regressionsmodelle mittels einer Vorwärts- und Rückwärtsanalyse erhalten und diese mit unserem Endmodell der NB2-Regression vergleichen. Eine große Anzahl von Regressoren, die in allen untersuchten Modellen auftauchen, sichert die Ähnlichkeit der verschiedenen Modellierungsansätze und den signifikanten Einfluß des Regressors auf die Schadenanzahl als Zielvariable.
Wir haben bereits erklärt, daß die Anzahl der von einem Versicherungsunternehmen registrierten Schäden nicht mit der Anzahl der tatsächlich eingetretenen Schäden übereinstimmen muß, da der Versicherungsnehmer Rückstufungen im Bonus-Malus-System vermeiden will. Dieses Verhalten des Versicherungsnehmers hängt von der Schadenhöhe ab: sobald diese einen bestimmten Schwellenwert überschritten hat, meldet der Versicherungsnehmer einen Schaden. Damit fassen

wir die Schadenhöhe als weiteren unbeobachteten Regressor auf, der die Schadenanzahl in Form einer Indikatorvariable beeinflußt mit dem Wert 1, falls die Schadenhöhe den Schwellenwert übersteigt, und dem Wert 0, wenn die Schadenhöhe unterhalb des Schwellenwerts bleibt. Wollen wir einen solchen Modellansatz wählen, dann ist ein diskretes gemischtes Poissonmodell geeignet. Auch hier handelt es sich bei dem Parameter der Poissonverteilung um keinen festen Wert, sondern um eine Zufallsvariable, die in Abhängigkeit von der unbeobachteten Indikatorvariable Schadenhöhe zwei verschiedene Werte annimmt. Im Gegensatz zum Regressionsmodell mit negativer Binomialverteilung, die nach Beispiel 2.20 als stetige gemischte Poissonverteilung mit gammaverteiltem Parameter aufgefaßt werden kann, besitzt der Parameter der Poissonverteilung in diesem Ansatz eine diskrete Dichte. Regressionsmodelle, die auf einer diskreten gemischten Poisson-Verteilung beruhen, werden z. B. in dem Artikel von Wang/Puterman/Cockburn/Le [1996, S. 381 ff] beschrieben.

Anhang A

A.1 Fisher-Information

Wenn X eine Zufallsvariable (oder einen Zufallsvektor) mit Dichte $f(x;\theta), \theta = (\theta_1, \ldots, \theta_k) \in I\!R^k$ ist, dann ist die Fisher-Information $\mathcal{I}(\theta)$ die Matrix mit den Einträgen

$$\mathcal{I}_{ij}(\theta) = E\left[\frac{\partial}{\partial\theta_i}\ln f(X;\theta)\frac{\partial}{\partial\theta_i}\ln f(X;\theta)\right] \qquad i,j = 1,\ldots,k. \tag{A.1}$$

Dennoch sollte der suggestive Begriff „Information“ nicht zu ernst genommen werden. Er wird hauptsächlich durch die Tatsache rechtfertigt, das unter den Regularitätsbedingungen $\mathcal{I}^{-1}(\theta)$ die kleinste asymptotische Varianz ist, die ein konsistenter Schätzer erreichen kann. Obwohl wir von der Information aus einer einzelnen Beobachtung X sprechen, beruht somit die Rechtfertigung auf der Annahme einer großen Stichprobe. Um zu erkennnen, was $\mathcal{I}(\theta)$ über den Informationsgehalt einer einzelnen Stichprobe aussagt, beachten wir, daß für $\theta \in I\!R$ die logarithmierte Ableitung

$$\frac{\partial}{\partial\theta}\ln f(X;\theta) = \frac{\frac{\partial}{\partial\theta}f(X;\theta)}{f(X;\theta)}$$

die relative Rate ist, mit der sich die Dichte $f(x;\theta)$ als Funktion von θ im Punkt x ändert. Die Größe $\mathcal{I}(\theta)$ ist gerade das erwartete Quadrat dieser Rate. Es ist plausibel, daß, je größer $\mathcal{I}(\theta)$ bei einem gegebenen Wert θ_0 ist, θ_0 umso einfacher von naheliegenden Werten θ zu unterscheiden ist und daß darum θ umso genauer in θ_0 geschätzt werden kann. Allgemeiner für $\theta \in I\!R^k$ ausgedrückt: Wenn es eine eindeutige Beobachtung von X gibt, die fast sicher einem Wert des Parameters θ entspricht, hat die Zufallsvariable maximale Information. Andererseits können wir, wenn die Zufallsvariable für alle Werte von θ dieselbe Verteilung hat, aufgrund des beobachteten Werts von X keine Aussagen über θ machen.
Beispiele, die widerlegen, daß es sich bei der Fisher-Information um ein umfassendes Maß für den Zusammenhang zwischen X und θ handelt, befinden sich in Lehmann [1999, S. 462 f].
Wir geben noch einige wichtige Eigenschaften der Fisher-Information ohne Beweis an. Die Beweise können in Rao [1973, S. 329 ff] nachgelesen werden.
Unter den Regularitätsannahmen aus 2.1.4 besitzt die Fisher-Information die alternative Darstellung

$$\mathcal{I}_{ij}(\theta) = -E\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\ln f(X;\theta)\right] \qquad i,j = 1,\ldots,k. \tag{A.2}$$

Die Gleichheit von (A.1) und (A.2) nennen wir auch Informationsgleichung.
Außerdem gilt die Additivität der Fisher-Information: Seien X und Y zwei unabhängige Zufallsvariablen mit Dichten, die den Regularitätsbedingungen genügen. Wenn $\mathcal{I}_1(\theta)$, $\mathcal{I}_2(\theta)$ und $\mathcal{I}(\theta)$ die Fisher-Informationen von X, Y und (X,Y) bezeichnen, dann gilt:

$$\mathcal{I}(\theta) = \mathcal{I}_1(\theta) + \mathcal{I}_2(\theta)$$

Daraus folgt für den Spezialfall, daß $X_1, \ldots, X_n$ iid Zufallsvariablen mit gemeinsamer Dichte, die den Regularitätsbedingungen genügt, sind, die Fisher-Information der Stichprobe $(X_1, \ldots, X_n)$ gerade gleich $n \cdot \mathcal{I}_1(\theta)$ ist, wobei $\mathcal{I}_1(\theta)$ die gemeinsame Fisher-Information einer einzelnen Zufallsvariable $X_i, i = 1, \ldots, n$, ist.

A.2 Wahrscheinlichkeitserzeugende Funktion

Sei X eine Zufallsvariable, die nichtnegative Werte annimmt, mit $P(X = i) = p_i, i \in \mathbb{N}_0$. Die wahrscheinlichkeitserzeugende Funktion $\mathcal{P}$ von X ist gegeben durch

$$\mathcal{P}^{(X)}(s) = \mathcal{P}(s) = E(s^X) = \sum_{i=0}^{\infty} s^i p_i.$$

Die Funktion $\mathcal{P}(s)$ wird durch die $p_i, i \in \mathbb{N}_0$, definiert und definiert ihrerseits die p_i, da die Taylorentwicklung eindeutig ist.

Beispiel A.1 *Sei $X \sim Poi(\mu)$. Die wahrscheinlichkeitserzeugende Funktion lautet*

$$\mathcal{P}(s) = \sum_{i=0}^{\infty} s^i \frac{\mu^i}{i!} e^{-\mu} = e^{-\mu} \sum_{i=0}^{\infty} \frac{(\mu s)^i}{i!} = e^{-\mu+\mu s}.$$

Beispiel A.2 *Die Zufallsvariable X habe eine logarithmische Verteilung, d. h. $P(X = k) = \alpha\theta^k/k$ mit $\alpha = -[\ln(1-\theta)]^{-1}, k \in \mathbb{N}_0, \theta \in]0,1[$. Wegen der Taylorreihe $\ln(1-z) = -\sum_{k=0}^{\infty} \frac{z^k}{k}$ von $\ln(1-z)$ in 0 gilt für die wahrscheinlichkeitserzeugende Funktion von X:*

$$\mathcal{P}(s) = \sum_{k=0}^{\infty} s^k \alpha \frac{\theta^k}{k} = \alpha \sum_{k=0}^{\infty} \frac{(\theta s)^k}{k} = -\alpha \ln(1-\theta s).$$

Eine wichtige Eigenschaft von $\mathcal{P}(s)$ ist, daß $\mathcal{P}(s)$ für $|s| \leq 1$ konvergiert, weil $\mathcal{P}(1) = \sum_{i=0}^{\infty} p_i = 1$. Insbesondere ermöglicht uns die absolute Konvergenz das gliedweise Differenzieren von $\mathcal{P}(s)$ und damit die Bestimmung der ersten beiden Momente:

$$E(X) = \sum_{i=0}^{\infty} i p_i = \mathcal{P}'(1). \tag{A.3}$$

Für die Varianz von X berechnen wir zuerst

$$E[X(X-1)] = \sum_{i=0}^{\infty} i(i-1) p_i = \mathcal{P}''(1)$$

und erhalten

$$Var\, X = [E[X(X-1)] + E(X) - [E(X)]^2 = \mathcal{P}''(1) + \mathcal{P}'(1) - [\mathcal{P}'(1)]^2. \tag{A.4}$$

A.3 Quelltext zum Programm Testen auf Überdispersion

```
/* Berechnung der Teststatistiken aus C.B. Dean: ''Testing for Overdispersion in
Poisson and Binomial Regression Models'', Journal of the American Statistical
Association, vol.87, 451-457
Die Bezeichnungen der drei Teststatistiken wird aus dem Artikel übernommen. Zur
Durchführung des Programms wird eine Datei benötigt, die bereits die beobachteten
und geschätzten Werte der Zielvariablen enthaelt, hier heißt sie modell7. Die
beobachteten Werte der Zielvariable tragen in diesem Programm die Bezeichnung
sanzahl und die ML-Schätzwerte der Poissonregression die Bezeichnung mue.

title1 ''Testen auf Überdispersion'';
title2 ''=========================='';

/* Im ersten DATA-Step werden Nenner und Zähler der verschiedenen Teststatistiken
für jede Beobachtung berechnet. */
data hilf;
set weg.modell7;
zaehlerA = (sanzahl - mue)**2 - mue;
zaehlerB = (sanzahl - mue)**2 - sanzahl;
nennerAB = mue*mue;
summandC = zaehlerB / mue;
keep zaehlerA zaehlerB nennerAB summandC;
run;

/* Im PROC-Step erfolgt die Summation von den Nennern und Zählern aus der Datei
hilf über alle Beobachtungen. */
proc means data=hilf noprint sum N;
output out=overdisp sum (zaehlerA zaehlerB nennerAB summandC) = zaehlerA zaehlerB
nennerAB summandC N=n;
run;

/* Endgültige Berechnung der drei Teststatistiken und der zugehörigen p-Werte */
data overtest;
set overdisp;
PA = zaehlerA / sqrt(2*nennerAB);
PB = zaehlerB / sqrt(2*nennerAB);
PC = summandC / sqrt(2*n);
pWertPA = 1 - prob(PA,1);
pWertPB = 1 - prob(PB,1);
```

```
pWertPC = 1 - prob(PC,1);
keep PA PB PC pWertPA pWertPB pWertPC;
run;

/* Ausgabe der Teststatistiken und der p-Werte */
proc print data=overtest noobs;
run;
```

Literaturverzeichnis

[1] R.E. Barlow, F. Proschan (1965), *Mathematical theory of reliability*, New York, John Wiley

[2] P.J. Bickel, K.A. Doksum (1977), *Mathematical statistics: basic ideas and selected topics*, Holden-Day

[3] A.C. Cameron, P.K. Trivedi (1990), „Regression-based tests for overdispersion in the Poisson model“, *Journal of Econometrics*, **46**, 347-364

[4] A.C. Cameron, P.K. Trivedi (1998), *Regression analysis of count data*, Cambridge University Press

[5] G. Casella, R.L. Berger (1990), *Statistical Interference*, Wadsworth & Brooks/Cole

[6] D.R. Cox (1966), *Erneuerungstheorie*, München, Wien, R. Oldenbourg Verlag

[7] D.R. Cox, D.V. Hinkley (1974), *Theoretical statistics*, London, Chapman and Hall

[8] C. Dean, J.F. Lawless (1989), „Tests for detecting overdispersion in Poisson regression models“, *Journal of the American Statistical Association*, **84**, 467-472

[9] C.D. Dean (1992), „Testing for overdispersion in Poisson and binomial regression models“, *Journal of the American Statistical Association*, **87**, 451-457

[10] C. Diepold (1996), *Statistische Analyse zur Rechtfertigung von Bonus-Malus-Systemen in der Krafthaftpflichtversicherung*, Diplomarbeit an der Katholischen Universität Eichstätt

[11] L. Fahrmeir, G. Tutz (1994), *Multivariate statistical modelling based on generalized linear models*, New York, Springer-Verlag

[12] L. Fahrmeir, R. Künstler, I. Pigeot, G. Tutz (1997), *Statistik: Der Weg zur Datenanalyse*, 2. verb. Auflage, Springer-Verlag

[13] M. Falk, R. Becker, F. Marohn (1995), *Angewandte Statistik mit SAS*, Berlin, Heidelberg, Springer-Verlag

[14] W. Feller (1968), *An introduction to probability theory and its applications (Vol. 1)*, 3rd edition, New York, John Wiley & Sons

[15] E.L. Lehmann (1999), *Elements of large-sample theory*, New York, Springer-Verlag

[16] J. Lemaire (1985), *Automobile insurance*, Boston, Kluwer Academic Publishers

[17] T. Mack (1997), *Schadenversicherungsmathematik*, Münchner Rück, Karlsruhe, Verlag Versicherungswirtschaft

[18] P. McCullagh, J.A. Nelder (1989), *Generalized linear models*, 2nd edition, Chapman and Hall

[19] P.A.P. Moran (1970), „Maximum-likelihood estimation in non-standard conditions“, *Proceedings of the Cambridge Philosophical Society*, **70**, 441-450

[20] R.H. Myers (1990), *Classical and modern regression with applications*, Duxbury Press

[21] C. Ortseifen (1997), *Der SAS-Kurs: Eine leicht verständliche Einführung*, 1. Auflage, Bonn, International Thompson Publishing GmbH

[22] C.R. Rao (1973), *Linear statistical inference and its applications*, 2nd edition, New York, Wiley

[23] S.I. Resnick (1992), *Adventures in stochastic processes*, Boston, Birkhäuser

[24] SAS Version 8(TS M0), SAS Online Doc, HTML-Format, Copyright 2000 SAS Institute Inc.

[25] P. Spector (1994), *An Introduction to S and S Plus*, Belmont, Calif., Duxbury Press

[26] P. Wang, M.L. Puterman, I. Cockburn, N. Le (1996), „Mixed Poisson regression models with covariate dependent rates“, *Biometrics*, **52**, 381-400

[27] R. Winkelmann (1996), *Econometric Analysis of count data*, 2nd reviewed and enlarged edition, Springer-Verlag

[28] H. Witting, U. Müller-Funk (1995), *Mathematische Statistik II*, Stuttgart, B.G. Teubner