Sofie WALTL, BSc

Standardfehler in Negativ Binomialmodellen mit zufälligen Effekten

MASTERARBEIT

zur Erlangung des akademischen Grades einer
Diplom-Ingenieurin

Masterstudium Finanz- und Versicherungsmathematik

Graz University of Technology
Technische Universität Graz

Betreuer:
Univ.-Prof. Dipl.-Ing. Dr.techn. Herwig FRIEDL

Institut für Statistik

Graz, im September 2013

Eidesstattliche Erklärung

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Graz, am ....

(Unterschrift)

ZUSAMMENFASSUNG

In R-Funktionen, welche in der Lage sind *Modelle mit zufälligen Effekten* zu schätzen, ist momentan noch keine Methode implementiert, um Standardfehler sinnvoll zu berechnen. Akkurate Standardfehler sind jedoch bei der Interpretation von Modellparametern von essentieller Bedeutung. Diese Unzulänglichkeit soll in der vorliegenden Arbeit behoben werden. Dazu werden Techniken vorgestellt, um Standardfehler in Modellen mit zufälligen Effekten basierend auf einer Verteilung aus der *einparametrigen, linearen Exponentialfamilie* beziehungsweise auf der *Negativen Binomialverteilung* herzuleiten. Diese werden abschließend durch eine Simulationsstudie validiert und an mikrobiologischen Daten angewandt.

ABSTRACT

Yet there are no methods implemented in R functions estimating *Random Effect Models* that calculate acceptable standard errors. However, accurate standard errors are of great importance when interpreting model parameters. This thesis shall eliminate those shortcomings. Therefore, methods to calculate standard errors in Random Effect Models based on distributions belonging to the *single-parameter, linear exponential family* or the *negative binomial distribution* are derived. The techniques are evaluated through a simulation studie and applied to microbiological data.

INHALTSVERZEICHNIS

1 EINLEITUNG

Ausgehend von *Generalisierten Linearen Modellen* wird die Theorie der *Modelle mit zufälligen Effekten* Schritt für Schritt aufgebaut. Dabei werden zunächst die einfachsten Modelle dieser Art - soganannte *Überdispersionsmodelle* - vorgestellt, welche zu *Modellen mit zufälligen Koeffizienten* und schließlich *Varianzkomponentenmodellen* erweitert werden. Es werden stets zwei Schätzmethoden behandelt: Ist die Annahme gerechtfertigt, dass die zufälligen Effekte normalverteilt sind, kann die Gauß-Hermite Quadratur herangezogen werden, um die ansonsten nicht analytisch lösbaren Integrale zu approximieren. Kann diese Annahme nicht getroffen werden, dient als Alternative eine nichtparametrische Schätzung der Dichte der Zufallseffekte. Es wird gezeigt, dass diese Dichte in Wahrheit eine diskrete Wahrscheinlichkeitsfunktion ist, welche durch die Angabe der Anzahl an Massepunkten sowie deren Lage und Gewicht vollständig spezifiziert ist. Während die Anzahl der Massepunkte a priori festgesetzt werden muss, werden die Lage und Gewichte der Massepunkte als Modellparameter aufgefasst und geschätzt.

Im ersten Teil dieser Arbeit wird der stufenweise Aufbau für Modelle basierend auf Verteilungen aus der *einparametrischen, linearen Exponentialfamilie* durchgeführt. Im zweiten Teil lassen wir diese Einschränkung fallen und erklären die Theorie für die *Negativ Binomialverteilung*. Diese Verteilung stellt eine gute Alternative zur Poissonverteilung bei der Modellierung von Anzahlen dar. Der große Vorteil dabei ist, dass Erwartungswert und Varianz nicht als ident angenommen werden müssen.
Teil drei stellt einen kurzen Überblick über die derzeit verfügbaren R-Pakete dar, welche in der Lage sind Modelle mit zufälligen Effekten zu schätzen.
Im vierten Teil wird eine Methode vorgestellt, um akkurate Standardfehler zu erhalten. Zur Zeit berechnen sämtliche Funktionen aus den R-Paketen, die in Teil drei genannt werden, Standardfehler über die Inverse der negativen vollständigen Hessematrix. Louis (1982) zeigt jedoch auf, dass sich zur Schätzung der Standardfehler die *beobachtete Fisher Informationsmatrix* weit besser eignet. In diesem Teil wird diese alternative Methode nachvollzogen. Als Anwendung werden die konkreten Formeln für Modelle basierend auf der Poisson- und der Negativ Binomialverteilung hergeleitet.
Im fünften und letzten Teil werden die vorgestellten Methoden schließlich validiert. Dazu führen wir eine *Monte Carlo Studie* durch,

die zum Ergebnis kommt, dass die Variabilität der Regressionsparameter durch die über die nach Louis (1982) benannten Formeln berechneten Standardfehler sehr gut beschreibt. Es wird auch deutlich, dass die in den R-Paketen bisher verwendete Methode die tatsächlichen Standardfehler deutlich unterschätzt. Als Ergänzung wird als Abschluss ein Beispiel anhand mikrobiologischer Daten vorgeführt.

Für die theoretischen Ausführungen werden drei Hauptquellen herangezogen. Die Ausführungen zu den Exponentialfamilienmodellen basieren haupsächlich auf Aitkin et al. (2009). Der Erweiterung auf die Negativ Binomialfamilie liegt hingegen Booth et al. (2003) zu Grunde. Als Hauptquelle für die Herleitung der Standardfehler in Exponentialfamilienmodellen dient Friedl und Kauermann (2000).

Der eben beschriebene Aufbau der Arbeit ist in Abbildung 1 kompakt zusammengefasst.

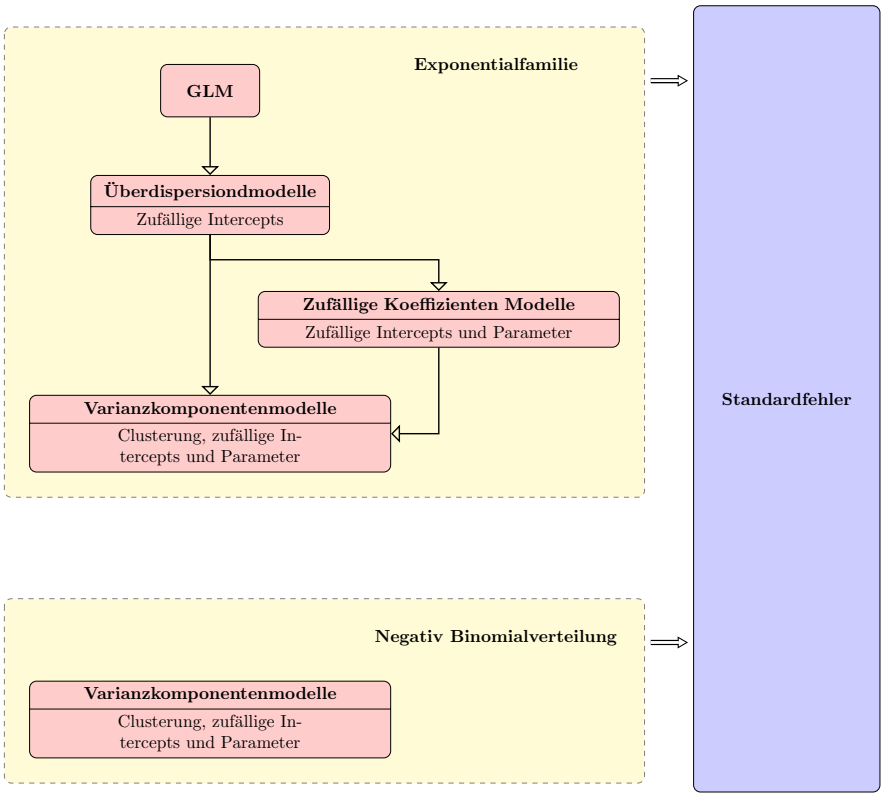

Abbildung 1: Aufbau der Arbeit

Teil I

EXPONENTIALFAMILIENMODELLE MIT ZUFÄLLIGEN EFFEKTEN

In diesem Abschnitt betrachten wir ausschließlich Modelle basierend auf der einparametrigen, linearen Exponentialfamilie. Zunächst wiederholen wir in aller Kürze die Theorie der *Generalisierten Linearen Modelle* und stellen das Prinzip des *EM-Algorithmus* vor. Im Anschluss daran bauen wir Schritt für Schritt mächtigere Modellklassen auf: Wir beginnen mit *Überdispersionsmodellen* - den einfachsten Modellen mit Zufallseffekten, gehen weiter zu *Zufälligen Koeffizientenmodellen* und schließen mit der alles bisherige umfassenden Klasse der *Varianzkomponentenmodellen*.

2

GRUNDLAGEN

2.1 GENERALISIERTE LINEARE MODELLE

Generalisierte Lineare Modelle stellen die Basis jener Modellklasse dar, die im Folgenden näher beschrieben werden soll: *Modelle mit zufälligen Effekten*. Darum werden hier kurz die definierenden Eigenschaften wiederholt. Die Notation dafür wird weitgehend aus Aitkin et al. (2009) übernommen.

Im Zentrum der Betrachtung steht eine besondere Klasse von Verteilungen, die *Exponentialfamilie*.

Definition 1 (Exponentialfamilie) *Die Verteilung einer Zufallsvariablen y stammt aus der **linearen, einparametrigen Exponentialfamilie**, falls deren Dichte- bzw. Wahrscheinlichkeitsfunktion in der folgenden kanonischen Form darstellbar ist:*

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

wobei $a(\cdot)$, $b(\cdot)$ und $c(\cdot)$ bekannte Funktionen sind und $a(\phi) > 0$. ϑ, der einzige freie Parameter, ist der kanonische Parameter der Verteilung und ϕ der Skalierungsparameter.

Wichtige Mitglieder der Exponentialfamilie sind beispielsweise die Normalverteilung, die Poissonverteilung oder die Gammaverteilung. Eine einfache Rechnung - vgl. Aitkin et al. (2009), Abschnitt 2.9 - liefert

$$\mathbb{E}[y] = b'(\vartheta), \tag{1}$$
$$\mathbb{V}ar[y] = a(\phi)b''(\vartheta). \tag{2}$$

Ein Generalisiertes Lineares Modell (GLM) wird nun durch die folgenden drei Komponenten definiert:

1. Eine Verteilung aus der Exponentialfamilie mit Parameter ϑ, aus der per Annahme die Response-Variable $y = (y_1, \ldots, y_n)^\top$ stammt. Die y_i, $i = 1, \ldots, n$, seien unabhängig.

2. Ein linearer Prädiktor $\eta = (\eta_1, \ldots, \eta_n)^\top$ in den erklärenden Variablen $x_j = (x_{1j}, \ldots, x_{nj})^\top$, $j = 1, \ldots, p-1$ - zusammengefasst zur Designmatrix $X = (x_0, \ldots, x_{p-1})^\top$ - und p Parametern $\beta = (\beta_0, \ldots, \beta_{p-1})^\top$,

$$\eta = X^\top \beta.$$

Die erste Spalte der Designmatrix beschreibt üblicherweise den Intercept. In diesem Fall gilt $x_0 = (1, \dots, 1)^\top$.

3. Eine *Link-Funktion* $g(\mu)$, welche den Zusammenhang zwischen dem linearen Prädiktor η und dem Vektor der Erwartungswerte $\mu = \mathbb{E}[y]$ beschreibt,

$$\eta = g(\mu).$$

Da der Erwartungswert in der Exponentialfamilie unabhängig vom Skalierungsparameter ϕ ist (siehe (1)), beeinflussen die erklärenden Variablen x_j im GLM nur den kanonische Parameter ϑ.

Mit $V_i = V(\mu_i)$ soll nun die Varianzfunktion und mit μ_i der Erwartungswert der i-ten Responsekomponente bezeichnet werden,

$$\mathbb{V}ar[y_i] = a_i(\phi)V(\mu_i) \text{ und } \mu_i := \mathbb{E}[y_i].$$

Im Zuge der Maximum-Likelihood Schätzung im GLM - vgl. zum Beispiel Aitkin et al. (2009), Abschnitt 2.9.3 - wird die Log-likelihoodfunktion $\ell(\beta, \vartheta|y)$ ermittelt,

$$\ell(\beta, \vartheta|y) = \sum_{i=1}^{n} \left(\frac{y_i\vartheta_i - b(\vartheta_i)}{a_i(\phi)} + c(y_i, \phi) \right).$$

Daraus erlangt man die Scorefunktion für β,

$$s(\beta) = \frac{\partial \ell}{\partial \beta} = \sum_{i=1}^{n} \frac{(y_i - \mu_i)x_i}{V_i g'(\mu_i)}, \tag{3}$$

wobei $x_i = (x_{i0}, \dots, x_{i(p-1)})^\top$ die i-te Zeile der Designmatrix bezeichnet. Eben diese Scorefunktion wird später bei der Maximum-Likelihood Schätzung in Modellen mit normalverteilten, zufälligen Effekten wieder vorkommen.

2.2 TECHNISCHE HILFSMITTEL

In diesem Abschnitt werden zwei völlig voneinander unabhängige Hilfsmittel - der *EM-Algorithmus* und die *Gauß-Quadratur* - vorgestellt, die im Weiteren für die Schätzung der Parameter in *Modellen mit zufälligen Effekten* hilfreich sein werden.

2.2.1 *EM-Algorithmus*

Der EM-Algorithmus *(Expectation-Maximization algorithm)* ist ein breit anwendbarer, iterativer Algorithmus, um Maximum-Likelihood Schätzungen im Falle unvollständiger Daten zu erhalten (Little und Rubin, 1987, Kapitel 7). Der Algorithmus basiert auf einer einfachen Idee, wie mit unvollständigen Daten umgegangen werden kann:

1. Ersetze fehlende Werte durch geschätzte Werte.
2. Schätze die Parameter des Modells basierend auf dem ergänzten Datensatz.
3. Schätze die fehlenden Werte erneut unter der Annahme, dass die geschätzen Parameter korrekt sind.
4. Wiederhole Schritt 2 und 3 bis Konvergenz eintritt.

Auf Grund der recht intuitiven Herangehensweise wurde der Algorithmus immer wieder in den verschiedensten Gebieten angewandt. Little und Rubin (1987) datieren die früheste Anwendung (im medizinischen Bereich) auf das Jahr 1926. 1977 wurde der Name *EM-Algorithmus* von Dempster, Laird und Rubin eingeführt und einige grundlegende Resultate bereitgestellt.

Wir gehen nun von folgender Situation aus: Es liegt ein Modell für die vollständigen Daten y mit Dichte $f(y|\vartheta)$ vor. Dabei ist der Parameter ϑ unbekannt. Die Daten setzen sich aus einem beobachteten Teil y_o und einem fehlenden Teil y_m zusammen, $y = (y_o, y_m)$. Der Algorithmus erlaubt es nun die Likelihoodfunktion zum beobachteten Teil

$$\mathcal{L}(\vartheta|y_o) = \int f(y_o, y_m|\vartheta)\,dy_m \tag{4}$$

über ϑ zu maximieren.

Ein Vorteil des EM-Algorithmus ist die zuverlässige Konvergenz. Dempster, Laird und Rubin haben 1977 gezeigt, dass bereits unter sehr allgemeinen Voraussetzungen in jedem Iterationsschritt die Loglikelihoodfunktion vergrößert wird. Allerdings hängt die Konvergenzrate von der Anzahl der fehlenden Beobachtungen ab: Je mehr

Beobachtungen fehlen, desto langsamer konvergiert das Verfahren.

Wie bereits weiter oben angedeutet, besteht jeder Iterationsschritt aus zwei Stufen: Im **M-Schritt** wird eine Maximum-Likelihood Schätzung für ϑ basierend auf den vollständigen Daten durchgeführt. Im **E-Schritt** wird der bedingte Erwartungswert der Loglikelihoodfunktion $\ell(\vartheta|y)$ gegeben die beobachteten Daten und die aktuelle Schätzung $\vartheta^{(t)}$ aus dem M-Schritt berechnet. Dieser Erwartungswert wird schließlich im M-Schritt wieder dafür verwendet, eine neue Schätzung $\vartheta^{(t+1)}$ zu erhalten. Im E-Schritt werden also nicht direkt die fehlenden Daten y_m durch Approximationen ersetzt. Vielmehr wird die einzige Funktion, in der y_m vorkommt, im E-Schritt aktualisiert.

Wir betrachten nun den Zusammenhang

$$f(y_o|\vartheta) = \frac{f(y|\vartheta)}{f(y_m|y_o,\vartheta)} \tag{5}$$

und wenden den Logarithmus darauf an

$$\ell(\vartheta|y_o) = \ell(\vartheta|y) - \log f(y_m|y_o,\vartheta). \tag{6}$$

Wir erhalten somit eine Verbindung zwischem dem Likelihood der *beobachteten Daten* $\ell(\vartheta|y_o)$ und dem Likelihood der *vollständigen Daten* $\ell(\vartheta|y)$. Wir betrachten nun den konditionalen Erwartungswert von $\ell(\vartheta|y_o)$ gegeben die beobachteten Daten für irgendeinen Wert ϑ_0:

$$\begin{aligned}\mathbb{E}\left[\ell(\vartheta|y_o)\middle|y_o,\vartheta_0\right] &= \int \ell(\vartheta|y_o) f(y_m|y_o,\vartheta_0)\,dy_m \\ &= \ell(\vartheta|y_o)\int f(y_m|y_o,\vartheta_0)\,dy_m = \ell(\vartheta|y_o).\end{aligned}$$

Gemeinsam mit (6) ergibt sich

$$\ell(\vartheta|y_o) = \mathbb{E}\left[\ell(\vartheta|y)\middle|y_o,\vartheta_0\right] - \mathbb{E}\left[\log f(y_m|y_o,\vartheta)\middle|y_o,\vartheta_0\right].$$

Überlicherweise wird die Bezeichnung

$$Q(\vartheta|\vartheta_0) = \mathbb{E}\left[\ell(\vartheta|y)\middle|y_o,\vartheta_0\right] \tag{7}$$

eingeführt. Sei nun $\vartheta^{(t)}$ die aktuelle Schätzung für ϑ. Im E-Schritt wird die Funktion $Q(\vartheta|\vartheta^{(t)})$ berechnet,

$$Q(\vartheta|\vartheta^{(t)}) = \int \ell(\vartheta|y) f(y_m|y_o,\vartheta = \vartheta^{(t)})\,dy_m. \tag{8}$$

Der M-Schritt legt im Anschluss $\vartheta^{(t+1)}$ durch Maximieren dieser erwarteten Loglikelihoodfunktion $Q(\vartheta|\vartheta^{(t)})$ fest:

$$Q(\vartheta|\vartheta^{(t+1)}) \geqslant Q(\vartheta|\vartheta^{(t)}).$$

Es bleibt noch zu zeigen, dass das Maximieren von $Q(\vartheta|\vartheta^{(t)})$ tatsächlich gleichbedeutend mit dem Maximieren der Loglikelihoodfunktion $\ell(\vartheta|y_o)$ ist. Dies kann mit Hilfe der Jensen Ungleichung gezeigt und in Robert und Casella (1999), Theorem 5.3.4, nachgelesen werden. Es gilt also in jedem Schritt

$$\mathcal{L}(\vartheta^{(t+1)}|y_o) \geqslant \mathcal{L}(\vartheta^{(t)}|y_o).$$

Gleichheit tritt genau dann ein, wenn

$$Q(\vartheta^{(t+1)}|\vartheta^{(t)}) = Q(\vartheta^{(t)}|\vartheta^{(t)}).$$

Somit konvergiert der EM-Algorithmus gegen einen stationären Punkt von $\mathcal{L}(\vartheta|y_o)$, welcher jedoch nicht unbedingt die Maximum-Likelihood Schätzung sein muss. In der Praxis schlagen Robert und Casella (1999) deshalb vor, den EM-Algorithmus mehrmals mit zufällig ausgewählten Startpunkten laufen zu lassen.

In weiterer Folge werden wir immer wieder auf die Funktion $Q(\vartheta|\vartheta^{(t)})$ zurückgreifen. Dabei wird insbesondere die Darstellung in (9) von Nutzen sein. Aus (5) erhalten wir

$$f(y_m|y_o,\vartheta) = \frac{f(y|\vartheta)}{f(y_o|\vartheta)} = \frac{f(y_o|y_m,\vartheta)f(y_m|\vartheta)}{f(y_o|\vartheta)}$$

und daraus schließlich

$$\begin{aligned} Q(\vartheta|\vartheta^{(t)}) &= \int \ell(\vartheta|y) f(y_m|y_o,\vartheta^{(t)}) dy_m \\ &= \int \ell(\vartheta|y) \frac{f(y_o|y_m,\vartheta^{(t)})}{f(y_o|\vartheta^{(t)})} f(y_m|\vartheta^{(t)}) dy_m \\ &= \frac{1}{f(y_0|\vartheta^{(t)})} \int \ell(\vartheta|y) f(y_o|y_m,\vartheta^{(t)}) f(y_m|\vartheta^{(t)}) dy_m. \end{aligned} \quad (9)$$

2.2.2 *Gauß-Hermite-Quadratur*

Allgemein ist die Gauß-Quadratur eine Methode, um Integrale der Form

$$\int_a^b f(z)\omega(z)dz$$

durch endliche Summen zu approximieren, wobei $a < b$ und ω eine positive Gewichtsfunktion ist.
Im Speziellen werden wir später an der sogenannten *Gauß-Hermite-Quadratur* interessiert sein. Dabei ist das Intervall, über das integriert wird, gleich $(-\infty, \infty)$ und die Gewichtsfunktion $\omega(z) = e^{-z^2}$,

$$\int_{-\infty}^{\infty} f(z) \exp(-z^2) dx \approx \sum_{j=1}^{k} \pi_j f(z_j). \quad (10)$$

Für die Approximation ist also die Berechnung von k Massepunkten z_j und zugehörigen Gewichten π_j notwendig. Dabei sind die Massepunkte die Nullstellen des Hermitepolynoms k-ten Grades.

Prinzipiell gibt es zwei unterschiedliche Definitionen der Hermitepolynome: Einmal bezüglich der Gewichtsfunktion $\omega(z) = e^{-z^2}$ *(physikalische Hermitepolynome,* H_n*)* und einmal bezüglich der Gewichtsfunktion $\omega'(z) = e^{\frac{-z^2}{2}}$ *(probabilistische Hermitepolynome,* He_n*)*. Da in weiterer Folge über die Normalverteilungsdichte integriert wird, sind in unserem Fall insbesondere die probabilistischen Hermitepolynome von Interesse. Allerdings wird bei den üblichen R-Funktionen - `gauss.quad()` aus Smyth (2013) bzw. `gqz()` aus Einbeck et al. (2012) - mit den physikalischen Hermitepolynomen gerechnet. Dies ist jedoch sofort in den Griff zu bekommen, da ein einfacher Zusammenhang zwischen den beiden Polynomtypen (Abramowitz und Stegun, 1964) besteht,

$$\begin{aligned} He_n(z) &= 2^{-\frac{n}{2}} H_n\left(\frac{z}{\sqrt{2}}\right), \text{ bzw.} \\ H_n(z) &= 2^{\frac{n}{2}} He_n(\sqrt{2}z). \end{aligned}$$

Damit ergibt sich zusammen mit (10) und der Definition der Dichte der Standardnormalverteilung $\phi(\cdot)$

$$\int_{-\infty}^{\infty} f(z)\phi(z)dz = \int_{-\infty}^{\infty} f(z)\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}dz \approx \sum_{j=1}^{k} \frac{\sqrt{2}\pi_j}{\sqrt{2\pi}} f\left(\sqrt{2}z_j\right)$$

und weiter die Transformation für die neuen Massepunkte z_j' und Gewichte π_j',

$$z_j' = \sqrt{2}z_j \quad \text{und} \quad \pi_j' = \frac{\pi_j}{\sqrt{\pi}}.$$

3

ÜBERDISPERSIONSMODELLE

Bei der Modellschätzung kann es vorkommen, dass die tatsächliche Varianz der Daten größer ist, als jene, die sich aus der Verteilungsannahme ergibt. Dieses Phänomen wird als *Überdispersion*[1] bezeichnet. Gründe dafür können zum einen positive Korrelation innerhalb der Daten oder zum anderen das Fehlen wichtiger erklärender Variablen im Modell sein. Oft ist auch nicht bekannt, *welche* erklärenden Variablen zusätzlich nötig wären, um ein adäquates Modell zu erhalten. Um diesem Problem Rechnung zu tragen, wird im Überdispersionsmodell angenommen, dass es zusätzlich zu den bereits im Modell berücksichtigten erklärenden Variablen $x_i = (1, x_{i1}, \cdots, x_{i(p-1)})^\top$ weitere q-dimensionale Vektoren $z_i = (z_{i1}, \ldots, z_{iq})^\top$ mit unbeobachteten Variablen gibt und dass die eigentlichen linearen Prädiktoren die Form $\eta_i = g(\mu_i) = x_i^\top \beta + z_i^\top \gamma$ haben. Dabei ist $\gamma = (\gamma_1, \ldots, \gamma_q)^\top$ der Koeffizientenvektor zu den unbeobachteten Variablen. Zu beachten ist, dass $z_i^\top \gamma$ skalarwertig ist. Somit kann ohne Beschränkung der Allgemeinheit angenommen werden, dass

$$\eta_i = x_i^\top \beta + Z_i,$$

wobei Z_i eine Zufallsvariable - genannt zufälliger Effekt - mit einer (voerst) nicht näher spezifizierten Verteilung ist. Diese zufälligen Effekte werden als unabhängig und identisch verteilt angenommen. Weiters soll deren Verteilung nicht von den anderen erklärenden Variablen abhängen.

Im Unterschied zum GLM kommen hier zufällige Prädiktoren vor, die nicht mehr direkt den Erwartungswert $\mu_i = \mathbb{E}[y_i]$ modellieren, sondern den bedingten Erwartungswert gegeben die zufälligen Effekte, $\mu_i = \mathbb{E}[y_i|Z_i]$.

3.1 NORMALVERTEILTE ZUFÄLLIGE EFFEKTE

In diesem Abschnitt gehen wir davon aus, dass die zufälligen Effekte Z_i unabhängige, standardnormalverteilte Zufallsvariablen sind. Der lineare Prädiktor für die i-te Beobachtung ist gegeben als $\eta_i = x_i^\top \beta + \sigma Z_i$ und die marginalen Dichten von y_i als

$$f(y_i|\vartheta) = \int f(y_i, Z_i|\vartheta)\,dZ_i = \int f(y_i|Z_i, \vartheta) f(Z_i)\,dZ_i,$$

1 engl. overdispersion

mit $\vartheta = (\beta, \sigma)$. Auf Grund der Normalverteilungsannahme können die marginalen Dichten geschrieben werden als

$$f(y_i|\vartheta) = \int f(y_i|Z_i, \vartheta)\phi(Z_i)dZ_i,$$

wobei $\phi(\cdot)$ die Dichte der Standardnormalverteilung bezeichnet. Da die zufälligen Effekte als unabhängig angenommen wurden, ist die Likelihoodfunktion der Stichprobe

$$\mathcal{L}(\vartheta|y) = \prod_{i=1}^{n} \int f(y_i|Z_i, \vartheta)\phi(Z_i)dZ_i.$$

Das Integral ist im Allgemeinen[2] nicht analytisch lösbar. Die in Abschnitt 2.2.2 vorgestellte Gauß-Hermite-Quadratur liefert als Approximation

$$\mathcal{L}(\vartheta|y) \approx \prod_{i=1}^{n} \sum_{j=1}^{k} \pi_j f(y_i|z_j, \vartheta), \text{ bzw.}$$
$$\ell(\vartheta|y) \approx \sum_{i=1}^{n} \log \left(\sum_{j=1}^{k} \pi_j f(y_i|z_j, \vartheta) \right).$$

Die Likelihoodfunktion ist also approximativ die einer endlichen Mischverteilung[3] mit Dichten aus der Exponentialfamilie. Hier sind die Gewichte π_j und die Massepunkte z_j für vorgegebenes k jedoch bekannt und müssen nicht mehr geschätzt werden. Der lineare Prädiktor der i-ten Beobachtung in der j-ten Mischkomponente ist

$$\eta_{ij} = x_i^\top \beta + \sigma z_j.$$

Aus den allgemeinen Resultaten für endliche Mischverteilungen in Aitkin et al. (2009), Kapitel 7 folgt nun direkt

$$\frac{\partial \ell}{\partial \beta} = \sum_{i=1}^{n} \frac{\sum_{j=1}^{k} \pi_j f_{ij} \frac{\partial \log f_{ij}}{\partial \beta}}{\sum_{l=1}^{k} \pi_l f_{il}} = \sum_{i=1}^{n} \sum_{j=1}^{k} w_{ij} s_{ij}(\beta).$$

Dabei bezeichnen $s_{ij}(\beta)$ die Scorefunktion für β im GLM, $f_{ij} := f(y_i|z_j, \vartheta)$ und

$$w_{ij} = \frac{\pi_j f_{ij}}{\sum_{l=1}^{k} \pi_l f_{il}}. \tag{11}$$

Die Scorefunktion für β ist nach (3)

$$s_{ij}(\beta) = \frac{y_i - \mu_{ij}}{V_{ij} g'(\mu_{ij})} x_i.$$

2 Einzig im Falle von konjugierten Verteilungen (Aitkin et al., 2009, Abschnitt 8.2) ist dieses Integral analytisch lösbar.

3 Details dazu findet man zum Beispiel in Aitkin et al. (2009), Kapitel 7.

Für den Parameter σ zum Zufallseffekt erhält man auf ähnliche Weise

$$s_{ij}(\sigma) = \frac{y_i - \mu_{ij}}{V_{ij} g'(\mu_{ij})} z_j.$$

Durch Nullsetzen dieser beiden Funktionen ergeben sich Scoregleichungen, die sich nur durch die Gewichtung mit w_{ij} von denen eines GLMs unterscheiden. Aus dieser Beobachtung kann sofort ein EM-Algorithmus abgeleitet werden: Im E-Schritt werden die Parameter mittels gewichtetem GLM[4] geschätzt. Im M-Schritt werden daraufhin unter Verwendung der zuvor geschätzten Parameter die Gewichte w_{ij} nach (11) aktualisiert.
Man beachte, dass in jedem M-Schritt das folgende große, gewichtete GLM geschätzt werden muss (Friedl, 1998):

Tabelle 1: Erweiterte Modellmatrix - GQ

y	w	β			σ
y_1	w_{11}	x_{11}	$\dots$	x_{1p}	z_1
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
y_n	w_{n1}	x_{n1}	$\dots$	x_{np}	z_1
y_1	w_{12}	x_{11}	$\dots$	x_{1p}	z_2
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
y_n	w_{n2}	x_{n1}	$\dots$	x_{np}	z_2
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
y_1	w_{1k}	x_{11}	$\dots$	x_{1p}	z_k
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
y_n	w_{nk}	x_{n1}	$\dots$	x_{np}	z_k

Die Response hat nun die Länge nk. Der Parameter zur letzten Spalte der Designmatrix ist σ und somit die Standardabweichung zum zufälligen Effekt.

Die Gewichte w_{ij} haben noch eine andere Interpretation: Es kann gezeigt werden, dass sie die a posteriori Wahrscheinlichkeiten dafür sind, dass ein gegebenes y_i aus einer Verteilung mit Dichte f_{ij} stammt.

Zum Abschluss dieses Abschnitts geben wir noch die zentrale Funktion im EM-Algorithmus $Q(\vartheta|\vartheta^{(t)})$ an. Aus (9) folgt in unserem Fall

$$Q(\vartheta|\vartheta^{(t)}) \approx Q_k(\vartheta|\vartheta^{(t)}) = \sum_{i=1}^{n} \sum_{j=1}^{k} w_{ij}^{(t)} [\log \pi_j + \log f(y_i|z_j, \vartheta)]. \quad (12)$$

Nähere Details zur Herleitung sind in Abschnitt A.1 zu finden.

4 In der Funktion `glm()` gibt es das Argument `weights`, über das vordefinierte Gewichte als Liste übergeben werden können.

3.2 NICHTPARAMETRISCHE SCHÄTZUNG

Oftmals entbehrt die Normalverteilungsannahme jeglicher Grundlage. In diesem Fall sollte die Verteilung *nichtparametrisch* geschätzt werden. Wir betrachten wieder ein Modell mit zufälligen Effekten,

$$\eta_i = x_i^\top \beta + Z_i,$$

wobei dieses Mal die Dichte der zufälligen Effekte $f(Z_i)$ nicht bekannt ist. Die marginale Dichte kann analog zu vorher geschrieben werden als

$$f(y_i|\vartheta) = \int f(y_i|Z_i,\vartheta) f(Z_i) dZ_i, \tag{13}$$

dabei ist ϑ ein Vektor, der alle Parameter des Modells enhält. Da wir keine Annahmen über $f(Z_i)$ treffen, behandeln wir die Dichte als eine weitere Unbekannte, die geschätzt werden soll. Die Likelihoodfunktion ist nun mit $Z = (Z_1, \ldots, Z_n)^\top$

$$\mathcal{L}(\vartheta, f(Z)|y) = \prod_{i=1}^{n} \int f(y_i|Z_i,\vartheta) f(Z_i) dZ_i.$$

Laird (1978) hat gezeigt, dass unter sehr allgemeinen Voraussetzungen, die nichtparametrische Maximum Likelihoodschätzung der Verteilung der zufälligen Effekte diskret mit einer endlichen Anzahl an Massepunkten ist. Die geschätzte Verteilungsfunktion ist demnach eindeutig durch die Lokalisierung dieser Massepunkte $\hat{z}_j$ und deren Gewichtung $\hat{\pi}_j$ für $j = 1, \ldots, \hat{k}$ charakterisiert. Somit ist die *Profile-Likelihoodfunktion* gleich

$$\mathcal{P}(\vartheta|y) = \prod_{i=1}^{n} \left(\sum_{j=1}^{\hat{k}} f(y_i|\hat{z}_j,\vartheta)\hat{\pi}_j \right).$$

Dabei sind $\hat{z}_j, \hat{\pi}_j$ und $\hat{k}$[5] Funktionen in ϑ. Dies kann umformuliert werden zu

$$\mathcal{L}(\vartheta, k, \pi_1, \ldots, \pi_{k-1}, z_1, \ldots, z_k|y) = \prod_{i=1}^{n} \left(\sum_{j=1}^{k} f(y_i|z_j,\vartheta)\pi_j \right). \tag{14}$$

Hier kommen in der Liste der zu schätzenden Parameter nur die $k-1$ Gewichte $\pi_1, \ldots, \pi_{k-1}$ vor, da der Parameter π_k durch die Bedingung $\sum_{j=1}^{k} \pi_j = 1$ festgelegt ist.
Die Likelihoodfunktion in (14) ist beinahe dieselbe wie zuvor bei der

5 Der Parameter k wird im Allgemeinen nicht geschätzt, sondern a priori festgelegt. Durch eine a posteriori Analyse der Modellgüte für verschiedene Werte für k kann ein finaler Wert recht gut ermittelt werden. Darum wird ab nun nicht mehr die Schätzung $\hat{k}$ sondern vielmehr der feste Wert k verwendet.

Annahme von normalverteilten zufälligen Effekten mit dem Unterschied, dass die Gewichte und Massepunkte nicht bekannt sind, sondern ebenfalls geschätzt werden müssen. Dieses Schätzproblem kann wieder mit Hilfe des EM-Algorithmus gelöst werden. Dazu betrachten wir den linearen Prädiktor zur i-ten Beobachtung und zum j-ten Massepunkt,

$$\eta_{ij} = x_i^\top \beta + z_j.$$

Dieser kann durch die Einführung eines k-stufigen Faktors umgeschrieben werden zu

$$\eta_{ij} = x_i^\top \beta + z_1 \cdot 0 + \cdots + z_{j-1} \cdot 0 + z_j \cdot 1 + z_{j+1} \cdot 0 + \cdots + z_k \cdot 0.$$

Nun sind die unbekannten Massepunkte z_j die Slope-Parameter zu diesen Dummy-Variablen (k-dimensionale kanonische Einheitsvektoren). Dabei ist zu beachten, dass es sinnvoller ist, in der Matrix der bekannten erklärenden Variablen X keinen Intercept einzubinden, da nicht zwischen diesem und dem Parameter zur ersten Faktorstufe der Dummy-Variablen z_1 unterschieden werden kann. Ähnlich wie zuvor wird nun der Datensatz erweitert. In jedem M-Schritt müssen nun die Schätzungen zum folgenden gewichteten GLM berechnet werden (Friedl, 1998):

Tabelle 2: Erweiterte Modellmatrix - NPML

y	w	β			z			
y_1	w_{11}	x_{11}	...	x_{1p}	1	0	...	0
⋮	⋮	⋮		⋮	⋮	⋮		⋮
y_n	w_{n1}	x_{n1}	...	x_{np}	1	0	...	0
y_1	w_{12}	x_{11}	...	x_{1p}	0	1	...	0
⋮	⋮	⋮		⋮	⋮	⋮		⋮
y_n	w_{n2}	x_{n1}	...	x_{np}	0	1	...	0
⋮	⋮	⋮		⋮	⋮	⋮		⋮
y_1	w_{1k}	x_{11}	...	x_{1p}	0	0	...	1
⋮	⋮	⋮		⋮	⋮	⋮		⋮
y_n	w_{nk}	x_{n1}	...	x_{np}	0	0	...	1

Im darauffolgenden E-Schritt werden wie zuvor die Gewichte w_{ij} auf Grundlage der neuen Schätzungen - also nach (11) - aktualisiert. Darin kommen jedoch die nun unbekannten Parameter π_j vor. Dazu maximieren wir die logarithmierte Likelihoodfunktion in (14), ℓ, unter der Bedingung $\sum_{j=1}^{k} \pi_j = 1$ bzgl. dieser Parameter. Unter Ver-

wendung eines Lagrangemultiplikators λ ergibt sich nun mit den Notationen $\pi = (\pi_1, \ldots, \pi_{k-1})$ und $z = (z_1, \ldots, z_k)$

$$\frac{\partial}{\partial \pi_j}\left(\ell(\vartheta, k, \pi, z|y) - \lambda\left(\sum_{j=1}^{k}\pi_j - 1\right)\right) = \sum_{i=1}^{n}\frac{f_{ij}}{\sum_{l=1}^{k}\pi_l f_{il}} - \lambda$$
$$= \sum_{i=1}^{n}\frac{w_{ij}}{\pi_j} - \lambda.$$

Durch Nullsetzen und Summation über j erhält man

$$\lambda\sum_{j=1}^{k}\pi_j = \sum_{j=1}^{k}\sum_{i=1}^{n}w_{ij} = \sum_{i=1}^{n}\sum_{j=1}^{k}w_{ij}$$

und mit $\sum_{j=1}^{k} w_{ij} = 1$ und $\sum_{j=1}^{k}\pi_j = 1$ schließlich $\lambda = n$.
Die Parameter π_j können also über

$$\hat{\pi}_j = \frac{1}{n}\sum_{i=1}^{n}w_{ij}$$

geschätzt werden.

3.3 VORHERSAGEN

Nun stellt sich noch die Frage, wie in diesen Modellen Vorhersagen getroffen werden können. Dabei gibt es verschiedene Ansätze, die nun besprochen werden.

Zunächst betrachten wir das *konditionale Modell*. Wir wollen also eine Schätzung für $\mu_{ij} = \mathbb{E}[y_i|z_j]$ für jede Beobachtung y_i, $i = 1, \dots, n$, und jeden Massepunkt z_j, $j = 1, \dots, k$. Dies führt wegen $g(\mu_{ij}) = \eta_{ij}$ unmittelbar zu

$$\hat{\eta}_{ij} = x_i^\top \hat{\beta} + \hat{z}_j \text{ und} \tag{15}$$

$$\hat{\mu}_{ij} = g^{-1}(\hat{\eta}_{ij}). \tag{16}$$

Mit Hilfe dieser Schätzer kann auch das *marginale Modell*, also $\mathbb{E}[y_i]$, geschätzt werden. Dazu wird die marginale Dichte aus (13) benötigt. Die Gauß-Hermite Quadratur bzw. die nichtparametrische Schätzung liefert dann

$$\begin{aligned}
\mathbb{E}[y_i] &= \int y_i f(y_i|\vartheta)\,dy_i = \int y_i \int f(y_i|Z_i,\vartheta) f(Z_i)\,dZ_i\,dy_i \\
&\approx \int y_i \sum_{j=1}^{k} f(y_i|z_j,\vartheta)\pi_j\,dy_i = \sum_{j=1}^{k} \pi_j \int y_i f(y_i|z_j,\vartheta)\,dy_i \\
&= \sum_{j=1}^{k} \pi_j\, \mathbb{E}[y_i|z_j] = \sum_{j=1}^{k} \pi_j \mu_{ij}.
\end{aligned}$$

Aus dieser Überlegung resultieren Schätzer für die Erwartungswerte

$$\hat{\mu}_i = \hat{\mathbb{E}}[y_i] = \sum_{j=1}^{k} \hat{\pi}_j \hat{\mu}_{ij}$$

und für lineare Prädiktoren

$$\hat{\eta}_i = g(\hat{\mu}_i).$$

Nun werden Bayes Schätzer hergeleitet. Dazu sind die folgenden Überlegungen hilfreich. Für alle $i = 1, \dots, n$ gilt

$$f(y_i, Z_i|\vartheta) = f(Z_i|y_i,\vartheta) f(y_i|\vartheta), \tag{17}$$

$$f(y_i, Z_i|\vartheta) = f(y_i|Z_i,\vartheta) f(Z_i), \tag{18}$$

$$f(y_i|\vartheta) = \int f(y_i, Z_i|\vartheta)\,dZ_i. \tag{19}$$

Daraus schließt man

$$\begin{aligned}
f(Z_i|y_i,\vartheta) &\overset{(17)}{=} \frac{f(y_i, Z_i|\vartheta)}{f(y_i|\vartheta)} \overset{(18)}{=} \frac{f(y_i|Z_i,\vartheta) f(Z_i)}{f(y_i|\vartheta)} \\
&\overset{(19)}{=} \frac{f(y_i|Z_i,\vartheta) f(Z_i)}{\int f(y_i, Z_i|\vartheta)\,dZ_i} \overset{(18)}{=} \frac{f(y_i|Z_i,\vartheta) f(Z_i)}{\int f(y_i|Z_i,\vartheta) f(Z_i)\,dZ_i}.
\end{aligned}$$

Der *a posteriori Mean* eines Zufallseffekts Z_i gegeben die Beobachtung y_i ist definiert durch $\mathbb{E}[Z_i|y_i] = \int Z_i f(Z_i|y_i, \vartheta) dZ_i$. Mit Hilfe der vorigen Überlegung ergibt sich

$$\begin{aligned}\mathbb{E}[Z_i|y_i] &= \int Z_i f(Z_i|y_i, \vartheta) dZ_i = \int Z_i \frac{f(y_i|Z_i, \vartheta) f(Z_i)}{\int f(y_i|Z_i, \vartheta) f(Z_i) dZ_i} dZ_i \\ &\approx \sum_{j=1}^{k} z_j \frac{f(y_i|z_j, \vartheta)\pi_j}{\sum_{l=1}^{k} f(y_i|z_l, \vartheta)\pi_l}.\end{aligned}$$

Der Bruch in der letzten Zeile enspricht genau den Gewichten w_{ij} von zuvor. Es ist also naheliegend im a posteriori Mean den Bruch durch die im letzten EM-Schritt berechneten Gewichte und die z_j durch die zuletzt geschätzten Massepunkte zu ersetzten. Dadurch erhält man den *empirischen Bayes Schätzer*

$$\tilde{\mathbb{E}}[Z_i|y_i] = \sum_{j=1}^{k} \hat{z}_j \hat{w}_{ij}. \tag{20}$$

Daraus lassen sich nun unter Verwendung von (15), (16) und der Tatsache $\sum_{j=1}^{k} \hat{w}_{ij} = 1$ wieder Schätzungen auf beiden Skalen (Beobachtungen und linare Prädiktoren) angeben

$$\begin{aligned}\tilde{\eta}_i &= x_i^\top \hat{\beta} + \tilde{\mathbb{E}}[Z_i|y_i] = \sum_{j=1}^{k} \hat{w}_{ij} \left(x_i^\top \hat{\beta} + \hat{z}_j \right) = \sum_{j=1}^{k} \hat{w}_{ij} \hat{\eta}_{ij}, \\ \tilde{\mu}_i &= g^{-1}(\tilde{\eta}_i).\end{aligned}$$

Schlussendlich ist es möglich im Falle normalverteilter zufälliger Effekte und konditional Poissonverteilter Response mit dem kanonischen log-Link analytische Formeln für Schätzungen anzugeben. Aus der Turmeigenschaft und den bekannten Erwartungswerten für Poisson- bzw. lognormalverteilte Zufallsvariablen folgt

$$\mu_i = \mathbb{E}[y_i] = \mathbb{E}[\mathbb{E}[y_i|Z_i]] = \mathbb{E}[e^{x_i^\top \beta + \sigma Z_i}] = e^{x_i^\top \beta + \frac{1}{2}\sigma^2}$$

und damit

$$\begin{aligned}&\hat{\mu}_i = e^{x_i^\top \hat{\beta} + \frac{1}{2}\hat{\sigma}^2} \text{ bzw.} \\ &\hat{\eta}_i = x_i^\top \hat{\beta} + \frac{1}{2}\hat{\sigma}^2.\end{aligned}$$

4

ZUFÄLLIGE KOEFFIZIENTEN MODELLE

Die bisher betrachteten Überdispersionsmodelle sind ein wichtiger Spezialfall der größeren Modellklasse der *zufälligen Koeffizientenmodelle*[1] (engl. random coefficient models). Im Überdispersionsmodell haben wir sozusagen nur einen zufälligen Intercept erlaubt, weswegen Überdispersionsmodelle in diesem neuen Kontext auch als *zufällige Interceptmodelle* (engl. random intercept models) bezeichnet werden. Betrachten wir das Modell von zuvor mit linearen Prädiktoren

$$\eta_i^{intercept} = x_i^\top \beta + Z_i.$$

Der zufällige Effekt Z_i kann mit dem Modellintercept β_0 kombiniert werden, sodass wir einen *zufälligen Intercept* $B_{i0} = \beta_0 + Z_i$ erhalten. B_{i0} hat dann Erwartungswert β_0 und Varianz σ^2, die Varianz des zufälligen Effekts. Der lineare Prädiktor kann also äquivalent als

$$\eta_i^{intercept} = B_{i0} + (x'_{1_i})^\top \beta'_1$$

geschrieben werden, wobei β'_1 alle Parameter mit Ausnahme des Intercepts enthält und x'_{1_i} die i-te Zeile der Modellmatrix ohne Interceptspalte bezeichnet. Dieses Konzept lässt sich derart erweitern, dass nicht nur ein zufälliger Intercept, sondern auch ein zufälliger Slope-Parameter zugelassen wird. Der zufällige Parameter B_{i1} zur erkllärenden Variablen x_{1i} soll Erwartungswert β_1 haben und sozusagen durch eine Zufallsvariable U_i mit Erwartung 0 gestört werden, $B_{i1} = \beta_1 + U_i$. Es bezeichne X'_2 analog zu vorher die Modellmatrix ohne Interceptspalte und ohne erklärende Variable x_1 und β'_2 den Parametervektor ohne Intercept β_0 und β_1, den Parameter zu x_1. Damit kann der lineare Prädiktor für dieses neue Modell geschrieben werden als

$$\begin{aligned}
\eta_i &= B_{i0} + B_{i1}x_{i1} + (x'_{2_i})^\top \beta'_2 \\
&= \beta_0 + Z_i + (\beta_1 + U_i)x_{i1} + (x'_{2_i})^\top \beta'_2 \\
&= \beta_0 + \beta_1 x_{i1} + (x'_{2_i})^\top \beta'_2 + Z_i + U_i x_{i1} \\
&= \underbrace{x_i^\top \beta}_{\text{deterministische Terme}} + \underbrace{Z_i + U_i x_{i1}}_{\text{stochastische Terme}}.
\end{aligned}$$

Die gemeinsame Dichte von Z_i und U_i sei $f(Z_i, U_i|\vartheta)$. Die Likelihoodfunktion ist dann

$$\mathcal{L}(\vartheta|y) = \prod_{i=1}^{n} \int\int f(y_i|Z_i, U_i, \vartheta) f(Z_i, U_i|\vartheta) dZ_i dU_i, \tag{21}$$

wobei ϑ alle im Modell vorkommenden Parameter enthält.

1 vgl. Aitkin et al. (2009), Abschnitt 8.8

4.1 GAUSS-QUADRATUR

Wir nehmen nun an, dass $f(Z_i, U_i|\vartheta)$ die Dichte einer zweidimensionalen Normalverteilung mit unbekannter Kovarianzmatrix ist. Z_i und U_i sind im Allgemeinen nicht unabhängig, weswegen es nicht so einfach ist, die Integrale in der Likelihoodfunktion numerisch zu berechnen. Darum transformieren wir Z_i und U_i zu unabhängigen Zufallsvariablen. Seien σ_Z^2 und σ_U^2 die Varianzen von Z_i bzw. U_i für alle $i = 1, \ldots, n$. Der lineare Prädiktor ist

$$\eta_i = x_i^\top \beta + \sigma_Z Z_i + \sigma_U U_i x_{i1},$$

wobei von nun an Z_i und U_i bivariat normalverteilt sind mit Erwartungswert 0, Varianz 1 und Korrelation $\mathrm{corr}(Z_i, U_i) = \rho$. Nun transformieren wir jedes Z_i zu

$$Z_i^* := \frac{Z_i - \rho U_i}{\sqrt{1-\rho^2}}.$$

Somit ist jedes Z_i^* standardnormalverteilt. Der lineare Prädiktor kann umformuliert werden zu

$$\begin{aligned}\eta_i &= x_i^\top \beta + \sigma_Z \left(\rho U_i + \sqrt{1-\rho^2} Z_i^*\right) + \sigma_U U_i x_{i1} \\ &= x_i^\top \beta + \lambda_0 Z_i^* + (\lambda_1 + \lambda_2 x_{i1}) U_i.\end{aligned}$$

Die neuen Parameter stehen mit den ursprünglichen Parametern in folgendem Zusammenhang:

$$\begin{aligned}\lambda_0 &= \sigma_Z \sqrt{1-\rho^2}, \\ \lambda_1 &= \sigma_Z \rho, \\ \lambda_2 &= \sigma_U.\end{aligned}$$

Berechnet man Schätzungen für λ_0, λ_1 und λ_2, lassen sich auch direkt Schätzungen für die ursprünglichen Parameter ρ, σ_Z und σ_U ableiten, da

$$\begin{aligned}\rho &= \pm\sqrt{\frac{\lambda_1^2}{\lambda_0^2 + \lambda_1^2}}, \\ \sigma_Z &= \frac{\lambda_1}{\rho}, \\ \sigma_U &= \lambda_2.\end{aligned}$$

Durch die Bedingung $\sigma_Z > 0$ wird ρ eindeutig festgelegt. Es gilt also

$$\rho = \mathrm{sign}(\lambda_1)\sqrt{\frac{\lambda_1^2}{\lambda_0^2 + \lambda_1^2}}.$$

Mit dem Verschiebungssatz für Kovarianzen und der Definition des Korrelationskoeffizienten folgt (Details sind in Abschnitt A.2 zu finden)

$$\mathrm{Cov}[Z_i^*, U_i] = 0.$$

Somit sind Z_i^* und U_i unkorreliert, was im Falle der Normalverteilung äquivalent zur Unabhängigkeit ist. Die Likelihoodfunktion aus (21) kann also geschrieben werden als

$$\mathcal{L}(\vartheta|y) = \prod_{i=1}^{n} \int \left(\int f(y_i|Z_i^*, U_i, \vartheta) \phi(Z_i^*) dZ_i^* \right) \phi(U_i) dU_i,$$

mit $\vartheta = (\beta, \lambda_0, \lambda_1, \lambda_2)$. Sowohl das innere als auch das äußere Integral können durch Gauß-Hermite Quadratur approximiert werden,

$$\mathcal{L}(\vartheta|y) \approx \prod_{i=1}^{n} \sum_{j=1}^{k} \sum_{l=1}^{k} f(y_i|z_j, u_l, \vartheta) \pi_j \pi_l.$$

Somit ist der lineare Prädiktor zu jedem Massepunkt z_j, $j = 1, \ldots, k$ und u_l, $l = 1, \ldots, k$ gleich

$$\eta_{ijl} \approx x_i^\top \beta + \lambda_0 z_j + (\lambda_1 + \lambda_2 x_{i1}) u_l.$$

Es müssen also zusätzlich zu den Parametern zum Intercept und den erklärenden Variablen auch noch jene zu den Massepunkten z_j und u_l sowie zur Interaktion $x_{i1} : u_l$ berechnet werden. Diese Schätzungen erfolgen nun analog zum Überdispersionsmodell mittels EM-Algorithmus und gewichteten GLMs.

Noch zu bemerken ist, dass bei R zufälligen Koeffizienten für die Integration k^R Terme aufsummiert werden müssen, weshalb dieser Ansatz für eine große Anzahl an zufälligen Koeffizienten impraktikabel ist.

4.2 NICHTPARAMETRISCHE SCHÄTZUNG

Soll keine Verteilungsannahme getroffen werden, kann diese wieder nichtparametrisch geschätzt werden. Die Likelihoodfunktion aus (21) ist

$$\begin{aligned}\mathcal{L}(\vartheta|y) &= \prod_{i=1}^{n} \int \int f(y_i|Z_i, U_i, \vartheta) f(Z_i, U_i) dZ_i dU_i \\ &= \prod_{i=1}^{n} \int f(y_i|Z_i, U_i, \vartheta) f(Z_i, U_i) d\lambda \otimes \lambda(Z_i, U_i)\end{aligned}$$

mit Lebesgue-Maß λ. Die gemeinsame Verteilung ist nun wieder eine diskrete Verteilung auf $k < \infty$ zweidimensionalen Massepunkten (z_j, u_j). Zu den k geschätzten Massepunkten $(\hat{z}_j, \hat{u}_j)$ gehören wiederum geschätzte Massen $\hat{\pi}_j$. Somit ist die Likelihoodfunktion ähnlich wie zuvor

$$\mathcal{L}(\theta|y) = \prod_{i=1}^{n} \sum_{j=1}^{k} f(y_i|z_j, u_j, \vartheta) \pi_j$$

mit $\theta = (\vartheta, k, \pi_1, \ldots, \pi_{k-1}, (z_1, u_1), \ldots, (z_k, u_k))$. Zur Modellierung können wieder $2k$ Dummy-Variablen eingeführt werden, sodass der lineare Prädiktor $\eta_i = x_i^\top \beta + Z_i + U_i x_{i1}$ komponentenweise geschrieben werden kann als

$$\begin{aligned}\eta_{ij} = x_i^\top \beta + z_1 \cdot 0 + \cdots + z_{j-1} \cdot 0 + z_j \cdot 1 + z_{j+1} \cdot 0 + \cdots + z_k \cdot 0 \\ + (u_1 \cdot 0 + \cdots + u_{j-1} \cdot 0 + u_j \cdot 1 + u_{j+1} \cdot 0 + \cdots + u_k \cdot 0) x_{i1}.\end{aligned}$$

Die Modellmatrix zum gewichteten GLM sieht nun so aus:

Tabelle 3: Erweiterte Modellmatrix - zufällige Koeffizienten

y	w	β			z				u			
y_1	w_{11}	x_{11}	$\ldots$	x_{1p}	1	0	$\ldots$	0	x_{11}	0	$\ldots$	0
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$		$\vdots$
y_n	w_{n1}	x_{n1}	$\ldots$	x_{np}	1	0	$\ldots$	0	x_{n1}	0	$\ldots$	0
y_1	w_{12}	x_{11}	$\ldots$	x_{1p}	0	1	$\ldots$	0	0	x_{11}	$\ldots$	0
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$		$\vdots$
y_n	w_{n2}	x_{n1}	$\ldots$	x_{np}	0	1	$\ldots$	0	0	x_{n1}	$\ldots$	0
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$		$\vdots$
y_1	w_{1k}	x_{11}	$\ldots$	x_{1p}	0	0	$\ldots$	1	0	0	$\ldots$	x_{11}
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$		$\vdots$
y_n	w_{nk}	x_{n1}	$\ldots$	x_{np}	0	0	$\ldots$	1	0	0	$\ldots$	x_{n1}

Durch Hinzunahme weiterer zufälliger Koeffizienten steigt die Dimension der Modellmatrix enorm. Auch große Werte für k wirken sich noch deutlicher als im Überdispersionsmodell auf die Dimension aus.

5 VARIANZKOMPONENTENMODELLE

Oft ist die Annahme, dass sämtliche Beobachtungen eines Datensatzes unkorreliert sind, nicht angebracht. Es könnten etwa hierarchische Daten vorliegen, in denen es mehrere Gruppen gibt, die in gewisser Weise homogen sind. Man könnte sich dazu eine Untersuchung von SchülerInnen aus verschiedenen Klassen vorstellen. Die Ergebnisse von SchülerInnen, die aus derselben Klasse stammen, werden wohl in gewisser Weise *ähnlich* sein. So haben ja beispielsweise alle SchülerInnen in einer Klasse dieselben LehrerInnen, dieselbe Klassengröße, et cetera. Vergleicht man jedoch zwei SchülerInnen aus unterschiedlichen Klassen, haben diese nicht mehr dieselben Voraussetzungen. Es ist also oft angebracht, die Daten zu mehr oder weniger *homogenen Gruppen* zusammenzufassen. Die Zusammenfassung sollte so geschehen, dass es Korrelationen innerhalb der Gruppen gibt, aber keine zwischen den verschiedenen Gruppen.
In Überdispersionsmodellen wurden zufällige Effekte unter anderem dazu eingeführt, um positive Korrelationen innerhalb der Daten (siehe Kapitel 3) zu beschreiben. Es ist also naheliegend, dass nun nicht mehr für alle Beobachtungen eigene Zufallseffekte vegeben werden, sondern jeder homogenen Gruppe derselbe Zufallseffekt zugeordnet wird.

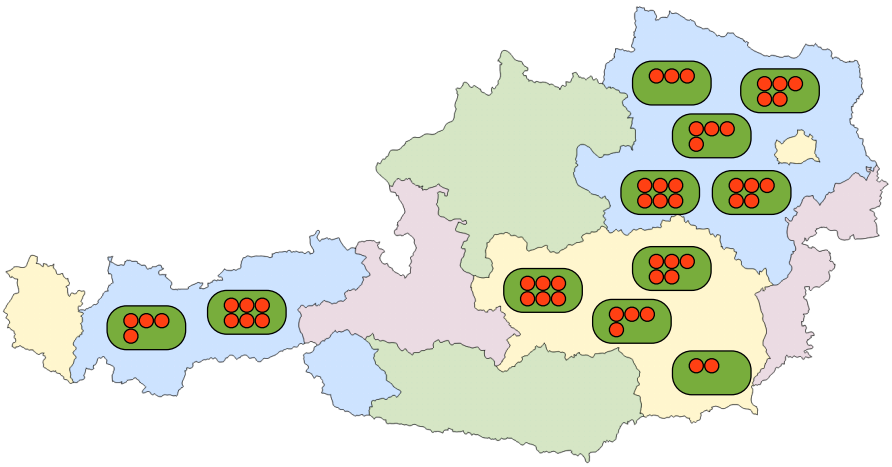

Abbildung 2: Untersuchung an Schulen[1]

1 Karte basierend auf `http://d-maps.com/carte.php?&num_car=17741&lang=de` von Daniel Dalet, 25.4.2013

Natürlich kann man sich auch weitere hierarchische Stufen überlegen: Beispielsweise könnte diese Untersuchung in mehreren Klassen an unterschiedlichen Schulen in verschiedenen Bundesländern durchgeführt werden. Dies wird in Abbildung 2 illustriert: Man könnte sich eine Untersuchung vorstellen, die in den Bundesländern Niederösterreich, Steiermark und Tirol durchgeführt wird. In jedem Bundesland werden unterschiedliche Anzahlen von Schulen (grüne Formen) mit jeweils unterschiedlich vielen Klassen (rote Kreise) analysiert. Nun sollten allen SchülerInnen aus einer Klasse, allen aus einer Schule und allen aus einem Bundesland ein gemeinsamer Zufallseffekt gegeben werden. Somit erhält jede und jeder SchülerIn drei Zufallseffekte.

Solch strukturierte Datensituationen können mit *Varianzkomponentenmodellen* erfasst werden. Zur Illustration beschränken wir uns nun auf Zwei-Stufen-Varianzkomponentenmodelle. Die Beobachtungen $y = (y_1, \ldots, y_n)$ werden in r Gruppen (*PSUs* - primary sampling units) zusammengefasst. Die i-te Gruppe für $i = 1, \ldots, r$ enthält n_i Mitglieder (*SSUs* - secondary sampling units), welche mit Hilfe des Index $j = 1, \ldots, n_i$ gezählt werden. Insgesamt gilt natürlich $\sum_{i=1}^{r} n_i = n$. Zu bemerken ist, dass das klassische Überdispersionsmodell ein Spezialfall der Varianzkomponentenmodelle mit $n_i = 1$ für alle i und $r = n$ ist. Oft wird r als *Clusteranzahl* und n_i als *Clustergröße* bezeichnet. Die r Cluster wären im vorherigen Beispiel etwa die verschiedenen Klassen und die Clustergrößen n_i die jeweilige Anzahl von SchülerInnen. Die erklärenden Variablen können sowohl auf dem Niveau der PSUs (x_{1i}) als auch auf dem der SSUs (x_{2ij}) gemessen werden. In unserem Beispiel könnte als erklärende Variable in x_{1i} etwa die Anzahl der Mathematikstunden pro Woche in der i-ten Klasse enthalten sein. In x_{2ij} könnten hingegen die Noten der letzten Mathematikschularbeit der j-ten SchülerIn vorkommen.
Die linearen Prädiktoren sind nun im Allgemeinen gegeben durch

$$\eta_{ij} = x_{ij}^\top \beta + Z_i, \quad Z_i \overset{iid}{\sim} F,$$

wobei F die Verteilung der Zufallseffekte mit Erwartungswert 0 ist und

$$x_{ij}^\top \beta = x_{1i}^\top \beta_1 + x_{2ij}^\top \beta_2.$$

5.1 NORMAL-NORMAL MODELL

Als Spezialfall - vgl. Aitkin et al. (2009), Abschnitt 9.2 - betrachten wir ein Modell, in dem sowohl die Beobachtungen als auch die Zufallseffekte normalverteilt sind. Die Linkfunktion sei $g(\cdot) = \mathrm{id}(\cdot)$,

$$\begin{aligned} y_{ij}|Z_i &\overset{iid}{\sim} N(\mu_{ij}, \sigma^2), \\ \mu_{ij} &= \eta_{ij} = x_{ij}^\top \beta + \sigma_Z Z_i, \\ Z_i &\overset{iid}{\sim} N(0,1). \end{aligned} \tag{22}$$

Nun kann das Modell als lineares Modell geschrieben werden: Definiere $\tilde{\varepsilon}_{ij} \overset{iid}{\sim} N(0, \sigma^2)$ und setze

$$\varepsilon_{ij} := \sigma_Z Z_i + \tilde{\varepsilon}_{ij}.$$

Dabei sind die $\tilde{\varepsilon}_{ij}$ unabhängig von den zufälligen Effekten Z_i. Damit gilt weiter $\varepsilon_{ij} \sim N(0, \sigma_Z^2 + \sigma^2)$ und

$$y_{ij} = x_{ij}^\top \beta + \varepsilon_{ij}.$$

Also ist $y_{ij} \sim N(x_{ij}^\top \beta, \sigma_Z^2 + \sigma^2)$. Für die Kovarianzen gilt (siehe Abschnitt A.3 für eine detaillierte Herleitung)

$$\mathrm{Cov}[y_{ij}, y_{kl}] = \sigma_Z^2 \mathbb{E}[Z_i Z_k].$$

und somit insgesamt

$$\mathrm{Cov}[y_{ij}, y_{kl}] = \begin{cases} \sigma_Z^2, & \text{falls } i = k \text{ und } j \neq l \\ 0, & \text{falls } i \neq k,\ j \text{ beliebig.} \end{cases}$$

Für $i = k$ und $j \neq l$ ist der Korrelationskoeffizient ungleich 0 und wird nun mit ρ bezeichnet,

$$\rho := \mathrm{corr}(y_{ij}, y_{il}) = \frac{\sigma_Z^2}{\sigma_Z^2 + \sigma^2}.$$

Wir haben also genau den Effekt erzielt, den wir erreichen wollten: Korrelation innerhalb der Gruppen, aber keine zwischen den Gruppen!

Wir führen noch eine andere Parametrisierung des Modells ein:

$$\begin{aligned} y_{ij}|Z_i &\sim N(x_{2ij}^\top \beta_2 + Z_i, \sigma^2), \\ Z_i &\sim N(x_{1i}^\top \beta_1, \sigma_Z^2). \end{aligned} \tag{23}$$

In dieser Parametrisierung wird die Zuordnung der erklärenden Variablen x_{1i} und x_{2ij} zu den jeweiligen Stufen des Modells klarer. Auffallend ist auch, dass in der Parametrisierung (22) σ_Z^2 als Regressionsparameter und in der Parametrisierung (23) hingegen als

Varianz auftritt. Aitkin et al. (2009) verweisen weiters darauf, dass die Konvergenzrate beim Modellfit mittels EM-Algorithmus bei der Parametrisierung (23) von σ_Z abhängt, bei der Parametrisierung (22) jedoch nicht. Weiters wird bemerkt, dass der EM-Algorithmus beim Normal-Normal-Modell nicht der schnellste Algorithmus ist. Für Modelle, bei denen nicht beide Male die Normalverteilung angenommen wird, sticht er jedoch heraus, da er sehr effizient und einfach zu programmieren ist.

Das Integral in der Likelihoodfunktion - siehe auch (24) - ist im Falle der Normal-Normal-Modelle analytisch lösbar und muss demnach nicht approximiert werden. Da dies jedoch ein seltener Spezialfall ist, wird darauf nicht näher eingegangen. Details dazu findet man beispielsweise in Aitkin et al. (2009), Abschnitt 9.2.

5.2 GAUSS-QUADRATUR

Nun gehen wir davon aus, dass die Verteilung der $y_{ij}|Z_i$ aus der Exponentialfamilie stammt. Die Verteilung der Zufallseffekte Z_i sei wieder die Standardnormalverteilung. Im Weiteren verwenden wir die Parametrisierung (22). Die Matrix der erklärenden Variablen enthalte nun sowohl die Variablen zu den PSUs als auch zu den SSUs und gegebenenfalls deren Interaktionen. Eine Spalte in der Matrix, die eine Variable x_i zu den PSUs enthält, hat die Gestalt

$$(\underbrace{x_1,\ldots,x_1}_{n_1-mal},\underbrace{x_2,\ldots,x_2}_{n_2-mal},\ldots,\underbrace{x_r,\ldots,x_r}_{n_r-mal})^\top,$$

während eine, die eine Variable x_{ij} zu den SSUs enthält, folgende Struktur hat

$$(x_{11},\ldots,x_{1n_1},x_{21},\ldots,x_{2n_2},\ldots,x_{r1},\ldots,x_{rn_r})^\top.$$

Unter diesen Annahmen folgt

$$f(y_i|Z_i,\vartheta) = \prod_{j=1}^{n_i} f(y_{ij}|Z_i,\vartheta),$$
$$f(y_i,Z_i|\vartheta) = f(y_i|Z_i,\vartheta)\phi(Z_i),$$

und damit die Likelihoodfunktion

$$\mathcal{L}(\vartheta|y) = \prod_{i=1}^{r}\left[\int\left(\prod_{j=1}^{n_i} f(y_{ij}|Z_i,\vartheta)\right)\phi(Z_i)\,dZ_i\right]. \tag{24}$$

Das Integral wird nun durch Gauß-Hermite Quadratur approximiert. Dazu verwenden wir wieder k Massepunkte z_l mit zugehörigen Gewichten π_l. Die Likelihoodfunktion ist nun annähernd

$$\mathcal{L}(\vartheta|y) \approx \prod_{i=1}^{r}\left[\sum_{l=1}^{k}\left(\prod_{j=1}^{n_i} f(y_{ij}|z_l,\vartheta)\right)\pi_l\right].$$

Wie bereits in Abschnitt 3.1 handelt es sich bei der Likelihoodfunktion approximativ um eine Likelihoodfunktion endlicher Mischungen von Exponentialfamiliendichten mit bekannten Massepunkten z_l und Mischungsanteilen π_l. Die linearen Prädiktoren sind nun

$$\eta_{ijl} = x_{ij}^\top \beta + \sigma_Z z_l.$$

Ähnlich wie zuvor - siehe Aitkin et al. (2009), Abschnitt 9.3 - erhält man

$$\begin{aligned} \frac{\partial \ell}{\partial \beta} &= \sum_{i=1}^{r} \sum_{j=1}^{n_i} \sum_{l=1}^{k} w_{il} s_{ijl}(\beta), \text{ mit} \\ w_{il} &= \frac{\pi_l m_{il}}{\sum_{m=1}^{k} \pi_m m_{im}} \end{aligned} \tag{25}$$

mit der Abkürzung

$$m_{il} = \prod_{j=1}^{n_i} f(y_{ij}|z_l, \vartheta)$$

und den Scorefunktionen

$$\begin{aligned} s_{ijl}(\beta) &= \frac{(y_{ij} - \mu_{ijl}) x_{ij}}{V_{ijl} g'(\mu_{ijl})} \text{ und} \\ s_{ijl}(\sigma_Z) &= \frac{(y_{ij} - \mu_{ijl}) z_l}{V_{ijl} g'(\mu_{ijl})}. \end{aligned}$$

Durch Nullsetzen der Scorefunktionen erhält man die Scoregleichungen eines gewichteten GLMs. Der EM-Algorithmus besteht nun abermals aus abwechslendem Aktualisieren der Gewichte und Schätzen der Parameter im GLM.

5.3 NICHTPARAMETRISCHE SCHÄTZUNG

Wie in Abschnitt 3.2 wird nun keine Annahme zur Verteilung der zufälligen Effekte getroffen, sondern diese nichtparametrisch geschätzt. Die linearen Prädiktoren sind

$$\eta_{ijl} = x_{ij}^\top \beta + z_l,$$

mit unbekannten Massepunkten z_l, $l = 1, \ldots, k$. Die Likelihoodfunktion ist wieder

$$\mathcal{L}(\vartheta, k, \pi_1, \ldots, \pi_{k-1}, z_1, \ldots, z_k | y) = \prod_{i=1}^{r} \left(\sum_{l=1}^{k} m_{il} \pi_l \right).$$

Der Parameter k wird wie bereits zuvor als bekannte Größe angenommen und sequentiell erhöht, bis das Maximum erreicht wird. Um die Gewichte π_l zu schätzen, wird die Loglikelihoodfunktion $\ell(\theta|y)$, $\theta = (\vartheta, \pi_1, \ldots, \pi_{k-1}, z_1, \ldots, z_k)$, unter der Nebenbedingung $\sum_{l=1}^{k} \pi_l = 1$ maximiert. Es gilt

$$\ell(\theta|y) = \sum_{i=1}^{r} \log \left(\sum_{l=1}^{k} m_{il} \pi_l \right).$$

Wir führen wieder einen Lagrangemultiplikator λ ein:

$$\frac{\partial}{\partial \pi_l} \left(\ell(\theta|y) - \lambda \left[\sum_{m=1}^{k} \pi_m - 1 \right] \right) = \sum_{i=1}^{r} \frac{w_{il}}{\pi_l} - \lambda.$$

Durch Nullsetzen und Summation über $l = 1, \ldots, k$ ergibt sich analog zu Abschnitt 3.2

$$\hat{\pi}_l = \frac{1}{r} \sum_{i=1}^{r} w_{il}. \tag{26}$$

Anzumerken bleibt, dass wir hier implizit immer von gleich großen Clustern ausgehen. Kommen unterschiedliche Größen vor, müsste beispielsweise in der Schätzung von π eine mit n_i gewichtete Summe berechnet werden.

5.4 ERWEITERUNGEN UND ALTERNATIVE ANSÄTZE

Grundsätzlich andere Ansätze - also nicht mittels nichtparametrischer Schätzung bzw. Approximation durch Gauß-Quadratur - werden in Aitkin et al. (2009), Abschnitt 9.4, kurz zusammengefasst. Dort werden auch die jeweiligen Quellen angegeben.

Die eben beschriebenen Modelle können als *Zufällige Interceptmodelle* aufgefasst werden. Ganz analog zu *Überdispersionsmodellen* können diese auf *Modelle mit zufälligen Koeffizienten* erweitert werden. Dies wird in Aitkin et al. (2009) kurz in Abschnitt 9.6 angerissen.

Teil II

MODELLE BASIEREND AUF DER NEGATIVEN BINOMIALVERTEILUNG

Möchte man Anzahlen modellieren, wird dazu oftmals das Poissonmodell mit dem kanonischen log-Link herangezogen. Jedoch ist dieses Modell recht starr, da Erwartungswert und Varianz stets gleich sind. Diese Einschränkung ist in vielen Anwendungen nicht angebracht und man möchte mehrparametrige Verteilungen für Anzahlen - wie etwa die Negative Binomialverteilung - verwenden. Allerdings ist diese Verteilung *kein* Mitglied der einparametrigen, linearen Exponentialfamilie, weshalb die bisherigen Überlegungen nicht direkt angewandt werden können.

6

DAS NEGATIV BINOMIALMODELL

Um die Theorie für Modelle basierend auf einer Negativen Binomialverteilung aufzubauen, führen wir nun zunächst den Begriff der *konjugierten Verteilung* ein.

6.1 KONJUGIERTE VERTEILUNGEN

Zu jeder Verteilung aus der Exponentialfamilie gibt es eine konjugierte Verteilung für den kanonischen Parameter ϑ (Aitkin et al., 2009). So ist die konjugierte Verteilung zu Z vom selben Typ wie die von $Y|Z$. Daraus resultieren dann oftmals einfache Formen der marginalen Verteilungen von Y. Zum Beispiel gilt für $Y|Z \sim N(\vartheta + Z, \sigma^2)$ und $Z \sim N(\mu, \varphi^2)$, dass $Y \sim N(\vartheta + \mu, \sigma^2 + \varphi^2)$.
Darüber hinaus sind auch solche Mischungen von großem Interesse, bei denen sowohl Z als auch $Y|Z$ und Y bekannte Verteilungen haben, die analytisch leicht handhabbar sind. Gerade bei Modellen mit zufälligen Effekten liegt die Hauptschwierigkeit in der Berechnung von Integralen der Form (siehe auch (13))

$$f(y|\vartheta) = \int f(y|Z, \vartheta) f(Z) dZ.$$

Betrachtet man zum Beispiel die Zusammensetzung $Y|Z \sim N(\mu, \sigma^2/Z)$ und $Z \sim \Gamma(r/2, 2/r)$, folgt, dass $(Y - \mu)/\sigma$ t-verteilt ist mit r Freiheitsgraden. Aus dem Spezialfall $Y|Z \sim N(0, 1/Z)$ resultiert also als marginale Verteilung von Y direkt die t-Verteilung mit r Freiheitsgraden.

6.2 DAS MODELL

Wir betrachten nun eine Mischung aus der Poisson- und Gammaverteilung und werden schließlich feststellen, dass daraus als marginale Verteilung die Negative Binomialverteilung resultiert. Die folgenden Ausführungen basieren auf Booth et al. (2003).
Wir verwenden die Dichte der Gammaverteilung in der folgenden Parametrisierung,

$$f(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x),$$

dabei bezeichnen $\Gamma(\cdot)$ die Gammafunktion, $\alpha > 0$ und $\beta > 0$ die Parameter. Die Gammafunktion ist definiert durch

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt.$$

Als Momente erhält man für $X \sim \Gamma(\alpha, \beta)$

$$\mathbb{E}[X] = \frac{\alpha}{\beta} \text{ und}$$
$$\mathbb{V}ar[X] = \frac{\alpha}{\beta^2}.$$

Nun bauen wir das gewünschte Modell auf. Es sei dazu $(Z_1, \ldots, Z_n)$ eine Zufallsstichprobe aus der Gammaverteilung mit Erwartungswert 1, das bedeutet

$$\forall\, i \in \{1, \ldots, n\} : \; Z_i \overset{iid}{\sim} \Gamma(\alpha, \alpha).$$

Weiters seien Y_i weitere Zufallsvariablen, sodass $Y_i|Z_i$ Poissonverteilt ist mit Parameter $Z_i\mu_i$,

$$f(y_i|Z_i, \mu_i) = \frac{(\mu_i Z_i)^{y_i}}{y_i!} e^{-\mu_i Z_i}.$$

Nun betrachten wir die marginale Dichte der Y_i. Diese ist

$$f(y_i|\mu_i) = \frac{\Gamma(y_i + \alpha)}{\Gamma(\alpha) y_i!} \left(\frac{\mu_i}{\mu_i + \alpha}\right)^{y_i} \left(\frac{\alpha}{\mu_i + \alpha}\right)^{\alpha}.$$

Eine detaillierte Rechnung ist in Abschnitt A.4 zu finden. Wir setzten nun $p_i = \frac{\alpha}{\mu_i + \alpha}$. Dann gilt $1 - p_i = \frac{\mu_i}{\mu_i + \alpha}$. Nach Abramowitz und Stegun (1964) gilt für $u, v \in \mathbb{N}$

$$\Gamma(v+1) = v! \quad \text{und} \quad \binom{u}{v} = \frac{\Gamma(u+1)}{\Gamma(v+1)\Gamma(u-v+1)}.$$

Setzt man $u = y_i + \alpha - 1$ und $v = y_i$, resultiert

$$\frac{\Gamma(y_i + \alpha)}{\Gamma(\alpha) y_i!} = \binom{y_i + \alpha - 1}{y_i},$$

woraus für $\alpha,\ y_i \in \mathbb{N}$ schließlich

$$f(y_i|\mu_i) = \binom{y_i + \alpha - 1}{y_i} p_i^{\alpha} (1 - p_i)^{y_i}$$

folgt. Dies ist die Wahrscheinlichkeitsfunktion einer Negativen Binomialverteilung. Wir schreiben von nun an $Y_i \overset{ind}{\sim} \mathrm{NegBin}(\alpha, \mu_i)$. Aus der Turmeigenschaft ergeben sich die ersten beiden Momente für Y_i (in Abschnitt A.4 ist deren Herleitung angegeben),

$$\mathbb{E}[Y_i] = \mu_i, \tag{27}$$
$$\mathbb{V}ar[Y_i] = \mu_i + \frac{\mu_i^2}{\alpha}. \tag{28}$$

Im Unterschied zum Poissonmodell kommt zur Varianz noch ein positiver Term hinzu. Über diesen zusätzlichen Term $\frac{\mu_i^2}{\alpha}$ ist es somit

möglich Überdispersion aus dem Poissonmodell in den Griff zu bekommen. Hier gibt der Parameter α den Grad der Überdispersion an. Ein kleiner Wert für α steht für große Überdispersion, wohingegen $\alpha = \infty$ keine Überdispersion induziert.

Nun kann ein loglineares Modell mit Negativ Binomialverteilter Response formuliert werden. Seien $Y_i \overset{ind}{\sim} \mathtt{NegBin}(\alpha, \mu_i)$. Der Erwartungswert μ_i sei über einen log-Link mit den erklärenden Variablen x_i verbunden, $\log \mu_i = x_i^\top \beta$, wobei β wieder den Parametervektor bezeichnet.

6.3 MAXIMUM LIKELIHOOD SCHÄTZUNG

Für das eben vorgestellte Modell wollen wir nun die Scoregleichungen für α und β herleiten. Diese Berechnungen sind auch in Booth et al. (2003), Anhang A, zu finden. Die Loglikelihoodfunktion für $y = (y_1, \ldots, y_n)$ ist gegeben durch

$$\begin{aligned}\ell(\alpha, \beta | y) &= \log \left[\prod_{i=1}^{n} \frac{\Gamma(y_i + \alpha)}{\Gamma(\alpha) y_i!} \left(\frac{\mu_i}{\mu_i + \alpha} \right)^{y_i} \left(\frac{\alpha}{\mu_i + \alpha} \right)^{\alpha} \right] \\ &= n[\alpha \log \alpha - \log \Gamma(\alpha)] + \sum_{i=1}^{n} \Big[\log \Gamma(y_i + \alpha) - \log y_i! \\ &\quad + y_i \log \left(\frac{\mu_i}{\mu_i + \alpha} \right) - \alpha \log(\mu_i + \alpha) \Big].\end{aligned}$$

Die Loglikelihoodfunktion muss nun bezüglich der Parameter (α, β) maximiert werden. Wir führen die folgenden Notationen ein

$$\begin{aligned}&\ell_\alpha(\alpha, \beta) = \frac{\partial \ell(\alpha, \beta | y)}{\partial \alpha}, \\ &\ell_\beta(\alpha, \beta) = \frac{\partial \ell(\alpha, \beta | y)}{\partial \beta}, \text{ sowie} \\ &\ell_{\alpha\alpha}(\alpha, \beta),\ \ell_{\alpha\beta}(\alpha, \beta) \text{ und } \ell_{\beta\beta}(\alpha, \beta) \text{ analog.}\end{aligned}$$

Die Schätzfunktion nach α ist

$$\begin{aligned}\ell_\alpha(\alpha, \beta) &= n(\log \alpha + 1 - \psi(\alpha)) \\ &\quad + \sum_{i=1}^{n} \left[\psi(y_i + \alpha) - \log(\mu_i + \alpha) - \frac{y_i + \alpha}{\mu_i + \alpha} \right], \qquad (29)\end{aligned}$$

dabei bezeichnet $\psi(z) = \frac{\partial}{\partial z} \log \Gamma(z)$. Die erste Scoregleichung ergibt sich schließlich durch Nullsetzen von (29).
Um die zweite Scoregleichung zu erhalten, muss die Loglikelihood-

funktion nach β abgeleitet werden. Dabei beachte man den Zusammenhang $\log \mu_i = x_i^\top \beta$:

$$\begin{aligned}\ell_\beta(\alpha,\beta) &= \frac{\partial}{\partial\beta}\sum_{i=1}^{n}\left[y_i x_i^\top\beta - y_i \log\left(e^{x_i^\top\beta}+\alpha\right) - \alpha\log\left(e^{x_i^\top\beta}+\alpha\right)\right] \\ &= \sum_{i=1}^{n}\left[y_i - \frac{\alpha+y_i}{\mu_i+\alpha}\mu_i\right]x_i = \sum_{i=1}^{n}\left[\frac{(y_i-\mu_i)x_i}{1+\frac{\mu_i}{\alpha}}\right]. \qquad (30)\end{aligned}$$

Für $\alpha \to \infty$ ist (30) identisch der Scoregleichung eines Generalisierten Linearen Modells basierend auf der Poissonverteilung.

Nun berechnen wir noch die zweiten Ableitungen. Dabei bezeichnet $\psi'(z) = \frac{\partial^2}{\partial z^2}\log\Gamma(z)$.

$$\begin{aligned}\ell_{\alpha\alpha}(\alpha,\beta) &= n\left(\frac{1}{\alpha}-\psi'(\alpha)\right)+\sum_{i=1}^{n}\left[\psi'(y_i+\alpha)-\frac{1}{\mu_i+\alpha}-\frac{\mu_i-y_i}{(\mu_i+\alpha)^2}\right], \\ \ell_{\alpha\beta}(\alpha,\beta) &= \sum_{i=1}^{n}\left[\frac{\mu_i(y_i-\mu_i)}{(\mu_i+\alpha)^2}\right]x_i, \\ \ell_{\beta\beta}(\alpha,\beta) &= -\alpha\sum_{i=1}^{n}\left[\frac{\mu_i(y_i+\alpha)}{(\mu_i+\alpha)^2}\right]x_i x_i^\top.\end{aligned}$$

Zur besseren Übersicht führen wir eine $n \times n$ Diagonalmatrix $D(\alpha,\beta)$ ein, welche als Diagonaleinträge für $i = 1,\dots,n$

$$d_{ii} = \frac{\alpha\mu_i(y_i+\alpha)}{(\mu_i+\alpha)^2}$$

enthält. Mit X bezeichnen wir wie üblich die Designmatrix. Damit lässt sich $\ell_{\beta\beta}(\alpha,\beta)$ umschreiben zu

$$\ell_{\beta\beta}(\alpha,\beta) = -X^\top D(\alpha,\beta)X.$$

Als nächstes berechnen wir die Fisher-Information für β.

$$\mathcal{I}_\beta(\alpha,\beta) = -\mathbb{E}\left[\ell_{\beta\beta}(\alpha,\beta)\right] = X^\top\mathbb{E}\left[D(\alpha,\beta)\right]X = X^\top W(\alpha,\beta)X,$$

wobei $W(\alpha,\beta)$ eine $n \times n$ Diagonalmatrix mit Diagonaleinträgen $i = 1,\dots,n$

$$w_{ii} = \mathbb{E}\left[d_{ii}\right] = \frac{\alpha\mu_i}{(\mu_i+\alpha)^2}\left(\mathbb{E}\left[y_i\right]+\alpha\right) = \frac{\alpha\mu_i}{\mu_i+\alpha} = \frac{\mu_i}{1+\frac{\mu_i}{\alpha}}.$$

Für $\alpha \to \infty$ entspricht w_{ii} der Varianz μ_i im log-linearen Poissonmodell.

6.4 NUMERISCHE VERFAHREN

In diesem Abschnitt wollen wir nun verschiedene Methoden vorstellen, um die Scoregleichungen numerisch zu lösen. In all diesen Methoden müssen die in den Scoregleichungen vorkommenden

Funktionen $\psi(z)$ bzw. $\psi'(z)$ berechnet werden. Die Funktion $\psi(z) = \frac{\partial}{\partial z} \log \Gamma(z)$ ist die sogenannte *Digammafunktion*, während $\psi'(z)$ als *Trigammafunktion* bezeichnet wird. Nach Abramowitz und Stegun (1964) können allgemein alle *Polygammafunktionen* über

$$\psi^{(n)}(z) = \frac{d^n}{dz^n}\psi(z) = \frac{d^{n+1}}{dz^{n+1}} \log \Gamma(z) = (-1)^{n+1} \int_0^\infty \frac{t^n e^{-zt}}{1 - e^{-t}} dt$$

berechnet werden. In `R` (R Core Team, 2012) stehen im Basispaket dafür die Funktionen `gamma(z)`, `digamma(z)` und `trigamma(z)` zur Verfügung. Weiters können mittels der Funktion `psigamma(z, deriv=0)` auch höhere Ableitungen von $\psi(z)$ durch Spezifizierung des Ableitungsgrades über `deriv` berechnet werden.

6.4.1 *Newton-Raphson*

Die wohl intuitivste Methode ist die direkte Anwendung des mehrdimensionalen *Newton-Raphson Algorithmus*:

$$\begin{pmatrix} \alpha^{(t+1)} \\ \beta^{(t+1)} \end{pmatrix} = \begin{pmatrix} \alpha^{(t)} \\ \beta^{(t)} \end{pmatrix} - \begin{pmatrix} \ell_{\alpha\alpha} & \ell_{\alpha\beta} \\ \ell_{\alpha\beta}^\top & \ell_{\beta\beta} \end{pmatrix}^{-1} \begin{pmatrix} \ell_\alpha \\ \ell_\beta \end{pmatrix},$$

wobei die Loglikelihoodableitungen auf der rechten Seite in $\left(\alpha^{(t)}, \beta^{(t)}\right)$ betrachtet werden.
Ein Nachteil dieser Methode ist das Fehlen eines naheliegenden Startwerts (eine mögliche Wahl wird in Abschnitt 6.4.2 dargestellt). Weiters tritt bei einer direkten Schätzung gelegentlich das Phänomen auf, dass negative Werte für α generiert werden. Dies lässt sich durch die Parametrisierung $\alpha^* = \exp(\alpha)$ vermeiden.

6.4.2 *Iteratively Reweighted Least Squares*

Wird der Parameter α als fest angenommen, lässt sich die Schätzung des Parametervektors β auf die Schätzung eines GLMs zurückführen. Dies ist möglich, da die Negative Binomialverteilung für ein fixes α Mitglied der einparametrigen, linearen Exponentialfamilie (vgl. Definition 1) ist. Die Varianzfunktion ist durch

$$V(\mu_i) = \mu_i \left(1 + \frac{\mu_i}{\alpha}\right)$$

gegeben - siehe zum Beispiel Hilbe (2011) für die Herleitung. Als Pseudobeobachtung resultiert der $(n \times 1)$-dimensionale Vektor $z(\beta)$, dessen i-te Komponete als $\log \mu_i + \frac{y_i - \mu_i}{\mu_i}$ gegeben ist. Die Iterationsvorschrift ist schließlich

$$\beta^{(t+1)} = \left(X^\top W(\alpha^{(t)}, \beta^{(t)}) X\right)^{-1} X^\top W(\alpha^{(t)}, \beta^{(t)}) z(\beta^{(t)}).$$

Wurde $\beta^{(t+1)}$ berechnet, wird im Anschluss ein neuer Wert für α durch den eindimensionalen Newton-Raphson Algorithmus berechnet. Dabei wird $\beta = \beta^{(t+1)}$ fixiert:

$$\alpha^{(t+1)} = \alpha^{(t)} - \frac{\ell_\alpha(\alpha^{(t)}, \beta^{(t+1)})}{\ell_{\alpha\alpha}(\alpha^{(t)}, \beta^{(t+1)})}.$$

Für $\alpha = \infty$ entspricht - wie bereits diskutiert - das Negative Binomialmodell einem Poissonmodell mit log-Link. Es ist also naheliegend als Startwerte $\beta^{(0)}$ die im Poissonmodell geschätzten Koeffizienten zu verwenden.
Für α wäre demzufolge der Startwert $\alpha^{(0)} = \infty$ angemessen, der jedoch nicht brauchbar ist. Alternativ könnte für $\alpha^{(0)}$ schlicht eine *große* Zahl verwendet werden. Eine weitere Möglichkeit besteht darin, α zunächst durch die Momentenmethode zu schätzen und das Ergebnis als Startwert heranzuziehen. Wir nehmen an, dass die Beobachtungen unabhängig und identisch verteilt sind mit Erwartungswert μ. Weiters gehen wir davon aus, dass die ersten beiden Momente durch (27) und (28) gegeben sind:

$$\begin{aligned} \bar{y} &= \frac{1}{n}\sum_{i=1}^{n} y_i \stackrel{!}{=} \mu, \\ s^2 &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 \stackrel{!}{=} \mu\left(1 + \frac{\mu}{\alpha}\right). \end{aligned}$$

Daraus ergibt sich schließlich der Startwert

$$\alpha^{(0)} = \frac{\bar{y}^2}{s^2 - \bar{y}}.$$

Es ist durchaus möglich, dass die empirische Varianz s^2 kleiner ist als das arithmetische Mittel $\bar{y}$. Dies würde zu einem negativen Startwert führen, weshalb man besser

$$\alpha^{(0)} = \max\left(\frac{\bar{y}^2}{s^2 - \bar{y}}, 0\right)$$

setzt.

In R (R Core Team, 2012) steht im Paket MASS (Ripley und Venables, 2002) die Funktion `theta.ml()` zur Verfügung, welche die Schätzung von α durchführt, zur Verfügung. Weiters existiert ebenfalls im Paket MASS die Funktion `glm.nb()`, welche Negativ Binomialmodelle nach der eben beschriebenen Methode schätzt. Darüber hinaus ist eine Implementierung dieser Methode in Abschnitt B.1 zu finden.

7

NORMALVERTEILTE ZUFÄLLIGE EFFEKTE

7.1 DAS ALLGEMEINE MODELL

Ähnlich wie bereits in Kapitel 3, Kapitel 4 bzw. Kapitel 5 können nun auch hier wieder zufällige Effekte zu den linearen Prädiktoren hinzugefügt werden. Nun betrachten wir gleich das allgemeine Varianzkomponentenmodell. Dazu seien y_{ij} für $i = 1, \ldots, r$ und $j = 1, \ldots, n_i$ die Beobachtungen. Für $n_i > 1$ für mindestens ein i sprechen wir wieder von clusterspezifischen und für $n_i = 1$ für alle i von beobachtungsspezifischen zufälligen Effekten. Die zufälligen Effekte werden mit Z_i bezeichnet. Da wir neben zufälligen Intercepts auch zufällige Parameter zulassen, sind die zufälligen Effekte q-dimensionale Zufallsvektoren. Dabei ist die Anzahl der zufälligen Steigungsparameter gleich $q - 1$. Weiters fassen wir die Beobachtungen zu $y_i = (y_{i1}, \ldots, y_{in_i})^\top$ zusammen. Es bezeichnen x_{ij} und u_{ij} erklärende Variablen zu y_{ij} und β den p-dimensionalen Parametervektor. Schlussendlich soll noch gelten, dass

$$\begin{aligned} y_{ij}|Z_i &\overset{ind}{\sim} \mathrm{NegBin}(\alpha, \mu_{ij}) \quad \text{und} \\ \mu_{ij} &= \mathbb{E}\left[y_{ij}|Z_i\right] = \exp\left(x_{ij}^\top \beta + u_{ij}^\top Z_i\right). \end{aligned}$$

7.2 NORMALVERTEILTE ZUFÄLLIGE EFFEKTE

Zunächst betrachten wir normalverteilte zufällige Effekte. Diese sind insbesondere von Interesse, da die Kovarianzstruktur flexibel aufgebaut werden kann. Es gelte nun also

$$Z_1, \ldots, Z_r \overset{iid}{\sim} N_q(0, \Sigma),$$

wobei Σ die Varianz-Kovarianzmatrix bezeichnet. Die Kovarianzstruktur sei bekannt (diese ist ja durch die Clusterung vorgegeben), die jeweiligen Varianzen jedoch nicht. Den Vektor der unbekannten Varianzen bezeichnen wir mit σ^2. Damit gilt $\sigma^2 = \mathtt{diag}(\Sigma)$ und wir schreiben $\Sigma = \Sigma(\sigma^2)$. Wir fassen nun alle zu schätzenden Parameter zu $\vartheta = (\alpha, \beta^\top, (\sigma^2)^\top)^\top$ zusammen.
Wie wir bereits bei den *Zufälligen Koeffizientenmodellen* in Kapitel 4 gesehen haben, ergeben sich bei der Maximum-Likelihood Schätzung r Integrale der Dimension q, welche im Allgemeinen auch numerisch nur durch großen Aufwand gelöst werden können. Bei hochdimensionalen Problemen eignen sich deshalb Monte Carlo Methoden besser.

7.3 DER MONTE CARLO EM ALGORITHMUS

Der von Wei und Tanner (1990) bzw. Booth und Hobert (1999) entworfene Monte Carlo EM Algorithmus (MCEM) umschifft das Problem der Berechnung hochdimensionaler Integrale. Erinnern wir uns zurück an den deterministischen EM-Algorithmus aus Abschnitt 2.2.1: Dabei muss im t-ten Schritt die Funktion

$$Q(\vartheta|\vartheta^{(t)}) = \mathbb{E}\left[\ell(\vartheta|y, Z)\middle|y, \vartheta^{(t)}\right] = \int \ell(\vartheta|y, Z) f(Z|y, \vartheta = \vartheta^{(t)}) dZ$$

bezüglich ϑ maximiert werden. $f(Z|y, \vartheta = \vartheta^{(t)})$ berechnet sich über

$$f(Z|y, \vartheta^{(t)}) = \frac{f(y, Z|\vartheta^{(t)})}{f(y|\vartheta^{(t)})}. \tag{31}$$

Der Nenner in (31) ist nichts anderes als die Likelihoodfunktion $\mathcal{L}(\vartheta^{(t)}|y)$, die jedoch nicht bekannt ist. Ebenso steht auch die gemeinsame Dichte $f(y, Z|\vartheta^{(t)})$ nicht dirket zur Verfügung. Eine analytische Lösung ist somit im Allgemeinen nicht möglich. Im Falle einer kleinen Dimension q der Zufallseffekte könnte das Integral wieder mittels Gauß-Hermite Quadratur approximiert werden. Für höhere Dimensionen wird dieser Schritt - der E-Schritt im EM-Algorithmus - nun durch eine Monte Carlo Simulation ersetzt.
Dazu muss zunächst eine Stichprobe $z_{t1}, \ldots, z_{tm}$ aus der konditionalen Verteilung $f(Z|y, \vartheta^{(t)})$ gezogen werden. Details dazu sind in Algorithmus 1 zu finden. Daraufhin wird die Funktion $Q(\vartheta|\vartheta^{(t)})$ durch

$$\hat{Q}(\vartheta|\vartheta^{(t)}) = \frac{1}{m}\sum_{l=1}^{m} \ell(\vartheta|y, z_{tl}) \tag{32}$$

approximiert und bezüglich ϑ maximiert. (32) zerfällt in zwei Teile, die getrennt voneinander maximiert werden können,

$$\hat{Q}(\vartheta|\vartheta^{(t)}) = \frac{1}{m}\sum_{l=1}^{m}\sum_{i=1}^{r}\sum_{j=1}^{n_i} \ell(\alpha, \beta|y_{ij}, z_{itl}) + \frac{1}{m}\sum_{l=1}^{m}\sum_{i=1}^{r} \ell(\Sigma|y_{ij}, z_{itl}).$$

Der erste Teil hängt nun ausschließlich von den Parametern α und β ab, wohingegen der zweite Teil nur den Parameter Σ mit den unbekannten Varianzen σ^2 enthält. Der vordere Teil kann mittels Newton-Raphson Algorithmus optimiert werden. Die dafür nötigen Ableitungen sind in Booth et al. (2003), Appendix B, zu finden. Zu bemerken ist, dass $\ell(\Sigma|y_{ij}, z_{itl}) = \log \phi(z_{itl}|\Sigma)$. Da sowohl über l als auch über i aufsummiert wird, ist der zweite Teil gleich der Loglikelihoodfunktion zu einer Zufallsstichprobe der Länge $r \cdot m$ aus der $N_q(0, \Sigma)$ - Verteilung. Als Schätzung für Σ ergibt sich im $t+1$-ten Schritt deswegen

$$\Sigma^{(t+1)} = \frac{1}{mr}\sum_{l=1}^{m}\sum_{i=1}^{r} z_{itl} z_{itl}^{\top}.$$

Robert und Casella (2010) weisen darauf hin, dass beim MCEM Algorithmus eine Fehleranalyse durgeführt werden sollte. Booth und Hobert (1999) haben dafür eine Approximation des Monte Carlo Fehlers hergeleitet.

7.3.1 *Verwerfungsmethode (Rejection Sampling)*

Booth und Hobert (1999) schlagen zur Gewinnung einer Zufallsstichprobe aus der konditionalen Verteilung die Verwerfungsmethode vor. Wir bezeichnen nun die konditionale Verteilung zu $Z|y$ mit $h(Z)$. Angenommen diese kann geschrieben werden als

$$h(Z) = c \cdot g(Z)e(Z),$$

wobei c eine normalisierende Konstante ist, die nicht von Z abhängt, und $e(\cdot)$ eine Dichtefunktion ist, aus der Stichproben gezogen werden können. Mit Algorithmus 1 kann nun aus $h(\cdot)$ simuliert werden.

Algorithmus 1 Verwerfungsmethode

Schritt 1:
Simuliere z aus e.
Simuliere unabhängig davon w aus $U(0,1)$.

Schritt 2:
Setzte $\tau = \sup_z\{g(z)\}$.
if $w \leqslant \frac{g(z)}{\tau}$ **then**
  Akzeptiere z.
else
  Gehe zu Schritt 1.
end if

Dieser Algorithmus geht auf Geweke (1996) zurück. Booth und Hobert (1999) weisen darauf hin, dass die Berechnung des Supremums in Schritt 2 äquivalent zur Bestimmung des Maximum Likelihood Regressionsparameters in einem GLM mit Offset ist. Weiters erklären Booth und Hobert (1999) in Abschnitt 4.1, dass es nicht immer notwendig ist, in jedem Schritt ein neues τ zu berechnen. Falls die Akzeptanzrate sehr gering und damit der Algorithmus sehr langsam ist, kann auf *Importance Sampling* (Booth und Hobert, 1999, Abschnitt 4.2) zurückgegriffen werden.
In unserem Fall ergibt sich wegen

$$f(Z|y,\vartheta) = \prod_{i=1}^{r} f(Z_i|y_i,\vartheta)$$

für alle $i = 1, \ldots, r$ die Zerlegung

$$\begin{aligned} f(Z_i|y_i, \vartheta) &= \frac{f(y_i|Z_i, \alpha, \beta)\varphi(Z_i, \Sigma)}{f(y_i|\vartheta)} \\ &= c \cdot \varphi(Z_i, \Sigma) \prod_{j=1}^{n_i} \left(\frac{\alpha}{\mu_{ij} + \alpha}\right)^{\alpha} \left(\frac{\mu_{ij}}{\mu_{ij} + \alpha}\right)^{y_{ij}}. \end{aligned}$$

Die Konstante enthält nun neben dem Binomialkoeffizienten aus der Negativ Binomialverteilung auch noch den von Z unabhängigen Term $\frac{1}{f(y_i|\vartheta)}$. Aus Gründen der Übersichtlichkeit wurde bei ϑ auf die Superskripts (t) verzichtet. In der obigen Notation gilt nun

$$e(Z_i) = \varphi(Z_i, \vartheta) \quad \text{und} \quad g(Z_i) = \prod_{j=1}^{n_i} \left(\frac{\alpha}{\mu_{ij} + \alpha}\right)^{\alpha} \left(\frac{\mu_{ij}}{\mu_{ij} + \alpha}\right)^{y_{ij}}.$$

Um im Schritt 2 das Supremum τ zu berechnen, muss also das Maximierungsproblem

$$\sup_{z_i} \prod_{j=1}^{n_i} \left(\frac{\alpha}{\mu_{ij} + \alpha}\right)^{\alpha} \left(\frac{\mu_{ij}}{\mu_{ij} + \alpha}\right)^{y_{ij}} \tag{33}$$

gelöst werden. (33) ist äquivalent zur Maximierung von

$$\begin{aligned} &\log\left(\prod_{j=1}^{n_i} \left(\frac{\alpha}{\mu_{ij} + \alpha}\right)^{\alpha} \left(\frac{\mu_{ij}}{\mu_{ij} + \alpha}\right)^{y_{ij}}\right) \\ &= n_i \alpha \log(\alpha) + \sum_{j=1}^{n_i} y_{ij} \log(\mu_{ij}) - \sum_{j=1}^{n_i} (\alpha + y_{ij}) \log(\mu_{ij} + \alpha). \end{aligned}$$

Man beachte, dass dies die Likelihoodfunktion des Negativ Binomialmodells ohne die von Z unabhängigen Terme ist. Die Berechnung von τ wird also tatächlich auf die Berechnung von Maximum Likelihood Schätzern zurückgeführt. Da $\mu_{ij} = \exp(x_{ij}^\top \beta + u_{ij}^\top Z_i)$, ist der Offset $x_{ij}^\top \beta$.
Booth et al. (2003) schlagen vor, dieses Optimierungsproblem mittels Newton-Raphson Verfahren zu lösen. Es ergeben sich die folgenden Ableitungen

$$\begin{aligned} \frac{\partial \log g(Z_i)}{\partial Z_i} &= \alpha \sum_{j=1}^{n_i} z_{ij} \left[\frac{y_{ij} - \mu_{ij}}{\mu_{ij} + \alpha}\right], \\ \frac{\partial^2 \log g(Z_i)}{\partial Z_i^2} &= -\sum_{j=1}^{n_i} z_{ij} \left[\frac{\alpha \mu_{ij}(\alpha + y_{ij})}{(\mu_{ij} + \alpha)^2}\right] z_{ij}^\top. \end{aligned}$$

8

NICHTPARAMETRISCHE SCHÄTZUNG

8.1 DAS MODELL

Nun lassen wir die Normalverteilungsannahme fallen und gehen nur noch davon aus, dass $Z = (Z_1, \ldots, Z_r)^\top$ eine Zufallsstichprobe aus einer unbekannten Verteilung mit Dichte $f(Z)$ sind. Die Likelihoodfunktion ist nun wie zuvor

$$\mathcal{L}(\alpha, \beta, f(Z)|y) = \prod_{i=1}^{r} \int f(y_i|Z_i, \alpha, \beta) f(Z_i) dZ_i.$$

Wieder können wir davon ausgehen, dass f die Massefunktion einer diskreten Verteilung an k Massepunkten $z = (z_1, \ldots, z_k)^\top$ mit zugehörigen Gewichten $\pi = (\pi_1, \ldots, \pi_k)^\top$ ist. Wir legen k a priori fest. Der Vektor der zu schätzenden Größen ist somit $\vartheta = (\alpha, \beta^\top, z^\top, \pi^\top)^\top$.

Wir betrachten nun den EM-Algorithmus. Wir bezeichnen mit $\alpha^{(t)}, \beta^{(t)}$ und $g^{(t)} = \left(z_l^{(t)}, \pi_l^{(t)}\right)_{l=1}^{k}$ die Ergebnisse des t-ten EM-Schritts. Wie zuvor - vgl. (25) - sind die a posteriori Wahrscheinlichkeiten für $l = 1, \ldots, k$ gegeben als

$$w_{il}^{(t)} = \frac{\pi_l^{(t)} \prod_{j=1}^{n_i} f(y_{ij}|Z_i, \alpha^{(t)}, \mu_{ijl}^{(t)})}{\sum_{m=1}^{k} \pi_m^{(t)} \left[\prod_{j=1}^{n_i} f(y_{ij}|Z_i, \alpha^{(t)}, \mu_{ijm}^{(t)})\right]},$$

wobei $f(y_{ij}|Z_i, \alpha^{(t)}, \beta^{(t)})$ die Dichtefunktion der $\mathrm{NegBin}(\alpha^{(t)}, \mu_{ijl}^{(t)})$-Verteilung mit $\mu_{ijl}^{(t)} = \exp(x_{ij}^\top \beta + u_{ij}^\top z_l^{(t)})$ bezeichnet. Es ist nun möglich, $Q_k(\vartheta|\vartheta^{(t)})$ in geschlossener Form darzustellen:

$$\begin{aligned} Q_k(\vartheta|\vartheta^{(t)}) &= \sum_{i=1}^{r} \sum_{j=1}^{n_i} \sum_{l=1}^{k} w_{il}^{(t)} \left(\log f(y_{ij}|z, \alpha, \mu_{ijl}) + \log \pi_l\right) \\ &= \sum_{i=1}^{r} \sum_{j=1}^{n_i} \sum_{l=1}^{k} w_{il}^{(t)} \log f(y_{ij}|z, \alpha, \mu_{ijl}) + \sum_{i=1}^{r} \sum_{l=1}^{k} w_{il}^{(t)} \log \pi_l. \end{aligned}$$

Die beiden Teile von $Q_k(\vartheta|\vartheta^{(t)})$ können nun getrennt voneinander maximiert werden. Wir betrachten zunächst den zweiten Teil. Wie bereits in den vorgegangenen Kapiteln wird π unter der Bedingung $\sum_{l=1}^{k} \pi_l = 1$ optimiert. Dazu führen wir abermals einen Lagrangemultiplikator λ ein. Es resultiert

$$\frac{\partial}{\partial \pi_l} \left[\sum_{i=1}^{r} \sum_{l=1}^{k} w_{il}^{(t)} \log \pi_l - \lambda \left(\sum_{m=1}^{k} \pi_m - 1\right)\right] = \frac{1}{\pi_l} \sum_{i=1}^{r} w_{il}^{(t)} - \lambda.$$

Nullsetzten führt zur bereits bekannten Formel

$$\pi_l^{(t+1)} = \frac{1}{r}\sum_{i=1}^{r} w_{il}^{(t)}.$$

Bei der Optimierung bzgl. α, β und z ist nur der erste Teil von $Q_k(\vartheta|\vartheta^{(t)})$ relevant. Zunächst berechnen wir die Ableitung nach α. Wegen $\sum_{l=1}^{k} w_{il}^{(t)} = 1$ gilt mit $\sum_{i=1}^{r} n_i = n$

$$\begin{aligned}
&\frac{\partial}{\partial\alpha}\sum_{i=1}^{r}\sum_{j=1}^{n_i}\sum_{l=1}^{k} w_{il}^{(t)} \log f(y_{ij}|z,\alpha,\mu_{ijl}) \\
&= \sum_{i=1}^{r}\sum_{j=1}^{n_i}\sum_{l=1}^{k} w_{il}^{(t)} \frac{\partial}{\partial\alpha}\log\left[\frac{\Gamma(y_{ij}+\alpha)}{\Gamma(\alpha)y_{ij}!}\left(\frac{\mu_{ijl}}{\mu_{ijl}+\alpha}\right)^{y_{ij}}\left(\frac{\alpha}{\mu_{ijl}+\alpha}\right)^{\alpha}\right] \\
&= \sum_{i=1}^{r}\sum_{j=1}^{n_i}\sum_{l=1}^{k} w_{il}^{(t)}\left[\psi(y_{ij}+\alpha) - \frac{y_{ij}+\alpha}{\mu_{ijl}+\alpha} - \log(\mu_{ijl}+\alpha)\right] \\
&\quad + n\left[\log\alpha + 1 - \psi(\alpha)\right]. \qquad (34)
\end{aligned}$$

Beim Vergleich mit (29) erkennt man, dass es sich bei (34) um die gewichtete Scoregleichung bezüglich α aus dem Negativ Binomialmodell handelt. Diese kann also mittels der in Kapitel 6 vorgestellten Methoden gelöst werden.
Die Scorefunktion für β ergibt sich als

$$\frac{\partial}{\partial\beta}\sum_{i=1}^{r}\sum_{j=1}^{n_i}\sum_{l=1}^{k} w_{il}^{(t)} \log f(y_{ij}|z,\alpha,\mu_{ijl}) = \sum_{i=1}^{r}\sum_{j=1}^{n_i}\sum_{l=1}^{k} w_{il}^{(t)}\left[\frac{y_{ij}-\mu_{ijl}}{1+\frac{\mu_{ijl}}{\alpha}}\right] x_{ij}. \qquad (35)$$

Für $l = 1, \ldots, k$ erhalten wir das System

$$\frac{\partial}{\partial z}\sum_{i=1}^{r}\sum_{j=1}^{n_i}\sum_{l=1}^{k} w_{il}^{(t)} \log f(y_{ij}|z,\alpha,\mu_{ijl}) = \sum_{i=1}^{r}\sum_{j=1}^{n_i}\sum_{l=1}^{k} w_{il}^{(t)}\left[\frac{y_{ij}-\mu_{ijl}}{1+\frac{\mu_{ijl}}{\alpha}}\right] u_{ij}. \qquad (36)$$

(35) und (36) sind wiederum gewichtete Versionen der Scorefunktionen aus dem Negativ Binomialmodell. Zu beachten ist, dass hier eine erweiterte Modellmatrix (vgl. beispielsweise Tabelle 2) resultiert, was einen deutlich größeren Rechenaufwand zur Folge hat.

Teil III

MODELLSCHÄTZUNG IN R

In diesem Kapitel werden R Pakete vorgestellt, mit deren Hilfe die in den vorherigen Kapiteln diskutierten Modelle geschätzt werden können.

9 R PAKETE

9.1 DAS PAKET npmlreg

Das R-Paket npmlreg *(Nonparametric maximum likelihood estimation for random effect models)* wurde von Jochen Einbeck, Ross Darnell und John Hinde (Einbeck et al., 2012) entwickelt und ermöglicht den Umgang mit Modellen basierend auf der Exponentialfamilie. Konkret stehen zur Zeit als konditionale Verteilungen die Normal-, Poisson-, Binomial- und Gammaverteilung zur Verfügung. Prinzipiell ist sowohl eine nichtparametrische Schätzung sowie die Approximation mittels Gauß-Hermite Quadratur möglich.

Die beiden Funktionen `alldist()` und `allvc()` stellen das Herzstück des Pakets dar. Mittels `alldist()` können *Zufällige Koeffizientenmodelle* (und natürlich der häufige Spezialfall der *Überdispersionsmodelle*) geschätzt werden, wohingegen `allvc()` für *Varianzkomponentenmodelle* gedacht ist.

Eine große Einschränkung besteht darin, dass bei Varianzkomponentenmodellen nur eine einzige Hierarchiestufe modellierbar ist. Weiters ist im Falle normalverteilter zufälliger Effekte nur die Modellierung eines zufälligen Intercepts möglich - zufällige Parameter können nur nichtparametrisch geschätzt werden.

Im nächsten Teil dieser Arbeit werden Methoden vorgestellt, um Standardfehler der zu schätzenden Parameter zu erhalten. Dies mag zunächst als überflüssig erscheinen, da `alldist()` und `allvc()` bereits Standardfehler zur Verfügung stellen. Diese können über einen `summary()`-Aufruf auf ein Objekt, das von einer der beiden Funktionen erzeugt wurde, abgerufen werden. Was hier jedoch als Standardfehler eines Modells mit zufälligen Effekten angegeben wird, sind schlicht die Standardfehler des im letzten EM-Schritt berechneten GLMs. Diese unterschätzen die wahren Standardfehler zumeist, da die zusätzliche, durch den EM-Algorithmus generierte Unsicherheit nicht beachtet wird.

9.2 DAS PAKET gamlss.mx

Das Paket `gamlss.mx` ist ein Zusatzpaket zu `gamlss` (Rigby und Stasinopoulos, 2005). Dieses Paket enthält die Funktion `gamlssNP()`, welche die Funktionen `alldist()` und `allvc()` aus dem Paket

npmlreg zusammenfasst und deutlich erweitert. Als konditionale Verteilung stehen sämtlich Verteilungen aus der sogenannten gamlss.family zur Verfügung. Unter anderem ist auch die Negative Binomialverteilung in dieser Familie enthalten. Eine vollständige Liste erlangt man durch den Aufruf ?gamlss.family.

Zusätzlich zur größeren Verteilungsauswahl ist mit gamlssNP() auch die Methode der Gauß-Quadratur zur Modellierung zufälliger Parameter möglich.

9.3 DIE FUNKTION glm.nb()

Im Paket MASS (Ripley und Venables, 2002) steht die Funktion glm.nb() zur Verfügung, welche das Negativ Binomialmodell aus Kapitel 6 ohne zufällige Effekte schätzt.

Teil IV

STANDARDFEHLER

Ein großer Nachteil bei der Verwendung des EM-Algorithmus ist jener, dass nicht automatisch eine Schätzung der Varianz-Kovarinazmatrix von $\hat{\beta}$ zur Verfügung steht. Im folgenden Abschnitt wird die von Friedl und Kauermann (2000) entwickelte Methode vorgestellt, um Standardfehler in Exponentialfamilienmodellen zu berechnen. Ergänzend dazu wird ein Konnex zum *Missing Information Principle* von Louis (1982) hergestellt. Als Abschluss dieses Teils leiten wir für zwei konkrete Modelle die Formeln für die Berechnung der Standardfehler her und geben Hinweise zu deren Implementierung in R.

10

EXPONENTIALFAMILIENMODELLE

10.1 NORMALVERTEILTE ZUFÄLLIGE EFFEKTE

Im ersten Schritt gehen wir von normalverteilten zufälligen Effekten aus. Wie für Überdispersionsmodelle (vgl. (12)) gilt auch für den Fall der Varianzkomponentenmodelle

$$Q_k(\vartheta|\vartheta^{(t)}) = \sum_{i=1}^{r}\sum_{l=1}^{k} w_{il}^{(t)}[\log \pi_l + \log f(y_i|z_l,\vartheta)].$$

Hier ist jedoch $f(y_i|z_i,\vartheta) = \prod_{j=1}^{n_i} f(y_{ij}|z_i,\vartheta)$. Zur besseren Übersicht führen wir für die durch Gauß-Hermite-Quadratur approximierten Dichtefunktion folgende Notation ein

$$f(y_i|\vartheta) = \int f(y_i|Z_i,\vartheta)\phi(Z_i)dz_i \approx \sum_{l=1}^{k} f(y_i|z_l,\vartheta)\pi_l =: f_k(y_i|\vartheta).$$

Damit lassen sich die Gewichte $w_{il}^{(t)}$ schreiben als

$$w_{il}^{(t)} = \frac{f(y_i|z_l,\vartheta^{(t)})\pi_l}{f_k(y_i|\vartheta^{(t)})}.$$

Im M-Schritt des EM-Algorithmuses muss jeweils Q_k maximiert werden. Dies führt zum System

$$\frac{\partial}{\partial\vartheta}Q_k(\vartheta|\vartheta^{(t)}) = \sum_{i=1}^{r}\sum_{l=1}^{k} w_{il}^{(t)} s_{il} = 0, \tag{37}$$

mit $s_{il} = \frac{\partial}{\partial\vartheta}\log f(y_i|z_l,\vartheta)$. Man beachte, dass hier $w_{il}^{(t)}$ nicht von ϑ abhängt, da wir die Massepunkte z_l und Gewichte π_l mittels Gauß-Hermite-Quadratur erhalten haben und sie damit bekannte Größen sind. So wie in Friedl und Kauermann (2000) setzen wir nun

$$g_\vartheta(\vartheta) = \frac{\partial}{\partial\widetilde{\vartheta}}Q_k(\widetilde{\vartheta}|\vartheta)\Big|_{\widetilde{\vartheta}=\vartheta}.$$

Offensichtlich gilt für alle EM-Schätzungen $g_\vartheta(\vartheta^{(t)}) = 0$. Wir betrachten nun den Erwartungswert von $g_\vartheta(\vartheta)$ für den wahren Parameter ϑ und approximieren dabei die wahre Dichte $f(y_i|\vartheta)$ durch $f_k(y_i|\vartheta)$. Dann gilt $\mathbb{E}[g_\vartheta(\vartheta)] \approx \mathbb{E}_k[g_\vartheta(\vartheta)]$. Dabei deutet der Subskript k an,

dass der Erwartungswert bzgl. der approximativen Dichte f_k gemeint ist. Es gilt

$$\begin{aligned}\mathbb{E}_k\left[g_\vartheta(\vartheta)\right] &= \sum_{i=1}^{r}\sum_{l=1}^{k}\int w_{il}s_{il}f_k(y_i|\vartheta)\,dy_i \\ &= \sum_{i=1}^{r}\sum_{l=1}^{k}\int \frac{\pi_l f(y_i|z_l,\vartheta)}{f_k(y_i|\vartheta)} f_k(y_i|\vartheta)\frac{\partial}{\partial\vartheta}\log f(y_i|z_l,\vartheta)\,dy_i \\ &= \sum_{i=1}^{r}\sum_{l=1}^{k}\pi_l\frac{\partial}{\partial\vartheta}\int f(y_i|z_l,\vartheta)\,dy_i = 0. \end{aligned} \tag{38}$$

Friedl und Kauermann (2000) definieren nun den wahren Parameter implizit über die Gleichung $\mathbb{E}_k\left[g_\vartheta(\vartheta)\right] = 0$. Als nächsten Schritt betrachten wir die Taylorentwicklung ersten Grades von $g_\vartheta(\vartheta^{(t)})$ um ϑ. Wegen $g_\vartheta(\vartheta^{(t)}) = 0$ gilt

$$\begin{aligned} 0 &\approx g_\vartheta(\vartheta) + \frac{\partial}{\partial\vartheta}g_\vartheta(\vartheta)\left(\vartheta^{(t)} - \vartheta\right) \\ \vartheta^{(t)} - \vartheta &\approx -\left(\frac{\partial}{\partial\vartheta}g_\vartheta(\vartheta)\right)^{-1} g_\vartheta(\vartheta). \end{aligned} \tag{39}$$

In der Berechnung von $\frac{\partial}{\partial\vartheta}g_\vartheta(\vartheta)$ kommt die Ableitung der Gewichte w_{il} nach ϑ vor. Diese wollen wir zunächst bestimmen und nutzen dabei aus, dass die Dichte $f(y_i|z_l,\vartheta)$ zur Exponentialfamilie gehört und somit eine Darstellung der Form wie in Definition 1 angegeben hat. Eine detailierte Berechnung findet sich in Abschnitt A.5.1. Als Ergebnis resultiert:

$$\frac{\partial w_{il}}{\partial\vartheta} = w_{il}\left(s_{il} - \sum_{m=1}^{k} w_{im}s_{im}\right).$$

Insgesamt erhalten wir für die Ableitung von $g_\vartheta(\vartheta)$

$$\begin{aligned}\frac{\partial}{\partial\vartheta}g_\vartheta(\vartheta) &= \sum_{i=1}^{r}\sum_{l=1}^{k}\frac{\partial}{\partial\vartheta}w_{il}s_{il} \\ &= \sum_{i=1}^{r}\sum_{l=1}^{k} w_{il}\left(s_{il}s_{il}^\top + \frac{\partial s_{il}}{\partial\vartheta}\right) - \sum_{i=1}^{r}\sum_{l=1}^{k}\sum_{m=1}^{k} w_{il}w_{im}s_{il}s_{im}^\top. \end{aligned} \tag{40}$$

Wir betrachten nun den Erwartungswert der ersten Komponente bzgl. der approximativen Dichte und erkennen, dass

$$\mathbb{E}_k\left[w_{il}s_{il}s_{il}^\top\right] = -\mathbb{E}_k\left[w_{il}\frac{\partial s_{il}}{\partial\vartheta}\right]$$

(Details sind in Abschnitt A.5.1 zu finden).

Daraus erhalten wir die (gewichtete) approximative Fisher-Information $F_k(\vartheta)$ als

$$\begin{aligned} F_k(\vartheta) &= \mathbb{E}_k\left[-\frac{\partial}{\partial\vartheta} g_\vartheta(\vartheta)\right] = \mathbb{E}_k\left[\sum_{i=1}^{r}\sum_{l=1}^{k}\sum_{m=1}^{k} w_{il}w_{im}s_{il}s_{im}^\top\right] \\ &= \mathbb{E}_k\left[g_\vartheta(\vartheta)g_\vartheta^\top(\vartheta)\right]. \end{aligned} \tag{41}$$

Als nächstes ersetzen wir in (39) die Hessematrix durch ihren approximativen Erwartungswert. Als Varianz erhalten wir schließlich

$$\mathbb{V}ar\left[\vartheta^{(t)}\right] \approx \mathbb{V}ar_k\left[\vartheta^{(t)}\right] = F_k^{-1}(\vartheta).$$

In der resultierenden Matrix sind die approximativen Varianzen aller geschätzten Parameter enthalten. Ist man beispielsweise nur an der Varianz zu $\beta^{(t)}$ interessiert, muss dazu die entsprechende Submatrix betrachtet werden.

10.1.1 *Berechnung von* $F_k(\vartheta)$

$F_k(\vartheta)$ hängt vom wahren Parameter ϑ ab, der natürlich nicht bekannt ist. Die naheliegendste Möglichkeit ist sicherlich ϑ schlicht durch die Schätzung $\hat{\vartheta}$ zu ersetzten[1] und anstelle der Information $F_k(\vartheta)$ deren empirische Version $\widehat{F}_k(\hat{\vartheta})$,

$$\widehat{F}_k(\hat{\vartheta}) := \sum_{i=1}^{r}\sum_{l=1}^{k}\sum_{m=1}^{k} \hat{w}_{il}\hat{w}_{im}\hat{s}_{il}\hat{s}_{im}^\top, \tag{42}$$

zu berechnen. Friedl und Kauermann (2000) weisen jedoch darauf hin, dass diese direkte Methode nicht immer angebracht ist. Ist das Modell unpassend, äußert sich dies in hohen Werten von $\hat{s}_{il}$, was schlussendlich dazu führt, dass die Varianzen in diesem Fall unterschätzt werden.
Eine alternative Bestimmung einer Schätzung von $F_k(\vartheta)$ ist, $F_k(\hat{\vartheta})$ durch Monte Carlo Simulation (siehe Abschnitt 10.3) zu berechnen. Dies führt zu $\widetilde{F}_k(\hat{\vartheta})$. Friedl und Kauermann (2000) schlagen schließlich vor, die Varianz durch eine Kombination dieser Alternativen *(Sandwich)* zu schätzen

$$\widehat{\mathbb{V}ar}(\hat{\vartheta}) := \frac{n}{df}\widetilde{F}_k^{-1}(\hat{\vartheta})\widehat{F}_k(\hat{\vartheta})\widetilde{F}_k^{-1}(\hat{\vartheta}). \tag{43}$$

Weiters gilt für den Freiheitsgrad $df = n - (p+1)$ mit der Dimension $p+1$ von ϑ. Ist das Modell recht gut, produzieren die Monte Carlo und die direkte Berechnung ähnliche Ergebnisse. In diesem Fall gilt also $\widetilde{F}_k^{-1}(\hat{\vartheta})\widehat{F}_k(\hat{\vartheta}) \approx \mathbf{I}$ und damit

$$\widetilde{F}_k^{-1}(\hat{\vartheta})\widehat{F}_k(\hat{\vartheta})\widetilde{F}_k^{-1}(\hat{\vartheta}) \approx \widetilde{F}_k^{-1}(\hat{\vartheta}) \approx \widehat{F}_k^{-1}(\hat{\vartheta}).$$

1 Hier und im folgenden bezeichnet $\hat{\vartheta}$ die im letzten EM-Schritt erzeugte Schätzung von ϑ. Auf analoge Weise sind auch andere Variablen mit Dach zu verstehen.

Da $\widehat{F}_k^{-1}(\hat{\vartheta})$ im Falle eines passenden Modells die Varianz gut approximiert, tut dies auch das gesamte Sandwich. Die Approximation wird durch Multiplikation mit $\frac{n}{df}$ um den Freiheitsgrad korrigiert, um vor allem bei einem kleinen Stichprobenumfang noch bessere Ergebnisse zu erhalten. Im Falle eines unpassenden Modells wird durch die Verwendung des Sandwichs die Gefahr der Unterschätzung der Varianz gemindert, da über den Monte Carlo-Teil auch empirische Werte miteinfließen.

10.2 NICHTPARAMETRISCHE SCHÄTZUNG

Nun treffen wir keinerlei Annahmen über die Verteilung der zufälligen Effekte, sondern schätzern diese mit. Die linearen Prädiktoren schreiben wir als

$$\eta_{ijl} = x_{ij}^\top \beta + z_l = x_{ij}^\top \beta + e_l^\top z,$$

wobei z der Vektor der geschätzten Massepunkte und e_l der k-dimensionale Einheitsvektor ist. Diese Schreibweise wird in weiterer Folge hilfreich sein. Wir approximieren $Q(\vartheta|\vartheta^{(t)})$ durch

$$Q_k(\vartheta|\vartheta^{(t)}) = \sum_{i=1}^{r}\sum_{l=1}^{k} w_{il}^{(t)}[\log \pi_l + \log f(y_i|z_l,\vartheta)].$$

Die Gewichte π_l können über die bereits bekannte Formel

$$\pi_l^{(t)} = \frac{1}{r}\sum_{i=1}^{r} w_{il}^{(t)} \tag{44}$$

geschätzt werden. Möchte man auch unterschiedliche Clustergrößen zulassen, erhält man

$$\pi_l^{(t)} = \frac{1}{n}\sum_{i=1}^{r} n_i w_{il}^{(t)}. \tag{45}$$

Mit $\sum_{i=1}^{r} n_i = n$ sieht man leicht, dass (44) ein Spezialfall von (45) mit $n_i = c$ für alle $i = 1,\dots,r$ und einer Konstanten $c \in \mathbb{N}$ ist. Wir erinnern uns zurück, dass bei der Schätzung der π_l die Restriktion $\sum_{l=1}^{k} \pi_l = 1$ beachtet werden muss, da es sich bei den π_l ja um Gewichte handelt. Um jedoch einen beschränkten Parameterraum zu vermeiden, führen wir gleich wie Friedl und Kauermann (2000) eine Reparametrisierung von π_l ein: Setze

$$\pi_l = \begin{cases} \exp(\varphi_l - \kappa(\varphi)), & \text{für } l = 1,\dots,k-1 \\ \exp(-\kappa(\varphi)), & \text{für } l = k. \end{cases}$$

Dabei ist $\kappa(\varphi)$ eine differenzierbare Funktion mit $\frac{\partial \kappa(\varphi)}{\partial \varphi_l} = \pi_l$ für alle $l = 1,\dots,k-1$. Damit sind mit allen $\varphi \in \mathbb{R}^{k-1}$ sämtliche Restriktionen an π erfüllt.

Wir haben nun zwei Parametervektoren: Zum einen den $p + k$-dimensionalen Vektor $\vartheta = (\beta^\top, z^\top)^\top$ und zum anderen den $k-1$-dimensionalen Vektor $\varphi = (\varphi_1, \ldots, \varphi_{k-1})^\top$. Gleich wie bei den normalverteilten zufälligen Effekten erhalten wir nun für ϑ die Schätzfunktion

$$g_\vartheta(\vartheta, \varphi) = \frac{\partial}{\partial \widetilde{\vartheta}} Q(\widetilde{\vartheta}|\vartheta)\Big|_{\widetilde{\vartheta}=\vartheta} = \sum_{i=1}^{r} \sum_{l=1}^{k} w_{il} s_{il}. \tag{46}$$

Aus den Definitionen kann man ablesen, dass w_{il} von beiden Parametervektoren abhängt, also $w_{il} = w_{il}(\vartheta, \varphi)$, s_{il} jedoch nur von ϑ, also $s_{il} = s_{il}(\vartheta)$.
Nun benötigen wir noch eine Schätzfunktion für φ, die ähnlich zu (46) ist. Dafür schreiben wir (45) folgendermaßen um: Für $l = 1, \ldots, k-1$ ist (45) äquivalent zu

$$\sum_{i=1}^{r} n_i \left(w_{il}^{(t)} - \pi_l^{(t)} \right) = 0.$$

Dies lässt sich zusammenfassen zu

$$\sum_{i=1}^{r} n_i \begin{pmatrix} w_{i1}^{(t)} - \pi_1^{(t)} \\ \vdots \\ w_{i(k-1)}^{(t)} - \pi_{k-1}^{(t)} \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Damit erhalten wir eine $(k-1)$-dimensionale Schätzfunktion für φ, nämlich

$$g_\varphi(\vartheta, \varphi) := \sum_{i=1}^{r} \sum_{l=1}^{k-1} n_i \widetilde{e}_l (w_{il} - \pi_l), \tag{47}$$

wobei $\widetilde{e}_l$ der $k-1$-dimensionale kanonische Einheitsvektor ist.
Die EM-Schätzungen definieren wir auf natürliche Art und Weise via $g(\vartheta^{(t)}, \varphi^{(t)}) = 0$ mit $g(\cdot) = \left(g_\vartheta(\cdot)^\top, g_\varphi(\cdot)^\top\right)^\top$.
Ananlog zu (38) erhalten wir, dass der wahre Parameter $\mathbb{E}_k\left[g_\vartheta(\vartheta, \varphi)\right] = 0$ erfüllt. Für φ erhalten wir dasselbe Ergebnis (siehe Abschnitt A.5.2 für Details). Damit ergibt sich insgesamt

$$\mathbb{E}_k\left[g(\vartheta, \varphi)\right] = 0.$$

Die Taylorentwicklung ersten Grades von $g(\vartheta^{(t)}, \varphi^{(t)})$ um (ϑ, φ) ist wegen $g(\vartheta^{(t)}, \varphi^{(t)}) = 0$ gleich

$$\begin{pmatrix} \vartheta^{(t)} - \vartheta \\ \varphi^{(t)} - \varphi \end{pmatrix} \approx - \begin{pmatrix} g_{\vartheta\vartheta} & g_{\vartheta\varphi} \\ g_{\varphi\vartheta} & g_{\varphi\varphi} \end{pmatrix}^{-1} \begin{pmatrix} g_\vartheta(\vartheta, \varphi) \\ g_\varphi(\vartheta, \varphi) \end{pmatrix},$$

wobei die Abkürzungen in der Jacobimatrix für $g_{\vartheta\vartheta} = \frac{\partial g_\vartheta(\vartheta, \varphi)}{\partial \vartheta^\top}$, $g_{\vartheta\varphi} = \frac{\partial g_\vartheta(\vartheta\varphi)}{\partial \varphi^\top}$, $g_{\varphi\vartheta} = \frac{\partial g_\varphi(\vartheta\varphi)}{\partial \vartheta^\top}$ und $g_{\varphi\varphi} = \frac{\partial g_\varphi(\vartheta\varphi)}{\partial \varphi^\top}$ stehen.

Die Ableitung $g_{\vartheta\vartheta}$ ist bereits aus (40) bekannt. Für die anderen Ableitungen benötigen wir zunächst die entsprechenden Ableitungen der Gewichte w_{il}. Deren Herleitung ist in Abschnitt A.5.2 zu finden. Für $l' = 1, \dots, k-1$ gilt

$$\frac{\partial w_{il}}{\partial \varphi_{l'}} = w_{il}\delta_{l=l'} - w_{il}w_{il'}$$

mit $\delta_{l=l'} = 1$ falls $l = l'$ und 0 sonst.
Damit erhalten wir die restlichen Ableitungen (Details in Abschnitt A.5.2)

$$\begin{aligned}
\frac{\partial g_\vartheta(\vartheta,\varphi)}{\partial \varphi^\top} &= \frac{\partial g_\varphi(\vartheta,\varphi)}{\partial \vartheta} \\
\frac{\partial g_\vartheta(\vartheta,\varphi)}{\partial \varphi^\top} &= \sum_{i=1}^{r}\sum_{m=1}^{k-1} w_{im}s_{im}\widetilde{e}_m^\top - \sum_{i=1}^{r}\sum_{l=1}^{k}\sum_{m=1}^{k-1} w_{il}w_{im}s_{il}\widetilde{e}_m^\top
\end{aligned} \tag{48}$$

$$\begin{aligned}
\frac{\partial g_\varphi(\vartheta,\varphi)}{\partial \varphi^\top} &= \sum_{i=1}^{r}\sum_{m=1}^{k-1} n_i\widetilde{e}_m\widetilde{e}_m^\top(w_{im}-\pi_m) \\
&\quad - \sum_{i=1}^{r}\sum_{l=1}^{k-1}\sum_{m=1}^{k-1} n_i\widetilde{e}_l\widetilde{e}_m^\top(w_{il}w_{im}-\pi_l\pi_m).
\end{aligned} \tag{49}$$

Wir schreiben die gesamte Fisher Informationsmatrix als $F_k(\vartheta,\varphi) = -\mathbb{E}_k\left[\frac{\partial g(\vartheta,\varphi)}{\partial(\vartheta^\top,\varphi^\top)}\right]$. Diese setzt sich aus vier Teilmatrizen zusammen,

$$F_k(\vartheta,\varphi) = \left(\begin{array}{c|c} F_{k,\vartheta\vartheta}(\cdot) & F_{k,\vartheta\varphi}(\cdot) \\ \hline F_{k,\varphi\vartheta}(\cdot) & F_{k,\varphi\varphi}(\cdot) \end{array}\right),$$

wobei $F_{k,\vartheta\varphi}(\cdot) = F_{k,\varphi\vartheta}(\cdot)^\top$. $F_{k,\vartheta\vartheta}(\cdot)$ ist bereits aus (41) bekannt. Als nächsten Schritt halten wir fest, dass $\mathbb{E}_k[w_{il}] = \pi_l$ und $\mathbb{E}_k[w_{il}s_{il}] = 0$ (dies wird in Abschnitt A.5.2 nachgerechnet). Daraus ergibt sich

$$\begin{aligned}
F_{k,\vartheta\varphi}(\vartheta,\varphi) &= \mathbb{E}_k\left[-\frac{\partial g_\vartheta(\vartheta,\varphi)}{\partial\varphi^\top}\right] \\
&= \sum_{i=1}^{r}\sum_{l=1}^{k}\sum_{m=1}^{k-1}\mathbb{E}_k\left[w_{il}w_{im}s_{il}\right]\widetilde{e}_m^\top \\
&= \mathbb{E}_k\left[g_\vartheta(\vartheta,\varphi)g_\varphi(\vartheta,\varphi)^\top\right].
\end{aligned}$$

Weiters erhalten wir wegen $\mathbb{E}_k[(w_{il}-\pi_l)(w_{im}-\pi_m)] = \mathbb{E}_k[w_{il}w_{im}] - \pi_l\pi_m$ auch die letzte Submatrix

$$\begin{aligned}
F_{k,\varphi\varphi}(\vartheta,\varphi) &= \mathbb{E}_k\left[-\frac{\partial g_\varphi(\vartheta,\varphi)}{\partial\varphi^\top}\right] \\
&= \sum_{i=1}^{r}\sum_{l=1}^{k-1}\sum_{m=1}^{k-1} n_i\widetilde{e}_l\widetilde{e}_m^\top\mathbb{E}_k\left[(w_{il}-\pi_l)(w_{im}-\pi_m)\right] \\
&= \mathbb{E}_k\left[g_\varphi(\vartheta,\varphi)g_\varphi(\vartheta,\varphi)^\top\right].
\end{aligned}$$

Damit gilt für die approximative Varianz von $(\vartheta^{(t)}, \varphi^{(t)})$

$$\mathbb{V}\mathrm{ar}\left[\vartheta^{(t)}, \varphi^{(t)}\right] \approx \mathbb{V}\mathrm{ar}_k\left[\vartheta^{(t)}, \varphi^{(t)}\right] = F_k(\vartheta, \varphi)^{-1}.$$

10.2.1 *Berechnung von* $F_k(\vartheta, \varphi)$

Wie bereits bei den normalverteilten zufälligen Effekten in Abschnitt 10.1.1 legen wir eine Varianzapproximation durch das Sandwich

$$\widehat{\mathbb{V}\mathrm{ar}}(\hat{\vartheta}, \hat{\varphi}) := \frac{n}{df}\widetilde{F}_k^{-1}(\hat{\vartheta}, \hat{\varphi})\widehat{F}_k(\hat{\vartheta}, \hat{\varphi})\widetilde{F}_k^{-1}(\hat{\vartheta}, \hat{\varphi}).$$

nahe. Dabei bezeichnet $\widehat{F}_k(\hat{\vartheta}, \hat{\varphi})$ analog zu vorher die empirische Version von $F_k(\vartheta, \varphi)$. Ebenso steht $\widetilde{F}_k(\hat{\vartheta}, \hat{\varphi})$ für die Monte Carlo Schätzung von $F_k(\vartheta, \varphi)$, die wie in Abschnitt 10.3 angegeben berechnet wird.

10.3 SCHÄTZUNG VON $F_k(\vartheta)$ BZW. $F_k(\vartheta, \varphi)$ DURCH MC SIMULATION

Wir wollen nun zwei Möglichkeiten vorstellen, wie eine Schätzung von $F_k(\vartheta)$ durch Monte Carlo Simulation erhalten werden kann. Die erste Variante ist die wohl intuitivste: Zunächst werden n zufällige Effekte z_i^* aus der diskreten Verteilung mit Massepunkten z_l und Gewichten π_l simuliert. Im Falle der Gauß-Hermite Quadratur sind die Masspunkte und Gewichte fixe, bekannte Größen. Bei der nichtparametrischen Schätzung werden dafür die zuletzt geschätzten Werte herangezogen. Gegeben diese simulierten zufälligen Effekte ist die Dichte $f(y|z_i^*, x_{ij}, \hat{\vartheta})$ ein bekanntes Mitglied der Exponentialfamilie. Unter Verwendung des final geschätzen Parameters $\hat{\vartheta}$ können schließlich daraus die y_{ij}^* gezogen werden. In jedem Simulationsschritt wird der Erwartungswertet in (41) durch die simulierten empirischen Momente ersetzt. Nach B Monte Carlo Schritten wird über alle Ergebnisse gemittelt. Dies ergibt schlussendlich $\widetilde{F}_k(\hat{\vartheta})$. Je größer B ist, desto genauere Ergebnisse werden erzielt. Gleichzeitig bedeutet ein großes B auch einen großen numerischen Aufwand. Aus diesem Grund schlagen Friedl und Kauermann (2000) eine andere Variante vor, bei der die Simulation der zufälligen Effekte z_i^* obsolet ist. Dabei wird ausgenutzt, dass $f_k(\cdot)$ eine gewichtete Summe der bekannten Dichten $f(\cdot|z_i, \vartheta)$ ist. Akzeptiert man den geschätzten Parametervektor, kann also direkt aus diesen Dichten simuliert werden. Als jeweilige Replikationszahl B_l wird $B_l = [1 + \pi_l B]$ herangezogen. Friedl und Kauermann (2000) gehen nun nach Algorithmus 2 vor:

Algorithmus 2 Schätzgun von $F_k(\vartheta)$ durch Monte Carlo Simulation

for $l = 1 \to k$ **do**
  for $b = 1 \to B_l$ **do**
    Simuliere y_{ij}^* aus $f(y|z_l, x_{ij}, \hat{\vartheta})$.
    Berechne w^* und s^* basierend auf y^*.
    Berechne $F_{l,b}^* = \sum_{i=1}^{r} \sum_{m=1}^{k} \sum_{m'=1}^{k} w_{im}^* w_{im'}^* s_{im}^* s_{im'}^{*\top}$ (Gauß-Quadratur) bzw. $F_{l,b}^* = F_k(\vartheta^*, \varphi^*)$ nach den Formeln in Abschnitt 10.2 (nichtparametrische Schätzung)[2] .
  end for
  Berechne $F_l^* = \frac{1}{B_l} \sum_{b=1}^{B_l} F_{l,b}^*$.
end for
Berechne $\widetilde{F}_k = \sum_{l=1}^{k} \pi_l F_l^*$.

2 Die Sternchennotation bedeutet, dass in den Erwartungswerten der entsprechenden Formeln w und s jeweils durch w^* bzw. s^* ersetzt werden.

11

STANDARDFEHLER IN NEGATIV BINOMIALMODELLEN

11.1 NORMALVERTEILTE ZUFÄLLIGE EFFEKTE

Da die Negativ Binomialverteilung kein Mitglied der Exponentialfamilie ist, können die Resultate aus Abschnitt 10.1 nicht zur Gänze übernommen werden. Die Gewichte w_{il} sind nun von der Form

$$w_{il} = \frac{\pi_l \prod_{j=1}^{n_i} \left(\frac{\mu_{ijl}}{\mu_{ijl}+\alpha}\right)^{y_{ij}} \left(\frac{\alpha}{\mu_{ijl}+\alpha}\right)^{\alpha}}{\sum_{m=1}^{k} \pi_m \prod_{j=1}^{n_i} \left(\frac{\mu_{ijm}}{\mu_{ijm}+\alpha}\right)^{y_{ij}} \left(\frac{\alpha}{\mu_{ijm}+\alpha}\right)^{\alpha}}.$$

Für die Ableitungen $\frac{\partial w_{il}}{\partial \vartheta}$ erhalten wir im Falle der Negativen Binomialverteilung

$$\frac{\partial w_{il}}{\partial \alpha} = w_{il}\left\{c_{il}^{\alpha} - \sum_{m=1}^{k} w_{im} c_{im}^{\alpha}\right\},$$

mit $c_{il}^{\alpha} = \sum_{j=1}^{n_i}\left[\frac{\mu_{ijl}-y_{ij}}{\mu_{ijl}+\alpha} + \log\left(\frac{\alpha}{\mu_{ijl}+\alpha}\right)\right]$. Für die Ableitung nach β gilt

$$\frac{\partial w_{il}}{\partial \beta} = w_{il}\left\{c_{il}^{\beta} - \sum_{m=1}^{k} w_{im} c_{im}^{\beta}\right\}$$

mit $c_{il}^{\beta} = \sum_{j=1}^{n_i} x_{ij}^{\top}\left(y_{ij} - \frac{\mu_{ijl}(y_{ij}+\alpha)}{\mu_{ijl}+\alpha}\right)$.

Eine detailierte Herleitung der obigen Ergebnisse findet sich in Abschnitt A.6.

Weiters benötigen wir die Scorefunktionen s_{il}. Diese wurden für das Modell ohne Zufallseffekte bereits in Kapitel 6 berechnet (vgl. (29) und (30)). Unter Berücksichtigung der Zufallseffekte sowie einer potentiellen Clusterung ergeben sich

$$\begin{aligned} s_{il}(\alpha) = \frac{\partial}{\partial \alpha} \log \prod_{j=1}^{n_i} f(y_{ij}|z_l, \vartheta) &= n_i(\log \alpha + 1 - \psi(\alpha)) \\ &+ \sum_{j=1}^{n_i}\left[\psi(y_{ij}+\alpha) - \log(\mu_{ijl}+\alpha) - \frac{y_{ij}+\alpha}{\mu_{ijl}+\alpha}\right] \end{aligned}$$

und

$$s_{il}(\beta) = \frac{\partial}{\partial \beta} \log \prod_{j=1}^{n_i} f(y_{ij}|z_l, \vartheta) = \sum_{j=1}^{n_i} \frac{(y_{ij}-\mu_{ijl})x_{ij}}{1+\frac{\mu_{ijl}}{\alpha}}.$$

Als nächsten Schritt berechnen wir die Ableitungen der Scorefunktionen und erhalten

$$\begin{aligned}
\frac{\partial s_{il}(\alpha)}{\partial \alpha} &= \sum_{j=1}^{n_i} \left[\psi'(y_{ij} + \alpha) - \psi'(\alpha) + \frac{\mu_{ijl}^2 + \alpha y_{ij}}{\alpha(\mu_{ijl} + \alpha)^2} \right], \\
\frac{\partial s_{il}(\alpha)}{\partial \beta} &= \sum_{j=1}^{n_i} \frac{(y_{ij} - \mu_{ijl})\mu_{ijl} x_{ij}}{(\mu_{ijl} + \alpha)^2} \quad \text{und} \\
\frac{\partial s_{il}(\beta)}{\partial \beta} &= -\sum_{j=1}^{n_i} x_{ij}^\top \frac{\mu_{ijl}(y_{ij} + \alpha)}{\alpha \left(1 + \frac{\mu_{ijl}}{\alpha}\right)^2} x_{ij}.
\end{aligned}$$

Wir setzten nun $g_\alpha(\vartheta) := \sum_{i=1}^{r} \sum_{l=1}^{k} w_{il} s_{il}(\alpha)$ und $g_\beta(\vartheta) := \sum_{i=1}^{r} \sum_{l=1}^{k} w_{il} s_{il}(\beta)$ und berechnen deren Ableitungen:

$$\begin{aligned}
\frac{\partial g_\alpha(\vartheta)}{\partial \alpha} &= \sum_{i=1}^{r} \sum_{l=1}^{k} \left(\frac{\partial w_{il}}{\partial \alpha} s_{il}(\alpha) + w_{il} \frac{\partial s_{il}(\alpha)}{\partial \alpha} \right), \\
\frac{\partial g_\alpha(\vartheta)}{\partial \beta} &= \sum_{i=1}^{r} \sum_{l=1}^{k} \left(\frac{\partial w_{il}}{\partial \beta} s_{il}(\alpha) + w_{il} \frac{\partial s_{il}(\alpha)}{\partial \beta} \right), \\
\frac{\partial g_\beta(\vartheta)}{\partial \alpha} &= \frac{\partial g_\alpha(\vartheta)}{\partial \beta} \text{ und} \\
\frac{\partial g_\beta(\vartheta)}{\partial \beta} &= \sum_{i=1}^{r} \sum_{l=1}^{k} \left(\frac{\partial w_{il}}{\partial \beta} s_{il}(\beta) + w_{il} \frac{\partial s_{il}(\beta)}{\partial \beta} \right).
\end{aligned}$$

Nachdem sich die Ausdrücke durch Einsetzen der zuvor berechneten Terme nicht wesentlich vereinfachen, sehen wir davon ab. Die approximative Fisher Information ist nun

$$F_k(\vartheta) = \left(\begin{array}{c|c} F_{k,\alpha\alpha} & F_{k,\alpha\beta} \\ \hline F_{k,\alpha\beta}^\top & F_{k,\beta\beta} \end{array} \right)$$

mit den Submatrizen

$$\begin{aligned}
F_{k,\alpha\alpha}(\vartheta) &= \mathbb{E}_k \left[-\frac{\partial g_\alpha(\vartheta)}{\partial \alpha} \right] = -\sum_{i=1}^{r} \sum_{l=1}^{k} \mathbb{E}_k \left[\frac{\partial w_{il}}{\partial \alpha} s_{il}(\alpha) + w_{il} \frac{\partial s_{il}(\alpha)}{\partial \alpha} \right], \\
F_{k,\alpha\beta}(\vartheta) &= \mathbb{E}_k \left[-\frac{\partial g_\alpha(\vartheta)}{\partial \beta} \right] = -\sum_{i=1}^{r} \sum_{l=1}^{k} \mathbb{E}_k \left[\frac{\partial w_{il}}{\partial \beta} s_{il}(\alpha) + w_{il} \frac{\partial s_{il}(\alpha)}{\partial \beta} \right], \\
F_{k,\beta\beta}(\vartheta) &= \mathbb{E}_k \left[-\frac{\partial g_\beta(\vartheta)}{\partial \beta} \right] = -\sum_{i=1}^{r} \sum_{l=1}^{k} \mathbb{E}_k \left[\frac{\partial w_{il}}{\partial \beta} s_{il}(\beta) + w_{il} \frac{\partial s_{il}(\beta)}{\partial \beta} \right].
\end{aligned}$$

Als approximative Varianz erhalten wir damit

$$\mathbb{V}ar\left[\vartheta^{(t)}\right] \approx \mathbb{V}ar_k\left[\vartheta^{(t)}\right] = F_k^{-1}(\vartheta).$$

Zur praktischen Schätzung von $F_k(\vartheta)$ sei auf Abschnitt 10.1.1 und Abschnitt 10.3 verwiesen.

11.2 NICHTPARAMETRISCHE SCHÄTZUNG

In diesem Abschnitt wollen wir Standardfehler im Falle der nichtparametrischen Schätzung im Negativ Binomialmodell herleiten. Bei den folgenden Ausführungen bauen wir direkt auf die Ergebnisse in Abschnitt 10.2 auf. Auch die Notationen werden aus diesem Kapitel übernommen. Da die einfache Struktur der einparametrigen, linearen Exponentialfamilie nicht mehr ausgenutzt werden kann, ist die Berechnung jedoch etwas aufwendiger. Die Grundideen bleiben aber dieselben.
Die a posteriori Wahrscheinlichkeiten w_{il} und deren Ableitungen nach α bzw. β sind uns im Falle der Negativen Binomialverteilung bereits aus Abschnitt 11.1 bekannt. Die Ableitung nach z erfolgt symmetrisch zu jener nach β und wir erhalten

$$\frac{\partial w_{il}}{\partial z} = w_{il}\left\{c_{il}^z - \sum_{m=1}^{k} w_{im} c_{im}^z\right\},$$

mit $c_{il}^z = \sum_{j=1}^{n_i} e_l^\top \left(y_{ij} - \frac{\mu_{ijl}(y_{ij}+\alpha)}{\mu_{ijl}+\alpha}\right)$.

Für die Gewichte π_l erhalten wir die Schätzformel

$$\pi_l^{(t)} = \frac{1}{n}\sum_{i=1}^{r} n_i w_{il}^{(t)}$$

- vergleiche dazu (45). Für π_l verwenden wir analog zu Abschnitt 10.2 eine Reparametrisierung, um einen eingeschränkten Parameterraum zu umgehen. Insgesamt arbeiten wir dieses Mal mit vier Parametervektoren: Zunächst das Skalar α, weiters der p-dimensionale Vektor β sowie der k-dimensionale Vektor z und schließlich der $k-1$-dimensionale Vektor $\varphi = (\varphi_1, \ldots, \varphi_{k-1})$. Aus Gründen der Übersichtlichkeit verwenden wir ab und an die Notation $\vartheta = (\alpha, \beta^\top, z^\top)^\top$.

Durch Umformungen ergibt sich für φ analog zu (47)

$$g_\varphi(\vartheta, \varphi) := \sum_{i=1}^{r}\sum_{l=1}^{k-1} n_i \widetilde{e}_l (w_{il} - \pi_l).$$

$g_\alpha(\vartheta, \varphi)$ und $g_\beta(\vartheta, \varphi)$ sind bereits aus Abschnitt 11.1 bekannt. Wir benötigen also nur noch $g_z(\vartheta, \varphi)$: Dafür leiten wir zunächst die Scorefunktion $s_{il}(z)$ her. Als Ergebnis erhalten wir

$$s_{il}(z) = \frac{\partial}{\partial z}\log\prod_{j=1}^{n_i} f(y_{ij}|z_l, \alpha, \beta) = \sum_{j=1}^{n_i} e_l^\top \frac{y_{ij} - \mu_{ijl}}{1 + \frac{\mu_{ijl}}{\alpha}}.$$

Als nächstes berechnen wir jene Ableitungen der Scorefunktionen, die wir noch nicht aus Abschnitt 11.1 kennen. Wir erhalten

$$\begin{aligned}
\frac{\partial s_{il}(\alpha)}{\partial z} &= \sum_{j=1}^{n_i} e_l^\top \frac{\mu_{ijl}(y_{ij} - \mu_{ijl})}{(\mu_{ijl} + \alpha)^2}, \\
\frac{\partial s_{il}(\beta)}{\partial z} &= -\sum_{j=1}^{n_i} e_l^\top \frac{\mu_{ijl}(y_{ij} + \alpha)}{\alpha\left(1 + \frac{\mu_{ijl}}{\alpha}\right)^2} x_{ij}, \text{ und} \\
\frac{\partial s_{il}(z)}{\partial z} &= -\sum_{j=1}^{n_i} e_l^\top \frac{\mu_{ijl}(y_{ij} + \alpha)}{\alpha\left(1 + \frac{\mu_{ijl}}{\alpha}\right)^2} e_l.
\end{aligned}$$

Nun stehen sämtliche Bausteine bereit, um die noch ausstehende Schätzfunktion

$$g_z(\vartheta, \varphi) = \sum_{i=1}^{r} \sum_{l=1}^{k} w_{il} s_{il}(z)$$

anzugeben. Wir fassen nun alle Schätzfunktionen zusammen zu $g(\cdot) := \left(g_\alpha(\cdot), g_\beta(\cdot)^\top, g_z(\cdot)^\top, g_\varphi(\cdot)^\top\right)^\top$. Die EM-Schätzungen definieren wir über $g(\alpha^{(t)}, \beta^{(t)}, z^{(t)}, \varphi^{(t)}) = 0$. Die Taylorentwicklung erster Ordnung ergibt nun

$$\begin{pmatrix} \alpha^{(t)} - \alpha \\ \beta^{(t)} - \beta \\ z^{(t)} - z \\ \varphi^{(t)} - \varphi \end{pmatrix} \approx - \begin{pmatrix} g_{\alpha\alpha} & g_{\alpha\beta} & g_{\alpha z} & g_{\alpha\varphi} \\ g_{\beta\alpha} & g_{\beta\beta} & g_{\beta z} & g_{\beta\varphi} \\ g_{z\alpha} & g_{z\beta} & g_{zz} & g_{z\varphi} \\ g_{\varphi\alpha} & g_{\varphi\beta} & g_{\varphi z} & g_{\varphi\varphi} \end{pmatrix}^{-1} \begin{pmatrix} g_\alpha(\vartheta, \varphi) \\ g_\beta(\vartheta, \varphi) \\ g_z(\vartheta, \varphi) \\ g_\varphi(\vartheta, \varphi) \end{pmatrix}.$$

Die Notation $g_{\cdot\cdot}$ ist wie in Abschnitt 10.2 zu verstehen. Dafür benötigen wir noch

$$\begin{aligned}
\frac{\partial g_\alpha(\vartheta, \varphi)}{\partial z} &= \sum_{i=1}^{r} \sum_{l=1}^{k} \left(\frac{\partial w_{il}}{\partial z} s_{il}(\alpha) + w_{il} \frac{\partial s_{il}(\alpha)}{\partial z} \right), \\
\frac{\partial g_\beta(\vartheta, \varphi)}{\partial z} &= \sum_{i=1}^{r} \sum_{l=1}^{k} \left(\frac{\partial w_{il}}{\partial z} s_{il}(\beta) + w_{il} \frac{\partial s_{il}(\beta)}{\partial z} \right), \text{ und} \\
\frac{\partial g_z(\vartheta, \varphi)}{\partial z} &= \sum_{i=1}^{r} \sum_{l=1}^{k} \left(\frac{\partial w_{il}}{\partial z} s_{il}(z) + w_{il} \frac{\partial s_{il}(z)}{\partial z} \right).
\end{aligned}$$

Die Ableitung $\frac{\partial g_\varphi(\vartheta, \varphi)}{\partial \varphi}$ ist bereits aus Abschnitt 10.2 - siehe (49) - bekannt. Auch die übrigen Ableitungen $\frac{\partial g_\alpha(\vartheta, \varphi)}{\partial \varphi}$, $\frac{\partial g_\beta(\vartheta, \varphi)}{\partial \varphi}$ und $\frac{\partial g_z(\vartheta, \varphi)}{\partial \varphi}$ kennen wir schon, da sie allesamt von der Gestalt (48) sind.

Somit können wir die Fisher Informationsmatrix über $F_k(\vartheta, \varphi) = -\mathbb{E}_k\left[\frac{\partial g(\vartheta,\varphi)}{\partial(\vartheta^\top,\varphi^\top)}\right]$ definieren. Diese Matrix sieht nun wie folgt aus:

$$F_k(\vartheta,\varphi) = \left(\begin{array}{c|c|c|c} F_{k,\alpha\alpha}(\cdot) & F_{k,\alpha\beta}(\cdot) & F_{k,\alpha z}(\cdot) & F_{k,\alpha\varphi}(\cdot) \\ \hline F_{k,\alpha\beta}(\cdot)^\top & F_{k,\beta\beta}(\cdot) & F_{k,\beta z}(\cdot) & F_{k,\beta\varphi}(\cdot) \\ \hline F_{k,\alpha z}(\cdot)^\top & F_{k,\beta z}(\cdot)^\top & F_{k,zz}(\cdot) & F_{k,z\varphi}(\cdot) \\ \hline F_{k,\alpha\varphi}(\cdot)^\top & F_{k,\beta\varphi}(\cdot)^\top & F_{k,z\varphi}(\cdot) & F_{k,\varphi\varphi}(\cdot) \end{array}\right),$$

dabei ist zu beachten, dass $F_{k,\alpha\alpha}(\cdot)$ skalarwertig, alle anderen Einträge jedoch Vektoren bzw. Matrizen sind. Die einzelnen Submatrizen sind analog zu Abschnitt 11.1 und Abschnitt 10.2 zu verstehen und werden nicht gesondert angeführt. Mit Hilfe dieser Matrix können wir nun die approximative Varianz über

$$\mathbb{V}\mathrm{ar}\left[\vartheta^{(t)},\varphi^{(t)}\right] \approx \mathbb{V}\mathrm{ar}_k\left[\vartheta^{(t)},\varphi^{(t)}\right] = F_k(\vartheta,\varphi)^{-1}$$

angeben.

12

DIE LOUIS FORMEL

In diesem Abschnitt wollen wir eine alternative Herangehensweise zur Herleitung der beobachteten Informationsmatrix diskutieren. Dieser Weg basiert auf Louis (1982), welcher die resultierende Formel *Missing Information Principle* nennt. Diesem Abschnitt liegt die nichtveröffentlichte Ausführung *Information Principles in Random Effects Models* von H. Friedl und G. Kauermann zu Grunde.

Wir betrachten zunächst den recht allgemeinen Fall von Daten, die sich aus einem beobachteten und einem nicht beobachtetem Teil zusammensetzten. Um keine Verwirrung zu stiften, verwenden wir nicht die naheligende Notation aus Abschnitt 2.2.1, wo der beobachtete Teil der Daten mit y_o, der nicht beobachtete Teil mit y_m und die vollständigen Daten mit $y = (y_o, y_m)$ bezeichnet werden, sondern bleiben dabei y als den beobachteten Teil und z als den nicht beobachteten Teil aufzufassen. Dies erlaubt in weiterer Folge einen einfachen Übergang zur Anwendung auf Modelle mit zufälligen Effekten, in denen wie üblich y den *beobachteten* Responsevektor und z den *nicht beobachtbaren* Zufallseffekt repräsentieren.

12.1 VOLLSTÄNDIGE INFORMATION

Wie üblich bezeichnet ϑ den Vektor aller unbekannten Parameter. Mit

$$\ell_c(\vartheta|y,z) = \log f(y,z|\vartheta)$$

bezeichnen wir die *vollständige* Loglikelihoodfunktion. Die vollständige Scorefunktion ist damit

$$S_c(y,z|\vartheta) = \frac{\partial}{\partial\vartheta}\ell_c(\vartheta|y,z) = \frac{f'(y,z|\vartheta)}{f(y,z|\vartheta)},$$

wobei mit $f'(y,z|\vartheta)$ stets die Ableitung nach ϑ gemeint ist. Die vollständige Fisher Information $\mathcal{I}_c$ erhält man über den Erwartungswert der negativen zweiten Ableitung der vollständigen Loglikelihoodfunktion. Dazu berechnen wir zunächst die negative zweite Ableitung und bezeichnen sie mit I_c,

$$\begin{aligned} I_c(y,z|\vartheta) &= -\frac{\partial^2}{\partial\vartheta\partial\vartheta^\top}\ell_c(\vartheta|y,z) \\ &= -\frac{f''(y,z|\vartheta)}{f(y,z|\vartheta)} + \frac{f'(y,z|\vartheta)f'(y,z|\vartheta)^\top}{f(y,z|\vartheta)^2} \\ &= -\frac{f''(y,z|\vartheta)}{f(y,z|\vartheta)} + S_c(y,z|\vartheta)S_c(y,z|\vartheta)^\top. \end{aligned} \tag{50}$$

12.2 BEOBACHTETE INFORMATION

Bereits in Abschnitt 2.2.1 haben wir gesehen, dass wir zum Erhalt der Schätzung $\hat{\vartheta}$ die *beobachtete Loglikelihoodfunktion*, wie die marginale Loglikelihoodfunktion in diesem Setting genannt wird,

$$\ell(\vartheta|y) = \log \int f(y, z|\vartheta)\,dz$$

maximieren müssen. Für die folgende Rechnung benötigen wir die Vertauschbarkeit von Integration und Differentiation, die in all unseren bisherigen Anwendungen stets gegeben war. Die beobachtete Scorefunktion ist nun

$$\begin{aligned} S(y|\vartheta) &= \frac{\partial}{\partial\vartheta}\ell(\vartheta|y) = \frac{\frac{\partial}{\partial\vartheta}\int f(y, z|\vartheta)\,dz}{\int f(y, z|\vartheta)\,dz} \\ &= \frac{\int f'(y, z|\vartheta)\,dz}{\int f(y, z|\vartheta)\,dz} = \int \frac{f'(y, z|\vartheta)}{f(y, z|\vartheta)} \frac{f(y, z|\vartheta)}{\int f(y, z|\vartheta)\,dz}\,dz \\ &= \int \frac{f'(y, z|\vartheta)}{f(y, z|\vartheta)} f(z|y, \vartheta)\,dz \\ &= \mathbb{E}\left[S_c(y, z|\vartheta)\middle|y\right]. \end{aligned} \tag{51}$$

Die negative zweite Ableitung ergibt sich als

$$\begin{aligned} I(y|\vartheta) &= -\frac{\partial^2}{\partial\vartheta\partial\vartheta^\top}\ell(\vartheta|y) \\ &= -\frac{\int f''(y, z|\vartheta)\,dz}{\int f(y, z|\vartheta)\,dz} + \frac{\int f'(y, z|\vartheta)\,dz \int f'(y, z|\vartheta)^\top\,dz}{\left(\int f(y, z|\vartheta)\,dz\right)^2} \\ &= -\int \frac{f''(y, z|\vartheta)}{f(y, z|\vartheta)} \frac{f(y, z|\vartheta)}{\int f(y, z|\vartheta)\,dz}\,dz + S(y|\vartheta)S(y|\vartheta)^\top \\ &= \mathbb{E}\left[-\frac{f''(y, z|\vartheta)}{f(y, z|\vartheta)}\middle|y\right] + S(y|\vartheta)S(y|\vartheta)^\top. \end{aligned}$$

12.3 MISSING INFORMATION PRINCIPLE

Aus (50) ergibt sich

$$-\frac{f''(y, z|\vartheta)}{f(y, z|\vartheta)} = I_c(y, z|\vartheta) - S_c(y, z|\vartheta)S_c(y, z|\vartheta)^\top$$

und damit gemeinsam mit (51)

$$\begin{aligned} I(y|\vartheta) &= \mathbb{E}\left[I_c(y, z|\vartheta)\middle|y\right] - \mathbb{E}\left[S_c(y, z|\vartheta)S_c(y, z|\vartheta)^\top\middle|y\right] \\ &\quad + \mathbb{E}\left[S_c(y, z|\vartheta)\middle|y\right]\mathbb{E}\left[S_c(y, z|\vartheta)^\top\middle|y\right] \\ &= \mathbb{E}\left[I_c(y, z|\vartheta)\middle|y\right] - \mathbb{V}\mathrm{ar}\left[S_c(y, z|\vartheta)\middle|y\right]. \end{aligned} \tag{52}$$

Diese Formel wird *Louis Formel* oder *Missing Information Principle* genannt.

Wir stellen zunächst fest, dass

$$\begin{aligned}\mathbb{E}\left[\mathbb{E}\left[-\frac{f''(y,z|\vartheta)}{f(y,z|\vartheta)}\middle|y\right]\right] &= -\int\left(\int\frac{f''(y,z|\vartheta)}{f(y,z|\vartheta)}\frac{f(y,z|\vartheta)}{f(y|\vartheta)}dz\right)f(y|\vartheta)dy\\ &= -\frac{\partial^2}{\partial\vartheta\partial\vartheta^\top}\int\int f(y,z|\vartheta)dzdy\\ &= 0\end{aligned}$$

und analog

$$\begin{aligned}\mathbb{E}^c\left[\mathbb{E}\left[-\frac{f''(y,z|\vartheta)}{f(y,z|\vartheta)}\middle|y\right]\right] &= -\int\int\left(\int\frac{f''(y,z|\vartheta)}{f(y,z|\vartheta)}\frac{f(y,z|\vartheta)}{f(y|\vartheta)}dz\right)f(y,z|\vartheta)dzdy\\ &= -\int\int\underbrace{\frac{f(y,z|\vartheta)}{f(y|\vartheta)}}_{=f(z|y,\vartheta)}\left(\frac{\partial^2}{\partial\vartheta\partial\vartheta^\top}\underbrace{\int f(y,z|\vartheta)dz}_{=f(y|\vartheta)}\right)dzdy\\ &= -\int f''(y|\vartheta)\left(\int f(z|y,\vartheta)dz\right)dy\\ &= -\frac{\partial^2}{\partial\vartheta\partial\vartheta^\top}\int f(y|\vartheta)dy.\\ &= 0,\end{aligned}$$

Dabei zeigt der Superskript c an, dass der Erwartungswert bezüglich der gemeinsamen Dichte $f(y,z|\vartheta)$ gemeint ist. Ein Erwartungswert ohne Superskript ist - konform zur restlichen Notation - bezüglich der marginalen Dichte $f(y|\vartheta)$ zu verstehen.
Damit ergeben sich nun aus (52) die beobachtete - $\mathcal{I}(\vartheta)$ - und die vollständige - $\mathcal{I}_c(\vartheta)$ - Fisher Informationsmatrix

$$\begin{aligned}\mathcal{I}(\vartheta) &= \mathbb{E}\left[I(y|\vartheta)\right]\\ &= \mathbb{E}\left[\mathbb{E}\left[I_c(y,z|\vartheta)\middle|y\right]\right] - \mathbb{E}\left[\mathbb{E}\left[S_c(y,z|\vartheta)S_c(y,z|\vartheta)^\top\middle|y\right]\right]\\ &\quad + \mathbb{E}\left[\mathbb{E}\left[S_c(y,z|\vartheta)\middle|y\right]\mathbb{E}\left[S_c(y,z|\vartheta)^\top\middle|y\right]\right]\\ &= \mathbb{E}\left[\mathbb{E}\left[-\frac{f''(y,z|\vartheta)}{f(y,z|\vartheta)}\middle|y\right]\right] + \mathbb{E}\left[S(y|\vartheta)S(y|\vartheta)^\top\right]\\ &= \mathbb{E}\left[S(y|\vartheta)S(y|\vartheta)^\top\right]\end{aligned}$$

und

$$\begin{aligned}\mathcal{I}_c(\vartheta) &= \mathbb{E}^c\left[\mathbb{E}\left[-\frac{f''(y,z|\vartheta)}{f(y,z|\vartheta)}\middle|y\right]\right] + \mathbb{E}^c\left[S_c(y,z|\vartheta)S_c(y,z|\vartheta)^\top\right]\\ &= \mathbb{E}^c\left[S_c(y,z|\vartheta)S_c(y,z|\vartheta)^\top\right].\end{aligned}$$

Häufig wird im *Missing Information Principle* in (52) der linke Teil als *vollständige Information* und der rechte Teil als *fehlende Information* bezeichnet. Diese Namensgebung ist durch die folgende Betrachtung

von McLachlan und Krishnan (1997) nachvollziehbar.
Aus $f(y, z|\vartheta) = f(y|\vartheta)f(z|y, \vartheta)$ folgt

$$\ell(\vartheta|y) = \ell_c(\vartheta|y, z) - \log f(z|y, \vartheta)$$

und damit

$$\begin{aligned} -\frac{\partial^2}{\partial\vartheta\partial\vartheta^\top}\ell(\vartheta|y) &= -\frac{\partial^2}{\partial\vartheta\partial\vartheta^\top}\ell_c(\vartheta|y, z) + \frac{\partial^2}{\partial\vartheta\partial\vartheta^\top}\log f(z|y, \vartheta) \\ I(y|\vartheta) &= I_c(y, z|\vartheta) + \frac{\partial^2}{\partial\vartheta\partial\vartheta^\top}\log f(z|y, \vartheta). \end{aligned}$$

Über die letzte Gleichung ziehen wir den Erwartungswert bezüglich der konditionalen Dichte $f(z|y, \vartheta)$ und erhalten

$$I(y|\vartheta) = \mathbb{E}\left[I_c(y, z|\vartheta)\middle|y\right] - \mathbb{E}\left[-\frac{\partial^2}{\partial\vartheta\partial\vartheta^\top}\log f(z|y, \vartheta)\middle|y\right]. \tag{53}$$

McLachlan und Krishnan (1997) definieren schließlich auf naheliegende Weise

$$\begin{aligned} \mathcal{I}_c(y|\vartheta) &:= \mathbb{E}\left[I_c(y, z|\vartheta)\middle|y\right], \\ \mathcal{I}_m(y|\vartheta) &:= \mathbb{E}\left[-\frac{\partial^2}{\partial\vartheta\partial\vartheta^\top}\log f(z|y, \vartheta)\middle|y\right]. \end{aligned}$$

$\mathcal{I}_c(y|\vartheta)$ wird nun als die *vollständige Information* und $\mathcal{I}_m(y|\vartheta)$ als die *fehlende Information* bezeichnet, da $\mathcal{I}_c(y|\vartheta)$ auf der gemeinsamen Dichte $f(y, z|\vartheta)$ und $\mathcal{I}_m(y|\vartheta)$ auf der konditionalen Dichte $f(z|y, \vartheta)$ der nicht beobachteten Daten gegeben die beobachteten Daten basiert.
$I(y|\vartheta)$ kann nun geschrieben werden als

$$I(y|\vartheta) = \mathcal{I}_c(y|\vartheta) - \mathcal{I}_m(y|\vartheta).$$

Vergleicht man diese Identiät mit (52) ergibt sich

$$\mathcal{I}_m(y|\vartheta) = \mathbb{V}\mathrm{ar}\left[S_c(y, z|\vartheta)\middle|y\right].$$

Schließlich betrachten wir noch

$$\begin{aligned} \mathbb{E}\left[\mathcal{I}_c(y|\vartheta)\right] &= \int\left(\int I_c(y, z|\vartheta)f(z|y, \vartheta)\,dz\right)f(y|\vartheta)\,dy \\ &= \int\int I_c(y, z|\vartheta)f(y, z|\vartheta)\,dz\,dy \\ &= \mathcal{I}_c(\vartheta). \end{aligned}$$

Mit (53) ergibt sich so

$$\mathcal{I}(\vartheta) = \mathcal{I}_c(\vartheta) - \mathbb{E}\left[\mathcal{I}_m(y|\vartheta)\right].$$

12.4 DIE LOUIS FORMEL IM EM-SETTING

Der EM-Algorithmus maximiert die Zielfunktion

$$Q(\vartheta|\vartheta^{(t)}) = \mathbb{E}\left[\ell_c(\vartheta|y,z)\middle|y,\vartheta^{(t)}\right]$$

aus (7). Zur Berechnung von Standardfehlern haben wir stets auf die Funktion

$$g_\vartheta(\vartheta) = \frac{\partial}{\partial\widetilde{\vartheta}} Q(\widetilde{\vartheta}|\vartheta)\bigg|_{\widetilde{\vartheta}=\vartheta}$$

zurückgegriffen. Betrachtet man diese Definition im Setting der Louis Formel, erkennt man, dass auf Grund der vorausgesetzten Vertauschbarkeit von Differentiation und Integration

$$g_\vartheta(\vartheta) = \mathbb{E}\left[S_c(y,z|\vartheta)|y\right] = S(y|\vartheta)$$

gilt. Weiters erhalten wir

$$-\frac{\partial g_\vartheta(\vartheta)}{\partial\vartheta} = -\frac{\partial S(y|\vartheta)}{\partial\vartheta} = I(y|\vartheta).$$

Hiermit schließt sich der Kreis, da wir in der direkten Berechnung der Standardfehler in Abschnitt 10.1 und Abschnitt 10.2 die Varianz-Kovarinzmatrix über die Inverse von $\mathbb{E}\left[-\frac{\partial g_\vartheta(\vartheta)}{\partial\vartheta}\right]$ und nach der Louis Formel durch die Inverse von $\mathcal{I}(\vartheta) = \mathbb{E}\left[I(y|\vartheta)\right]$ berechnen.

Abschließend wollen wir noch einige praktische Anmerkungen zur Berechnung der Varianz-Kovarianzmatrix mit Hilfe der Louis Formel machen. Die Erwartungswerte in der Louis Formel können analog zu Abschnitt 10.1 und Abschnitt 10.2 über Erwartungswerte bezüglich der approximativen Dichte $f_k(y|\vartheta)$ angenähert werden. Tut man dies erhält man

$$\mathcal{I}_k(\vartheta) = F_k(\vartheta).$$

Um eine Schätzung von $\mathcal{I}_k(\vartheta)$ zu erhalten, wird dies nun wieder durch die empirische Version $\widehat{\mathcal{I}}_k(\hat{\vartheta})$ mit $\hat{\vartheta} = \vartheta^{(\infty)}$ ersetzt. Für deren Berechnung ist zu beachten, dass

$$-\frac{\partial^2}{\partial\vartheta\partial\vartheta^\top} Q(\vartheta|\vartheta^{(t)})\bigg|_{\vartheta=\vartheta^{(t)}} = \mathbb{E}\left[I_c(y,z|\vartheta^{(t)})|y,\vartheta^{(t)}\right] = \mathcal{I}_c(y|\vartheta^{(t)}).$$

Für die fehlende Information gilt für alle EM-Schätzungen $\vartheta^{(t)}$ zunächst

$$\begin{aligned} 0 = g_\vartheta(\vartheta^{(t)}) &= \frac{\partial}{\partial\vartheta}\mathbb{E}\left[\ell_c(\vartheta^{(t)}|y,z)\middle|y,\vartheta^{(t)}\right] \\ &= \mathbb{E}\left[S_c(y,z|\vartheta^{(t)})\middle|y,\vartheta^{(t)}\right] = S(y|\vartheta^{(t)}). \end{aligned}$$

Damit erhalten wir

$$\begin{aligned}\mathcal{I}_m(y,\vartheta^{(t)}) &= \mathbb{V}ar\left[S_c(y,z|\vartheta)\middle|y\right]\Big|_{\vartheta=\vartheta^{(t)}} \\ &= \mathbb{E}\left[S_c(y,z|\vartheta)S_c(y,z|\vartheta)^\top\middle|y\right]\Big|_{\vartheta=\vartheta^{(t)}}. \end{aligned} \tag{54}$$

Insgesamt gilt

$$\begin{aligned}\widehat{\mathcal{I}}(\hat{\vartheta}) &= \widehat{\mathcal{I}}_c(y|\hat{\vartheta}) + \mathbb{E}_k\left[\widehat{\mathcal{I}}_m(y|\hat{\vartheta})\right] \\ &= \mathbb{E}\left[I_c(y,z|\hat{\vartheta})|y,\hat{\vartheta}\right] + \mathbb{E}_k\left[\mathbb{E}\left[S_c(y,z|\vartheta)S_c(y,z|\vartheta)^\top\middle|y\right]\Big|_{\vartheta=\hat{\vartheta}}\right].\end{aligned}$$

$\widehat{\mathcal{I}}_c(y|\hat{\vartheta})$ wird von Programmen, die eine EM-Schätzung durchführen im Allgemeinen mitgeliefert. Bei den Funktionen `alldist()` und `allvc()` aus dem `npmlreg`-Paket von Einbeck et al. (2012) erhält man diese Matrix - als QR-Zerlegung im `qr`-Format - über `object$lastglm$qr`. Bei der Funktion `gamlssNP()` aus dem Paket `gamlss.mx` von Rigby und Stasinopoulos (2005) ist die Matrix im selben Format abgespeichert und kann über `object$mu.qr` abgerufen werden.
Die eben zitierten Funktionen ziehen zur Schätzung der Standardfehler nicht die gesamte Matrix $\widehat{\mathcal{I}}(\hat{\vartheta})$, welche auch die fehlende Information beachtet, heran, sondern ausschließlich die Matrix $\widehat{\mathcal{I}}_c(y|\hat{\vartheta})$. Efron und Hinkley (1978) weisen jedoch darauf hin, dass dies keine gute Herangehensweise ist.

Um die Qualität der geschätzten Standardfehler zu verbessern, sollte auch hier wieder eine Kombination aus Monte Carlo Simulation und direkter Schätzung - also das sogenannte *Sandwich* - verwendet werden.

13 ANWENDUNGEN

Wir wollen nun die Louis Formel heranziehen, um eine Schätzung der Standardfehler zu erhalten. Dazu betrachten wir in Abschnitt 13.1.1 bzw. Abschnitt 13.2.1 das Poisson Modell und in Abschnitt 13.1.2 bzw. Abschnitt 13.2.2 das Negativ Binomialmodell.

13.1 NORMALVERTEILTE ZUFÄLLIGE EFFEKTE

Zunächst gehen wir von normalverteilten zufälligen Effekten aus und geben konkrete Formeln zur Berechnung der Standardfehler an. Ergänzt werden diese mit Hinweise zur Implementierung.

13.1.1 *Das Poisson Modell*

Wir betrachten ein Modell, in dem $y_i|Z_i$ Poissonverteilt ist mit Parameter $\mu_i = \exp(x_i^\top \beta + \sigma Z_i)$. Weiters seien die zufälligen Effekte Z_i standardnormalverteilt. Für den Parametervektor gilt somit $\vartheta = (\beta^\top, \sigma)^\top$. Es ist sinnvoll, eine erweiterte Systemmatrix $\widetilde{X}$ einzuführen, die neben der ursprünglichen Designmatrix X auch den Vektor der zufälligen Effekte Z beinhaltet. Damit kann

$$\mu_i = \exp(x_i^\top \beta + \sigma Z_i) = \exp(\tilde{x}_i^\top \vartheta)$$

formuliert werden und es gilt

$$\frac{\partial \mu_i}{\partial \vartheta} = \tilde{x}_i \mu_i.$$

Wenn wir solch ein Modell mit Hilfe der R-Funktionen `alldist()` bzw. `allvc()` oder `gamlssNP()` schätzen, erhalten wir auf sehr einfache Art und Weise die *vollständige Information* (vgl. Abschnitt 12.4). Dennoch wollen wir die Formeln zur Berechnung der vollständigen Information herleiten, da wir diese beispielsweise zur Berechnung des Sandwichschätzer brauchen werden. Dazu benötigen wir zunächst die vollständige Scorefunktion

$$S_c(y, Z|\vartheta) = \sum_{i=1}^{r} \tilde{x}_i^\top (y_i - \mu_i). \tag{55}$$

Die Herleitung dieses Ergebnisses ist in Abschnitt A.7 zu finden. Aus dieser Herleitung ist auch ersichtlich, dass

$$S_c(y, Z|\vartheta) = \frac{f'(y|Z, \vartheta)}{f(y|Z, \vartheta)} = \frac{\partial}{\partial \vartheta} \log f(y|Z, \vartheta).$$

Damit gilt gemeinsam mit (55)

$$\begin{aligned}-\frac{\partial^2}{\partial\vartheta\partial\vartheta^\top}\log f(y|Z,\vartheta) &= -\frac{\partial}{\partial\vartheta}(S_c(y,Z|\vartheta))^\top \\ &= -\frac{\partial}{\partial\vartheta}\sum_{i=1}^{r}\tilde{x}_i^\top(y_i-\mu_i) \\ &= \sum_{i=1}^{r}\mu_i\tilde{x}_i\tilde{x}_i^\top\end{aligned}$$

und damit

$$\begin{aligned}\mathcal{I}_c(y|\vartheta) &= \mathbb{E}\left[-\frac{\partial^2}{\partial\vartheta\partial\vartheta^\top}\log f(y|Z,\vartheta)\middle| y\right] \\ &\approx \sum_{i=1}^{r}\sum_{l=1}^{k}\mu_{il}\tilde{x}_{il}\tilde{x}_{il}^\top w_{il},\end{aligned}$$

dabei ist $\tilde{x}_{il} = (x_i^\top, z_l)^\top$. Wir werden in der Monte Carlo Simulationsstudie diese Herangehensweise testen. Dabei fällt auf, dass es zu kleinen Abweichungen zwischen der direkt über den Modelloutput generierten und der über die eben hergeleitete Formel berechneten vollständigen Information kommt. Diese Ungenauigkeiten resultieren aus numerischen Fehlern und sind klein genug, um für unsere Zwecke vernachlässigt zu werden.

Wir sind nun noch daran interessiert, wie wir möglichst effizient zur *fehlenden Information* gelangen. Für die fehlende Information gilt nach (54)

$$\mathcal{I}_m(y,\vartheta) = \mathbb{E}\left[S_c(y,Z|\vartheta)S_c(y,Z|\vartheta)^\top\middle|y\right] - \mathbb{E}\left[S_c(y,Z|\vartheta)\middle|y\right]\mathbb{E}\left[S_c(y,Z|\vartheta)^\top\middle|y\right].$$

Die bedingten Erwartungswerte approximieren wir nun wieder mit Hilfe der Gauß-Hermite Quadratur, indem wir die bedingte Dichte $f(Z_i|y_i,\vartheta)$ durch w_{il} annähern. Damit erhalten wir

$$\begin{aligned}&\mathbb{E}\left[S_c(y,Z|\vartheta)S_c(y,Z|\vartheta)^\top\middle|y\right] \\ &= \mathbb{E}\left[\sum_{i=1}^{r}\tilde{x}_i\tilde{x}_i^\top(y_i-\mu_i)^2 + \sum_{i=1}^{r}\sum_{\substack{d=1\\d\neq i}}^{r}\tilde{x}_i\tilde{x}_d^\top(y_i-\mu_i)(y_d-\mu_d)\middle| y\right] \\ &\approx \sum_{i=1}^{r}\sum_{l=1}^{k}\tilde{x}_{il}\tilde{x}_{il}^\top(y_i-\mu_{il})^2 w_{il} \\ &\quad + \sum_{i=1}^{r}\sum_{\substack{d=1\\d\neq i}}^{r}\sum_{l=1}^{k}\sum_{m=1}^{k}\tilde{x}_{il}\tilde{x}_{dm}^\top(y_i-\mu_{il})(y_d-\mu_{dm})w_{il}w_{dm}\end{aligned}$$

und

$$\begin{aligned}&\mathbb{E}\left[S_c(y, Z|\vartheta)\middle|y\right]\mathbb{E}\left[S_c(y, Z|\vartheta)^\top\middle|y\right]\\ \approx&\left(\sum_{i=1}^{r}\sum_{l=1}^{k}\tilde{x}_{il}(y_i-\mu_{il})w_{il}\right)\left(\sum_{d=1}^{r}\sum_{m=1}^{k}\tilde{x}_{dm}^\top(y_d-\mu_{dm})w_{dm}\right).\end{aligned}$$

Wir wissen zwar bereits, dass der zweite Teil $\mathbb{E}\left[S_c(y, Z|\vartheta)\middle|y\right]\mathbb{E}\left[S_c(y, Z|\vartheta)^\top\middle|y\right]$ für alle EM-Schätzungen $\vartheta^{(t)}$ gleich 0 ist. Dennoch lohnt es sich, diesen Teil zu berücksichtigen, da man dadurch eine schönere Darstellung der fehlenden Information erhält. Es gilt nämlich insgesamt

$$\begin{aligned}\mathcal{I}_m(y,\vartheta)\approx&\sum_{i=1}^{r}\sum_{l=1}^{k}\tilde{x}_{il}\tilde{x}_{il}^\top(y_i-\mu_{il})^2w_{il}\\ &-\sum_{i=1}^{r}\sum_{l=1}^{k}\sum_{m=1}^{k}\tilde{x}_{il}\tilde{x}_{im}^\top(y_i-\mu_{il})(y_i-\mu_{im})w_{il}w_{im}.\end{aligned}\tag{56}$$

13.1.1.1 *Implementierung*

In Abschnitt 12.4 haben wir bereits darauf hingewiesen, dass die *vollständige Information* ein Nebenprodukt der Modellschätzung mit `alldist()` bzw. `allvc()` und `gamlssNP()` ist. Wir wollen also noch angeben, wie man auf effiziente Weise zur *fehlenden Information* kommt. Dazu gehen wir davon aus, dass ein Modell mittels der Funktion `gamlssNP()` geschätzt und unter `object` abgespeichert wurde. (Für die anderen beiden Funktionen ist die Herangehensweise analog.) Laut (56) müssen wir eine Doppel- und eine Dreifachsumme berechnen. Jeder Summand besteht aus einer Matrix $\tilde{x}_{il}\tilde{x}_{dm}^\top$[1], die mit einem Skalar $(y_i-\mu_{il})^2w_{il}$ bzw. $(y_i-\mu_{il})(y_i-\mu_{im})w_{il}w_{im}$ gewichtet wird. Die Matrix $\tilde{X}$, die als Zeilen jeweils die Vektoren $\tilde{x}^\top$ enthält, ist über `x.tilde=object$mu.x` abrufbar. Die Matrizen $\tilde{x}_{il}\tilde{x}_{dm}^\top$ erhält man schließlich über das Kroneckerprodukt $\tilde{X}\otimes\tilde{X}$, das in `R` via `kronecker(x.tilde,x.tilde)` berechnet werden kann. Das Ergebnis ist eine $k^2\cdot n^2$-dimensionale Matrix. In den Zeilen dieser Kroneckermatrix steht nun jeweils eine als $(p+1)$-dimensionaler Vektor umformatierte Matrix $\tilde{x}_{il}\tilde{x}_{dm}^\top$:

$$\begin{pmatrix}kr_{11} & \dots & kr_{1q}\\ \vdots & & \vdots\\ kr_{q1} & \dots & kr_{qq}\end{pmatrix}\rightsquigarrow\begin{pmatrix}kr_{11},\dots,kr_{q1},\dots,kr_{1q},\dots,kr_{qq}\end{pmatrix},$$

1 In (56) kommen zwar nur Terme vor, die nicht mehr vom Index d abhängen, dennoch werden wir sehen, dass es leichter ist, sämtliche Terme zu berechnen und danach erst jene zu filtern, die tatsälich für die Berechnung benötigt werden.

dabei ist q die tatsächliche Anzahl der zu schätzenden Parameter, also $q = p$ bei Modellen ohne Intercept ($p - 1$ Regressionsparameter und σ) und $q = p + 1$ bei Modellen mit Intercept (p Regressionsparameter und σ). Die genaue Anordnung der Zeilen lässt sich leicht an Hand der Definition von Kroneckerprodukten nachvollziehen und ist in Tabelle 4 in der Spalte *Matrix* $\tilde{x}_{il}\tilde{x}_{dm}^{\top}$ festgehalten. Durch das Kroneckerprodukt erhalten wir sämtliche Kombinationen von $\tilde{x}_{il}\tilde{x}_{dm}^{\top}$. Zu den jeweiligen Kombinationen fügen wir vier Spalten hinzu, welche die jeweiligen Werte der Indizes i, l, d und m angeben. Dies legen wir als `data.frame` ab und fügen außerdem noch die zur jeweiligen Ausprägung der Indizes passenden Werte von y, μ und w hinzu. Das Schema des gesamten Datenframes, welchen wir einfach mit `mat` bezeichnen, ist in Tabelle 4 dargestellt. Mit Hilfe der letzten fünf Spalten lassen sich die Gewichte $(y_i - \mu_{il})^2 w_{il}$ bzw. $(y_i - \mu_{il})(y_i - \mu_{im})w_{il}w_{im}$ leicht berechnen.

Zeile	i	l	d	m	Matrix $\tilde{x}_{il}\tilde{x}_{dm}^{\top}$	y_i	μ_{il}	μ_{im}	w_{il}	w_{im}
1	1	1	1	1	— $\tilde{x}_{11}\tilde{x}_{11}^{\top}$ —	y_1	μ_{11}	μ_{11}	w_{11}	w_{11}
2	2	1	1	1	— $\tilde{x}_{21}\tilde{x}_{11}^{\top}$ —	y_2	μ_{21}	μ_{21}	w_{21}	w_{21}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	n	1	1	1	— $\tilde{x}_{n1}\tilde{x}_{11}^{\top}$ —	y_n	μ_{n1}	μ_{n1}	w_{n1}	w_{n1}
$n+1$	1	2	1	1	— $\tilde{x}_{12}\tilde{x}_{11}^{\top}$ —	y_1	μ_{12}	μ_{11}	w_{12}	w_{11}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$2n$	n	2	1	1	— $\tilde{x}_{n2}\tilde{x}_{11}^{\top}$ —	y_n	μ_{n2}	μ_{n1}	w_{n2}	w_{n1}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
nk	n	k	1	1	— $\tilde{x}_{nk}\tilde{x}_{11}^{\top}$ —	y_n	μ_{nk}	μ_{n1}	w_{nk}	w_{n1}
$nk+1$	1	1	2	1	— $\tilde{x}_{11}\tilde{x}_{21}^{\top}$ —	y_1	μ_{11}	μ_{11}	w_{11}	w_{11}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$nk+n$	n	1	2	1	— $\tilde{x}_{n1}\tilde{x}_{21}^{\top}$ —	y_n	μ_{n1}	μ_{n1}	w_{n1}	w_{n1}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$2nk$	n	k	2	1	— $\tilde{x}_{nk}\tilde{x}_{21}^{\top}$ —	y_n	μ_{nk}	μ_{n1}	w_{nk}	w_{n1}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n^2k	n	k	n	1	— $\tilde{x}_{nk}\tilde{x}_{n1}^{\top}$ —	y_n	μ_{nk}	μ_{n1}	w_{nk}	w_{n1}
n^2k+1	1	1	1	2	— $\tilde{x}_{11}\tilde{x}_{12}^{\top}$ —	y_1	μ_{11}	μ_{12}	w_{11}	w_{12}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n^2k^2	n	k	n	k	— $\tilde{x}_{nk}\tilde{x}_{nk}^{\top}$ —	y_n	μ_{nk}	μ_{nk}	w_{nk}	w_{nk}

Tabelle 4: Berechnungsschema der *Fehlenden Information*

Zur Berechnung der Doppel- bzw. Dreifachsumme benötigen wir jedoch nicht sämtliche Kombinationen von $\tilde{x}_{il}\tilde{x}_{dm}^{\top}$. Deshalb

bilden wir zwei Subdatenframes: Für die Doppelsumme benötigen wir alle $k \cdot n$ Kombinationen, für die gilt $i = d$ und $l = m$. Diese können einfach durch `mat.1 = subset(mat, (i==d & l==m))` gefiltert werden. Die Einträge werden schließlich mit dem entsprechenden Gewichtsvektor zeilenweise (das heißt alle Einträge einer Zeile erhalten dasselbe Gewicht) multipliziert. Daraufhin werden über `apply(mat.1,2,sum)` spaltenweise Summen gebildet und der resultierende q-dimensionale Vektor wieder zu einer $q \times q$-Matrix umgewandelt.
Bei der Dreifachsumme geht man analog vor. Man beachte jedoch, dass hier $k^2 \cdot n$ Kombinationen benötigt werden, welche über `mat.2 = subset(mat,i==d)` gefiltert werden können. Schlussendlich werden die beiden resultierenden Matrizen subtrahiert und man erhält die Matrix der *fehlenden Information*.

Um die in Abschnitt 10.3 vorgestellte Sandwichschätzung zu erhalten, adaptieren wir Algorithmus 2 und gehen nun nach Algorithmus 3 vor.

Algorithmus 3 $I(y|\vartheta)$ durch Monte Carlo Simulation

Setze `mu` $=$ `exp(object$mu.lp)`.
Setze `mass = object$prob`.
Setze `I.MC = I.temp = matrix(0,q,q)`.
for $l = 1 \to k$ **do**
  for $b = 1 \to B_l$ **do**
    Simuliere `y.star` $=$ `rpois(n, mu[, l])`.
    Berechne `w.star` - die a posteriori Wahrscheinlichkeiten - basierend auf `mu` und `y.star`.
    Berechne die beobachtete Information I basierend auf $y.star$ und $w.star$.
    Setze `I.temp` $=$ `I.temp` $+$ `I`.
  end for
  Berechne `I.MC` $=$ `I.MC` $+$ `mass`$[l]/B_l *$ `I.temp`.
  Setze `I.temp = matrix(0,q,q)`.
end for

13.1.2 *Das Negativ Binomialmodell*

Nun betrachten wir das Modell, in dem $y_i|z_i$ Negativ Binomialverteilt ist mit Parametern α und $\mu_i = \exp(x_i^\top \beta + \sigma Z_i)$. Die Zufallseffekte Z_i seien wieder unabhängige standardnormalverteilte Zufallsvariablen. Praktischerweise schreiben wir $\tilde{x}_i = (x_i^\top, Z_i)^\top$ und in weiterer Folge auch $\tilde{x}_{il} = (x_i^\top, z_l)^\top$. Damit kann $\mu_i = \exp(\tilde{x}_i^\top \vartheta)$ mit $\vartheta = (\beta^\top, \sigma)^\top$ und $\theta = (\alpha, \vartheta^\top)^\top$ geschrieben werden. Wir erhalten so-

fort die vollständigen Scorefunktionen direkt aus (29) und (30), wobei dort x_i jeweils durch $\tilde{x}_i$ ersetzt werden muss:

$$S_c^{\alpha}(y, Z|\theta) = r(\log\alpha + 1 - \psi(\alpha)) + \sum_{i=1}^{r}\left[\psi(y_i+\alpha) - \log(\mu_i+\alpha) - \frac{y_i+\alpha}{\mu_i+\alpha}\right] \tag{57}$$

$$S_c^{\vartheta}(y, Z|\theta) = \sum_{i=1}^{r}\left[\frac{(y_i-\mu_i)\tilde{x}_i}{1+\frac{\mu_i}{\alpha}}\right]. \tag{58}$$

Es sind nun zwei Herangehensweisen denkbar: Zum einen könnten die Standardfehler von ϑ und α getrennt voneinamder geschätzt werden - so arbeitet beispielsweise die Funktion `gamlssNP()`. Dies entspreche der Annahme unabhänigiger Maximum Likelihood Schätzer. Die Unabhängigkeit ist jedoch nicht gegeben, weshalb zu beachten ist, dass bei einem derartigen Vorgehen gewisse Wechselwirkungen außer Acht gelassen werden. Zum anderen könnten die Scorefunktionen zusammengefügt

$$S_c(y, Z|\theta) = \left(S_c^{\alpha}(y, Z|\theta), S_c^{\vartheta}(y, Z|\theta)^{\top}\right)^{\top}$$

und gemeinsam weiter verwertet werden. Diese Variante ist natürlich die *korrekte*, aber auch die aufwändigere. Wir werden nun beide Herangehensweisen weiterverfolgen und in der Simulationsstudie schließlich testen, ob ein wesentlicher Unterschied zwischen den beiden Ansätzen besteht.

13.1.2.1 *Variante A: Getrennte Schätzung*

Zunächst halten wir fest, dass die Scorefunktion $S_c^{\vartheta}(y, z|\vartheta)$ nur eine mit $\frac{1}{1+\frac{\mu_i}{\alpha}} = \frac{\alpha}{\alpha+\mu_i}$ gewichtete Variante der Scorefunktion im Poissonmodell ist. Damit kann die *fehlende Information* direkt aus den bereits bekannten Formeln für das Poisson Modell hergeleitet werden. Es gilt also

$$\begin{aligned}\mathcal{J}_m^{\vartheta}(y, \theta) \approx & \sum_{i=1}^{r}\sum_{l=1}^{k}\left(\frac{\alpha}{\alpha+\mu_{il}}\right)^2 \tilde{x}_{il}\tilde{x}_{il}^{\top}(y_i-\mu_{il})^2 w_{il} \\ & - \sum_{i=1}^{r}\sum_{l=1}^{k}\sum_{m=1}^{k}\frac{\alpha}{\alpha+\mu_{il}}\frac{\alpha}{\alpha+\mu_{im}}\tilde{x}_{il}\tilde{x}_{im}^{\top} \\ & \cdot (y_i-\mu_{il})(y_i-\mu_{im})w_{il}w_{im}.\end{aligned} \tag{59}$$

Bei der Scorefunktion $S_c^{\alpha}(y, Z|\theta)$ handelt es sich um ein Skalar. Mit den Abkürzungen $c := r(\log\alpha + 1 - \psi(\alpha))$ und $\Sigma := \sum_{i=1}^{r}\left[\psi(y_i+\alpha) - \log(\mu_i+\alpha) - \frac{y_i+\alpha}{\mu_i+\alpha}\right]$ erhalten wir

$$\begin{aligned}\mathbb{E}\left[(S_c^{\alpha}(y, Z|\theta))^2|y\right] &= \mathbb{E}\left[(c+\Sigma)^2|y\right] \\ &= c^2 + 2c\mathbb{E}\left[\Sigma|y\right] + \mathbb{E}\left[\Sigma^2|y\right]\end{aligned}$$

und

$$\begin{aligned}\mathbb{E}\left[S_c^{\alpha}(y, Z|\theta)|y\right]^2 &= (c + \mathbb{E}\left[\Sigma|y\right])^2 \\ &= c^2 + 2c\mathbb{E}\left[\Sigma|y\right] + \mathbb{E}\left[\Sigma|y\right]^2.\end{aligned}$$

Die fehlende Information ist also

$$\begin{aligned}\mathcal{I}_m^{\alpha}(y, \theta) &= \mathbb{E}\left[(S_c^{\alpha}(y, Z|\theta))^2\middle|y\right] - \mathbb{E}\left[S_c^{\alpha}(y, Z|\theta)|y\right]^2 \\ &= \mathbb{E}\left[\Sigma^2\middle|y\right] - \mathbb{E}\left[\Sigma|y\right]^2 \\ &= \mathbb{E}\left[\sum_{i=1}^{r} s_i^2 + \sum_{i=1}^{r}\sum_{\substack{d=1\\d\neq i}}^{r} s_i s_d \middle| y\right] - \mathbb{E}\left[\sum_{i=1}^{r} s_i \middle| y\right]^2 \\ &\approx \sum_{i=1}^{r}\sum_{l=1}^{k} s_{il}^2 w_{il} + \sum_{i=1}^{r}\sum_{\substack{d=1\\d\neq i}}^{r}\sum_{l=1}^{k}\sum_{m=1}^{k} s_{il} s_{dm} w_{il} w_{dm} \\ &\quad - \left(\sum_{i=1}^{r}\sum_{l=1}^{k} s_{il} w_{il}\right)\left(\sum_{d=1}^{r}\sum_{m=1}^{k} s_{dm} w_{dm}\right) \\ &= \sum_{i=1}^{r}\sum_{l=1}^{k} s_{il}^2 w_{il} - \sum_{i=1}^{r}\sum_{l=1}^{k}\sum_{m=1}^{k} s_{il} s_{im} w_{il} w_{im},\end{aligned}$$

mit $s_i := \psi(y_i + \alpha) - \log(\mu_i + \alpha) - \frac{y_i+\alpha}{\mu_i+\alpha}$ bzw. $s_{il} := \psi(y_i + \alpha) - \log(\mu_{il} + \alpha) - \frac{y_i+\alpha}{\mu_{il}+\alpha}$.

Zur Implementierung erstellen wir wieder eine $k^2 \cdot n^2$-dimensionale Matrix mit den Spalten $i, d, l, m, s_{il}, s_{im}, w_{il}$, und w_{im}. Die Reihenfolge der Zeilen ist wieder dieselbe wie die in Tabelle 4. Damit lassen sich die Doppel- und Dreifachsumme ganz analog wie in Abschnitt 13.1.1.1 berechnen.
Außerdem ist darauf zu achten, dass die in `gamlssNP()` verwendete Negative Binomialverteilung `NBI()` anders parametrisiert ist als hier. `NBI()` kennt den Parameter μ - genau so wie er hier verstanden wird - und den Parameter $\sigma = \frac{1}{\alpha}$. Es ist also $\hat{\alpha} =$ `1/object$sigma.coef`. Deshalb muss auch die vollständige Information für α selbst berechnet werden. Ähnlich wie bei der Poissonverteilung erhalten wir mit (57)

$$\begin{aligned}\frac{\partial}{\partial\alpha} S_c^{\alpha}(y, Z|\theta) &= \frac{\partial}{\partial\alpha}\left(r(\log\alpha + 1 - \psi(\alpha))\right. \\ &\qquad \left. + \sum_{i=1}^{r}\left[\psi(y_i + \alpha) - \log(\mu_i + \alpha) - \frac{y_i + \alpha}{\mu_i + \alpha}\right]\right) \\ &= r\left(\frac{1}{\alpha} - \psi'(\alpha)\right) + \sum_{i=1}^{r}\left[\psi'(y_i + \alpha) - \frac{2\mu_i + \alpha - y_i}{(\mu_i + \alpha)^2}\right]\end{aligned}$$

und damit

$$\begin{aligned}
\mathcal{I}_c^{\alpha}(y|\theta) &= \mathbb{E}\left[-\frac{\partial^2}{\partial\alpha^2}\log f(y|Z,\theta)\big|y\right] \\
&= \mathbb{E}\left[-\frac{\partial}{\partial\alpha}S_c^{\alpha}(y,Z|\theta)\big|y\right] \\
&\approx -r\left(\frac{1}{\alpha}-\psi'(\alpha)\right) \\
&\quad -\sum_{i=1}^{r}\sum_{l=1}^{k}\left[\psi'(y_i+\alpha)-\frac{2\mu_{il}+\alpha-y_i}{(\mu_{il}+\alpha)^2}\right]w_{il}.
\end{aligned}$$

13.1.2.2 *Variante B: Gemeinsame Schätzung*

Wir betrachten stets alle Parameter gemeinsam und widmen uns zunächst der *fehlenden Information*. Dazu benötigen wir zuallererst die vollständige Scorefunktion, welche durch

$$S_c(y,Z|\theta) = \left(S_c^{\alpha}(y,Z|\theta), S_c^{\vartheta}(y,Z|\theta)^{\top}\right)^{\top}$$

gegeben ist. Aus Gründen der Übersichtlichkeit verwenden wir in diesem Abschnitt nun die Abkürzungen $S_c^{\alpha} = S_c^{\alpha}(y,Z|\theta)$ und $S_c^{\vartheta} = S_c^{\vartheta}(y,Z|\theta)$. Damit gilt

$$S_c(y,Z|\theta)S_c(y,Z|\theta)^{\top} = \begin{pmatrix} (S_c^{\alpha})^2 & S_c^{\alpha}(S_c^{\vartheta})^{\top} \\ S_c^{\alpha}S_c^{\vartheta} & S_c^{\vartheta}(S_c^{\vartheta})^{\top} \end{pmatrix}.$$

Die Terme $(S_c^{\alpha})^2$ und $S_c^{\vartheta}(S_c^{\vartheta})^{\top}$ sind uns bereits aus Abschnitt 13.1.2.1 bekannt. Da S_c^{α} ein Skalar ist, gilt weiters $S_c^{\alpha}(S_c^{\vartheta})^{\top} = \left(S_c^{\alpha}S_c^{\vartheta}\right)^{\top}$. Um zur fehlenden Information zu gelangen, bedarf es also nur noch einer Betrachtung von $S_c^{\alpha}S_c^{\vartheta}$:

$$\begin{aligned}
S_c^{\alpha}S_c^{\vartheta} &= r(\log\alpha + 1 - \psi(\alpha))\sum_{i=1}^{r}\frac{(y_i-\mu_i)\tilde{x}_i}{1+\frac{\mu_i}{\alpha}} \\
&\quad + \sum_{i=1}^{r}\left[\psi(y_i+\alpha)-\log(\mu_i+\alpha)-\frac{y_i+\alpha}{\mu_i+\alpha}\right]\frac{(y_i-\mu_i)\tilde{x}_i}{1+\frac{\mu_i}{\alpha}} \\
&\quad + \sum_{i=1}^{r}\sum_{\substack{d=1\\d\neq i}}^{r}\left[\psi(y_i+\alpha)-\log(\mu_i+\alpha)-\frac{y_i+\alpha}{\mu_i+\alpha}\right]\frac{(y_d-\mu_d)\tilde{x}_d}{1+\frac{\mu_d}{\alpha}}.
\end{aligned}$$

Wir approximieren nun analog zu allen bisherigen Fällen fehlende Information für die gemischen Terme über

$$\begin{aligned}
\mathcal{I}_m^{(\alpha,\vartheta)}(y,\theta) &= \mathbb{E}\left[S_c^{\alpha} S_c^{\vartheta}|y\right] - \mathbb{E}\left[S_c^{\alpha}|y\right]\mathbb{E}\left[S_c^{\vartheta}|y\right] \\
&\approx r(\log\alpha + 1 - \psi(\alpha)) \sum_{i=1}^{r}\sum_{l=1}^{k} \frac{(y_i-\mu_{il})\tilde{x}_{il}}{1+\frac{\mu_{il}}{\alpha}} w_{il} \\
&\quad + \sum_{i=1}^{r}\sum_{l=1}^{k}\left[\psi(y_i+\alpha) - \log(\mu_{il}+\alpha) - \frac{y_i+\alpha}{\mu_{il}+\alpha}\right] \\
&\quad \cdot \frac{(y_i-\mu_{il})\tilde{x}_{il}}{1+\frac{\mu_{il}}{\alpha}} w_{il} - \sum_{i=1}^{r}\sum_{l=1}^{k}\sum_{m=1}^{k}\Bigg[\psi(y_i+\alpha) \\
&\quad - \log(\mu_{il}+\alpha) - \frac{y_i+\alpha}{\mu_{il}+\alpha}\Bigg] \frac{(y_i-\mu_{im})\tilde{x}_{im}}{1+\frac{\mu_{im}}{\alpha}} w_{il} w_{im}. \\
&= \sum_{i=1}^{r}\sum_{l=1}^{k} (c^1 + c_{il}^2) \frac{(y_i-\mu_{il})\tilde{x}_{il}}{1+\frac{\mu_{il}}{\alpha}} w_{il} \\
&\quad - \sum_{i=1}^{r}\sum_{l=1}^{k}\sum_{m=1}^{k} c_{il}^2 \frac{(y_i-\mu_{im})\tilde{x}_{im}}{1+\frac{\mu_{im}}{\alpha}} w_{il} w_{im}, \text{ mit}
\end{aligned}$$

$$\begin{aligned}
c^1 &= r(\log\alpha + 1 - \psi(\alpha)), \\
c_{il}^2 &= \psi(y_i+\alpha) - \log(\mu_{il}+\alpha) - \frac{y_i+\alpha}{\mu_{il}+\alpha}.
\end{aligned}$$

Die fehlende Information ist schließlich

$$\mathcal{I}_m(y,\theta) = \begin{pmatrix} \mathcal{I}_m^{\alpha}(y,\theta) & \mathcal{I}_m^{(\alpha,\vartheta)}(y,\theta)^{\top} \\ \mathcal{I}_m^{(\alpha,\vartheta)}(y,\theta) & \mathcal{I}_m^{\vartheta}(y,\theta) \end{pmatrix}.$$

Nun betrachten wir noch die *vollständige Information*, da wir diese leider aus dem Output von `gamlssNP()` im konkreten Fall nicht direkt mitgeliefert bekommen. Nach (50) gilt für die vollständige Information

$$\begin{aligned}
\mathcal{I}_c(y|\theta) &= \mathbb{E}\left[I_c(y,Z|\theta)\middle|y\right] \\
&= \mathbb{E}\left[-\frac{f''(y,Z|\theta)}{f(y,Z|\theta)}\middle|y\right] + \mathbb{E}\left[S_c(y,Z|\theta)S_c(y,Z|\theta)^{\top}\middle|y\right].
\end{aligned}$$

Da für EM-Schätzungen $\theta^{(t)}$ stets $\mathbb{E}\left[S_c(y,Z|\theta^{(t)})\middle|y,\theta^{(t)}\right] = 0$ gilt, ist

$$\begin{aligned}
\mathcal{I}_c(y|\theta^{(t)}) &= \left(\mathbb{E}\left[-\frac{f''(y,Z|\theta)}{f(y,Z|\theta)}\middle|y\right] + \mathbb{V}ar\left[S_c(y,Z|\theta)\middle|y\right]\right)\Bigg|_{\theta=\theta^{(t)}} \\
&= \mathbb{E}\left[-\frac{f''(y,Z|\theta)}{f(y,Z|\theta)}\middle|y\right]\Bigg|_{\theta=\theta^{(t)}} + \mathcal{I}_m(y|\theta^{(t)}).
\end{aligned}$$

Es muss also nur noch der Erwartungswert $\mathbb{E}\left[-\frac{f''(y,Z|\theta)}{f(y,Z|\theta)}\middle|y\right]$ berechnet werden. Für die Hessematrix gilt

$$f''(y,Z|\theta) = \begin{pmatrix} \frac{\partial^2}{\partial\alpha^2} f(y,Z|\theta) & \frac{\partial^2}{\partial\alpha\partial\vartheta^\top} f(y,Z|\theta) \\ \left(\frac{\partial^2}{\partial\alpha\partial\vartheta^\top} f(y,Z|\theta)\right)^\top & \frac{\partial^2}{\partial\vartheta\partial\vartheta^\top} f(y,Z|\theta) \end{pmatrix}.$$

Wir kennen bereits aus Abschnitt 13.1.2.1 die nach α und ϑ getrennt berechneten vollständigen und fehlenden Informationen. Es gilt

$$\begin{aligned} \mathbb{E}\left[-\frac{\frac{\partial^2 f(y,Z|\theta)}{\partial\alpha^2}}{f(y,Z|\theta)}\middle|y\right]\Bigg|_{\theta=\theta^{(t)}} &\approx \mathcal{I}_c^\alpha(y,\theta^{(t)}) - \mathcal{I}_m^\alpha(y,\theta^{(t)}), \\ \mathbb{E}\left[-\frac{\frac{\partial^2 f(y,Z|\theta)}{\partial\vartheta\partial\vartheta^\top}}{f(y,Z|\theta)}\middle|y\right]\Bigg|_{\theta=\theta^{(t)}} &\approx \mathcal{I}_c^\vartheta(y,\theta^{(t)}) - \mathcal{I}_m^\vartheta(y,\theta^{(t)}). \end{aligned}$$

Somit müssen wir nur noch $\mathbb{E}\left[-\frac{\frac{\partial^2 f(y,Z|\theta)}{\partial\alpha\partial\vartheta^\top}}{f(y,Z|\theta)}\middle|y\right]$ berechnen. Dazu halten wir fest, dass

$$\frac{f''(y,Z|\theta)}{f(y,Z|\theta)} = \frac{f''(y|Z,\theta)\varphi(Z)}{f(y|Z,\theta)\varphi(Z)} = \frac{f''(y|Z,\theta)}{f(y|Z,\theta)}.$$

Für alle $i = 1,\dots,r$ erhalten wir (vergleiche Abschnitt A.7)

$$\begin{aligned} &\frac{\partial^2}{\partial\alpha\partial\vartheta^\top} f(y_i|Z_i,\theta) = f(y_i|Z_i,\theta)R_i \quad \text{mit} \\ &R_i = \frac{(y_i-\mu_i)\tilde{x}_i}{\mu_i+\alpha}\left[\frac{\alpha\psi(y_i+\alpha)}{\Gamma(y_i+\alpha)^2} - \alpha\psi(\alpha) + \frac{\mu_i+\alpha\mu_i-y_i\alpha}{\mu_i+\alpha}\right]. \end{aligned}$$

Wegen

$$\begin{aligned} \frac{\partial^2}{\partial\alpha\partial\vartheta^\top}\prod_{i=1}^r f(y_i|Z_i,\theta) &= R_1 f(y_1|Z_1,\vartheta) f(y_2|Z_2,\vartheta)\dots f(y_r|Z_r,\vartheta) \\ &\quad + f(y_1|Z_1,\vartheta) R_2 f(y_2|Z_2,\vartheta)\dots f(y_r|Z_r,\vartheta) \\ &\quad + \dots \\ &\quad + f(y_1|Z_1,\vartheta) f(y_2|Z_2,\vartheta)\dots R_r f(y_r|Z_r,\vartheta) \\ &= f(y|Z,\theta)\sum_{i=1}^r R_i \end{aligned}$$

ergibt sich damit

$$\begin{aligned} \frac{\frac{\partial^2}{\partial\alpha\partial\vartheta^\top} f(y|Z,\theta)}{f(y|Z,\theta)} &= \frac{\frac{\partial^2}{\partial\alpha\partial\vartheta^\top}\prod_{i=1}^r f(y_i|Z_i,\theta)}{\prod_{i=1}^r f(y_i|Z_i,\theta)} \\ &= \frac{f(y|Z,\theta)\sum_{i=1}^r R_i}{f(y|Z,\theta)} \\ &= \sum_{i=1}^r R_i. \end{aligned}$$

Somit gilt

$$\mathbb{E}\left[-\frac{\frac{\partial^2 f(y,Z|\theta)}{\partial\alpha\partial\vartheta^\top}}{f(y,Z|\theta)}\middle| y\right] = \mathbb{E}\left[-\sum_{i=1}^{r} R_i \middle| y\right] \approx -\sum_{i=1}^{r}\sum_{l=1}^{k} R_{il} w_{il},$$

mit

$$R_{il} = \frac{(y_i-\mu_{il})\tilde{x}_{il}}{\mu_{il}\alpha}\left[\frac{\alpha\psi(y_i+\alpha)}{\Gamma(y_i+\alpha)^2} - \alpha\psi(\alpha) + \frac{\mu_{il}+\alpha\mu_{il}-y_i\alpha}{\mu_{il}+\alpha}\right].$$

Da

$$\begin{aligned} I(y|\theta^{(t)}) &= \mathcal{I}_c(y|\theta^{(t)}) - \mathcal{I}_m(y|\theta^{(t)}) \\ &= \mathbb{E}\left[-\frac{f''(y,Z|\theta)}{f(y,Z|\theta)}\middle| y\right]\Bigg|_{\theta=\theta^{(t)}} + \mathcal{I}_m(y|\theta^{(t)}) - \mathcal{I}_m(y|\theta^{(t)}) \\ &= \mathbb{E}\left[-\frac{f''(y,Z|\theta)}{f(y,Z|\theta)}\middle| y\right]\Bigg|_{\theta=\theta^{(t)}} \end{aligned}$$

benötigten wir zur Berechnung der Standardfehler die fehlende Information nicht explizit. Um jedoch Vergleiche zwischen den einzelnen Methoden anstellen zu können, ist sie dennoch hilfreich.

13.2 NICHTPARAMETRISCHE SCHÄTZUNG

Nun lassen wir die Annahme der normalverteilten zufälligen Effekte fallen und gehen von einer nichtparametrischen Schätzung aus.

Ganz allgemein betrachten wir nun Modelle mit $\mu_i = \exp(x_i^\top\beta + Z_i)$, wobei die Verteilung der zufälligen Effekte Z_i unbekannt ist. Diese unbekannte Verteilung wird durch eine diskrete Verteilung mit k Massepunkten $z = (z_1,\dots,z_k)^\top$ und zugehörigen Gewichten $\pi = (\pi_1,\dots,\pi_k)^\top$ geschätzt. Das bedeutet, dass für alle $i = 1,\dots,r$ und $l = 1,\dots,k$ der Schätzer durch

$$\mathbb{P}(Z_i = z_l|\pi) = \pi_l \tag{60}$$

charakterisiert ist. Für die Gewichte gilt weiterhin $\sum_{l=1}^{k}\pi_l = 1$ und $\pi_l > 0$ für alle $l = 1,\dots,k$. Der konditionale Erwartungswert μ_i kann nun geschrieben werden als

$$\mu_i = \exp(x_i^\top\beta + \epsilon_i^\top z) = \exp(\tilde{x}_i^\top\vartheta)$$

mit $\tilde{x}_i = (x_i^\top, \epsilon_i^\top)^\top$, $\vartheta = (\beta^\top, z^\top)^\top$ und $\epsilon_i = (\epsilon_{i1},\dots,\epsilon_{ik})^\top \overset{iid}{\sim} MN(1,\pi)$. ϵ_i ist also multinomialverteilt mit Umfang 1 und Wahrscheinlichkeitsverteilung π. Damit ist (60) gleichbedeutend mit

$$\mathbb{P}(Z_i = z_l|\pi) = \mathbb{P}(\epsilon_i = e_l|\pi) = \prod_{l=1}^{k}\pi_l^{e_{il}},$$

wobei e_l eine spezielle Realisation von ϵ_i ist mit $e_l = (e_{i1}, \ldots, e_{ik})^\top = (0, \ldots, 1, \ldots, 0)^\top$. Dabei ist die Eins an der l-ten Stelle. Wir sehen hier nun explizit, dass die diskrete Schätzung der Diche der zufälligen Effekte $f(Z)$ nur von π abhängt. Hingegen ist die konditionale Dichte $f(y|Z)$ wieder bekannt - in unseren Anwendungen also die Wahrscheinlichkeitsfunktion einer Poisson- bzw. Negativ Binomialverteilung - mit Parameter ϑ. Die beiden Wahrscheinlichkeitsfunktionen hängen also von unterschiedlichen Parametersets ab, weswegen sie getrennt voneinander behandelt werden können.

13.2.1 *Das Poisson Modell*

Wir nehmen nun an, dass $f(y_i|Z_i)$ die Wahrscheinlichkeitsfunktion einer Poissonverteilung ist mit Parameter μ_i. Wir haben bereits festegestellt, dass wir die Parametersets ϑ und π getrennt voneinander betrachten können und wenden uns zunächst ϑ zu. Die *vollständige Information* kann wieder direkt aus dem durch `gamlssNP()` geschätzten Modell entnommen werden. Für die *fehlende Information* benötigen wir die vollständige Scorefuntkion, die nach (55) durch

$$S_c^\vartheta(y, Z|\vartheta) = \frac{\frac{\partial}{\partial\vartheta} f(y|Z,\vartheta) f(Z|\pi)}{f(y|Z,\vartheta) f(Z|\pi)} = \frac{\partial}{\partial\vartheta} \log f(y|Z,\vartheta) = \sum_{i=1}^{r} \tilde{x}_i^\top (y_i - \mu_i)$$

gegeben ist. Somit kann die fehlende Information analog (56) berechnet werden und wir erhalten

$$\begin{aligned} \mathcal{I}_m^\vartheta(y,\vartheta) \approx & \sum_{i=1}^{r}\sum_{l=1}^{k} \tilde{x}_{il}\tilde{x}_{il}^\top (y_i - \mu_{il})^2 w_{il} \\ & - \sum_{i=1}^{r}\sum_{l=1}^{k}\sum_{m=1}^{k} \tilde{x}_{il}\tilde{x}_{im}^\top (y_i - \mu_{il})(y_i - \mu_{im}) w_{il} w_{im}. \end{aligned}$$

Nun betrachten wir den Prametervektor π. Auf Grund der Bedingung $\sum_{l=1}^{k} \pi_l = 1$ schreiben wir nun ohne Beschränkung der Allgemeinheit $\pi_k = 1 - \pi_1 - \cdots - \pi_{k-1} = 1 - \sum_{l=1}^{k-1} \pi_l$. Damit erhalten wir

$$\begin{aligned} \log f(Z|\pi) &= \log \prod_{i=1}^{r} f(Z_i|\pi) = \sum_{i=1}^{r} \log \prod_{l=1}^{k} \pi_l^{e_{il}} \\ &= \sum_{i=1}^{r} \left[\sum_{l=1}^{k-1} e_{il} \log \pi_l + e_{ik} \log \left(1 - \sum_{l=1}^{k-1} \pi_l \right) \right]. \end{aligned}$$

Die vollständige Scorefunktion ist damit

$$S_c^\pi(y, Z|\pi) = \frac{\frac{\partial}{\partial\pi} f(y|Z,\vartheta) f(Z|\pi)}{f(y|Z,\vartheta) f(Z|\pi)} = \frac{\partial}{\partial\pi} \log f(Z|\pi).$$

Für $l = 1, \ldots, k$ gilt

$$\frac{\partial}{\partial\pi_l} \log f(Z|\pi) = \sum_{i=1}^{r} \left(\frac{e_{il}}{\pi_l} - \frac{e_{ik}}{\pi_k} \right).$$

Insgesamt resultiert

$$S_c^{\pi}(y, Z|\pi) = \sum_{i=1}^{r} \sum_{l=1}^{k-1} e_l \left(\frac{e_{il}}{\pi_l} - \frac{e_{ik}}{\pi_k} \right),$$

wobei e_l der $k-1$-dimensionale kanonische Einheitsvektor (1 an der l-ten Stelle und Nullen sonst) ist. Die zweiten Ableitungen sind

$$\begin{aligned} \frac{\partial^2}{\partial \pi_l \partial \pi_m} \ell_c(\pi|y, Z) &= \frac{\partial}{\partial \pi_m} \sum_{i=1}^{r} \left(\frac{e_{il}}{\pi_l} - \frac{e_{ik}}{\pi_k} \right) \\ &= \begin{cases} -\sum_{i=1}^{r} \left(\frac{e_{il}}{\pi_l^2} + \frac{e_{ik}}{\pi_k^2} \right), & \text{falls } m = l, \\ -\sum_{i=1}^{r} \frac{e_{ik}}{\pi_k^2}, & \text{falls } m \neq l. \end{cases} \end{aligned}$$

Somit erhalten wir

$$\begin{aligned} I_c^{\pi}(y, Z|\pi) &= \left(-\frac{\partial^2}{\partial \pi_l \partial \pi_m} \ell_c(\pi|y, Z) \right)_{l,m} \\ &= \sum_{i=1}^{r} \sum_{l=1}^{k-1} \sum_{m=1}^{k-1} e_l \frac{e_{ik}}{\pi_k^2} e_m^{\top} + \sum_{i=1}^{r} \sum_{l=1}^{k} e_l \frac{e_{il}}{\pi_l^2} e_l^{\top}. \end{aligned}$$

Nach dem Satz von Bayes und der Definition von w_{il} gilt für den bedingten Erwartungswert

$$\begin{aligned} \mathbb{E}[e_{il}|y_i] &= \sum_{j=0}^{1} j \mathbb{P}(e_{il} = j|y_i) = \mathbb{P}(e_{il} = 1|y_i) = \frac{\pi_l f(y_i|e_{il} = 1)}{f(y_i)} \\ &= \frac{\pi_l f(y_i|z_l, \vartheta)}{f(y_i|\vartheta)} \approx w_{il}. \end{aligned}$$

Damit gilt für die *vollständige Information*

$$\begin{aligned} \mathcal{I}_c^{\pi}(y, Z|\pi) &= \mathbb{E}\left[I_c^{\pi}(y, Z|\pi) \middle| y \right] \\ &= \sum_{i=1}^{r} \sum_{l=1}^{k-1} \sum_{m=1}^{k-1} e_l \mathbb{E}\left[\frac{e_{ik}}{\pi_k^2} \middle| y_i \right] e_m^{\top} + \sum_{i=1}^{r} \sum_{l=1}^{k-1} e_l \mathbb{E}\left[\frac{e_{il}}{\pi_l^2} \middle| y_i \right] e_l^{\top} \\ &\approx \sum_{i=1}^{r} \sum_{l=1}^{k-1} \sum_{m=1}^{k-1} e_l \frac{w_{ik}}{\pi_k^2} e_m^{\top} + \sum_{i=1}^{r} \sum_{l=1}^{k-1} e_l \frac{w_{il}}{\pi_l^2} e_l^{\top}. \end{aligned}$$

Nun betrachten wir die *fehlende Information*. Diese ist (vgl. Abschnitt A.7)

$$\begin{aligned} \mathcal{I}_m^{\pi}(y|\pi) = &\sum_{i=1}^{r} \sum_{l=1}^{k-1} \sum_{m=1}^{k-1} e_l \mathbb{E}\left[\left(\frac{e_{il}}{\pi_l} - \frac{e_{ik}}{\pi_k} \right) \left(\frac{e_{im}}{\pi_m} - \frac{e_{ik}}{\pi_k} \right) \middle| y_i \right] e_m^{\top} \\ &- \sum_{i=1}^{r} \sum_{l=1}^{k-1} \sum_{m=1}^{k-1} e_l \mathbb{E}\left[\left(\frac{e_{il}}{\pi_l} - \frac{e_{ik}}{\pi_k} \right) \middle| y_i \right] \\ &\cdot \mathbb{E}\left[\left(\frac{e_{im}}{\pi_m} - \frac{e_{ik}}{\pi_k} \right) \middle| y_i \right] e_m^{\top}. \end{aligned}$$

Man beachte weiters, dass

$$\mathbb{E}\left[e_{il}e_{im}|y\right] = \begin{cases} 0, & \text{für } l \neq m \\ \mathbb{E}\left[e_{il}|y\right], & \text{für } l = m. \end{cases}$$

Somit erhalten wir insgesamt

$$\begin{aligned} \mathcal{I}^{\pi}_{m}(y|\pi) = & \sum_{i=1}^{r}\sum_{l=1}^{k-1} e_l \mathbb{E}\left[\frac{e_{il}}{\pi_l^2}\middle| y_i\right] e_l^{\top} + \sum_{i=1}^{r}\sum_{l=1}^{k-1}\sum_{m=1}^{k-1} e_l \mathbb{E}\left[\frac{e_{ik}}{\pi_k^2}\middle| y_i\right] e_m^{\top} \\ & - \sum_{i=1}^{r}\sum_{l=1}^{k-1}\sum_{m=1}^{k-1} e_l \mathbb{E}\left[\left(\frac{e_{il}}{\pi_l} - \frac{e_{ik}}{\pi_k}\right)\middle| y_i\right] \\ & \cdot \mathbb{E}\left[\left(\frac{e_{im}}{\pi_m} - \frac{e_{ik}}{\pi_k}\right)\middle| y_i\right] e_m^{\top} \\ \approx & \sum_{i=1}^{r}\sum_{l=1}^{k-1} e_l \frac{w_{il}}{\pi_l^2} e_l^{\top} + \sum_{i=1}^{r}\sum_{l=1}^{k-1}\sum_{m=1}^{k-1} e_l \frac{w_{ik}}{\pi_k^2} e_m^{\top} \\ & - \sum_{i=1}^{r}\sum_{l=1}^{k-1}\sum_{m=1}^{k-1} e_l \left(\frac{w_{il}}{\pi_l} - \frac{w_{ik}}{\pi_k}\right)\left(\frac{w_{im}}{\pi_m} - \frac{w_{ik}}{\pi_k}\right) e_m^{\top}. \end{aligned}$$

Für die *beobachtete Information* bezüglich π gilt schließlich

$$I^{\pi}(y|\pi) \approx \sum_{i=1}^{r}\sum_{l=1}^{k-1}\sum_{m=1}^{k-1} e_l \left(\frac{w_{il}}{\pi_l} - \frac{w_{ik}}{\pi_k}\right)\left(\frac{w_{im}}{\pi_m} - \frac{w_{ik}}{\pi_k}\right) e_m^{\top}.$$

Abschließend können wir die *beobachtete Information* bezüglich aller Parameter schreiben als

$$I(y|\vartheta,\pi) = \begin{pmatrix} I^{\vartheta}(y|\vartheta) & 0 \\ 0 & I^{\pi}(y|\pi) \end{pmatrix}.$$

In den Nebendiagonalen steht aufgrund der Unabhängigkeit der Parametersets ϑ und π jeweils die Null.

13.2.2 *Das Negativ Binomialmodell*

Auch hier lässt sich die *beobachtete Information* wieder getrennt nach π und ϑ berechnen. Die beobachtete Information bezüglich ϑ ist bereits aus Abschnitt 13.1.2 bekannt, jene bezüglich π wird auf dieselbe Art und Weise berechnet wie in Abschnitt 13.2.1.

Teil V

NUMERISCHE ERGEBNISSE

In diesem Teil werden die Methoden zur Berechnung von Standardfehlern, welche im letzten Teil vorgestellt wurden, zunächst durch eine Monte Carlo Simulationsstudie validiert. Dabei werden die Methoden exemplarisch an einem Mitglied der linearen, einparametrigen Exponentialfamilie - der Poissonverteilung - sowie der Negativ Binomialverteilung getestet. Daraufhin werden die Methoden an einem konkreten Datenbeispiel angewandt.

14 MONTE CARLO STUDIE

14.1 STANDARDMODELL

Um die Methoden der vorangegangenen Kapiteln zu evaluieren, legen wir ein Modell vollständig fest, das bedeutet wir spezifizieren die Modellklasse, den Stichprobenumfang, die erklärenden Variablen sowie sämtliche Parameter. Aus diesem Modell - dem Standardmodell - simulieren wir schließlich den Responsevektor y. Daraufhin behandeln wir die Parameter wieder als unbekannt und schätzen diese. Somit sind sowohl die *wahren* Werte als auch die *Parameterschätzer* bekannt und Vergleiche können angestellt werden. Wiederholt man diesen Vorgang oftmals können auch Monte Carlo Schätzungen berechnet werden.

Wir legen das Standardmodell nun wie folgt fest: Die 40 elementige Responsevariable y wird durch x erklärt. Dabei ist x äquidistant aus dem Intervall $[2, 4]$. In unseren Modellen verzichten wir im Falle der nichtparametrischen Schätzung auf einen Intercept. Dies ist sinnvoll, da es dadurch zu keinerlei Identifikationsproblemen mit Massepunkten kommt.

Insgesamt werden wir vier verschiedene Varianten des Standardmodells betrachten: Einerseits unterscheiden wir zwischen normalverteilten Zufallseffekten (`gq`) und Zufallseffekten ohne Verteilungsannahme (`npml`) und andererseits zwischen der Poisson- und Negativ Binomialverteilung als konditionale Verteilung der Response. Die verschiedenen Varianten sind in Abbildung 3 dargestellt.

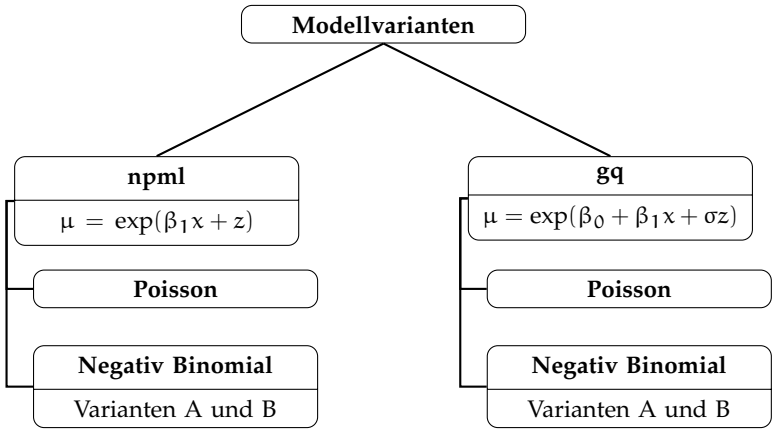

Abbildung 3: Modellvarianten

Bei der nichtparametrische Schätzung simulieren wir die Zufallseffekte aus einer zuvor spezifizierten Verteilung und nehmen dies als Grundlage für die Schätzung.

Weiters verwenden wir stets die kanonische Linkfunktion zur Poissonverteilung - also den log-Link. Wir wollen ausschließlich einen zufälligen Intercept modellieren und von zufälligen Parametern absehen. Weiters wollen wir zu Demonstrationszwecken nur Überdispersionsmodelle betrachten. Dies erhalten wir dadurch, dass die Anzahl der Cluster gleich der Anzahl der Beobachtungen gewählt wird. Schlussendlich fassen wir sämtliche Parametervorgaben sowie weitere Spezifikationen der Monte Carlo Studie in Tabelle 5 zusammen.

Parameter	Wert	*Anmerkung*
n	40	*Anzahl der Beobachtungen*
r	n	*Anzahl der Cluster*
β_0	1	*Intercept (falls verwendet)*
β_1	0	*Slopeparameter*
σ	0.5	*Standardabweichung des zufälligen Intercepts*
α	0.5	*Parameter der Negativ Binomialverteilung*
B	1000	*Monte Carlo Replikationszahl (für die Poissonverteilung)*
B_{sw}	30	*Monte Carlo Replikationszahl zur Berechnung des Sandwichschätzers*
k	10	*Anzahl der Gauß-Hermite Quadratur Punkte*

Tabelle 5: Modellspezifikationen

Um nun einen Responsevektor y zu erzeugen, der den obigen Anforderungen genügt, gehen wir nach Algorithmus 4 vor. Dieser Algorithmus erlaubt neben der Simulation aus einfachen Überdispersionsmodellen auch jene aus Varianzkomponentenmodellen.

Algorithmus 4 ist in der Funktion `rand.y(X, distribution, random.distribution='normal', sigma, beta, alpha, r, seed)` implementiert. Über die Argumente `beta`, `sigma` und `alpha` werden die gewünschten Parameter spezifiziert. Das Arguement `X` steht für die Designmatrix - egal ob mit oder ohne Interceptspalte (diese wird wenn nötig - also wenn `beta` einen Parameter mehr als die Spaltenanzahl von `X` enthält - vom Programm hinzuge-

Algorithmus 4 Zufallszahlengenerator

Wähle r, X, β, σ, Verteilungstyp für nichtparametrische Schätzung und ggf. α.

if GQ **then**
  Simuliere $Z_i \sim N(0, \sigma^2)$ für $i = 1, \ldots, r$.
else if NPML **then**
  Simuliere $Z_i \sim F(0)$ für $i = 1, \ldots, r$
end if

Setzte $Z = (Z_1, \ldots, Z_1, Z_2, \ldots, Z_2, \ldots, Z_r)$ mit Clusterlänge n/r.
Setze $\mu = X^\top \beta + Z$.

if Poisson **then**
  Simuliere $y_i \sim Poisson(\mu)$.
else if NegBin **then**
  Simuliere $y_i \sim NegBin(\mu, \alpha)$.
end if

fügt). Über `distribution` kann man zwischen `'poisson'` und `'negbin'` als konditionale Verteilung für y wählen. Das Argument `random.distribution` gibt die Verteilung der Zufallseffekte an. Implementiert sind `'normal'` für die Normalverteilung, `'t'` für die Studentsche t-Verteilung, die Gleichverteilung auf $[0, 1]$, die logistische Verteilung sowie die Lognormalverteilung. Dabei nehmen wir stets einen Erwartungswert von 0. Im Falle der Normal-, t- und der logistischen Verteilung wird der Parameter σ dahingehend verwendet, dass die Varianz gleich σ^2 ist. Im Falle der Lognormalverteilung ist σ der zweite Parameter. Der erste Parameter μ wird so berechnet, dass die Verteilung einen Erwartungswert von 1 hat. Daraufhin werden die generierten Zufallszahlen um -1 verschoben, sodass wir insgesamt wieder einen Erwartungswert von 0 erreichen.
Die Clusteranzahl wird über `r` festgesetzt. Dabei ist zu beachten, dass nur gleich große Cluster zugelassen sind. Möchte man einen bestimmten Seed festlegen, kann man dies über das Argument `seed` tun.
Als Rückgabewerte stehen `y`, `type`, `Intercept`, `distribution`, `random.distribution`, `mu` und `p` zur Verfügung. Der wichtigste Wert ist klarerweise `y`, der generierte Responsevektor. Die übrigen Werte dienen hauptsächlich der Information. So ist `Intercept=TRUE`, falls im Laufe der Berechnung eine Interceptspalte zur Designmatrix hinzugefügt wurde. In `type` wird angegeben, ob es sich um ein *Überdispersionsmodell* (also $r = n$) oder ein *Varianzkomponentenmodell* handelt. Der Wert `p` gibt die Anzahl der Regressionsparameter - also die Dimension von β - an. Im Rückgabewert `mu` sind die realisierten

konditionalen Erwartungswerte $\mu_i = g^{-1}(x_i^\top \beta + \sigma Z_i)$ abgespeichert. Der Code zu `rand.y()` ist in Abschnitt B.2 zu finden.

14.2 ERGEBNISSE

Zur Validierung der Studienergebnisse berechnen wir mehrere Kennzahlen. Dies ist zum einen das Monte Carlo Mittel der geschätzten Parameter - $\texttt{mean}(\hat{\beta})$ sowie gegebenenfalls $\texttt{mean}(\hat{\sigma})$ und $\texttt{mean}(\hat{\alpha})$. Dadurch lässt sich erkennen, wir gut die einzelnen Modelle im Schnitt die wahren Parameter reproduzieren. Weiters berechnen wir deren jeweilige Monte Carlo Standardabweichungen $\texttt{sd}(\hat{\beta})$ und gegebenenfalls $\texttt{sd}(\hat{\sigma})$ und $\texttt{sd}(\hat{\alpha})$. Diese Werte vergleichen wir mit den mittleren durch die Sandwichschätzer ermittelten Standardfehler $\texttt{sd}_{\texttt{sand}}$, jenem durch die beobachtete Information berechnetem Standardfehler $\texttt{sd}_o$ sowie jenem, den die Funktion `gamlssNP()` zur Zeit verwendet, $\texttt{sd}_c$. Weiters wird unter *EM* angegeben, in wie vielen Fällen der EM-Algorithmus konvergiert hat.

14.2.1 *Ergebnisse: Poisson - GQ*

Die Ergebnisse für den Fall normalverteilter zufälliger Effekte sind in Tabelle 6 zusammengefasst. Die empirischen Mittel der geschätzten Parameter passen sehr gut zu den *wahren Werten*. Einzig im Fall von σ trifft der Schätzer im Mittel den wahren Wert nicht so gut. Es bestätigt sich die These, dass die Schätzung der Standardfehler ausschließlich basierend auf der *vollständigen Information* $\texttt{sd}_c$ keine brauchbaren Werte liefert. Im Vergleich zu diesen sind die über die *beobachtete Information* berechnete Schätzung $\texttt{sd}_o$ sowie der Sandwichschätzer $\texttt{sd}_{\texttt{sand}}$ deutlich besser. Diese Aussagen gelte zumindest für die relevanten Parameter β_0 und β_1.

Poisson - GQ: EM: 999 $(k = 10)$

	$\beta_0 = 1$	$\texttt{mean}(\hat{\beta}_0) = 0.947$	$\beta_1 = 0$	$\texttt{mean}(\hat{\beta}_1) = 0.019$
	$\sigma = 0.5$	$\texttt{mean}(\hat{\sigma}) = 0.392$		

	$\texttt{sd}(\hat{\vartheta})$	$\texttt{sd}_{\texttt{sand}}$	$\texttt{sd}_o$	$\texttt{sd}_c$
β_0	0.869	0.634	0.600	0.479
β_1	0.283	0.210	0.195	0.156
σ	0.187	0.357	0.171	0.093

Tabelle 6: Ergebnisse der Monte Carlo Studie

In Abbildung 4 sind die Monte Carlo Histogramme der Parameterschätzer $\hat{\beta}_0$ und $\hat{\beta}_1$ dargestellt. Man erkennt deutlich, dass im Mittel

die *wahren* Parameter gut geschätzt wurden. Weiters gab es auch keine extremen Werte. Solche extremen Schätzungen können vor allem dann vorkommen, wenn der EM-Algotihmus nicht innerhalb der vorgegebenen maximalen Iterationszahl 200 konvergiert.

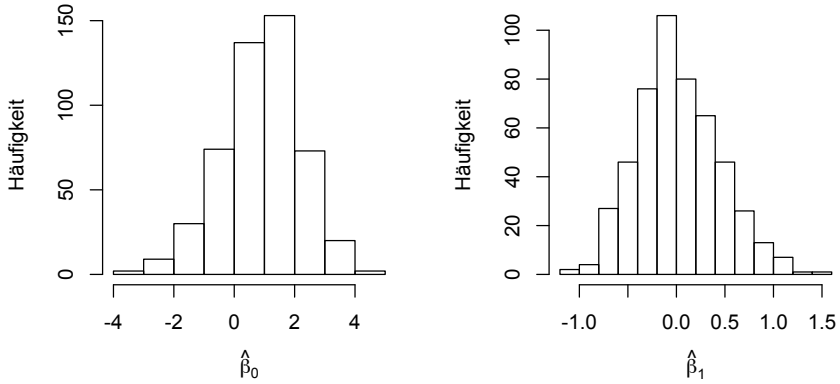

Abbildung 4: Monte Carlo Verteilung der Parameterschätzer

14.2.2 *Ergebnisse: Poisson - NP*

Wir betrachten nun drei verschiedene Szenarien. Dabei unterscheiden sich die Szenarien darin, aus welcher Verteilung die Zufallseffekte gezogen werden. Details sind in Tabelle 7 zu finden.

Szenario 1:	$N(0, 0.5^2)$
Szenario 2:	$U(0, 1)$
Szenario 3:	$N(0.5, 0.7^2)$

Tabelle 7: Szenarien

Bei der Annahme normalverteilter zufälliger Effekte legen wir die Anzahl der Gauß-Hermite Quadraturpunkte, mit der die Integrale approximiert werden, vorab fest. Im Falle der nichtparametrischen Schätzung ist die Anzahl der zu schätzenden Massepunkte k nicht a priori klar und müsste eigentlich mitgeschätzt werden. Da dies jedoch nicht möglich ist, führen wir in jedem Monte Carlo Schritt einen Test durch, um einen sinnvollen Wert für k zu ermitteln. Dazu schätzen wir zunächst ein Modell mit $k = 2$ und berechnen die Disparität[1], $-2 \cdot \ell(\hat{u}|y)$. Daraufhin erhöhen wir k in jedem weiteren

1 Diese Disparität wir auch *Globale Devianz* genannt.

Schritt um 1. Dies tun wir solange, bis die relative Veränderung der Disparität des Modells mit $k+1$ Massepunkten im Vergleich zum Modell mit k Massepunkten kleiner als die vorgegebene Schranke $\varepsilon = 0.01$ ist. Falls der EM-Algorithmus im Modell mit $k+1$ Massepunkten keine Konvergenz erreicht hat, wird ebenfalls das Modell mit k Massepunkten präferiert. Insgesamt ist eine Höchstanzahl von $k = 10$ vorgegeben.
In Tabelle 8 geben wir eine Auswertung an, mit welchen Werten für k in der Monte Carlo Studie schließlich gerechnet wurde.

Replikationszahl: 1000

	$k=2$	$k=3$	$k=4$
Szenario 1	97.8%	2.2%	0%
Szenario 2	97.9%	2.1%	0%
Szenario 3	87.3%	12.1%	0.6%

Tabelle 8: Anzahl der Massepunkte k

Nun betrachten wir die Standardfehler für den Regressionsparameter $\beta_1 = 0$. Diese Ergebnisse sind in Tabelle 12 zusammengefasst[2]. Es ist deutlich erkennbar, dass der Sandwichschätzer $\mathtt{sd}_{\mathtt{sand}}$ die empirische Standardabweichung am besten erreicht. Insbesondere ist auch darauf hinzuweisen, dass die bisher verwendeten Standardfehler $\mathtt{sd}_c$, die ausschließlich auf der *vollständigen Information* basieren, die tatsächliche Variablität ungemein unterschätzen und nicht für Analysen herangezogen werden sollten. Der Standardfehler basierend auf der *beobachteten Information* $\mathtt{sd}_o$ liefert auch bereits eine gute Näherung. Dieser ist insbesondere interessant, da die Berechnung kaum numerischen Aufwand in sich birgt. Dieser zusätzliche Aufwand bei der Berechnung von $\mathtt{sd}_{\mathtt{sand}}$ kann insbesondere bei großen Modellen (großer Stichprobenumfang, viele erklärende Variablen, großes k) eine numerische Hürde darstellen.

Die Verteilung des geschätzten Regressionsparameters $\hat{\beta}_1$ ist in Abbildung 5 (beispielhaft für Szenario 1) als Histogramm dargestellt. Wieder sind keine extremen Werte erkennbar und die Verteilung ist zentriert um den wahren Parameterwert 0.

2 Es werden stets nur jene Ergebnisse herangezogen, wo der EM Algorithmus Konvergenz erreicht hat.

Poisson - NP

		$\texttt{mean}(\hat{\beta}_1)$	$\texttt{sd}(\hat{\beta}_1)$	$\texttt{sd}_{\texttt{sand}}$	$\texttt{sd}_{\texttt{o}}$	$\texttt{sd}_{\texttt{c}}$	EM
Szenario 1	β_1	−0.014	0.372	0.307	0.292	0.262	996
Szenario 2	β_1	−0.012	0.258	0.236	0.222	0.208	995
Szenario 3	β_1	−0.009	0.431	0.248	0.244	0.197	999

Tabelle 9: Ergebnisse der Monte Carlo Studie

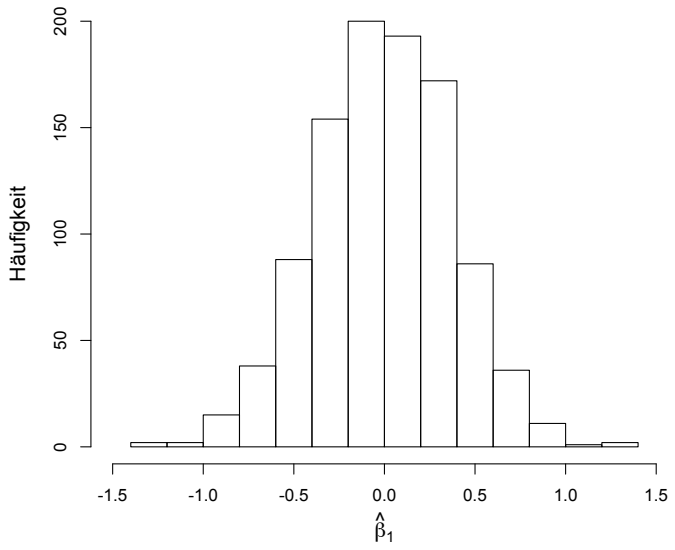

Abbildung 5: Monte Carlo Verteilung von $\hat{\beta}_1$ im Szenario 1

14.2.3 *Ergebnisse: Negativ Binomial - GQ*

Die Durchführung einer Simulationsstudie im Falle der Negativen Binomialverteilung erwies sich als schwieriger als für die Poissonverteilung. Dies hat mehrere Gründe: Das Funktion `gamlssNP()`, welche zur Schätzung der Modelle verwendet wird, lässt auch negative Schätzungen des Parameters α zu. Die Negative Binomialverteilung ist jedoch nur für strikt positive Werte von α definiert. In diesen Fällen ist somit auch keine Berechnung von Sandwichschätzern mehr möglich. Weiters treten vergleichsweise viele Fälle mit Konvergenzschwierigkeiten - sowohl Divergenz bezüglich des Parameters α also auf Divergenz des EM-Algorithmus - auf. Ein weiteres Problem ist dadurch gegeben, dass hin und wieder recht hohe Werte für die Response-Variable y simuliert wurden. In diesen Fällen ist die numerische Auswertung der Gamma-, Digamma und Trigammafunktion nicht mehr möglich und `R` setzt die Resultate auf `Inf`. Nachdem somit für die Analyse nur noch ein kleiner Teil der Ergebnisse verwendbar ist, wurde bei der Gauß-Quadratur die Replikationszahl auf 5000 erhöht. Damit erhalten wir 4335 brauchbare Datensätze. Bei der nichtparametrischen Schätzung traten die zuvor genannten Probleme vermehrt auf, weswegen eine alternative Herangehensweise gewählt wurde: Die Replikationszahl wurde nicht mehr zu Beginn fix vorgegeben. Vielmehr wurde diese solange erhöht, bis zumindest 500 vollständige und gültige Datensätze vorlagen.
Auf Grund all dieser Probleme muss darauf hingewiesen werden, dass die präsentierten Ergebnisse mit Vorsicht zu genießen sind. Zukünftige Untersuchungen betreffend der Negativen Binomialverteilung sollten zur Modellschätzung nicht die Funktion `gamlssNP()` heranziehen beziehungsweise diese zunächst dahingegen adaptieren, dass zumindest der gewaltigste Nachteil - die potenzielle negative Schätzung von α - nicht auftritt. Diesem Problem könnte man etwa durch eine interne Reparametrisierung $\alpha^* = \exp(\alpha)$ Herr werden.
In Tabelle 10 sind die Ergebnisse der Simulationsstudie angegeben. Auf Grund der zuvor beschriebenen Unzulänglichkeiten werdern keine Ergebnisse für den Parameter α angegeben. Ansonsten zeigt sich wieder ein ähnliches Bild wie bereits bei der Poissonverteilung: Die Standardfehler, die einzig auf der vollständigen Information basierend berechnet werden, unterschätzen die tatsächlichen Standardfehler (ausgedrückt durch die Monte Carlo Standardabweichung) gewaltig. So unterschätzt etwa `sd.c` für den Regressionsparameter β_1 bei der Variante B die *wahre Standardabweichung* um 23.7%. Die Schätzung durch `sd.o` - also basierend auf der beobachteten Information - liegt im gleichen Setting jedoch nur um 5.1% daneben.
Bei der Negativen Binomialverteilung vergleichen wir weiters zwei Schätzvarianten: Dabei muss man - wie in Abschnitt 13.1.2 dargelegt - erwarten, dass Variante B stets bessere Ergebnisse liefert als Varian-

te A. Diese Erwartungen werden für die Regressionsparameter auch durchgehend erfüllt!
Wie wir bereits bei der Poissonverteilung gesehen haben, sind die Ergebnisse für die beiden Regressionsparameter β_0 und β_1 sehr gut, wohingegen für die Schätzung der Varianz der Zufallseffekte σ^2 nicht so gute Ergebnisse erzielt werden. Dies ist auch nicht weiter verwunderlich, da es sich um einen gewissermaßen *impliziten* bzw. *versteckten* Parameter handelt. Darüber hinaus ist die Schätzung dieses Parameters sowie die genaue Kenntnis dessen Standardfehlers nicht sonderlich relevant fur die Betreibung statistischer Inferenz.

Negativ Binomial - GQ: # 4335 / 5000

				Variante A		Variante B	
	ϑ	$\texttt{mean}(\hat{\vartheta})$	$\texttt{sd}(\hat{\vartheta})$	`sd.o`	`sd.c`	`sd.o`	`sd.c`
β_0	1	0.984	1.598	1.483	1.284	1.505	1.209
β_1	0	−0.005	0.521	0.487	0.420	0.495	0.398
σ	0.5	0.225	0.327	0.304	0.250	0.458	0.247

Tabelle 10: Ergebnisse der Monte Carlo Studie

14.2.4 *Ergebnisse: Negativ Binomial - NP*

Wie bereits im letzten Abschnitt erläutert wird bei der nichtparametrischen Schätzung nicht mehr eine fixe Replikationszahl vorgegeben. Vielmehr wird diese solange erhöht, bis zumindest 500 verwertbare Datensätze vorliegen. Die Anzahl der benötigten Replikationen ist zum Teil sehr groß und wird in Tabelle 11 angegeben.

	Anzahl der gültigen Datensäzte	Replikationszahl
Szenraio 1	519	125,000
Szenraio 2	535	22,500
Szenraio 3	607	20,000

Tabelle 11: Replikationszahlen

Die Ergebnisse der Simulationsstudie sind in Tabelle 12 zusammengefasst. Bei allen gültigen Simulationsschritten wurde - ähnlich wie bei der Poissonverteilung - schlussendlich mit $k = 2$ gerechnet. Der wahre Parameter $\beta_0 = 0$ wurde im Mittel in allen drei Szenarien sehr gut getroffen. Bei den Standardfehlern werden bei der Betrachtung der beobachteten Information bessere Ergebnisse erzielt als bei der Betrachtung der vollständigen Information. Dies ist konsistent zu allen bisher präsentierten Simulationsergebnissen und untermauert die

in dieser Arbeit hervorgehobene Relevanz der Hinzunahme der fehlenden Information. Darüber hinaus werden bei Variante B wiederum - wie erwartet - bessere Ergebnisse erzielt als bei Variante A.

Negativ Binomial - NP

				Variante A		Variante B	
		$\texttt{mean}(\hat{\beta}_1)$	$\texttt{sd}(\hat{\beta}_1)$	$\texttt{sd}_o$	$\texttt{sd}_c$	$\texttt{sd}_o$	$\texttt{sd}_c$
Szenario 1	β_1	0.022	0.398	0.360	0.359	0.381	0.219
Szenario 2	β_1	0.021	0.405	0.351	0.350	0.381	0.209
Szenario 3	β_1	0.031	0.501	0.354	0.353	0.379	0.204

Tabelle 12: Ergebnisse der Monte Carlo Studie

In Abbildung 6 ist beispielhaft für Szenario 1 die Verteilung von $\hat{\beta}_1$ dargestellt. Es zeigt sich deutlich, dass die geschätzen Werte um den wahren Parameterwert 0 zentriert sind.

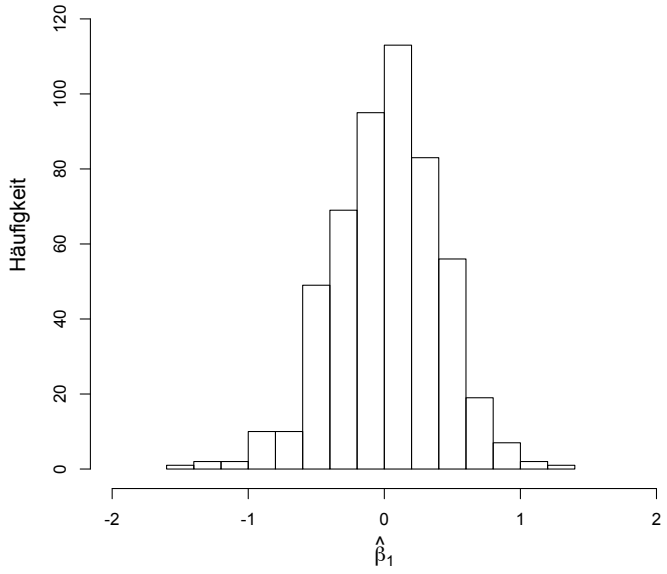

Abbildung 6: Monte Carlo Verteilung von $\hat{\beta}_1$ im Szenario 1

14.3 CONCLUSIO

Zusammenfassend ist festzuhalten, dass bei der Berechnung von Standardfehlern basierend auf der beobachteteten Information sowohl bei der Poisson- als auch bei der Negativen Binomialverteilung sehr gute Ergebnisse erzielt werden. Dies ist insbesondere bei den zur Betreibung statistischer Inferenz wichtigen Regressionsparametern der Fall. Zum Vergleich wurden auch stets Standardfehler basierend auf der vollständigen Information berechnet, da dies die aktuell übliche Methode darstellt. Bei allen Simulationsergebnissen erzielte diese Methode (zum Teil sehr deutlich) schlechtere Ergebnisse, was die Relevanz der hier vorgestellten Methode untermauert.
Verwendet man die Funktion `gamlssNP()`, um Negativ Binomialmodelle zu schätzen, ist darauf zu achten, dass in einer großen Anzahl an Fällen negative Werte für α geschätzt werden. Da die Negative Binomialverteilung jedoch nur für strikt poisitve α definiert ist, stellt dies erhebliche Probleme dar. Eine Korrektur der Funktion `gamlssNP()` ist daher notwendig, um die Ergebnisse der Funktion direkt zur Berechnung der Standardfehler heranziehen zu können.

15 DATENBEISPIEL

15.1 DATEN

Wir wollen nun unsere Methoden zur Berechnung von Standardfehlern an einem konkreten Datenbeispiel demonstrieren. Dazu wird auf das Ergebnis einer Studie des Landeshygienikers für die Steiermark zurückgegriffen. In dieser Studie wurden über ein Jahr[1] hinweg alle zwei Wochen an insgesamt sieben verschiedenen Orten in Graz Messungen durchgeführt. Gemessen wurden in dieser Studie die Bakterien- und Pilzkonzentration in der Außenluft. Die sieben verschiedenen Orte zeichnen sich durch je eine Besonderheit bzgl. Bakterien- bzw. Pilzbelastung aus. Für unser Beispiel werden nur die Daten zur Bakterienbelastung an zwei Orten (A und B) verwendet. Diese beiden Orte liegen sehr nahe beieinander und unterscheiden sich nur durch ein Merkmal: In unmittelbarer Nähe zu Ort B befindet sich eine Kompostierungsanlage.

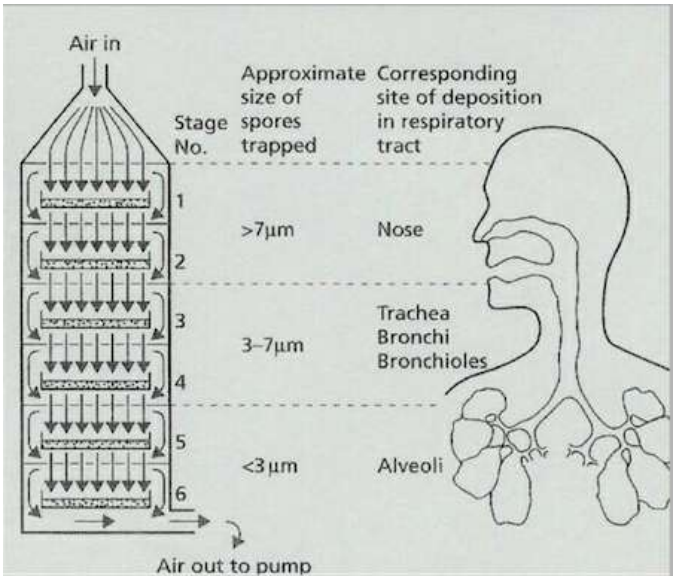

Abbildung 7: Sechsstufiger Luftsammler[2]

Die Bakterienkonzentration wurde mit Hilfe eines sechsstufigen Luftsammlers (Andersen Air Sampler) gemessen. Dabei wird die Außenluft durch sechs nach unten hin immer feiner werdende Filter gezogen. Die größten Mikroorganismen bleiben demnach in den

1 8. März 1995 bis 21. Februar 1996

2 http://archive.bio.ed.ac.uk/jdeacon/FungalBiology/chap10_3.htm, 12.4.2013

obersten Filtern hängen, wohingegen kleinere Mikroorganismen erst durch die unteren Filter aufgefangen werden. Ein ähnliches Filtersystem ist auch das menschliche Atemsystem - je kleiner die Keime in der Luft sind, desto weiter können sie vordringen. Insbesondere stellen jene Keime, die auf den Stufen 4-6 aufgefangen werden, für den Menschen eine Gefahr dar, da sie bis in die Lunge vordringen. Die Funktionsweise des Luftsammlers ist auch in Abbildung 7 dargestellt.

Da die Anzahl der kolonienbildenden, luftgetragenen Mikroorganismen, *CFU*[3], vom Wetter abhängt, enthält der Datensatz weiters auch die jeweilige Temperatur (in Grad Celsius), `temp`, und relative Luftfeuchtigkeit (in %), `humi`.

Der Datensatz wurde als `data.frame` unter dem Namen `data` abgespeichert. Dabei steht `site` für den Ort und `date` für den Tag (im Format dd.mm) der Messung. Die Variablen b_i, für $i = 1, \ldots, 6$ stehen für die Anzahl der CFUs auf der Stufe i. Der vollständige Datensatz ist im Abschnitt C.1 zu finden.

Für die späteren Beispiele ist es oft notwendig, die Struktur des Datensatzes anzupassen: Ziel ist es, dass die Bakterienanzahl in nur einer Variablen (`bac`) abgespeichert ist und eine neue Variable `stage` (die Stufe im Luftsammler) eingeführt wird. Zusätzlich sollen die Variablen `stage` und `site` als Faktoren abgespeichert werden.

Code 1: Datentransformation

```
data.bac <- rbind.data.frame(cbind(data[,1:4], stage = 4, bac =
    data$b4), cbind(data[,1:4],stage = 5, bac = data$b5), cbind(
    data[,1:4], stage = 6, bac = data$b6))
data.bac$stage = as.factor(data.bac$stage)
data.bac$site = as.factor(data.bac$site)
```

In den Daten gibt es einen unrealistisch hohen Wert (Bakterienanzahl von 36 auf Stufe 6), weshalb beide Beobachtungen an dem Tag mit der extremen Messung nicht verwendet werden.

```
falsch <- subset(data.bac, data.bac$bac >20)$date
data.bac <- data.bac[data.bac$date != falsch,]
```

15.2 MODELL

Wir gehen davon aus, dass die Bakterienanzahlen bei einer Messung auf den unterschiedlichen Stufen des Luftsammlers nicht unabhängig sind. Um deren Korrelation zu berücksichtigen, bekommen die am

3 colonies forming units

selben Tag und am selben Ort durchgeführt wurden einen gemeinsamen Zufallseffekt. Unsere Clustergrößen sind demnach $n_i = 3$. Die folgenden vier Code-Zeilen zeigen, wie eine geeignete Cluster-Zeile erzeugt werden kann.

Code 2: Clusterung

```
o <- order(data.bac$site,data.bac$date)
data.bac <- data.bac[o,]
cluster <- gl(50,3)
data.cluster <- data.frame(data.bac,cluster)
```

Wir betrachten nun das Modell

$$\texttt{bac} \sim 1 + \texttt{stage} * \texttt{site} + \texttt{temp} + \texttt{temp}^2.$$

Dabei nehmen wir an, dass `bac` konditional Poisson- bzw. Negativ Binomialverteilt ist. Wir schätzen die Verteilung der Zufallseffekte sowohl nichtparametrisch also auch parametrisch unter Annahme einer Normalverteilung. Der folgende Code zeigt den Aufruf zur Modellschätzung im Falle der Poissonverteilung. Für die Negativ Binomialverteilung muss `family=NBI()` gesetzt werden.

Code 3: Modellschätzung

```
mod.np <- gamlssNP(bac~stage*site+temp+I(temp^2)-1, family=PO(),
    random=~1|cluster,data=data.cluster, mixture="np", K=2)
mod.gq <- gamlssNP(bac~stage*site+temp+I(temp^2), family=PO(),
    random=~1|cluster, data=data.cluster, mixture="gq", K=5)
```

In unserem Modell werden die Messstelle, in deren Nähe sich keine Kompostierungsanlage befindet, sowie die Stufe 4 des Luftsammlers als Referenzgrößen genommen.

Anzumerken ist noch, dass bei einer Modellschätzung mittels `gamlssNP()` bei der nichtparametrischen Schätzung ein Intercept hinzugenommen werden muss. Dieser entspricht dann dem ersten Massepunkt. Bei Modellen, die mit `alldist()` oder `allvc()` geschätzt werden, ist ein Intercept bei der nichtparametrischen Schätzung ausgeschlossen. Die Parameter zu den Massepunkten werden also direkt als Slopeparameter geschätzt. Schätzt man dasselbe Modell einmal mit `gamlssNP() - mod1 -` und einmal mit `alldist()` bzw. `allvc() - mod2`, erhält man die Massepunkte von `mod2`, indem man den geschätzten Intercept aus `mod1` zu den geschätzten Parametern der jeweiligen Massepunkte hinzuzählt. Es ist jedoch darauf zu achten, dass beide Modelle nicht exakt dieselben Werte liefern. Die Unterschiede sind allerdings meist marginal.

15.3 ERGEBNISSE

Die Ergebnisse unserer Schätzung sind in Tabelle 13 zusammengefasst. Der Wert für k im Falle der nichtparametrischen Schätzung wurde vorab sequentiell bestimmt: Begonnen wird mit $k = 2$. Daraufhin wird k schrittweise erhöht, bis sich die Disparität[4] nicht mehr verkleinert. Dies tritt dann ein, wenn Massepunkte nicht mehr unterschieden werden können. Durch diese Methode wurde im Poissonmodell ein $k = 2$ als *optimal* bzw. *gut genug* identifiziert. Die Disparitäten und AIC-Werte sind in Tabelle 14 zusammengefasst. Im obigen Teil der Abbildung ist die Entwicklung der Disparität im EM-Algorithmus dargestellt. Im unteren Teil sind die geschätzten Massepunkte zu sehen. Es ist deutlich erkennbar, dass die beiden Massepunkte klar unterscheidbar und somit für die Modellierung relevant sind.
Das Modell ist mit 150 Beobachtungen und insgesamt 8 Regressionsparametern recht groß. Dies merkt man insbesondere bei der Berechnung der Sandwichschätzer. Auf Grund des Umfangs wurde im Falle der Schätzung mittels Gauß-Hermite Quadratur *nur* ein $k = 5$ gewählt.

Poisson

	NPML (k=2)			**GQ (k=5)**		
Effect	$\hat{\beta}$	$\mathtt{sd}_{\mathtt{sand}}$	$\mathtt{sd}_{\mathtt{c}}$	$\hat{\beta}$	$\mathtt{sd}_{\mathtt{sand}}$	$\mathtt{sd}_{\mathtt{c}}$
`Intercept`[5]	−0.045	0.280	0.196	0.214	0.291	0.190
`stage 5`	0.353	0.253	0.176	0.398	0.325	0.197
`stage 6`	−0.340	0.295	0.216	−0.296	0.348	0.233
`site B`	−0.033	0.322	0.213	0.026	0.341	0.212
`stage 5:site B`	−0.442	0.436	0.289	−0.442	0.473	0.289
`stage 6:site B`	0.418	0.456	0.309	0.418	0.481	0.310
`temp`	0.076	0.026	0.018	0.069	0.026	0.018
$\mathtt{temp}^2$	−0.003	0.0009	0.0006	−0.002	0.0009	0.0006
σ				0.595	0.082	0.069
Disparität:	525.02			529.185		
AIC:	545.02			547.185		

Tabelle 13: Ergebnisse

4 Im Zweifelsfall kann zusätzlich das AIC-Kriterium herangezogen werden.
5 Im Falle der nichtparametrischen Schätzung wurde ohne Intercept modelliert, um Identifikationsprobleme zu umgehen. Die präsentierten Ergebnisse sind also für `stage 4` zu verstehen.

In den Ergebnissen ist erkennbar, dass der Sandwichschätzer stets höhere Werte liefert als der bisher verwendete Schätzer $\texttt{sd}_c$. Dies ist ein Indikator dafür, dass $\texttt{sd}_c$ die tatsächlichen Standardfehler unterschätzt.

	$k = 2$	$k = 3$	$k = 4$
Disparität	525.02	525.029	525.03
AIC	545.02	549.029	553.03

Tabelle 14: A priori Festlegung von k im Poissonmodell

Im Falle der nichtparametrischen Schätzug werden auch zwei Massepunkte geschätzt. Allerdings wird nur der zweite (also der größere der beiden) im Output angezeigt, da der erste im Intercept (also im Schätzer zur Variablen `stage 4`) enthalten ist. Die Ergebnisse sind in Tabelle 15 angegeben.

	`MASS1`	`MASS2`
Massepunkte		1.089
Gewichte	0.792	0.208

Tabelle 15: Geschätzte Massepunkte und Gewichte wie in Output

Möchte man die beiden Massepunkte explizt haben, kann diese leicht über den Zusammenhang

$$\texttt{MASS1.exp} = \texttt{Intercept},$$
$$\texttt{MASS2.exp} = \texttt{MASS2} + \texttt{Intercept}.$$

berechnet werden. Damit erhält man die in Tabelle 16 angegebenen Ergebnisse.

	`MASS1.exp`	`MASS2.exp`
Massepunkte	−0.045	1.044
Gewichte	0.792	0.208

Tabelle 16: Geschätzte Massepunkte und Gewichte (explizit)

In Abbildung 8 sind die Entwicklungen der Disparität sowie der geschätzen Masspunkte im Laufe des EM-Algorithmus dargestellt.

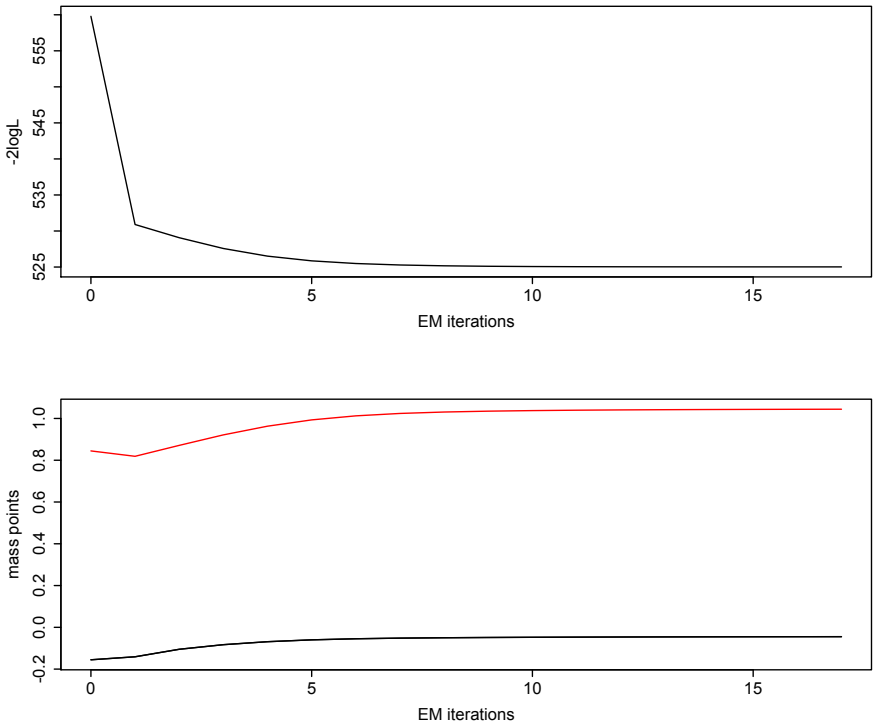

Abbildung 8: Poissonmodell: NP

Schätzt man das Modell unter der Annahme einer konditionalen Negativen Binomialverteilung, erhält man sowohl bei der Schätzung mittels Gauß-Hermite Quadratur als auch bei der nichtparametrischen Schätzung negative Werte für $\hat{\alpha}$. Wie bereits in den vorherigen Kapiteln diskutiert, ist in diesem Fall die Berechnung der Standardfehler problematisch, weswegen auf die Angabe der Ergebnisse verzichtet wird.

Teil VI

APPENDIX

Der Appendix gliedert sich in drei Teile. Im ersten Teil werden Ergänzungen - insbesondere ausführliche Rechnungen und Herleitungen - zu den vorangegangenen Kapiteln angegeben. Dies soll das schnelle und einfache Verständnis der Theorie erleichtern. Der zweite Teil ist Implementierungen gewidmet. Einige Methoden, die in dieser Arbeit vorgestellt werden, wurden in R implementiert und sind - sofern es der Code-Umfang zulässt - in diesem Abschnitt zu finden. Der letzte Teil sind die Daten, mit denen gearbeitet wurde, abgedruckt. Dadurch ist die Nachvollziehbarkeit der Ergebnisse gewährleistet.

A

BERECHNUNGEN

A.1 ÜBERDISPERSIONSMODELLE

In diesem Abschnitt sind Details und Ergänzungen zu Kapitel 3 zu finden.
In Abschnitt 3.1 wird die für den EM Algorithmus zentrale Funktion zitiert. Die Herleitung dieser wird an dieser Stelle nachgeholt.

$$\begin{aligned}
Q(\vartheta|\vartheta^{(t)}) &= \sum_{i=1}^{n} \frac{1}{f(y_i|\vartheta^{(t)})} \int \log f(y_i, Z_i|\vartheta) f(y_i|Z_i, \vartheta^{(t)}) \phi(Z_i) dZ_i \\
&\approx \sum_{i=1}^{n} \frac{\sum_{j=1}^{k} \log f(y_i, z_j|\vartheta) f(y_i|z_j, \vartheta^{(t)}) \pi_j}{\sum_{j=1}^{k} f(y_i|z_j, \vartheta^{(t)}) \pi_j} \\
&= \sum_{i=1}^{n} \sum_{j=1}^{k} w_{ij}^{(t)} \log f(y_i, z_j|\vartheta) \\
&= \sum_{i=1}^{n} \sum_{j=1}^{k} w_{ij}^{(t)} [\log \pi_j + \log f(y_i|z_j, \vartheta)],
\end{aligned}$$

wobei $w_{ij} = \frac{\pi_j f_{ij}}{\sum_{l=1}^{k} \pi_l f_{il}}$ und $f_{ij} := f(y_i|z_j, \vartheta)$.
Da wir stets nur die Approximation von $Q(\vartheta|\vartheta^{(t)})$ benötigen werden, führen wir die Notation

$$Q_k(\vartheta|\vartheta^{(t)}) = \sum_{i=1}^{n} \sum_{j=1}^{k} w_{ij}^{(t)} [\log \pi_j + \log f(y_i|z_j, \vartheta)]$$

ein.

A.2 ZUFÄLLIGE KOEFFIZIENTENMODELLE

Hier werden Ergänzungen zu Kapitel 4 angegeben. Die Berechnung der Kovarianz der transformierten zufälligen Effekte in Abschnitt 4.1 kann in der folgenden Rechnung nachvollzogen werden.

$$\begin{aligned}
\mathrm{Cov}[Z_i^*, U_i] &= \mathbb{E}[Z_i^* U_i] - \mathbb{E}[Z_i^*] \underbrace{\mathbb{E}[U_i]}_{=0} = \mathbb{E}\left[\frac{Z_i U_i - \rho U_i^2}{\sqrt{1-\rho^2}}\right] \\
&= \frac{1}{\sqrt{1-\rho^2}} \left(\mathbb{E}[Z_i U_i] - \rho \underbrace{\mathbb{V}\mathrm{ar}[U_i]}_{=1} \right) \\
&= \frac{1}{\sqrt{1-\rho^2}} \left(\mathrm{Cov}[Z_i, U_i] + \underbrace{\mathbb{E}[Z_i]}_{=0} \underbrace{\mathbb{E}[U_i]}_{=0} - \rho \right) \\
&= \frac{1}{\sqrt{1-\rho^2}} \left(\rho \underbrace{\sqrt{\mathbb{V}\mathrm{ar}[Z_i]}}_{=1} \underbrace{\sqrt{\mathbb{V}\mathrm{ar}[U_i]}}_{=1} - \rho \right) = 0.
\end{aligned}$$

A.3 VARIANZKOMPONENTENMODELLE

Im Normal-Normal Modell gilt $y_{ij} \sim N(x_{ij}^\top \beta, \sigma_Z^2 + \sigma^2)$. Für die Kovarianzen ergibt sich daraus schließlich

$$\begin{aligned}
\mathrm{Cov}[y_{ij}, y_{kl}] &= \mathbb{E}[y_{ij} y_{kl}] - \mathbb{E}[y_{ij}]\mathbb{E}[y_{kl}] \\
&= \mathbb{E}\left[(x_{ij}^\top \beta)(x_{kl}^\top \beta) + x_{ij}^\top \beta \varepsilon_{kl} + x_{kl}^\top \beta \varepsilon_{ij} + \varepsilon_{ij}\varepsilon_{kl} \right] \\
&\quad - (x_{ij}^\top \beta)(x_{kl}^\top \beta) \\
&= x_{ij}^\top \beta \underbrace{\mathbb{E}[\varepsilon_{kl}]}_{=0} + x_{kl}^\top \beta \underbrace{\mathbb{E}[\varepsilon_{ij}]}_{=0} + \mathbb{E}[\varepsilon_{ij}\varepsilon_{kl}] \\
&= \mathbb{E}\left[(\sigma_Z Z_i + \tilde{\varepsilon}_{ij})(\sigma_Z Z_k + \tilde{\varepsilon}_{kl}) \right] \\
&= \sigma_Z^2 \mathbb{E}[Z_i Z_k].
\end{aligned}$$

A.4 DAS NEGATIV BINOMIALMODELL

Dieser Abschnitt dient als Ergänzung zu Kapitel 6. Die marginale Dichte der Y_i im Negativ Binomialmodell kann wie folgt hergeleitet werden:

$$\begin{aligned}
f(y_i|\mu_i) &= \int_0^\infty f(y_i|Z_i,\mu_i)f(Z_i)dZ_i \\
&= \frac{\alpha^\alpha}{\Gamma(\alpha)y_i!}\int_0^\infty (\mu_i Z_i)^{y_i} e^{-Z_i\mu_i} Z_i^{\alpha-1} e^{-\alpha Z_i} dZ_i \\
&= \frac{\alpha^\alpha \mu_i^{y_i}}{\Gamma(\alpha)y_i!}\int_0^\infty e^{-Z_i(\mu_i+\alpha)} Z_i^{y_i+\alpha-1} dZ_i \\
&= \frac{\alpha^\alpha \mu_i^{y_i}}{\Gamma(\alpha)y_i!}\frac{\Gamma(y_i+\alpha)}{(\mu_i+\alpha)^{y_i+\alpha}} \\
&= \frac{\Gamma(y_i+\alpha)}{\Gamma(\alpha)y_i!}\left(\frac{\mu_i}{\mu_i+\alpha}\right)^{y_i}\left(\frac{\alpha}{\mu_i+\alpha}\right)^{\alpha}.
\end{aligned}$$

Für $Y_i \overset{ind}{\sim} NegBin(\alpha,\mu_i)$ erhält man aus der Turmeigenschaft die ersten beiden Momente über,

$$\begin{aligned}
\mathbb{E}[Y_i] &= \mathbb{E}\left[\mathbb{E}[Y_i|Z_i]\right] = \mu_i\mathbb{E}[Z_i] = \mu_i\frac{\alpha}{\alpha} = \mu_i, \\
\mathbb{V}ar[Y_i] &= \mathbb{E}\left[\mathbb{V}ar\left[Y_i|Z_i\right]\right] + \mathbb{V}ar\left[\mathbb{E}\left[Y_i|Z_i\right]\right] = \mu_i\mathbb{E}\left[Z_i\right] + \mu_i^2\mathbb{V}ar\left[Z_i\right] \\
&= \mu_i\frac{\alpha}{\alpha} + \mu_i^2\frac{\alpha}{\alpha^2} = \mu_i + \frac{\mu_i^2}{\alpha}.
\end{aligned}$$

A.5 STANDARDFEHLER IM EXPONENTIALFAMILIENMODELL

A.5.1 *Normalverteilte zufällige Effekte*

Die folgenden Herleitungen sind Abschnitt 10.1 zuzuordnen. Zunächst ist eine detaillierte Berechnung der Ableitung der Gewichte w_{il} nach dem Parameter ϑ (vgl. Abschnitt 10.1) angegeben.

$$\begin{aligned}
\frac{\partial w_{il}}{\partial\vartheta} &= \frac{\partial}{\partial\vartheta}\frac{\exp\left(\frac{\vartheta y_i - b(\vartheta)}{a(\phi)} + c(y_i,\phi)\right)\pi_l}{\sum_{m=1}^k f(y_i|z_m,\vartheta)\pi_m} \\
&= \frac{f(y_i|z_l,\vartheta)\pi_l}{f_k(y_i|\vartheta)}\left[\frac{y_i - b'(\vartheta)}{a(\phi)} - \frac{\sum_{m=1}^k \pi_m \frac{\partial}{\partial\vartheta} f(y_i|z_m,\vartheta)}{f_k(y_i|\vartheta)}\right] \\
&= \frac{\pi_l}{f_k(y_i|\vartheta)}\left(\frac{\partial}{\partial\vartheta}f(y_i|z_m,\vartheta) - \sum_{m=1}^k \pi_m \frac{f(y_i|z_l,\vartheta)}{f_k(y_i|\vartheta)}\frac{\partial}{\partial\vartheta}f(y_i|z_m,\vartheta)\right) \\
&= w_{il}\left(\frac{\frac{\partial}{\partial\vartheta}f(y_i|z_m,\vartheta)}{f(y_i|z_l,\vartheta)} - \sum_{m=1}^k w_{im}\frac{\frac{\partial}{\partial\vartheta}f(y_i|z_m,\vartheta)}{f(y_i|z_l,\vartheta)}\right) \\
&= w_{il}\left(s_{il} - \sum_{m=1}^k w_{im}s_{im}\right).
\end{aligned}$$

Im Falle von Varianzkomponentenmodellen, muss die Clusterung beachtet werden. Also gilt $f(y_i|z_l,\vartheta) = \prod_{j=1}^{n_i} \exp\left(\frac{\vartheta y_{ij}-b(\vartheta)}{a(\phi)} + c(y_{ij},\phi)\right)$. Die Ableitung von w_{il} nach ϑ findet man in diesem Fall jedoch analog zur obigen Rechnung. Die Ergebnisse sind dabei ident.
Als nächstes wird nachgewiesen, dass

$$\mathbb{E}_k\left[w_{il}s_{il}s_{il}^\top\right] = -\mathbb{E}_k\left[w_{il}\frac{\partial s_{il}}{\partial\vartheta}\right].$$

Wir berechnen zunächst den Erwartungswert auf der linken Seite.

$$\begin{aligned}
\mathbb{E}_k\left[w_{il}s_{il}s_{il}^\top\right] &= \int w_{il}s_{il}s_{il}^\top f_k(y_i|\vartheta)\,dy_i \\
&= \int \frac{f(y_i|z_l,\vartheta)\pi_l}{f_k(y_i|\vartheta)}\left(\frac{\partial}{\partial\vartheta}\log f(y_i|z_l,\vartheta)\right)^2 f_k(y_i|\vartheta)\,dy_i \\
&= \int \frac{f(y_i|z_l,\vartheta)\pi_l}{f(y_i|z_l,\vartheta)^2}\left(\frac{\partial}{\partial\vartheta}f(y_i|z_l,\vartheta)\right)^2 dy_i \\
&= \int \pi_l s_{il}\frac{\partial}{\partial\vartheta}f(y_i|z_l,\vartheta)\,dy_i.
\end{aligned}$$

Man beachte, dass bei den quadratischen Ausdrücken das jeweilige Vektorprodukt gemeint ist. Zum Beispiel steht $\left(\frac{\partial}{\partial\vartheta}\log f(y_i|z_l,\vartheta)\right)^2$ für $\frac{\partial}{\partial\vartheta}\log f(y_i|z_l,\vartheta)\left[\frac{\partial}{\partial\vartheta}\log f(y_i|z_l,\vartheta)\right]^\top$. Der Erwartungswert auf der rechten Seite vereinfacht sich zu

$$\begin{aligned}
-\mathbb{E}_k\left[w_{il}\frac{\partial s_{il}}{\partial\vartheta}\right] &= -\int w_{il}\frac{\partial s_{il}}{\partial\vartheta}f_k(y_i|\vartheta)\,dy_i \\
&= -\int f(y_i|z_l,\vartheta)\pi_l\frac{\partial}{\partial\vartheta}\left(\frac{\partial}{\partial\vartheta}\log f(y_i|z_l,\vartheta)\right)dy_i \\
&= -\int \pi_l\frac{f(y_i|z_l,\vartheta)f''(y_i|z_l,\vartheta) - f'(y_i|z_l,\vartheta)^2}{f(y_i|z_l,\vartheta)}dy_i \\
&= -\int \pi_l f''(y_i|z_l,\vartheta)\,dy_i + \int \pi_l s_{il}\frac{\partial}{\partial\vartheta}f(y_i|z_l,\vartheta)\,dy_i \\
&= \int \pi_l s_{il}\frac{\partial}{\partial\vartheta}f(y_i|z_l,\vartheta)\,dy_i.
\end{aligned}$$

In der obigen Notation bezeichnen die nicht näher spezifizierten Ableitungen stets Ableitungen nach ϑ.

A.5.2 *Nichtparametrische Schätzung*

Die Herleitungen in diesem Abschnitt sind Abschnitt 10.2 zuzuordnen.
Zunächst berechnen wir $\mathbb{E}_k[g_\varphi(\vartheta,\varphi)]$. Wegen $\sum_{m=1}^k \pi_m = 1$ gilt

$$
\begin{aligned}
\mathbb{E}_k\left[g_\varphi(\vartheta,\varphi)\right] &= \sum_{i=1}^{r}\sum_{l=1}^{k-1} n_i \widetilde{e}_l \int (w_{il}-\pi_l)\,dy_i \\
&= \sum_{i=1}^{r}\sum_{l=1}^{k-1} n_i \widetilde{e}_l \int \left(\frac{\pi_l f(y_i|z_l,\vartheta)}{f_k(y_i|\vartheta)} - \pi_l\right) f_k(y_i|\vartheta)\,dy_i \\
&= \sum_{i=1}^{r}\sum_{l=1}^{k-1} n_i \widetilde{e}_l \pi_l \left[\int f(y_i|z_l,\vartheta)\,dy_i - \int f_k(y_i|\vartheta)\,dy_i\right] \\
&= \sum_{i=1}^{r}\sum_{l=1}^{k-1} n_i \widetilde{e}_l \pi_l \left[1 - \sum_{m=1}^{k} \pi_m \int f(y_i|z_m,\vartheta)\,dy_i\right] \\
&= \sum_{i=1}^{r}\sum_{l=1}^{k-1} n_i \widetilde{e}_l \pi_l \left[1 - \sum_{m=1}^{k} \pi_m\right] \\
&= 0.
\end{aligned}
$$

Hier ist die Berechnung der Gewichte w_{il} nach den Parametern $\varphi_{l'}$ für $l' = 1,\ldots,k-1$ angegeben. Aus Gründen der Übersichtlichkeit geben wir zunächst die Ableitung von $f_k(y_i|\vartheta)$ an.

$$
\begin{aligned}
\frac{\partial f_k(y_i|\vartheta)}{\partial \varphi_{l'}} &= \frac{\partial}{\partial \varphi_{l'}} \sum_{m=1}^{k} f(y_i|z_m,\vartheta)\pi_m \\
&= \frac{\partial}{\partial \varphi_{l'}} \left[f(y_i|z_k,\vartheta)e^{-\kappa(\varphi)} + \sum_{m=1}^{k-1} f(y_i|z_m,\vartheta)e^{\varphi_m-\kappa(\varphi)}\right] \\
&= \left(-\frac{\partial \kappa(\varphi)}{\partial \varphi_{l'}}\right)\left[f(y_i|z_k,\vartheta)e^{-\kappa(\varphi)}\right. \\
&\quad \left. + \sum_{m=1}^{k-1} f(y_i|z_m,\vartheta)e^{\varphi_m-\kappa(\varphi)}\right] + f(y_i|z_{l'},\vartheta)e^{\varphi_{l'}-\kappa(\varphi)} \\
&= \pi_{l'} f(y_i|z_{l'},\vartheta) - \pi_{l'} \sum_{m=1}^{k} f(y_i|z_m,\vartheta)\pi_m \\
&= \pi_{l'}\left[f(y_i|z_{l'},\vartheta) - f_k(y_i|\vartheta)\right].
\end{aligned}
$$

Für $l' \neq l$ gilt damit

$$
\begin{aligned}
\frac{\partial w_{il}}{\partial \varphi_{l'}} &= \frac{\partial}{\partial \varphi_{l'}} \frac{\pi_l f(y_i|z_l,\vartheta)}{f_k(y_i|\vartheta)} \\
&= \frac{-\pi_l f(y_i|z_l,\vartheta)\pi_{l'} f(y_i|z_{l'},\vartheta)}{f_k(y_i|\vartheta)^2} \\
&= -w_{il}w_{il'}
\end{aligned}
$$

und für $l' = l$

$$\begin{aligned}\frac{\partial w_{il}}{\partial \varphi_l} &= \frac{\partial}{\partial \varphi_l}\frac{\pi_l f(y_i|z_l,\vartheta)}{f_k(y_i|\vartheta)} \\ &= \frac{\pi_l(1-\pi_l)f(y_i|z_l,\vartheta)f_k(y_i|\vartheta) - \pi_l^2 f(y_i|z_l,\vartheta)[f(y_i|z_l,\vartheta) - f_k(y_i|\vartheta)]}{f_k(y_i|\vartheta^2)} \\ &= \frac{\pi_l f(y_i|z_l,\vartheta)[f_k(y_i|\vartheta) - \pi_l f(y_i|z_l,\vartheta)]}{f_k(y_i|\vartheta)^2} \\ &= w_{il} - w_{il}^2.\end{aligned}$$

Die obigen Resultate verwenden wir nun, um die folgenden Ableitungen zu erhalten. Für ein $m \in \{1,\dots,k-1\}$ gilt

$$\begin{aligned}\frac{\partial g_\vartheta(\vartheta,\varphi)}{\partial \varphi_m} &= \sum_{i=1}^{r}\sum_{l=1}^{k}\left(\frac{\partial w_{il}}{\partial \varphi_m}\right)s_{il} \\ &= \sum_{i=1}^{r} w_{im}s_{im} - \sum_{i=1}^{r}\sum_{l=1}^{k} w_{il}w_{im}s_{il}.\end{aligned}$$

Insgesamt erhalten wir also

$$\frac{\partial g_\vartheta(\vartheta,\varphi)}{\partial \varphi^\top} = \sum_{i=1}^{r}\sum_{m=1}^{k-1} w_{im}s_{im}\widetilde{e}_m^\top - \sum_{i=1}^{r}\sum_{l=1}^{k}\sum_{m=1}^{k-1} w_{il}w_{im}s_{il}\widetilde{e}_m^\top.$$

Wir betrachten wieder zunächst die Ableitung nach φ_m für ein $m \in \{1,\dots,k-1\}$:

$$\begin{aligned}\frac{\partial g_\varphi(\vartheta,\varphi)}{\partial \varphi_m} &= \sum_{i=1}^{r}\sum_{l=1}^{k-1} n_i\widetilde{e}_l\frac{\partial}{\partial \varphi_m}(w_{il}-\pi_l) \\ &= \sum_{i=1}^{r}\sum_{l=1}^{k-1} n_i\widetilde{e}_l\left[w_{il}\delta_{l=m} - w_{il}w_{im} - \pi_l(\delta_{l=m}-\pi_m)\right] \\ &= \sum_{i=1}^{r} n_i\widetilde{e}_m(w_{im}-\pi_m) - \sum_{i=1}^{r}\sum_{l=1}^{k-1} n_i\widetilde{e}_l(w_{il}w_{im}-\pi_l\pi_m).\end{aligned}$$

Für die Ableitung nach dem gesamten Vektor φ ergibt sich damit

$$\begin{aligned}\frac{\partial g_\varphi(\vartheta,\varphi)}{\partial \varphi} &= \sum_{i=1}^{r}\sum_{m=1}^{k-1} n_i\widetilde{e}_m\widetilde{e}_m^\top(w_{im}-\pi_m) \\ &\quad - \sum_{i=1}^{r}\sum_{l=1}^{k-1}\sum_{m=1}^{k-1} n_i\widetilde{e}_l\widetilde{e}_m^\top(w_{il}w_{im}-\pi_l\pi_m).\end{aligned}$$

Als Hilfsmittel benötigen wir die folgenden beiden Resultate:

$$\mathbb{E}_k[w_{il}] = \int w_{il}f_k(y_i|\vartheta)\,dy_i = \pi_l\int f(y_i|z_l,\vartheta)\,dy_i = \pi_l$$

und

$$\begin{aligned}\mathbb{E}_k[w_{il}s_{il}] &= \pi_l \int \frac{f(y_i|z_l,\vartheta)}{f_k(y_i|\vartheta)} \frac{\partial \log f(y_i|z_l,\vartheta)}{\partial \vartheta} f_k(y_i|\vartheta)\,dy_i \\ &= \pi_l \frac{\partial}{\partial \vartheta} \int f(y_i|zl,\vartheta)\,dy_i \\ &= 0.\end{aligned}$$

A.6 STANDARDFEHLER IM NEGATIV BINOMIALMODELL

Die folgenden Ergänzungen beziehen sich auf Kapitel 11. Zunächst berechnen wir die Ableitung der a posteriori Wahrscheinlichkeiten w_{il}. Allgemein gilt

$$\begin{aligned}w_{il} &= \frac{\pi_l \prod_{j=1}^{n_i} f(y_{ij}|z_l,\vartheta)}{f_k(y_i|\vartheta)} \\ &= \frac{\pi_l \prod_{j=1}^{n_i} \frac{\Gamma(y_{ij}+\alpha)}{\Gamma(\alpha)y_{ij}!} \left(\frac{\mu_{ijl}}{\mu_{ijl}+\alpha}\right)^{y_{ij}} \left(\frac{\alpha}{\mu_{ijl}+\alpha}\right)^{\alpha}}{\sum_{m=1}^{k} \pi_m \prod_{j=1}^{n_i} \frac{\Gamma(y_{ij}+\alpha)}{\Gamma(\alpha)y_{ij}!} \left(\frac{\mu_{ijm}}{\mu_{ijm}+\alpha}\right)^{y_{ij}} \left(\frac{\alpha}{\mu_{ijm}+\alpha}\right)^{\alpha}} \\ &= \frac{\pi_l \prod_{j=1}^{n_i} \left(\frac{\mu_{ijl}}{\mu_{ijl}+\alpha}\right)^{y_{ij}} \left(\frac{\alpha}{\mu_{ijl}+\alpha}\right)^{\alpha}}{\sum_{m=1}^{k} \pi_m \prod_{j=1}^{n_i} \left(\frac{\mu_{ijm}}{\mu_{ijm}+\alpha}\right)^{y_{ij}} \left(\frac{\alpha}{\mu_{ijm}+\alpha}\right)^{\alpha}}.\end{aligned}$$

Es bezeichne nun

$$\Sigma := \sum_{m=1}^{k} \pi_m \prod_{j=1}^{n_i} \left(\frac{\mu_{ijm}}{\mu_{ijm}+\alpha}\right)^{y_{ij}} \left(\frac{\alpha}{\mu_{ijm}+\alpha}\right)^{\alpha} = f_k(y_i|\vartheta) \prod_{j=1}^{n_i} \frac{\Gamma(\alpha)y_{ij}!}{\Gamma(y_{ij}+\alpha)}.$$

Außerdem schreiben wir

$$\exp_l := \prod_{j=1}^{n_i} \left(\frac{\mu_{ijl}}{\mu_{ijl}+\alpha}\right)^{y_{ij}} \left(\frac{\alpha}{\mu_{ijl}+\alpha}\right)^{\alpha}.$$

Wegen $\exp_l = e^{\log \exp_l}$ gilt

$$\exp_l = \exp\left[\sum_{j=1}^{n_i} y_{ij} \log \mu_{ijl} - y_{ij} \log(\mu_{ijl}+\alpha) + \alpha \log \alpha - \alpha \log(\mu_{ijl}+\alpha)\right].$$

Die Ableitung nach α ist dann weiter

$$\begin{aligned}
\frac{\partial w_{il}}{\partial \alpha} &= \frac{\partial}{\partial \alpha} \frac{\pi_l \exp_l}{\sum_{m=1}^{k} \pi_m \exp_m} \\
&= \Sigma^{-1} \pi_l \exp_l \cdot \left\{ \sum_{j=1}^{n_i} \left[\frac{\mu_{ijl} - y_{ij}}{\mu_{ijl} + \alpha} + \log\left(\frac{\alpha}{\mu_{ijl} + \alpha} \right) \right] \right. \\
&\quad \left. - \Sigma^{-1} \sum_{m=1}^{k} \pi_m \exp_m \sum_{j=1}^{n_i} \left[\frac{\mu_{ijm} - y_{ij}}{\mu_{ijm} + \alpha} + \log\left(\frac{\alpha}{\mu_{ijm} + \alpha} \right) \right] \right\} \\
&= \frac{\pi_l \left(\prod_{j=1}^{n_i} \frac{\Gamma(\alpha) y_{ij}!}{\Gamma(y_{ij} + \alpha)} \right) f(y_i | z_l, \vartheta)}{\left(\prod_{j=1}^{n_i} \frac{\Gamma(\alpha) y_{ij}!}{\Gamma(y_{ij} + \alpha)} \right) f_k(y_i | \vartheta)} \\
&\quad \cdot \left\{ c_{il}^{\alpha} - \frac{\sum_{m=1}^{k} \pi_m \left(\prod_{j=1}^{n_i} \frac{\Gamma(\alpha) y_{ij}!}{\Gamma(y_{ij} + \alpha)} \right) f(y_i | z_m, \vartheta) c_{im}^{\alpha}}{\left(\prod_{j=1}^{n_i} \frac{\Gamma(\alpha) y_{ij}!}{\Gamma(y_{ij} + \alpha)} \right) f_k(y_i | \vartheta)} \right\} \\
&= w_{il} \left\{ c_{il}^{\alpha} - \frac{\sum_{m=1}^{k} \pi_m f(y_i | z_m, \vartheta) c_{im}^{\alpha}}{f_k(y_i | \vartheta)} \right\} \\
&= w_{il} \left\{ c_{il}^{\alpha} - \sum_{m=1}^{k} w_{im} c_{im}^{\alpha} \right\},
\end{aligned}$$

mit $c_{il}^{\alpha} = \sum_{j=1}^{n_i} \left[\frac{\mu_{ijl} - y_{ij}}{\mu_{ijl} + \alpha} + \log\left(\frac{\alpha}{\mu_{ijl} + \alpha} \right) \right]$. Für die Ableitung nach β stellen wir zunächst fest, dass

$$\frac{\partial \mu_{ijl}}{\partial \beta} = \frac{\partial}{\partial \beta} \exp\left(x_{ij}^{\top} \beta + u_{ij}^{\top} z_l \right) = x_{ij}^{\top} \mu_{ijl}.$$

Damit gilt

$$\begin{aligned}
\frac{\partial w_{il}}{\partial \beta} &= \frac{\pi_l \exp_l}{\sum_{m=1}^{k} \pi_m \exp_m} \\
&= \Sigma^{-1} \pi_l \exp_l \left\{ \sum_{j=1}^{n_i} \frac{\partial \mu_{ijl}}{\partial \beta} \left(\frac{y_{ij}}{\mu_{ijl}} - \frac{y_{ij}+\alpha}{\mu_{ijl}+\alpha} \right) \right. \\
&\quad \left. - \Sigma^{-1} \sum_{m=1}^{k} \pi_m \exp_m \sum_{j=1}^{n_i} \frac{\partial \mu_{ijm}}{\partial \beta} \left(\frac{y_{ij}}{\mu_{ijm}} - \frac{y_{ij}+\alpha}{\mu_{ijm}+\alpha} \right) \right\} \\
&= w_{il} \left\{ \sum_{j=1}^{n_i} x_{ij}^{\top} \left(y_{ij} - \frac{\mu_{ijl}(y_{ij}+\alpha)}{\mu_{ijl}+\alpha} \right) \right. \\
&\quad \left. - \frac{\sum_{m=1}^{k} \pi_m f(y_i|z_m,\vartheta) \sum_{j=w}^{n_i} x_{ij}^{\top} \left(y_{ij} - \frac{\mu_{ijm}(y_{ij}+\alpha)}{\mu_{ijm}+\alpha} \right)}{f_k(y_i|\vartheta)} \right\} \\
&= w_{il} \left\{ c_{il}^{\beta} - \frac{\sum_{m=1}^{k} \pi_m f(y_i|z_m,\vartheta) c_{im}^{\beta}}{f_k(y_i|\vartheta)} \right\} \\
&= w_{il} \left\{ c_{il}^{\beta} - \sum_{m=1}^{k} w_{im} c_{im}^{\beta} \right\}
\end{aligned}$$

mit $c_{il}^{\beta} = \sum_{j=1}^{n_i} x_{ij}^{\top} \left(y_{ij} - \frac{\mu_{ijl}(y_{ij}+\alpha)}{\mu_{ijl}+\alpha} \right)$.

A.7 DIE LOUIS FORMEL

Hier geben wir Ergänzgunen zu Kapitel 13 an. Für unsere erste Anwendung in Abschnitt 13.1.1 - namentlich die Herleitung der Standardfehler über die Louis Formel im Poissonmodell mit normalverteilten zufälligen Effekten - benötigen wir die vollständige Scorefunktion. Diese erhält man über

$$\begin{aligned}
S_c(y, Z|\vartheta) &= \frac{f'(y, Z|\vartheta)}{f(y, Z|\vartheta)} = \frac{f'(y|Z,\vartheta)\varphi(Z)}{f(y|Z,\vartheta)\varphi(Z)} \\
&= \frac{f'(y|Z,\vartheta)}{f(y|Z,\vartheta)} = \frac{\partial}{\partial \vartheta} \sum_{i=1}^{r} \log f(y_i|Z,\vartheta) \\
&= \sum_{i=1}^{r} \frac{\frac{\partial}{\partial \vartheta} \frac{\mu_i^{y_i}}{y_i!} e^{-\mu_i}}{\frac{\mu_i^{y_i}}{y_i!} e^{-\mu_i}} = \sum_{i=1}^{r} \frac{\frac{1}{y_i!} e^{-\mu_i} \frac{\partial \mu_i}{\partial \vartheta} \mu_i^{y_i - 1} (y_i - \mu_i)}{\frac{\mu_i^{y_i}}{y_i!} e^{-\mu_i}} \\
&= \sum_{i=1}^{r} \frac{\frac{\partial \mu_i}{\partial \vartheta} (y_i - \mu_i)}{\mu_i} = \sum_{i=1}^{r} \frac{\tilde{x}_i^{\top} \mu_i (y_i - \mu_i)}{\mu_i} \\
&= \sum_{i=1}^{r} \tilde{x}_i^{\top} (y_i - \mu_i),
\end{aligned}$$

wobei $\varphi(\cdot)$ die Dichtefunktion der Standardnormalverteilung ist.

Im Negativ Binomialmodell benötigen wir in Abschnitt 13.1.2.2 den Ausdruck $\frac{\partial^2 f(y_i|Z_i,\theta)/\partial\alpha\partial\vartheta^\top}{f(y_i|Z_i,\theta)}$. Dazu berechnen wir zunächst nur die gemischte Ableitung $\frac{\partial^2 f(y_i|Z_i,\theta)}{\partial\alpha\partial\vartheta^\top}$:

$$\begin{aligned}
&\frac{\partial^2}{\partial\alpha\partial\vartheta^\top}\frac{\Gamma(y_i+\alpha)}{\Gamma(\alpha)y_i!}\left(\frac{\mu_i}{\mu_i+\alpha}\right)^{y_i}\left(\frac{\alpha}{\mu_i+\alpha}\right)^{\alpha}\\
&=\frac{\partial}{\partial\alpha}\frac{\Gamma(y_i+\alpha)}{\Gamma(\alpha)y_i!}\left(\frac{\mu_i}{\mu_i+\alpha}\right)^{y_i-1}\left(\frac{\alpha}{\mu_i+\alpha}\right)^{\alpha-1}\frac{\alpha\mu_i\tilde{x}_i}{(\mu_i+\alpha)^2}\frac{\alpha(y_i-\mu_i)}{\mu_i+\alpha}\\
&=\frac{\partial}{\partial\alpha}\frac{\Gamma(y_i+\alpha)}{\Gamma(\alpha)y_i!}\left(\frac{\mu_i}{\mu_i+\alpha}\right)^{y_i}\left(\frac{\alpha}{\mu_i+\alpha}\right)^{\alpha}\frac{\alpha(y_i-\mu_i)\tilde{x}_i}{\mu_i+\alpha}\\
&=f(y_i|z_i,\theta)\frac{(y_i-\mu_i)\tilde{x}_i}{\mu_i+\alpha}\left[\frac{\alpha\psi(y_i+\alpha)}{\Gamma(y_i+\alpha)^2}-\alpha\psi(\alpha)+\frac{\mu_i+\alpha\mu_i-y_i\alpha}{\mu_i+\alpha}\right]\\
&=f(y_i|Z_i,\theta)R_i\quad\text{mit}
\end{aligned}$$

$$R_i=\frac{(y_i-\mu_i)\tilde{x}_i}{\mu_i+\alpha}\left[\frac{\alpha\psi(y_i+\alpha)}{\Gamma(y_i+\alpha)^2}-\alpha\psi(\alpha)+\frac{\mu_i+\alpha\mu_i-y_i\alpha}{\mu_i+\alpha}\right].$$

Schätzt man die Verteilung der Zufallseffekt nichtparametrisch, muss die fehlende Information bezüglich π berechnet werden (vgl. Abschnitt 13.2.1). Dazu halten wir zunächst fest, dass wegen der Unabhängigkeit

$$\begin{aligned}
\mathbb{E}\left[S_c^\pi S_c^\top\middle|y\right]=&\sum_{i=1}^{r}\sum_{l=1}^{k-1}\sum_{m=1}^{k-1}e_l\mathbb{E}\left[\left(\frac{e_{il}}{\pi_l}-\frac{e_{ik}}{\pi_k}\right)\left(\frac{e_{im}}{\pi_m}-\frac{e_{ik}}{\pi_k}\right)\middle|y_i\right]e_m^\top\\
&+\sum_{i=1}^{r}\sum_{\substack{d=1\\d\neq i}}^{r}\sum_{l=1}^{k-1}\sum_{m=1}^{k-1}e_l\mathbb{E}\left[\left(\frac{e_{il}}{\pi_l}-\frac{e_{ik}}{\pi_k}\right)\middle|y_i\right]\\
&\cdot\mathbb{E}\left[\left(\frac{e_{dm}}{\pi_m}-\frac{e_{dk}}{\pi_k}\right)\middle|y_d\right]e_m^\top
\end{aligned}$$

und

$$\begin{aligned}
\mathbb{E}\left[S_c^\pi|y\right]\mathbb{E}\left[S_c^\top|y\right]=&\left(\sum_{i=1}^{r}\sum_{l=1}^{k-1}e_l\mathbb{E}\left[\frac{e_{il}}{\pi_l}-\frac{e_{ik}}{\pi_k}\middle|y_i\right]\right)\\
&\cdot\left(\sum_{d=1}^{r}\sum_{m=1}^{k-1}\mathbb{E}\left[\frac{e_{dm}}{\pi_m}-\frac{e_{dk}}{\pi_k}\middle|y_d\right]e_m^\top\right).
\end{aligned}$$

Damit erhalten wir

$$\begin{aligned}
\mathcal{I}_m^\pi(y|\pi)=&\sum_{i=1}^{r}\sum_{l=1}^{k-1}\sum_{m=1}^{k-1}e_l\mathbb{E}\left[\left(\frac{e_{il}}{\pi_l}-\frac{e_{ik}}{\pi_k}\right)\left(\frac{e_{im}}{\pi_m}-\frac{e_{ik}}{\pi_k}\right)\middle|y_i\right]e_m^\top\\
&-\sum_{i=1}^{r}\sum_{l=1}^{k-1}\sum_{m=1}^{k-1}e_l\mathbb{E}\left[\left(\frac{e_{il}}{\pi_l}-\frac{e_{ik}}{\pi_k}\right)\middle|y_i\right]\\
&\cdot\mathbb{E}\left[\left(\frac{e_{im}}{\pi_m}-\frac{e_{ik}}{\pi_k}\right)\middle|y_i\right]e_m^\top.
\end{aligned}$$

B IMPLEMENTIERUNGEN

B.1 IRLS-METHODE

Code 4: IRLS

```
IRLS = function(y,x,weights,offset,beta0,alpha0,tol,max_iter){

  require(MASS)

  n <- length(y)
  p <- dim(x)[2]
  alpha <- alpha0
  beta <- beta0
  iter <- 1
  cond <- TRUE
  alpha.dev <- matrix(,nrow=1, ncol=max_iter)
  alpha.dev[1] <- alpha
  beta.dev <- matrix(, nrow=p, ncol=max_iter)
  beta.dev[,1] <- beta

  while(cond){
    mu <- exp(x%*%beta + offset)
    w <- as.numeric(mu/(1+mu/alpha))
    W <- diag(w)
    z <- log(mu) + (y-mu)/mu*weights

    beta.old <- beta
    mat <- t(x)%*% W %*% x

    e <- sqrt(.Machine$double.eps)
    Xsvd <- svd(mat)
    Positive <- Xsvd$d > max(e * Xsvd$d[1L], 0)
    if(!all(Positive)) warning(gettextf("Matrix (almost) singular
         in iteration %d! Moore Penrose Inverse is used.", iter))

    beta <- ginv(mat)%*%t(x)%*%W%*%z

    alpha.old <- alpha

    psi1 <- digamma(alpha)
    psi2 <- digamma(y+alpha)
    psi3 <- trigamma(alpha)
    psi4 <- trigamma(y+alpha)

    la <- sum(weights*(log(alpha)+1-psi1+psi2-log(mu+alpha)-(y+
        alpha)/(mu+alpha)))
```

```
40      laa <- sum(weights*(1/alpha-psi3+psi4-1/(mu+alpha)-(mu-y)/(mu
            +alpha)^2))
41
42      alpha <- alpha.old-la/laa
43
44      iter <- iter+1
45      alpha.dev[iter] <- alpha
46      beta.dev[,iter] <- beta
47
48      old <- rbind(alpha.old, beta.old)
49      new <- rbind(alpha, beta)
50      delta <- norm(old-new)
51
52      cond <- (iter<max_iter) && (delta > tol)
53    }
54
55  fit <- list(alpha=alpha, beta=beta, iter=iter, delta=delta, alpha
        .dev=alpha.dev, beta.dev=beta.dev)
56
57  }
```

B.2 ZUFALLSZAHLENGENERATOR

Für die Durchführung der Monte Carlo Studie benötigen wir einen Zufallszahlengenerator, der einen Responsevektor, der vorgegebenen Modellspezifikationen genügt, erzeugen kann. Dafür steht die Funktion `rand.y()` zur Verfügung, welche in Abschnitt 14.1 vorgestellt wird.

Code 5: `rand.gq()`

```
1  rand.y <- function(X, distribution, random.distribution='normal',
       sigma, beta, alpha, r, seed){
2    require(npmlreg)
3    Call <- match.call(expand.dots=TRUE)
4
5    if(!missing(seed)){set.seed(seed)}
6    n <- dim(X)[1]
7    if(missing(r)){r <- n}
8    size <- n/r
9
10   if(r == n){type <- "overdispersion"}
11   if((size-floor(size)==0)&(r<n)){type <- "variance component"}
12   if(r>n){stop("number of clusters must be smaller than sample
         size")}
13   if(size-floor(size)!=0){stop("only clusters of the same size
         are supported")}
14
15   if(!(distribution %in% c("poisson", "negbin"))){stop("only '
         poisson' and 'negbin' supported")}
16   if(distribution=="negbin"&&missing(alpha)){stop("alpha must be
         specified for 'negbin'")}
```

```
17
18   if(!(random.distribution %in% c("normal", "t", "uniform", "
         logistic", "lognorm"))){stop("random.distribution not
         supported")}
19   if(random.distribution=="t"&sigma<1)(stop("sigma must be
         greater than 1 in case of Student's t-distribution"))
20
21   p <- length(beta)
22   d <- dim(X)[2]
23
24   if(d==p){Intercept=FALSE}
25   if(d==p-1){Intercept=TRUE}
26   if(!(d %in% c(p,p-1))){stop("beta and X do not fit")}
27
28   if(Intercept){X <- cbind(rep(1,n),X)}
29
30   if(random.distribution=="normal"){Z <- rnorm(r,0,sigma)}
31   if(random.distribution=="t"){
32     df <- 2*sigma^2/(sigma^2-1)
33     Z <- rt(r,df)
34     }
35   if(random.distribution=="uniform"){Z <- runif(r,-1,1)}
36   if(random.distribution=="logistic"){
37     scale <- pi^2/(3*sigma^2)
38     Z <- rlogis(r,location=0,scale=scale)
39   }
40   if(random.distribution=="lognorm"){
41     mu <- -sigma^2/2
42     Z <- rlnorm(r,mu,sigma)-1
43   }
44
45   Z <- matrix(expand(t(Z),size),n,1)
46
47   mu <- exp(as.matrix(X)%*%beta+Z)
48
49   if(distribution=="poisson"){y <- rpois(n,mu)}
50   if(distribution=="negbin"){y <- rnbinom(n,size=alpha,mu=mu)}
51
52   list(Call=Call, y=y, type=type, Intercept=Intercept,
         distribution=distribution, random.distribution=random.
         distribution, mu=mu, p=p)
53 }
```

C

DATEN

C.1 DATENBEISPIEL

Für unser Datenbeispiel verwednen wir die Ergebnisse einer Studie zur Bakterienbelastung. Der vollständige Datensatz ist hier nun abgedruckt. Details zu den Bezeichnungen sind in Kapitel 15 zu finden.

Code 6: `Datenbeispiel`

```
> data
     date site.all humi temp b1 b2 b3 b4 b5 b6
6    8.03        6   29   10  2  1  0  2  1  1
7    8.03        7   23   10  1  0  0  1  1  0
13  22.03        6   37    7  2  1  2  1  3  2
14  22.03        7   38    7  1  2  2  4  2  0
20   5.04        6   31   12  3  2  4  5  1  1
21   5.04        7   32   12  3  0  1  0  2  0
27  19.04        6   33   21  9  3  1  1  2  1
28  19.04        7   38   20  2  1  1  2  2  7
34   2.05        6   25   22  2  1  1  2  1  1
35   2.05        7   27   19  1  0  2  3  0  2
40  16.05        6   30   19  0  3  3  4 12  1
41  16.05        7   36   20  0  1  2  3  2  3
47   7.06        6   72   17  1  0  2  1  3  3
48   7.06        7   71   17  0  0  2  0  1  1
54  13.06        6   45   20  2  1  1  0  1  1
55  13.06        7   54   18  3  2  3  1  4  2
61  27.06        6   47   24  3  2  4  1  3  3
62  27.06        7   45   25  3  5  1  1  0  3
68  12.07        6   40   30  0  0  0  1  1  1
69  12.07        7   39   30  2  0  1  1  0  1
75  25.07        6   33   31 11  7  0  0  0  1
76  25.07        7   35   30  5  3  4  4  2  1
82   8.08        6   53   28  0  1  0  2  0  0
83   8.08        7   48   28  2  2  1  2  3  1
89  22.08        6   43   29  1  0  3  0  2  0
90  22.08        7   45   29  3  6  6  2  1  2
96   5.09        6   36   21  9  1  4  1  3  2
97   5.09        7   42   20  1  1  2  4  0  1
103 19.09        6   83   17  2  1  1  2  2  0
104 19.09        7   85   16  2  1  1  2  2  0
110  4.10        6   44   24  1  1  1  1  0  5
111  4.10        7   48   24  1  2  3  1  5  8
116 17.10        6   51   17  2  2  0  6  7  3
117 17.10        7   58   15  5  0  3  2  2  1
123 31.10        6   88   12  0  1  2  2  2  0
124 31.10        7   88   12  3  0  2  3  2  3
```

130	14.11	6	65	11	3	4	0	0	2	3
131	14.11	7	71	10	6	5	0	5	6	5
136	28.11	6	67	8	2	0	1	0	0	4
137	28.11	7	69	9	2	1	1	2	12	36
143	13.12	6	69	-1	1	1	1	1	0	0
144	13.12	7	74	-1	0	1	0	1	1	6
149	3.01	6	86	-1	3	2	1	1	0	1
150	3.01	7	85	-2	4	3	3	3	3	2
156	9.01	6	86	3	3	1	0	1	4	2
157	9.01	7	87	3	1	1	2	0	2	1
163	23.01	6	77	-5	1	0	1	4	1	0
164	23.01	7	81	-6	3	0	0	0	1	2
170	6.02	6	51	-2	0	1	1	1	1	0
171	6.02	7	56	-2	0	0	1	1	0	0
177	21.02	6	58	3	0	1	2	3	12	0
178	21.02	7	60	5	0	2	0	0	0	0

LITERATURVERZEICHNIS

Abramowitz, M. und Stegun, I. A. (1964). *Handbook of Mathematical Functions*. Dover, New York, USA.

Aitkin, M., Francis, B., Hinde, J., und Darnell, R. (2009). *Statistical Modelling in R*. Oxford University Press, Oxford, GB.

Booth, J. G., Casella, G., Friedl, H., und Hobert, J. (2003). Negative binomial loglinear mixed models. *Statistical Modelling*, 3:179–191.

Booth, J. G. und Hobert, J. P. (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society*, 61:265–285.

Efron, B. und Hinkley, D. (1978). The observed versus expected information. *Biometrika*, 65:457–487.

Einbeck, J., Darnell, R., und Hinde, J. (2012). *npmlreg: Nonparametric maximum likelihood estimation for random effect models*. R package version 0.45-1. Unter: `http://CRAN.R-project.org/package=npmlreg`.

Friedl, H. (1998). Nichtparametrische Maximum-Likelihood-Schätzung bei Generalisierten Linearen Mischmodellen. *Austrian Journal of Statistics*, 26:7–30.

Friedl, H. und Kauermann, G. (2000). Standard errors for em estimates in generalized linear modeks with random effects. *Biometrics*, 56:761–767.

Geweke, J. (1996). Monte Carlo Simulation and Numerical Integration. *Handbook of Computational Economics*, 80:125–165.

Hilbe, J. (2011). *Negative Binomial Regression*. Cambridge University Press, Cambridge, Großbritannien.

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

Little, R. J. und Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. Wiley, New York, USA.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 44:226–233.

McLachlan, G. J. und Krishnan, T. (1997). *The EM Algorithm and Extensions*. John Wiley & Sons, Inc, New York, USA.

R Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Unter: `http://www.R-project.org/`.

Rigby, R. A. und Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape,(with discussion). *Applied Statistics*, 54:507–554.

Ripley, B. D. und Venables, W. N. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition. Unter: `http://www.stats.ox.ac.uk/pub/MASS4`.

Robert, C. und Casella, G. (1999). *Monte Carlo Statistical Methods*. Springer, New York, USA.

Robert, C. und Casella, G. (2010). *Introducing Monte Carlo Methods with R*. Springer, New York, USA.

Smyth, G. (2013). *statmod: Statistical Modeling*. R package version 1.4.17. Unter: `http://CRAN.R-project.org/package=statmod`.

Wei, G. C. G. und Tanner, M. A. (1990). A monte carlo implementation of the em algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, 85:699–704.