

Technische Universität München

ZENTRUM MATHEMATIK

**Quantifizierung
von Überdispersion**

Diplomarbeit

von

Inka - Diana Sikora

Themensteller: Prof. Dr. C. Czado

Betreuer: Prof. Dr. C. Czado

Abgabetermin: 8. April 2002

Hiermit erkläre ich, daß ich die Diplomarbeit selbständig angefertigt und nur die angegebenen Quellen verwendet habe.

München, 8. April 2002

Inhaltsverzeichnis

1	Einleitung	1
2	Theoretische Grundlagen	3
2.1	Die Poissonverteilung	4
2.2	Einführung in die Theorie der GLM's	7
2.2.1	Anforderungen an ein Modell	7
2.2.2	Terminologie	9
2.2.3	Die exponentielle Familie	9
2.2.4	Komponenten eines GLM's	11
2.2.5	Offset in der Poissonregression	14
2.2.6	Maximum Likelihood (ML)- Schätzungen	15
2.2.7	Asymptotische Hypothesentests in GLM's	24
2.2.8	Devianzanalyse	26
2.2.9	Residualanalyse	29
2.3	Die Negativ-Binomial-Verteilung	30
2.3.1	Die NB-Verteilung und ihre Eigenschaften	31
2.3.2	Unbeobachtete Heterogenität	42
2.3.3	Die NB-Verteilung als GLM	45
3	Quantifizierung von Überdispersion	56
3.1	Einführung in die Theorie der Hypothesentests	56
3.1.1	Grundlagen der Testtheorie	56
3.1.2	Most-Powerful-Tests für einfache Hypothesen und Alternativen	62
3.1.3	Uniformly-most-powerful (UMP)-Tests	67
3.2	Bereichstest für Überdispersion	69
3.3	Theoretische Grundlagen der Simulationsstudie	72

3.3.1	Konzept der Simulationsdaten	73
3.3.2	Powerfunktion	75
3.3.3	P-Wert-Kurven bezüglich Dispersionsindex a_0	80
3.3.4	p-Wert-Kurven bezüglich Abstandsmass d_0	82
3.3.5	Interpretation und Bestimmung von p-Wert-Kurven	84
4	Simulationsstudie	86
4.1	Simulation der Powerfunktion	87
4.2	Simulation der p-Wert-Kurven	95
4.2.1	p-Wert-Kurven für a_0	96
4.2.2	p-Wert-Kurven für d_0	105
4.2.3	Zusammenfassung	115
5	Anwendungsbeispiele	117
5.1	Schwedische Kfz-Versicherung	117
5.1.1	Datenbeschreibung	118
5.1.2	Explorative Datenanalyse der Kfz-Daten	119
5.1.3	Poissonregression der Kfz-Daten	121
5.1.4	NB-Regression und p-Wert-Kurven der Kfz-Daten	127
5.2	Patent-Daten	133
5.2.1	Explorative Datenanalyse der Patent-Daten	134
5.2.2	Poissonregression der Patent-Daten	135
5.2.3	NB-Regression und p-Wert-Kurven der Patent-Daten	139
6	Zusammenfassung	144
	Anhang	146
	Literaturverzeichnis	152
	Abbildungsverzeichnis	154
	Tabellenverzeichnis	156

Kapitel 1

Einleitung

Um die Motivation für diese Arbeit zu verdeutlichen, betrachte man einige Fragestellungen aus der Sicht eines Versicherungsunternehmens. Speziell im Bereich der Kfz-Versicherung ist man z.B. daran interessiert zu wissen, wieviele Schäden im Durchschnitt für das nächste Jahr zu erwarten sind. Eine ebenfalls wichtige Fragestellung vor allem für die Tarifbildung ist, welche Gruppen von Fahrzeughaltern deutlich weniger Schäden verursachen als andere Gruppen. Um diese Fragen beantworten zu können, muss zunächst ein statistisches Modell erstellt werden, das so gewählt werden sollte, dass es die beobachteten Daten in ihrer wesentlichen Struktur wiedergibt. Bleibt man beim Beispiel der Kfz-Versicherung, so wird versucht, die Anzahl der Schäden in Abhängigkeit von Grössen wie das Alter des Fahrers, Baujahr des Fahrzeugs, Geschlecht des Fahrers, u.s.w. zu erklären. Da die zu untersuchende Grösse „*Schadenanzahl*“ eine Zählgrösse ist, wird in diesen Fällen üblicherweise das Poissonmodell als Grundlage verwendet, da dieses den Eigenschaften eines Zählmodells am besten entspricht. Im ersten Abschnitt des 2. Kapitels wird daher die Poissonverteilung kurz mit ihren charakteristischen Merkmalen zusammengefasst. Eine der wichtigsten Eigenschaften der Poissonverteilung ist die Gleichheit von Erwartungswert und Varianz. Modelliert man nun die Zufallsgrösse „*Schadenanzahl*“ mit Hilfe des Poissonmodells, so kann es vorkommen, dass das Modell die Daten nicht ausreichend anpasst. Dies kann z.B. erkannt werden, wenn die Devianz oder Pearson-Statistik grösser als die Anzahl der Freiheitsgrade ist. In solchen Fällen muss schliesslich mittels einer Residuenanalyse überprüft werden, ob die systematische Komponente des Modells richtig spezifiziert wurde und gegebenenfalls müssen noch weitere relevante Grössen, die die Schadenanzahl mit beeinflussen, in den linearen Prädiktor aufgenommen werden. Erst wenn nach diesen Überprüfungen ein Fehler in der systematischen Komponente des

Modells ausgeschlossen werden kann, ist davon auszugehen, dass die Ursache für eine schlechte Datenanpassung in der Modellannahme selber zu finden ist. Die bei der Verwendung des Poissonmodells am häufigsten auftretende Modellverletzung ist die sogenannte Überdispersion, die entsteht, wenn die Zählraten, also die Schadenanzahl im Beispiel der Kfz-Versicherung, in einem grösseren Masse um ihren Erwartungswert streuen, als es durch das Poissonmodell erwartet wird. Damit ist die charakteristische Eigenschaft der Gleichheit von Erwartungswert und Varianz bei der Poissonverteilung verletzt, wodurch gleichzeitig eine Verletzung der Modellannahme gegeben ist. Im Kapitel „*Theoretische Grundlagen*“ wird dieser Sachverhalt genauer beschrieben und es wird auf die Möglichkeit eingegangen, bei Vorliegen von Überdispersion diese mit Hilfe eines Wechsels vom Poisson- zum Negativ-Binomial-Modell zu berücksichtigen. Da in der Literatur schon viel über die Ursachen und Konsequenzen von Überdispersion und deren Beseitigung geschrieben wurde, ist es hier nicht das Ziel, das Auftreten von Überdispersion zu diskutieren, sondern es soll stattdessen versucht werden, die vorliegende Überdispersion zu quantifizieren. Die Frage, die dieser Arbeit zugrundeliegt, ist:

„Wie gross muss die vorliegende Überdispersion sein, um einen Wechsel vom Poissonmodell zum wesentlich aufwendigeren Negativ-Binomial-Modell zu rechtfertigen?“

Zur Beantwortung dieser Frage wird ein Bereichstest konstruiert, auf den im 3. Kapitel eingegangen wird. Im Prinzip handelt es sich um den Vergleich des Poisson- und Negativ-Binomial-Modells, wobei verwendet wird, dass das Negativ-Binomial-Modell mit dem Poissonmodell übereinstimmt, falls keine Überdispersion vorliegt. Das Ziel ist schliesslich, die Quantifizierung der Überdispersion anhand von sogenannten p-Wert-Kurven darzustellen, d.h. das klassische Konzept der p-Werte wird durch Kurven ersetzt, die es einem ermöglichen, die „*Evidenz*“ des Poisson- bzw. Negativ-Binomial-Modells simultan für unterschiedliche Akzeptanzgrenzen der Überdispersion graphisch darzustellen. Zunächst wird an einer Simulationsstudie der Bereichstest mit der Erstellung der p-Wert-Kurven durchgeführt, während anschliessend die Anwendung auf reale Datenbeispiele erfolgt.

Kapitel 2

Theoretische Grundlagen

In diesem Kapitel sollen die wichtigsten theoretischen Grundlagen zusammengefasst werden, die die Basis für diese Arbeit darstellen. In erster Linie handelt es sich um die sogenannten generalisierten linearen Modelle, kurz GLM's, die im 2. Abschnitt beschrieben werden. Die GLM's umfassen eine grosse Anzahl an Verteilungen, jedoch wird hier hauptsächlich der Spezialfall der Poissonverteilung betrachtet, da diese Grundlage für das Standardmodell für Zähldaten ist, die in dieser Arbeit verwendet werden.

Wie die linearen Modelle besitzen Zähldaten-Modelle eine Struktur aus zwei wesentlichen Komponenten: die systematische Komponente, die durch die Regression modelliert wird, und die zufällige Komponente, die der Grund für die Abweichung der beobachteten Daten vom Erwartungswert ist. Die Verteilungsannahme des zufälligen Effektes hat die Nichtnegativität der Zähldaten sowie deren Ganzzahligkeitscharakter zu berücksichtigen. Die bekanntesten Verteilungen, die diese Kriterien erfüllen, sind die Poisson- und die Negativ-Binomialverteilung. Aus diesem Grund gehören das Poisson- und das Negativ-Binomialregressions-Modell zu den verbreitetsten Zähldaten-Modellen. Liegen Zähldaten vor, so wird jedoch am häufigsten die Poissonverteilung zur Modellierung der Daten verwendet, weshalb sie auch als eine Art Richtlinie dient. Im ersten Abschnitt dieses Kapitels werden deshalb die wichtigsten Eigenschaften der Poissonverteilung wiederholt. Anders als die Normalverteilung besteht sie aus nur einem Parameter, der gleichzeitig Erwartungswert und Varianz ist. Diese Forderung der Gleichheit von Erwartungswert und Varianz ist oft zu stark. In der Praxis wird eher beobachtet, dass die Varianz grösser als der Erwartungswert ist, was mit Überdispersion bezeichnet wird. Eine Möglichkeit, diesem eben angesprochenen Problem auszuweichen, ist, zu dem Modell der Negativ-Binomialverteilung überzugehen, welches eine so starke Forderung bezüglich der Beziehung von Erwartungswert und Varianz nicht beinhaltet.

Die Eigenschaften der Negativ-Binomialverteilung sowie deren Zusammenhänge zur Poissonverteilung werden im letzten Abschnitt dieses Kapitels behandelt.

2.1 Die Poissonverteilung

Die Poissonverteilung gehört zur Klasse der diskreten Verteilungen, z.B. kann man die Anzahl der eingehenden Telefonanrufe in einer Vermittlungsstelle pro Stunde mit dieser Verteilung modellieren. Die Wahrscheinlichkeitsfunktion, die momenterzeugende Funktion und einige nützliche Eigenschaften der Poissonverteilung werden in diesem Abschnitt kurz zusammengefasst.

Definition 2.1.1 (Poissonverteilung) Sei Y eine Zufallsvariable mit einer diskreten Verteilung, die definiert ist auf $\mathbb{N} \cup \{0\} = \{0, 1, 2, \dots\}$.

Y ist dann poissonverteilt mit Parameter λ , kurz $Y \sim \text{Poi}(\lambda)$, falls die Wahrscheinlichkeitsfunktion folgende Form hat:

$$P(Y = y) = \frac{e^{-\lambda} \lambda^y}{y!} \quad \text{für } \lambda \in \mathbb{R} \quad \text{und } y = 0, 1, 2, \dots$$

Die Poissonverteilung zeichnet sich durch eine besondere Beziehung zwischen Erwartungswert und Varianz von anderen Verteilungen ab, es gilt hier nämlich

$$E(Y) = \text{Var}(Y) = \lambda.$$

Diese Eigenschaft der Gleichheit von Erwartungswert und Varianz bezeichnet man mit „Äquidispersion“. Falls die Varianz grösser ist als der Erwartungswert, spricht man von „Überdispersion“, die in der späteren Diskussion noch eine wichtige Rolle spielen wird. Eine andere Möglichkeit der Abweichung von der Äquidispersion ist die sogenannte „Unterdispersion“, die dadurch gekennzeichnet ist, dass in diesem Fall der Erwartungswert die Varianz übersteigt.

Satz 2.1.2 (Momenterzeugende Funktion der Poissonverteilung) Die momenterzeugende Funktion der Poissonverteilung ist gegeben durch

$$M_Y(t) = E[e^{tY}] = e^{\lambda(e^t - 1)}.$$

Beweis:

$$M_Y(t) = E[e^{tY}] = \sum_{y=0}^{\infty} e^{ty} e^{-\lambda} \frac{\lambda^y}{y!} = e^{-\lambda} \sum_{y=0}^{\infty} \frac{e^{ty} \lambda^y}{y!} = e^{-\lambda} \sum_{y=0}^{\infty} \frac{(e^t \lambda)^y}{y!} = e^{-\lambda} e^{e^t \lambda} = e^{\lambda(e^t - 1)}.$$

□

Die ersten k Momente erhält man daraus durch k -maliges differenzieren, d.h.

$$E[Y^k] = M^{(k)}(0) = \frac{\partial^k M(0)}{\partial t^k} \quad \text{für } k = 0, 1, 2, \dots \quad (2.1)$$

Konkret ergeben sich für die Poissonverteilung folgende erste 3 Momente:

$$\begin{aligned} E[Y] &= M'(0) = e^{\lambda(e^t-1)} \lambda e^t|_{t=0} = \lambda \\ E[Y^2] &= M''(0) = \lambda e^t e^{\lambda(e^t-1)} + \lambda e^t e^{\lambda(e^t-1)} \lambda e^t|_{t=0} = e^{\lambda(e^t-1)} (\lambda e^t + \lambda^2 e^{2t})|_{t=0} = \lambda + \lambda^2 \\ E[Y^3] &= M'''(0) = e^{\lambda(e^t-1)} \lambda e^t (\lambda e^t + \lambda^2 e^{2t}) + e^{\lambda(e^t-1)} (\lambda e^t + \lambda^2 e^{2t} 2)|_{t=0} \\ &= e^{\lambda(e^t-1)} (\lambda^2 e^{2t} + \lambda^3 e^{3t} + \lambda e^t + 2\lambda^2 e^{2t})|_{t=0} = \lambda + 3\lambda^2 + \lambda^3. \end{aligned}$$

Aus dem ersten Moment lässt sich der Erwartungswert direkt ablesen, während für die Varianz nach Springers Formelsammlung (1997, S.420) [17] allgemein die Formel

$$\text{Var}(Y) = M''(0) - [M'(0)]^2 \quad (2.2)$$

gilt. Im Fall der Poissonverteilung kann also die Varianz wie folgt überprüft werden:

$$\text{Var}(Y) = M''(0) - [M'(0)]^2 = \lambda + \lambda^2 - [\lambda]^2 = \lambda.$$

Als weitere angenehme Eigenschaft der Poissonverteilung kann die sogenannte Additivität angegeben werden, die im folgenden Satz gezeigt wird.

Satz 2.1.3 Seien $Y_i \sim \text{Poi}(\lambda_i)$ $i = 1, 2, \dots$ unabhängige Zufallsvariablen und es gelte $\sum \lambda_i < \infty$, dann gilt:

$$Z = \sum Y_i \sim \text{Poi}(\sum \lambda_i).$$

Beweis: Da die Y_i unabhängig sind, gilt für die momenterzeugenden Funktionen die Gleichung

$$M_Z(t) = M_{\sum Y_i}(t) = \prod_i M_{Y_i}(t).$$

Die einzelnen Y_i sind alle poissonverteilt, also gilt

$$M_Z(t) = \prod_i M_{Y_i}(t) = \prod_i e^{\lambda_i(e^t-1)} = e^{\sum \lambda_i(e^t-1)},$$

was aber gerade die momenterzeugende Funktion einer poissonverteilten Zufallsvariable mit Parameter $\sum \lambda_i$ ist.

□

Für die Charakterisierung der Poissonverteilung gibt es viele Möglichkeiten. Eine ist z.B. der Poisson-Zähl-Prozess, auf den hier aber nicht näher eingegangen werden soll. Eine kurze Einführung in die Theorie der stochastischen Prozesse kann aber z. B. in Resnick (1992) [18] nachgelesen werden, wo in Kapitel 4 speziell auf den Poissonprozess eingegangen wird.

Eine andere Möglichkeit ist die Charakterisierung der Poissonverteilung als Grenzfall der Binomialverteilung, d.h. die gesamte Anzahl der Ereignisse folgt approximativ der Poissonverteilung, falls ein Ereignis in einer grossen Zahl von Versuchen auftritt, die Wahrscheinlichkeit jedoch, dass dieses Ereignis eintritt, sehr klein ist.

Satz 2.1.4 (Poisson-und Binomialverteilung) *Dieser Satz ist auch unter dem Titel „Law of rare events“ bekannt. Sei Y die Gesamtanzahl der Erfolge in einer grossen Folge von unabhängigen Bernoulli-Experimenten mit Erfolgswahrscheinlichkeit π , die in jedem Versuch sehr klein ist, d.h. man erhält eine Binomialverteilung mit den Parametern n und π , kurz $B(n, \pi)$, deren Wahrscheinlichkeitsfunktion gegeben ist durch*

$$P(Y = k) = \binom{n}{k} \pi^k (1 - \pi)^{n-k}, \quad k = 0, 1, 2, \dots, n.$$

Für $\lambda > 0$ geht im Grenzfall für $n \rightarrow \infty$ und $\pi := \frac{\lambda}{n} \rightarrow 0$ die Binomialverteilung in die Poissonverteilung mit Parameter λ über.

Beweis: Der Beweis hierzu findet sich in Kredler (1998, S. 102) [12].

Man betrachte zunächst für $\lambda > 0$ die $B(n, \frac{\lambda}{n})$ -verteilte Zufallsvariable Y_n .

$$P(Y_n = k) = \frac{n!}{k!(n-k)!} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k}.$$

Durch Umsortieren dieser Gleichung erhält man

$$\begin{aligned} P(Y_n = k) &= \frac{\lambda^k}{k!} \left(1 - \frac{\lambda}{n}\right)^n \frac{n(n-1) \dots (n-k+1)}{\left(1 - \frac{\lambda}{n}\right)^k n^k} \\ &= \frac{\lambda^k}{k!} \left(1 - \frac{\lambda}{n}\right)^n \frac{1(1 - \frac{1}{n}) \dots (1 - \frac{k-1}{n})}{\left(1 - \frac{\lambda}{n}\right)^k}. \end{aligned}$$

Für $n \rightarrow \infty$ gilt

$$\lim_{n \rightarrow \infty} \left(1 - \frac{\lambda}{n}\right)^n \rightarrow e^{-\lambda} \quad \text{und} \quad \lim_{n \rightarrow \infty} \frac{1(1 - \frac{1}{n}) \dots (1 - \frac{k-1}{n})}{\left(1 - \frac{\lambda}{n}\right)^k} \rightarrow 1.$$

Somit ergibt sich insgesamt

$$\lim_{n \rightarrow \infty} P(Y_n = k) = \frac{\lambda^k}{k!} e^{-\lambda},$$

was die Wahrscheinlichkeitsfunktion einer poissonverteilten Zufallsvariable mit Parameter λ ist.

□

2.2 Einführung in die Theorie der GLM's

In den letzten Jahren hat die Theorie der GLM's immer mehr an Bedeutung gewonnen und so gibt es eine Vielzahl an Literatur, die sich mit diesem Thema beschäftigt. Eine übersichtliche Einführung für alle GLM's ist z.B. in McCullagh & Nelder (1989) [15] gegeben und weiterführende Details speziell für Zähldaten können z.B. in Cameron & Trivedi (1998) [2] nachgelesen werden.

Generalisierte lineare Modelle haben ihren Namen daher erhalten, da sie die klassischen linearen Modelle, die auf der Normalverteilung basieren, erweitern. Die Erweiterung besteht aus zwei Aspekten:

Zum einen können den GLM's eine Vielzahl von unterschiedlichen Verteilungen zugrundeliegen, sofern diese zu der exponentiellen Familie gehören. Zum anderen erlauben diese Modelle als Erweiterung zum linearen Modell Transformationen des Erwartungswertes, was mit der sogenannten Linkfunktion beschrieben wird.

Zunächst wird im folgenden erläutert, welche Anforderungen allgemein an ein Modell gestellt werden und welche Terminologie in dieser Arbeit verwendet wird. In den anschließenden Unterabschnitten werden die einzelnen Bestandteile eines GLM's kurz zusammengefasst. Auf die Methode der Maximum-Likelihood-Schätzung und die Residualanalyse wird in den letzten Unterabschnitten eingegangen.

2.2.1 Anforderungen an ein Modell

Gegeben sei eine Anzahl an Messungen oder Zähldaten. Zusätzlich seien strukturelle Informationen, wie die Reihenfolge der Messungen oder die Art der Messinstrumente, die benutzt wurden oder Informationen über die Bedingungen, unter welchen die Daten erhoben wurden, gegeben.

Um nun die Daten interpretieren zu können, sucht man nach bestimmten Strukturen innerhalb der Daten, z.B. dass ein Instrument regelmässig höhere Werte liefert als andere. Dafür muss zunächst ein Modell erstellt werden, das den Daten zugrundeliegt. Das einfachste ist das lineare Modell $\mathbf{y} = \alpha + \beta\mathbf{x}$ mit \mathbf{y} und $\mathbf{x} \in \mathbb{R}^n$. Das bedeutet,

dass mit dem Parameterpaar (α, β) eine einfache Beziehung zwischen \mathbf{x} und \mathbf{y} hergestellt werden kann. Gibt es mehrere Einflussgrössen $\mathbf{x}_1, \dots, \mathbf{x}_p$ auf die zu untersuchende Grösse \mathbf{y} , so schreibt man $\mathbf{y} = \beta_0 + \beta_1 \mathbf{x}_1 + \dots + \beta_p \mathbf{x}_p$, wobei das α durch β_0 ersetzt wird. In der Praxis gilt die lineare Beziehung zwischen $\mathbf{x}_1, \dots, \mathbf{x}_p$ und \mathbf{y} allerdings nur approximativ. Daher wird in das Modell noch ein sogenannter Fehlerterm ε eingefügt, so dass sich schliesslich das folgende stochastische Modell ergibt:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \varepsilon_i, \quad i = 1, \dots, n. \quad (2.3)$$

Von dem Fehlerterm ε_i nimmt man in linearen Modellen an, dass er normalverteilt ist mit Erwartungswert Null und konstanter Varianz σ^2 , kurz $\varepsilon_i \sim N(0, \sigma^2)$. Eine weitere Annahme ist, dass die ε_i identisch und unabhängig verteilt sind, d.h. sie sind iid (identically independent distributed). Das Ziel ist nun, die zu untersuchende Grösse \mathbf{y} durch die Einflussgrössen $\mathbf{x}_1, \dots, \mathbf{x}_p$ zu erklären. Dazu müssen die Parameter $(\beta_0, \beta_1, \dots, \beta_p)$ so geschätzt werden, dass sie die Beziehung zwischen \mathbf{y} und $\mathbf{x}_1, \dots, \mathbf{x}_p$ bestmöglichst wiedergeben. Die geschätzten Parameter werden im folgenden mit $(\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p)$ beschrieben. Die sich damit ergebenden Werte $\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \dots + \hat{\beta}_p x_{ip}$, $i = 1, \dots, n$, die man auch mit \hat{y}_i oder $\hat{\mu}_i$ bezeichnet, heissen dann „angepasste“ bzw. „gefittete Werte“, die mit Hilfe der Daten und des gewählten Modells erzeugt werden.

Ziel der Modellerstellung ist es also, die gegebenen Daten \mathbf{y} durch einen Datensatz von gefitteten $\hat{\mathbf{y}}$ zu ersetzen. Dabei spielt die Anzahl der Einflussgrössen, die in das Modell mit aufgenommen werden, eine grosse Rolle. Es gibt im wesentlichen drei Fälle zu unterscheiden: im ersten Fall spricht man vom sogenannten „Nullmodell“. Hier gehen keine Einflussgrössen in das Modell ein, d.h. es wird nur ein Parameter β_0 verwendet, der den gemeinsamen Erwartungswert μ repräsentiert. Dieses Modell hat den Nachteil, dass es zu einfach ist und dass keine Beziehung zwischen \mathbf{x} und \mathbf{y} modelliert wird. Das andere Extrem ist das „volle Modell“. In diesem Fall werden so viele Parameter verwendet wie Beobachtungen vorhanden sind. Damit sind die beobachteten Daten \mathbf{y} exakt reproduzierbar. Dies ist zwar wünschenswert, doch der Nachteil hierbei ist, dass dieses Modell keine Informationen über den Erklärungsgehalt von \mathbf{y} durch die Einflussgrössen $\mathbf{x}_1, \dots, \mathbf{x}_p$ enthält. Das volle Modell ist demnach nicht aussagekräftig genug, denn es fasst die Daten nicht in ihrer wesentlichen Struktur zusammen, sondern gibt sie exakt wieder. Im letzten Fall betrachtet man Modelle, deren Anzahl von Einflussgrössen grösser als Null (Nullmodell), aber kleiner als die Anzahl der Beobachtungen (volles Modell) ist.

Zu der bestmöglichen Anpassung der gefitteten an die beobachteten Werte kommt also

als weitere Anforderung hinzu, die Komplexität der Struktur des Modells so gering wie möglich zu halten, was erreicht werden kann, wenn keine überflüssigen Parameter verwendet werden. Diese grundlegenden Ideen der Modellbildung sind hier sehr allgemein gefasst und gelten sowohl für die linearen als auch für die generalisierten linearen Modelle.

2.2.2 Terminologie

Wie im vorherigen Abschnitt angesprochen, wird allgemein bei einer Regression versucht, eine zu untersuchende Grösse \mathbf{y} durch Einflussgrössen $\mathbf{x}_1, \dots, \mathbf{x}_p$ zu erklären. Dabei wird üblicherweise in der Literatur die Grösse \mathbf{y} mit „*Response*“ oder „*abhängige Variable*“ bezeichnet, während man die Einflussvektoren $\mathbf{x}_1, \dots, \mathbf{x}_p$ „*Kovariablen*“ oder auch „*erklärende Variablen*“ nennt. Die Datensituation lässt sich auch kurz durch

$$(y_i \mid \mathbf{x}_i), \quad i = 1, \dots, n$$

beschreiben, d.h. gegeben den Kovariablenvektor \mathbf{x}_i der i -ten Beobachtung beobachtet man den Wert y_i , wobei $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip})^t$ ist, was auch mit „*Design-Vektor*“ oder „*Kovariablenvektor*“ bezeichnet wird. Die Kovariablen können quantitativ oder qualitativ sein. Quantitative Kovariablen nehmen numerische Werte an, qualitative bestehen aus nichtnumerischen Werten. Die abhängige Variable kann stetig oder diskret sein (im Fall der Zähldatenregression nimmt die abhängige Variable nur diskrete Werte an). In Matrixschreibweise werden die Beobachtungen durch einen Spaltenvektor $\mathbf{y} = (y_1, \dots, y_n)^t$ beschrieben und die p Kovariablen werden in einer $(n \times p)$ -Matrix \mathbf{X} zusammengefasst, die man „*Design*“ - oder „*Datenmatrix*“ nennt. Jede Zeile davon beschreibt eine Beobachtung, jede Spalte gehört zu einer Kovariablen. In der Regel fügt man zu der Matrix \mathbf{X} noch eine Einserspalte hinzu, die den Intercept bestimmt. Damit erhöht sich die Dimension der Matrix von $(n \times p)$ auf $(n \times (p + 1))$.

Zu jeder Kovariable gehört ein Parameter, der in der Regel unbekannt ist. Die Parameter werden in einem Vektor der Länge p zusammengefasst und kurz mit $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^t$ bezeichnet. Nimmt man den Intercept hinzu, ergibt sich der Parametervektor $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^t$ der Länge $p + 1$.

2.2.3 Die exponentielle Familie

Wie schon anfangs erwähnt, basieren GLM's auf Wahrscheinlichkeitsfunktionen bzw. Dichten, die zur exponentiellen Familie gehören.

Definition 2.2.1 (Lineare exponentielle Familie) Eine Dichte $f(y; \theta, \phi)$ gehört zu der linearen exponentiellen Familie mit natürlichem oder kanonischem Parameter θ und Stör- oder Skalenparameter ϕ , falls sie sich in folgender Form schreiben lässt:

$$f(y; \theta, \phi) = \exp \left\{ \frac{\theta y - b(\theta)}{a(\phi)} + c(y, \phi) \right\}.$$

Dabei beschreibt $c(y, \phi)$ eine normalisierende Konstante und die Funktionen $a(\cdot)$ und $b(\cdot)$ bestimmen die jeweilige Verteilung.

Für den Erwartungswert und die Varianz gelten folgende Formeln:

$$E(Y | x) = \mu = b'(\theta) \quad (2.4)$$

$$Var(Y | x) = b''(\theta)a(\phi). \quad (2.5)$$

Die Varianz von Y ergibt sich also aus einem Produkt zweier Funktionen. Die erste ist $b''(\theta)$, eine Funktion, die nur vom kanonischen Parameter abhängt und mit „Varianzfunktion“ bezeichnet wird. Die Varianzfunktion wird häufig als Funktion von μ betrachtet, weshalb sie in der Regel mit $V(\mu)$ abgekürzt wird. Die zweite Funktion ist $a(\phi)$, die unabhängig von θ , dafür aber abhängig vom Dispersionsparameter ϕ ist. Einen Beweis zu den obigen Formeln kann man in McCullagh & Nelder (1989, S. 28) [15] finden. Die Formeln lassen sich aus den bekannten Gleichungen (siehe Casella, Berger (1990, S. 309ff)

$$E \left(\frac{\partial \mathcal{L}}{\partial \theta} \right) = 0 \quad (2.6)$$

und

$$E \left(\frac{\partial^2 \mathcal{L}}{\partial \theta^2} \right) + E \left(\frac{\partial \mathcal{L}}{\partial \theta} \right)^2 = 0 \quad (2.7)$$

ableiten, wobei hier $\mathcal{L}(\theta, \phi; y) = \ln f(y; \theta, \phi)$ die Log-Likelihood-Funktion darstellt. Ausgehend von der Dichte der exponentiellen Familie erhält man die Log-Likelihood-Funktion

$$\mathcal{L}(\theta, \phi; y) = \frac{\theta y - b(\theta)}{a(\phi)} + c(y, \phi).$$

Als erste und zweite Ableitungen ergeben sich

$$\frac{\partial \mathcal{L}}{\partial \theta} = \frac{y - b'(\theta)}{a(\phi)} \quad \text{und} \quad \frac{\partial^2 \mathcal{L}}{\partial \theta^2} = -\frac{b''(\theta)}{a(\phi)}.$$

Für den Erwartungswert hat man nun die Gleichung (2.6) zu lösen, d.h.

$$E\left(\frac{\partial \mathcal{L}}{\partial \theta}\right) = \frac{\mu - b'(\theta)}{a(\phi)} = 0,$$

woraus sich

$$E(Y) = \mu = b'(\theta)$$

ergibt. Für die Varianz löst man die Gleichung (2.7), d.h.

$$E\left(\frac{\partial^2 \mathcal{L}}{\partial \theta^2}\right) + E\left(\frac{\partial \mathcal{L}}{\partial \theta}\right)^2 = -\frac{b''(\theta)}{a(\phi)} + \frac{\text{Var}(Y)}{a(\phi)^2} = 0,$$

und erhält daraus

$$\text{Var}(Y) = b''(\theta)a(\phi).$$

Beispiel 2.2.2 (Poissonverteilung) *Die Poissonverteilung gehört mit vielen anderen Verteilungen zu der exponentiellen Familie, denn ihre Dichte bzw. Wahrscheinlichkeitsfunktion lässt sich wie folgt umschreiben:*

$$f_{\text{poi}}(y; \theta, \phi) = \exp\{-\lambda + y \ln \lambda - \ln y!\}.$$

Dabei ist $a(\phi) = 1$, $\theta = \ln \lambda$, $b(\theta) = e^\theta = e^{\ln \lambda} = \lambda$ und $c(y, \phi) = -\ln y!$.

Für den Erwartungswert ergibt sich nach obiger Formel

$$E(Y) = \mu = b'(\theta) = e^\theta = \lambda. \quad (2.8)$$

Die Varianz überprüft man analog nach den oben angegebenen Formeln mit

$$\text{Var}(Y) = b''(\theta)a(\phi) = e^\theta = \lambda.$$

Es ergibt sich also wie erwartet die schon angesprochene Eigenschaft der Äquidispersion, d.h. Gleichheit von Erwartungswert und Varianz bei der Poissonverteilung.

2.2.4 Komponenten eines GLM's

Sei \mathbf{y} , ein Vektor von Beobachtungen aus n Komponenten, die Realisierung einer Zufallsvariablen Y . Um nun ein GLM definieren zu können, benötigt man folgende Annahmen:

(i) Verteilungsannahme:

Gegeben den Kovariablenvektor \mathbf{x}_i sind die y_i , $i = 1, \dots, n$ bedingt unabhängig

verteilt nach einer Verteilung, die sich als Exponentialfamilie darstellen lässt mit Erwartungswert

$$E(Y_i | \mathbf{x}_i) =: \mu_i, \quad i = 1, \dots, n.$$

Die Verteilungsannahme bildet die zufällige Komponente des Modells. Die flexible Wahl der Verteilung ist ein Aspekt der Erweiterung der GLM's gegenüber den klassischen linearen Modellen, in welchen ausschliesslich von der Annahme der Normalverteilung ausgegangen wird. Der andere Aspekt ist die Möglichkeit der Transformation des Erwartungswertes, die im nächsten Punkt näher erläutert wird.

(ii) **Strukturelle Annahme:**

Im folgenden sei i der Index der n verschiedenen Beobachtungen und j sei der Index der p relevanten Kovariablen. Die Kovariablenvektoren $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_p$ definieren dann den sogenannten „linearen Prädiktor“ $\boldsymbol{\eta}$ mit $\boldsymbol{\eta} = (\eta_1, \dots, \eta_n)^t$, der gegeben ist durch

$$\boldsymbol{\eta} = \sum_{j=1}^p \mathbf{x}_j \beta_j = \mathbf{X} \boldsymbol{\beta} \quad \text{mit} \quad \mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_p) \in \mathbb{R}^{n \times p} \quad \text{und} \quad \boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^t$$

bzw. in Komponentenschreibweise $\eta_i = \mathbf{x}_i^t \boldsymbol{\beta}$ mit $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip})^t$, $i = 1, \dots, n$. Anschliessend wird der Erwartungswert μ_i über eine „Responsefunktion“ h mit dem linearen Prädiktor η_i verbunden, d.h.

$$\mu_i = h(\eta_i) = h(\mathbf{x}_i^t \boldsymbol{\beta}).$$

Oft verwendet man auch die inverse Responsefunktion $g := h^{-1}(\mu_i)$, die mit „Linkfunktion“ bezeichnet wird:

$$g(\mu_i) = \eta_i = \mathbf{x}_i^t \boldsymbol{\beta}.$$

Die Linkfunktion verbindet also die zufällige Komponente mit der systematischen, welche eine Spezifikation für den Vektor $\boldsymbol{\mu}$ in Termen aus einer kleinen Anzahl von unbekannten Parametern β_1, \dots, β_p ist, deren Werte durch die Daten geschätzt werden müssen.

In den klassischen linearen Modellen sind der Erwartungswert und der lineare Prädiktor identisch, d.h. es gilt $\boldsymbol{\eta} = \boldsymbol{\mu}$. GLM's erlauben aber eine Erweiterung, so dass die Linkfunktion eine beliebige, monotone, differenzierbare Funktion sein kann. Wenn man z.B. mit Zähldaten arbeitet und die zugrundeliegende Verteilung

der Daten eine Poissonverteilung ist, muss man $\boldsymbol{\mu} > \mathbf{0}$ voraussetzen. Ein additives Modell für $\boldsymbol{\mu}$, d.h. $\boldsymbol{\mu} = \boldsymbol{\eta} = \sum_{j=1}^p \mathbf{x}_j \beta_j$, ist hier nicht befriedigend, denn $\boldsymbol{\eta}$ kann für bestimmte Parameterwerte jederzeit negativ werden, $\boldsymbol{\mu}$ soll aber positiv sein, womit die Identität als Linkfunktion keinen Sinn mehr macht.

Modelle für Zähldaten führen daher oft zu multiplikativen Effekten, was durch die sogenannte log-link-Funktion ausgedrückt wird. Dies bedeutet formal

$$\boldsymbol{\eta} = \ln(\boldsymbol{\mu}) \quad \text{bzw.} \quad \boldsymbol{\mu} = e^{\boldsymbol{\eta}}. \quad (2.9)$$

Oft wird in der Literatur statt von Poissonregression auch von log-linearen Modellen gesprochen, da der Logarithmus des bedingten Erwartungswertes linear in den Parametern ist. Durch diese Wahl von Linkfunktionen werden nun additive Effekte bzgl. $\boldsymbol{\eta}$ zu multiplikativen Effekten bzgl. $\boldsymbol{\mu}$ und $\boldsymbol{\mu}$ ist strikt positiv. Jede Verteilung hat eine spezielle Linkfunktion für die eine suffiziente Statistik existiert. Dabei heisst eine Statistik $T(X)$ „suffizient“ für einen Parameter θ , genau dann wenn die bedingte Verteilung von X gegeben $T(X) = t$ nicht von θ abhängt. Solche Linkfunktionen heissen „kanonische Linkfunktionen“ und sind dadurch definiert, dass $\boldsymbol{\theta} = \boldsymbol{\eta}$ gilt. Mit Hilfe des Faktorisierungstheorems (siehe Bickel & Doksum (1977, S. 65) [1]) kann nachgewiesen werden, dass in einem GLM mit kanonischer Linkfunktion $(\sum_{i=1}^n x_{i1} y_i, \dots, \sum_{i=1}^n x_{ip} y_i)^t$ eine suffiziente Statistik für $(\beta_1, \dots, \beta_p)$ ist.

(iii) **Kovariablenvektor:**

Um ein GLM eindeutig definieren zu können, fehlt zum Abschluss noch die Angabe des Kovariablenvektors \mathbf{x}_i .

Beispiel 2.2.3 (Poissonregression) *Da die Poissonregression ein Modell für Zähldaten mit Kovariablen ist, sollen hier noch einmal die oben genannten Annahmen für diesen Spezialfall zusammengefasst werden. Die Verteilung der Y_i gegeben die Kovariablenvektoren \mathbf{x}_i ist eine Poissonverteilung, d.h. $(Y_i | \mathbf{x}_i) \sim \text{Poi}(\mu_i)$. Ausserdem seien die Beobachtungen alle unabhängig. Die kanonische Linkfunktion ist in diesem Fall die log-link-Funktion, d.h. $\boldsymbol{\eta} = \ln(\boldsymbol{\mu})$, denn nach (2.8) gilt*

$$\boldsymbol{\eta} = \ln(\boldsymbol{\mu}) = \ln(\exp(\boldsymbol{\theta})) = \boldsymbol{\theta}.$$

Damit ist der bedingte Erwartungswert der Beobachtungen Y_i gegeben die Kovariablen \mathbf{x}_i gegeben durch $E(Y_i | \mathbf{x}_i) = \exp(\eta_i) = \exp(\mathbf{x}_i^t \boldsymbol{\beta})$, $i = 1, \dots, n$.

Fasst man die beiden Annahmen zusammen, erhält man die folgende Wahrscheinlichkeitsfunktion

$$P(Y_i = y \mid \mathbf{x}_i) = \frac{\exp(-\exp(\mathbf{x}_i^t \boldsymbol{\beta})) \exp(\mathbf{x}_i^t \boldsymbol{\beta} y)}{y!}, \quad y = 0, 1, 2, \dots$$

2.2.5 Offset in der Poissonregression

Bisher wurde angenommen, dass bei der Poissonregression die Beobachtungszeiträume für den Parameter μ_i , der die erwartete Anzahl von Ereignissen im Beobachtungszeitraum beschreibt, für alle $i = 1, \dots, n$ einheitlich sind. Nun kann es in der Praxis häufig vorkommen, dass Daten analysiert werden müssen, deren Beobachtungszeiträume variieren. Aus diesem Grund muss dann eine zusätzliche Variable t_i eingeführt werden, die sich auf den jeweiligen Zeitraum der i -ten Beobachtung bezieht, um die Zeiteinheit zu standardisieren. Die zu untersuchende Variable Y_i ist damit poissonverteilt mit dem Parameter $t_i \mu_i^*$, d.h.

$$P(Y_i = y_i \mid \mathbf{x}_i) = \frac{e^{-t_i \mu_i^*} (t_i \mu_i^*)^{y_i}}{y_i!}, \quad i = 1, \dots, n, \quad y = 0, 1, 2, \dots,$$

wobei $\mu_i^* = \exp(\mathbf{x}_i^t \boldsymbol{\beta})$ ist. Schreibt man diese Dichte in die Form einer exponentiellen Familie um, so ergibt sich

$$f_{poi}(y; \theta, \phi) = \exp\{-t_i \mu_i^* + y_i \ln(t_i \mu_i^*) - \ln(y_i!)\}.$$

Dabei muss wie in Beispiel 2.2.2 $a(\phi) = \phi = 1$ und $b(\theta_i) = e^{\theta_i}$ gewählt werden, aber der Parameter θ_i ist nun gegeben durch $\theta_i = \ln(t_i \mu_i^*)$. Damit ergibt sich für den Erwartungswert

$$E(Y_i) = \mu_i = b'(\theta_i) = e^{\theta_i} = t_i \mu_i^* = t_i \exp(\mathbf{x}_i^t \boldsymbol{\beta}) = \exp(\ln(t_i) + \mathbf{x}_i^t \boldsymbol{\beta}).$$

Für den linearen Prädiktor gilt schliesslich

$$\eta_i = \mathbf{x}_i^t \boldsymbol{\beta} = \ln(\mu_i) - \ln(t_i) = \theta_i - \underbrace{\ln(t_i)}_{\text{offset}},$$

wobei der letzte Term $\ln(t_i)$ mit „offset“ bezeichnet und als bekannt vorausgesetzt wird. Ist $t_i = 1 \forall i$, so erhält man wie in Beispiel 2.2.3 gezeigt, die kanonische Linkfunktion $g(\mu_i) = \ln(\mu_i)$, während man für den Fall, dass $t_i \neq 1$ die kanonische Linkfunktion

$$g(\mu_i) = \ln(\mu_i) - \ln(t_i)$$

erhält. Im folgenden wird allerdings von einem einheitlichen Beobachtungszeitraum ausgegangen, d.h. es wird $t_i = 1 \forall i$ gesetzt.

2.2.6 Maximum Likelihood (ML)- Schätzungen

Nach der Wahl eines geeigneten Modells ist man an der Schätzung der Parameter interessiert und möchte auch, wenn möglich, deren Genauigkeit beurteilen. Erst dann können mit Hilfe des Modells und den Schätzungen Prognosen für die Zukunft erstellt werden, die in der Regel das eigentliche Ziel sind. In diesem Abschnitt soll es nun um die Schätzung der Parameter gehen.

In linearen Modellen schätzt man die Parameter mit der Methode der kleinsten Quadrate, d.h. man minimiert die Abstandsquadrate zwischen den beobachteten und den gefitteten Daten. Diese Methode entspricht unter Voraussetzung der Normalverteilung genau der Maximum-Likelihood-Schätzung, bei der man versucht, die Parameter $\beta_0, \beta_1, \dots, \beta_p$ so zu ermitteln, dass die Wahrscheinlichkeit, die gegebenen Daten zu erhalten, maximiert wird. In GLM's dagegen sind diese beiden Schätzmethoden nicht äquivalent. Da im Gegensatz zum kleinsten Quadrate-Ansatz bei der Maximum-Likelihood-Methode alle Informationen der Daten enthalten sind und diese Methode deshalb eine modellbasierende Schätzung ist, wird in der Regel in GLM's nur Maximum-Likelihood-Schätzung betrieben, weshalb hier nur auf diese eingegangen wird.

Um ML-Schätzer für die unbekannten Parameter zu erhalten, wird vorausgesetzt, dass die zugrundeliegende Verteilung vollständig bekannt ist und dass die Design-Matrix \mathbf{X} der Kovariablen einen vollen Spaltenrang p besitzt.

Sei nun allgemein $f(\mathbf{y} \mid \boldsymbol{\theta})$ die Wahrscheinlichkeits- bzw. Dichtefunktion eines Zufallsvariablenvektors $\mathbf{Y} = (Y_1, \dots, Y_n)^t$, wobei der Parametervektor $\boldsymbol{\theta} = (\theta_1, \dots, \theta_m)^t$ in dem zulässigen Parameterraum Θ liegt. Betrachtet man dann $f(\mathbf{y} \mid \boldsymbol{\theta})$ als eine Funktion von $\boldsymbol{\theta}$ für feste Realisierungen \mathbf{y} , so nennt man diese Funktion die „*Likelihood-Funktion*“ und schreibt $L(\boldsymbol{\theta})$. Für den Spezialfall, dass die Zufallsvariablen Y_1, \dots, Y_n gegeben die Kovariablenvektoren $\mathbf{x}_i, i = 1, \dots, n$ unabhängig diskret oder stetig verteilt sind mit Wahrscheinlichkeitsfunktion bzw. Dichte $f(y_i \mid \mathbf{x}_i, \boldsymbol{\theta})$, so ergibt sich wegen der Unabhängigkeit der Y_i die Likelihood-Funktion

$$L(\boldsymbol{\theta}) = \prod_{i=1}^n f(y_i \mid \mathbf{x}_i, \boldsymbol{\theta}).$$

Da in dieser Arbeit nur unabhängige Zufallsvariablen betrachtet werden, kann für die weiteren Ausführungen von diesem Spezialfall ausgegangen werden. Nach der Maximum-Likelihood-Methode (siehe Bickel & Doksum (1977, S. 99) [1] und Kredler (1998, S. 173) [12]) wählt man nun die Schätzungen $\hat{\boldsymbol{\theta}}$ für den unbekannten Parameter $\boldsymbol{\theta}$ so, dass

das Eintreten der beobachteten Stichprobe maximale Wahrscheinlichkeit besitzt, d.h.

$$L(\hat{\boldsymbol{\theta}}_{ML}) \geq L(\tilde{\boldsymbol{\theta}}) \quad \forall \tilde{\boldsymbol{\theta}} \in \Theta,$$

wobei Θ den Parameterraum bezeichnet. Um das Maximieren zu erleichtern, geht man häufig zur sogenannten „Log-Likelihood-Funktion“ über, die man durch Logarithmieren der Likelihood-Funktion erhält:

$$\mathcal{L}(\boldsymbol{\theta}) = \ln L(\boldsymbol{\theta}) = \sum_{i=1}^n \ln f(y_i | \mathbf{x}_i, \boldsymbol{\theta}) = \sum_{i=1}^n \mathcal{L}_i(\boldsymbol{\theta}).$$

Da die Logarithmusfunktion streng monoton ist, stimmen die Maximalstellen von L und \mathcal{L} überein. In Kredler (1998, S. 173) [12] wird somit der Maximum-Likelihood-Schätzer wie folgt definiert:

Definition 2.2.4 (ML-Schätzer) Die Zufallsvariablen Y_1, \dots, Y_n seien unabhängig verteilt mit Wahrscheinlichkeitsfunktion bzw. Dichte $f(y_i | \mathbf{x}_i, \boldsymbol{\theta}), i = 1, \dots, n$. Mit $L(\boldsymbol{\theta}) = \prod_{i=1}^n f(y_i | \mathbf{x}_i, \boldsymbol{\theta})$ sei ihre Likelihood-Funktion, mit $\mathcal{L}(\boldsymbol{\theta}) = \ln L(\boldsymbol{\theta})$ ihre Log-Likelihood-Funktion bezeichnet. Jede Maximalstelle

$$\hat{\boldsymbol{\theta}}_{ML} = \max\{L(\hat{\boldsymbol{\theta}}); \hat{\boldsymbol{\theta}} \in \Theta\} = \max\{\mathcal{L}(\hat{\boldsymbol{\theta}}); \hat{\boldsymbol{\theta}} \in \Theta\}$$

heisst dann Maximum-Likelihood-Schätzer bzw. kurz ML-Schätzer für $\boldsymbol{\theta}$. $\hat{\boldsymbol{\theta}}_{ML}$ muss nicht immer existieren und im Falle der Existenz nicht immer eindeutig sein.

Im Fall der GLM's ergibt sich ausgehend von einer Dichte der exponentiellen Familie

$$f(y_i; \theta_i, \phi) = \exp \left\{ \frac{\theta_i y_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right\}$$

die folgende Log-Likelihood-Funktion

$$\mathcal{L}_{GLM}(\boldsymbol{\beta}) = \sum_{i=1}^n \left\{ \frac{\theta_i(\mathbf{x}_i^t \boldsymbol{\beta}) y_i - b(\theta_i(\mathbf{x}_i^t \boldsymbol{\beta}))}{a(\phi)} + c(y_i, \phi) \right\},$$

wobei zu beachten ist, dass das $\boldsymbol{\beta}$ durch die Beziehungen $\theta_i = \theta_i(\mu_i)$, $\mu_i = h(\eta_i)$ und $\eta_i = \mathbf{x}_i^t \boldsymbol{\beta}$ in die Funktion eingeht. Um diesen Zusammenhang zu verdeutlichen, wird die Schreibweise $\theta_i = \theta_i(\mathbf{x}_i^t \boldsymbol{\beta})$ verwendet.

Die Maximalstelle, und damit die ML-Schätzung $\hat{\boldsymbol{\beta}}$ von $\boldsymbol{\beta}$, erhält man durch Differenzieren der Log-Likelihood-Funktion nach $\boldsymbol{\beta}$ und anschliessendes Nullsetzen, d.h.

$$\frac{\partial \mathcal{L}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^n \frac{\partial \ln f_i}{\partial \boldsymbol{\beta}} = \mathbf{0},$$

wobei $f_i = f(y_i \mid \mathbf{x}_i, \boldsymbol{\theta})$ und $\frac{\partial \mathcal{L}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$ ein $(p \times 1)$ -Vektor ist. Die p Gleichungen sind allerdings nichtlinear in den p unbekannten $\boldsymbol{\beta}$'s, so dass eine analytische Lösung des Gleichungssystems nicht möglich ist. Trotzdem kann man durch iteratives Vorgehen zu einer Lösung gelangen. Bekannte Verfahren sind z.B. das „*Newton-Raphson*“- bzw. das „*Fisher-scoring-Verfahren*“ (siehe McCullagh & Nelder (1989, S. 43) [15], und Winkelmann (1996, S. 60) [19]). Um das Prinzip des Fisher-Scoring-Verfahrens zu erklären, müssen zunächst noch einige Begriffe eingeführt werden.

Definition 2.2.5 (Score-Funktion) Die *Score-Funktion* $s(\boldsymbol{\beta})$, oder auch *Score-Vektor* genannt, ist der Vektor der Ableitungen der Log-Likelihood-Funktion nach $\boldsymbol{\beta}$:

$$s(\boldsymbol{\beta}) = \frac{\partial \mathcal{L}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^n \frac{\partial \ln f_i}{\partial \boldsymbol{\beta}} \in \mathbb{R}^p.$$

Die Maximum-Likelihood-Gleichungen lauten dann $s(\boldsymbol{\beta}) = \mathbf{0}$, deren Lösung der ML-Schätzer ist.

Definition 2.2.6 (Fisher-Informationsmatrix) Die beobachtete Fisher-Informationsmatrix wird durch

$$FI_{obs}(\boldsymbol{\beta}) = - \left(\frac{\partial^2 \mathcal{L}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^t} \right)$$

definiert, wobei $\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^t$ die vektorwertige Differentiation abkürzt. Die erwartete Fisher-Informationsmatrix dagegen ist durch die Kovarianzmatrix des Score-Vektors definiert:

$$FI(\boldsymbol{\beta}) = \text{Cov}(s(\boldsymbol{\beta})) = E(s(\boldsymbol{\beta})s(\boldsymbol{\beta})^t).$$

Satz 2.2.7 (Fisher-Informationsmatrix) Die erwartete Fisher-Informationsmatrix lässt sich durch folgende Beziehung aus der beobachteten Fisher-Informationsmatrix berechnen:

$$FI(\boldsymbol{\beta}) = E(FI_{obs}(\boldsymbol{\beta})) = -E \left(\frac{\partial^2 \mathcal{L}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^t} \right).$$

Beweis: Da die erwartete Fisher-Informationsmatrix als Kovarianzmatrix des Score-Vektors definiert wurde, bleibt die Gleichung $E(FI_{obs}(\boldsymbol{\beta})) = \text{Cov}(s(\boldsymbol{\beta}))$ zu zeigen. Dazu betrachte man zunächst

$$\begin{aligned} \frac{\partial^2 \mathcal{L}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^t} &= \frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^t} \ln L(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}^t} \left[\frac{\frac{\partial}{\partial \boldsymbol{\beta}} L(\boldsymbol{\beta})}{L(\boldsymbol{\beta})} \right] = \frac{\frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^t} L(\boldsymbol{\beta}) L(\boldsymbol{\beta}) - \frac{\partial}{\partial \boldsymbol{\beta}} L(\boldsymbol{\beta}) \frac{\partial}{\partial \boldsymbol{\beta}^t} L(\boldsymbol{\beta})}{(L(\boldsymbol{\beta}))^2} \\ &= \frac{\frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^t} L(\boldsymbol{\beta})}{L(\boldsymbol{\beta})} - \underbrace{\frac{\frac{\partial}{\partial \boldsymbol{\beta}} L(\boldsymbol{\beta}) \frac{\partial}{\partial \boldsymbol{\beta}^t} L(\boldsymbol{\beta})}{L(\boldsymbol{\beta}) L(\boldsymbol{\beta})}}_{\frac{\partial}{\partial \boldsymbol{\beta}} \ln L(\boldsymbol{\beta}) \frac{\partial}{\partial \boldsymbol{\beta}^t} \ln L(\boldsymbol{\beta})}. \end{aligned}$$

Damit ergibt sich

$$\begin{aligned}
E(FI_{obs}(\beta)) &= - \int \left[\frac{\partial^2}{\partial \beta \partial \beta^t} \ln L(\beta) \right] L(\beta) dy \\
&= - \int \left[\frac{\frac{\partial^2}{\partial \beta \partial \beta^t} L(\beta)}{L(\beta)} - \frac{\partial}{\partial \beta} \ln L(\beta) \frac{\partial}{\partial \beta^t} \ln L(\beta) \right] L(\beta) dy \\
&= - \int \frac{\partial^2}{\partial \beta \partial \beta^t} L(\beta) dy + \underbrace{\int \frac{\partial}{\partial \beta} \ln L(\beta)}_{s(\beta)} \underbrace{\frac{\partial}{\partial \beta^t} \ln L(\beta) L(\beta) dy}_{s(\beta)^t} \\
&\stackrel{(*)}{=} 0 + \int s(\beta) s(\beta)^t L(\beta) dy \\
&= E(s(\beta) s(\beta)^t) \\
&= Cov(s(\beta)), \quad \text{da } E(s(\beta)) = 0.
\end{aligned}$$

(*): Für das erste Integral gilt unter Regularitätsbedingungen, die am Ende dieses Abschnittes aufgeführt und erläutert werden, dass Integration und Differentiation vertauscht werden dürfen. Damit erhält man

$$- \int \frac{\partial^2}{\partial \beta \partial \beta^t} L(\beta) dy = - \frac{\partial^2}{\partial \beta \partial \beta^t} \underbrace{\int L(\beta) dy}_{=1, \text{ da } L(\beta) \text{ eine Dichte ist}} = - \frac{\partial^2}{\partial \beta \partial \beta^t} 1 = 0.$$

□

Nun sind alle notwendigen Begriffe definiert, um die Fisher-Scoring-Methode zu erklären. Diese Methode ist ein iterativer Algorithmus zur Bestimmung des ML-Schätzers $\hat{\beta}$. Er verwendet die Taylorentwicklung erster Ordnung, um die nichtlinearen Gleichungen zu linearisieren und so eine schnelle Approximation zu gewährleisten. Sei $\hat{\beta}^{(s)}$ die vorläufige Approximation an den ML-Schätzer $\hat{\beta}$ in der s-ten Iteration. Der Rechenschritt zur Bestimmung der verbesserten Approximation in der s+1-ten Iteration lautet dann nach Fahrmeir & Tutz (1994, S. 40) [7]:

$$\hat{\beta}^{(s+1)} = \hat{\beta}^{(s)} + FI^{-1}(\hat{\beta}^{(s)}) s(\hat{\beta}^{(s)}).$$

Es wird solange iteriert, bis ein bestimmtes Abstandskriterium unter einer vorgegebenen Schranke ε bleibt. So ein Kriterium könnte z.B.

$$\frac{\|\hat{\beta}^{(s+1)} - \hat{\beta}^{(s)}\|}{\|\hat{\beta}^{(s)}\|} < \varepsilon, \quad \varepsilon > 0$$

sein, wobei $\|\cdot\|$ die euklidische Norm bezeichnet.

Hat man nun einen Schätzer $\hat{\beta}$ errechnet, bleibt zur Kontrolle des Maximums noch die sogenannte „Hesse-Matrix“ $H(\beta)$ zu überprüfen. Ist

$$H(\beta) = \frac{\partial^2 \mathcal{L}(\beta)}{\partial \beta \partial \beta^t}$$

an der Stelle $\beta = \hat{\beta}$ negativ-definit, so ist der errechnete Schätzer $\hat{\beta}$ wirklich ein Maximum. In der Regel sind die errechneten $\hat{\beta}$ keine globalen Maxima der Log-Likelihood-Funktion, sondern nur Lösungen der Score-Gleichungen $s(\hat{\beta}) = 0$, was einem lokalen Maximum entspricht, wenn die Hesse-Matrix negativ-definit ist. Für viele Modelle ist die Log-Likelihood-Funktion aber konkav, so dass das lokale mit dem globalen Maximum übereinstimmt. Für strikt konkave Log-Likelihood-Funktionen ist das Maximum sogar eindeutig, wenn es existiert.

Satz 2.2.8 *Sei V eine nicht-leere konkave Menge $V \subset \mathbb{R}^m$ und $f : V \rightarrow \mathbb{R}$ eine strikt konkave Funktion auf V , so gibt es höchstens einen Maximalpunkt von f auf V .*

Beweis: *Eine strikt konkave Funktion ist für alle Paare $\mathbf{x}_1, \mathbf{x}_2 \in V$ mit $\mathbf{x}_1 \neq \mathbf{x}_2$ und jede Zahl $\lambda \in (0, 1)$ definiert durch die Ungleichung*

$$f(\lambda \mathbf{x}_1 + (1 - \lambda) \mathbf{x}_2) > \lambda f(\mathbf{x}_1) + (1 - \lambda) f(\mathbf{x}_2).$$

Angenommen, es gäbe nun zwei verschiedene Maximalpunkte $\hat{\mathbf{x}}_1$ und $\hat{\mathbf{x}}_2 \in V$ von f , dann gilt für die Verbindungsgerade von $\hat{\mathbf{x}}_1$ und $\hat{\mathbf{x}}_2$

$$\lambda f(\hat{\mathbf{x}}_1) + (1 - \lambda) f(\hat{\mathbf{x}}_2) = f(\hat{\mathbf{x}}_1) = f(\hat{\mathbf{x}}_2).$$

Für jedes $\lambda \in (0, 1)$ und für jedes $\mathbf{x} = \lambda \hat{\mathbf{x}}_1 + (1 - \lambda) \hat{\mathbf{x}}_2 \in V$, das zwischen $\hat{\mathbf{x}}_1$ und $\hat{\mathbf{x}}_2$ liegt, ergibt sich aber nach der Definition der Konkavität

$$f(\mathbf{x}) = f(\lambda \hat{\mathbf{x}}_1 + (1 - \lambda) \hat{\mathbf{x}}_2) > \lambda f(\hat{\mathbf{x}}_1) + (1 - \lambda) f(\hat{\mathbf{x}}_2) = f(\hat{\mathbf{x}}_1) = f(\hat{\mathbf{x}}_2),$$

was aber ein Widerspruch dazu ist, dass $\hat{\mathbf{x}}_1$ und $\hat{\mathbf{x}}_2$ Maxima sind.

□

Desweiteren gilt nach Krabs (1983, S. 59) [11] der folgende Satz:

Satz 2.2.9 *Eine Funktion $f : V \rightarrow \mathbb{R}$ auf einer nicht-leeren offenen und konkaven Teilmenge des \mathbb{R}^n , die für jedes $\mathbf{x} \in V$ stetige zweite partielle Ableitungen besitzt, ist genau dann strikt konkav auf V , wenn ihre Hesse-Matrix $H(\mathbf{x})$ für jedes $\mathbf{x} \in V$ negativ definit ist.*

Im folgenden Beispiel wird für den Spezialfall der Poissonverteilung nun gezeigt, dass die Hesse-Matrix negativ-definit und somit die Log-Likelihood-Funktion strikt konkav ist, woraus nach Satz 2.2.8 die Eindeutigkeit des Maximums folgt.

Beispiel 2.2.10 (ML-Schätzung bei Poissonverteilung) Für dieses Beispiel sei die kanonische Linkfunktion $\mu_i = E(Y_i | \mathbf{x}_i) = \exp(\mathbf{x}_i^t \boldsymbol{\beta})$ gewählt. Ausgehend von der Wahrscheinlichkeitsfunktion der Poissonverteilung für $i = 1, \dots, n$ unabhängige Zufallsvariablen Y_i erhält man die Log-Likelihood-Funktion

$$\mathcal{L}(\boldsymbol{\beta}) = \sum_{i=1}^n \{-\mu_i + y_i \ln \mu_i - \ln(y_i!)\} = \sum_{i=1}^n \{-\exp(\mathbf{x}_i^t \boldsymbol{\beta}) + y_i \mathbf{x}_i^t \boldsymbol{\beta} - \ln(y_i!)\}.$$

Dieser Term muss nun bezüglich $\boldsymbol{\beta}$ abgeleitet und nullgesetzt werden.

$$\begin{aligned} \frac{\partial \mathcal{L}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^n \{-\mathbf{x}_i \exp(\mathbf{x}_i^t \boldsymbol{\beta}) + y_i \mathbf{x}_i\} = \sum_{i=1}^n \{y_i - \exp(\mathbf{x}_i^t \boldsymbol{\beta})\} \mathbf{x}_i \\ &= \begin{pmatrix} \sum_{i=1}^n \{y_i - \exp(\mathbf{x}_i^t \boldsymbol{\beta})\} x_{i1} \\ \vdots \\ \sum_{i=1}^n \{y_i - \exp(\mathbf{x}_i^t \boldsymbol{\beta})\} x_{ip} \end{pmatrix} = \mathbf{0}. \end{aligned}$$

Um daraus die Hesse-Matrix zu berechnen, werden zunächst durch partielles Ableiten die einzelnen Matrixeinträge ermittelt.

$$\frac{\partial^2 \mathcal{L}(\boldsymbol{\beta})}{\partial \beta_r \partial \beta_s} = \frac{\partial}{\partial \beta_s} \sum_{i=1}^n \{y_i - \exp(\mathbf{x}_i^t \boldsymbol{\beta})\} x_{ir} = - \sum_{i=1}^n \exp(\mathbf{x}_i^t \boldsymbol{\beta}) x_{ir} x_{is}, \quad r, s = 1, \dots, p.$$

Die gesamte Hesse-Matrix lässt sich demnach durch

$$H(\boldsymbol{\beta}) = \frac{\partial^2 \mathcal{L}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^t} = - \sum_{i=1}^n \exp(\mathbf{x}_i^t \boldsymbol{\beta}) \underbrace{\mathbf{x}_i \mathbf{x}_i^t}_{\in \mathbb{R}^{p \times p}}$$

ausdrücken. Für den Nachweis, dass auch wirklich ein Maximum vorliegt, muss noch gezeigt werden, dass die Hesse-Matrix an der Stelle $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$ negativ definit ist, d.h. es ist $\mathbf{z}^t H(\hat{\boldsymbol{\beta}}) \mathbf{z} < 0 \quad \forall \mathbf{z} \neq \mathbf{0}$ zu überprüfen. Kann aber gezeigt werden, dass die Hesse-Matrix für beliebige $\boldsymbol{\beta}$ negativ definit ist, dass also $\mathbf{z}^t H(\boldsymbol{\beta}) \mathbf{z} < 0 \quad \forall \mathbf{z} \neq \mathbf{0}$ gilt, so ist dies nach Satz 2.2.9 gleichzeitig der Nachweis für eine global strikt konkave Log-Likelihood-Funktion und nach Satz 2.2.8 ist das Maximum dann eindeutig.

$$\mathbf{z}^t H(\boldsymbol{\beta}) \mathbf{z} = -\mathbf{z}^t \sum_{i=1}^n \underbrace{\exp(\mathbf{x}_i^t \boldsymbol{\beta}) \mathbf{x}_i \mathbf{x}_i^t}_{\text{Skalar}} \mathbf{z} = - \sum_{i=1}^n \exp(\mathbf{x}_i^t \boldsymbol{\beta}) \mathbf{z}^t \mathbf{x}_i \mathbf{x}_i^t \mathbf{z}$$

$$\begin{aligned}
&= - \sum_{i=1}^n \exp(\mathbf{x}_i^t \boldsymbol{\beta}) \underbrace{\mathbf{z}^t \mathbf{x}_i}_{=: w_i \in \mathbb{R}} (\mathbf{z}^t \mathbf{x}_i)^t = - \sum_{i=1}^n \exp(\mathbf{x}_i^t \boldsymbol{\beta}) w_i w_i^t \\
&= - \sum_{i=1}^n \exp(\mathbf{x}_i^t \boldsymbol{\beta}) w_i^2 \leq 0.
\end{aligned}$$

Es bleibt zu zeigen, dass der Wert gleich Null nur dann angenommen wird, wenn $\mathbf{z} = \mathbf{0}$ ist. Da die Exponentialfunktion strikt positiv ist und somit $\exp(\mathbf{x}_i^t \boldsymbol{\beta}) > 0$ gilt, ist $-\sum_{i=1}^n \exp(\mathbf{x}_i^t \boldsymbol{\beta}) w_i^2 = 0$, genau dann wenn $w_i = 0$, $\forall i = 1, \dots, n$. Dies gilt aber genau dann, wenn $\mathbf{x}_i^t \mathbf{z} = 0$, $\forall i = 1, \dots, n$. Ausführlich in Matrixschreibweise ergibt dies

$$\begin{pmatrix} \mathbf{x}_1^t \\ \vdots \\ \mathbf{x}_n^t \end{pmatrix} \mathbf{z} = \begin{pmatrix} x_{11}, & \dots, & x_{1p} \\ \vdots & \vdots & \vdots \\ x_{n1}, & \dots, & x_{np} \end{pmatrix} \begin{pmatrix} z_1 \\ \vdots \\ z_p \end{pmatrix} = \mathbf{0}.$$

Nun hat aber die Matrix nach Voraussetzung (siehe Beginn des Abschnittes) einen vollen Spaltenrang p und es gilt $p \leq n$, denn es können höchstens nur so viele Parameter geschätzt werden, wie Beobachtungen vorhanden sind. Damit gibt es p linear unabhängige Zeilen in der Matrix, deren Indexmenge mit $K = \{k_1, \dots, k_p\}$ bezeichnet wird. Dann gilt

$$\mathbf{X}^* \mathbf{z} = \underbrace{\begin{pmatrix} \mathbf{x}_{k_1}^t \\ \vdots \\ \mathbf{x}_{k_p}^t \end{pmatrix}}_{\in \mathbb{R}^{p \times p}} \mathbf{z} = \mathbf{0},$$

was aber nur möglich ist, wenn $\mathbf{z} = \mathbf{0}$, denn die Matrix \mathbf{X}^* hat vollen Rang p . Damit ist gezeigt, dass $\mathbf{z}^t H(\boldsymbol{\beta}) \mathbf{z} < 0 \forall \mathbf{z} \neq \mathbf{0}$, d.h. die Hesse-Matrix ist für beliebige $\boldsymbol{\beta}$ negativ definit. Die Log-Likelihood-Funktion ist daher global konkav und somit sind die Bedingungen für ein globales Maximum erfüllt.

Sobald die Schätzungen durchgeführt sind, geht es darum, die Eigenschaften der Schätzer zu überprüfen. Zu den wichtigsten gehören Unverzerrtheit, Konsistenz und asymptotische Normalverteilung. Die Unverzerrtheit und Konsistenz werden in Kredler (1998, S. 169f) [12] folgendermassen definiert:

Definition 2.2.11 (Unverzerrtheit bzw. Erwartungstreue eines Schätzers) Eine Schätzfunktion $T : \mathbb{R}^n \rightarrow \mathbb{R}^m$ heisst erwartungstreu oder unverzerrt, falls

$$E[T(X_1, \dots, X_n)] = \boldsymbol{\theta}.$$

Die Abweichung $E[T(X_1, \dots, X_n)] - \boldsymbol{\theta}$ heisst Bias des Schätzers T .

Für die Definition der Konsistenz wird T_n anstatt T geschrieben, um die Abhängigkeit des Schätzers vom Stichprobenumfang anzudeuten. Ausserdem wird die Definition auf eindimensionale Schätzer mit Parameter θ eingeschränkt, um eine einfache Darstellung zu erhalten.

Definition 2.2.12 (Konsistenz eines Schätzers) Ein Schätzer $T_n : \mathbb{R}^n \rightarrow \mathbb{R}$ heisst konsistent für θ , falls

$$\lim_{n \rightarrow \infty} P(|T_n(X_1, \dots, X_n) - \theta| \leq \varepsilon) = 1, \quad \forall \varepsilon > 0.$$

Konsistenz bedeutet, dass der Schätzer T_n in Wahrscheinlichkeit gegen θ konvergiert.

Diese gewünschten Eigenschaften eines Schätzers sind allerdings nur unter bestimmten Voraussetzungen, den sogenannten „Regularitätsbedingungen“ gegeben.

In Cameron & Trivedi (1998, S. 23) [2] sind diese wie folgt zusammengefasst:

- (i) Die Dichte $f(y | \mathbf{x}_i, \boldsymbol{\theta})$ ist global definiert und es gilt $f(y | \mathbf{x}_i, \boldsymbol{\theta}_1) \neq f(y | \mathbf{x}_i, \boldsymbol{\theta}_2)$ für alle $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2$.
- (ii) Es gilt $\boldsymbol{\theta} \in \Theta$, wobei der Parameterraum Θ endlichdimensional, abgeschlossen und kompakt ist.
- (iii) Es existieren stetige und beschränkte Ableitungen von $\mathcal{L}(\boldsymbol{\theta})$ bis zur 3. Ordnung.
- (iv) Die Reihenfolge von Differentiation und Integration der Likelihood-Funktion darf vertauscht werden.
- (v) Die Kovariablenvektoren $\mathbf{x}_i, i = 1, \dots, n$ erfüllen folgende Bedingungen

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

$$(b) \quad \frac{E(\omega_i^2)}{\sum_i E(\omega_i^2)} = 0 \quad \forall i, \text{ wobei } \omega_i \equiv \mathbf{x}_i^t \frac{\partial \ln f(y_i | \mathbf{x}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

$$(c) \quad \lim_{n \rightarrow \infty} \frac{\sum_{i=1}^n E(\omega_i^2 | \Omega_{i-1})}{\sum_{i=1}^n E(\omega_i^2)} = 1, \text{ wobei } \Omega_{i-1} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{i-1}).$$

Die erste Bedingung sichert, dass ein eindeutiges Maximum existiert. Die zweite Voraussetzung schliesst mögliche Probleme am Rand des Parameterraumes Θ aus und kann weggelassen werden, wenn z.B. \mathcal{L} global konkav ist. Die dritte Bedingung kann oft etwas gelockert werden, indem man die Existenz der Ableitungen nur bis zur zweiten Ordnung fordert. Der vierte Punkt ist eine wichtige Bedingung, die solche Dichten ausschliesst, bei denen der Wertebereich von y_i von $\boldsymbol{\theta}$ abhängt. Durch den letzten Punkt wird jede

Beobachtung ausgeschlossen, deren Anteil an der Likelihood-Funktion zu gross ist. Da für den ML-Schätzer $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\beta}}$ keine analytischen Lösungen existieren, können nur asymptotische Eigenschaften gefolgert werden, allerdings nur unter der Voraussetzung, dass die oben beschriebenen Regularitätsbedingungen gelten. Nach Fahrmeir & Tutz (1994, S. 43) [7] lassen sich die wichtigsten Eigenschaften des ML-Schätzers $\hat{\boldsymbol{\beta}}$ folgendermassen zusammenfassen:

(i) **Asymptotische Existenz und Eindeutigkeit:**

Die Wahrscheinlichkeit, dass $\hat{\boldsymbol{\beta}}$ existiert und lokal eindeutig ist, konvergiert gegen 1 für $n \rightarrow \infty$.

(ii) **Konsistenz:**

Falls $\boldsymbol{\beta}$ den wahren Wert beschreibt, dann konvergiert $\hat{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{\beta}$ für $n \rightarrow \infty$ in Wahrscheinlichkeit.

(iii) **Asymptotische Normalverteilung:**

Die Verteilung des ML-Schätzers ist eine Normalverteilung für $n \rightarrow \infty$, bzw. formal

$$\hat{\boldsymbol{\beta}} \overset{a}{\sim} N(\boldsymbol{\beta}, FI^{-1}(\hat{\boldsymbol{\beta}})), \quad (2.10)$$

was gleichbedeutend ist mit

$$FI^{\frac{1}{2}}(\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} N(\mathbf{0}, I_p),$$

wobei d die Konvergenz in Verteilung andeutet und I_p eine p -dimensionale Einheitsmatrix abkürzt. $\hat{\boldsymbol{\beta}}$ ist also asymptotisch normalverteilt mit asymptotischer Kovarianzmatrix

$$Cov(\hat{\boldsymbol{\beta}}) \overset{a}{=} FI^{-1}(\hat{\boldsymbol{\beta}}),$$

wobei $FI^{-1}(\hat{\boldsymbol{\beta}})$ die Inverse der Fisher-Informationsmatrix ausgewertet an der Stelle $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$ darstellt.

Es sind z.B. in Fahrmeir & Kaufmann (1985, S. 364) [5] diese Eigenschaften bewiesen worden wie auch in Pruscha (1989) [16], der in den Kapiteln 6 und 7 auf die Beweise der asymptotischen Eigenschaften eines ML-Schätzers eingeht. Speziell werden die Existenz eines konsistenten ML-Schätzers und seine asymptotische Normalverteilung in GLM's mit kanonischer Linkfunktion auf S. 273f bewiesen, während auf S. 281 der Beweis für GLM's mit nicht notwendig kanonischen Linkfunktionen gezeigt wird.

2.2.7 Asymptotische Hypothesentests in GLM's

Zur statistischen Modellbildung gehört zunächst die explorative Datenanalyse, bei der erste geeignete sinnvolle Modelle vorgeschlagen werden. Anschliessend erfolgt eine Überprüfung des gefundenen Modells mit Hilfe von statistischen Hypothesentests. Erst danach schliesst sich die Residualanalyse an, durch welche die Modellannahmen kontrolliert werden. Nun sollen in diesem Abschnitt die wichtigsten Hypothesentests in GLM's kurz vorgestellt werden, während auf die grundlegende Testtheorie ausführlich in Kapitel 3 eingegangen wird.

Die meisten Testprobleme für den Parametervektor β sind lineare Hypothesen der Form

$$H : C\beta = \xi \quad \text{gegen} \quad K : C\beta \neq \xi, \quad (2.11)$$

wobei von der Matrix C vorausgesetzt wird, dass sie einen vollen Zeilenrang $q \leq p + 1$ besitzt. Von diesen Testproblemen gibt es einen wichtigen Spezialfall

$$H : \beta_r = \mathbf{0} \quad \text{gegen} \quad K : \beta_r \neq \mathbf{0}. \quad (2.12)$$

Dabei beschreibt β_r einen Teilvektor von β . Mit den Hypothesentests (2.12) überprüft man die Signifikanz der zu β_r gehörenden Effekte, indem man das Teilmodell, das durch $\beta_r = \mathbf{0}$ definiert ist, mit dem vollen Modell vergleicht. Zum Testen der genannten Hypothesen können die Likelihood-Quotienten-, Score- und Wald-Statistik verwendet werden, die nun nacheinander erläutert werden.

(i) **Likelihood-Quotienten-Test:**

Die Likelihood-Quotienten-Statistik

$$lq = -2\{\mathcal{L}(\tilde{\beta}) - \mathcal{L}(\hat{\beta})\}$$

vergleicht das unrestringierte Maximum $\mathcal{L}(\hat{\beta})$ des Log-Likelihoods mit dem Maximum $\mathcal{L}(\tilde{\beta})$, wobei $\tilde{\beta}$ der restringierte ML-Schätzer unter der Nullhypothese $H : C\beta = \xi$ ist. Falls das unrestringierte Maximum $\mathcal{L}(\hat{\beta})$ signifikant grösser ist als $\mathcal{L}(\tilde{\beta})$, was gleichbedeutend mit einer sehr grossen Teststatistik lq ist, so wird H zu Gunsten von K verworfen. Der Likelihood-Quotienten-Test für das Testproblem (2.12), das der Überprüfung eines Teilmodells, welches durch $\beta_r = \mathbf{0}$ definiert ist, entspricht, erfordert die ML-Schätzung von β_r im entsprechenden Teilmodell und damit weitere Iterationen im Fisher-Scoring-Verfahren. Allerdings

ist dieser Aufwand noch sehr klein im Vergleich zur Berechnung des ML-Schätzers $\tilde{\beta}$ im Testproblem (2.11), in welchem durch die allgemeineren linearen Gleichungsrestriktionen der numerische Aufwand erheblich erhöht wird.

(ii) **Wald-Test:**

Die Wald-Statistik

$$w = (\mathbf{C}\hat{\beta} - \boldsymbol{\xi})^t [\mathbf{C}FI^{-1}(\hat{\beta})\mathbf{C}^t]^{-1} (\mathbf{C}\hat{\beta} - \boldsymbol{\xi})$$

misst die durch die asymptotische Kovarianzmatrix $\mathbf{C}FI^{-1}(\hat{\beta})\mathbf{C}^t$ von $\mathbf{C}\hat{\beta}$ gewichtete Distanz zwischen dem unrestringierten Schätzer $\mathbf{C}\hat{\beta}$ von $\mathbf{C}\beta$ und dem hypothetischen Wert $\boldsymbol{\xi} = \mathbf{C}\beta$ unter der Nullhypothese.

(iii) **Score-Test:**

Die Score-Statistik ist gegeben durch

$$u = s(\tilde{\beta})^t FI^{-1}(\tilde{\beta}) s(\tilde{\beta}).$$

Der dazugehörige Test basiert auf der folgenden Idee: Die Score-Funktion $s(\beta)$ ist für das unrestringierte Modell gleich dem Nullvektor $\mathbf{0}$, falls sie an dem unrestringierten ML-Schätzer $\hat{\beta}$ ausgewertet wird. Ersetzt man nun $\hat{\beta}$ durch den ML-Schätzer $\tilde{\beta}$ unter H , so wird sich $s(\tilde{\beta})$ signifikant von Null unterscheiden, falls H nicht wahr ist. Die Distanz zwischen $s(\tilde{\beta})$ und Null wird eben durch die Score-Statistik u gemessen, wobei die Inverse der Fisher-Informationsmatrix als Gewichtung fungiert. Betrachtet man den Spezialfall (2.12) des allgemeinen Testproblems, so reduzieren sich sowohl die Wald- als auch die Score-Statistik auf eine stark vereinfachte Form. Für die Wald-Statistik gilt dann

$$w = \hat{\beta}_r^t \hat{A}_r^{-1} \hat{\beta}_r,$$

während die Score-Statistik durch

$$u = \tilde{s}_r^t \tilde{A}_r \tilde{s}_r$$

gegeben ist. Dabei ist A_r die Teilmatrix von $A = FI^{-1}$, die zu den entsprechenden Elementen von β_r gehört, und s_r ist der dazugehörige Teilvektor von s .

Nach Fahrmeir & Tutz (1994, S. 46) [7] ist allen drei Teststatistiken gemeinsam, dass sie unter der Nullhypothese von (2.11) asymptotisch äquivalent sind und dieselbe χ^2 -Grenzverteilung mit q Freiheitsgraden besitzen, d.h. $lq, w, u \stackrel{a}{\sim} \chi_q^2$, und analog unter

der Hypothese H von (2.12) die χ^2 -Grenzverteilung mit r Freiheitsgraden aufweisen. Ein ausführlicher Beweis für die asymptotische χ^2 -Verteilung der Teststatistiken wurde in der Diplomarbeit von A. Franzmann (2001, S. 24f) [9] ausgearbeitet.

2.2.8 Devianzanalyse

Nach der vorläufigen Modellerstellung und der Überprüfung durch geeignete Hypothesentests müssen nun die Modellannahmen kontrolliert werden. Dazu gehören sowohl die nachfolgend erläuterten Anpassungskriterien als auch die im nächsten Abschnitt beschriebene Residualanalyse. Den Prozess, ein Modell zu erstellen, kann man als Methode betrachten, einen gegebenen Datensatz \mathbf{y} durch gefittete Daten $\hat{\mathbf{y}}$ bzw. $\hat{\boldsymbol{\mu}}$ zu ersetzen, die vom Modell unter Verwendung einer kleinen Anzahl an Parametern abgeleitet werden. In der Regel werden aber die $\hat{\mu}_i$ von den gegebenen Daten y_i abweichen. Die Frage ist dann, ob die Abweichung noch tolerierbar ist oder nicht. Benötigt wird demnach ein Mass, das angibt, wie gut das erarbeitete Modell an die Datenstruktur angepasst wird. In diesem Abschnitt sollen nun Methoden aufgezeigt werden, die die Abweichung zwischen den wahren und den gefitteten Daten messen.

Das bekannteste Kriterium hierfür ist die „Devianz“. Die skalierte Devianz ist gegeben durch

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

$\mathcal{L}(\mathbf{y}, \mathbf{y})$ beschreibt hierbei die Log-Likelihood-Funktion, die man in einem vollen Modell mit n Parametern erhält, d.h. man erreicht die maximal mögliche Anpassung, da die Daten exakt reproduziert werden. $\mathcal{L}(\hat{\boldsymbol{\mu}}, \mathbf{y})$ dagegen ist die Log-Likelihood-Funktion des aktuellen Modells mit $p < n$ Parametern. Da $\mathcal{L}(\mathbf{y}, \mathbf{y})$ nicht von Parametern abhängt, ist die Maximierung von $\mathcal{L}(\hat{\boldsymbol{\mu}}, \mathbf{y})$ äquivalent zur Minimierung von $D^*(\mathbf{y}, \hat{\boldsymbol{\mu}}, \phi)$ bezüglich $\hat{\boldsymbol{\mu}}$. Der Faktor 2 dient nur der Normalisierung, damit im Fall der Normalverteilung die Devianz mit der Abstandsquadratsumme übereinstimmt.

Wird $D^*(\mathbf{y}, \hat{\boldsymbol{\mu}}, \phi)$ mit dem Parameter $a(\phi)$, der als bekannt oder zumindest als fest vorausgesetzt wird, multipliziert, erhält man die Devianz

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

Beachtet man, dass $\mathcal{L} = \ln L$ ist, dann ist die Devianz also ein Mass, das aus dem Logarithmus des Likelihood-Quotienten gebildet wird.

Ein anderes bedeutendes Mass zur Bestimmung der Abweichung ist die sogenannte

„*Pearson-Statistik*“, die durch

$$\chi^2 = \sum_{i=1}^n \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}$$

definiert wird. Dabei ist $V(\hat{\mu}_i)$ die geschätzte Varianzfunktion der unterstellten Verteilung, die in (2.5) definiert wurde. Sind Erwartungswert und Varianz korrekt spezifiziert, so gilt $E(\chi^2) = E\left[\sum_{i=1}^n \frac{(y_i - \mu_i)^2}{\text{Var}(\mu_i)}\right] = n$, da $E\left[\frac{(y_i - \mu_i)^2}{\text{Var}(\mu_i)}\right] = 1$ ist. Nun müssen aber p Parameter noch geschätzt werden, womit sich die Anzahl der Freiheitsgrade von n auf $n - p$ reduziert. Offensichtlich liegt bei einer Normalverteilung der y_i eine exakte χ^2_{n-p} -Verteilung vor, da sich die Pearson-Statistik in diesem Fall aus der Summe von quadrierten standardisierten normalverteilten Zufallsvariablen ergibt. In anderen Fällen ist nach Fahrmeir & Tutz (1994, S. 48) [7] sowohl die Devianz als auch die Pearson-Statistik zumindest asymptotisch χ^2_{n-p} verteilt mit $n - p$ Freiheitsgraden. Ein Beweis hierfür steht in Pruscha (1989) [16], der auf S. 243 die Asymptotik für die Pearson-Statistik zeigt und auf S. 273 in Satz 3.4 die asymptotische χ^2 -Verteilung für die skalierte Devianz in GLM's mit kanonischer Linkfunktion und auf S. 281 für allgemeine GLM's beweist.

Mit Hilfe der Devianz und ihrer asymptotischen χ^2 -Verteilung können zusätzlich zu den in Abschnitt 2.2.7 beschriebenen Hypothesentests zwei weitere Tests durchgeführt werden, die die Anpassungsgüte des gefundenen Modells überprüfen. Der „*Residual Deviance-Test*“ oder auch „*Test of Fit*“ genannt, testet die allgemeine Nullhypothese

$$H : \text{Modell passt} \quad \text{gegen} \quad K : \text{Modell passt nicht.}$$

Man verwirft dann die Nullhypothese, d.h. man geht davon aus, dass die Anpassung des erarbeiteten Modells an die Datenstruktur noch nicht ausreichend ist, genau dann wenn

$$\text{Devianz} > \chi^2_{n-p, 1-\alpha}$$

gilt, wobei n die Anzahl der Beobachtungen, p die Anzahl der zu schätzenden Parameter, α das Signifikanzniveau beschreibt und $\chi^2_{n-p, 1-\alpha}$ das $1-\alpha$ Quantil der χ^2 -Verteilung ist. Der Residual-Deviance-Test gehört zu den asymptotischen Tests, da die Devianz, wie vorher erläutert, asymptotisch χ^2 -verteilt ist. Im Gegensatz zu diesem Test, der das gesamte Modell überprüft, wird beim „*Partial-Deviance-Test*“ die Signifikanz von nur einer oder mehreren Kovariablen gemessen. Dies wird erreicht, indem man zwei Modelle M_1 und M_2 miteinander in Beziehung bringt, wobei in Modell M_2 im Vergleich zu Modell M_1 genau diejenigen Parameter zusätzlich enthalten sind, deren Einfluss getestet

werden soll. Der Regressionsparametervektor β wird also aufgespalten in $\beta = (\beta_1, \beta_2)^t$, wobei Modell M_1 aus den Parametern, die in β_1 enthalten sind, besteht, und Modell M_2 zusätzlich die Parameter aus β_2 aufweist. Der dazugehörige Test lautet dann

$$H : \beta_2 = \mathbf{0} \quad \text{gegen} \quad K : \beta_2 \neq \mathbf{0}.$$

Die Nullhypothese wird verworfen, sobald

$$\text{Devianz}(M_1) - \text{Devianz}(M_2) > \chi_{p_2-p_1, 1-\alpha}^2$$

gilt. Dabei ist p_1 die Anzahl der Parameter in Modell M_1 und entsprechend p_2 die Anzahl der Parameter in Modell M_2 . Auch hier beschreibt wieder $\chi_{p_2-p_1, 1-\alpha}^2$ das $1-\alpha$ Quantil der $\chi_{p_2-p_1}^2$ -Verteilung. Wird die Nullhypothese abgelehnt, so erhält man einen Nachweis für den signifikanten Einfluss der zu den Parametern aus β_2 gehörenden Kovariablen. Zu bemerken bleibt, dass sich, wenn man mit V das volle Modell bezeichnet, beim Partial-Deviance-Test eine allgemeine Testgrösse

$$\begin{aligned} \text{Devianz}(M_1) - \text{Devianz}(M_2) &= -2a(\phi) (\mathcal{L}(M_1) - \mathcal{L}(V)) + 2a(\phi) (\mathcal{L}(M_2) - \mathcal{L}(V)) \\ &= -2a(\phi) (\mathcal{L}(M_1) - \mathcal{L}(M_2)) \end{aligned}$$

ergibt, die bis auf den Vorfaktor $a(\phi)$ der Testgrösse des Likelihood-Quotienten-Tests in Abschnitt 2.2.7 entspricht, wenn M_2 das unrestringierte und M_1 das Modell unter der Nullhypothese abkürzt.

Beispiel 2.2.13 (Devianz und Pearson-Statistik bei Poissonverteilung) Für die Poissonregression ergibt sich die folgende Devianz, die hier schrittweise hergeleitet wird: Zunächst werden die Log-Likelihood-Funktionen berechnet. Zu beachten ist dabei, dass $\mathcal{L} = \sum_{i=1}^n \mathcal{L}_i$ gilt.

$$\mathcal{L}_i(\hat{\mu}_i, y_i) = -\hat{\mu}_i + y_i \ln \hat{\mu}_i - \ln(y_i!) \quad \text{und} \quad \mathcal{L}_i(y_i, y_i) = -y_i + y_i \ln y_i - \ln(y_i!).$$

Die Devianz lautet dann:

$$\begin{aligned} D_{Poi}(\mathbf{y}, \hat{\boldsymbol{\mu}}) &= -2 \sum_{i=1}^n \mathcal{L}_i(\hat{\mu}_i, y_i) - \mathcal{L}_i(y_i, y_i) \\ &= -2 \sum_{i=1}^n [-\hat{\mu}_i + y_i \ln \hat{\mu}_i - \ln(y_i!) - (-y_i + y_i \ln y_i - \ln(y_i!))] \\ &= -2 \sum_{i=1}^n (-y_i \ln \frac{y_i}{\hat{\mu}_i} + y_i - \hat{\mu}_i) \\ &= 2 \sum_{i=1}^n (y_i \ln \frac{y_i}{\hat{\mu}_i} - (y_i - \hat{\mu}_i)). \end{aligned}$$

Dabei wird die Konvention unterstellt, dass $y_i \ln y_i = 0$, falls $y_i = 0$. Zu bemerken bleibt, dass nach Beispiel 2.2.2 $a(\phi) = 1$ gilt und somit die Devianz D mit der skalierten Devianz D^* übereinstimmt.

Die generalisierte Pearson-Statistik kann man sofort angeben. In Beispiel 2.2.2 wurde gezeigt, dass $\text{Var}(Y_i) = b''(\theta_i)a(\phi) = e^{\theta_i} = \mu_i$ ist. Somit lautet die Pearson-Statistik im Spezialfall der Poissonverteilung

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

Die Pearson-Statistik spielt speziell für die Poissonverteilung eine wichtige Rolle. Sie kann nämlich als Indikator für die theoretisch geforderte Gleichheit von Erwartungswert und Varianz verwendet werden, da sie das Verhältnis der empirischen Varianz zum empirischen Erwartungswert angibt. Nach McCullagh & Nelder (1989, S. 200) [15] lässt sich der Dispersionsparameter ϕ , der nach Beispiel 2.2.2 im Poissonfall 1 ist, durch $\frac{\chi_{Poi}^2}{n-p}$ schätzen. Somit lässt sich $\chi_{Poi}^2 > n - p$ als Hinweis auf Überdispersion deuten, während man entsprechend den Fall $\chi_{Poi}^2 < n - p$ so interpretiert, dass die Varianz kleiner als der Erwartungswert ist.

Sowohl die Devianz als auch die Pearson-Statistik sollten nie als alleinige Kriterien für die Anpassung herangezogen werden. Sie dienen vielmehr dazu, eine Vorentscheidung zu bekommen, ob ein Modell geeignet ist, während die genaue Überprüfung des Modells eine detailliertere Untersuchung erfordert, die auch die nachstehend erläuterte Residualanalyse beinhaltet.

2.2.9 Residualanalyse

Residuen, die im klassischen Modell als Differenz zwischen wahrem und gefittetem Wert definiert werden, d.h. $r_i = (y_i - \hat{\mu}_i)$, können dazu verwendet werden, die Anpassung des Modells bezüglich der Wahl der Varianzfunktion, der Linkfunktion und der Terme im linearen Prädiktor zu untersuchen. Für die klassische lineare Regression mit normalverteilten, homoskedastischen Fehlern gilt $(y_i - \mu_i) \sim N(0, \sigma^2)$, was bedeutet, dass in grossen Stichproben die Residuen symmetrisch um 0 mit konstanter Varianz σ^2 verteilt sind. Dagegen ist für Zähldaten $(y_i - \mu_i)$ heteroskedastisch und asymmetrisch. Es gibt in diesen Fällen kein Residuum, das die gewünschten Eigenschaften wie Erwartungswert gleich 0, konstante Varianz und symmetrische Verteilung gleichzeitig erfüllt. Dies führt zu unterschiedlichen Definitionen von Residuen, je nachdem auf welche Eigenschaft am meisten Wert gelegt wird. Hier sollen aber nur die zwei wichtigsten vorgestellt werden:

(i) *Pearson-Residuum:*

Das Pearson-Residuum r_i^P ist durch

$$r_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}}$$

definiert. Dies ist das klassische Residuum, das noch mit der geschätzten Standardabweichung zur Korrektur der Heteroskedastizität skaliert wird. In grossen Stichproben hat dieses Residuum einen Erwartungswert 0 und ist homoskedastisch (mit Varianz 1), allerdings ist es asymmetrisch verteilt. Im Fall der Poissonverteilung gilt die folgende Beziehung zwischen der Pearson-Statistik und dem Pearson-Residuum:

$$\sum_{i=1}^n (r_i^P)^2 = \chi^2.$$

Durch diese Beziehung mit der Pearson-Statistik hat dieses Residuum auch seinen Namen erhalten.

(ii) *Devianz-Residuum:*

Wird die Devianz als Anpassungsmass in einem GLM verwendet, dann trägt jede Einheit einen bestimmten Teil d_i zum gesamten Mass bei, d.h. $\sum_{i=1}^n d_i = D$. Damit kann man nun das Devianz-Residuum bilden:

$$r_i^D = \text{sign}(y_i - \hat{\mu}_i) \sqrt{d_i}, \quad \text{mit } \text{sign}(x) = \begin{cases} -1 & \text{für } x < 0 \\ 0 & \text{für } x = 0 \\ 1 & \text{für } x > 0. \end{cases}$$

Dieses Mass wächst mit $y_i - \hat{\mu}_i$ und es gilt $\sum_{i=1}^n (r_i^D)^2 = D$. Für den Spezialfall der Poissonverteilung ergibt sich das Residuum zu

$$r_i^D = \text{sign}(y_i - \hat{\mu}_i) 2 \left[y_i \ln \frac{y_i}{\hat{\mu}_i} - (y_i - \hat{\mu}_i) \right]^{1/2}.$$

2.3 Die Negativ-Binomial-Verteilung

In den bisherigen Abschnitten wurde besonders auf die Poissonverteilung als Spezialfall der GLM's eingegangen, da diese Verteilung zur Modellierung von Zähldaten am häufigsten verwendet wird. Allerdings treten im Zusammenhang mit dieser Verteilung oftmals Schwierigkeiten bei der Modellbildung auf, da die für die Poissonverteilung charakteristische Eigenschaft der Äquidispersion, d.h. die Gleichheit von Erwartungswert und Varianz, durch die Daten in der Regel verletzt wird. Üblicherweise weisen die

Daten in der Praxis Überdispersion auf, d.h. ihre Variabilität ist höher als durch das Poissonmodell erwartet. Im Gegensatz zu anderen Modellen reicht hier die Verletzung der Äquidispersion schon aus, um die Poissonmodellannahme zu verletzen. Mögliche Ursachen der Überdispersion sind z.B., dass noch weitere Regressoren im Modell fehlen, die Linkfunktion könnte falsch gewählt worden sein, oder die zugrundeliegende Beobachtungseinheit (Zeit, Volumen, etc.) ist nicht fest, sondern variabel. Während die ersten beiden Ursachen in jedem beliebigen Regressionsmodell auftreten können, ist der letzte Punkt typisch für die Poissonverteilung. Eine Möglichkeit, mit dem Problem der Überdispersion umzugehen ist, bei der Annahme der Poissonverteilung zu bleiben, obwohl die Voraussetzung der Gleichheit von Erwartungswert und Varianz nicht erfüllt ist. Dies kann zu falschen Schätzungen der Teststatistiken und zu erhöhten optimistischen Schlussfolgerungen bezüglich der Signifikanz der Regressoren führen. Allerdings, solange hierbei der Erwartungswert richtig spezifiziert ist, sind die Schätzer laut Cameron & Trivedi (1998, S. 60) [2] immer noch konsistent. Schätzungen der Standardfehler jedoch müssen dann korrigiert werden.

Eine Alternative zur Poissonverteilung wäre, die Überdispersion mit Hilfe einer anderen, allgemeineren Verteilung zu modellieren, die keine Äquidispersion erfordert. Die in solchen Fällen verwendete Verteilung ist die Negativ-Binomial-Verteilung, deren Varianz eine quadratische Funktion des Erwartungswertes ist.

Im nächsten Unterabschnitt wird diese Verteilung zunächst definiert, es werden kurz einige wichtige Eigenschaften aufgezeigt und es wird erläutert, inwiefern diese Verteilung eine Alternative zur Poissonverteilung darstellt.

2.3.1 Die NB-Verteilung und ihre Eigenschaften

Im Gegensatz zur Poissonverteilung erlaubt die NB-Verteilung mehr Flexibilität in der Modellierung der Varianz und besteht nicht nur aus einem, sondern aus zwei Parametern. Zunächst wird die NB-Verteilung als diskrete Verteilung mit ganzzahligem Parameter r eingeführt, während später auf die Erweiterung mit stetigem Parameter eingegangen wird.

Definition 2.3.1 (NB-Verteilung mit $r \in \mathbb{N}$) Die Zufallsvariable Y heisst negativ-binomial-verteilt mit Parametern r und p , kurz $Y \sim NB(r, p)$, falls sich ihre Wahrscheinlichkeitsfunktion wie folgt darstellen lässt:

$$P(Y = y) = \binom{y + r - 1}{y} p^r (1 - p)^y, \quad y = 0, 1, 2, \dots, \quad r \in \mathbb{N}, \quad 0 < p < 1.$$

Ähnlich wie die Binomial- und auch die Poissonverteilung lässt sich die NB-Verteilung in dieser Form leicht anschaulich interpretieren. Bei einer Zufallsvariablen Y , die $NB(r, p)$ -verteilt ist, bedeutet $Y = y$ in einer Reihe von unabhängigen Bernoulli-Versuchen mit Erfolgswahrscheinlichkeit p die Anzahl der Misserfolge vor dem r -ten Erfolg.

Beispiel 2.3.2 *Eine defekte Pistole schiesst in 5 von 6 Fällen nicht. Die Wahrscheinlichkeit, dass es z.B. 10 Fehlversuche vor dem 3. Erfolg gibt, ist dann gegeben durch*

$$P(Y = 10) = \binom{10 + 3 - 1}{10} \left(\frac{1}{6}\right)^3 \left(1 - \frac{1}{6}\right)^{10}.$$

Y wäre in diesem Fall $NB(3, \frac{1}{6})$ -verteilt.

In der Abbildung 2.1 sind Wahrscheinlichkeitsfunktionen von NB-Verteilungen abgebildet mit jeweils unterschiedlichen Parametern. Man erkennt, dass je kleiner die Erfolgswahrscheinlichkeit p ist, desto weiter ist die Funktion nach rechts verschoben, da mehr Versuche benötigt werden, bis wieder ein Erfolgseignis eintritt.

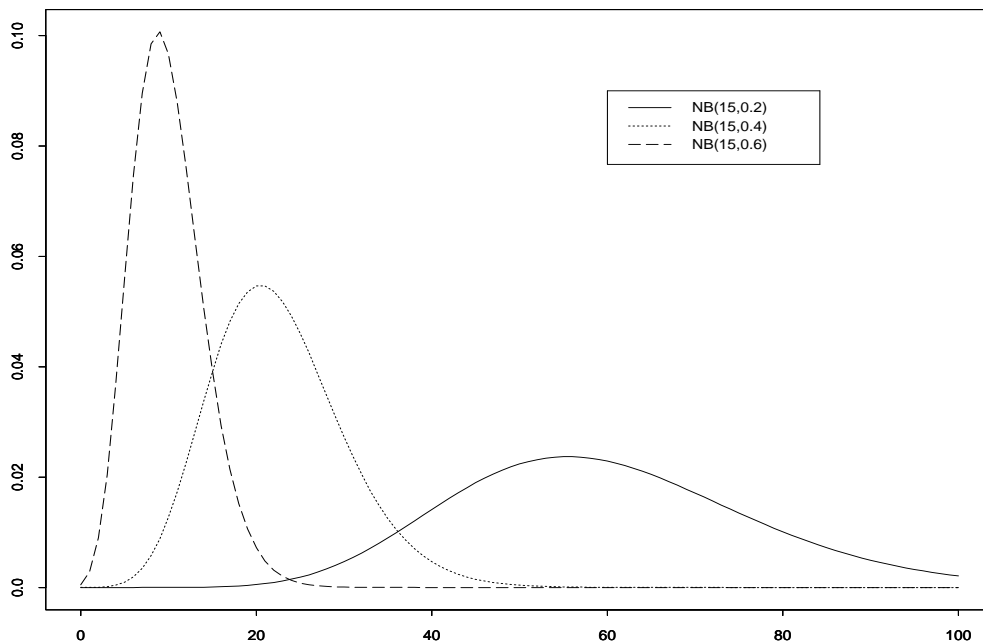


Abbildung 2.1: Verschiedene NB-Verteilungen

Satz 2.3.3 (Momenterzeugende Funktion der NB-Verteilung) *Die momenterzeugende Funktion der NB-Verteilung lautet*

$$M_Y(t) = E(e^{tY}) = \left(\frac{p}{1 - (1-p)e^t} \right)^r, \quad t < |\ln(1-p)|.$$

Beweis: Am einfachsten sieht man den Beweis ein, wenn man zunächst den Spezialfall $r=1$ betrachtet, d.h. man beobachtet in einer Reihe von unabhängigen Bernoulli-Versuchen die Anzahl der Fehlversuche vor dem ersten Erfolg. Dies entspricht genau der geometrischen Verteilung:

$$P(X = x) = p(1-p)^x.$$

Die momenterzeugende Funktion der geometrischen Verteilung lässt sich folgendermaßen berechnen:

$$\begin{aligned} M_X(t) = E(e^{tX}) &= \sum_{x=0}^{\infty} p(1-p)^x e^{tx} \\ &= p \sum_{x=0}^{\infty} ((1-p)e^t)^x \\ &= \frac{p}{1 - (1-p)e^t}. \end{aligned}$$

Nun kann eine negativ-binomial-verteilte Zufallsvariable Y mit Parameter r durch $Y = \sum_{i=1}^r X_i$ aus r unabhängigen geometrisch verteilten Zufallsvariablen X_i gebildet werden. Dies soll hier für $r = 2$ gezeigt werden. Seien dazu X_1 und X_2 zwei geometrisch verteilte Zufallsvariablen mit Erfolgswahrscheinlichkeitsparameter p . Dann gilt nach Kredler (1998, S. 70) [12] für die Wahrscheinlichkeitsfunktion der Zufallsvariable $Y = X_1 + X_2$

$$\begin{aligned} f_{X_1+X_2}(y) &= \sum_{x_1} f_{X_1}(x_1) f_{X_2}(y - x_1) \\ &= \sum_{i=0}^y P(X_1 = i) P(X_2 = y - i) \\ &= \sum_{i=0}^y p(1-p)^i p(1-p)^{y-i} = \sum_{i=0}^y p^2(1-p)^y \\ &= (y+1)p^2(1-p)^y = \binom{y+2-1}{y} p^2(1-p)^y, \end{aligned}$$

was der Wahrscheinlichkeitsfunktion der NB-Verteilung mit Parametern $r = 2$ und p gleicht. Die obige Behauptung kann dann wegen der Unabhängigkeit und der Gedächtnis-

nislosigkeit der geometrischen Verteilung leicht auch für $r > 2$ Zufallsvariablen nachvollzogen werden. Damit ergibt sich aber für die momenterzeugende Funktion der NB-Verteilung

$$M_Y(t) = M_{\sum_{i=1}^r X_i}(t) = \prod_{i=1}^r M_X(t) = \left(\frac{p}{1 - (1-p)e^t} \right)^r.$$

Die Produktformel für die momenterzeugende Funktion darf angewendet werden, da die X_i unabhängig vorausgesetzt werden.

□

Mit Hilfe der momenterzeugenden Funktion können nun aus den Formeln (2.1) und (2.2) der Erwartungswert und die Varianz berechnet werden.

$$\begin{aligned} E(Y) = \frac{dM(0)}{dt} &= r \left(\frac{p}{1 - (1-p)e^t} \right)^{r-1} \frac{p(1-p)e^t}{(1 - (1-p)e^t)^2} \Big|_{t=0} \\ &= r \left(\frac{p}{1 - (1-p)} \right)^{r-1} \frac{p(1-p)}{(1 - (1-p))^2} \\ &= r \frac{p(1-p)}{p^2} \\ &= r \frac{1-p}{p}. \end{aligned}$$

Für die Varianz gilt

$$\begin{aligned} \text{Var}(Y) &= \frac{d^2 M(0)}{dt^2} - \left(\frac{dM(0)}{dt} \right)^2 \\ &= \frac{d}{dt} \left[r \left(\frac{p}{1 - (1-p)e^t} \right)^{r-1} \frac{p(1-p)e^t}{(1 - (1-p)e^t)^2} \right] \Big|_{t=0} - \left(r \frac{1-p}{p} \right)^2 \\ &= r(r-1) \left(\frac{p}{1 - (1-p)e^t} \right)^{r-2} \left(\frac{p(1-p)e^t}{(1 - (1-p)e^t)^2} \right)^2 \Big|_{t=0} + r \left(\frac{p}{1 - (1-p)e^t} \right)^{r-1} \\ &\quad \cdot \frac{p(1-p)e^t(1 - (1-p)e^t)^2 + p(1-p)e^t 2(1 - (1-p)e^t)(1-p)e^t}{(1 - (1-p)e^t)^4} \Big|_{t=0} \\ &\quad - \left(r \frac{1-p}{p} \right)^2 \\ &= r(r-1) \left(\frac{p(1-p)}{p^2} \right)^2 + r \frac{p(1-p)p^2 + p(1-p)2p(1-p)}{p^4} - \frac{r^2(1-p)^2}{p^2} \\ &= \frac{r(r-1)(1-p)^2 + rp(1-p) + 2r(1-p)^2 - r^2(1-p)^2}{p^2} \\ &= \frac{r^2(1-p)^2 - r(1-p)^2 + rp(1-p) + 2r(1-p)^2 - r^2(1-p)^2}{p^2} \end{aligned}$$

$$\begin{aligned}
&= r(1-p) \frac{p + (1-p)}{p^2} \\
&= r \frac{1-p}{p^2}.
\end{aligned}$$

Zusammenfassend gelten also für negativ-binomial-verteilte Zufallsvariablen Y die Formeln

$$E(Y) = r \frac{1-p}{p} \quad \text{und} \quad \text{Var}(Y) = r \frac{1-p}{p^2}. \quad (2.13)$$

An diesen Werten kann man sofort erkennen, dass die Varianz immer grösser als der Erwartungswert ist, da $0 < p < 1$. Aus diesem Grund eignet sich die NB-Verteilung besonders gut zur Modellierung von Daten, die Überdispersion aufweisen.

Als weitere Eigenschaft kann analog zur Poissonverteilung ein Additivitätssatz angegeben werden, der nachfolgend erläutert wird.

Satz 2.3.4 (Additivität der NB-Verteilung) *Für die Zufallsvariable $Z = X + Y$ mit $X \sim NB(r, p)$ und $Y \sim NB(s, p)$ und X und Y unabhängig gilt:*

$$Z = X + Y \sim NB(r + s, p).$$

Beweis: *Da die Zufallsvariablen X und Y unabhängig sind, gilt für die momenterzeugende Funktion*

$$\begin{aligned}
M_Z(t) &= M_{X+Y}(t) = M_X(t) M_Y(t) \\
&= \left(\frac{p}{1 - (1-p)e^t} \right)^r \left(\frac{p}{1 - (1-p)e^t} \right)^s = \left(\frac{p}{1 - (1-p)e^t} \right)^{r+s},
\end{aligned}$$

was gerade die momenterzeugende Funktion einer $NB(r + s, p)$ -verteilten Zufallsvariable ist.

□

Ähnlich wie für die Poissonverteilung gibt es auch hier eine Vielzahl von verschiedenen Charakterisierungsmöglichkeiten. Die mitunter wichtigste und geläufigste ist die Darstellung der NB-Verteilung als eine Poisson-Gamma-Mischung, worauf im anschließenden Unterabschnitt noch näher eingegangen wird. Eggenberger und Polya (1923) fanden eine noch andere Methode. Ihre Idee beruht auf der Charakterisierung der NB-Verteilung als ein Grenzwert eines Urnenmodells. In Johnson & Kotz (1993, S. 205) [10] ist dieser Zusammenhang folgendermassen beschrieben:

Satz 2.3.5 (NB-Verteilung als Grenzwert der Polya-Verteilung) *Gegeben sei eine Urne, die Np weiße und $N(1 - p)$ schwarze Kugeln enthält. Nun werde eine Stichprobe der Länge n gezogen, wobei nach jedem einzelnen Zug die jeweils gezogene Kugel wieder in die Urne zurückgelegt wird und $c = N\beta$ Kugeln derselben Farbe wie die gezogene Kugel in die Urne hinzugefügt werden. Y sei dann die Anzahl der weißen Kugeln in n Zügen. Die Verteilung von Y folgt der Polya-Verteilung, die durch*

$$P(Y = y) = \binom{n}{y} \left(\frac{p}{\beta}\right)^{[y]} \left(\frac{q}{\beta}\right)^{[n-y]} / \left(\frac{1}{\beta}\right)^{[n]}$$

gegeben ist, wobei $a^{[y]} = a(a+1)\dots(a+y-1)$ abkürzt, und $q = 1 - p$ ist. Damit lässt sich der obige Ausdruck umschreiben in die gängige Form

$$\begin{aligned} P(Y = y) &= \binom{n}{y} \frac{\left(\frac{p}{\beta}\right) \left(\frac{p}{\beta} + 1\right) \cdots \left(\frac{p}{\beta} + y - 1\right) \left(\frac{q}{\beta}\right) \left(\frac{q}{\beta} + 1\right) \cdots \left(\frac{q}{\beta} + n - y - 1\right)}{\left(\frac{1}{\beta}\right) \left(\frac{1}{\beta} + 1\right) \cdots \left(\frac{1}{\beta} + n - 1\right)} \\ &= \binom{n}{y} \frac{p(p+\beta) \cdots (p+y\beta-\beta) q(q+\beta) \cdots (q+n\beta-y\beta-\beta)}{1(1+\beta) \cdots (1+n\beta-\beta)}, \end{aligned}$$

für $y = 0, 1, \dots, n$.

Betrachtet man nun den Grenzfall für $n \rightarrow \infty$, $p \rightarrow 0$, $\beta \rightarrow 0$ mit $np = \eta k$ und $n\beta = \eta$, dann ergibt sich daraus die NB-Verteilung

$$P(Y = y) = \binom{k+y-1}{k-1} \left(\frac{\eta}{1+\eta}\right)^y \left(1 - \frac{\eta}{1+\eta}\right)^k,$$

mit den Parametern k und $p = 1 - \frac{\eta}{1+\eta} = \frac{1}{1+\eta}$.

Beweis: Einen Beweis hierfür kann man in Boswell and Patil, 1970 oder in Feller (1968, S.118-145) [8] nachlesen.

□

In der Literatur wird in der Regel nur die bisher angegebene Form der NB-Verteilung mit ganzzahligem Parameter r und p angesprochen, doch häufig wird bei der Verwendung von NB-Modellen eine Erweiterung erwünscht, die nicht nur ganzzahlige, sondern auch stetige Parameter zulässt. Allerdings hat die erweiterte Form dann den Nachteil, dass sie sich nicht mehr anschaulich interpretieren lässt, wie das bisher der Fall war. Trotz des stetigen Parameters gehört aber die erweiterte Form der NB-Verteilung immer noch zu der Klasse der diskreten Verteilungen, da die Zufallsvariable Y weiterhin nur diskrete Werte annimmt.

Definition 2.3.6 (Negativ-Binomial-Verteilung, erweiterte Form) Die ZV Y heisst negativ-binomial-verteilt, falls ihre Wahrscheinlichkeitsfunktion folgende Form hat:

$$\begin{aligned} P(Y = y) &= \frac{\Gamma(y + a^{-1})}{\Gamma(a^{-1})y!} \left(\frac{a\mu}{1 + a\mu} \right)^y \left(\frac{1}{1 + a\mu} \right)^{a^{-1}} \\ &= \frac{\Gamma(y + a^{-1})}{\Gamma(a^{-1})y!} \left(\frac{\mu}{a^{-1} + \mu} \right)^y \left(\frac{a^{-1}}{a^{-1} + \mu} \right)^{a^{-1}} \end{aligned}$$

für $y = 0, 1, 2, \dots$. Der Parameter $a \geq 0$ wird hierbei oft mit Dispersionsindex bezeichnet. Eine andere, auch sehr häufig auftretende Parametrisierung der Wahrscheinlichkeitsfunktion benutzt statt a^{-1} den Parameter $k = a^{-1}$. Hier soll aber im folgenden die erstere Parametrisierung verwendet werden. Als Abkürzung für die Negativ-Binomial-Verteilung wird nun $Y \sim NB(\mu, a)$ geschrieben. Um zu zeigen, dass die angegebene Funktion auch wirklich eine Wahrscheinlichkeitsfunktion ist, muss bewiesen werden, dass sich die Summe über Y zu 1 aufaddiert.

Beweis:

$$\begin{aligned} \lim_{y \rightarrow \infty} P(Y \leq y) &= \sum_{y=0}^{\infty} \frac{\Gamma(y + a^{-1})}{\Gamma(a^{-1})y!} \left(\frac{a\mu}{1 + a\mu} \right)^y \left(\frac{1}{1 + a\mu} \right)^{a^{-1}} \\ &\stackrel{(*)}{=} \frac{1}{\Gamma(a^{-1})} \left(\frac{1}{1 + a\mu} \right)^{a^{-1}} \sum_{y=0}^{\infty} \int_0^{\infty} s^{y+a^{-1}-1} e^{-s} ds \frac{1}{y!} \left(\frac{a\mu}{1 + a\mu} \right)^y \\ &\stackrel{(**)}{=} \frac{1}{\Gamma(a^{-1})} \left(\frac{1}{1 + a\mu} \right)^{a^{-1}} \int_0^{\infty} \sum_{y=0}^{\infty} s^{y+a^{-1}-1} e^{-s} \frac{1}{y!} \left(\frac{a\mu}{1 + a\mu} \right)^y ds \\ &= \frac{1}{\Gamma(a^{-1})} \left(\frac{1}{1 + a\mu} \right)^{a^{-1}} \int_0^{\infty} s^{a^{-1}-1} e^{-s} \sum_{y=0}^{\infty} \left(s \frac{a\mu}{1 + a\mu} \right)^y \frac{1}{y!} ds \\ &= \frac{1}{\Gamma(a^{-1})} \left(\frac{1}{1 + a\mu} \right)^{a^{-1}} \int_0^{\infty} s^{a^{-1}-1} e^{-s} e^{s \frac{a\mu}{1 + a\mu}} ds \\ &= \frac{1}{\Gamma(a^{-1})} \left(\frac{1}{1 + a\mu} \right)^{a^{-1}} \int_0^{\infty} s^{a^{-1}-1} e^{-s(1 - \frac{a\mu}{1 + a\mu})} ds \\ &= \frac{1}{\Gamma(a^{-1})} \left(\frac{1}{1 + a\mu} \right)^{a^{-1}} \int_0^{\infty} s^{a^{-1}-1} e^{-s(\frac{1}{1 + a\mu})} ds \\ &\stackrel{(***)}{=} \frac{1}{\Gamma(a^{-1})} \left(\frac{1}{1 + a\mu} \right)^{a^{-1}} \frac{\Gamma(a^{-1})}{\left(\frac{1}{1 + a\mu} \right)^{a^{-1}}} = 1. \end{aligned}$$

(*): Die Gammafunktion ist definiert als

$$\Gamma(x) = \int_0^{\infty} s^{x-1} e^{-s} ds.$$

(**): Nach dem Satz von der majorisierten Konvergenz (A. Leutbecher, Analysis-Skriptum 3, 1998/99, S. 123) gilt:

Es seien f_n und f auf $U \in \mathbb{R}^n$ definiert und lokal Riemann-integrierbar und es gelte $\forall x \in U$:

$$\lim_{n \rightarrow \infty} f_n(x) = f(x).$$

Sei weiter g auf U nicht negativ und uneigentlich Riemann-integrierbar mit $|f_n| < g \forall n \in \mathbb{N}$. Dann existieren die folgenden uneigentlichen Integrale und es gilt zwischen ihnen die folgende Relation

$$\lim_{n \rightarrow \infty} \int_U f_n(x) dx = \int_U f(x) dx.$$

Hier sei nun $U := [0, \infty)$, $f_n(s) = \sum_{y=0}^n s^{y+a^{-1}-1} e^{-s} \frac{1}{y!} \left(\frac{a\mu}{1+a\mu} \right)^y$ und

$$\begin{aligned} f(s) &= \lim_{n \rightarrow \infty} f_n(s) = \lim_{n \rightarrow \infty} \sum_{y=0}^n s^{y+a^{-1}-1} e^{-s} \frac{1}{y!} \left(\frac{a\mu}{1+a\mu} \right)^y \\ &= s^{a^{-1}-1} e^{-s} \lim_{n \rightarrow \infty} \sum_{y=0}^n \left(s \frac{a\mu}{1+a\mu} \right)^y \frac{1}{y!} \\ &= s^{a^{-1}-1} e^{-s} e^{s \frac{a\mu}{1+a\mu}} = s^{a^{-1}-1} e^{-s \left(\frac{1}{1+a\mu} \right)}. \end{aligned}$$

Da die Summanden von $f_n(s)$ alle positiv sind, ist $f_n(s)$ monoton steigend und es gilt $f_1 \leq f_2 \leq \dots \leq f$, woraus $|f_n| \leq f \forall n \in \mathbb{N}$ folgt. Die Grenzfunktion f ist ausserdem uneigentlich Riemann integrierbar, so dass die Voraussetzungen für den Satz von der majorisierten Konvergenz erfüllt sind und der Satz angewendet werden kann. Damit gilt schliesslich

$$\begin{aligned} \int_0^\infty f(s) &= \lim_{n \rightarrow \infty} \int_0^\infty f_n(s) ds \\ &= \lim_{n \rightarrow \infty} \int_0^\infty \sum_{y=0}^n s^{y+a^{-1}-1} e^{-s} \frac{1}{y!} \left(\frac{a\mu}{1+a\mu} \right)^y ds \\ &\stackrel{n \text{ endlich}}{=} \lim_{n \rightarrow \infty} \sum_{y=0}^n \int_0^\infty s^{y+a^{-1}-1} e^{-s} \frac{1}{y!} \left(\frac{a\mu}{1+a\mu} \right)^y ds \\ &= \sum_{y=0}^\infty \int_0^\infty s^{y+a^{-1}-1} e^{-s} \frac{1}{y!} \left(\frac{a\mu}{1+a\mu} \right)^y ds, \end{aligned}$$

womit die Vertauschbarkeit von Summation und Integration gezeigt ist.

(***) : Laut Springers Formelsammlung (1997, S. 177) [17] gilt die Formel

$$\int_0^\infty x^n e^{-ax} dx = \frac{\Gamma(n+1)}{a^{n+1}}, \quad n > -1, \quad a > 0.$$

□

Um den Erwartungswert und die Varianz für die erweiterte Form zu erhalten, wird zunächst analog wie für die vereinfachte Darstellung die momenterzeugende Funktion hergeleitet.

Satz 2.3.7 (Momenterzeugende Funktion der NB-Verteilung, erweiterte Form)

Die momenterzeugende Funktion für die NB-Verteilung aus Definition 2.3.6 ist gegeben durch

$$M_Y(t) = \left(\frac{\frac{a^{-1}}{a^{-1}+\mu}}{1 - (1 - \frac{a^{-1}}{a^{-1}+\mu})e^t} \right)^{a^{-1}}.$$

Beweis: Bei diesem Beweis werden analog zum Beweis der Wahrscheinlichkeitsfunktion in Definition 2.3.6 die Umformungen der Gammafunktion in eine Integraldarstellung, die Vertauschbarkeit von Summation und Integration und die Integralformel nach Springers Formelsammlung (1997, S. 177) [17] verwendet.

$$\begin{aligned} M_Y(t) &= E(e^{tY}) = \sum_{y=0}^{\infty} e^{ty} \frac{\Gamma(y+a^{-1})}{\Gamma(a^{-1})y!} \left(\frac{\mu}{a^{-1}+\mu} \right)^y \left(\frac{a^{-1}}{a^{-1}+\mu} \right)^{a^{-1}} \\ &= \frac{1}{\Gamma(a^{-1})} \left(\frac{a^{-1}}{a^{-1}+\mu} \right)^{a^{-1}} \sum_{y=0}^{\infty} \int_0^{\infty} s^{y+a^{-1}-1} e^{-s} ds \frac{1}{y!} \left(\frac{\mu}{a^{-1}+\mu} \right)^y e^{ty} \\ &= \frac{1}{\Gamma(a^{-1})} \left(\frac{a^{-1}}{a^{-1}+\mu} \right)^{a^{-1}} \int_0^{\infty} e^{-s} s^{a^{-1}-1} \sum_{y=0}^{\infty} \left(se^t \frac{\mu}{a^{-1}+\mu} \right)^y \frac{1}{y!} ds \\ &= \frac{1}{\Gamma(a^{-1})} \left(\frac{a^{-1}}{a^{-1}+\mu} \right)^{a^{-1}} \int_0^{\infty} e^{-s} s^{a^{-1}-1} e^{se^t \frac{\mu}{a^{-1}+\mu}} ds \\ &= \frac{1}{\Gamma(a^{-1})} \left(\frac{a^{-1}}{a^{-1}+\mu} \right)^{a^{-1}} \int_0^{\infty} s^{a^{-1}-1} e^{-s(1-e^t \frac{\mu}{a^{-1}+\mu})} ds \\ &= \frac{1}{\Gamma(a^{-1})} \left(\frac{a^{-1}}{a^{-1}+\mu} \right)^{a^{-1}} \frac{\Gamma(a^{-1})}{(1 - e^t \frac{\mu}{a^{-1}+\mu})^{a^{-1}}} \\ &= \left(\frac{\frac{a^{-1}}{a^{-1}+\mu}}{1 - (1 - \frac{a^{-1}}{a^{-1}+\mu})e^t} \right)^{a^{-1}}. \end{aligned}$$

□

Aus der momenterzeugenden Funktion werden nun Erwartungswert und Varianz nach den bekannten Formeln (2.1) und (2.2) berechnet.

$$\begin{aligned} E(Y) &= \frac{dM(0)}{dt} \\ &= a^{-1} \left(\frac{\frac{a^{-1}}{a^{-1}+\mu}}{1 - (1 - \frac{a^{-1}}{a^{-1}+\mu})e^t} \right)^{a^{-1}-1} \frac{\frac{a^{-1}}{a^{-1}+\mu} \left(1 - \frac{a^{-1}}{a^{-1}+\mu} \right) e^t}{\left(1 - (1 - \frac{a^{-1}}{a^{-1}+\mu})e^t \right)^2} \Big|_{t=0} \end{aligned}$$

$$\begin{aligned}
&= a^{-1} \frac{1 - \frac{a^{-1}}{a^{-1} + \mu}}{\frac{a^{-1}}{a^{-1} + \mu}} = \frac{\mu}{a^{-1} + \mu} (a^{-1} + \mu) = \mu \quad \text{und} \tag{2.14} \\
\text{Var}(Y) &= \frac{d^2 M(0)}{dt^2} - \frac{dM(0)}{dt} \\
&= a^{-1} (a^{-1} - 1) \left(\frac{\frac{a^{-1}}{a^{-1} + \mu}}{1 - (1 - \frac{a^{-1}}{a^{-1} + \mu}) e^t} \right)^{a^{-1} - 2} \left(\frac{\frac{a^{-1}}{a^{-1} + \mu} (1 - \frac{a^{-1}}{a^{-1} + \mu}) e^t}{\left(1 - (1 - \frac{a^{-1}}{a^{-1} + \mu}) e^t\right)^2} \right)^2 \Big|_{t=0} \\
&\quad + a^{-1} \left(\frac{\frac{a^{-1}}{a^{-1} + \mu}}{1 - (1 - \frac{a^{-1}}{a^{-1} + \mu}) e^t} \right)^{a^{-1} - 1} \left[\frac{\frac{a^{-1}}{a^{-1} + \mu} \left(1 - \frac{a^{-1}}{a^{-1} + \mu}\right) e^t \left(1 - (1 - \frac{a^{-1}}{a^{-1} + \mu}) e^t\right)^2}{\left(1 - (1 - \frac{a^{-1}}{a^{-1} + \mu}) e^t\right)^4} \right. \\
&\quad \left. + \frac{\frac{a^{-1}}{a^{-1} + \mu} \left(1 - \frac{a^{-1}}{a^{-1} + \mu}\right) e^t 2 \left(1 - (1 - \frac{a^{-1}}{a^{-1} + \mu}) e^t\right) \left(1 - \frac{a^{-1}}{a^{-1} + \mu}\right) e^t}{\left(1 - (1 - \frac{a^{-1}}{a^{-1} + \mu}) e^t\right)^4} \right] \Big|_{t=0} - \mu^2 \\
&= a^{-1} (a^{-1} - 1) \left(\frac{1 - \frac{a^{-1}}{a^{-1} + \mu}}{\frac{a^{-1}}{a^{-1} + \mu}} \right)^2 \\
&\quad + a^{-1} \frac{\left(\frac{a^{-1}}{a^{-1} + \mu}\right)^3 \left(1 - \frac{a^{-1}}{a^{-1} + \mu}\right) + 2 \left(\frac{a^{-1}}{a^{-1} + \mu}\right)^2 \left(1 - \frac{a^{-1}}{a^{-1} + \mu}\right)^2}{\left(\frac{a^{-1}}{a^{-1} + \mu}\right)^4} - \mu^2 \\
&= \frac{(a^{-1})^2 \left(1 - \frac{a^{-1}}{a^{-1} + \mu}\right)^2 - a^{-1} \left(1 - \frac{a^{-1}}{a^{-1} + \mu}\right)^2 + a^{-1} \left(\frac{a^{-1}}{a^{-1} + \mu}\right) \left(1 - \frac{a^{-1}}{a^{-1} + \mu}\right)}{\left(\frac{a^{-1}}{a^{-1} + \mu}\right)^2} \\
&\quad + \frac{2a^{-1} \left(1 - \frac{a^{-1}}{a^{-1} + \mu}\right)^2 - \mu^2 \left(\frac{a^{-1}}{a^{-1} + \mu}\right)^2}{\left(\frac{a^{-1}}{a^{-1} + \mu}\right)^2} \\
&= \frac{a^{-1} \left(\frac{a^{-1}}{a^{-1} + \mu}\right) \left(1 - \frac{a^{-1}}{a^{-1} + \mu}\right) + a^{-1} \left(1 - \frac{a^{-1}}{a^{-1} + \mu}\right)^2}{\left(\frac{a^{-1}}{a^{-1} + \mu}\right)^2} \\
&= a^{-1} \left(1 - \frac{a^{-1}}{a^{-1} + \mu}\right) \frac{\frac{a^{-1}}{a^{-1} + \mu} + 1 - \frac{a^{-1}}{a^{-1} + \mu}}{\left(\frac{a^{-1}}{a^{-1} + \mu}\right)^2} \\
&= \frac{a^{-1} \left(1 - \frac{a^{-1}}{a^{-1} + \mu}\right)}{\left(\frac{a^{-1}}{a^{-1} + \mu}\right)^2} = \frac{a^{-1} \mu}{a^{-1} + \mu} \frac{(a^{-1} + \mu)^2}{(a^{-1})^2} \\
&= \frac{\mu(a^{-1} + \mu)}{a^{-1}} = \mu + a\mu^2. \tag{2.15}
\end{aligned}$$

Für den Spezialfall, dass der nun stetige Parameter a diskrete Werte annimmt, kann durch leichtes Umformen die Äquivalenz beider Darstellungsformen der NB-Verteilung gezeigt werden. Im Vergleich zur ersten Darstellung wird hier statt des Parameters r der Parameter a^{-1} verwendet und p wird durch den Bruch $\frac{a^{-1}}{a^{-1}+\mu}$ ersetzt. Schliesslich bleibt zu zeigen, dass der Binomialkoeffizient aus der ersteren Darstellung dem Bruch der Gammafunktionen in der erweiterten Form entspricht. Es gilt

$$\frac{\Gamma(y + a^{-1})}{\Gamma(a^{-1})y!} = \frac{(y + a^{-1} - 1)!}{(a^{-1} - 1)!y!} = \binom{y + a^{-1} - 1}{y}, \text{ für } a^{-1} \in \mathbb{N},$$

womit die Äquivalenz beider Formen gezeigt ist.

Zusammenfassend lässt sich sagen, dass die NB-Verteilung ebenso zur Erstellung von Zähldatenmodellen geeignet ist wie die Poissonverteilung, nur mit dem Vorteil, dass die NB-Verteilung in der Modellierung der Varianz wesentlich mehr Flexibilität zulässt. Wie eng beide Modelle miteinander zusammenhängen, kann man daran erkennen, dass die NB-Verteilung die Poissonverteilung als Spezialfall enthält. Wird nämlich der Dispersionsparameter $a = 0$ gewählt, was gleichbedeutend damit ist, dass keine Überdispersion vorliegt, so ergibt sich daraus die Poissonverteilung.

Satz 2.3.8 (Poissonverteilung als Grenzfall der NB-Verteilung) *Sei Y eine ZV, die negativ-binomial-verteilt ist mit Parametern μ und $a = 0$. Dann ergibt sich daraus die Poissonverteilung mit Parameter μ .*

Beweis:

$$\begin{aligned} P(Y = y) &= \frac{\Gamma(y + a^{-1})}{\Gamma(a^{-1})y!} \left(\frac{a\mu}{1 + a\mu} \right)^y \left(\frac{1}{1 + a\mu} \right)^{a^{-1}} \\ &\stackrel{\alpha=a^{-1}}{=} \frac{\Gamma(y + \alpha)}{\Gamma(\alpha)y!} \left(\frac{\frac{1}{\alpha}\mu}{1 + \frac{1}{\alpha}\mu} \right)^y \left(\frac{1}{1 + \frac{1}{\alpha}\mu} \right)^{\alpha} \\ &= \frac{\Gamma(y + \alpha)}{\Gamma(\alpha)y!} \left(\frac{\mu}{\alpha + \mu} \right)^y \left(\frac{\alpha}{\alpha + \mu} \right)^{\alpha} \\ &\stackrel{(*)}{=} \left(\prod_{j=0}^{y-1} (j + \alpha) \right) \frac{1}{y!} \left(\frac{1}{\alpha + \mu} \right)^y \mu^y \left(\frac{\alpha}{\alpha + \mu} \right)^{\alpha} \\ &\stackrel{(**)}{=} \left(\prod_{j=0}^{y-1} \frac{j + \alpha}{\alpha + \mu} \right) \frac{1}{y!} \mu^y \left(\frac{\alpha}{\alpha + \mu} \right)^{\alpha} \\ &= \left(\prod_{j=0}^{y-1} \frac{\frac{j}{\alpha} + 1}{1 + \frac{\mu}{\alpha}} \right) \left(\frac{1}{1 + \frac{\mu}{\alpha}} \right)^{\alpha} \frac{\mu^y}{y!} \\ \lim_{a \rightarrow 0} P(Y = y) &= \lim_{\alpha \rightarrow \infty} P(Y = y) \stackrel{(***)}{=} 1 e^{-\mu} \frac{\mu^y}{y!}. \end{aligned}$$

Dies ist die Wahrscheinlichkeitsfunktion der Poissonverteilung mit Parameter μ .

$$(*) \quad \frac{\Gamma(y + \alpha)}{\Gamma(\alpha)} = \frac{(y + \alpha - 1)!}{(\alpha - 1)!} = (y + \alpha - 1)(y + \alpha - 2) \dots (y + \alpha - y) = \prod_{j=0}^{y-1} (j + \alpha)$$

$$(**) \quad \left(\frac{1}{\alpha + \mu} \right)^y = \underbrace{\left(\frac{1}{\alpha + \mu} \right) \left(\frac{1}{\alpha + \mu} \right) \dots \left(\frac{1}{\alpha + \mu} \right)}_{y\text{-mal}} = \prod_{j=0}^{y-1} \left(\frac{1}{\alpha + \mu} \right)$$

$$(* *) \quad \left(\prod_{j=0}^{y-1} \frac{\frac{j}{\alpha} + 1}{1 + \frac{\mu}{\alpha}} \right) \xrightarrow{\alpha \rightarrow \infty} 1, \quad \text{und es gilt}$$

$$(* *) \quad \left(\frac{1}{1 + \frac{\mu}{\alpha}} \right)^\alpha = \left(\frac{\alpha}{\alpha + \mu} \right)^\alpha = \left(1 + \frac{\mu}{\alpha} \right)^{-\alpha} \xrightarrow{\alpha \rightarrow \infty} e^{-\mu}.$$

□

2.3.2 Unbeobachtete Heterogenität

Bisher wurde die Überdispersion als Grund für das Scheitern der Poissonregression bei der Modellierung von Zähldaten schon genannt. Nun soll untersucht werden, durch welche Ursachen diese entstehen kann, wobei hier nur auf die wichtigsten Gründe eingegangen wird.

Wie zu Beginn dieses Abschnittes erläutert, kann Überdispersion daher rühren, dass z.B. noch relevante Regressoren fehlen, oder es kann der für die Poissonverteilung typische Fall eintreten, dass die sogenannte „*unbeobachtete Heterogenität*“ vorliegt. Um diese zu modellieren, wählt man den einzigen Parameter der Verteilung nicht fest, sondern zufällig, d.h. man fügt im bedingten Erwartungswert der Poissonverteilung noch einen multiplikativen Fehlerterm ein, um zufällige Variation zu berücksichtigen. Dies führt zu der Familie der gemischten Modelle. In einer Poissonregression ohne Heterogenität wird die Dichte bzw. die Wahrscheinlichkeitsfunktion von $(Y_i \mid \mathbf{x}_i)$ bedingt durch die Kovariablenvektoren \mathbf{x}_i spezifiziert. Dies ist gleichbedeutend mit der Spezifizierung des bedingten Erwartungswertes als eine nichtstochastische Funktion von \mathbf{x}_i . In gemischten Modellen dagegen wird die Dichte von $(Y_i \mid \mathbf{x}_i, \nu_i)$ spezifiziert, wobei ν_i einen unbeobachteten Heterogenitätsterm für die i -te Beobachtung darstellt. Man nimmt also an, dass die einzelnen Beobachtungen in einer Weise zufällig variieren, die nicht vollständig durch die beobachteten Kovariablen erklärt werden kann.

Obwohl eine multiplikative Heterogenitätsannahme sehr speziell ist, ist sie mathematisch gesehen attraktiver als ein additiver Fehlerterm, der zu einer Verletzung der

Nichtnegativitätsannahme der Y_i führen könnte. Deshalb werden in der Regel die Y_i und (\mathbf{x}_i, ν_i) über den exponentiellen Erwartungswert mit multiplikativem Fehlerterm miteinander verbunden, d.h.

$$E(Y_i | \mathbf{x}_i, \nu_i) = \exp(\mathbf{x}_i^t \boldsymbol{\beta}) \nu_i = \mu_i \nu_i.$$

Dabei wird vorausgesetzt, dass die ν_i unabhängig von den Regressoren \mathbf{x}_i und iid-verteilt sind, und es gilt $\nu_i = \exp(\varepsilon_i)$, wobei ε_i den Fehlerterm des Modells, wie in (2.3) beschrieben, für die i -te Beobachtung angibt. Mit den obigen Bezeichnungen kann nun gezeigt werden, dass unbeobachtete Heterogenität zu Überdispersion führt, vorausgesetzt, dass $\text{Var}(\nu_i) > 0$ gilt. Geht man also von der Poissonverteilung aus, so gilt

$$(Y_i | \mathbf{x}_i, \nu_i) \sim \text{Poi}(\lambda_i) \quad \text{mit } \lambda_i = \exp(\mathbf{x}_i^t \boldsymbol{\beta}) \nu_i = \mu_i \nu_i$$

und ν_i iid und unabhängig von Y_i . Der Parameter λ_i der Poissonverteilung ist hier nun nicht mehr ein fester Wert, sondern selbst eine Zufallsvariable. Man stellt die Modellannahme, dass für den bedingten Erwartungswert $E(\lambda_i | \nu_i) = \mu_i \nu_i$ gilt, wobei nach Herausintegrieren des multiplikativen Fehlerterms ν_i

$$E(\lambda_i) = \mu_i E(\nu_i) = \mu_i \tag{2.16}$$

gelten soll, wodurch man die Identifizierbarkeitsbedingung $E(\nu_i) = 1$ erhält. Zu zeigen ist nun, dass damit die Varianz der entstandenen Mischverteilung den Erwartungswert übersteigt.

Man setzt dazu also $E(\nu_i) = 1$. Weiter sei $\text{Var}(\nu_i) = \sigma_\nu^2$ und die ν_i seien iid vorausgesetzt. Dann gilt mit den bekannten Regeln für bedingten Erwartungswert und Varianz

$$\begin{aligned} E(Y_i | \mathbf{x}_i) &= E_{\nu_i} \left[\underbrace{E(Y_i | \mathbf{x}_i, \nu_i)}_{\lambda_i = \mu_i \nu_i} \right] = E_{\nu_i}(\mu_i \nu_i) = \mu_i \\ \text{Var}(Y_i | \mathbf{x}_i) &= E_{\nu_i} \left[\underbrace{\text{Var}(Y_i | \mathbf{x}_i, \nu_i)}_{\lambda_i = \mu_i \nu_i} \right] + \text{Var}_{\nu_i} \left[\underbrace{E(Y_i | \mathbf{x}_i, \nu_i)}_{\lambda_i = \mu_i \nu_i} \right] \\ &= E_{\nu_i}(\mu_i \nu_i) + \text{Var}_{\nu_i}(\mu_i \nu_i) \\ &= \mu_i + \sigma_\nu^2 \mu_i^2. \end{aligned}$$

Damit ist $\text{Var}(Y_i | \mathbf{x}_i) > E(Y_i | \mathbf{x}_i)$ gezeigt, sofern $\sigma_\nu^2 > 0$ gilt.

Ist die Verteilung von ν_i bekannt, so erhält man die Verteilung des Mischmodells durch Integration der gemischten Dichte bezüglich ν_i . Konkret sei nun $f(y_i | \mathbf{x}_i, \nu_i)$ die Dichte

bzw. die Wahrscheinlichkeitsfunktion der Y_i und $g(\nu_i)$ die Dichte der Fehler ν_i . Dann ergibt sich daraus die gemischte Marginaldichte von $(Y_i | \mathbf{x}_i)$ durch Integrieren bezüglich ν_i , d.h.

$$h(y_i | \mathbf{x}_i) = \int f(y_i | \mathbf{x}_i, \nu_i) g(\nu_i) d\nu_i. \quad (2.17)$$

Ein Beispiel für ein gemischtes Modell ist das Poisson-Gamma-Modell, das man erhält, wenn der Parameter der Poissonverteilung nicht fest vorausgesetzt wird, sondern nach der Gammaverteilung variiert. Das Ergebnis dieser Mischung ist dann eine NB-Verteilung.

Beispiel 2.3.9 (Poisson-Gamma-Mischmodell) (siehe Cameron & Trivedi (1998, S. 101) [2]).

Die Zählvariable Y sei bedingt poissonverteilt mit Parameter λ :

$$f(y_i | \lambda_i) = \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!}, \quad y_i = 0, 1, \dots$$

Der Parameter λ_i habe einen multiplikativen Fehlerterm $\exp(\varepsilon_i)$, was mit

$$\lambda_i = \exp(\mathbf{x}_i^t \boldsymbol{\beta} + \varepsilon_i) = \exp(\mathbf{x}_i^t \boldsymbol{\beta}) \exp(\varepsilon_i) = \mu_i \nu_i \quad (2.18)$$

ausgedrückt wird. Dabei ist $\boldsymbol{\beta} = (\beta_0, \dots, \beta_p)^t$, wobei β_0 den Intercept-Term bestimmt.

Sei nun der Fehlerterm ν_i gammaverteilt mit den Parametern q und δ , also

$$g(\nu_i, q, \delta) = \frac{\delta^q}{\Gamma(q)} \nu_i^{q-1} e^{-\delta \nu_i}, \quad q > 0, \delta > 0.$$

Dann gilt für den Erwartungswert $E(\nu_i) = \frac{q}{\delta}$ und für die Varianz $\text{Var}(\nu_i) = \frac{q}{\delta^2}$.

Wegen der Identifizierbarkeitsbedingung $E(\nu_i) = 1$ aus (2.16) muss $q = \delta$ gewählt werden, womit sich hier die Gammaverteilung auf nur einen Parameter reduziert.

Betrachtet man ausgehend von der bekannten Gammaverteilung von ν_i die Verteilung von λ , so ergibt sich durch die Transformation $\nu = \frac{\lambda}{\mu}$ nach (2.18) die folgende Dichte für λ :

$$\begin{aligned} f(\lambda | \mu, q) = \frac{1}{\mu} g\left(\nu = \frac{\lambda}{\mu}\right) &= \frac{1}{\mu} \frac{q^q}{\Gamma(q)} \left(\frac{\lambda}{\mu}\right)^{q-1} e^{-q \frac{\lambda}{\mu}} \\ &= \frac{\left(\frac{q}{\mu}\right)^q}{\Gamma(q)} \lambda^{q-1} e^{-q \frac{\lambda}{\mu}}, \end{aligned}$$

was einer Gammaverteilung mit den Parametern q und $\frac{q}{\mu}$ entspricht. Zusammenfassend kann also gefolgert werden:

Sei $Y \mid \lambda \sim \text{Poi}(\lambda)$, und $\lambda \sim \Gamma(q, \frac{q}{\mu})$, dann ergibt sich daraus als Mischverteilung eine NB-Verteilung mit den Parametern μ und $\frac{1}{q} = a$, d.h. $NB(\mu, \frac{1}{q} = a)$.

Beweis: Da keine Verwechslungen möglich sind, wird hier im folgenden der Index i weggelassen. Nach (2.17) gilt dann

$$\begin{aligned}
 h(y \mid \mu, q) &= \int_0^\infty f(y \mid \mu, \nu) g(\nu) d\nu \\
 &\stackrel{\nu=\frac{\lambda}{\mu}}{=} \int_0^\infty f(y \mid \mu, \lambda) g(\lambda) \frac{1}{\mu} d\lambda \\
 &= \int_0^\infty \frac{e^{-\lambda} \lambda^y}{y!} \frac{1}{\mu} \frac{q^q}{\Gamma(q)} \left(\frac{\lambda}{\mu}\right)^{q-1} e^{-q\frac{\lambda}{\mu}} d\lambda \\
 &= \frac{1}{y! \mu} \frac{q^q}{\Gamma(q)} \frac{1}{\mu^{q-1}} \int_0^\infty e^{-\lambda(1+\frac{q}{\mu})} \lambda^{y+q-1} d\lambda \\
 &\stackrel{(*)}{=} \frac{1}{\Gamma(q)y!} \left(\frac{q}{\mu}\right)^q \frac{\Gamma(y+q)}{(1+\frac{q}{\mu})^{y+q}} \\
 &= \frac{\Gamma(y+q)}{\Gamma(q)y!} \left(\frac{1}{1+\frac{q}{\mu}}\right)^y \left(\frac{q}{\mu} \frac{1}{1+\frac{q}{\mu}}\right)^q \\
 &= \frac{\Gamma(y+q)}{\Gamma(q)y!} \left(\frac{\mu}{\mu+q}\right)^y \left(\frac{q}{\mu+q}\right)^q.
 \end{aligned}$$

$$(*) \text{ Es gilt } \int_0^\infty t^{a-1} e^{-bt} dt = \frac{\Gamma(a)}{b^a} \quad \text{für jedes } b > 0.$$

Setzt man nun $q = \frac{1}{a}$, so erhält man genau die Darstellung der NB-Verteilung, wie sie in Definition 2.3.6 angegeben ist.

□

2.3.3 Die NB-Verteilung als GLM

Im zweiten Abschnitt dieses Kapitels wurde in Beispiel 2.2.2 bereits gezeigt, dass die Poissonverteilung zur exponentiellen Familie gehört und dass sie somit als zugrundeliegende Verteilung eines generalisierten linearen Modells gewählt werden kann. Analog lässt sich dies nun auch für die NB-Verteilung zeigen, wobei im folgenden immer von der erweiterten Darstellungsform mit stetigem Parameter ausgegangen wird. Diese Wahrscheinlichkeitsfunktion aus Definition 2.3.6 kann dann zu

$$f_{NB}(y; \theta, \phi) = \exp \left\{ \ln \left(\frac{\Gamma(y + a^{-1})}{\Gamma(a^{-1})y!} \right) + y \ln \left(\frac{a\mu}{1 + a\mu} \right) - a^{-1} \ln(1 + a\mu) \right\}$$

umgeformt werden. Für den Fall, dass a bekannt ist, gehört die NB-Verteilung zu der linearen exponentiellen Familie, wie sie in Definition 2.2.1 beschrieben wurde. Dafür

ist $\theta = \ln\left(\frac{a\mu}{1+a\mu}\right)$, $a(\phi) = \phi = 1$, $c(y, \phi) = c(y) = \ln\left(\frac{\Gamma(y+a^{-1})}{\Gamma(a^{-1})y!}\right)$ und

$$b(\theta) = a^{-1} \ln(1 + a\mu) \quad (2.19)$$

zu wählen. Die Funktion $c(y, \phi) = c(y)$ hängt in diesem Fall nur von y ab und der Parameter ϕ spielt keine Rolle. Um $b(\theta)$ in Abhängigkeit von θ zu berechnen, muss zunächst θ nach μ aufgelöst und das Ergebnis in (2.19) eingesetzt werden. Aus

$$\begin{aligned} \theta &= \ln\left(\frac{a\mu}{1+a\mu}\right) \quad \text{erhält man für } \mu \\ \mu &= \frac{e^\theta}{a(1-e^\theta)}. \end{aligned}$$

Wird dieses Ergebnis nun in (2.19) eingesetzt, so ergibt sich

$$\begin{aligned} b(\theta) &= a^{-1} \ln(1 + a\mu) = a^{-1} \ln\left(1 + a \frac{e^\theta}{a(1-e^\theta)}\right) \\ &= a^{-1} \ln\left(1 + \frac{e^\theta}{1-e^\theta}\right) = a^{-1} \ln\left(\frac{1}{1-e^\theta}\right). \end{aligned}$$

Zur Überprüfung wird der Erwartungswert nach der bekannten Formel in (2.4) berechnet:

$$\begin{aligned} E(Y) = b'(\theta) &= a^{-1} \frac{e^\theta}{(1-e^\theta)^2} \bigg/ \frac{1}{1-e^\theta} = a^{-1} \frac{e^\theta}{(1-e^\theta)^2} (1-e^\theta) \\ &= a^{-1} \frac{e^\theta}{1-e^\theta} = \mu. \end{aligned}$$

Die Varianz war in (2.5) angegeben und berechnet sich hier als

$$\begin{aligned} \text{Var}(Y) = b''(\theta)a(\phi) &= a^{-1} \frac{e^\theta(1-e^\theta) + e^\theta e^\theta}{(1-e^\theta)^2} \\ &= a^{-1} \frac{e^\theta}{1-e^\theta} + a^{-1} \frac{(e^\theta)^2}{(1-e^\theta)^2} = \mu + a\mu^2. \end{aligned}$$

Sowohl der Erwartungswert als auch die Varianz stimmen somit wie erwartet mit den in den Gleichungen (2.14) und (2.15) berechneten Ergebnissen überein. Ist nun aber zusätzlich auch a unbekannt, so ist die NB-Verteilung dann zwar kein GLM, aber es kann nach Cameron & Trivedi (1998, S. 73) [2] gezeigt werden, dass sie zumindest zu der Klasse der linearen exponentiellen Familie mit einem Störparameter ϕ gehört, die zu den GLM's sehr ähnlich ist (siehe Cameron & Trivedi (1998, S. 33) [2]). Die exponentielle Familie mit Störparameter ϕ ist in Cameron & Trivedi (1998, S.32) [2] definiert durch

$$f_{EFN}(y; \mu, \phi) = \exp\{\tilde{a}(\mu, \phi) + \tilde{b}(y, \mu) + \tilde{c}(\mu, \phi)y\},$$

wobei $\mu = E(y)$, $\omega(\mu, \phi) = Var(y)$ und $\phi = \Psi(\mu, \omega)$ ist und Ψ eine differenzierbare Funktion von μ und ω darstellt. Ausserdem definiert Ψ für jedes gegebene μ eine eins-zu-eins Beziehung zwischen ϕ und ω . Bei der NB-Verteilung entspricht nun der unbekannte Parameter a dem Störparameter ϕ und es gilt damit

$$\tilde{c}(\mu, a) = \ln \left(\frac{a\mu}{1 + a\mu} \right), \quad \tilde{b}(y, a) = \ln \left(\frac{\Gamma(y + a^{-1})}{\Gamma(a^{-1})y!} \right) \quad \text{und} \quad \tilde{a}(\mu, a) = a^{-1} \ln(1 + a\mu).$$

Weiter ist $\omega(\mu, a) = \mu + a\mu^2$. Aufgelöst nach dem Störparameter a ergibt dies $a = \frac{\omega - \mu}{\mu^2} = \Psi(\mu, \omega)$, wobei klar ersichtlich ist, dass die Funktion Ψ sowohl in μ als auch in ω differenzierbar ist und dass sie für ein festes μ eine eindeutige Beziehung zwischen a und ω darstellt, womit die Zugehörigkeit der NB-Verteilung zur exponentiellen Familie mit Störparameter gezeigt ist.

Das Ziel dieser Arbeit ist nun im wesentlichen, das Poissonregressionsmodell mit dem NB-Regressionsmodell zu vergleichen, wobei nach Satz 2.3.8 der Dispersionsindex a eine entscheidende Rolle spielt. Schliesslich soll ein Test erstellt werden, der prüft, ob der Parameter a grösser als eine vorgegebene Schranke a_0 ist, oder nicht. Im ersten Fall würde man sich für das NB-Modell entscheiden, das die offensichtlich vorhandene Überdispersion modelliert, während man im Fall $a < a_0$ die Überdispersion durch die Wahl des Poissonmodells vernachlässigen würde. Die notwendigen Grundlagen und die Herleitung des Tests folgen im nächsten Kapitel. Dazu sind jedoch die Eigenschaften der Schätzer für die Parameter β und a zu untersuchen, wofür nach (2.10) die Fisher-Informationsmatrix benötigt wird. Diese soll daher im folgenden schrittweise hergeleitet werden.

Zunächst wird die Likelihood-Funktion berechnet, wobei zu beachten ist, dass für die weiteren Ausführungen die log-link-Funktion für das NB-Modell gewählt wird, d.h. $Y_i \sim NB(\mu_i, a)$, $i = 1, \dots, n$ unabhängig mit $\mu_i = \exp(\mathbf{x}_i^t \beta)$. Für andere Linkspezifikationen ergeben sich analoge Ergebnisse. Die Likelihood-Funktion ist gegeben durch

$$L(\beta, a) = \prod_{i=1}^n \frac{\Gamma(y_i + a^{-1})}{\Gamma(a^{-1})y_i!} \left(\frac{a\mu_i}{1 + a\mu_i} \right)^{y_i} \left(\frac{1}{1 + a\mu_i} \right)^{a^{-1}}.$$

Da $\Gamma(x+1) = x\Gamma(x)$ für $x > 0$, gilt weiter

$$\frac{\Gamma(y_i + a^{-1})}{\Gamma(a^{-1})} = (y_i + a^{-1} - 1) \dots (y_i + a^{-1} - y_i) = \prod_{j=0}^{y_i} (j + a^{-1}),$$

wobei $y_i^* = y_i - 1$ ist und $\prod_{j=0}^{y_i^*}$ Null wird, wenn $y_i^* < 0$. Deshalb kann die Likelihood-Funktion umgeschrieben werden zu

$$\begin{aligned} L(\boldsymbol{\beta}, a) &= \prod_{i=1}^n \left[\left(\prod_{j=0}^{y_i^*} (j + a^{-1}) \right) \frac{1}{y_i!} \left(\frac{a\mu_i}{1 + a\mu_i} \right)^{y_i} \left(\frac{1}{1 + a\mu_i} \right)^{a^{-1}} \right] \\ &= \prod_{i=1}^n \left[\left(\frac{1}{a^{y_i}} \prod_{j=0}^{y_i^*} (aj + 1) \right) \frac{1}{y_i!} a^{y_i} \left(\frac{\mu_i}{1 + a\mu_i} \right)^{y_i} \left(\frac{1}{1 + a\mu_i} \right)^{a^{-1}} \right] \\ &= \prod_{i=1}^n \left[\left(\prod_{j=0}^{y_i^*} (aj + 1) \right) \frac{1}{y_i!} \left(\frac{\mu_i}{1 + a\mu_i} \right)^{y_i} \left(\frac{1}{1 + a\mu_i} \right)^{a^{-1}} \right]. \end{aligned}$$

Daraus ergibt sich durch Logarithmieren die Log-Likelihood-Funktion

$$\mathcal{L}(\boldsymbol{\beta}, a) = \ln L(\boldsymbol{\beta}, a) = \sum_{i=1}^n \left[\sum_{j=0}^{y_i^*} \ln(aj + 1) - \ln(y_i!) + y_i \ln \mu_i - (y_i + a^{-1}) \ln(1 + a\mu_i) \right].$$

Um die Fisher-Informationsmatrix zu erhalten, werden ausgehend von der Log-Likelihood-Funktion zunächst die ersten Ableitungen bezüglich der Parameter gebildet, also die partiellen Ableitungen nach β_1 bis β_p und die Ableitung nach a . Bei den Ableitungen nach den $\boldsymbol{\beta}$'s muss beachtet werden, dass zwar formal in der Log-Likelihood-Funktion keine $\boldsymbol{\beta}$'s auftauchen, aber die Funktion wegen $\mu_i = \exp(\mathbf{x}_i^t \boldsymbol{\beta})$ implizit $\boldsymbol{\beta}$ enthält. Ersetzt man daher in der Log-Likelihood-Funktion die μ_i durch $\mu_i = \exp(\mathbf{x}_i^t \boldsymbol{\beta})$, kann man wie gewohnt nach β_r , $r = 1, \dots, p$ ableiten.

$$\begin{aligned} \frac{\partial \mathcal{L}(\boldsymbol{\beta}, a)}{\partial \beta_r} &= \frac{\partial \mathcal{L}(\boldsymbol{\beta}, a)}{\partial \beta_r} \sum_{i=1}^n \left[\sum_{j=0}^{y_i^*} \ln(aj + 1) - \ln(y_i!) + y_i \ln(\exp(\mathbf{x}_i^t \boldsymbol{\beta})) - \right. \\ &\quad \left. - (y_i + a^{-1}) \ln(1 + a \exp(\mathbf{x}_i^t \boldsymbol{\beta})) \right] \\ &= \sum_{i=1}^n \left(y_i x_{ir} - (y_i + a^{-1}) \frac{1}{1 + a\mu_i} a \underbrace{\exp(\mathbf{x}_i^t \boldsymbol{\beta})}_{\mu_i} x_{ir} \right) \\ &= \sum_{i=1}^n \left(y_i x_{ir} - \frac{y_i a \mu_i x_{ir} + \mu_i x_{ir}}{1 + a\mu_i} \right) \\ &= \sum_{i=1}^n \left(\frac{y_i x_{ir} + y_i x_{ir} a \mu_i - y_i a \mu_i x_{ir} - \mu_i x_{ir}}{1 + a\mu_i} \right) \\ &= \sum_{i=1}^n \frac{x_{ir} (y_i - \mu_i)}{1 + a\mu_i} \quad \text{für } r = 1, \dots, p. \end{aligned}$$

$$\frac{\partial \mathcal{L}(\boldsymbol{\beta}, a)}{\partial a} = \sum_{i=1}^n \left[\sum_{j=0}^{y_i^*} \left(\frac{j}{aj+1} \right) + \frac{1}{a^2} \ln(1 + a\mu_i) - (y_i + a^{-1}) \frac{\mu_i}{1 + a\mu_i} \right].$$

Da die Fisher-Informationsmatrix nach Satz 2.2.7 der Erwartungswert der negativen Hesse-Matrix ist, muss diese nun im folgenden Schritt bestimmt werden. Dazu werden die zweiten Ableitungen benötigt, weshalb zuerst die zweite Ableitung nach $\boldsymbol{\beta}$, also $\frac{\partial^2 \mathcal{L}(\boldsymbol{\beta}, a)}{\partial \beta_r \partial \beta_s}$ berechnet, dann die gemischte Ableitung $\frac{\partial^2 \mathcal{L}(\boldsymbol{\beta}, a)}{\partial \beta_r \partial a}$ und schliesslich die zweite Ableitung nach a , d.h. $\frac{\partial^2 \mathcal{L}(\boldsymbol{\beta}, a)}{\partial a^2}$ ermittelt wird.

Um nicht immer μ_i durch $\mu_i = \exp(\mathbf{x}_i^t \boldsymbol{\beta})$ ersetzen zu müssen, wird für die weiteren Berechnungen die Gleichheit

$$\frac{\partial \mu_i}{\partial \beta_s} = \frac{\partial \exp(\mathbf{x}_i^t \boldsymbol{\beta})}{\partial \beta_s} = \exp(\mathbf{x}_i^t \boldsymbol{\beta}) x_{is} = \mu_i x_{is}$$

verwendet. Damit ergibt sich

$$\begin{aligned} \frac{\partial^2 \mathcal{L}(\boldsymbol{\beta}, a)}{\partial \beta_r \partial \beta_s} &= \frac{\partial}{\partial \beta_s} \sum_{i=1}^n \frac{x_{ir}(y_i - \mu_i)}{1 + a\mu_i} \\ &= \sum_{i=1}^n \frac{-x_{ir}\mu_i x_{is}(1 + a\mu_i) - x_{ir}(y_i - \mu_i)a\mu_i x_{is}}{(1 + a\mu_i)^2} \\ &= \sum_{i=1}^n \frac{-x_{ir}x_{is}\mu_i - x_{ir}x_{is}\mu_i^2 a - x_{ir}x_{is}y_i a\mu_i + x_{ir}x_{is}a\mu_i^2}{(1 + a\mu_i)^2} \\ &= \sum_{i=1}^n \frac{-x_{ir}x_{is}\mu_i(1 + ay_i)}{(1 + a\mu_i)^2} \end{aligned} \quad (2.20)$$

$$\frac{\partial^2 \mathcal{L}(\boldsymbol{\beta}, a)}{\partial \beta_r \partial a} = \sum_{i=1}^n \frac{-x_{ir}(y_i - \mu_i)\mu_i}{(1 + a\mu_i)^2} \quad (2.21)$$

$$\begin{aligned} \frac{\partial^2 \mathcal{L}(\boldsymbol{\beta}, a)}{\partial a^2} &= \sum_{i=1}^n \left[\sum_{j=0}^{y_i^*} \frac{-j^2}{(aj+1)^2} - 2\frac{1}{a^3} \ln(1 + a\mu_i) + 2\frac{1}{a^2} \frac{\mu_i}{1 + a\mu_i} + (y_i + a^{-1}) \frac{\mu_i^2}{(1 + a\mu_i)^2} \right] \\ &= - \sum_{i=1}^n \left[\sum_{j=0}^{y_i^*} \left(\frac{j}{aj+1} \right)^2 + \frac{2}{a^3} \ln(1 + a\mu_i) - \frac{2}{a^2} \frac{\mu_i}{1 + a\mu_i} - \frac{(y_i + a^{-1})\mu_i^2}{(1 + a\mu_i)^2} \right] \end{aligned} \quad (2.22)$$

Das Ziel ist nun, aus diesen Ableitungen die Fisher-Informationsmatrix FI herzuleiten, die in Satz 2.2.7 definiert wurde. Die Fisher-Informationsmatrix ist in diesem Fall eine $((p+1) \times (p+1))$ -Matrix, deren erster Block aus den Erwartungswerten der negativen zweiten Ableitungen nach $\boldsymbol{\beta}$ besteht und mit $FI_{r,s}(\boldsymbol{\beta}, a)$, $r, s = 1, \dots, p$ abgekürzt wird. In der letzten Spalte, und wegen der Symmetrie der Fisher-Informationsmatrix analog auch in der letzten Zeile, sind die gemischten Ableitungen eingetragen, die mit

$FI_{r,p+1}(\beta, a)$ bezeichnet werden. Schliesslich ergibt sich der letzte Eintrag an der Stelle $(p+1, p+1)$ der Matrix aus dem Erwartungswert der negativen zweiten Ableitung nach a , kurz $FI_{p+1,p+1}(\beta, a)$. Aus (2.20) erhält man also

$$\begin{aligned} FI_{r,s}(\beta, a) &= E \left[\sum_{i=1}^n \frac{x_{ir} x_{is} \mu_i (1 + a y_i)}{(1 + a \mu_i)^2} \right] \\ &= \sum_{i=1}^n \frac{(1 + a E(y_i)) \mu_i x_{ir} x_{is}}{(1 + a \mu_i)^2} = \sum_{i=1}^n \frac{\mu_i x_{ir} x_{is}}{1 + a \mu_i} \quad \text{für } r, s = 1, \dots, p. \end{aligned} \quad (2.23)$$

Um die Einträge der letzten Spalte und Zeile zu erhalten, berechnet man ausgehend von (2.21)

$$\begin{aligned} FI_{r,p+1}(\beta, a) &= E \left[\sum_{i=1}^n \frac{\mu_i (y_i - \mu_i) x_{ir}}{(1 + a \mu_i)^2} \right] \\ &= \sum_{i=1}^n \frac{\mu_i (E(y_i) - \mu_i) x_{ir}}{(1 + a \mu_i)^2} = 0 \quad \text{für } r = 1, \dots, p. \end{aligned} \quad (2.24)$$

Der $(p+1, p+1)$ -Eintrag der Fisher-Informationsmatrix ist wegen des Erwartungswertes einer Summe, deren Index von y abhängt, etwas schwieriger zu berechnen. Der einfachste Weg das Ergebnis zu erhalten ist, $L(\beta, a)$ in eine Form umzuschreiben, die aus den Parametern β und $k = a^{-1}$ besteht. Daraus bildet man die Log-Likelihood-Funktion $\mathcal{L}(\beta, k)$ und berechnet sodann die zweite Ableitung nach k , also $\frac{\partial^2 \mathcal{L}}{\partial k^2}$. Hieraus lässt sich schliesslich durch Bildung des Erwartungswertes der Eintrag in der Fisher-Informationsmatrix ermitteln, wobei der folgende Satz verwendet wird.

Satz 2.3.10 *Es gilt die Gleichheit*

$$E \left(-\frac{\partial^2 \mathcal{L}(\beta, a)}{\partial a^2} \right) = \frac{1}{a^4} E \left(-\frac{\partial^2 \mathcal{L}(\beta, k)}{\partial k^2} \right),$$

wobei $k = \frac{1}{a}$ ist.

Beweis: Da $\mathcal{L} = \ln L(\beta, a)$ die Log-Likelihood-Funktion beschreibt, lässt sich die linke Seite zunächst ohne den Erwartungswert wie folgt berechnen:

$$\frac{\partial^2 \mathcal{L}(\beta, a)}{\partial a^2} = \frac{\partial^2}{\partial a^2} (\ln L(\beta, a)) = \frac{\partial}{\partial a} \left(\frac{L'(\beta, a)}{L(\beta, a)} \right) = \frac{L''(\beta, a) L(\beta, a) - [L'(\beta, a)]^2}{[L(\beta, a)]^2},$$

wobei L' und L'' jeweils die erste bzw. zweite Ableitung nach a abkürzt. Wird die Likelihood-Funktion nun mit $k = g(a) = \frac{1}{a}$ umparametrisiert, so lässt sie sich als eine Funktion $L(\beta, k) = L(\beta, g(a))$ auffassen, für die allgemein die Formeln

$$\frac{\partial}{\partial a} L(\beta, k) = \frac{\partial}{\partial a} L(\beta, g(a)) = \frac{\partial}{\partial g(a)} L(\beta, g(a)) \frac{\partial}{\partial a} g(a) = L'(\beta, g(a)) g'(a) \quad (2.25)$$

und

$$\begin{aligned}
\frac{\partial^2}{\partial a^2} L(\beta, k) &= \frac{\partial^2}{\partial a^2} L(\beta, g(a)) \\
&= \frac{\partial^2}{\partial (g(a))^2} L(\beta, g(a)) \left[\frac{\partial}{\partial a} g(a) \right]^2 + \frac{\partial}{\partial g(a)} L(\beta, g(a)) \frac{\partial^2}{\partial a^2} g(a) \\
&= L''(\beta, g(a)) [g'(a)]^2 + L'(\beta, g(a)) g''(a)
\end{aligned} \tag{2.26}$$

gelten. Die Ableitungen von $g(a)$ ergeben sich zu

$$g'(a) = -\frac{1}{a^2} \quad \text{und} \quad g''(a) = \frac{2}{a^3}. \tag{2.27}$$

Mit dieser Upparametrisierung berechnet sich die linke Seite der Behauptung ohne Beachtung des Erwartungswertes wie folgt:

$$\begin{aligned}
\frac{\partial^2 \mathcal{L}(\beta, k)}{\partial a^2} &= \frac{\partial^2 \mathcal{L}(\beta, g(a))}{\partial a^2} = \frac{\partial^2}{\partial a^2} (\ln L(\beta, g(a))) = \frac{\partial}{\partial a} \left(\frac{\frac{\partial}{\partial a} L(\beta, g(a))}{L(\beta, g(a))} \right) \\
&\stackrel{(2.25)}{=} \frac{\partial}{\partial a} \left(\frac{L'(\beta, g(a)) g'(a)}{L(\beta, g(a))} \right) \\
&\stackrel{(2.26)}{=} \frac{[L''(\beta, g(a)) [g'(a)]^2 + L'(\beta, g(a)) g''(a)] L(\beta, g(a)) - [L'(\beta, g(a)) g'(a)]^2}{[L(\beta, g(a))]^2} \\
&\stackrel{(2.27)}{=} \frac{1}{a^4} \frac{[L''(\beta, g(a)) + 2a L'(\beta, g(a))] L(\beta, g(a)) - [L'(\beta, g(a))]^2}{[L(\beta, g(a))]^2}.
\end{aligned} \tag{2.28}$$

Ermittelt man nun davon den Erwartungswert, so kann dieser in

$$\frac{1}{a^4} E \left(\frac{L''(\beta, g(a)) L(\beta, g(a)) - [L'(\beta, g(a))]^2}{[L(\beta, g(a))]^2} \right) + \frac{1}{a^4} E \left(\frac{2a L'(\beta, g(a)) L(\beta, g(a))}{[L(\beta, g(a))]^2} \right)$$

aufgespalten werden. Für den 2. Term gilt speziell

$$\begin{aligned}
\frac{1}{a^4} E \left(\frac{2a L'(\beta, g(a)) L(\beta, g(a))}{[L(\beta, g(a))]^2} \right) &= \frac{2}{a^3} E \left(\frac{L'(\beta, g(a))}{L(\beta, g(a))} \right) \\
&= \frac{2}{a^3} \int \frac{L'(\beta, g(a))}{L(\beta, g(a))} L(\beta, g(a)) \, dy \\
&= \frac{2}{a^3} \int L'(\beta, g(a)) \, dy \\
&= \frac{2}{a^3} \frac{\partial}{\partial g(a)} \underbrace{\int L(\beta, g(a)) \, dy}_{=1, \text{ da } L \text{ Dichte}} = 0,
\end{aligned} \tag{2.29}$$

wobei im vorletzten Schritt die Vertauschung von Integration und Differenzierung unter Regularitätsbedingungen durchgeführt werden darf. Mit diesem Resultat ergibt sich

schliesslich zusammenfassend

$$\begin{aligned}
E\left(-\frac{\partial^2 \mathcal{L}(\boldsymbol{\beta}, a)}{\partial a^2}\right) &= E\left(-\frac{L''(\boldsymbol{\beta}, a)L(\boldsymbol{\beta}, a) - [L'(\boldsymbol{\beta}, a)]^2}{[L(\boldsymbol{\beta}, a)]^2}\right) \\
&\stackrel{(2.28)}{=} \frac{1}{a^4} E\left(-\frac{[L''(\boldsymbol{\beta}, g(a)) + 2aL'(\boldsymbol{\beta}, g(a))] L(\boldsymbol{\beta}, g(a)) - [L'(\boldsymbol{\beta}, g(a))]^2}{[L(\boldsymbol{\beta}, g(a))]^2}\right) \\
&\stackrel{(2.29)}{=} \frac{1}{a^4} E\left(-\frac{L''(\boldsymbol{\beta}, g(a))L(\boldsymbol{\beta}, g(a)) - [L'(\boldsymbol{\beta}, g(a))]^2}{[L(\boldsymbol{\beta}, g(a))]^2}\right) \\
&= \frac{1}{a^4} E\left(-\frac{\partial^2 \mathcal{L}(\boldsymbol{\beta}, g(a))}{\partial (g(a))^2}\right) = \frac{1}{a^4} E\left(-\frac{\partial^2 \mathcal{L}(\boldsymbol{\beta}, k)}{\partial k^2}\right).
\end{aligned}$$

□

Zunächst wird daher die Likelihood-Funktion als eine Funktion von $\boldsymbol{\beta}$ und $k = a^{-1}$ umgeschrieben:

$$\begin{aligned}
L(\boldsymbol{\beta}, k) &= \prod_{i=1}^n \frac{\Gamma(y_i + k)}{\Gamma(k)y!} \left(\frac{\mu_i}{1 + \frac{\mu_i}{k}}\right)^{y_i} \left(\frac{1}{1 + \frac{\mu_i}{k}}\right)^k \\
&= \prod_{i=1}^n \left[\prod_{j=0}^{y_i^*} (j + k) \frac{1}{y!} \left(\frac{\mu_i}{k + \mu_i}\right)^{y_i} \left(\frac{1}{1 + \frac{\mu_i}{k}}\right)^k \right].
\end{aligned}$$

Durch Logarithmieren ergibt sich nun die Log-Likelihood-Funktion

$$\mathcal{L}(\boldsymbol{\beta}, k) = \sum_{i=1}^n \left[\sum_{j=0}^{y_i^*} \ln(j + k) - \ln(y!) + y_i \ln \mu_i - y_i \ln(k + \mu_i) - k \ln\left(1 + \frac{\mu_i}{k}\right) \right].$$

Daraus erhält man die Ableitungen

$$\begin{aligned}
\frac{\partial \mathcal{L}(\boldsymbol{\beta}, k)}{\partial k} &= \sum_{i=1}^n \left[\sum_{j=0}^{y_i^*} \frac{1}{j + k} - y_i \frac{1}{k + \mu_i} - \ln\left(1 + \frac{\mu_i}{k}\right) - k \frac{1}{1 + \frac{\mu_i}{k}} \left(-\frac{\mu_i}{k^2}\right) \right] \\
&= \sum_{i=1}^n \left[\sum_{j=0}^{y_i^*} \frac{1}{j + k} - \frac{y_i}{k + \mu_i} - \ln\left(1 + \frac{\mu_i}{k}\right) + \frac{\mu_i}{k + \mu_i} \right] \quad \text{und} \\
\frac{\partial^2 \mathcal{L}(\boldsymbol{\beta}, k)}{\partial k^2} &= \sum_{i=1}^n \left[\sum_{j=0}^{y_i^*} \frac{-1}{(j + k)^2} + \frac{y_i}{(k + \mu_i)^2} - \frac{-\frac{\mu_i}{k^2}}{1 + \frac{\mu_i}{k}} + \frac{-\mu_i}{(k + \mu_i)^2} \right] \\
&= -\sum_{i=1}^n \left[\sum_{j=0}^{y_i^*} \frac{1}{(j + k)^2} - \frac{y_i}{(k + \mu_i)^2} - \frac{\mu_i}{(k + \mu_i)k} + \frac{\mu_i}{(k + \mu_i)^2} \right] \quad (2.30)
\end{aligned}$$

Für die Berechnung des letzten Eintrags in der Fisher-Informationsmatrix ermittelt man aus (2.30) den Erwartungswert, der sich zu

$$\begin{aligned}
 FI_{p+1,p+1}(\beta, a) &= E \left[-\frac{\partial^2 \mathcal{L}(\beta, a)}{\partial a^2} \right] = a^{-4} E \left[-\frac{\partial^2 \mathcal{L}(\beta, k)}{\partial k^2} \right] \\
 &= a^{-4} E \left[\sum_{i=1}^n \left[\sum_{j=0}^{Y_i^*} \frac{1}{(j+k)^2} - \frac{y_i}{(k+\mu_i)^2} - \frac{\mu_i}{(k+\mu_i)k} + \frac{\mu_i}{(k+\mu_i)^2} \right] \right] \\
 &= a^{-4} \left[\sum_{i=1}^n \left[E \left(\sum_{j=0}^{Y_i^*} \frac{1}{(j+k)^2} \right) - \frac{E(y_i)}{(k+\mu_i)^2} - \frac{\mu_i}{(k+\mu_i)k} + \frac{\mu_i}{(k+\mu_i)^2} \right] \right] \\
 &= a^{-4} \left[\sum_{i=1}^n \left[E \left(\sum_{j=0}^{Y_i^*} \frac{1}{(j+a^{-1})^2} \right) - \frac{\mu_i}{(a^{-1}+\mu_i)a^{-1}} \right] \right] \quad (2.31)
 \end{aligned}$$

ergibt. Der i-te Term in obiger Formel lässt sich nach Lawless (1987) [13] durch

$$a^{-4} \left(\sum_{j=0}^{\infty} \frac{1}{(j+a^{-1})^2} P_{NB}(Y_i \geq j) - \frac{a\mu_i}{\mu_i + a^{-1}} \right) \quad (2.32)$$

berechnen.

Beweis: Für diesen Beweis ist speziell die Gleichheit

$$E \left(\sum_{j=0}^{Y_i^*} \frac{1}{(j+a^{-1})^2} \right) = \sum_{j=0}^{\infty} \frac{1}{(j+a^{-1})^2} P_{NB}(Y_i \geq j)$$

zu zeigen, wobei $Y_i^* = Y_i - 1$ ist. Nach Resnick (1992, S. 2) [18] gilt allgemein für eine nichtnegative ganzzahlige Zufallsvariable Y die Formel

$$E(f(Y)) = \sum_{y=0}^{\infty} f(y) p_y,$$

wobei $p_y = P(Y = y)$ abkürzt. Damit gilt

$$\begin{aligned}
 E(f(Y)) &= E \left(\sum_{j=0}^{Y^*} \frac{1}{(j+a^{-1})^2} \right) \\
 &= \sum_{y^*=0}^{\infty} \sum_{j=0}^{y^*} \frac{1}{(j+a^{-1})^2} P_{NB}(Y = y^*) \\
 &= \sum_{j=0}^0 \frac{1}{(j+a^{-1})^2} P_{NB}(Y = 0) + \sum_{j=0}^1 \frac{1}{(j+a^{-1})^2} P_{NB}(Y = 1) + \dots \\
 &= \frac{1}{(0+a^{-1})^2} (P_{NB}(Y = 0) + P_{NB}(Y = 1) + \dots)
 \end{aligned}$$

$$\begin{aligned}
& + \frac{1}{(1 + a^{-1})^2} (P_{NB}(Y = 1) + P_{NB}(Y = 2) + \dots) + \dots \\
& = \frac{1}{(0 + a^{-1})^2} P_{NB}(Y \geq 0) + \frac{1}{(1 + a^{-1})^2} P_{NB}(Y \geq 1) + \dots \\
& = \sum_{j=0}^{\infty} \frac{1}{(j + a^{-1})^2} P_{NB}(Y \geq j).
\end{aligned}$$

□

Nun sind alle Einträge berechnet, so dass die Fisher-Informationsmatrix in Blockform angegeben werden kann.

$$FI(\boldsymbol{\beta}, a) = \begin{bmatrix} FI_{r,s}(\boldsymbol{\beta}, a) & \mathbf{0} \\ \mathbf{0}^t & FI_{p+1,p+1}(\boldsymbol{\beta}, a) \end{bmatrix}, \quad \text{für } r, s = 1, \dots, p.$$

Bei dieser Darstellung ist zu beachten, dass $FI_{r,s}$ hier eine Blockmatrix darstellt, während $FI_{p+1,p+1}$ ein skalarer Wert ist. $\mathbf{0}$ beschreibt einen Nullvektor der Länge p . Das Ziel ist nun, Aussagen über die Schätzer $(\hat{\boldsymbol{\beta}}, \hat{a})$ von $(\boldsymbol{\beta}, a)$ im NB-Modell treffen zu können. Allerdings treten bei der Ermittlung der ML-Schätzer zusätzliche Schwierigkeiten auf, da der Parameter a nicht bekannt ist, sondern ebenfalls geschätzt werden muss. Die einfachste Methode, die Schätzer $(\hat{\boldsymbol{\beta}}, \hat{a})$ zu erhalten ist, $\mathcal{L}(\boldsymbol{\beta}, a)$ bezüglich $\boldsymbol{\beta}$ für bestimmte, feste Werte a zu maximieren. Damit ergeben sich die Schätzung $\hat{\boldsymbol{\beta}}(a)$ und der sogenannte „*Profile-Likelihood*“ $\mathcal{L}(\hat{\boldsymbol{\beta}}(a), a)$ in Abhängigkeit von a , von welchem aus man schliesslich \hat{a} durch eindimensionales Maximieren bestimmen kann. Die Maximierung von $\mathcal{L}(\boldsymbol{\beta}, a)$ bezüglich $\boldsymbol{\beta}$ kann wie in Abschnitt 2.2.6 beschrieben z.B. mit dem Fisher-Scoring-Algorithmus oder dem Newton-Raphson-Verfahren gelöst werden. Das wichtigste Resultat betrifft schliesslich die Asymptotik der Schätzer. Wie am Ende des Abschnittes 2.2.6 erläutert, gilt auch hier für sie eine asymptotische Verteilung. Lawless (1987) [13] beschreibt ohne konkreten Beweis, dass man für $a > 0$ und unter schwachen Bedingungen für die Kovariablenvektoren \mathbf{x}_i , die die Approximation von $n^{-1}FI(\boldsymbol{\beta}, a)$ gegen einen positiv definiten Grenzwert für $n \rightarrow \infty$ sichern, die ML-Schätzer $(\hat{\boldsymbol{\beta}}, \hat{a})$ als asymptotisch normalverteilt behandeln kann, d.h. man kann für grosse n zeigen, dass

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}, \hat{a} - a) \sim N(\mathbf{0}, nFI(\hat{\boldsymbol{\beta}}, \hat{a})^{-1}) \quad (2.33)$$

gilt, wobei die Kovarianzmatrix gegeben ist durch

$$nFI(\hat{\boldsymbol{\beta}}, \hat{a})^{-1} = n \begin{bmatrix} FI_{r,s}^{-1}(\hat{\boldsymbol{\beta}}, \hat{a}) & \mathbf{0} \\ \mathbf{0}^t & FI_{p+1,p+1}^{-1}(\hat{\boldsymbol{\beta}}, \hat{a}) \end{bmatrix}$$

mit den Einträgen $FI_{r,s}^{-1}(\hat{\boldsymbol{\beta}}, \hat{a})$ und $FI_{p+1,p+1}^{-1}(\hat{\boldsymbol{\beta}}, \hat{a})$, die sich aus den Gleichungen (2.23) und (2.31) ausgewertet an $(\boldsymbol{\beta}, a) = (\hat{\boldsymbol{\beta}}, \hat{a})$ ergeben. Da der $(p+1, p+1)$ -Eintrag ein skalarer Wert ist, lässt sich $FI_{p+1,p+1}^{-1}(\hat{\boldsymbol{\beta}}, \hat{a})$ durch einfaches Invertieren berechnen. Fahrmeir & Kaufmann (1985) [5] zeigen dieses asymptotische Verhalten der ML-Schätzer für normale GLM's bei festem Design, während Czado & Munk (2000) [4] auf eine Erweiterung für GLM's mit nichtkanonischen Linkfunktionen eingehen. Sie zeigen, dass eine asymptotische Normalverteilung für den ML-Schätzer, der sowohl den Schätzer für den Regressionsparametervektor $\boldsymbol{\beta}$, als auch den Schätzer für den Linkparameter umfasst, vorliegt. Da hier ebenfalls der ML-Schätzer $(\hat{\boldsymbol{\beta}}, \hat{a})$ aus einem zusätzlichen Parameter a besteht, muss für einen Beweis der asymptotischen Normalverteilung analog wie in Czado & Munk (2000) [4] vorgegangen werden.

Es bleibt anzumerken, dass die Kovarianzmatrix eine Diagonal-Blockmatrix ist, da nach (2.24) die Parameter a und $\boldsymbol{\beta}$ asymptotisch unkorreliert sind.

Kapitel 3

Quantifizierung von Überdispersion

Dieses Kapitel befasst sich mit dem zentralen Thema dieser Arbeit, nämlich der Quantifizierung von Überdispersion. Dazu wird ein Hypothesentest erstellt, der in den nachfolgenden Abschnitten noch näher erläutert wird. In den ersten Abschnitten dieses Kapitels werden die dazu nötigen Grundlagen der Testtheorie zusammengefasst, die auf Lehmann (1986, Kapitel 3) [14] und Bickel & Doksum (1977, Kapitel 5,6) [1] beruhen, während im darauffolgenden Abschnitt schliesslich der spezielle Test zur Quantifizierung von Überdispersion beschrieben wird. Abschliessend erfolgt eine kurze Zusammenfassung der notwendigen Grundlagen zur Durchführung einer Simulation, deren Ergebnisse im 4. Kapitel dargestellt sind.

3.1 Einführung in die Theorie der Hypothesentests

3.1.1 Grundlagen der Testtheorie

Die Grundlage für Hypothesentests ist der Wunsch, eine Entscheidung darüber zu treffen, ob eine erstellte Hypothese wahr oder falsch ist. Die Auswahl liegt in solchen Fällen zwischen nur zwei Entscheidungen: entweder ist die Hypothese anzunehmen, oder sie ist abzulehnen. Eine Entscheidungsprozedur für solche Probleme heisst „*Test*“ der in Frage gestellten Hypothese.

Die Entscheidung basiert auf dem Wert einer Zufallsvariablen X mit Verteilung P_θ , wobei P_θ zu der Klasse $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$ von Verteilungen gehört. Man setzt weiter voraus, dass man mit Sicherheit weiss, ob die Hypothese wahr oder falsch ist, sobald θ bekannt ist. Die Verteilungen von \mathcal{P} können dann in zwei Gruppen eingeteilt werden, nämlich einmal in die, für welche die Hypothese wahr, und in diejenige, für welche die

Hypothese falsch ist. Die resultierenden zwei disjunkten Gruppen werden mit H und K bezeichnet. Der „Parameterraum“ Θ spaltet sich analog in die jeweils dazugehörigen Räume Θ_H und Θ_K auf, so dass $H \cup K = \mathcal{P}$ und $\Theta_H \cup \Theta_K = \Theta$. Mit diesen Bezeichnungen gilt dann

$$H : \theta \in \Theta_H \quad \text{und} \quad K : \theta \in \Theta_K.$$

Mathematisch ist eine Hypothese H äquivalent zu der Aussage, dass P_θ ein Element aus H ist. Die Verteilungen, die in K liegen, werden mit „Alternativen“ zu H bezeichnet, d.h. K ist die Klasse der Alternativen. Das Problem ist nun, eine „Entscheidungsvorschrift“ zu finden, die einem angibt, welche Entscheidung zwischen Annahme und Ablehnung der Hypothese zu treffen ist. Mathematisch wird eine solche Entscheidungsvorschrift durch eine Funktion δ ausgedrückt, die zu jedem möglichen Wert x der Zufallsvariable X eine Entscheidung $d = \delta(x)$ zuordnet. δ ist also eine Funktion, deren Definitionsbereich alle möglichen Werte von X umfasst und deren Wertebereich aus den möglichen Entscheidungen besteht. Die Menge der Entscheidungen wird mit D abgekürzt und die Entscheidungen der Annahme oder Ablehnung der Hypothese H werden mit d_0 und d_1 bezeichnet, d.h. $D = \{d_0, d_1\}$. Die Funktion $d = \delta(x)$ kann nun so interpretiert werden, dass zu einem Wert x von X eine Entscheidungsregel $\delta(x)$ angewendet wird, die schliesslich zu der tatsächlichen Entscheidung d führt. Durch die Zuordnung einer Entscheidung zu jedem möglichen Wert x von X wird der Zustandsraum von X in zwei disjunkte Regionen S_0 und S_1 getrennt. Falls x in S_0 fällt, wird die Hypothese akzeptiert, ansonsten wird sie abgelehnt. S_0 wird dann mit „Annahmebereich“ bezeichnet, wogegen S_1 „Ablehnungsbereich“ oder „kritischer Bereich“ heisst. In der folgenden Definition ist dieser Sachverhalt noch einmal kurz zusammengefasst.

Definition 3.1.1 (Deterministischer, statistischer Test) *Es seien mit x die Realisierungen einer Zufallsvariablen X bezeichnet, deren Verteilung P_θ von einem unbekannten Parameter $\theta \in \Theta$ abhängt. Die Hypothese sei durch $H : \theta \in \Theta_H$, die Alternative durch $K : \theta \in \Theta_K$ gegeben. Ein nichtzufälliger bzw. deterministischer Test ist dann eine Entscheidungsfunktion $\delta(x)$ aus dem Zustandsraum der Zufallsvariable X in die Menge $D = \{d_0, d_1\}$ der Entscheidungen, derart dass gilt:*

$$\delta(x) = \begin{cases} d_0 & \text{falls } x \in S_0, \text{ d.h. } H \text{ wird akzeptiert} \\ d_1 & \text{falls } x \in S_1, \text{ d.h. } H \text{ wird abgelehnt.} \end{cases}$$

Bei einer Testdurchführung ist allerdings zu beachten, dass nicht immer die wahren Entscheidungen getroffen werden, d.h. es besteht die Gefahr von zwei möglichen Fehlentscheidungen.

(i) Fehler 1.Art:

Die Hypothese H wird abgelehnt, obwohl sie wahr ist. Die Wahrscheinlichkeit im Modell P_θ einen solchen Fehler zu begehen, wird mit α abgekürzt, d.h.

$$\alpha(\theta) = P_\theta(\text{Fehler 1.Art}).$$

(ii) Fehler 2.Art:

Die Hypothese H wird akzeptiert, obwohl sie falsch ist. In diesem Fall bezeichnet man die Wahrscheinlichkeit für eine Fehlentscheidung im Modell P_θ mit β , d.h.

$$\beta(\theta) = P_\theta(\text{Fehler 2.Art}).$$

Die Konsequenzen dieser Fehler sind oft sehr unterschiedlich. Möchte man z.B. prüfen, ob ein Patient an einer bestimmten Krankheit leidet, so könnte man

$$H : \text{Patient gesund} \quad \text{gegen} \quad K : \text{Patient krank}$$

testen. Der Fehler 1.Art ist dann, die Hypothese abzulehnen, obwohl sie wahr ist, d.h. man erklärt den Patienten für krank, obwohl er in Wahrheit gesund ist. Diese Fehlentscheidung würde dem Patienten höchstens Unannehmlichkeiten und eventuell einen finanziellen Verlust für eine Behandlung bringen, die nicht notwendig wäre. Wird dagegen der Patient für gesund gehalten, obwohl er krank ist, so entspricht dies dem Fehler 2.Art, da die Hypothese akzeptiert wird, obwohl sie falsch ist. Im schlimmsten Fall könnte diese Fehlentscheidung den Tod des Patienten zur Folge haben. Wünschenswert wäre daher ein Test, der die Wahrscheinlichkeit sowohl für den Fehler 1.Art, als auch für den Fehler 2.Art minimiert, jedoch können in der Realität bei gegebenen Beobachtungsdaten nicht beide Wahrscheinlichkeiten gleichzeitig kontrolliert werden. Deshalb ist es üblich, eine Schranke für die Wahrscheinlichkeit, einen Fehler 1.Art zu begehen, d.h. dass H zu Unrecht abgelehnt wird, anzugeben und unter dieser Bedingung die Fehlerwahrscheinlichkeiten 2.Art zu minimieren. Man wählt also eine Schranke α zwischen 0 und 1 und erhält somit für den Fehler 1.Art die Bedingung

$$P_\theta\{\delta(X) = d_1\} = P_\theta\{X \in S_1\} \leq \alpha \quad \forall \theta \in \Theta_H. \quad (3.1)$$

α heisst dabei das „*Signifikanzniveau*“. Damit wird die Betrachtung auf diejenigen Tests eingeschränkt, deren Wahrscheinlichkeit für eine Ablehnung der Hypothese kleiner oder gleich α für alle $\theta \in \Theta_H$ ist. Diese Tests werden dann „ *α -Level-Tests*“ genannt und man spricht von einer Ablehnung der Nullhypothese H zum Level α . Nun ist aber ein Test

zum Level α ebenso ein Test zum Level $\alpha' > \alpha$, denn wenn $P_\theta\{\delta(X) = d_1\} \leq \alpha \ \forall \theta \in \Theta_H$ erfüllt ist, so gilt auch $P_\theta\{\delta(X) = d_1\} \leq \alpha < \alpha' \ \forall \theta \in \Theta_H$. Daher ist es sinnvoll, den kleinsten Signifikanzlevel eines Tests auszuzeichnen. Dieser Wert wird dann mit „size“ eines Tests bezeichnet und ist offensichtlich die maximale Wahrscheinlichkeit des Fehlers 1. Art, d.h.

$$\text{size} = \sup_{\Theta_H} P_\theta\{\delta(X) = d_1\} = \sup_{\Theta_H} P_\theta\{X \in S_1\}.$$

Unter der Bedingung (3.1) wird nun versucht, die Fehlerwahrscheinlichkeiten 2. Art $P_\theta\{\delta(X) = d_0\} = 1 - P_\theta\{\delta(X) = d_1\}$ für $\theta \in \Theta_K$ zu minimieren, was äquivalent zur Maximierung von

$$P_\theta\{\delta(X) = d_1\} = P_\theta\{X \in S_1\} \quad \forall \theta \in \Theta_K \quad (3.2)$$

ist. Diese Grösse ist in der anschliessenden Definition noch genauer beschrieben.

Definition 3.1.2 (Power bzw. Güte und Powerfunktion) *Die Power oder auch Güte eines Tests gegen die Alternative θ ist die Wahrscheinlichkeit, die Hypothese H abzulehnen, wenn θ der wahre Wert ist. Damit gilt für die Power*

$$\text{Power an der Stelle } \theta \in \Theta_K = 1 - P_\theta(\text{Fehler 2. Art}).$$

Betrachtet man die Wahrscheinlichkeit in (3.2) als eine Funktion von θ , jedoch für alle θ aus dem gesamten Parameterraum Θ , so heisst

$$\begin{aligned} \beta(\theta, \delta) &= P_\theta(\text{Ablehnung der Hypothese } H) = P_\theta\{\delta(X) = d_1\} \\ &= \begin{cases} \text{Güte des Tests} & \text{für } \theta \in \Theta_K \\ \text{Wahrscheinlichkeit für Fehler 1. Art} & \text{für } \theta \in \Theta_H \end{cases} \end{aligned}$$

Power-Funktion. Es sind also sowohl die Güte eines Tests als auch die Wahrscheinlichkeit, einen Fehler 1. Art zu begehen, in der Power-Funktion enthalten.

Eine weitere, oft sehr nützliche und aussagekräftige statistische Grösse ist der sogenannte „p-Wert“, der als kleinstes Signifikanzniveau α definiert wird, zu dem die Hypothese gerade noch abgelehnt werden kann. Dieser Wert gibt einen Hinweis darauf, inwieweit die beobachteten Daten der Hypothese widersprechen.

Bisher wurde nur von nichtzufälligen bzw. deterministischen Tests ausgegangen, doch im folgenden wird nun mit Hilfe der sogenannten „kritischen Funktion“ die Struktur eines „zufälligen Tests“ eingeführt.

Definition 3.1.3 (Kritische Funktion Φ) Nach Definition 3.1.1 ist ein Test eine Entscheidungsregel $\delta : \mathcal{X} \times D \rightarrow D$, wobei \mathcal{X} den Zustandsraum der Zufallsvariable X beschreibt, und $D = \{d_0, d_1\}$ die Menge der Entscheidungen ist. Dann kann jedem Test eine umkehrbar eindeutig messbare Funktion

$$\Phi : \mathcal{X} \rightarrow D \quad \text{mit} \quad x \mapsto \delta(x, \{d_1\}) =: \Phi(x)$$

zugeordnet werden. Dabei gibt $\delta(x, \{d_1\}) =: \Phi(x)$ die Wahrscheinlichkeit an, bei der Beobachtung x die Nullhypothese H zu verwerfen und sich für die Alternative K zu entscheiden. Wird Φ auf diese Weise mit δ identifiziert, so kann ein Test als eine messbare Abbildung

$$\Phi : \mathcal{X} \rightarrow D$$

aufgefasst werden und man bezeichnet dann Φ als kritische Funktion.

Es ist zu bemerken, dass es nur eine Konvention ist, mit $\Phi(x)$ die Wahrscheinlichkeit für die Ablehnung und entsprechend mit $1 - \Phi(x)$ die Wahrscheinlichkeit für die Annahme der Hypothese zu bezeichnen. Falls Φ nur die Werte 0 und 1 annimmt, entspricht dies gerade einem nichtzufälligen Test, d.h. wenn die Realisierung x von X vorliegt und Φ durch

$$\Phi(x) = \begin{cases} 1 & \text{falls } x \in S_1 \\ 0 & \text{falls } x \in S_0, \end{cases}$$

gegeben ist, dann wird die Hypothese für $\Phi(x) = 1$ abgelehnt, während sie für $\Phi(x) = 0$ akzeptiert wird. Die Menge der Punkte x , für die $\Phi(x) = 1$ ist, ist dann gerade der Ablehnungsbereich, so dass in nichtzufälligen Tests Φ genau der Indikatorfunktion des kritischen Bereichs entspricht. Um den Unterschied zwischen einem deterministischen und einem zufälligen Test zu verdeutlichen, betrachte man folgendes Beispiel:

Beispiel 3.1.4 X sei eine normalverteilte Zufallsvariable. Es soll getestet werden, ob der Erwartungswert μ der Zufallsvariable X einen bestimmten vorgegebenen Wert μ_0 übersteigt, d.h. der Entscheidungsraum ist gegeben durch $D = \{d_0, d_1\}$, wobei d_0 die Entscheidung für die Hypothese und d_1 die Entscheidung für die Alternative bezeichnet. Getestet wird die Hypothese $H : \mu > \mu_0$ gegen die Alternative $K : \mu \leq \mu_0$. Weiter seien x_1, x_2, \dots, x_m Realisierungen von X mit empirischem Mittelwert $\bar{x} = \frac{1}{m} \sum_{i=1}^m x_i$. Ein möglicher deterministischer Test ist dann durch

$$\delta(x) = \begin{cases} d_1 & \text{falls } \bar{x} \leq \mu_0, \text{ d.h. } H \text{ wird abgelehnt} \\ d_0 & \text{falls } \bar{x} > \mu_0, \text{ d.h. } H \text{ wird akzeptiert} \end{cases}$$

gegeben. Es wird hier vorausgesetzt, dass abhängig von den Realisierungen von X eine konkrete Entscheidung getroffen wird. Bei einem zufälligen Test dagegen wird d_1 , d.h. die Ablehnung der Hypothese nur mit bestimmten Wahrscheinlichkeiten $\Phi(x)$ gewählt. Ein möglicher zufälliger Test könnte zum Beispiel durch

$$\Phi(x) = \begin{cases} \frac{3}{4} & \text{falls } \bar{x} < \mu_0 \\ \frac{1}{2} & \text{falls } \bar{x} = \mu_0 \\ \frac{1}{4} & \text{falls } \bar{x} > \mu_0 \end{cases}$$

gegeben sein. Dies bedeutet, dass man sich im Fall $\bar{x} < \mu_0$ im Gegensatz zum deterministischen Test nicht sicher für die Ablehnung der Hypothese entscheidet, sondern nur mit einer Wahrscheinlichkeit von $\frac{3}{4}$. Im Grenzfall $\bar{x} = \mu_0$ wird die Hypothese H nur mit einer Wahrscheinlichkeit von $\frac{1}{2}$ verworfen, während man sie nur mit einer Wahrscheinlichkeit von $\frac{1}{4}$ im Fall $\bar{x} > \mu_0$ ablehnt.

Ist die Verteilung von X gegeben durch P_θ und wird die kritische Funktion Φ verwendet, so ist die Wahrscheinlichkeit für eine Ablehnung der Hypothese durch

$$E_\theta \Phi(X) = \int \Phi(x) dP_\theta(x) \quad (3.3)$$

gegeben. Die Schwierigkeit ist, Φ so zu wählen, dass zum einen die Power

$$\beta_\Phi(\theta) = E_\theta \Phi(X) \quad \forall \theta \in \Theta_K \quad (3.4)$$

maximiert wird, und zum anderen gleichzeitig die Schranke für die Wahrscheinlichkeit des Fehlers 1. Art

$$E_\theta \Phi(X) \leq \alpha \quad \forall \theta \in \Theta_H \quad (3.5)$$

nicht überschritten wird. Ein damit verbundenes Problem ist, dass sich zu jedem statistischen Test viele Entscheidungsfunktionen $\Phi(x)$ finden lassen, die diese Bedingungen erfüllen. Die Aufgabe der mathematischen Statistik ist es daher, zu einem vorliegenden Entscheidungsproblem eine optimale Lösung anzugeben. Das Ziel ist schliesslich, einen „gleichmässig optimalen Test“ zu konstruieren, der unabhängig von der Alternative die Power maximiert. Diese Tests werden im nachfolgenden Abschnitt noch genauer untersucht. Man spricht dagegen nur von einem „optimalen Test“, wenn der Test die Power maximiert, aber nur unter Betrachtung einer bestimmten Alternative, d.h. wenn die Alternative einfach ist.

Definition 3.1.5 (Einfache und zusammengesetzte Hypothesen) Falls der Parameterraum Θ_H der Hypothese nur aus einem einzigen Wert θ_0 besteht, d.h.

$\Theta_H = \{\theta_0\}$, so heissen Θ_H und die Hypothese H einfach. Enthält dagegen die der Hypothese zugeordnete Teilmenge Θ_H von Θ mehrere Werte, so heissen Θ_H und die Hypothese H zusammengesetzt.

Die gleichen Konventionen gelten ebenfalls für den Teilraum Θ_K und die Alternativen K .

Ein optimaler Test für H gegen K zum Niveau α lässt sich häufig so bestimmen, dass man aus K einen speziellen Wert $\theta_1 \in \Theta_K$ herausgreift und zunächst einen optimalen Test Φ^* für H gegen die einfache Alternative $\Theta_K = \{\theta_1\}$ ermittelt.

Definition 3.1.6 (Optimaler (most-powerful) Test) Ein α -level Test Φ^* heisst optimal (most powerful), wenn er unter Betrachtung einer bestimmten Alternative $\Theta_K = \{\theta_1\}$ die Power maximiert, d.h.

$$\beta_{\Phi^*}(\theta_1) \geq \beta_{\Phi}(\theta_1)$$

für jeden beliebigen α -level Test Φ mit Alternative Θ_K .

Die Hypothesentests, für die K einfach ist, sind dann vollständig durch (3.4) und (3.5) spezifiziert. Die Lösung ist in diesen Fällen relativ einfach zu erhalten und kann sogar explizit angegeben werden, wenn auch H einfach ist.

3.1.2 Most-Powerful-Tests für einfache Hypothesen und Alternativen

Es seien die Verteilungen einer Zufallsvariablen X mit Realisierungen x unter einer einfachen Hypothese $H : \theta = \theta_0$ und Alternative $K : \theta = \theta_1$ durch P_0 und P_1 gegeben. Ausserdem nehme man an, dass diese Verteilungen diskret seien mit $P_i(X = x) = P_i(x)$ für $i = 0, 1$.

Betrachtet man nun zunächst nichtzufällige Tests, so ist ein optimaler Test definiert durch den Ablehnungsbereich S_1 , der

$$P_0\{X \in S_1\} = \sum_{x \in S_1} P_0(x) \leq \alpha \quad \text{und} \quad (3.6)$$

$$P_1\{X \in S_1\} = \sum_{x \in S_1} P_1(x) = \text{maximum} \quad (3.7)$$

erfüllt. Die Gleichung (3.6) bedeutet, dass die Wahrscheinlichkeit des Fehlers 1. Art, d.h. die Ablehnung der Hypothese ($x \in S_1$), obwohl sie wahr ist ($P_0(x)$), durch α beschränkt ist. Dagegen beschreibt die Gleichung (3.7) die Tatsache, dass die Power, d.h. die richtige Ablehnung der Hypothese, maximiert werden soll. Zu jeder Realisierung x gehören also zwei Werte, nämlich einmal die Wahrscheinlichkeit unter P_0 und einmal unter P_1 . Dabei dürfen die Werte x , die für S_1 gewählt werden, einerseits einen bestimmten Wert α nicht überschreiten, andererseits müssen sie zu einem möglichst hohen Wert führen. Solche Situationen tauchen häufig auf, wenn man z.B. mit einem begrenzten Budget versucht, einen möglichst hohen Gewinn zu erzielen. Eine andere Situation ist, wenn man in einer begrenzten Zeit versucht, eine möglichst weite Strecke zu bewältigen. In einer solchen Situation würde man sich dasjenige Transportmittel wählen, das die meisten km/Std zurücklegt. Analog wählt man im obigen Testproblem diejenigen Realisierungen x für S_1 , die zu einem grossen Wert von

$$r(x) = \frac{P_1(x)}{P_0(x)}$$

führen. Es werden also die Realisierungen der Reihe nach bezüglich des grössten Wertes von $r(x)$ sortiert und in S_1 aufgenommen, solange sie die Bedingung (3.6) erfüllen.

Formal bedeutet dies, dass S_1 aus denjenigen Punkten x besteht, für die $r(x) > c$ gilt, wobei c bestimmt wird durch die Gleichung

$$P_0\{X \in S_1\} = \sum_{x:r(x)>c} P_0(x) = \alpha.$$

Hier kann es zu dem Problem kommen, dass bei einem bestimmten Wert x das Signifikanzniveau α noch nicht erreicht ist, aber mit der Hinzunahme des bezüglich der Sortierung nachfolgenden x -Wertes α überstiegen wird. Es kann also passieren, dass der Wert α nicht exakt angenommen wird und damit hätte das Optimierungsproblem keine explizite Lösung.

Diese Schwierigkeit kann umgangen werden, wenn man nicht von deterministischen, sondern von zufälligen Tests ausgeht. Das Fundamentallemma von Neyman und Pearson, welches in Lehmann (1986, S. 74) [14] nachgelesen werden kann, beschreibt diesen Sachverhalt. Das Lemma geht von dem Fall aus, dass sowohl die Hypothese als auch die Alternative einfach sind. Dies ist zwar eher nur von theoretischem Interesse, doch ist die Lösung solcher Probleme von grosser Bedeutung in Bezug auf die Herleitung von gleichmässig optimalen (uniformly most powerful) Tests, auf die im folgenden Abschnitt genauer eingegangen wird.

Satz 3.1.7 (Fundamentallemma von Neyman-Pearson) Seien P_0 und P_1 Verteilungen mit dazugehörigen Dichten p_0 und p_1 bezüglich eines Masses μ . Dann gilt:

(i) *Existenz:*

Für die Hypothese $H : p_0$ gegen die Alternative $K : p_1$ gibt es einen Test Φ und eine Konstante k , derart, dass

$$E_0 \Phi(X) = \alpha \quad \text{und} \quad (3.8)$$

$$\Phi(X) = \begin{cases} 1 & \text{falls } \frac{p_1(x)}{p_0(x)} > k \\ 0 & \text{falls } \frac{p_1(x)}{p_0(x)} < k. \end{cases} \quad (3.9)$$

(ii) *Hinreichende Bedingung für einen most-powerful-Test:*

Falls ein Test die Gleichungen (3.8) und (3.9) für ein k erfüllt, dann ist er optimal (most-powerful) für den Test p_0 gegen p_1 zum Level α .

(iii) *Notwendige Bedingung für einen most-powerful-Test:*

Falls Φ ein optimaler Test zum Level α für den Test $H : p_0$ gegen $K : p_1$ ist, dann erfüllt er für ein k auch die Gleichung (3.9) fast überall bezüglich des Masses μ . Er erfüllt auch die Gleichung (3.8), ausgenommen wenn es einen Test mit $\text{size} < \alpha$ und mit Power 1 gibt.

Beweis: Für die Spezialfälle $\alpha = 0$ und $\alpha = 1$ ist leicht zu zeigen, dass ein solcher Test existiert, vorausgesetzt, der Wert $k = \infty$ wird zugelassen, denn $\alpha = 0$ bedeutet, dass die Wahrscheinlichkeit, H abzulehnen, obwohl H wahr ist, gleich 0 ist, d.h. H wird nie abgelehnt. Damit muss

$$E_0 \Phi(x) = P \left\{ \frac{p_1(x)}{p_0(x)} > k \right\} \stackrel{!}{=} 0$$

gelten, was für $k = \infty$ zutrifft. Analog kann für den Fall $\alpha = 1$ argumentiert werden, nur dass hier

$$E_0 \Phi(x) = P \left\{ \frac{p_1(x)}{p_0(x)} < k \right\} \stackrel{!}{=} 1$$

gelten muss, was mit einem sehr kleinen k gewährleistet ist. Im folgenden Beweis sei also $0 < \alpha < 1$ vorausgesetzt.

(i) Sei nun $\alpha(c) = P_0 \left\{ \frac{p_1(X)}{p_0(X)} > c \right\}$ gesetzt. Da diese Wahrscheinlichkeit nur unter P_0 betrachtet wird, d.h. unter der Voraussetzung, dass H wahr ist, braucht man nur die Fälle $p_0(x) > 0$ zu betrachten. $\alpha(c)$ ist also die Wahrscheinlichkeit, dass die Zufallsvariable $\frac{p_1(X)}{p_0(X)}$ den Wert c übersteigt. Es gilt

$$\alpha(c) = P_0 \left\{ \frac{p_1(X)}{p_0(X)} > c \right\} = 1 - P_0 \left\{ \frac{p_1(X)}{p_0(X)} \leq c \right\}, \quad (3.10)$$

wobei $P_0\{\frac{p_1(X)}{p_0(X)} \leq c\}$ eine Wahrscheinlichkeitsfunktion darstellt und damit monoton wachsend und rechtsseitig stetig ist. Aufgrund (3.10) ist auch $\alpha(c)$ rechtsseitig stetig und aufgrund $1 - P_0\{\frac{p_1(X)}{p_0(X)} \leq c\}$ ist $\alpha(c)$ nicht wie die Wahrscheinlichkeitsfunktion monoton wachsend, sondern monoton fallend. Insgesamt ist also $\alpha(c)$ eine nichtwachsende, rechtsseitig stetige Funktion mit $\alpha(-\infty) = 1$ und $\alpha(\infty) = 0$. Wegen der Rechtsstetigkeit gilt

$$\alpha(c-0) - \alpha(c) = P_0\left\{\frac{p_1(X)}{p_0(X)} = c\right\}.$$

Für ein gegebenes α mit $0 < \alpha < 1$ sei nun ein c_0 so gewählt, dass $\alpha(c_0) \leq \alpha \leq \alpha(c_0 - 0)$ gilt. Dann betrachte man den zufälligen Test

$$\Phi(x) = \begin{cases} 1 & \text{falls } \frac{p_1(x)}{p_0(x)} > c_0 \\ \frac{\alpha - \alpha(c_0)}{\alpha(c_0 - 0) - \alpha(c_0)} & \text{falls } \frac{p_1(x)}{p_0(x)} = c_0 \\ 0 & \text{falls } \frac{p_1(x)}{p_0(x)} < c_0. \end{cases} \quad (3.11)$$

Der mittlere Ausdruck ist sinnvoll definiert, solange $\alpha(c_0 - 0) \neq \alpha(c_0)$ ist. Da nun aber $P_0\{\frac{p_1(X)}{p_0(X)} = c_0\} = 0$, ist der Test fast überall definiert.

Die size des Tests berechnet man aus

$$\begin{aligned} E_0\Phi(X) &= P_0\left\{\frac{p_1(X)}{p_0(X)} > c_0\right\} + \frac{\alpha - \alpha(c_0)}{\alpha(c_0 - 0) - \alpha(c_0)} P_0\left\{\frac{p_1(X)}{p_0(X)} = c_0\right\} \\ &= P_0\left\{\frac{p_1(X)}{p_0(X)} > c_0\right\} + \frac{\alpha - P_0\left\{\frac{p_1(X)}{p_0(X)} > c_0\right\}}{P_0\left\{\frac{p_1(X)}{p_0(X)} = c_0\right\}} P_0\left\{\frac{p_1(X)}{p_0(X)} = c_0\right\} = \alpha. \end{aligned}$$

Das c_0 kann schliesslich durch das k im Lemma ersetzt werden und somit ist die Existenz eines Tests, wie im Lemma angegeben, gezeigt.

- (ii) Angenommen, Φ sei ein Test, der (3.8) und (3.9) erfüllt und Φ^* sei ein beliebiger anderer Test mit $E_0\Phi^*(X) \leq \alpha$. Mit S^+ und S^- seien die Mengen im Zustandsraum von X bezeichnet, für die $\Phi(x) - \Phi^*(x) > 0$ bzw. < 0 ist. Falls x in S^+ liegt, folgt damit, dass $\Phi(x) > 0$ gilt und es ist $p_1(x) \geq kp_0(x)$. Analog folgt $p_1(x) \leq kp_0(x)$ für alle x , die in S^- liegen. Damit ergibt sich für die Differenz der Wahrscheinlichkeiten der Ablehnung der beiden Tests nach (3.3)

$$\int (\Phi(x) - \Phi^*(x))(p_1(x) - kp_0(x)) d\mu(x) = \int_{S^+ \cup S^-} (\Phi(x) - \Phi^*(x))(p_1(x) - kp_0(x)) d\mu(x) \geq 0,$$

denn für x in S^+ ist $\Phi(x) - \Phi^*(x) > 0$ und $p_1(x) - kp_0(x) \geq 0$, während für x in S^- $\Phi(x) - \Phi^*(x) < 0$ und gleichzeitig aber auch $p_1(x) - kp_0(x) \leq 0$ gilt, was

insgesamt wieder zu einem Integral führt, das grösser oder gleich 0 ist.

Spaltet man nun das obige Integral nach $p_1(x)$ und $p_0(x)$ auf, so erhält man

$$\int (\Phi(x) - \Phi^*(x))p_1(x) d\mu(x) \geq k \int (\Phi(x) - \Phi^*(x))p_0(x) d\mu(x) \geq 0.$$

Das linke Integral in dieser Ungleichung entspricht der Differenz der Power bei der Tests, womit gezeigt ist, dass die Power von Φ unter Betrachtung der Alternative p_1 immer grösser als die Power eines beliebigen Tests Φ^* ist, d.h. $\beta_\Phi(p_1) > \beta_{\Phi^*}(p_1)$. Somit ist der Test Φ most-powerful, was zu beweisen war.

(iii) Dieser Teil des Lemmas wird mit Hilfe eines Widerspruchsbeweises gezeigt.

Sei Φ^* most-powerful für $H : p_0$ gegen $K : p_1$ zum Level α und sei Φ ein weiterer Test, der (3.8) und (3.9) erfüllt. Sei S nun die Schnittmenge von $S^+ \cup S^-$ (auf der sich Φ und Φ^* unterscheiden) mit der Menge $\{x : p_1(x) \neq kp_0(x)\}$.

Angenommen, es sei $\mu(S) > 0$, dann gilt wegen $(\Phi(x) - \Phi^*(x))(p_1(x) - kp_0(x)) > 0$ auf S

$$\int_S (\Phi(x) - \Phi^*(x))(p_1(x) - kp_0(x)) d\mu(x) > 0.$$

Dies würde aber bedeuten, dass die Power von Φ grösser als die von Φ^* bezüglich des Tests p_0 gegen p_1 wäre, was ein Widerspruch dazu wäre, dass Φ^* ein most-powerful-Test ist. Die Annahme $\mu(S) > 0$ kann also nicht aufrecht erhalten bleiben und es folgt $\mu(S) = 0$, d.h. es gibt keine Menge mit Mass $\mu > 0$, auf der sich Φ und Φ^* voneinander unterscheiden. Damit ist der most-powerful-Test Φ^* auch genau der Test, der die Gleichungen (3.8) und (3.9) erfüllt.

Der Beweis von Teil iii) dieses Lemmas zeigt, dass der most-powerful-Test durch (3.8) und (3.9) eindeutig bestimmt ist, ausser auf der Menge, für die $p_1(x) = kp_0(x)$ gilt. Auf dieser Menge kann Φ beliebig gewählt werden, vorausgesetzt, dass der resultierende Test eine size = α hat. Es wurde bisher gezeigt, dass es immer möglich ist, Φ über eine beschränkte Menge konstant zu wählen. Im einfachsten Fall existiert ein Test mit Power gleich 1, wobei die Konstante k aus (3.9) gleich 0 ist und man akzeptiert die Hypothese H für alle Punkte, für die $p_1(x) = kp_0(x)$ ist, auch wenn der Test dann möglicherweise eine size $< \alpha$ hat.

Es folgt also abschliessend, dass der most-powerful-Test eindeutig (bis auf Nullmengen) durch (3.8) und (3.9) definiert ist, wenn immer die Menge, für die $p_1(x) = kp_0(x)$ gilt, das μ -Mass Null hat. Dieser Test ist dann deterministisch. Die Einführung von Zufälligkeit ist nicht erforderlich, ausser möglicherweise auf

der Menge, für die $p_1(x) = kp_0(x)$ gilt, auf der Zufälligkeit notwendig ist, um einen Test mit size genau gleich α zu erhalten. Denn wenn ein Test mit Power=1 existiert, dann bestimmen (3.8) und (3.9) einen most-powerful-Test, der aber nicht unbedingt eindeutig sein muss in dem Sinn, dass es einen weiteren most-powerful-Test gibt, der (3.8) und (3.9) erfüllt, aber mit einem $\alpha' < \alpha$.

□

3.1.3 Uniformly-most-powerful (UMP)-Tests

Bisher wurde nur der Spezialfall von einfachen Hypothesen und Alternativen betrachtet, wobei dies allerdings eher nur von theoretischem Interesse ist, denn in der Praxis treten häufig zusammengesetzte Hypothesen auf. In solchen Fällen wird ein Test optimal, wenn er die Power unabhängig von der Alternative K maximiert.

Definition 3.1.8 (Gleichmässig optimaler (UMP)-Test) Ein α -Level-Test Φ^* ist ein gleichmässig optimaler bzw. ein uniformly most powerful-Test (UMP) für die Hypothese $H : \theta \in \Theta_H$ gegen die Alternative $K : \theta \in \Theta_K$, genau dann, wenn

$$\beta_{\Phi^*}(\theta) \geq \beta_{\Phi}(\theta) \quad \forall \theta \in \Theta_K$$

für jeden beliebigen α -Level-Test Φ .

In vielen Anwendungsfällen hängen die Verteilungen von nur einem einzigen reellwertigen Parameter θ ab und die Hypothese ist einseitig, d.h. die Hypothese könnte z.B. $H : \theta \leq \theta_0$ lauten. Der optimale Test von H gegen eine Alternative $K : \theta = \theta_1$ mit $\theta_1 > \theta_0$ hängt dann von θ_1 ab und ist somit nicht mehr gleichmässig optimal (UMP). Man kann jedoch in Fällen von einseitigen Hypothesen und Alternativen einen UMP-Test konstruieren, wenn eine zusätzliche Bedingung erfüllt ist. Dies führt zu Verteilungen mit monotonem Likelihood-Quotienten.

Definition 3.1.9 (Verteilungen mit monotonem Likelihood-Quotienten) Eine Familie von Dichten $p_{\theta}(x)$ mit reellen Parametern hat einen monotonen Likelihood-Quotienten, falls eine reellwertige Funktion $T(x)$ existiert, derart, dass für jedes $\theta < \theta'$ die Verteilungen P_{θ} und $P_{\theta'}$ unterschiedlich sind und der Quotient $\frac{p_{\theta'}(x)}{p_{\theta}(x)}$ eine nichtfallende Funktion von $T(x)$ ist.

Nach Lehmann (1986, S. 78) [14] kann nun mit Hilfe dieser Definition ein UMP-Test konstruiert werden, was im nachstehenden Satz erläutert wird.

Satz 3.1.10 (UMP-Test für einseitiges Testproblem) Sei θ ein reeller Parameter und die Zufallsvariable X habe eine Dichte bzw. Wahrscheinlichkeitsfunktion $p_\theta(x)$ mit monotonem Likelihood-Quotienten in $T(x)$. Dann gilt:

(i) Für den Test $H : \theta \leq \theta_0$ gegen $K : \theta > \theta_0$ existiert ein UMP-Test, der gegeben ist durch

$$\Phi(x) = \begin{cases} 1 & \text{für } T(x) > C, \\ \gamma & \text{für } T(x) = C, \\ 0 & \text{für } T(x) < C, \end{cases} \quad (3.12)$$

wobei C und γ durch $E_{\theta_0} \Phi(X) = \alpha$ bestimmt werden.

(ii) Die Power-Funktion

$$\beta(\theta) = E_\theta \Phi(X) \quad (3.13)$$

dieses Tests ist strikt monoton wachsend für alle θ für die $0 < \beta(\theta) < 1$ gilt.

Für den Beweis dieses Satzes wird ein zusätzliches Korollar benötigt, welches hier allerdings nur zitiert, aber nicht bewiesen werden soll. Der Beweis dazu lässt sich in Lehmann (1986, S. 76) [14] nachlesen.

Korollar 3.1.11 Es seien P_0 und P_1 Verteilungsfunktionen mit dazugehörigen Dichte- bzw. Wahrscheinlichkeitsfunktionen p_0 und p_1 . Weiter sei mit β die Power des most-powerful-Tests $H : P_0$ gegen $K : P_1$ zum Level α bezeichnet. Dann gilt $\alpha < \beta$, ausser wenn $P_0 = P_1$.

Beweis von Satz 3.1.10:

(i) und (ii): Man betrachte zuerst das Testen der einfachen Hypothese $H : \theta = \theta_0$ gegen die einfache Alternative $K : \theta = \theta_1$ mit $\theta_1 > \theta_0$, wobei θ_1 ein beliebiger Wert aus dem Parameterraum Θ_K der Alternativen ist. Zu diesem Testproblem gibt es nach dem Fundamentallemma einen optimalen Test, nämlich (3.11) mit $E_{\theta_0}(\Phi(X)) = \alpha$. Aufgrund der Voraussetzung der Monotonie des Likelihood-Quotienten $\frac{p_{\theta_1}(x)}{p_{\theta_0}(x)}$ ist dieser Test mit (3.12), wobei $E_{\theta_0}(\Phi(X)) = \alpha$ gilt, äquivalent. Dabei werden C und γ durch $E_{\theta_0} \Phi(x) = P_{\theta_0}(T(x) > C) + \gamma P_{\theta_0}(T(x) = C) = \alpha$, also unabhängig von $\theta_1 \in \Theta_K$, bestimmt. Somit ist der angegebene Test (3.12) für das Testproblem mit einfacher Hypothese und Alternative ein UMP-Test. Für den Beweis des allgemeinen Falls $H : \theta \leq \theta_0$ gegen $K : \theta > \theta_0$ wird die Aussage

(ii) benötigt, die aber direkt aus dem Korollar 3.1.11 gefolgert werden kann. Da also angenommen werden kann, dass $\beta(\theta)$ eine nichtfallende Funktion ist, erfüllt der Test

$$\beta(\theta) = E_{\theta}\Phi(X) \leq \alpha \quad \text{für } \theta \leq \theta_0. \quad (3.14)$$

Nun gilt aber, dass die Klasse der Tests, die (3.14) erfüllen, in der Klasse derjenigen Tests enthalten ist, die

$$\beta(\theta_0) = E_{\theta_0}\Phi(X) \leq \alpha \quad (3.15)$$

erfüllen, denn wegen der Monotonie von $\beta(\theta)$ folgt aus (3.15) direkt (3.14). Für eine beliebige Alternative $\theta_1 \in \Theta_K$ maximiert der gegebene Test Φ aus (3.12) nach dem ersten Teil dieses Beweises die Power $\beta(\theta_1)$ innerhalb der Klasse (3.15), womit gleichzeitig $\beta(\theta_1)$ ebenfalls innerhalb der kleineren Klasse von Tests (3.14) maximiert wird. Da aber die Maximierung der Power innerhalb der Klasse (3.14) unabhängig von der speziell gewählten Alternative θ_1 erfolgt, ist gezeigt, dass der angegebene Test auch ein UMP-Test für die Hypothese $H : \theta \leq \theta_0$ gegen $K : \theta > \theta_0$ ist.

□

Eine wichtige Bemerkung ist, dass man durch Vertauschen aller Ungleichungen im obigen Satz in analoger Weise die Lösung für das sogenannte „duale Problem“

$$H : \theta \geq \theta_0 \quad \text{gegen} \quad K : \theta < \theta_0 \quad (3.16)$$

erhält.

3.2 Bereichstest für Überdispersion

Ausgehend von den bisher hergeleiteten allgemeinen Hypothesentests kann nun ein spezieller Test bezüglich Überdispersion konstruiert werden.

Zunächst soll aber die Motivation, die zu diesem Test führt, kurz erläutert werden:

In vielen Anwendungsbereichen ist es gängig, gegebene Zählraten mit Hilfe der Poisson-regression zu modellieren, obwohl bekannt ist, dass die Daten Überdispersion aufweisen. Der Grund, weshalb man oft bei dem Poissonmodell bleibt, ist, dass es wegen seiner

Struktur her einfacher zu handhaben ist und eine weniger komplexe Wahrscheinlichkeitsfunktion besitzt als die NB-Verteilung, die in solchen Fällen eigentlich verwendet werden sollte, um die Überdispersion zu berücksichtigen. Zusätzlich ist die Tatsache, dass für die Modellierung der Daten mittels der Poissonverteilung Standard-Software-Pakete zur Verfügung stehen, was für die Negativ-Binomial-Verteilung nicht zutrifft, ein wesentlicher Grund, die Poissonverteilung als Modellgrundlage zu bevorzugen. Man vernachlässigt also die Tatsache, dass die Daten Überdispersion aufweisen zu Gunsten einer einfacheren Modellierung.

Die Frage, die man sich dann stellen kann, ist, ab wann der Aufwand gerechtfertigt wäre, das Poissonmodell aufgrund zu hoher Überdispersion zurückzuweisen und stattdessen die Daten mit der NB-Verteilung zu modellieren.

Um dies zu untersuchen, macht man sich den Zusammenhang zwischen der Poisson- und der NB-Verteilung zu Nutzen, denn wie in Satz 2.3.8 gezeigt, ist die Poissonverteilung ein Grenzfall der NB-Verteilung, nämlich genau dann, wenn der Überdispersionsparameter $a = 0$ ist. Aus diesem Grund versucht man, ein leicht interpretierbares sogenanntes „*Abstandsmass*“ zu wählen, welches den Überdispersionsparameter a mit dem Spezialfall $a = 0$ in Beziehung bringt, um Aussagen darüber zu erhalten, wie gross der Abstand zum Poissonmodell ist. Die Möglichkeiten der Wahl eines Abstandsmasses werden im nächsten Abschnitt erläutert.

Um die obige Frage beantworten zu können, wird ein „*Bereichstest*“ erstellt, der die Hypothese $H : a > a_0$ gegen die Alternative $K : a \leq a_0$ testet, wobei a_0 eine beliebige vorgegebene Schranke für a darstellt. Ein Bereichstest und die dazugehörigen p-Wert-Kurven, die im folgenden Abschnitt definiert werden, geben nicht nur Informationen über Ablehnung oder Annahme der Hypothese, sondern sie ermöglichen auch die Quantifizierung der Abweichung vom Poissonmodell. Das Ziel ist also, ausgehend von dem gewählten Abstandsmass p-Wert-Kurven zu erstellen, die auf dem Bereichstest basieren. Die p-Wert-Kurven geben einem schliesslich Auskunft darüber, welchen Wert des Abstandsmasses man akzeptieren müsste, um das Poissonmodell bei einem Signifikanzniveau α beibehalten zu können.

Im folgenden wird nun der Bereichstest konstruiert, während die Simulationsstudie mit den p-Wert-Kurven im nächsten Kapitel ausgeführt wird.

Zur Erstellung eines asymptotischen Bereichstests für $H : a > a_0$ gegen $K : a \leq a_0$ benötigt man die asymptotische Eigenschaft des ML- Schätzers $(\hat{\beta}, \hat{a})$ im NB-Modell.

Nach (2.33) in Abschnitt 2.3.3 gilt

$$\sqrt{n}(\hat{\beta} - \beta, \hat{a} - a) \stackrel{a}{\sim} N_{p+1}(\mathbf{0}, nFI(\hat{\beta}, \hat{a})^{-1}),$$

was gleichbedeutend ist mit

$$FI(\hat{\beta}, \hat{a})^{\frac{1}{2}}(\hat{\beta} - \beta, \hat{a} - a) \stackrel{\mathcal{D}}{\rightarrow} N_{p+1}(\mathbf{0}, I_{p+1}).$$

Die Einträge der Fisher-Informationsmatrix $FI(\hat{\beta}, \hat{a})$ wurden ebenfalls im Abschnitt 2.3.3 hergeleitet.

Satz 3.2.1 (Bereichstest für Überdispersion) *Unter der Annahme, dass $FI(\hat{\beta}, \hat{a})^{\frac{1}{2}}(\hat{\beta} - \beta, \hat{a} - a) \stackrel{\mathcal{D}}{\rightarrow} N_{p+1}(\mathbf{0}, I_{p+1})$ gilt, ist ein asymptotischer α -Level-Test für $H : a > a_0$ gegen $K : a \leq a_0$ gegeben durch den Ablehnungsbereich*

$$\mathcal{C} = \left\{ \hat{a} : \Phi \left(\frac{\hat{a} - a_0}{\hat{\sigma}(\hat{\beta}, \hat{a})} \right) \leq \alpha \right\}, \quad (3.17)$$

wobei $\hat{\sigma}^2$ einen konsistenten Schätzer von σ^2 beschreibt. Ein solcher konsistenter Schätzer ist z.B. durch den $(p+1, p+1)$ -Eintrag in der Inversen der Fisher-Informationsmatrix gegeben.

Beweis: Zunächst muss gezeigt werden, dass ein monotoner Likelihood-Quotient gemäss Definition 3.1.9 vorliegt. Nach Voraussetzung des Satzes geht man davon aus, dass der Schätzer \hat{a} normalverteilt ist, d.h. $\hat{a} \sim N(a, \hat{\sigma})$, wobei $\hat{\sigma}$ der $p+1, p+1$ -Eintrag in der Inversen der Fisher-Informationsmatrix darstellt und hier als bekannt vorausgesetzt wird. Der einzige Parameter, der die Verteilung von \hat{a} bestimmt, ist damit a . Sei also $a_1 > a_2$ gewählt. Dann gilt für den Dichtequotienten

$$\begin{aligned} \frac{p_{a_1}(\hat{a})}{p_{a_2}(\hat{a})} &= \frac{\frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{1}{2\sigma}(\hat{a} - a_1)^2\}}{\frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{1}{2\sigma}(\hat{a} - a_2)^2\}} \\ &= \exp \left\{ -\frac{1}{2\sigma} ((\hat{a} - a_1)^2 - (\hat{a} - a_2)^2) \right\} \\ &= \exp \left\{ -\frac{1}{2\sigma} (\hat{a}^2 - 2\hat{a}a_1 + a_1^2 - \hat{a}^2 + 2\hat{a}a_2 - a_2^2) \right\} \\ &= \exp \left\{ \frac{1}{2\sigma} (2\hat{a}(a_1 - a_2) + a_2^2 - a_1^2) \right\}. \end{aligned}$$

Da wegen $a_1 > a_2$ die Differenz $a_1 - a_2 > 0$ gilt, ist der errechnete Dichtequotient eine wachsende Funktion in \hat{a} . Nun kann daher Satz 3.1.10 angewendet werden, da ein monotoner Likelihood-Quotient in \hat{a} vorliegt. Es wird allerdings hier zum Beweis des

Bereichstests das duale Problem (3.16) benötigt, da der Bereichstest die Nullhypothese $H : a > a_0$ gegen die Alternative $K : a \leq a_0$ testet. Der erste Teil des Satzes 3.1.10(i) formuliert sich dann um zu:

(i*) Für den Test $H : \theta > \theta_0$ gegen $K : \theta \leq \theta_0$ existiert ein UMP-Test, der gegeben ist durch

$$\Phi(x) = \begin{cases} 1 & \text{für } T(x) < C, \\ \gamma & \text{für } T(x) = C, \\ 0 & \text{für } T(x) > C, \end{cases}$$

wobei C und γ durch $E_{\theta_0}\Phi(x) = \alpha$ bestimmt werden. Damit ergibt sich für den Bereichstest $H : a > a_0$ gegen $K : a \leq a_0$ ein Ablehnungsbereich von

$$\mathcal{C} = \{\hat{a} : \hat{a} < C\}.$$

Das C wird hier durch die Gleichung

$$\hat{P}(C, a_0, \hat{\sigma}) = \Phi\left(\frac{C - a_0}{\hat{\sigma}}\right) = \alpha \quad (3.18)$$

bestimmt, wobei $\Phi(\cdot)$ in diesem Fall die Standardnormalverteilung bedeutet. Da die Funktion $\Phi\left(\frac{C - a_0}{\hat{\sigma}}\right)$ in C monoton fällt, wenn C kleiner wird, ist (3.18) äquivalent zum Ablehnungsbereich

$$\mathcal{C} = \{\hat{a} : \Phi\left(\frac{\hat{a} - a_0}{\hat{\sigma}(\hat{\beta}, \hat{a})}\right) \leq \alpha\},$$

wie im Satz behauptet.

□

Die bisher erläuterte Theorie soll nun an Simulationsdaten getestet werden. Während die Durchführung, Beschreibung und Interpretation der Simulation im nächsten Kapitel erfolgt, wird zunächst in den folgenden Abschnitten die grundlegende Theorie zur Erstellung von Simulationsdaten, Powerfunktionen und p-Wert-Kurven zusammengefasst.

3.3 Theoretische Grundlagen der Simulationsstudie

Das Ziel der Simulation ist es, Informationen darüber zu erhalten, welche Einflussgrößen und Parameter, die in das Modell eingehen, eine entscheidende Wirkung auf

den in Satz 3.2.1 beschriebenen Bereichstest haben. Dazu wird speziell die Zusammensetzung der Daten untersucht, d.h. die Grösse der Parametersätze und die Schwankungsbreite des theoretischen Erwartungswertes sollen zwischen den verschiedenen Parametersätzen variieren. Durch den Vorteil der Simulationsdaten, dass die dem Modell zugrundeliegenden Parameter bekannt sind, kann weiter geprüft werden, inwiefern die Ergebnisse, die der Hypothesentest liefert, der Wahrheit entsprechen und es kann somit die Zuverlässigkeit des Tests beurteilt werden. Zunächst müssen jedoch erst die Simulationsdaten erstellt werden, wozu häufig der sogenannte „*Signal-to-noise-Quotient*“ betrachtet wird, der nachstehend beschrieben wird. Anschliessend wird kurz auf die Powerfunktion, wie sie in Definition 3.1.2 eingeführt wurde, eingegangen und es werden die p-Wert-Kurven am Ende dieses Kapitels erklärt.

3.3.1 Konzept der Simulationsdaten

Für die Wahl von Simulationsdaten ist es zunächst einmal entscheidend zu wissen, welche Parameter gewählt und festgelegt werden müssen. In diesem Fall sollen negativ-binomial-verteilte Zufallsvariablen Y erzeugt werden, deren Verteilung nach Definition 2.3.6 durch zwei Parameter, nämlich durch den Dispersionsparameter a und μ bestimmt wird. Im folgenden wird wie im Abschnitt 2.3.3 von der log-link-Funktion ausgegangen, so dass nach (2.9) sich der Erwartungswert mit dem Regressionsparameter β und dem Kovariablenvektor \mathbf{x}_i durch die Formel $\mu = \exp(\mathbf{x}_i^t \beta)$ bestimmen lässt. Schliesslich ist noch die Länge n des Zufallsvariablenvektors festzulegen. Es soll also $Y_i \sim NB(\mu_i, a)$ gelten, wobei nach (2.14) und (2.15) der Erwartungswert und die Varianz durch

$$E(Y_i) = \mu_i \quad \text{und} \quad Var(Y_i) = \mu_i + a\mu_i^2$$

gegeben sind, mit $\mu_i = \exp(\mathbf{x}_i^t \beta)$. Es sind somit zusätzlich die Kovariablen $\mathbf{x}_i, i = 1, \dots, n$ zu wählen. Für die Simulationsstudie wird nur von einem Kovariablenvektor ausgegangen, d.h. es wird $p = 1$ gesetzt. Zusammen mit dem Intercept erhält man damit eine Design-Matrix der Gestalt

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}.$$

Der Regressionsparametervektor β besteht daher aus nur 2 Komponenten β_0 und β_1 , die es ebenfalls zu wählen gilt.

In dieser Arbeit werden zwei Gruppen von Simulationsdaten $(a, \mu, n, \mathbf{x}_i, \boldsymbol{\beta})$ gebildet. Die eine enthält Zufallsvektoren der Länge $n = 25$, die andere der Länge $n = 100$. Diese Aufspaltung soll Aufschluss darüber geben, ob, und falls um wieviel, die Schätzungen der zweiten Gruppe im Vergleich zu der Gruppe, die weniger Daten enthält, genauer ausfallen. Desweiteren werden die Parameter β_0 und β_1 und der Kovariablenvektor \mathbf{x} einmal so gewählt, dass der daraus resultierende Erwartungswertvektor $\boldsymbol{\mu}$ in einem kleinen Intervall schwankt, d.h.

$$\frac{1}{2}\bar{\mu} < \mu_i < \frac{3}{2}\bar{\mu}, \quad \forall i = 1, \dots, n, \quad (3.19)$$

im anderen Fall wird das Schwankungsintervall um

$$\frac{1}{4}\bar{\mu} < \mu_i < \frac{7}{4}\bar{\mu}, \quad \forall i = 1, \dots, n \quad (3.20)$$

vergrössert. Dabei bedeutet $\bar{\mu}$ den empirischen Mittelwert, d.h.

$$\bar{\mu} = \frac{1}{n} \sum_{i=1}^n \mu_i.$$

Die Schwierigkeit liegt nun in der Bestimmung der Parameter a und $\boldsymbol{\mu}$. Dazu wird der „*Signal-to-noise-Quotient*“ betrachtet, der durch

$$SN(a, \mu_i) = \frac{E(Y_i)}{\sqrt{Var(Y_i)}} \quad (3.21)$$

definiert ist. Für den Fall, dass $Y_i \sim NB(a, \mu_i)$ verteilt ist, ergibt sich ein Signal-to-noise-Quotient von

$$SN_{NB}(a, \mu_i) = \frac{E(Y_i)}{\sqrt{Var(Y_i)}} = \frac{\mu_i}{\sqrt{\mu_i + a\mu_i^2}} = \frac{\mu_i}{\sqrt{\mu_i}\sqrt{1 + a\mu_i}} = \sqrt{\frac{\mu_i}{1 + a\mu_i}}.$$

Lässt man sich diese Funktion in Abhängigkeit von a und μ_i darstellen, so ergibt sich die Abbildung 3.1. Hieran lässt sich erkennen, dass der Signal-to-noise-Quotient schnell gegen 0 konvergiert, sobald a Werte annimmt, die grösser als 0.6 sind. Grosse Werte kann der Quotient nur dann annehmen, wenn ein kleines a in Kombination mit einem grossen μ vorliegt.

Für die Simulationsdaten werden nun die Werte 1.25 und 2 des Signal-to-noise-Quotienten für die weitere Betrachtung untersucht. Um daraus die jeweiligen verschiedenen Kombinationen für a und μ abzuleiten, wird der sogenannte „*Konturenplot*“ verwendet (Abbildung 3.2). In der linken Graphik sind die Werte von μ auf einen Bereich bis 10 eingeschränkt, während rechts der Wertebereich von μ bis 100 läuft. Wie auch

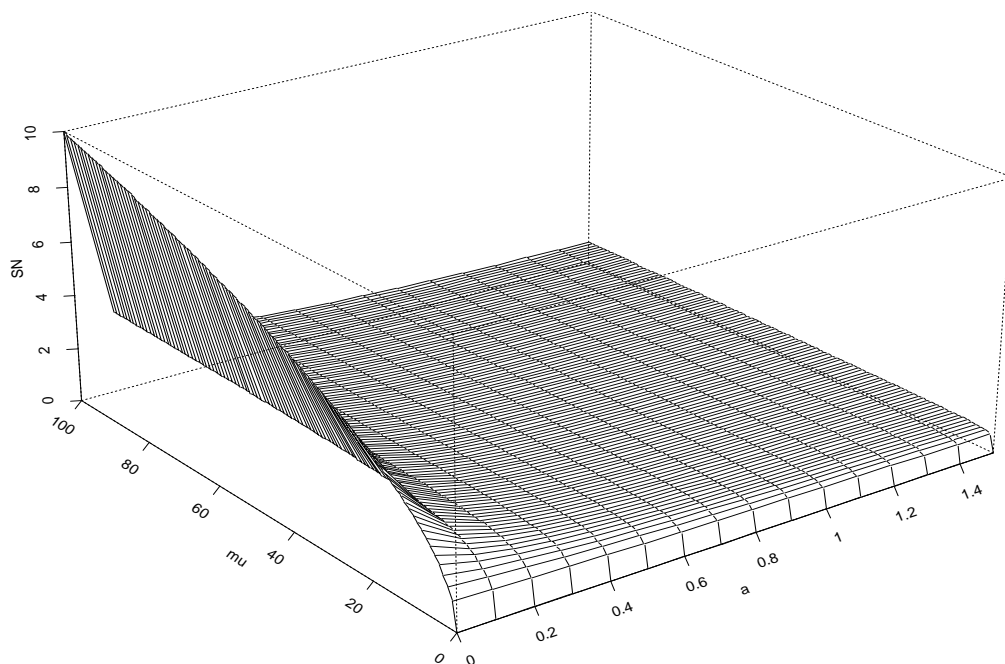


Abbildung 3.1: Signal-to-noise-Quotient für NB-Verteilung

rechnerisch nachgeprüft werden kann, ergeben sich schliesslich z.B. die Paare $a = 0.3$ und $\bar{\mu} = 3$ bzw. $a = 0.63$ und $\bar{\mu} = 100$ für einen Signal-to-noise-Quotienten von 1.25, während sich für einen Wert von 2 die Paare $a = 0.15$ und $\bar{\mu} = 10$ bzw. $a = 0.24$ und $\bar{\mu} = 100$ ergeben. Ausgehend von diesen Ergebnissen werden nun Kombinationen von β und \mathbf{x} gesucht, so dass $\bar{\mu} = \frac{\sum_{i=1}^n \mu_i}{n}$ mit $\mu_i = \exp(\mathbf{x}_i^t \beta)$ einmal den Wert 3 bzw. 100 für einen Signal-to-noise-Quotienten von 1.25 und einmal den Wert 10 bzw. 100 für einen Signal-to-noise-Quotienten von 2 annimmt. Dabei müssen jeweils die zwei Fälle unterschieden werden, für die μ_i im kleinen Intervall $\frac{1}{2}\bar{\mu} < \mu_i < \frac{3}{2}\bar{\mu}$ und einmal im grossen Intervall $\frac{1}{4}\bar{\mu} < \mu_i < \frac{7}{4}\bar{\mu}$ schwankt. Die ermittelten Parametersätze $(a, \bar{\mu}, n, \mathbf{x}, \beta)$, die die geforderten Bedingungen weitgehendst erfüllen, sind im nächsten Kapitel in Tabelle 4.1 systematisch zusammengefasst.

3.3.2 Powerfunktion

Als erster Schritt der Simulationsstudie soll die Powerfunktion zum Hypothesentest für Überdispersion betrachtet werden. Um diese anhand der ermittelten Daten zu überprüfen, muss ein Programm erzeugt werden, das nun schrittweise erläutert wird.

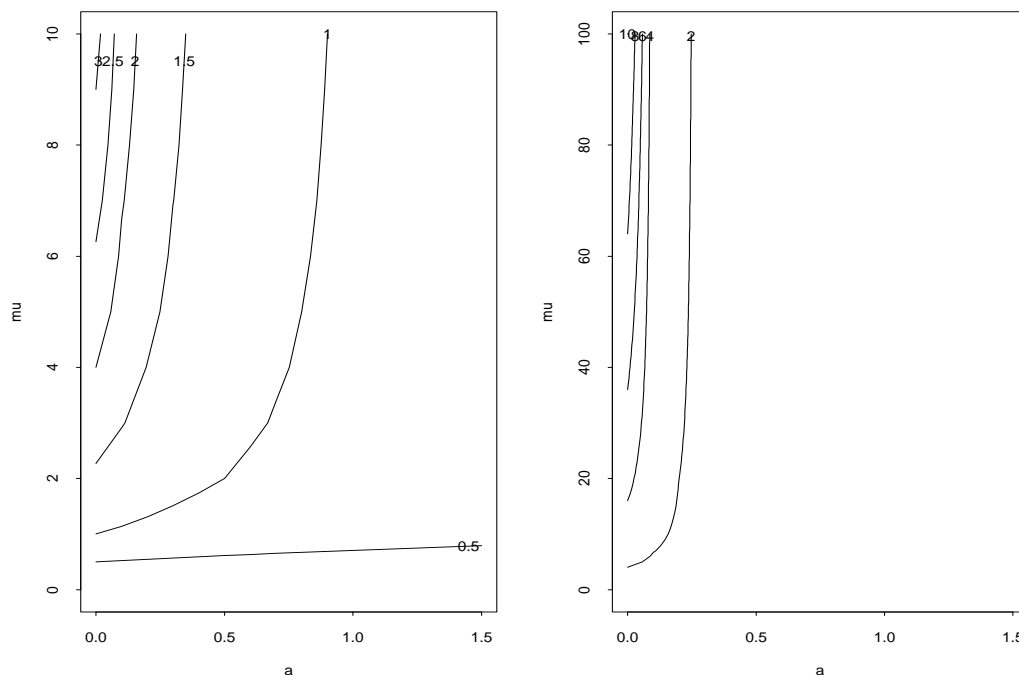


Abbildung 3.2: Konturenplot des Signal-to-noise-Quotienten bei der NB-Verteilung

Zunächst wird noch einmal der zu untersuchende Test betrachtet. Nach Satz 3.2.1 in Kapitel 3.2 ist der Bereichstest für Überdispersion gegeben durch die Nullhypothese $H : a > a_0$ gegen die Alternative $K : a \leq a_0$, wobei a der Dispersionsparameter der NB-Verteilung ist und a_0 eine beliebig wählbare Schranke darstellt. Der Ablehnungsbereich oder die kritische Region ist dann gegeben durch

$$\mathcal{C} = \left\{ \hat{a} : \Phi \left(\frac{\hat{a} - a_0}{\hat{\sigma}(\hat{\beta}, \hat{a})} \right) \leq \alpha \right\}.$$

Die Powerfunktion wurde in Definition 3.1.2 angegeben und ergibt übertragen auf dieses spezielle Problem die Powerfunktion $\beta(a)$ mit

$$\beta(a) = P(H_0 \text{ ablehnen} \mid a \text{ wahrer Parameter}). \quad (3.22)$$

Im folgenden wird stets ein Signifikanzniveau von $\alpha = 0.05$ angenommen.

Für die Simulation der Powerfunktion ist es nun zunächst notwendig, einen festen Wert a_0 zu wählen und den Parameter a festzusetzen. Anschliessend simuliert man z.B. 300 mal negativ-binomial-verteilte Zufallsvariablen mit eben diesem festen Parameter a . Schliesslich führt man den Test für $H : a > a_0$ gegen $K : a \leq a_0$ aus Satz 3.2.1 für jede

der 300 erzeugten Zufallsvariablen durch, wozu noch das $\hat{\sigma}$ berechnet werden muss. Dafür werden die Parameter a und β geschätzt, deren Werte nachfolgend mit \hat{a} und $\hat{\beta}$ abgekürzt werden. Zuletzt muss gezählt werden, wie oft bei den 300 Durchführungen der Test verworfen wird, d.h. wie oft

$$\Phi\left(\frac{\hat{a} - a_0}{\hat{\sigma}(\hat{\beta}, \hat{a})}\right) \leq \alpha$$

auftritt. Dividiert man diese Anzahl durch die Gesamtanzahl der Durchführungen, so erhält man eine Schätzung für die Power (3.22) zu dem festen Parameterwert a . Dieser Vorgang muss schliesslich bei gleichem a_0 für mehrere Werte von a wiederholt werden, bevor man eine Powerfunktion graphisch darstellen kann.

Das Programm „*powerfunktion*“, das die einzelnen Schritte durchführt und schliesslich die Powerfunktion berechnet, ist im Anhang zu finden. Es ist, wie alle weiteren Programme auch, in der Statistik-Programmiersprache Splus geschrieben. Es ist zu bemerken, dass Splus im Standard-Paket zwar ein Programm zur Berechnung von NB-Modellen enthält, doch lässt dieses nur diskrete Werte des Dispersionsparameters zu (siehe Definition 2.3.1). Um nun aber auch Modelle mit stetigem Parameter a berechnen zu können (siehe Definition 2.3.6), wurde von der Internetseite

[http : //lib.stat.cmu.edu/S/](http://lib.stat.cmu.edu/S/)

das Programm „*negbin*“ von Bill Venables installiert. Dabei ist zu beachten, dass die NB-Verteilung in Splus eine andere Parametrisierung verwendet als in dieser Arbeit in Definition 2.3.6 angegeben ist. Splus benutzt die Bezeichnungen

$$\begin{aligned} P(Y = y) &= \frac{\Gamma(\theta + y)}{\Gamma(\theta)y!} \frac{\theta^\theta \mu^y}{(\theta + \mu)^{\theta+y}} \\ &= \frac{\Gamma(\theta + y)}{\Gamma(\theta)y!} \left(\frac{\theta}{\theta + \mu}\right)^\theta \left(\frac{\mu}{\theta + \mu}\right)^y, \end{aligned} \quad (3.23)$$

was im Vergleich zur Definition 2.3.6 einer Parametrisierung von $\theta = \frac{1}{a}$ entspricht. Die grösste Schwierigkeit in der Programmierung der Powerfunktion ist hier die Berechnung von $\hat{\sigma}(\hat{\beta}, \hat{a})$ in der asymptotisch normalverteilten Teststatistik. So wie in Satz 3.2.1 beschrieben, ist $\hat{\sigma}^2(\hat{\beta}, \hat{a})$ der Schätzer von $\sigma^2(\beta, a)$, dem letzten Eintrag in der Inversen der Fisher-Informationsmatrix. Damit erhält man

$$\hat{\sigma}(\hat{\beta}, \hat{a}) = \frac{1}{\sqrt{FI_{p+1,p+1}(\hat{\beta}, \hat{a})}},$$

wobei $FI_{p+1,p+1}(\hat{\beta}, \hat{a})$ nach (2.31) durch

$$FI_{p+1,p+1}(\hat{\beta}, \hat{a}) = \hat{a}^{-4} \left[\sum_{i=1}^n \left[E \left(\sum_{j=0}^{y_i^*} \frac{1}{(j + \hat{a}^{-1})^2} \right) - \frac{\hat{\mu}_i}{(\hat{a}^{-1} + \hat{\mu}_i)\hat{a}^{-1}} \right] \right] \quad (3.24)$$

gegeben ist mit $\hat{\mu}_i = \exp(\mathbf{x}_i^t \boldsymbol{\beta})$, ausgewertet an der Stelle $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$.

Nach (2.32) lässt sich schliesslich der Erwartungswert in obiger Formel wie folgt für die i -te Beobachtung berechnen:

$$\begin{aligned} FI_{p+1,p+1}(\hat{\beta}, \hat{a})_i &= \hat{a}^{-4} \left(\sum_{j=0}^{\infty} \frac{1}{(j + \hat{a}^{-1})^2} P_{NB}(Y_i \geq j) - \frac{\hat{a}\hat{\mu}_i}{\hat{\mu}_i + \hat{a}^{-1}} \right) \\ &= \hat{a}^{-4} \left(\sum_{j=0}^{\infty} (\hat{a}^{-1} + j)^{-2} (1 - P(Y_i \leq j - 1)) - \frac{\hat{a}\hat{\mu}_i}{\hat{\mu}_i + \hat{a}^{-1}} \right). \end{aligned} \quad (3.25)$$

Im Programm „*powerfunktion*“ wird also zuerst der wahre Erwartungswertvektor $\boldsymbol{\mu}$ aus den Daten \mathbf{x} und $\boldsymbol{\beta}$ der Simulationsdaten durch $\mu_i = \exp(\mathbf{x}_i^t \boldsymbol{\beta})$ gebildet. Anschliessend werden in die Matrix Z der Länge n , was der Länge des Vektors \mathbf{x} entspricht, und der Breite 300 die durch `rnegbin(mu, theta=1/a)` erzeugten Zufallsvariablen geschrieben. In dem Splus-Befehl `out_glm.nb(Z[,k] ~ x, link=log)` werden die Regressionsparameter β_0 , β_1 und a geschätzt, die für die Berechnung von $\hat{\sigma}(\hat{\boldsymbol{\beta}}, \hat{a})$ notwendig sind. Wie schon angesprochen, lässt sich $\hat{\sigma}(\hat{\boldsymbol{\beta}}, \hat{a})$ mit Hilfe der Gleichung (3.25) berechnen, wobei die unendliche Summe über j in der Simulationsstudie mit einer Summe bis 150 approximiert wird, da für $j > 150$ die Wahrscheinlichkeit $P(Y_i \geq 150) = 1 - P(Y_i < 150)$ gegen Null konvergiert, was gleichbedeutend damit ist, dass $P(Y_i < 150)$ gegen 1 strebt. Dieser Sachverhalt ist in den Graphiken in Abbildung 3.3 dargestellt, in denen für bestimmte feste Werte $\boldsymbol{\mu}$ die Verteilungsfunktion der NB-Verteilung $P_{NB}(Y < j)$ in Abhängigkeit von j und a abgebildet ist. Man erkennt daran, dass bei den Parametersätzen mit $\mu = 3$ und $\mu = 10$ eine Approximation der Summe bis $j = 50$ schon ausreichen würde, da unabhängig von a in diesen Fällen die Verteilungsfunktion bei $j = 50$ bereits fast den Wert 1 annimmt. Im Fall $\mu = 100$ wird allerdings deutlich, dass eine Approximation der unendlichen Summe bis $j = 50$ nicht ausreicht; stattdessen muss j mindestens 150 gewählt werden. Wird als obere Grenze für j genau 150 angenommen, so wird zwar nur etwa ein Wert von 0.85 der Verteilungsfunktion $P_{NB}(Y < j)$ erreicht, doch ist dies für die Gesamtberechnung von $\hat{\sigma}(\hat{\boldsymbol{\beta}}, \hat{a})$ ausreichend, da die Summanden $(\hat{a}^{-1} + j)^{-2} (1 - P(Y_i \leq j - 1))$ für $j > 150$ so kleine Beträge ergeben, dass sie für das Gesamtergebnis $\hat{\sigma}$ vernachlässigbar sind. Insgesamt reicht es also für alle

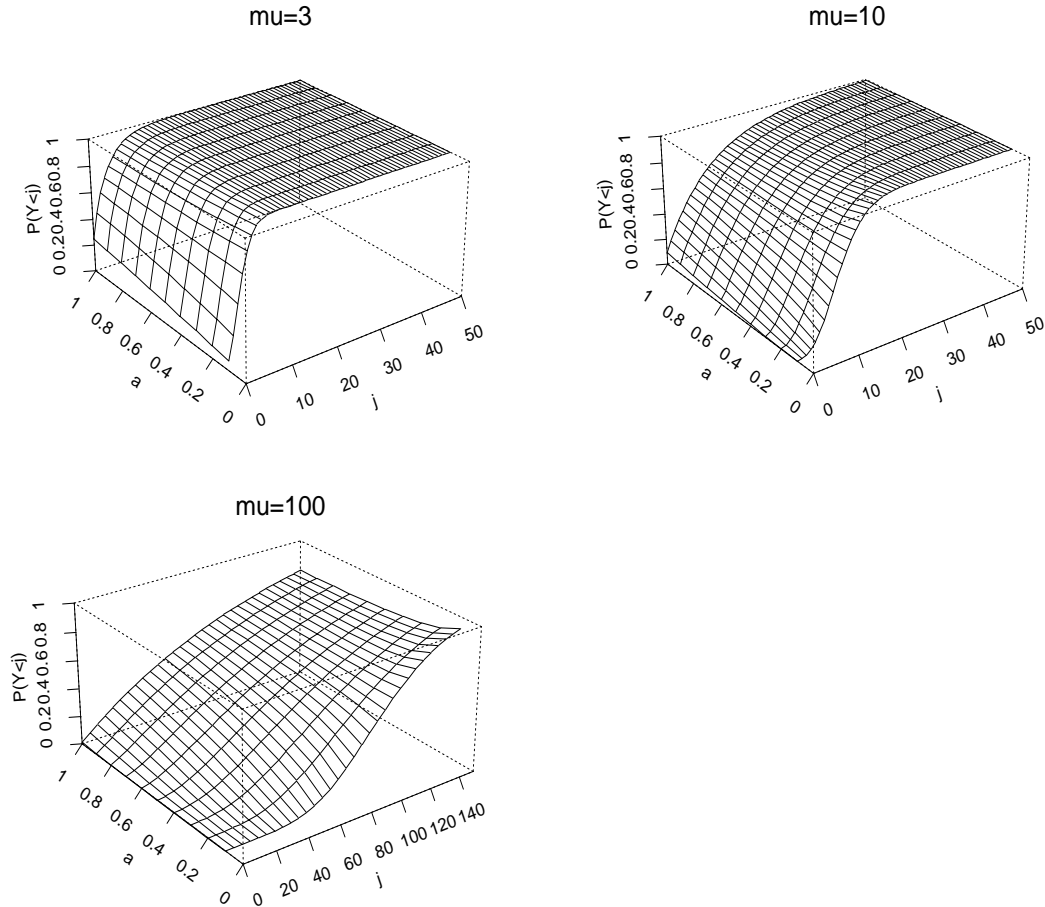


Abbildung 3.3: Grenzverhalten der NB-Verteilungsfunktion $P(Y \leq j)$ für $Y \sim NB(\mu, a)$

betrachteten Fälle aus, die unendliche Summe mit einer Summe bis 150 zu approximieren. Schliesslich wird eine Matrix aus n Zeilen für die Beobachtungen 1 bis n und 151 Spalten für die Summeneinträge $j = 0, \dots, 150$ gebildet. Summiert man nun über die Zeilen, entspricht dies der Summe über j in (3.25). Wird der resultierende Vektor „innensum“ der Länge n aufsummiert, so erhält man das Ergebnis der Summe über i in (3.24). Anschliessend wird noch \hat{a}^{-4} an das Zwischenergebnis heranmultipliziert, der Kehrwert gebildet und die Wurzel gezogen, womit das endgültige Ergebnis $\hat{\sigma}(\hat{\beta}, \hat{a})$ berechnet ist. Am Ende wird die Testgrösse $\Phi\left(\frac{\hat{a}-a_0}{\hat{\sigma}(\hat{\beta}, \hat{a})}\right)$ berechnet und getestet, ob diese kleiner oder grösser $\alpha = 0.05$ ist. Zuletzt wird die Anzahl der Verwerfungen durch 300 dividiert, womit man eine Schätzung für die Wahrscheinlichkeit für eine Ablehnung von H erhält, unter der Bedingung, dass der Parameter a vorliegt.

3.3.3 P-Wert-Kurven bezüglich Dispersionsindex a_0

In diesem Abschnitt wird nun beschrieben, wie eine p-Wert-Kurve erstellt werden kann, die Aufschluss darüber geben soll, ob der Aufwand, von einem Poissonmodell zu einem NB-Modell zu wechseln, um Überdispersion zu berücksichtigen, gerechtfertigt ist. Dazu wird zunächst die p-Wert-Kurve bezüglich des Dispersionsindizes a_0 betrachtet, während später noch auf die Betrachtung eines besser interpretierbaren Abstandsmasses zwischen beiden Verteilungen eingegangen wird.

Wie schon in den vorangegangenen Kapiteln erwähnt, ist das allgemeine Testproblem durch die Hypothese $H : a > a_0$ gegen die Alternative $K : a \leq a_0$ gegeben. Dazu gibt es, wie in Kapitel 3.1 eingeführt, zwei verschiedene Arten, einen Fehler zu begehen. Der erste Fehler liegt vor, wenn die Nullhypothese abgelehnt wird, obwohl sie wahr ist. Der zweite Fehler dagegen ist die Annahme der Nullhypothese, obwohl sie falsch ist. Das Problem bei derartigen Tests ist, dass nicht beide Fehler gleichzeitig kontrolliert werden können. Allgemein sind nun die Tests so aufgebaut, dass nur der Fehler 1. Art betrachtet wird, da dieser in der Regel als der schwerwiegendere gilt. Es gibt aber Fälle, in denen nicht klar ist, welches der Fehler 1. Art bzw. 2. Art sein soll. In solchen Situationen ist es häufig hilfreich, die p-Wert-Kurve zu betrachten, die nicht nur angibt, ob eine Hypothese zu verwerfen oder anzunehmen ist, sondern auch Informationen darüber liefert, wie weit man im Falle einer Annahme von der Ablehnung der Nullhypothese entfernt liegt. Dazu sei

$$\hat{P}(\hat{a}, \hat{\sigma}, a_0) = \Phi \left(\frac{\hat{a} - a_0}{\hat{\sigma}(\hat{\beta}, \hat{a})} \right) \quad (3.26)$$

mit „*asymptotischer p-Wert*“ bezeichnet. In der Abbildung 3.4 sind die p-Wert-Kurven für die zwei Testprobleme $H : a > a_0$ gegen $K : a \leq a_0$ und $K : a \leq a_0$ gegen $H : a > a_0$ als Funktionen von a_0 abgebildet.

Ist ein fester Datensatz an Beobachtungen gegeben, so gibt $\hat{P} = \alpha$ in der linken Graphik für den Test $H : a > a_0$ gegen $K : a \leq a_0$ den minimalen Wert a_0 an, zu dem die Nullhypothese H bei einem Signifikanzniveau α verworfen werden kann. In der rechten Graphik ergibt sich bei $\hat{P} = \alpha$ der maximale Wert a_0 des Testproblems $K : a \leq a_0$ gegen $H : a > a_0$, zu dem die Nullhypothese K bei einem Signifikanzniveau α verworfen werden kann. Zu bemerken ist, dass sich die rechte Kurve durch eine Achsenspiegelung an $\hat{P} = 0.5$ aus der linken Kurve erzeugen lässt, womit die beiden getrennten Testprobleme auch vereinfacht in einer Graphik betrachtet werden können. In der gemeinsamen p-Wert-Kurve für beide Tests kann wie gewohnt der minimale Wert a_{0u} (u steht für

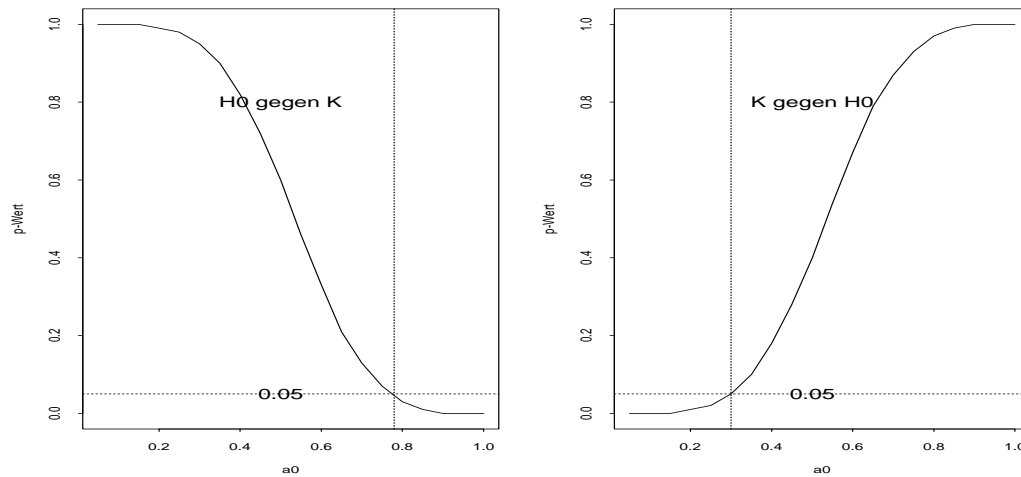


Abbildung 3.4: p-Wert-Kurven für die Testprobleme H gegen K und K gegen H

„upper“), für den die Nullhypothese H zu einem Signifikanzniveau α verworfen werden kann bei einem Level $\hat{P} = \alpha$ abgelesen werden. Andererseits ergibt sich nun auch bei einem Level $\hat{P} = 1 - \alpha$ zusätzlich der maximale Wert a_{0l} (l steht für „lower“), zu dem die Hypothese K des Tests $K : a \leq a_0$ gegen $H : a > a_0$ zu einem Signifikanzniveau α verworfen werden kann. Dieser Zusammenhang soll in Abbildung 3.5 verdeutlicht werden.

Zur Interpretation dieser Abbildung betrachte man nur den Bereich ganz links der Diskriminierung und den Bereich rechts der Validierung des Modells. Ist $\hat{P} = \alpha = 0.05$, so ergibt sich ein minimaler Wert a_{0u} von 0.77, zu dem H verworfen werden kann und analog bei einem Level von $\hat{P} = 1 - \alpha = 0.95$ ergibt sich ein maximaler Wert a_{0l} von 0.3, zu dem die Hypothese K des zweiten Tests verworfen werden kann. Kurz zusammengefasst lässt sich die Kurve folgendermassen beschreiben:

- (i) **Modell-Validierung** ($H : a > a_0$ gegen $K : a \leq a_0$):

Ab einem Wert von $a_{0u} = 0.77$ kann die Nullhypothese zu einem Signifikanzniveau α verworfen werden, d.h. man kann $K : a \leq a_0$ annehmen und somit das Poissonmodell als Regressionsmodell wählen.

- (ii) **Modell-Diskriminierung** ($K : a \leq a_0$ gegen $H : a > a_0$):

Bis höchstens $a_{0l} = 0.3$ kann $K : a \leq a_0$ verworfen werden zu einem Signifikanzniveau α , d.h. man nimmt $H : a > a_0$ an und wechselt vom Poissonmodell zum NB-Modell.

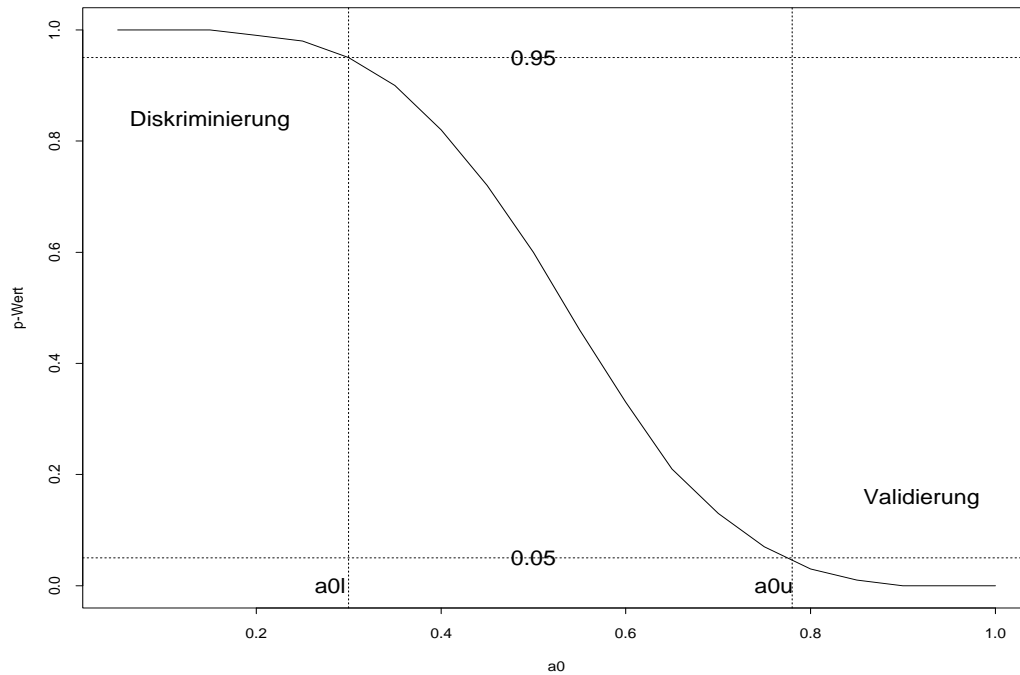


Abbildung 3.5: Allgemeine p-Wert-Kurve

Abschliessend lässt sich also sagen, dass ein Poissonmodell gerechtfertigt ist, wenn man bereit ist, einen Dispersionsindex a_{0u} von 0.77 als Schranke zu akzeptieren. Andererseits kann das Poissonmodell mit einem signifikant feststellbaren Abstand von $a_{0l} = 0.3$ abgelehnt werden. Akzeptiert man also höchstens einen Dispersionsindex a_{0l} von 0.3, so wäre in diesem Fall der Aufwand gerechtfertigt, das Poissonmodell zu verwerfen und zum NB-Modell überzugehen.

3.3.4 p-Wert-Kurven bezüglich Abstandsmass d_0

Der Unterschied zu einer p-Wert-Kurve bezüglich des Dispersionsindizes a_0 ist hier, dass eine Abstandsmassfunktion d definiert wird, die leichter zu interpretieren ist als der Parameter a_0 und die somit den Abstand zwischen dem Poisson- und dem NB-Modell anschaulich wiedergibt. Es ist dann zu beachten, dass sich der Test damit ändert und nun die Form

$$H : d(a) > d_0 \quad \text{gegen} \quad K : d(a) \leq d_0$$

hat, wobei die Nullhypothese genau dann gilt, wenn $a > d^{-1}(d_0) = a_0$ ist, woraus sich wieder der Zusammenhang mit dem Dispersionsindex a_0 ergibt.

Bevor nun eine p-Wert-Kurve bezüglich eines Abstandsmasses gebildet werden kann, muss dieses zunächst einmal festgelegt werden. Es gibt viele Möglichkeiten, ein Mass zu definieren, das den gewünschten Forderungen entspricht. Eine der wichtigsten Bedingungen, die das Mass erfüllen sollte, ist die Monotoniebedingung, d.h. das Mass sollte entweder streng monoton fallend oder steigend sein, da sonst Interpretationsschwierigkeiten auftreten könnten. Ausserdem sollte die zu wählende Funktion eine Beziehung zwischen der Poisson- und der NB-Verteilung herstellen. Die einfachste Funktion, die diesen Bedingungen genügt, ist das Verhältnis der Varianzen beider Verteilungen:

$$d(a) = \frac{Var(\text{NB-Verteilung})}{Var(\text{Poissonverteilung})} = \frac{\mu + a\mu^2}{\mu} = 1 + a\mu. \quad (3.27)$$

Da in der Regel mehrere Beobachtungen $Y_i, i = 1, \dots, n$ mit Erwartungswerten $\mu_i, i = 1, \dots, n$ vorliegen, wird der Eindeutigkeit halber die Abstandsmassfunktion auf

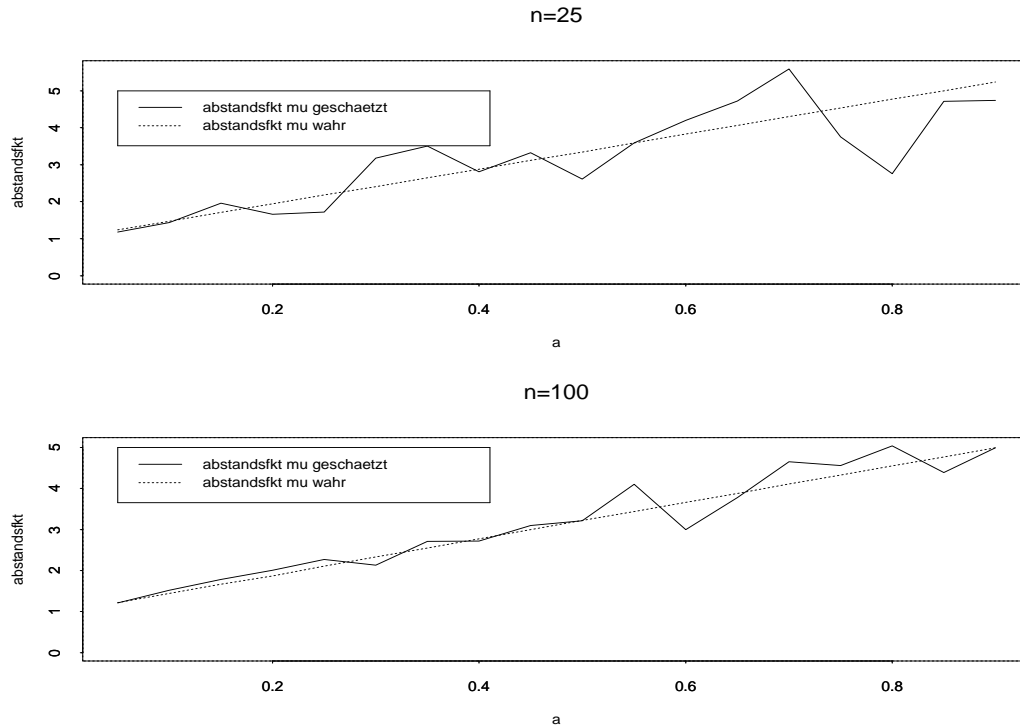
$$d(a) = 1 + a \exp \left\{ \max_i \mathbf{x}_i^t \boldsymbol{\beta} \right\} = 1 + a \max_i \{\mu_i\} \quad (3.28)$$

festgelegt. Dieses Mass gibt dann das maximale Verhältnis der Varianzen aus dem Poisson- und dem NB-Modell wieder. Je näher der Wert des Abstandsmasses bei 1 liegt, desto eher gilt für die maximalen Varianzen der Zusammenhang

$$Var(\text{Poissonverteilung}) = Var(\text{NB-Verteilung}),$$

was ein Hinweis darauf ist, dass die Daten keine Überdispersion aufweisen und dass das Poissonmodell damit gerechtfertigt ist. Ergibt sich dagegen für das Abstandsmass ein sehr grosser Wert, so bedeutet dies, dass die maximale Varianz der NB-Verteilung sehr viel grösser als die maximale Varianz der Poissonverteilung ist. Daraus kann dann gefolgert werden, dass die Daten einen hohen Grad an Überdispersion aufweisen, weshalb eine Modellierung mittels der Poissonverteilung nicht gerechtfertigt erscheint, sondern stattdessen die NB-Verteilung als Grundlage des Modells gewählt werden sollte. In der Abbildung 3.6 ist diese Funktion für den ersten Parametersatz mit $n = 25$ und für den zweiten Parametersatz mit $n = 100$ abgebildet.

Die gestrichelte Linie ist dabei das wahre $d(a)$, d.h. es wird aus dem Regressionsparametervektor $\boldsymbol{\beta}$ und dem Kovariablenvektor \mathbf{x} das wahre feste $\max_i \{\mu_i\}$ berechnet, während für die geschätzte Abstandsmassfunktion zunächst Zufallsvariablen erzeugt und anschliessend die Parameter β_0 und β_1 geschätzt werden. Das Problem dabei ist, dass sich mit variierendem a auch die Schätzungen ändern, so dass für jeden Wert von a eine neue Erzeugung von Zufallsvariablen durchgeführt werden muss. Aus diesem

Abbildung 3.6: Abstandsmassfunktion $d(a)$

Grund ist die geschätzte Abstandsmassfunktion nicht ganz monoton, jedoch nähert sie sich für grosse n an die Gerade der wahren Abstandsmassfunktion an. Die p-Wert-Kurven bezüglich des Abstandsmasses $d(a)$ sind dann analog zu (3.26) durch

$$\hat{P}(\hat{a}, \hat{\sigma}, d_0) = \Phi \left(\frac{\hat{a} - d^{-1}(d_0)}{\hat{\sigma}(\hat{\beta}, \hat{a})} \right) \quad (3.29)$$

definiert und lassen sich wie die Kurven bezüglich a_0 interpretieren, wie es im vorangegangenen Abschnitt beschrieben wurde.

3.3.5 Interpretation und Bestimmung von p-Wert-Kurven

Im folgenden sollen nun die einzelnen Schritte zur Erstellung der p-Wert-Kurven erläutert werden. Die entscheidenden interessierenden Werte einer Kurve sind die Schnittpunkte a_{0u} und a_{0l} der Kurve mit den Levels $\hat{P} = \alpha = 0.05$ und $\hat{P} = 1 - \alpha = 0.95$. Da die Kurven aus erzeugten Zufallsvariablen entstehen, können, obwohl dieselben Parameter zugrundeliegen, bei mehreren Durchführungen sehr unterschiedliche Kurven, und damit auch unterschiedliche Ergebnisse von a_{0u} und a_{0l} auftreten. Deshalb werden für jeden Parametersatz 20 Kurven erzeugt, die Schnittpunkte a_{0u} und a_{0l} mit den Le-

vels $\hat{P} = 0.05$ und $\hat{P} = 0.95$ berechnet und zuletzt ein Mittelwert der Schnittpunkte ausgegeben. Zur Berechnung der Schnittpunkte a_{0u} und a_{0l} aus

$$\Phi\left(\frac{\hat{a} - a_{0u}}{\hat{\sigma}(\hat{\beta}, \hat{a})}\right) = 0.05 \quad \text{und} \quad \Phi\left(\frac{\hat{a} - a_{0l}}{\hat{\sigma}(\hat{\beta}, \hat{a})}\right) = 0.95$$

muss die Umkehrfunktion der Normalverteilung Φ berechnet werden. Es gelten nach Springers Formelsammlung (1997, S. 451) [17] die Werte $\Phi^{-1}(0.95) = 1.645$ und $\Phi^{-1}(0.05) = -1.645$. Damit ergibt sich

$$\frac{\hat{a} - a_{0u}}{\hat{\sigma}(\hat{\beta}, \hat{a})} = -1.645, \quad \text{woraus} \quad a_{0u} = \hat{a} + 1.645 \hat{\sigma}(\hat{\beta}, \hat{a}) \quad (3.30)$$

folgt. Analog berechnet man a_{0l} durch

$$a_{0l} = \hat{a} - 1.645 \hat{\sigma}(\hat{\beta}, \hat{a}). \quad (3.31)$$

Sowohl für die Kurve bezüglich a_0 , als auch für die Kurve bezüglich des Abstandsmasses $d_0 = d(a_0)$ werden zunächst, ausgehend von den Simulationsdaten, Zufallsvariablen aus a und $\mu_i = \exp(\mathbf{x}_i^t \boldsymbol{\beta})$ erzeugt. Anschliessend werden wie bei der Powerfunktion die Parameter a und $\boldsymbol{\beta}$ geschätzt. Es folgt dann die Berechnung von $\hat{\sigma}(\hat{\beta}, \hat{a})$ in Abhängigkeit des geschätzten Parametervektors $\hat{\beta}$ und des geschätzten Dispersionsparameters \hat{a} , woran sich die Berechnung der Schnittpunkte a_{0u} und a_{0l} nach den Formeln (3.30) und (3.31) anschliesst. Zuletzt werden die 20 verschiedenen Werte a_{0u} und a_{0l} gemittelt, womit die entscheidenden Daten zur Erstellung einer p-Wert-Kurve berechnet sind. Diese Prozedur wird im Programm „*a0quer*“ durchgeführt, welches im Anhang nachgelesen werden kann. Die p-Wert-Kurven selber werden mittels des Programms „*pwertaneu*“ erzeugt, welches ebenfalls im Anhang zu finden ist. Dabei wird zuerst der Vektor a_0 , der die x-Achse der Graphik darstellt, angegeben. Zu jedem dieser Werte a_0 wird dann $\hat{P} = \Phi\left(\frac{\hat{a} - a_0}{\hat{\sigma}(\hat{\beta}, \hat{a})}\right)$ berechnet, wobei die geschätzten Werte $\hat{\beta}$, \hat{a} und $\hat{\sigma}(\hat{\beta}, \hat{a})$ aus den Ergebnissen des Programms „*a0quer*“ entnommen werden. Für die p-Wert-Kurven bezüglich des Abstandsmasses d_0 erfolgt zuvor noch die Umrechnung

$$d_0 = 1 + a_0 \max_i \{\hat{\mu}_i\}.$$

Auch dieses Programm „*pwertdneu*“ zur Erzeugung der p-Wert-Kurven bezüglich d_0 ist im Anhang beigefügt.

Kapitel 4

Simulationsstudie

In diesem Kapitel werden nun die Ergebnisse der Simulation, wie sie in Abschnitt 3.3 beschrieben wurde, dargestellt. Das Ziel der Simulation ist es, im wesentlichen die folgenden Fragen zu beantworten:

- Wann ist der Aufwand gerechtfertigt, vom Poissonmodell zum NB-Modell aufgrund zu hoher Überdispersion zu wechseln, bzw.
- Wie gross ist der statistische Nachweis für das Poissonmodell?

Dabei soll ausserdem untersucht werden, welchen Einfluss die Datenstruktur auf die Ergebnisse des Bereichstests hat, d.h. es bestehen zusätzlich die Fragen:

- Hat der Datenumfang n einen Einfluss auf die Testergebnisse?
- Hat die Schwankungsbreite (range) der Erwartungswerte μ_i einen Einfluss auf die Testergebnisse?
- Inwiefern beeinflusst ein unterschiedlicher Wert des Signal-to-noise-Quotienten (3.21) die Testergebnisse?

Um diese Fragen beantworten zu können, sind Datensätze der Länge $n = 25$ und $n = 100$ erzeugt worden und es wurde das Schwankungsintervall (range) der μ_i 's einmal klein (siehe (3.19)) und einmal gross (siehe (3.20)) gewählt. In Tabelle 4.1 sind die Parameter, die zur Erzeugung der Daten in der Simulation verwendet werden, systematisch zusammengefasst.

Die erste Zeile eines jeden Parametersatzes enthält die theoretischen Werte, während in der zweiten Zeile die Ergebnisse für den Signal-to-noise-Quotienten ($SN(a, \bar{\mu})$), für $\bar{\mu}$

und für das Schwankungsintervall (range) aufgelistet sind, die sich mit den gewählten Werten für \mathbf{x} und β ergeben. Für die gesamte Simulation der Powerfunktionen und der p-Wert-Kurven werden ausschliesslich diese Parametersätze verwendet. Die 16 verschiedenen Parametersätze sind in 4 Gruppen eingeteilt. Innerhalb einer Gruppe liegt den Daten stets derselbe Parameter a und der gemittelte Erwartungswert $\bar{\mu}$ zugrunde, woraus sich ein einheitlicher Wert des Signal-to-noise-Quotienten ergibt. Die 4 Parametersätze innerhalb einer Gruppe ergeben sich durch die Aufspaltung in Daten mit $n = 25$ und mit $n = 100$ und in die Aufspaltung nach unterschiedlichen Schwankungsintervallen für μ_i . An dieser Stelle wird angemerkt, dass im folgenden die Abkürzung PS für Parametersatz verwendet wird.

4.1 Simulation der Powerfunktion

Die Powerfunktion ist nach (3.22) eine Funktion des Parameters a , d.h. a kann variieren, während der Index a_0 als fest vorausgesetzt wird. In dieser Arbeit werden vier verschiedene Werte von a_0 gewählt, die den Werten von a in den Simulationsdaten entsprechen, also $a_0 = 0.3$, $a_0 = 0.63$, $a_0 = 0.15$ und $a_0 = 0.24$. Schliesslich wird zu jedem PS das Programm „*powerfunktion*“ durchgeführt, d.h. zur Erzeugung der negativ-binomial-verteilten Zufallsvariablen mit festem Parameter a werden je die verschiedenen Regressionsparameter β mit dem zugehörigen Kovariablenvektor \mathbf{x} verwendet. Damit erhält man 16 verschiedene Powerfunktionen, die in 4 Gruppen mit gleichem a_0 zusammengefasst sind. Zunächst werden die Powerfunktionen für ein Signifikanzniveau von $\alpha = 0.05$ berechnet und in Abbildung 4.1 dargestellt. In Abbildung 4.2 dagegen sind die Powerfunktionen bezüglich eines Signifikanzniveaus von $\alpha = 0.1$ und in Abbildung 4.3 bezüglich eines Signifikanzniveaus von $\alpha = 0.15$ abgebildet. Das Signifikanzniveau ist in den einzelnen Graphiken jeweils mit einer gestrichelten horizontalen Linie gekennzeichnet.

Interpretation der Powerfunktionen

(i) Powerfunktionen mit Signifikanzniveau $\alpha = 0.05$:

In allen 4 Gruppen lässt sich feststellen, dass der Bereichstest eher ein liberaler Test ist, d.h. man erlaubt an der Stelle $a = a_0$, die in den Graphiken mit einer gestrichelten vertikalen Linie gekennzeichnet ist, einen Fehler von $\alpha = 0.05$, was bedeutet, dass an dieser Stelle der Wert der Powerfunktion höchstens 0.05 be-

Parametersatz	$SN(a, \bar{\mu})$	a	$\bar{\mu}$	Schwankungsintervall μ_i	$d = 1 + a \max_i(\mu_i)$	n	\mathbf{x}	β
1 Ziel erreicht	1.25 1.20	0.3	3 2.76	[1.5 ; 4.5] [1.42; 4.71]	2.41	25	seq(0.05, 1.25, 0.05)	c(0.3, 1)
2 Ziel erreicht	1.25 1.22	0.3	3 2.82	[1.5 ; 4.5] [1.65 ; 4.44]	2.33	100	seq(-5, 4.9, 0.1)	c(1, 0.1)
3 Ziel erreicht	1.25 1.17	0.3	3 2.77	[0.75 ; 5.25] [0.98 ; 5.90]	2.77	25	seq(-1.05, 2.6, 0.15)	c(0.5, 0.5)
4 Ziel erreicht	1.25 1.18	0.3	3 2.89	[0.75 ; 5.25] [0.91 ; 6.62]	2.99	100	seq(-10.9, 9, 0.2)	c(1, 0.1)
5 Ziel erreicht	1.25 1.25	0.63	100 95.70	[50 ; 150] [56.83 ; 148.41]	94.50	25	seq(15.2, 20, 0.2)	c(1, 0.2)
6 Ziel erreicht	1.25 1.25	0.63	100 94.28	[50 ; 150] [55.15 ; 148.41]	94.50	100	seq(30.1, 40, 0.1)	c(1, 0.1)
7 Ziel erreicht	1.25 1.25	0.63	100 91.09	[25 ; 175] [40.45 ; 170.72]	108.55	25	seq(11, 18.2, 0.3)	c(1.5, 0.2)
8 Ziel erreicht	1.25 1.25	0.63	100 84.31	[25 ; 175] [36.60 ; 161.58]	102.79	100	seq(16, 30.85, 0.15)	c(2, 0.1)
9 Ziel erreicht	2 1.96	0.15	10 9.60	[5 ; 15] [5.70 ; 14.88]	3.23	25	seq(-4.6, 5, 0.4)	c(2.2, 0.1)
10 Ziel erreicht	2 1.95	0.15	10 9.45	[5 ; 15] [5.53 ; 14.88]	3.23	100	seq(-4.9, 5, 0.1)	c(2.2, 0.1)
11 Ziel erreicht	2 1.95	0.15	10 10.09	[2.5 ; 17.5] [4.48 ; 18.91]	3.84	25	seq(-3.5, 3.7, 0.3)	c(2.2, 0.2)
12 Ziel erreicht	2 1.91	0.15	10 9.07	[2.5 ; 17.5] [3.94 ; 17.34]	3.61	100	seq(-1.3, 13.6, 0.15)	c(1.5, 0.1)
13 Ziel erreicht	2 1.99	0.24	100 91.30	[50 ; 150] [46.99 ; 156.02]	38.45	25	seq(1.7, 4.1, 0.1)	c(3, 0.5)
14 Ziel erreicht	2 1.99	0.24	100 94.28	[50 ; 150] [55.15 ; 148.41]	36.62	100	seq(20.1, 30, 0.1)	c(2, 0.1)
15 Ziel erreicht	2 2.00	0.24	100 100.90	[25 ; 175] [51.94 ; 172.43]	42.38	25	seq(39, 63, 1)	c(2, 0.05)
16 Ziel erreicht	2 1.99	0.24	100 89.98	[25 ; 175] [39.06 ; 172.43]	42.38	100	seq(33.3, 63.2, 0.3)	c(2, 0.05)

Tabelle 4.1: Parametersätze für die Simulationsstudie

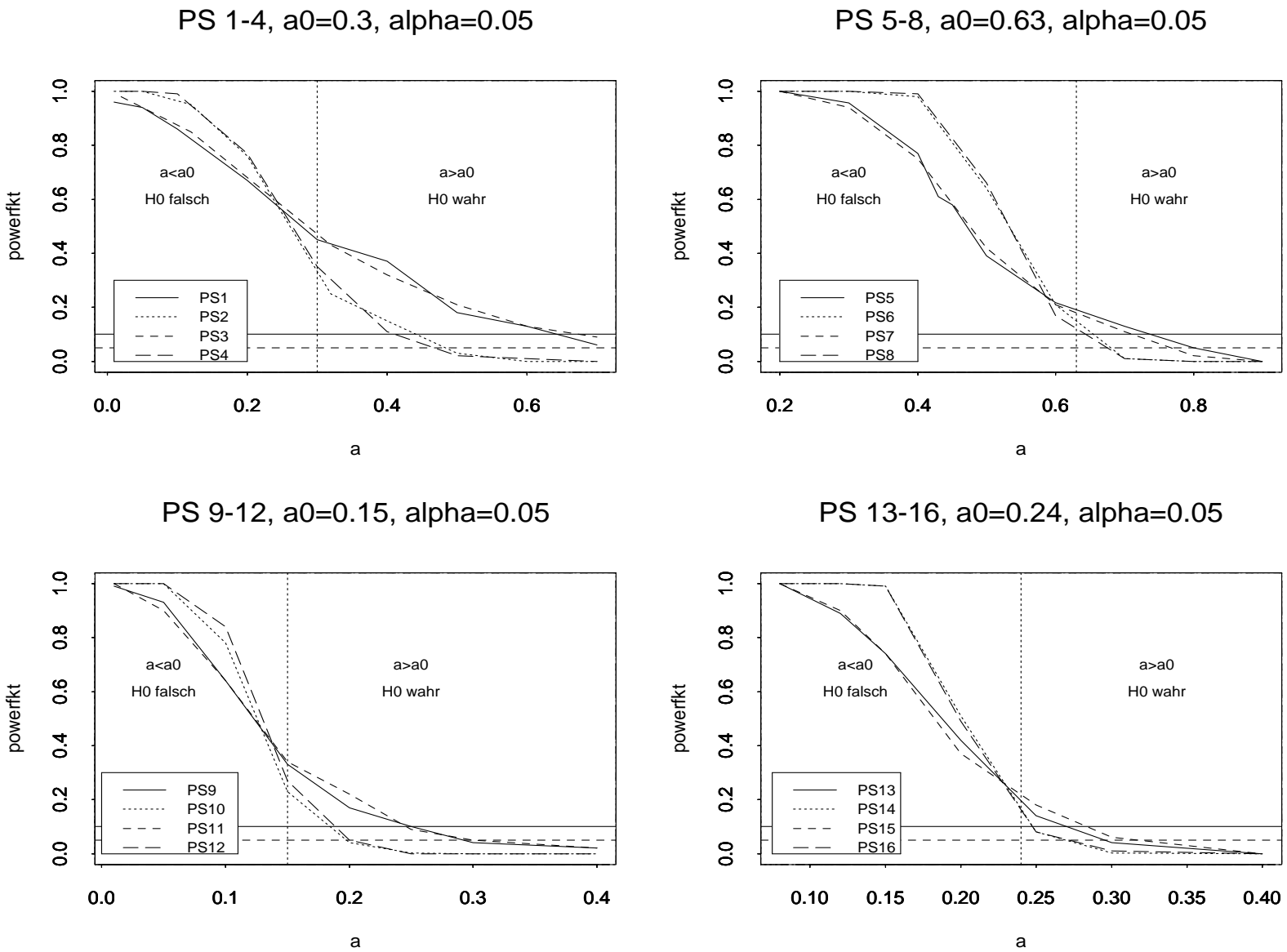


Abbildung 4.1: Powerfunktionen für Test (3.17) mit $\alpha = 0.05$ (gestrichelte Linie) und Test (4.1) mit $\alpha = 0.1$ (durchgezogene Linie)

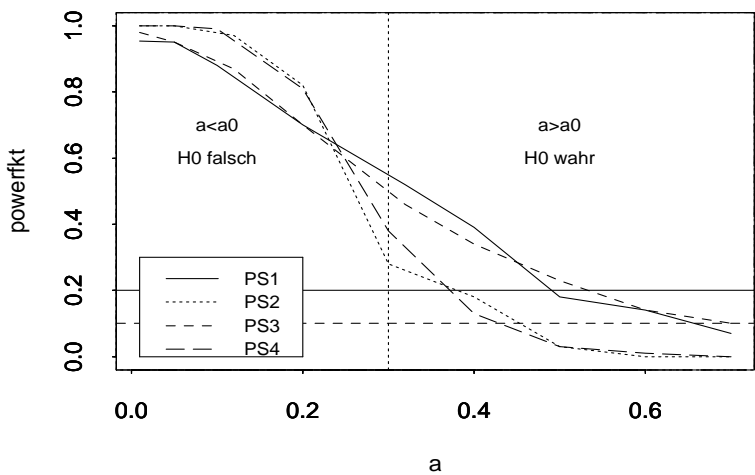
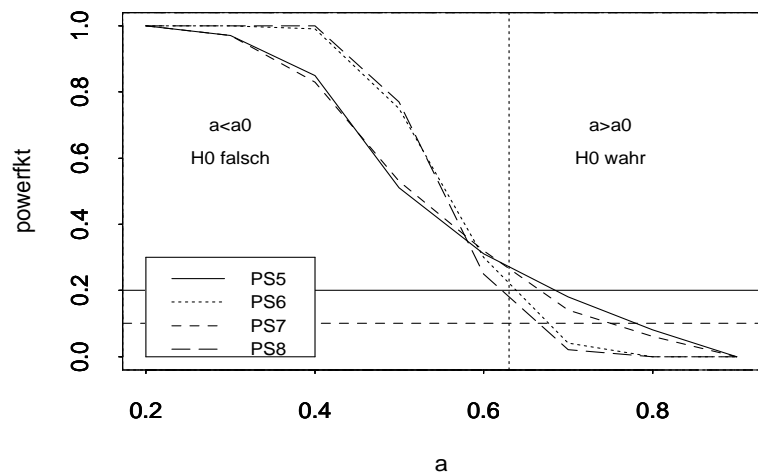
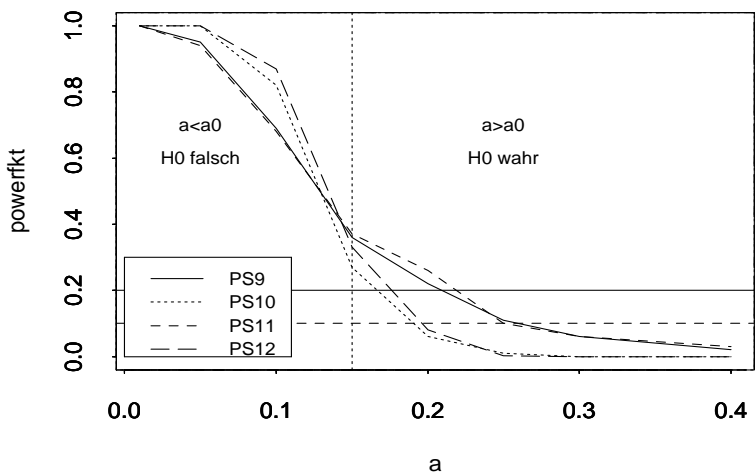
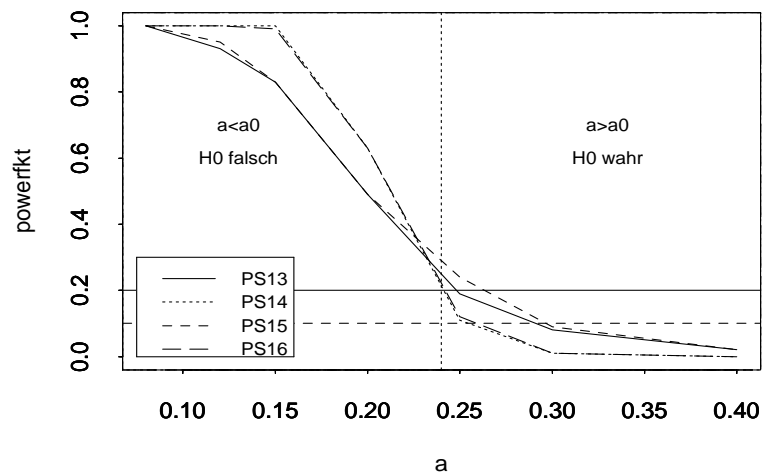
PS 1-4, $a_0=0.3$, $\alpha=0.1$ PS 5-8, $a_0=0.63$, $\alpha=0.1$ PS 9-12, $a_0=0.15$, $\alpha=0.1$ PS 13-16, $a_0=0.24$, $\alpha=0.1$ 

Abbildung 4.2: Powerfunktionen für Test (3.17) mit $\alpha = 0.1$ (gestrichelte Linie) und Test (4.1) mit $\alpha = 0.2$ (durchgezogene Linie)

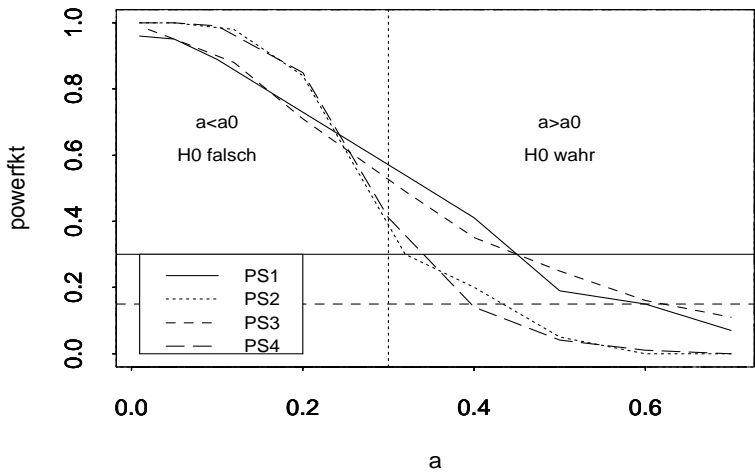
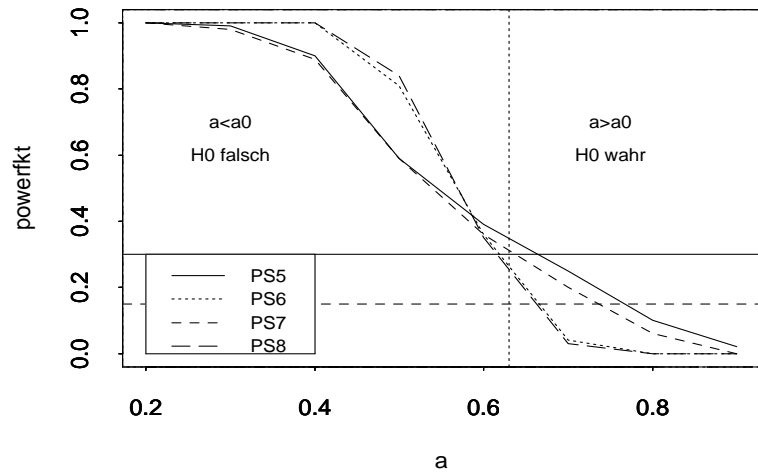
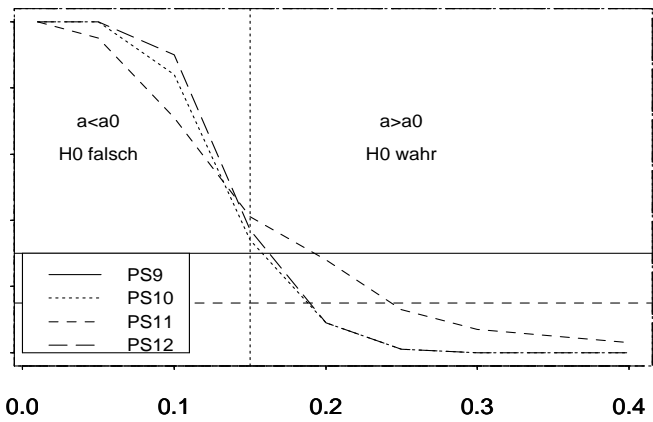
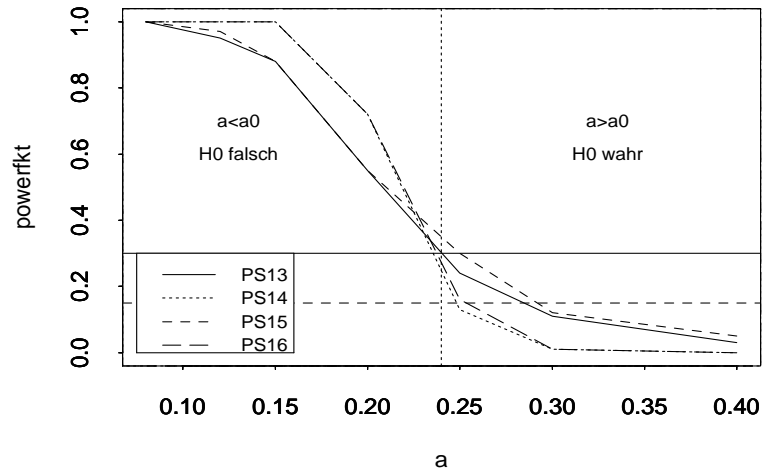
PS 1-4, $a_0=0.3$, $\alpha=0.15$ PS 5-8, $a_0=0.63$, $\alpha=0.15$ PS 13-16, $a_0=0.24$, $\alpha=0.15$ 

Abbildung 4.3: Powerfunktionen für Test (3.17) mit $\alpha = 0.15$ (gestrichelte Linie) und Test (4.1) mit $\alpha = 0.3$ (durchgezogene Linie)

tragen dürfte. In allen 4 Gruppen ist jedoch der Fehler an dieser Stelle deutlich höher, weshalb man von einem liberalen Test spricht.

- Einfluss von n :

Es ist in allen 4 Gruppen gleichermassen zu erkennen, dass die Stichprobenlänge n einen mehr oder weniger deutlichen Einfluss auf die Powerfunktionen ausübt. In der 1. Gruppe beispielsweise verlaufen die Kurven, die zu den PSen 2 und 4 mit $n = 100$ gehören, wesentlich steiler als die Kurven, die zu den PSen 1 und 3 mit $n = 25$ gehören. Durch den steileren Verlauf schneiden die Kurven zu den PSen 2 und 4 die Achse $a = a_0$ deutlich tiefer bei ca. 0.37, während die Kurven mit $n = 25$ die Achse $a = a_0$ bei etwa 0.49 schneiden. In der 2. Gruppe ist der Unterschied an der Grenze $a = a_0$ zwischen den Kurven mit $n = 25$ und $n = 100$ nicht ganz so extrem wie in Gruppe 1, jedoch kann auch hier klar ein steilerer Verlauf der Kurven zu den PSen 6 und 8 mit $n = 100$ ausgemacht werden. Analog dazu können die Gruppen 3 und 4 interpretiert werden, d.h. es lässt sich zusammenfassend feststellen, dass eine grössere Stichprobenlänge n zu steileren Kurven führt, die die zunächst stark vorhandene Liberalität des Tests reduzieren.

- Einfluss der Schwankungsbreite (range) des Erwartungswertes:

Auch bezüglich des Einflusses der Schwankungsbreite ergibt sich in allen 4 Gruppen ein einheitliches Bild. Nachdem in allen 4 Graphiken jeweils die 1. und 3. bzw. 2. und 4. Kurven einer Gruppe nahe aneinander liegen, kann daraus gefolgert werden, dass die Schwankungsbreite des Erwartungswertes keinen relevanten Einfluss auf die Powerfunktionen ausübt, denn sonst müssten sich jeweils die 1. und 3. bzw. 2. und 4. Kurven deutlicher voneinander unterscheiden.

- Einfluss des Signal-to-noise-Quotienten (SN-Quotient):

Da der SN-Quotient in den Gruppen 1 und 2 bzw. 3 und 4 jeweils übereinstimmt, werden nun die oberen 2 Graphiken mit den unteren 2 Graphiken verglichen. Allerdings stellt sich heraus, dass nicht direkt der Wert des SN-Quotienten eine Rolle spielt, sondern eher ein Unterschied zwischen Gruppe 1 und 2 bzw. 3 und 4 festzustellen ist, was vermutlich mit der Wahl von $\bar{\mu}$ zusammenhängt. Sowohl in Gruppe 1 und Gruppe 3 ergibt sich aus den Kurven ein weitaus liberalerer Test als dies für die Gruppen 2 und 4 der Fall ist, d.h. das Fehlerniveau an der Stelle $a = a_0$ ist in den Gruppen 2

und 4 wesentlich niedriger (ca. zwischen 0.15 und 0.2 in Gruppe 2 und 4) als in den Gruppen 1 und 3 (ca. zwischen 0.35 und 0.49 in Gruppe 1 und zwischen 0.22 und 0.35 in Gruppe 3). Dabei ist zu bemerken, dass für die Gruppe 1 ein durchschnittlicher Erwartungswert von $\bar{\mu} = 3$ und in Gruppe 3 ein $\bar{\mu} = 10$ zugrunde gelegt wurde, während bei den Gruppen 2 und 4 jeweils ein $\bar{\mu} = 100$ vorliegt. Daraus lässt sich folgern, dass nicht direkt der SN-Quotient die Power des Bereichstests beeinflusst, aber der damit zusammenhängende Wert von $\bar{\mu}$ sehr wohl den Verlauf der Powerfunktionen beeinträchtigt. Je grösser nämlich der durchschnittliche Erwartungswert ist, desto tiefer schneiden die Kurven die Achse $a = a_0$, d.h. desto weniger liberal wird der Test.

(ii) **Powerfunktionen mit Signifikanzniveau $\alpha = 0.1$ und $\alpha = 0.15$:**

- Einfluss von n :

Analog zu den Powerfunktionen mit Signifikanzniveau $\alpha = 0.05$ kann ein deutlicher Einfluss der Stichprobenlänge n festgestellt werden, der daran zu erkennen ist, dass die Kurven zu den PSen mit $n = 100$ steiler verlaufen und somit ein geringeres Fehlerniveau an der Achse $a = a_0$ verursachen.

- Einfluss der Schwankungsbreite (range) des Erwartungswertes:

Durch die Vergrösserung des Signifikanzniveaus ändert sich die Tatsache nicht, dass die Schwankungsbreite des Erwartungswertes keinen Einfluss auf den Verlauf der Kurven ausübt, denn analog zu den Kurven mit Signifikanzniveau $\alpha = 0.05$ liegen jeweils die 1. und 3. bzw. 2. und 4. Kurven dicht aneinander, so dass kein relevanter Unterschied festzustellen ist. Höchstens in der 1. Gruppe in Abbildung 4.2 können leichte Differenzen der Fehlerniveaus an $a = a_0$ erkannt werden, jedoch schneidet beim Vergleich der PS 1 ($n = 25$, kleine Schwankungsbreite) mit PS 3 ($n = 25$, grosse Schwankungsbreite) die Kurve des PSes 1 die Achse $a = a_0$ bei einem höheren Niveau als die Kurve des PSes 3, während beim Vergleich der 2. und 4. PS die Kurve des PSes 4 mit einer grossen Schwankungsbreite des Erwartungswertes die Achse an einem höheren Niveau schneidet. Es ist also keine Regelmässigkeit beim Vergleich der Kurven mit einer grossen bzw. kleinen Schwankungsbreite zu erkennen, so dass diese kleinen Abweichungen von den Beobachtungen aus den anderen Gruppen auf die Zufälligkeit zurückzuführen sind.

- Einfluss des SN-Quotienten:

Beim Vergleich der oberen 2 Graphiken mit den unteren ist ebenso wie in Abbildung 4.1 mit einem Signifikanzniveau von $\alpha = 0.05$ festzustellen, dass nicht der Wert des SN-Quotienten relevant ist, sondern die Wahl des durchschnittlichen Erwartungswertes $\bar{\mu}$, der, je höher er ist, zu Kurven mit einem niedrigeren Fehlerniveau an $a = a_0$ führt, was an den Gruppen 2 und 4 mit einem $\bar{\mu} = 100$ im Vergleich zu den Gruppen 1 und 3 mit $\bar{\mu} = 3$ bzw. $\bar{\mu} = 10$ zu erkennen ist.

(iii) **Einfluss der Wahl des Signifikanzniveaus:**

Um beurteilen zu können, welches Signifikanzniveau gewählt werden sollte, werden nun alle 3 Abbildungen 4.1, 4.2 und 4.3 mit den unterschiedlichen Niveaus $\alpha = 0.05$, $\alpha = 0.1$ und $\alpha = 0.15$ miteinander verglichen. In allen 3 Abbildungen ist zu erkennen, dass ein liberaler Test vorliegt, d.h. der tatsächliche Fehler ist an der Stelle $a = a_0$ grösser als das vorgegebene Signifikanzniveau. Man kann jedoch beobachten, dass bei Vergrößerung des Signifikanzniveaus von $\alpha = 0.05$ auf $\alpha = 0.1$ sich die Kurven wie erwartet leicht nach oben verschieben, aber dies geschieht vor allem in den Gruppen 2 und 4 in einer äusserst geringen Masse. Während bei einem Niveau von $\alpha = 0.05$ in Gruppe 2 die Schnittpunkte mit der Achse $a = a_0$ zwischen etwa 0.12 und 0.2 liegen, ergeben sich bei einem Niveau von $\alpha = 0.1$ Werte zwischen 0.18 und 0.25. Analog steigen bei Gruppe 4 die Fehlerniveaus von zwischen 0.15 und 0.21 bei $\alpha = 0.05$ auf nur 0.2 bis 0.25 bei $\alpha = 0.1$. Bei den Gruppen 1 und 3 liegen die Fehlerniveaus generell wie schon angesprochen etwas höher, so dass in diesen Fällen stets von einem liberalen Test ausgegangen werden muss. Jedoch ist auch in diesen Gruppen zu erkennen, dass sich mit einer Verdoppelung des Signifikanzniveaus von $\alpha = 0.05$ auf $\alpha = 0.1$ die Fehlerniveaus nur geringfügig verschlechtern. Erhöht man nun das Signifikanzniveau noch einmal auf $\alpha = 0.15$ (Abbildung 4.3), so ergeben sich die kleinsten Fehlerniveaus für die Gruppe 4, deren Werte zwischen ca. 0.2 und 0.32 liegen. Bei einer erlaubten Fehlerwahrscheinlichkeit $\alpha = 0.15$ liefert der Test also ein Fehlerniveau zwischen 0.2 und 0.32, d.h. dass der Test zwar noch immer liberal ist, allerdings deutlich weniger stark als zunächst bei einem Signifikanzniveau von $\alpha = 0.05$ angenommen werden musste.

Zusammenfassung

Es ergibt sich also zusammenfassend, dass man statt eines Signifikanzniveaus von $\alpha = 0.05$, das etwas zu klein ist, ein etwas grösseres Signifikanzniveau von z.B. $\alpha = 0.1$ oder $\alpha = 0.15$ wählen müsste, um einigermaßen einen α -level-Test zu erhalten. Dies könnte auch gerechtfertigt werden, da der zugrundeliegende Bereichstest ein asymptotischer Test ist und deshalb die Fehlerwahrscheinlichkeiten etwas grösser ausfallen können, bis die Asymptotik zum Tragen kommt. Doch trotz Vergrösserung des Signifikanzniveaus verhält sich der Test immer noch in den meisten Fällen sehr liberal. Aus diesem Grund ist es sinnvoll, den Ablehnungsbereich des Bereichstests für Überdispersion auf die folgende Form abzuändern:

Verwerfe die Hypothese $H : a > a_0$ gegen die Alternative $K : a \leq a_0$ zum Signifikanzniveau α , genau dann wenn

$$\Phi \left(\frac{\hat{a} - a_0}{\hat{\sigma}(\hat{\beta}, \hat{a})} \right) \leq \frac{\alpha}{2} \quad (4.1)$$

gilt. Mit dieser Veränderung verhält sich der Test deutlich weniger liberal, so dass die Fehlerniveaus an der Stelle $a = a_0$ etwa dem vorgegebenen Signifikanzniveau entsprechen. Zu erkennen ist dies in den Abbildungen an der durchgezogenen horizontalen Linie, die das neue Signifikanzniveau angibt. Beispielsweise sind die Kurven in Abbildung 4.1 zu einem $\alpha = 0.05$ berechnet. Dieses α entspricht aber nun nach der neuen Testvorschrift (4.1) dem $\frac{\alpha}{2}$ aus dem neuen Ablehnungsbereich. Das dazugehörige Signifikanzniveau beträgt daher dann $\alpha = 0.1$, das mit der durchgezogenen Linie gekennzeichnet ist. In Abbildung 4.1 ist die Verbesserung durch die neue Testvorschrift noch eher gering, während sie schon in Abbildung 4.2 deutlich zu erkennen ist. Die Kurven sind hier zu einem $\alpha = 0.1$ gezeichnet, was bedeutet, dass nach (4.1) das dazugehörige Signifikanzniveau 0.2 beträgt. Dies ist auch der Bereich, in dem vor allem die Kurven, die zu den PSen mit $n = 100$ und grossen Erwartungswerten $\bar{\mu}$ gehören (siehe Gruppe 2 und 4), die Achse $a = a_0$ schneiden. Somit liefert der Test mit der neuen Testvorschrift (4.1) an der Stelle $a = a_0$ in etwa das Fehlerniveau, das durch das Signifikanzniveau vorgegeben wird.

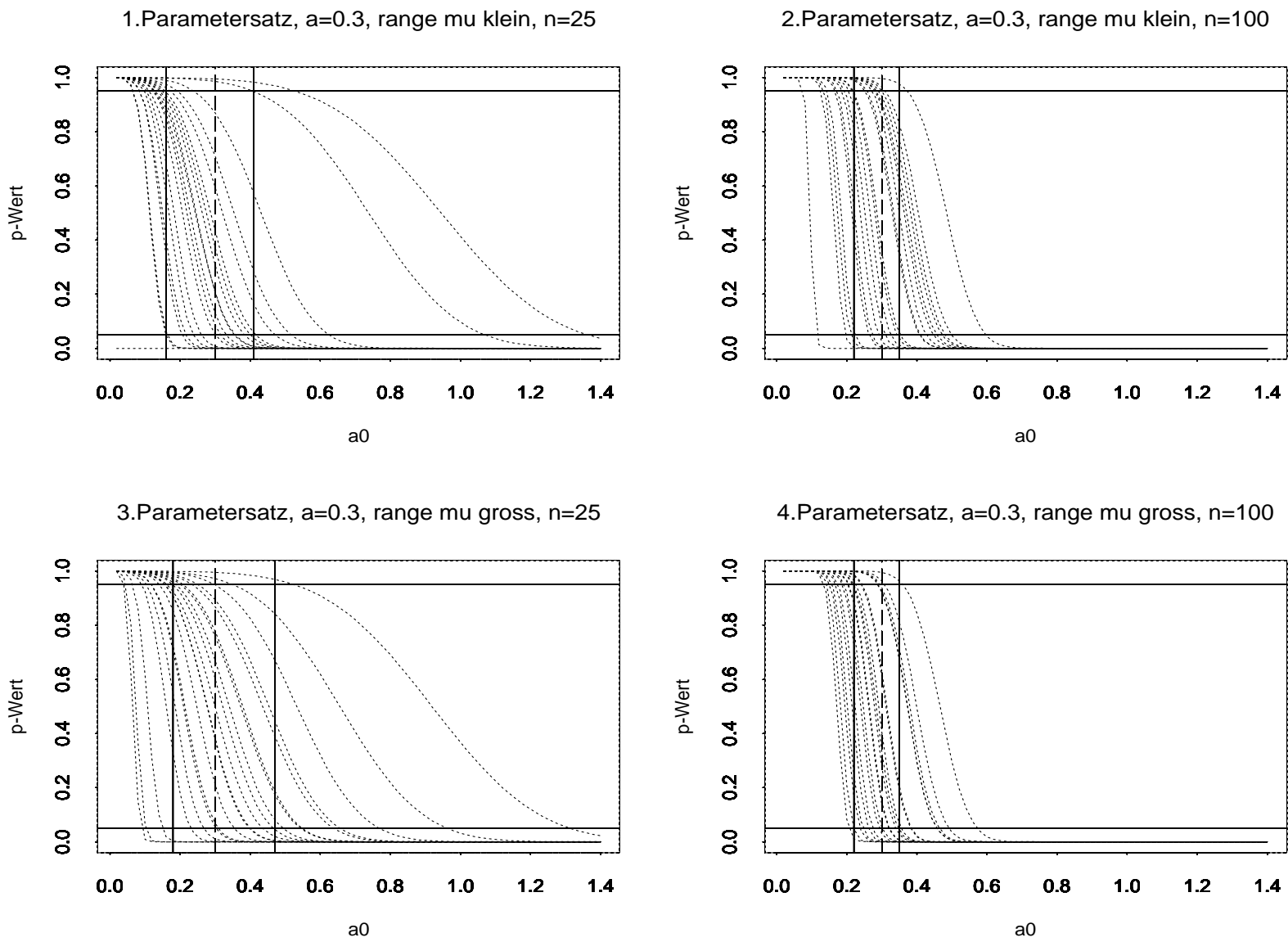
4.2 Simulation der p-Wert-Kurven

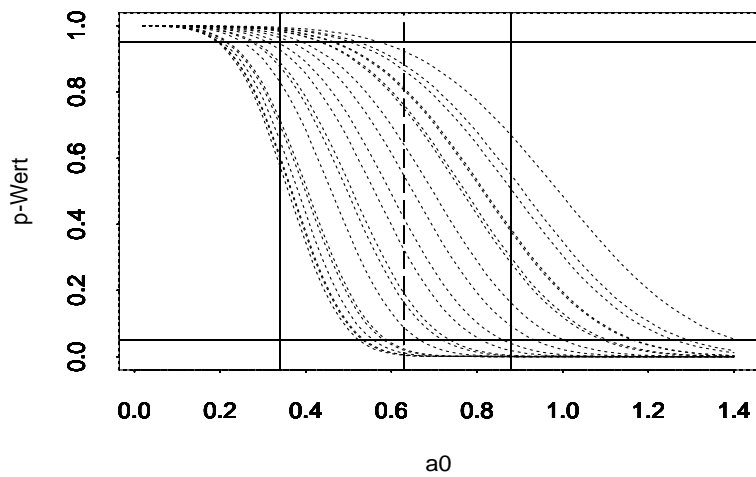
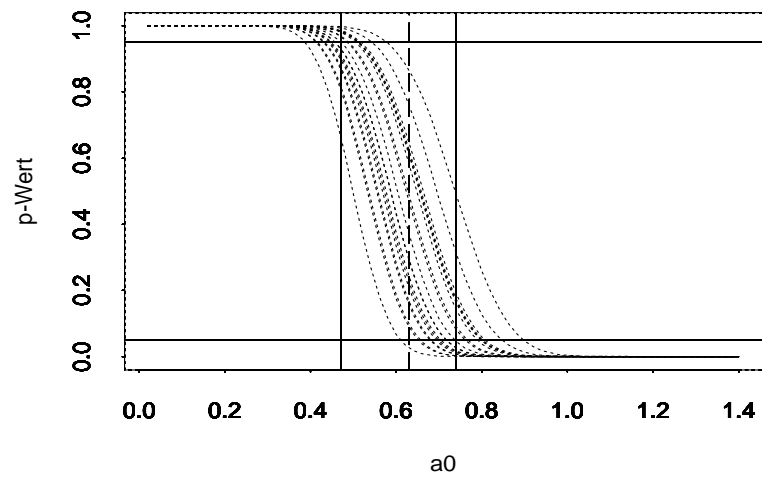
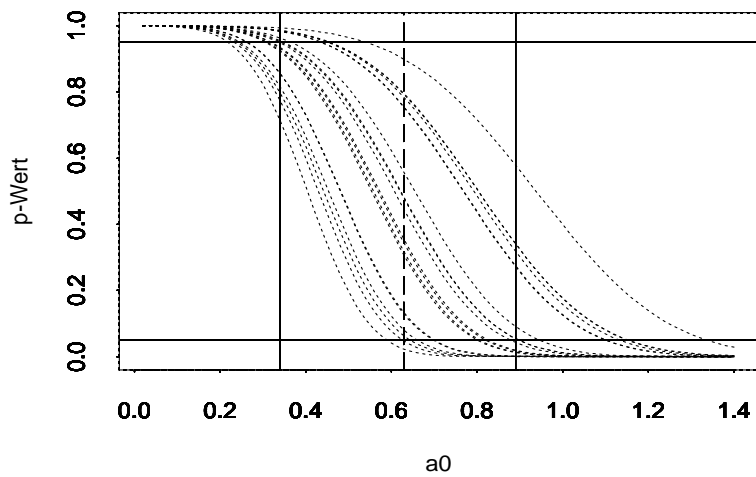
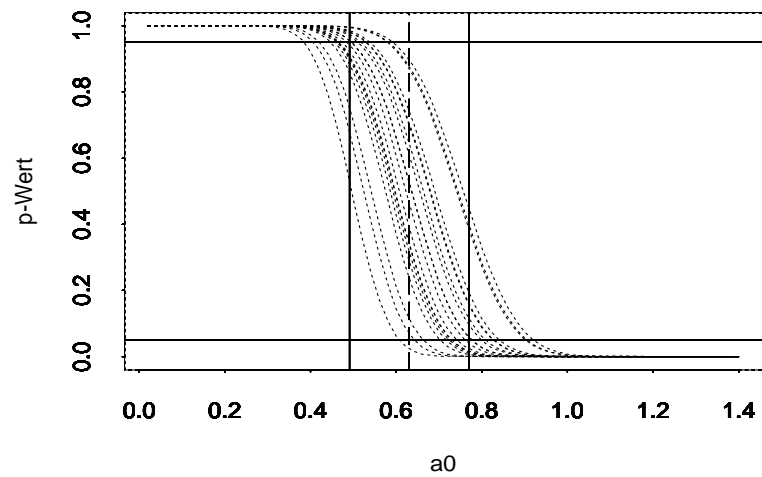
Wie in Abschnitt 3.3.4 erläutert, werden für jeden einzelnen PS 20 p-Wert-Kurven erzeugt, die in einer Graphik pro PS zusammen abgebildet sind. Dieser Vorgang wird

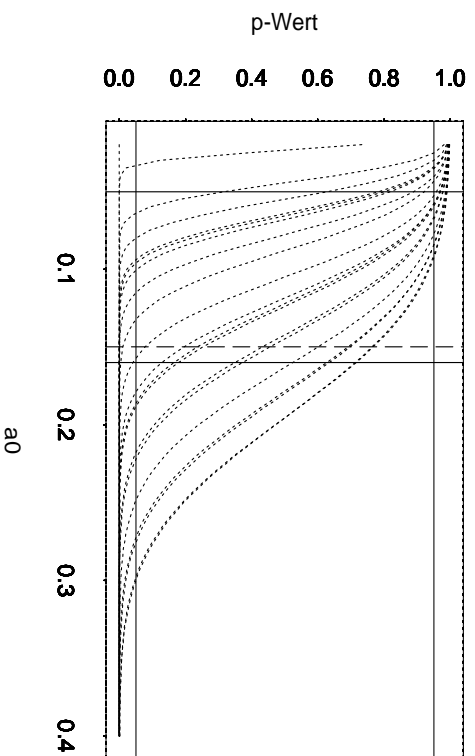
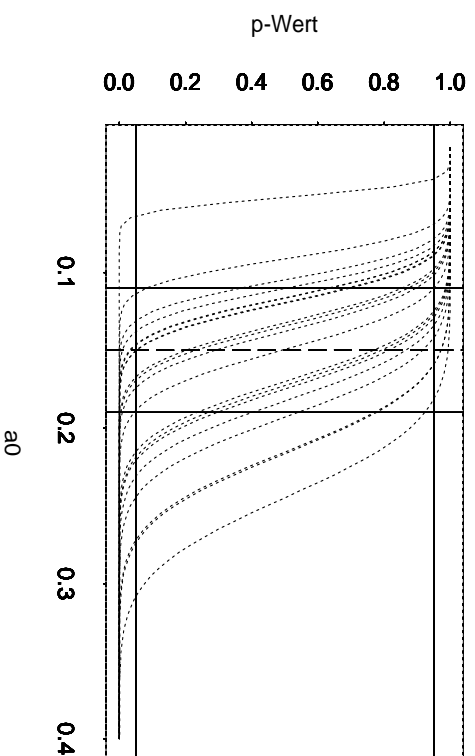
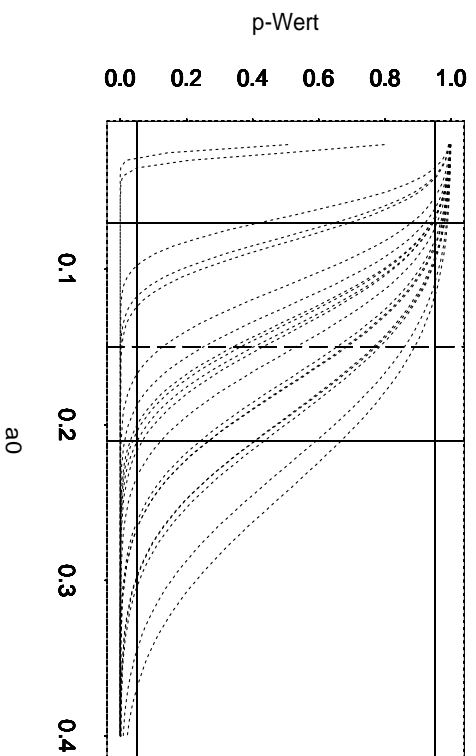
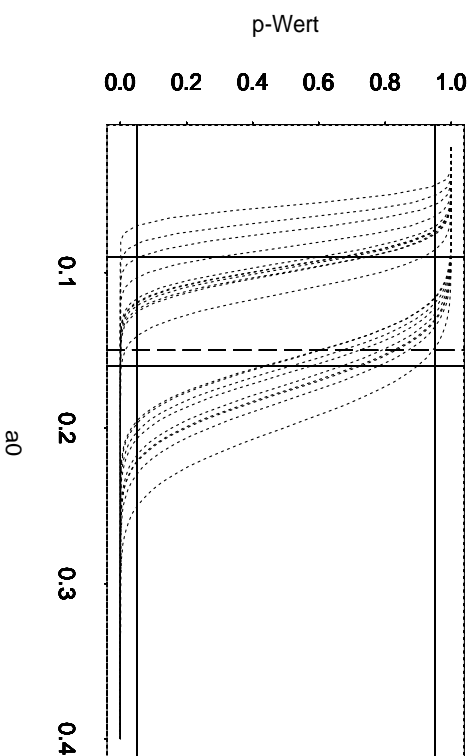
sowohl für Kurven bezüglich a_0 , als auch für Kurven bezüglich des Abstandsmasses d_0 durchgeführt, und es werden wie bei den Powerfunktionen die PSe in 4 Gruppen mit je gleichem Parameter a und gleichem gemittelten Erwartungswert $\bar{\mu}$ eingeteilt.

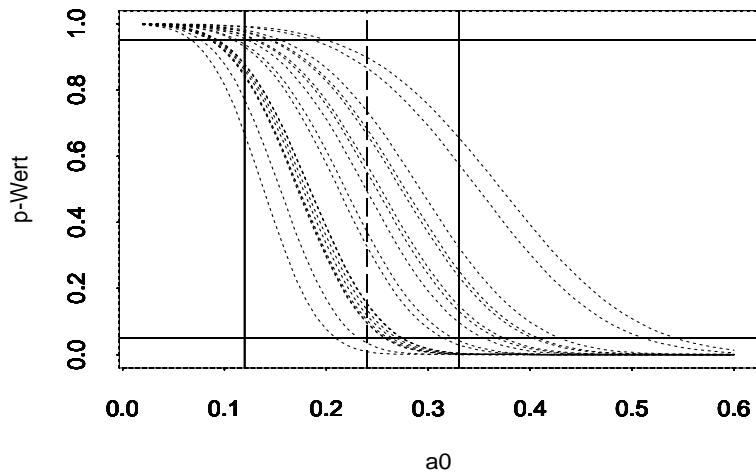
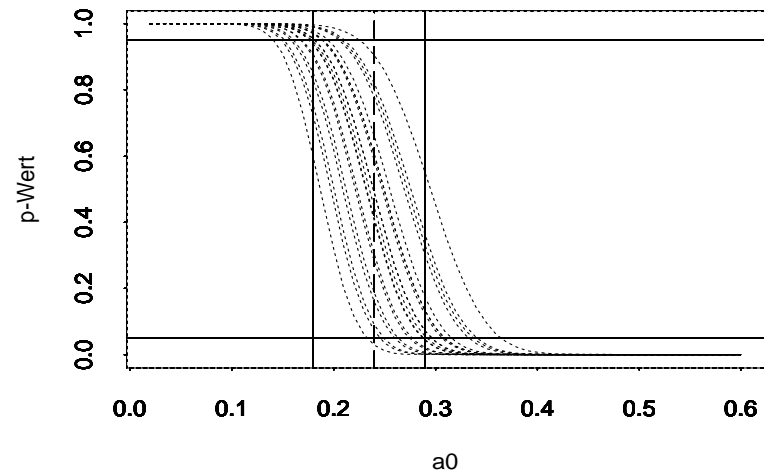
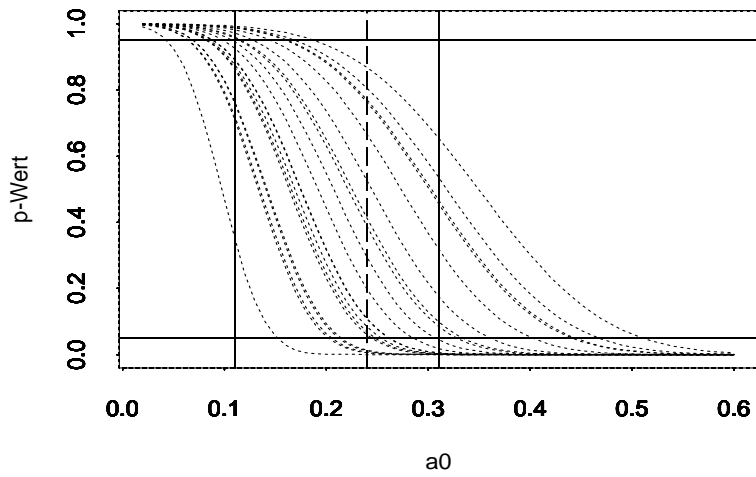
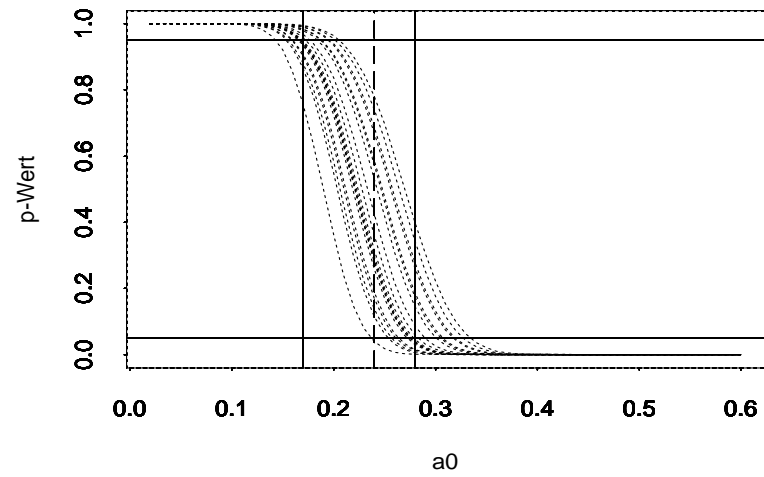
4.2.1 p-Wert-Kurven für a_0

In diesem Unterabschnitt werden die Ergebnisse der p-Wert-Kurven bezüglich a_0 (siehe 3.26) zusammengefasst. Zunächst folgen die Abbildungen 4.4, 4.5, 4.6 und 4.7, in denen die je 20 p-Wert-Kurven pro PS dargestellt sind, wobei zusätzlich als gestrichelte Linie der wahre Wert a , und als durchgezogene Linien die gemittelten Werte \bar{a}_{0l} und \bar{a}_{0u} eingetragen sind. In der anschliessenden Tabelle 4.2 sind die Ergebnisse, die sich für jeden PS für die gemittelten Schnittpunkte \bar{a}_{0l} und \bar{a}_{0u} der Kurven mit den Levels $\hat{P} = 0.95$ und $\hat{P} = 0.05$ ergeben, eingetragen. Zusätzlich sind die Abweichungen $a - \bar{a}_{0l}$ und $\bar{a}_{0u} - a$ in % angegeben, wobei der feste Wert a 100% entspricht. Die Levels $\hat{P} = 0.95$ und $\hat{P} = 0.05$ ergeben sich nach Abschnitt 3.3.3 aus der Verwendung eines Signifikanzniveaus von $\alpha = 0.05$. Da sich erst im Nachhinein anhand der Powerfunktionen ergab, dass besser ein grösseres Signifikanzniveau verwendet werden sollte (siehe vorangegangener Abschnitt), wurden die Simulationen der p-Wert-Kurven weiterhin mit einem kleinen $\alpha = 0.05$ durchgeführt. Es ist aber zu bemerken, dass die nachfolgende Interpretation auch bei der Verwendung eines grösseren Signifikanzniveaus immer noch richtig bleibt und sich nur geringfügig in den konkreten Werten von \bar{a}_{0l} und \bar{a}_{0u} ändert.

Abbildung 4.4: p-Wert-Kurven der Parametersätze 1-4 bezüglich a_0

5.Parametersatz, $a=0.63$, range μ klein, $n=25$ 6.Parametersatz, $a=0.63$, range μ klein, $n=100$ 7.Parametersatz, $a=0.63$, range μ gross, $n=25$ 8.Parametersatz, $a=0.63$, range μ gross, $n=100$ Abbildung 4.5: p-Wert-Kurven der Parametersätze 5-8 bezüglich a_0

9. Parametersatz, $a=0.15$, range μ klein, $n=25$ 10. Parametersatz, $a=0.15$, range μ klein, $n=100$ 11. Parametersatz, $a=0.15$, range μ gross, $n=25$ 12. Parametersatz, $a=0.15$, range μ gross, $n=100$ Abbildung 4.6: p-Wert-Kurven der Parametersätze 9-12 bezüglich a_0

13. Parametersatz, $a=0.24$, range μ klein, $n=25$ 14. Parametersatz, $a=0.24$, range μ klein, $n=100$ 15. Parametersatz, $a=0.24$, range μ gross, $n=25$ 16. Parametersatz, $a=0.24$, range μ gross, $n=100$ Abbildung 4.7: p-Wert-Kurven der Parametersätze 13-16 bezüglich a_0

						Diskriminanz zu Poisson		Akzeptanz zu Poisson	
PS	n	$SN(a, \bar{\mu})$	$\bar{\mu}$	Schwankungs- intervall μ_i	wahres a	\bar{a}_{0l}	$a - \bar{a}_{0l}$ in %	\bar{a}_{0u}	$\bar{a}_{0u} - a$ in %
1	25	1.25	3	[1.5 ; 4.5]	0.3	0.16	47%	0.41	37%
2	100	1.25	3	[1.5 ; 4.5]	0.3	0.22	27%	0.35	17%
3	25	1.25	3	[0.75 ; 5.25]	0.3	0.18	40%	0.47	57%
4	100	1.25	3	[0.75 ; 5.25]	0.3	0.22	27%	0.35	17%
5	25	1.25	100	[50 ; 150]	0.63	0.34	46%	0.88	40%
6	100	1.25	100	[50 ; 150]	0.63	0.47	25%	0.74	17%
7	25	1.25	100	[25 ; 175]	0.63	0.34	46%	0.89	41%
8	100	1.25	100	[25 ; 175]	0.63	0.49	22%	0.77	22%
9	25	2	10	[5 ; 15]	0.15	0.05	67%	0.16	7%
10	100	2	10	[5 ; 15]	0.15	0.11	27%	0.19	27%
11	25	2	10	[2.5 ; 17.5]	0.15	0.07	53%	0.21	40%
12	100	2	10	[2.5 ; 17.5]	0.15	0.09	40%	0.16	7%
13	25	2	100	[50 ; 150]	0.24	0.12	50%	0.33	38%
14	100	2	100	[50 ; 150]	0.24	0.18	25%	0.29	21%
15	25	2	100	[25 ; 175]	0.24	0.11	54%	0.31	29%
16	100	2	100	[25 ; 175]	0.24	0.17	29%	0.28	17%

Tabelle 4.2: Ergebnisse der p-Wert-Kurven bezüglich a_0 , d.h. $\hat{P}(\hat{a}, \hat{\sigma}, a_0) = \Phi\left(\frac{\hat{a}-a_0}{\hat{\sigma}(\hat{\beta}, \hat{a})}\right)$

Interpretation der p-Wert-Kurven bezüglich a_0

(i) Gruppe 1: Parametersätze 1 bis 4

- Einfluss von n :

Um einen Überblick über den Einfluss der Stichprobenlänge n auf die Testergebnisse zu erhalten, müssen die PSe 1 mit 2 und 3 mit 4 verglichen werden. Dabei fällt auf, dass beim 1. PS die Differenzen von $\bar{a}_{0l} = 0.16$ und $\bar{a}_{0u} = 0.41$ zum wahren Wert $a = 0.3$ (in Prozent ergeben sich damit die Werte 47% und 37%) sehr gross im Vergleich zum 2. PS sind, bei dem $\bar{a}_{0l} = 0.22$ und $\bar{a}_{0u} = 0.35$, was einer Abweichung vom wahren Wert von nur 27% und 17% entspricht. Die Werte \bar{a}_{0l} und \bar{a}_{0u} liegen also beim 2. PS mit einem Da-

tenumfang von $n = 100$ sehr viel näher am wahren Wert a , als beim PS mit $n = 25$. Sowohl an den angegebenen Zahlen in der Tabelle 4.2, als auch an den Kurven in Abbildung 4.4 ist zu erkennen, dass sich die Abweichungen von \bar{a}_{0l} und \bar{a}_{0u} zum wahren Wert a im Durchschnitt etwa halbieren, wenn man die PSe 1 mit 2 vergleicht, d.h. wenn sich die Datenlänge von $n = 25$ auf $n = 100$ vergrössert. Analoge Ergebnisse treten für die PSe 3 und 4 auf. Während beim 3. PS die Werte $\bar{a}_{0l} = 0.18$ und $\bar{a}_{0u} = 0.47$ (entspricht einer Differenz zu $a = 0.3$ von 40% und 57%) weit von $a = 0.3$ entfernt liegen, nähern sie sich im Schnitt beim 4. PS mit $n = 100$ stark an den wahren Wert an ($\bar{a}_{0l} = 0.22$, was einer Differenz von 27% entspricht, und $\bar{a}_{0u} = 0.35$, was einer Differenz von nur 17% entspricht). Mit wachsendem n werden die Kurven also steiler und liegen enger aneinander, womit sich das Intervall $[\bar{a}_{0l}, \bar{a}_{0u}]$ um den wahren Wert a stark verkleinert. Betrachtet man auch hier wieder die Kurven des 3. und 4. PSes in Abbildung 4.4, so kann optisch die starke Verkleinerung des Intervalls $[\bar{a}_{0l}, \bar{a}_{0u}]$ festgestellt werden, die wie bei den PSen 1 und 2 in etwa eine Halbierung ergibt, wenn vom PS 3 mit $n = 25$ zum PS 4 mit $n = 100$ gewechselt wird. Bezüglich der Modellwahl kann nun festgehalten werden, dass man im 1. und 3. PS schliesslich einen Wert von $\bar{a}_{0u} = 0.41$ bzw. $\bar{a}_{0u} = 0.47$ akzeptieren müsste, um das Poissonmodell beibehalten zu können, während ein signifikanter Abstand von nur $\bar{a}_{0l} = 0.16$ bzw. $\bar{a}_{0l} = 0.18$ gemessen wird, d.h. wenn man höchstens eine Schranke a_0 von 0.16 bzw. 0.18 akzeptiert, so ist in diesem Falle der Aufwand gerechtfertigt, vom Poisson- zum NB-Modell zu wechseln. Bei den PSen 2 und 4 dagegen ist der Unterschied zwischen \bar{a}_{0l} und \bar{a}_{0u} sehr gering, was zu dem positiven Effekt führt, dass der Bereich zwischen \bar{a}_{0l} und \bar{a}_{0u} , in dem keine Aussagen bezüglich Annahme oder Ablehnung des Poissonmodells getroffen werden können, stark verkleinert wird. In diesen Fällen muss dann überprüft werden, ob der Wert \bar{a}_{0u} (hier $\bar{a}_{0u} = 0.35$ im 2. PS und 4. PS), der die Akzeptanzgrenze des Poissonmodells angibt, noch so klein ist, dass er angenommen werden kann. Ist dies der Fall, so kann das Poissonmodell beibehalten werden, während man zum NB-Modell wechseln muss, sobald einem die Grenze $\bar{a}_{0u} = 0.35$ zu gross erscheint. An dieser Stelle ist jedoch zu bemerken, dass es generell sehr schwierig ist, die Werte des Dispersionsindizes a_0 einzuschätzen und zu beurteilen, welcher Wert \bar{a}_{0u} als zu gross bzw. als klein genug für das Poissonmodell eingestuft werden kann.

- Einfluss der Schwankungsbreite (range) des Erwartungswertes:

Dazu müssen die Ergebnisse der PSe 1 mit 3 und 2 mit 4 verglichen werden. Allerdings sind hier keine nennenswerten Unterschiede auszumachen, weshalb in diesen Fällen gefolgert werden kann, dass die Schwankungsbreite der Erwartungswerte keinen relevanten Einfluss auf die Testergebnisse hat.

(ii) **Gruppe 2: Parametersätze 5 bis 8:**

- Einfluss von n :

Hierfür werden zunächst die Ergebnisse der PSe 5 mit 6 und 7 mit 8 verglichen. Bei PS 5 ergeben sich $\bar{a}_{0l} = 0.34$ und $\bar{a}_{0u} = 0.88$, was prozentual einer relativ grossen Abweichung von 46% und 40% von $a = 0.63$ entspricht, im Gegensatz zum 6. PS mit $\bar{a}_{0l} = 0.47$ und $\bar{a}_{0u} = 0.74$, bei dem die Abweichung nur 25% und 17% beträgt, und damit im Vergleich zum 5. PS mehr als halbiert wird. Der deutliche Einfluss der Datenlänge n und die durchschnittliche Halbierung der Intervalle $[\bar{a}_{0l}; \bar{a}_{0u}]$ lassen sich auch klar beim Vergleich der Kurven von PS 5 mit PS 6 in Abbildung 4.5 erkennen. Eine ähnliche Situation ergibt sich für den 7. und 8. PS. Während beim 7. PS \bar{a}_{0l} und \bar{a}_{0u} mit $\bar{a}_{0l} = 0.34$ und $\bar{a}_{0u} = 0.89$ stark von $a = 0.63$ abweichen (46% und 41%), liegen sie beim 8. PS mit $\bar{a}_{0l} = 0.49$ und $\bar{a}_{0u} = 0.77$ deutlich näher am wahren Wert a (Abweichung von je 22%). Auch hier ergibt sich beim Übergang von PS 7 mit $n = 25$ zum PS 8 mit $n = 100$ in etwa eine Halbierung der Abweichungen, die in Abbildung 4.5 noch einmal graphisch verdeutlicht werden. Analog zu den Daten der 1. Gruppe würde man für die PSe 5 und 7 mit $n = 25$ einen Wechsel vom Poisson- zum NB-Modell rechtfertigen können, da ein Wert $\bar{a}_{0u} = 0.88$ bzw. $\bar{a}_{0u} = 0.89$, den man für den Erhalt des Poissonmodells akzeptieren müsste, nicht vertretbar ist. Beim 6. und 8. PS ist dagegen der Unterschied zwischen \bar{a}_{0l} und \bar{a}_{0u} so gering, dass nur der Wert \bar{a}_{0u} zu betrachten ist und abgewogen werden muss, ob dieser Wert für ein Poissonmodell noch akzeptabel ist. Sollte dies nicht der Fall sein, so müsste man den Aufwand in Kauf nehmen, die Daten mittels des NB-Modells zu modellieren.

- Einfluss der Schwankungsbreite (range) des Erwartungswertes:

Analog wie bei der Gruppe 1 ist hier ebenfalls kein relevanter Unterschied beim Vergleich der PSe 5 mit 7 und 6 mit 8 zu erkennen.

(iii) **Gruppe 3: Parametersätze 9 bis 12 und Gruppe 4: Parametersätze 13 bis 16:**

Da sich sowohl für die 3. Gruppe, als auch für die Gruppe 4 analoge Ergebnisse herausstellen wie für die bisher behandelten Gruppen 1 und 2, werden hier die einzelnen Resultate nicht explizit angegeben, sondern können in der Ergebnistabelle 4.2 nachgelesen werden. Der einzige auffällige Unterschied zu den Gruppen 1 und 2 ist die Asymmetrie der Abweichungen von \bar{a}_{0l} und \bar{a}_{0u} zu a in Gruppe 3. Während sich bei den Gruppen 1 und 2 die Abweichungsintervalle $[\bar{a}_{0l}; \bar{a}_{0u}]$ von den PSen mit $n = 25$ auf die PSe mit $n = 100$ ungefähr halbieren oder zumindest stark verkleinern, fällt in Gruppe 3 beim Vergleich der PSe 9 mit 10 auf, dass sich die Differenz von \bar{a}_{0u} zu a nicht wie in allen anderen Fällen verringert, sondern von $\bar{a}_{0u} = 0.16$ im PS 9 auf $\bar{a}_{0u} = 0.19$ vergrößert. Dieser Sachverhalt ist in den Graphiken zu den PSen 9 und 10 in Abbildung 4.6 deutlich zu erkennen. Wie ebenfalls graphisch festzustellen ist, liegt auch beim 12. PS eine auffällige Asymmetrie des Intervalls $[\bar{a}_{0l}; \bar{a}_{0u}]$ um a vor, allerdings entspricht der 12. PS immer noch den Beobachtungen aus den anderen Gruppen, dass sich mit grösserem Datenumfang n das Abweichungsintervall verkleinert, was hier trotz starker Asymmetrie der Fall ist. Vergleicht man nun noch in der Gruppe 3 die PSe 9 mit 11 und 10 mit 12 um den Einfluss der Schwankungsbreite des Erwartungswertes zu beurteilen, so erkennt man, dass sich die Gruppe 3 von allen anderen Gruppen, die diesbezüglich keinen Einfluss aufweisen, unterscheidet. Beim Vergleich der PSe 9 mit 11 fällt an den Kurven in Abbildung 4.6 auf, dass der Wert von \bar{a}_{0l} in beiden Fällen in etwa gleich bleibt, während sich der Wert von \bar{a}_{0u} im 11. PS weiter nach rechts verschiebt, d.h. mit einer grösseren Schwankungsbreite des Erwartungswertes im PS 11 zentriert sich das Abweichungsintervall $[\bar{a}_{0l}; \bar{a}_{0u}]$ besser um den wahren Wert a . Damit unterscheidet sich auch die Breite des Abweichungsintervalls, das im 9. PS durch die Asymmetrie kleiner als im 11. PS ist. Betrachtet man nun die Breite des Intervalls $[\bar{a}_{0l}; \bar{a}_{0u}]$ für die PSe 10 und 12, so ergibt sich anhand der Graphiken, dass diese zwar in beiden Fällen ungefähr übereinstimmt, allerdings ist das Intervall im 12. PS leicht nach links verschoben, woraus die schon angesprochene Asymmetrie resultiert. Im Gegensatz zu allen anderen Gruppen ist hier nun zusammenfassend ein Unterschied zwischen den PSen mit grossem und mit kleinem Schwankungsintervall des Erwartungswertes zu erkennen, der jedoch nicht in eine Regel zusammengefasst werden kann, da die Effekte zwischen den PSen 9 und 11 bzw. 10 und 12 jeweils unterschied-

lich sind. Deshalb ist zu vermuten, dass diese Abweichungen von den bisherigen Beobachtungen aus den anderen Gruppen durch die Zufälligkeit der Daten zu erklären ist.

(iv) **Einfluss des Signal-to-noise-Quotienten (SN-Quotient):**

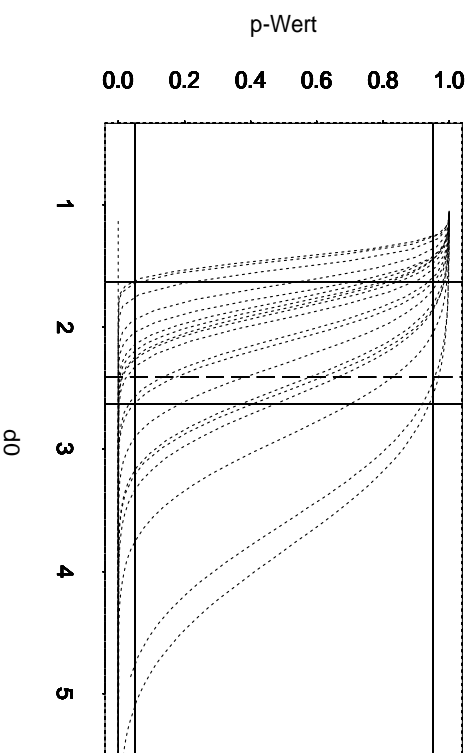
Da der SN-Quotient jeweils in den Gruppen 1 und 2 bzw. 3 und 4 übereinstimmt, werden nun die ersten beiden mit den letzten beiden Gruppen verglichen. Die Betrachtung an den Abbildungen 4.4, 4.5, 4.6 und 4.7 ist hier allerdings mit Vorsicht zu geniessen, da nur die ersten beiden Gruppen bezüglich der x-Achse a_0 einen gleichen Massstab aufweisen, weshalb der Vergleich mit den anderen beiden Gruppen durch die optische Verzerrung erschwert wird. Es kann aber festgestellt werden, dass ähnlich wie bei den Powerfunktionen sich keine relevanten Unterschiede zwischen den Gruppen 1,2 mit 3,4 ergeben, woraus zu schliessen ist, dass der Wert des SN-Quotienten nicht direkt den Verlauf der Kurven beeinflusst. Vielmehr ist beim Vergleich der Gruppen 1 und 2 an den Abbildungen 4.4 und 4.5 zu erkennen, dass sich für alle 4 PSe der 1. Gruppe mit $a = 0.3$ und $\bar{\mu} = 3$ die Breite des Intervalls $[\bar{a}_{0l}; \bar{a}_{0u}]$ beim Übergang zur 2. Gruppe verdoppelt, was mit der Verdoppelung des Dispersionsparameters von $a = 0.3$ auf $a = 0.63$ zusammenhängt und woraus resultiert, dass die prozentualen Abweichungen in beiden Gruppen in etwa wieder übereinstimmen. Ein analoges Verhalten kann für die Gruppen 3 und 4 beobachtet werden. Der Dispersionsparameter vergrössert sich dabei von $a = 0.15$ auf $a = 0.24$ und entsprechend werden die Intervalle $[\bar{a}_{0l}; \bar{a}_{0u}]$ in der 4. Gruppe breiter, was jedoch wegen der unterschiedlichen Massstäbe nicht an den Abbildungen 4.6 und 4.7 zu erkennen ist, aber an den konkreten Werten von \bar{a}_{0l} und \bar{a}_{0u} in der Tabelle 4.2 nachvollzogen werden kann. Auch hier gilt, dass die prozentualen Abweichungen von beiden Gruppen ungefähr gleich gross sind, so dass insgesamt kein Effekt des SN-Quotienten bzw. der Wahl von a und $\bar{\mu}$ zu vermuten ist.

4.2.2 p-Wert-Kurven für d_0

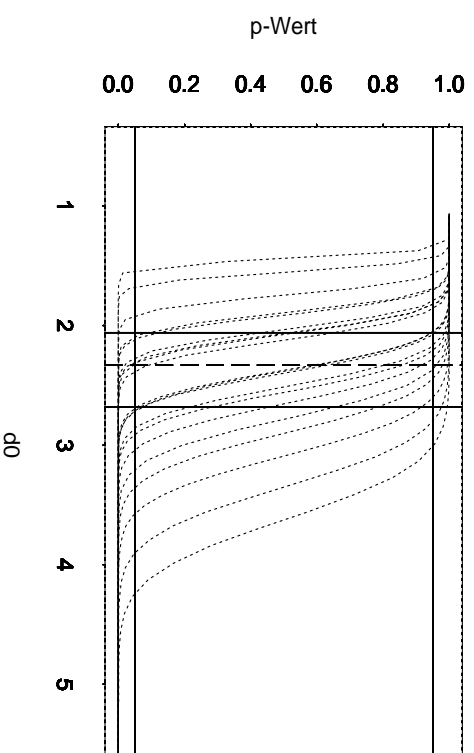
Die Schwierigkeit bei den bisher betrachteten p-Wert-Kurven bezüglich a_0 ist, diese richtig zu interpretieren, denn man kann kaum Aussagen darüber treffen, was z.B. ein Wert $\bar{a}_{0l} = 0.18$ für die Beziehung zwischen Poisson- und NB-Modell bedeutet. Aus diesem Grund werden nun die p-Wert-Kurven bezüglich des Abstandsmasses d betrachtet, das nach (3.27) und (3.28) das Verhältnis der maximalen Varianzen der beiden Mo-

delle angibt, womit eine einfachere Interpretation der Kurven ermöglicht wird. Analog wie im vorherigen Unterabschnitt werden in den Abbildungen 4.8, 4.9, 4.10 und 4.11 zunächst die je 20 Kurven bezüglich d_0 eines PSes in einer Graphik zusammengefasst und zusätzlich mit einer gestrichelten Linie der wahre Wert d , sowie mit einer durchgezogenen Linie die Werte \bar{d}_{0l} und \bar{d}_{0u} eingetragen. Anschliessend werden die Ergebnisse, die sich nun für die Kurven bezüglich des Abstandsmasses $d_0 = d(a_0)$ ergeben, in der Tabelle 4.3 angegeben. Hier beschreibt d den wahren Wert des Abstandsmasses, das in (3.28) definiert wurde. Die Werte \bar{d}_{0l} und \bar{d}_{0u} ergeben sich analog zu \bar{a}_{0l} und \bar{a}_{0u} aus den gemittelten Werten der Schnittpunkte der Kurven mit den Levels $\hat{P} = 0.95$ und $\hat{P} = 0.05$. Ebenso wie in Tabelle 4.2 sind auch hier die Differenzen $d - \bar{d}_{0l}$ und $\bar{d}_{0u} - d$ je in % angegeben, wobei der wahre Wert d 100% entspricht.

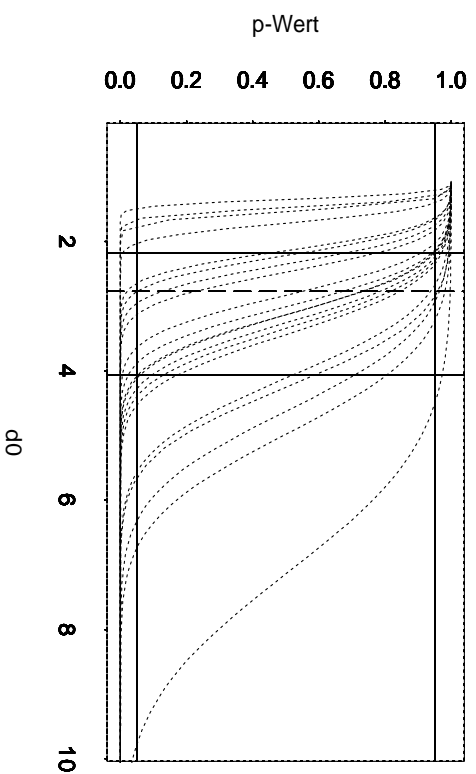
1. Parametersatz, $d=2.41$, range μ klein, $n=25$



2. Parametersatz, $d=2.33$, range μ klein, $n=100$



3. Parametersatz, $d=2.77$, range μ gross, $n=25$



4. Parametersatz, $d=2.99$, range μ gross, $n=100$

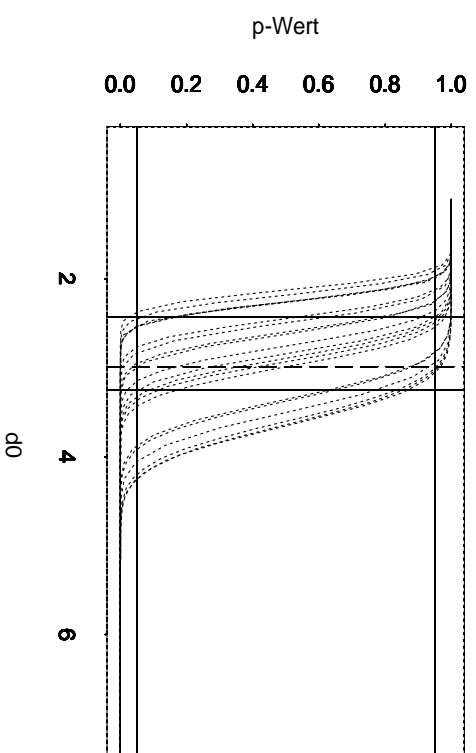
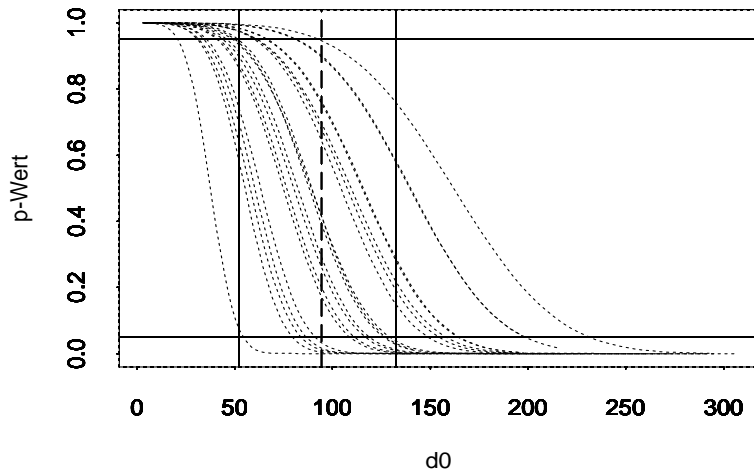
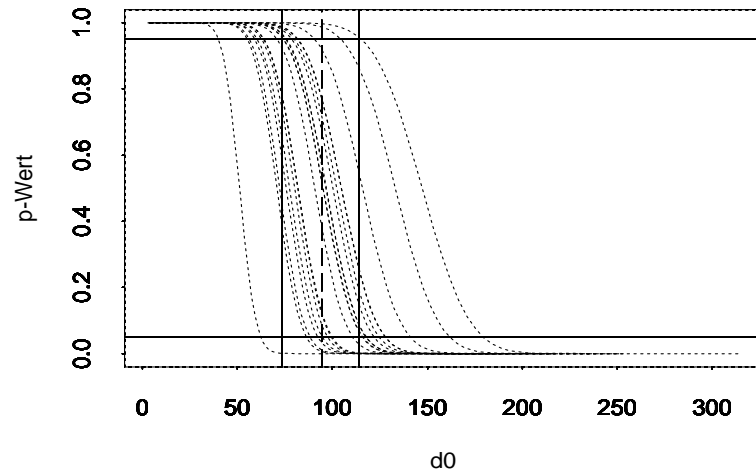
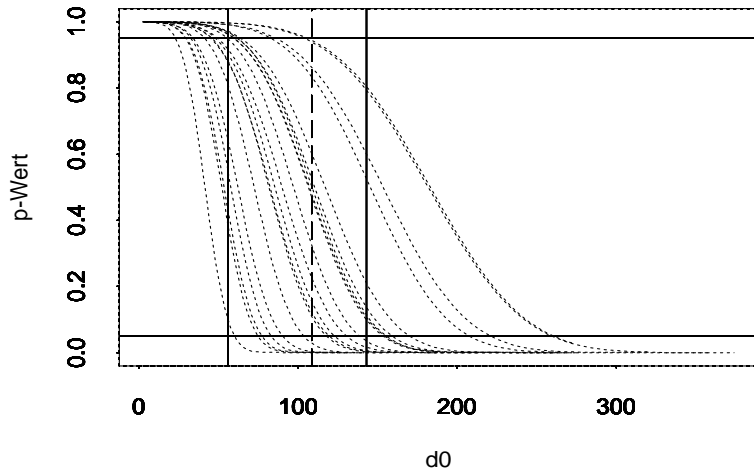
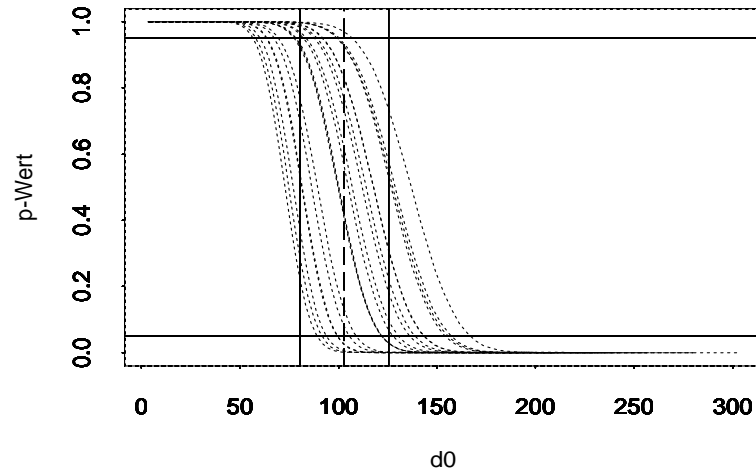
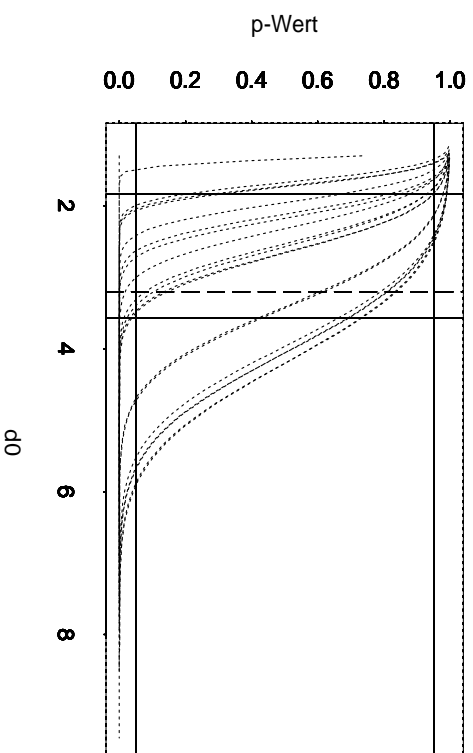


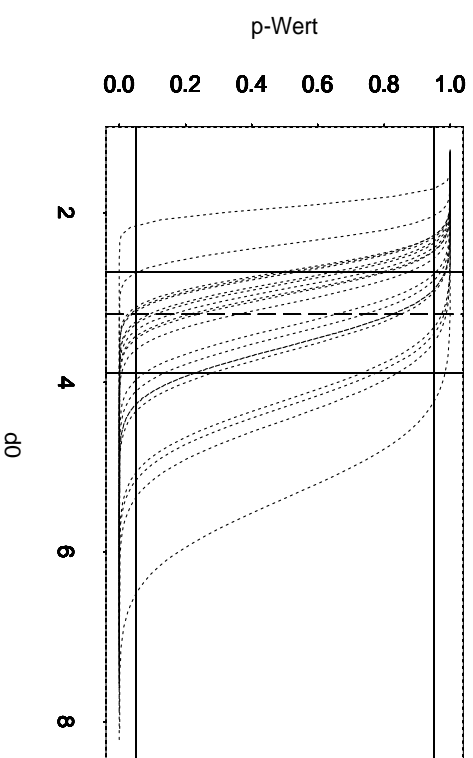
Abbildung 4.8: p-Wert-Kurven der Parametersätze 1-4 bezüglich d_0

5. Parametersatz, $d=94.5$, range μ klein, $n=25$ 6. Parametersatz, $d=94.5$, range μ klein, $n=100$ 7. Parametersatz, $d=108.55$, range μ gross, $n=25$ 8. Parametersatz, $d=102.8$, range μ gross, $n=100$ Abbildung 4.9: p-Wert-Kurven der Parametersätze 5-8 bezüglich d_0

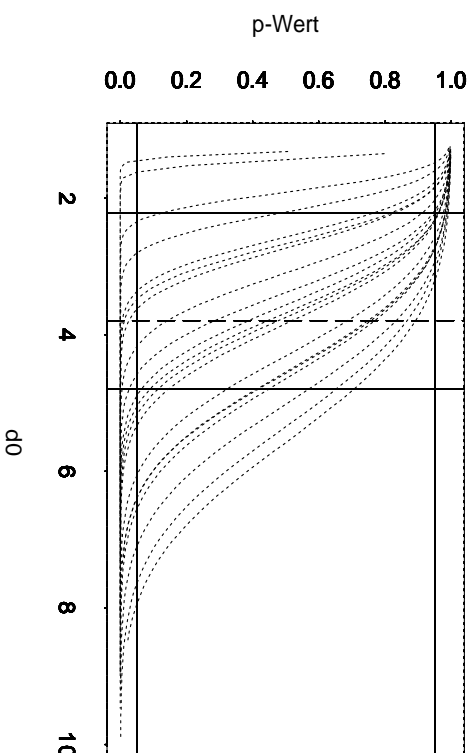
9. Parametersatz, $d=3.2$, range μ klein, $n=25$



10. Parametersatz, $d=3.2$, range μ klein, $n=100$



11. Parametersatz, $d=3.8$, range μ gross, $n=25$



12. Parametersatz, $d=3.6$, range μ gross, $n=100$

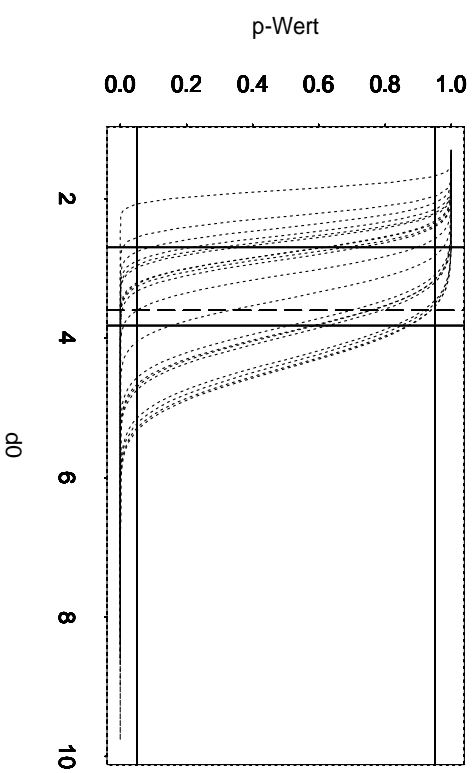
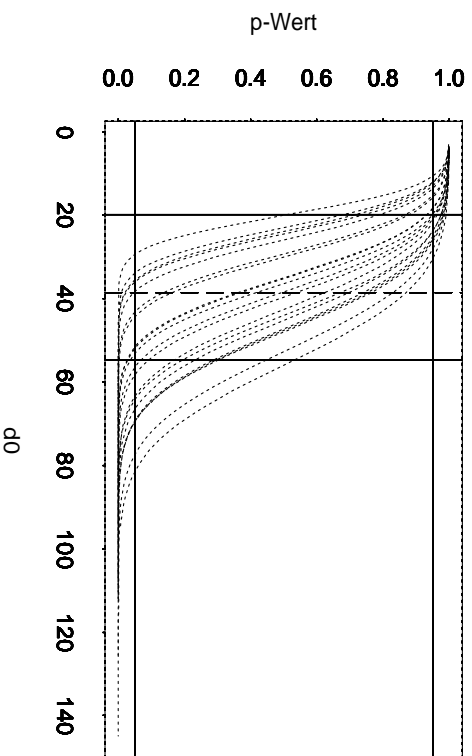
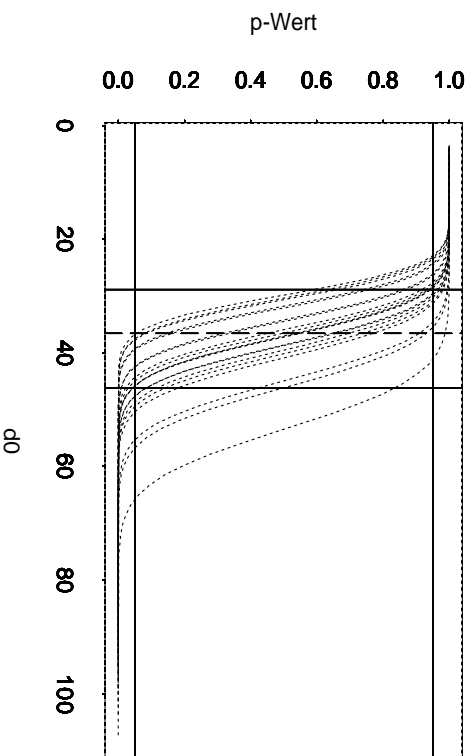
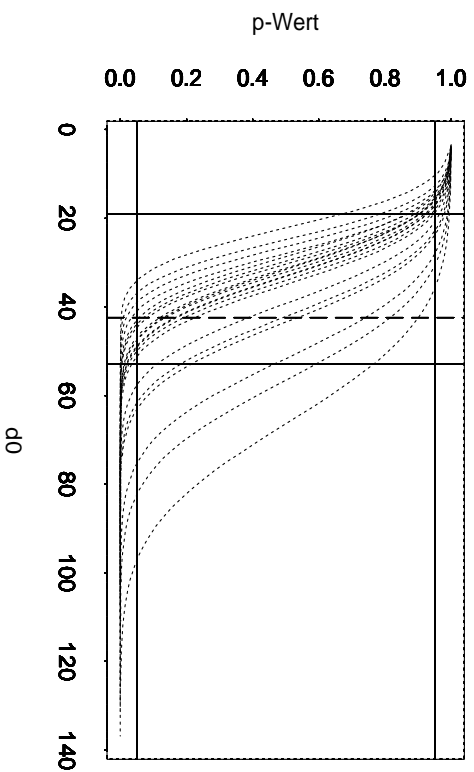
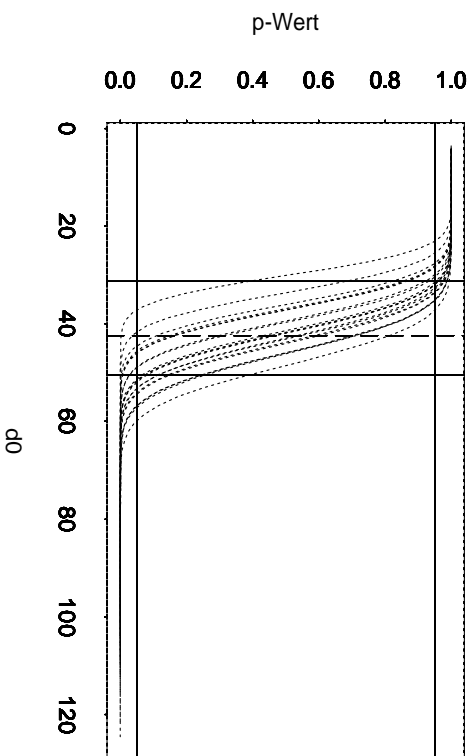


Abbildung 4.10: p-Wert-Kurven der Parametersätze 9-12 bezüglich d_0

13. Parametersatz, $d=38.5$, range μ klein, $n=25$ 14. Parametersatz, $d=36.6$, range μ klein, $n=100$ 15. Parametersatz, $d=42.4$, range μ gross, $n=25$ 16. Parametersatz, $d=42.4$, range μ gross, $n=100$ Abbildung 4.11: p-Wert-Kurven der Parametersätze 13-16 bezüglich d_0

						Diskriminanz zu Poisson		Akzeptanz zu Poisson	
PS	n	$SN(a, \bar{\mu})$	$\bar{\mu}$	Schwankungs- intervall μ_i	wahres d	\bar{d}_{0l}	$d - \bar{d}_{0l}$ in %	\bar{d}_{0u}	$\bar{d}_{0u} - d$ in %
1	25	1.25	3	[1.5 ; 4.5]	2.41	1.63	32%	2.63	9%
2	100	1.25	3	[1.5 ; 4.5]	2.33	2.06	12%	2.69	15%
3	25	1.25	3	[0.75 ; 5.25]	2.77	2.19	21%	4.06	47%
4	100	1.25	3	[0.75 ; 5.25]	2.99	2.43	19%	3.25	9%
5	25	1.25	100	[50 ; 150]	94.50	52.31	45%	132.49	40%
6	100	1.25	100	[50 ; 150]	94.50	73.28	22%	114.06	21%
7	25	1.25	100	[25 ; 175]	108.55	56.17	48%	142.89	32%
8	100	1.25	100	[25 ; 175]	102.79	80.45	22%	125.31	22%
9	25	2	10	[5 ; 15]	3.23	1.84	43%	3.57	11%
10	100	2	10	[5 ; 15]	3.23	2.70	16%	3.89	20%
11	25	2	10	[2.5 ; 17.5]	3.84	2.22	42%	4.80	25%
12	100	2	10	[2.5 ; 17.5]	3.61	2.70	25%	3.82	6%
13	25	2	100	[50 ; 150]	38.45	19.80	49%	54.64	42%
14	100	2	100	[50 ; 150]	36.62	28.79	21%	46.14	26%
15	25	2	100	[25 ; 175]	42.38	18.99	55%	52.97	25%
16	100	2	100	[25 ; 175]	42.38	31.30	26%	50.49	19%

Tabelle 4.3: Ergebnisse der p-Wert-Kurven bezüglich d_0 , d.h. $\hat{P}(\hat{a}, \hat{\sigma}, d_0) = \Phi\left(\frac{\hat{a} - d^{-1}(d_0)}{\hat{\sigma}(\hat{\beta}, \hat{a})}\right)$

Interpretation der p-Wert-Kurven bezüglich d_0

(i) Gruppe 1: Parametersätze 1 bis 4

- Einfluss von n :

Betrachtet man die Kurven bezüglich des Abstandsmasses d_0 , so ergeben sich ähnliche Schlussfolgerungen wie bei den Kurven bezüglich a_0 . Beim 1. PS liegen die Werte $\bar{d}_{0l} = 1.63$ und $\bar{d}_{0u} = 2.63$ (entspricht einer Differenz von 32% und 9% zum wahren Wert $d = 2.41$) ebenso weit von $d = 2.41$ entfernt wie beim 3. PS mit $\bar{d}_{0l} = 2.19$ und $\bar{d}_{0u} = 4.06$ (entspricht einer Differenz von 21% und 47% zu $d = 2.77$). Dagegen wird das Intervall $[\bar{d}_{0l}, \bar{d}_{0u}]$ sowohl im 2., als auch im 4. PS durch die Vergrößerung des Datenumfangs von $n = 25$

auf $n = 100$ stark verkleinert, denn für den 2. PS ergeben sich die Werte $\bar{d}_{0l} = 2.06$ und $\bar{d}_{0u} = 2.69$, was einer Abweichung von 12% bzw. 15% vom wahren Wert $d = 2.33$ entspricht. Für den 4. PS beobachtet man die Werte $\bar{d}_{0l} = 2.43$ und $\bar{d}_{0u} = 3.25$, woraus sich eine Differenz zu $d = 2.99$ von 19% und 9% errechnet. Eine kleine Unregelmässigkeit ist beim Übergang vom 1. PS zum 2. PS bei \bar{d}_{0u} einzuräumen, denn hier vergrössert sich der Abstand wie schon angesprochen von 9% im 1. auf 15% im 2. PS. Es ist aber an der Abbildung 4.8 deutlich zu erkennen, dass sich insgesamt die Breite des Intervalls $[\bar{d}_{0l}; \bar{d}_{0u}]$ verkleinert und die eben angesprochene Unregelmässigkeit eine Folge der leichten Asymmetrie im 1. PS ist. Es lässt sich also auch hier wieder zusammenfassen, dass mit einem grösseren Datenumfang die Kurven steiler verlaufen und somit das Intervall $[\bar{d}_{0l}, \bar{d}_{0u}]$ kleiner wird. Zu der Modellwahl kann hier gesagt werden, dass ein Poissonmodell gerade noch zu rechtfertigen ist, wenn man bereit ist, ein Abstandsmass von maximal $\bar{d}_{0u} = 4.06$ zu akzeptieren. Dies bedeutet wegen $d = \frac{\text{Var}(\text{NB-Verteilung})}{\text{Var}(\text{Poissonverteilung})}$, dass die Varianz des NB-Modells etwa 4 mal so gross ist wie die Varianz des Poissonmodells. Dieser Wert ist zwar schon sehr hoch, da in der Praxis maximal etwa eine 2-fache Abweichung in den Varianzen toleriert wird, doch könnte man im Grenzfall den Wert 4 noch zulassen, womit das Poissonmodell gerechtfertigt wäre.

- Einfluss der Schwankungsbreite (range) des Erwartungswertes:

Um den Einfluss der Schwankungsbreite zu erkennen, müssen die PSe 1 mit 3 und 2 mit 4 verglichen werden. Im Gegensatz zu den Kurven bezüglich a_0 ist hier ein deutlicher Unterschied auszumachen. Während im 1. PS $\bar{d}_{0l} = 1.63$ und $\bar{d}_{0u} = 2.63$ sind, liegen \bar{d}_{0l} und \bar{d}_{0u} im 3. PS bei $\bar{d}_{0l} = 2.19$ und $\bar{d}_{0u} = 4.06$. Eine solche Rechtsverschiebung der Kurven durch eine grössere Schwankungsbreite lässt sich auch beim Vergleich der PSe 2 mit 4 erkennen, denn es gilt $\bar{d}_{0l} = 2.06$ bzw. $\bar{d}_{0u} = 2.69$ im 2. und $\bar{d}_{0l} = 2.43$ bzw. $\bar{d}_{0u} = 3.25$ im 4. PS. Es kann also festgestellt werden, dass zwar keine grossen prozentualen Unterschiede der Intervalle $[\bar{d}_{0l}, \bar{d}_{0u}]$ auftreten, jedoch bewirkt die Vergrösserung der Schwankungsbreite des Erwartungswertes eine Gesamtverschiebung der Kurven tendenziell nach rechts, d.h. je grösser das Schwankungsintervall ist, desto grösser werden die Werte des Abstandsmasses. Dies hat dann zur Folge, dass die Rechtfertigung des Poissonmo-

dells immer schwieriger wird, denn je grösser die Werte des Abstandsmasses werden, desto mehr unterscheidet sich die Varianz des Poissonmodells von der des NB-Modells, was für einen Wechsel vom Poisson- zum NB-Modell spricht. Damit kann zusammengefasst werden, dass, je grösser das Schwankungsintervall des Erwartungswertes ist, desto eher muss das Poissonmodell zu Gunsten des NB-Modells abgelehnt werden.

(ii) **Gruppe 2: Parametersätze 5 bis 8:**

- Einfluss von n :

Anhand der Ergebnisse dieser PSe wird noch einmal bestätigt, dass der Datenumfang n einen relevanten Einfluss auf die Testergebnisse hat. Man kann an der Tabelle 4.3 ablesen, dass sich beim Vergleich des 5. und 6. PSe die prozentualen Abweichungen von \bar{d}_{0l} zu d und \bar{d}_{0u} zu d in etwa halbieren ($d - \bar{d}_{0l} = 45\%$ im 5. und $d - \bar{d}_{0l} = 22\%$ im 6. PS, ebenso $\bar{d}_{0u} - d = 40\%$ im 5. und $\bar{d}_{0u} - d = 21\%$ im 6. PS). Entsprechende Ergebnisse liegen bei den PSen 7 und 8 vor. Es kann also analog zu Gruppe 1 gefolgert werden, dass die Kurven steiler und die Intervalle $[\bar{d}_{0l}, \bar{d}_{0u}]$ sehr viel kleiner werden, sobald der Datenumfang n zunimmt. Ein wesentlicher Unterschied zu den Daten der Gruppe 1 ist, dass die reinen Werte sowohl von d , als auch von \bar{d}_{0l} und \bar{d}_{0u} sehr viel grösser sind als diejenigen der 1. Gruppe. Dies hängt allerdings mit der Wahl des relativ grossen Erwartungswertes $\bar{\mu} = 100$ zusammen (im Vergleich zu $\bar{\mu} = 3$ bei Gruppe 1), der in die Funktion des Abstandsmasses d nach (3.28) eingeht und somit die Werte d in der Gruppe 2 im Vergleich zu Gruppe 1 sehr gross werden lässt. Diese Tatsache hat wiederum Folgen für die Wahl des Modells. Aufgrund der extrem hohen Werte von \bar{d}_{0l} ist in allen 4 Fällen der PSe 5-8 ein Poissonmodell nicht vertretbar, da die Varianz des NB-Modells mindestens das 52-fache ($\bar{d}_{0l} = 52.31$ und grösser) der Varianz des Poissonmodells beträgt.

- Einfluss der Schwankungsbreite (range) des Erwartungswertes:

Der Einfluss der Schwankungsbreite ist in dieser Gruppe zwar andeutungsweise noch zu erkennen, jedoch längst nicht so deutlich wie in Gruppe 1. Da an den Kurven in Abbildung 4.9 optisch kaum Veränderungen sichtbar sind, ist es sinnvoller, die konkreten Werte aus der Tabelle 4.3 für die Analyse heranzuziehen. Aus ihnen geht hervor, dass mit einem grösseren Schwankungsintervall des Erwartungswertes (PSe 7 und 8) die Kurven leicht nach rechts

verschoben sind. So ergeben sich z.B. im 5. PS die Werte $\bar{d}_{0l} = 52.31$ und $\bar{d}_{0u} = 132.49$, während im 7. PS sich die Werte $\bar{d}_{0l} = 56.17$ und $\bar{d}_{0u} = 142.89$ herausstellen. Ebenso kann man beim Vergleich des 6. ($\bar{d}_{0l} = 73.28$ und $\bar{d}_{0u} = 114.06$) mit dem 8. PSes ($\bar{d}_{0l} = 80.45$ und $\bar{d}_{0u} = 125.31$) eine leichte Tendenz der Rechtsverschiebung erkennen, deren Konsequenz bezüglich der Modellwahl schon bei den PSen der Gruppe 1 angesprochen wurde.

(iii) Gruppe 3: Parametersätze 9 bis 12 und Gruppe 4: Parametersätze 13 bis 16:

Für die 3. und 4. Gruppe ergeben sich analoge Ergebnisse bezüglich des Einflusses des Datenumfangs n , sowie des Einflusses der Schwankungsbreite des Erwartungswertes wie in den Gruppen 1 und 2. Daher werden die einzelnen Resultate nicht mehr ausführlich beschrieben, sondern können in Tabelle 4.3 nachgelesen und analog zu Gruppe 1 und 2 interpretiert werden. Es ist allerdings anzumerken, dass in Gruppe 3 der PS 9 in gleicher Weise eine leichte Asymmetrie des Intervalls $[\bar{d}_{0l}; \bar{d}_{0u}]$ zu d aufweist, wie PS 1 in Gruppe 1. Daher ergibt sich auch hier eine kleine Unregelmässigkeit in Bezug auf den Effekt eines grösseren Datenumfangs n . Während in allen anderen Fällen die Abstände von \bar{d}_{0l} und \bar{d}_{0u} zu d bei grösserem n geringer werden, liegt hier eine Vergrösserung von $\bar{d}_{0u} = 3.57$ im 9. PS auf $\bar{d}_{0u} = 3.89$ im 10. PS vor. Es ist aber trotzdem eine Verkleinerung des Gesamtintervalls in Abbildung 4.10 erkennbar, so dass die bisherige Beobachtung der Intervallverkleinerung bei grossem Datenumfang n auch hier Gültigkeit hat. Weiterhin ist bei Gruppe 3 in Bezug auf die Modellwahl zu bemerken, dass wie in Gruppe 1 gerade noch ein Poissonmodell gerechtfertigt werden könnte, da maximal eine Abweichung von $\bar{d}_{0u} = 4.8$ akzeptiert werden müsste, um das Poissonmodell beibehalten zu können. Dagegen sind in Gruppe 4 die Werte d , \bar{d}_{0l} und \bar{d}_{0u} wie in Gruppe 2 aufgrund eines grossen Erwartungswertes $\bar{\mu} = 100$ viel zu hoch, als dass man ein Poissonmodell vertreten könnte. In allen 4 PSen der Gruppe 4 müsste also das NB-Modell gewählt werden.

(iv) Einfluss des Signal-to-noise-Quotienten (SN-Quotient):

Um den Einfluss des SN-Quotienten zu beurteilen, ist es ebenso wie bei den p-Wert-Kurven bezüglich a_0 schwierig, die Kurven in den Abbildungen 4.8, 4.9, 4.10 und 4.11 miteinander zu vergleichen, da sie alle unterschiedliche Massstäbe d_0 auf der x-Achse aufweisen. Daher werden hauptsächlich die Daten aus Tabelle 4.3 betrachtet. Im Gegensatz zu den Kurven bezüglich a_0 sind hier kleine

Unterschiede in den prozentualen Abweichungen von \bar{d}_{0l} und \bar{d}_{0u} zu d in den verschiedenen Gruppen vorhanden. Dies bedeutet aber, dass nicht der Wert des SN-Quotienten relevant ist, sondern die Wahl des gemittelten Erwartungswertes $\bar{\mu}$. Die prozentualen Abweichungen der 2. und 4. Gruppe, in denen ein $\bar{\mu} = 100$ vorliegt, entsprechen sich mit nur geringfügigen Unterschieden (ca. 43% mit $n = 25$ und ca. 20% mit $n = 100$). Dagegen sind die prozentualen Abweichungen in der Gruppe 1 mit $\bar{\mu} = 3$ gegenüber den Gruppen 3 (mit $\bar{\mu} = 10$), 2 und 4 (je $\bar{\mu} = 100$) am kleinsten. Es lässt sich damit vermuten, dass die Grösse von $\bar{\mu}$ die prozentualen Abweichungen leicht beeinflusst, d.h. je grösser $\bar{\mu}$ ist, desto grösser werden die Abweichungen \bar{d}_{0l} und \bar{d}_{0u} von d . Abgesehen von den prozentualen Abweichungen ist an den direkten Werten \bar{d}_{0l} und \bar{d}_{0u} deutlich der Einfluss von $\bar{\mu}$ zu erkennen. Da, wie schon erwähnt, der Erwartungswert in die Berechnung des Abstandsmasses d eingeht, spiegelt sich der grosse Wert $\bar{\mu} = 100$ in d , \bar{d}_{0l} und \bar{d}_{0u} in den Gruppen 2 und 4 wider. Daraus kann gefolgert werden, dass, je grösser der Erwartungswert ist, desto grösser wird das Abstandsmass d und desto eher wird das Poissonmodell aufgrund eines zu hohen Wertes \bar{d}_{0u} abgelehnt.

4.2.3 Zusammenfassung

Insgesamt lässt sich feststellen, dass in jeder Gruppe der Datensatzumfang n einen entscheidenden Einfluss auf die Testergebnisse hat, denn während die Abweichungen von \bar{a}_{0l} und \bar{a}_{0u} vom wahren Wert a bei PSen mit $n = 25$ im Schnitt bei 40% liegen, ergeben sich für PSe mit $n = 100$ prozentuale Differenzen von nur 25%. Diese Beobachtung gilt gleichermassen für alle Gruppen. Analog dazu kann auch bei den Kurven bezüglich d_0 ein relevanter Einfluss des Datenumfangs n in allen Gruppen festgestellt werden. Weiterhin ergibt sich, dass eine grössere Schwankungsbreite des Erwartungswertes bei den Kurven bezüglich a_0 keinen relevanten Einfluss hat, doch bei den Kurven bezüglich des Abstandsmasses d_0 eine tendenzielle Verschiebung nach rechts bewirkt. Aus der Rechtsverschiebung folgt dann eine Vergrösserung der relevanten Werte \bar{d}_{0u} , was zur Konsequenz hat, dass das Poissonmodell nicht akzeptiert werden kann und stattdessen das NB-Modell gewählt werden muss. Bemerkenswert ist noch, dass sich zwischen den Gruppen grosse Differenzen bezüglich des Abstandsmasses d ergeben. Während sich z.B. der Parameter a vom 1. zum 5. PS von $a = 0.3$ auf $a = 0.63$ nur verdoppelt, wird d um etwa das 50-fache vergrössert. Dies hängt aber mit dem sich ebenfalls veränderndem $\bar{\mu}$ zusammen, das sich von $\bar{\mu} = 3$ auf $\bar{\mu} = 100$ vergrössert, womit schliesslich

gewährleistet ist, dass der Signal-to-noise-Quotient in beiden Gruppen denselben Wert $SN = 1.25$ annimmt. Eine ähnliche Veränderung kann zwischen den Gruppen 3 und 4 festgestellt werden, bei denen sich das a von 0.15 auf 0.24 nicht ganz verdoppelt, aber das Abstandsmass d sich von im Schnitt 3.3 auf 36 mehr als verzehnfacht. Es gilt also zusammenfassend, dass weder die Kurven bezüglich a_0 noch die bezüglich d_0 direkt vom Wert des SN-Quotienten beeinflusst werden, aber die Kurven bezüglich des Abstandsmasses d_0 von der Wahl des gemittelten Erwartungswertes $\bar{\mu}$ abhängen.

Zu der Frage, wann der Aufwand eines Wechsels vom Poisson- zum NB-Modell gerechtfertigt ist, lässt sich sagen, dass man bei der Betrachtung der p-Wert-Kurven bezüglich a_0 Schwierigkeiten hat, die relevanten Werte \bar{a}_{0u} für die Akzeptanz des Poissonmodells zu beurteilen. Generell sind solche Kurven angebracht, die sich für die PSe mit $n = 100$ ergeben, bei denen die Werte \bar{a}_{0l} und \bar{a}_{0u} nahe aneinander liegen. Denn wünschenswert sind solche Kurven, deren Bereich zwischen \bar{a}_{0l} und \bar{a}_{0u} , in dem keine Aussagen über die Modellwahl getroffen werden können, so klein wie möglich ist. Es muss dann in jedem Fall individuell entschieden werden, ob der Wert \bar{a}_{0u} für das Poissonmodell noch akzeptiert werden kann. Wegen der grossen Interpretationsschwierigkeit des Dispersionsparameters a ist es sinnvoller, die Kurven bezüglich des Abstandsmasses d zu betrachten. Hier ergibt sich in den Gruppen 1 und 3 ein noch akzeptabler durchschnittlicher Wert von $\bar{d}_{0u} = 4$, so dass man für alle PSe dieser Gruppen ein Poissonmodell tolerieren kann. Bei den Gruppen 2 und 4 dagegen ergibt sich für alle PSe ein so hoher Wert \bar{d}_{0l} , dass nur ein NB-Modell vertretbar ist, denn ein hoher Wert \bar{d}_{0l} ist gleichbedeutend damit, dass sich die Varianz des NB-Modells stark von der Varianz des Poissonmodells unterscheidet. Mit Abschnitt 3.3 Abbildung 3.5 kann dann gefolgert werden, dass bis zu einer maximalen Akzeptanzgrenze $\bar{a}_{0l} = d^{-1}(\bar{d}_{0l})$ die Hypothese $K : a \leq a_0$ gegen $H : a > a_0$ verworfen wird, womit man annimmt, dass $a > a_0$ gilt, und somit das NB-Modell zu wählen ist.

Abschliessend ist noch zu bemerken, dass speziell beim 3. PS extreme Konvergenzschwierigkeiten bei der Berechnung des Schätzers für a auftraten, jedoch konnten stückweise 20 Zufallsvektoren gefunden werden, für die die Berechnung aller Schätzer innerhalb der vorgegebenen Iterationsgrenze konvergierte. Für den 3. PS wurden also nicht 20 Zufallsvektoren auf einmal berechnet, wie im Programm „a0quer“ angegeben, sondern es wurden jeweils einzeln Zufallsvektoren erzeugt, die Schätzungen durchgeführt, und wenn die Schätzung speziell für den Parameter a konvergierte, so wurde dieser Zufallsvektor für die weiteren Berechnungen verwendet. Dieses Verfahren wurde solange durchgeführt, bis 20 verwendbare Zufallsvektoren gegeben waren.

Kapitel 5

Anwendungsbeispiele

In diesem letzten Kapitel wird schliesslich der Bereichstest, der im 3. Kapitel hergeleitet und im 4. Kapitel anhand einer Simulationsstudie getestet wurde, auf 2 konkrete Datenbeispiele angewendet. Dazu werden einerseits Daten aus dem Bereich der Kfz-Versicherung und andererseits Daten aus dem Bereich der wirtschaftswissenschaftlichen Forschung untersucht. Zunächst wird mittels einer explorativen Datenanalyse ein Poissonmodell angepasst und anschliessend gezeigt, dass die Daten Überdispersion aufweisen. Zuletzt wird dann das NB-Modell auf die Daten angewendet und es wird damit eine p-Wert-Kurve erstellt, wie sie im vorangegangenen Kapitel an der Simulationsstudie beschrieben wurde. Das Ziel ist letztlich, anhand der p-Wert-Kurve die Überdispersion zu quantifizieren und zu beurteilen, welches Modell zur Datenanpassung gewählt werden sollte.

5.1 Schwedische Kfz-Versicherung

Die zu untersuchenden Daten sind auf der Internetseite

[http : //www.statsci.org/data/general/motorins.txt](http://www.statsci.org/data/general/motorins.txt)

erhältlich und entstammen einer schwedischen Kfz-Haftpflicht-Versicherung aus dem Jahr 1977. Da in Schweden alle Kfz-Versicherungen dieselben Grössen zur Klassifizierung ihrer Kunden verwenden, können ihre Portfolios und ihre Schadensstatistiken miteinander kombiniert werden. Die zugrundeliegenden Daten wurden durch ein schwedisches Komitee gebildet, deren Aufgabe es war, die Einflussgrössen auf die Schadenanzahl zu identifizieren und diese mit der Struktur des damals aktuellen Tarifs zu vergleichen.

5.1.1 Datenbeschreibung

Der Datensatz besteht aus den 6 Variablen Kilometer, Zone, Bonus, Automarke, Anzahl der Versicherten und Schadenanzahl. Die Grösse „*Kilometer*“ beschreibt, wieviele Kilometer der Versicherte im Durchschnitt jährlich zurücklegt und ist in 5 Kategorien eingeteilt, die sich wie folgt ergeben:

- 1 : < 1.000 km
- 2 : 1.000 km – 15.000 km
- 3 : 15.000 km – 20.000 km
- 4 : 20.000 km – 25.000 km
- 5 : > 25.000 km

Die Variable „*Zone*“ hat 7 Ausprägungen, die den geographischen Aufenthaltsort des Versicherten angeben:

- 1: Stockholm, Göteborg, Malmö mit Umland
- 2: Weitere Grossstädte mit Umland
- 3: Kleinere Städte mit Umland im Süden Schwedens
- 4: Ländliche Gebiete im Süden Schwedens
- 5: Kleinere Städte mit Umland im Norden Schwedens
- 6: Ländliche Gebiete im Norden Schwedens
- 7: Gotland

Die Variable „*Bonus*“ enthält 7 Kategorien und entspricht der Anzahl der Versicherungsjahre plus 1 seit dem letzten Schaden. Schliesslich gibt die Grösse „*Automarke*“ in 9 verschiedenen Kategorien die untersuchten Automarken an, wobei in den Klassen 1-8 die gängigsten 8 Automodelle enthalten sind, während alle anderen Automarken in der Klasse 9 zusammengefasst sind. Es bleibt die Variable „*Versicherte*“, die die Anzahl der Versicherten in einem Versicherungsjahr pro Merkmalskombination angibt. Diese Grösse wird als „*offset*“ verwendet werden, auf den später noch genauer eingegangen wird. Den Response bildet die „*Schadenanzahl*“, die nun mittels der beschriebenen Kovariablen erklärt werden soll. Die Daten liegen in aggregierter Form vor, d.h. alle Kovariablenkombinationen treten höchstens einmal auf, woraus sich eine Länge des Datensatzes von 2182 ergibt. Damit wird die Schadenanzahl nicht bezüglich einzelner Individuen, sondern bezüglich einer Gruppe von Versicherungsnehmern mit identischen Merkmalen modelliert. Es kann allerdings vorkommen, dass die Grösse der einzelnen

Gruppen sehr unterschiedlich ist, weshalb man die Anzahl der Versicherten innerhalb einer Gruppe als Gewichtung verwendet, um die Gruppen der verschiedenen Merkmalskombinationen miteinander vergleichbar zu machen. Bevor nun dieser Datensatz näher untersucht wird, soll zunächst seine äussere Struktur betrachtet werden, um ihn in eine Gruppe der Parametersätze aus der Simulationsstudie einordnen zu können. Aus der Spalte „*Schadenanzahl*“ der Kfz-Daten geht hervor, dass die Schadenanzahlen zwischen 0 und Werten über 3000 variieren. Die durchschnittliche Schadenanzahl beträgt in etwa 52, so dass man von einem gemittelten Erwartungswert $\bar{\mu} = 52$ ausgehen kann. Zusammen mit dem im nachfolgend erläuterten Modell (5.2) geschätzten Dispersionsparameter a ergibt sich weiter für diese Daten ein Signal-to-noise-Quotient von 5.95. Wegen des relativ grossen SN-Quotienten sollte man den Vergleich mit den Parametersätzen auf die Gruppen 3 und 4 beschränken, die einen SN-Quotienten von 2 aufweisen. Die Datenlänge ist zwar mit $n = 2182$ sehr viel höher als die maximale Länge $n = 100$ der Parametersätze der Simulationsstudie, trotzdem ist dieses Beispiel am besten mit denjenigen Parametersätzen vergleichbar, deren Datenlänge $n = 100$ ist, deren mittlerer Erwartungswert $\bar{\mu} = 100$ beträgt (Gruppe 4) und die eine grosse Schwankungsbreite des Erwartungswertes aufweisen (Parametersatz 16). Bei der Betrachtung der Powerfunktionen in Abschnitt 4.1 ergab sich für diesen Parametersatz ein wenig liberaler Test, so dass auch hier (vor allem wegen der extrem grossen Datenlänge) bei der Durchführung des Bereichstests ein α -level-Test zu erwarten ist.

5.1.2 Explorative Datenanalyse der Kfz-Daten

Mit Hilfe der explorativen Datenanalyse soll nun versucht werden, einen ersten Überblick über die Daten zu erhalten und ein mögliches Modell zu erstellen. Dazu werden zunächst die Kovariablen jeweils gegen die zu untersuchende Grösse „*Schadenanzahl*“, die im folgenden mit S abgekürzt wird, abgetragen. Allerdings ist zu beachten, dass aus bereits angesprochenen Gründen die Schadenanzahl noch mit der Anzahl der Versicherten (V) gewichtet wird, woraus sich die Grösse der Schadenhäufigkeit

$$SH = \frac{\text{Schadenanzahl}}{\text{Anzahl Versicherte}} = \frac{S}{V}$$

ergibt. Es wird also zunächst ein Poissonmodell mit der log-link-Funktion an die Daten angepasst, wobei die Variable „*Versicherte*“ als Gewichtung fungiert. Nach Abschnitt 2.2.5 gilt dann:

$S_i = \text{Schadenanzahl} \sim \text{Poi}(\mu_i)$, $i = 1, \dots, 2128$ mit

$$\mu_i = V_i \exp(\underbrace{\mathbf{x}_i^t \boldsymbol{\beta}}_{\text{offset}}) = \exp(\ln(V_i) + \mathbf{x}_i^t \boldsymbol{\beta}) \quad \text{bzw.} \quad \ln(\mu_i) = \ln(V_i) + \mathbf{x}_i^t \boldsymbol{\beta}.$$

Löst man diese Gleichung nach $\mathbf{x}_i^t \boldsymbol{\beta}$ auf, so erhält man $\ln\left(\frac{\mu_i}{V_i}\right) = \mathbf{x}_i^t \boldsymbol{\beta}$, d.h. dass sich die logarithmierte Schadenhäufigkeit linear zu den Kovariablen verhält, weshalb in den folgenden Abbildungen 5.1 auf der x-Achse die einzelnen Kovariablen und auf der y-Achse die logarithmierte Schadenhäufigkeit $\ln SH = \ln \frac{S}{V}$ abgetragen sind. An der Gra-

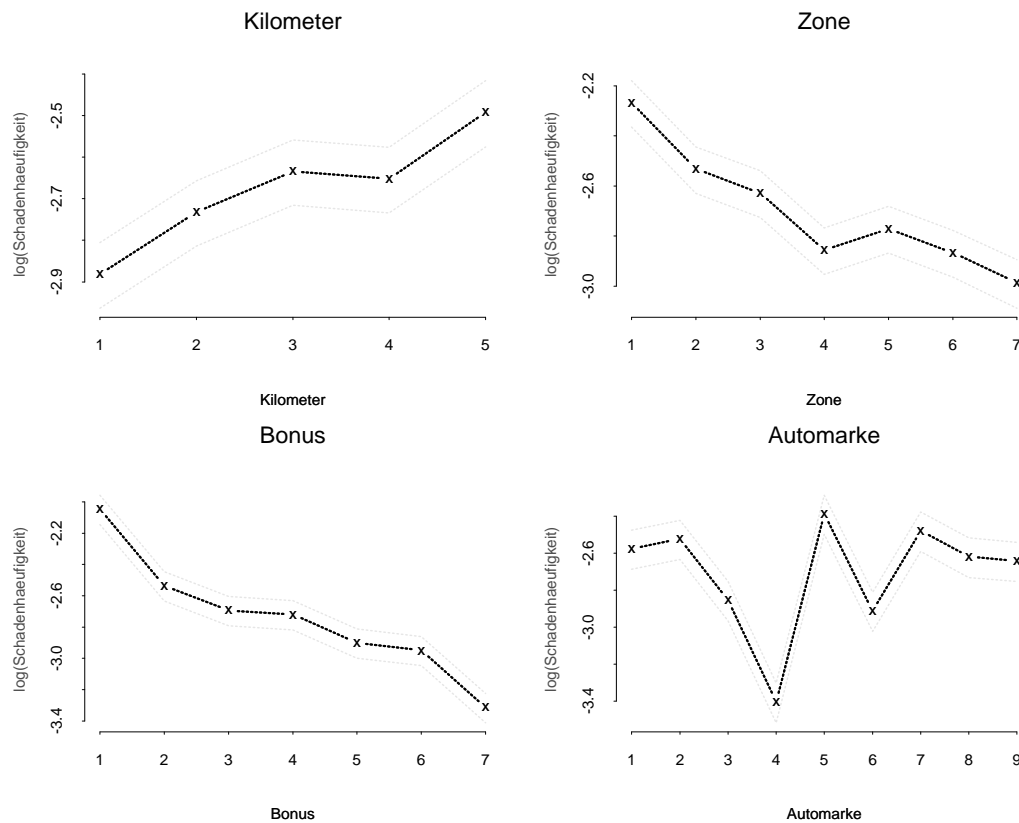


Abbildung 5.1: Explorative Datenanalyse der Kfz-Daten

phik für die Variable „*Kilometer*“ ist deutlich zu erkennen, dass Versicherungsnehmergruppen, die jährlich grosse Kilometerstrecken zurücklegen (Kategorie 5), eine höhere durchschnittliche Schadenhäufigkeit aufweisen als diejenigen Gruppen, die nur wenige Kilometer jährlich fahren (Kategorien 1 und 2). Ähnlich dazu kann an der Graphik für die Variable „*Zone*“ festgestellt werden, dass in Grossstädten (Kategorie 1 und 2) eine höhere Schadenhäufigkeit vorliegt als in Kleinstädten (Kategorien 3 und 5) bzw. ländlichen Gebieten (Kategorien 4 und 6). Anhand der Graphik für die Grösse „*Bonus*“ ist

deutlich der negative lineare Zusammenhang zwischen schadenfreien Jahren und Schadenhäufigkeit zu erkennen, der der Struktur der Beitragsermässigung entspricht, d.h. je mehr schadenfreie Jahre der Versicherungsnehmer vorzuweisen hat, desto geringer ist die Schadenhäufigkeit und umso höher fällt für ihn der Bonus aus. In der Abbildung für die Variable „*Automarke*“ fällt besonders die Kategorie 4 durch eine extrem kleine Schadenhäufigkeit im Vergleich zu den anderen Kategorien auf. Es ist bekannt, dass die Klasse 4 die Automarke VW 1200 enthält, da deren Produktion kurz nach 1977 eingestellt wurde. Die anderen Kategorien wurden nicht identifiziert, da sie sonst die Verkaufsstatistiken der untersuchten Automarken beeinflusst hätten. Ansonsten lässt sich dieser Graphik entnehmen, dass offensichtlich die Automarke einen wesentlichen Einfluss auf die Schadenhäufigkeit hat, denn abgesehen von der Kategorie 4 unterscheiden sich die Werte der Schadenhäufigkeiten aus beispielsweise den Kategorien 2 und 6 relevant.

5.1.3 Poissonregression der Kfz-Daten

Da die vier betrachteten Kovariablen „*Kilometer*“, „*Zone*“, „*Bonus*“ und „*Automarke*“ offensichtlich alle einen Einfluss auf die Zielvariable Schadenanzahl S ausüben, werden sie alle in das zu bildende Poissonmodell aufgenommen. Aus der Abbildung 5.1 geht hervor, dass die Variablen „*Kilometer*“, „*Zone*“ und „*Bonus*“ einen relativ linearen Verlauf zeigen, wodurch die Möglichkeit besteht, diese nicht als kategorielle Variablen, sondern als metrisch zu betrachten. Desweiteren können Interaktionen vorhanden sein, d.h. Zusammenhänge zwischen den verschiedenen Kovariablen. Das Modell, das wie eben beschrieben, alle Variablen (bis auf „*Automarke*“) metrisch verwendet, und auch Interaktionen berücksichtigt, liefert allerdings eine extrem schlechte Anpassung, weshalb hier auf ein Poissonmodell, das alle Kovariablen als kategorielle Variablen enthält, zurückgegriffen wird. Es werden auch weiter keine Interaktionen betrachtet, da sie zu einem hohen Verlust an Freiheitsgraden führen, der in keiner Relation zur verbesserten Anpassung durch die Verwendung von Interaktionen steht. Damit ergibt sich schliesslich das folgende Modell, das in Splus berechnet wird:

$$\begin{aligned} &\text{insurancepoi_glm}(S \sim \text{offset}(\log(\text{Versicherte})) + \text{kilometerf} + \text{zonef} + \\ &\text{bonusf} + \text{automarkef}, \text{family} = \text{poisson}, \text{link} = \log, \text{x} = \text{T}), \end{aligned} \quad (5.1)$$

wobei zu erwähnen ist, dass die Kovariablen *kilometerf*, *zonef*, *bonusf* und *automarkef* die vorher beschriebenen Variablen darstellen, allerdings als Faktor, d.h. als kategorielle

Coefficients:

	Value	Std. Error	t value	Pr(> t)
(Intercept)	-1.812840	0.013755	-131.795	< 2e-16 ***
kilometerf2	0.212586	0.007523	28.259	< 2e-16 ***
kilometerf3	0.320226	0.008660	36.979	< 2e-16 ***
kilometerf4	0.404657	0.012050	33.581	< 2e-16 ***
kilometerf5	0.575955	0.012826	44.906	< 2e-16 ***
zonef2	-0.238168	0.009494	-25.085	< 2e-16 ***
zonef3	-0.386395	0.009669	-39.964	< 2e-16 ***
zonef4	-0.581902	0.008653	-67.252	< 2e-16 ***
zonef5	-0.326128	0.014526	-22.451	< 2e-16 ***
zonef6	-0.526234	0.011874	-44.317	< 2e-16 ***
zonef7	-0.730935	0.040474	-18.059	< 2e-16 ***
bonusf2	-0.478993	0.012092	-39.614	< 2e-16 ***
bonusf3	-0.693172	0.013505	-51.325	< 2e-16 ***
bonusf4	-0.827397	0.014580	-56.748	< 2e-16 ***
bonusf5	-0.925632	0.013965	-66.284	< 2e-16 ***
bonusf6	-0.993457	0.011627	-85.445	< 2e-16 ***
bonusf7	-1.327406	0.008683	-152.867	< 2e-16 ***
automarkef2	0.076245	0.021230	3.591	0.000329 ***
automarkef3	-0.247407	0.025056	-9.874	< 2e-16 ***
automarkef4	-0.653523	0.024171	-27.037	< 2e-16 ***
automarkef5	0.154924	0.020225	7.660	1.86e-14 ***
automarkef6	-0.335581	0.017371	-19.318	< 2e-16 ***
automarkef7	-0.055939	0.023325	-2.398	0.016473 *
automarkef8	-0.043921	0.031533	-1.393	0.163666
automarkef9	-0.068054	0.009954	-6.837	8.11e-12 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 34070.6 on 2181 degrees of freedom

Residual deviance: 2966.1 on 2157 degrees of freedom

Tabelle 5.1: *summary*-Tabelle des Poissonmodells (5.1) der Kfz-Daten

Variable kodiert. Lässt man sich dann das „*summary*“ dieses Modells ausgeben, so erhält man die in Tabelle 5.1 unter der Spalte „*value*“ angegebenen Ergebnisse für die Schätzungen der einzelnen Regressionsparameter. Aus dieser Tabelle geht hervor, dass alle durch die Faktorisierung entstandenen 25 Kovariablen hoch signifikant sind, da sie (bis auf die vorletzte Kategorie 8 in Automarke) einen extrem kleinen p-Wert aufweisen. Dies bedeutet, dass für jedes $j = 1, \dots, 25$ der Test $H_j : \beta_j = 0$ gegen $K_j : \beta_j \neq 0$ die Hypothese zu einem Signifikanzniveau von $\alpha = 0.001$ und kleiner verworfen werden kann, womit die statistische Signifikanz der zu β_j gehörenden Kovariable bewiesen ist. Aus der Tabelle kann ebenfalls abgelesen werden, dass die Devianz mit 2966,1 grösser als die Anzahl der Freiheitsgrade (2157) ist, was ein Zeichen für Überdispersion sein könnte. Bevor man aber davon ausgehen kann, dass Überdispersion vorliegt, müssen zunächst noch die Residuen und die Linkspezifikation überprüft werden, um auszuschliessen, dass sie die mangelnde Anpassung verursachen. Es wurde bereits in Abschnitt 2.2.9 angesprochen, dass es für Zählraten kein Residuum gibt, das den gewünschten Eigenschaften wie Erwartungswert 0, konstante Varianz und symmetrische Verteilung um den Erwartungswert gleichzeitig entspricht. In McCullagh & Nelder (1989, S. 39) [15] wird daher für die Residuenanalyse die Betrachtung der Devianzresiduen empfohlen, da sie am besten mit standard-normalverteilten Residuen vergleichbar sind, die die obigen Eigenschaften erfüllen. Für die Residuenanalyse sind in der Abbildung 5.2 vier verschiedene Graphiken dargestellt. Die linke obere Graphik zeigt die Devianzresiduen r_i^D (siehe Abschnitt 2.2.9) für alle Merkmalskombinationen $i = 1, \dots, 2182$. Es wird daran deutlich, dass die Residuen wie gewünscht zufällig um 0 streuen. In der oberen rechten Graphik sind die Devianzresiduen gegen den geschätzten Erwartungswert der Schadenanzahl abgetragen. Hieran lässt sich ebenfalls erkennen, dass die Residuen zufällig um 0 streuen, ohne eine erkennbare Form aufzuweisen, die die Abhängigkeit der Residuen von der Schadenanzahl bedeuten würde. Aus beiden oberen Graphiken geht weiter hervor, dass alle Schadenanzahlen mit einer betragsmässig maximalen Abweichung von 6 geschätzt werden, doch ist auffällig, dass die meisten Residuen in einem Band von -2 bis 2 enthalten sind (siehe linke obere Graphik). Angesichts einer Schadenanzahl bis zu über 3000 sprechen diese kleinen Residuenwerte für eine relativ gute Anpassung. Da sich die Graphiken der einzelnen Kovariablen gegen die Residuen kaum unterscheiden, ist hier in der unteren linken Graphik stellvertretend für alle anderen Variablen die Graphik der Devianzresiduen gegen die Kovariable „*Kilometer*“ abgebildet. Aus ihr kann abgelesen werden, dass wie schon angesprochen, die Residuen einen maximalen Wert von 6 annehmen. Die letzte Graphik zeigt einen

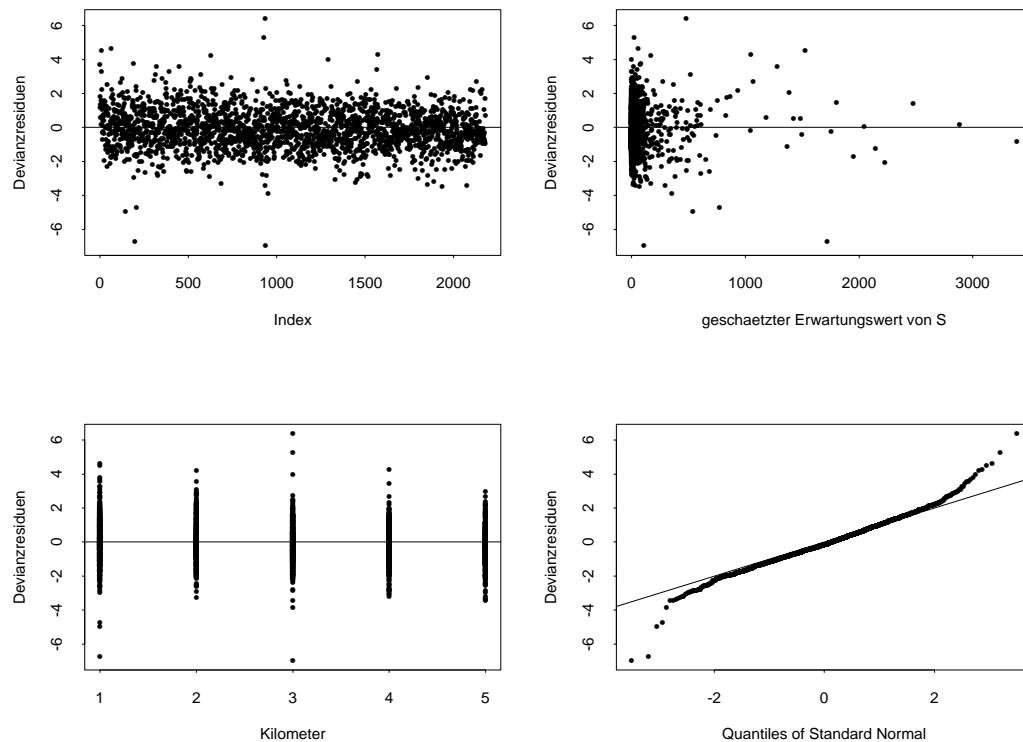


Abbildung 5.2: Devianzresiduen des Poissonmodells (5.1) der Kfz-Daten

sogenannten Q-Q-Plot, der angibt, inwiefern sich die Residuen nach der erwünschten Normalverteilung verhalten. Liegen die Residuen vollständig auf der als durchgezogene Linie eingezeichneten Winkelhalbierenden, so liegt eine Normalverteilung der Residuen vor. Die hier eingetragenen Devianzresiduen weisen für mittlere Werte annähernd eine Normalverteilung auf, während für sehr kleine (< -3) und grosse Residuen (> 3) eine starke Abweichung von der Winkelhalbierenden erkennbar ist. Nach Fahrmeir & Künstler (1999, S. 94) [6] kann daraus gefolgert werden, dass die Residuen zwar eine symmetrische Verteilung haben, die aber im Vergleich zur Normalverteilung dickere Enden links und rechts aufweist. Insgesamt lässt sich aus der Residuenanalyse schließen, dass die Gesamtanpassung des Poissonmodells zwar nicht optimal ist, aber auch nicht so schlecht, als dass sie nicht akzeptiert werden könnte. Als nächstes wird nun die Wahl der Linkfunktion untersucht. Dazu ist in Abbildung 5.3 in der oberen Graphik die Schadenanzahl gegen den Schätzer des linearen Prädiktors abgetragen. Da für das Poissonmodell die log-link-Funktion gewählt wurde (2.9), bedeutet dies, dass die Schadenanzahl nun zufällig um die Exponentialfunktion streuen sollte, was augenscheinlich

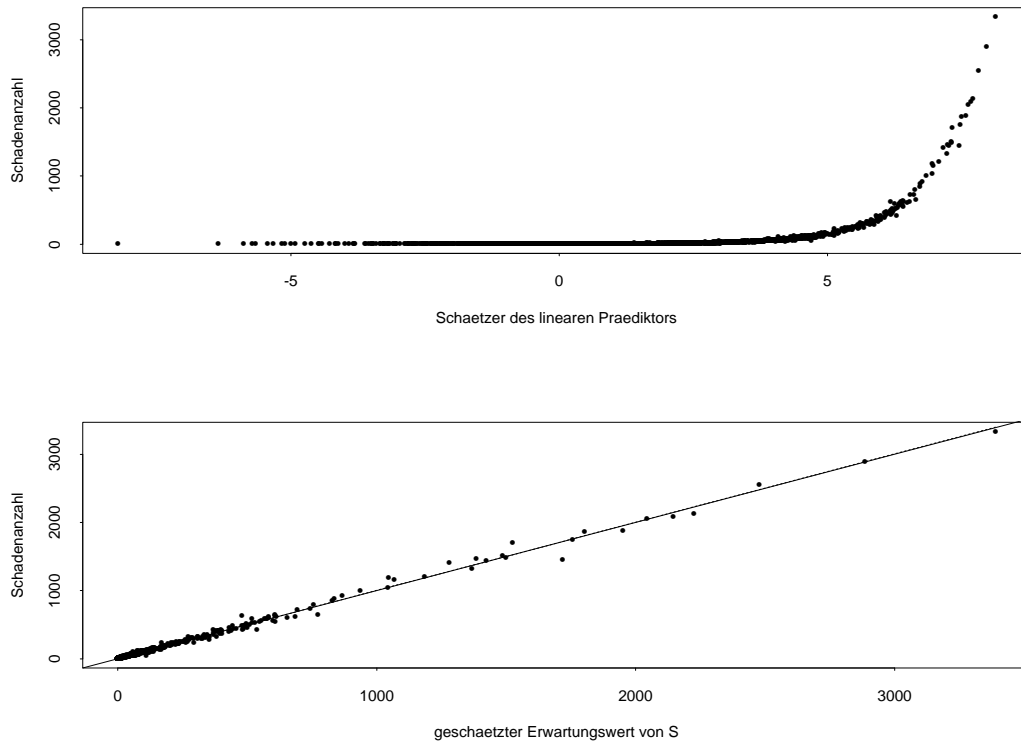


Abbildung 5.3: Linkspezifikation im Poissonmodell der Kfz-Daten

der Fall ist. Allerdings lässt sich generell ein linearer Zusammenhang graphisch leichter erkennen, weshalb in der unteren Graphik die Schadenanzahlen gegen den geschätzten Erwartungswert der Schadenanzahl abgetragen sind, d.h. statt des geschätzten linearen Prädiktors $\mathbf{x}_i^t \hat{\boldsymbol{\beta}}$ werden die exponentiellen Werte $\exp(\mathbf{x}_i^t \hat{\boldsymbol{\beta}})$ verwendet, die gerade den geschätzten Erwartungswert ergeben. Da offensichtlich alle Punkte im Rahmen der statistischen Ungenauigkeit auf der Winkelhalbierenden liegen (durchgezogene Linie), ist nachgewiesen, dass der geschätzte Erwartungswert den Schadenanzahlen entspricht, woraus gefolgert werden kann, dass die Spezifizierung der Linkfunktion durch den Logarithmus geeignet ist. Zuletzt wird nun versucht, ein mögliches Vorhandensein von Überdispersion graphisch darzustellen. Dazu werden in der folgenden Abbildung 5.4 die geschätzten Erwartungswerte $\exp(\mathbf{x}_i^t \hat{\boldsymbol{\beta}})$ auf der x-Achse gegen die geschätzten Varianzen auf der y-Achse abgetragen. Als Schätzer der Varianzen werden hier die quadrierten Rohresiduen $(y - \hat{\mu}_i)^2$ verwendet. Da im Poissonmodell nach Voraussetzung die Gleichung $E(Y_i) = \text{Var}(Y_i)$, $i = 1, \dots, n$ gilt, sollten alle Werte in der Graphik auf der Winkelhalbierenden (durchgezogene Linie) liegen. Allerdings ist der Graphik eine Ver-

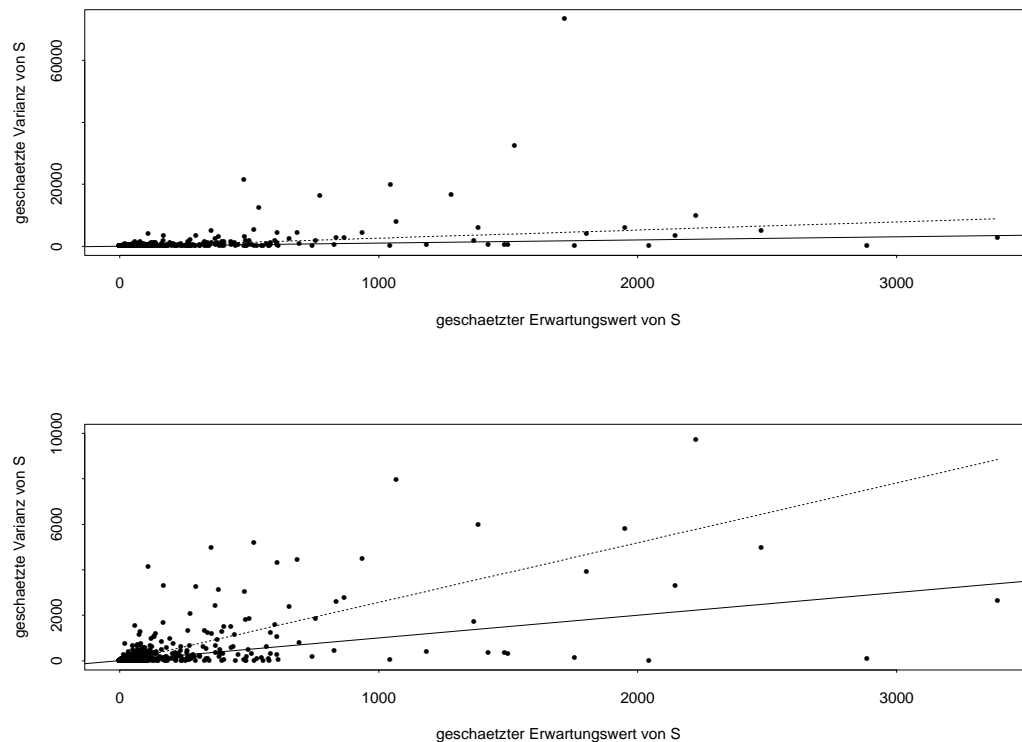


Abbildung 5.4: Überdispersion im Poissonmodell der Kfz-Daten

letzung der Modellannahme zu entnehmen, denn die mit dem Befehl „*lowess*“ in Splup erzeugte Glättungsfunktion der einzelnen Werte (hier mit einer gestrichelten Linie gekennzeichnet) liegt deutlich oberhalb der Winkelhalbierenden. Dies bedeutet, dass im Mittel die Varianzen der einzelnen Beobachtungsgruppen die geschätzten Erwartungswerte übersteigen und somit ein Hinweis für Überdispersion vorliegt. Diese Tatsache wird noch einmal in der unteren Graphik verdeutlicht, bei der alle Beobachtungsgruppen, deren geschätzte Varianzen 10.000 übersteigen, vernachlässigt werden. Hierdurch wird eine Vergrößerung des Massstabes im Bereich 0-10.000 erreicht. Aufgrund der Tatsachen, dass im „*summary*“ eine Differenz zwischen Devianz und Anzahl der Freiheitsgrade auftrat, was für eine nicht optimale Anpassung spricht, das Verhalten der Residuen nicht ganz optimal war und nun in der Abbildung 5.4 der Varianzfunktion ein Hinweis auf Überdispersion zu erkennen war, werden im folgenden die zu untersuchenden Daten mittels der Negativ-Binomial-Regression modelliert.

5.1.4 NB-Regression und p-Wert-Kurven der Kfz-Daten

Um den Vergleich zum Poissonmodell zu gewährleisten, wird nun dasselbe Modell wie bei der Poissonregression mit allen Haupteffekten ohne Interaktionen mit der NB-Verteilung modelliert. Damit ergibt sich in Splus das Modell

```
insurancenb_glm.nb(S~offset(log(Versicherte))+kilometerf+zonef+
bonusf+automarkef,link=log,x=T). (5.2)
```

Im dazugehörigen „*summary*“ (Tabelle 5.2) sind dann die geschätzten Werte der Regressionsparameter β_j , $j = 1, \dots, 25$ unter der Spalte „*value*“ eingetragen. Die statistische Signifikanz aller Kovariablen bleibt wie im Poissonmodell erhalten, da nach wie vor alle p-Werte < 0.05 sind (bis auf Automarke 8). Die Devianz ist nun von 2966 im Poissonmodell auf 2231 gesunken, woraus geschlossen werden kann, dass das NB-Modell eine bessere Anpassung liefert als bisher mit dem Poissonmodell erreicht werden konnte. Dies bestätigt sich auch bei der Betrachtung der Residuen für das NB-Modell. Es werden wieder dieselben Graphiken dargestellt, wie sie bei der Erstellung des

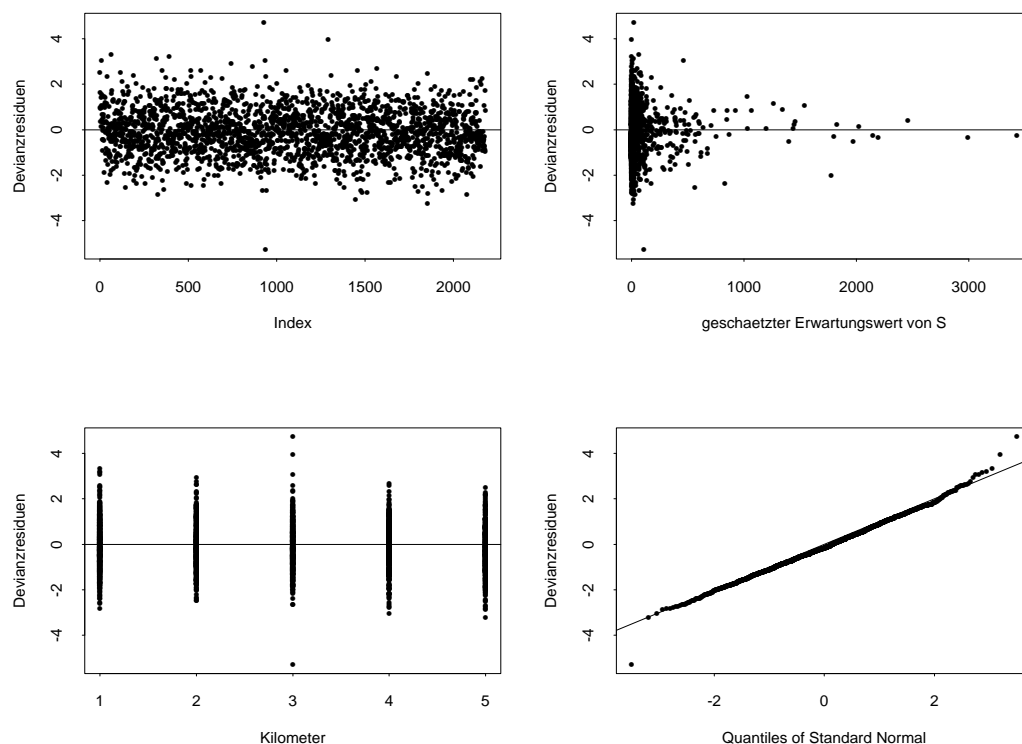


Abbildung 5.5: Devianzresiduen des NB-Modells (5.2) der Kfz-Daten

Coefficients:

	Value	Std. Error	t value	Pr(> t)
(Intercept)	-1.78219839	0.02372733	-75.111614	0
kilometerf2	0.18856823	0.01502698	12.548642	0
kilometerf3	0.27537027	0.01617412	17.025359	0
kilometerf4	0.35211554	0.01928738	18.256263	0
kilometerf5	0.51678723	0.02005369	25.770180	0
zonef2	-0.22424254	0.01788669	-12.536838	0
zonef3	-0.38285860	0.01803328	-21.230669	0
zonef4	-0.55590741	0.01696209	-32.773511	0
zonef5	-0.33855491	0.02252520	-15.030053	0
zonef6	-0.52249743	0.01998301	-26.147080	0
zonef7	-0.73067681	0.04622826	-15.805847	0
bonusf2	-0.44284018	0.02134377	-20.747979	0
bonusf3	-0.68053311	0.02238521	-30.401018	0
bonusf4	-0.82037411	0.02309783	-35.517366	0
bonusf5	-0.91903955	0.02237480	-41.074755	0
bonusf6	-0.99315080	0.02029865	-48.926931	0
bonusf7	-1.32594264	0.01768959	-74.956114	0
automarkef2	0.06726367	0.02559771	2.627723	0.008595848
automarkef3	-0.23520926	0.02943322	-7.991287	1.332268e-15
automarkef4	-0.68356843	0.02931592	-23.317311	0
automarkef5	0.15226372	0.02450209	6.214316	5.15487e-10
automarkef6	-0.36330053	0.02209394	-16.443445	0
automarkef7	-0.07909344	0.02768327	-2.857085	0.004275513
automarkef8	-0.04117853	0.03516840	-1.170896	0.2416406
automarkef9	-0.09024748	0.01566068	-5.762679	8.278916e-09

(Dispersion Parameter for Negative Binomial family taken to be 1)

Null Deviance: 11603.73 on 2181 degrees of freedom

Residual Deviance: 2231.168 on 2157 degrees of freedom

Number of Fisher Scoring Iterations: 1

Theta: 111.05799

Std. Err.: 14.89298

2 x log-likelihood: 1100220.48205

Tabelle 5.2: *summary*-Tabelle des NB-Modells (5.2) der Kfz-Daten

Poissonmodells schon beschrieben wurden. In den ersten 3 Abbildungen ist zu erkennen, dass weiterhin eine gleichmässige Streuung um 0 vorliegt, aber die betragsmässig maximale Abweichung ist nun von 6 auf 4 gesunken. Die letzte Graphik rechts unten zeigt deutlich die verbesserte Anpassung der Daten durch das NB-Modell, da die Devianzresiduen dort fast alle auf der Winkelhalbierenden liegen und somit eine Normalverteilung der Residuen angenommen werden kann. Das „summary“ liefert nun zusätzlich noch den geschätzten Dispersionsparameter „*theta*“, der nach (3.23) dem Bruch $\frac{1}{a}$ entspricht. Damit ergibt sich ein geschätzter Dispersionsparameter \hat{a} von 0.009, der angesichts der Tatsache, dass ein $a = 0$ zu einem Poissonmodell ohne Überdispersion führt, eine relativ geringe Überdispersion in den Daten vermuten lässt. Doch auch hier sei angemerkt, dass der Wert des Dispersionsparameters allein Schwierigkeiten bereitet, den Grad der Überdispersion zu quantifizieren. Um dies zu verdeutlichen sind in der Abbildung 5.6 die beiden p-Wert-Kurven bezüglich des Dispersionsparameters a_0 und des Abstandsmasses d_0 abgebildet. Zusätzlich zu den Kurven sind die Levels $\hat{P} = 0.05$ und $\hat{P} = 0.95$

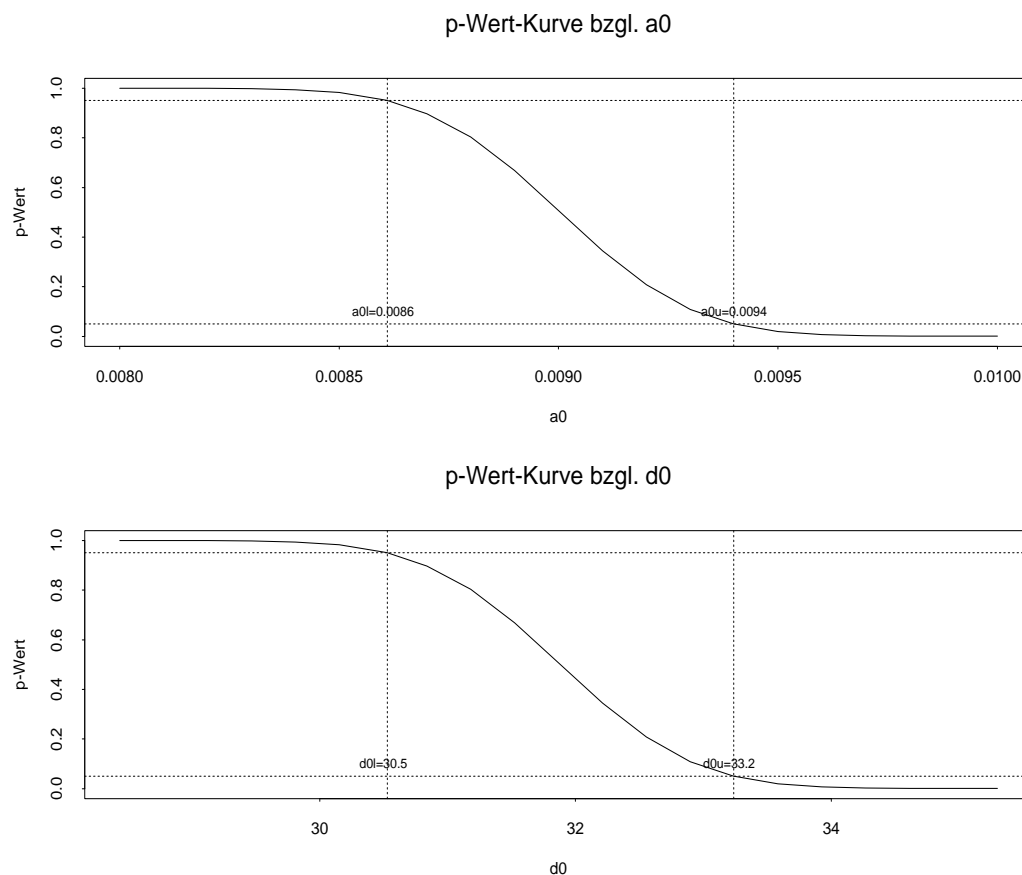


Abbildung 5.6: p-Wert-Funktionen für schwedische Kfz-Daten

mit einer horizontalen gestrichelten Linie eingezeichnet. Die p-Wert-Kurven werden also zu einem Signifikanzniveau von $\alpha = 0.05$ betrachtet, wobei zu bemerken ist, dass mit der Anwendung der neuen Testvorschrift (4.1), die in Kapitel 4 eingeführt wurde, die Ergebnisse wie für ein Signifikanzniveau von $\alpha = 0.1$ zu interpretieren sind. Die obere Kurve in Abbildung 5.6 gibt einem die Akzeptanzgrenze für das Poissonmodell von $a_{0u} = 0.0094$ an, die, wie schon angesprochen, so gering erscheint, dass man dazu geneigt ist, das Poissonmodell zu akzeptieren. Betrachtet man dagegen die untere Kurve bezüglich des Abstandsmasses d_0 , so erhält man eine Akzeptanzgrenze für das Poissonmodell von $d_{0u} = 33.2$. Dieser hohe Wert von d_{0u} bedeutet aber wegen (3.27), dass die maximale geschätzte Varianz des NB-Modells das 33.2-fache der geschätzten Varianz des Poissonmodells beträgt. Eine solch grosse Abweichung in den maximalen Varianzen ist nicht zu akzeptieren, weshalb unter Betrachtung der p-Wert-Kurven bezüglich d_0 gefolgert werden muss, dass das Poissonmodell nicht zu rechtfertigen ist. Wie schon in der Simulationsstudie zu sehen war, haben die extrem unterschiedlichen Aussagen der Kurven bezüglich a_0 und d_0 ihre Ursache in den grossen Werten der geschätzten Erwartungswerte der Schadenanzahlen (hier bis über 3000), die in der Abbildung 5.7 noch einmal dargestellt sind. Da der maximale Wert des geschätzten Erwartungswertes

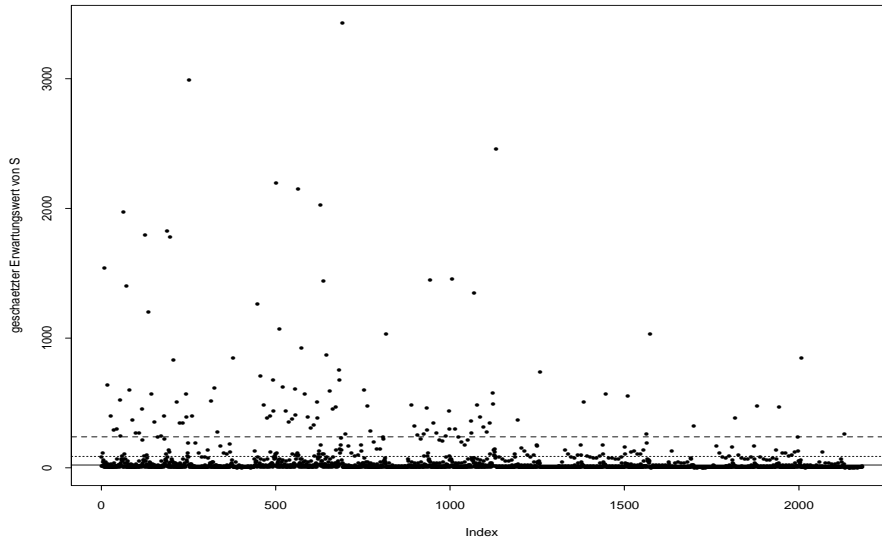


Abbildung 5.7: geschätzter Erwartungswert für S im NB-Modell (5.2)

tes $\max_i \{\hat{\mu}_i\}$ in die Berechnung des Abstandsmasses nach (3.28) durch die Formel $d_{0u} = 1 + a_{0u} \max_i \{\hat{\mu}_i\}$ einfließt, führt dies trotz eines sehr kleinen a_{0u} 's zu einem extrem grossen Wert d_{0u} , der eine Rechtfertigung des Poissonmodells nicht zulässt. In

der Abbildung 5.7 ist allerdings zu erkennen, dass die Schwankungsbreite des Erwartungswertes sehr hoch ist (von fast 0 bis über 3000) und dass die einzelnen Werte $\hat{\mu}_i$ im Mittel sehr viel kleiner als der maximale Wert $\max_i \{\hat{\mu}_i\}$ sind. Weiterhin sind in der Graphik drei horizontale Linien eingezeichnet, die das 75%-, das 90%- und das 95%-Quantil angeben. Daran lässt sich ablesen, dass 90% der geschätzten Erwartungswerte nicht über einen Wert von 87 hinausgehen und 95% der $\hat{\mu}_i$'s zwischen 0 und 238 liegen. Aus diesem Grund ist es sinnvoll, die bisher angegebene Abstandsmassfunktion leicht zu verändern, indem nun nicht mehr der maximale Erwartungswert der gesamten Daten betrachtet wird, sondern man schränkt sich auf einen Teilbestand der Daten ein. Dazu wird in der Definition der Abstandsmassfunktion statt des maximalen geschätzten Erwartungswertes z.B. das 90%-Quantil eingesetzt. Daraus ergibt sich die neue Abstandsmassfunktion

$$d_0^{Quantil} = 1 + a_0 \hat{\mu}^{Quantil}. \quad (5.3)$$

Mit der Verwendung dieser Abstandsmassfunktion ändern sich die Ergebnisse der p-Wert-Kurven drastisch, da der Wert des Erwartungswertes in der Abstandsmassfunktion sich vom maximalen Wert 3429 auf nur 87.3 bei Verwendung des 90%-Quantils und auf 238 bei Verwendung des 95%-Quantils verringert. In der Abbildung 5.8 sind schliesslich die p-Wert-Kurven bezüglich dieses neuen Abstandsmasses für die 90%- und 95%-Quantile abgebildet. Aus der oberen Graphik geht hervor, dass nun eine Akzeptanzgrenze für das Poissonmodell von nur $d_{0u} = 1.82$ vorliegt. Dieser Wert bedeutet, dass die geschätzte Abweichung in den Varianzen zwischen dem Poisson- und dem NB-Modell für 90% der Daten nicht grösser als 1.82 wird, d.h. für 90% der Daten ist die Varianz des NB-Modell nur 1.82 mal so gross wie die des Poissonmodells. Bei der unteren Kurve ergibt sich schon ein etwas grösseres d_{0u} von 3.24. Möchte man also die Betrachtung von 95% der Daten einschliessen, so muss eine 3.24-fache Varianz des NB-Modells gegenüber der des Poissonmodells akzeptiert werden, was gerade noch angenommen werden kann. In beiden Fällen führt also die Betrachtung der Abweichung in den Varianzen für einen Teilbestand der Daten (siehe (5.3)) in diesem Beispiel zu einer Annahme des Poissonmodells. An diesem Beispiel wird deutlich, dass die Betrachtung des Dispersionsparameters a allein nicht ausreicht, um zu entscheiden, ob das Poisson- oder das NB-Modell gewählt werden sollte, da die Werte von a nicht bezüglich der Quantifizierung von Überdispersion interpretiert werden können. Vielmehr sollten die Kurven bezüglich des Abstandsmasses d betrachtet werden, aus denen konkrete Werte d_{0u} abgelesen werden können, die eine einfache Interpretation zulassen.

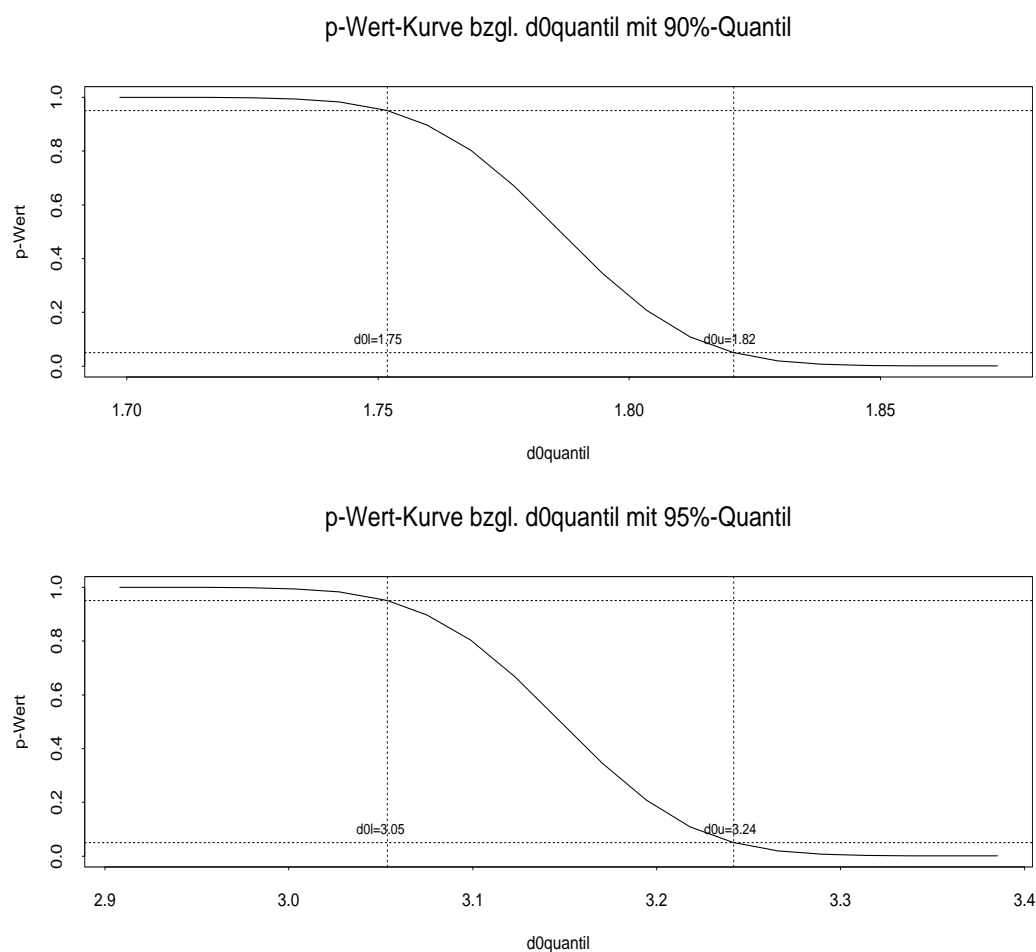


Abbildung 5.8: p-Wert-Funktionen mit neuer Abstandsmassfunktion $d_0^{Quantil}$

Dieses Beispiel zeigt weiter, dass aufgrund von grossen geschätzten Erwartungswerten die Kurve bezüglich des Abstandsmasses d zu anderen Ergebnissen führt, als zunächst die Kurve bezüglich a vermuten lässt. Verwendet man als Abstandsmass die maximale Abweichung in den Varianzen unter Berücksichtigung der gesamten Daten, so führt dies in diesem Beispiel zur deutlichen Ablehnung des Poissonmodells. Doch muss vor allem bei Daten, die grosse Erwartungswerte mit grosser Streuung aufweisen, die Struktur der Erwartungswerte beachtet werden, um zu entscheiden, ob die maximale Abweichung als Abstandsmass geeignet ist, oder ob die maximale Abweichung eines Teildatenbestandes ausreichend ist. Da hier 90% der Erwartungswerte in einem Bereich bis maximal 87 liegen, was deutlich unter dem maximalen Erwartungswert liegt, kann die Betrachtung auf die 90% der Daten eingeschränkt werden, woraus nach der Abbildung 5.8 die Annahme des Poissonmodells folgt.

5.2 Patent-Daten

Die nachfolgend untersuchten Daten sind auf der Internetseite

<http://coe.ubc.ca/users/marty/patent.data>

erhältlich und beschreiben Patent-Daten von amerikanischen High-Tech- Firmen aus dem Jahr 1976. Es sind im wesentlichen nur 3 Spalten vorhanden. Die erste Spalte „*patente*“ gibt die Anzahl der angemeldeten Patente einer Firma an und wird hier als Response verwendet. Die zweite Spalte „*rdsales*“ gibt das Verhältnis der Geldbeträge an, die für Forschung und Entwicklung (research and development) investiert wurden im Vergleich zu den Einnahmen, die sich durch die Entwicklung der Patente ergaben (R&D/Sales). Die dritte Spalte ist hier als „*log(rd)*“ angegeben, d.h. die Geldbeträge für Forschung und Entwicklung liegen in einer durch den Logarithmus transformierten Form vor. Da hier im folgenden aber auch die reinen Investitionsgelder für Forschung und Entwicklung benötigt werden, wird nicht die Spalte „*log(rd)*“, sondern die durch die Exponentialfunktion zurücktransformierte Spalte „*rd*“ verwendet. Diese Spalte ist in der Einheit von 1 Mio.Dollar angegeben. Der Originaldatensatz enthält noch eine zusätzliche Spalte „*(log(rd))²*“, auf die aber hier nicht weiter eingegangen wird. Die Daten liegen nicht gruppiert vor, sondern basieren auf Einzeldaten, d.h. jede Zeile des Datensatzes kann genau einer Firma zugeordnet werden. Insgesamt wurden 70 Firmen untersucht, d.h. die Länge der Daten ist auf 70 beschränkt. Analog zum schwedischen Kfz-Beispiel kann auch dieser Datensatz zum Vergleich mit den Parametersätzen in der Simulationsstudie in die Gruppen 3 oder 4 mit einem SN-Quotienten von 2 eingestuft werden, da sich für die Patent-Daten mit dem nachfolgend in (5.5) ermittelten Dispersionsparameters a und einem gemittelten Erwartungswert $\bar{\mu} = 27.1$ ein SN-Quotient von 1.75 ergibt. Aus der Spalte „*patente*“ geht hervor, dass die Anzahlen der angemeldeten Patente von 0 bis 173 variieren, d.h. es liegt eine grosse Schwankungsbreite des Erwartungswertes vor. Dieser Datensatz ist also zusammenfassend am besten mit demjenigen Parametersatz der Simulationsstudie vergleichbar, dessen Erwartungswert $\bar{\mu} = 10$ ist (Gruppe 3), der eine grosse Schwankungsbreite des Erwartungswertes aufweist und dessen Datenlänge $n = 100$ beträgt (Parametersatz 12). Für diesen Parametersatz ergab sich bei der Untersuchung der Powerfunktionen eher ein liberaler Test, doch da der gemittelte Erwartungswert der Patent-Daten mit $\bar{\mu} = 27.1$ schon deutlich grösser als der des Parametersatzes 12 mit $\bar{\mu} = 10$ ist, kann für die Patent-Daten analog zum schwedischen Kfz-Beispiel ein weniger liberaler Test erwartet werden.

5.2.1 Explorative Datenanalyse der Patent-Daten

Ähnlich wie im vorangegangenen Datenbeispiel wird nun anhand von ersten Abbildungen versucht herauszufinden, mit welchen Kovariablen und eventuell transformierten Kovariablen die zu untersuchende Zählvariable „Anzahl der Patente“ erklärt werden kann. Da in diesem Datenbeispiel keine Gewichtung des Responses notwendig ist, entfällt die Anwendung eines „offsets“. Ausserdem ist anzumerken, dass im Gegensatz zum Beispiel der schwedischen Kfz-Daten hier keine kategoriellen Daten vorliegen, d.h. alle Kovariablen werden als metrisch angesehen. Verwendet man nun zunächst die Poissonverteilung mit der log-link-Funktion als Linkspezifizierung zur Modellierung der Daten, so ergibt sich:

Patente $\sim Poi(\mu_i)$, $i = 1, \dots, 70$ mit

$$\mu_i = \exp(\mathbf{x}_i^t \boldsymbol{\beta}) \quad \text{bzw.} \quad \ln(\mu_i) = \mathbf{x}_i^t \boldsymbol{\beta}.$$

Lässt man sich also den Logarithmus der Patente auf der y-Achse gegen die Kovariablen auf der x-Achse antragen, so müsste sich zwischen ihnen ein linearer Zusammenhang

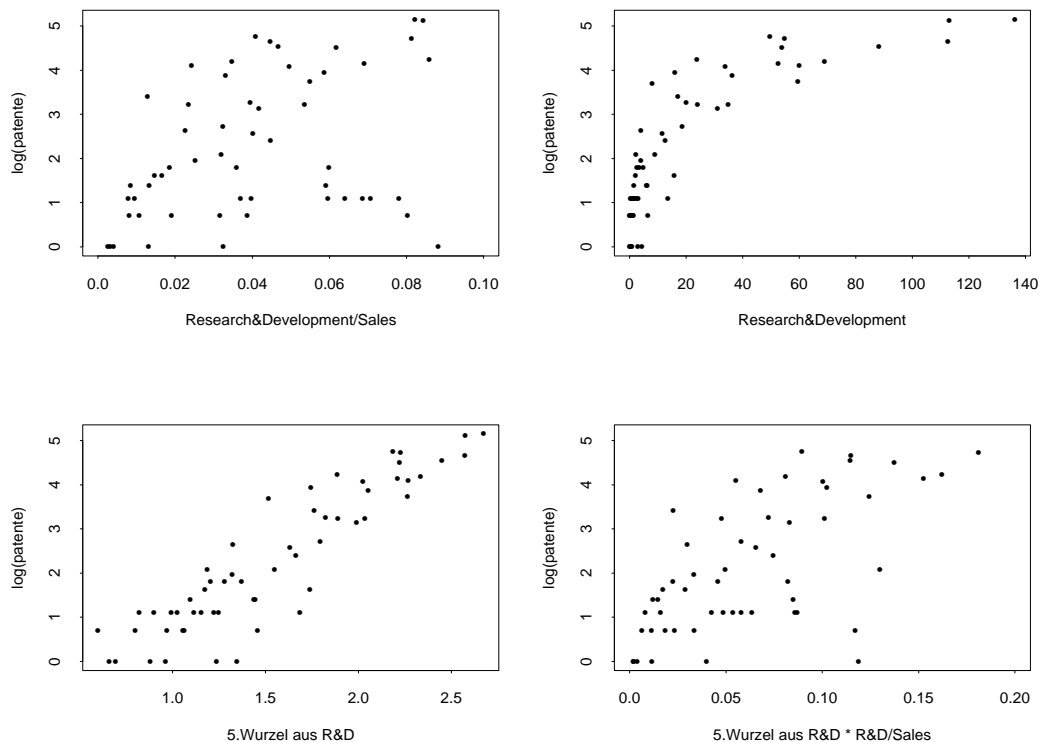


Abbildung 5.9: Explorative Datenanalyse der Patent-Daten

ergeben. In Abbildung 5.9 zeigt die obere linke Graphik zwar nur undeutlich eine Linearität, doch wird zunächst versucht, die Variable „*rdsales*“ (R&D/Sales) untransformiert in das Modell aufzunehmen. Die obere rechte Graphik zeigt jedoch durch die leichte Bogenform, dass hier eine Transformation der Variable „*rd*“ (R&D) notwendig ist. Um eine solche Bogenform zu linearisieren, kann versucht werden, die Variable mit der Logarithmusfunktion zu transformieren, oder man verwendet eine Wurzelfunktion. An dieser Stelle sei erwähnt, dass die Transformation mit Hilfe der Logarithmusfunktion untersucht wurde, doch konnte in der Graphik der logarithmierten Spalte „*rd*“ gegen „*log(patente)*“ noch immer keine Linearität festgestellt werden. Aus diesem Grund wird hier die Wurzelfunktion als Transformation im weiteren Verlauf verwendet. Bildet man die 5. Wurzel aus der Variable „*rd*“, so kann an der unteren linken Graphik erkannt werden, dass diese Transformation vernünftig erscheint, da nun eine deutliche Linearität zwischen der neuen Variable „*rdw5*“ (5. Wurzel aus *rd*) und „*log(patente)*“ zu erkennen ist. Die letzte Graphik zeigt eine Interaktion zwischen den beiden Kovariablen „*rdw5*“ und „*rdsales*“, die aufgrund der leicht vorhandenen Linearität noch als relevant betrachtet und in das zu erstellende Modell mit aufgenommen wird. Mit den anhand der Abbildung 5.9 gefundenen Ergebnissen wird nun ein Poissonmodell erstellt, das die Kovariablen „*rdsales*“ und „*rdw5*“ und zusätzlich die Interaktion zwischen beiden Kovariablen enthält.

5.2.2 Poissonregression der Patent-Daten

In Splus wird daher das folgende Poissonmodell berechnet:

```
patentpoi_glm(patente~rdsales+rdw5+rdsales*rdw5,family=poisson,
link=log,x=T).
```

(5.4)

In dem dazugehörigen „*summary*“ lassen sich dann die geschätzten Werte der Regressionsparameter β_j für $j = 1, \dots, 4$ unter der Spalte „*value*“ ablesen.

Coefficients:

	Value	Std. Error	t value	Pr(> t)	
(Intercept)	-1.1658	0.1850	-6.300	2.98e-10	***
rdsales	-5.5006	1.9909	-2.763	0.00573	**
rdw5	2.2997	0.1036	22.202	< 2e-16	***
rdsales:rdw5	3.9034	1.3119	2.975	0.00293	**

```

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for poisson family taken to be 1)

Null deviance: 3324.77  on 69  degrees of freedom
Residual deviance:  377.18  on 66  degrees of freedom
Number of Fisher Scoring iterations: 4

```

Tabelle 5.3: *summary*-Tabelle des Poissonmodells (5.4) der Patent-Daten

Aus der Tabelle 5.3 geht hervor, dass alle Kovariablen hoch signifikant sind, da alle p-Werte, die in der letzten Spalte aufgeführt sind, sehr viel kleiner als 0.05 sind, d.h. für alle $j = 1, \dots, 4$ können die Hypothesen $H_j : \beta_j = 0$, $j = 1, \dots, 4$ zu einem sehr kleinen Signifikanzniveau verworfen werden, woraus zu folgern ist, dass die zu β_j gehörenden Kovariablen einen signifikanten Einfluss auf die Zielvariable „Anzahl der Patente“ haben. Aus der Tabelle kann aber ebenfalls abgelesen werden, dass die Anpassung der Daten durch ein Poissonmodell als weniger gelungen eingestuft werden kann, da die Devianz mit 377.2 deutlich höher ist als die Anzahl der Freiheitsgrade (66). Diese Beobachtung bestätigt sich in der Betrachtung der Residuen. In der Abbildung 5.10 zeigt

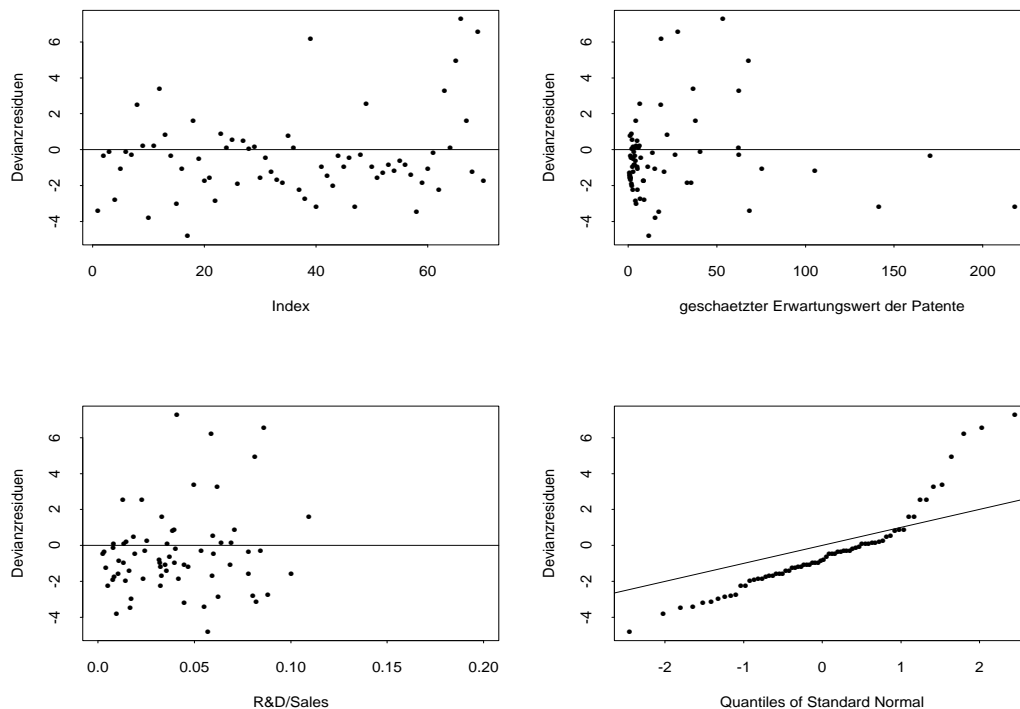


Abbildung 5.10: Devianzresiduen des Poissonmodells (5.4) der Patent-Daten

zwar die erste Graphik links oben, dass die Residuen zufällig um 0 streuen, doch sind die Werte der Residuen bis maximal 6 relativ hoch im Vergleich zu den geschätzten Erwartungswerten. In der zweiten Graphik erkennt man, dass vor allem die kleineren Erwartungswerte schlecht geschätzt werden, da für die Erwartungswerte im Bereich von 0 bis 75 viele Residuen einen Wert annehmen, der betragsmässig grösser als 4 ist. Die untere linke Graphik zeigt die Devianzresiduen, die hier gegen die Kovariable „rdw5“ abgetragen sind. Hieran lässt sich erkennen, dass zwar die Residuen unregelmässig um 0 streuen, jedoch ist eine leichte Tendenz in den negativen Bereich offensichtlich. Dies bedeutet nach Definition des Devianzresiduums $r_i^D = \text{sign}(y_i - \hat{\mu}_i) \sqrt{d_i}$ (siehe Abschnitt 2.2.9), dass die Vorzeichenfunktion „sign“ vorwiegend negativ ist, woraus wiederum zu schliessen ist, dass das $\hat{\mu}_i$ grösser ist als das y_i , d.h. dass in der Kovariable „rdw5“ die meisten Werte überschätzt werden. Der letzten Graphik ist schliesslich zu entnehmen, dass die Residuen sich keinesfalls nach der Normalverteilung verhalten, da sie deutlich sichtbar von der als durchgezogene Linie eingezeichneten Winkelhalbierenden abweichen. Anhand der „summary“-Tabelle und der Residuenanalyse kann bisher davon ausgegangen werden, dass das Poissonmodell die Daten offensichtlich nicht optimal anpasst. Um nun die Linkspezifizierung zu überprüfen, werden analog wie im vorange-

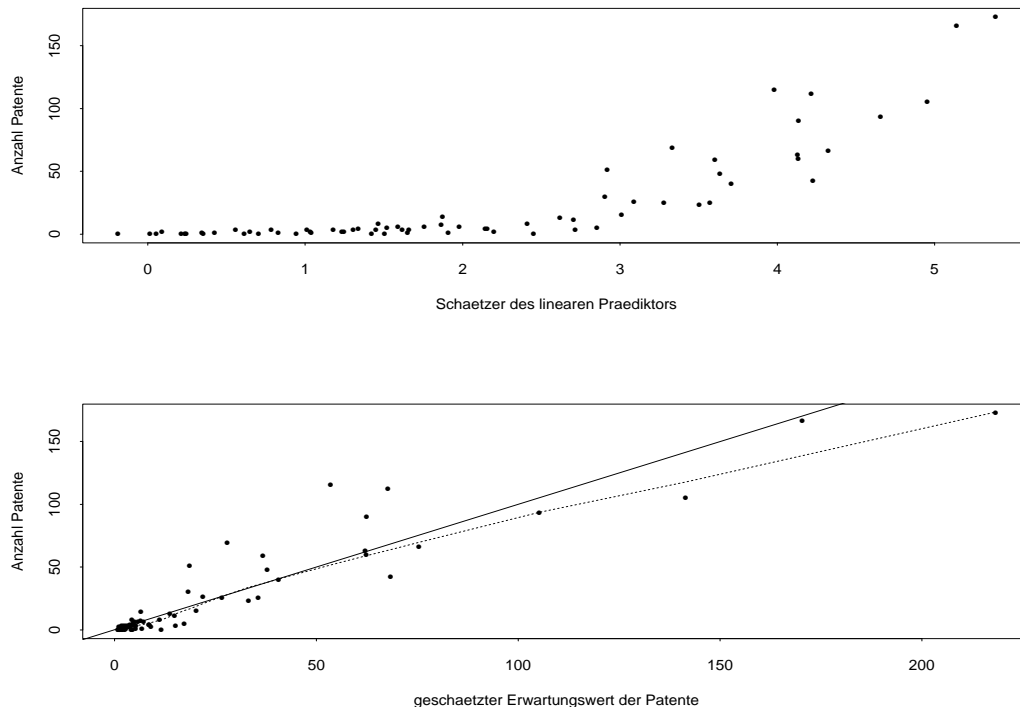


Abbildung 5.11: Linkspezifikation im Poissonmodell (5.4) der Patent-Daten

gangenen Beispiel zwei Graphiken erstellt, die in der Abbildung 5.11 dargestellt sind. Dabei zeigt die obere Graphik wie im Beispiel der schwedischen Kfz-Daten die Streuung der Zielvariable „patente“ gegenüber dem geschätzten linearen Prädiktor. Es ist an dieser Graphik zu erkennen, dass die Anzahl der Patente um die Kurve einer Exponentialfunktion streut, so wie dies auch bei einer richtigen Spezifizierung der log-link-Funktion gefordert wird. Um die untere Graphik zu erhalten, wird der lineare Prädiktor mit der Exponentialfunktion transformiert, woraus der geschätzte Erwartungswert resultiert. Durch die Transformation sollte nun ein linearer Zusammenhang zwischen der Anzahl der Patente und dem geschätzten Erwartungswert der Patente ersichtlich sein, was für kleine Anzahlen der Patente in einem Bereich bis zu 100 in etwa zutrifft, denn die mit „lowess“ erzeugte Glättung der einzelnen Punkte (gestrichelte Linie) stimmt in diesem Intervall mit der Winkelhalbierenden (durchgezogene Linie) überein. Für grössere Werte der Patente allerdings weicht die Glättungsfunktion deutlich von der Winkelhalbierenden ab, was aber auch damit erklärt werden kann, dass in diesem Bereich zuwenig Daten vorliegen. Zuletzt wird noch der geschätzte Erwartungswert gegen die geschätzte Varianz abgetragen, um die möglicherweise vorhandene Überdispersion, die die Ursache für die schlechte Anpassung sein könnte, graphisch darzustellen. Die

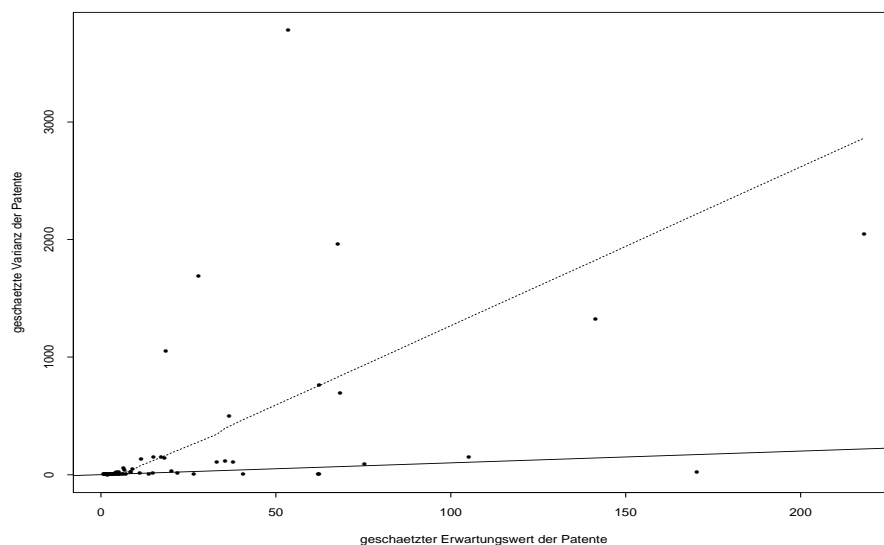


Abbildung 5.12: Überdispersion im Poissonmodell (5.4) der Patent-Daten

Abbildung 5.12 lässt keinen Zweifel daran, dass die Daten Überdispersion aufweisen, denn die gestrichelte Linie, die die Glättungsfunktion darstellt, verläuft weit oberhalb der als durchgezogene Linie eingezeichneten Winkelhalbierenden. Damit ist gezeigt,

dass die Varianz der einzelnen Beobachtungen deutlich grösser als ihr Erwartungswert ist, womit eine Modellannahme der Poissonverteilung verletzt ist. Dies ist auch die Erklärung dafür, dass im „*summary*“ die Devianz stark von der Anzahl der Freiheitsgrade abweicht, wodurch eine schlechte Anpassung der Daten durch das Poissonmodell indiziert ist. Im anschliessenden Unterabschnitt wird also versucht, die Anpassung der Daten mittels der NB-Verteilung zu verbessern.

5.2.3 NB-Regression und p-Wert-Kurven der Patent-Daten

Das NB-Modell, das nun zur Modellierung der Patent-Daten verwendet wird, enthält wie das Poissonmodell beide Kovariablen „*rdsales*“ und „*rdw5*“ zusammen mit der Interaktion der Kovariablen, um die Vergleichbarkeit mit dem Poissonmodell zu gewährleisten. Es wird in Splus also das folgende Modell berechnet:

```
patentnb_glm.nb(patente~rdsales+rdw5+rdsales*rdw5,link=log,x=T). (5.5)
```

Im dazugehörigen „*summary*“ (Tabelle 5.4) erkennt man, dass zwar die Kovariable „*rdsales*“ und die Interaktion nicht mehr einen so deutlichen Einfluss auf die Zielvariable „*Anzahl Patente*“ haben, da ihr p-Wert über 0.05 liegt, jedoch soll die Vergleichbarkeit mit dem Poissonmodell erhalten bleiben, weshalb die nichtsignifikanten Kovariablen nicht aus dem Modell entfernt werden. Anhand des Wertes für „*theta*“, der hier mit 3.4

Coefficients:

	Value	Std. Error	t value	Pr(> t)
(Intercept)	-1.855308	0.3980121	-4.661435	3.140122e-06
rdsales	-8.234623	5.9975079	-1.373007	0.1697502
rdw5	2.635982	0.2513850	10.485834	0
rdsales:rdw5	5.742834	3.9627112	1.449218	0.1472767

(Dispersion Parameter for Negative Binomial family taken to be 1)

Null Deviance: 524.4383 on 69 degrees of freedom

Residual Deviance: 78.55312 on 66 degrees of freedom

Number of Fisher Scoring Iterations: 1

Theta: 3.40194

Std. Err.: 0.92558

2 x log-likelihood: 10233.61781

Tabelle 5.4: *summary*-Tabelle des NB-Modells (5.5) der Patent-Daten

angegeben ist, errechnet sich ein Überdispersionsparameter \hat{a} mit $\hat{a} = \frac{1}{\theta} = \frac{1}{3.4} = 0.29$. Dieser Wert ist im Vergleich zum vorherigen Datenbeispiel relativ gross, so dass dies zusätzlich ein Anzeichen dafür ist, dass die Daten Überdispersion aufweisen und daher das NB-Modell zur Modellierung der Patent-Daten gut geeignet ist. Bemerkenswert ist auch, dass die Anpassung im Vergleich zum Poissonmodell nun deutlich besser geworden ist, was das Verhältnis der Devianz und der Anzahl der Freiheitsgrade betrifft. Die Devianz lässt sich aus der Tabelle mit einem Wert von 78.5 ablesen, während 66 Freiheitsgrade zur Verfügung stehen. Diese starke Senkung der Devianz von 377 im Poissonmodell auf nun 78.5 spricht dafür, dass das NB-Modell die Daten wesentlich besser als das Poissonmodell anpasst. Es bleibt dann zu überprüfen, ob sich die vermutete bessere Anpassung auch in den Residuen widerspiegelt. Die Abbildung 5.13 zeigt

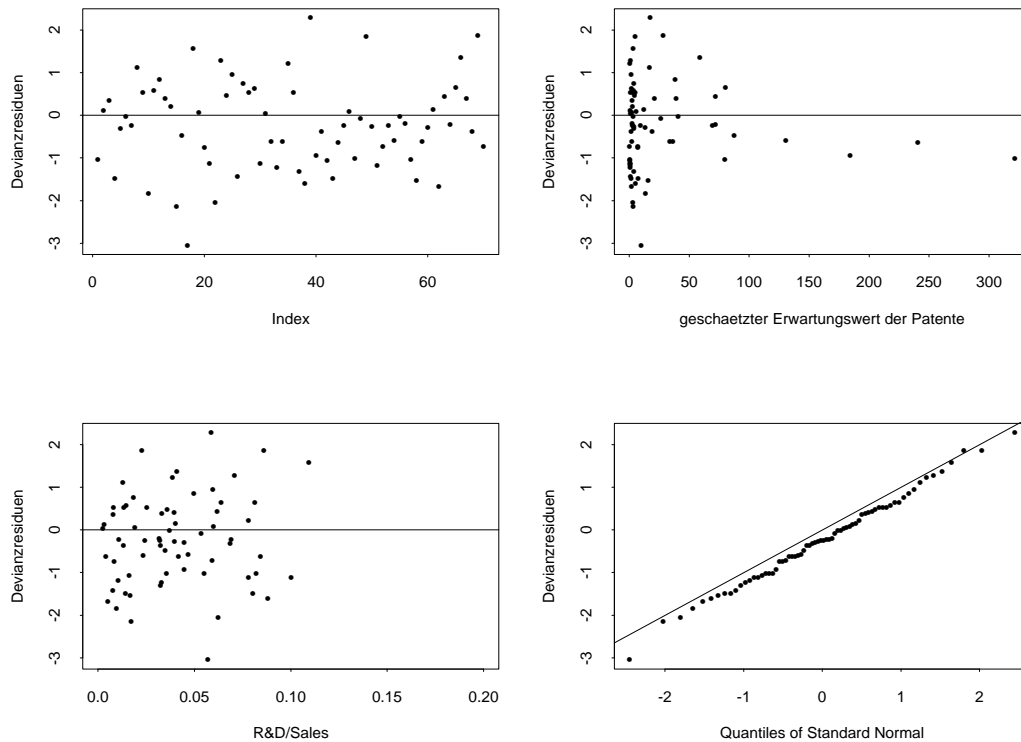


Abbildung 5.13: Devianzresiduen des NB-Modells (5.5) der Patent-Daten

deutlich, dass sich die Residuen mit der NB-Verteilung verbessern. Während bei der Poissonverteilung die maximalen Werte der Residuen grösser als 6 waren, sind hier maximale Abweichungen von nur 2 zu verzeichnen. Ausserdem fällt an der unteren linken Graphik auf, dass die Residuen nun gleichmässig um 0 verteilt sind, d.h. die Tendenz in den negativen Bereich, der in der Poissonregression auftrat, ist hier ausgeglichen.

Betrachtet man zuletzt die untere rechte Graphik, so liegen die Datenpunkte zwar nicht ganz auf der Winkelhalbierenden, aber die Abweichung ist deutlich geringer im Vergleich zum Poissonmodell, d.h. hier kann eher eine Normalverteilung der Residuen angenommen werden. Fasst man die momentanen Ergebnisse zusammen, so ergibt sich, dass die NB-Verteilung für die Struktur der Daten besser geeignet ist als die Poissonverteilung. Ob aber die Verbesserung durch die Verwendung der NB-Verteilung gross genug ist, um das Poissonmodell zu verwerfen, wird abschliessend anhand der p-Wert-Kurven überprüft. Zunächst sind in der Abbildung 5.14 die beiden p-Wert-Kurven bezüglich a_0 und d_0 abgebildet. Es wird angemerkt, dass die Kurven bezüglich eines

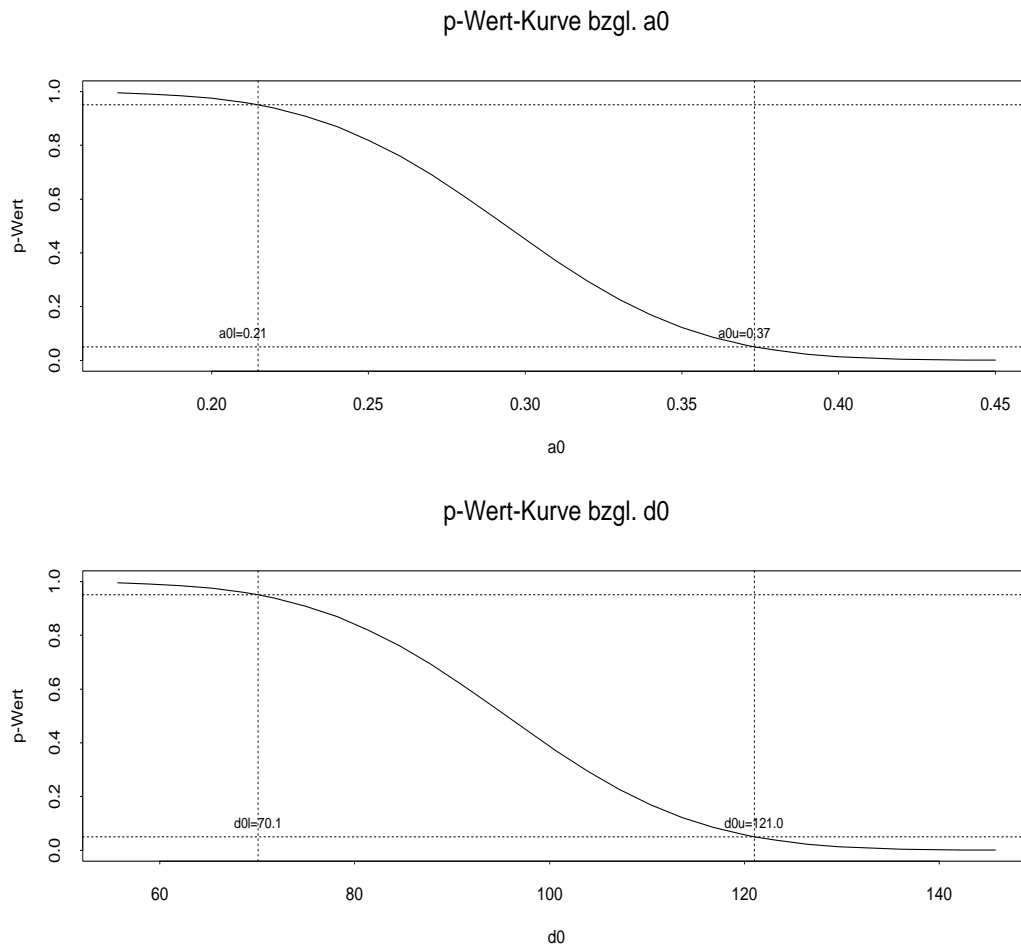


Abbildung 5.14: p-Wert-Funktionen für Patent-Daten

Signifikanzniveaus von $\alpha = 0.05$ gezeichnet und auch die Schnittpunkte der Kurven mit den Levels $\hat{P} = 0.05$ und $\hat{P} = 0.95$ berechnet sind, doch lassen sich diese Werte wie zu einem Signifikanzniveau $\alpha = 0.1$ interpretieren, wenn die neue Testvorschrift wie in Kapitel 4 (4.1) beschrieben, verwendet wird, um einen weniger liberalen Test

zu garantieren. Aus der oberen Kurve bezüglich a_0 kann abgelesen werden, dass zur Akzeptanz des Poissonmodells ein Dispersionsindex von $a_{0u} = 0.37$ akzeptiert werden müsste, der im Vergleich zum Kfz-Datenbeispiel sehr gross ist. Besser erkennt man jedoch an der unteren Kurve bezüglich des Abstandsmasses d_0 , dass das Poissonmodell nicht gerechtfertigt ist, da sich die Varianzen mit einem Faktor von $d_{0u} = 121$ unterscheiden, d.h. die Varianz des NB-Modells ist 121 mal so gross wie die des Poissonmodells. Beide Kurven lassen daher darauf schliessen, dass die Überdispersion so gross ist, dass das Poissonmodell verworfen werden muss. Es bleibt die Frage zu untersuchen, wie die p-Wert-Kurven verlaufen, wenn statt des maximalen geschätzten Erwartungswertes nur 90% oder 95% der Daten berücksichtigt werden. Dafür wird die in (5.3) definierte neue Abstandsmassfunktion verwendet. Dies ist dadurch zu rechtfertigen, da auch hier, ähnlich wie beim Kfz-Daten-Beispiel die geschätzten Erwartungswerte eine relativ grosse Schwankungsbreite aufweisen, und die meisten Erwartungswerte doch deutlich geringer als der maximale Wert sind. In der folgenden Abbildung 5.15 sind die geschätzten Erwartungswerte abgebildet und zusätzlich sind mit horizontalen Linien

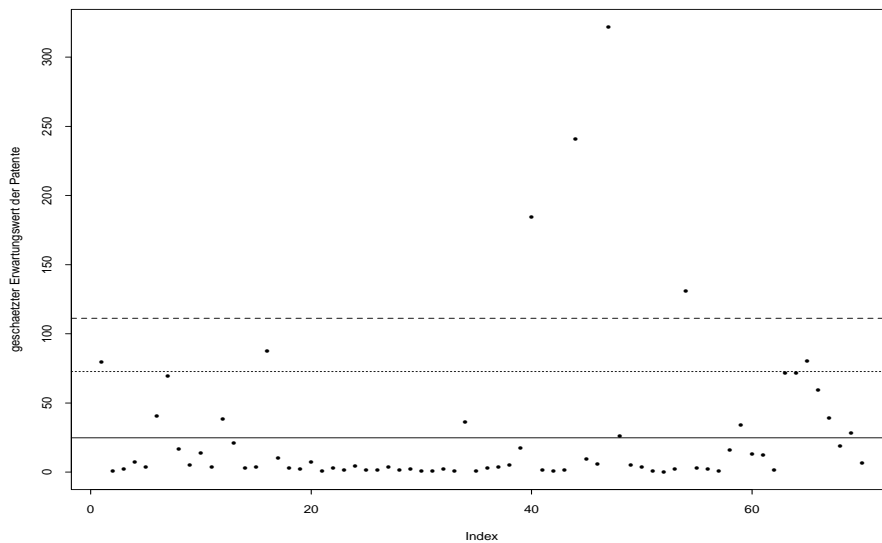


Abbildung 5.15: geschätzter Erwartungswert für Patente im NB-Modell (5.5)

die 75%-, 90%- und 95%- Quantile eingetragen. Akzeptiert man, dass nur 90% der Daten die nachfolgend berechnete Schranke d_{0u} nicht überschreiten, so kann in der Berechnung des Abstandsmasses statt des maximalen Wertes $\max_i \{\hat{\mu}_i\} = 321.7$ nun das 90%-Quantil 72.6 verwendet werden, bzw. wenn man die maximale Abweichung im Abstandsmass von 95% der Daten garantieren möchte, so muss das 95%-Quantil (111.2)

der geschätzten Erwartungswerte verwendet werden. Daraus ergeben sich schliesslich die p-Wert-Kurven bezüglich des Abstandsmasses $d_0^{Quantil}$, die in der Abbildung 5.16 dargestellt sind. Doch hier ist in beiden Graphiken im Gegensatz zum Beispiel der

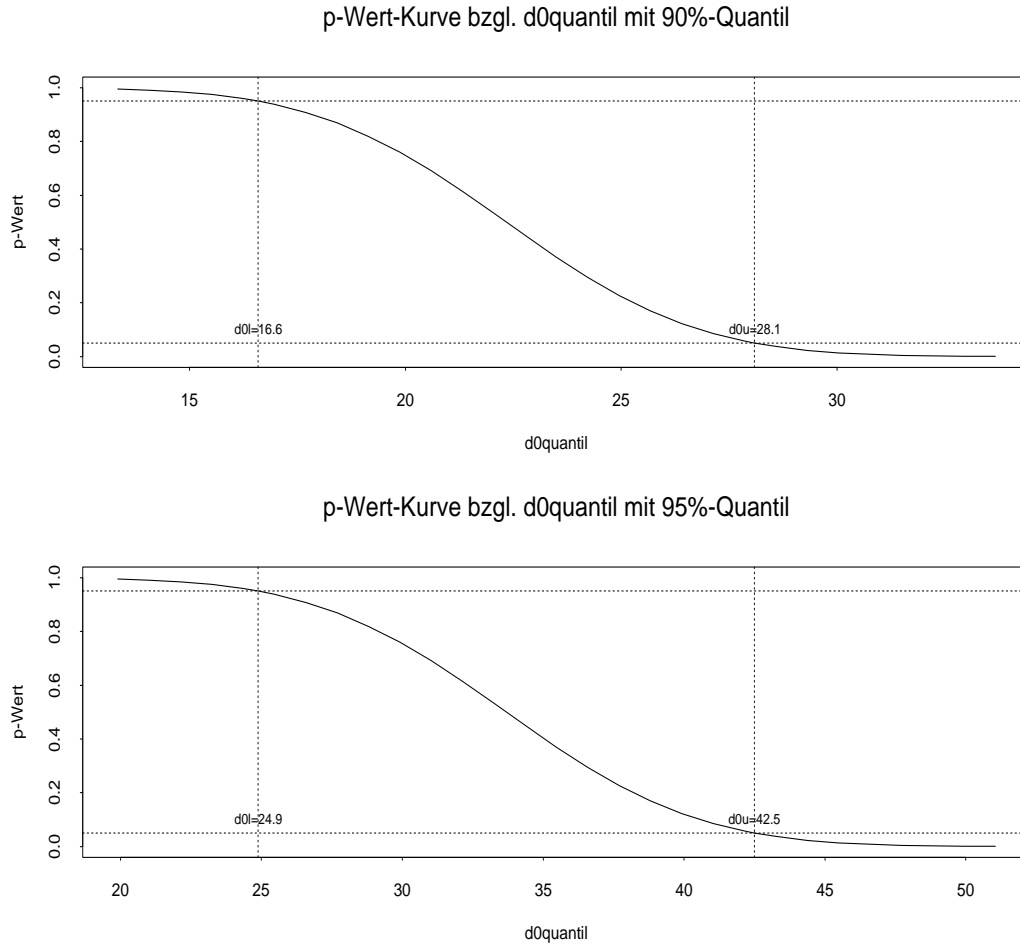


Abbildung 5.16: p-Wert-Funktionen mit neuer Abstandsmassfunktion $d_0^{Quantil}$

schwedischen Kfz-Daten keine Verbesserung zu erkennen. Selbst wenn nur die maximale Abweichung von 90% der Daten durch den Wert $d_{0u} = 28.1$ garantiert wird, ist dieser Wert immer noch viel zu hoch, als dass er eine Akzeptanz des Poissonmodells rechtfertigen würde. In der unteren Graphik erkennt man, dass für 95% der Daten eine Abweichung der Varianzen von $d_{0u} = 42.5$ vorliegt, d.h. in 95% der Fälle beträgt die Varianz des NB-Modells das 42.5-fache der Varianz des Poissonmodells. Alle p-Wert-Kurven haben also die aus dem „summary“ und den Residuen gewonnene Vermutung bestätigt, dass die Anpassung dieses Datensatzes nur mit der NB-Verteilung gerechtfertigt ist.

Kapitel 6

Zusammenfassung

Zum Abschluss dieser Arbeit sollen hier noch einmal die wichtigsten Erkenntnisse und Ergebnisse zusammengefasst werden. Da die Poissonverteilung wesentlich einfacher zu modellieren ist als die NB-Verteilung, ist es generell erwünscht, Zählraten wenn möglich mit einem Poissonmodell anzupassen. Das Problem, das dabei entstehen kann, ist, dass die Daten sehr häufig Überdispersion aufweisen und somit eine wichtige Modellannahme des Poissonmodells (die Gleichheit von Erwartungswert und Varianz) verletzen. Das Ziel dieser Arbeit war es nun, den Grad der Überdispersion zu quantifizieren, um konkrete Aussagen treffen zu können, ob das Poissonmodell noch akzeptabel ist, oder ob das NB-Modell gewählt werden muss, um die zu hohe Überdispersion zu berücksichtigen. Dazu wurde zunächst der Zusammenhang beider Verteilungen ausgenutzt, d.h. es wurde die Eigenschaft verwendet, dass die NB-Verteilung die Poissonverteilung als Spezialfall enthält, nämlich genau dann, wenn der Dispersionsparameter der NB-Verteilung $a = 0$ ist. Mit Hilfe dieser Eigenschaft wurde dann ein Bereichstest konstruiert, der testet, ob der geschätzte Dispersionsparameter grösser als eine vorgegebene Schranke a_0 ist. Zu diesem Test wurden dann zunächst in einer Simulationsstudie anhand von zufällig erzeugten NB-Daten die Powerfunktionen betrachtet, deren Ergebnisse auf einen stark liberalen Test schliessen liessen. Es wurde daher die Möglichkeit angesprochen, die Liberalität des Tests zu reduzieren, indem der Ablehnungsbereich des Bereichstests statt durch das Signifikanzniveau α durch $\frac{\alpha}{2}$ beschränkt wird. Mit Einführung dieser neuen Testvorschrift konnte die Liberalität des Tests stark verkleinert werden. Anschliessend erfolgte dann anhand der Simulationsdaten die Untersuchung der sogenannten p-Wert-Kurven, die den wichtigsten Teil dieser Arbeit darstellen. Mit Hilfe der p-Wert-Kurven kann konkret angegeben werden, welche Schranken a_0 des Dispersionsparameters zu akzeptieren sind, um das

Poissonmodell zu rechtfertigen. Es ergab sich aber aus den Kurven, dass der Dispersionsparameter Schwierigkeiten bereitet, den Grad der Überdispersion zu interpretieren, weshalb ein Abstandsmass d_0 eingeführt wurde, das eine anschauliche Interpretation zulässt. Das Abstandsmass wurde so gewählt, dass es das Verhältnis der maximalen Varianzen beider Modelle widerspiegelt, so dass es bei Fehlen von Überdispersion 1 ergibt und bei Vorhandensein von Überdispersion einen Wert grösser 1 annimmt. Die Untersuchung der p-Wert-Kurven aus der Simulationsstudie ergab, dass die Berechnungen für Parametersätze mit einer grossen Datenlänge genauer ausfielen, was auch damit zusammenhängt, dass der Bereichstest ein asymptotischer Test ist. Weiter ergab sich anhand der Kurven, dass speziell die Ergebnisse bei der Betrachtung der p-Wert-Kurven bezüglich des Abstandsmasses stark von der Grösse des geschätzten Erwartungswertes abhängen. Sind die Erwartungswerte sehr hoch, so führt dies schnell zu einem grossen Wert des Abstandsmasses, was wiederum die Ablehnung des Poissonmodells zur Folge hat. Abschliessend wurde der Bereichstest auf 2 konkrete Datenbeispiele angewendet und jeweils die p-Wert-Kurven gebildet. Im ersten Beispiel führte zunächst die p-Wert-Kurve bezüglich des Abstandsmasses zu einer deutlichen Ablehnung des Poissonmodells, obwohl ein äusserst geringer Dispersionsparameter vorlag. Die Ursache für die unterschiedlichen Ergebnisse in den p-Wert-Kurven bezüglich a_0 und d_0 lag in den sehr grossen Erwartungswerten. Durch die Einführung eines neuen Abstandsmasses, das die maximale Abweichung in den Varianzen von nur 90% der Daten berücksichtigt, sind allerdings p-Wert-Kurven entstanden, aus denen Werte für d_0 abgelesen werden konnten, die eine Annahme des Poissonmodells rechtfertigten. Das zweite Beispiel dagegen lieferte p-Wert-Kurven, die allesamt gegen das Poissonmodell sprachen. Durch das Abstandsmass d_0 ergab sich nämlich, dass die Varianz des NB-Modells ein grosses Vielfaches der Varianz des Poissonmodells betrug, was ein deutlicher Hinweis auf zu grosse Überdispersion in den Daten war, weshalb in diesem Fall der Wechsel vom einfachen Poissonmodell zum aufwendigeren NB-Modell unbedingt erforderlich wurde.

Anhang

Das folgende Programm dient zur Erstellung der Powerfunktion, wie in Abschnitt 3.3.2 beschrieben.

```
powerfunktion_function(a0,a,beta,x,alpha)
{
  n_length(x)
  now <- proc.time()
  X_matrix(1,nrow=n,ncol=2)
  X[,2]_x
  mu_rep(0,n)
  muhut_matrix(0,nrow=n,ncol=300)
  ahut_rep(0,300)
  betahut_matrix(0,nrow=2,ncol=300)
  Z_matrix(0,nrow=n,ncol=300)
  sigmahut_rep(0,300)
  phi_rep(0,300)
  kr_rep(0,300)
  #Berechnung des wahren Erwartungswertes
  for ( i in (1:n))
  {
    mu[i]_exp(X[i,]%*%beta)
  }
  for ( k in (1:300))
  {
    #Erzeugung der NB-Zufallsvariablen
    Z[,k]_rnegbin(mu,theta=1/a)
    #Schaetzung der Parameter a und beta
    out_glm.nb(Z[,k]~x,link=log,control=glm.control(maxit=500))
  }
}
```

```

    ahut[k]_1/out$theta
    betahut[,k]_c(out$coef[1],out$coef[2])
    #Berechnung des geschätzten Erwartungswertes
    for ( i in (1:n))
    {
    muhut[i,k]_exp(X[i,]%%betahut[,k])
    }
    #Berechnung von sigmahut
    innen_matrix(0,nrow=n,ncol=151)
    innensum_rep(0,n)
    for ( i in (1:n))
    {
    innen[i,1]_((1/ahut[k])^(-2))
    for ( j in (1:150))
    {
    innen[i,j+1]_((1/ahut[k]+j)^(-2))*(1-NBvtlg(j-1,ahut[k],muhut[i,k]))
    }
    innensum[i]_sum(innen[i,])-ahut[k]*muhut[i,k]/(muhut[i,k]+1/ahut[k])
    }
    ibetaazw_ahut[k]^(-4)*sum(innensum)
    ibetaainv_1/ibetaazw
    sigmahut[k]_sqrt(ibetaainv)
    #Berechnung der Teststatistik
    phi[k]_pnorm((ahut[k]-a0)/sigmahut[k])
    if (phi[k]<= alpha)  kr[k]_1
    else kr[k]_0
    }
    poweraa0_sum(kr)/300
    speed <- proc.time() - now
    print("time spent")
    print(speed)
    return(list(ahut=ahut,betahut=betahut,muhut=muhut,sigmahut=sigmahut,
    phi=phi,poweraa0=poweraa0,speed=speed))
}

```

Für das eben beschriebene Programm werden noch die Programme „*NBVtlg*“ und „*fnbinom*“ benötigt, wobei „*fnbinom*“ die Dichte, und „*NBVtlg*“ die Verteilungsfunktion der NB-Verteilung für stetige Parameterwerte a berechnet.

```
fnbinom_function(y,a,mu)
{
  zaehler_gamma(y+1/a)
  nenner_gamma(1/a)*fak(y)
  if (zaehler==Inf) zaehler_1e300
  if (nenner==Inf) nenner_1e300
  erg_zaehler/nenner*(mu/(1/a+mu))^y*((1/a)/(1/a+mu))^(1/a)
return(erg)
}
```

```
NBVtlg_function(y, a, mu)
{
  vertvektor <- rep(0, y + 1)
  for(i in (0:y)) {
    vertvektor[i + 1] <- fnbinom(i, a, mu)
  }
  summe <- sum(vertvektor)
  return(summe)
}
```

Das nun folgende Programm „*a0quer*“ erzeugt die notwendigen Daten zur Erstellung von 20 verschiedenen p-Wert-Kurven pro Parametersatz. Es gibt zusätzlich die durchschnittlichen Schnittpunkte \bar{a}_{0l} und \bar{a}_{0u} der p-Wert-Kurven mit den Levels $\hat{P} = 0.95$ und $\hat{P} = 0.05$ aus. Dieses Programm ist in Abschnitt 3.3.4 näher erläutert.

```
a0quer_function(x,beta,a)
{
  n_length(x)
  Y_matrix(0,nrow=n,ncol=20)
  ahut_rep(0,20)
  betahut_matrix(0,nrow=2,ncol=20)
  muhut_matrix(0,nrow=n,ncol=20)
  sigmahut_rep(0,20)
```

```

a0l_rep(0,20)
a0u_rep(0,20)
for (k in (1:20))
{
  #Erzeugung der NB-Zufallsvariablen
  Y[,k]_nbzv(x,beta,a)
  #Schaetzung der Parameter beta und a
  out_glm.nb(Y[,k]~x,link=log)
  ahut[k]_1/out$theta
  betahut[,k]_c(out$coef[1],out$coef[2])
  #Erzeugung der Design-Matrix X
  X_matrix(1,nrow=n,ncol=2)
  X[,2]_x
  #Berechnung des geschaetzten Erwartungswertes
  for ( i in (1:n))
  {
    muhut[i,k]_exp(X[i,]%%betahut[,k])
  }
  #Berechnung von sigmahut
  innen_matrix(0,nrow=n,ncol=151)
  innensum_rep(0,n)
  for ( i in (1:n))
  {
    innen[i,1]_((1/ahut[k])^(-2))
    for ( j in (1:150))
    {
      innen[i,j+1]_((1/ahut[k]+j)^(-2))*(1-NBVtlg(j-1,ahut[k],muhut[i,k]))
    }
    innensum[i]_sum(innen[i,])-ahut[k]*muhut[i,k]/(muhut[i,k]+1/ahut[k])
  }
  ibetaazw_ahut[k]^(-4)*sum(innensum)
  ibetaainv_1/ibetaazw
  sigmahut[k]_sqrt(ibetaainv)
  #Berechnung der Schnittpunkte a0l und a0u
  a0l[k]_ahut[k]-(1.645*sigmahut[k])
}

```

```

    a0u[k]_ahut[k]+(1.645*sigmahut[k])
  }
  #Berechnung der durchschnittlichen Schnittpunkte
  a0lquer_sum(a0l)/20
  a0uquer_sum(a0u)/20
  return(list(Y=Y,ahut=ahut,betahut=betahut,muhut=muhut,sigmahut=sigmahut,
    a0l=a0l,a0u=a0u,a0lquer=a0lquer,a0uquer=a0uquer))
}

```

In diesem Programm erfolgt die Erzeugung der NB-Zufallsvariablen über einen Programmaufruf „*nbzv*“, in welchem zunächst der wahre Erwartungswert berechnet wird und schliesslich durch den in Splus implementierten Befehl „*rnegbin*“ die Zufallsvariablen mit den gewünschten Parametern erzeugt werden.

```

nbzv_function(x,beta,a)
{
  n_length(x)
  #Berechnung des wahren Erwartungswertes mu
  X_matrix(1,nrow=n,ncol=2)
  X[,2]_x
  mu_rep(0,n)
  for ( i in (1:n))
  {
    mu[i]_exp(X[i,]*beta)
  }
  #Erzeugung von Zufallsvariablen mit Parameter a und mu
  Y_rnegbin(mu,theta=1/a)
  return(Y)
}

```

Die Programme „*pwertaneu*“ und „*pwertdneu*“ dienen schliesslich der Erstellung der p-Wert-Kurven, einmal bezüglich des Parameters a_0 und einmal bezüglich des Abstandsmasses d_0 . Voraussetzung ist, dass die Daten, die im vorherigen Programm „*a0quer*“ berechnet wurden, unter einem Namen „-out“ abgespeichert wurden. Auch diese Programme wurden bereits in Abschnitt 3.3.4 beschrieben. Da sich das Programm „*pwertaneu*“ von „*pwertdneu*“ kaum unterscheidet (es entfällt nur die Berechnung für d_0), ist hier nur das Programm „*pwertdneu*“ für den 1.PS als Beispielprogramm aufgeführt.

```

pwertdneu_function(out)
{
  n_25
  a0_seq(0.02,1.4,0.02)
  m_length(a0)
  Phi_matrix(0,ncol=20,nrow=m)
  D0_matrix(0,ncol=20,nrow=m)
  for (k in (1:20)){
    for (j in (1:m)){
      Phi[j,k]_pnorm((out$ahut[k]-a0[j])/out$sigmahut[k])
      D0[j,k]_1+a0[j]*max(out$muhut[,k])
    }
  }
  yr_range(0,1)
  dr_range(D0)
  phi1_Phi[,1]
  plot(D0[,1],phi1,xlab="d0",ylab="p-Wert",xlim=dr/2,ylim=yr,type="l",lty=2)
  box(n=1)
  title(cex=0.8,main="1.Parametersatz, d=2.41, range mu klein, n=25")
  for (k in (2:20)){
    phi_Phi[,k]
    par(new=T)
    plot(D0[,k],phi,xlab="",ylab="",main="",xlim=dr/2,ylim=yr,type="l",lty=2)
    abline(0.05,0,lty=1)
    abline(0.95,0,lty=1)
    abline(v=2.41,lty=4)
    abline(v=1.63,lty=1)
    abline(v=2.63,lty=1)
  }
  #return(Phi)
}

```

Literaturverzeichnis

- [1] P.J. Bickel, K.A. Doksum (1977): *Mathematical Statistics: basic ideas and selected topics*, Holden-Day
- [2] A.C. Cameron, P.K. Trivedi (1998): *Regression analysis of count data*, Cambridge University Press
- [3] G. Casella, R.L. Berger (1990): *Statistical Inference*, Wadsworth & Brooks/Cole
- [4] C. Czado, A. Munk (2000): *Noncanonical links in generalized linear models - when is the effort justified?*, Journal of Statistical Planning and Inference, 87, 317-345
- [5] L. Fahrmeir, H. Kaufmann (1985): *Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models*, Annals of Statistic 13, 342-368
- [6] L. Fahrmeir, R. Künstler, I. Pigeot, G. Tutz (1999): *Statistik, der Weg zur Datenanalyse*, zweite verbesserte Auflage, Springer-Verlag
- [7] L. Fahrmeir, G. Tutz (1994): *Multivariate statistical modelling based on generalized linear models*, New York, Springer-Verlag
- [8] W. Feller (1968): *An introduction to probability theory and its applications*, Vol 1, 3rd edition, New York, John Wiley & Sons
- [9] A. Franzmann (2001): *Multivariate Klassifikation in der Kraftfahrzeughaftpflichtversicherung*, Diplomarbeit an der Johannes Gutenberg-Universität Mainz
- [10] N.L. Johnson, S. Kotz, A.W. Kemp (1993): *Univariate discrete distributions*, 2nd edition, New York, John Wiley & Sons
- [11] W. Krabs (1983): *Einführung in die lineare und nichtlineare Optimierung*, Stuttgart, Teubner-Verlag

-
- [12] C. Kredler (1998): *Einführung in die Wahrscheinlichkeitsrechnung und Statistik*, TU-München
 - [13] J.F. Lawless (1987): *Negative binomial and mixed Poisson regression*, The Canadian Journal of Statistics, Vol 15, No.3, 209-225
 - [14] E.L. Lehmann (1986): *Testing statistical hypotheses*, 2nd edition, New York, John Wiley & Sons
 - [15] P. McCullagh, J.A. Nelder (1989): *Generalized linear models*, 2nd edition, Chapman and Hall
 - [16] H. Pruscha (1989): *Angewandte Methoden der mathematischen Statistik*, Stuttgart, Teubner-Verlag
 - [17] L. Rade, B. Westergren (1997): *Springers Mathematische Formeln*, 2.Auflage, Springer-Verlag
 - [18] S.I. Resnick (1992): *Adventures in stochastic processes*, Boston, Birkhäuser
 - [19] R. Winkelmann (1996): *Econometric analysis of count data*, 2nd reviewed and enlarged edition, Springer-Verlag

Abbildungsverzeichnis

2.1	Verschiedene NB-Verteilungen	32
3.1	Signal-to-noise-Quotient für NB-Verteilung	75
3.2	Konturenplot des Signal-to-noise-Quotienten bei der NB-Verteilung . .	76
3.3	Grenzverhalten der NB-Verteilungsfunktion $P(Y \leq j)$ für $Y \sim NB(\mu, a)$	79
3.4	p-Wert-Kurven für die Testprobleme H gegen K und K gegen H . . .	81
3.5	Allgemeine p-Wert-Kurve	82
3.6	Abstandsmassfunktion $d(a)$	84
4.1	Powerfunktionen für Test (3.17) mit $\alpha = 0.05$ (gestrichelte Linie) und Test (4.1) mit $\alpha = 0.1$ (durchgezogene Linie)	89
4.2	Powerfunktionen für Test (3.17) mit $\alpha = 0.1$ (gestrichelte Linie) und Test (4.1) mit $\alpha = 0.2$ (durchgezogene Linie)	90
4.3	Powerfunktionen für Test (3.17) mit $\alpha = 0.15$ (gestrichelte Linie) und Test (4.1) mit $\alpha = 0.3$ (durchgezogene Linie)	91
4.4	p-Wert-Kurven der Parametersätze 1-4 bezüglich a_0	97
4.5	p-Wert-Kurven der Parametersätze 5-8 bezüglich a_0	98
4.6	p-Wert-Kurven der Parametersätze 9-12 bezüglich a_0	99
4.7	p-Wert-Kurven der Parametersätze 13-16 bezüglich a_0	100
4.8	p-Wert-Kurven der Parametersätze 1-4 bezüglich d_0	107
4.9	p-Wert-Kurven der Parametersätze 5-8 bezüglich d_0	108
4.10	p-Wert-Kurven der Parametersätze 9-12 bezüglich d_0	109
4.11	p-Wert-Kurven der Parametersätze 13-16 bezüglich d_0	110
5.1	Explorative Datenanalyse der Kfz-Daten	120
5.2	Devianzresiduen des Poissonmodells (5.1) der Kfz-Daten	124
5.3	Linkspezifikation im Poissonmodell der Kfz-Daten	125

5.4	Überdispersion im Poissonmodell der Kfz-Daten	126
5.5	Devianzresiduen des NB-Modells (5.2) der Kfz-Daten	127
5.6	p-Wert-Funktionen für schwedische Kfz-Daten	129
5.7	geschätzter Erwartungswert für S im NB-Modell (5.2)	130
5.8	p-Wert-Funktionen mit neuer Abstandsmassfunktion $d_0^{Quantil}$	132
5.9	Explorative Datenanalyse der Patent-Daten	134
5.10	Devianzresiduen des Poissonmodells (5.4) der Patent-Daten	136
5.11	Linkspezifikation im Poissonmodell (5.4) der Patent-Daten	137
5.12	Überdispersion im Poissonmodell (5.4) der Patent-Daten	138
5.13	Devianzresiduen des NB-Modells (5.5) der Patent-Daten	140
5.14	p-Wert-Funktionen für Patent-Daten	141
5.15	geschätzter Erwartungswert für Patente im NB-Modell (5.5)	142
5.16	p-Wert-Funktionen mit neuer Abstandsmassfunktion $d_0^{Quantil}$	143

Tabellenverzeichnis

4.1	Parametersätze für die Simulationsstudie	88
4.2	Ergebnisse der p-Wert-Kurven bezüglich a_0 , d.h. $\hat{P}(\hat{a}, \hat{\sigma}, a_0) = \Phi\left(\frac{\hat{a}-a_0}{\hat{\sigma}(\hat{\beta}, \hat{a})}\right)$ 101	
4.3	Ergebnisse der p-Wert-Kurven bezüglich d_0 , d.h. $\hat{P}(\hat{a}, \hat{\sigma}, d_0) = \Phi\left(\frac{\hat{a}-d^{-1}(d_0)}{\hat{\sigma}(\hat{\beta}, \hat{a})}\right)$	111
5.1	<i>summary</i> -Tabelle des Poissonmodells (5.1) der Kfz-Daten	122
5.2	<i>summary</i> -Tabelle des NB-Modells (5.2) der Kfz-Daten	128
5.3	<i>summary</i> -Tabelle des Poissonmodells (5.4) der Patent-Daten	136
5.4	<i>summary</i> -Tabelle des NB-Modells (5.5) der Patent-Daten	139