M A S T E R A R B E I T

Die Extended-Quasi-Likelihood-Funktion in Generalisierten Linearen Modellen

aus der Studienrichtung
Technische Mathematik, Operations Research und Statistik

ausgeführt am
Institut für Statistik der Technischen Universität Graz

betreut von
Ao. Univ.-Prof. DI Dr.techn. Herwig Friedl

durch
Thorn Thaler, Bakk.techn.
Matr.-Nr. 9930585

Graz, 7. September 2009

Zusammenfassung

Das *Generalisierte Lineare Modell (GLM)* ist eine Erweiterung des *Linearen Modells* und erlaubt auch andere Verteilungsannahmen als die der Normalverteilung. Da die Schätzung der Koeffizienten eines solchen Modells nur von den ersten beiden Momenten abhängt, stellt die Einführung der *Quasi-Likelihood-Funktion (QL-Funktion)* eine natürliche Verallgemeinerung dar. Dabei wird nicht mehr die vollständige Verteilung spezifiziert, sondern nur der funktionale Zusammenhang zwischen dem Erwartungswert μ und der Varianz, wie er durch die Varianzfunktion $\mathrm{var}(Y) \propto V(\mu)$ ausgedrückt wird, angegeben. Nelder und Pregibon (1987) führen die *Extended-Quasi-Likelihood-Funktion (EQL-Funktion)* ein, die es u. a. nun auch ermöglicht, unterschiedliche Varianzfunktionen miteinander zu vergleichen. Bettet man die Varianzfunktion in eine parametrische Varianzfamilie ein, ist es mit diesem Ansatz möglich, die Parameter der Familie zu bestimmen.

Diese Arbeit beschäftigt sich im folgenden mit der Theorie der EQL-Idee. Dabei wird der Bogen vom GLM, zu den allgemeinen Eigenschaften der EQL-Funktion einschließlich des Zusammenhangs mit der Sattelpunkt-Approximation gespannt und mündet schließlich in der Implementierung eines R Pakets zur Schätzung der Parameter einer Familie von Varianzfunktionen mittels des EQL-Ansatzes.

Abstract

The *Generalized Linear Model (GLM)* is an extension of the *Linear Model*. As opposed to a Linear Model, a GLM does not only allow for the usage of the normal distribution, but for all distributions from the exponential family. Estimators of the coefficients of a GLM are determined completely by the specification of the first two moments. Thus, a natural extension is to use a *quasi-likelihood* approach, where one does not specify the whole distribution, but the functional relationship between the mean μ and the variance, as expressed by the variance function $\text{var}(Y) \propto V(\mu)$. Nelder und Pregibon (1987) introduce the *Extended Quasi-Likelihood function (EQL function)* which, inter alia, permits comparisons between different variance functions.

In this thesis we will elaborate on the theory of EQL. Starting from the properties of a GLM, we will discuss the characteristics of the EQL approach including the close relationship between the EQL function and the saddlepoint approximation. An R library which reduces the introduced methods to practice completes the work.

Ich erkläre, dass ich diese Arbeit selbst verfasst, alle verwendeten Quellen zitiert und mich keiner unerlaubten Hilfsmittel bedient habe.

Graz, 7. September 2009 Thorn Thaler

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Quellcodeverzeichnis

Abkürzungsverzeichnis

AIC	Akaike-Informationskriterium (engl. *Akaike's information criterion*, ursprünglich *An information criterion*)
ANOVA	Analysis-of-Variance
BIC	Bayessche Informationskriterium (engl. *Bayesian information criterion*)
BLUE	Best Linear Unbiased Estimator
EQL	Extended-Quasi-Likelihood
GLM	Generalisiertes Lineares Modell
IWLS	Iterative Weighted Least-Squares
LM	Lineares Modell
LS	Least-Squares
ML	Maximum-Likelihood
PL	Pseudo-Likelihood
QL	Quasi-Likelihood
REML	Restricted-Maximum-Likelihood (auch Residual-Maximum-Likelihood)
SSE	Fehlerquadratsumme (engl. *Sum of Squared Errors*)
SSR	Residuenquadratsumme (engl. *Sum of Squared Residuals*)
SST	Gesamtquadratsumme (engl. *Total Sum of Squares*)
WLS	Weighted Least-Squares

1 Einleitung

Diese Arbeit beschäftigt sich mit dem Einsatz der *Extended-Quasi-Likelihood-Funktion* in Generalisierten Linearen Modellen (GLM). Generalisierte Lineare Modelle gehen auf Nelder und Wedderburn (1972) zurück und werden dazu verwendet, um ausgehend von einem Satz von erklärenden Variablen, den Erwartungswert einer ausgezeichneten Größe (der Response-Variable) zu modellieren. Sie stellen dabei eine Verallgemeinerung der Linearen Modelle in zweierlei Hinsicht dar:

1. Die Verteilungsannahme wird insofern erweitert, dass anstelle von der Annahme einer Normalverteilung auch Verteilungen aus der Exponentialfamilie Berücksichtigung finden können. Dadurch wird die funktionale Beziehung zwischen Erwartungswert und Varianz implizit festgelegt.

2. Der Erwartungswert wird nicht direkt linear modelliert, sondern eine Funktion desselben wird als Linearkombination angesetzt.

Um die Verteilungsannahme selbst zu erweitern, werden *Quasi-Likelihood-Funktionen* (Wedderburn, 1974) verwendet, die anstatt von einer vollen Verteilungsannahme nur von den ersten beiden Momenten und deren funktionalem Zusammenhang ausgehen.

Mit dem Quasi-Likelihood-Ansatz ist es allerdings nicht möglich, unterschiedliche Varianzfunktionen miteinander zu vergleichen. Daher wird dieses Konzept ein weiteres Mal erweitert und führt uns zu der Verwendung der auf Nelder und Pregibon (1987) zurückgehenden *Extended-Quasi-Likelihood-Funktion*, die eben diese Vergleiche zulässt. Die simultane Modellierung von Erwartungswert und Dispersion ist mit diesem Ansatz ebenfalls möglich.

Es stellt sich heraus, dass die Extended-Quasi-Likelihood-Funktion auch aus der *Sattelpunkt-Approximation* hergeleitet werden kann, die eine Approximation der Dichte des arithmetischen Mittels einer Folge von Zufallsvariablen darstellt. Die Sattelpunkt-Approximation ist dabei eine Erweiterung der Edgeworth-Approximation und zeichnet sich dadurch aus, dass sie nicht nur im Zentrum gute Annäherungen liefert.

Der Rest der Arbeit gliedert sich wie folgt: Das Kapitel 2 beschäftigt sich mit der Theorie der Generalisierten Linearen Modelle, um den allgemeinen Rahmen für die weitere Arbeit abzustecken. Dabei wird der zentrale Begriff der Exponentialfamilie eingeführt und insbesondere die Maximum-Likelihood-Schätzung der Modellparameter besprochen.

In Kapitel 3 wird das Generalisierte Lineare Modell vorerst durch die Einführung der Quasi-Likelihood-Funktion erweitert, um auch Erwartungswert-Varianz-Strukturen untersuchen zu können, die keiner Verteilung aus der Exponentialfamilie zuzuordnen sind. Die Quasi-Likelihood-Funktion wird dann zu der Extended-Quasi-Likelihood-Funktion erweitert, um auch verschiedene Varianzfunktionen berücksichtigen zu können. Außerdem wird die gemeinsame Modellierung von Erwartungswert und Dispersion behandelt.

Die Abhandlung verschiedener Informationskriterien bzw. der Fehlerrate schließt dieses Kapitel ab.

Das Kapitel 4 befasst sich vorerst mit der Edgeworth-Approximation der Dichte der (gewichteten) Summe von Zufallsvariablen. Die Schwächen dieser Approximation vor allem in den Schwänzen einer Verteilung führt zu einer verbesserten Variante, der Sattelpunkt-Approximation. Diese beruht auf der Verwendung einer exponentiell verschobenen Dichte. Die spezielle Form der Sattelpunkt-Approximation für Verteilungen aus der Exponentialfamilie selbst, sowie der Zusammenhang mit der Extended-Quasi-Likelihood-Funktion werden ebenfalls behandelt.

Das Kapitel 5 beschreibt die Implementierung der EQL-Funktion für das Statistik-Programm R und erläutert die Verwendung anhand einiger Beispiele.

Das Kapitel 6 fasst die gewonnenen Ergebnisse zusammen und gibt einen Ausblick über mögliche weiterführende Ansätze.

Im Anhang findet sich schließlich ein kurzer Abriss über die Theorie der Linearen Modelle, sowie die Dokumentation des R-Codes der beiden entwickelten R-Pakete *ttutils* und *EQL*.

Zu guter Letzt obliegt mir die angenehme Pflicht, Herrn Univ.-Prof. DI Dr. Friedl für die Unterstützung und die vielen wertvollen Anregungen meinen aufrichtigen Dank auszusprechen. Auch möchte ich Frau Djuras danken, die mir zielführende Hinweise bei der Lösung einiger Integrale geben konnte.

2 Das Generalisierte Lineare Modell

2.1 Vom Linearen Modell zum Generalisierten Linearen Modell

Das klassische Lineare Modell (LM)[1] setzt voraus, dass die n Response-Variablen Y_i unabhängig normalverteilt sind und die Varianz $\text{var}(Y_i) = \sigma^2$ konstant ist. Es modelliert dann die Erwartungswerte $\mathbb{E}(Y_i) = \mu_i$ als Linearkombination der erklärenden Variablen $\boldsymbol{x}_i \in \mathbb{R}^p$:

$$\mu_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}, \tag{2.1}$$

wobei $\boldsymbol{\beta}$ den p-dimensionalen Vektor der Modellparameter bezeichnet. Diese restriktiven Annahmen sind in der Praxis oft nicht erfüllt, man stelle sich z. B. ganzzahlige Response-Variablen vor. Allein die Tatsache, dass es sich bei der zugrunde liegenden Verteilung um eine diskrete handelt, lässt die Annahme einer Normalverteilung nicht zu. Auch die Voraussetzung der konstanten Varianz ist nicht immer haltbar.

Ein möglicher Ansatz um dieses Problem zu umgehen, stellt die Verwendung von geeigneten Transformationen dar. Bei der *Box-Cox-Transformation* (Box und Cox, 1964) werden die Y_i derart transformiert, dass sie approximativ normalverteilt sind und eine konstante Varianz aufweisen.

Das Generalisierte Lineare Modell (GLM) geht einen anderen Weg, indem es – anstelle die Daten zu transformieren – die Annahmen des LM verallgemeinert. Anstatt sich auf normalverteilte Y_i zu beschränken, erlaubt man Response-Variablen, die einer Verteilung aus der *Exponentialfamilie* folgen. Außerdem wird die strikte Forderung einer konstanten Varianz zugunsten einer *Varianzfunktion*, die eine Abhängigkeitsstruktur zwischen dem Erwartungswert μ und der Varianz $\text{var}(Y) \propto V(\mu)$ erlaubt, aufgegeben. Schließlich ermöglicht das GLM im Gegensatz zum klassischen LM, das den Erwartungswert selbst durch eine Linearkombination der Prädiktor-Variablen darstellt (wie in Gleichung (2.1)), auch die Modellierung einer *Funktion* des Erwartungswertes (Linkfunktion).

Die folgende Diskussion über die Eigenschaften von Generalisierten Linearen Modellen orientiert sich an McCullagh und Nelder (1989).

2.1.1 Die Exponentialfamilie

Zuerst wollen wir die Verteilungsannahme erweitern, um auch Modelle betrachten zu können, deren Response-Variablen nicht normalverteilt sind. Dazu führen wir zunächst den Begriff der Exponentialfamilie ein (vgl. Lehmann und Romano, 2004).

[1] Die Theorie des LM wird in Anhang A besprochen.

Definition 2.1 (k-parametrische Exponentialfamilie)**.** Lässt sich für den Parametervektor $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^\top$ eine Dichte in der Form

$$f_Y(y, \boldsymbol{\theta}) = c(\boldsymbol{\theta}) \exp\left\{ \sum_{j=1}^{k} b_j(\boldsymbol{\theta}) T_j(y) \right\} h(y) \tag{2.2}$$

mit bekannten Funktionen $c(\,\cdot\,), b_j(\,\cdot\,), T_j(\,\cdot\,)$ und $h(\,\cdot\,)$ schreiben, so ist f_Y eine Dichte einer Verteilung aus der k-parametrischen Exponentialfamilie. Dabei bezeichnet $\boldsymbol{\theta}$ den Parametervektor der Exponentialfamilie.

Bemerkung 1. Definition 2.1 ist äquivalent zu der Forderung, dass die Dichte f_Y folgende Form hat:

$$f_Y(y, \boldsymbol{\theta}) = \exp\left\{ \tilde{c}(\boldsymbol{\theta}) + \sum_{j=1}^{k} b_j(\boldsymbol{\theta}) T_j(y) + \tilde{h}(y) \right\}.$$

Bemerkung 2. Spezielle Parameter der Verteilung, die nicht näher von Interesse sind, bezeichnet man als *Nuisance-Parameter*.

Bemerkung 3. Gilt zusätzlich $b_j(\boldsymbol{\theta}) = \theta_j$ dann spricht man von einer *Exponentialfamilie in kanonischer Form*.

Bemerkung 4. Aus (2.2) folgt mit dem Faktorisierungssatz von Neyman unmittelbar, dass die Statistik $\boldsymbol{T} = (T_1(y), \ldots, T_k(y))^\top$ suffizient für $\boldsymbol{\theta}$ ist.

Kann man einen Parameter ϕ als bekannten Nuisance-Parameter betrachten, geben McCullagh und Nelder (1989) eine alternative Formulierung für die einparametrische Exponentialfamilie mit kanonischem Parameter θ.

Definition 2.2 (1-parametrische lineare kanonische Exponentialfamilie)**.** Als einparametrische lineare Exponentialfamilie bezeichnet man für ein bekanntes ϕ eine Menge von Wahrscheinlichkeitsverteilungen, die eine Dichtefunktion der Form

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

besitzen, wobei die Funktionen $a(\,\cdot\,)$, $b(\,\cdot\,)$ und $c(\,\cdot\,)$ bekannt sind. Der Parameter θ bezeichnet dabei den *kanonischen Parameter*.

Bemerkung. Sollte ϕ unbekannt sein, kann Definition 2.2 als Definition für eine 2-parametrische Exponentialfamilie aufgefasst werden.

Um die Momente einer Verteilung aus der Exponentialfamilie zu bestimmen, betrachten wir die allgemeinen Eigenschaften der Score-Funktion.

Satz 2.3 (Eigenschaften der Score-Funktion)**.** *Es bezeichne $f_Y(y, \theta)$ eine beliebige differenzierbare Dichtefunktion mit Parameter θ. Dann gilt für den Erwartungswert und die Varianz der Score-Funktion:*

$$\begin{aligned} \mathbb{E}\left(\frac{\partial \log f_Y(y, \theta)}{\partial \theta} \right) &= 0, \\ \mathbb{E}\left(\frac{\partial \log f_Y(y, \theta)}{\partial \theta} \right)^2 &= \mathbb{E}\left(-\frac{\partial^2 \log f_Y(y, \theta)}{\partial \theta^2} \right). \end{aligned} \tag{2.3}$$

Beweis. Für die Ableitung der log-Likelihood-Funktion gilt

$$\frac{\partial \log f_Y(y,\theta)}{\partial \theta} = \frac{1}{f_Y(y,\theta)} \frac{\partial f_Y(y,\theta)}{\partial \theta} \tag{$*$}$$

und da $f_Y(y,\theta)$ eine Dichte ist, gilt außerdem

$$\int_{\mathbb{R}} f_Y(y,\theta) dy = 1. \tag{$**$}$$

Damit folgt dann

$$\begin{aligned}
\mathbb{E}\left(\frac{\partial \log f_Y(y,\theta)}{\partial \theta}\right) &\overset{(*)}{=} \mathbb{E}\left(\frac{1}{f_Y(y,\theta)} \frac{\partial f_Y(y,\theta)}{\partial \theta}\right) \\
&= \int_{\mathbb{R}} \frac{1}{f_Y(y,\theta)} \frac{\partial f_Y(y,\theta)}{\partial \theta} \cdot f_Y(y,\theta) dy \\
&= \frac{\partial}{\partial \theta} \underbrace{\int_{\mathbb{R}} f_Y(y,\theta)\, dy}_{\overset{(**)}{=} 1} \\
&= 0.
\end{aligned}$$

Andererseits gilt für die zweite Ableitung von $\log f_Y(y,\theta)$

$$\frac{\partial^2 \log f_Y(y,\theta)}{\partial \theta^2} = -\frac{1}{f_Y(y,\theta)^2} \left(\frac{\partial f_Y(y,\theta)}{\partial \theta}\right)^2 + \frac{1}{f_Y(y,\theta)} \frac{\partial^2 f_Y(y,\theta)}{\partial \theta^2}$$

und damit folgt für den Erwartungswert

$$\begin{aligned}
\mathbb{E}\left(-\frac{\partial^2 \log f_Y(y,\theta)}{\partial \theta^2}\right) &= \mathbb{E}\left(\frac{1}{f_Y(y,\theta)^2} \left(\frac{\partial f_Y(y,\theta)}{\partial \theta}\right)^2 - \frac{1}{f_Y(y,\theta)} \frac{\partial^2 f_Y(y,\theta)}{\partial \theta^2}\right) \\
&= \int_{\mathbb{R}} \frac{f_Y(y,\theta)}{(f_Y(y,\theta))^2} \left(\frac{\partial f_Y(y,\theta)}{\partial \theta}\right)^2 dy - \int_{\mathbb{R}} \frac{f_Y(y,\theta)}{f_Y(y,\theta)} \frac{\partial^2 f_Y(y,\theta)}{\partial \theta^2} dy \\
&\overset{(*)}{=} \int_{\mathbb{R}} \left(\frac{\partial \log f_Y(y,\theta)}{\partial \theta}\right)^2 f_Y(y,\theta)\, dy - \frac{\partial^2}{\partial \theta^2} \underbrace{\int_{\mathbb{R}} f_Y(y,\theta)\, dy}_{\overset{(**)}{=} 1} \\
&= \mathbb{E}\left(\frac{\partial \log f_Y(y,\theta)}{\partial \theta}\right)^2.
\end{aligned}$$

□

Bemerkung. Im Beweis von Satz 2.3 werden Integral und Differentiation vertauscht. Diese Vertauschung lässt sich im Speziellen dann rechtfertigen, wenn es für eine in θ differenzierbare Funktion $f(y,\theta)$ eine Funktion $g(y,\theta)$ und eine Konstante $\delta_0(\theta) > 0$ gibt, für die

$$\begin{gathered}
\int_{-\infty}^{\infty} g(y,\theta)\, dy < \infty, \\
\left|\frac{\partial}{\partial \theta} f(y,\theta)\Big|_{\theta=\theta'}\right| \leq g(y,\theta) \quad \text{für alle } \theta' \text{ mit } |\theta' - \theta| \leq \delta_0(\theta)
\end{gathered} \tag{2.4}$$

gilt (vgl. Casella und Berger, 2002, S. 70). In der weiteren Diskussion wollen wir diese Regularitätsbedingung voraussetzen.

Damit lassen sich die ersten beiden Momente einer Verteilung aus der Exponentialfamilie leicht bestimmen.

Satz 2.4 (Momente der Exponentialfamilie)**.** *Ist Y eine Zufallsvariable mit einer Dichtefunktion $f_Y(y, \theta)$ aus der Exponentialfamilie, dann gilt für die ersten beiden Momente:*

$$\begin{aligned} \mathbb{E}(Y) &= b'(\theta), \\ \operatorname{var}(Y) &= a(\phi)b''(\theta). \end{aligned}$$

Beweis. Dies ist eine unmittelbare Folgerung aus Satz 2.3:

$$\mathbb{E}\left(\frac{\partial \log f_Y(y,\theta)}{\partial \theta}\right) = \mathbb{E}\left(\frac{Y - b'(\theta)}{a(\phi)}\right) \overset{(2.3)}{=} 0 \Rightarrow \mathbb{E}(Y) = b'(\theta)$$

und

$$\begin{aligned} \mathbb{E}\left(\frac{\partial \log f_Y(y,\theta)}{\partial \theta}\right)^2 &\overset{(2.3)}{=} -\mathbb{E}\left(\frac{\partial^2 \log f_Y(y,\theta)}{\partial \theta^2}\right) \\ \overbrace{\frac{\mathbb{E}\left(Y - b'(\theta)\right)^2}{a(\phi)^2}}^{=\operatorname{var}(Y)} &= \frac{b''(\theta)}{a(\phi)}, \\ \operatorname{var}(Y) &= a(\phi)b''(\theta). \end{aligned}$$

□

Um die Abhängigkeit der Varianz vom Erwartungswert $\mu := \mathbb{E}(Y)$ zu verdeutlichen, führt man die *Varianzfunktion* $V(\mu)$ ein und erhält somit für das zweite Moment

$$\operatorname{var}(Y) = a(\phi)b''(\theta) = a(\phi)\frac{\partial b'(\theta)}{\partial \theta} = a(\phi)\frac{\partial \mu}{\partial \theta} =: a(\phi)V(\mu), \tag{2.5}$$

wobei $a(\phi)$ unabhängig vom Erwartungswert μ ist. Der Parameter ϕ bezeichnet dabei den *Dispersionsparameter*. Liegen mehrere unabhängige Beobachtungen y_i vor, kann der Fall auftreten, dass die Funktionen $a_i(\phi)$ die Form

$$a_i(\phi) = w_i \cdot \phi$$

aufweisen. Dabei ist ϕ konstant für alle Beobachtungen und nur die a-priori bekannten Gewichte w_i unterscheiden sich von Beobachtung zu Beobachtung.

Die Funktion $b(\theta)$ nennt man die *Kumulantenfunktion*; der Name Kumulantenfunktion ist der Eigenschaft geschuldet, dass die k-te Kumulante $\kappa_k(\theta)$ durch

$$\kappa_k(\theta) = a(\phi)^{k-1}\frac{\partial^k b(\theta)}{\partial \theta^k} \tag{2.6}$$

gegeben ist, wie man leicht mittels der Identitäten

$$M(t) = \mathbb{E}\left(e^{ty}\right), \qquad \text{(Momenterzeugende)}$$
$$K(t) = \log M(t), \qquad \text{(Kumulantenerzeugende)}$$
$$\kappa_k(\theta) = \left.\frac{\partial^k K(t)}{\partial t^k}\right|_{t=0} \qquad \text{(Kumulante)}$$

überprüfen kann:

$$\begin{aligned}
M(t) &= \mathbb{E}\left(e^{ty}\right) \\
&= \int_{\mathbb{R}} \exp(ty) \cdot \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\} dy \\
&= \exp\left\{-\frac{b(\theta)}{a(\phi)}\right\} \exp\left\{\frac{b\left(\theta + t\cdot a(\phi)\right)}{a(\phi)}\right\} \cdot \\
&\quad \underbrace{\int_{\mathbb{R}} \exp\left\{\frac{y\cdot(\theta + t\cdot a(\phi)) - b\left(\theta + t\cdot a(\phi)\right)}{a(\phi)} + c(y,\phi)\right\} dy}_{=\int_{\mathbb{R}} \exp\{\frac{y\tilde{\theta} - b(\tilde{\theta})}{a(\phi)} + c(y,\phi)\}\, dy = 1} \\
&= \exp\left(\frac{b\left(\theta + t\cdot a(\phi)\right) - b(\theta)}{a(\phi)}\right), \qquad (2.7a)
\end{aligned}$$

$$K(t) = \frac{b\left(\theta + t\cdot a(\phi)\right) - b(\theta)}{a(\phi)}, \qquad (2.7b)$$

$$\kappa_k(\theta) = a(\phi)^{k-1}\frac{\partial^k b(\theta)}{\partial \theta^k}. \qquad (2.7c)$$

Zusammenfassend lässt sich festhalten, dass das GLM durch den Einsatz einer Verteilung aus der Exponentialfamilie nicht wie das LM ausschließlich auf die konstante Varianzfunktion $V(\mu) = 1$ beschränkt ist, sondern eine Variabilität erlaubt, die vom Erwartungswert μ abhängt.

2.1.2 Die Linkfunktion

Die zweite wesentliche Änderung betrifft die Modellierung des Erwartungswertes. Während das LM den Erwartungswert μ direkt durch eine Linearkombination der Prädiktoren $\boldsymbol{x}$ beschreibt, verwendet das GLM eine *Linkfunktion* und modelliert eine Funktion des Erwartungswertes μ:

$$\eta := g(\mu) = \boldsymbol{x}^\top \boldsymbol{\beta}.$$

Diese Verallgemeinerung trägt der Tatsache Rechnung, dass der Erwartungswert für Zufallsvariablen, die bestimmten Einschränkungen hinsichtlich ihres Wertebereiches unterliegen (die z.B keine negativen Werte annehmen), auch diese Einschränkungen widerspiegeln sollte. Würde man beispielsweise für strikt positive Zufallsvariablen den Erwartungswert direkt wie in Formel (2.1) modellieren, schließt man im speziellen negative Erwartungswerte nicht dezidiert aus.[2]

[2] Sind die Parameter für ein Modell einmal geschätzt, kann eine „ungünstige“ Kombination von Kovariaten zu einer Vorhersage eines negativen Erwartungswertes führen.

Die Linkfunktion hat also sicherzustellen, dass die modellierten Erwartungswerte in einem plausiblen Bereich bleiben und modelliert so die Struktur der zugrunde liegenden Zufallsvariablen. Die Linkfunktion ist im Prinzip eine beliebige streng monotone und differenzierbare Funktion. Die Existenz der Inversen, die für die Berechnung des Erwartungswertes $\mu = g^{-1}(\eta)$ benötigt wird, folgt dabei aus der strengen Monotonie und der Stetigkeit.

Eine spezielle Bedeutung kommt der sogenannten *kanonischen Linkfunktion* zu. Dabei wird der Zusammenhang zwischen dem kanonischen Parameter θ und dem linearen Prädiktor $\eta = \boldsymbol{x}^\top \boldsymbol{\beta}$ – schematisch in (2.8a) dargestellt – dadurch vereinfacht, dass man wie in (2.8b) den kanonischen Parameter und den linearen Prädiktor gleichsetzt.

$$\theta \xleftarrow[\text{Transformation}]{b'(\theta)=\mu} \mu \xrightarrow[\text{Transformation}]{g(\mu)=\eta} \eta \tag{2.8a}$$

$$\theta \qquad = \qquad \eta. \tag{2.8b}$$

Der kanonische Link ist u. a. deshalb von besonderem Interesse, da für ihn eine minimal suffiziente Statistik für den Vektor der Regressionsparameter existiert und die Maximum-Likelihood (ML)-Schätzung vereinfacht wird.

Die Wahl der Linkfunktion ist problemabhängig und muss von Fall zu Fall untersucht werden. Die kanonische Linkfunktion erlaubt zwar eine einfache Interpretation der Regressionsparameter,[3] gewährleistet aber nicht automatisch die beste Anpassung an die Daten. Czado und Munk (2000) zeigen, wann der erhöhte Rechenaufwand für eine nicht kanonische Linkfunktion durch eine bessere Anpassung gerechtfertigt wird.

Beispiel 2.5 (Binomialverteilte Response-Variablen)**.** Wir betrachten den Fall, dass eine Zufallsvariable Y einen relativen Anteil beschreibt.[4] Dann ist das n-fache dieser Zufallsvariable binomialverteilt:

$$ny \sim \mathcal{B}(n, \mu).$$

Damit folgt für die Momente von ny:

$$\mathbb{E}(ny) = n\mu \qquad\qquad \operatorname{var}(ny) = n\mu(1-\mu),$$

oder alternativ für y selbst:

$$\mathbb{E}(y) = \mu \qquad\qquad \operatorname{var}(y) = \frac{1}{n}\mu(1-\mu).$$

Der Parameter μ beschreibt eine Wahrscheinlichkeit und liegt somit zwischen Null und Eins. Daher sollte eine Linkfunktion – oder eigentlich deren Inverse – gewährleisten, dass jede Linearkombination der Prädiktor-Variablen (die ja einen beliebigen reellen Wert annehmen kann) auf das Intervall $(0, 1)$ abgebildet wird.

[3] Man denke etwa an den Logit-Link – den kanonischen Link bei binomial-verteilten Zufallsgrößen. Durch ihn wird das *odds ratio* modelliert, das wiederum einfach zu interpretieren ist

[4] Eine ausführliche Diskussion der standardisierten Binomialverteilung findet man in Abschnitt 2.4.1.

Eine mögliche Wahl für eine Funktion, die das Intervall $(0, 1)$ auf die reellen Zahlen abbildet, wäre beispielsweise die Inverse einer Verteilungsfunktion und so verwendet man für binomialverteilte Response-Variablen als Linkfunktion u. a. die Inverse der Verteilungsfunktion der Standardnormalverteilung $g(\mu) = \Phi^{-1}(\mu)$, den sogenannten *Probit-Link*. Eine andere Wahl wird durch die *Logistische Verteilung* induziert, deren Verteilungsfunktion durch

$$F(x, \alpha, \beta) = \frac{1}{1 + e^{-(x-\alpha)/\beta}}$$

mit den Parametern α und $\beta > 0$ gegeben ist. Für die spezielle Parameterwahl $\alpha = 0$ und $\beta = 1$ folgt als Inverse der Verteilungsfunktion der *Logit-Link*, der gleichzeitig den kanonische Link der standardisierten Binomialverteilung darstellt:

$$g(\mu) = F^{-1}(\mu) = \log \frac{\mu}{1 - \mu}. \qquad \Diamond$$

Will man unterschiedliche Linkfunktionen miteinander vergleichen, bettet man die Linkfunktion in eine parametrische Familie von Linkfunktionen $\mathcal{F}_{\boldsymbol{\lambda}}(\mu)$ ein und schätzt zusätzlich zu den Modellparametern $\boldsymbol{\beta}$ den Parametervektor $\boldsymbol{\lambda}$ der Linkfunktion. So unterscheidet man beispielsweise für Response-Variablen mit positivem Erwartungswert die Potenzfamilie:

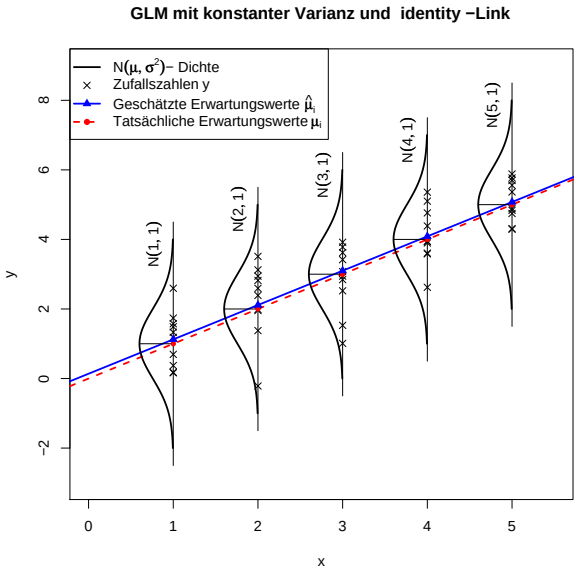
$$\mathcal{F}_\lambda(\mu) = \begin{cases} \mu^\lambda & \lambda \neq 0 \\ \log \mu & \lambda = 0 \end{cases},$$

oder speziell für binäre Response-Variablen eine Familie von asymmetrischen Linkfunktionen (vgl. Preisser, 2002):

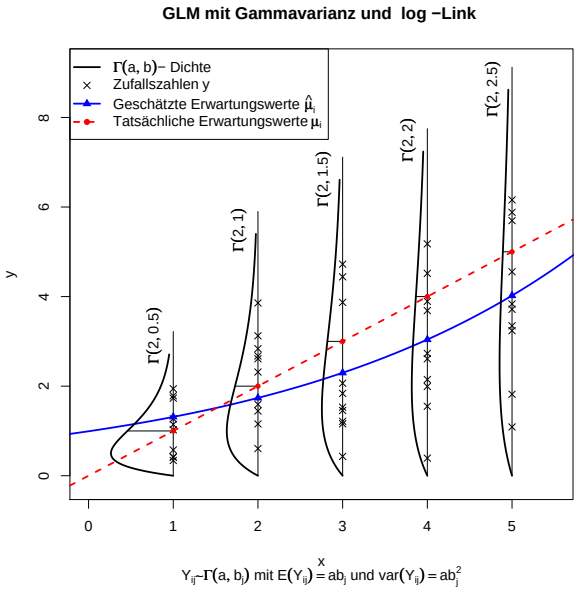
$$\mathcal{F}_\lambda(\mu) = \log \left\{ \frac{1}{\lambda} \left[\frac{1}{(1-\mu)^\lambda} - 1 \right] \right\}.$$

Man beachte, dass der Identitätslink $g(\mu) = \mu$ ein Mitglied der ersten und der Logit-Link $g(\mu) = \log \frac{\mu}{1-\mu}$ ein Mitglied der zweiten Familie darstellt (setze jeweils $\lambda = 1$). Eine ausführliche Diskussion über die Auswahl der Linkfunktion findet man bei Pregibon (1980).

Die Unterschiede zwischen einem LM und einem GLM werden in der Abbildung 2.1 verdeutlicht. Während das LM in Abbildung 2.1(a) auf einer Normalverteilungsannahme mit konstanter Varianz beruht und die Erwartungswerte linear modelliert, erweitert das GLM in Abbildung 2.1(b) diese Prämissen. Durch die Annahme einer Verteilung aus der Exponentialfamilie erlaubt es eine Abhängigkeitsstruktur zwischen dem Erwartungswert und der Varianz. Man erkennt eine deutliche Zunahme der Variabilität mit steigendem Erwartungswert – eine der Gammaverteilung inhärente Eigenschaft ($V(\mu) = \mu^2$). Die Varianz beim LM bleibt hingegen konstant. Da außerdem nicht die Erwartungswerte selbst sondern eine Funktion derselben modelliert wird, liegen beim GLM die geschätzten Erwartungswerte $\hat{\mu}_i$ nicht mehr auf einer Geraden – wie es beim LM der Fall ist – sondern auf einer Kurve, die der Inversen der Linkfunktion (in diesem Fall des log-Links) entspricht. Auffallend ist außerdem, dass trotz der gleichen Anzahl an Datenpunkten die Schätzung der Erwartungswerte beim LM wesentlich genauer ist, als es beim GLM der



(a) Generalisiertes Lineares Modell mit konstanter Varianz und Identitätslink (=LM)



(b) Generalisiertes Lineares Modell mit Gammavarianz und log-Link

Abbildung 2.1: Unterschiede zwischen einem Linearen und einem Generalisierten Linearen Modell

Fall ist. Dieser Umstand hängt natürlich auch damit zusammen, dass die gewählte Linkfunktion $g(\mu_i) = \log \mu_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} \Rightarrow \mu_i = \exp\left\{\boldsymbol{x}_i^\top \boldsymbol{\beta}\right\}$ einen exponentiellen Zusammenhang zwischen den erklärenden Variablen und den Erwartungswerten annimmt, während der tatsächliche Zusammenhang linearer Natur ist.

2.1.3 Beispiele für Verteilungen aus der Exponentialfamilie

Bevor wir uns der ML-Schätzung widmen, geben wir einige Beispiele für Verteilungen aus der Exponentialfamilie zusammen mit ihren kanonischen Linkfunktionen an.

Normalverteilung

Die Dichte einer normalverteilten Zufallsvariable $y \sim \mathcal{N}\left(\mu, \sigma^2\right)$ mit Erwartungswert μ und Varianz σ^2 ist durch

$$f(y, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\},\ y \in \mathbb{R}$$

gegeben und lässt sich schreiben als:

$$f(y, \mu, \sigma^2) = \exp\left\{\frac{y\mu - \mu^2/2}{\sigma^2} - \frac{y^2}{2\sigma^2} - \frac{1}{2}\log\left(2\pi\sigma^2\right)\right\}.$$

Mit der Parametrisierung $\theta = \mu$ und dem Nuisance-Parameter $\phi = \sigma^2$ ergibt das gerade eine Exponentialfamilie mit den Funktionen

$$a(\phi) = \phi, \qquad b(\theta) = \theta^2/2, \qquad c(y, \phi) = -\frac{y^2}{2\phi} - \frac{1}{2}\log\left(2\pi\phi\right).$$

Die ersten beiden Momente ergeben sich aus Satz 2.4 und die höheren Kumulanten aus (2.6) und lauten

$$\begin{aligned}\mathbb{E}\left(y\right) &= \theta = \mu,\\ \operatorname{var}\left(y\right) &= \phi = \sigma^2,\\ \kappa_k(\theta) &= 0,\ \forall k > 2.\end{aligned}$$

Damit ergeben sich für die Varianzfunktion $V(\mu)$ und die kanonische Linkfunktion $\eta = \theta$ folgende Identitäten:

$$V(\mu) = 1, \qquad \eta = \mu.$$

Gammaverteilung

Für eine Zufallsvariable y aus der Gammaverteilung $y \sim \Gamma(\mu, \nu)$ mit dem Erwartungswert $\mathbb{E}\left(y\right) = \mu$ und Varianz $\operatorname{var}\left(y\right) = \mu^2/\nu$ gilt für die Dichte

$$f(y, \mu, \nu) = \exp\left\{-\frac{\nu}{\mu}y\right\}\left(\frac{\nu}{\mu}\right)^\nu y^{\nu-1}\frac{1}{\Gamma(\nu)},\ y \in \mathbb{R}^+,$$

die sich zu

$$f(y,\mu,\nu) = \exp\left\{-\frac{\nu}{\mu}y + \nu\log\nu - \nu\log\mu + (\nu-1)\log y - \log\Gamma(v)\right\}$$
$$= \exp\left\{\frac{y\left(-\frac{1}{\mu}\right) + \log\frac{1}{\mu}}{1/\nu} + \nu\log\nu + (\nu-1)\log y - \log\Gamma(\nu)\right\}.$$

umformen lässt. Damit erhalten wir vermöge der Parametrisierung $\theta = -1/\mu$ und des Nuisance-Parameters $\phi = 1/\nu$ eine Exponentialfamilie mit den Funktionen

$$a(\phi) = \phi, \quad b(\theta) = -\log(-\theta), \quad c(y,\phi) = \frac{1}{\phi}\log\frac{1}{\phi} + \left(\frac{1}{\phi} - 1\right)\log y - \log\Gamma\left(\frac{1}{\phi}\right).$$

Die Momente und Kumulanten lassen sich wieder aus Satz 2.4 und Formel (2.6) bestimmen:

$$\mathbb{E}(y) = -\frac{1}{\theta} = \mu,$$
$$\operatorname{var}(y) = \phi\frac{1}{\theta^2} = \frac{1}{\nu}\mu^2,$$
$$\kappa_k(\theta) = (k-1)!\nu\left(\frac{\mu}{\nu}\right)^k, \forall k > 2.$$

Die Gammaverteilung weist also eine quadratische Varianzfunktion und eine reziproke kanonische Linkfunktion auf:

$$V(\mu) = \mu^2, \qquad \eta = \frac{1}{\mu}.$$

Poissonverteilung

Auch diskrete Verteilungen wie die Poissonverteilung passen in das Schema der Exponentialfamilie. Die Wahrscheinlichkeitsfunktion für eine Zufallsvariable $y \sim \mathcal{P}(\mu)$ aus der Poissonverteilung ist durch

$$f(y,\mu) = \frac{\mu^y}{y!}e^{-\mu},\ y \in \mathbb{N}_0$$

definiert und lässt sich in Form einer Verteilung aus der Exponentialfamilie schreiben:

$$f(y,\mu) = \exp\{y\log\mu - \mu - \log y!\}.$$

Mit $\theta = \log\mu$ und fixem $\phi = 1$ erhält man die Funktionen

$$a(\phi) = \phi, \qquad b(\theta) = e^\theta, \qquad c(y,\phi) = -\log y!$$

und für die Momente und die Kumulanten gilt dann:

$$\begin{aligned}\mathbb{E}(y) &= e^\theta = \mu,\\ \operatorname{var}(y) &= e^\theta = \mu,\\ \kappa_k(\theta) &= e^\theta = \mu,\ \forall k > 2.\end{aligned}$$

Damit ist die Varianzfunktion $V(\mu)$ der Poissonverteilung linear im Erwartungswert μ und als kanonische Linkfunktion ergibt sich der logarithmische Link:

$$V(\mu) = \mu, \qquad\qquad \eta = \log \mu.$$

Bemerkung. Für poissonverteilte Zufallsgrößen gilt also, dass die Varianz mit dem Erwartungswert wächst ($V(\mu) = \mu$). In praktischen Anwendungen ist dieser Zusammenhang aber oft nicht exakt erfüllt, da die Varianz der Daten größer ist als es das Modell zulassen würde. Das führt zu einer Varianz-Annahme bei der der Dispersionsparameter ϕ größer als Eins ist:

$$V(\mu) = \phi \cdot \mu, \text{ mit } \phi > 1.$$

Die Daten sind dann in diesem Fall nicht mehr poissonverteilt. Man spricht von einem *Überdispersionsmodell*.[5] Hinde und Demétrio (1998) zeigen, wie die Modellparameter in diesem Fall zu bestimmen sind.

2.2 Maximum Likelihood Schätzung

Damit haben wir alle Bausteine für das GLM erläutert und können nun selbiges einführen. Ein Generalisiertes Lineares Modell wird unter der Annahme der Existenz von $\mathbb{E}(Y_i)$ und $\operatorname{var}(Y_i)$ durch

$$\begin{aligned} Y_i &\overset{\text{ind.}}{\sim} F(y, \theta_i), \\ F &\in \text{Exponentialfamilie}, \end{aligned} \tag{2.9a}$$

$$\eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}, \tag{2.9b}$$

$$g(\mu_i) = \eta_i \tag{2.9c}$$

spezifiziert.[6] Dabei bezeichnet der Vektor $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$ eine Realisierung des Zufallsvektors $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$, dessen unabhängige Komponenten Y_i aus der gleichen Verteilung der einparametrischen Exponentialfamilie mit dem jeweiligen kanonischen Parameter θ_i stammen und die Momente

$$\mathbb{E}(Y_i) = \mu_i, \qquad\qquad \operatorname{var}(Y_i) = a(\phi_i)V(\mu_i)$$

[5] Dementsprechend heißen Modelle mit einem Dispersionsparameter $\phi < 1$ Unterdispersionsmodelle.

[6] Der Erwartungswert der Y_i hängt vom Parameter θ_i der Exponentialfamilie ab, um diesen Umstand zu verdeutlichen, schreibt man gelegentlich auch $\mu_i = \mu_i(\theta_i)$.

aufweisen. Der Vektor $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^\top$ bezeichnet den p-dimensionalen Vektor der bekannten Prädiktor-Variablen für die i-te Beobachtung. Die Menge der n Vektoren $\boldsymbol{x}_i$ wird zu der *Designmatrix* $X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\top \in \mathbb{R}^{n \times p}$ zusammengefasst.

Ausgehend von Modell (2.9) mit einer stochastischen (2.9a), einer systematischen Komponente (2.9b) und einer Linkfunktion (2.9c), sind wir vor allem daran interessiert, einen Schätzer für den Parametervektor $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ zu finden. Dazu bedienen wir uns der ML-Schätzung.

Die Y_i sind nach Voraussetzung unabhängig, daher ist die log-Likelihood-Funktion $\mathscr{L}$ der Stichprobe $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$ durch den Logarithmus des Produkts der Likelihood-Funktionen $\ell(\theta_i, y_i)$ der einzelnen y_i gegeben, wobei $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^\top$ den Vektor der n unbekannten Verteilungsparameter beschreibt:

$$\mathscr{L}(\boldsymbol{\theta}, \boldsymbol{y}) = \log \left\{ \prod_{i=1}^{n} \ell(\theta_i, y_i) \right\} = \sum_{i=1}^{n} \log \ell(\theta_i, y_i). \tag{2.10}$$

Da die Y_i einer Verteilung aus der Exponentialfamilie folgen, können wir die Definition der Dichte aus Definition 2.2 in (2.10) einsetzen und erhalten:

$$\mathscr{L}(\boldsymbol{\theta}, \boldsymbol{y}) = \sum_{i=1}^{n} \left(\frac{y_i \theta_i - b(\theta_i)}{a(\phi_i)} + c(y_i, \phi_i) \right), \tag{2.11}$$

dabei bezeichnet $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_n) \top$ den Vektor der beobachtungsspezifischen Dispersionsparameter, den wir zunächst als Vektor bekannter Nuisance-Parameter betrachten.

Bemerkung. Die Funktionen $a(\phi_i)$ können auch als beobachtungsspezifische Funktionen $a_i(\phi)$ mit einem gemeinsamen Dispersionsparameter ϕ aufgefasst werden.

Durch die Linkfunktion $g(\mu_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}$ wird ein Zusammenhang zwischen den Modellparametern $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ und den Erwartungswerten $\mu_i = \mu_i(\boldsymbol{\beta})$ hergestellt. Damit kann (2.11) als eine Funktion in $\boldsymbol{\beta}$ aufgefasst werden und nach dem ML-Prinzip erhält man einen Schätzer für die β_j als Lösung der Score-Gleichungen:

$$\frac{\partial \mathscr{L}(\boldsymbol{\theta}, \boldsymbol{y})}{\partial \beta_j} \stackrel{!}{=} 0, \quad j = 1, \ldots, p. \tag{2.12}$$

Um die impliziten Abhängigkeiten $\theta_i = \theta_i(\mu_i), \mu_i = \mu_i(\eta_i)$ und $\eta_i = \eta_i(\boldsymbol{\beta})$ zu berücksichtigen erweitern wir (2.12) unter Zuhilfenahme der Kettenregel und erhalten

$$\frac{\partial \mathscr{L}(\boldsymbol{\theta}, \boldsymbol{y})}{\partial \beta_j} = \frac{\partial \mathscr{L}(\boldsymbol{\theta}, \boldsymbol{y})}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j} \stackrel{!}{=} 0. \tag{2.13}$$

Mit den Identitäten

$$\frac{\partial \mu_i}{\partial \theta_i} = \frac{\partial b'(\theta_i)}{\theta_i} = b''(\theta_i) \stackrel{(2.5)}{=} V(\mu_i),$$
$$\frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial g(\mu_i)} x_{ij} = \frac{x_{ij}}{g'(\mu_i)}$$

folgt für die Score-Gleichung (2.13) für alle $j = 1, \ldots, p$:

$$\begin{aligned}\frac{\partial \mathscr{L}(\boldsymbol{\theta}, \boldsymbol{y})}{\partial \beta_j} &= \sum_{i=1}^{n} \frac{\partial}{\partial \theta_i} \left(\frac{y_i \theta_i - b(\theta_i)}{a(\phi_i)} + c(y_i, \phi_i) \right) \frac{1}{V(\mu_i)} \frac{x_{ij}}{g'(\mu_i)} \\ &= \sum_{i=1}^{n} \frac{y_i - \mu_i}{a(\phi_i) V(\mu_i)} \frac{x_{ij}}{g'(\mu_i)} \stackrel{!}{=} 0. \end{aligned} \tag{2.14}$$

Verwendet man die kanonische Linkfunktion $g(\mu_i) = \theta_i$, vereinfacht sich die Score-Gleichung (2.14) wegen $g'(\mu_i) = 1/V(\mu_i)$ zu:

$$\frac{\partial \mathscr{L}(\boldsymbol{\theta}, \boldsymbol{y})}{\partial \beta_j} = \sum_{i=1}^{n} \frac{y_i - \mu_i}{a(\phi_i)} x_{ij} \stackrel{!}{=} 0, \quad j = 1, \ldots, p. \tag{2.15}$$

Die Lösung der Gleichungssysteme (2.14) für beliebige Linkfunktionen oder (2.15) für den kanonischen Link nach β_j liefert den ML-Schätzer $\hat{\beta}_j$.

Bemerkung. Die β_j stecken in den Gleichungssystemen (2.14) und (2.15) implizit in den μ_i, da $\mu_i = g^{-1}(\eta_i) = g^{-1}(\boldsymbol{x}_i^\top \boldsymbol{\beta})$ gilt. Die Gleichungsysteme sind nur für normalverteilte Response-Variablen mit der Identität als Linkfunktion linear in $\boldsymbol{\beta}$ und müssen daher iterativ gelöst werden.

2.2.1 Iterative Weighted Least-Squares

Um die Gleichungssysteme (2.14) und (2.15) zu lösen, fasst man sie als eine Instanz eines Weighted Least-Squares (WLS) Problems auf (McCullagh und Nelder, 1989, S. 41 ff.). Um diese Vorgehensweise zu motivieren, wendet man zunächst das Newton-Verfahren auf das Gleichungssystem an, um die Nullstelle iterativ bestimmen zu können. Die k-te Iterierte des gesuchten Parametervektors $\hat{\boldsymbol{\beta}}$ lautet dann:

$$\boldsymbol{\beta}^{(k)} = \boldsymbol{\beta}^{(k-1)} - H_{k-1}^{-1} \nabla \mathscr{L}_{k-1}, \quad k = 1, 2, \ldots, \tag{2.16}$$

wobei $\nabla \mathscr{L}_{k-1}$ und H_{k-1} den Gradienten respektive die Hessematrix – jeweils ausgewertet an der letzten Iterierten $\boldsymbol{\beta}^{(k-1)}$ – bezeichnen:

$$\nabla \mathscr{L}_{k-1} = \left(\begin{array}{c} \frac{\partial \mathscr{L}(\boldsymbol{\theta}, \boldsymbol{y})}{\partial \beta_1} \\ \vdots \\ \frac{\partial \mathscr{L}(\boldsymbol{\theta}, \boldsymbol{y})}{\partial \beta_p} \end{array} \right) \Bigg|_{\boldsymbol{\beta} = \boldsymbol{\beta}^{(k-1)}},$$

$$H_{k-1} = \left(\begin{array}{cccc} \frac{\partial^2 \mathscr{L}(\boldsymbol{\theta}, \boldsymbol{y})}{\partial \beta_1 \partial \beta_1} & \frac{\partial^2 \mathscr{L}(\boldsymbol{\theta}, \boldsymbol{y})}{\partial \beta_1 \partial \beta_2} & \cdots & \frac{\partial^2 \mathscr{L}(\boldsymbol{\theta}, \boldsymbol{y})}{\partial \beta_1 \partial \beta_p} \\ \frac{\partial^2 \mathscr{L}(\boldsymbol{\theta}, \boldsymbol{y})}{\partial \beta_2 \partial \beta_1} & \frac{\partial^2 \mathscr{L}(\boldsymbol{\theta}, \boldsymbol{y})}{\partial \beta_2 \partial \beta_2} & \cdots & \frac{\partial^2 \mathscr{L}(\boldsymbol{\theta}, \boldsymbol{y})}{\partial \beta_2 \partial \beta_p} \\ \vdots & & \ddots & \vdots \\ \frac{\partial^2 \mathscr{L}(\boldsymbol{\theta}, \boldsymbol{y})}{\partial \beta_p \partial \beta_1} & \cdots & \cdots & \frac{\partial^2 \mathscr{L}(\boldsymbol{\theta}, \boldsymbol{y})}{\partial \beta_p \partial \beta_p} \end{array} \right) \Bigg|_{\boldsymbol{\beta} = \boldsymbol{\beta}^{(k-1)}}.$$

Bemerkung. Da die Bestimmung der Inversen H_{k-1}^{-1} in (2.16) numerisch aufwendig ist, bestimmt man stattdessen oft die Lösung $\delta\boldsymbol{\beta}^{(k-1)}$ des Gleichungssystems

$$H_{k-1}\delta\boldsymbol{\beta}^{(k-1)} = \nabla\mathscr{L}_{k-1}$$

und ersetzt (2.16) durch

$$\boldsymbol{\beta}^{(k)} = \boldsymbol{\beta}^{(k-1)} - \delta\boldsymbol{\beta}^{(k-1)}.$$

Die j-te Komponente des Gradienten wie sie in (2.14) definiert ist, lässt sich zu

$$\begin{aligned}
(\nabla\mathscr{L})_j &= \sum_{i=1}^{n} \frac{y_i - \mu_i}{a(\phi_i)V(\mu_i)} \frac{x_{ij}}{g'(\mu_i)} \\
&= \sum_{i=1}^{n} (y_i - \mu_i) \underbrace{\frac{1}{a(\phi_i)V(\mu_i)(g'(\mu_i))^2}}_{=:\omega_i} \overbrace{g'(\mu_i)}^{=:\delta_i} x_{ij} \\
&= \sum_{i=1}^{n} (y_i - \mu_i)\omega_i\delta_i x_{ij}
\end{aligned}$$

umformen. Es sei daran erinnert, dass wir ϕ_i als bekannten Nuisance-Parameter betrachten. Wir fassen die Hilfsgrößen $(\omega_1, \ldots, \omega_n)$ und $(\delta_1, \ldots, \delta_n)$ zu den Diagonalmatrizen $W = \operatorname{diag}(\omega_i)$ und $D = \operatorname{diag}(\delta_i)$ zusammen. Damit ergibt sich für den Gradienten $\nabla\mathscr{L}$:

$$\nabla\mathscr{L} = X^\top DW(\boldsymbol{y} - \boldsymbol{\mu}). \tag{2.17}$$

Die Hessematrix erhalten wir, indem wir jede Komponente des Gradienten wiederum nach β_j für $1 \le j \le p$ ableiten:

$$\begin{aligned}
H &= \frac{\partial^2\mathscr{L}(\boldsymbol{\theta}, \boldsymbol{y})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top} = \frac{\partial^2\mathscr{L}(\boldsymbol{\theta}, \boldsymbol{y})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\eta}^\top}\frac{\partial\boldsymbol{\eta}}{\partial\boldsymbol{\beta}} \\
&= X^\top \left(\operatorname{diag}\left(\frac{\partial\delta_i\omega_i}{\partial\eta_i}(y_i - \mu_i) \right) - DW\frac{\partial\boldsymbol{\mu}}{\partial\boldsymbol{\eta}^\top} \right) X \\
&= -X^\top \underbrace{\left(W - \operatorname{diag}\left(\frac{\partial\delta_i\omega_i}{\partial\eta_i}(y_i - \mu_i) \right) \right)}_{=:W'} X \\
&= -X^\top W' X.
\end{aligned} \tag{2.18}$$

Bemerkung. Die Einträge der Diagonalmatrix W' lassen sich durch

$$\begin{aligned}
\omega_i' &= \omega_i - (y_i - \mu_i)\frac{\partial\delta_i\omega_i}{\partial\eta_i} \\
&= \omega_i - (y_i - \mu_i)\frac{\partial\mu_i}{\partial\eta_i}\frac{\partial}{\partial\mu_i}\frac{1}{a(\phi_i)V(\mu_i)g'(\mu_i)} \\
&= \omega_i - (y_i - \mu_i)\frac{1}{g'(\mu_i)}\left(-\frac{a(\phi_i)V'(\mu_i)g'(\mu_i) + a(\phi_i)V(\mu_i)g''(\mu_i)}{[a(\phi_i)V(\mu_i)g'(\mu_i)]^2} \right) \\
&= \omega_i + (y_i - \mu_i)\frac{V'(\mu_i)g'(\mu_i) + V(\mu_i)g''(\mu_i)}{a(\phi_i)[V(\mu_1)]^2[g'(\mu_i)]^3}
\end{aligned} \tag{2.19}$$

bestimmen. Die Matrizen W, W' und D hängen vom Vektor $\boldsymbol{\mu}$ ab, der wiederum von der aktuellen Iterierten $\boldsymbol{\beta}^{(k)}$ abhängt. Um diese Abhängigkeit zu verdeutlichen schreiben wir auch W'_k, W_k und D_k bzw. $\boldsymbol{\mu}^{(k)}$.[7] Außerdem gilt für den Erwartungswert von W' : $\mathbb{E}(W') = W$.

Setzen wir nun (2.17) und (2.18) in (2.16) ein, folgt die Iterationsvorschrift für den Vektor der Schätzer für die Modellparameter $\hat{\boldsymbol{\beta}}$:

$$\boldsymbol{\beta}^{(k)} = \boldsymbol{\beta}^{(k-1)} + (X^\top W'_{k-1} X)^{-1} X^\top D_{k-1} W_{k-1} (\boldsymbol{y} - \boldsymbol{\mu}^{(k-1)}). \tag{2.20}$$

Der Zusammenhang mit WLS-Problemen ergibt sich durch die Einführung von *Pseudobeobachtungen* $\boldsymbol{z}_k$:

$$\boldsymbol{z}_k = X\boldsymbol{\beta}^{(k)} + {W'}_k^{-1} D_k W_k (\boldsymbol{y} - \boldsymbol{\mu}^{(k)}).$$

Damit lässt sich (2.20) zu

$$\boldsymbol{\beta}^{(k)} = \left(X^\top W'_{k-1} X\right)^{-1} X^\top W'_{k-1} \boldsymbol{z}_{k-1} \tag{2.21a}$$

oder äquivalent

$$X^\top W'_{k-1} X \boldsymbol{\beta}^{(k)} = X^\top W'_{k-1} \boldsymbol{z}_{k-1} \tag{2.21b}$$

umformen. Gleichung (2.21b) entspricht dabei der *Normalgleichung* eines WLS-Problems mit Gewicht W'_{k-1}. Die Methode hat iterativ zu erfolgen, da sowohl die Pseudobeobachtungen $\boldsymbol{z}_{k-1}$ als auch das Gewicht W'_{k-1} über $\boldsymbol{\mu}^{(k-1)} = \boldsymbol{\mu}(\boldsymbol{\beta}^{(k-1)})$ vom zu bestimmenden Parametervektor $\boldsymbol{\beta}$ abhängen, der zum Zeitpunkt k eben durch $\boldsymbol{\beta}^{(k)}$ angenähert werden muss.[8] Daher spricht man in diesem Fall auch von einer Methode zur Lösung eines Iterative Weighted Least-Squares (IWLS) Problems. Charnes, Frome und Yu (1976) zeigen, dass die IWLS Schätzer für den Fall einer Verteilung aus der Exponentialfamilie auch tatsächlich mit den ML-Schätzern übereinstimmen.

2.2.2 Kanonische Links und Fisher-Scoring

Für kanonische Linkfunktionen lässt sich Gleichung (2.19) wegen der Beziehungen $g(\mu) = \eta = \theta, \mu = b'(\theta)$ und

$$\begin{aligned} g'(\mu) &= \frac{\partial g(\mu)}{\partial \mu} = \frac{\partial \theta}{\partial b'(\theta)} = \frac{1}{b''(\theta)} = \frac{1}{V(\mu)}, \\ g''(\mu) &= \qquad\qquad\qquad\qquad = -\frac{V'(\mu)}{[V(\mu)]^2} \end{aligned}$$

[7] Um die k-te Iterierte für den Vektor der Erwartungswerte nicht mit der k-ten Komponente desselbigen zu verwechseln, verwenden wir $\boldsymbol{\mu}^{(k)}$ für die k-te Iterierte und μ_k für die k-te Komponente des Vektors.

[8] Die Normalgleichung beim LM ist durch $X^\top X\hat{\boldsymbol{\beta}} = X^\top \boldsymbol{y}$ definiert. Da hier weder das Gewicht ($W' = I$) noch die Beobachtungen $\boldsymbol{y}$ von $\boldsymbol{\beta}$ abhängen, kann beim LM der Schätzer direkt bestimmt und muss nicht iterativ ermittelt werden.

weiter vereinfachen:

$$\begin{aligned}\omega_i' &= \omega_i + (y_i - \mu_i)\frac{V'(\mu_i)g'(\mu_i) + V(\mu_i)g''(\mu_i)}{a(\phi_i)[V(\mu_1)]^2[g'(\mu_i)]^3} \\ &= \omega_i + (y_i - \mu_i)\frac{V'(\mu_i)/V(\mu_i) - V'(\mu_i)/V(\mu_i)}{a(\phi_i)/V(\mu_i)} \\ &= \omega_i. \end{aligned} \tag{2.22}$$

Für die Pseudobeobachtungen $\boldsymbol{z}_k$ folgt dann mit (2.22) und der Tatsache, dass die Multiplikation der Diagonalmatrizen D und W kommutativ ist:

$$\boldsymbol{z}_k = X\boldsymbol{\beta}^{(k)} + W_k^{-1}D_kW_k(\boldsymbol{y} - \boldsymbol{\mu}^{(k)}) = X\boldsymbol{\beta}^{(k)} + D_k(\boldsymbol{y} - \boldsymbol{\mu}^{(k)}). \tag{2.23}$$

Und damit lautet die Iterationsvorschrift zur Bestimmung des Schätzers $\hat{\boldsymbol{\beta}}$:

$$\boldsymbol{\beta}^{(k)} = \left(X^\top W_{k-1}X\right)^{-1} X^\top W_{k-1}\boldsymbol{z}_{k-1} \quad \text{bzw.} \quad X^\top W_{k-1}X\boldsymbol{\beta}^{(k)} = X^\top W_{k-1}\boldsymbol{z}_{k-1}. \tag{2.24}$$

Durch die Beziehung $W' = W$ für kanonische Linkfunktionen vereinfacht sich also die Berechnung der Pseudobeobachtungen. Diese kann man auch für nicht kanonische Linkfunktionen erreichen, indem man statt der beobachteten Hessematrix in Gleichung (2.16) deren Erwartungswert – also die Informationsmatrix – heranzieht. Der Erwartungswert der Hessematrix lautet mit Gleichung (2.18):

$$\begin{aligned}\mathbb{E}(H) &= \mathbb{E}\left(-X^\top W'X\right) \\ &= -X^\top WX.\end{aligned}$$

Setzt man die Pseudobeobachtungen wieder wie in Gleichung (2.23) an, erhält man die gleiche Iterationsvorschrift (2.24) wie für kanonische Linkfunktionen. Diese Technik bezeichnet man als *Fisher-Scoring.*

Bemerkung 1. Die Pseudobeobachtungen $\boldsymbol{z}$ beim Fisher-Scoring entsprechen einer Linearisierung der Linkfunktion durch eine Taylorreihen-Entwicklung:

$$\begin{aligned}g(y_i) &\approx g(\mu_i) + (y_i - \mu_i)g'(\mu_i) \\ &= \eta_i + (y_i - \mu_i)\delta_i \\ &= \boldsymbol{x}_i^\top\boldsymbol{\beta} + (y_i - \mu_i)\delta_i \Rightarrow \\ \boldsymbol{z} &= X\boldsymbol{\beta} + D(\boldsymbol{y} - \boldsymbol{\mu}).\end{aligned}$$

Bemerkung 2. Für die Momente von z gilt $\mathbb{E}(\boldsymbol{z}) = X\boldsymbol{\beta}$ und $\operatorname{var}(\boldsymbol{z}) = D\operatorname{var}(\boldsymbol{y})D = W^{-1}$. Damit ist die Gewichtungsmatrix gerade die Inverse der Kovarianzmatrix der (Pseudo-)Beobachtungen und damit hat das nach (2.24) bestimmte $\hat{\boldsymbol{\beta}}$ nach dem Satz von Aitken (Aitken, 1935) minimale Varianz[9] in der Klasse aller unverzerrten linearen Schätzer (Best Linear Unbiased Estimator (BLUE)).

[9] „Minimale Varianz“ ist in dem Sinne zu verstehen, dass – falls $\tilde{\boldsymbol{\beta}}$ einen weiteren linearen, unverzerrten Schätzer von $\boldsymbol{\beta}$ bezeichnet – für die Varianz von $\operatorname{var}(\tilde{\boldsymbol{\beta}}) = \operatorname{var}(\hat{\boldsymbol{\beta}}) + P$ mit einer positiv definiten Matrix P gilt.

Bemerkung 3. Das Fisher-Scoring legt einen einfachen Algorithmus für die iterative Bestimmung des Schätzers $\hat{\boldsymbol{\beta}}$ nahe:

1. Als erste Schätzung für den Vektor der Erwartungswerte bestimmt man beispielsweise $\boldsymbol{\mu}^{(0)} = \boldsymbol{y} + c$ mit einer nicht negativen Konstante $c \geq 0$.
2. Für $\boldsymbol{\mu}^{(0)}$ berechnet man $\boldsymbol{\eta}^{(0)} = g(\boldsymbol{\mu}^{(0)}), \boldsymbol{z}_0 = \boldsymbol{\eta}^{(0)}$ und W_0.
3. Danach bestimmt man die nächste Iterierte $\boldsymbol{\beta}^{(1)}$ gemäß (2.24).
4. Mit der neuen Iterierten $\boldsymbol{\beta}^{(1)}$ bestimmt man $W_1, D_1, \boldsymbol{\eta}^{(1)}, \boldsymbol{\mu}^{(1)}$ und $\boldsymbol{z}_1$.
5. So bestimmt man so lange iterativ $\boldsymbol{\beta}^{(k)}$ bis die relativen Änderungen

$$\frac{\|\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}^{(k-1)}\|}{\|\boldsymbol{\beta}^{(k)}\|}$$

hinreichend klein sind.

Bemerkung 4. Da die Gewichtsmatrix W_k eine Diagonalmatrix ist, kann man das IWLS-Problem (2.24) auch als ein Least-Squares (LS)-Problem auffassen:

$$\underbrace{X^\top \sqrt{W_{k-1}}}_{:=X'^\top} \underbrace{\sqrt{W_{k-1}} X}_{=:X'} \boldsymbol{\beta}^{(k)} = \underbrace{X^\top \sqrt{W_{k-1}}}_{=:X'^\top} \underbrace{\sqrt{W_{k-1}} \boldsymbol{z}_{k-1}}_{=:\boldsymbol{z'}_k}$$
$$X'^\top X' \boldsymbol{\beta}^{(k)} = X'^\top \boldsymbol{z'}_k,$$

wobei $\sqrt{W_{k-1}} = \operatorname{diag}\left(\sqrt{\omega_i}\right)$ bezeichnet. Man beachte aber, dass $X' = X'(\boldsymbol{\beta}^{(k-1)})$ und $\boldsymbol{z}' = \boldsymbol{z}'(\boldsymbol{\beta}^{(k-1)})$ auch von der Iterierten $\boldsymbol{\beta}^{(k-1)}$ abhängen und somit das Gleichungssystem trotzdem iterativ gelöst werden muss.

Bemerkung 5. Zur effektiven numerischen Lösung des IWLS-Problems für das GLM verwendet man u. a. die Choleski- oder die QR-Zerlegung.

2.3 Goodness-of-Fit

Sind die Modellparameter einmal geschätzt, ist man in weiterer Folge an einem Maß für die Güte der Modellanpassung interessiert. Die Aufgabe eines GLM ist es ja, für eine Menge an erklärenden Variablen X einen Schätzvektor $\hat{\boldsymbol{\mu}}$ für die Menge der tatsächlichen Erwartungswerte $\boldsymbol{\mu}$ der beobachteten Response-Variablen $\boldsymbol{y}$ zu liefern. Dabei stimmen die Werte des Vektors $\hat{\boldsymbol{\mu}}$ mit den Werten des Response-Vektors $\boldsymbol{y}$ i. Allg. nicht exakt überein, zumal das GLM ja auch den Erwartungswert der Response-Variablen und nicht diese selbst modelliert. Es liegt daher nahe, die Größenordnung der Abweichung der geschätzten Erwartungswerte $\hat{\boldsymbol{\mu}}$ von den Werten des Responsevektors $\boldsymbol{y}$ zu bestimmen und zu bewerten.

Um diese Bewertung vornehmen zu können, bedarf es eines Maßes. Dabei sollte das gewählte Maß derart beschaffen sein, dass statistische Aussagen über die Größenordnung der Abweichung getroffen werden können. Im folgenden werden wir zwei solcher Maße betrachten: die *Deviance* und die *Pearson-Statistik*.

2.3.1 Deviance

Ein GLM für einen n-dimensionalen Responsevektor $\boldsymbol{y}$ modelliert den Erwartungswert unter Zuhilfenahme des p-dimensionalen Parametervektors $\boldsymbol{\beta}$. Im einfachsten Fall hat das Modell genau einen Parameter

$$g(\mu_i) = \beta_0, \quad 1 \leq i \leq n.$$

Man spricht in diesem Fall von dem *Null-Modell*. Das Null-Modell postuliert, dass es einen gemeinsamen Erwartungswert μ_0 für alle y_i gibt. Die Varianz der y_i schreibt das Modell daher vollkommen der stochastischen Komponente (2.9a) zu.

Im Gegensatz dazu weist das *volle* oder *saturierte Modell* gerade so viele Parameter auf, wie es Beobachtungen gibt, es folgt also

$$g(\mu_i) = \beta_i, \quad 1 \leq i \leq n.$$

Die β_i werden so bestimmt, dass sich die Beobachtungen y_i und die Schätzer für den Erwartungswert $\hat{\mu}_i$ decken, es gilt also $y_i = \hat{\mu}_i$. Die Varianz der y_i wird in diesem Modell vollständig durch die systematische Komponente (2.9b) erklärt.

Das volle Modell ist i. Allg. uninteressant, da es die Daten nur wiedergibt und nicht aggregiert, während das Null-Modell zu stark vereinfacht. Daher wird man ein Modell dazwischen suchen, das mit möglichst wenig Parametern möglichst viel „Information“ enthält. Um die Güte dieser Information zu bestimmen, bedient man sich der log-Likelihood-Funktion $\mathscr{L}(\tilde{\boldsymbol{\mu}}, \boldsymbol{y})$, ist sie ja ein Maß für die Güte einer bestimmten Parameterwahl $\tilde{\boldsymbol{\mu}}$. Als Referenzpunkt bestimmt man den Wert der log-Likelihood-Funktion des vollen Modells $\mathscr{L}(\boldsymbol{y}, \boldsymbol{y})$, das dadurch charakterisiert ist, dass die geschätzten Erwartungswerte $\hat{\mu}_i$ gerade den Beobachtungen y_i entsprechen.

Definition 2.6 (Deviance)**.** Sei $\mathscr{L}(\boldsymbol{\mu}, \boldsymbol{y})$ die log-Likelihood-Funktion der n unabhängigen Zufallsvariablen Y_i mit den Realisierungen $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$. Bezeichne ϕ weiters den Dispersionsparameter, dann heißt die Statistik

$$\frac{1}{\phi} D(\boldsymbol{y}, \boldsymbol{\mu}) = -2 \left[\mathscr{L}(\boldsymbol{\mu}, \boldsymbol{y}) - \mathscr{L}(\boldsymbol{y}, \boldsymbol{y}) \right] \tag{2.25}$$

die *skalierte Deviance*.

Ziel ist es, die Deviance als Maß für den „Lack of Fit“ zu minimieren. Bezeichne $\hat{\boldsymbol{\mu}}$ den Schätzvektor für den Erwartungswert, den man durch Maximierung von $\mathscr{L}(\boldsymbol{\mu}, \boldsymbol{y})$ über $\boldsymbol{\beta}$ erhält, dann minimiert $\hat{\boldsymbol{\mu}}$ auch die Deviance, da $\mathscr{L}(\boldsymbol{y}, \boldsymbol{y})$ ein von $\hat{\boldsymbol{\mu}}(\boldsymbol{\beta})$ unabhängiger Wert ist.

Bemerkung. Fasst man ein GLM als Hypothese auf, testet man also

$$H_0 : \boldsymbol{\mu} = g^{-1}(X\boldsymbol{\beta}) \qquad \text{vs.} \qquad H_A : \boldsymbol{\mu} \neq g^{-1}(X\boldsymbol{\beta}),$$

erhält man die Gleichungen:

$$\max_{H_0} f(\boldsymbol{y}, \boldsymbol{\mu}) = f(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) \qquad \text{vs.} \qquad \max f(\boldsymbol{y}, \boldsymbol{\mu}) = f(\boldsymbol{y}, \boldsymbol{y}),$$

Tabelle 2.1: Deviance-Terme ausgewählter Verteilungen

Verteilung	Deviance $D(\boldsymbol{y}, \hat{\boldsymbol{\mu}})$
Normalverteilung	$\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2$
Gammaverteilung	$2\sum_{i=1}^{n}\left[-\log(y_i/\hat{\mu}_i) + (y_i - \hat{\mu}_i)/\hat{\mu}_i\right]$
Poissonverteilung	$2\sum_{i=1}^{n}\left[y_i \log(y_i/\hat{\mu}_i) + (y_i - \hat{\mu}_i)\right]$.

deren Quotient durch

$$\Lambda(\boldsymbol{y}) := \frac{\max_{H_0} f(\boldsymbol{y}, \boldsymbol{\mu})}{\max f(\boldsymbol{y}, \boldsymbol{\mu})} = \frac{f(\boldsymbol{y}, \hat{\boldsymbol{\mu}})}{f(\boldsymbol{y}, \boldsymbol{y})}$$

gegeben ist. Der Quotient $\Lambda(\boldsymbol{y})$ entspricht also einer Likelihood-Quotienten-Teststatistik. Wiewohl die exakte Verteilung von $\Lambda(\boldsymbol{y})$ i. Allg. schwierig zu bestimmen ist, kennt man die asymptotischen Verteilungseigenschaften von $-2\log\Lambda(\boldsymbol{y})$, das unter gewissen Regularitätsbedingungen asymptotisch[10] χ^2_q verteilt ist, wobei die Anzahl der Freiheitsgrade q gerade der Differenz der Dimensionen der Parameterräume in H_0 und H_A entspricht und somit $q = n - p$ gilt. Außerdem folgt

$$\begin{aligned} -2\log\Lambda(\boldsymbol{y}) &= -2\log\frac{f(\boldsymbol{y}, \hat{\boldsymbol{\mu}})}{f(\boldsymbol{y}, \boldsymbol{y})} \\ &= -2\left[\mathscr{L}(\boldsymbol{\mu}, \boldsymbol{y}) - \mathscr{L}(\boldsymbol{y}, \boldsymbol{y})\right] \\ &= \frac{1}{\phi} D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) \end{aligned}$$

und somit erhalten wir asymptotische Verteilungsaussagen für die Deviance.

Diese sind allerdings mit Vorsicht zu interpretieren, da sie auf der Asymptotik für $n \to \infty$ beruhen. Die Dimension von H_A ist gerade gleich n und damit sind die asymptotischen Aussagen ihrer Grundlage beraubt. Die χ^2-Approximation für die Deviance kann aber u. a. für Poisson-Modelle mit großen μ_i, für (nicht standardisierte) Binomial-Modelle mit großem m_i und für Gamma-Modelle mit kleinem ϕ trotzdem passend sein (vgl. Firth, 1991; Jørgensen, 1987).

Beispiel 2.7. Für die Verteilungen aus der Exponentialfamilie die in Abschnitt 2.1.3 besprochen wurden, findet sich die jeweilige (unskalierte) Deviance $D(\,\cdot\,)$ in der Tabelle 2.1 (vgl. McCullagh und Nelder, 1989, S. 34 ff.).

Die Deviance der Normalverteilung entspricht offenbar genau der Fehlerquadratsumme $\mathrm{SSE}(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2$ des LM und auch für das GLM folgt hier exakt:

$$\frac{1}{\phi} D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2 \sim \chi^2_{n-p}. \qquad \Diamond$$

[10] Für normalverteilte Responses folgt sogar eine exakte χ^2-Verteilung.

2.3.2 Pearson-Statistik

Ein anderes Maß für den Goodness-of-Fit ist die *Pearson-Statistik*, welche in Anlehnung an den Schätzer für die Varianz σ^2 beim LM gebildet wird.

Definition 2.8 (Pearson-Statistik)**.** Für die n unabhängigen Zufallsvariablen Y_i mit den Realisierungen $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$ gelte $\operatorname{var}(y_i) = \phi \cdot a_i V(\mu_i)$ mit beobachtungsspezifischen a_i und einem gemeinsamen und festen Dispersionsparameter ϕ, dann heißt die Statistik

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

die *generalisierte Pearson-Statistik*.

Die Pearson-Statistik kann auch dazu verwendet werden, den Dispersionsparameter ϕ zu schätzen. Die Größe

$$\frac{1}{n} \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{a_i V(\mu_i)} \tag{2.26}$$

ist erwartungstreu für ϕ:

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{n} \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{a_i V(\mu_i)}\right] &= \frac{1}{n} \sum_{i=1}^{n} \frac{1}{a_i V(\mu_i)} \mathbb{E}\left((y_i - \mu_i)^2\right) \\
&= \frac{1}{n} \sum_{i=1}^{n} \frac{1}{a_i V(\mu_i)} \operatorname{var}(y_i) \\
&= \frac{1}{n} \sum_{i=1}^{n} \frac{\phi \cdot a_i V(\mu_i)}{a_i V(\mu_i)} \\
&= \phi.
\end{aligned}$$

Die Statistik (2.26) hängt aber vom unbekannten Parametervektor $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^\top$ ab. Daher schätzt man die Dispersion indem man den Schätzer $\hat{\boldsymbol{\mu}}$ verwendet und erhält so als Schätzer für die Dispersion die gemittelte Pearson-Statistik:

$$\hat{\phi} = \frac{1}{n - p} X^2. \tag{2.27}$$

Bemerkung 1. Die Pearson-Statistik ist für normalverteilte Größen äquivalent mit der skalierten Deviance.

Bemerkung 2. Die nicht quadrierten Summanden der Pearson-Statistik

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

heißen *Pearson-Residuen*.

Bemerkung 3. Die Pearson-Statistik folgt für normalverteilte Response-Variablen einer χ^2-Verteilung. Für andere Verteilungen trifft dies nur mehr asymptotisch zu. Die Einschränkungen hinsichtlich der asymptotischen Aussagen über die Verteilung der Deviance gilt für die Pearson-Statistik sinngemäß.

2.3.3 Analysis-of-Deviance

Analog zur Analysis-of-Variance (ANOVA) kann man eine „Analysis-of-Deviance“ durchführen, um sukzessive ein Modell zu bestimmen, in dem diejenigen erklärenden Variablen, die nur wenig zum Goodness-of-Fit beitragen, eliminiert wurden. Dazu testet man für $1 \leq q < p$ ineinandergeschachtelte Modelle („nested models“):

$$\begin{aligned} \eta &= \beta_1 x_1 + \cdots + \beta_q x_q + \beta_{q+1} x_{q+1} + \cdots + \beta_p x_p \qquad \text{(Modell 1)}, \\ \eta &= \qquad\qquad\qquad\quad \beta_{q+1} x_{q+1} + \cdots + \beta_p x_p. \qquad \text{(Modell 2)}. \end{aligned}$$

Die Idee ist nun, den Goodness-of-Fit für beide Modelle zu bestimmen und in Relation zu bringen. Dies geschieht über die Differenz der jeweiligen Deviance Werte. Seien dazu $\hat{\boldsymbol{\mu}}_1$ und $\hat{\boldsymbol{\mu}}_2$ die Schätzer für $\boldsymbol{\mu}$ unter dem ersten bzw. dem zweiten Modell, dann gilt für die Differenz der skalierten Deviance Terme:

$$\begin{aligned} \frac{1}{\phi}\left[D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}_2) - D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}_1)\right] &= -2\left[\mathscr{L}(\hat{\boldsymbol{\mu}}_2, \boldsymbol{y}) - \mathscr{L}(\boldsymbol{y}, \boldsymbol{y})\right] + 2\left[\mathscr{L}(\hat{\boldsymbol{\mu}}_1, \boldsymbol{y}) - \mathscr{L}(\boldsymbol{y}, \boldsymbol{y})\right] \\ &= -2\left[\mathscr{L}(\hat{\boldsymbol{\mu}}_2, \boldsymbol{y}) - \mathscr{L}(\hat{\boldsymbol{\mu}}_1, \boldsymbol{y})\right], \end{aligned} \tag{2.28}$$

was wiederum einer Likelihood-Quotienten-Teststatistik entspricht und damit asymptotisch χ^2_{df} verteilt ist, wobei die Anzahl der Freiheitsgrade wieder der Reduktion der Dimensionen in den Parameterräumen entspricht und damit $\mathrm{df} = q$ gilt.

Bemerkung. Auch hier kann die asymptotische $\chi^2_{p-(p-q)}$-Verteilung die tatsächliche Verteilung schlecht beschreiben, dann nämlich, wenn p nahe bei n liegt (vgl. Firth, 1991).

Formel (2.28) setzt voraus, dass der Dispersionsparameter ϕ bekannt ist. Trifft dies nicht zu und muss man ihn aus den Daten schätzen, kann man analog zum LM die Teststatistik

$$\frac{D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}_2) - D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}_1)}{\hat{\phi} \cdot q} \overset{\text{asy.}}{\sim} F_{q,n-p}$$

verwenden, wobei $\hat{\phi}$ aus dem größeren Modell nach (2.27) zu bestimmen ist (also hier Modell 1).

2.4 Beispiele

2.4.1 Logistische Regression

Die Ursache des Absturzes der Raumfähre Challenger 1986 war ein Ausfall eines Dichtungsringes *(O-Rings)* und dem dadurch entstehenden seitlichen Flammenaustritts. Der Ausfall der Dichtung ließ sich auf verminderte Elastizitätseigenschaften zurückführen, die durch tiefe Temperaturen in der Nacht vor und am Tag der Katastrophe begünstigt wurden.

Casella und Berger (2002) untersuchen einen auf Dalal u. a. (1989) zurückgehenden Datensatz, der die Temperatur bei verschiedenen Starts zusammen mit einem $(0, 1)$ kodierten Faktor, der den Ausfall eines O-Ringes charakterisiert, beschreibt (siehe Tabelle 2.2).

Tabelle 2.2: Challenger Daten (vgl. Casella und Berger, 2002, S. 594)

Nr.	°F	Ausfall	Nr.	°F	Ausfall
1	66	0	13	67	0
2	70	1	14	53	1
3	69	0	15	67	0
4	68	0	16	75	0
5	67	0	17	70	0
6	72	0	18	81	0
7	73	0	19	76	0
8	70	0	20	79	0
9	57	1	21	75	1
10	63	1	22	76	0
11	70	1	23	58	1
12	78	0			

0 = kein Ausfall, 1 = Ausfall des O-Rings

Mit dem Statistikprogramm R[11] wollen wir nun ein GLM an die Daten anpassen. Das GLM modelliert den Erwartungswert einer Zielgröße Y. Um die entwickelte Theorie anwenden zu können, muss die Verteilung von Y außerdem aus der Exponentialfamilie stammen.

Ein weit verbreitetes Verteilungsmodell für eine nach oben beschränkte, nicht negative Anzahl Y ist die *Binomialverteilung* $\mathcal{B}(\pi, m)$, deren Wahrscheinlichkeitsfunktion sich für $y = 0, 1, \ldots, m$ mit

$$Y \sim \mathcal{B}(\pi, m) :\Leftrightarrow f_Y(y, \pi, m) = \mathbb{P}(Y = y) = \binom{m}{y} \pi^y (1-\pi)^{m-y}$$

angeben lässt. Da der Wertebereich der Zufallsvariable Y vom Nuisance-Parameter m abhängt, standardisieren wir Y indem wir die relativen Häufigkeiten $Y' := Y/m$ betrachten und erhalten so für $y' = 0, 1/m, 2/m, \ldots, 1$ die *standardisierte Binomialverteilung* $\mathcal{B}'(\pi, m)$:

$$Y' \sim \mathcal{B}'(\pi, m) :\Leftrightarrow f_{Y'}(y', \pi, m) = \mathbb{P}\left(Y' = y'\right) = \binom{m}{my'} \pi^{my'} (1-\pi)^{m-my'}.$$

Damit lässt sich der Zusammenhang zwischen der Binomialverteilung und der standardisierten Binomialverteilung über

$$Y = mY' \sim \mathcal{B}(\pi, m) \Leftrightarrow Y' \sim \mathcal{B}'(\pi, m)$$

ausdrücken.

[11] Die hier verwendete Version trägt die Versionsnummer `2.8.1` und ist über `http://www.r-project.org` zu beziehen.

Bemerkung 1. Die standardisierte Binomialverteilung ist ein Mitglied der einparametrischen Exponentialfamilie. Definiert man

$$\theta = \log \frac{\pi}{1-\pi}, \quad \phi = 1/m,$$
$$a(\phi) = \phi, \quad b(\theta) = \log(1 + \exp\theta), \quad c(y', \phi) = \log \binom{m}{my'},$$

folgt für die Wahrscheinlichkeitsfunktion der standardisierten Binomialverteilung sofort:

$$\begin{aligned}
\mathbb{P}\left(Y' = y'\right) &= \exp\left\{\log \binom{m}{my'} + my' \log \pi + m(1-y')\log(1-\pi)\right\} \\
&= \exp\left\{\frac{y' \log \frac{\pi}{1-\pi} - \log \frac{1}{1-\pi}}{1/m} + \log \binom{m}{my'}\right\} \\
&= \exp\left\{\frac{y'\theta - b(\theta)}{a(\phi)} + c(y', \phi)\right\}.
\end{aligned}$$

Damit gilt für die ersten beiden Momente von $\mathcal{B}'(\pi, m)$:

$$\mathbb{E}\left(Y'\right) = \pi, \qquad \operatorname{var}\left(Y'\right) = \frac{1}{m}\pi(1-\pi).$$

Bemerkung 2. Die Binomialverteilung selbst ist kein Mitglied der einparametrischen Exponentialfamilie. Dies rührt daher, dass der Wertebereich von y von m abhängt (vgl. Friedl, 1991).

Bemerkung 3. Die kanonische Linkfunktion $\eta = \theta$ der (standardisierten) Binomialverteilung lässt sich mit

$$\operatorname{logit}(\pi) = \log \frac{\pi}{1-\pi}$$

angeben und wird als *Logit-Link* bezeichnet.

Mit der standardisierten Binomialverteilung sind wir nun in der Lage, ein Modell für die Wahrscheinlichkeit eines O-Ring Ausfalles anzugeben. Bezeichne dazu x_i die Temperatur und π_i die Ausfallswahrscheinlichkeit. Ein GLM mit Intercept lautet somit:

$$\operatorname{logit}(\pi_i) := \alpha + \beta x_i.$$

Die folgende Sequenz von R-Befehlen berechnet die ML-Schätzer des Modells, erstellt die „Analysis-of-Deviance“ (die wir der Einfachheit halber als ANOVA bezeichnen) und schätzt die Wahrscheinlichkeit für einen Ausfall des O-Rings für die Temperatur von 31°F, wie sie am Unglückstag vorlag:

```
> temp <- c(66,70,69,68,67,72,73,70,57,63,70,78,67,53,67,75,70,81,
+           76,79,75,76,58)
> fail <- c(0,1,0,0,0,0,0,0,1,1,1,0,0,1,0,0,0,0,0,0,1,0,1)

> mod <- glm(fail ~ temp, family=binomial())
```

```
> anova(mod, test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: fail

Terms added sequentially (first to last)


     Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                    22    28.2672
temp  1   7.9520        21    20.3152    0.0048

> (co <- coef(mod))
(Intercept)        temp
 15.0429016  -0.2321627

> predict(mod, data.frame(temp=31), type="response")
        1
0.9996088
```

Die ANOVA liefert für die Nullhypothese $H_0 : \beta = 0$ einen p-Wert von 0.0048, was ein starkes Indiz dafür ist, dass die Nullhypothese H_0 falsch ist. Das heißt in anderen Worten, dass die Temperatur einen signifikanten Einfluss auf die Ausfallswahrscheinlichkeit zu haben scheint. Damit lautet das geschätzte Modell:

$$\operatorname{logit}(\hat{\pi}_i) = 15.043 - 0.232 \cdot x_i.$$

Das negative Vorzeichen des zu der Temperatur gehörigen Parameters bewirkt, dass höhere Temperaturen eine geringere Ausfallswahrscheinlichkeit induzieren.

Bemerkung. Da die π_i Wahrscheinlichkeiten beschreiben, liegt ihr Wertebereich in $[0, 1]$. Durch den Logit-Link

$$\operatorname{logit} : (0,1) \mapsto \mathbb{R} \qquad \text{bzw. dessen Inversen} \qquad \operatorname{logit}^{-1} : \mathbb{R} \mapsto (0,1)$$

wird erreicht, dass der lineare Prädiktor $\alpha + \beta x_i$, der ja *beliebige* Werte in $\mathbb{R}$ annehmen kann, auf das passende Intervall $(0, 1)$ abgebildet wird.

Am Tag der Challenger Katastrophe herrschte eine Temperatur von 31°F. Die geschätzte Wahrscheinlichkeit unter dem Modell für einen Ausfall eines O-Rings lautet somit:

$$\hat{\pi} = \operatorname{logit}^{-1}(15.043 - 0.232 \cdot 31) = \operatorname{logit}^{-1}(7.846) = 0.9996.$$

Um die Temperatur zu bestimmen, deren Unterschreitung eine Ausfallswahrscheinlichkeit von mindestens 50% bedeutet, müssen wir das Modell lediglich für $\pi_0 = 0.5$ nach x_0 auflösen und erhalten:

$$x_0 = \frac{\operatorname{logit}(\pi_0) - \alpha}{\beta}.$$

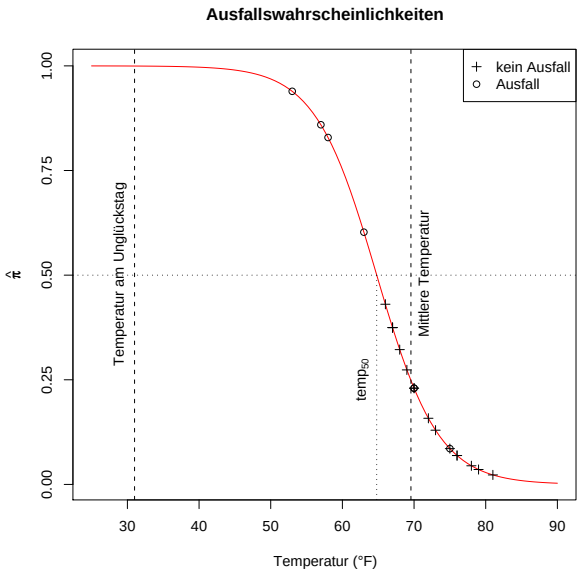


Abbildung 2.2: Logistische Regression für die Ausfallswahrscheinlichkeiten des O-Rings der Challenger Raumfähre bei unterschiedlichen Temperaturen

In R lässt sich die Temperatur entweder direkt oder unter Zuhilfenahme der Funktion `dose.p` aus dem Paket *MASS* berechnen:

```
> library(MASS)
> dose.p(mod)
             Dose       SE
p = 0.5: 64.79464 2.80903

> (mod$family$linkfun(0.5) - co["(Intercept)"]) / co["temp"]
(Intercept)
   64.79464
```

Das heißt, dass für $x \leq 64.79°\text{F}$ die geschätzte Ausfallswahrscheinlichkeit $\hat{\pi}$ größer gleich 0.5 ist.

Die Abbildung 2.2 zeigt die geschätzte Ausfallswahrscheinlichkeit $\hat{\pi}$ in Abhängigkeit der Temperatur. Die mittlere Temperatur und die am Tag der Katastrophe sind als strichlierte Linien eingezeichnet. Man erkennt, dass die Wahrscheinlichkeit für einen Ausfall des O-Rings bei einer Temperatur unter 50°F fast bei Eins liegt. Über 50°F fällt die Wahrscheinlichkeit schnell ab und ab einer Temperatur von 80°F liegt die Ausfallswahrscheinlichkeit nahezu bei Null. Diese sigmoide Form der Kurve ist charakteristisch für den Logit-Link.

Die geschätzten Erwartungswerte des Modells entsprechen den erwarteten Ausfallswahrscheinlichkeiten. Um eine Schätzung auf der Skala der Response-Variable Y zu er-

halten, definiert man einen Grenzpunkt C_0 und erhält so über die Bildungsvorschrift

$$x \mapsto \begin{cases} 1 & \hat{\pi}(x) > C_0, \\ 0 & \hat{\pi}(x) \leq C_0 \end{cases},$$

einen Schätzer $\hat{y}$, wobei $\hat{\pi}(x)$ die geschätzte Ausfallswahrscheinlichkeit für die Temperatur x unter dem Modell bezeichnet. Vergleicht man nun den Schätzer $\hat{y}$ mit den tatsächlichen Beobachtungen y lässt sich feststellen, wie oft das Modell richtige Vorhersagen trifft:

```
> fail.hat <- as.numeric(predict(mod, type="response") > 0.5)
> (pred.table <- table(fail, fail.hat))
    fail.hat
fail  0  1
   0 16  0
   1  3  4

> sum(diag(pred.table)) / length(fail)
[1] 0.8695652
```

Das Modell liefert also bei einem Grenzpunkt von $C_0 = 0.5$ in 20 von 23 Fällen eine korrekte Schätzung für den Ausfall, d. h. dass das Modell im Mittel in 86.96% der Fälle eine richtige Vorhersage trifft. Das Modell sagt bei diesem Grenzpunkt außerdem keinen Ausfall falsch voraus. Diese Tatsache ist auch in Abbildung 2.2 zu erkennen, da über der Grenzlinie bei $C_0 = 0.5$ nur Ausfallsbeobachtungen liegen.

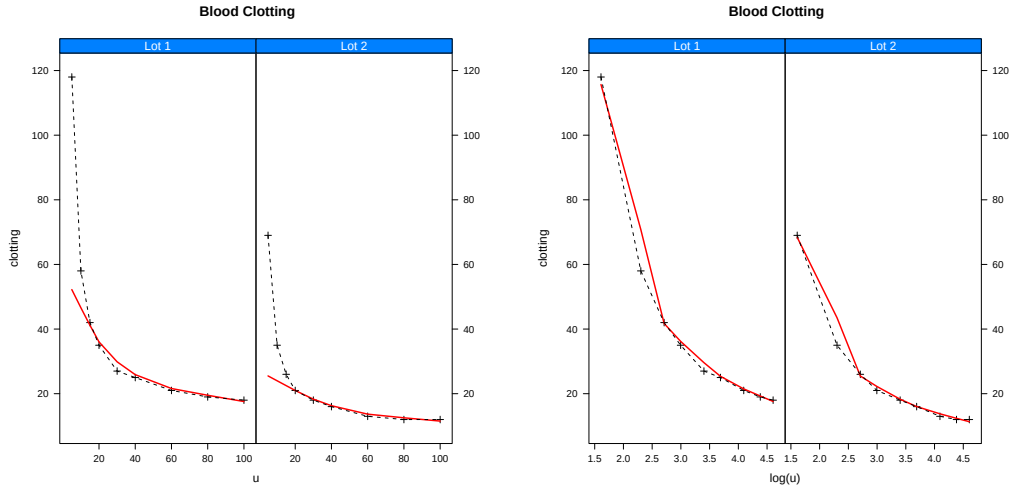
Bemerkung. Die hier vorgestellte Variante der Fehlerabschätzung unterschätzt den tatsächlichen Fehler, da der gleiche Datensatz zur Modellbestimmung und zur Fehlerabschätzung verwendet wird. Eine ausführlichere Diskussion über diese Problematik findet sich in Abschnitt 3.2.4.

2.4.2 Gamma Modell

Ein auf Hurn u. a. (1945) zurückgehender Datensatz beschäftigt sich mit den Gerinnungszeiten von Blutplasma für zwei verschiedene Thromboplastin Gruppen. Die Zeiten wurden dabei für verschiedene Verdünnungen mit Prothrombrin freien Plasma erhoben und finden sich in Tabelle 2.3.

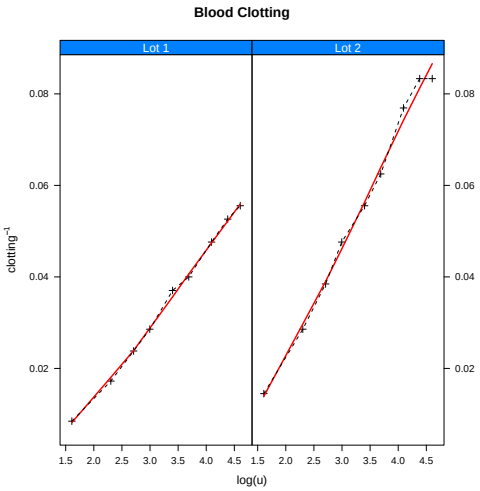
Um eine ungefähre Vorstellung der Art des Zusammenhangs zwischen der Plasma Konzentration und der Gerinnungszeit zu erhalten, tragen wir die beiden Größen gegeneinander auf. Die Abbildung 2.3 zeigt die Relation für verschiedene Skalenniveaus mit den jeweiligen Glättungen (rote Linie).

Während in der Abbildung 2.3(a) das Skalenniveau dem Wertebereich der jeweiligen Variablen entspricht, wird in Abbildung 2.3(b) für die Konzentration eine logarithmische Skala verwendet. Auf der Originalskala ist eine leichte Krümmung in der Glättung zu erkennen, die – wenngleich auch weniger ausgeprägt – auch auf der logarithmischen Skala zu erkennen ist. Da das GLM die Möglichkeit bietet, eine Funktion des Erwartungswertes (linear) zu modellieren, betrachten wir in Abbildung 2.3(c) nicht die Response-Variable



(a) Standard Skala für die x-Achse

(b) Logarithmische Skala für die x-Achse



(c) Logarithmische Skala für die x-Achse und inverse Skala für die y-Achse

Abbildung 2.3: Verschiedende Zusammenhänge zwischen der Plasma Konzentration und der Gerinnungszeit

Tabelle 2.3: Gerinnungszeiten von Blutplasma (vgl. McCullagh und Nelder, 1989, S. 302)

	Gerinnungszeit	
u	Gruppe 1	Gruppe 2
5	118	69
10	58	35
15	42	26
20	35	21
30	27	18
40	25	16
60	21	13
80	19	12
100	18	12

u = Plasma Konzentration

Tabelle 2.4: Untermodelle für die Modellierung der Gerinnungszeiten

	Modell	Bedeutung
(1)	$\mu_{ij}^{-1} = \alpha_0$	weder „Gruppe“ noch „Konzentration“ hat einen Effekt
(2a)	$\mu_{ij}^{-1} = \alpha_0 + \beta \log u_i$	nur „Konzentration“ hat einen Einfluss
(2b)	$\mu_{ij}^{-1} = \alpha_j$	nur „Gruppe“ hat einen Einfluss
(3)	$\mu_{ij}^{-1} = \alpha_j + \beta \log u_i$	„Gruppe“ und „Konzentration“ haben einen Einfluss, es gibt aber keine Wechselwirkung
(4)	$\mu_{ij}^{-1} = \alpha_j + \beta_j \log u_i$	„Gruppe“ und „Konzentration“ haben einen Einfluss, zusätzlich existiert eine Wechselwirkung

„Gerinnungszeit“ sondern deren Inverse. Der Zusammenhang mit der logarithmischen Konzentration ist für diesen Fall annähernd linear und wir vermuten daher ein Modell der Form:

$$\mu_{ij}^{-1} = \alpha_j + \beta_j \log u_i, \quad j = 1, 2,\ i = 1, \ldots, 9, \qquad \text{(volles Modell)}$$

wobei wir den Berechnungen eine Gamma-Varianz $\operatorname{var}(Y_{ij}) \propto \mu_{ij}^2$ zugrunde legen.

In der Tabelle 2.4 findet man die Untermodelle des vollen Modells. Um nun zu überprüfen, ob eines dieser bereits ausreichend wäre, generiert der folgende R-Code die entsprechenden Modelle und führt eine ANOVA durch:

```
> clotting <- data.frame(u=rep(c(5*1:4, 10*3:4, 20*3:5), 2),
+                        lot=rep(factor(paste("Lot", 1:2)), each=9),
+                        clotting=c(118,58,42,35,27,25,21,19,18,
+                          69,35,26,21,18,16,13,12,12))
```

```
> model.1    <- glm(clotting ~ 1, data=clotting, family=Gamma())
> model.lu   <- glm(clotting ~ log(u), data=clotting, family=Gamma())
> model.lot  <- glm(clotting ~ lot, data=clotting, family=Gamma())
> model.main <- glm(clotting ~ log(u) + lot, data=clotting,
+                   family=Gamma())
> model.full <- glm(clotting ~ log(u)*lot, data=clotting,
+                   family=Gamma())

> anova(model.1, model.lu, model.main, model.full, test="F")
Analysis of Deviance Table

Model 1: clotting ~ 1
Model 2: clotting ~ log(u)
Model 3: clotting ~ log(u) + lot
Model 4: clotting ~ log(u) * lot
  Resid. Df Resid. Dev Df Deviance       F    Pr(>F)
1        17     7.7087
2        16     1.0183  1   6.6904 3141.47 < 2.2e-16 ***
3        15     0.3004  1   0.7178  337.06 3.420e-11 ***
4        14     0.0294  1   0.2710  127.26 2.059e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> anova(model.1, model.lot, model.main, model.full, test="F")
Analysis of Deviance Table

Model 1: clotting ~ 1
Model 2: clotting ~ lot
Model 3: clotting ~ log(u) + lot
Model 4: clotting ~ log(u) * lot
  Resid. Df Resid. Dev Df Deviance       F    Pr(>F)
1        17     7.7087
2        16     6.6314  1   1.0773  505.84 2.178e-12 ***
3        15     0.3004  1   6.3310 2972.69 < 2.2e-16 ***
4        14     0.0294  1   0.2710  127.26 2.059e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> coef(model.full)
    (Intercept)          log(u)        lotLot 2 log(u):lotLot 2
   -0.016554382     0.015343115    -0.007354088     0.008256099
```

Wie in Kapitel 2.3.3 besprochen, folgt die Deviance Reduktion beim Übergang von in-

einander geschachtelten Modellen – passend skaliert – asymptotisch einer F-Verteilung. Ist der p-Wert klein, ist die Nullhypothese[12] zu verwerfen und die Terme des größeren Modells haben einen signifikanten Einfluss.

Das heißt, dass wir ausgehend von der ANOVA keine Teilmenge von Termen aus dem vollen Modell entfernen können und erhalten somit das Modell:

$$\begin{aligned}\hat{\mu}_{i1}^{-1} &= -0.01655 + 0.01534 \cdot \log u_i, && \text{(Gruppe 1)}\\ \hat{\mu}_{i2}^{-1} &= (-0.01655 - 0.00735) + (0.01534 + 0.00826) \cdot \log u_i \\ &= -0.02391 + 0.02360 \cdot \log u_i. && \text{(Gruppe 2)}\end{aligned}$$

Mit diesem Modell ist es uns jetzt möglich, den Erwartungswert $\hat{\mu}_0$ auch für nicht beobachtete Konzentrationen u_0 zu schätzen:

```
> u <- seq(4.5, 150, length=5)
> l <- rep("Lot 1", length(u))

# Vorhersage auf der Skala des linearen Prädiktors
> (eta <- predict(model.full, data.frame(u=u, lot=l), type="link"))
         1          2          3          4          5
0.00652285 0.04037653 0.05014286 0.05606310 0.06032437

# Vorhersage auf der originalen Skala
> (mu <- predict(model.full, data.frame(u=u, lot=l), type="response"))
        1         2         3         4         5
153.30721  24.76686  19.94302  17.83704  16.57705

> all.equal(model.full$family$linkinv(eta), mu)
[1] TRUE
```

Aufgrund der Inversen als verwendete Linkfunktion $g(\mu) = \mu^{-1}$, bedeutet der positive Parameter für $\log u_i$, dass für größere Konzentrationen die mittlere Gerinnungszeit tatsächlich *sinkt*.

Um für eine Vorhersage $\hat{\mu}_0$ auch ein Konfidenzintervall zu erhalten, können wir uns zweier verschiedender Ansätze bedienen. Wie in Abschnitt 3.1.1 gezeigt wird, folgt der Schätzer $\hat{\boldsymbol{\beta}}$ für den Parametervektor $\boldsymbol{\beta}$ asymptotisch einer multivariaten Normalverteilung mit Erwartungswertvektor $\boldsymbol{\beta}$ und Varianz-Kovarianz-Matrix $\operatorname{cov}(\hat{\boldsymbol{\beta}})$. Da eine affin lineare Transformation (multivariat) normalverteilter Größen, wieder (multivariat) normalverteilt ist, können wir mit der Tatsache, dass $\hat{\eta}_i$ asymptotisch erwartungstreu für η_i ist, ein Konfidenzintervall für η_i angeben:

$$\mathrm{KIV}(\eta_i) = \hat{\eta}_i \pm z_{1-\frac{\alpha}{2}} \cdot \mathrm{s.e.}(\hat{\eta}_i),$$

[12] Die Nullhypothese H_0 lautet dabei, dass die Parameter, die zu den nur im größeren Modell vorkommenden Termen gehören, jeweils Null sind und somit keinen signifikanten Einfluss auf die Response ausüben.

wobei $\text{s.e.}(\hat{\eta}_i) = \sqrt{\widehat{\text{var}}\,(\hat{\eta}_i)} = \sqrt{\boldsymbol{x}_i^\top \widehat{\text{var}}(\hat{\boldsymbol{\beta}})\boldsymbol{x}_i}$ die Standardabweichung von $\hat{\eta}_i$ und $z_{1-\frac{\alpha}{2}}$ das $1-\alpha/2$ Quantil der Standardnormalverteilung bezeichnet. Damit erhalten wir ein Konfidenzintervall auf der Skala des linearen Prädiktors, welches wir mit der Inversen der Linkfunktion auf die Skala der Response-Variable transformieren:

$$\text{KIV}(\mu_i) = g^{-1}\left(\hat{\eta}_i \pm z_{1-\frac{\alpha}{2}} \cdot \text{s.e.}(\hat{\eta}_i)\right). \qquad (*)$$

Ein anderes Intervall lässt sich über die *Delta Methode* (vgl. Casella und Berger, 2002, S. 240) herleiten. Dazu fassen wir den Erwartungswert μ als Funktion $\mu(\eta) = g^{-1}(\eta)$ des linearen Prädiktors η auf und erhalten für die Varianz von $\hat{\mu}_i$ aus eben dieser Delta Methode:

$$\text{var}\,(\hat{\mu}_i) = \text{var}\left(g^{-1}(\hat{\eta}_i)\right) = \left[\left(g^{-1}(\hat{\eta}_i)\right)'\right]^2 \text{var}\,(\hat{\eta}_i)\,,$$

wobei $\left(g^{-1}(\,\cdot\,)\right)'$ die Ableitung der inversen Linkfunktion bezeichnet. Somit lässt sich das Konfidenzintervall für $\hat{\mu}_i$ auch direkt konstruieren und wir erhalten:

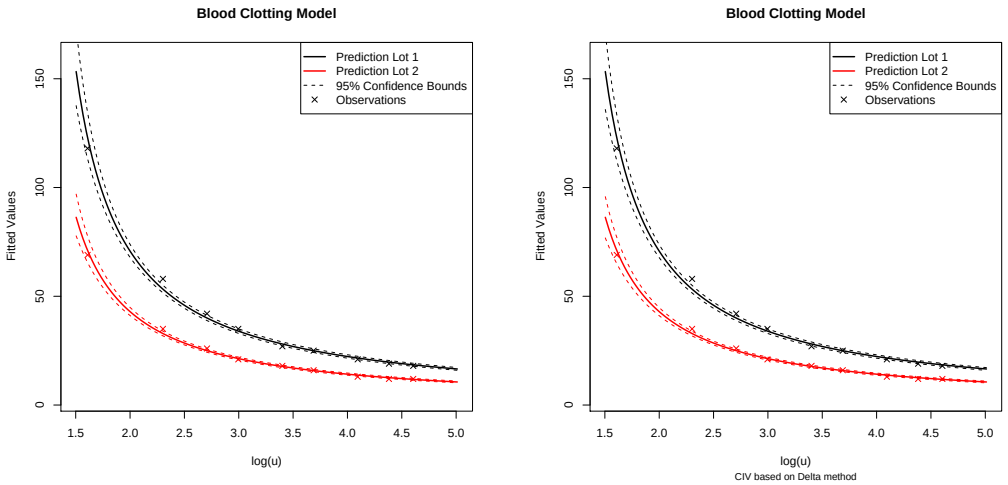
$$\text{KIV}(\mu_i) = \hat{\mu}_i \pm z_{1-\frac{\alpha}{2}} \cdot \text{s.e.}(\hat{\mu}_i). \qquad (**)$$

Für die zum Einsatz kommende Linkfunktion $\eta_i = g(\mu_i) = 1/\mu_i \Leftrightarrow \mu_i = g^{-1}(\eta_i) = 1/\eta_i$ erhält man für die Ableitung $\left(g^{-1}(\eta_i)\right)' = -1/\eta_i^2$. In R lassen sich die Konfidenzintervalle für $(*)$ und $(**)$ wie folgt bestimmen:

```
> n <- length(clotting$clotting)
> civ <- delta.civ <- matrix(nrow=n, ncol=3)

> colnames(civ) <- c("lower", "mu.hat", "upper")
> colnames(delta.civ) <- c("lower.delta", "mu.hat", "upper.delta")

> eta.hat <- predict(model.full, type="link", se.fit=TRUE)
> mu.hat  <- predict(model.full, type="response", se.fit=TRUE)
> inv <- model.full$family$linkinv

> civ[,c(3,1)] <- inv(eta.hat$fit + c(rep(-1, n), rep(1, n)) *
+                     eta.hat$se.fit * qnorm(0.975))
> delta.civ[,c(1,3)] <- mu.hat$fit  + c(rep(-1, n), rep(1, n)) *
+   mu.hat$se.fit * qnorm(0.975)
> civ[,2] <- delta.civ[,2] <- mu.hat$fit
> both.civ <- cbind(civ, delta.civ)

# KIV für "Gruppe 1":
> both.civ[1:9,]
          lower    mu.hat     upper lower.delta    mu.hat upper.delta
 [1,] 113.16062 122.85904 134.37570   112.32942 122.85904   133.38867
 [2,]  51.57904  53.26389  55.06252    51.52401  53.26389    55.00377
```

```
 [3,]  38.83616  40.00713  41.25091    38.80086  40.00713    41.21341
 [4,]  32.98756  34.00264  35.08217    32.95632  34.00264    35.04895
 [5,]  27.18438  28.06578  29.00624    27.15581  28.06578    28.97575
 [6,]  24.15935  24.97221  25.84166    24.13200  24.97221    25.81241
 [7,]  20.87905  21.61432  22.40328    20.85315  21.61432    22.37549
 [8,]  19.04256  19.73182  20.47285    19.01762  19.73182    20.44603
 [9,]  17.82574  18.48317  19.19096    17.80149  18.48317    19.16485

# KIV für "Gruppe 2":
> both.civ[10:18,]
         lower   mu.hat    upper lower.delta   mu.hat upper.delta
 [1,] 65.49079 71.05806 77.65980    65.01752 71.05806    77.09860
 [2,] 31.80377 32.86152 33.99205    31.76859 32.86152    33.95445
 [3,] 24.26856 25.00038 25.77772    24.24649 25.00038    25.75428
 [4,] 20.73673 21.37279 22.04912    20.71721 21.37279    22.02837
 [5,] 17.18714 17.74399 18.33812    17.16910 17.74399    18.31888
 [6,] 15.31950 15.83627 16.38912    15.30207 15.83627    16.37047
 [7,] 13.28116 13.75235 14.25820    13.26444 13.75235    14.24026
 [8,] 12.13415 12.57800 13.05555    12.11791 12.57800    13.03808
 [9,] 11.37187 11.79664 12.25437    11.35600 11.79664    12.23727
```

Die beiden Intervalle unterscheiden sich also kaum, wobei die Intervallgrenzen der Delta Methode jeweils kleiner sind. Die Abbildung 2.4 zeigt die Vorhersagen für das Modell mit den Konfidenzintervallen $(*)$ und $(**)$. Man erkennt, dass beide Verfahren annähernd die gleichen Ergebnisse liefern. Zu sehen ist auch, dass eine höhere log-Konzentration eine Abnahme der Gerinnungszeit induziert.



(a) Modell für die Gerinnungszeit mit Konfidenzintervall nach $(*)$

(b) Modell für die Gerinnungszeit mit Konfidenzintervall nach $(**)$

Abbildung 2.4: Modell für die Gerinnungszeit mit verschiedenen Konfidenzintervallen für die Vorhersage

3 Die Extended-Quasi-Likelihood-Funktion

3.1 Die Quasi-Likelihood-Funktion

Bis jetzt sind wir von einer expliziten Verteilungsannahme aus der Exponentialfamilie ausgegangen und haben dadurch die Varianzfunktion $V(\mu)$ erhalten. Die systematische Komponente (2.9b) und die Linkfunktion (2.9c) stellten einen Zusammenhang zwischen den erklärenden Variablen – in der Designmatrix X zusammengefasst – und den Erwartungswerten μ_i her. Über die Score-Gleichungen (2.14) erhielten wir einen Schätzvektor $\hat{\boldsymbol{\beta}}$ für den Vektor der Modellparameter $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$.

Nun wollen wir einen anderen Weg gehen. Betrachten wir die Score-Funktion für den Parameter μ_i für n unabhängige Zufallsvariablen $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$ mit den Realisierungen $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$ aus einer Verteilung aus der Exponentialfamilie mit den Erwartungswerten $\mathbb{E}(Y_i) = \mu_i$, so erhalten wir für $1 \leq i \leq n$:

$$\begin{aligned} \frac{\partial \mathscr{L}(\boldsymbol{\theta}, \boldsymbol{y})}{\partial \mu_i} &= \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial}{\partial \theta_i} \sum_{j=1}^{n} \left(\frac{y_j \theta_j - b(\theta_j)}{a(\phi_j)} + c(y_j, \phi_j) \right) \\ &= \frac{y_i - \mu_i}{a(\phi_i) V(\mu_i)}. \end{aligned} \tag{3.1}$$

Weiters gelte ohne Beschränkung der Allgemeinheit, dass es einen gemeinsamen Dispersionsparameter $\phi_i = \phi$, $1 \leq i \leq n$ gibt und die Funktion $a(\,\cdot\,)$ der identischen Abbildung $a(\phi_i) = a(\phi) = \phi$ entspricht. Man erkennt, dass die Score-Funktion (3.1) für gegebene Beobachtungen $\boldsymbol{y}$ nur von der Varianzfunktion $V(\,\cdot\,)$ abhängt.

Die Score-Funktion (3.1) ergibt sich aus der Ableitung der Likelihood-Funktion, zu deren Bestimmung i. Allg. die Verteilung bekannt sein muss. Um die Verteilung bestimmen zu können, bedarf es entweder genauer Kenntnisse über den die Zufallsgrößen generierenden Prozess oder grundlegender Erfahrungen aus früheren Experimenten. In praktischen Anwendungen ist die genaue Spezifizierung der exakten Verteilung meist nicht möglich, wenngleich man aber über gewisse Informationen, wie beispielsweise den groben Zusammenhang zwischen Erwartungswert und Varianz oder den Einfluss gewisser Größen auf den Erwartungswert, verfügt (vgl. McCullagh und Nelder, 1989, S. 323).

In diesem Abschnitt wollen wir daher der Frage nachgehen, welche Auswirkung die Annahme einer durch die Daten motivierten Varianz–Erwartungswert-Beziehung (wie sie durch die Varianzfunktion $V(\mu)$ ausgedrückt wird), die keiner Verteilung aus der Exponentialfamilie zuzuordnen ist, auf die Theorie des GLM hat. Dazu definieren wir zuerst die *log-Quasi-Likelihood-Funktion* wie sie Wedderburn (1974) eingeführt hat.

Definition 3.1 (log-Quasi-Likelihood-Funktion, Wedderburn (1974))**.** Sei Y eine Zufallsvariable mit Erwartungswert $\mathbb{E}(Y) = \mu$ und $\operatorname{var}(Y) = \phi V(\mu)$ mit bekannter Varianzfunktion $V(\mu)$. Bezeichne y eine Realisation dieser Zufallsvariable. Dann bezeichnet die durch

$$Q(\mu, y) = \int^{\mu} \frac{y - t}{\phi V(t)}\, dt + \text{Funktion in } y \text{ und } \phi \tag{3.2a}$$

definierte Funktion $Q(\mu, y)$ die *log-Quasi-Likelihood-Funktion*. Die Ableitung der log-Quasi-Likelihood-Funktion

$$\frac{\partial Q(\mu, y)}{\partial \mu} = \frac{y - \mu}{\phi V(\mu)} \tag{3.2b}$$

heißt dementsprechend *Quasi-Score-Funktion*.

Bemerkung 1. Wedderburn (1974) bezeichnet $Q(\mu, y)$ als *Quasi-Likelihood-Funktion*. Ist also im folgenden von der Quasi-Likelihood (QL)-Funktion die Rede, ist also streng genommen die log-Quasi-Likelihood-Funktion gemeint.

Bemerkung 2. McCullagh und Nelder (1989) geben für $Q(\mu, y)$ die Beziehung

$$Q(\mu, y) = \int_{y}^{\mu} \frac{y - t}{\phi V(t)}\, dt \tag{3.3}$$

an. Damit unterscheidet sich die QL-Funktion in (3.3) von der Definition in (3.2a) nur durch die Angabe einer expliziten unteren Integrationsgrenze. Löst man das Integral auf, erhält man so zusätzlich eine Funktion in y.[1] Insbesondere sind die aus (3.2a) und (3.3) abgeleiteten Quasi-Score-Funktionen identisch.

Bemerkung 3. Sind die Komponenten y_i des Vektors $\boldsymbol{y} = (y_1, \ldots, y_n)^{\top}$ unabhängig, ergibt sich die QL-Funktion $Q(\boldsymbol{\mu}, \boldsymbol{y})$ der gesamten Stichprobe $\boldsymbol{y}$ aus der Summe der einzelnen QL-Funktionen und lautet:

$$Q(\boldsymbol{\mu}, \boldsymbol{y}) = \sum_{i=1}^{n} Q(\mu_i, y_i).$$

Bemerkung 4. Eine verallgemeinerte Definition der QL-Funktion, die auf der Annahme von

$$\mathbb{E}(\boldsymbol{Y}) = \boldsymbol{\mu}, \qquad \operatorname{cov}(\boldsymbol{Y}) = \phi \boldsymbol{V}(\boldsymbol{\mu})$$

beruht, wobei $\boldsymbol{V}$ eine Varianz-Kovarianz-Matrix in Abhängigkeit von $\boldsymbol{\mu}$ beschreibt, findet man bei McCullagh (1983) und lautet:

$$\frac{\partial Q(\boldsymbol{\mu}, \boldsymbol{y})}{\partial \boldsymbol{\mu}} = \frac{1}{\phi} \boldsymbol{V}^{-}(\boldsymbol{\mu})(\boldsymbol{y} - \boldsymbol{\mu}).$$

Dabei bezeichnet $\boldsymbol{V}^{-}$ die verallgemeinerte Inverse von $\boldsymbol{V}$. Diese verallgemeinerte Annahme erlaubt auch korrelierte Beobachtungen. Setzen wir unkorrelierte Beobachtungen voraus, führt das zu $\boldsymbol{V}(\boldsymbol{\mu}) = \operatorname{diag}(V(\mu_1), \ldots, V(\mu_n))$ und resultiert in der Definition der QL-Funktion von Wedderburn.

[1] Funktionen in y werden auch in der ursprünglichen Definition berücksichtigt.

Setzt man den Erwartungswert wie im vorherigen Abschnitt mit den erklärenden Variablen über $g(\mu) = \boldsymbol{x}^\top \boldsymbol{\beta}$ in Relation, ist man ausgehend von einer gegebenen QL-Funktion daran interessiert, einen Schätzer $\hat{\boldsymbol{\beta}}_{\mathrm{QL}}$ so zu bestimmen, dass die QL-Funktion maximiert wird. Vergleicht man nun die zur einparametrischen Exponentialfamilie gehörige Score-Funktion (3.1) mit der Quasi-Score-Funktion (3.2b) erkennt man, dass beide die selbe Gestalt aufweisen. Folglich können zur Bestimmung von $\hat{\boldsymbol{\beta}}_{\mathrm{QL}}$ die gleichen Methoden wie zur Bestimmung des ML-Schätzers $\hat{\boldsymbol{\beta}}_{\mathrm{ML}}$, der als Lösung der Score-Gleichungen (2.14) die Score-Funktion für $\boldsymbol{\beta}$ maximiert, herangezogen werden.

3.1.1 Eigenschaften der Quasi-Likelihood-Funktion

Quasi-Likelihood-Funktionen haben einige Eigenschaften mit gewöhnlichen Likelihood-Funktionen gemein, wie folgender Satz zeigt.

Satz 3.2 (Eigenschaften der QL-Funktion, (Wedderburn, 1974))**.** *Sei $Q(\mu, y)$ die Quasi-Likelihood-Funktion einer Zufallsvariable Y mit Erwartungswert $\mathbb{E}(Y) = \mu$ und $\operatorname{var}(Y) = \phi V(\mu)$. Der Erwartungswert lasse sich weiters als eine Funktion mit den Parametern $\beta_1, \ldots, \beta_p$ darstellen. Dann hat die Quasi-Likelihood-Funktion unter einfachen Regularitätsbedingungen wie in* (2.4) *folgende Eigenschaften:*

$$\mathbb{E}\left(\frac{\partial Q(\mu, y)}{\partial \mu}\right) = 0, \tag{3.4a}$$

$$\mathbb{E}\left(\frac{\partial Q(\mu, y)}{\partial \beta_j}\right) = 0, \tag{3.4b}$$

$$\mathbb{E}\left(\frac{\partial Q(\mu, y)}{\partial \mu}\right)^2 = -\mathbb{E}\left(\frac{\partial^2 Q(\mu, y)}{\partial \mu^2}\right) = \frac{1}{\phi V(\mu)}, \tag{3.4c}$$

$$\mathbb{E}\left(\frac{\partial Q(\mu, y)}{\partial \beta_j}\frac{\partial Q(\mu, y)}{\partial \beta_k}\right) = -\mathbb{E}\left(\frac{\partial^2 Q(\mu, y)}{\partial \beta_j \partial \beta_k}\right) = \frac{1}{\phi V(\mu)}\frac{\partial \mu}{\partial \beta_j}\frac{\partial \mu}{\partial \beta_k}. \tag{3.4d}$$

Beweis. (a) Trivial.

(b) Folgt aus $\mathbb{E}\left(\frac{\partial Q(\mu,y)}{\partial \beta_j}\right) = \mathbb{E}\left(\frac{\partial Q(\mu,y)}{\partial \mu}\frac{\partial \mu}{\partial \beta_j}\right)$, (a) und der Vertauschbarkeit von Integration und Differentiation.

(c) Ist ein Spezialfall von (d).

(d) Für die linke Seite von (3.4d) gilt mit $\mathbb{E}\left((y-\mu)^2\right) = \operatorname{var}(y) = \phi V(\mu)$:

$$\begin{aligned}
\mathbb{E}\left(\frac{\partial Q(\mu, y)}{\partial \beta_j}\frac{\partial Q(\mu, y)}{\partial \beta_k}\right) &= \mathbb{E}\left[\left(\frac{\partial Q(\mu, y)}{\partial \mu}\right)^2\right]\frac{\partial \mu}{\partial \beta_j}\frac{\partial \mu}{\partial \beta_k} \\
&= \mathbb{E}\left(\frac{(y-\mu)^2}{\phi^2 V(\mu)^2}\right)\frac{\partial \mu}{\partial \beta_j}\frac{\partial \mu}{\partial \beta_k} \\
&= \frac{1}{\phi V(\mu)}\frac{\partial \mu}{\partial \beta_j}\frac{\partial \mu}{\partial \beta_k}.
\end{aligned}$$

Andererseits folgt für die rechte Seite der ersten Gleichung in (3.4d):

$$\begin{aligned}
-\mathbb{E}\left(\frac{\partial^2 Q(\mu,y)}{\partial\beta_j\partial\beta_k}\right) &= -\mathbb{E}\left[\frac{\partial\mu}{\partial\beta_j}\frac{\partial}{\partial\mu}\left(\frac{y-\mu}{\phi V(\mu)}\frac{\partial\mu}{\partial\beta_k}\right)\right] \\
&= -\mathbb{E}\left[-(y-\mu)\frac{V'(\mu)}{\phi V(\mu)^2}\frac{\partial\mu}{\partial\beta_j}\frac{\partial\mu}{\partial\beta_k} - \frac{1}{\phi V(\mu)}\frac{\partial\mu}{\partial\beta_j}\frac{\partial\mu}{\partial\beta_k}\right] \\
&= \frac{1}{\phi V(\mu)}\frac{\partial\mu}{\partial\beta_j}\frac{\partial\mu}{\partial\beta_k}.
\end{aligned}$$

□

Korollar 3.3. *Falls die log-Likelihood-Funktion $\mathscr{L}$ von y in Abhängigkeit von μ definiert werden kann, gilt:*

$$-\mathbb{E}\left(\frac{\partial^2 Q(\mu,y)}{\partial\mu^2}\right) \leq -\mathbb{E}\left(\frac{\partial^2 \mathscr{L}(\mu,y)}{\partial\mu^2}\right). \tag{3.5}$$

Beweis. Siehe Wedderburn (1974). □

Bemerkung. Die Eigenschaften (3.4a)–(3.4d) treffen auch auf die „klassische“ Score-Funktion zu.

Der Vorteil der QL-Funktion liegt auf der Hand. Anstatt die Likelihood-Funktion bestimmen zu müssen, die die Kenntnis der Verteilung voraussetzt, erhält man nur über die Spezifizierung der Relation zwischen Erwartungswert und Varianz bezüglich der Score-Funktion die gleichen Resultate. Dazu muss aber die zugrunde liegende Verteilung *nicht* vollständig spezifiziert werden und damit müssen auch weniger Annahmen getroffen werden. In weiterer Konsequenz heißt das, dass die Auswahl einer Varianz- und einer Linkfunktion bereits ausreichend ist, um Schätzer für die Koeffizienten eines GLM zu bestimmen. Dies stellt eine Analogie zu dem Verhältnis zwischen der Theorie des LM und der LS-Schätzung dar: beide Methoden liefern die gleichen Schätzer, wobei die LS-Schätzung sich mit der Annahme des zweiten Moments begnügt, während das LM auf der vollständigen Spezifizierung der (Normal-)Verteilung beruht (vgl. Blough u. a., 1999).

Viele Asymptotik-Aussagen erster Ordnung für Likelihood-Funktionen basieren auf den Eigenschaften (3.4a)–(3.4d), daher hat die QL-Funktion viele asymptotische Eigenschaften mit der Likelihood-Funktion gemein (vgl. McCullagh und Nelder, 1989, S. 325). So gilt für die Taylorentwicklung der Score-Funktion um den wahren Parameter $\boldsymbol{\beta}$ für $g(\boldsymbol{\mu}) = X\boldsymbol{\beta}$ beispielsweise:

$$\left.\frac{\partial\mathscr{L}(\boldsymbol{\mu},\boldsymbol{y})}{\partial\boldsymbol{\beta}}\right|_{\hat{\boldsymbol{\beta}}_{\mathrm{ML}}} \approx \left.\frac{\partial\mathscr{L}(\boldsymbol{\mu},\boldsymbol{y})}{\partial\boldsymbol{\beta}}\right|_{\boldsymbol{\beta}} + \left.\frac{\partial^2\mathscr{L}(\boldsymbol{\mu},\boldsymbol{y})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top}\right|_{\boldsymbol{\beta}} (\hat{\boldsymbol{\beta}}_{\mathrm{ML}} - \boldsymbol{\beta}). \tag{3.6}$$

Da $\hat{\boldsymbol{\beta}}_{\mathrm{ML}}$ der ML-Schätzer von $\mathscr{L}(\boldsymbol{\mu},\boldsymbol{y})$ ist, ist die linke Seite der Approximation (3.6) gerade gleich Null. Wir ersetzen die Ableitung der Score-Funktion durch ihren Erwartungswert, womit sich mit (2.17) und (2.18) die Approximation (3.6) zu

$$\mathbf{0} \approx X^\top DW(\boldsymbol{y}-\boldsymbol{\mu}) - \underbrace{\mathbb{E}\left(X^\top\left(W - \operatorname{diag}\left(\frac{\partial\delta_i\omega_i}{\partial\eta_i}(y_i-\mu_i)\right)\right)X\right)}_{=-X^\top WX}(\hat{\boldsymbol{\beta}}_{\mathrm{ML}} - \boldsymbol{\beta}) \Rightarrow$$

$$\hat{\boldsymbol{\beta}}_{\mathrm{ML}} - \boldsymbol{\beta} \approx (X^\top WX)^{-1}X^\top DW(\boldsymbol{y}-\boldsymbol{\mu})$$

umschreiben lässt, wobei die Diagonalmatrizen W und D durch

$$W = \operatorname{diag}\left(1/[\phi V(\mu_i)(g'(\mu_i))^2]\right) \qquad \text{bzw.} \qquad D = \operatorname{diag}\left(g'(\mu_i)\right)$$

definiert sind. Damit lassen sich der Erwartungswert und die Varianz-Kovarianz-Matrix des Schätzers $\hat{\boldsymbol{\beta}}_{\mathrm{ML}}$ asymptotisch bestimmen und lauten:

$$\mathbb{E}(\hat{\boldsymbol{\beta}}_{\mathrm{ML}}) \approx (X^\top W X)^{-1} X^\top D W \underbrace{\mathbb{E}\left(\boldsymbol{y} - \boldsymbol{\mu}\right)}_{=0} + \boldsymbol{\beta} \qquad = \boldsymbol{\beta}, \tag{3.7a}$$

$$\operatorname{cov}(\hat{\boldsymbol{\beta}}_{\mathrm{ML}}) \approx (X^\top W X)^{-1} X^\top D W \underbrace{\operatorname{var}\left(y\right)}_{=(DWD)^{-1}} W D X (X^\top W X)^{-1} = (X^\top W X)^{-1}, \tag{3.7b}$$

wobei W im unbekannten Parameter $\boldsymbol{\beta}$ ausgewertet wird.

Betrachtet man nun analog die Taylorentwicklung der Quasi-Score-Funktion wieder um den wahren Parameter $\boldsymbol{\beta}$ erhält man:

$$\left.\frac{\partial Q(\boldsymbol{\mu}, \boldsymbol{y})}{\partial \boldsymbol{\beta}}\right|_{\hat{\boldsymbol{\beta}}_{\mathrm{QL}}} \approx \left.\frac{\partial Q(\boldsymbol{\mu}, \boldsymbol{y})}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}} + \left.\frac{\partial^2 Q(\boldsymbol{\mu}, \boldsymbol{y})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top}\right|_{\boldsymbol{\beta}} (\hat{\boldsymbol{\beta}}_{\mathrm{QL}} - \boldsymbol{\beta}).$$

Ersetzt man die Ableitung der Quasi-Score-Funktion durch ihren Erwartungswert, erhalten wir unter Verwendung von Gleichung (3.4d), $\tilde{W} := \operatorname{diag}\left(1/\phi V(\mu_i)\right)$, der Jacobi-Matrix

$$J := J_{\boldsymbol{\mu}}(\boldsymbol{\beta}) = \begin{pmatrix} \frac{\partial \mu_1}{\partial \beta_1} & \cdots & \frac{\partial \mu_1}{\partial \beta_p} \\ \vdots & & \vdots \\ \frac{\partial \mu_n}{\partial \beta_1} & \cdots & \frac{\partial \mu_n}{\partial \beta_p} \end{pmatrix}$$

und mit der Tatsache, dass $\hat{\boldsymbol{\beta}}_{\mathrm{QL}}$ die QL-Funktion maximiert, die Approximation:

$$\begin{gathered} \mathbf{0} \approx J^\top \tilde{W}(\boldsymbol{y} - \boldsymbol{\mu}) - J^\top \tilde{W} J(\hat{\boldsymbol{\beta}}_{\mathrm{QL}} - \boldsymbol{\beta}) \Rightarrow \\ \hat{\boldsymbol{\beta}}_{\mathrm{QL}} - \boldsymbol{\beta} \approx (J^\top \tilde{W} J)^{-1} J^\top \tilde{W}(\boldsymbol{y} - \boldsymbol{\mu}). \end{gathered}$$

Damit lauten der Erwartungswert und die Varianz-Kovarianz-Matrix des Schätzers $\hat{\boldsymbol{\beta}}_{\mathrm{QL}}$:

$$\mathbb{E}(\hat{\boldsymbol{\beta}}_{\mathrm{QL}}) \approx (J^\top \tilde{W} J)^{-1} J^\top \tilde{W} \underbrace{\mathbb{E}\left(\boldsymbol{y} - \boldsymbol{\mu}\right)}_{=0} + \boldsymbol{\beta} \qquad = \boldsymbol{\beta}, \tag{3.8a}$$

$$\operatorname{cov}(\hat{\boldsymbol{\beta}}_{\mathrm{QL}}) \approx (J^\top \tilde{W} J)^{-1} J^\top \tilde{W} \underbrace{\operatorname{var}\left(y\right)}_{=\tilde{W}^{-1}} \tilde{W} J (J^\top \tilde{W} J)^{-1} = (J^\top \tilde{W} J)^{-1}. \tag{3.8b}$$

Der QL-Schätzer $\hat{\boldsymbol{\beta}}_{\mathrm{QL}}$ ist damit auch asymptotisch erwartungstreu und die Varianz-Kovarianz-Matrix hängt vom funktionalen Zusammenhang zwischen den Erwartungswerten μ_i und den Parametern $\beta_1, \ldots, \beta_p$ ab. Wird ein GLM modelliert, also ein linearer

Zusammenhang zwischen einer Funktion des Erwartungswertes und den erklärenden Variablen ($g(\boldsymbol{\mu}) = X\boldsymbol{\beta}$) angenommen, gilt für die Jacobi-Matrix $J_{\boldsymbol{\mu}}(\boldsymbol{\beta})$:

$$J_{\boldsymbol{\mu}}(\boldsymbol{\beta}) = \begin{pmatrix} \frac{x_{11}}{g'(\mu_1)} & \cdots & \frac{x_{1p}}{g'(\mu_1)} \\ \vdots & & \vdots \\ \frac{x_{n1}}{g'(\mu_n)} & \cdots & \frac{x_{np}}{g'(\mu_n)} \end{pmatrix}$$

und damit folgt für diesen Fall wie erwartet $X^\top W X = J^\top \tilde{W} J$ und damit gilt für (3.7b) und (3.8b) Gleichheit.

Bemerkung 1. Die Gleichheit der asymptotischen Varianz-Kovarianz-Matrizen ergibt sich aus der Tatsache, dass die QL-Funktion und die log-Likelihood-Funktion die Eigenschaften (3.4) gemein haben.

Bemerkung 2. Für die Momente der Quasi-Score-Funktion $U_{\boldsymbol{\beta}} := J^\top \tilde{W}(\boldsymbol{y} - \boldsymbol{\mu})$ für den Parameter $\boldsymbol{\beta}$ folgt aus den Gleichungen (3.4b) und (3.4d):

$$\mathbb{E}\,(U_{\boldsymbol{\beta}}) = \mathbf{0} \qquad\qquad \operatorname{cov}(U_{\boldsymbol{\beta}}) = J^\top \tilde{W} J.$$

Die Varianz-Kovarianz-Matrix von $U_{\boldsymbol{\beta}}$ entspricht also gerade der Inversen der asymptotischen Varianz-Kovarianz-Matrix des Parameter-Schätzers $\hat{\boldsymbol{\beta}}_{\mathrm{QL}}$. Dies stellt somit eine Analogie zur Score-Funktion dar: die Varianz-Kovarianz-Matrix der Score-Funktion (die Fisher-Information) entspricht bei ML-Schätzung auch der Inversen der asymptotischen Varianz-Kovarianz-Matrix des Schätzvektors $\hat{\boldsymbol{\beta}}_{\mathrm{ML}}$.

Wenn wir mit $i_{\boldsymbol{\beta}}$ die Varianz-Kovarianz-Matrix von $U_{\boldsymbol{\beta}}$ bezeichnen und annehmen, dass der Grenzwert von $n^{-1} i_{\boldsymbol{\beta}}$ positiv definit ist, ergeben sich unter der Voraussetzung, dass die dritten Momente von $\boldsymbol{Y}$ endlich sind, die nachstehenden Asymptotik-Aussagen (McCullagh, 1983):[2]

$$\frac{1}{\sqrt{n}} U_{\boldsymbol{\beta}} \sim \mathtt{N}_p\big(\mathbf{0}, i_{\boldsymbol{\beta}}/n\big) + \mathcal{O}_p\big(n^{-1/2}\big),$$
$$\sqrt{n}\big(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\big) \sim \mathtt{N}_p\big(\mathbf{0}, i_{\boldsymbol{\beta}}^{-1} \cdot n\big) + \mathcal{O}_p\big(n^{-1/2}\big).$$

Analog zur Deviance kann auch für QL-Funktionen eine *Quasi-Deviance* definiert werden, die für die Bestimmung eines Maßes für den Goodness- bzw. den Lack-of-Fit herangezogen werden kann.

Definition 3.4 (Quasi-Deviance)**.** Seien $Q(\mu_i, y_i)$ die Quasi-Likelihood-Funktionen der n unabhängigen Zufallsvariablen $Y_i, 1 \leq i \leq n$, mit Erwartungswert $\mathbb{E}\,(Y_i) = \mu_i$ und $\operatorname{var}(Y_i) = \phi V(\mu_i)$. Dann bezeichnet man die Größe

$$D(\boldsymbol{y}, \boldsymbol{\mu}) = -2\phi\,(Q(\boldsymbol{\mu}, \boldsymbol{y}) - Q(\boldsymbol{y}, \boldsymbol{y})) = -2\sum_{i=1}^{n} \int_{y_i}^{\mu_i} \frac{y_i - t}{V(t)}\, dt \tag{3.9}$$

als *Quasi-Deviance*.

[2] Dabei bedeutet die sogenannte Mann-Wald-Notation $Y_n = \mathcal{O}_p(a_n)$ für eine Folge von Zufallsvariablen Y_n, dass $|Y_n/a_n|$ für $n \to \infty$ in der Wahrscheinlichkeit beschränkt ist, d. h. für $\epsilon > 0 : \exists k < \infty, n_0 \in \mathbb{N}$ sodass $\forall n > n_0 : \mathbb{P}\,(|Y_n/a_n| < k) > 1 - \epsilon$ gilt.

Bemerkung. Verwendet man für $Q(\mu, y)$ die Alternative Formulierung (3.3) ergibt sich für die Quasi-Deviance der Zusammenhang:

$$D(\boldsymbol{y}, \boldsymbol{\mu}) = -2\phi \cdot Q(\boldsymbol{\mu}, \boldsymbol{y}). \tag{3.10}$$

Ziel ist auch hier die Minimierung der Quasi-Deviance, was wiederum durch den Maximum-Quasi-Likelihood-Schätzer $\hat{\boldsymbol{\beta}}_{\mathrm{QL}}$ erreicht wird. Die asymptotische Verteilung der skalierten Deviance konnte über die Konstruktion eines Likelihood-Quotienten-Tests ermittelt werden und es stellt sich auch für die skalierte Quasi-Deviance heraus, dass diese asymptotisch χ^2_{n-p}-verteilt ist, wobei die Anzahl der Freiheitsgrade wieder der Reduktion der Dimensionen der Parameter-Räume entspricht.

McCullagh (1983) gibt den Fehler der asymptotischen Aussagen für die Deviance bzw. die Quasi-Deviance mit

$$\begin{aligned} -2\left[\mathscr{L}(\boldsymbol{\mu}, \boldsymbol{y}) - \mathscr{L}(\boldsymbol{y}, \boldsymbol{y})\right] &\sim \phi\chi^2_{n-p} + \mathcal{O}_p(n^{-1}), \\ -2\left[Q(\boldsymbol{\mu}, \boldsymbol{y}) - Q(\boldsymbol{y}, \boldsymbol{y})\right] &\sim \phi\chi^2_{n-p} + \mathcal{O}_p(n^{-1/2}) \end{aligned}$$

an, wobei die erste Aussage auf Hayakawa (1977) zurückgeht und für die Annahme einer stetig verteilten Zufallsvariable Y gilt.

Die Stärke des QL-Ansatzes beruht auf der Möglichkeit, nur durch die Angabe der Varianz- und der Linkfunktion Schätzer für ein GLM bestimmen zu können, ohne die volle Verteilung angeben zu müssen. Dies kann einerseits hilfreich sein, wenn die Varianzfunktion nicht zu einer Verteilung aus der Exponentialfamilie gehört. Andererseits können Varianz-Annahmen $\operatorname{var}(y) = \phi V(\mu)$, die zwar zu einer Verteilung aus der Exponentialfamilie gehören, aber eigentlich einen fixen Dispersionsparameter von $\phi = 1$ vorsehen, relaxiert werden, indem der Dispersionsparameter ϕ nicht als fix betrachtet wird (vgl. Nelder und Lee, 1992). Die beiden folgenden Beispiele demonstrieren die Einsatzmöglichkeiten der QL-Funktion.

Beispiel 3.5. Abbildung 3.1 zeigt ein GLM mit angenommener konstanter Varianz und identischer Linkfunktion (also wie bei einer Normalverteilungsannahme). Die Zufallszahlen kommen in diesem Beispiel allerdings aus drei verschiedenen Verteilungen (wobei die t-Verteilung nicht einmal Mitglied der Exponentialfamilie ist):

$Y_{i1} \sim t_4$	$Y_{i2} \sim \mathcal{N}(1, 2)$	$Y_{i3} \sim \Gamma(2, 1)$
$\mathbb{E}(Y_{i1}) = 0$	$\mathbb{E}(Y_{i2}) = 1$	$\mathbb{E}(Y_{i3}) = 2$
$\operatorname{var}(Y_{i1}) = 2$	$\operatorname{var}(Y_{i3}) = 2$	$\operatorname{var}(Y_{i3}) = 2.$

Da die Varianz für alle drei Gruppen tatsächlich konstant ist, erlaubt ein GLM mit konstanter Varianz-Annahme auch für diesen Fall eine Schätzung, die bereits für 30 Datenpunkte erstaunlich genau ist. Das Beispiel verdeutlicht, dass für die Schätzung beim GLM weniger die korrekte Spezifizierung der Verteilung der Zufallszahlen selbst eine Rolle spielt,[3] als vielmehr eine korrekte Spezifizierung der Erwartungswert-Varianz-Funktion, wie sie durch die QL-Funktion erreicht wird. ◊

[3] Was in diesem konstruierten Beispiel auch schwer möglich ist, da die Zufallszahlen eben nicht aus einer einzigen Verteilung kommen.

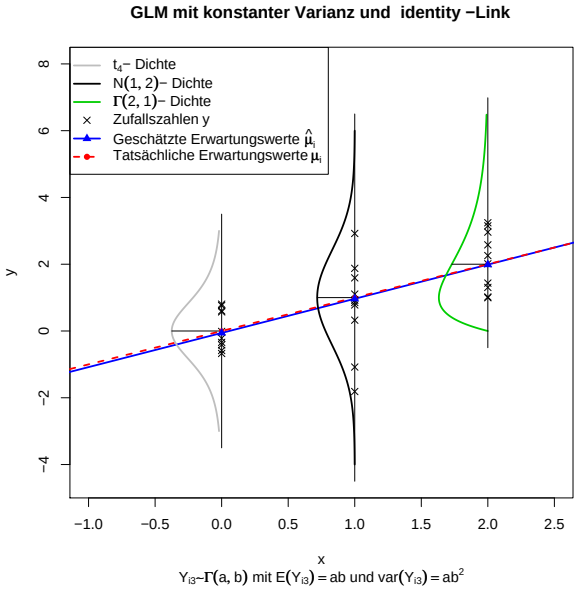


Abbildung 3.1: Generalisiertes Lineares Modell mit verschiedenen Verteilungen und konstanter Varianz

Beispiel 3.6. Ein weiteres Beispiel für einen QL-Ansatz findet man in der Annahme von negativ-binomialverteilten Response-Variablen. Für Zufallszahlen $Y \sim \texttt{NegB}\,(k, \mu)$ aus einer Negativ-Binomialverteilung ergibt sich nämlich für $\mu > 0, k > 0, y \geq 0$ die Varianz $\text{var}\,(Y) = \mu + \mu^2/k$. Wählen wir also $V(\mu) = \mu + \mu^2/k$ erhalten wir für die QL-Funktion:

$$Q(\mu, y) = \int^{\mu} \frac{y - t}{t + t^2/k}\, dt = y \log \left| \frac{\mu}{\mu + k} \right| - k \log|\mu + k|.$$

Bemerkung 1. Die Negativ-Binomialverteilung ist nur für *bekanntes* k ein Mitglied der Exponentialfamilie.

Bemerkung 2. Beinhaltet die QL-Funktion einen unbekannten Parameter k – wie für die Negativ-Binomialverteilung – ist es mit dem QL-Ansatz *nicht* möglich, diesen zu schätzen. ◊

3.1.2 Die Quasi-Dichte

Die Äquivalenz der Score-Funktion für Verteilungen aus der einparametrischen Exponentialfamilie und der Quasi-Score-Funktion legt die Frage nahe, in welchen Fällen die Quasi-Likelihood-Funktion mit einer log-Likelihood-Funktion übereinstimmt. Es stellt sich heraus, dass die Äquivalenz tatsächlich nur für Verteilungen aus der einparametrischen Exponentialfamilie gegeben ist, wie folgender Satz beweist.

Satz 3.7 (Wedderburn (1974))**.** *Bezeichne $\mathscr{L}(\mu, y)$ die log-Likelihood-Funktion einer Beobachtung y der Zufallsvariable Y mit $\mathbb{E}(Y) = \mu$ und $\operatorname{var}(Y) = \phi V(\mu)$. Dann erfüllt die Score-Funktion die Gleichung*

$$\frac{\partial \mathscr{L}(\mu, y)}{\partial \mu} = \frac{y - \mu}{\phi V(\mu)} \tag{3.11}$$

dann und nur dann, wenn die Verteilung von Y aus der einparametrischen Exponentialfamilie stammt.

Beweis. „⇒“: Es erfülle die log-Likelihood-Funktion die Gleichung (3.11). Integration über μ ergibt dann:

$$\begin{aligned}
\mathscr{L}(\mu, y) &= \int \frac{\partial \mathscr{L}(\mu, y)}{\partial \mu}\, d\mu = \int \frac{y - \mu}{\phi V(\mu)}\, d\mu \\
&= \frac{y}{\phi} \underbrace{\int \frac{1}{V(\mu)}\, d\mu}_{:=\theta} - \frac{1}{\phi} \underbrace{\int \frac{\mu}{V(\mu)}\, d\mu}_{:=b(\theta)} \\
&= \frac{y\theta - b(\theta)}{\phi} + c(y, \phi) \Rightarrow \\
f_Y(y, \mu) &= \exp\left\{ \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right\}.
\end{aligned}$$

Damit stammt die Verteilung von Y aus der einparametrischen Exponentialfamilie.

„⇐“: Für die Score-Funktion in μ gilt für eine Verteilung aus der Exponentialfamilie mit der Kettenregel und $\partial\theta/\partial\mu = 1/V(\mu)$:

$$\begin{aligned}
\frac{\partial \mathscr{L}(\mu, y)}{\partial \mu} &= \frac{\partial \theta}{\partial \mu} \frac{\partial}{\partial \theta} \left(\frac{y\theta - b(\theta)}{\phi} + c(y, \phi) \right) \\
&= \frac{y - b'(\theta)}{\phi V(\mu)} = \frac{y - \mu}{\phi V(\mu)}.
\end{aligned}$$

□

Bemerkung. Mit der Ungleichung (3.5) erhalten wir für die Varianz der Score-Funktion (die *Fisher-Information*) bezüglich des Parameters μ eine untere Schranke. Für Verteilungen aus der Exponentialfamilie folgt mit Satz 3.7 sofort:

$$-\mathbb{E}\left(\frac{\partial^2 Q(\mu, y)}{\partial \mu^2} \right) = -\mathbb{E}\left(\frac{\partial^2 \mathscr{L}(\mu, y)}{\partial \mu^2} \right).$$

Das heißt, dass für Verteilungen aus der Exponentialfamilie der ML-Schätzer für μ gerade so bestimmt wird, dass die Fisher-Information minimiert wird.

Betrachtet man die Varianz der Quasi-Score-Funktion als Maß der Informationsgüte, die die Beobachtung y über den Parameter μ liefert, wenn nur die Erwartungswert-Varianz-Beziehung bekannt ist, erhält man über

$$0 \leq \frac{1}{\operatorname{var}(Y)} = -\mathbb{E}\left(\frac{\partial^2 Q(\mu, y)}{\partial \mu^2}\right) \leq -\mathbb{E}\left(\frac{\partial^2 \mathscr{L}(\mu, y)}{\partial \mu^2}\right) \Rightarrow$$
$$\mathbb{E}\left(\frac{\partial^2 \left[Q(\mu, y) - \mathscr{L}(\mu, y)\right]}{\partial \mu^2}\right) \geq 0$$

ein Maß für die zusätzliche Information, die die vollständige Spezifizierung der Verteilung zu geben imstande ist. In anderen Worten heißt das auch, dass für Verteilungen aus der Exponentialfamilie die Spezifizierung der Varianzfunktion $V(\mu)$ allein bereits das Maximum an Information liefert. Die vollständige Spezifizierung der Verteilung liefert in diesem Fall keine zusätzliche Information (vgl. Wedderburn, 1974).

Satz 3.7 sagt aus, dass die Likelihood-Funktion und die QL-Funktion nur für Varianzfunktionen, die zu einer Verteilungen aus der Exponentialfamilie gehören, übereinstimmen. Definiert man nun eine QL-Funktion $Q(\mu, y)$ mittels einer Varianz-Annahme, die keiner Verteilung aus der Exponentialfamilie zuzuordnen ist, erhält man demnach keine Dichte-Funktion. Es ist aber möglich, zu einer gegebenen QL-Funktion $Q(\mu, y)$ eine sogenannte *Quasi-Dichte* zu konstruieren, indem man $Q(\mu, y)$ normalisiert (Nelder und Lee, 1992). Die folgende Definition gibt an, wie eine solche Dichte definiert werden kann.

Definition 3.8 (Quasi-Dichte)**.** Sei $Q(\mu, y)$ eine QL-Funktion. Mit der Normalisierungsfunktion $\omega(\mu) = \int \exp Q(\mu, y)\, dy$ heißt die durch

$$f_q(y, \mu) := \exp\{Q(\mu, y)\}/\omega(\mu) \tag{3.12}$$

definierte Funktion *Quasi-Dichte*.

Bemerkung. Mit Satz 3.7 folgt sofort $\omega(\mu) = 1$, falls die QL-Funktion auf einer Varianz-Annahme beruht, die zu einer Verteilung aus der Exponentialfamilie gehört.

Natürlich kann zu der Quasi-Dichte auch eine log-Likelihood-Funktion gebildet werden:

$$\mathscr{L}_q(\mu, y) = \log f_q(y, \mu) = Q(\mu, y) - \log \omega(\mu),$$

womit sich die Differenz aus Quasi-Score- und Score-Funktion aus

$$\begin{aligned}\frac{\partial Q(\mu, y)}{\partial \mu} - \frac{\partial \mathscr{L}_q(\mu, y)}{\partial \mu} &= \frac{\partial \log \omega(\mu)}{\partial \mu} \\ &= \frac{1}{\omega(\mu)} \frac{\partial \omega(\mu)}{\partial \mu} \\ &= \frac{1}{\omega(\mu)} \frac{\partial}{\partial \mu} \int \exp\{Q(\mu, y)\}\, dy\end{aligned}$$

ergibt, und bei Vertauschbarkeit von Integration und Differentiation folgt weiters:

$$\begin{aligned}\frac{\partial Q(\mu,y)}{\partial \mu} - \frac{\partial \mathscr{L}_q(\mu,y)}{\partial \mu} &= \int \frac{1}{\omega(\mu)} \frac{\partial Q(\mu,y)}{\partial \mu} \exp\{Q(\mu,y)\}\, dy \\ &= \int \frac{y-\mu}{\phi V(\mu)} \underbrace{\frac{\exp\{Q(\mu,y)\}}{\omega(\mu)}}_{\overset{(3.12)}{=} f_q(y,\mu)}\, dy \\ &= \mathbb{E}_q\left[\frac{y-\mu}{\phi V(\mu)}\right] = \frac{\mu_q - \mu}{\phi V(\mu)}.\end{aligned}$$

Der Erwartungswert μ_q bezüglich der Quasi-Dichte heißt auch *Quasi-Mean*. Ist die Differenz zwischen Quasi-Mean und Erwartungswert, $\mu_q - \mu$, verglichen mit der Differenz $y - \mu$ klein, liegt der Maximum-Quasi-Likelihood-Schätzer nahe beim ML-Schätzer bezüglich der Quasi-Dichte.

3.2 Eine Erweiterung der Quasi-Likelihood-Funktion

Die Einführung der QL-Funktion für das GLM erlaubt die Spezifizierung von Varianzfunktionen, die keiner Verteilung aus der Exponentialfamilie zuzuordnen sind. Dabei muss die Varianzfunktion bis auf eine multiplikative Konstante exakt angegeben werden. Stammt die Varianzfunktion $V(\mu)$ aus einer Familie von Funktionen $\mathcal{F}_{\boldsymbol{\theta}}$ mit einem *unbekannten* Parametervektor $\boldsymbol{\theta}$, sind die bis jetzt vorgestellten Methoden *nicht* in der Lage, den Parametervektor $\boldsymbol{\theta}$ simultan mit dem Parametervektor $\boldsymbol{\beta}$ des linearen Modells zu schätzen. Das heißt mit anderen Worten, dass der direkte Vergleich unterschiedlicher Varianzfunktionen mit den bisher diskutierten Verfahren nicht möglich ist. Außerdem erlauben die QL-Methoden auch nicht die Schätzung von nicht konstanten und unbekannten Dispersionsparametern.

Um auch verschiedene Varianzfunktionen miteinander vergleichen zu können, bedarf es einer weiteren Anpassung. Die Extended-Quasi-Likelihood (EQL)-Funktion geht auf Nelder und Pregibon (1987) zurück und erlaubt derartige Vergleiche.

Definition 3.9 (EQL-Funktion, Nelder und Pregibon (1987))**.** Sei Y eine Zufallsvariable mit Erwartungswert $\mathbb{E}(Y) = \mu$ und $\operatorname{var}(Y) = \phi V(\mu)$ mit bekannter Varianzfunktion $V(\mu)$. Bezeichne y eine Realisation dieser Zufallsvariable mit der QL-Funktion $Q(\mu, y) = \int^{\mu} (y-t)/(\phi V(t))\, dt$ und der Quasi-Deviance $D(y,\mu) = -2\phi\,(Q(\mu,y) - Q(y,y))$. Dann heißt die durch

$$Q^{+}(\mu,\phi,y) = -\frac{1}{2}\log\{2\pi\phi V(y)\} - \frac{1}{2} D(y,\mu)/\phi \tag{3.13}$$

definierte Funktion die *Extended-Quasi-Likelihood-Funktion*.

Bemerkung 1. Die EQL-Funktion beruht wie die QL-Funktion nur auf der Annahme der ersten beiden Momente von Y. Insbesondere muss die Verteilung von Y *nicht* vollständig spezifiziert sein.

Bemerkung 2. Für die n unabhängigen Zufallsvariablen Y_i ist die EQL-Funktion für die gesamte Stichprobe $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$ durch

$$Q^+(\boldsymbol{\mu}, \phi, \boldsymbol{y}) = \sum_{i=1}^{n} Q^+(\mu_i, \phi, y_i)$$

gegeben, wobei $\mathbb{E}(Y_i) = \mu_i$ gilt.

Bemerkung 3. Manche Autoren bezeichnen $-2Q^+(\boldsymbol{\mu}, \phi, \boldsymbol{y})$ auch als *Extended-Quasi-Deviance* (vgl. z.B Nelder und Lee, 1991).

Lässt sich der Erwartungswert μ durch eine Funktion mit dem Parametervektor $\boldsymbol{\beta}$ beschreiben, kann ein Schätzer für diesen Vektor durch die Maximierung von Q^+ bestimmt werden. Der erste Term von $Q^+(\mu, \phi, y)$ ist unabhängig von μ (und damit auch von $\boldsymbol{\beta}$), daher entspricht eine Maximierung von $Q^+(\mu, \phi, y)$ der Minimierung der (Quasi-) Deviance $D(y, \mu)$. Das heißt mit anderen Worten: derjenige Schätzer von $\boldsymbol{\beta}$, der die EQL-Funktion $Q^+(\mu, \phi, y)$ maximiert, ist der gleiche, den man erhält, wenn man die QL-Funktion $Q(\mu, y)$ maximiert.

Einen Schätzer für den (gemeinsamen) Dispersionsparameter ϕ erhält man durch die Maximierung von $Q^+(\boldsymbol{\mu}, \phi, \boldsymbol{y})$ nach ϕ:

$$\begin{aligned}
\frac{\partial Q^+(\boldsymbol{\mu}, \phi, \boldsymbol{y})}{\partial \phi} &= \sum_{i=1}^{n} \left(-\frac{1}{2} \frac{2\pi V(y_i)}{2\pi\phi V(y_i)} + \frac{1}{2} \frac{D(y_i, \mu_i)}{\phi^2} \right) \\
&= -\frac{n}{2\phi} + \frac{D(\boldsymbol{y}, \boldsymbol{\mu})}{2\phi^2} \stackrel{!}{=} 0 \Rightarrow \\
\hat{\phi} &= D(\boldsymbol{y}, \boldsymbol{\mu})/n.
\end{aligned}$$

Da der wahre Parameter $\boldsymbol{\mu}$ natürlich nicht bekannt ist, verwendet man stattdessen den Schätzer $\hat{\boldsymbol{\mu}}$, und man erhält somit als Schätzer für den Dispersionsparameter ϕ die mittlere Deviance:

$$\hat{\phi} = D(\boldsymbol{y}, \hat{\boldsymbol{\mu}})/n. \tag{3.14}$$

Bemerkung. Anstatt der mittleren Deviance findet auch die Bias korrigierte Variante Verwendung. Diese lautet:

$$\hat{\phi} = D(\boldsymbol{y}, \hat{\boldsymbol{\mu}})/(n - p),$$

wobei p die Anzahl der Parameter im linearen Prädiktor beschreibt.

Für Verteilungen aus der Exponentialfamilie gilt mit Satz 3.7 für die Quasi-Score- und die Score-Funktion Äquivalenz. Für die EQL-Funktion ist das nicht zwingend der Fall. Es stellt sich daher die Frage, in welchen Fällen auch die EQL-Funktion mit einer log-Likelihood-Funktion übereinstimmt.

Nelder und Pregibon (1987) zeigen, dass für bestimmte Verteilungen aus der Exponentialfamilie tatsächlich Äquivalenz gilt und geben für die anderen Verteilungen an, inwieweit sich die log-Likelihood-Funktion von der EQL-Funktion unterscheidet:

- Für die *Normalverteilung* und die *Inverse-Gauß-Verteilung* gilt Äquivalenz:

$$\begin{aligned}
V(\mu)=1,\phi=\sigma^2: \qquad Q^+(\mu,\phi,y) &= -\frac{1}{2}\log\{2\pi\phi V(y)\} - \frac{1}{2}D(y,\mu)/\phi \\
&= \log\frac{1}{\sqrt{2\pi\cdot\sigma^2\cdot 1}} - \frac{1}{2\sigma^2}(y-\mu)^2 \\
&\widehat{=} \text{ log-Likelihood einer Normalverteilung,} \\
V(\mu)=\mu^3,\phi=\sigma^2: \qquad Q^+(\mu,\phi,y) &= -\frac{1}{2}\log\{2\pi\phi V(y)\} - \frac{1}{2}D(y,\mu)/\phi \\
&= \log\frac{1}{\sqrt{2\pi\cdot\sigma^2\cdot y^3}} - \frac{1}{2\sigma^2}\frac{(y-\mu)^2}{\mu^2 y} \\
&\widehat{=} \text{ log-Likelihood einer Inversen-Gauß-Verteilung.}
\end{aligned}$$

  Damit ist der Schätzer für $\hat{\phi}$ in Gleichung (3.14) für diese beiden Verteilungen auch der ML-Schätzer.

- Für die *Poisson-*, die *standardisierte Binomial-* und die *Negativ-Binomialverteilung* erhält man die EQL-Funktion aus der jeweiligen log-Likelihood-Funktion in dem man die enthaltenen Fakultäten jeweils durch die Stirling-Approximation gemäß

$$n! \approx n^n \mathrm{e}^{-n}\sqrt{2\pi n} \qquad (*)$$

  ersetzt. Exemplarisch gilt daher für die Poissonverteilung mit $V(\mu)=\mu, \phi=1$:

$$\begin{aligned}
Q^+(\mu,\phi,y) &= -\frac{1}{2}\log\{2\pi\phi V(y)\} - \frac{1}{2}D(y,\mu)/\phi \\
&= -\frac{1}{2}\log(2\pi\cdot 1\cdot y) - y\log\frac{y}{\mu} + y - \mu.
\end{aligned}$$

  Betrachtet man die log-Likelihood-Funktion der Poissonverteilung erhält man:

$$\begin{aligned}
\mathscr{L}(\mu,y) &= y\log\mu - \mu - \log y! \\
&\overset{(*)}{\approx} y\log\mu - \mu - \log\left(y^y\mathrm{e}^{-y}\sqrt{2\pi y}\right) \\
&= y\log\mu - \mu - y\log y + y - \frac{1}{2}\log(2\pi y) \\
&= -\frac{1}{2}\log(2\pi y) - y\log\frac{y}{\mu} + y - \mu,
\end{aligned}$$

  was wiederum der EQL-Funktion entspricht.

- Für die *Gammaverteilung* mit der Parametrisierung wie in Abschnitt 2.1.3 unterscheidet sich die EQL-Funktion von der log-Likelihood-Funktion um einen Faktor,

der nur von ϕ abhängt. Mit $V(\mu) = \mu^2, \phi = 1/\nu$ gilt:

$$\begin{aligned} Q^+(\mu, \phi, y) &= -\frac{1}{2}\log\{2\pi\phi V(y)\} - \frac{1}{2}D(y,\mu)/\phi \\ &= -\frac{1}{2}\log\left(2\pi \cdot 1/\nu \cdot y^2\right) - \nu\left(-\log\frac{y}{\mu} + \frac{y-\mu}{\mu}\right) \\ &= -\log\sqrt{2\pi} + \log\sqrt{\nu} - \log y + \nu\log y - \nu\log\mu - \nu\frac{y}{\mu} + \nu \\ &= -\nu\frac{y}{\mu} - \nu\log\mu + (\nu-1)\log y + \underbrace{\log\sqrt{\nu} + \nu - \log\sqrt{2\pi}}_{=:g(\nu)}. \end{aligned}$$

Betrachtet man die log-Likelihood-Funktion der Gammaverteilung erhält man:

$$\mathscr{L}(\mu, \nu, y) = -\nu\frac{y}{\mu} - \nu\log\mu + (\nu-1)\log y + \underbrace{\nu\log\nu - \log\Gamma(\nu)}_{=:\tilde{g}(\nu)}$$

und damit unterscheiden sich die beiden Funktionen nur um die Größe $g(\nu) - \tilde{g}(\nu)$. Verwendet man auch für die Gamma-Funktion die Stirling-Approximation

$$\Gamma(x) = x^x \mathrm{e}^{-x}\sqrt{2\pi/x}\,(1 + \mathcal{O}(1/x)), \qquad (**)$$

erhält man für die Differenz zwischen der log-Likelihood- und der EQL-Funktion:

$$\begin{aligned} g(\nu) - \tilde{g}(\nu) &= \log\sqrt{\nu} + \nu - \log\sqrt{2\pi} - (\nu\log\nu - \log\Gamma(\nu)) \\ &\overset{(**)}{\approx} \log\sqrt{\nu} + \nu - \log\sqrt{2\pi} - \left[\nu\log\nu - \log\left(\nu^\nu \mathrm{e}^{-\nu}\sqrt{2\pi/\nu}\right)\right] \\ &= \log\sqrt{\nu} + \nu - \log\sqrt{2\pi} - \left[\nu\log\nu - \nu\log\nu + \nu - \log\sqrt{2\pi} + \log\sqrt{\nu}\right] \\ &= 0. \end{aligned}$$

Bemerkung. Für die Poisson-, die (nicht standardisierte) Binomial- und die Negativ-Binomialverteilung gilt insbesondere für $y = 0 : V(y) = 0$. Die EQL-Funktion in Gleichung (3.13) ist aber für $V(y) = 0$ nicht definiert, während die jeweiligen Likelihood-Funktionen in $y = 0$ sehr wohl definiert sind. Die Ursache dafür liegt in der Verwendung der Stirling–Approximation. Diese ergibt für $y = 0$ die Annäherung $0! \approx 0$. Zieht man stattdessen die modifizierte Form

$$n! \approx n^n \mathrm{e}^{-n}\sqrt{2\pi(n+c)}$$

heran, erhält man für $c = 1/6 : 0! \approx 1.023$. Verwendet man die modifizierte Stirling-Approximation, muss man sich je nach Verteilung anderer Varianzfunktionen bedienen. Tabelle 3.1 zeigt, wie die Varianzfunktion modifiziert werden muss, wobei diese (lineare) Transformation die Schätzer für $\boldsymbol{\beta}$ und ϕ nicht ändert (vgl. Nelder und Pregibon, 1987).

Tabelle 3.1: Modifizierte Varianz-Funktionen

Verteilung	$V(y)$	$V(y,c)$
Poisson	y	$y+c$
Binomial	$\frac{y(n-y)}{n}$	$\frac{(y+c)(n-y+c)}{n+c}$
Negativ-Binomial	$\frac{y(y+\nu)}{\nu}$	$\frac{(y+\nu)^2(y+c)(\nu+c)}{\nu^2(y+\nu+c)}$

Nelder und Pregibon (1987) motivieren die Verwendung der EQL-Funktion mit deren „partial likelihood flavour“ (also in etwa „Likelihood-ähnlichem Verhalten“), das darauf beruht, dass man – wie auch schon für die QL-Funktion – für die EQL-Funktion eine Dichte finden kann, indem man $\exp\{Q^+\}$ normalisiert.[4] Die Verwendung der nicht normalisierten EQL-Funktion, lässt sich dann dadurch rechtfertigen, dass sich der Normalisierungsfaktor (der in der Regel von μ, ϕ und allfälligen weiteren Parametern θ der Varianzfunktion abhängt) in manchen Fällen auch für große Änderung in den Parametern nur geringfügig ändert. Nelder und Lee (1992) vergleichen beispielsweise die Schätzer für die Dispersion, die bei ML- und EQL-Schätzung bei einer NBα-Verteilung[5] und einer Poisson-Inverse-Gauß-Mischverteilung resultieren und kommen zu dem Ergebnis, dass der EQL-Schätzer in den meisten von ihnen getesteten Parameterkonfigurationen sogar einen kleineren mittleren quadratischen Fehler aufweist als der ML-Schätzer.

Eine andere Herleitung findet man bei McCullagh und Nelder (1989, S. 349 ff.) und ergibt sich aus der Idee, die QL-Funktion $Q(\mu, y)$ derart zu einer Funktion $Q^+(\mu, \phi, y)$ zu erweitern, dass Q^+ einerseits für einen bekannten Dispersionsparameter ϕ die gleichen Eigenschaften besitzt wie die Quasi-Likelihood-Funktion selbst und andererseits für unbekanntes ϕ die Eigenschaften einer gewöhnlichen log-Likelihood-Funktion bezüglich ϕ „erbt“. Für die folgenden Überlegungen wollen wir die Existenz der Momente bis zur vierten Ordnung annehmen. Außerdem legen wir die Definition der Quasi-Deviance wie in (3.10) zugrunde.

Für die Erweiterung der QL-Funktion wählen wir den Ansatz:

$$\begin{aligned} Q^+(\mu, \phi, y) &= Q(\mu, y) + h(\phi, y) \\ &= -\frac{D(y, \mu)}{2\phi} + h(\phi, y), \end{aligned}$$

für eine Funktion $h(\phi, y) = -h_1(\phi)/2 - h_2(y)$. Damit sich Q^+ bezüglich ϕ wie eine log-

[4] Das heißt man bestimmt einen Faktor ω für den $\int_{\mathbb{R}} \omega \exp\{Q^+\}\, dy = 1$ gilt.

[5] Die NBα-Verteilung ist eine Variante der Negativ-Binomialverteilung. Letztere kann man als stetige Mischung einer Poissonverteilung mit Parameter μ auffassen, wobei μ selbst einer Gammaverteilung mit den Parametern α und ν folgt. Für ein GLM betrachtet man ν als fix, während μ variiert. Somit ergibt sich für die Negativ-Binomialverteilung ein Erwartungswert von $\mathbb{E}(Y) = \mu = \alpha\nu$ und eine Varianz von $\operatorname{var}(Y) = \mu + \mu^2/\nu$. Nimmt man hingegen an, dass μ und ν variieren, während α konstant bleibt, erhält man $\operatorname{var}(Y) = \mu(1+\alpha)$, was einer Poissonvarianz mit Überdispersion entspricht. Diese Verteilung wird als NBα-Verteilung bezeichnet (vgl. Nelder und Lee, 1992).

Likelihood-Funktion verhält, muss für die erwartete Score-Funktion bezüglich ϕ

$$\mathbb{E}\left[\frac{\partial Q^+(\mu,\phi,y)}{\partial\phi}\right] = \frac{1}{2\phi^2}\mathbb{E}\left[D(y,\mu)\right] - \frac{1}{2}\frac{\partial h_1(\phi)}{\partial\phi} \stackrel{!}{=} 0$$

gelten. Daraus folgt unmittelbar

$$\phi^2\frac{\partial h_1(\phi)}{\partial\phi} = \mathbb{E}\left[D(y,\mu)\right].$$

Eine Taylorentwicklung für $D = D(y) = D(y,\mu)$ in der ersten Komponente liefert für den Entwicklungspunkt $y = \mu$ die Abschätzung:

$$\begin{aligned} D \approx D(\mu) + (y-\mu)\left.\frac{\partial D(y)}{\partial y}\right|_{y=\mu} + \frac{1}{2}(y-\mu)^2\left.\frac{\partial^2 D(y)}{\partial y^2}\right|_{y=\mu} + \frac{1}{6}(y-\mu)^3\left.\frac{\partial^3 D(y)}{\partial y^3}\right|_{y=\mu} \\ + \frac{1}{24}(y-\mu)^4\left.\frac{\partial^4 D(y)}{\partial y^4}\right|_{y=\mu}. \end{aligned} \tag{3.15}$$

Die Deviance lässt sich zu

$$\begin{aligned} D(y,\mu) &= 2\int_\mu^y \frac{y-t}{V(t)}\,dt = 2y\int_\mu^y \underbrace{\frac{1}{V(t)}}_{=:f_1(t)}\,dt - 2\int_\mu^y \underbrace{\frac{t}{V(t)}}_{=:f_2(t)}\,dt \\ &= 2y(F_1(y) - F_1(\mu)) - 2(F_2(y) - F_2(\mu)) \end{aligned}$$

umschreiben, wobei $F_1(\,\cdot\,)$ und $F_2(\,\cdot\,)$ die Stammfunktionen von $f_1(\,\cdot\,)$ respektive $f_2(\,\cdot\,)$ bezeichnen. Die Ableitungen der Deviance lauten mit $V'(\mu) = \partial V(\mu)/\partial\mu$ und $V''(\mu) = \partial^2 V(\mu)/\partial\mu^2$ daher:

$$\begin{aligned} \frac{\partial D(y)}{\partial y} &= 2(F_1(y) - F_1(\mu)) + \underbrace{2yf_1(y) - 2f_2(y)}_{=0} = 2(F_1(y) - F_1(\mu)), \\ \frac{\partial^2 D(y)}{\partial y^2} &= 2f_1(y) = \frac{2}{V(y)}, \\ \frac{\partial^3 D(y)}{\partial y^3} &= 2f_1'(y) = -\frac{2V'(y)}{V(y)^2}, \\ \frac{\partial^4 D(y)}{\partial y^4} &= 2f_1''(y) = \frac{2}{V(y)^3}(2V'(y)^2 - V(y)V''(y)). \end{aligned} \tag{3.16}$$

Setzt man die Identitäten (3.16) in die Approximation (3.15) ein, erhält man mit $D(\mu) = 2\mu(F_1(\mu) - F_1(\mu)) - 2(F_2(\mu) - F_2(\mu)) = 0$ und $D'(\mu) = 2(F_1(\mu) - F_1(\mu)) = 0$:

$$D \approx (y-\mu)^2\frac{1}{V(\mu)} - \frac{1}{3}(y-\mu)^3\frac{V'(\mu)}{V(\mu)^2} + \frac{1}{12}(y-\mu)^4\frac{2V'(\mu)^2 - V(\mu)V''(\mu)}{V(\mu)^3}. \tag{3.17}$$

Wir bezeichnen mit $\mu_k = \mathbb{E}\left((y-\mu)^k\right)$ das k-te zentrale Moment; somit gilt $\operatorname{var}(Y) = \mu_2 = \phi V(\mu)$. Für den Erwartungswert von (3.17) folgt daher für $V = V(\mu)$:

$$\begin{aligned}\mathbb{E}(D) &\approx \phi V \frac{1}{V} - \frac{1}{3}\frac{V'}{V^2}\mu_3 + \frac{1}{12}\frac{2V'^2 - VV''}{V^3}\mu_4 \\ &= \phi + \frac{1}{12V^2}\left(-4\mu_3 V' + 2\mu_4 V'^2/V - \mu_4 V''\right). \end{aligned} \tag{3.18}$$

Vernachlässigt man die Momente höherer Ordnung, erhält man aus Gleichung (3.18) die approximative Aussage $\mathbb{E}[D(y,\mu)] \approx \phi$ und daraus folgt:

$$\begin{aligned} h_1(\phi) &= \log\phi + \text{Konstante}, \\ Q^+(\mu,\phi,y) &\approx -\frac{1}{2}D(y,\mu)/\phi - \frac{1}{2}\log\phi. \end{aligned} \tag{3.19}$$

Die Approximation (3.19) unterscheidet sich von Gleichung (3.13) also nur um eine additive Funktion in y.

Verzichtet man in Gleichung (3.18) nicht auf die höheren Momente, lassen sich noch genauere Ergebnisse erzielen. Über die Beziehungen $\mu_4 = \kappa_4 + 3\kappa_2^2 = \kappa_4 + 3\phi^2V^2$ und $\mu_3 = \kappa_3$, wobei κ_i die i-te Kumulante bezeichnet, erhält man aus der Approximation (3.18):

$$\mathbb{E}(D) \approx \phi + \frac{1}{12V^2}\left(-4\kappa_3 V' + 6\phi^2 VV'^2 - 3\phi^2V^2V'' + 2\kappa_4 V'^2/V - \kappa_4 V''\right),$$

und nach Weglassung der Kumulante vierter Ordnung resultiert:

$$\mathbb{E}(D) \approx \phi + \frac{1}{12V^2}\left(-4\kappa_3 V' + 6\phi^2 VV'^2 - 3\phi^2V^2V''\right). \tag{3.20}$$

Lässt sich zusätzlich die Beziehung

$$\kappa_{k+1} = \frac{\partial\kappa_k}{\partial\mu}\kappa_2 = \kappa_k'\kappa_2, \quad \text{für } k \geq 2, \tag{3.21}$$

für Kumulanten bis zur vierten Ordnung rechtfertigen, lässt sich die Approximation (3.20) noch weiter vereinfachen und wir erhalten mit der k-ten standardisierten Kumulante $\rho_k = \kappa_k/\kappa_2^{k/2}$:

$$\begin{aligned}\mathbb{E}(D) &\approx \phi + \frac{1}{12V^2}\left(-4\kappa_3 V' + 6\phi^2 VV'^2 - 3\phi^2V^2V''\right) \\ &= \phi + \frac{1}{12V^2}\left(-4\phi^2 VV'^2 + 6\phi^2 VV'^2 - 3\phi^2V^2V''\right) \\ &= \phi + \left(5\phi^2V'^2/V - 3\phi^2V'' - 3\phi^2V'^2/V\right)/12 \\ &= \phi + \phi(5\underbrace{\phi V'^2/V}_{=\rho_3^2} - 3\underbrace{\phi(V''V + V'^2)/V}_{=\rho_4})/12 \\ &= \phi\left\{1 + \phi(5\rho_3^2 - 3\rho_4)/12\right\}. \end{aligned}$$

Bemerkung 1. Für Verteilungen aus der Exponentialfamilie trifft die Eigenschaft (3.21) stets zu. Für die k-te Kumulante einer Verteilung aus der Exponentialfamilie gilt nämlich mit Gleichung (2.6):

$$\begin{aligned}\kappa_k &= a(\phi)^{k-1}\frac{\partial^k b(\theta)}{\partial\theta^k},\\ &= \underbrace{a(\phi)^{k-2}\frac{\partial^k b(\theta)}{\partial\theta^k}.\frac{\partial\theta}{\partial\mu}}_{=\kappa'_{k-1}} \underbrace{a(\phi)\frac{\partial\mu}{\partial\theta}}_{=a(\phi)b''(\theta)=\mathrm{var}(Y)=\kappa_2}\\ &= \frac{\partial\kappa_{k-1}}{\partial\mu}\kappa_2.\end{aligned}$$

Bemerkung 2. Trifft die Eigenschaft (3.21) zu, ergeben sich für die (standardisierten) Kumulanten folgende Abschätzungen:

$$\kappa_3 = \mathcal{O}(\phi^2) \qquad \rho_3 = \mathcal{O}(\sqrt{\phi}),$$
$$\kappa_4 = \mathcal{O}(\phi^3) \qquad \rho_4 = \mathcal{O}(\phi).$$

Analog lassen sich nun auch die Varianz $\mathrm{var}\,(D)$ und die Kovarianz-Matrix $\mathrm{cov}\,(D,Y)$ approximieren:

$$\mathrm{var}\,(D) \approx 2\kappa_2^2/V^2 = 2\phi^2 \qquad \mathrm{cov}\,(D,Y) \approx (\kappa_3 - \kappa_2'\kappa_2)/V. \tag{3.22}$$

Bemerkung. Für Verteilungen aus der Exponentialfamilie folgt aus $\kappa_3 = \kappa_2'\kappa_2$, dass die Kovarianz $\mathrm{cov}\,(D,Y)$ verschwindet.

Aus der Approximation (3.19) lassen sich nun die ϕ- und die μ-Score-Funktion bestimmen:

$$\frac{\partial Q^+(\mu,\phi,y)}{\partial\mu} = \frac{y-\mu}{\phi V(\mu)}, \tag{3.23a}$$
$$\frac{\partial Q^+(\mu,\phi,y)}{\partial\phi} = \frac{D(y,\mu)}{2\phi^2} - \frac{1}{2\phi}, \tag{3.23b}$$

wobei Gleichung (3.23a) allgemein und Gleichung (3.23b) nur für hinreichend kleine ϕ gilt.[6] Die Score-Funktion (3.23a) entspricht dabei der bereits bekannten (Quasi-) Score-Funktion für μ. Beide Scores haben einen Erwartungswert von Null und für deren Varianz respektive deren Kovarianz gilt für $Q_\mu^+ := \partial Q^+/\partial\mu$ und $Q_\phi^+ := \partial Q^+/\partial\phi$ mit (3.22):

$$\begin{aligned}\mathrm{var}\left(Q_\mu^+\right) &= \mathrm{var}\left(\frac{y-\mu}{\phi V(\mu)}\right) = \frac{\phi V(\mu)}{\phi^2 V(\mu)^2} = \frac{1}{\phi V(\mu)},\\ \mathrm{var}\left(Q_\phi^+\right) &= \mathrm{var}\left(\frac{D(y,\mu)}{2\phi^2} - \frac{1}{2\phi}\right) \approx \frac{2\phi^2}{4\phi^4} = \frac{1}{2\phi^2},\\ \mathrm{cov}\left(Q_\mu^+, Q_\phi^+\right) &= \mathrm{cov}\left(\frac{y-\mu}{\phi V(\mu)}, \frac{D(y,\mu)}{2\phi^2} - \frac{1}{2\phi}\right) \approx \frac{\kappa_3 - \kappa_2'\kappa_2}{2\phi^3 V(\mu)^2}.\end{aligned} \tag{3.24}$$

[6] Widrigenfalls ließe sich die Abschätzung $\mathbb{E}\,(D) = \phi$ nicht rechtfertigen.

Die zweiten Ableitungen der EQL-Funktion lauten außerdem:

$$\begin{aligned}\frac{\partial^2 Q^+(\mu,\phi,y)}{\partial\mu^2} &= \frac{-\phi V(\mu) - (y-\mu)\phi V(\mu)}{\phi^2 V(\mu)^2},\\ \frac{\partial^2 Q^+(\mu,\phi,y)}{\partial\phi^2} &= -\frac{D(y,\mu)}{\phi^3} + \frac{1}{2\phi^2},\\ \frac{\partial^2 Q^+(\mu,\phi,y)}{\partial\mu\partial\phi} &= -\frac{y-\mu}{\phi^2 V(\mu)},\end{aligned} \tag{3.25}$$

womit sich schließlich der negative Erwartungswert der zu einer Matrix zusammengefassten Elemente in Gleichung (3.25) durch

$$\begin{pmatrix} \frac{1}{\phi V(\mu)} & 0 \\ 0 & \frac{1}{2\phi^2} \end{pmatrix} \tag{3.26}$$

ausdrücken lässt. Die Diagonalelemente der Matrix (3.26) entsprechen dabei genau den Varianzen der Score-Funktionen in Gleichung (3.24). In Analogie zu der Fisher-Information bezeichnen wir diese Größe als *Quasi-Fisher-Information*.

Gilt für die zweite und dritte Kumulante die Abschätzung

$$\kappa_3 - \kappa_2'\kappa_2 = \mathcal{O}(\phi^2),$$

folgt für die Korrelation der beiden Score-Funktionen Q_μ^+ und Q_ϕ^+:

$$\begin{aligned}\operatorname{cor}\left(Q_\mu^+, Q_\phi^+\right) &= \frac{\operatorname{cov}\left(Q_\mu^+, Q_\phi^+\right)}{\sqrt{\operatorname{var}\left(Q_\mu^+\right)\operatorname{var}\left(Q_\phi^+\right)}}\\ &= \frac{\kappa_3 - \kappa_2'\kappa_2}{2\phi^3 V(\mu)^2}\cdot 2\phi\sqrt{\phi V(\mu)}\\ &= \mathcal{O}(\sqrt{\phi}),\end{aligned}$$

was eine zu vernachlässigende Größe darstellt. Zusammenfassend heißt das also, dass unter den Annahmen

1. Der Dispersionsparameter ϕ ist hinreichend klein.
2. Die Kumulanten κ_k sind von der Größenordnung $\mathcal{O}(\phi^{k-1})$

$Q^+(\mu,\phi,y)$ die Eigenschaften der QL-Funktion $Q(\mu,y)$ bezüglich *beider* Parameter μ und ϕ besitzt.

Bemerkung. Eine weitere Herleitung ergibt sich aus der Anwendung der *Sattelpunkt-Approximation*, welche in Kapitel 4 besprochen wird.

In den folgenden Abschnitten werden nun verschiedene Einsatzmöglichkeiten der EQL-Funktion diskutiert. Dabei hilft die EQL-Funktion insbesondere die Parameter zu schätzen, die Nelder und Pregibon (1987) als „nicht lineare Parameter“ bezeichnen – also all jene Parameter, die nicht im linearen Prädiktor vorkommen.

3.2.1 Gemeinsame Modellierung von Erwartungswert und Dispersion

Bis jetzt sind wir von einem (unbekannten) konstanten Dispersionsparameter ϕ, wenigstens aber von beobachtungsspezifischen aber bekannten Parametern ϕ_i ausgegangen. Die EQL-Funktion erlaubt nun eine simultane Modellierung sowohl der Erwartungswerte μ_i als auch der *unbekannten* Dispersionsparameter ϕ_i.

Dazu betrachten wir zwei von einander *abhängige* Modelle:

$$\begin{aligned} \mathbb{E}(Y_i) &= \mu_i, & \mathbb{E}(t_i) &= \phi_i, \\ \operatorname{var}(Y_i) &= \phi_i V(\mu_i), & \operatorname{var}(t_i) &= \psi V_D(\phi_i), \\ g(\boldsymbol{\mu}) &= X\boldsymbol{\beta}, & h(\boldsymbol{\phi}) &= Z\boldsymbol{\gamma}, \end{aligned} \tag{3.27}$$

mit $\boldsymbol{\beta} \in \mathbb{R}^p, X \in \mathbb{R}^{n\times p}$ und $\boldsymbol{\gamma} \in \mathbb{R}^q, Z \in \mathbb{R}^{n\times q}$. Dabei bezeichnen $g(\cdot)$ und $h(\cdot)$ die jeweiligen Linkfunktionen, X und Z die Designmatrizen der jeweiligen erklärenden Variablen, $\boldsymbol{\beta}$ und $\boldsymbol{\gamma}$ die Modellparameter, $V(\cdot)$ und $V_D(\cdot)$ die jeweiligen Varianzfunktionen und ψ den (festen) Dispersionsparameter des Dispersionsmodells. Wir nehmen an, dass die Dispersionsparameter ϕ_i von den Erwartungswerten μ_i funktional unabhängig sind. Das heißt, dass die Abhängigkeit der Varianz vom Erwartungswert vollständig durch die Varianzfunktion $V(\mu)$ erklärt ist.

Bemerkung. Verwendet man für die Modellierung des Erwartungswertes und der Dispersion den gleichen Satz an erklärenden Variablen (gilt also $X = Z$), können implizite Abhängigkeiten auftreten (vgl. Nelder und Pregibon, 1987).

Die Response-Variable des Dispersionmodells t_i muss derart gewählt sein, dass $\mathbb{E}(t_i) = \phi_i$ und $\operatorname{var}(t_i) = \psi V_D(\phi_i)$ gilt. Eine mögliche Wahl ist die i-te Komponente der Deviance d_i, wobei dann $\mathbb{E}(d_i) \approx \phi_i$ nur approximativ gilt, wie im vorherigen Abschnitt gezeigt wurde.

Um die Wahl von $t_i = d_i$ zu motivieren, untersuchen wir die Eigenschaften der EQL-Funktion. Diese lautet für nicht konstante Dispersionsparameter:

$$Q^+(\boldsymbol{\mu}, \boldsymbol{\phi}, \boldsymbol{y}) = -\frac{1}{2}\sum_{i=1}^{n}\left\{\log\left(2\pi\phi_i V(y_i)\right) + D(y_i, \mu_i)/\phi_i\right\}. \tag{3.28}$$

Betrachtet man die ϕ_i als bekannt, erhält man über die Ableitung von (3.28) nach $\beta_j, 1 \le j \le p$ die bereits bekannte Score-Gleichung:

$$\sum_{i=1}^{n}\frac{y_i - \mu_i}{\phi_i V(\mu_i)}\frac{\partial\mu_i}{\partial\beta_j} = 0.$$

Hält man umgekehrt μ_i fest, liefert die Ableitung von (3.28) nach $\gamma_k, 1 \le k \le q$ für $d_i = D(y_i, \mu_i)$ die Gleichung:

$$\sum_{i=1}^{n}\frac{d_i - \phi_i}{2\phi_i^2}\frac{\partial\phi_i}{\partial\gamma_k} = 0. \tag{3.29}$$

Gleichung (3.29) entspricht also einem Quasi-Likelihood-Ansatz für die Response-Variable $\boldsymbol{d} = (d_1, \ldots, d_n)^\top$ mit Erwartungswert $\mathbb{E}(d_i) = \phi_i$, quadratischer Varianzfunktion $V_D(\phi_i) = \phi_i^2$ und Dispersionsparameter $\psi = 2$.

Um nun die Parametervektoren $\boldsymbol{\beta}$ und $\boldsymbol{\gamma}$ zu bestimmen, bedient man sich des *Seesaw-Algorithmus'* (Sägezahn-Algorithmus). Dazu berechnet man zuerst das GLM für den Erwartungswert, wobei man mit $1 \leq i \leq n : \phi_i = 1$ startet. Mit den geschätzten Erwartungswerten $\hat{\boldsymbol{\mu}}$ erhält man auch die Deviance $\boldsymbol{d}$, die als Response im Dispersionsmodell dienen soll. Aus dieser bestimmt man dann eine Schätzung $\hat{\boldsymbol{\phi}}$ für die Dispersionsparameter. Selbige verwendet man nun wiederum als a-priori Gewichte im GLM für den Erwartungswert. Dieser Vorgang wird wiederholt, wobei in der Regel vier bis fünf Iterationen ausreichend sind (vgl. Nelder und Lee, 1998). Einen Algorithmus zur Bestimmung der Parametervektoren $\boldsymbol{\beta}$ und $\boldsymbol{\gamma}$ findet man bei Smyth (1989).

Für die aus der EQL-Funktion abgeleitete gemeinsame Modellierung von Erwartungswert und Dispersion verwendet man an zwei Stellen Approximationen. Zum einen nimmt man an, dass $\mathbb{E}(d_i) = \phi_i$ gilt. Der Bias für d_i ist i. Allg. klein, außer für extreme Fälle, wie für Poisson-verteilte Fehler mit kleinen μ-Werten. Da ein allfälliger Bias des Schätzers d_i allerdings nur als a-priori Gewicht mit $\boldsymbol{\mu}$ interferiert, ist sein Einfluss auf die Schätzung von $\boldsymbol{\beta}$ gering (vgl. Nelder und Lee, 1998).

Zum anderen nimmt man unabhängig von der zugrunde liegenden Verteilung für das Erwartungswertmodell eine Gamma-Varianz für die Dispersion an. Sind die Y_i normalverteilt, folgt, dass die $d_i := D(y_i, \mu_i)$ eine $\phi_i \chi^2_1$-Verteilung aufweisen. Die χ^2_n-Verteilung ist äquivalent zu einer $\Gamma(n/2, 2)$-Verteilung und damit lässt sich für $d_i/\phi_i \sim \chi^2_1 \Leftrightarrow d_i/\phi_i \sim \Gamma(1/2, 2)$[7] die Varianz durch $\operatorname{var}(d_i/\phi_i) = 2 \Rightarrow \operatorname{var}(d_i) = 2\phi_i^2$ ausdrücken und damit beschreibt die Gamma-Varianz die Varianz der Deviance-Komponenten bei normalverteilten Response-Variablen exakt.

Für nicht normalverteilte Response-Variablen trifft dies allerdings nicht mehr zu. Pierce und Schafer (1986) zeigen aber, dass die Berechnung der Deviance-Residuen $r_i^D := \operatorname{sign}(y_i - \hat{\mu}_i)\sqrt{D(y_i, \hat{\mu}_i)}$ der „best normalizing transformation" (der besten Transformation auf Normalverteilung) der GLM-Verteilung sehr ähnlich ist, womit die i-te (skalierte) Deviance-Komponente (die dem i-ten quadrierten Deviance-Residuum entspricht) als Quadrat einer approximativ standardnormalverteilten Größe χ^2_1-verteilt ist, was – wie bereits gezeigt – einer $\Gamma(1/2, 2)$-Verteilung entspricht. Somit lässt sich auch für nicht normalverteilte Response-Variablen die Verwendung der Gamma-Varianz rechtfertigen.

Bemerkung 1. Eine Variante der EQL-Funktion ist die Pseudo-Likelihood (PL)-Funktion (Davidian und Carroll, 1988). Dazu betrachtet man die log-Likelihood-Funktion einer Stichprobe $\boldsymbol{y}$ unabhängig normalverteilter Zufallszahlen mit unterschiedlichen Erwartungswerten μ_i und (unterschiedlichen) Varianzen σ_i^2, wobei man die Varianz als Funk-

[7] Dabei ist die Parametrisierung derart gewählt, dass $\mathbb{E}(d_i/\phi_i) = 2 \cdot 1/2 = 1$ und $\operatorname{var}(d_i/\phi_i) = 4 \cdot 1/2 = 2$ gilt.

tion des Erwartungswertes über $\operatorname{var}(Y_i) = \phi_i V(\mu_i) = \sigma_i^2 \cdot 1$ ausdrückt:

$$\begin{aligned}\mathscr{L}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{y}) &= -\frac{1}{2}\sum_{i=1}^{n} \log\left(2\pi\sigma_i^2\right) - \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{2\sigma_i^2} \\ &= -\frac{1}{2}\sum_{i=1}^{n} \log\left(2\pi\phi_i V(\mu_i)\right) - \frac{1}{2\phi_i}\sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{V(\mu_i)} \\ &= -\frac{1}{2}\sum_{i=1}^{n} \left\{\log\left(2\pi\phi_i V(\mu_i)\right) + r_i^2/\phi_i\right\}, \end{aligned} \tag{3.30}$$

dabei bezeichnen die r_i die Pearson-Residuen. Die Gleichung (3.30) ist für normalverteilte Variablen eine exakte log-Likelihood-Funktion. Verwendet man sie jedoch auch für nicht normalverteilte Variablen (verwendet man also eine andere Varianzfunktion), trifft dies nicht mehr zu. Wir bezeichnen die Größe (3.30) als *Pseudo-Likelihood-Funktion*.

Die PL-Funktion unterscheidet sich von der EQL-Funktion also in der Verwendung der (quadrierten) Pearson-Residuen anstatt der Deviance und im Argument der Varianzfunktion im ersten Term. Verwendet man als Response-Variable im Dispersionsmodell (3.27) die quadrierten Pearson-Residuen,[8] erhält man die Schätzer für $\boldsymbol{\gamma}$ über Maximierung der PL-Funktion.

Bemerkung 2. Da ML-Schätzer für die Varianz in Regressionsmodellen verzerrt sein können, greift man beispielsweise beim LM auf Restricted-Maximum-Likelihood (REML)-Schätzer zurück.[9] Damit erreicht man zumindest asymptotische Unverzerrtheit. Smyth und Verbyla (1999) erweitern die Idee der REML-Schätzer auch für nicht normalverteilte Response-Variablen.

Bemerkung 3. Eine andere Zugangsweise zur Lösung von Problemen mit nicht konstanten Dispersionsparametern findet man bei Efron (1986a). Durch die Einführung der *Double Exponential Family* („Doppelte Exponentialfamilie“) – die bezüglich der Dispersionsparameter $\boldsymbol{\phi}$ ähnliche Eigenschaften wie für den Vektor der Erwartungswerte $\boldsymbol{\mu}$ besitzt – wird auch eine Modellierung der Dispersionsparameter ermöglicht. Beide Ansätze liefern dabei die gleichen Resultate (vgl. Lee und Nelder, 2000).

Zusammenfassend lässt sich die Idee der gemeinsamen Modellierung von Erwartungswert und Dispersionsparameter also auf die EQL- bzw. die PL-Funktion zurückführen, da die jeweiligen Score-Funktionen nach ϕ die Gestalt einer Score-Funktion eines GLM aufweisen. Die Tabelle 3.2 fasst die Komponenten der beiden, aus der EQL-Funktion abgeleiteten Modelle noch einmal zusammen (vgl. Nelder, 1998).

3.2.2 Parametrisierte Varianzfunktionen

Eine weitere Anwendungsmöglichkeit der EQL-Funktion besteht in der Verwendung von Varianzfunktionen aus einer parametrisierten Familie $\mathcal{F}_{\boldsymbol{\theta}} \ni V_{\boldsymbol{\theta}}(\mu) := V(\mu, \boldsymbol{\theta})$, wobei $\boldsymbol{\theta}$

[8] Für die Pearson-Residuen r_i gilt im Gegensatz zu den Deviance-Komponenten $d_i : \mathbb{E}\left(r_i^2\right) = \phi_i$ exakt.
[9] REML wird auch als *Residual-Maximum-Likelihood* bezeichnet.

Tabelle 3.2: EQL-Modelle für die gemeinsame Modellierung von Erwartungswert und Dispersion

Komponente	Erwartungswertmodell GLM für μ	Dispersionsmodell GLM für ϕ
Response	y	d
Erwartungswert	μ	ϕ
Varianz	$\phi V(\mu)$	$2\phi^2$
Linkfunktion	$\eta = g(\mu)$	$\zeta = h(\phi)$
Linearer Prädiktor	$\eta = X\beta$	$\zeta = Z\gamma$
Deviance Komponente	d	$2\{-\log(d/\phi) + (d-\phi)/\phi\}$
a-priori Gewichte	$1/\phi$	1

einen Vektor unbekannter Parameter der Varianzfunktion beschreibt. Die EQL-Funktion lautet dafür:

$$Q^+_{\boldsymbol{\theta}}(\boldsymbol{\mu}, \phi, \boldsymbol{y}) = -\frac{1}{2}\sum_{i=1}^{n}\left\{\log\left(2\pi\phi V_{\boldsymbol{\theta}}(y_i)\right) + D_{\boldsymbol{\theta}}(y_i, \mu_i)/\phi\right\}, \tag{3.31}$$

wobei die Deviance $D_{\boldsymbol{\theta}}(y_i, \mu_i)$ über die QL-Funktion von der Varianzfunktion $V_{\boldsymbol{\theta}}$ und damit von $\boldsymbol{\theta}$ selbst abhängt. Das Ziel ist nun die Bestimmung jenes Parameter-Schätzers $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})$ der die EQL-Funktion (3.31) maximiert.

Um statistische Aussagen über die nicht linearen Parameter treffen zu können, schlagen Nelder und Pregibon (1987) vor, sogenannte *Profile-Quasi-Likelihood-Intervalle* zu verwenden. Dazu hält man den Parameter $\boldsymbol{\theta}$ auf einem Wert $\boldsymbol{\theta}_0$ fest und berechnet den Schätzer $\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}_0)$ mittels der bereits vorgestellten Methoden. Diesen Vorgang wiederholt man für verschiedene Werte aus einem zu definierendem Intervall I (das auch mehrdimensional sein kann) und bestimmt damit den Maximalwert der EQL-Funktion aus diesem Intervall:

$$Q^+_{\max} = \max_{\boldsymbol{\theta}\in I} Q^+_{\boldsymbol{\theta}}. \tag{3.32a}$$

Definiert man weiters den Wert

$$Q^+_{\mathrm{PL}} = Q^+_{\max} - Q^+_{\boldsymbol{\theta}}, \tag{3.32b}$$

lässt sich über die Gleichungen (3.32) eine Menge Θ finden, für dessen Werte die Differenz (3.32b) kleiner als eine vorgegebene Schranke q ist. Die Menge Θ entspricht für den Fall, dass $\boldsymbol{\theta} \in \mathbb{R}$ gilt (die Varianzfamilie $\mathcal{F}_\theta$ hängt also nur von dem Skalar θ ab), einem Intervall. Dieses Intervall wird als Profile-Quasi-Likelihood-Intervall bezeichnet. Eine mögliche Wahl für q ist beispielsweise ein $1/2 \cdot \chi^2_{\mathrm{df}}$-Quantil, wobei die Anzahl der Freiheitsgrade der Dimension von $\boldsymbol{\theta}$ entspricht. Bei Candy (2004) findet man eine praktische Anwendung der Profile-QL-Methode zur Bestimmung des Parameters θ der Varianzfunktion $V(\mu) = \mu^\theta$ im Zuge der Modellierung von Fischfangquoten.

Für Verteilungen aus der Exponentialfamilie findet man asymptotische Eigenschaften für Profile-QL-Intervalle bei Jørgensen (1983). Trifft man nur Annahmen bezüglich der ersten beiden Momente, regen Nelder und Pregibon (1987) für bestimmte Datensätze Bootstrap-Verfahren an, um über die dadurch gewonnene „Resampling-Verteilung" Intervalle für $\boldsymbol{\theta}$ zu bestimmen.

Im folgenden untersuchen wir unterschiedliche Familien von Varianzfunktionen und leiten uns die jeweiligen QL-Funktionen her, die in die Deviance einfließen.

Potenz-Ansatz

Verwenden wir für die Varianzfunktion einen Potenz-Ansatz $V(\mu) = \mu^k$, erhalten wir je nach Wahl von k unterschiedliche Resultate, wie sie im folgenden zusammengefasst sind.

- $k = 0$: Die Varianzfunktion lautet somit $V(\mu) = 1$ und entspricht der Annahme einer Normalverteilung. Es gilt also für $y, \mu \in \mathbb{R}, \phi = \sigma^2$:

$$Q(\mu, y) = \int^{\mu} \frac{y - t}{\sigma^2}\, dt + \text{Funktion in } y = -\frac{(y-\mu)^2}{2\sigma^2}.$$

- $k = 1$: Der Ansatz von $V(\mu) = \mu$ entspricht bei einem fixen Dispersionsparameter von $\phi = 1$ einer Poisson-Verteilungsannahme. Die QL-Funktion ergibt damit für $\mu > 0, y \geq 0$:

$$Q(\mu, y) = \int^{\mu} \frac{y - t}{t}\, dt = y \log \mu - \mu.$$

- $k = 2$: Die quadratische Varianzfunktion $V(\mu) = \mu^2$ ergibt sich bei einer Gammaverteilung $Y \sim \Gamma(\mu, 1)$. Für $\mu > 0, y \geq 0$ gilt:

$$Q(\mu, y) = \int^{\mu} \frac{y - t}{t^2}\, dt = -\frac{y}{\mu} - \log \mu.$$

- $k \in \mathbb{R} \setminus \{1, 2\}$: Die Varianz-Annahme $V(\mu) = \mu^k$ führt für $k \in \mathbb{R} \setminus \{1, 2\}$ bei $\mu > 0, y \geq 0$ zu der QL-Funktion

$$Q(\mu, y) = \int^{\mu} \frac{y - t}{t^k}\, dt = \frac{1}{(k-2)\mu^{k-2}} - \frac{y}{(k-1)\mu^{k-1}}.$$

Bemerkung 1. Die Potenzfunktion $V(\mu) = \mu^k$ führt also für $k \in \{0, 1, 2, 3\}$ zu einer Verteilung aus der Exponentialfamilie.[10]

Bemerkung 2. Die Familie der *Tweedie-Verteilungen* ist eine Familie von Verteilungen aus der Klasse der *Exponential Dispersion Models* (exponentielle Dispersionsmodelle),[11] die auf Jørgensen (1987) zurückgeht. Dabei charakterisiert eine Tweedie-Verteilung einen Erwartungswert von $\mathbb{E}(Y) = \mu$ und eine Varianz von $\text{var}(Y) = \phi\mu^k$ mit $k \leq 0 \vee k \geq 1$ und

[10] Dabei gehört die Varianz-Annahme $V(\mu) = \mu^3$ zu einer Inversen-Gauß-Verteilung.
[11] Die Familie der „Exponential Dispersion Models" selbst ist eine Erweiterung der Exponentialfamilie.

damit sind insbesondere die Normal-, die Poisson-, die Gamma- und die Inverse-Gauß-Verteilung Mitglieder der Tweedie-Familie. Für den Potenz-Ansatz $V(\mu) = \mu^k$ bei einem EQL-Modell existiert somit für $k \leq 0 \vee k \geq 1$ eine Verteilung aus der Tweedie-Familie.

Das R-Paket *tweedie* (Dunn, 2007) stellt die Funktion *tweedie.profile* zur Verfügung, die den Schätzer für den Parameter k der Tweedie-Verteilung über einen Profile-Likelihood-Ansatz bestimmt. Der Dispersionsparameter ϕ wird in diesem Paket über die Maximierung der log-Likelihood-Funktion bestimmt und ist somit ein ML-Schätzer. Diese Vorgehensweise unterscheidet sich von dem durch die EQL-Funktion induzierten Verfahren, indem sie stärkere Annahmen[12] trifft und somit einen anderen Zugang wählen kann.

Binomialähnlicher Ansatz

Die Varianzfunktion einer standardisierten Binomialverteilung ist durch $V(\mu) = \mu(1-\mu)$ gegeben. Diese Annahme wollen wir erweitern, indem wir zwei zusätzliche Parameter $k \in \mathbb{N}$ und $l \in \mathbb{N}$ einführen und den Ansatz $V(\mu) = \mu^k(1-\mu)^l$ wählen.

Um die QL-Funktion bestimmen zu können, betrachten wir vorerst das Integral

$$\int \frac{1}{\mu^k(1-\mu)^l}\, d\mu. \tag{3.33}$$

Mittels der Partialbruchzerlegung lässt sich das Integral (3.33) zu

$$\int \frac{1}{\mu^k(1-\mu)^l}\, d\mu = \int \left(\frac{A_0}{\mu^k} + \cdots + \frac{A_{k-1}}{\mu} + \frac{B_0}{(1-\mu)^l} + \cdots + \frac{B_{l-1}}{1-\mu} \right) d\mu,$$
$$(A_i)_{i=0}^{k-1} := \binom{l-1+i}{l-1}, \quad (B_j)_{j=0}^{l-1} := \binom{k-1+j}{k-1}$$

umschreiben. Mit der Symmetrie des Binomialkoeffizienten folgt sofort

$$A_{k-1} = \binom{k+l-2}{l-1} = \binom{k+l-2}{k+l-2-(l-1)} = \binom{k+l-2}{k-1} = B_{l-1}.$$

Damit lässt sich das Integral (3.33) für $k \geq 1, l \geq 1$ nun lösen:

$$I_{k,l}(\mu) := \int \frac{1}{\mu^k(1-\mu)^l}\, d\mu = \binom{k+l-2}{k-1} \log\left|\frac{\mu}{1-\mu}\right| - \sum_{i=0}^{k-2} A_i \frac{1}{(k-i-1)\mu^{k-i-1}} + \sum_{j=0}^{l-2} B_j \frac{1}{(l-j-1)(1-\mu)^{l-j-1}},$$

wobei wir für $k = 0$ oder $l = 0$ die folgenden Identitäten verwenden:

$$I_{0,l}(\mu) = \begin{cases} \mu & l = 0, \\ -\log|1-\mu| & l = 1, \\ 1/\left[(l-1)(1-\mu)^{l-1}\right] & l \geq 2 \end{cases} \quad I_{k,0}(\mu) = \begin{cases} \mu & k = 0, \\ \log|\mu| & k = 1, . \\ -1/\left[(k-1)\mu^{k-1}\right] & k \geq 1 \end{cases}$$

[12] Es wird die Kenntnis der Verteilung zugrunde gelegt, während sich der EQL-Ansatz mit der Existenz der ersten beiden Momente begnügt.

Die QL-Funktion für den Ansatz $V(\mu) = \mu^k(1-\mu)^l$ für $k \geq 1$ und $l \in \mathbb{N}$ lautet daher für $0 < \mu < 1, 0 \leq y \leq 1$:

$$Q(\mu, y) = \int^{\mu} \frac{y-t}{t^k(1-t)^l}\, dt = y \cdot I_{k,l}(\mu) - I_{k-1,l}(\mu). \tag{3.34}$$

Für die Varianzfunktion der standardisierten Binomialverteilung $V(\mu) = \mu(1-\mu)$ ergibt sich aus Gleichung (3.34) also beispielsweise mit $k = 1$ und $l = 1$:

$$\begin{aligned} Q(\mu, y) &= y \cdot I_{1,1}(\mu) - I_{0,1}(\mu) \\ &= y \log \left| \frac{\mu}{1-\mu} \right| + \log |1-\mu| . \end{aligned}$$

Für logistische Regressionsmodelle, deren Variabilität durch die Varianzfunktion der Binomialverteilung allerdings überschätzt wird, verwendet man auch gerne den Ansatz $V(\mu) = \mu^2(1-\mu)^2$ (vgl. McCullagh und Nelder, 1989, S. 330). Die zugehörige QL-Funktion ergibt sich wieder aus Gleichung (3.34) und lautet:

$$\begin{aligned} Q(\mu, y) &= y \cdot I_{2,2}(\mu) - I_{1,2}(\mu) \\ &= 2y \log \left| \frac{\mu}{1-\mu} \right| - \frac{y}{\mu} + \frac{y}{1-\mu} - \log \left| \frac{\mu}{1-\mu} \right| - \frac{1}{1-\mu} \\ &= (2y-1) \log \left| \frac{\mu}{1-\mu} \right| - \frac{y}{\mu} - \frac{1-y}{1-\mu}. \end{aligned}$$

Relaxiert man die Bedingung $k \in \mathbb{N}$ und $l \in \mathbb{N}$ indem man auch reelle Exponenten erlaubt, erhält man für das bestimmte Integral für $k \in \mathbb{R}$ und $l \in \mathbb{R}$ vermöge der Substitution $t = \mu p, p \in (0,1), dt/dp = \mu$:

$$\begin{aligned} \int_0^{\mu} \frac{1}{t^k(1-t)^l}\, dt &= \int_0^1 \frac{1}{(\mu p)^k(1-\mu p)^l} \mu\, dp \\ &= \frac{1}{\mu^{k-1}} \int_0^1 p^{-k}(1-\mu p)^{-l}\, dp \\ &= \frac{1}{\mu^{k-1}} \underbrace{\int_0^1 p^{-k} \overbrace{(1-p)^{2-k-(1-k)-1}}^{=(1-p)^0=1} (1-\mu p)^{-l}\, dp}_{=\,{}_2\mathrm{F}_1(1-k,l;2-k;\mu)\frac{\Gamma(1-k)\Gamma(1)}{\Gamma(2-k)} = \frac{1}{1-k}\,{}_2\mathrm{F}_1(1-k,l,2-k,\mu)} \\ &= \frac{\mu^{1-k}}{1-k}\, {}_2\mathrm{F}_1(1-k, l; 2-k; \mu), \end{aligned} \tag{3.35}$$

wobei ${}_2\mathrm{F}_1(1-k, l; 2-k; \mu)$ die *hypergeometrische Funktion* beschreibt, die allgemein durch

$${}_p\mathrm{F}_q(a_1, \ldots, a_p; b_1, \ldots, b_q; z) = \sum_{k=0}^{\infty} \prod_{i=1}^{p} \frac{\Gamma(k+a_i)}{\Gamma(a_i)} \prod_{j=1}^{q} \frac{\Gamma(k+b_j)}{\Gamma(b_j)} \frac{z^k}{k!}$$

gegeben ist. Für $p = 2, q = 1$ ergeben sich außerdem mit $\mathrm{B}(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$ die folgenden äquivalenten Darstellungen:

$$ {}_2\mathrm{F}_1(a_1, a_2; b; x) = \mathrm{B}(a_1, b-a_1)^{-1} \int_0^1 u^{a_1-1}(1-u)^{b-a_1-1}(1-xu)^{a_2}\, du, \quad b > a_1 > 0, $$

(Integraldarstellung)

$$ {}_2\mathrm{F}_1(a_1, a_2; b; x) = \sum_{j=0}^{\infty} \frac{(a_1)_j (a_2)_j}{(b)_j} \frac{x^j}{j!}, \quad b \notin \mathbb{Z} \setminus (\mathbb{N} \cup \{0\}), \qquad \text{(Reihendarstellung)} $$

wobei $(x)_j$ die *Pochhammer-Notation* (oder *steigende Faktorielle*) bezeichnet und durch

$$ (x)_n := x(x+1)(x+2)\cdots(x+n-1) = \frac{\Gamma(x+n)}{\Gamma(x)}, $$
$$ (x)_0 := 1 $$

definiert ist.

Bemerkung. Die Lösung des Integrals (3.35) beruht auf der Integraldarstellung der hypergeometrischen Funktion, die nur für $b > a_1 > 0$ (bzw. $b > a_2 > 0$)[13] definiert ist. Damit folgt sofort, dass $2-k > 1-k > 0 \Leftrightarrow 1 > k$ bzw. $2-k > l \Leftrightarrow 2 > k+l$ gelten muss.

3.2.3 Informationskriterien

Eine zentrale Frage bei der Modellierung von Daten ist, welche Teilmenge an erklärenden Variablen ausgewählt werden soll. Wie im Abschnitt 2.3 bereits besprochen, besteht eine Möglichkeit darin, mittels einer Analysis-of-Deviance aus ineinander verschachtelten Modellen sukzessive Terme zu entfernen, bis man bei einem Modell angelangt ist, aus dem sich keine Terme mehr entfernen lassen. Das dabei auftretende Problem ist die Balance zwischen Modellen, die mit vielen Termen viel erklären aber kompliziert sind und einfachen Modellen, die mitunter einen großen Lack-of-Fit aufweisen, zu finden.

Ein Maß, das dieser Bemühung Rechnung trägt, indem es einen von der Anzahl an Termen abhängigen Strafterm enthält, ist das *Akaike-Informationskriterium (AIC)* (Akaike, 1973) und lautet allgemein:

$$ \mathrm{AIC} = 2p - 2\hat{\mathscr{L}}, \tag{3.36} $$

wobei p die Anzahl der Terme im Modell und $\hat{\mathscr{L}}$ die maximierte log-Likelihood-Funktion unter dem Modell bezeichnet. Um nun zwischen zwei Modellen eines auszuwählen, vergleicht man die zugehörigen AIC-Werte und gibt demjenigen Modell den Vorzug, das den kleineren AIC-Wert aufweist.

[13] Aus der Reihendarstellung ist sofort ersichtlich, dass die Reihenfolge der Argumente der ersten Komponente $(a_1, \ldots, a_p)$ bzw. der zweiten Komponente $(b_1, \ldots, b_q)$ nicht relevant ist und so gilt beispielsweise ${}_2\mathrm{F}_1(a_1, a_2; b; x) = {}_2\mathrm{F}_1(a_2, a_1; b; x)$.

Bemerkung. Ein Nachteil des AIC ist, dass es den Stichprobenumfang unberücksichtigt lässt. Schwarz (1978) schlägt deshalb das *Bayessche Informationskriterium (BIC)* vor, das durch

$$\mathrm{BIC} = p \log n - 2\hat{\mathscr{L}}$$

gegeben ist. Durch die Berücksichtigung des Stichprobenumfangs tendiert das BIC öfter zu Modellen mit weniger Parametern als das AIC.

Beispiel 3.10. Mit der Fehlerquadratsumme $\mathrm{SSE}(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2$ und dem ML-Schätzer für die Varianz $\hat{\sigma}^2 = \mathrm{SSE}(\hat{\boldsymbol{\beta}})/n$ folgt für ein LM aus Gleichung (3.36):

$$\begin{aligned}
\mathrm{AIC} &= 2p - 2\hat{\mathscr{L}} \\
&= 2p - 2\log f_Y(\boldsymbol{y}, \hat{\boldsymbol{\mu}}, \hat{\sigma}^2) \\
&= 2p - 2\log\left\{\prod_{i=1}^{n} f_{Y_i}(y_i, \hat{\mu}_i, \hat{\sigma}^2)\right\} \\
&= 2p - 2\sum_{i=1}^{n}\log\left\{\frac{1}{\sqrt{2\pi\hat{\sigma}^2}}\exp\left(-\frac{(y_i - \hat{\mu}_i)^2}{2\hat{\sigma}^2}\right)\right\} \\
&= 2p - 2\left\{-\frac{n}{2}\log(2\pi\hat{\sigma}^2) - \frac{1}{2\hat{\sigma}^2}\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2\right\} \\
&= 2p + n\left\{\log(2\pi\,\mathrm{SSE}(\hat{\boldsymbol{\beta}})/n) + 1\right\}.
\end{aligned}$$

◊

Das AIC ist als approximativ unverzerrter Schätzer für die erwartete *Kullback-Leibler-Information* (Kullback und Leibler, 1951) entworfen worden, die ein Maß für die Unterschiedlichkeit zweier auf der gleichen Ereignismenge definierten Verteilungen F und G beschreibt. Im stetigen Fall lässt sie sich durch

$$\mathrm{KL}(F, G) = \int_{\mathbb{R}} f(x)\log\frac{f(x)}{g(x)}\,dx$$

beschreiben. Dabei bezeichnen $f(x)$ und $g(x)$ die jeweiligen Dichtefunktionen. Üblicherweise beschreibt $F(\,\cdot\,)$ die präzise Verteilung, während $G(\,\cdot\,)$ eine Approximation oder ein Modell derselben darstellt.

Hurvich und Tsai (1989) zeigen, dass das AIC beim LM für kleine Stichprobenumfänge allerdings stark verzerrt sein kann und entwickeln für diesen Fall einen weniger verzerrten Schätzer AIC_c:

$$\mathrm{AIC}_c = n\log\hat{\sigma}^2 + \frac{n(n+p)}{n-p-2}. \tag{3.37}$$

Ein anderes Informationskriterium für das LM ist *Mallows' C_p-Statistik* (Mallows, 2000), die mit der üblichen Notation durch

$$C_p = \frac{1}{\hat{\sigma}^2}\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2 + 2p - n \tag{3.38}$$

definiert ist.

Um ein Kriterium für das GLM zu erhalten, modifizieren Hosmer u. a. (1989) die C_p-Statistik in Gleichung (3.38) unter Zuhilfenahme der Pearson-Statistik X^2 und erhalten so die Größe

$$\tilde{C}_p = (n-q)X_p^2/X_q^2 + 2p - n. \tag{3.39}$$

Dabei enthält das komplexere Modell q und das zu testende Untermodell p Variablen und X_p^2 und X_q^2 bezeichnen die jeweiligen Pearson-Statistiken. Die Verwendung der Deviance anstelle der Pearson-Statistik führt zu

$$C_p^* = (n-q)D_p/D_q + 2p - n \tag{3.40}$$

und geht auf Pregibon (1979) zurück.

Die für das GLM bis jetzt vorgestellten Informationskriterien beinhalten keine Möglichkeit den Dispersionsparameter zu berücksichtigen. McCullagh und Nelder (1989) geben ein auf der C_p-Statistik beruhendes Kriterium an, das den Dispersionsparameter berücksichtigt, das aber bei kleinen Stichproben zu *overfitted models* (überangepassten Modellen) führen kann. Hurvich und Tsai (1995) erweitern deshalb speziell für kleine Stichprobenumfänge ihr AIC_c aus Gleichung (3.37) auch für die Verwendung bei EQL-Modellen:

$$\mathrm{AIC}_c^{\mathrm{EQL}} = n\left\{\log\hat{\phi} + 1\right\} + \frac{2n(p+1)}{n-p-2}. \tag{3.41}$$

In einer vergleichenden Monte-Carlo Studie der Kriterien (3.36), (3.39), (3.40) und (3.41) für logistische Modelle stellen sie weiters fest, dass ihr $\mathrm{AIC}_c^{\mathrm{EQL}}$ häufiger das richtige Modell auszuwählen im Stande ist als die anderen Kriterien. Außerdem zeigen sie, dass das $\mathrm{AIC}_c^{\mathrm{EQL}}$ als Schätzer für die erwartete Kullback-Leibler-Information weniger verzerrt ist als das AIC.

3.2.4 Fehlerrate

Unabhängig von dem zur Modellfindung herangezogenen Fehlermaß, stellt sich die Frage, wie genau das Modell in der Lage ist, neue Daten vorherzusagen. Diese Qualität der Vorhersage lässt sich ebenfalls messen. Passt man etwa für binäre Daten $y_i \in \{0,1\}$ mit den zugehörigen Wahrscheinlichkeiten $\pi_i \in (0,1)$ ein Modell an, erhält man mit der logistischen Regression geschätzte Erfolgswahrscheinlichkeiten $\hat{\pi}_i \in (0,1)$. Um nun eine Vorhersage $\hat{\xi}_i \in \{0,1\}$ auf der Skala der Response-Variablen selbst zu erhalten, wählt man einen Grenzpunkt C_0 (üblicherweise $C_0 = 0.5$) und definiert

$$\hat{\xi}_i = \begin{cases} 1 & \hat{\pi}_i > C_0, \\ 0 & \hat{\pi}_i \leq C_0 \end{cases}.$$

Ein Vergleich mit den beobachteten Daten ergibt so eine geschätzte mittlere Fehlerrate von:

$$\hat{\rho}(\boldsymbol{y}, \hat{\boldsymbol{\xi}}(\hat{\boldsymbol{\pi}})) = \frac{1}{n}\Big|\{\hat{\xi}_i \neq y_i\}\Big|. \tag{3.42}$$

Da für die Erstellung des Modells und die Abschätzung des Fehlers die gleichen Daten verwendet werden, unterschätzt der Schätzer $\hat{\rho}$ die tatsächliche Fehlerrate ρ und ist somit verzerrt.

Efron (1986b) untersucht den Bias für die Fehlerrate, wobei eine verallgemeinerte Variante des Fehlermaßes verwendet wird:

$$R(y_i, \hat{\mu}_i) = r(\hat{\mu}_i) + r'(\hat{\mu}_i)(y_i - \hat{\mu}_i). \tag{3.43}$$

Dabei bezeichnet $r(\,\cdot\,)$ eine konkave Funktion mit der Ableitung $r'(\,\cdot\,)$. Die geschätzte Fehlerrate ergibt sich aus der gemittelten Summe und lautet:

$$\hat{\rho}(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = \frac{1}{n} \sum_{i}^{n} R(y_i, \hat{\mu}_i) = \frac{1}{n} R(\boldsymbol{y}, \hat{\boldsymbol{\mu}}). \tag{3.44}$$

Bemerkung 1. Die Ableitung von $r(\,\cdot\,)$ wird – für den Fall, dass sich linksseitige und rechtsseitige Ableitung unterscheiden – durch Festsetzung auf die linksseitige Ableitung

$$r'(x) := \lim_{\xi \to x^-} \frac{r(x) - r(\xi)}{x - \xi}$$

eindeutig definiert.

Bemerkung 2. Da der Schätzer $\hat{\boldsymbol{\mu}}$ von den Beobachtungen $\boldsymbol{y}$ abhängt, könnte man die geschätzte Fehlerrate auch mit $\hat{\rho}(\boldsymbol{y})$ bezeichnen. Da wir aber für den Moment noch keine Annahmen über die Bildungsvorschrift $\boldsymbol{y} \mapsto \hat{\boldsymbol{\mu}}$ treffen wollen, bleiben wir bei der Notation $\hat{\rho}(\boldsymbol{y}, \hat{\boldsymbol{\mu}})$.

Bemerkung 3. Für das Fehlermaß aus Gleichung (3.42) lässt sich die Funktion $r(\,\cdot\,)$ für $C_0 = 0.5$ durch $r(x) := \min(x, 1-x), x \in (0,1)$, definieren. Damit gilt für die Ableitung von $r(\,\cdot\,)$:

$$r'(x) = \begin{cases} 1 & x \leq 0.5, \\ -1 & x > 0.5 \end{cases},$$

womit für den i-ten Beitrag zum Fehlermaß

$$\begin{aligned} R(y_i, \hat{\pi}_i) &= \begin{cases} \hat{\pi}_i + (y_i - \hat{\pi}_i) & \hat{\pi}_i \leq 0.5, \\ 1 - \hat{\pi}_i - (y_i - \hat{\pi}_i) & \hat{\pi}_i > 0.5 \end{cases} \\ &= \begin{cases} y_i & \hat{\pi}_i \leq 0.5, \\ 1 - y_i & \hat{\pi}_i > 0.5 \end{cases} \end{aligned}$$

gilt. Da y_i nur die Werte Null oder Eins annehmen kann folgt das gewünschte Ergebnis:

$$R(y_i, \hat{\pi}_i) = \begin{cases} 0 & (y_i = 0 \wedge \hat{\pi}_i \leq 0.5) \vee (y_i = 1 \wedge \hat{\pi}_i > 0.5), \\ 1 & (y_i = 0 \wedge \hat{\pi}_i > 0.5) \vee (y_i = 1 \wedge \hat{\pi}_i \leq 0.5) \end{cases}.$$

Bemerkung 4. Andere Fehlermaße für die logistische Regression sind beispielsweise der *quadrierte Fehler* und die *Deviance* und sind in Tabelle 3.3 zusammengefasst (vgl. Efron, 1986b).

Tabelle 3.3: Fehlermaße für die logistische Regression

Name	$r(\hat{\mu}_i)$	$R(y_i, \hat{\mu}_i)$
Quadrierter Fehler	$\hat{\mu}_i(1-\hat{\mu}_i)$	$(y_i - \hat{\mu}_i)^2$
Deviance	$-2\left[\hat{\mu}_i \log \hat{\mu}_i + (1-\hat{\mu}_i)\log(1-\hat{\mu}_i)\right]$	$-2\log\left(\hat{\mu}_i^{y_i}(1-\hat{\mu}_i)^{1-y_i}\right)$

Mit der allgemeinen Definition des Fehlermaßes in Gleichung (3.43) lässt sich nun auch die erwartete tatsächliche Fehlerrate definieren:

$$\rho(\boldsymbol{y}, \boldsymbol{\mu}) = \mathbb{E}_{\tilde{\boldsymbol{y}}}\left[R(\tilde{\boldsymbol{y}}, \hat{\boldsymbol{\mu}})\right], \tag{3.45}$$

dabei bezeichnet $\tilde{\boldsymbol{y}}$ einen (hypothetischen) Datenvektor der – unabhängig von dem beobachteten Datenvektor $\boldsymbol{y}$ – aus der gleichen Verteilung kommt wie dieser.

Bemerkung. Die erwartete Fehlerrate $\rho(\boldsymbol{y}, \boldsymbol{\mu})$ misst also den Fehler für einen fixen Wert $\hat{\boldsymbol{\mu}}$ wenn der tatsächliche (üblicherweise unbekannte) Parametervektor mit $\boldsymbol{\mu}$ bezeichnet wird.

Der Bias für die Fehlerrate ergibt sich aus der Differenz der Größen (3.45) und (3.44) und lautet somit:

$$B(\boldsymbol{y}, \boldsymbol{\mu}) = \rho(\boldsymbol{y}, \boldsymbol{\mu}) - \hat{\rho}(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) > 0. \tag{3.46}$$

Der Bias in Gleichung (3.46) bezieht sich noch immer auf einen fixen Wert von $\hat{\boldsymbol{\mu}}$. Ein allgemeiner Bias für die *Bildungsvorschrift* $\boldsymbol{y} \mapsto \hat{\boldsymbol{\mu}}$ lässt sich über

$$\omega(\boldsymbol{\mu}) = \mathbb{E}_{\boldsymbol{y}}\left[B(\boldsymbol{y}, \boldsymbol{\mu})\right] \tag{3.47}$$

finden. Die Größe (3.47) ist noch von dem *tatsächlichen* Wert des Parametervektors $\boldsymbol{\mu}$ und natürlich von dem gewählten Fehlermaß abhängig. Da der wahre Parameter $\boldsymbol{\mu}$ i. Allg. nicht bekannt ist, greift man auf Schätzungen für den Bias zurück. Die Tabelle 3.4 fasst die Ergebnisse noch einmal zusammen.

Verwendet man die gemittelte unskalierte Deviance als geschätzte Fehlerrate

$$\hat{\rho}(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = -\frac{2}{n}\left[\mathscr{L}(\hat{\boldsymbol{\mu}}, \boldsymbol{y}) - \mathscr{L}(\boldsymbol{y}, \boldsymbol{y})\right],$$

zeigt Efron (1986b), dass für Verteilungen aus der Exponentialfamilie bei Verwendung des ML-Schätzers $\hat{\boldsymbol{\mu}}$ für den Erwartungswertvektor $\boldsymbol{\mu}$ des GLM die Größe (3.47) approximativ durch

$$\omega(\boldsymbol{\mu}) \approx 2p/n \tag{3.48}$$

gegeben ist, wobei p der Dimension des Parametervektors $\boldsymbol{\beta}$ entspricht.

Tabelle 3.4: Übersicht über die definierten Fehlerterme

Name	Zeichen	Bedeutung
Fehlermaß	$R(y_i, \hat{\mu}_i)$	Beitrag der i-ten Beobachtung zum gesamten Fehler
Geschätzte Fehlerrate	$\hat{\rho}(\boldsymbol{y}, \hat{\boldsymbol{\mu}})$	mittlere Fehlerrate basierend auf den Beobachtungen $\boldsymbol{y}$ und der gegebenen Schätzung $\hat{\boldsymbol{\mu}}$, (verzerrter) Schätzer für $\rho(\boldsymbol{y}, \boldsymbol{\mu})$
Erwartete tatsächliche Fehlerrate	$\rho(\boldsymbol{y}, \boldsymbol{\mu})$	tatsächliche Fehlerrate für einen fixen Wert von $\hat{\boldsymbol{\mu}}$, setzt Kenntnis des wahren Parameters $\boldsymbol{\mu}$ voraus
Bias	$B(\boldsymbol{y}, \boldsymbol{\mu})$	Bias zwischen der geschätzten und der tatsächlichen Fehlerrate für einen fixen Wert von $\hat{\boldsymbol{\mu}}$
Erwarteter Bias	$\omega(\boldsymbol{\mu})$	erwarteter Bias zwischen der geschätzten und der tatsächlichen Fehlerrate bei gegebener Bildungsvorschrift $\boldsymbol{y} \mapsto \hat{\boldsymbol{\mu}}$

Mit der Näherung (3.48) können wir nun eine approximative Formel für einen Schätzer der erwarteten Fehlerrate angeben:

$$\begin{aligned}
\tilde{\rho}(\boldsymbol{y}, \boldsymbol{\mu}) &= \hat{\rho}(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) + \omega(\boldsymbol{\mu}) \\
&\approx -\frac{2}{n}\left[\mathscr{L}(\hat{\boldsymbol{\mu}}, \boldsymbol{y}) - \mathscr{L}(\boldsymbol{y}, \boldsymbol{y})\right] + \frac{2p}{n} \\
&= \frac{1}{n}\underbrace{\left[-2\mathscr{L}(\hat{\boldsymbol{\mu}}, \boldsymbol{y}) + 2p\right]}_{=\mathrm{AIC}} + \frac{2}{n}\mathscr{L}(\boldsymbol{y}, \boldsymbol{y}) \\
&= \frac{1}{n}\,\mathrm{AIC} + \frac{2}{n}\mathscr{L}(\boldsymbol{y}, \boldsymbol{y}). \qquad (3.49)
\end{aligned}$$

Da der zweite Term in Gleichung (3.49) konstant ist, ist eine Modellsuche auf Basis der Minimierung des AIC äquivalent zur Minimierung der Fehlerrate.

Bemerkung. Die Approximation (3.49) ist für normalverteilte Response-Variablen exakt. Außerdem folgt in diesem Fall für den Schätzer der erwarteten tatsächliche Fehlerrate bei Verwendung der gemittelten skalierten Deviance

$$\hat{\rho}(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = \frac{1}{n\sigma^2}\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2$$

als geschätzte Fehlerrate mit dem Bias in Gleichung (3.48) (der in diesem Fall sogar exakt

ist) die Beziehung:

$$\begin{aligned}\tilde{\rho}(\boldsymbol{y}, \boldsymbol{\mu}) &= \frac{1}{n}\left(\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2 + 2p\right) \\ &= \frac{1}{n}\left(\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2 + 2p - n + n\right) \\ &= \frac{1}{n}\Bigg(\underbrace{\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2 + 2p - n}_{=C_p}\Bigg) + 1 \\ &= \frac{1}{n}C_p + 1,\end{aligned}$$

wobei die Größe C_p wieder Mallows' C_p-Statistik, wie sie in Gleichung (3.38) definiert ist, bezeichnet.

3.3 Beispiele

3.3.1 Nicht konstante Dispersionsparameter

Für die gemeinsame Modellierung von Erwartungswert und Dispersion erzeugen wir zuerst je $n = 100$ Zufallszahlen gemäß:

$$\left.\begin{array}{c}\mu_i \in \{1, 10\} \\ \sigma_j^2 \in \{0.25, 4, 25\}\end{array}\right\} Y_{ij} \sim \mathcal{N}\left(\mu_i, \sigma_j^2\right).$$

Dabei beschreibt σ_j^2 die jeweilige (nicht konstante) Varianz. Wir erhalten so $100 \times 2 \times 3 = 600$ normalverteilte, heteroskedastische Zufallszahlen, die wir folgendermaßen zu einem Vektor $\boldsymbol{y}$ zusammenfassen:

$$\begin{gathered}y_k^{ij} \text{ ist eine Realisation von } Y_{ij} \text{ für } 1 \leq k \leq 100, \\ \boldsymbol{y}_{ij} = (y_1^{ij}, \ldots, y_{100}^{ij}), \\ \boldsymbol{y} = (\boldsymbol{y}_{11}, \boldsymbol{y}_{12}, \boldsymbol{y}_{13}, \boldsymbol{y}_{21}, \boldsymbol{y}_{22}, \boldsymbol{y}_{23})^\top.\end{gathered}$$

Wir suchen nach einem Modell ohne Intercept-Term und verwenden der Einfachheit halber die tatsächlichen Erwartungswerte als erklärende Variablen. Somit erhalten wir für die Designmatrix $X \in \mathbb{R}^{600 \times 1}$:

$$X = (\underbrace{1, \ldots, 1}_{300 \text{ mal}}, \underbrace{10, \ldots, 10}_{300 \text{ mal}})^\top.$$

Für den Vektor $\boldsymbol{y} = (y_1, \ldots, y_{600})^\top$ passen wir vorerst ein LM an (und nehmen damit implizit eine konstante Varianz an):

$$\begin{gathered}\mathbb{E}(Y_i) = x_i\beta, \beta \in \mathbb{R}, \\ \operatorname{var}(Y_i) = \sigma^2,\end{gathered}$$

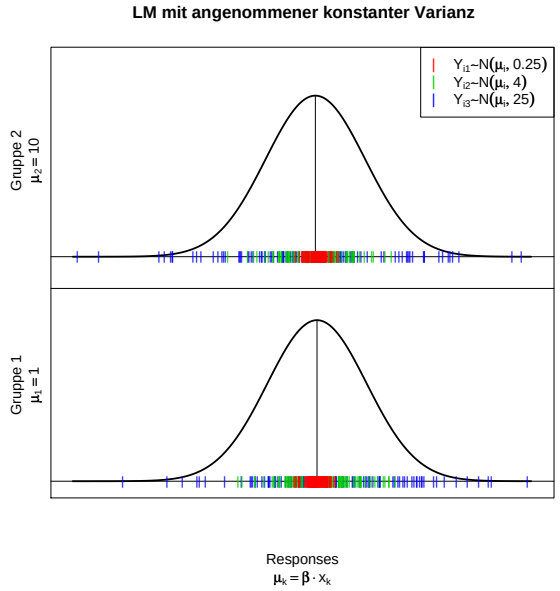


Abbildung 3.2: LM für heteroskedastische Daten

wobei x_i die i-te Zeile[14] der Designmatrix X beschreibt. Das tatsächliche β ist gleich Eins und für den Schätzer $\hat{\beta}$ erhalten wir $\hat{\beta} = 0.9884$. Das LM nimmt eine konstante Varianz an, wofür das Modell einen Schätzer von $\hat{\sigma}^2 = 9.883142$ liefert.

Die Abbildung 3.2 verdeutlicht die Schwäche dieses Ansatzes. Für den gesamten Zufallsvektor $\boldsymbol{y}$ wurde eine konstante Varianz angenommen, man erkennt aber deutlich, dass die einzelnen Komponenten des Vektors $\boldsymbol{y}$ unterschiedlich stark streuen. Die Zufallszahlen aus der roten Gruppe streuen am wenigsten, während die Streuungsbreite der Zufallszahlen aus der blauen Gruppe am größten ist. Da das LM aber von einer konstanten Varianz ausgeht, wird diesem Umstand nicht Rechnung getragen und es wird lediglich eine gemeinsame Varianz geschätzt. Will man nun etwa Vorhersagen für die rote Gruppe treffen, wird das Ergebnis zu breit streuen.

Über die EQL-Methode ist man nun in der Lage, Dispersion und Erwartungswert simultan zu modellieren. Dazu bedarf es neben dem Erwartungswertmodell eines Modells für die Dispersion und insbesondere eines Satzes von erklärenden Variablen. Wir greifen in unserem Beispiel auf einen dreistufigen Faktor zurück und modellieren die Dispersion wieder ohne Intercept.

[14] Im vorliegenden Fall entspricht die i-te Zeile einem Skalar.

Damit lautet die Designmatrix für das Dispersionsmodell $Z \in \mathbb{R}^{600\times 3}$:

$$Z_1 = \begin{pmatrix} 1 & 0 & 0 \\ \vdots & & \vdots \\ 1 & 0 & 0 \end{pmatrix}, Z_2 = \begin{pmatrix} 0 & 1 & 0 \\ \vdots & & \vdots \\ 0 & 1 & 0 \end{pmatrix}, Z_3 = \begin{pmatrix} 0 & 0 & 1 \\ \vdots & & \vdots \\ 0 & 0 & 1 \end{pmatrix},$$

$$Z = \begin{pmatrix} Z_1 \\ Z_2 \\ Z_3 \\ Z_1 \\ Z_2 \\ Z_3 \end{pmatrix}.$$

Für das Dispersionsmodell treffen wir die Annahme einer Gammavarianz und verwenden die Deviance-Komponenten d_i als erklärende Variablen. Wir erhalten so ein Gammamodell und entscheiden uns für die log-Linkfunktion. Das heißt wir suchen einen Schätzer $\hat{\boldsymbol{\gamma}} \in \mathbb{R}^3$ für das Modell:[15]

$$\log\{\mathbb{E}(d_i)\} = \log\phi_i = \boldsymbol{z}_i\boldsymbol{\gamma}, \boldsymbol{\gamma} \in \mathbb{R}^3,$$
$$\operatorname{var}(d_i) = 2\phi_i^2.$$

Mithilfe des Sägezahn-Algorithmus' ermitteln wir nun die Schätzer $\hat{\boldsymbol{\beta}}$ für das Erwartungswert- und $\hat{\boldsymbol{\gamma}}$ für das Dispersionsmodell. Der Dispersionsparameter bei normalverteilten Variablen entspricht der Varianz, somit lauten in unserem Fall die tatsächlichen Werte für die jeweiligen Gruppen $0.25, 4$ bzw. 25. Das Dispersionsmodell liefert die Werte $0.287, 4.099$ und 25.303.

Die Abbildung 3.3 verdeutlicht den Vorteil der gemeinsamen Modellierung von Dispersion und Erwartungswert. Die unterschiedlichen Varianzen werden in diesem Modell berücksichtigt. Die geschätzten Dispersionsparameter sind für beide Erwartungswertgruppen gleich, da funktionale Unabhängigkeit zwischen der Dispersion und dem Erwartungswert angenommen wurde.

3.3.2 Parametrisierte Varianzfunktion

Nelder und Pregibon (1987) untersuchen einen Datensatz über das Verhalten von Wolle, der auf Box und Cox (1964) zurückgeht. Dabei wird die Wolle wiederholt belastet, wobei die Anzahl der Wiederholungen gezählt wird bis die Wolle reißt und wird als Response-Variable y aufgezeichnet. Die erklärenden Variablen werden mit $x_i, 1 \leq i \leq 3$, bezeichnet und beschreiben die Länge des Wollfadens, die Amplitude der Belastung und die Belastung selbst.

Bemerkung. Die Kodierung für die Variablen x_i ist etwas eigentümlich gewählt. Die Variablen x_i sind als Faktoren ausgelegt, die die Werte $\{-1, 0, 1\}$ annehmen können. In der Berechnung wird x_i aber als Variable betrachtet.

[15] Der feste Dispersionsparameter kann in der verwendeten R-Bibliothek *JointModeling* allerdings nicht vorgegeben werden, sondern wird geschätzt und man erhält $\hat{\psi} = 2.029$.

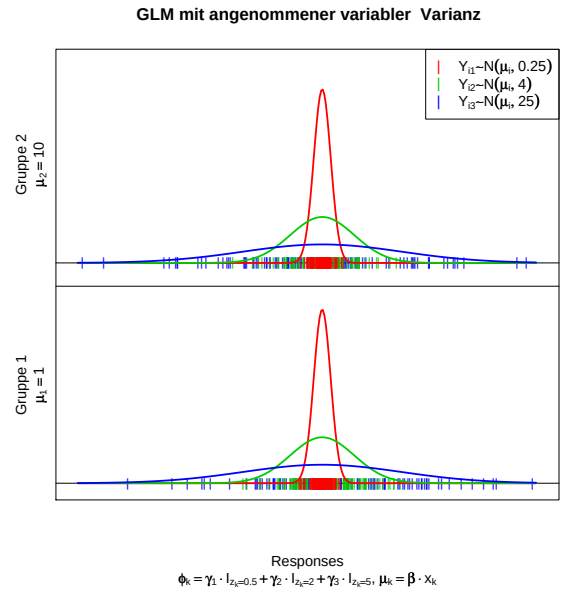


Abbildung 3.3: GLM für heteroskedastische Daten

Nelder und Pregibon verwenden dazu das parametrisierte EQL-Modell

$$\begin{aligned} \log \mu &= x_1\beta_1 + x_2\beta_2 + x_3\beta_3, \\ V_\theta(\mu) &= \mu^\theta \end{aligned} \tag{3.50}$$

und geben ein 95% Konfidenzintervall für den Parameter θ basierend auf dem EQL-Ansatz mit $\theta \in (1.75, 3.35)$ an.

Bemerkung. Obwohl in der Modellformulierung (3.50) nicht ersichtlich, verwenden Nelder und Pregibon auch einen Intercept-Term und das Modell lautet eigentlich

$$\begin{aligned} \log \mu &= \alpha + x_1\beta_1 + x_2\beta_2 + x_3\beta_3, \\ V_\theta(\mu) &= \mu^\theta. \end{aligned}$$

Wie in Abschnitt 3.3.2 erwähnt, gilt für den hier für die Varianzfunktion gewählten Potenz-Ansatz, dass eine Verteilung aus der Tweedie-Familie existiert. Diese Eigenschaft wollen wir uns zunutze machen, um ein alternatives Konfidenzintervall anzugeben. Die Abbildung 3.4 zeigt den Profile-Plot für den Parameter θ. Im Gegensatz zum EQL-Ansatz beruht diese Vorgehensweise auf der Maximierung der (vollen) Likelihood-Funktion.

Dieser alternative Ansatz liefert ein Konfidenzintervall von $\theta \in (1.75, 2.97)$. Das Intervall ist kleiner als das von Nelder und Pregibon bestimmte, was nicht weiter verwundert, da stärkere Annahmen getroffen wurden. Insbesondere wurde die vollständige Verteilung spezifiziert. Das Intervall enthält im übrigen auch den Wert von $\theta = 2$. Der Maximalwert liegt bei $\theta_{\text{opt}} = 2.51$.

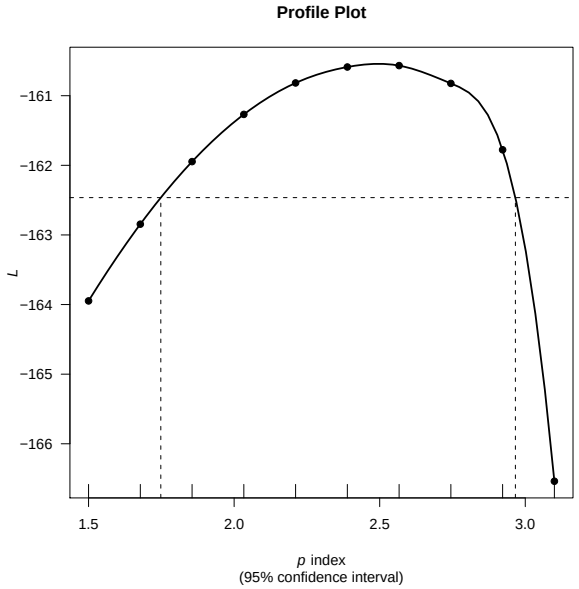
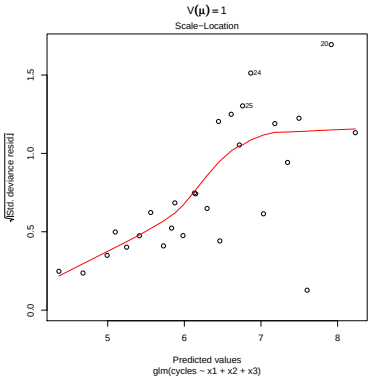
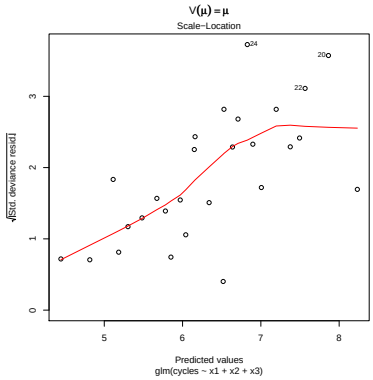


Abbildung 3.4: Profile-Plot für den Parameter einer Tweedie-Verteilung

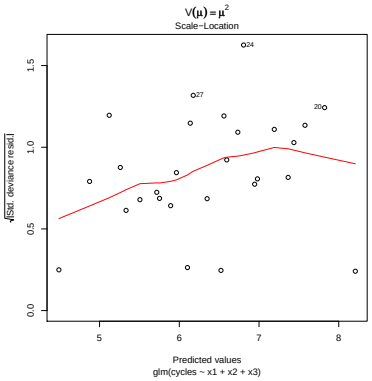
Macht man in einem Plot, der die angepassten Werte gegen die Wurzel der standardisierten Deviance Residuen aufträgt, einen wachsenden (oder fallenden) Trend in der Variabilität der Residuen aus, lässt das auf eine schlecht spezifizierte Varianzfunktionen schließen. Speziell für eine Varianzfunktion $V(\mu) = \mu^\theta$ aus der Potenzfamilie deutet ein fallender Trend auf ein zu großes θ hin, während ein wachsender Trend ein Anzeichen dafür ist, dass θ zu klein gewählt wurde. Für das vorliegende Modell zeigt die Abbildung 3.5 die Residuenplots für verschiedene Werte von θ. Man erkennt deutlich die wachsende Variabilität der Residuen für die Normal- ($\theta = 0$) und die Poissonverteilung ($\theta = 1$), ebenso die abnehmende Streuung der Residuen für die Inverse-Gauß- ($\theta = 3$) und die Tweedieverteilung mit $\theta = 4$. Im Gegensatz dazu ist die Variabilität der Residuen für die Gammaverteilung ($\theta = 2$) und die optimale Tweedieverteilung mit $\theta = 2.51$ unauffälliger.



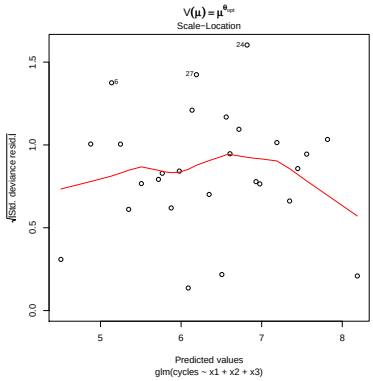
(a) Normalverteilung



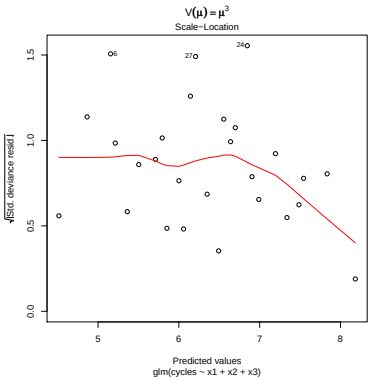
(b) Poissonverteilung



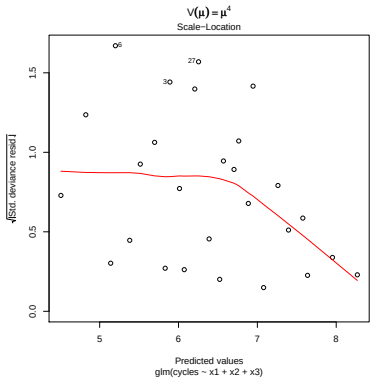
(c) Gammaverteilung



(d) Tweedieverteilung mit optimalem Parameter $\theta = 2.51$



(e) Inverse-Gauß-Verteilung



(f) Tweedieverteilung mit Parameter $\theta = 4$

Abbildung 3.5: Residuenplots für verschiedene Varianzfunktionen aus der Potenzfamilie

4 Die Sattelpunkt-Approximation

4.1 Grundlagen

In diesem Abschnitt untersuchen wir die Eigenschaften der *Sattelpunkt-Approximation*. Insbesondere interessiert dabei der Zusammenhang mit der EQL-Funktion.

Die Sattelpunkt-Approximation geht auf Daniels (1954) zurück und kann als eine spezielle Variante einer (verschobenen) *Edgeworth-Approximation* angesehen werden.[1] Ziel der Sattelpunkt-Approximation ist es, eine Approximation der Dichte des arithmetischen Mittels $\bar{Y} = 1/n \cdot \sum_{i=1}^{n} Y_i$ von unabhängig identisch verteilten Zufallszahlen $Y_1, \ldots, Y_n$ mit $\mathbb{E}(Y_i) = \mu$ und $\operatorname{var}(Y_i) = \sigma^2 < \infty$ anzugeben. Allgemein lässt sich die Dichte von $\bar{Y}$ durch den Zentralen Grenzwertsatz für $n \to \infty$ durch die Normalverteilung approximieren:

$$\bar{Y}' := \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \xrightarrow{D} \mathcal{N}(0,1)\,. \tag{4.1}$$

Die Sattelpunkt-Approximation versucht nun diese Approximation dadurch zu verbessern, dass auch höhere Momente in der Approximation Einzug finden. Die folgenden Herleitungen orientieren sich an Barndorff-Nielsen und Cox (1989), Friedl (1991), Reid (1991) und Jensen (1995).

4.1.1 Edgeworth-Approximation

Bevor wir uns der Sattelpunkt-Approximation selbst widmen, untersuchen wir die Edgeworth-Approximation, aus deren verschobenen Variante sich die Sattelpunkt-Approximation herleiten lässt. Dazu betrachten wir die momenterzeugende und die kumulantenerzeugende Funktion der Zufallsvariablen Y, die – Existenz des Erwartungswertes für t in einem offenen Intervall um Null vorausgesetzt – durch

$$M_Y(t) = \mathbb{E}(\exp\{tY\}) \tag{4.2a}$$

$$K_Y(t) = \log M_Y(t) \tag{4.2b}$$

gegeben sind. Wir bezeichnen das k-te Moment mit $\mu'_k = \mathbb{E}(Y^k)$ und das zentrierte k-te Moment wieder mit $\mu_k = \mathbb{E}((Y - \mu'_1)^k)$. Wenn wir für den Ausdruck (4.2a) eine

[1] Eine andere Herleitung beruht auf der inversen Fourier Transformation. Dabei wird der Integrationspfad derart gewählt, dass er entlang der Richtung des steilsten Abstiegs (engl. *steepest descent*) durch den Sattelpunkt des Integranden geht, daher der Name Sattelpunkt-Approximation.

Taylorentwicklung um Null durchführen, erhalten wir:

$$\begin{aligned} M_Y(t) &= \mathbb{E}\left(\exp\{tY\}\right) \\ &= \mathbb{E}\left(1 + tY + \frac{1}{2!}(tY)^2 + \cdots + \frac{1}{r!}(tY)^r + \cdots\right) \\ &= 1 + t\mu_1' + \frac{t^2}{2!}\mu_2' + \cdots + \frac{t^r}{r!}\mu_r' + \cdots . \end{aligned}$$

Für die Taylorentwicklung der kumulantenerzeugenden Funktion (4.2b) erhalten wir mit der Definition der k-ten Kumulante

$$\kappa_k = \left.\frac{\partial^k K_Y(t)}{\partial t^k}\right|_{t=0}$$

die folgende Beziehung:

$$\begin{aligned} K_Y(t) &= K_Y(0) + t\left.\frac{\partial K_Y(t)}{\partial t}\right|_{t=0} + \frac{t^2}{2}\left.\frac{\partial^2 K_Y(t)}{\partial t^2}\right|_{t=0} + \cdots + \frac{t^r}{r!}\left.\frac{\partial^r K_Y(t)}{\partial t^r}\right|_{t=0} + \cdots \\ &= t\kappa_1 + \frac{t^2}{2!}\kappa_2 + \cdots + \frac{t^r}{r!}\kappa_r + \cdots . \end{aligned} \tag{4.3}$$

Da die Y_i unabhängig identisch verteilt sind, lässt sich die momenterzeugende Funktion für die Summe $S_n = Y_1 + \cdots + Y_n$ über

$$\begin{aligned} M_{S_n}(t) &= \mathbb{E}\left(\exp\{tS_n\}\right) = \mathbb{E}\left(\exp\{t(Y_1 + \cdots + Y_n)\}\right) \\ &= \mathbb{E}\left(\prod_{i=1}^{n}\exp\{tY_i\}\right) \overset{\text{ind.}}{=} \prod_{i=1}^{n}\mathbb{E}\left(\exp\{tY_i\}\right) \\ &= \prod_{i=1}^{n} M_{Y_i}(t) \overset{\text{id.}}{=} \left[M_{Y_1}(t)\right]^n \end{aligned} \tag{4.4}$$

berechnen. Damit folgt für die kumulantenerzeugende Funktion von S_n sofort ein additiver Zusammenhang:

$$\begin{aligned} K_{S_n}(t) &= \log M_{S_n}(t) \overset{(4.4)}{=} \log\left[M_{Y_1}(t)\right]^n \\ &= n\log M_{Y_1}(t) = nK_{Y_1}(t). \end{aligned} \tag{4.5}$$

Bemerkung. Da die Y_i unabhängig identisch verteilt sind, gilt $\forall i, j : M_{Y_i}(t) = M_{Y_j}(t)$. Wir schreiben daher für die momenterzeugende Funktion der $Y_i : M_Y(t)$ und lassen den Index weg.

Nun sind wir in der Lage die kumulantenerzeugende Funktion für den standardisierten Mittelwert (4.1) anzugeben:

$$\begin{aligned}
K_{\bar{Y}'}(t) &= \log M_{\bar{Y}'}(t) \\
&= \log\left[\mathbb{E}\left(\exp\left\{\frac{1/n\cdot S_n-\mu}{\sigma/\sqrt{n}}\cdot t\right\}\right)\right] \\
&= \log\left[\exp\left\{\frac{-\mu t}{\sigma/\sqrt{n}}\right\}\right] + \underbrace{\log\left[\mathbb{E}\left(\exp\left\{\frac{t}{\sigma\sqrt{n}}S_n\right\}\right)\right]}_{=K_{S_n}(\frac{t}{\sigma\sqrt{n}})} \\
&= -\frac{\mu t}{\sigma/\sqrt{n}} + K_{S_n}\left(\frac{t}{\sigma\sqrt{n}}\right) \overset{(4.5)}{=} -\frac{\mu t}{\sigma/\sqrt{n}} + nK_Y\left(\frac{t}{\sigma\sqrt{n}}\right).
\end{aligned}$$

Verwenden wir nun die Taylorreihenentwicklung (4.3) und brechen nach dem vierten Glied ab erhalten wir mit $\kappa_1=\mu, \kappa_2=\sigma^2$ und den standardisierten Kumulanten $\rho_k = \kappa_k/\kappa_2^{k/2} = \kappa_k/\sigma^k$ für $n\to\infty$ die Approximation:

$$\begin{aligned}
K_{\bar{Y}'}(t) &= -\frac{\mu t}{\sigma/\sqrt{n}} + nK_Y\left(\frac{t}{\sigma\sqrt{n}}\right) \\
&= -\frac{\mu t}{\sigma/\sqrt{n}} + n\left[\frac{t}{\sigma\sqrt{n}}\kappa_1 + \frac{1}{2!}\frac{t^2}{\sigma^2 n}\kappa_2 + \frac{1}{3!}\frac{t^3}{\sigma^3 n^{3/2}}\kappa_3 + \frac{1}{4!}\frac{t^4}{\sigma^4 n^2}\kappa_4 + \mathcal{O}(n^{-5/2})\right] \\
&= \frac{t^2}{2} + \frac{t^3}{6\sqrt{n}}\frac{\kappa_3}{\sigma^3} + \frac{t^4}{24n}\frac{\kappa_4}{\sigma^4} + \mathcal{O}(n^{-3/2}) \\
&= \frac{t^2}{2} + \frac{t^3}{6\sqrt{n}}\rho_3 + \frac{t^4}{24n}\rho_4 + \mathcal{O}(n^{-3/2}).
\end{aligned} \tag{4.6}$$

Aus (4.6) lässt sich nun auch eine Approximation für die momenterzeugende Funktion selbst finden, indem wir für $\exp\{\cdot\}$ wieder eine Taylorreihenentwicklung durchführen:

$$\begin{aligned}
M_{\bar{Y}'}(t) &= \exp\left\{K_{\bar{Y}'}(t)\right\} \\
&= \exp\left\{\frac{t^2}{2}\right\}\exp\left\{\frac{t^3}{6\sqrt{n}}\rho_3 + \frac{t^4}{24n}\rho_4 + \mathcal{O}(n^{-3/2})\right\} \\
&= \exp\left\{\frac{t^2}{2}\right\}\left(1 + \frac{t^3}{6\sqrt{n}}\rho_3 + \frac{t^4}{24n}\rho_4 + \frac{1}{2!}\left[\frac{t^3}{6\sqrt{n}}\rho_3 + \frac{t^4}{24n}\rho_4\right]^2 + \mathcal{O}(n^{-3/2})\right) \\
&= \exp\left\{\frac{t^2}{2}\right\}\left(1 + \frac{t^3}{6\sqrt{n}}\rho_3 + \frac{t^4}{24n}\rho_4 + \frac{t^6}{72n}\rho_3^2 + \mathcal{O}(n^{-3/2})\right).
\end{aligned} \tag{4.7}$$

Bezeichne $\varphi(x)$ die Dichte der Standardnormalverteilung, dann sind die Hermite-Polynome $H_k(x)$ durch

$$H_k(x) = (-1)^k\frac{1}{\varphi(x)}\frac{\partial^k\varphi(x)}{\partial x^k} = (-1)^k\,\mathrm{e}^{x^2/2}\,\frac{\partial^k}{\partial x^k}\,\mathrm{e}^{-x^2/2}$$

definiert. Um aus der momenterzeugenden Funktion die Dichte zu erhalten, bedienen wir uns der folgenden Identität:

$$\int_{-\infty}^{\infty} \mathrm{e}^{tz}\, H_k(z)\varphi(z)\, dz = t^k\, \mathrm{e}^{t^2/2}, \tag{4.8}$$

die sich durch wiederholte Anwendung der partiellen Integration wie folgt ergibt:

$$\begin{aligned}
\int_{-\infty}^{\infty} \mathrm{e}^{tz}\, H_k(z)\varphi(z)\, dz &= \int_{-\infty}^{\infty} \mathrm{e}^{tz}(-1)^k \frac{\partial^k \varphi(z)}{\partial z^k}\, dz \\
&= \underbrace{(-1)^k\, \mathrm{e}^{tz}\, \frac{\partial^{k-1}\varphi(z)}{\partial z^{k-1}}\bigg|_{-\infty}^{\infty}}_{=-\,\mathrm{e}^{tz}\, H_{k-1}(z)\varphi(z)\big|_{-\infty}^{\infty}=0} + (-1)^{k+1} \int_{-\infty}^{\infty} t\, \mathrm{e}^{tz}\, \frac{\partial^{k-1}\varphi(z)}{\partial z^{k-1}}\, dz \\
&= \underbrace{(-1)^{k+1} t\, \mathrm{e}^{tz}\, \frac{\partial^{k-2}\varphi(z)}{\partial z^{k-2}}\bigg|_{-\infty}^{\infty}}_{=0} + (-1)^{k+2} \int_{-\infty}^{\infty} t^2\, \mathrm{e}^{tz}\, \frac{\partial^{k-2}\varphi(z)}{\partial z^{k-2}}\, dz \\
&\;\;\vdots \\
&= (-1)^{2k} t^k \underbrace{\int_{-\infty}^{\infty} \mathrm{e}^{tz}\, \varphi(z)\, dz}_{=M_\varphi(t)=\exp(t^2/2)} \\
&= t^k\, \mathrm{e}^{t^2/2}\,.
\end{aligned}$$

Bemerkung. Die Tatsache, dass der erste Term der partiellen Integration stets verschwindet, lässt sich darauf zurückführen, dass die Exponentialfunktion schneller wächst als jedes Polynom:

$$\begin{aligned}
\lim_{z\to\pm\infty} \mathrm{e}^{tz}\, H_m(z)\varphi(z) &= \frac{1}{\sqrt{2\pi}} \lim_{z\to\pm\infty} \mathrm{e}^{tz}\, H_m(z)\, \mathrm{e}^{-z^2/2} \\
&= \frac{1}{\sqrt{2\pi}} \lim_{z\to\pm\infty} \mathrm{e}^{tz-z^2/2} \cdot \text{Polynom } m\text{-ten Grades in } z \\
&= \frac{1}{\sqrt{2\pi}} \lim_{z\to\pm\infty} \frac{1}{\mathrm{e}^{z^2/2-tz}} \cdot \text{Polynom } m\text{-ten Grades in } z \\
&= 0.
\end{aligned}$$

Mit der Identität (4.8) lässt sich nun Gleichung (4.7) umschreiben:

$$\begin{aligned}
M_{\bar{Y}'}(t) &= \exp\left\{\frac{t^2}{2}\right\}\left(1+\frac{t^3}{6\sqrt{n}}\rho_3+\frac{t^4}{24n}\rho_4+\frac{t^6}{72n}\rho_3^2+\mathcal{O}(n^{-3/2})\right)\\
&= \int_{-\infty}^{\infty} \mathrm{e}^{t\bar{y}'}\, H_0(\bar{y}')\varphi(\bar{y}')\,d\bar{y}' + \frac{\rho_3}{6\sqrt{n}}\int_{-\infty}^{\infty} \mathrm{e}^{t\bar{y}'}\, H_3(\bar{y}')\varphi(\bar{y}')\,d\bar{y}' +\\
&\quad \frac{\rho_4}{24n}\int_{-\infty}^{\infty} \mathrm{e}^{t\bar{y}'}\, H_4(\bar{y}')\varphi(\bar{y}')\,d\bar{y}' + \frac{\rho_3^2}{72n}\int_{-\infty}^{\infty} \mathrm{e}^{t\bar{y}'}\, H_6(\bar{y}')\varphi(\bar{y}')\,d\bar{y}' + \mathcal{O}(n^{-3/2})\\
&= \int_{-\infty}^{\infty} \mathrm{e}^{t\bar{y}'}\left[\varphi(\bar{y}')\Big(H_0(\bar{y}')+\frac{\rho_3}{6\sqrt{n}}H_3(\bar{y}')+\frac{\rho_4}{24n}H_4(\bar{y}')+\right.\\
&\qquad\qquad \left.\left.\frac{\rho_3^2}{72n}H_6(\bar{y}')+\mathcal{O}(n^{-3/2})\Big)\right]d\bar{y}'. \qquad (4.9)
\end{aligned}$$

Ein Koeffizientenvergleich für $M_{\bar{Y}'}(t) = \int_{-\infty}^{\infty} \mathrm{e}^{t\bar{y}'} f_{\bar{Y}'}(\bar{y}')\,d\bar{y}'$ mit der Gleichung (4.9) ergibt mit $H_0(\bar{y}') = 1$ die Edgeworth-Approximation für die Dichte des standardisierten Mittelwertes $\bar{Y}'$:

$$f_{\bar{Y}'}(\bar{y}') = \varphi(\bar{y}')\left[1+\frac{\rho_3}{6\sqrt{n}}H_3(\bar{y}')+\frac{\rho_4}{24n}H_4(\bar{y}')+\frac{\rho_3^2}{72n}H_6(\bar{y}')+\mathcal{O}(n^{-3/2})\right]. \qquad (4.10)$$

Bemerkung 1. Eine Approximation für die Dichte der Summe S_n lässt sich mit dem Transformationssatz für Dichten aus der Dichte für $\bar{Y}'$ in Gleichung (4.10) bestimmen und lautet (vgl. Brazzale u. a., 2007, S. 209, dort fehlt aber der Vorfaktor $1/\sigma$):

$$f_{S_n}(s_n) = f_{\bar{Y}'}(g(s_n))\left|\frac{\partial g(s_n)}{\partial s_n}\right| \quad \text{mit} \quad z = g(s_n) = \frac{s_n - n\mu}{\sqrt{n\sigma^2}},$$

$$f_{S_n}(s_n) = \frac{1}{\sqrt{n\sigma^2}}\,\varphi(z)\left\{1+\frac{\rho_3}{6\sqrt{n}}H_3(z)+\frac{\rho_4}{24n}H_4(z)+\frac{\rho_3^2}{72n}H_6(z)+\mathcal{O}(n^{-3/2})\right\}. \qquad (4.11)$$

Eine Approximation für die gewichtete Summe $\bar{Y} = S_n/n$ lässt sich analog bestimmen:

$$z = g(\bar{y}) = \frac{n\bar{y} - n\mu}{\sqrt{n\sigma^2}},$$

$$f_{\bar{Y}}(\bar{y}) = \frac{\sqrt{n}}{\sqrt{\sigma^2}}\varphi(z)\left\{1+\frac{\rho_3}{6\sqrt{n}}H_3(z)+\frac{\rho_4}{24n}H_4(z)+\frac{\rho_3^2}{72n}H_6(z)+\mathcal{O}(n^{-3/2})\right\}. \qquad (4.12)$$

Bemerkung 2. Der führende Term in der Edgeworth-Approximation (4.10) entspricht der (standardisierten) Normalverteilung, was mit der Tatsache einhergeht, dass $\bar{Y}'$ in der Verteilung für $n \to \infty$ gegen $\mathcal{N}(0,1)$ strebt. Die Edgeworth-Approximation berücksichtigt aber eben auch die höheren Momente, wobei die Schiefe durch den Term $n^{-1/2}$ skaliert wird, während der Faktor n^{-1} die Schiefe und die Wölbung gleichzeitig anpasst (vgl. Reid, 1991).

Bemerkung 3. Eine Approximation der Verteilungsfunktion von $\bar{Y}'$ ergibt sich unter der Verwendung von

$$-\frac{\partial}{\partial z}\Big[\varphi(z)H_k(z)\Big] = -\frac{\partial}{\partial z}\Big[(-1)^k\frac{\partial^k\varphi(z)}{\partial z^k}\Big] = (-1)^{k+1}\frac{\partial^{k+1}\varphi(z)}{\partial z^{k+1}} = H_{k+1}(z)\varphi(z) \qquad (*)$$

durch Integration der Dichte:

$$
\begin{aligned}
F_{\bar{Y}'}(\bar{y}') &= \int_{-\infty}^{\bar{y}'} f_{\bar{Y}'}(y)\,dy \\
&\approx \int_{-\infty}^{\bar{y}'} \varphi(y)\left[1 + \frac{\rho_3}{6\sqrt{n}}H_3(y) + \frac{1}{n}\left[\frac{\rho_4}{24}H_4(y) + \frac{\rho_3^2}{72}H_6(y)\right]\right] dy \\
&= \int_{-\infty}^{\bar{y}'} \varphi(y)\,dy + \frac{\rho_3}{6\sqrt{n}}\int_{-\infty}^{\bar{y}'} H_3(y)\varphi(y)\,dy + \frac{\rho_4}{24n}\int_{-\infty}^{\bar{y}'} H_4(y)\varphi(y)\,dy + \\
&\quad \frac{\rho_3^2}{72n}\int_{-\infty}^{\bar{y}'} H_6(y)\varphi(y)\,dy \\
&\overset{(*)}{=} \Phi(\bar{y}') - \frac{\rho_3}{6\sqrt{n}}H_2(\bar{y}')\varphi(\bar{y}') - \frac{\rho_4}{24n}H_3(\bar{y}')\varphi(\bar{y}') - \frac{\rho_3^2}{72n}H_5(\bar{y}')\varphi(\bar{y}') \\
&= \Phi(\bar{y}') - \varphi(\bar{y}')\left[\frac{\rho_3}{6\sqrt{n}}H_2(\bar{y}') + \frac{\rho_4}{24n}H_3(\bar{y}') + \frac{\rho_3^2}{72n}H_5(\bar{y}')\right].
\end{aligned}
$$

Bemerkung 4. Die in der Herleitung der Dichte verwendete Taylorentwicklung wurde beim vierten Glied abgebrochen. Prinzipiell ist es möglich noch weitere Terme in die Entwicklung miteinzubeziehen. Abgesehen von der Notwendigkeit dann auch höhere Momente bestimmen zu müssen, sind asymptotische Entwicklungen keine konvergenten Reihen. Das heißt, dass die Hinzunahme von weiteren Termen nicht unbedingt eine höhere Genauigkeit bedeutet (vgl. Brazzale u. a., 2007, S. 2).

Blinnikov und Moessner (1998) geben einen Algorithmus an, mit dem die höheren Terme der Edgeworth-Approximation leicht bestimmt werden können.

Bemerkung 5. Existiert das dritte zentrierte Moment, gibt der *Satz von Berry-Esseen* (Berry, 1941; Esseen, 1945) für den Zentralen Grenzwertsatz die Güte der Approximation mit $\mathcal{O}(n^{-1/2})$ an.

Bemerkung 6. Alle mit ungeraden Potenzen von $n^{-1/2}$ gewichteten Hermite-Polynome in Gleichung (4.10) weisen ebenfalls einen ungeraden Grad auf – für ebensolche gilt $H_{2l+1}(0) = 0$. Berechnen wir also die Näherung für $f_{\bar{Y}'}(0)$ (was dem Erwartungswert von $\bar{Y}'$ entspricht), erhalten wir eine Entwicklung in Potenzen von n^{-1} anstatt nur in Potenzen von $n^{-1/2}$. Die Edgeworth-Approximation ist im Erwartungswert also genauer als die Approximation durch die Normalverteilung, wie sie durch den Zentralen Grenzwertsatz motiviert wird. Auf der anderen Seite ist die Approximation in den Schwänzen aus dem selben Grund schlechter und kann sogar negative Werte liefern (vgl. Barndorff-Nielsen und Cox, 1979).

Bemerkung 7. Blinnikov und Moessner (1998) vergleichen drei verschiedene auf Hermite-Polynomen beruhenden Approximationen (Gram-Charlier, Gauß-Hermite und Edgeworth-Approximation) für Verteilungen, die der Normalverteilung ähnlich sind. Sie kommen zu dem Schluss, dass die Edgeworth-Approximation in diesem Fall die besten Ergebnisse liefert.

4.1.2 Exponential Tilting und die Sattelpunkt-Approximation

Die Schwäche der Edgeworth-Approximation in den Schwänzen führt zu der Idee der *verschobenen Edgeworth-Approximation*. Dazu betrachten wir wieder unabhängig identisch verteilte Zufallszahlen $Y_1, \ldots, Y_n$ mit Erwartungswert $\mathbb{E}(Y_i) = \mu$, Varianz $\operatorname{var}(Y_i) = \sigma^2 < \infty$ und Dichte $f_Y(y)$. Wir wollen weiters annehmen, dass die momenterzeugende und die kumulantenerzeugende Funktion durch

$$M_Y(t) = \mathbb{E}(\exp\{tY\}) =: \exp b(t) \qquad K_Y(t) = \log M_Y(t) = b(t)$$

gegeben sind.

Das auf Esscher (1932) zurückgehende *Exponential Tilting* (in etwa: exponentielles Verschieben) ist ein Verfahren, bei dem die Dichte in eine *exponentiell gewichtete Dichte* mit dem Parameter θ eingebettet wird:

$$f_{Y^*}(y^*;\theta) \propto \exp\{\theta y^*\} f_Y(y^*). \tag{4.13}$$

Bemerkung. Mit der Wahl von $\theta = 0$ erhält man aus (4.13) sofort wieder die ursprüngliche Dichte:

$$f_{Y^*}(y^*;0) = \exp\{0\} f_Y(y^*) = f_Y(y^*).$$

Der Ausdruck (4.13) ist streng genommen noch keine Dichte, weil das Integral von $f_{Y^*}(y^*;\theta)$ nicht notwendigerweise Eins ergibt. Daher normalisieren wir $f_{Y^*}(y^*;\theta)$ und erhalten:

$$c\int_{-\infty}^{\infty} f_{Y^*}(y^*;\theta)\,dy^* = c\underbrace{\int_{-\infty}^{\infty} \exp\{\theta y^*\} f_Y(y^*)\,dy^*}_{=M_Y(\theta)=\exp b(\theta)} \stackrel{!}{=} 1 \Rightarrow c = \exp\{-b(\theta)\},$$

$$f_{Y^*}(y^*;\theta) = \exp\{\theta y^* - b(\theta)\} f_Y(y^*) = \exp\{\theta y^* - b(\theta) + \log f_Y(y^*)\}. \tag{4.14}$$

Die Dichte in Gleichung (4.14) weist für $a(\phi) = 1$ und $c(y,\phi) = \log f_Y(y^*)$ die Gestalt einer Dichte aus der Exponentialfamilie (vgl. Definition 2.2) auf, daher wird die Familie der Verteilungen mit Dichte $f_{Y^*}(y^*;\theta)$ auch als *konjugierte Exponentialfamilie* bezeichnet.

Somit können wir aus den Gleichungen (2.7a), (2.7b) und (2.7c) sofort die momenterzeugende und die kumulantenerzeugende Funktion, sowie die Kumulanten selbst bestimmen:

$$\begin{aligned} M_{Y^*}(t;\theta) &= \exp\{b(\theta+t) - b(\theta)\} &&= M_Y(\theta+t)/M_Y(\theta), \\ K_{Y^*}(t;\theta) &= b(\theta+t) - b(\theta) &&= K_Y(\theta+t) - K_Y(\theta), \\ \kappa_k^*(\theta) &= \left.\frac{\partial^k K_{Y^*}(t;\theta)}{\partial t^k}\right|_{t=0} &&= \frac{\partial^k K_Y(\theta)}{\partial \theta^k}. \end{aligned}$$

Betrachtet man wieder die Summe $S_n^* = Y_1^* + \cdots + Y_n^*$ von unabhängig identisch verteilten Zufallszahlen Y_i^* mit der Dichte $f_{Y^*}(y^*;\theta)$, erhält man mittels der Gleichungen (4.4) und (4.5) die zugehörigen erzeugenden Funktionen $M_{S_n^*}(t;\theta)$ und $K_{S_n^*}(t;\theta)$:

$$M_{S_n^*}(t;\theta) = [M_{Y^*}(t;\theta)]^n = \exp\{n[b(\theta+t) - b(\theta)]\}, \tag{4.15a}$$

$$K_{S_n^*}(t;\theta) = n \cdot K_{Y^*}(t;\theta) = n[b(\theta+t) - b(\theta)]. \tag{4.15b}$$

Bemerkung. Die Dichte $f_{S_n^*}(s_n^*;\theta)$ der Summe lässt sich durch Faltung bestimmen und lautet mit Gleichung (4.14):

$$\begin{aligned} f_{S_n^*}(s_n^*;\theta) &= \int_{y_1+\cdots+y_n=s_n^*} \cdots \int \prod_{i=1}^{n} f_{Y_i^*}(y_i;\theta)\, dy_n \cdots dy_1 \\ &= \int_{y_1+\cdots+y_n=s_n^*} \cdots \int \exp\Big\{\theta \underbrace{(y_1+\cdots+y_n)}_{=s_n^*} - nb(\theta)\Big\} \prod_{i=1}^{n} f_{Y_i}(y_i)\, dy_n \cdots dy_1 \\ &= \exp\{\theta s_n^* - nb(\theta)\} \int_{y_1+\cdots+y_n=s_n^*} \cdots \int \prod_{i=1}^{n} f_{Y_i}(y_i)\, dy_n \cdots dy_1 \\ &= \exp\{\theta s_n^* - nK_Y(\theta)\}\, f_{S_n}(s_n^*). \end{aligned} \tag{4.16}$$

Um für den Mittelwert $\bar{Y}^* = S_n^*/n$ die erzeugenden Funktionen bestimmen zu können, benutzen wir die Identität

$$M_{\bar{Y}^*}(t;\theta) = M_{S_n^*/n}(t;\theta) = \mathbb{E}\left[\exp\{S_n^* t/n\}\right] = M_{S_n^*}(t/n;\theta)$$

und erhalten somit schließlich die Beziehungen:

$$M_{\bar{Y}^*}(t;\theta) = M_{S_n^*}(t/n;\theta) \qquad = \exp\{n[b(\theta+t/n)-b(\theta)]\}, \tag{4.17a}$$

$$K_{\bar{Y}^*}(t;\theta) = \log M_{\bar{Y}^*}(t;\theta) \qquad = n[b(\theta+t/n)-b(\theta)], \tag{4.17b}$$

$$\bar{\kappa}_k^*(\theta) = \left.\frac{\partial^k K_{\bar{Y}^*}(t;\theta)}{\partial t^k}\right|_{t=0} = \frac{1}{n^{k-1}} \frac{\partial^k K_Y(\theta)}{\partial \theta^k}. \tag{4.17c}$$

Die Tabelle 4.1 fasst noch einmal alle bis jetzt definierten Größen zusammen.

Bemerkung 1. Die momenterzeugende Funktion in Gleichung (4.17a) entspricht einer momenterzeugenden Funktion einer Dichte aus der Exponentialfamilie mit $a(\phi) = 1/n$. Da die momenterzeugende Funktion die Verteilung eindeutig spezifiziert (vgl. Curtiss, 1942), ist somit auch die Dichte der gemittelten Summe $f_{\bar{Y}^*}(\bar{y}^*;\theta)$ ein Mitglied der Exponentialfamilie.

Bemerkung 2. Der Zusammenhang zwischen der verschobenen Dichte von $\bar{Y}^*$ und der ursprünglichen Dichte von $\bar{Y}$ lässt sich wieder mittels des Transformationssatzes für Dichten bestimmen und lautet mit Gleichung (4.16):

$$\begin{aligned} g(\bar{y}^*) &= n\bar{y}^* = s_n^* \\ f_{\bar{Y}^*}(\bar{y}^*;\theta) &= f_{S_n^*}\left(g(\bar{y}^*);\theta\right) \left|\frac{\partial g(\bar{y}^*)}{\partial \bar{y}^*}\right| \\ &= \exp\{\theta n\bar{y}^* - nK_Y(\theta)\} \underbrace{f_{S_n}(n\bar{y}^*)n}_{=f_{\bar{Y}}(\bar{y})} \\ &= \exp\{n[\theta\bar{y}^* - K_Y(\theta)]\}\, f_{\bar{Y}}(\bar{y}). \end{aligned} \tag{4.18}$$

Tabelle 4.1: Übersicht der eingeführten Terme für die Edgeworth- und Sattelpunkt-Approximation

	Y	Y^*	S_n^*	$\bar{Y}^*$
Name	Zufallsvariable	Eingebettete ZV	Summe	Gewichtete Σ
Dichte	$f_Y(y)$	$f_{Y^*}(y^*;\theta)$	$f_{S_n^*}(s_n^*;\theta)$	$f_{\bar{Y}^*}(\bar{y}^*;\theta)$
$M(t)$	$M_Y(t)$	$M_{Y^*}(t;\theta)$	$M_{S_n^*}(t;\theta)$	$M_{\bar{Y}^*}(t;\theta)$
Wert	$\mathrm{e}^{b(t)}$	$\mathrm{e}^{b(\theta+t)-b(\theta)}$	$\mathrm{e}^{n[b(\theta+t)-b(\theta)]}$	$\mathrm{e}^{n[b(\theta+t/n)-b(\theta)]}$
$K(t)$	$K_Y(t)$	$K_{Y^*}(t;\theta)$	$K_{S_n^*}(t;\theta)$	$K_{\bar{Y}^*}(t;\theta)$
Wert	$b(t)$	$b(\theta+t)-b(\theta)$	$n[b(\theta+t)-b(\theta)]$	$n[b(\theta+\frac{t}{n})-b(\theta)]$
κ_k	κ_k	$\kappa_k^*(\theta)$	$\kappa_{S_n;k}^*(\theta)$	$\bar{\kappa}_k^*(\theta)$
Wert	$\frac{\partial^k K_Y(t)}{\partial t^k}\Big\vert_{t=0}$	$\frac{\partial^k K_Y(\theta)}{\partial \theta^k}$	$n\frac{\partial^k K_Y(\theta)}{\partial \theta^k}$	$\frac{1}{n^{k-1}}\frac{\partial^k K_Y(\theta)}{\partial \theta^k}$
$\mathbb{E}(\cdot)$	μ	$\mu^*(\theta)=K_Y'(\theta)$	$nK_Y'(\theta)$	$K_Y'(\theta)$
$\mathrm{var}(\cdot)$	σ^2	$\sigma^{*2}(\theta)=K_Y''(\theta)$	$nK_Y''(\theta)$	$\frac{1}{n}K_Y''(\theta)$

ZV=Zufallsvariable, Σ=Summe

Führt man nun für die (exponentiell verschobene) Dichte von $\bar{Y}^* = S_n^*/n$ eine Edgeworth-Approximation durch, erhalten wir mit $\sigma^*(\theta) = \sqrt{\mathrm{var}\,(Y^*;\theta)}, \mu^*(\theta) = \mathbb{E}\,(Y^*;\theta)$ und $z = (n\bar{y} - n\mu^*(\theta))\,/(\sqrt{n}\sigma^*(\theta))$ aus Gleichung (4.12) die Beziehung:

$$f_{\bar{Y}^*}(\bar{y}^*;\theta) = \frac{\sqrt{n}}{\sigma^*(\theta)}\,\varphi(z)\left\{1 + \frac{\rho_3^*(\theta)}{6\sqrt{n}}H_3(z) + \frac{\rho_4^*(\theta)}{24n}H_4(z) + \frac{\rho_3^*(\theta)^2}{72n}H_6(z) + \mathcal{O}(n^{-3/2})\right\}, \tag{4.19}$$

wobei $\rho_k^*(\theta)$ die k-te standardisierte Kumulante von Y^* bezeichnet.

Bemerkung. Für die standardisierten Kumulanten $\bar{\rho}_k^*(\theta)$ von $\bar{Y}^*$ gilt:

$$\bar{\kappa}_k^*(\theta) = \frac{1}{n^{k-1}}\frac{\partial^k K_Y(\theta)}{\partial\theta^k}, \qquad \bar{\rho}_k^*(\theta) = \frac{\bar{\kappa}_k^*(\theta)}{(\bar{\kappa}_2^*(\theta))^{k/2}},$$

$$\bar{\rho}_3^*(\theta) = \frac{\frac{1}{n^2}K_Y'''(\theta)}{\left(\frac{1}{n}K_Y''(\theta)\right)^{3/2}} = \frac{\rho_3^*(\theta)}{\sqrt{n}}, \qquad \bar{\rho}_4^*(\theta) = \frac{\frac{1}{n^3}K_Y''''(\theta)}{\left(\frac{1}{n}K_Y''(\theta)\right)^{2}} = \frac{\rho_4^*(\theta)}{n}.$$

Außerdem gilt

$$\mathrm{var}\left(\bar{Y}^*;\theta\right) = \bar{\kappa}_2^*(\theta) = \frac{1}{n}K_Y''(\theta) = \frac{1}{n}\,\mathrm{var}\,(Y^*;\theta) = \frac{\sigma^*(\theta)^2}{n},$$
$$\mathbb{E}\left(\bar{Y}^*;\theta\right) = \bar{\kappa}_1^*(\theta) = K_Y'(\theta) = \mathbb{E}\,(Y^*;\theta) = \mu^*(\theta).$$

Somit lässt sich Gleichung (4.19) auch zu

$$f_{\bar{Y}^*}(\bar{y}^*;\theta) = \frac{1}{\sqrt{\mathrm{var}\left(\bar{Y}^*;\theta\right)}}\,\varphi(z)\left\{1+\frac{\bar{\rho}_3^*(\theta)}{6}H_3(z)+\frac{\bar{\rho}_4^*(\theta)}{24}H_4(z)+\frac{\bar{\rho}_3^*(\theta)^2}{72}H_6(z)+\mathcal{O}(n^{-3/2})\right\}$$

umschreiben, wobei

$$z = \frac{\bar{y}^* - \mathbb{E}\left(\bar{Y}^*; \theta\right)}{\sqrt{\operatorname{var}\left(\bar{Y}^*; \theta\right)}}$$

gilt.

Durch die Einbettung der ursprünglichen Dichte $f_Y(y)$ in eine exponentielle gewichtete Dichte $f_{Y^*}(y^*; \theta)$ mit dem Parameter θ sind wir nun in der Lage, die Schwäche der Edgeworth-Approximation zu beheben. Diese liefert im Zentrum um den Erwartungswert gute Approximationen, während die Annäherung in den Schwänzen der Verteilung schlecht ist. Der Mittelwert der verschobenen Dichte $f_{Y^*}(y^*; \theta)$ hängt vom Parameter θ ab, weil $\mathbb{E}\left(Y^*\right) = K'_Y(\theta)$ gilt. Da der Parameter θ aber beliebig ist, wählen wir ihn so, dass der Erwartungswert gerade den Wert ergibt, an dem die Dichte ausgewertet werden soll. Es soll also der Sattelpunkt $\hat{\theta}$ so bestimmt werden, dass

$$\mu^*(\hat{\theta}) = K'(\hat{\theta}) = \bar{y}^*$$

gilt. Damit wird die Edgeworth-Approximation stets im Erwartungswert ausgewertet und es gilt außerdem:

$$z = \frac{n\bar{y}^* - n\mu^*(\hat{\theta})}{\sqrt{n}\sigma^*(\hat{\theta})} = 0.$$

Die Hermite-Polynome ergeben an dieser Stelle die Werte

$$H_3(0) = 0, \qquad H_4(0) = 3, \qquad H_6(0) = -15$$

und somit erhält man für die im Sattelpunkt ausgewertete Edgeworth-Approximation die Beziehung:

$$\begin{aligned} f_{\bar{Y}^*}(\bar{y}^*; \hat{\theta}) &= \frac{\sqrt{n}}{\sigma^*(\hat{\theta})}\,\varphi(0)\left\{1 + \frac{\rho_3^*(\hat{\theta})}{6\sqrt{n}}H_3(0) + \frac{\rho_4^*(\hat{\theta})}{24n}H_4(0) + \frac{\rho_3^*(\hat{\theta})^2}{72n}H_6(0) + \mathcal{O}(n^{-3/2})\right\} \\ &= \frac{\sqrt{n}}{\sigma^*(\hat{\theta})\sqrt{2\pi}}\left\{1 + \frac{3\rho_4^*(\hat{\theta}) - 5\rho_3^*(\hat{\theta})^2}{24n} + \mathcal{O}(n^{-3/2})\right\}. \end{aligned}$$

Für die ursprüngliche Dichte der gewichteten Summe $f_{\bar{Y}}(\bar{y})$ folgt mit Gleichung (4.18) und $\sigma^*(\hat{\theta}) = \sqrt{K''_Y(\hat{\theta})}$ schließlich die *Sattelpunkt-Approximation*:

$$f_{\bar{Y}}(\bar{y}) = \sqrt{\frac{n}{2\pi K''_Y(\hat{\theta})}}\exp\left\{n[K_Y(\hat{\theta}) - \hat{\theta}\bar{y})]\right\}\left\{1 + \frac{3\rho_4^*(\hat{\theta}) - 5\rho_3^*(\hat{\theta})^2}{24n} + \mathcal{O}(n^{-3/2})\right\}. \quad (4.20)$$

Bemerkung 1. In der Herleitung der Edgeworth-Approximation haben wir in der Gleichung (4.7) eine Taylorentwicklung für die Exponentialfunktion durchgeführt und bei Termen der Ordnung $\mathcal{O}(n^{-3/2})$ abgebrochen. Führt man diese Entwicklung bis zur Ordnung $\mathcal{O}(n^{-2})$ durch, sind die zu $n^{-3/2}$ gehörigen Hermite-Polynome vom Grad sieben

und neun. Mit $H_7(0) = 0$ und $H_9(0) = 0$ erhält man somit eine höhere Genauigkeit und es gilt:

$$f_{\bar{Y}}(\bar{y}) = \sqrt{\frac{n}{2\pi K_Y''(\hat{\theta})}} \exp\left\{n[K_Y(\hat{\theta}) - \hat{\theta}\bar{y})]\right\} \left\{1 + \frac{3\rho_4^*(\hat{\theta}) - 5\rho_3^*(\hat{\theta})^2}{24n} + \mathcal{O}(n^{-2})\right\}. \quad (4.21)$$

Bemerkung 2. Die Approximation in Gleichung (4.20) kann (re-)normalisiert werden. Dazu betrachten wir

$$f_{\bar{Y}}(\bar{y}) \approx c\sqrt{\frac{n}{K_Y''(\hat{\theta})}} \exp\left\{n[K_Y(\hat{\theta}) - \hat{\theta}\bar{y})]\right\} \quad (4.22)$$

und bestimmen den Faktor c derart, dass das Integral der rechten Seite der Approximation (4.22) Eins ergibt (vgl. Reid, 1991).

Bemerkung 3. Die höheren Momente in Gleichung (4.20) hängen über $\hat{\theta}$ vom Punkt $\bar{y}$ ab an dem die Dichte ausgewertet werden soll. In der Näherungsformel (4.22) werden diese Momente aber für die Bestimmung des Normalisierungsfaktors c nicht berücksichtigt. Sind $\rho_3^*(\hat{\theta})$ und $\rho_4^*(\hat{\theta})$ allerdings konstant in y, ist die Approximation (4.22) exakt (vgl. Reid, 1991). Daniels (1980) zeigt, dass dies im skalaren Fall nur für die Normal-, die Gamma- und die Inverse-Gauß-Verteilung zutrifft.

Bemerkung 4. Im allgemeinen variiert der Korrekturterm $(3\rho_4^*(\hat{\theta}) - 5\rho_3^*(\hat{\theta})^2)/24n$ in Gleichung (4.21) für unterschiedliche Werte von y nur langsam und der relative Fehler der (re-)normalisierten Sattelpunkt-Approximation ist $\mathcal{O}(n^{-3/2})$ (Reid, 1991).

4.2 Anwendungen der Sattelpunkt-Approximation

4.2.1 Approximation von Verteilungen aus der Exponentialfamilie

Von besonderem Interesse ist nun aufgrund der einfachen Gestalt der Kumulantenfunktion die Approximation von Verteilungen aus der Exponentialfamilie. Da eine Dichte aus der Exponentialfamilie bereits die Form einer exponentiell gewichteten Dichte aufweist, resultiert für die Sattelpunkt-Approximation eine einfache Vorschrift.

Seien dazu $Y_1, \ldots, Y_n$ unabhängig identisch verteilte Zufallsvariablen mit einer Dichte aus der 1-parametrischen Exponentialfamilie (vgl. Definition 2.2):

$$f_Y(y; \theta) = \exp\{y\theta - b(\theta) + c(y)\},$$

dabei bezeichnet $b(t)$ die Kumulantenfunktion und es gilt mit Gleichung (2.7a) $K_Y(t) = b(\theta + t) - b(\theta)$. Wenn wir den Parameter θ für den Moment als einen Einbettungsparameter, wie er in Gleichung (4.14) definiert wird, betrachten, folgt für die Dichte $f_{S_n}(s_n; \theta)$ der Summe $S_n = Y_1 + \cdots + Y_n$ aus der Faltung in Gleichung (4.16):

$$f_{S_n}(s_n; \theta) = \exp\{\theta s_n - nb(\theta) + h(s_n)\}, \quad (4.23)$$

mit einer passend gewählten Funktion $h(s_n)$.

Eine Edgeworth-Approximation für die linke Seite von Gleichung (4.23) liefert mit Formel (4.11) die Beziehung:

$$z = \frac{s_n - n\mu(\theta)}{\sqrt{n\sigma^2(\theta)}},$$
$$f_{S_n}(s_n;\theta) = \frac{1}{\sqrt{n\sigma^2(\theta)}}\,\varphi(z)\left\{1+\mathcal{O}(n^{-1})\right\}. \tag{4.24}$$

Der Erwartungswert und die Varianz hängen dabei natürlich vom Parameter θ ab und es gilt außerdem:

$$\begin{aligned} \mathbb{E}(Y) &= \mu(\theta) \quad = b'(\theta), & \operatorname{var}(Y) &= \sigma^2(\theta) \quad = b''(\theta), \\ \mathbb{E}(S_n) &= n\mu(\theta) = nb'(\theta), & \operatorname{var}(S_n) &= n\sigma^2(\theta) = nb''(\theta). \end{aligned}$$

Um die Sattelpunkt-Approximation zu erhalten, wählen wir θ gerade so, dass die Gleichung (4.24) im Erwartungswert ausgewertet wird. Es soll also

$$s_n \stackrel{!}{=} n\mu(\hat{\theta}) = nb'(\hat{\theta})$$

gelten. Mit dieser speziellen Parameterwahl erhalten wir für die Dichte der Summe im Sattelpunkt ausgewertet die Gleichung:

$$f_{S_n}(s_n;\hat{\theta}) = \frac{1}{\sqrt{2\pi n b''(\hat{\theta})}}\left[1+\mathcal{O}(n^{-1})\right]. \tag{4.25}$$

Mit Gleichung (4.25) sind wir nun in der Lage eine Approximation für $\exp\{h(s_n)\}$ anzugeben. Die Gleichung (4.23) liefert für den Sattelpunkt $\hat{\theta}$:

$$f_{S_n}(s_n;\hat{\theta}) = \exp\left\{\hat{\theta}s_n - nb(\hat{\theta})\right\}\exp\{h(s_n)\}.$$

Somit gilt für $\exp\{h(s_n)\}$ die Beziehung:

$$\begin{aligned} \exp\{h(s_n)\} &= f_{S_n}(s_n;\hat{\theta})\exp\left\{nb(\hat{\theta}) - \hat{\theta}s_n\right\} \\ &= \frac{1}{\sqrt{2\pi n b''(\hat{\theta})}}\exp\left\{nb(\hat{\theta}) - \hat{\theta}s_n\right\}\left[1+\mathcal{O}(n^{-1})\right]. \end{aligned} \tag{4.26}$$

Setzen wir nun Gleichung (4.26) in die Gleichung (4.23) ein, erhalten wir schließlich die Sattelpunkt-Approximation für die Dichte $f_{S_n}(s_n;\theta)$:

$$f_{S_n}(s_n;\theta) = \frac{1}{\sqrt{2\pi n b''(\hat{\theta})}}\exp\left\{(\theta-\hat{\theta})s_n - n[b(\theta)-b(\hat{\theta})]\right\}\left[1+\mathcal{O}(n^{-1})\right]. \tag{4.27}$$

Bemerkung. Höhere Momente können in Gleichung (4.27) natürlich analog zur Herleitung der Sattelpunkt-Approximation in Abschnitt 4.1.2 berücksichtigt werden.

Die Wahl des Sattelpunkts $\hat{\theta}$ lässt sich auch anders motivieren. Betrachten wir die log-Likelihood-Funktion der Dichte $f_{S_n}(s_n;\theta)$ für eine einzelne Beobachtung s_n, erhalten wir:

$$\mathscr{L}(\theta, s_n) = \theta s_n - nb(\theta) + h(s_n).$$

Die Score-Gleichung für θ lautet somit:

$$\left.\frac{\partial \mathscr{L}(\theta, s_n)}{\partial \theta}\right|_{\theta=\hat{\theta}} = s_n - nb'(\hat{\theta}) \stackrel{!}{=} 0,$$

woraus wieder die Sattelpunktgleichung $s_n = nb'(\hat{\theta})$ folgt. Damit ist die Forderung, dass $\hat{\theta}$ der ML-Schätzer von θ ist, äquivalent mit der Sattelpunktgleichung (vgl. Friedl, 1991).

Die log-Likelihood-Funktion der Dichte $f_{S_n}(s_n;\theta)$ ausgewertet im Sattelpunkt lässt sich durch

$$\mathscr{L}(\hat{\theta}, s_n) = \hat{\theta} s_n - nb(\hat{\theta}) + h(s_n) = \sup_{\theta}\{\theta s_n - nb(\theta) + h(s_n)\}$$

angeben. Mit der Likelihood-Funktion $\ell(\theta) = \exp\{\mathscr{L}(\theta, s_n)\}$ und der beobachteten Fisher-Information

$$i(\theta) = -\frac{\partial^2 \mathscr{L}(\theta, s_n)}{\partial \theta^2} = nb''(\theta)$$

lässt sich die Sattelpunkt-Approximation in Gleichung (4.27) zu

$$f_{S_n}(s_n;\theta) = \frac{1}{\sqrt{2\pi}} (i(\hat{\theta}))^{-\frac{1}{2}} \frac{\ell(\theta)}{\ell(\hat{\theta})} \left[1 + \mathcal{O}(n^{-1})\right] \tag{4.28}$$

umschreiben. Gleichung (4.28) stellt also eine Beziehung zwischen der Sattelpunkt-Approximation und dem Likelihood-Quotienten her (vgl. Daniels, 1958).

4.2.2 Zusammenhang mit der Extended-Quasi-Likelihood-Funktion

Zwischen der EQL-Funktion und der Sattelpunkt-Approximation besteht ein enger Zusammenhang. Die EQL-Funktion aus Gleichung (3.13) lautet:

$$Q^+(\mu, \phi, y) = -\frac{1}{2} \log\{2\pi\phi V(y)\} - \frac{1}{2} D(y, \mu)/\phi,$$

dabei bezeichnet $D(y, \mu)/\phi$ die gewichtete Deviance, die gemäß Gleichung (2.25) durch

$$\frac{1}{\phi} D(y, \mu) = -2\left[\mathscr{L}(\mu, y) - \mathscr{L}(y, y)\right]$$

gegeben ist. Die folgenden Rechnungen beziehen sich auf eine Dichte aus der Exponentialfamilie der Form:

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

Betrachtet man nun die Exponentialfunktion der zur Dichte $f_Y(y;\theta,\phi)$ zugehörigen EQL-Funktion $\exp\{Q^+(\mu,\phi,y)\}$, erhält man:

$$\begin{aligned}\exp\left\{Q^+(\mu,\phi,y)\right\} &= \exp\left\{-\frac{1}{2}\log\left\{2\pi\phi V(y)\right\} - \frac{1}{2}D(y,\mu)/\phi\right\} \\ &= \frac{1}{\sqrt{2\pi\phi V(y)}}\exp\left\{-\frac{1}{2}D(y,\mu)/\phi\right\} \\ &= \frac{1}{\sqrt{2\pi\phi V(y)}}\exp\left\{\mathscr{L}(\mu,y) - \mathscr{L}(y,y)\right\}.\end{aligned}$$

Für den Erwartungswert μ gilt $\mathbb{E}(Y) = \mu = b'(\theta)$, wobei θ den kanonischen Parameter der Exponentialfamilie bezeichnet, während die Varianz durch $\operatorname{var}(Y) = \phi V(\mu) = \phi b''(\theta)$ gegeben ist.

Bemerkung. Um den Umstand zu verdeutlichen, dass der Erwartungswert vom kanonischen Parameter abhängt, schreiben wir anstatt $\mathscr{L}(\mu,y)$ von nun an $\mathscr{L}(\theta,y)$.

Setzen wir für den Erwartungswert μ die Beobachtung y selbst ein (berechnen wir also die Dichte im Sattelpunkt), ist das äquivalent zu der Forderung, dass wir die Dichte im ML-Schätzer $\hat\theta$ auswerten. Es folgt also mit $V(y) = b''(\hat\theta)$ für $\exp\{Q^+(\mu,\phi,y)\}$:

$$\exp\left\{Q^+(\mu,\phi,y)\right\} = \frac{1}{\sqrt{2\pi\phi\cdot b''(\hat\theta)}}\exp\big\{\mathscr{L}(\theta,y) - \mathscr{L}(\hat\theta,y)\big\}. \qquad (*)$$

Für den führenden Term der Sattelpunkt-Approximation für $n = 1$ (woraus $S_n = S_1 = Y$ folgt) gilt mit Gleichung (4.27) unter Berücksichtigung des Dispersionsparameters ϕ:

$$\begin{aligned}f_Y(y;\theta,\phi) &\approx \frac{1}{\sqrt{2\pi n\phi\cdot b''(\hat\theta)}}\exp\left\{\frac{(\theta-\hat\theta)y - n[b(\theta)-b(\hat\theta)]}{\phi}\right\} \\ &= \frac{1}{\sqrt{2\pi\phi\cdot b''(\hat\theta)}}\exp\big\{\mathscr{L}(\theta,y) - \mathscr{L}(\hat\theta,y)\big\}. \qquad (**)\end{aligned}$$

Vergleicht man nun $(*)$ und $(**)$ erkennt man Äquivalenz und somit entspricht die EQL-Funktion der nicht normalisierten Variante der Sattelpunkt-Approximation einer Verteilung aus der Exponentialfamilie (vgl. Nelder und Pregibon, 1987).

4.3 Beispiel

Wir betrachten die gewichtete Summe $\bar Y = \frac{1}{n}\sum_{i=1}^n Y_i$ von n unabhängig identisch verteilten Zufallsvariablen aus einer χ^2_k-Verteilung mit Freiheitsgrad k. Die Summe $S_n = Y_1 + \cdots + Y_n$ selbst ist eine χ^2_{nk}-verteilte Größe mit $n\cdot k$ Freiheitsgraden. Um die Dichte

der gewichteten Summe $f_{\bar{Y}}(\bar{y})$ zu bestimmen, verwenden wir den Transformationssatz für Dichten und erhalten somit die exakte Dichte:

$$\begin{aligned} g(\bar{y}) &= n\bar{y} = s_n, \\ f_{\bar{Y}}(\bar{y}) &= f_{S_n}(g(\bar{y})) \left| \frac{\partial g(\bar{y})}{\partial \bar{y}} \right| \\ &= n f_{S_n}(n\bar{y}). \end{aligned} \qquad \text{(Exakte Dichte)}$$

Für die Sattelpunkt-Approximation müssen wir die kumulantenerzeugende Funktion bzw. die Kumulanten bestimmen:

$$M_Y(t) = (1-2t)^{-k/2}, \qquad K_Y(t) = -\frac{k}{2}\log(1-2t), \qquad \kappa_m^*(\theta) = \frac{2^{m-1}k(m-1)!}{(1-2\theta)^m}.$$

Der Sattelpunkt $\hat{\theta}$ lässt sich über $K'(\hat{\theta}) = \bar{y}$ bestimmen und lautet somit:

$$K'(\hat{\theta}) \stackrel{!}{=} \bar{y} \quad \Rightarrow \quad \hat{\theta} = \frac{\bar{y}-k}{2\bar{y}} = \frac{1}{2} - \frac{k}{2\bar{y}}. \qquad \text{(Sattelpunkt)}$$

Für die standardisierten Kumulanten folgt aus $\rho_m^*(\theta) = \kappa_m^*(\theta)/(\kappa_2^*(\theta))^{m/2}$ die Beziehung:

$$\begin{aligned} \rho_m^*(\theta) &= \frac{2^{m-1}k(m-1)!}{(1-2\theta)^m} \Bigg/ \left(\frac{2k}{(1-2\theta)^2}\right)^{m/2} \\ &= \frac{|1-2\theta|^m\, 2^{m-1}k(m-1)!}{(2k)^{m/2}(1-2\theta)^m} \\ &= [\operatorname{sign}(1-2\theta)]^m\, (k/2)^{1-m/2}(m-1)!. \end{aligned} \qquad \text{(Standardisierte Kumulante)}$$

Bemerkung. Die standardisierten Kumulanten werden in der Sattelpunkt-Approximation im Sattelpunkt ausgewertet. Für diesen folgt:

$$\begin{aligned} \hat{\theta} = \frac{1}{2} - \frac{k}{2\bar{y}} < \frac{1}{2} \Rightarrow \\ \rho_m^*(\hat{\theta}) = \underbrace{[\operatorname{sign}(1 - \underbrace{2\hat{\theta}}_{<1})]^m}_{=1}\, 2^{m/2-1}k^{1-m/2}(m-1)! \\ = 2^{m/2-1}k^{1-m/2}(m-1)!. \end{aligned}$$

Das heißt, dass die standardisierten Kumulanten für alle $\bar{y}$ konstant sind. Die renormalisierte Sattelpunkt-Approximation ist also exakt, was daher rührt, dass die χ^2-Verteilung ein Spezialfall der Gammaverteilung ist.

Die Abbildung 4.1 zeigt für verschiedene Werte von n jeweils die Edgeworth- und die Sattelpunkt-Approximation der Dichte der gemittelten Summe von χ_2^2-verteilten Größen. Zusätzlich zeigt die Abbildung die exakte Dichte sowie eine Approximation durch die Normalverteilung.

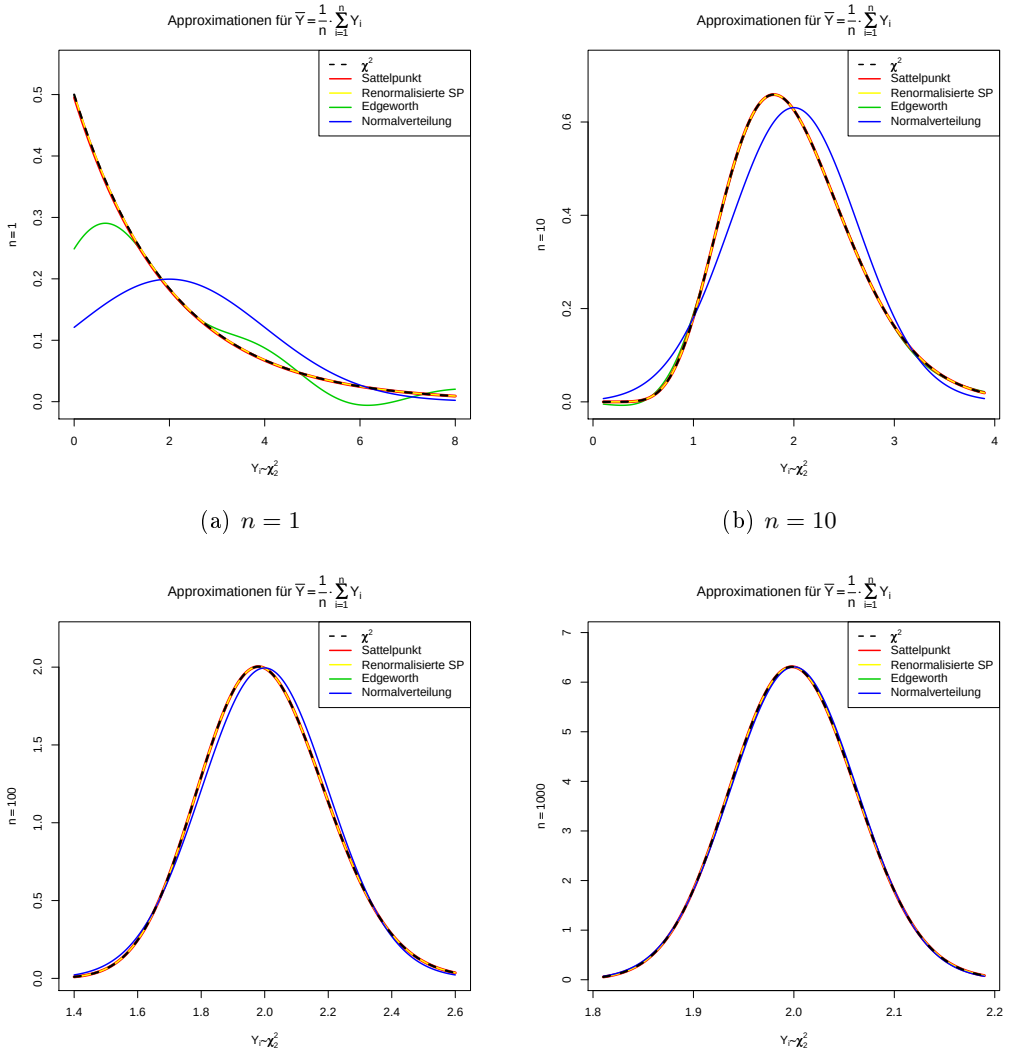
Die Sattelpunkt-Approximation (rote Kurve bzw. gelbe Kurve für die renormalisierte[2] Version) ist bereits für $n = 1$ in Abbildung 4.1(a) kaum von der originalen Dichte zu unterscheiden. Man erkennt aber die Schwächen der Edgeworth-Approximation (grüne Kurve), die im Zentrum um den Erwartungswert eine gute Annäherung liefert, während sie in den Schwänzen zum Teil stark von der originalen Dichte abweicht. Für $n = 10$ erkennt man in Abbildung 4.1(b) kaum noch einen Unterschied zwischen der Edgeworth- und der Sattelpunkt-Approximation.

Die Approximation durch die Normalverteilung (blaue Kurve), wie sie durch den zentralen Grenzwertsatz motiviert wird, ist naturgemäß für kleines n schlecht. Je größer n wird desto mehr nähert sich auch die Normalverteilung der exakten Dichte an, die für wachsendes n selbst einer Normalverteilungsdichte immer ähnlicher wird. Für $n = 1000$ in Abbildung 4.1(d) ist kaum noch ein Unterschied zwischen den Approximationen auszumachen.

Für die Edgeworth-Approximation bedeutet wachsendes n, dass die höheren Momente stark an Einfluss verlieren und der dominierende Term (also die Normalverteilungsdichte) die Form der Kurve bestimmt.

Zusammenfassend kann gesagt werden, dass die Sattelpunkt-Approximation bereits für kleines n im ganzen Wertebereich von y eine sehr gute Anpassung liefert, während die Edgeworth-Approximation die exakte Dichte nur im Zentrum annähert. Mit wachsendem n dominiert die Normalverteilungsdichte und man kann keinen Unterschied mehr zwischen den Approximationen erkennen.

[2] Der Normalisierungsfaktor c wurde dabei numerisch bestimmt.



(a) $n = 1$

(b) $n = 10$

(c) $n = 100$

(d) $n = 1000$

Abbildung 4.1: Sattelpunkt- und Edgeworth Approximationen für die Dichte der gemittelten Summe von χ^2_2-verteilten Größen

5 Die Implementierung der Extended-Quasi-Likelihood-Funktion

5.1 Der Algorithmus

In diesem Kapitel erläutern wir die Implementierung der EQL-Funktion zur Schätzung des Parametervektors $\boldsymbol{\theta}$ einer Varianzfamilie $\mathcal{F}_{\boldsymbol{\theta}}$ für das Statistikprogramm R. Dazu wurde die R-Bibliothek *EQL* entwickelt, deren Leistungsumfang auch Funktionen zur Sattelpunkt- bzw. Edgeworth-Approximation umfasst. Während sich eine vollständige Dokumentation aller entwickelten Funktionen im Anhang B befindet, wird in den anschließenden Abschnitten im speziellen die Arbeitsweise der Funkion *eql*, deren Signatur man in Listing 5.1 findet, näher beleuchtet. Außerdem wird die Möglichkeit vorgestellt, mittels Diagnoseplots Konfidenzbereiche für den Parametervektor $\boldsymbol{\theta}$ explorativ zu ermitteln.

Im Folgenden verwenden wir für *Klassen* eine kursive Schrift, für *Bibliotheken* eine kursive und für *Funktionen* eine geneigte Schreibmaschinenschrift und für Parameter eine serifenlose Schrift.

5.1.1 Modell und Varianzfamilie

Die wichtigsten Schnittstellen für die EQL-Funktion sind einerseits die Modellformel formula und andererseits die Varianzfamilie $\mathcal{F}_{\boldsymbol{\theta}}$, die über den Parameter family an *eql* übergeben wird. Die Syntax für formula gleicht der bei der Funktion *glm* zum Einsatz kommenden.

Um eine Varianzfamilie zu definieren, kann man entweder auf die beiden bereits vordefinierten Funktionen *powerVarianceFamily* oder *extBinomialVarianceFamily* zurückgreifen, oder aber die Funktion *varianceFamily* nutzen, deren Signaturen in Listing 5.2 dargestellt sind.

Listing 5.1: Signatur *eql*

```
eql <- function(formula, param.space, family=powerVarianceFamily(),
                phi.method=c("pearson", "mean.dev"),
                include.model=TRUE, smooth.grid=10,
                do.smooth=dim(family)==1, verbose=1, ...)
```

Während die ersten beiden Funktionen den im Abschnitt 3.3.2 vorgestellten Potenz-Ansatz $\left(V_\theta(\mu) = \mu^\theta\right)$ bzw. den binomialähnlichen Ansatz $\left(V_{k,l}(\mu) = \mu^k(1-\mu)^l\right)$ beschreiben, erlaubt *varianceFamily* die Definition einer beliebigen Varianzfamilie.

Alle drei Funktionen geben ein Objekt der Klasse[1] *varianceFamily* zurück, das die Funktion *family* beinhaltet. Ruft man *family* mit einer speziellen Parameterwahl $\boldsymbol{\theta}_0$ auf, erhält man ein *extFamily*-Objekt, das eine Erweiterung der Klasse *family*, wie man sie von der Verwendung mit dem Befehl *glm* aus R kennt, darstellt.

Beispiel 5.1. Der Potenz-Ansatz $V_\theta(\mu) = \mu^\theta$ beinhaltet u. a. die Normalverteilung ($\theta = 0$) und die Poissonverteilung ($\theta = 1$). Der folgende R-Code generiert die Varianzfamilie für den Potenz-Ansatz und wertet die Varianzfunktion für $\theta_0 = 0$ und $\theta_0 = 1$ aus.

```
> pf <- powerVarianceFamily()
> pf$family(0)

Family: Power-Family
Link function: log

Family-Parameters:
theta = 0
> pf$family(1)

Family: Power-Family
Link function: log

Family-Parameters:
theta = 1

> pf$family(0)$variance(1:5)
[1] 1 1 1 1 1
> gaussian()$variance(1:5)
[1] 1 1 1 1 1

> pf$family(1)$variance(1:5)
[1] 1 2 3 4 5
> poisson()$variance(1:5)
[1] 1 2 3 4 5
```

◊

Um ein *varianceFamily* Objekt zu erstellen, bedarf es im Prinzip nur der Kenntnis der zugrunde liegenden Varianzfunktion $V_{\boldsymbol{\theta}}(\mu)$, da die Deviance durch das Integral in Gleichung (3.9) vollständig determiniert wird. Über den Parameter varf (der eine Funktion

[1] Die Implementierung der Bibliotheken erfolgte dabei unter Verwendung der sogenannten „S3-Klassenhierarchie“.

Listing 5.2: Signaturen der Varianzfamilien erzeugenden Funktionen

```
varianceFamily <- function(varf, devf=NULL, link="log", initf=NULL,
                           validmuf=NULL, name="default")
powerVarianceFamily <- function(link="log")
extBinomialVarianceFamily <- function(link="logit")
```

Listing 5.3: Beispiel für die Varianz- und Deviancefunktion einer Varianzfamilie

```
varf <- function(y, theta) y^theta

devf <- function(y, mu, theta) {
  if(theta == 1) {
    return(2 * (y * log(ifelse(y == 0, 1, y / mu)) - (y - mu)))
  } else if(theta == 2) {
    return(2 * (y / mu - log(ifelse(y == 0, 1, y / mu)) - 1))
  } else {
    return(1/ ((1 - theta) * (2 - theta)) * 2 *
           (y ^ (2 - theta) - (2 - theta) * y * mu ^ (1 - theta) +
           (1 - theta) * mu ^ (2 - theta)))
  }
}
```

beschreibt) bestimmt man die allgemeine Form der Varianzfamilie. Bei der Definition der Varianzfamilie ist darauf zu achten, dass das erste Argument der Funktion `varf` stets dem Argument entspricht, an dem die Funktion später für eine bestimmte Familienparameterwahl $\boldsymbol{\theta}_0$ ausgewertet wird. Die Familienparameter selbst sind im Anschluss daran anzuführen.

Kann man die Deviance-Funktion explizit angeben, empfiehlt es sich, diese über den Parameter `devf` an die Funktion *`varianceFamily`* zu übergeben. Dies hat erstens den Vorteil, dass die Berechnungen genauer werden, weil das Integral nicht numerisch gelöst werden muss, und beschleunigt zweitens später die Berechnung der EQL-Werte. Die ersten beiden Argumente der Deviance müssen y bzw. μ entsprechen und die nachfolgenden müssen – für den Fall, dass $\boldsymbol{\theta}$ nicht nur aus einem Skalar besteht – in der gleichen Reihenfolge wie für `varf` angegeben werden. Ein vollständiges Beispiel für Varianzfunktion und Deviance findet man in Listing 5.3. Es beschreibt die Varianzfamilie, wie sie durch den Potenz-Ansatz induziert wird.

Bemerkung. Die Funktion *`extBinomialVarianceFamily`* für den binomialähnlichen Ansatz berechnet die Deviance numerisch. Dadurch dauern Berechnungen für den binomialähnlichen Ansatz bedeutend länger als für den Potenz-Ansatz, für den eine analytische

Lösung für die Deviance vorliegt. Außerdem sei an dieser Stelle darauf hingewiesen, dass die vordefinierten Varianzfamilien definitionsgemäß für $y = 0 : V(y) = 0$ liefern. Da die EQL-Funktion für diesen Fall nicht definiert ist, sind `extBinomialVarianceFamily` und `powerVarianceFamily` im speziellen *nicht* in der Lage mit exakten Null-Werten zu arbeiten.

Die Parameter initf und validmuf dienen der Initialisierung der Varianzfamilie bzw. der Überprüfung gültiger Werte für μ. Die Validierungsfunktion validmuf spielt dabei die selbe Rolle wie die Funktion `validmu` für *family*-Objekte. Die Initialisierungsfunktion initf ist `initialize` nachempfunden, das für das gewöhnliche GLM die Aufgabe innehat, die Startwerte für die jeweilige Verteilungsannahme zu setzen. Sowohl validmuf als auch initf können die Familienparameter als zusätzliche Parameter akzeptieren, um gegebenenfalls Initialisierungen bzw. Validitätsüberprüfungen in Abhängigkeit der speziellen Parameterwahl $\boldsymbol{\theta}_0$ durchführen zu können.[2] Ein Beispiel für diese beiden Funktionen ist in Listing 5.4 zu finden und beschreibt die Initialisierung bzw. die Validitätsüberprüfung für den Potenz-Ansatz.

Ein Objekt der Klasse *varianceFamily* bildet die gesamte *Varianzfamilie* ab. Will man für ein spezielles $\boldsymbol{\theta}_0$ die zugehörige *Varianzfunktion* bestimmen, verwendet man – wie bereits in Beispiel 5.1 erläutert – die Funktion `family`.

Bemerkung. Der Name „family“ ist in diesem Kontext mehrdeutig. Die Funktion `glm` verwendet ein Objekt der Klasse *family*, um für die gewählte Verteilungsannahme u. a. die Varianzfunktion und die Deviance zu bestimmen. Wenn wir aber von der Varianzfamilie sprechen, meinen wir eine Menge von Varianzfunktionen, die alle die gleiche Struktur aufweisen. Für eine bestimmte Parameterwahl $\boldsymbol{\theta}_0$ erhalten wir aus der Varianzfamilie eine spezielle Varianzfunktion, für die wir – damit wir sie für `glm` verwenden können – ein Objekt der Klasse *family* konstruieren, indem wir in dem Objekt neben der Varianzfunktion $V(\mu)$ selbst, auch alle diejenigen Funktionen (wie z. B. die Deviance) kapseln, die für die Berechnung mittels `glm` notwendig sind.[3]

Die restlichen Parameter link und name beschreiben schließlich die gewählte Linkfunktion bzw. den Namen der Varianzfamilie.

5.1.2 Gitter Suche

Die Idee der Funktion `eql` ist es, für eine gegebene Menge $\Theta = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m\}$ von Familienparametern jeweils den Wert der EQL-Funktion zu bestimmen und von dieser Menge

[2] So ist beispielsweise beim Potenz-Ansatz $V(\mu, k) = \mu^k$ für $k = 0$ jedes μ zulässig, während für $k > 0 : \mu > 0$ gelten muss.

[3] In den Hilfeseiten von R findet man mit dem Befehl `?family` detaillierte Informationen über die Struktur eines *family* Objekts.

Listing 5.4: Beispiel für eine Initialisierungs- und Validierungsfunktion

```
init <- function(theta,...) {
  if (theta == 0) {
    return(expression({
      mustart <- y
      n <- rep.int(1, nobs)}))
  } else if (theta == 1) {
    return(expression({
      if (any(y < 0)) {
        stop("'y' must not be negative for theta=1!")
      }
      mustart <- y + 0.1
      n <- rep.int(1, nobs)}))
  } else {
    return(expression({
      if (any(y <= 0)) {
        stop("'y' must be postive!")
      }
      mustart <- y
      n <- rep.int(1, nobs)}))
  }
}

validmuf <- function(mu, theta) {
  if (theta == 0) {
    return(TRUE)
  } else if (theta == 1 | theta == 2 | theta == 3) {
    return(all(mu > 0))
  } else {
    return(all(mu > 0))
  }
}
```

Listing 5.5: Beispiele für ein Suchgitter

```
ps.power <- list(seq(1, 4, length = 20))

ps.binomial <- list(seq(1, 2.2, length=20), seq(1, 3, length=20))
```

dann den Maximalwert zu ermitteln:

$$Q^+(\boldsymbol{\theta}) := \sum_{i=1}^{n} \left[-\frac{1}{2} \log \left\{ 2\pi \hat{\phi}(\boldsymbol{\theta}) V_{\boldsymbol{\theta}}(y_i) \right\} - \frac{1}{2} D_{\boldsymbol{\theta}}(y_i, \hat{\mu}_i(\boldsymbol{\theta}))/\hat{\phi}(\boldsymbol{\theta}) \right],$$

$$\boldsymbol{\theta}_{\max} := \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} Q^+(\boldsymbol{\theta}),$$

$$Q^+_{\max} := Q^+(\boldsymbol{\theta}_{\max}),$$

dabei hängen die Schätzer $\hat{\boldsymbol{\mu}}(\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}_0))$ und $\hat{\phi}(\boldsymbol{\theta}_0)$ natürlich vom jeweiligen Parameter $\boldsymbol{\theta}_0$ ab. Dieser Ansatz folgt der Idee von Nelder und Pregibon (1987) und man erhält so eine „Profile-QL-Menge“,[4] indem man diejenigen Punkte $\boldsymbol{\theta}$ bestimmt, für die die Differenz $Q^+_{\max} - Q^+(\boldsymbol{\theta})$ kleiner als ein passend gewählter Wert d ist.

Diese Vorgehensweise impliziert aber auch, dass man bei einem schlecht gewähltem Suchgitter Gefahr läuft, anstatt eines globalen Maximums nur ein lokales zu finden. Das heißt, es liegt in besonderem Maß in der Verantwortung des Benutzers, das Suchgitter „passend“ zu wählen. Liegt das bestimmte Maximum etwa am Rand des gewählten Gitters, kann das ein Indiz dafür sein, ein größeres Gitter in Betracht zu ziehen.

Der Vorteil dieser Methode liegt einerseits in der einfachen Implementierung und andererseits in der Möglichkeit bereits vorhandene Methoden zur Bestimmung des Schätzers für den Parametervektor $\boldsymbol{\beta}$ heranziehen zu können, da sich das Problem für fixiertes $\boldsymbol{\theta}_0$ auf ein gewöhnliches GLM mit QL-Ansatz reduziert.

Bezeichne k die Anzahl der Parameter der Varianzfamilie[5] und $1 \leq i \leq k : \theta_i$ den i-ten Parameter der Varianzfamilie. Um das Suchgitter für `eql` zu definieren, übergibt man über den Parameter `param.space` eine Liste von Mengen. Dabei muss diese Liste für jeden Parameter θ_i der Varianzfamilie eine Menge L_i enthalten, die die interessierenden Werte für den i-ten Parameter enthält. Die Funktion `eql` bildet aus den Mengen L_i das kartesische Produkt und wir erhalten so das Suchgitter

$$\Theta = L_1 \times \cdots \times L_k,$$

über welches die Maximierung erfolgt. Das Listing 5.5 zeigt ein Beispiel für die Definition des Suchgitters für den Potenz- und den binomialähnlichen Ansatz.

[4] Im eindimensionalen Fall erhält man dementsprechend ein Intervall.
[5] Es gilt also $\boldsymbol{\theta} \in \mathbb{R}^k$.

Bemerkung. Die „Feinmaschigkeit“ des Gitters beeinflusst neben der Genauigkeit einer Lösung natürlich auch die Laufzeit des Algorithmus'. Zu beachten ist außerdem, dass sich die Anzahl der Gitterpunkte gemäß der Vorschrift

$$|\Theta| = |L_1| \cdot |L_2| \cdot \cdots \cdot |L_k|$$

ergibt. Für das Beispiel aus Listing 5.5 haben wir somit $|\Theta_{\mathrm{P}}| = 20$ bzw. $|\Theta_{\mathrm{B}}| = 20 \cdot 20 = 400$, wobei wir das Suchgitter für den Potenz- und den binomialähnlichen Ansatz mit Θ_{P} bzw. Θ_{B} bezeichnen. Das heißt, dass die EQL-Funktion für den Potenz-Ansatz neben dem Geschwindigkeitsgewinn durch die Kenntnis der exakten Deviance Funktion im Regelfall auch an weniger Stellen berechnet werden muss.

5.1.3 Sonstige Steuerparameter und Rückgabewert

Über den Parameter phi.method kann man `eql` mitteilen, welche Methode zur Schätzung des Dispersionsparameters ϕ verwendet werden soll. Zur Auswahl stehen die Schätzung über die Pearson-Statistik und die Schätzung über die mittlere Deviance.

Der Parameter include.model gibt an, ob das GLM für die Varianzfunktion mit dem gefundenen Parametervektor $\boldsymbol{\theta}_{\max}$ im Rückgabewert inkludiert werden soll.

Die Menge an Information die während des Bearbeitungsprozesses am Bildschirm ausgegeben werden soll, wird über den Parameter verbose gesteuert. Da die Prozedur für große Parametermengen bzw. nicht explizit vorliegende Deviance Funktionen längere Zeit in Anspruch nehmen kann, zeigt die Funktion `eql` in der Standardeinstellung einen Fortschrittsbalken an, der den Anteil der bereits abgearbeiteten Parameter darstellt. Man kann sowohl detailliertere Informationen ausgeben lassen, als auch die Fortschrittsanzeige vollkommen ausblenden.

Für eindimensionale Varianzfamilien ($\boldsymbol{\theta} \in \mathbb{R}$) besteht die Möglichkeit, zwischen den explizit berechneten EQL-Werten zu interpolieren. Der Parameter do.smooth gibt an, ob eine Interpolation vorgenommen werden soll. Die Anzahl der Interpolationspunkte zwischen zwei berechneten EQL-Werten wird `eql` über den Parameter smooth.grid mitgeteilt.

Bemerkung. Für höherdimensionale Varianzfamilien ($\boldsymbol{\theta} \in \mathbb{R}^k, k \geq 2$) ist eine Interpolation *nicht* möglich.

Zusätzliche Parameter die über „...“ übergeben werden, werden an die `glm` Funktion weitergereicht. So ist es möglich, den vollen Funktionsumfang von `glm` nutzen zu können.

Die Funktion `eql` gibt ein Objekt der Klasse *eql* zurück, dessen Komponenten in Tabelle 5.1 zusammengefasst sind. Für den Fall, dass für eine bestimmte Parameterwahl $\boldsymbol{\theta}_0$ das GLM nicht berechnet werden kann, wird eine Warnung ausgegeben und der zugehörige EQL-Wert wird mit `NA` belegt, das für „nicht verfügbar“ (engl. *not available*) steht.

Tabelle 5.1: Komponenten eines *eql* Objekts

Komponente	Bedeutung
eql	die EQL-Werte
param	die Parameter-Werte an denen die EQL-Funktion ausgewertet wurde
eql.max	der maximale EQL-Wert im Suchgitter
param.max	der Parameter-Wert an dem das Maximum erreicht wurde
dim	die Dimension der Varianzfamilie
smooth	eine Boolesche Variable, die angibt ob eine Interpolation durchgeführt wurde
is.smoothed	ein Vektor von Booleschen Werten, die angeben ob der korrespondierende EQL-Wert interpoliert wurde
smooth.grid	die Anzahl der Interpolationspunkte zwischen zwei berechneten EQL-Werten

Listing 5.6: Signatur der EQL-Plot-Funktion

```
plot.eql <- function(x, do.points=(dim(x) == 1 &&
                                   sum(!x$is.smoothed) <= 20),
                     do.ci=TRUE, alpha=0.95, do.bw=TRUE,
                     show.max=TRUE, ...)
```

5.2 Der EQL-Plot

Um die Werte die man mittels der Funktion `eql` erhält, weiter zu untersuchen, stellt die Bibliothek `EQL` auch einen `plot` Befehl[6] zur Verfügung, dessen Signatur man in Listing 5.6 findet. In Abhängigkeit von der Dimension von $\boldsymbol{\theta}$ wird entweder ein Profile-Plot erzeugt, der die Werte des Parameters gegen die jeweiligen EQL-Werte aufträgt, oder ein Kontur-Plot generiert, der für jede Parameter Kombination die Größenordnung des dazugehörigen EQL-Wertes mittels abgestufter Farben kodiert. Die Bedeutung der Parameter findet sich in Tabelle 5.2.

5.2.1 Profile-Plot

Besteht der Vektor $\boldsymbol{\theta} \in \mathbb{R}$ nur aus einem Skalar, liegt es nahe, die EQL-Werte in Abhängigkeit verschiedener Werte dieses Skalars zu betrachten. Die Funktion `plot.eql` trägt dazu die Werte des Suchgitters gegen die jeweiligen EQL-Werte auf. Um eine möglichst glatte Kurve zu erhalten, muss man entweder ein engmaschiges Suchgitter wählen, oder man greift auf interpolierte Werte zwischen den Knoten des Gitters zurück.

[6] Der Idee von S3-Klassen folgend, ist `plot` für *eql* Objekte in der Funktion `plot.eql` implementiert.

Tabelle 5.2: Parameter von `plot.eql`

Parameter	Bedeutung
x	ein Objekt der Klasse *eql*
do.points	ein Boolescher Wert, der bestimmt, ob die berechneten EQL-Werte im Plot extra markiert werden sollen
do.ci	ein Boolescher Wert, der bestimmt, ob ein Konfidenzbereich eingezeichnet werden soll
alpha	das Signifikanzniveau für den Konfidenzbereich
show.max	ein Boolescher Wert, der bestimmt, ob der maximale EQL-Wert im Plot speziell ausgezeichnet werden soll
...	zusätzliche Grafikparameter die an die jeweiligen Grafikfunktionen weitergeleitet werden

Das $1-\alpha$-Konfidenzintervall enthält all jene Werte des Suchgitters, deren zugehörige EQL-Werte größer als $Q^+_{\max} - \frac{1}{2}\chi^2_{1;1-\alpha}$ sind, wobei $\chi^2_{1;1-\alpha}$ das $(1-\alpha)$-Quantil einer χ^2_1-Verteilung bezeichnet. Für ein Suchgitter[7] $\Theta = (\theta_1, \ldots, \theta_m)$ mit $\theta_i < \theta_j$ für $i < j$ werden die Intervallgrenzen θ_l und θ_r dann folgendermaßen ermittelt:

1. Sei $z := Q^+_{\max} - \frac{1}{2}\chi^2_{1;1-\alpha}$. Bestimme die Indizes l_1 und r_1 so, dass
$$l_1 = \min\left\{i \,|\, Q^+(\theta_i) \geq z\right\}$$
$$r_1 = \max\left\{i \,|\, Q^+(\theta_i) \geq z\right\}$$
gilt. Somit ist l_1 der kleinste Index, so dass der EQL-Wert von θ_{l_1} gerade noch größer oder gleich z ist.

2. Setze $l_2 = l_1 - 1$ und $r_2 = r_1 + 1$.

3. Damit lassen sich nun die Intervallgrenzen θ_l und θ_r bestimmen:
$$\theta_l = \theta_{l_2} + (Q^+(\theta_{l_2}) - z)\frac{\theta_{l_1} - \theta_{l_2}}{Q^+(\theta_{l_2}) - Q^+(\theta_{l_1})},$$
$$\theta_r = \theta_{r_1} + (Q^+(\theta_{r_1}) - z)\frac{\theta_{r_2} - \theta_{r_1}}{Q^+(\theta_{r_1}) - Q^+(\theta_{r_2})}.$$

Bemerkung. Für das zur Berechnung der Intervallgrenzen eingesetzte Interpolationsverfahren setzen wir voraus, dass $\forall\theta \in [\theta_{l_1}, \theta_{r_1}] : Q^+(\theta) \geq z$ gilt.

5.2.2 Kontur-Plot

Hat die Varianzfamilie zwei Parameter ($\boldsymbol{\theta} \in \mathbb{R}^2$), bedarf es einer dreidimensionalen Grafik, um die Abhängigkeit des EQL-Wertes von den Parametern darzustellen. Ein Konturplot kodiert eine Variable mittels Farben und trägt sie gegen die beiden anderen auf.

[7] Das Suchgitter Θ enthält dabei auch jene Werte, für die der EQL-Wert nicht berechnet sondern interpoliert wurde.

Für zweidimensionale Varianzfamilien greift die Funktion `plot.eql` auf den Konturbzw. den Levelplot aus der Bibliothek `lattice` (Sarkar, 2008) zurück, welche eine alternative Schnittstelle für Grafiken in R anbietet (für weiterführende Informationen über das Paket vgl. Murrell, 2005).

Im Gegensatz zum Profile-Plot besteht die Konfidenz-Menge nicht aus einem Intervall, sondern aus einem einer Ellipse ähnlichen Bereich. Um diesen zu erhalten, wird die Funktion `contourplot` verwendet, die eine Konturlinie für den EQL-Wert $Q^+_{\max} - \frac{1}{2}\chi^2_{2;1-\alpha}$ einzeichnet. Der Bereich, den diese Linie einschließt, enthält alle Parameter-Kombinationen für die der EQL-Wert größer als $Q^+_{\max} - \frac{1}{2}\chi^2_{2;1-\alpha}$ ist.

5.3 Beispiele

In diesem Abschnitt wollen wir die Funktionalität der Bibliothek `EQL` anhand einiger Beispiele vorstellen.

5.3.1 Potenzfamilie

Anschauungsbeispiel

Zuerst wollen wir in einem konstruierten Beispiel die Plausibilität der Funktion `eql` überprüfen. Dazu generieren wir zuerst je zwei Gruppen zu 100 poissonverteilten Zufallsvariablen $Y_{ij} \sim \mathcal{P}(\mu_{ij}), 1 \leq i \leq 100, j = 1, 2$ mit den Erwartungswerten $\mu_{i1} = 10$ und $\mu_{i2} = 100$. Die Poissonverteilung ist durch die Varianzfunktion $V(\mu) = \mu$ charakterisiert und ist somit ein Mitglied der Potenzfamilie $\mathcal{F}_\theta = \{V_\theta(\mu) = \mu^\theta\}$. Wir wollen das Modell

$$1 \leq i \leq 100 : x_{ij} = \begin{cases} 1 & j = 1, \\ 2 & j = 2 \end{cases},$$
$$\log \mu_{ij} = \beta \cdot x_{ij} \quad \text{mit} \quad V_\theta(\mu) = \mu^\theta$$

an die Daten anpassen, wobei wir den EQL-Ansatz verwenden um den Parameter θ zu schätzen. Da die Daten aus einer Poissonverteilung kommen, erwarten wir ein Ergebnis für $\hat{\theta}$ in der Nähe von Eins. Der tatsächliche Wert des Parameters β liegt bei $\beta = \log 10 = 2.3026$, da gilt:

$$\begin{aligned} \log 10 = \beta \quad &\Rightarrow \quad \beta = \qquad\qquad\qquad = 2.3026, \\ \log 100 = 2\beta \quad &\Rightarrow \quad \beta = \frac{1}{2}\log 100 = \log 10 = 2.3026. \end{aligned}$$

Als Suchgitter wählen wir ein äquidistantes Gitter im Intervall $[0, 2]$. Der folgende R-Code fasst die Berechnungen zusammen:

```
> library(EQL)
Lade nötiges Paket: ttutils

> set.seed(1234)
```

```
> y <- c(rpois(100, 10), rpois(100, 100))
> x <- rep(1:2, each=100)

> ps <- list(seq(0, 2, length=20))

> eq <- eql(y ~ x - 1, ps, powerVarianceFamily(), include.model=TRUE)
~~ Extended Quasi Likelihood ~~
* Dimension of parameter space:  20
* Starting loop:

0%       50%        100%
+---------+---------+
||||||||||||||||||||

* End loop
* Start smoothing:
Smoothing on an equidistant grid between EQL values with 10 nodes
* End smoothing

> eq
EQL-Max: -633.4521 at: theta = 0.959
EQL Summary:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -698.2  -669.7  -650.5  -655.0  -638.0  -633.5

> eq$mod$coef
       x
2.305751
```

Der maximale EQL-Wert wird im Gitter an der Stelle $\hat{\theta} = 0.959$ angenommen. Dieser Wert liegt nahe beim tatsächlichen Wert $\theta = 1$ und ein Blick auf den Profile-Plot in Abbildung 5.1 zeigt, dass $\theta = 1$ im 95%-Konfidenzintervall liegt. Der Schätzer für β schließlich liegt bei $\hat{\beta} = 2.3058$ und ist somit auch in der Nähe des tatsächlichen Parameterwertes.

Textildaten

Den in Abschnitt 3.3.2 vorgestellten Datensatz über die Strapazierfähigkeit von Textilfasern von Box und Cox (1964) verwenden Nelder und Pregibon (1987) um mit Hilfe der EQL-Funktion den Parameter der Potenzfamilie zu bestimmen. Wir wollen diesen Datensatz ebenfalls analysieren und die erhaltenen Schätzer mit Schätzern vergleichen, die der Tweedie-Ansatz und eine direkte numerische Optimierung der EQL-Funktion liefert.

Bemerkung 1. Der Unterschied der direkten Optimierung zur Funktionalität von `eql` besteht darin, dass in der direkten Optimierung versucht wird, die linearen und die nicht

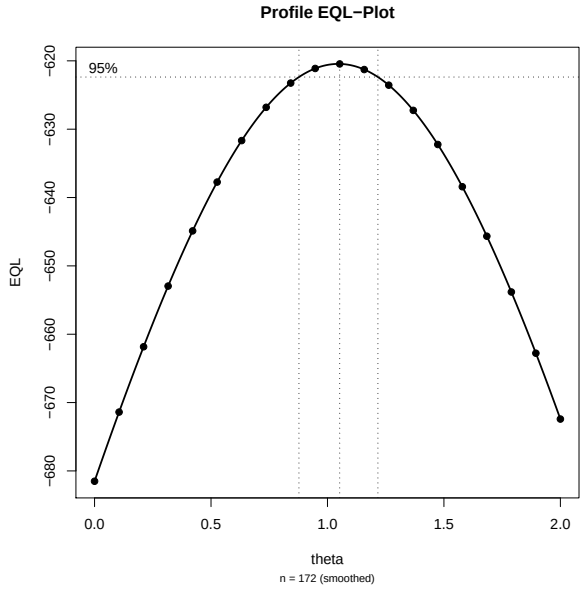


Abbildung 5.1: Profile-Plot für den Potenz-Ansatz für einen konstruierten Datensatz

linearen Parameter des Modells unter Zuhilfenahme eines numerischen Optimierungsverfahrens *direkt* über die Maximierung der EQL-Funktion zu bestimmen, während `eql` die linearen Parameter (also den Parametervektor $\boldsymbol{\beta}$) mit den Methoden des Generalisierten Linearen Modells bestimmt und den nicht linearen Parameter (also den Parameter θ der Varianzfamilie) über die Gittersuche ermittelt.

Bemerkung 2. Im Allgemeinen verlangt die Verwendung eines numerischen Optimierungsverfahrens eine Spezifizierung eines Startvektors für die Parameter, über die optimiert werden soll. Daher erweist sich diese Methode zur Bestimmung der Parameter als wenig praktikabel, da sie eine ungefähre Kenntnis der Größerordnung aller im Modell vorkommenden Parameter voraussetzt.

Wie bereits in Abschnitt 3.3.2 betrachten wir das Modell

$$\log \mu = \alpha + x_1\beta_1 + x_2\beta_2 + x_3\beta_3, \quad V_\theta(\mu) = \mu^\theta,$$

wobei wir für den Tweedie-Ansatz noch zusätzlich $Y \sim$ Tweedie-Verteilung annehmen.

Das Suchgitter für die Funktionen `tweedie.profile` und `eql` besteht in beiden Fällen aus einer äquidistanten Folge im Intervall $[1.1, 4]$. Die Startwerte für die direkte Optimierung setzen wir mit $(\alpha^0, \beta_1^0, \beta_2^0, \beta_3^0, \theta^0) = (4, 0, 0, 0, 1)$ fest. Die resultierenden Schätzer sind in Tabelle 5.3 zusammengefasst. Man erkennt, dass alle Ansätze in etwa die gleichen Ergebnisse liefern, wobei der Ansatz über die Gittersuche den höchsten EQL-Wert liefern kann. Allerdings unterscheiden sich die Schätzer dieser Methode von den Schätzern der anderen Methoden nur kaum. Die Schätzer für die linearen Parameter beim Tweedie- und beim EQL-Ansatz sind nahezu identisch, wenngleich die EQL-Funktion einen leicht grö-

Tabelle 5.3: Parameterschätzer für den Textildatensatz

Methode	$\hat{\alpha}$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\theta}$	EQL-Wert
EQL	6.3477	0.8408	−0.6288	−0.3707	2.4906	−160.5579
Tweedie	6.3478	0.8407	−0.6286	−0.3718	2.4612	−160.5612
Direkte Optimierung	6.3487	0.8433	−0.6267	−0.3727	2.5360	−160.5726

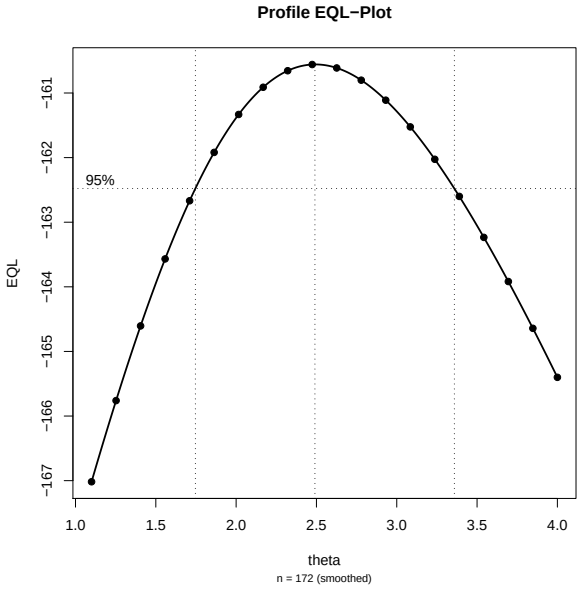


Abbildung 5.2: Profile-Plot für den Potenz-Ansatz für den Textildatensatz

ßeren Wert für θ schätzt, der allerdings selbst wiederum kleiner als der von der direkten Optimierung bestimmte Wert ist.

Vergleicht man die log-Likelihood-Funktion der Tweedie-Verteilung in Abbildung 3.4 mit dem Profile-EQL-Plot in Abbildung 5.2 erkennt man, dass das Konfidenzintervall bei letzterem größer ist, was sich damit erklären lässt, dass der Tweedie-Ansatz – wie bereits erläutert – restriktivere Annahmen trifft und damit stärkere Aussagen zu machen im Stande ist.

Bemerkung. Die direkte Optimierung reagiert äußerst sensitiv auf die Wahl der Startwerte. Verwendet man beispielsweise den Vektor $(\alpha^0, \beta_1^0, \beta_2^0, \beta_3^0, \theta^0) = (0, 0, 0, 0, 1)$ erhält man die Schätzer $(\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\theta}) = (4.9780, 1.0205, -2.0455, -0.3083, 0.3881)$ mit einem zugehörigen EQL-Wert von -207.4769 als Lösung. Daher ist die direkte Optimierung nur dann zuverlässig, wenn man die Größenordnung der zu schätzenden Parameter kennt.

Tabelle 5.4: Anteil [in %] der mit der Blattfleckenkrankheit befallenen Blattfläche verschiedener Gerstenarten

	Gerstenart									
Gebiet	1	2	3	4	5	6	7	8	9	10
1	0.05	0.00	0.00	0.10	0.25	0.05	0.50	1.30	1.50	1.50
2	0.00	0.05	0.05	0.30	0.75	0.30	3.00	7.50	1.00	12.70
3	1.25	1.25	2.50	16.60	2.50	2.50	0.00	20.00	37.50	26.25
4	2.50	0.50	0.01	3.00	2.50	0.01	25.00	55.00	5.00	40.00
5	5.50	1.00	6.00	1.10	2.50	8.00	16.50	29.50	20.00	43.50
6	1.00	5.00	5.00	5.00	5.00	5.00	10.00	5.00	50.00	75.00
7	5.00	0.10	5.00	5.00	50.00	10.00	50.00	25.00	50.00	75.00
8	5.00	10.00	5.00	5.00	25.00	75.00	50.00	75.00	75.00	75.00
9	17.50	25.00	42.50	50.00	37.50	95.00	62.50	95.00	95.00	95.00

5.3.2 Erweiterte Binomialfamilie

In Tabelle 5.4 findet sich ein Datensatz (vgl. Wedderburn, 1974), der die von der Blattfleckenkrankheit (*Rhynchosporium secalis*) befallene Blattfläche verschiedener Gerstenarten für unterschiedliche Anbaugebiete beschreibt. McCullagh und Nelder (1989) untersuchen diesen Datensatz unter Zuhilfenahme des QL-Ansatzes $V(\mu) = \mu^2(1-\mu)^2$, da die Binomialvarianz $V(\mu) = \mu(1-\mu)$ speziell für sehr kleine bzw. sehr große Flächenanteile zu groß sein scheint.

Bemerkung. Der Datensatz enthält exakte Nullen und für die erweiterte Binomialvarianz gilt $\forall k, l : V(0) = 0$. Da die EQL-Funktion für Varianzfunktionen mit $V(\cdot) = 0$ nicht definiert ist, belegen wir die entsprechenden Zellen mit dem Wert `NA`. Das heißt im speziellen auch, dass die Ergebnisse von denen von McCullagh und Nelder abweichen.

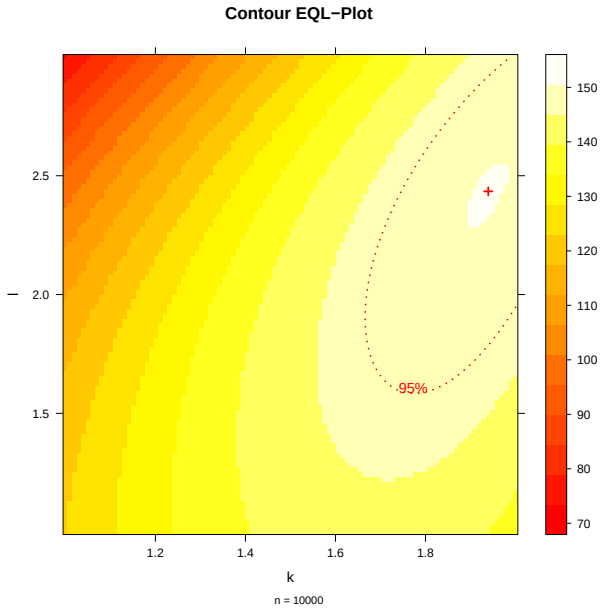
Mit Hilfe des EQL-Ansatzes wollen wir nun untersuchen, welche Werte für k und l plausibel sind. Dazu passen wir ein Modell mit den beiden Faktoren „Anbaugebiet“ und „Gerstenart“ an und untersuchen mit dem EQL-Konturplot welche Kombinationen von Werten für k und l innerhalb des Konfidenzbereichs liegen.

Für das Suchgitter bilden wir das kartesische Produkt der äquidistanten Folgen $L_1 \subseteq [1, 2]$ und $L_2 \subseteq [1, 3]$ mit $|L_1| = |L_2| = 100$. Damit ergibt sich ein Suchgitter Θ mit $|L_1| \cdot |L_2| = 10000$ Gitterpunkten.

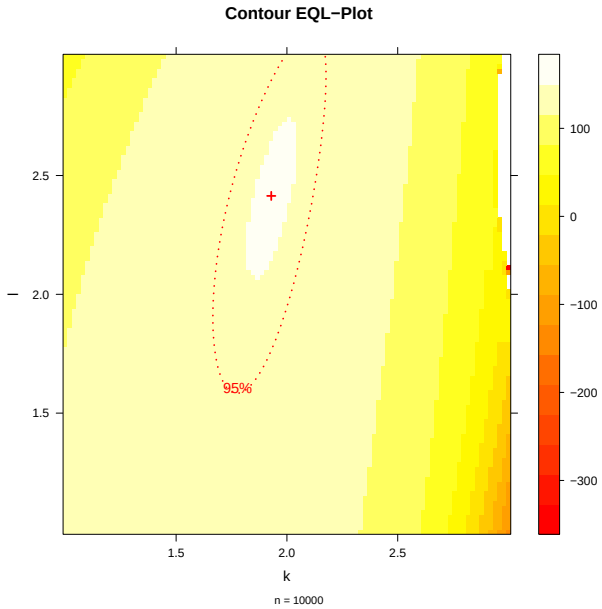
Der maximale EQL-Wert im Gitter von 150.6194 wird für $\hat{k} = 1.939$ und $\hat{l} = 2.434$ erreicht. Die Abbildung 5.3(a) zeigt den EQL-Konturplot und man erkennt, dass die von McCullagh und Nelder gewählte Parameterwahl $k = l = 2$ im Konfidenzbereich liegt, während die Binomialvarianz $k = l = 1$ von den Daten ganz deutlich *nicht* unterstützt wird.

Die Deviance bei der erweiterten Binomialvarianz wird – wie bereits besprochen – numerisch bestimmt. Für gewisse Parameter-Kombinationen konvergiert das Integrationsverfahren allerdings nicht. Verschiebt man das Suchgitter etwas nach links und verwendet

eine äquidistante Folge $L_1 \subseteq [1, 3]$, können gewisse Kombinationen nicht berechnet werden. In der Abbildung 5.3(b) erkennt man am rechten Rand einen weißen Bereich von Parameter-Kombinationen, für die kein EQL-Wert berechnet werden konnte.



(a) $L_1 \subseteq [1, 2]$



(b) $L_1 \subseteq [1, 3]$

Abbildung 5.3: EQL-Konturplots für den Blattflecken-Datensatz

6 Zusammenfassung und Ausblick

In dieser Arbeit wurde der Bogen von den Generalisierten Linearen Modellen über den Quasi-Likelihood-Ansatz zu der Extended-Quasi-Likelihood-Funktion gespannt. Dabei erlauben Generalisierte Lineare Modelle als Erweiterung des Linearen Modells einerseits die Verteilungsannahme zu verallgemeinern und andererseits durch die Verwendung der Linkfunktion auch nicht lineare Zusammenhänge zwischen den Prädiktoren und den Erwartungswerten herzustellen. Durch die Einbettung der Linkfunktion in eine Familie ist es möglich, unterschiedliche Linkfunktionen miteinander zu vergleichen. Somit ist man mit den Methoden des GLM in der Lage, sowohl die linearen Parameter des Modells als auch die Parameter der Linkfunktion zu schätzen. Will man allerdings die Varianzfunktion selbst parametrisieren und die Parameter schätzen, stößt man an die Grenzen des Generalisierten Linearen Modells.

Die Idee der Extended-Quasi-Likelihood-Funktion, die das Hauptaugenmerk dieser Arbeit bildete, schließt diese Lücke. Die EQL-Funktion baut auf dem Quasi-Likelihood-Ansatz auf, der die Verteilungsannahme des GLM ein weiteres mal relaxiert, indem er – anstatt die Kenntnis der vollständigen Verteilung vorauszusetzen – nur mehr den funktionalen Zusammenhang zwischen den ersten beiden Momenten spezifiziert. Über die Maximierung der EQL-Funktion ist man in der Lage, auch die Parameter der Varianzfunktion zu schätzen. Eine weitere Anwendungsmöglichkeit der EQL-Funktion ergibt sich in der simultanen Modellierung von Dispersion und Erwartungswert, womit Modelle betrachtet werden können, die nicht konstante Dispersionsparameter aufweisen.

Eine der Herleitungen der EQL-Funktion ergibt sich aus der Idee, die QL-Funktion derart zu erweitern, dass sie einerseits für einen bekannten Dispersionsparameter die gleichen Eigenschaften wie die QL-Funktion selbst besitzt und andererseits für einen unbekannten Dispersionsparameter die Eigenschaften mit der klassischen log-Likelihood-Funktion teilt. Die EQL-Funktion kann aber auch als ein Spezialfall der Sattelpunkt-Approximation angesehen werden, der ein weiterer Abschnitt dieser Arbeit gewidmet war. Approximiert man die Dichte einer Verteilung aus der Exponentialfamilie, erhält man wiederum die EQL-Funktion.

Den Abschluss dieser Arbeit bildete die Implementierung einer R-Bibliothek zur Schätzung der Parameter einer Varianzfamilie. Der gewählte Ansatz einer Gittersuche in einem vorgegebenen Suchbereich erlaubt zwar die Verwendung bereits gut dokumentierter Algorithmen, birgt aber die Gefahr anstatt globaler Maxima nur lokale zu finden. Der Vergleich mit den Ergebnissen eines generischer Optimierungsalgorithmus' zeigte, dass neben den Schwierigkeiten, hinreichend gute Startwerte für die Optimierung wählen zu müssen, die Schätzer ohnedies denen gleichen, die unser Ansatz liefert.

Es erwies sich außerdem, dass die Schätzung der Parameter der Varianzfamilie $\mathcal{F}_{k,l} = \{V(\mu) = \mu^k(1-\mu)^l\}$ zeitintensiver ist, als die Schätzung der Parameter des Potenzan-

satzes $\mathcal{F}_\theta = \{V(\mu) = \mu^\theta\}$. Eine Ursache wurde in dem tendenziell größeren Suchgitter identifiziert, während der Hauptgrund in der numerischen Integration der Deviance zu liegen scheint. Die exakte Lösung – so vorhanden – des Integrals stellt somit eine lohnende Verbesserung des Progammcodes dar.

Die Möglichkeiten der *Exponential Dispersion Models* im Allgemeinen und der Tweedie-Verteilungen im speziellen zur Schätzung des Parameters für die Varianzfamilie $\mathcal{F}_\theta = \{V(\mu) = \mu^\theta\}$ wurden angeschnitten, aber nicht erschöpfend ausgeführt. Eine mögliche Weiterführung dieser Arbeit wäre es also beispielsweise, diese unterschiedlichen Ansätze miteinander zu vergleichen.

Anhang

A Das Lineare Modell

In diesem Anhang wollen wir die wichtigsten Ergebnisse der Theorie des *Linearen Modells* zusammenfassen. Die folgenden Ausführungen orientieren sich dabei hauptsächlich an Casella und Berger (2002), Rawlings u. a. (2001) und Davison (2003).

Definition A.1 (Lineares Modell)**.** Bezeichne $1 \leq i \leq n : \boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{i,p-1})^\top$ einen Vektor von bekannten Größen (den *erklärenden Variablen*) und Y_i eine von $\boldsymbol{x}_i$ abhängige Zufallsvariable mit Erwartungswert $\mathbb{E}(Y_i) = \mu_i$ (die *Response-Variable*). Bezeichne $\boldsymbol{\beta} \in \mathbb{R}^p$ weiters den Vektor der unbekannten Modellparameter. Fasst man die Vektoren $\boldsymbol{x}_i$ zu der Designmatrix $X \in \mathbb{R}^{n \times p}$ zusammen, postuliert das Lineare Modell über

$$\boldsymbol{\mu} = X\boldsymbol{\beta} \tag{1}$$

einen linearen Zusammenhang zwischen der Designmatrix X und dem Vektor der Erwartungswerte $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^\top$.

Bemerkung. Setzen wir die Y_i als unabhängig voraus, lässt sich mit dem nicht beobachtbaren *statistischen Fehler* ϵ_i mit Erwartungswert $\mathbb{E}(\epsilon_i) = 0$ der Zusammenhang zwischen den erklärenden Variablen und der Response auch durch

$$Y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \epsilon_i$$

angeben, wobei $\boldsymbol{x}_i^\top$ die i-te Zeile der Designmatrix X bezeichnet.

Um statistische Aussagen erhalten zu können, trifft man üblicherweise Annahmen bezüglich des statistischen Fehlers. Wenn wir für die Varianz-Kovarianz-Matrix des Vektors $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\top$ von der Beziehung

$$\operatorname{cov}(\boldsymbol{\epsilon}) = \begin{pmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & & \sigma^2 \end{pmatrix}$$

mit unbekanntem aber festem σ^2 ausgehen,[1] und außerdem annehmen, dass $\boldsymbol{\epsilon}$ einer n-variaten Normalverteilung entstammt, folgt – unter der Annahme, dass das Modell korrekt spezifiziert wurde – auch der Vektor der Responsevariablen $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$ einer n-variaten Normalverteilung:

$$\boldsymbol{\epsilon} \sim \mathcal{N}_n(\boldsymbol{0}, \sigma^2 I_n) \Leftrightarrow \boldsymbol{Y} \sim \mathcal{N}_n(X\boldsymbol{\beta}, \sigma^2 I_n), \tag{2}$$

[1] Wir nehmen also *Homoskedastizität* des statistischen Fehlers an.

wobei $I_n \in \mathbb{R}^{n \times n}$ die Einheitsmatrix beschreibt.

Um einen Schätzer für $\boldsymbol{\beta}$ zu erhalten, können wir uns zweier verschiedener Ansätze bedienen, die unterschiedlich starke Annahmen treffen:

Kleinste Quadrate Wir treffen nur die Annahme hinsichtlich des linearen Zusammenhangs zwischen dem Erwartungswert und der Designmatrix (1). Den Schätzer $\hat{\boldsymbol{\beta}}$ erhält man über Minimierung der Fehlerquadratsumme (SSE):

$$\mathrm{SSE}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}\right)^2 = (\boldsymbol{y} - X\boldsymbol{\beta})^\top (\boldsymbol{y} - X\boldsymbol{\beta}).$$

Maximum Likelihood Zusätzlich nehmen wir eine Normalverteilung für $\boldsymbol{\epsilon}$ (und damit für $\boldsymbol{Y}$) wie in (2) an und erhalten so einen Schätzer über Maximierung der log-Likelihood-Funktion:

$$\begin{aligned}\mathscr{L}(\boldsymbol{\beta}, \sigma^2, \boldsymbol{y}) &= -\frac{n}{2}\log 2\pi - \frac{n}{2}\log \sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}\right)^2 \\ &= \text{Konstante} - \frac{1}{2\sigma^2}(\boldsymbol{y} - X\boldsymbol{\beta})^\top (\boldsymbol{y} - X\boldsymbol{\beta}).\end{aligned}$$

Da der relevante Teil der log-Likelihood-Funktion proportional zur Fehlerquadratsumme ist, liefern beide Ansätze den gleichen Schätzer für $\boldsymbol{\beta}$. Mittels Nullsetzen der Ableitung nach $\boldsymbol{\beta}$, erhält man die sogenannte *Normalgleichung*:

$$\frac{\partial \,\mathrm{SSE}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2X^\top \boldsymbol{y} + 2X^\top X\boldsymbol{\beta} \stackrel{!}{=} 0 \Rightarrow X^\top X\hat{\boldsymbol{\beta}} = X^\top \boldsymbol{y}.$$

Bemerkung. Die Normalgleichung hat nur dann eine eindeutige Lösung, falls $X^\top X$ regulär ist, was bedeutet, dass X vollen Spaltenrang hat (vgl. Rawlings u. a., 2001, S. 79). Der Parameter-Schätzer lautet in diesem Fall:

$$\hat{\boldsymbol{\beta}} = (X^\top X)^{-1} X^\top \boldsymbol{y}. \tag{3}$$

Um einen Schätzer $\hat{\sigma}^2$ für die Varianz von Y_i zu erhalten, müssen wir die log-Likelihood-Funktion nach σ^2 ableiten und Null setzen:

$$\frac{\partial}{\partial \sigma^2}\mathscr{L}(\boldsymbol{\beta}, \sigma^2, \boldsymbol{y}) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}\left(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}\right)^2 \stackrel{!}{=} 0,$$

wobei wir für den unbekannten Parametervektor $\boldsymbol{\beta}$ dessen Schätzer $\hat{\boldsymbol{\beta}}$ einsetzen. Somit erhalten wir:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}\right)^2 = \frac{1}{n}\,\mathrm{SSE}(\hat{\boldsymbol{\beta}}).$$

Mit dem Schätzer $\hat{\beta}$ ist es uns jetzt möglich, einen Schätzer $\hat{\boldsymbol{\mu}}$ für den Vektor der Erwartungswerte $\boldsymbol{\mu}$ zu erhalten:

$$\hat{\boldsymbol{\mu}} = X\hat{\boldsymbol{\beta}} \stackrel{(3)}{=} X(X^\top X)^{-1} X^\top \boldsymbol{y} =: H\boldsymbol{y}.$$

Tabelle A.1: Momente einiger Größen eines Linearen Modells

ZV	$\mathbb{E}(\cdot)$	$\mathrm{cov}(\cdot)$
$\boldsymbol{Y}$	$\boldsymbol{\mu}$	$\sigma^2 I_n$
$\boldsymbol{\epsilon}$	$\mathbf{0}$	$\sigma^2 I_n$
$\hat{\boldsymbol{\beta}}$	$\boldsymbol{\beta}$	$\sigma^2 (X^\top X)^{-1}$
$\hat{\boldsymbol{\mu}}$	$\boldsymbol{\mu}$	$\sigma^2 H$
$\boldsymbol{r}$	$\mathbf{0}$	$\sigma^2 (I_n - H)$

Die Matrix H bezeichnet dabei die sogenannte *Hat-Matrix*, die symmetrisch und idempotent ist:

$$\begin{aligned} H^\top &= \left(X(X^\top X)^{-1}X^\top\right)^\top = X\left((X^\top X)^{-1}\right)^\top X^\top = X\left((X^\top X)^\top\right)^{-1} X^\top = H, \\ HH &= X(X^\top X)^{-1} \underbrace{X^\top X (X^\top X)^{-1}}_{=I_n} X^\top = X(X^\top X)^{-1}X^\top = H, \end{aligned}$$

außerdem folgt sofort aus der Definition:

$$HX = X(X^\top X)^{-1}X^\top X = X.$$

Der Vektor der *Residuen* $\boldsymbol{r} = (r_1, \ldots, r_n)^\top$ beschreibt die Abweichung der beobachteten Werte $\boldsymbol{y}$ von dem Vektor der geschätzten Erwartungswerte $\hat{\boldsymbol{\mu}}$:

$$\boldsymbol{r} = \boldsymbol{y} - \hat{\boldsymbol{\mu}} = \boldsymbol{y} - H\boldsymbol{y} = (I_n - H)\boldsymbol{y}.$$

Bemerkung. Auch die Matrix $I_n - H$ ist symmetrisch und idempotent, wie leicht überprüft werden kann. Außerdem gilt $(I_n - H)X = X - HX = \mathbf{0}$.

Unter der Annahme, dass das Modell durch Gleichung (1) korrekt spezifiziert ist, lassen sich die Momente der eingeführten Terme berechnen. Gilt außerdem die Normalverteilungsannahme (2), sind die Terme $\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\mu}}$ und $\boldsymbol{r}$ als Linearkombinationen von (multivariat) normalverteilten Größen ebenfalls (multivariat) normalverteilt. Die Tabelle A.1 fasst die Momente der vorgestellten Größen zusammen, wobei wir normalverteilte statistische Fehler $\boldsymbol{\epsilon}$ annehmen.

Bemerkung. Die Varianz der Schätzer $\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\mu}}$ und $\boldsymbol{r}$ hängt einerseits von der unbekannten Größe σ^2 und andererseits von der Designmatrix X ab. In einem Versuchsaufbau, in dem die erklärenden Variablen kontrolliert werden können, kann so Einfluss auf die Varianz der Schätzer genommen werden (vgl. Rawlings u. a., 2001, S. 92).

Mit den folgenden Sätzen, lassen sich auch die Momente der Fehlerquadratsumme bestimmen (vgl. Davison, 2003, S. 370 ff.).

Satz A.2. *Der statistische Fehler $\boldsymbol{\epsilon}$ sei multivariat normalverteilt mit $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 I_n)$. Dann sind die Statistiken $\hat{\boldsymbol{\beta}}$ und $\boldsymbol{r}$ unabhängig.*

Beweis. Mit $\boldsymbol{\epsilon}$ sind auch $\hat{\boldsymbol{\beta}}$ und $\boldsymbol{r}$ normalverteilt, daher reicht es zu zeigen, dass die Kovarianz-Matrix $\operatorname{cov}(\hat{\boldsymbol{\beta}}, \boldsymbol{r})$ gerade Null ergibt. Mit $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$ gilt:

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (X^\top X)^{-1} X^\top y = (X^\top X)^{-1} X^\top (X\boldsymbol{\beta} + \boldsymbol{\epsilon}) = \boldsymbol{\beta} + (X^\top X)^{-1} X^\top \boldsymbol{\epsilon}, \\ \boldsymbol{r} &= (I_n - H)y \qquad = (I_n - H)(X\boldsymbol{\beta} + \boldsymbol{\epsilon}) \qquad = (I_n - H)\boldsymbol{\epsilon}.\end{aligned}$$

Damit folgt für die Kovarianz-Matrix:

$$\begin{aligned}\operatorname{cov}(\hat{\boldsymbol{\beta}}, \boldsymbol{r}) &= \operatorname{cov}(\boldsymbol{\beta} + (X^\top X)^{-1} X^\top \boldsymbol{\epsilon}, (I_n - H)\boldsymbol{\epsilon}) \\ &= (X^\top X)^{-1} X^\top \underbrace{\operatorname{cov}(\boldsymbol{\epsilon}, \boldsymbol{\epsilon})}_{=\operatorname{var}(\boldsymbol{\epsilon}) = \sigma^2 I_n} (I_n - H) \\ &= \sigma^2 (X^\top X)^{-1} X^\top - \sigma^2 (X^\top X)^{-1} \underbrace{X^\top X (X^\top X)^{-1}}_{=I_n} X^\top \\ &= \mathbf{0}.\end{aligned}$$

□

Korollar A.3. *Es gelten die gleichen Voraussetzungen wie in Satz A.2. Dann sind* $\hat{\boldsymbol{\beta}}$ *und* $\operatorname{SSE}(\hat{\boldsymbol{\beta}}) = \boldsymbol{r}^\top \boldsymbol{r}$ *unabhängig.*

Satz A.4. *Es gelten die gleichen Voraussetzungen wie in Satz A.2. Dann gilt*

$$\frac{1}{\sigma^2} \boldsymbol{r}^\top \boldsymbol{r} = \frac{1}{\sigma^2} \operatorname{SSE}(\hat{\boldsymbol{\beta}}) \sim \chi^2_{n-p}.$$

Beweis. Mit $\boldsymbol{r} = (I - H)\boldsymbol{y}$ betrachten wir vorerst:

$$\begin{aligned}\boldsymbol{\epsilon}^\top \boldsymbol{\epsilon}/\sigma^2 &= 1/\sigma^2 \left(\boldsymbol{y} - X\hat{\boldsymbol{\beta}} + X\hat{\boldsymbol{\beta}} - X\boldsymbol{\beta}\right)^\top \left(\boldsymbol{y} - X\hat{\boldsymbol{\beta}} + X\hat{\boldsymbol{\beta}} - X\boldsymbol{\beta}\right) \\ &= 1/\sigma^2 \left(\boldsymbol{r} + X(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right)^\top \left(\boldsymbol{r} + X(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right) \\ &= \boldsymbol{r}^\top \boldsymbol{r}/\sigma^2 - 2/\sigma^2 \boldsymbol{r}^\top X(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + 1/\sigma^2 (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top X^\top X(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\ &= \boldsymbol{r}^\top \boldsymbol{r}/\sigma^2 - 2/\sigma^2 \boldsymbol{y}^\top \underbrace{(I_n - H)^\top X}_{=\mathbf{0}} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + 1/\sigma^2 (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top X^\top X(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\ &= \boldsymbol{r}^\top \boldsymbol{r}/\sigma^2 + \frac{1}{\sigma^2} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top X^\top X(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}). \qquad (*)\end{aligned}$$

Da $\hat{\boldsymbol{\beta}}$ p-variat normalverteilt mit Erwartunsgwert $\mathbb{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ und Varianz $\operatorname{var}(\hat{\boldsymbol{\beta}}) = \sigma^2 (X^\top X)^{-1}$ ist, folgt:

$$\frac{1}{\sigma^2} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top X^\top X(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim \chi^2_p,$$

und analog gilt für $\boldsymbol{\epsilon}^\top \boldsymbol{\epsilon}/\sigma^2$:

$$\boldsymbol{\epsilon}^\top \boldsymbol{\epsilon}/\sigma^2 \sim \chi^2_n.$$

Die momenterzeugende Funktion einer χ^2_k-verteilten Größe lautet

$$M_{\chi^2_k}(t) = (1 - 2t)^{-k/2}.$$

Die momenterzeugende Funktion einer Summe zweier unabhängiger Zufallsvariablen ist das Produkt der jeweiligen momenterzeugenden Funktionen (vgl. Casella und Berger, 2002, S. 155). Somit können wir wegen der Unabhängigkeit von $\hat{\boldsymbol{\beta}}$ und $\boldsymbol{r}$ aus Satz A.2 die momenterzeugende Funktion der Summe in $(*)$ und damit die Verteilung von $\mathrm{SSE}(\hat{\boldsymbol{\beta}})/\sigma^2$ bestimmen:

$$\begin{aligned}(1-2t)^{-n/2} &= \mathbb{E}\big(\exp\{t\cdot \boldsymbol{r}^\top \boldsymbol{r}/\sigma^2\}\big)(1-2t)^{-p/2} \Rightarrow \\ \mathbb{E}\big(\exp\{t\cdot \boldsymbol{r}^\top \boldsymbol{r}/\sigma^2\}\big) &= (1-2t)^{-(n-p)/2} \Rightarrow \\ \boldsymbol{r}^\top \boldsymbol{r}/\sigma^2 &\sim \chi^2_{n-p}.\end{aligned}$$

□

Bemerkung 1. Aus Satz A.4 lässt sich ein unverzerrter Schätzer für σ^2 angeben. Aus der Verteilung von $\mathrm{SSE}(\hat{\boldsymbol{\beta}})/\sigma^2$ folgt für den Erwartungswert der Fehlerquadratsumme $\mathbb{E}(\mathrm{SSE}(\hat{\boldsymbol{\beta}})) = \sigma^2(n-p)$. Somit kann ein erwartungstreuer Schätzer für die Varianz durch

$$s^2 = \frac{1}{n-p}\,\mathrm{SSE}(\hat{\boldsymbol{\beta}})$$

definiert werden.

Bemerkung 2. Es lässt sich zeigen, dass auch für den Fall, dass $\boldsymbol{Y}$ nicht normalverteilt ist, die Beziehung

$$\mathbb{E}(\mathrm{SSE}(\hat{\boldsymbol{\beta}})) = \sigma^2(n-p)$$

gilt.

Um einen einzelnen Parameter β_j auf seine statistische Signifikanz zu testen, kann man bei bekannter Varianz die Teststatistik

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma^2 c_{jj}}} \sim \mathcal{N}(0,1)$$

verwenden, wobei c_{jj} das zu β_j gehörige Diagonalelement[2] von $C = (X^\top X)^{-1}$ bezeichnet. Will man also überprüfen, ob x_j für das Modell relevant ist, testet man die Nullhypothese $H_0 : \beta_j = 0$ und erhält über die Quantile der Standardnormalverteilung den entsprechenden kritischen Bereich.

Ist die Varianz nicht bekannt, kann man aufgrund der Unabhängigkeit von $\hat{\boldsymbol{\beta}}$ und $\mathrm{SSE}(\hat{\boldsymbol{\beta}})$ aus Korrolar A.3 die Statistik

$$T = \frac{\frac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma^2 c_{jj}}}}{\sqrt{\frac{n-p}{\sigma^2}s^2/(n-p)}} = \frac{\hat{\beta}_j - \beta_j}{\sqrt{s^2 c_{jj}}} \sim t_{n-p}$$

verwenden. Die Verteilung ergibt sich aus der Tatsache, dass der Quotient aus einer standardnormalverteilten Größe und der Wurzel einer mit ihren Freiheitsgraden gewichteten

[2] Bei der hier verwendeten Notation steht c_{jj} an $j+1$-ter Stelle in den Diagonalelementen von $C = (X^\top X)^{-1}$.

χ^2-verteilten Zufallsvariable einer Student-t-Verteilung folgt (vgl. Casella und Berger, 2002, S. 223).

Will man die Relevanz mehrerer erklärenden Größen simultan bestimmen, bedient man sich einer „Analysis-of-Variance“ (ANOVA). Dazu kann man folgende Zerlegung verwenden:

$$\begin{aligned} \sum_{i=1}^{n}(y_i - \bar{y})^2 &= \sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2 + \sum_{i=1}^{n}(\hat{\mu}_i - \bar{y})^2 \\ \mathrm{SST} &= \mathrm{SSE}(\hat{\boldsymbol{\beta}}) \qquad + \mathrm{SSR}(\hat{\boldsymbol{\beta}}). \end{aligned} \tag{4}$$

Dabei bezeichnet SSE die Fehlerquadratsumme, SSR die Residuenquadratsumme und SST die Gesamtquadratsumme. Da die Größe „SST“ unabhängig vom Modell und damit für eine gegebene Stichprobe fest ist und da $\hat{\boldsymbol{\beta}}$ gerade so gewählt wurde, dass SSE minimal ist, folgt, dass die Residuenquadratsumme SSR maximal in $\hat{\boldsymbol{\beta}}$ ist.

Bemerkung. Die Zerlegung in Gleichung (4) lässt sich mit $\bar{\boldsymbol{y}} := (\bar{y}, \ldots, \bar{y})^\top$ wie folgt zeigen:

$$\begin{aligned} \mathrm{SST} = (\boldsymbol{y} - \bar{\boldsymbol{y}})^\top(\boldsymbol{y} - \bar{\boldsymbol{y}}) &= \big((\boldsymbol{y} - \hat{\boldsymbol{\mu}}) - (\bar{\boldsymbol{y}} - \hat{\boldsymbol{\mu}})\big)^\top \big((\boldsymbol{y} - \hat{\boldsymbol{\mu}}) - (\bar{\boldsymbol{y}} - \hat{\boldsymbol{\mu}})\big) \\ &= (\boldsymbol{y} - \hat{\boldsymbol{\mu}})^\top(\boldsymbol{y} - \hat{\boldsymbol{\mu}}) + 2(\boldsymbol{y} - \hat{\boldsymbol{\mu}})^\top(\hat{\boldsymbol{\mu}} - \bar{\boldsymbol{y}}) + (\hat{\boldsymbol{\mu}} - \bar{\boldsymbol{y}})^\top(\hat{\boldsymbol{\mu}} - \bar{\boldsymbol{y}}) \\ &= \mathrm{SSE}(\hat{\boldsymbol{\beta}}) + 2(\boldsymbol{y}^\top\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}^\top\hat{\boldsymbol{\mu}} - \boldsymbol{y}^\top\bar{\boldsymbol{y}} + \hat{\boldsymbol{\mu}}^\top\bar{\boldsymbol{y}}) + \mathrm{SSR}(\hat{\boldsymbol{\beta}}). \qquad (*) \end{aligned}$$

Um zu zeigen dass der mittlere Term gerade Null ergibt, verwenden wir die folgende Eigenschaft (vgl. Fahrmeir u. a., 2007, S. 97):

$$\bar{\hat{\mu}} = \frac{1}{n}\sum_{i=1}^{n}\hat{\mu}_i = \bar{y}. \qquad (**)$$

Damit ergibt sich mit $\hat{\boldsymbol{\mu}} = H\boldsymbol{y}$ und der Idempotenz der (symmetrischen) Hat-Matrix H für den gemischten Term in $(*)$ die Identität:

$$\begin{aligned} 2(\boldsymbol{y}^\top\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}^\top\hat{\boldsymbol{\mu}} - \boldsymbol{y}^\top\bar{\boldsymbol{y}} + \hat{\boldsymbol{\mu}}^\top\bar{\boldsymbol{y}}) &= 2(\underbrace{\boldsymbol{y}^\top H\boldsymbol{y} - \boldsymbol{y}^\top HH\boldsymbol{y}}_{=\boldsymbol{y}^\top H\boldsymbol{y} - \boldsymbol{y}^\top H\boldsymbol{y} = 0}) + 2\sum_{i=1}^{n}(\hat{\mu}_i\bar{y} - y_i\bar{y}) \\ &= 2n\bar{y}\bar{\hat{\mu}} - 2n\bar{y}^2 \\ &\overset{(**)}{=} 2n\bar{y}^2 - 2n\bar{y}^2 = 0. \end{aligned}$$

Mit Hilfe dieser Zerlegung kann man ein Gütemaß für ein Modell bestimmen. Dazu verwendet man eine der beiden Größen:

$$R^2 = 1 - \frac{\mathrm{SSE}(\hat{\boldsymbol{\beta}})}{\mathrm{SST}}, \qquad R^2_{\mathrm{adj}} = 1 - \frac{n-1}{n-p}\frac{\mathrm{SSE}(\hat{\boldsymbol{\beta}})}{\mathrm{SST}},$$

wobei n die Anzahl der Beobachtungen und p die Anzahl der Parameter im linearen Prädiktor beschreibt. Lässt man die Anzahl der Parameter p anwachsen, wird die Anpassung besser und R^2 strebt gegen Eins. Daher ist die unskalierte Version nur bedingt

aussagekräftig. Der Skalierungsterm für R^2_{adj} berücksichtigt eben auch die Anzahl der Parameter und kann bei schlechten Modellen auch negativ sein.

Um die statistische Notwendigkeit einer Teilmenge an erklärenden Variablen in einem Modell zu überprüfen, testet man formal für das Modell

$$\begin{gathered}\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon} = \begin{pmatrix} X_1 & X_2 \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} + \boldsymbol{\epsilon}, \\ X \in \mathbb{R}^{n\times p}, X_1 \in \mathbb{R}^{n\times q}, X_2 \in \mathbb{R}^{n\times(p-q)}, \\ \boldsymbol{\beta} \in \mathbb{R}^p, \boldsymbol{\beta}_1 \in \mathbb{R}^q, \boldsymbol{\beta}_2 \in \mathbb{R}^{p-q}, \boldsymbol{\epsilon} \in \mathbb{R}^n\end{gathered}$$

die Nullhypothese $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$. Kann man H_0 nicht verwerfen, ist das einfachere Modell $\boldsymbol{y} = X_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}$ bereits ausreichend.

Um eine Teststatistik zu finden, zerlegen wir die Fehlerquadratsumme des kleineren Modells und erhalten so die Beziehung:

$$\text{SSE}(\hat{\boldsymbol{\beta}}_1) = \text{SSE}(\hat{\boldsymbol{\beta}}) + \big[\text{SSE}(\hat{\boldsymbol{\beta}}_1) - \text{SSE}(\hat{\boldsymbol{\beta}})\big].$$

Es lässt sich zeigen, dass die Differenz $\text{SSE}(\hat{\boldsymbol{\beta}}_1) - \text{SSE}(\hat{\boldsymbol{\beta}})$ unabhängig von $\text{SSE}(\hat{\boldsymbol{\beta}})$ ist. Unter der Annahme, dass das einfache Modell richtig ist,[3] folgt mit Satz A.4 für die Fehlerquadratsummen $\text{SSE}(\hat{\boldsymbol{\beta}})$ und $\text{SSE}(\hat{\boldsymbol{\beta}}_1)$:

$$\text{SSE}(\hat{\boldsymbol{\beta}}) \sim \sigma^2\chi^2_{n-p}, \qquad\qquad \text{SSE}(\hat{\boldsymbol{\beta}}_1) \sim \sigma^2\chi^2_{n-q}.$$

Mit der Unabhängigkeit der beiden Fehlerquadratsummen lässt sich über die momenterzeugende Funktion leicht zeigen, dass auch der Differenzterm $\text{SSE}(\hat{\boldsymbol{\beta}}_1) - \text{SSE}(\hat{\boldsymbol{\beta}})$ einer χ^2-Verteilung mit $p-q$ Freiheitsgraden folgt. Der mit den jeweiligen Freiheitsgraden skalierte Quotient zweier χ^2-verteilten Größen folgt einer F-Verteilung, wobei die Anzahl der Freiheitsgrade gerade den Freiheitsgraden der χ^2-verteilten Zufallsvariablen entspricht (vgl. Casella und Berger, 2002, S. 225). Somit erhalten wir die Teststatistik:

$$F = \frac{\big(\text{SSE}(\hat{\boldsymbol{\beta}}_1) - \text{SSE}(\hat{\boldsymbol{\beta}})\big)/(p-q)}{\text{SSE}(\hat{\boldsymbol{\beta}})/(n-p)} \sim F_{p-q;n-p}. \tag{5}$$

Mit der Statistik (5) lässt sich die ANOVA durchführen, indem man sukzessive Terme in ein Modell aufnimmt und den p-Wert für die mittlere Reduktion der Fehlerquadratsumme aus der entsprechenden F-Verteilung bestimmt. Ist der p-Wert kleiner als das Signifikanzniveau α ist die entsprechende Nullhypothese zu verwerfen und die erklärende Größe ist statistisch signifikant.

[3] Ist das kleinere Modell korrekt, ist auch das größere korrekt, da man in diesem Fall nur $\boldsymbol{\beta}_2 = \mathbf{0}$ setzen muss.

B R-Dokumentation

In diesem Anhang befindet sich die Dokumentation der R-Bibliotheken *ttutils* und *EQL*, die im Rahmen dieser Arbeit erstellt wurden. Die Bibliotheken selbst werden auf dem *Comprehensive R Archive Network* (CRAN)[4] bereitgestellt und können somit in R über den Befehl `install.packages` von einem CRAN-Spiegel bezogen werden. Da *EQL* einige Funktionen von *ttutils* benutzt, muss man, um den Funktionsumfang von *EQL* nutzen zu können, auch die Bibliothek *ttutils* installieren.

[4] Das CRAN ist unter `http://cran.r-project.org/index.html` zu finden.

B.1 Paket `ttutils`

`ttutils-package` *Utility Functions*

Description

The package **ttutils** contains some auxiliary functions.

See section 'Index' for a list of exported functions. Section 'Internals' lists the internal functions of the package, which are not exported but may be referenced by `ttutils:::.functionName`.

Details

Version:	1.0-0
Date:	2009-06-18
License:	GPL-2
Built:	R 2.8.1; ; 2009-06-22 15:18:40; unix

Index

`check`	: Generic function to check the validity of a given object
`interval`	: Interval class
`isInteger`	: Test for integrity
`liesWithin`	: Test for interval coverage
`merge.list`	: Merge two lists
`plotAndSave`	: Display and save a plot

Internals

`.parseRelation`	: Parse a relation symbol and return the result of the comparison
`.savePdf`	: Save a plot as "pdf"
`.savePs`	: Save a plot as "eps"

Author(s)

Thorn Thaler <thorn.thaler@thothal.com>

Maintainer: Thorn Thaler <thorn.thaler@thothal.com>

`check` *Check Objects*

Description

`check` is a generic function that tests the validity of a given object.

Usage

```
check(object, ...)
```

Arguments

`object`	an object to be tested for validity.
`...`	further arguments to be passed to the particular dispatched function.

Details

`check` tests if a given object meets the formal requirements of being a valid object of its class. If the test fails, additional warnings should be provided, giving some information why the test failed.

Value

returns `TRUE` if `object` passes the validity test for the specific class and `FALSE` otherwise.

Note

R's dispatching mechanism determines the class of a given object and then calls the function `check.<class-name>`. If no specific `check` function is found, `check.default` is called. The function `check.default` does not make much sense, for the purpose of `check` is to test the validity for a *specific class*. Hence, `check.default` simply returns `FALSE` together with a warning message that no specific `check.<class-name>` function was found.

The dispatching mechanism has immediately two consequences:

1. a class specific `check` routine need not to check whether the object belongs to the class itself, because if it would not, the function would not have been called.
2. if no specific `check` routine is found, the result for a call of `check` will be `FALSE`, since in this case the default function is called which will return `FALSE` in any case.

Author(s)

Thorn Thaler

`interval` *Interval Class*

Description

`interval` constructs an object of class `interval` representing an interval.

`liesWithin` checks if a number lies within a given interval.

Usage

```
interval(lower, upper, left=c(">=", ">"), right=c("<=", "<"))

liesWithin(x, int)
```

Arguments

`lower`	the lower boundary of the interval. Can be set to `-Inf`.
`upper`	the upper boundary of the interval. Can be set to `Inf`.
`left, right`	a comparison symbol. Must be one of (“>=”, “>”) for `left` and (“<=”, “<”) for `right`, respectively. Determines whether the boundary values are included in the interval or not. The default is “>=” and “<=”, respectively.
`x`	a numeric vector or array giving the numbers to be checked.
`int`	an interval object.

Value

`interval` returns an object of class `interval` containing the following components:

`lower`	the lower boundary of the interval
`upper`	the upper boundary of the interval
`left`	the left comparison operator
`right`	the right comparison operator

`liesWithin` returns `TRUE` if the given number lies within the interval and `FALSE` otherwise.

Author(s)

Thorn Thaler

Examples

```
i <- interval(-3, 3, left=">")

liesWithin(-3:5, i)
```

isInteger *Test For Integrity*

Description

isInteger tests if a given number is an integer.

Usage

```
isInteger(n, tol = .Machine$double.eps)
```

Arguments

n	a vector or an array of values to be tested.
tol	a numeric value giving the tolerance level.

Details

As opposed to is.integer this function tests for integrity of a given value, rather than being of *type* integer.

In R integers are specified by the suffix L (e.g. 1L), whereas all other numbers are of class numeric independent of their value. The function is.integer does not test whether a given variable has an integer value, but whether it belongs to the class integer.

In contrast, the function isInteger compares the difference between its argument and its rounded argument. If it is smaller than some predefined tolerance level, the variable is regarded as integer.

Value

TRUE if the argument n has an integer value, FALSE otherwise.

Note

The R function c concatenates its argument and forms a vector. In doing so, it coerces the values to a common type. Hence, attention has to be paid, because isInteger may give some unexpected results in this case. The R command list, however, does not coerce its arguments (see the example).

Author(s)

Thorn Thaler

See Also

is.integer

Examples

```
# isInteger tests if the _value_ of a variable is an integer
# 'c' as opposed to 'list' coerces its arguments!
isInteger(c("test", 1, 2, 2.1))     # FALSE FALSE FALSE FALSE
isInteger(list("test", 1, 2, 2.1))  # FALSE TRUE TRUE FALSE

class(1L)  # integer
typeof(1L) # integer
class(1)   # numeric
typeof(1)  # double

# is.integer tests if the _class_ of a variable is 'integer'
is.integer(c("test", 1, 2))     # FALSE
is.integer(list("test", 1, 2)) # FALSE
is.integer(1)                  # FALSE
is.integer(1L)                 # TRUE
```

---

merge.list *Merge Two Lists*

Description

`merge.list` merges two lists. If there are identical names in both lists, only the elements of the first list are considered.

Usage

```
## S3 method for class 'list':
merge(x, y = NULL, mergeUnnamed = TRUE, ...)
```

Arguments

`x`	a list of possibly named elements. All of these are in the merged list.
`y`	a list of possibly named elements or any object, which can be coerced to `list`. If an element has a name occuring also in the argument `x`, it will not be included in the merged list to avoid duplicate names. If `NULL`, `x` is returned.
`mergeUnnamed`	logical. If `TRUE` (the default) unnamed elements in the second list are always included.
`...`	arguments to be passed to or from methods.

Details

The purpose of this function is to merge two lists (e.g. argument lists). If a named element is found as well in the first list as in the second, only the value of the element in the first list is considered. One can think of the second list as a list of default values, which should be considered only if they are not set explicitly in the first list.

Unnamed elements in `y` are included in the merged list only if `mergeUnnamed` is `TRUE`.

Value

a list containing all elements of the argument `x` and those of `y` having names not occuring in `x`.

Author(s)

Thorn Thaler

Examples

```
merge(list(a=1, b="test"), list(3, b=2)) # list(a=1, b="test", 3)
merge(list(1), "test")                   # list(1, "test")
merge(list(1), "test", FALSE)            # list(1)
merge(list(1))                           # list(1)
merge(list(1, a=2, b=3), list(2, b=4))   # list(1, a=2, b=3, 2)
merge(list(1), list(2, b=3), FALSE)      # list(1, b=3)

a <- list(1, 2, 3)
b <- list("a", "b", "c")
names(a)[2] <- names(b)[2] <- "z"
all.equal(merge(a, b), list(1, z=2, 3, "a", "c")) # TRUE
```

---

plotAndSave *Display And Save A Plot*

Description

`plotAndSave` saves a plot as "pdf" and/or "eps" and additionally displays the plot.

Usage

```
plotAndSave(plot.func, plot.name, ..., folder = getwd(),
            format = c("eps", "pdf"),
            ps.options=list(onefile = TRUE, horizontal = FALSE,
                 paper = "special", width = 7, height = 7),
            pdf.options = list(onefile = TRUE),
            do.plot = TRUE, do.return = do.plot)
```

Arguments

`plot.func`	either a function or a non-empty character string naming the plotting function to be called.
`plot.name`	a character string (without any suffix such as ".pdf" or ".eps") giving the name of the file where the plot should be saved to.
`...`	additional arguments to be passed to the plotting function.
`folder`	a character string giving the name of the folder to which the plot should be saved. The default is the current directory.

format	output format. Must be a subset of ("eps", "pdf"). The default is to produce both an eps-file and a pdf-file. Can be abbreviated.
ps.options	named list of options to be passed to the PostScript device driver. See postscript for further details.
pdf.options	named list of options to be passed to the PDF device driver. See pdf for further details.
do.plot	logical. If TRUE (the default) the plot is displayed.
do.return	logical. If TRUE the return value of the plotting function is returned. Defaults to the value of the parameter do.plot.

Details

The purpose of this function is to produce a plot on the monitor and to save it to a file simultaneously.

The file name must be given without any file-suffix. Depending on the argument format the function then generates either a PDF-file, an EPS-file or both with the appropriate suffix. The path should not be included in the file name, since the location where the files should be saved to is controlled by the parameter folder.

The function needs a plotting function to be defined, which actually does the plotting itself. Additional arguments (e.g. further graphical parameters) can be passed to plotAndSave, which in turn, passes these arguments down to the plotting function,

The parameters of the PostScript and the PDF device are controlled by the arguments ps.options and pdf.options, respectively.

Value

the return value of the plotting function.

Note

When using Trellis plots from package **lattice** one has to assure that the plotting function actually *does* the plotting. Since the default behaviour of Trellis plots is just to return the Trellis object, one should wrap the call to the particular **lattice** function in a call of the function print. The generic function print ensures that the plot is displayed and not just returned as an object.

Author(s)

Thorn Thaler

See Also

pdf, postscript

Examples

```
## Not run:
## Plotting Function
# For 'lattice' graphics:
# WRONG:
```

```
# f <- function(x, ...) xyplot(x~sin(x), ...)
# CORRECT:
# f <- function(x, ...) print(xyplot(x~sin(x), ...))

f <- function(x, ...) plot(x, sin(x), col=2, type="l", ...)

# Save the plot as "Sine_Function.pdf" in the current folder
# and add a title to the plot

plotAndSave(f, "Sine_Function", x=seq(-pi, pi, length=100),
            main="Sine-Function", format="p")

## End(Not run)
```

B.2 Paket EQL

`EQL-package` *Extended Quasi-Likelihood Function (EQL)*

Description

The package **EQL** contains functions for

- computation of the EQL for a given family of variance functions
- Edgeworth approximations
- Saddlepoint approximations
- related auxiliary functions (e.g. Hermite polynomials)

See section 'Index' for a list of exported functions. Section 'Internals' lists the internal functions of the package, which are not exported but may be referenced by `EQL:::functionName`.

Details

Version:	1.0-0
Date:	2009-06-18
Depends:	ttutils(>= 0.1-0)
Imports:	lattice(>= 0.17-17)
License:	GPL-2
Built:	R 2.8.1; ; 2009-06-22 15:24:08; unix

Index

Internals

`.eql`	: Computes a single EQL value
`.getFactor`	: Calculates the normalizing factor for the saddlepoint approximation
`.missingFormals`	: Check if a list contains all the arguments of a particular function

Author(s)

Thorn Thaler <thorn.thaler@thothal.com>

Maintainer: Thorn Thaler <thorn.thaler@thothal.com>

See Also

ttutils

`approximation` *An Approximation Class*

Description

An object of class `approximation` stores the approximation nodes together with the approximation itself. Some meta information is saved as well.

Usage

```
approximation(y, approx, n,
              type = c("standardized", "mean", "sum"),
              approx.type = c("Edgeworth", "Saddlepoint"))

## S3 method for class 'approximation':
plot(x, do.annotate = TRUE, ...)
```

Arguments

`y`	a numeric vector or array giving the approximation nodes.
`approx`	a numeric vector or array giving the approximated values at `y`.
`n`	a positive integer giving the number of i.i.d. random variables in the sum.
`type`	a character string giving the type of approximation, i.e. which kind of sum is to be approximated. Must be one of (“standardized”, “mean”, “sum”), representing the shifted and scaled sum, the weighted sum and the raw sum. Can be abbreviated.
`approx.type`	a character string giving the approximation routine used. Must be one of of (“Edgeworth”, “Saddlepoint”) and can be abbreviated.

x	an approximation object.
do.annotate	logical. If TRUE (the default) the value of the argument n is added to the plot.
...	other parameters to be passed through to the plotting function. Giving a named argument for any of • main • sub • type • xlab • ylab overrides the default values in plot.approximation.

Value

An object of class approximation contains the following components:

y	a numeric vector of values at which the approximation is evaluated (the approximation nodes).
approx	a numeric vector containing the approximated values at the approximation nodes y.
type	a character string giving the type of sum considered, i.e. one of ("standardized", "mean", "sum").
n	a positive integer giving the number of i.i.d. random variables in the sum.
approx.type	a character string giving the type of approximation.

Author(s)

Thorn Thaler

See Also

edgeworth, saddlepoint

---

cumulants *Cumulants Class For Saddlepoint Approximations*

Description

A cumulants object contains all the cumulant functions that are needed to calculate the saddlepoint approximation.

The predefined functions

- gammaCumulants,
- gaussianCumulants and
- inverseGaussianCumulants

compute the cumulant functions for the normal, gamma and inverse gaussian distribution, respectively.

Usage

```
cumulants(saddlef, cgf = NULL, kappa2f = NULL, rho3f = NULL,
          rho4f = NULL, cgf.deriv = NULL,
          domain = interval(-Inf, Inf), ...)

gammaCumulants(shape, scale)
gaussianCumulants(mu, sigma2)
inverseGaussianCumulants(lambda, nu)

## S3 method for class 'cumulants':
check(object, ...)
```

Arguments

saddlef	the saddlepoint function. Corresponds to the inverse of the first derivative of the cumulant generating function (CGF).
cgf, cgf.deriv	cgf is the cumulant generating function. If NULL (the default), it will be derived from cgf.deriv (the generic derivative function of the cgf).
kappa2f	the variance function. If NULL (the default), it will be derived from the function cgf.deriv.
rho3f, rho4f	the 3rd and the 4th standardized cumulant function, respectively. If NULL (the default), the functions will be derived from cgf.deriv if supplied. If neither the cumulants nor cgf.deriv are supplied, a warning will be displayed and a flag is set in the output. In this case, saddlepoint approximations cannot make use of the correction term (see saddlepoint for further details).
domain	an object of type interval giving the domain of the random variable. Will be needed to calculate the normalizing factor. See interval for further information.
...	additional parameters to be passed to the cumulant functions, respectively function check. See section 'Details' for further information.
shape, scale	shape and scale parameter for the gamma distribution.
mu, sigma2	mean and variance parameter for the normal distribution.
lambda, nu	parameters for the inverse Gaussian distribution.
object	an object to be tested whether or not it meets the formal requirements.

Details

Basically, there are two ways to specify the cumulant functions using cumulants. The first one is to specify each of the following functions seperately:

- cgf
- kappa2f
- rho3f
- rho4f

Since the functions may (and probably will) depend on some additional parameters, it is necessary to include these parameters in the respective argument lists. Thus, these additional parameters must be passed to `cumulants` as *named* parameters as well. To be more specific, if one of the above functions has an extra parameter `z`, say, the particular value of `z` must be passed to the function `cumulants` as well (see the example). In any case, the first argument of the cumulant functions must be the value at which the particular function will be evaluated.

The other way to specify the cumulant functions is to specify the generic derivative of the CGF `cgf.deriv`. Its first argument must be the order of the derivative and its second the value at which it should be evaluated, followed by supplementary arguments. `cgf.deriv` must be capable to return the CGF itself, which corresponds to the zeroth derivative.

The function `cumulants` performs a basic check to test if all needed additional parameters are supplied and displays a warning if there are extra arguments in the cumulant functions, which are not specified.

The generic function `check` for the class `cumulants` tests if

- an object has the same fields as an `cumulants` object and
- the cumulant functions are properly vectorized, i.e. if they return a vector whenever the argument is a vector.

Value

`cumulants` returns an object of class `cumulants` containing the following components:

`K`	the cumulant function.
`mu.inv`	the saddlepoint function.
`kappa2`	the variance function.
`rho3, rho4`	the 3rd and the 4th standardized cumulant functions.
`domain`	an interval giving the domain of the random variable.
`extra.params`	extra parameter passed to `cumulants`, typically parameters of the underlying distribution.
`type`	character string equating either to “explicit” or “implicit” indicating whether the cumulant functions were passed explicitly or were derived from the generic derivative of the CGF.
`missing`	logical. If `TRUE`, the 3rd and/or the 4th cumulant function were not defined.

`gammaCumulants`, `gaussianCumulants` and `inverseGaussianCumulants` return a `cumulants` object representing the cumulant functions of the particular distribution.

Note

If it happens that one of the cumulant functions `f`, say, does not need any extra arguments while the others do, one have to define these extra arguments for `f` nonetheless. The reason is that `cumulants` passes any additional arguments to all defined cumulant functions and it would end up in an error, if a function is not capable of dealing with additional arguments.

Hence, it is good practice to define all cumulant functions for the same set of arguments, needed or not. An alternative is to add `...` to the argument list in order to absorb any additional arguments.

The functions must be capable of handling vector input properly.

Supplementary arguments *must not* be named similar to the arguments of `cumulants` (especially any abbreviation must be avoided), for the argument matching may match an argument (thought to be an extra argument for one of the cumulant function) to an argument of `cumulants`. The same problem may arise, if additional cumulant function parameters are not named.

Author(s)

Thorn Thaler

References

Reid, N. (1991). Approximations and Asymptotics. *Statistical Theory and Modelling*, London: Chapman and Hall.

See Also

`edgeworth`, `saddlepoint`

Examples

```
# Define cumulant functions for the normal distribution

saddlef <- function(x, mu, sigma2) (x-mu)/sigma2
cgf <- function(x, mu, sigma2) mu*x+sigma2*x^2/2

## Not run:

# cgf, saddlef, kappa2, rho3 and rho4 must have the same argument lists!
# Functions are _not_ properly vectorized!
kappa2 <- function(x, sigma2) sigma2
rho3 <- function(x) 0
rho4 <- function(x) 0

cc <- cumulants(saddlef, cgf, kappa2, rho3, rho4, mu=0, sigma2=1)

check(cc) # FALSE

## End(Not run)

kappa2 <- function(x, mu, sigma2)
    rep(sigma2, length(x))
rho3 <- function(x, mu, sigma2)   # or function(x, ...)
    rep(0, length(x))
rho4 <- function(x, mu, sigma2)   # or function(x, ...)
    rep(0, length(x))

cc <- cumulants(saddlef, cgf, kappa2, rho3, rho4, mu=0, sigma2=1)

cc$K(1:2)       # 0.5 2
cc$kappa2(1:2)  # 1 1
cc$mu.inv(1:2)  # 1 2
```

```
cc$rho3(1:2)    # 0 0
cc$rho4(1:2)    # 0 0

check(cc) # TRUE

# The same using the generic derivative of the cgf

K.deriv <- function(n, x, mu, sigma2) {
  if (n <= 2) {
    switch(n + 1,
           return(mu * x + sigma2 * x ^ 2 / 2), # n == 0
           return(mu + sigma2 * x),             # n == 1
           return(rep(sigma2, length(x))))      # n == 2
  } else {
    return(rep(0, length(x)))                   # n >= 3
  }
}

cc <- cumulants(saddlef, cgf.deriv=K.deriv, mu=0, sigma2=1)

cc$K(1:2)       # 0.5 2
cc$kappa2(1:2)  # 1 1
cc$mu.inv(1:2)  # 1 2
cc$rho3(1:2)    # 0 0
cc$rho4(1:2)    # 0 0

check(cc) # TRUE

# The same using a predefined function
cc <- gaussianCumulants(0, 1)

cc$K(1:2)       # 0.5 2
cc$kappa2(1:2)  # 1 1
cc$mu.inv(1:2)  # 1 2
cc$rho3(1:2)    # 0 0
cc$rho4(1:2)    # 0 0

check(cc) # TRUE
```

`edgeworth` *Edgeworth Approximation*

Description

Computes the Edgeworth expansion of either the standardized mean, the mean or the sum of i.i.d. random variables.

Usage

```
edgeworth(x, n, rho3, rho4, mu, sigma2, deg=3,
          type = c("standardized", "mean", "sum"))
```

Arguments

`x`	a numeric vector or array giving the values at which the approximation should be evaluated.
`n`	a positive integer giving the number of i.i.d. random variables in the sum.
`rho3`	a numeric value giving the standardized 3rd cumulant. May be missing if `deg <= 1`.
`rho4`	a numeric value giving the standardized 4th cumulant. May be missing if `deg <= 2`.
`mu`	a numeric value giving the mean. May be missing if `type = "standardized"`, since it is only needed for transformation purposes.
`sigma2`	a positive numeric value giving the variance. May be missing if `type= "standardized"`.
`deg`	an integer value giving the order of the approximation: • `deg=1`: corresponds to a normal approximation • `deg=2`: takes 3rd cumulant into account • `deg=3`: allows for the 4th cumulant as well. The default value is 3.
`type`	determines which sum should be approximated. Must be one of ("standardized", "mean", "sum"), representing the shifted and scaled sum, the weighted sum and the raw sum. Can be abbreviated.

Details

The Edgeworth approximation (EA) for the density of the standardized mean $Z = \frac{S_n - n\mu}{\sqrt{n\sigma^2}}$, where

- $S_n = Y_1 + \ldots + Y_n$ denotes the sum of i.i.d. random variables,
- μ denotes the expected value of Y_i,
- σ^2 denotes the variance of Y_i

is given by:

$$f_Z(s) = \varphi(s)[1 + \frac{\rho_3}{6\sqrt{n}} H_3(s) + \frac{\rho_4}{24n} H_4(s) + \frac{\rho_3^2}{72n} H_6(s)],$$

with φ denoting the density of the standard normal distribution and ρ_3 and ρ_4 denoting the 3rd and the 4th standardized cumulants of Y_i respectively. $H_n(x)$ denotes the nth Hermite polynomial (see `hermite` for details).

The EA for the mean and the sum can be obtained by applying the transformation theorem for densities. In this case, the expected value `mu` and the variance `sigma2` must be given to allow for an appropriate transformation.

Value

edgeworth returns an object of the class approximation. See approximation for further details.

Author(s)

Thorn Thaler

References

Reid, N. (1991). Approximations and Asymptotics. *Statistical Theory and Modelling*, London: Chapman and Hall.

See Also

approximation,hermite,saddlepoint

Examples

```
# Approximation of the mean of n iid Chi-squared(2) variables

n <- 10
df <- 2
mu <- df
sigma2 <- 2*df
rho3 <- sqrt(8/df)
rho4 <- 12/df
x <- seq(max(df-3*sqrt(2*df/n),0), df+3*sqrt(2*df/n), length=1000)
ea <- edgeworth(x, n, rho3, rho4, mu, sigma2, type="mean")
plot(ea, lwd=2)

# Mean of n Chi-squared(2) variables is n*Chi-squared(n*2) distributed
lines(x, n*dchisq(n*x, df=n*mu), col=2)
```

---

eql *The Extended Quasi-Likelihood Function*

Description

Computes the Extended Quasi Likelihood (EQL) function for a given set of parameters for a particular variance family.

Usage

```
eql(formula, param.space, family = powerVarianceFamily(),
    phi.method = c("pearson", "mean.dev"), include.model = TRUE,
    smooth.grid = 10, do.smooth = dim(family) == 1,
    verbose = 1, ...)
```

```
## S3 method for class 'eql':
plot(x, do.points = (dim(x) == 1 && sum(!x$is.smoothed) <= 20),
     do.ci = TRUE, alpha = 0.95, do.bw = TRUE,
     show.max = TRUE, ...)
```

Arguments

formula	an object of class formula (or one that can be coerced to that class): a symbolic description of the model to be used to determine the parameters of the variance function.
param.space	a list of parameters for which the EQL value should be evaluated. If provided as a named list, the names must equal the names of the parameters defined by the variance family.
family	an object of class varianceFamily giving a parameterized family of variance functions. See varianceFamily for further details.
phi.method	a character string giving the name of the method used to estimate the dispersion parameter ϕ. Must be one of ("pearson", "mean.dev") representing the estimation of ϕ by the mean Pearson's statistic or by the mean deviance, respectively.
include.model	logical. If TRUE (the default) the final model is included in the output.
x	an object of class eql.
do.smooth, smooth.grid	do.smooth is a logical value and smooth.grid is an integer value giving the number of nodes for the smoothing process. If do.smooth is TRUE, smoothing is carried out by cubic splines on an equidistant grid with an amount of nodes equals to smooth.grid between two adjacent EQL values. Smoothing is currently only available for one-dimensional variance families, i.e. families that depend only on one parameter.
verbose	the amount of feedback requested: '0' or FALSE means no feedback, '1' or TRUE means some feedback (the default), and '2' means to show all available feedback. For the default setting, a progress bar will be displayed to give a rough estimation of the remaining calculation time. Full feedback prints the EQL value for each parameter combination.
...	further arguments to be passed to the glm routine and the plotting routine, respectively.
do.points, show.max	logical. If do.points is TRUE, the computed EQL values are marked in the plot. If show.max is TRUE, the maximum of the EQL function is emphasized in the plot.
do.ci, alpha	do.ci is a logical value, if TRUE an α confidence interval (respectively confidence ellipsoid) is added to the plot.
do.bw	logical. If TRUE (the default) a "black and white" plot is produced, otherwise colours are used.

Details

The EQL function as defined by *Nelder and Pregibon* (see 'References') is given by:

$$Q_\theta^+(y,\mu) = -\frac{1}{2}\log[2\pi\phi V_\theta(y)] - \frac{1}{2\phi}D_\theta(y,\mu),$$

where $D_\theta()$ and $V_\theta()$ denote the deviance function and the variance function, respectively, determined by the particular choice of the variance family.

The goal is to maximize the EQL function over μ and the not necessarily one-dimensional space of parameters θ. The function `eql` takes a particular finite set of candidate parameters and computes the corresponding EQL value for each of these parameters and returns the maximum EQL value for the given set. That implies that the function is only capable of capturing local maxima. If the maximum occurs at the boundary of the set, the set of parameters may be badly chosen and one should consider a larger set with the found maximum as an interior point.

The `plot` function is an important tool to investigate the structure of the EQL function. Confidence intervals and confidence ellipsoids give an idea of plausible parameter values for the variance function. The contour plot used for two-dimensional variance families is generated using the package **lattice**, which in turn relies on so called `trellis` plots. Hence, for two-dimensional families the `plot` function does not only generate the plot, but also returns the plot object to allow for further modifications of the plot. This is not true for one-dimensional variance models, which are plotted using the R standard graphical engine.

For large parameter sets the computation may take a long time. If no feedback is chosen, the function seems to be hung up, because the function does not provide any textual feedback while computing. Hence, a minimal feedback (including a progress bar) should be chosen to have an idea of the remaining calculation time.

An explicitely given deviance function speeds up calculation. A rather large amount of the total calculation time is used to determine the numerical values of the integral in the deviance function.

Value

`eql` returns an object of class `eql`, which contains the following components:

`eql`	a numerical vector with the computed eql values for the given set of parameter values. For one-dimensional variance families (i.e. those families with only one parameter), a smoothing operation can be performed to obtain intermediate values.
`param`	a `data.frame` containing the values of the parameters at which the eql function was evaluated.
`eql.max`	the maximum value of the eql function in the considered range.
`param.max`	a `data.frame` containing the values of the parameters at which the maximum is obtained.
`dim`	an integer value giving the dimension of the parameters in the underlying variance family.
`smooth`	a logical value indicating whether a smoothing operation was performed.
`is.smoothed`	a vector of logical values of the same length as `eql` indicating if the particular EQL value was obtained by smoothing or was calculated directly.

smooth.grid	an integer value giving the number of points used in the smoothing process or NULL if no smoothing was performed.
model	if include.model is TRUE, the GLM for which the maximum EQL value was archieved, NULL otherwise.

Note

The EQL for variance functions with $V_\theta(0) = 0$ becomes infinite. Hence, if there are exact zeros in the data, one should provide a variance family, which do not equate to zero at the origin. *Nelder and Pregibon* propose some adjustment of $V(y)$ at the origin, which leads to a modified variance function.

The predefined families powerVarianceFamily and extBinomialVarianceFamily are, however, *not* capable of dealing with exact zeros, for there is no general mechanism to modify the variance function for all possible values of the particular variance family.

The confidence interval for one-dimensional variance families is not calculated exactly, but depends on the amount of EQL values available. Hence, if one is interested in a confidence interval, one should allow for smoothing.

The function eql does not use a direct maximization routine, but rather do a simple maximization over a finite set. Hence, all obtained values including confidence intervals and confidence ellipsoids have a "local flavour" and should not be regarded as global solutions.

The confidence bounds are determined rather empirically and do heavily depend on the amount of parameter values under consideration.

Author(s)

Thorn Thaler

References

Nelder, J.A. and Pregibon, D. (1987). An extended quasi-likelihood function. *Biometrika*, **74**, 221–232.

See Also

varianceFamily, glm

Examples

```
## Power Variance Family
# Data from Box and Cox (1964)
x <- (-1:1)
y <- c(674,370,292,338,266,210,170,118,90,1414,1198,634,1022,620,438,
       442,332,220,3636,3184,2000,1568,1070,566,1140,884,360)
yarn.raw <- data.frame(expand.grid(x3=x, x2=x, x1=x), cycles=y)
yarn <- data.frame(x1=yarn.raw$x1, x2=yarn.raw$x2, x3=yarn.raw$x3,
   cycles=yarn.raw$cycles)
attach(yarn)

ps.power <- list(theta=seq(1, 4, length = 20))
```

```
eq.power <- eql(cycles~x1+x2+x3, param.space=ps.power,
   family=powerVarianceFamily("log"), smooth.grid=500)
plot(eq.power)

## Not run:
## Extended Binomial Variance Family
# Data from McCullagh & Nelder: GLM, p. 329
# (zeros replaced by 'NA')

site <- rep(1:9, each=10)
variety <- rep(1:10, 9)
resp <- c(0.05,NA,NA,0.10,0.25,0.05,0.50,1.30,1.50,1.50,
   NA,0.05,0.05,0.30,0.75,0.30,3,7.50,1,12.70,1.25,1.25,
   2.50,16.60,2.50,2.50,NA,20,37.50,26.25,2.50,0.50,0.01,
   3,2.50,0.01,25,55,5,40,5.50,1,6,1.10,2.50,8,16.50,
   29.50,20,43.50,1,5,5,5,5,5,10,5,50,75,5,0.10,5,5,
   50,10,50,25,50,75,5,10,5,5,25,75,50,75,75,75,17.50,
   25,42.50,50,37.50,95,62.50,95,95,95) / 100

ps.binomial <- list(seq(1, 2.2, length=32), seq(1, 3, length=32))
eq.binomial <- eql(resp~site*variety, param.space=ps.binomial,
   family=extBinomialVarianceFamily())
plot(eq.binomial)
## End(Not run)
```

`hermite` *Hermite Polynomials*

Description

Computes the Hermite polynomial $H_n(x)$.

Usage

```
hermite(x, n, prob = TRUE)
```

Arguments

`x`	a numeric vector or array giving the values at which the Hermite polynomial should be evaluated.
`n`	an integer vector or array giving the degrees of the Hermite polynomials. If `length(x) != 1`, `n` must be either of the same length as `x` or a single value.
`prob`	logical. If `TRUE` (the default) the probabilistic version of the Hermite polynomial is evaluated, otherwise the physicists' Hermite polynomials are used. See the 'Details' section below for further information.

Details

The Hermite polynomials are given by:

- $H_{n+1}(x) = xH_n(x) - nH_{n-1}(x)$, with $H_0(x) = 1$ and $H_1(x) = x$, (Probabilists' version $H_n^{Pr}(x)$)
- $H_{n+1}(x) = 2xH_n(x) - 2nH_{n-1}(x)$, with $H_0(x) = 1$ and $H_1(x) = 2x$. (Physicists' version $H_n^{Ph}(x)$)

and the relationship between the two versions is given by

$$H_n^{Ph}(x) = 2^{n/2} H_n^{Pr}(\sqrt{2}x).$$

The term 'probabilistic' is motivated by the fact that in this case the Hermite polynomial $H_n(x)$ can be as well defined by

$$H_n(x) = (-1)^n \frac{1}{\varphi(x)} \varphi^{(n)}(x),$$

where $\varphi(x)$ denotes the density function of the standard normal distribution and $\varphi^{(k)}(x)$ denotes the kth derivative of $\varphi(x)$ with respect to x.

If the argument `n` is a vector it must be of the same length as the argument `x` or the length of the argument `x` must be equal to one. The Hermite polynomials are then evaluated either at x_i with degree n_i or at x with degree n_i, respectively.

Value

the Hermite polynomial (either the probabilists' or the physicists' version) evaluated at `x`.

Author(s)

Thorn Thaler

References

Fedoryuk, M.V. (2001). Hermite polynomials. *Encyclopaedia of Mathematics*, Kluwer Academic Publishers.

Examples

```
2^(3/2)*hermite(sqrt(2)*5, 3)     # = 940
hermite(5, 3, FALSE)              # = 940
hermite(2:4, 1:3)                 # H_1(2), H_2(3), H_3(4)
hermite(2:4, 2)                   # H_2(2), H_2(3), H_2(4)
hermite(2, 1:3)                   # H_1(2), H_2(2), H_3(2)
## Not run:
hermite(1:3, 1:4)                 # Error!
## End(Not run)
```

`saddlepoint` *Saddlepoint Approximation*

Description

Computes the (normalized) saddlepoint approximation of the mean of n i.i.d. random variables.

Usage

```
saddlepoint(x, n, cumulants, correct = TRUE, normalize = FALSE)
```

Arguments

`x`	a numeric vector or array with the values at which the approximation should be evaluated.
`n`	a positive integer giving the number of i.i.d. random variables in the sum.
`cumulants`	a `cumulants` object giving the cumulant functions and the saddlepoint function. See `cumulants` for further information.
`correct`	logical. If `TRUE` (the default) the correction term involving the 3rd and the 4th standardized cumulant functions is included.
`normalize`	logical. If `TRUE` the renormalized version of the saddlepoint approximation is calculated. The renormalized version does neither make use of the 3rd nor of the 4th cumulant function so setting `correct=TRUE` will result in a warning. The default is `FALSE`.

Details

The saddlepoint approximation (SA) for the density of the mean $Z = S_n/n$ of i.i.d. random variables Y_i with $S_n = \sum_{i=1}^{n} Y_i$ is given by:

$$f_Z(z) \approx c\sqrt{\frac{n}{2\pi K_Y''(s)}}\exp\{n[K_Y(s) - sz]\},$$

where c is an appropriatly chosen correction term, which is based on higher cumulants. The function $K_Y(\cdot)$ denotes the cumulant generating function and s denotes the *saddlepoint* which is the solution of the saddlepoint function:

$$K'(s) = z.$$

For the renormalized version of the SA one chooses c such that $f_Z(z)$ integrates to one, otherwise it includes the 3rd and the 4th standardized cumulant.

The saddlepoint approximation is an improved version of the Edgeworth approximation and makes use of 'exponential tilted' densities. The weakness of the Edgeworth method lies in the approximation in the tails of the density. Thus, the saddlepoint approximation embed the original density in the "conjugate exponential family" with parameter θ. The mean of the embeded density depends now on θ which allows for evaluating the Edgeworth approximation at the mean, where it is known to give reasonable results.

Value

`saddlepoint` returns an object of class `approximation`. See function `approximation` for further details.

Author(s)

Thorn Thaler

References

Reid, N. (1991). Approximations and Asymptotics. *Statistical Theory and Modelling*, London: Chapman and Hall.

See Also

`approximation`, `cumulants`, `edgeworth`

Examples

```
# Saddlepoint approximation for the density of the mean of n Gamma
# variables with shape=1 and scale=1
n <- 10
shape <- scale <- 1
x <- seq(0, 3, length=1000)
sp <- saddlepoint(x, n, gammaCumulants(shape, scale))
plot(sp, lwd=2)

# Mean of n Gamma(1,1) variables is n*Gamma(n,1) distributed
lines(x, n*dgamma(n*x, shape=n*shape, scale=scale), col=2)
```

`varianceFamily` *Variance Family Class For The EQL-Method*

Description

`varianceFamily` provides a class for a parameterized family of variance functions to be used with `eql`.

The predefined functions `powerVarianceFamily` and `extBinomialVarianceFamily` compute the variance family defined by the parametric variance functions $V_\theta(\mu) = \mu^\theta$ and $V_{k,l}(\mu) = \mu^k(1-\mu)^l$, respectively.

Usage

```
varianceFamily(varf, devf = NULL, link = "log", initf = NULL,
               validmuf = NULL, name = "default")

powerVarianceFamily(link = "log")
extBinomialVarianceFamily(link = "logit")
```

Arguments

varf	the parameterized variance function.
devf	the deviance function. If NULL (the default) it will be determined numerically from the variance function varf.
link	the link function.
initf	a function returning an object of class expression. The expression object should give a sequence of initializing commands for the glm routine such as setting the starting values. If NULL (the default), a very rudimentary initialize function is chosen, which may not be appropriate. See family for further details.
validmuf	a function giving TRUE if its argument is a valid value for μ and FALSE otherwise. If NULL (the default), all μ are supposed to be valid.
name	a character string giving the name of the variance family.

Details

The purpose of the function varianceFamily is to provide a convenient way to specify families of variance functions. An extended glm family object for a particular choice of a parameter vector can be obtained via the class member family.

The minimal specification for a varianceFamily object is the variance function $V_\theta(\mu)$ with θ describing the vector of family parameters. If not given explicitly, the deviance function is determined numerically.

The family parameter of powerVarianceFamily is 'theta', while the names of the parameters of extBinomialVarianceFamily are 'k' and 'l'.

Value

varianceFamily returns an object of class varianceFamily containing the following components:

family	a function which computes an extFamily object, which is an extension of the family object known from classical glm. extFamily inherits from class family and contains an additional field holding the value of the particular parameters at which the family was evaluated.
name	a character string giving the name of the variance family.
params	a list of the parameters of the variance family.
type.dev	a character string. Equals either "explicit" or "numerical" depending on how the deviance function was determined.

Note

Those arguments passed to varianceFunction that are functions, are supposed to accept the variance family's parameter as an argument. The idea is that any of these functions may give different results for different values of the family's parameters. Even if any of these functions do not depend on these parameters, they must be contained in the function's argument list.

Author(s)

Thorn Thaler

References

Nelder, J.A. and Pregibon, D. (1987). An Extended Quasi-Likelihood Function. *Biometrika*, **74**, 221–232.

See Also

family, eql

Examples

```
# The extended binomial variance family
# (the deviance is determined numerically)

# init does not depend on k and l but it must accept
# these parameters anyways
init <- function(k, l) {
  return(expression({
    mustart <- (weights * y + 0.5)/(weights + 1)
    n <- rep.int(1, nobs)}))
}
validmuf <- function(mu, k, l) {
  return(all(mu > 0) && all(mu < 1))
}
varf <- function(y, k, l)  y^k*(1-y)^l
suppressWarnings(vf <- varianceFamily(varf=varf, link="log", initf=init,
                        validmuf=validmuf,
                        name="Extended-Binomial-Family"))
vf$family(1,1) # corresponds to binomial()

y <- runif(10, 0, 1)
mu <- runif(10, 0, 1)

all.equal(vf$family(1,1)$dev.resids(y,mu,1),     # TRUE
          binomial()$dev.resids(y,mu,1))
```

Literaturverzeichnis

Aitken, A. C.: On least squares and linear combinations of observations. *Proceedings of the Royal Society of Edinburgh* **55** (1935), 42–48

Akaike, H.: Information theory and an extension of the maximum likelihood principle. In: Petrov, B. N. (Hrsg.) und Csaki, F. (Hrsg.): *Proceedings of the 2nd International Symposium on Information Theory*, 1973, 267–281

Barndorff-Nielsen, O. und Cox, D. R.: Edgeworth and saddle-point approximations with statistical applications. *Journal of the Royal Statistical Society. Series B (Methodological)* **41** (1979), 279–312

Barndorff-Nielsen, O. E. und Cox, D. R.: *Asymptotic Techniques for Use in Statistics.* Chapman and Hall, 1989

Berry, A. C.: The accuracy of the Gaussian approximation to the sum of independent variates. *Transactions of the American Mathematical Society* **49** (1941), 122–136

Blinnikov, S. und Moessner, R.: Expansions for nearly Gaussian distributions. *Astronomy and Astrophysics Supplement Series* **130** (1998), 193–205

Blough, D. K., Madden, C. W. und Hornbrook, M. C.: Modeling risk using generalized linear models. *Journal of Health Economics* **18** (1999), 153–171

Box, G. E. P. und Cox, D. R.: An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)* **26** (1964), 211–252

Brazzale, A. R., Davison, A. C. und Reid, N.: *Applied Asymptotics: Case Studies in Small-Sample Statistics.* Cambridge University Press, 2007

Candy, S. G.: Modelling catch and effort data using generalised linear models, the Tweedie distribution, random vessel effects and random stratum-by-year effects. *CCAMLR Science* **11** (2004), 59–80

Casella, G. und Berger, R. L.: *Statistical Inference.* Duxbury, 2002

Charnes, A., Frome, E. L. und Yu, P. L.: The equivalence of generalized least squares and maximum likelihood estimates in the exponential family. *Journal of the American Statistical Association* **71** (1976), 169–171

Curtiss, J. H.: A note on the theory of moment generating functions. *The Annals of Mathematical Statistics* **13** (1942), 430–433

Czado, C. und Munk, A.: Noncanonical links in generalized linear models - when is the effort justified? *Journal of Statistical Planning and Inference* **87** (2000), 317–345

Dalal, S. R., Fowlkes, E. B. und Hoadley, B.: Risk analysis of the space shuttle: pre-challenger prediction of failure. *Journal of the American Statistical Association* **84** (1989), 945–957

Daniels, H. E.: Saddlepoint approximations in statistics. *The Annals of Mathematical Statistics* **25** (1954), 631–650

Daniels, H. E.: Discussion of „The regression analysis of binary sequences“ by D. R. Cox. *Journal of the Royal Statistical Society. Series B (Methodological)* **20** (1958), 236–238

Daniels, H. E.: Exact saddlepoint approximations. *Biometrika* **67** (1980), 59–63

Davidian, M. und Carroll, R. J.: A note on extended quasi-likelihood. *Journal of the Royal Statistical Society. Series B (Methodological)* **50** (1988), 74–82

Davison, A. C.: *Statistical Models.* Cambridge University Press, 2003

Dunn, P. K.: *tweedie: Tweedie exponential family models.* 2007. – R package version 1.5.2

Efron, B.: Double exponential families and their use in generalized linear regression. *Journal of the American Statistical Association* **81** (1986a), 709–721

Efron, B.: How biased is the apparent error rate of a prediction rule? *Journal of the American Statistical Association* **81** (1986b), 461–470

Esscher, F.: On the probability function in the collective theory of risk. *Skandinavisk Aktuarietidskrift* **15** (1932), 175–195

Esseen, C. G.: Fourier analysis of distribution functions. A mathematical study of the Laplace-Gaussian law. *Acta Mathematica* **77** (1945), 1–125

Fahrmeir, L., Kneib, T. und Lang, S.: *Regression.* Springer, 2007

Firth, D.: Generalized linear models. In: Hinkley, D. V. (Hrsg.), Reid, N. (Hrsg.) und Snell, E. J. (Hrsg.): *Statistical Theory and Modelling.* Chapman and Hall, 1991

Friedl, H.: *Verallgemeinerte logistische Modelle in der Analyse von Zervix-Karzinomen*, Technische Universität Graz, Dissertation, 1991

Hayakawa, T.: The likelihood ratio criterion and the asymptotic expansion of its distribution. *Annals of the Institute of Statistical Mathematics* **29** (1977), 359–378

Hinde, J. und Demétrio, C. G. B.: Overdispersion: models and estimation. *Computational Statistics & Data Analysis* **27** (1998), 151–170

Hosmer, D. W., Jovanovic, B. und Lemeshow, S.: Best subsets logistic regression. *Biometrics* **45** (1989), 1265–1270

Hurn, M. W., Barker, N. W. und Magath, T. D.: The determination of prothrombin time following the administration of dicumarol with specific reference to thromboplastin. *Journal of Laboratory & Clinical Medicine* **30** (1945), 432–437

Hurvich, C. M. und Tsai, C. T.: Regression and time series model selection in small samples. *Biometrika* **76** (1989), 297–307

Hurvich, C. M. und Tsai, C.T.: Model selection for extended quasi-likelihood models in small samples. *Biometrics* **51** (1995), 1077–1084

Jensen, J. L.: *Saddlepoint Approximations.* Clarendon Press, 1995 (Oxford statistical science series)

Jørgensen, B.: Maximum likelihood estimation and large-sample inference for generalized linear and nonlinear regression models. *Biometrika* **70** (1983), 19–28

Jørgensen, B.: Exponential dispersion models. *Journal of the Royal Statistical Society. Series B (Methodological)* **49** (1987), 127–162

Kullback, S. und Leibler, R. A.: On information and sufficiency. *The Annals of Mathematical Statistics* **22** (1951), 79–86

Lee, Y. und Nelder, J. A.: The relationship between double-exponential families and extended quasi-likelihood families, with application to modelling Geissler's human sex ratio data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **49** (2000), 413–419

Lehmann, E. L. und Romano, J. P.: *Testing Statistical Hypotheses.* Springer, Berlin, 2004

Mallows, C. L.: Some comments on C_P. *Technometrics* **42** (2000), 87–94

McCullagh, P.: Quasi-likelihood functions. *The Annals of Statistics* **11** (1983), 59–67

McCullagh, P. und Nelder, J. A.: *Generalized Linear Models.* Chapman and Hall, 1989 (Monographs on Statistics and Applied Probability)

Murrell, P.: *R Graphics.* Chapman & Hall, 2005

Nelder, J. A.: A large class of models derived from generalized linear models. *Statistics in Medicine* **17** (1998), 2747–2753

Nelder, J. A. und Lee, Y.: Generalized linear models for the analysis of taguchi-type experiments. *Applied Stochastic Models and Data Analysis* (1991), 107–120

Nelder, J. A. und Lee, Y.: Likelihood, quasi-likelihood and pseudolikelihood: some comparisons. *Journal of the Royal Statistical Society. Series B (Methodological)* **54** (1992), 273–284

Nelder, J. A. und Lee, Y.: Joint modeling of mean and dispersion. *Technometrics* **40** (1998), 168–175

Nelder, J. A. und Pregibon, D.: An extended quasi-likelihood function. *Biometrika* **74** (1987), 221–232

Nelder, J. A. und Wedderburn, R. W. M.: Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)* **135** (1972), 370–384

Pierce, D. A. und Schafer, D. W.: Residuals in generalized linear models. *Journal of the American Statistical Association* **81** (1986), 977–986

Pregibon, D.: *Data analytic methods for generalized linear models*, University of Toronto, Dissertation, 1979

Pregibon, D.: Goodness of link tests for generalized linear models. *Applied Statistics* **29** (1980), 15–24

Preisser, J. S.: Quasi-likelihood analysis of patient satisfaction with medical care. *Health Services and Outcomes Research Methodology* **3** (2002), 233–245

Rawlings, J., Pantula, S. und Dickey, D.: *Applied Regression Analysis: A Research Tool.* Springer, Berlin, 2001

Reid, N.: Approximations and asymptotics. In: Hinkley, D. V. (Hrsg.), Reid, N. (Hrsg.) und Snell, E. J. (Hrsg.): *Statistical Theory and Modelling.* Chapman and Hall, 1991

Sarkar, D.: *lattice: Lattice Graphics.* 2008. – R package version 0.17-17

Schwarz, G.: Estimating the dimension of a model. *The Annals of Statistics* **6** (1978), 461–464

Smyth, G. K.: Generalized linear models with varying dispersion. *Journal of the Royal Statistical Society. Series B (Methodological)* **51** (1989), 47–60

Smyth, G. K. und Verbyla, A. P.: Adjusted likelihood methods for modelling dispersion in generalized linear models. *Environmetrics* **10** (1999), 695–709

Wedderburn, R. W. M.: Quasi-likelihood functions, generalized linear models, and the Gauss–Newton method. *Biometrika* **61** (1974), 439–447