Nichtparametrische Maximum-Likelihood-Schätzung bei Generalisierten Linearen Mischmodellen

Herwig Friedl
Institut für Statistik, Technische Universität Graz

Zusammenfassung: Die Arbeit untersucht algorithmische Aspekte des EM-Algorithmus in Generalisierten Linearen Mischmodellen mit unbekannter Effekt-Verteilung. Die nichtparametrische Maximum-Likelihood Schätzung entspricht der Aufnahme zusätzlicher Prädiktor-Parameter und kann durch eine künstliche Datenreplikation realisiert werden.

Abstract: Computational details of the EM algorithm in generalised linear mixed models with unknown mixing distribution are considered. Nonparametric maximum likelihood estimation corresponds to using additional parameters in the linear predictor and replicating the data.

Schlüsselwörter: Zufallseffekte, EM-Algorithmus, Gauß-Quadratur.

1 Einleitung

Modelle mit zufälligen Effekten stellen eine flexible Klasse dar, durch die Überdispersion wegen der speziellen Abhängigkeitsstruktur in den Variablen berücksichtigt werden kann. Sehr einfach erweist sich die Maximum-Likelihood-Schätzung (ML-Schätzung), falls die Responsevariablen und die zufälligen Komponenten einer Normalverteilung genügen. Deshalb zählen diese Modelle bereits zum erweiterten Standardangebot sehr vieler Softwarepakete.

Für normalverteilte Responses und Effekte oder für ein anderes konjugiertes Dichtepaar resultiert eine analytisch geschlossene Form der Likelihood-Funktion deren Integral direkt maximiert werden kann. Die Annahme normalverteilter Effekte ist für viele Anwendungen angebracht, jedoch impliziert dies auch, daß das Integrationsproblem nur für dazu konjugierte normalverteilte Responses analytisch lösbar ist. Bei anderen Verteilung ist diese Auswertung ein Problem. Oft werden numerische Verfahren eingesetzt oder die Likelihood-Funktion wird indirekt maximiert.

Bei Poisson- (Binomial-) Responses ist eine direkte Maximierung nur für loggamma- (beta-) Effekte möglich. Diese maßgeschneiderten Strukturen erscheinen sehr künstlich und sind in der Praxis schwer motivierbar. Ist es nicht möglich, konkrete Annahmen über die Verteilung der Effekte zu treffen, dann wäre es zielführend, die Parameter und die Verteilung zu schätzen. Einen Hinweis darauf findet man in Anderson und Hinde (1988), wo der iterative EM-Algorithmus von Dempster, Laird und Rubin (1977) als indirekte Methode für normalverteilte Mischungen von Poissonvariablen eingesetzt wird. Aitkin und Francis (1995) bieten GLIM-Makros an, welche die Schätzer für Response-Verteilungen aus der Exponentialfamilie bei unbekannter Effektverteilung berechnen. Anwendung dieser Technik zur Analyse von Überdispersion in Generalisierten Linearen Modellen (GLMs) sind in ANDERSON (1988) und AITKIN (1994, 1996A) gegeben.

Zwar gibt es ausgezeichnete Bücher, welche die verschiedenen Schätztechniken beschreiben und auch teilweise deren algorithmische Umsetzung diskutieren wie LONGFORD (1993), FAHRMEIR und TUTZ (1994) oder DIGGLE, LIANG und ZEGER (1995), jedoch sind die Verfahren wegen der fehlenden Software nur selten verwendbar. Bei GLMs mit Zufallseffekten fällt dieser Mangel besonders auf.

Es wird nun in die Klasse der GLMs mit zufälligen Effekten eingeführt und die nichtparametrische ML-Schätzung motiviert. Den Schwerpunkt bildet die Diskussion und die algorithmische Beschreibung dieser Schätzprozedur. Dazu wird zunächst das Lineare Mischmodell vorgestellt und auf einige Mitglieder dieser Modellklasse hingewiesen. Die Verallgemeinerung auf die Klasse der GLMs folgt in Abschnitt 3. Im vierten Abschnitt wird gezeigt, daß die Gauß-Quadratur wie auch die nichtparametrische ML-Schätzung eine Likelihood-Funktion ergibt, wie sie auch beim diskreten Mischmodell resultiert. Die Konstruktion eines Schätzers für den Standardfehler wird im fünften Abschnitt ausgeführt. Abschließend wird anhand der Analyse zweier Datensätze der Prozeß der Modellfindung klargemacht und Hinweise auf diagnostische Aspekte gegeben.

2 Lineare Mischmodelle

Hier liegen n Cluster von korrelierten Beobachtungen $y_i = (y_{i1}, \ldots, y_{in_i})^t$ vor. Diese Responsevektoren hängen linear von $p+1$ unbekannten aber festen Effekten $\beta = (\beta_1, \ldots, \beta_{p+1})^t$ und von n nicht beobachtbaren q-dimensionalen clusterspezifischen Zufallseffekten $b_i = (b_{i1}, \ldots, b_{iq})^t$ ab, also

$$y_{ij} = u_{ij}^t \beta + w_{ij}^t b_i + \epsilon_{ij}. \tag{1}$$

$u_{ij} = (u_{ij1}, \ldots, u_{ijp+1})^t$ und $w_{ij} = (w_{ij1}, \ldots, w_{ijq})^t$ sind Designvektoren, die man clusterweise zu den Matrizen $u_i = (u_{i1}, \ldots, u_{in_i})^t$ und $w_i = (w_{i1}, \ldots, w_{in_i})^t$ zusammenfaßt. Sehr oft ist w_i eine Teilmenge von u_i. Ferner enthalte u_i die Interceptspalte.

Die $\epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{in_i})^t$ sind n_i-dimensionale Fehler mit $\epsilon_i \stackrel{iid}{\sim} N(0, \sigma_\epsilon^2 I_{n_i})$ und die Effekte b_i variieren unabhängig von Cluster zu Cluster mit $E(b_i) = 0$ und unbekannter Varianz-Kovarianzmatrix $Var(b_i) = \Sigma_b > 0$. Für gewöhnlich wird angenommen, daß

$$b_i \stackrel{iid}{\sim} N(0, \Sigma_b) \tag{2}$$

und daß ϵ_{ij} und b_i vollständig unabhängig sind. Es resultiert das Lineare Modell

$$y_i = u_i \beta + \epsilon_i^*, \tag{3}$$

wobei unter (2)

$$\epsilon_i^* = w_i b_i + \epsilon_i \stackrel{iid}{\sim} N(0, \sigma_\epsilon^2 I_{n_i} + w_i \Sigma_b w_i^t) \tag{4}$$

gilt. Daraus ist ersichtlich, daß durch dieses Modell Überdispersion relativ zur Varianz der Fehlerterme ϵ_{ij} beschreibbar ist. Da hier auf ein Konditionieren bezüglich der b_i verzichtet wird, repräsentiert (3) zusammen mit (4) die marginale Version eines Regressionsmodells mit zufälligen Effekten.

Zufällige Konstanten: In manchen Situationen scheint die Annahme gerechtfertigt zu sein, daß zu jedem y_i eine individuelle, um einen festen Wert τ zufällig variierende Basisgröße (Intercept) τ_i gehört. Man betrachtet daher das Modell

$$y_{ij} = \tau_i + x_{ij}^t \gamma + \epsilon_{ij}$$

mit $E(\tau_i) = \tau$. Der Vektor $x_{ij} = (x_{ij1}, \ldots, x_{ijp})^t$ beinhaltet hierbei und in allen weiteren Beispielen nur die p erklärenden Kovariablen zur j-ten Beobachtung des i-ten Clusters und nicht den Interceptterm. Die dazugehörenden Steigungs-Parameter γ seien fest. Natürlich sind die Abweichungen zwischen dem Populationsmittel τ und den für die Cluster spezifischen Realisationen τ_i nicht beobachtbar. Diese Differenzen können als Effekte fehlender Variablen interpretiert werden. Mit

$$\beta = (\tau, \gamma^t)^t, \quad u_{ij} = (1, x_{ij}^t)^t, \quad w_{ij} = 1, \quad b_i = (\tau_i - \tau) \overset{iid}{\sim} N(0, \sigma_b^2)$$

ergibt sich ein Modell der Form (1) und (2). In der Literatur findet man für diese Modellklasse oft die Bezeichnung „Varianzkomponenten Modell“, da hier als Spezialfall von (4) die Varianzzerlegung $Var(y_i) = Var(\epsilon_i^*) = \sigma_\epsilon^2 I_{n_i} + \sigma_b^2 1_{n_i} 1_{n_i}^t$ betrachtet wird. Somit ist die Korrelation eines Beobachtungspaares in einem beliebigen Cluster i gegeben durch das Kovarianz-Varianz-Verhältnis $\sigma_b^2/(\sigma_b^2 + \sigma_\epsilon^2)$.

Zufällige Steigungen: Variierte zuvor nur der Interceptterm, so kann zusätzlich auch eine Heterogenitätsannahme für die Steigungs-Parameter getroffen werden, d.h.

$$y_{ij} = \tau_i + x_{ij}^t \gamma_i + \epsilon_{ij}.$$

Bezeichnet $\beta_i = (\tau_i, \gamma_i^t)^t$ den Vektor aller zufälligen Effekte im Modell, dann gilt

$$\beta = E(\beta_i) = (\tau, \gamma^t)^t, \quad u_{ij} = w_{ij} = (1, x_{ij}^t)^t, \quad b_i = (\beta_i - \beta) \overset{iid}{\sim} N(0, \Sigma_b).$$

Der feste Vektor β beschreibt die Erwartung der b_i. Ein Modell mit ausschließlich zufälligen Parametern wird „Zufalls-Koeffizienten Modell“ genannt.

Feste und zufällige Steigungen: Es seien die Vektoren $u_{ij} = (u_{ij}^{1t}, u_{ij}^{2t})^t$ derart partitioniert, daß sich die Komponenten aus u_{ij}^1 auf jene Terme aus $(1, x_{ij}^t)$ beziehen, die individuell variierende Parameter haben, und daß u_{ij}^2 die restlichen Größen mit festen Effekten beinhaltet. Ferner liege die entsprechende Partitionierung auch für $\beta_i = (\beta_i^{1t}, \beta^{2t})^t$ vor. Somit bezeichnet β_i^1 alle variierenden Effekte mit $E(\beta_i^1) = \beta^1$, $Var(\beta_i^1) = \Sigma_b$ und β^2 die festen Koeffizienten und es ist

$$\beta = (\beta^{1t}, \beta^{2t})^t, \quad u_{ij} = (u_{ij}^{1t}, u_{ij}^{2t})^t, \quad w_{ij} = u_{ij}^1, \quad b_i = (\beta_i^1 - \beta^1) \overset{iid}{\sim} N(0, \Sigma_b).$$

Da man hier feste und zufällige Parameter mischt, spricht man von einem „Linearen Mischmodell“.

3 Generalisierte Lineare Mischmodelle

Stammen die y_{ij} nicht aus einer Normalverteilung, so kann man allgemein keine Darstellung der Form (3) und (4) erhalten. Infolgedessen werden wir für diese Fälle eine Modellstruktur wie bei den GLMs (vgl. McCullagh und Nelder, 1989) annehmen. Diese Verallgemeinerung impliziert die Notwendigkeit, lineare Modelle mit Zufallseffekten über deren bedingte Verteilung zu definieren. Seien dazu für einen gegebenen Zufallseffekt b_i die Beobachtungen y_{ij} konditional unabhängige, normalverteilte Zufallsvariablen, d.h.

$$y_{ij}|b_i \stackrel{ind}{\sim} N(\mu_{ij}, \sigma_\epsilon^2), \tag{5}$$

wobei der bedingte Erwartungswert $\mu_{ij} = E(y_{ij}|b_i)$ dargestellt wird durch

$$\mu_{ij} = u_{ij}^t \beta + w_{ij}^t b_i. \tag{6}$$

Ferner muß getrennt davon eine Verteilungsannahme für die zufälligen Effekte getroffen werden. Entsprechend (2) wird vorerst wieder angenommen, daß

$$b_i \stackrel{iid}{\sim} N(0, \Sigma_b). \tag{7}$$

Das durch (5), (6) und (7) spezifizierte Modell ist äquivalent dem Modell (3) und (4).

Will man auch andere Response-Verteilungen in Betracht ziehen und wird der Erwartungswert nicht ausschließlich linear modelliert, so muß (5) und (6) verallgemeinert werden. Dazu sei $f(y_{ij}|b_i)$ die bedingte Dichte mit bedingter Erwartung

$$\mu_{ij} = E(y_{ij}|b_i) = h(\eta_{ij}) \quad \text{mit} \quad \eta_{ij} = u_{ij}^t \beta + w_{ij}^t b_i. \tag{8}$$

Hier bezeichnet $h(\cdot)$ die Inverse einer bekannten Linkfunktion und η den linearen Prädiktor. Die b_i seien unabhängige und ident verteilte Größen mit $E(b_i) = 0$ und mit unbekannter Varianz-Kovarianzmatrix $Var(b_i) = \Sigma_b > 0$. Sehr oft werden wiederum normalverteilte Effekte angenommen. Im allgemeinen sollte dafür aber eine Dichte oder Wahrscheinlichkeitsfunktion $g(b_i)$ vorausgesetzt werden, die eine Verteilungsfunktion mit Erwartung Null und Varianz Σ_b beschreibt.

4 Parameterschätzungen

Das Hauptinteresse besteht nun darin, die ML-Schätzer sämtlicher Parameter unter Einbeziehung der Verteilung der Zufallseffekte zu berechnen. Zu diesen unbekannten Parametern zählen vor allem die festen Effekte β im linearen Prädiktor (8) aber auch die Varianzparameter Σ_b in (7), oder allgemeiner die verteilungsspezifizierenden Größen der zufälligen Effekte. All diese Parameter seien im Vektor θ subsummiert.

Bezeichnet $f(y_i, b_i; \theta)$ die gemeinsame Dichte einer vollständigen Beobachtung (y_i, b_i), dann resultiert als marginale Log-Likelihood-Funktion

$$l(y_i; \theta) = \log f(y_i; \theta) = \log \int f(y_i, b_i; \theta) db_i. \tag{9}$$

Um den marginalen ML-Schätzer von θ zu bestimmen, wird für gewöhnlich die Funktion (9) direkt maximiert. Wünschenswert wäre es aber, θ so zu schätzen, daß dadurch die vollständige Log-Likelihood-Funktion $\log f(y_i, b_i; \theta)$ maximal wird. Da jedoch die zufälligen Effekte gar nicht beobachtbar sind, ist dies nicht unmittelbar möglich. Die bedingte Dichte

$$k(b_i|y_i; \theta) = \frac{f(y_i, b_i; \theta)}{f(y_i; \theta)}$$

beschreibt gerade den Zusammenhang zwischen der marginalen und der gemeinsamen Likelihood-Funktion und man erhält damit

$$\log f(y_i, b_i; \theta) = \log k(b_i|y_i; \theta) + \log f(y_i; \theta).$$

Im $(t+1)$-ten Schritt des iterativen EM-Algorithmus von Dempster, Laird und Rubin (1977) wird der bedingte Erwartungswert der gemeinsamen Log-Likelihood-Funktion bezüglich der bedingten Dichte $k(b_i|y_i; \theta^{(t)})$ berechnet (E-Schritt). Dazu verwendet man den Parameter $\theta^{(t)}$ des t-ten Iterationsschritts und erhält

$$\begin{aligned} E\Big(\log f(y_i, b_i; \theta)\Big|y_i; \theta^{(t)}\Big) &= E\Big(\log k(b_i|y_i; \theta)\Big|y_i; \theta^{(t)}\Big) + E\Big(\log f(y_i; \theta)\Big|y_i; \theta^{(t)}\Big) \\ \int \log f(y_i, b_i; \theta) k(b_i|y_i; \theta^{(t)}) db_i &= \int \log k(b_i|y_i; \theta) k(b_i|y_i; \theta^{(t)}) db_i \\ &\quad + \int \log f(y_i; \theta) k(b_i|y_i; \theta^{(t)}) db_i \\ Q(\theta|\theta^{(t)}) &= H(\theta|\theta^{(t)}) + l(y_i; \theta). \end{aligned} \tag{10}$$

Da für einen Bereich S mit $\int_S (k(\theta) - k(\theta^{(t)})) d\mu \geq 0$ die Abschätzung

$$\int_S k(\theta) \log k(\theta) d\mu \geq \int_S k(\theta) \log k(\theta^{(t)}) d\mu,$$

gilt (Jensen-Ungleichung), folgt

$$H(\theta|\theta) \geq H(\theta^{(t)}|\theta). \tag{11}$$

Sei nun $\theta^{(t+1)}$ jener Wert von θ, der $Q(\theta|\theta^{(t)})$ bei festem $\theta^{(t)}$ maximiert. Dann ist

$$Q(\theta^{(t+1)}|\theta^{(t)}) - Q(\theta^{(t)}|\theta^{(t)}) \geq 0$$

und für die Differenz der Log-Likelihood-Funktionen resultiert wegen (11)

$$l(y_i; \theta^{(t+1)}) - l(y_i; \theta^{(t)}) = Q(\theta^{(t+1)}|\theta^{(t)}) - Q(\theta^{(t)}|\theta^{(t)}) + H(\theta^{(t)}|\theta^{(t)}) - H(\theta^{(t+1)}|\theta^{(t)}) \geq 0.$$

Damit ist gezeigt, daß durch die Maximierung von $Q(\theta|\theta^{(t)})$ (M-Schritt) automatisch auch $l(y_i; \theta)$ maximiert wird oder zumindest konstant bleibt.

Nimmt man nun an, daß beim Generalisierten Linearen Mischmodell die Dichte $g(\cdot)$ der Zufallseffekte durch eine einfache Reparametrisierung von unbekannten Parametern befreit werden kann, so resultiert

$$f(y_i; \theta) = \int f(y_i, b_i; \theta) db_i = \int f(y_i|b_i; \theta) g(b_i) db_i \tag{12}$$

und

$$k(b_i|y_i;\theta) = \frac{f(y_i|b_i;\theta)g(b_i)}{f(y_i;\theta)}.$$

Im E-Schritt wird daher bei n unabhängigen Stichprobenelementen die Funktion

$$\begin{aligned} Q(\theta|\theta^{(t)}) &= \sum_{i=1}^{n} \int \log f(y_i, b_i;\theta) k(b_i|y_i;\theta^{(t)}) db_i \\ &= \sum_{i=1}^{n} \frac{1}{f(y_i;\theta^{(t)})} \int \log f(y_i, b_i;\theta) f(y_i|b_i;\theta^{(t)}) g(b_i) db_i \end{aligned} \tag{13}$$

berechnet und dann im M-Schritt bezüglich aller Komponenten von θ maximiert.

Noch immer können sowohl $f(y_i;\theta^{(t)})$ als auch das zweite Integral in (13) sehr unangenehm sein, da beide nur selten analytisch geschlossen darstellbar sind. Zwei mögliche Approximationsansätze werden nun diskutiert. Im ersten Fall nehmen wir an, daß $g(\cdot)$ der Standard-Normalverteilungsdichte $\phi(\cdot)$ entspricht und approximieren beide Integrale durch eine K-Punkt Gauß-Quadratur. Falls diese Verteilungsannahme nicht getroffen werden kann, wird im zweiten Ansatz die Dichte $g(\cdot)$ durch den nichtparametrischen ML-Schätzer $\hat{g}$ ersetzt und die dadurch resultierende Zielfunktion maximiert. Beide Vorgangsweisen werden im folgenden anhand unterschiedlicher Modelle näher diskutiert.

4.1 Überdispersionsmodelle

Mit dem einfachsten Fall eines Varianzkomponenten Modells, in dem keine gruppierten Beobachtungen vorliegen ($n_i = 1$), kann Überdispersion in GLMs erklärt werden. Ist die durch das Modell spezifizierte Varianz geringer als die beobachtete Variabilität der Daten, so könnte dies darin begründet sein, daß es unberücksichtigte variierende Größen gibt, welche nicht im Modell enthalten aber mit der Response Variablen assoziiert sind. Eine Möglichkeit, dieses Defizit zu korrigieren, stellt die Hinzunahme eines nichtbeobachtbaren skalaren Zufallseffektes als zusätzliche Varianzquelle in das Modell dar.

Sind die zufälligen Intercepts τ_i unabhängig normalverteilt mit Erwartung τ und Varianz σ_τ^2, dann erhält man ein Überdispersionsmodell der Form

$$\eta_i = \tau_i + x_i^t\gamma = \tau + x_i^t\gamma + \sigma_\tau z_i \qquad \text{mit} \quad z_i \overset{iid}{\sim} N(0,1). \tag{14}$$

Die Parameter τ und σ_τ gehen somit als zusätzliche Parameter in den Prädiktor ein. Während τ als globaler Interceptparameter interpretierbar ist, könnte σ_τ, falls die standardisierten Effekte z_i beobachtbar wären, wie eine weitere Komponente von γ behandelt werden. Der Vektor θ beinhaltet somit alle interessierenden Parameter $\beta = (\tau, \gamma^t)^t$ und σ_τ. Da im allgemeinen die marginale Dichte (12) in (13) nicht analytisch ausgewertet werden kann, verwendet Hinde (1982) die Gauß-Quadratur mit K Quadraturpunkten, um das Integral durch eine endliche Summe zu approximieren. Als Näherung erhält man für die i-te Beobachtung

$$f(y_i;\theta^{(t)}) = \int f(y_i|z_i;\theta^{(t)})\phi(z_i)dz_i \approx \sum_{k=1}^{K} f(y_i|z_k;\theta^{(t)})\pi_k. \tag{15}$$

Die nichtbeobachtbaren z_i werden hierbei durch eine Folge fester Stellen z_k ersetzt, welche so wie π_k speziellen Tabellen zu entnehmen sind. Bei einer polynomialen Funktion $f(y_i|z_i)$ mit maximalem Grad $2K-1$ liefert die K-Punkt Gauß-Quadratur das Integral exakt. Dies motiviert die Vorgangsweise, in der man durch Vergrößern von K die Approximation beliebig genau machen will.

Wendet man dieses Approximationsverfahren auch auf das zweite Integral in (13) an, so resultiert

$$\begin{aligned} Q(\theta|\theta^{(t)}) &\approx \sum_{i=1}^{n} \frac{\sum_{k=1}^{K} \log f(y_i, z_k;\theta) f(y_i|z_k;\theta^{(t)})\pi_k}{\sum_{k=1}^{K} f(y_i|z_k;\theta^{(t)})\pi_k} = \sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik} \log f(y_i, z_k;\theta) \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\Big(\log f(y_i|z_k;\theta) + \log \pi_k\Big) \end{aligned} \tag{16}$$

mit den Gewichten

$$w_{ik} = \frac{f(y_i|z_k;\theta^{(t)})\pi_k}{\sum_{j=1}^{K} f(y_i|z_j;\theta^{(t)})\pi_j}, \tag{17}$$

die in $\theta = \theta^{(t)}$ ausgewertet sind und daher in der folgenden Maximierung als feste Größen betrachtet werden. Diese Gewichte können auch aus der Sicht eines formalen Bayes-Modells interpretiert werden. Wegen

$$P(z_k|y_i) = \frac{P(z_k)P(y_i|z_k)}{\sum_{j=1}^{K} P(z_j)P(y_i|z_j)} = \frac{\pi_k f(y_i|z_k)}{\sum_{j=1}^{K} \pi_j f(y_i|z_j)} = w_{ik}$$

haben diese Gewichte folgende Bedeutung: Wähle die Komponente z_k zufällig mit Wahrscheinlichkeit π_k. Ziehe nun y_i aus dieser Komponente, also aus einer Verteilung mit Dichte $f(y_i|z_k)$. Unter gegebenen y_i ist die a-posteriori Wahrscheinlichkeit, daß die Komponente z_k gewählt wurde, gleich w_{ik}.

Da bei dieser Darstellung außer den w_{ik} auch die Terme π_k fest sind und jedes einzelne z_i in (14) durch dieselben K bekannten Größen z_k ersetzt wird, stellt die Maximierung von (16) eine mit (17) gewichtete ML-Schätzung bezüglich einer $n \times K$-dimensionalen Stichprobe dar. Diese erweiterte Stichprobe erhält man, wenn jeder Beobachtung y_i genau die K Prädiktoren

$$\eta_{ik} = \tau + x_i^t\gamma + \sigma_\tau z_k$$

zugeordnet werden. Im M-Schritt werden also die ML-Schätzer $\hat{\tau}$, $\hat{\gamma}$ und $\hat{\sigma}_\tau$ bezüglich der Response-Design-Struktur

y	w	τ	γ	σ_τ
y_1	w_{11}	1	$x_{11} \ldots x_{1p}$	z_1
$\vdots$	$\vdots$	$\vdots$	$\vdots \quad \vdots$	$\vdots$
y_n	w_{n1}	1	$x_{n1} \ldots x_{np}$	z_1
y_1	w_{12}	1	$x_{11} \ldots x_{1p}$	z_2
$\vdots$	$\vdots$	$\vdots$	$\vdots \quad \vdots$	$\vdots$
y_n	w_{n2}	1	$x_{n1} \ldots x_{np}$	z_2
$\vdots$	$\vdots$	$\vdots$	$\vdots \quad \vdots$	$\vdots$
y_1	w_{1K}	1	$x_{11} \ldots x_{1p}$	z_K
$\vdots$	$\vdots$	$\vdots$	$\vdots \quad \vdots$	$\vdots$
y_n	w_{nK}	1	$x_{n1} \ldots x_{np}$	z_K

berechnet. Im darauf folgenden E-Schritt werden nur noch die w_{ik} aktualisiert.

Da die Likelihood-Funktion durch die Form von $g(\cdot)$ mitbestimmt wird, können Fehlspezifikationen wesentliche Konsequenzen haben. Falls bezüglich der Effektverteilung keine parametrisierte Verteilungsannahme getroffen werden kann, sollte zumindest ein Verteilungsschätzer als Information in die Likelihood-Funktion eingehen. Entsprechend (14) verwendet man zunächst als Prädiktormodell

$$\eta_i = x_i^t\gamma + z_i \qquad \text{mit} \quad z_i \overset{iid}{\sim} G. \tag{18}$$

Jetzt ist $\theta = (\gamma, G)$ und die Verteilungsfunktion G soll gleichzeitig mit γ geschätzt werden. Kiefer und Wolfowitz (1956) zeigten die Konsistenz dieser Schätzung. Den Abhandlungen von Laird (1978) und Lindsay (1983) ist weiters zu entnehmen, daß der nichtparametrische ML-Schätzer $\hat{g}$ im Falle von Mischverteilungen eine diskrete Wahrscheinlichkeitsfunktion darstellt, welche auf einer endlichen Anzahl K von Massepunkten z_k mit Wahrscheinlichkeitsmassen π_k definiert ist. Für einen fest vorgegebenen Wert K liefert das zur Gauß-Quadratur analoge Vorgehen wiederum (16) als zu maximierende Funktion mit Gewichten (17). Jedoch sind nun außer γ auch z_k und π_k unbekannt. Die Schätzung der z_k ist üblicherweise sehr rechenintensiv. Wood und Hinde (1987) zeigten aber, daß die z_k für festes K wie zuvor bestimmt werden können.

Mit der Bedingung $\sum_k \pi_k = 1$ folgt

$$\frac{\partial}{\partial \pi_k}\Big(Q(\theta|\theta^{(t)}) - \lambda\Big(\sum_k \pi_k - 1\Big)\Big) = \frac{1}{\pi_k}\sum_i w_{ik} - \lambda,$$

also $\hat{\pi}_k = \sum_i \hat{w}_{ik}/\hat{\lambda}$. Wegen $\sum_k w_{ik} = 1$ resultiert $\sum_k \hat{\pi}_k = \sum_i \sum_k \hat{w}_{ik}/\hat{\lambda} = n/\hat{\lambda}$ und somit $\hat{\lambda} = n$. Die Maximierung von (16) liefert daher unter der obigen Bedingung den expliziten Schätzer

$$\hat{\pi}_k = \frac{1}{n}\sum_{i=1}^{n} \hat{w}_{ik}. \tag{19}$$

Weiters kann für die Maximierung von Q bezüglich γ und z_k der lineare Prädiktor formuliert werden als

$$\eta_{ik} = x_i^t \gamma + z_k,$$

wobei jetzt die z_k zu schätzen sind. Mit einem K-stufigen Faktor entspricht dies dem Modell

$$\eta_{ik} = x_i^t \gamma + z_1 \cdot 0 + \ldots + z_{k-1} \cdot 0 + z_k \cdot 1 + z_{k+1} \cdot 0 + \ldots + z_K \cdot 0.$$

Die unbekannten Stellen z_k können also als Parameter von K zusätzlichen Dummy-Variablen interpretiert werden. Die erweiterte Response-Design-Struktur ist somit gleich

$$\begin{array}{cc|ccc cccc}
y & w & & \gamma & & & z & & \\
\hline
y_1 & w_{11} & x_{11} & \ldots & x_{1p} & 1 & 0 & \ldots & 0 \\
\vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\
y_n & w_{n1} & x_{n1} & \ldots & x_{np} & 1 & 0 & \ldots & 0 \\
y_1 & w_{12} & x_{11} & \ldots & x_{1p} & 0 & 1 & \ldots & 0 \\
\vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\
y_n & w_{n2} & x_{n1} & \ldots & x_{np} & 0 & 1 & \ldots & 0 \\
\vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\
y_1 & w_{1K} & x_{11} & \ldots & x_{1p} & 0 & 0 & \ldots & 1 \\
\vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\
y_n & w_{nK} & x_{n1} & \ldots & x_{np} & 0 & 0 & \ldots & 1
\end{array}$$

Im folgenden E-Schritt werden wiederum nur die Gewichte w_{ik} mit den aktuellen Schätzungen von γ, z_k und π_k neu berechnet.

4.2 Zufälliger Steigungsparameter

Im Gegensatz zu den Überdispersionsmodellen soll jetzt die Annahme getroffen werden, daß der Intercept fest ist und genau ein Steigungsparameter über die Population variiert. Sei dies die zur j-ten erklärenden Variablen x_{ij} gehörende Komponente γ_j. Dann ist das Modell unter Normalverteilungsannahme in reparametrisierter Form gegeben durch

$$\eta_i = \tau + x_i^t \gamma + \sigma_{\gamma_j} x_{ij} z_i \qquad \text{mit} \quad z_i \overset{iid}{\sim} N(0,1). \tag{20}$$

Ein zufälliger Steigungsparameter kann daher als Wechselwirkung der j-ten Variablen mit der standardisierten Form des Zufallseffektes interpretiert werden. Die Gauß-Quadratur liefert wiederum (16) als Zielfunktion und die Gewichte (17). Für die Maximierung von Q betrachtet man somit die linearen Prädiktoren

$$\eta_{ik} = \tau + x_i^t \gamma + \sigma_{\gamma_j} x_{ij} z_k.$$

Die dazu korrespondierende erweiterte Response-Design-Struktur ist gegeben durch

y	w	τ	γ	σ_{γ_j}
y_1	w_{11}	1	$x_{11} \ldots x_{1p}$	$x_{1j}z_1$
$\vdots$	$\vdots$	$\vdots$	$\vdots \quad \vdots$	$\vdots$
y_n	w_{n1}	1	$x_{n1} \ldots x_{np}$	$x_{nj}z_1$
y_1	w_{12}	1	$x_{11} \ldots x_{1p}$	$x_{1j}z_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots \quad \vdots$	$\vdots$
y_n	w_{n2}	1	$x_{n1} \ldots x_{np}$	$x_{nj}z_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots \quad \vdots$	$\vdots$
y_1	w_{1K}	1	$x_{11} \ldots x_{1p}$	$x_{1j}z_K$
$\vdots$	$\vdots$	$\vdots$	$\vdots \quad \vdots$	$\vdots$
y_n	w_{nK}	1	$x_{n1} \ldots x_{np}$	$x_{nj}z_K$

Falls bezüglich der Effekte keine Normalverteilungsannahme getroffen werden kann, betrachtet man allgemeiner

$$\eta_i = x_i^t\gamma + x_{ij}z_i \qquad \text{mit} \quad z_i \overset{iid}{\sim} G. \tag{21}$$

Die obige Überlegung ergibt bei der Verwendung eines nichtparametrischen ML-Schätzers

$$\eta_{ik} = x_i^t\gamma + z_1 \cdot 0 + \ldots + z_{k-1} \cdot 0 + z_k \cdot x_{ij} + z_{k+1} \cdot 0 + \ldots + z_K \cdot 0$$

und führt zur Response-Design-Struktur

y	w	γ	z			
y_1	w_{11}	$x_{11} \ldots x_{1p}$	x_{1j}	0	$\ldots$	0
$\vdots$	$\vdots$	$\vdots \quad \vdots$	$\vdots$	$\vdots$		$\vdots$
y_n	w_{n1}	$x_{n1} \ldots x_{np}$	x_{nj}	0	$\ldots$	0
y_1	w_{12}	$x_{11} \ldots x_{1p}$	0	x_{1j}	$\ldots$	0
$\vdots$	$\vdots$	$\vdots \quad \vdots$	$\vdots$	$\vdots$		$\vdots$
y_n	w_{n2}	$x_{n1} \ldots x_{np}$	0	x_{nj}	$\ldots$	0
$\vdots$	$\vdots$	$\vdots \quad \vdots$	$\vdots$	$\vdots$		$\vdots$
y_1	w_{1K}	$x_{11} \ldots x_{1p}$	0	0	$\ldots$	x_{1j}
$\vdots$	$\vdots$	$\vdots \quad \vdots$	$\vdots$	$\vdots$		$\vdots$
y_n	w_{nK}	$x_{n1} \ldots x_{np}$	0	0	$\ldots$	x_{nj}

Wiederum entsprechen die Parameterschätzer der Faktorstufen genau den gesuchten Massestellen $\hat{z}_k$.

4.3 Höherdimensionale Zufallseffekte

Bis jetzt wurden ausschließlich Modelle betrachtet, in denen nur ein skalarer Effekt variiert. Möchte man Intercept und Steigung oder mehrere Steigungsparameter als

Zufallseffekte in das Modell aufnehmen, so können die obigen Ansätze nicht unmittelbar übernommen werden. Bei der Analyse wird sich zeigen, daß eine K-fache Vervielfältigung der Daten nicht immer sinnvolle Approximationen der Zielfunktion Q liefert. Stellvertretend für Modelle mit einem multivariaten Zufallseffekt wird hier der zweidimensionale Fall diskutiert. Eine Verallgemeinerung auf eine höhere Dimension ist einfach möglich.

Für zwei unabhängige normalverteilte Zufallseffekte $z_i = (z_{i1}, z_{i2})$ mit Korrelationskoeffizienten $\rho_z = 0$ ergibt die Gauß-Quadratur für eine Beobachtung

$$\begin{aligned} f(y_i;\theta^{(t)}) &= \int\int f(y_i|z_i;\theta^{(t)})\phi(z_i)dz_i \approx \int \sum_{k=1}^{K_2} f(y_i|(z_{i1}, z_k);\theta^{(t)})\pi_k\phi(z_{i1})dz_{i1} \\ &\approx \sum_{k=1}^{K_1}\sum_{l=1}^{K_2} f(y_i|(z_k, z_l);\theta^{(t)})\pi_k\pi_l. \end{aligned}$$

Hier wird die bivariate Standardnormalverteilungsdichte zweier unabhängiger Effekte über ein Raster von $K_1 \times K_2$ Massepunkten (z_k, z_l) beschrieben, die auch schon zuvor bei den Modellen mit eindimensionalen Effekten verwendet wurden. Somit haben wir

$$Q(\theta|\theta^{(t)}) \approx \sum_{i=1}^{n}\sum_{k=1}^{K_1}\sum_{l=1}^{K_2} w_{ikl}\Big(\log f(y_i|(z_k, z_l);\theta^{(t)}) + \log \pi_k + \log \pi_l\Big)$$

mit den Gewichten

$$w_{ikl} = \frac{f(y_i|(z_k, z_l);\theta^{(t)})\pi_k\pi_l}{\sum_{j=1}^{K_1}\sum_{m=1}^{K_2} f(y_i|(z_j, z_m);\theta^{(t)})\pi_j\pi_m}.$$

Im Unterschied zu den Modellen mit einem zufälligen Effekt hat jetzt der erweiterte Datensatz die Dimension $n \times K_1 \times K_2$. Wie man sieht, wächst hier die Anzahl der Massestellen exponentiell mit der Dimension des Zufallseffektes.

Verteilt man die Massepunkte nur auf der ersten Mediane, so folgt die Approximation

$$f(y_i;\theta^{(t)}) \approx \sum_{k=1}^{K} f(y_i|(z_k, z_k);\theta^{(t)})\pi_k,$$

was einer exakt linearen Abhängigkeit der beiden Effekte entspricht, also nur für $\rho_z = 1$ brauchbare Ergebnisse liefert. Jedoch können wiederum dieselben Werte $\{z_k\}$ mit den gleichen Massen $\{\pi_k\}$ verwendet werden.

In der Regel sind jedoch gerade Modelle mit abhängigen Effekten interessant. Für diese Situation ist die nichtparametrischen ML-Schätzung günstig, wo die K Massestellen (z_{1k}, z_{2k}) geschätzt und dadurch automatisch optimal positioniert werden. Nehmen wir wiederum an, daß der Schätzer $\hat{g}$ im zweidimensionalen Fall aus der K-elementigen Menge $\{(\hat{z}_{1k}, \hat{z}_{2k})\}$ mit den entsprechenden Massen $\{\hat{\pi}_k\}$ gebildet wird. Dann folgt

$$f(y_i;\theta^{(t)}) \approx \int\int f(y_i|z_i;\theta^{(t)})\hat{g}(z_i)dz_i = \sum_{k=1}^{K} f(y_i|(z_{1k}, z_{2k});\theta^{(t)})\pi_k$$

und der bereits diskutierte Schätz-Algorithmus kann wiederum ohne Modifikation übernommen werden. Da hier eine höherdimensionale Dichte geschätzt wird, sollte auch K entsprechend groß gewählt sein.

Stellen beispielsweise in einem Modell sowohl der Intercept als auch der zur einzigen erklärenden Variablen x_i gehörende Steigungsparameter zufällige Effekte (aus einer unbekannten bivariaten Verteilung G_2) dar, d.h.

$$\eta_i = z_{i1} + x_i z_{i2} \qquad \text{mit} \quad z_i = (z_{i1}, z_{i1}) \overset{iid}{\sim} G_2, \tag{22}$$

so ergibt sich als Approximation

$$\begin{aligned}\eta_{ik} = {} & z_{11} \cdot 0 + \ldots + z_{1,k-1} \cdot 0 + z_{1k} \cdot 1 + z_{1,k+1} \cdot 0 + \ldots + z_{1K} \cdot 0 \\ & + z_{21} \cdot 0 + \ldots + z_{2,k-1} \cdot 0 + z_{2k} \cdot x_i + z_{2,k+1} \cdot 0 + \ldots + z_{2k} \cdot 0.\end{aligned}$$

Man betrachtet daher das Design

$$\begin{array}{cc|cccccccc}
y & w & & z_1 & & & & z_2 & & \\
\hline
y_1 & w_{11} & 1 & 0 & \ldots & 0 & x_1 & 0 & \ldots & 0 \\
\vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\
y_n & w_{n1} & 1 & 0 & \ldots & 0 & x_n & 0 & \ldots & 0 \\
y_1 & w_{11} & 0 & 1 & \ldots & 0 & 0 & x_1 & \ldots & 0 \\
\vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\
y_n & w_{n1} & 0 & 1 & \ldots & 0 & 0 & x_n & \ldots & 0 \\
\vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\
y_1 & w_{11} & 0 & 0 & \ldots & 1 & 0 & 0 & \ldots & x_1 \\
\vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\
y_n & w_{n1} & 0 & 0 & \ldots & 1 & 0 & 0 & \ldots & x_n
\end{array}$$

welches durch zwei Faktoren mit jeweils K Stufen aufgebaut ist. Der zweite Faktor beschreibt den zufälligen Steigungsparameter und geht nur als Wechselwirkung mit x in dieses approximative Modell ein. Wie zuvor beschreiben die $2K$ Parameter gerade die unbekannten Massestellen (z_{1k}, z_{2k}) und können als Parameter des linearen Prädiktors aufgefaßt werden.

4.4 Varianzkomponentenmodelle

Die Ergebnisse bei den Überdispersionsmodellen ($n_i = 1$) können fast unmittelbar auf die Klasse der Varianzkomponentenmodelle mit Clustergrößen $n_i \geq 1$ verallgemeinert werden. Generell hat hierbei der lineare Prädiktor die Form

$$\eta_{ij} = \tau + x_{ij}^t \gamma + \sigma_\tau z_i \qquad \text{mit} \quad z_i \overset{iid}{\sim} N(0,1)$$

oder

$$\eta_{ij} = x_{ij}^t \gamma + z_i \qquad \text{mit} \quad z_i \overset{iid}{\sim} G.$$

Dies bedeutet für beide Situationen, daß alle Beobachtungen eines Clusters den gleichen Zufallseffekt teilen. Bezeichnet beispielsweise y_{ij} die Beobachtung des j-ten Patienten im i-ten Krankenhaus, so sind unter diesem Modell die Beobachtungen an den Patienten desselben Krankenhauses korreliert, während Patienten unterschiedlicher Krankenhäuser unkorrelierte Beobachtungen liefern. Der zufällige Effekt ist also nur für ein Krankenhaus spezifisch und nicht für die einzelnen Patienten.

Um für dieses Modell die Funktion Q zu berechnen, verwendet man die Resultate des Überdispersionsmodells und trifft die zusätzliche Annahme der bedingten Unabhängigkeit der Beobachtungen innerhalb eines Clusters, also

$$f(y_i|z_i) = \prod_{j=1}^{n_i} f(y_{ij}|z_i).$$

Damit erhält man sowohl bei der Gauß-Quadratur als auch bei Anwendung des nichtparametrischen ML-Schätzers wiederum approximativ

$$Q(\theta|\theta^{(t)}) \approx \sum_{i=1}^{n} \frac{\sum_{k=1}^{K} \log f(y_i, z_k; \theta) f(y_i|z_k; \theta^{(t)})\pi_k}{\sum_{k=1}^{K} f(y_i|z_k; \theta^{(t)})\pi_k},$$

wofür jetzt aber gilt

$$Q(\theta|\theta^{(t)}) \approx \sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik} \log f(y_i, z_k; \theta) = \sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{j=1}^{n_i} w_{ik}\Big(\log f(y_{ij}|z_k; \theta) + \log \pi_k\Big) \quad (23)$$

mit den Gewichten

$$w_{ik} = \frac{\pi_k \prod_{j=1}^{n_i} f(y_{ij}|z_k; \theta^{(t)})}{\sum_{l=1}^{K} \prod_{j=1}^{n_i} f(y_{ij}|z_l; \theta^{(t)})\pi_l}. \quad (24)$$

Der einzige Unterschied zu (16) liegt in der zusätzlichen Summation über alle Elemente eines jeden Clusters. Weiters sei hier noch bemerkt, daß die Gewichte für alle Beobachtungen eines Clusters dieselben sind und daher den Termen in (17) entsprechen.

Sind die π_k unbekannt, so folgt mit der Bedingung $\sum_k \pi_k = 1$

$$\frac{\partial}{\partial \pi_k}\Big(Q(\theta|\theta^{(t)}) - \lambda\Big(\sum_k \pi_k - 1\Big)\Big) = \frac{1}{\pi_k}\sum_i n_i w_{ik} - \lambda,$$

also $\hat{\pi}_k = \sum_i n_i \hat{w}_{ik}/\hat{\lambda}$. Mit $\sum_k w_{ik} = 1$ resultiert $\sum_k \hat{\pi}_k = \sum_i n_i/\hat{\lambda}$. Daher ergibt die Maximierung von (23) den Schätzer

$$\hat{\pi}_k = \frac{\sum_{i=1}^{n} n_i \hat{w}_{ik}}{\sum_{i=1}^{n} n_i}, \quad (25)$$

was wiederum dem Ergebnis (19) entspricht.

Im E-Schritt werden daher die für die Kliniken spezifischen Gewichte auf dem i-Niveau berechnet, die dort für alle n_i Beobachtungen gleich sind. Im folgenden M-Schritt wendet man dieselben Gewichte w_{ik} auf alle Daten des j-Niveaus an.

Um (23) zu maximieren, führt man eine mit den festen Größen (24) gewichtete ML-Schätzung bezüglich einer erweiterten Response-Design-Struktur an, welche im wesentlichen der Struktur in den Überdispersionsmodellen entspricht. Da jetzt n_i Replikationen innerhalb der i-ten Gruppe vorliegen, wird beispielsweise jene Zeile, die der ersten Beobachtung für $k = 2$ entspricht, bei der Gauß-Quadratur ersetzt durch die Matrix

$$\begin{array}{cc|cccc} y & w & \tau & & \gamma & \sigma_\tau \\ \hline y_{11} & w_{12} & 1 & x_{111} & \dots\ x_{11p} & \tau_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ y_{1n_1} & w_{12} & 1 & x_{1n_11} & \dots\ x_{1n_1p} & \tau_2 \end{array}$$

oder im Falle einer unbekannten Mischverteilung durch

$$\begin{array}{cc|ccc} y & w & & \gamma & z \\ \hline y_{11} & w_{12} & x_{111} & \dots\ x_{11p} & 0\ 1 \dots 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots\ \vdots \quad \vdots \\ y_{1n_1} & w_{12} & x_{1n_11} & \dots\ x_{1n_1p} & 0\ 1 \dots 0 \end{array}$$

Eine Anwendung der Gauß-Quadratur bei binären Responsevariablen unter einem Varianzkomponentenmodell ist in Anderson und Aitkin (1985) gegeben.

5 Deviance und Parametervarianz

Die Deviance ist ein Maß für die Güte der Modellanpassung und kann hier analog wie bei GLMs definiert werden. Bis jetzt wurden alle unbekannten Parameter zum Vektor θ zusammengefaßt. Andererseits gilt für die im vorigen Abschnitt diskutierten Modelle, daß die Komponenten von θ unter Verwendung der Gauß-Quadratur oder der nichtparametrischen ML-Schätzung nur im Prädiktor $\eta = h(\mu)$ enthalten sind. Deshalb kann für diese Modelle die approximative marginale Likelihood-Funktion in Termen von μ betrachtet werden, d.h.

$$L_K(y;\mu) = \prod_{i=1}^{n} \sum_{k=1}^{K} f(y_i|\mu_{ik})\pi_k, \qquad \text{mit } \mu_{ik} = h(\eta_{ik}).$$

Das saturierte Modell erhält man in $\mu_{ik} = y_i$. Dazu korrespondiert der maximale Likelihood-Wert

$$L_K(y;y) = \prod_{i=1}^{n} \sum_{k=1}^{K} f(y_i|\mu_{ik} = y_i)\pi_k = \prod_{i=1}^{n} f(y_i|y_i) \sum_{k=1}^{K} \pi_k = \prod_{i=1}^{n} f(y_i|y_i) = L(y;y),$$

welcher unabhängig von den Parametern und allen erklärenden Größen ist. Mit $l_K(y;\mu) = \log L_K(y;\mu)$ ist die skalierte Deviance proportional dem durch das Modell nicht erklärten Anteil der Likelihood-Funktion. Als Approximation verwenden wir die Differenz

$$-2\Big(l_K(y;\mu) - l(y;y)\Big).$$

Falls die bedingte Verteilung der Beobachtungen aus der Exponentialfamilie stammt (siehe McCullagh und Nelder (1989)), so ist

$$f(y_i|\mu_i) = \exp\left(\frac{y_i\psi_i - b(\psi_i)}{a_i(\sigma^2)} - c(y_i,\sigma^2)\right)$$

mit der Kumulantenfunktion $b(.)$ und dem kanonischen Parameter ψ. Wird eine kanonische Linkfunktion verwendet, so erhält man ein lineares Modell für den kanonischen Parameter $\psi = \eta = h^{-1}(\mu)$. Mit $a_i(\sigma^2) = \sigma^2/v_i$ folgt

$$l(y;y) = -\sum_{i=1}^{n} c(y_i,\sigma^2) + \sum_{i=1}^{n} \frac{y_i h^{-1}(y_i) - b(h^{-1}(y_i))}{\sigma^2/v_i}$$

sowie

$$\begin{aligned} l_K(y;\mu) &= \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k \exp\left(\frac{y_i\psi_{ik} - b(\psi_{ik})}{\sigma^2/v_i} - c(y_i,\sigma^2)\right) \\ &= -\sum_{i=1}^{n} c(y_i,\sigma^2) + \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k \exp\left(\frac{y_i h^{-1}(\mu_{ik}) - b(h^{-1}(\mu_{ik}))}{\sigma^2/v_i}\right). \end{aligned}$$

In der Berechnung der skalierten Deviance verschwindet daher die Normierungsfunktion $c(\cdot)$ und es ergibt sich

$$\begin{aligned} \frac{1}{\sigma^2} D_K(y;\mu) = -2 \Bigg(&\sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k \exp\left(\frac{y_i h^{-1}(\mu_{ik}) - b(h^{-1}(\mu_{ik}))}{\sigma^2/v_i}\right) \\ &- \sum_{i=1}^{n} \frac{y_i h^{-1}(y_i) - b(h^{-1}(y_i))}{\sigma^2/v_i} \Bigg) \end{aligned}$$

Beispiel: Seien $y_i|z_i$ unabhängige $Poiss(\mu_i)$ Variablen mit $\log\mu_i = x_i^t\gamma + z_i$ und $z_i \overset{iid}{\sim} G$. Dann ist

$$\begin{aligned} l_K(y;\mu) &= -\sum_{i=1}^{n} \log y_i! + \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k \exp(y_i \log\mu_{ik} - \mu_{ik}) \\ l(y;y) &= -\sum_{i=1}^{n} \log y_i! + \sum_{i=1}^{n} (y_i \log y_i - y_i) \end{aligned}$$

und die (skalierte) Deviance hat die Form

$$D_K(y;\mu) = -2\Big(\sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k \exp(y_i \log\mu_{ik} - \mu_{ik}) - \sum_{i=1}^{n} (y_i \log y_i - y_i)\Big).$$

Leider liefert der EM-Algorithmus nicht automatisch Schätzer für die Standardfehler der Parameter mit. Dietz und Böhning (1995) diskutieren daher verschiedene Ansätze, wovon einer auf der asymptotischen Äquivalenz der quadrierten Waldstatistik und der Likelihood-Quotientenstatistik beruht. In diesem Sinne gilt

$$\left(\frac{\hat\gamma_j}{S.E.(\hat\gamma_j)}\right)^2 \approx -2\Big(l_K(y;\hat\mu_{(j)}) - l_K(y;\hat\mu)\Big) = \frac{1}{\sigma^2}\left(D_K(y;\hat\mu_{(j)}) - D_K(y;\hat\mu)\right).$$

Die Likelihood-Quotientenstatistik entspricht also der skalierten Deviance-Differenz zwischen dem reduzierten Modell $\mu_{(j)}$, ohne der j-ten erklärenden Variablen ($\gamma_j = 0$), und dem übergeordneten Modell μ, in dem γ_j geschätzt wird. Diese Überlegung führt zu

$$\widehat{Var}(\hat{\gamma}_j) = \sigma^2 \frac{\hat{\gamma}_j^2}{D_K(y;\hat{\mu}_{(j)}) - D_K(y;\hat{\mu})}.$$

Ist σ^2 unbekannt, so kann man die mittlere Pearson-Statistik als Schätzer heranziehen.

Offen bleibt noch die Diskussion der Wahl von K. Bis jetzt wurde die unbekannte Anzahl K immer als feste Größe angenommen. Betrachtet man eine andere Approximation oder Schätzung von (9), für die $K' > K$ Massepunkte verwendet werden, so resultiert im Falle $\sigma^2 = 1$

$$D_K(y;\hat{\mu}) - D_{K'}(y;\hat{\mu}) = -2(l_K(y;\hat{\mu}) - l_{K'}(y;\hat{\mu})).$$

Verwendet man die Gauß-Quadratur mit K und K' Quadraturpunkten, so stellt die Deviance-Differenz keine Likelihood-Quotientenstatistik im herkömmlichen Sinn dar, sondern beschreibt nur die Differenz zweier unterschiedlicher Approximationen von ein und derselben Größe $l(y;\hat{\mu})$. Die Anzahl der Parameter im Modell hängt auch nicht von der Anzahl der Massepunkte K ab. Eigentlich vergleicht man hier zwei Mischverteilungsmodelle, in denen die jeweiligen Massepunkte mit den Wahrscheinlichkeitsmassen als bekannt angenommen werden.

Wird die Verteilung der Zufallseffekte nichtparametrisch geschätzt, so ergibt die Vergrößerung von K auf K' weitere Modellparameter. Diese beschreiben die Koordinaten der zusätzlichen Massepunkte. Natürlich zählt zur vollständigen nichtparametrischen ML-Schätzung auch die Bestimmung von K. Zumindest für Modelle, deren linearer Prädiktor keine erklärenden Variablen beinhaltet, gibt es dafür bereits Algorithmen und Programme (siehe Böhning et al. (1992)). Für ein Modell mit Einflußgrößen stellt die obige Deviance-Differenz eine Likelihood-Quotientenstatistik dar, mit der die Hypothese $H_0 : \pi_{K+1} = \ldots = \pi_{K'} = 0$ getestet werden kann. Da jedoch dieses Nullmodell am Rande des Parameterraumes unter H_1 liegt, ist diese Teststatistik unter H_0 auch nicht mehr χ^2-verteilt, sondern weist eine degenerierte Verteilung auf. Wie schon zuvor ist nicht einmal gewährleistet, daß sich beim Übergang von K zu K' die Deviance reduziert.

Aitkin (1994) empfiehlt für beide Ansätze, mit $K = 1$ (keine zufälligen Effekte) zu starten und diese Anzahl solange zu erhöhen bis sich das Maximum der Likelihood-Funktion stabilisiert. Dann scheint zumindest die Approximation ausreichend gut zu sein.

Es sei hier noch erwähnt, daß die Zulässigkeit einer Modellreduktion durch Weglassen erklärender Größen im linearen Prädiktor sehr wohl mit der entsprechenden Likelihood-Quotientenstatistik (Deviance-Differenz) getestet werden kann und diese wie gewöhnlich unter H_0 auch asymptotisch χ^2-verteilt ist.

Tabelle 1: Jährliche Schwangerschaftsanzahlen m_i mit Anteil Jugendlicher r_i, sowie geschätzte Erwartungen $\tilde{\mu}_i$ und Gewichte $\hat{w}_{ik}$ bzgl. der nichtparametrischen Lösung mit $K = 4$.

	County	total m_i	Rate r_i	$\tilde{\mu}_i$	$\hat{w}_{ik}$			
1	Alachua	2848	32.2	30.92	.000	.000	1.00	.000
2	Bradford	344	48.1	46.21	.000	.989	.011	.000
3	Clay	1617	22.7	22.95	.000	.000	.007	.993
4	Columbia	688	50.4	46.38	.000	1.00	.000	.000
5	Dixie	160	43.8	42.93	.001	.777	.216	.006
6	Gilchrist	133	20.0	27.88	.000	.029	.536	.435
7	Hamilton	171	79.8	78.20	.976	.024	.000	.000
8	Lafayette	66	35.5	36.99	.004	.430	.469	.097
9	Levy	350	28.6	29.85	.000	.014	.827	.159
10	Marion	2753	29.4	30.91	.000	.000	1.00	.000
11	Putman	982	43.8	46.37	.000	.999	.001	.000
12	Suwanee	351	36.1	35.65	.000	.312	.677	.011
13	Union	135	54.3	46.68	.033	.915	.051	.001

6 Beispiele

6.1 Überdispersion

In Booth (1995) und Aitkin (1996b) wird die Häufigkeit von Schwangerschaften bei Teenagern untersucht. Das verwendete Datenmaterial beschreibt die Situation in $n = 13$ Counties von North Central Florida und basiert auf den Beobachtungszeitraum 1989-1991. In der Tabelle 1 sind die durchschnittlichen Anzahlen m_i von Geburten innerhalb eines Jahres und die Anteile r_i (in Promille) jener Mütter angegeben, welche jünger als 17 Jahre alt sind. Primär soll untersucht werden, ob der Anteil Jugendlicher unabhängig vom County ist.

Unter der Annahme, daß die Anzahl Jugendlicher über alle drei Jahre eine binomialverteilte Größe ist, also $y_i = 3m_i \cdot r_i/1000 \overset{ind}{\sim} Bin(3m_i, \pi_i)$, ergibt das logistische Unabhängigkeits-Modell $logit(\pi_i) = \tau$ eine Gesamtrate von 33.9 ($\hat{\tau} = -3.350$) für alle 13 Counties mit einer Deviance von 89.48 bei 12 Freiheitsgraden.

Da keine weiteren erklärenden Variablen vorliegen und man durch die Hinzunahme des County-Faktors das volle Modell mit Deviance Null erhält, können anstelle des einen festen Effektes τ auch n normalverteilte Zufallseffekte τ_i verwendet werden, um diese ausgeprägte Überdispersion in das Modell einzubeziehen. Für das Modell $logit(\pi_i) = \tau_i$ mit $\tau_i \overset{iid}{\sim} N(\tau, \sigma_\tau^2)$ liefert die 6-Punkt Gauß-Quadratur die Schätzungen $\hat{\tau} = -3.22$ und $\hat{\sigma}_\tau = 0.326$. Für dieses Logit-Normal Modell, welches auch von Andersen und Hinde (1988) in ähnlicher Form diskutiert wird, reduziert sich die Deviance auf 33.02.

Das nichtparametrisch geschätzte Mischmodell ergibt für $K = 4$ als Deviance

31.09. Mit den Realisationen der Schätzer von den Massepunkten und von den Mischungsanteilen konstruiert man hier Schätzer für den Erwartungswert und für die Standardabweichung der Effektverteilung. Man erhält dafür die Werte $\hat{\tau} = -3.230$ und $\hat{\sigma}_\tau = 0.343$. Diese beiden Resultate sind erstaunlicherweise sehr ähnlich den vorigen Ergebnissen unter der Normalverteilungsannahme. Die einzelnen geschätzten Massen und Massestellen (zentriert um $\hat{\tau}$) sind

k	1	2	3	4
$\hat{\tau}_k$	0.7744	0.2070	-0.2147	-0.5236
$\hat{\pi}_k$	0.0781	0.4219	0.3691	0.1309

Interessant ist es auch, die Gewichte $\hat{w}_{ik}$ zu betrachten. Diese beschreiben die Zugehörigkeit der i-ten Beobachtung zur k-ten Mischkomponente beschreibt. Aus der Tabelle 1 geht hervor, daß die erste Komponente ausschließlich die Situation in Hamilton beschreibt, wo eine extrem hohe Rate von 7.98% beobachtet wurde. Dies entspricht in etwa wie auch $\hat{\pi}_1 = 7.81\%$ dem Stichprobenanteil von $1/n$.

Von weiterem Interesse ist auch die Betrachtung der K konditionalen Modelle. Allgemein erhält man für jede Mischkomponente

$$\hat{\eta}_{ik} = x_i^t\hat{\gamma} + \hat{\tau}_k.$$

Die konditionalen Erwartungswert-Modelle $\hat{\mu}_{ik} = h(\hat{\eta}_{ik})$ ergeben die Raten 79.0 (Hamilton) sowie 46.4, 30.9 und 22.9 (vor allem Clay und Gilchrist mit sehr niedrigen Raten).

Mit der geschätzten Version der Approximation (15) für die marginale Dichte der Response-Variablen resultiert für

$$E(y_i) = \int y_i f(y_i; \theta) dy_i$$

das geschätzte marginale (populationsgemittelte) Modell

$$\hat{E}(y_i) = \int y_i \sum_{k=1}^{K} \hat{\pi}_k f(y_i|\hat{\tau}_k; \hat{\gamma}) dy_i = \sum_{k=1}^{K} \hat{\pi}_k \hat{E}(y_i|\hat{\tau}_k) = \sum_{k=1}^{K} \hat{\pi}_k \hat{\mu}_{ik}.$$

Hier ergibt sich für jede Beobachtung eine Rate von 40.2. Zusammen mit den vier konditionalen Modellen ist dieses Ergebnis in der Abbildung 1 dargestellt.

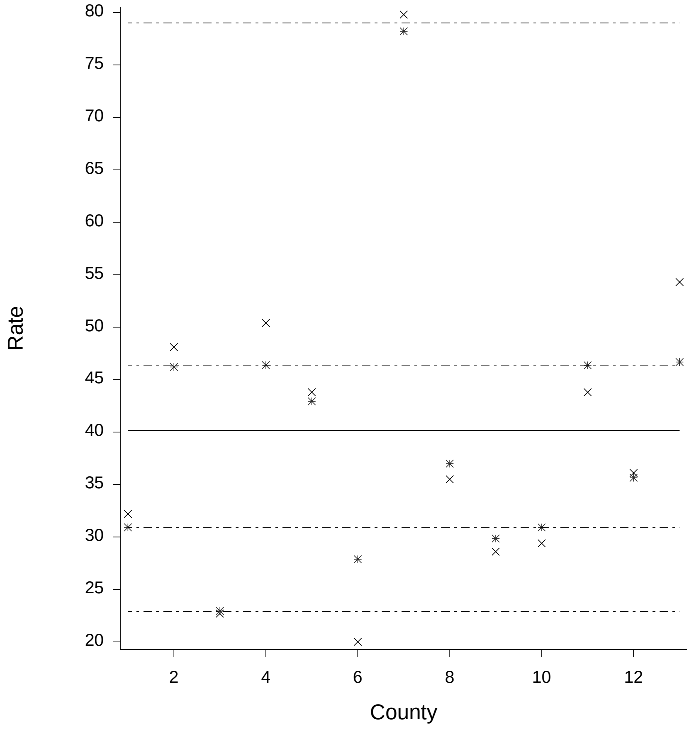
Approximiert man analog die a-posteriori Erwartung des Zufallseffektes durch

$$E(\tau_i|y_i) = \frac{\int \tau_i f(y_i|\tau_i) g(\tau_i) d\tau_i}{\int f(y_i|\tau_i) g(\tau_i) d\tau_i} \approx \frac{\sum_{k=1}^{K} \tau_k f(y_i|\tau_k)\pi_k}{\sum_{k=1}^{K} f(y_i|\tau_k)\pi_k},$$

so resultiert als Plug-in Schätzer der posteriori gewichtete Mittelwert der geschätzten Massestellen

$$\hat{E}(\tau_i|y_i) = \sum_{i=1}^{K} \hat{\tau}_k \frac{\hat{f}(y_i|\hat{\tau}_k)\hat{\pi}_k}{\sum_{k=1}^{K} \hat{f}(y_i|\hat{\tau}_k)\hat{\pi}_k} = \sum_{i=1}^{K} \hat{\tau}_k \hat{w}_{ik}.$$

Abbildung 1: Daten ($\times$), konditionale Modelle (strichlierte Linien) und marginales Modell (durchgezogene Linie) sowie empirische Bayes-Schätzungen (*).



Verwendet man diesen geschätzten Erwartungswert anstelle der nichtbeobachtbaren τ_i im linearen Prädiktor, führt dies wegen $\sum_k \hat{w}_{ik} = 1$ zu den empirischen Bayes-Schätzungen

$$\tilde{\eta}_i = x_i^t\hat{\gamma} + \hat{E}(\tau_i|y_i) = \sum_{k=1}^{K}(x_i^t\hat{\gamma} + \hat{\tau}_k)\hat{w}_{ik} = \sum_{k=1}^{K}\hat{\eta}_{ik}\hat{w}_{ik}.$$

Für die entsprechend geschätzten Erwartungswerte, welche in der Tabelle 1 und in der Abbildung 1 das Schätzergebnis darstellen, folgt

$$\tilde{\mu}_i = \sum_{k=1}^{K}\hat{\mu}_{ik}\hat{w}_{ik}.$$

6.2 Hierarchische Modelle

Mehta, Patel und Senchaudhuri (1988) vergleichen an 22 Kliniken die Wirkung eines neuen Medikamentes mit einer Standardtherapie. Dazu wird in jeder Klinik jeweils eine Patientengruppe mit einem der beiden Mitteln behandelt. Die neue Behandlung wird als erfolgreich bezeichnet, falls eine unerwünschte Nebenwirkung selten auftritt. Von Interesse ist daher die Betrachtung des Behandlungseffektes γ_T. In den Daten der Tabelle 2 fällt besonders die extreme Responserate bei der Standardbehandlung in der Klinik 15 auf.

Tabelle 2: Beobachtete absolute und relative Mißerfolgshäufigkeiten an 22 Kliniken in Abhängigkeit vom Medikament, sowie empirische Bayes-Schätzungen und Gewichte für das Modell mit klinikspezifischen Intercepts und Behandlungseffekten.

Klinik	neu	alt	neu	alt	neu	alt	
i	y_{i1}/m_{i1}	y_{i2}/m_{i2}	y_{i1}/m_{i1}	y_{i2}/m_{i2}	$\tilde{\mu}_{i1}$	$\tilde{\mu}_{i2}$	$\hat{w}_{ik}$
1	0/15	0/15	.000	.000	.015	.059	0 .161 .839
2	0/39	6/38	.000	.158	.034	.127	0 .933 .067
3	1/21	3/21	.048	.143	.034	.125	0 .919 .081
4	1/15	2/17	.067	.118	.032	.120	0 .855 .145
5	1/21	2/21	.048	.095	.030	.113	0 .775 .225
6	0/12	2/12	.000	.167	.030	.110	0 .749 .251
7	3/52	10/52	.058	.192	.036	.132	0 1.00 .000
8	0/19	2/19	.000	.105	.025	.094	0 .558 .442
9	1/15	0/15	.067	.000	.021	.080	0 .400 .600
10	2/28	2/29	.071	.069	.031	.117	0 .821 .179
11	0/19	2/20	.000	.100	.024	.091	0 .535 .466
12	0/12	1/12	.000	.083	.023	.086	0 .476 .524
13	0/24	5/24	.000	.208	.035	.129	0 .960 .040
14	2/12	2/13	.167	.154	.035	.130	0 .970 .030
15	0/14	11/14	.000	.786	.000	.786	1 .000 .000
16	0/53	4/52	.000	.077	.015	.061	0 .189 .811
17	0/20	0/20	.000	.000	.013	.053	0 .094 .906
18	0/21	0/21	.000	.000	.013	.052	0 .084 .916
19	1/51	1/49	.020	.020	.011	.047	0 .031 .969
20	0/13	1/14	.000	.071	.021	.082	0 .422 .578
21	0/13	1/14	.000	.071	.021	.082	0 .422 .578
22	0/21	0/21	.000	.000	.013	.052	0 .084 .916

Die Werte der Deviance für die logistischen Modelle mit festen Parametern sind

Modell	Dev/df	$\hat{\gamma}_T$ ($\hat{\sigma}_T$)
τ	129.05/43	
$\tau + \gamma_T$	95.32/42	1.646 (0.324)
$\tau + \gamma_T + \gamma_K$	29.47/21	1.780 (0.339)

Besonders auffällig ist der signifikante Behandlungseffekt (Deviance-Reduktion von 33.73/1). Trotzdem scheint auch Überdispersion vorzuliegen. Zusätzlich scheint es große Unterschiede in den Behandlungserfolgen an den einzelnen Kliniken zu geben (Deviance-Reduktion von 65.85/21).

Dies motiviert ein Varianzkomponentenmodell, in dem bei festem Parameter γ_T der Intercept τ_i der i-ten Klinik als zufälliger Effekt betrachtet wird. Nichtparametrische ML-Schätzungen liefern für verschiedene Werte von K

K	Dev	$\hat{\tau}$ $(\hat{\sigma}_\tau)$	$\hat{\gamma}_T$ $(\hat{\sigma}_T)$	$(\hat{\tau}_k)$ $\hat{\pi}_k$
2	81.2	-3.92 (0.518)	1.69 (0.288)	(1.21) .156; (-0.22) .844
3	71.3	-3.99 (0.816)	1.76 (0.294)	(2.59) .046; (0.43) .520; (-0.78) .434
4	71.3	-4.01 (0.851)	1.76 (0.294)	(2.61) .046; (0.48) .472; (-0.56) .407; (-1.56) .075

Der Standarfehler des Interceptparameters wird hier über die Massestellen und deren Massen geschätzt, während für diese Schätzung beim festen Effekt die Methode aus Abschnitt 5 verwendet wird. Für eine größere Anzahl von Massepunkten ändert sich weder der Wert der Deviance noch der geschätzte Behandlungseffekt.

Wir wollen noch ein Modell diskutieren, in dem zusätzlich zum Intercept auch der Behandlungsparameter einen klinikspezifischen Zufallseffekt darstellt. Dafür resultiert

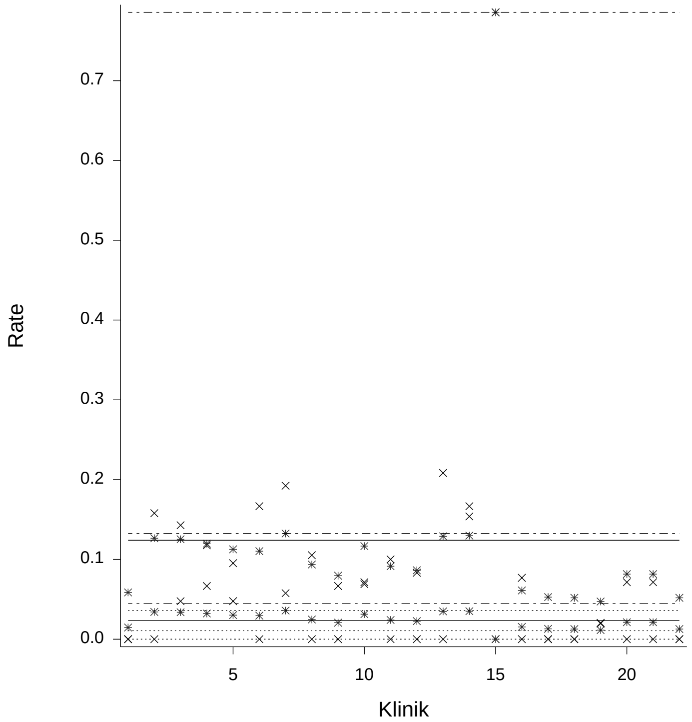
K	Dev	$\hat{\tau}$ $(\hat{\sigma}_\tau)$	$\hat{\gamma}_T$ $(\hat{\sigma}_T)$	$(\hat{\tau}_k, \hat{\gamma}_k)$ $\hat{\pi}_k$
2	66.4	-3.98 (1.294)	1.86 (2.042)	(-5.93, 9.35) .046; (0.28,-0.45) .954
3	61.8	-4.13 (1.399)	1.88 (2.036)	(-5.78, 9.33) .046; (0.84,-0.47) .521; (-0.40,-0.41) .433
4	60.1	-5.53 (3.085)	3.17 (3.642)	(-5.38, 9.04) .046; (2.52,-2.07) .438; (-4.30, 4.51) .271; (1.25,-2.97) .246

Alle drei Ergebnisse sind bezüglich der ersten Mischungskomponente sehr ähnlich. Offensichtlich wird damit gerade die extreme Beobachtung in der Klinik 15 erfaßt. Dies wird auch durch die Werte in der Spalte $\hat{w}_{i1}$ der Tabelle 2 bestätigt. Diese Gewichte beziehen sich auf die Lösung mit $K = 3$. Die sechs konditionalen (neu: 0.0001, 0.0358, 0.0106; alt: 0.7857, 0.1324, 0.0445) und die beiden marginalen Modelle (neu: 0.0233; alt: 0.1240) zu dieser Lösung sind in der Abbildung 2 eingezeichnet. Die empirischen Bayes-Schätzer für beide Behandlungsarten sind in der Tabelle 2 angegeben.

7 Diskussion

Die Approximation der Likelihood-Funktion im EM-Algorithmus durch die Gauß-Quadratur oder durch die nichtparametrische ML-Schätzung resultiert in Generalisierten Linearen Mischmodellen zu einer gewichteten Version einer ML-Schätzung wie sie auch bei den herkömmlichen GLMs resultiert. Prinzipiell ist daher für die Berechnung der Schätzer jede Software verwendbar, mit der eine iterativ gewichtete Kleinste-Quadrate Schätzung durchgeführt werden kann. Hier wurden für die Berechnungen beim Überdispersionsmodell im vorigen Abschnitt das in Aitkin und Francis (1995) beschriebene GLIM-Makro verwendet. Ein weiteres, von denselben Autoren zu Verfügung gestelltes, Makro erlaubt die Parameterschätzung bei hierarchischen Modellen mit höherdimensionaler Effektverteilung. Letzteres stellt eine Verallgemeinerung des Ansatzes in Anderson und Aitkin (1985) dar.

Abbildung 2: Daten (×), konditionale (strichlierte und punktierte Linien) und marginale Modelle (durchgezogene Linien), sowie empirische Bayes-Schätzungen (*).



Zu erwähnen ist noch, daß durch dieses Verfahren vor allem die unbekannten Parameter im Modell geschätzt werden können. Keinesfalls sollte es jedoch für die Schätzer und der unbekannten Effektverteilung eingesetzt werden. Die Ursache dafür liegt hauptsächlich in der doch oft kritischen Wahl von K. Möglicherweise kann die in DerSimonian (1986) verwendete Konstruktion für Mischverteilungen auch auf diese Modellklasse angewendet werden.

Die hier beschriebene Technik ist ein sehr allgemeines Verfahren und äußerst vielseitig verwendbar. Natürlich stellen die im vierten Abschnitt diskutierten Modelle nur einen kleinen Ausschnitt aller möglichen Anwendungsgebiete dar. Weitere Einsatzmöglichkeiten dieses Zuganges findet man in Aitkin (1995) beschrieben.

Danksagung

Mein ganz besonderer Dank gilt R. Hatzinger für die Organisation und Durchführung eines gleichnamigen Seminars. Weiters sei an dieser Stelle noch allen übrigen Referenten für die wertvollen Kommentare zu dieser Thematik in ihren Seminarbeiträgen gedankt, die mich motivierten, diesen Aufsatz in der vorliegenden Form zu verfassen.

Literatur

AITKIN, M. (1994). An EM algorithm for overdispersion in generalized linear models. In: *Proceedings of the 9th International Workshop on Statistical Modelling, Exeter 1994.*

AITKIN, M. (1995). NPML estimation of the mixing distribution in general statistical models with unobserved random effects. In: *Statistical Modelling* (G.U.H. Seeber et al., eds.), 1–9. New York: Springer-Verlag.

AITKIN, M. (1996A). A general maximum likelihood analysis of overdispersion in generalized linear models. *Statistics and Computing* **6**, 251–262.

AITKIN, M. (1996B). Empirical Bayes shrinkage using posterior random effect means from nonparametric maximum likelihood estimation in general random effect models. In: *Statistical Modelling - Proceedings of the 11th International Workshop on Statistical Modelling* (A. Forcina et al., eds.), 87–94. Citta di Castello: Graphos.

AITKIN, M., FRANCIS, B.J. (1995). Fitting overdispersed generalized linear models by nonparametric maximum likelihood. *GLIM Newsletter* **25**, 37–45.

ANDERSON, D.A. (1988). Some models for overdispersed binomial data. *The Australian Journal of Statistics* **30**, 125–148.

ANDERSON, D.A., AITKIN, M. (1985). Variance component models with binary response: interviewer variability. *Journal of the Royal Statistical Society B* **47** 203–210.

ANDERSON, D.A., HINDE, J.P. (1988). Random effects in generalized linear models and the EM algorithm. *Communications in Statistics – Theory and Methods* **17**, 3847–3856.

BÖHNING, D., SCHLATTMANN, P., LINDSAY, B. (1992). Computer-assisted analysis of mixtures (C.A.MAN): Statistical algorithms. *Biometrics* **48**, 283–303.

BOOTH, J. (1995). Bootstrap methods for generalized linear mixed models with applications to small area estimation. In: *Statistical Modelling* (G.U.H. Seeber et al., eds.), 43–51. New York: Springer-Verlag.

DEMPSTER, A.P., LAIRD, N.M., RUBIN, D.B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm. (with discussion). *Journal of the Royal Statistical Society B* **39**, 1–38.

DERSIMONIAN, R. (1986). Maximum likelihood estimation of a mixing distribution. *Applied Statistics* **35**, 302–309.

DIETZ, E., BÖHNING, D. (1995). Statistical inference based on a general model of unobserved heterogeneity. In: *Statistical Modelling* (G.U.H. Seeber et al., eds.), 75–82. New York: Springer-Verlag.

Diggle, P.J., Liang, K.-Y., Zeger, S. (1995). *The Analysis of Longitudinal Data.* Oxford: Clarendon Press.

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

Hinde, J. (1982). Compound Poisson regression models. In: *GLIM 82* (R. Gilchrist, ed.), 109–121. New York: Springer-Verlag.

Kiefer, J., Wolfowitz, J.(1956). Consistency of the maximum likelihood estimator in the presence of infinitely many nuisance parameters. *The Annals of Mathematical Statistics* **27**, 887–906.

Laird, N. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association* **73**, 805–811.

Lindsay, B.G. (1983). The geometry of mixture likelihoods: a general theory. *The Annals of Statistics* **11**, 86–94.

Longford, N.T. (1993). *Random Coefficient Models.* Oxford: Clarendon Press.

McCullagh, P., Nelder, J.A. (1989). *Generalized Linear Models. Second Edition.* London, New York: Chapman and Hall.

Mehta, C.R., Patel, N.R., Senchaudhuri, P. (1988). Importance sampling for estimating exact probabilities in permutational inference. *Journal of the American Statistical Association* **83**, 999–1005.

Wood, A., Hinde, J.P. (1987). Binomial variance component models with a nonparametric assumption concerning random effects. In: *Longitudinal Data Analysis* (R. Crouchley, ed.), 110–128. Aldershot, Hants: Avebury.

Adresse des Autors:

Dr. Herwig Friedl
Institut für Statistik, Technische Universität Graz
Lessingstraße 27, A–8010 Graz
E-Mail: `Friedl@Stat.tu-graz.ac.at`