

Carl von Ossietzky  
Universität Oldenburg  
Diplomstudiengang Mathematik



# DIPLOMARBEIT

## Modellierung und Versuchsplanung für wiederholte Zählraten

vorgelegt von:

Mouna Akacha

vorgelegt am:

02. Mai 2007

Betreuende Gutachterin:

Prof. Dr. Christine Müller

Zweiter Gutachter:

Prof. Dr. Dietmar Pfeifer

Auswärtiger Berater bei

Themenstellung und Betreuung: Dr. Norbert Benda



# Inhaltsverzeichnis

Tabellenverzeichnis . . . . .	v
Abbildungsverzeichnis . . . . .	vii
Symbolverzeichnis . . . . .	ix
<b>1 Einleitung</b>	<b>1</b>
<b>2 Grundlagen</b>	<b>5</b>
2.1 Das Modell . . . . .	5
2.1.1 Verallgemeinerte lineare Modelle . . . . .	7
2.1.2 Verteilungen . . . . .	8
2.2 Notationen . . . . .	9
2.2.1 Ableitungen . . . . .	9
2.2.2 Likelihood-Funktion, Score-Funktion, ML-Schätzer . . . . .	9
2.2.3 Konvergenzarten . . . . .	10
2.2.4 Landau-Symbole . . . . .	11
2.2.5 Matrizen . . . . .	11
<b>3 Das Grundmodell und Modellierungsmöglichkeiten</b>	<b>13</b>
3.1 Marginale Modelle . . . . .	15
3.1.1 Schätzmethoden für das marginale Modell . . . . .	17
3.1.1.1 Verallgemeinerte Schätzgleichungen (GEE1) . . . . .	18
3.1.1.2 Fisher-Scoring-Methode . . . . .	22
3.1.1.3 Schätzer für $\alpha$ und $\phi$ . . . . .	27
3.1.1.4 Eigenschaften des GEE-Schätzers . . . . .	29
3.1.1.5 Pseudo-Maximum-Likelihood-Methode und Quasi- Generalisierte-Pseudo-Maximum-Likelihood-Methode	32
3.1.1.5.1 PML-Methode . . . . .	32
3.1.1.5.2 QGPML-Methode . . . . .	49
3.1.1.5.3 PML-, QGPML-Methode und GEE . . . . .	51
3.2 Subjekt-spezifische Modelle . . . . .	55

3.2.1	Drei Möglichkeiten für die Handhabung der $U_i$ . . . . .	56
3.2.1.1	Die subjekt-spezifischen Größen als feste unbekannte Parameter . . . . .	56
3.2.1.2	Die subjekt-spezifischen Größen als Störparameter . . . . .	57
3.2.1.3	Die subjekt-spezifischen Größen als zufällige Para- meter . . . . .	57
3.2.2	Schätzmethoden für das subjekt-spezifische Modell . . . . .	59
3.2.2.1	Normalverteilte subjekt-spezifische Größen . . . . .	60
3.2.2.1.1	Hermite-Gauß-Quadratur . . . . .	63
3.2.2.1.2	Adaptive Hermite-Gauß-Quadratur . . . . .	66
<b>4</b>	<b>Ein Vergleich der Modelle</b> . . . . .	<b>73</b>
4.1	Epilepsie-Studie . . . . .	74
4.2	Marginales Modell . . . . .	75
4.3	Subjekt-spezifisches Modell . . . . .	76
4.4	Vergleich der Resultate . . . . .	77
<b>5</b>	<b>Implementierung und Beispiele</b> . . . . .	<b>81</b>
5.1	Implementierung für das marginale Modell . . . . .	81
5.2	Implementierung für das subjekt-spezifische Modell . . . . .	84
5.3	Beispiele . . . . .	89
5.3.1	Medikament-A-Studie . . . . .	90
5.3.2	Medikament-B-Studie . . . . .	91
5.3.3	Modellierung im Fall eines marginalen Modells . . . . .	92
5.3.4	Auswertung der Beispiele im Fall marginaler Modelle . . . . .	93
5.3.4.1	Auswertung der Medikament-A-Studie . . . . .	93
5.3.4.2	Auswertung der Medikament-B-Studie . . . . .	99
5.3.5	Modellierung im Fall eines subjekt-spezifischen Modells . . . . .	103
5.3.6	Auswertung mittels subjekt-spezifischer Modelle . . . . .	104
5.4	Vergleich der Ergebnisse . . . . .	108
5.4.1	Alternatives subjekt-spezifisches Modell . . . . .	109
<b>6</b>	<b>Versuchsplanung</b> . . . . .	<b>113</b>
6.1	Der Stichprobenumfang im Fall von zwei Stichproben . . . . .	120
6.2	Der Stichprobenumfang und seine Einflussgrößen . . . . .	126
<b>7</b>	<b>Simulation korrelierter Zähldaten</b> . . . . .	<b>135</b>
7.1	Simulation von Daten für das marginale Modell . . . . .	135

7.1.1	Der Algorithmus . . . . .	137
7.2	Auswertung der simulierten Studien . . . . .	146
<b>8</b>	<b>Zusammenfassung und Ausblick</b>	<b>151</b>
<b>Anhang</b>		<b>159</b>
<b>A</b>	<b>Orthogonale Polynome</b>	<b>159</b>
<b>B</b>	<b>Die Multivariate Hermite-Gauß-Quadratur</b>	<b>169</b>
<b>C</b>	<b>Die SAS - Anweisungen zum Zeichnen der Abbildungen in Kapitel 4</b>	<b>175</b>
<b>D</b>	<b>Die Endtableaus für die Analyse von Modell (5.4) und Modell (5.5)</b>	<b>181</b>
	D.1 Endtableaus für die Analyse des Modells (5.4) . . . . .	181
	D.2 Endtableaus für die Analyse des Modells (5.5) . . . . .	183
<b>E</b>	<b>Die SAS-Anweisung für die Berechnung des marginalisierten Erwartungswertes im Abschnitt 5.3.6</b>	<b>187</b>
<b>F</b>	<b>Die Endtableaus für die alternative Analyse des subjekt- spezifischen Modells in Abschnitt 5.4.1</b>	<b>189</b>
<b>G</b>	<b>Die R-Funktionen zur Verdeutlichung der Einflussgrößen des benötigten Stichprobenumfangs</b>	<b>191</b>
<b>H</b>	<b>Das R-Programm für den Algorithmus in Abschnitt 7.1.1</b>	<b>195</b>
	H.1 Detaillierte Implementierung . . . . .	195
	H.2 Das R -Programm . . . . .	197
<b>I</b>	<b>Die R-Programme zur Daten-Erzeugung</b>	<b>203</b>
	I.1 Das R -Programm für den Fall $n = 1$ . . . . .	203
	I.2 Das R -Programm für den Fall von $n > 1$ . . . . .	204
<b>J</b>	<b>Das SAS-Programm zur Auswertung der simulierten Datensätze</b>	<b>207</b>



# Tabellenverzeichnis

3.1	Die gängigsten Schätzer für $\alpha$ bei unterschiedlichen Abhängigkeitsstrukturen . . . . .	27
5.1	Die Regressionsparameter für unterschiedliche Korrelationsstrukturen	99
5.2	Schätzer und p-Werte für die Analyse von Modell (5.4) . . . . .	101
5.3	Schätzer und p-Werte für die Analyse von Modell (5.5). . . . .	102
5.4	Die signifikanten multiplikativen Behandlungseffekte für die Analyse des alternativen Modells (5.5) . . . . .	102
5.5	Ermittelte Schätzer für das Modell in Gleichung (5.6) mit unterschiedlichen Anzahlen an Quadraturpunkten $Q$ . . . . .	107
7.1	Ergebnisse der simulierten Studien für $n = 1$ . . . . .	147
7.2	Ergebnisse der simulierten Studien für $n > 1$ . . . . .	149



# Abbildungsverzeichnis

3.1	Gauß-Quadratur versus adaptive Gauß-Quadratur . . . . .	67
4.1	Marginale Entwicklung versus Marginalisierte Entwicklung . . . . .	78
4.2	Subjekt-spezifische und marginalisierte Entwicklungen . . . . .	79
5.1	Entwicklung der Hitzewallungsanzahl im Design 3 . . . . .	103
6.1	Wiederholungsanzahl $n$ versus Stichprobengröße $m$ für $\beta_1 = \ln(2)$ . .	127
6.2	Wiederholungsanzahl $n$ versus Stichprobengröße $m$ für $\beta_1 = \ln(1.1)$ .	128
6.3	Korrelation versus Stichprobenumfang $m$ . . . . .	130
6.4	Überdispersion versus Stichprobenumfang $m$ . . . . .	131
6.5	Behandlungsunterschied versus Stichprobenumfang $m$ . . . . .	132
6.6	Der Regressionsparameter $\beta_0$ versus Stichprobenumfang $m$ . . . . .	133



# Symbolverzeichnis

$\{\dots\}$	Menge
$\mathbb{N}, \mathbb{N}_0$	Menge der natürlichen Zahlen $\{1, 2, 3, \dots\}$ bzw. $\mathbb{N} \cup \{0\}$
$\mathbb{Z}$	Menge der ganzen Zahlen $\{-n, n   n \in \mathbb{N}\} \cup \{0\}$
$\mathbb{R}, \mathbb{R}^+$	Menge der reellen Zahlen bzw. $\{x \in \mathbb{R}   x > 0\}$
$\mathbb{R}^k$	reeller Standardraum $\{x = (x_1, \dots, x_k)   x_i \in \mathbb{R}, i \in \{1, \dots, k\}\}$ der Dimension $k \in \mathbb{N}$
$(a, b)$	für $a < b$ offenes Intervall $I = \{x \in \mathbb{R}   a < x < b\}$ , sonst $I := \emptyset$
$[a, b]$	für $a \leq b$ abgeschlossenes Intervall $I = \{x \in \mathbb{R}   a \leq x \leq b\}$ , sonst $I := \emptyset$
$(a, b]$	für $a \leq b$ links offenes, rechts abgeschlossenes Intervall $I = \{x \in \mathbb{R}   a < x \leq b\}$ , sonst $I := \emptyset$
$[a, b)$	für $a \leq b$ links abgeschlossenes, rechts offenes Intervall $I = \{x \in \mathbb{R}   a \leq x < b\}$ , sonst $I := \emptyset$
$\forall$	für alle
$\exists$	es existiert
$\partial$	partieller Ableitungsoperator
$\lceil x \rceil$	Gauß-Klammern mit $\lceil x \rceil := \min\{z \in \mathbb{Z}   x \leq z\}$ für $x \in \mathbb{R}$
$\mathbb{1}_A(x)$	Indikatorfunktion ( $= 1$ , falls $x \in A$ , $= 0$ , sonst)
$\exp(\cdot)$	Exponentialfunktion
$\ln(\cdot)$	natürliche Logarithmusfunktion
$\ \cdot\ $	euklidische Norm
$\langle \cdot, \cdot \rangle_\phi$	mit der Funktion $\phi$ gewichtetes Skalarprodukt
$m$	Anzahl der Subjekte
$n_i$	Anzahl der wiederholten Beobachtungen zu Subjekt $i$ , $i \in \{1, \dots, m\}$

$Y_{ij}$	Zufallsvariable, $i \in \{1, \dots, m\}$ und $j \in \{1, \dots, n_i\}$
$y_{ij}$	Realisierung der Zufallsvariable $Y_{ij}$ , $i \in \{1, \dots, m\}$ und $j \in \{1, \dots, n_i\}$
$Y_i$	$n_i$ -dimensionaler Zufallsvektor $(Y_{i1}, \dots, Y_{in_i})^\top$ , $i \in \{1, \dots, m\}$
$y_i$	Realisierung des Zufallsvektors $Y_i$ , $i \in \{1, \dots, m\}$
$Y$	$(n_1 + \dots + n_m)$ -dimensionaler Zufallsvektor $(Y_1^\top, \dots, Y_m^\top)^\top$
$y$	Realisierung des Zufallsvektors $Y$
$x_{ij}$	erklärender $q$ -dimensionaler Variablenvektor
$X_i$	$(n_i \times q)$ -dimensionale Prädiktormatrix von Subjekt $i$ , $i \in \{1, \dots, m\}$
$U_i$	$p$ -dimensionale subjekt-spezifische Größe, $i \in \{1, \dots, m\}$
$u_i$	Realisierung der Größe $U_i$ , im Fall zufälliger subjekt-spezifischer Größen
$U$	$pm$ -dimensionaler Vektor $(U_1^\top, \dots, U_m^\top)^\top$
$R_i$	Korrelationsmatrix des Zufallsvektors $Y_i$ , $i \in \{1, \dots, m\}$
$\Theta$	Parameterraum, $\subset \mathbb{R}^q$
$\beta$	$q$ -dimensionaler Regressionsvektor, $\in \Theta$
$\hat{\beta}$	Schätzer für $\beta$ , $\in \mathbb{R}^q$
$\beta^*$	wahrer Regressionsvektor im Rahmen der PML-Methode, $\in \Theta$
$\tilde{\beta}$	Wahrscheinlichkeit für den Fehler 2. Art
$\alpha$	Parameter der Abhängigkeitsstruktur, $\in \mathbb{R}$
$\alpha^*$	Wahrscheinlichkeit für den Fehler 1. Art (Signifikanzniveau)
$\phi$	Dispersions-Parameter, $\in \mathbb{R}$
$d$	kleinste bedeutungsvolle Differenz, $\in \mathbb{R}$
$U_\beta$	Umgebung von $\beta$ , $\subset \mathbb{R}^q$
$0_{1 \times q}$	$q$ -dimensionaler Zeilenvektor $(0, \dots, 0)$
$1_{1 \times q}$	$q$ -dimensionaler Zeilenvektor $(1, \dots, 1)$
$0_q$	$q$ -dimensionaler Spaltenvektor $(0, \dots, 0)^\top$
$1_q$	$q$ -dimensionaler Spaltenvektor $(1, \dots, 1)^\top$
$I_n$	$(n \times n)$ -dimensionale Einheitsmatrix
$P(\cdot)$	Wahrscheinlichkeitsverteilung
$E(\cdot)$	Erwartungswert
$\text{Var}(\cdot)$	Varianz
$\text{Corr}(\cdot)$	Korrelation

$\text{Cov}(\cdot)$	Kovarianz
$F_Z(z)$	Verteilungsfunktion von $Z$
$f_Z(z)$	Dichtefunktion von $Z$
$f_{Z,W}(z, w)$	gemeinsame Dichtefunktion von $Z$ und $W$
$f_{Z W=w}(z)$	Dichtefunktion von $Z$ bedingt unter $W = w$
$f^a(\cdot)$	angenommene Dichtefunktion im Rahmen der PML-Methode
$f^w(\cdot)$	wahre Dichtefunktion im Rahmen der PML-Methode
$E_f(\cdot)$	Erwartungswert von $Z$ bezüglich der Dichte $f$
$h(\cdot)$	Linkfunktion
$S(\beta)$	Score-Funktion
$L_y(\beta)$	Maximum-Likelihood-Funktion von $Y$
$\ell_y(\beta)$	Loglikelihood-Funktion von $Y$



# Kapitel 1

## Einleitung

Die vorliegende Arbeit thematisiert die Modellierung und Versuchsplanung wiederholter Zähl­daten.

Von wiederholten Daten wird dann gesprochen, wenn bei einem Subjekt (z.B. bei einem Patienten oder einem Krankenhaus) oder einem Cluster (z.B. einer Familie oder einer Stadt) das gleiche Merkmal mehrfach beobachtet wird.

Diese Datenart gewinnt in verschiedenen Bereichen – wie etwa in der Wirtschaft, Meteorologie, Geographie, Medizin oder Politik – zunehmend an Bedeutung. Für das Vorkommen und die Relevanz wiederholter Daten können daher zahlreiche Beispiele angeführt werden. So liefern Studien, die bei Familien die Verbreitung von vererbba­ren Krankheiten untersuchen, wiederholte Daten. Auch statistische Erhebungen der Überlebenschance von Kleinunternehmen oder der Migrationsstruktur von Städten erzeugen wiederholte Daten. Beispiele für wiederholte Zähl­daten sind die Anzahl an Panikattacken, die über einen bestimmten Zeitraum nach einer Behandlung gegen die Beschwerden einer psychologischen Krankheit erhoben wurden, die Anzahl an Geschlechtspartnern während einer Erhebung im Laufe eines HIV-Präventions­Programms oder die Häufigkeit von Kindstod vor und nach der Einführung eines Mutterschaftsvorsorge-Programms.<sup>1</sup>

Zwei statistische Studien erfahren in dieser Arbeit eine besondere Betrachtung: Sie liefern wiederholte Zähl­daten als Anzahl an Hitzewallungen bei Frauen in den Wechseljahren bzw. als Anzahl an Epilepsieanfällen bei Epilepsie-Patienten in einem vor­gegebenen Zeitraum.

---

<sup>1</sup>Vgl. Diggle u. a. (1996), S. 160.

Klassische Annahmen der Statistik – dies sind insbesondere Unabhängigkeit und Normalverteilung – sind im Fall wiederholter Daten im Allgemeinen nicht erfüllt.<sup>2</sup> Sobald für verschiedene Subjekte oder Cluster mehr als eine Beobachtung vorliegt, treten Abhängigkeitsstrukturen innerhalb der Daten eines Subjekts bzw. Clusters auf. Bei einer klinischen Studie zum Beispiel, in der die Wirksamkeit eines Medikaments zur Reduzierung von Beschwerden bei mehreren Epilepsie-Erkrankten erforscht werden soll, ist die Anzahl der Epilepsieanfälle eines Patienten an einem bestimmten Tag abhängig von der Anzahl der Epilepsieanfälle an den vorherigen Tagen. Dies liegt an der subjekt-spezifischen Anfälligkeit und der gesundheitlichen Verfassung dieser Person. Anstelle von Patienten können auch Krankenhäuser, Schulklassen oder Haushalte betrachtet werden.

Die Missachtung der Abhängigkeiten zwischen den Beobachtungen eines Subjekts bzw. eines Clusters ist bedenklich, da es nach Sherman und Le Cessie (1997) zu Verzerrungen der Resultate im Sinne von Überschätzungen kommen kann.

Die besondere Herausforderung bei der Analyse wiederholter Daten stellt folglich die Abhängigkeitsstruktur der Daten dar, die berücksichtigt und modelliert werden muss. Einfache Zusammenhänge wie die Darstellung der Maximum-Likelihood-Funktion als Produktdichte der Einzeldichten im Fall unabhängiger Größen gelten bei der Betrachtung wiederholter Daten nicht mehr. Dementsprechend werden in der wissenschaftlichen Literatur verschiedene Ansätze für die Modellierung dieser Strukturen diskutiert.

Die in diesem Zusammenhang am häufigsten genannten Modelle sind die *marginalen Modelle*, die *subjekt-spezifischen Modelle* und die *bedingten Modelle*. Jede dieser Modellfamilien basiert auf den von McCullagh und Nelder (1989) vorgestellten *verallgemeinerten linearen Modellen* und liefert im gewissen Sinne eine Erweiterung für korrelierte Daten.

Bei der statistischen Erhebung von wiederholten Daten liegt am Ende der Studie für jedes Subjekt bzw. jeden Cluster ein Beobachtungsvektor vor.

Im *marginalen Modell* wird der Regressionszusammenhang zwischen dem Beobachtungsvektor und den erklärenden Variablen (z.B. den Versuchsbedingungen) durch die Modellierung des marginalen Erwartungswertes in Abhängigkeit von den erklärenden Variablen beschrieben. Die Abhängigkeitsstruktur wird durch Zusammen-

---

<sup>2</sup>Vgl. Dahmen und Ziegler (2004).

---

hangsmaße, wie etwa der Korrelation, separat modelliert.

Für die Analyse mit Hilfe eines *subjekt-spezifischen Modells* werden die erklärenden Variablen dagegen durch zusätzliche subjekt-spezifische Größen, die die subjekt-spezifischen Effekte charakterisieren, ergänzt. Eine zusätzliche Modellierung der Abhängigkeitsstruktur entfällt bei diesem Vorgehen.

Die dritte Modellfamilie – die *bedingten Modelle* – beschreibt die Verteilung der einzelnen Beobachtungen sowohl in Abhängigkeit der erklärenden Variablen als auch in Abhängigkeit von den vorherigen Beobachtungen. Dabei können die Beobachtungen beispielsweise zeitlich angeordnet sein.

Die Untersuchungen in dieser Arbeit konzentrieren sich auf die *marginalen* und die *subjekt-spezifischen Modelle* mit dem Ziel, diese beiden Modellfamilien vorzustellen, zu analysieren und zu vergleichen.

In diesem Zusammenhang werden im zweiten Kapitel zunächst das grundlegende Modell und die in der Arbeit verwendeten Notationen eingeführt. Zudem werden theoretische Grundlagen bereitgestellt.

Im ersten Teil des dritten Kapitels werden dann die *marginalen Modelle* definiert und erläutert. Es wird gezeigt, dass die von Liang und Zeger (1986) eingeführten *Verallgemeinerten Schätzgleichungen* (*generalized estimating equations*, GEE) zur Analyse und Auswertung marginaler Modelle genutzt werden können. Darüber hinaus wird durch das *Fisher-Scoring-Verfahren* eine iterative Methode zur Schätzung des gesuchten Regressionsvektors vorgestellt. Neben der Theorie der GEE werden die *Pseudo-Maximum-Likelihood-Methode* und die *Quasi-Generalisierte-Pseudo-Maximum-Likelihood-Methode* behandelt und basierend auf den Untersuchungen von Ziegler (1999) und White (1981) der Beweis zur Konsistenz des GEE-Schätzers skizziert. Darüber hinaus werden die von Liang und Zeger (1986) eingeführten *Arbeitskorrelationsmatrizen* betrachtet und deren Einfluss in der Analyse *marginaler Modelle* diskutiert.

Im zweiten Teil von Kapitel 3 werden die *subjekt-spezifischen Modelle* unter der Einbeziehung einer subjekt-spezifischen Größe vorgestellt. Für diese subjekt-spezifischen Größen werden unterschiedliche Handhabungen diskutiert, wobei der Schwerpunkt auf zufälligen, identisch verteilten subjekt-spezifischen Größen liegt. Es wird gezeigt, dass das subjekt-spezifische Modell in diesem Fall durch ein sogenanntes *verallgemeinertes lineares gemischtes Modell* (*generalized linear mixed models*, GLMM) dargestellt werden kann. Analog zum Vorgehen bei *marginalen Modellen* werden im Anschluss, basierend auf der Theorie der GLMM, Schätzmöglichkeiten für die ge-

suchten Parameter vorgestellt. In diesem Zusammenhang werden die *Hermite-Gauß-Quadratur* und die *adaptive Hermite-Gauß-Quadratur* eine Anwendung finden.

Ein Vergleich beider Modellfamilien wird in Kapitel 4 exemplarisch an einer Studie, die wiederholte Zählraten enthält, vorgenommen.

Das Kapitel 5 besteht aus zwei Teilen: Im ersten Teil werden bestehende Implementierungen für die Auswertung von *marginalen* und *subjekt-spezifischen Modellen* vorgestellt. Die Betrachtungen konzentrieren sich hierbei auf das Softwarepaket **SAS**, da dieses in Bereichen der Biometrie breite Anwendung findet und von vielen Nutzern als Standard angesehen wird.<sup>3</sup>

Im Anschluss werden zwei von der Schering AG (Berlin) durchgeführte Studien mittels *marginaler* und *subjekt-spezifischer Modelle* sowie den vorgestellten Prozeduren ausgewertet. In beiden Studien wird die Anzahl der täglichen Hitzewallungen von Frauen in den Wechseljahren erhoben; es liegen somit wiederholte Zählraten vor.

Im sechsten Kapitel wird für die Analyse *marginaler Modelle* auf das Themenfeld der Versuchsplanung eingegangen. Basierend auf der Arbeit von Liu und Liang (1997), wird eine Formel zur Berechnung des benötigten Stichprobenumfangs zur Einhaltung eines bestimmten Niveaus und einer festen Power angegeben. Diese Formel wird dann in Abhängigkeit von ihren Einflussgrößen untersucht.

Im siebten Kapitel wird überprüft, ob die in Kapitel 6 berechnete Fallzahlformel für unterschiedliche Parameterkonstellationen das vorgegebene Niveau und die vorgegebene Power einhält. Darüber hinaus wird die Güte der Auswertung durch die vorgestellten **SAS**-Prozeduren untersucht. Zu diesem Zweck ist es unerlässlich, korrelierte poissonverteilte Daten zu simulieren und ein entsprechendes Programm zur Erzeugung ebendieser aufzustellen.

Die Arbeit schließt im achten Kapitel mit einer Bewertung der Resultate und einem Ausblick ab. Zusätzliche Resultate und Methoden sind im Anhang zu finden.

---

<sup>3</sup>Vgl. Molenberghs und Verbeke (2005), S. 203.

# Kapitel 2

## Grundlagen

In diesem Kapitel werden die in der Arbeit verwendeten Notationen und Modelle eingeführt und einige theoretische Grundlagen bereitgestellt.

### 2.1 Das Modell

Bei der Betrachtung wiederholter Daten wird davon ausgegangen, dass eine bestimmte Anzahl von Beobachtungen für mehrere Subjekte wie Patienten oder Unternehmen vorliegen. Im Folgenden wird angenommen, dass  $m$  Subjekte untersucht und für das  $i$ -te Subjekt  $n_i$  Beobachtungen erhoben werden. Der  $n_i$ -dimensionale Beobachtungsvektor  $y_i$  sei nun eine Realisierung eines  $n_i$ -dimensionalen Zufallsvektors  $Y_i$ . Die einzelnen Komponenten des Zufallsvektors  $Y_i$  genügen einer Verteilung aus einer *1-parametrischen natürlichen Exponentialfamilie mit Dispersions-Parameter  $\zeta$* .

**Definition 2.1.** (1-parametrische natürliche Exponentialfamilie)

Besitzt  $Y_{ij}$  die Dichte

$$f_{\theta_{ij}, \zeta}(y_{ij}) = \exp\{[y_{ij} \theta_{ij} - \psi(\theta_{ij})] \zeta^{-1} + c(y_{ij}, \zeta)\},$$

mit unbekanntem  $\theta_{ij} \in \mathbb{R}$  und  $\zeta \in \mathbb{R}^+$  sowie bekannten Funktionen  $\psi : \mathbb{R} \rightarrow \mathbb{R}$  und  $c : \mathbb{R}^2 \rightarrow \mathbb{R}$ , dann gehört die Verteilung von  $Y_{ij}$  einer 1-parametrischen natürlichen Exponentialfamilie mit Dispersions-Parameter  $\zeta$  an.

Beispiele solcher Familien sind die Normalverteilungen, Poissonverteilungen, Gammaverteilungen und Binomialverteilungen. Ist  $Y_{ij}$  beispielsweise poissonverteilt mit

$Y_{ij} \sim \mathcal{P}(\lambda_{ij})$ , so lässt sich die Dichte von  $Y_{ij}$  darstellen als

$$\begin{aligned} f_{\lambda_{ij}}(y_{ij}) &= \frac{\lambda_{ij}^{y_{ij}} \exp(-\lambda_{ij})}{y_{ij}!} \\ &= \exp\{y_{ij} \ln(\lambda_{ij}) - \lambda_{ij} - \ln(y_{ij}!)\} \\ &= \exp\{[y_{ij} \theta_{ij} - \psi(\theta_{ij})] \zeta^{-1} + c(y_{ij}, \zeta)\} \end{aligned}$$

mit  $\theta_{ij} = \ln(\lambda_{ij})$ ,  $\psi(\theta_{ij}) = e^{\theta_{ij}}$ ,  $\zeta = 1$  und  $c(y_{ij}, \zeta) = -\ln(y_{ij}!)$ .

Für Zufallsvariablen mit einer Verteilung aus einer 1-parametrischen natürlichen Exponentialfamilie mit Dispersions-Parameter  $\zeta$  gilt ein besonderer Zusammenhang zwischen den ersten beiden Momenten von  $Y_{ij}$  und der Funktion  $\psi$  aus Definition 1.

**Satz 2.2.**

*Besitzt  $Y_{ij}$  eine Verteilung aus einer 1-parametrischen natürlichen Exponentialfamilie und lassen sich die Differentiation nach  $\theta_{ij}$  und die Integration nach  $y_{ij}$  vertauschen, so gilt*

$$E(Y_{ij}) = \psi'(\theta_{ij}) \quad \text{und} \quad \text{Var}(Y_{ij}) = \psi''(\theta_{ij}) \zeta.$$

Beweis: Siehe McCullagh und Nelder (1989), S. 28f.

Die Vertauschbarkeit der Integration und Differentiation kann oft mit Hilfe des Satzes über die *dominierte Konvergenz* gezeigt werden.

**Satz 2.3. (Satz über die dominierte Konvergenz)<sup>1</sup>**

*Sei  $(\Omega, \mathcal{A}, \varphi)$  ein Maßraum und sei  $(f_n)_{n \in \mathbb{N}}$  eine Folge von Funktionen, die auf  $\Omega$   $\varphi$ -integrierbar sind. Zudem werde  $(f_n)_{n \in \mathbb{N}}$  von einer  $\varphi$ -integrierbaren Funktion  $g$  auf  $\Omega$  dominiert, d.h. für alle  $n \in \mathbb{N}$  gilt  $|f_n| \leq g$   $\varphi$ -fast-überall.*

*Weiter konvergiere die Folge  $(f_n)_{n \in \mathbb{N}}$   $\varphi$ -fast-überall gegen eine  $\varphi$ -messbare Funktion  $f$ . Dann ist auch  $f$   $\varphi$ -integrierbar und es gilt:*

$$\lim_{n \rightarrow \infty} \int f_n d\varphi = \int f d\varphi.$$

Für poissonverteilte Zufallsvariablen lässt sich für die Folge der Differenzenquotienten mit Hilfe des Mittelwertsatzes eine Majorante  $g$  ermitteln, so dass sich aus Satz 2.2 die bekannte Beziehung zwischen dem Erwartungswert und der Varianz ergibt:

$$E(Y_{ij}) = (e^{\theta_{ij}})' = e^{\theta_{ij}} = \lambda_{ij}$$

und

$$\text{Var}(Y_{ij}) = (e^{\theta_{ij}})'' \zeta = e^{\theta_{ij}} \cdot 1 = \lambda_{ij}.$$

---

<sup>1</sup>Vgl. Bröcker (1995), S.98.

### 2.1.1 Verallgemeinerte lineare Modelle

Mit Hilfe von *verallgemeinerten linearen Modellen* können Zusammenhänge zwischen den beobachteten Daten und bekannten Einflussgrößen untersucht werden. Sei  $x_{ij}$  der  $q$ -dimensionale Vektor der Einflussgrößen bzw. der erklärenden Variablen. Ein lineares Modell ist dann gegeben durch

$$E(Y_{ij}) = x_{ij}^\top \beta,$$

mit dem  $q$ -dimensionalen Regressionsvektor  $\beta \in \Theta$ , wobei  $\Theta \subset \mathbb{R}^q$  der Parameterraum ist. Dieser Ansatz ist allerdings nur für einige Verteilungen, wie z.B. die Normalverteilung, realistisch, da die rechte Seite der Gleichung im Allgemeinen jede reelle Zahl annehmen kann, die linke Seite diese Voraussetzung allerdings nicht für jede Verteilung erfüllt. Bei der Poissonverteilung liegt der Erwartungswert in  $\mathbb{R}^+$ , so dass es in diesem Fall nicht ausreicht, lediglich das lineare Modell zu betrachten. Vielmehr muss zur Beschreibung des Regressionszusammenhangs im Fall poissonverteilter Daten das lineare Modell zu einem verallgemeinerten linearen Modell (*generalized linear model*, GLM) erweitert werden.

Ein verallgemeinertes lineares Modell charakterisiert den Zusammenhang zwischen den Beobachtungen  $y_{i1}, \dots, y_{in_i}$  und den erklärenden Variablen (z.B. den Versuchsbedingungen)  $x_{i1}, \dots, x_{in_i} \in \mathbb{R}^q$  über einen unbekanntem Regressionsvektor  $\beta \in \mathbb{R}^q$  und einer bekannten *Linkfunktion*  $h : A \rightarrow \mathbb{R}$  mit  $A \subseteq \mathbb{R}$  gemäß folgender Gleichung:

$$h(E(Y_{ij})) = x_{ij}^\top \beta \quad \text{für alle } i \in \{1, \dots, m\}, j \in \{1, \dots, n_i\}.$$

Als Linkfunktionen kommen je nach zugrundeliegender Verteilung verschiedene Funktionen in Frage. Aufgrund von günstigen Eigenschaften wird aber des Öfteren die *kanonische* Linkfunktion  $h$  mit

$$h(E(Y_{ij})) = \theta_{ij},$$

d.h.  $\theta_{ij} = x_{ij}^\top \beta$  bevorzugt. Aus der Darstellung für das verallgemeinerte lineare Modell und Satz 2.2 folgt für die *kanonische* Linkfunktion die Beziehung  $h = (\psi')^{-1}$ . Für poissonverteilte Daten entspricht der natürliche Logarithmus gerade der kanonischen Linkfunktion.

Somit ist ein verallgemeinertes lineares Modell gegeben, wenn

- die Dichten der unabhängigen Zufallsvariablen  $Y_{ij}$  für alle  $i \in \{1, \dots, m\}$ ,  $j \in \{1, \dots, n_i\}$  einer 1-parametrischen natürlichen Exponentialfamilie entstammen und

- der Erwartungswert von  $Y_{ij}$  mit dem Regressionsvektor  $\beta \in \Theta$  über eine eindeutige dreimal stetig differenzierbare und streng monotone Linkfunktion  $h$  gemäß

$$h(\mathbb{E}(Y_{ij})) = x_{ij}^\top \beta$$

verbunden ist.

Die Matrix  $X_i = (x_{i1}, \dots, x_{in_i})^\top$  wird auch *Prädiktormatrix* genannt.

## 2.1.2 Verteilungen

Im Folgenden werden die Verteilungen betrachtet, die im Rahmen dieser Arbeit eine Anwendung finden.

### Definition 2.4. (Poissonverteilung)

Als *Poissonverteilung* mit Erwartungswert  $\lambda$  und Varianz  $\lambda$ , kurz  $\mathcal{P}(\lambda)$ , wird die Verteilung  $P$  der Zufallsvariablen  $Y$  mit der Dichte

$$f_\lambda(y) = \frac{\lambda^y \exp(-\lambda)}{y!}$$

bezeichnet.

### Definition 2.5. (Normalverteilung)

Als (*eindimensionale*) *Normalverteilung* mit Erwartungswert  $\mu$  und Varianz  $\sigma^2$ , kurz  $\mathcal{N}(\mu, \sigma^2)$ , wird die Verteilung  $P$  der Zufallsvariablen  $Y$  mit der Dichte

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

bezeichnet.

### Definition 2.6. (Gammaverteilung)

Eine Zufallsvariable  $Y$  besitzt eine *Gammaverteilung* mit Parametern  $\alpha \in \mathbb{R}^+$  und  $\beta \in \mathbb{R}^+$ , oder kurz  $\Gamma(\alpha, \beta)$ , falls ihre Dichte die Form

$$f_{(\alpha, \beta)}(y) = \begin{cases} \frac{\alpha^{-\beta}}{\Gamma(\beta)} y^{\beta-1} e^{-\frac{y}{\alpha}} & y \geq 0 \\ 0 & y < 0 \end{cases}$$

besitzt. Dabei ist die *Gammafunktion*  $\Gamma(x)$  definiert durch

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

**Bemerkung 2.7.**

Für  $Y \sim \Gamma(\alpha, \beta)$  gilt

$$E(Y) = \alpha \beta \quad \text{und} \quad \text{Var}(Y) = \alpha \beta^2.$$

**Definition 2.8. (Negativ-Binomialverteilung)**

Die Zufallsvariable  $Y$  mit dem Träger  $\mathbb{N}_0$  besitzt eine Negativ-Binomialverteilung mit Parametern  $r \in \mathbb{R}^+$  und  $p \in (0, 1)$ , kurz  $\mathcal{NB}(r, p)$ , falls die Dichte von  $Y$

$$f_{(r,p)}(y) = \frac{\Gamma(r+y)}{y! \Gamma(r)} p^r (1-p)^y$$

entspricht.

**Bemerkung 2.9.**

Für die ersten beiden Momente einer negativ-binomialverteilten Zufallsvariable  $Y$  gilt

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

## 2.2 Notationen

### 2.2.1 Ableitungen

Für  $X \subset \mathbb{R}^q, Y \subset \mathbb{R}^p$  und eine differenzierbare Funktion  $f : Y \times X \rightarrow \mathbb{R}$  bezeichne

$$\frac{\partial f(y, x)}{\partial x} = \left( \frac{\partial f(y, x)}{\partial x_1}, \dots, \frac{\partial f(y, x)}{\partial x_q} \right) \in \mathbb{R}^{1 \times q}$$

die *partielle Ableitung* nach  $x$  und für  $k \in \mathbb{N}$  bezeichne entsprechend

$$\frac{\partial^k f(y, x)}{\partial^k x}$$

die *k-te partielle Ableitung* nach  $x$ . Für  $k = 2$  gilt

$$\frac{\partial^2 f(y, x)}{\partial^2 x} = \left( \frac{\partial^2 f(y, x)}{\partial x_i \partial x_j} \right)_{\substack{i \in \{1, \dots, q\} \\ j \in \{1, \dots, q\}}} \in \mathbb{R}^{q \times q}.$$

### 2.2.2 Likelihood-Funktion, Score-Funktion, ML-Schätzer

Seien  $Y_1, \dots, Y_n$  unabhängige, identisch verteilte Zufallsvariablen mit Dichte  $f_\theta(y_i)$ ,  $\theta \in \Theta \subset \mathbb{R}^q$  und  $Y = (Y_1^\top, \dots, Y_n^\top)^\top$ . Die Funktion

$$L_y(\theta) := \prod_{i=1}^n f_\theta(y_i)$$

bezeichne die *Likelihood-Funktion* und entsprechend

$$l_y(\theta) := \ln \left( \prod_{i=1}^n f_\theta(y_i) \right) = \sum_{i=1}^n \ln (f_\theta(y_i))$$

die *Loglikelihood-Funktion*.

Ein Schätzer  $\hat{\theta}$ , der die Bedingung

$$\hat{\theta} = \operatorname{argmax}_{\theta \in \Theta} L_y(\theta)$$

erfüllt, heißt *Maximum-Likelihood-Schätzer*. Aufgrund der Monotonie des Logarithmus gilt äquivalent

$$\hat{\theta} = \operatorname{argmax}_{\theta \in \Theta} l_y(\theta).$$

Zudem sei die *Score-Funktion* definiert durch

$$S(\beta) = \left( \frac{\partial l_y(\theta)}{\partial \theta} \right)^\top \in \mathbb{R}^q.$$

### 2.2.3 Konvergenzarten

Sei  $(Y_n)_{n \in \mathbb{N}}$  eine Folge von Zufallsvariablen. Dann konvergiert diese Folge *fast sicher* gegen  $Y$ , oder kurz  $Y_n \xrightarrow{f.s.} Y$ , falls

$$\mathbb{P} \left( |Y_n - Y| \xrightarrow{n \rightarrow \infty} 0 \right) = 1$$

gilt.

Die Folge  $(Y_n)_{n \in \mathbb{N}}$  konvergiert *stochastisch* oder *in Wahrscheinlichkeit* gegen  $Y$ , falls für alle  $\varepsilon > 0$

$$\mathbb{P} (|Y_n - Y| > \varepsilon) \xrightarrow{n \rightarrow \infty} 0$$

gilt, kurz  $Y_n \xrightarrow{\mathcal{P}} Y$ .

Die Folge  $(Y_n)_{n \in \mathbb{N}}$  konvergiert *in Verteilung* gegen  $Y$ , oder kurz  $Y_n \xrightarrow{\mathcal{D}} Y$ , falls für alle Stetigkeitspunkte  $y$  der Verteilungsfunktion  $F$  von  $Y$

$$F_n(y) = \mathbb{P}(Y_n \leq y) \xrightarrow{n \rightarrow \infty} \mathbb{P}(Y \leq y) = F(y)$$

gilt.

Für die soeben genannten Konvergenzbegriffe gilt folgender Zusammenhang:

$$Y_n \xrightarrow{f.s.} Y \implies Y_n \xrightarrow{\mathcal{P}} Y \implies Y_n \xrightarrow{\mathcal{D}} Y.$$

## 2.2.4 Landau-Symbole

Die *Landau-Symbole* werden verwendet, um das asymptotische Verhalten von Funktionen und Folgen zu beschreiben. Sie lassen sich formal für zwei Folgen  $(a_n)_{n \in \mathbb{N}}$ ,  $(b_n)_{n \in \mathbb{N}}$  mit  $b_n \neq 0$  für alle  $n \in \mathbb{N}$  folgendermaßen definieren:

$$a_n \in o(b_n) : \iff \frac{a_n}{b_n} \xrightarrow{n \rightarrow \infty} 0$$

und

$$a_n \in O(b_n) : \iff 0 \leq \limsup_{n \rightarrow \infty} \frac{a_n}{b_n} < \infty.$$

Für zwei Folgen von Zufallsvariablen  $(X_n)_{n \in \mathbb{N}}$ ,  $(Y_n)_{n \in \mathbb{N}}$  mit  $P(Y_n \neq 0) = 1$  für alle  $n \in \mathbb{N}$  seien

$$X_n \in o_p(Y_n) : \iff \frac{X_n}{Y_n} \xrightarrow{\mathcal{P}} 0$$

und

$$X_n \in O_p(Y_n) : \iff \forall \varepsilon > 0 \exists M, \text{ so dass } \sup_{n \in \mathbb{N}} P\left(\left|\frac{X_n}{Y_n}\right| > M\right) < \varepsilon$$

definiert. Damit bezeichnet  $O_p(1)$  die stochastische Beschränktheit einer Folge von Zufallsvariablen und  $o_p(1)$  die stochastische Konvergenz gegen Null.

## 2.2.5 Matrizen

Es sei

$$\text{diag}(\lambda_1, \dots, \lambda_n) := \begin{pmatrix} \lambda_1 & 0 & \cdots & \cdots & 0 \\ 0 & \lambda_2 & 0 & \cdots & 0 \\ \vdots & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & 0 & \cdots & 0 & \lambda_n \end{pmatrix}$$

die  $n \times n$  Diagonalmatrix mit Einträgen  $\lambda_i \in \mathbb{R}$ ,  $i \in \{1, \dots, n\}$ .



# Kapitel 3

## Das Grundmodell und Modellierungsmöglichkeiten

Wiederholte Daten sind in den Naturwissenschaften weit verbreitet und treten dann auf, wenn bei verschiedenen Subjekten bzw. Clustern mehr als eine einzige Beobachtung desselben Merkmals pro Subjekt bzw. Cluster erhoben wird. Diese Beobachtungen werden unter bestimmten Versuchsbedingungen, wie etwa der Einnahme eines Medikaments oder der Durchführung einer Behandlung, gewonnen. Im Folgenden sollen diese Beobachtungen Zähldaten sein. Diese Art von Daten wird üblicherweise als poissonverteilt angenommen.<sup>1</sup> Je nach Anzahl der Subjekte und der jeweiligen Beobachtungen gibt es unterschiedliche Modelle zur Untersuchung des Zusammenhangs zwischen den Beobachtungen und den Versuchsbedingungen.

Existiert pro Subjekt  $i \in \{1, \dots, m\}$  nur eine Beobachtung, dann können alle Beobachtungen als unabhängig angenommen werden und der Zusammenhang zwischen den Beobachtungen und den Versuchsbedingungen  $x_i \in \mathbb{R}^q$  lässt sich mit Hilfe des verallgemeinerten linearen Modells bestimmen:

$$\ln(\mathbb{E}(Y_i)) = x_i^\top \beta$$

für alle  $i \in \{1, \dots, m\}$ , mit dem Regressionsvektor  $\beta \in \Theta \subset \mathbb{R}^q$ .

Existiert andererseits nur ein Subjekt (d.h.  $m = 1$ ) und mehr als eine Beobachtung zu diesem Subjekt, dann sind die Beobachtungen bedingt auf den Subjekteinflüssen unabhängig und es kann erneut ein verallgemeinertes lineares Modell angepasst werden.

Wie verhalten sich jedoch wiederholte Zähldaten, wenn  $m > 1$  und  $n_i > 1$  für mindestens ein  $i$  ist?

---

<sup>1</sup>Vgl. Diggle u. a. (1996), S. 186.

Um diesen Fall genauer zu untersuchen, wird im Folgenden angenommen, dass  $m > 1$  gilt und ein  $i$  existiert, so dass  $n_i > 1$  ist. Für jedes Subjekt  $i \in \{1, \dots, m\}$  liegt also ein  $n_i$ -dimensionaler Beobachtungsvektor  $y_i$  vor, welcher eine Realisierung eines  $n_i$ -dimensionalen Zufallsvektors  $Y_i$  ist. Es wird angenommen, dass die einzelnen Komponenten  $Y_{ij}$ ,  $j \in \{1, \dots, n_i\}$ , des Zufallsvektors  $Y_i$  einer Verteilung aus einer 1-parametrischen natürlichen Exponentialfamilie genügen.

Zur Analyse des Regressionszusammenhangs zwischen den Versuchsbedingungen und den Beobachtungen reicht ein verallgemeinertes lineares Modell nicht mehr aus, da die Beobachtungen eines Subjekts abhängig voneinander sind und die Abhängigkeitsstruktur der Daten somit berücksichtigt werden muss.

In der aktuellen Fachliteratur werden drei Modelle für diesen Fall diskutiert, wobei alle auf den von McCullagh und Nelder (1989) vorgestellten verallgemeinerten linearen Modellen basieren. In den folgenden Abschnitten werden allerdings lediglich zwei Modellfamilien, das marginale und das subjekt-spezifische Modell, vorgestellt. Beide Modelle ermöglichen es durch die Bestimmung des Regressionsvektors  $\beta$ , Aussagen über die Auswirkungen der erklärenden Variablen (z.B. ein Medikament) auf den Populationsdurchschnitt (z.B. die Anzahl der Hitzewallungen im Populationsdurchschnitt) zu treffen.

Im marginalen Modell werden die marginalen Verteilungen benutzt, um die Beobachtungen in Abhängigkeit von den erklärenden Variablen zu beschreiben. Die Abhängigkeitsstruktur wird durch eine Menge von Abhängigkeitsparametern, wie etwa der Korrelation oder den Odds Ratios, erfasst und getrennt vom Regressionszusammenhang modelliert.

Alternativ wird die Abhängigkeitsstruktur der Daten im subjekt-spezifischen Modell durch die zugrundeliegende Heterogenität zwischen den einzelnen Subjekten charakterisiert. Dieses Modell kann auch dann genutzt werden, wenn Rückschlüsse über einzelne Individuen, nicht aber über den Populationsdurchschnitt gezogen werden sollen. Im subjekt-spezifischen Modell werden die erklärenden Variablen durch zusätzliche subjekt-spezifische Größen ergänzt. Diese subjekt-spezifischen Größen können auf drei unterschiedliche Weisen gehandhabt werden, so dass drei verschiedene Teilmodelle der subjekt-spezifischen Modelle unterschieden werden (vgl. Abschnitt 3.2.1).

Welches der Modelle in welchem Fall angewendet wird, hängt im Wesentlichen von der Fragestellung ab. Liegt das Interesse am Populationsdurchschnitt, wie bei der Untersuchung der durchschnittlichen Hitzewallungsanzahl, dann sind beide Model-

le geeignet. Trotzdem wird nach Molenberghs und Verbeke (2005) in diesem Fall oftmals das marginale Modell dem subjekt-spezifischen Modell vorgezogen. Dies könnte unter anderem an der leichteren Implementierung in gängigen Softwarepaketen für die marginalen Modelle liegen. Ein weiterer Grund könnte sein, dass beim subjekt-spezifischen Modell Aussagen über den Populationsdurchschnitt nicht direkt aufgrund der geschätzten Regressionsparameter möglich sind, sondern weitere Berechnungen durchgeführt werden müssen - hierauf wird im Laufe der Arbeit eingegangen, siehe Kapitel 4.

Sind die Subjekte aber unterschiedlichen, unbeobachtbaren Einflüssen, wie z.B. der Luftverschmutzung oder dem Wetter, ausgesetzt, erscheint ein subjekt-spezifisches Modell sinnvoller.

### 3.1 Marginale Modelle

Das Prinzip der marginalen Modelle besteht darin, für jedes  $i \in \{1, \dots, m\}$  die Regression von  $Y_i = (Y_{i1}, \dots, Y_{in_i})$  in Abhängigkeit von der Prädiktormatrix  $X_i = (x_{i1}, \dots, x_{in_i})^\top$  unabhängig von der Abhängigkeitsstruktur von  $Y_i$  zu modellieren. Damit besteht das marginale Modell für ein Subjekt  $i$  im Wesentlichen aus zwei Gleichungskomplexen, die getrennt modelliert werden.

Die Regression des marginalen Erwartungswertes wird hierbei als Funktion der erklärenden Variablen beschrieben. Dies entspricht der bekannten Vorgehensweise bei unkorrelierten Daten, so dass marginale Modelle als natürliches Analogon zu den GLM für korrelierte Daten angesehen werden können.

Allgemein wird nach Diggle u. a. (1996) ein marginales Modell durch drei Eigenschaften beschrieben:

- Der Zusammenhang zwischen der marginalen Erwartung von  $Y_{ij}$  und dem erklärenden Variablenvektor  $x_{ij} \in \mathbb{R}^q$  wird durch folgende Gleichungen charakterisiert:

$$\begin{aligned} - E(Y_{ij}) &= \mu_{ij} \quad \forall i \in \{1, \dots, m\}, j \in \{1, \dots, n_i\} \\ - h(\mu_{ij}) &= x_{ij}^\top \beta \quad \forall i \in \{1, \dots, m\}, j \in \{1, \dots, n_i\}. \end{aligned}$$

Dabei ist  $h$  eine bekannte Linkfunktion,  $Y_{ij}$  eine Zufallsvariable mit einer Verteilung aus einer 1-parametrischen natürlichen Exponentialfamilie und  $\beta \in \Theta \subset \mathbb{R}^q$  der Regressionsvektor.

- Die marginale Varianz hängt vom marginalen Erwartungswert gemäß der Gleichung

$$\text{Var}(Y_{ij}) = v(\mu_{ij})\phi \quad (3.1)$$

ab, wobei  $v$  eine bekannte Varianzfunktion und  $\phi \in \mathbb{R}$  ein Parameter ist, der oft geschätzt werden muss.

- Die Korrelation zwischen  $Y_{ij}$  und  $Y_{ik}$  ist eine Funktion der marginalen Erwartungswerte  $\mu_{ij}$ ,  $\mu_{ik}$  und eines zusätzlichen Parameters  $\alpha$ , also

$$\text{Corr}(Y_{ij}, Y_{ik}) = \rho(\mu_{ij}, \mu_{ik}, \alpha), \quad (3.2)$$

wobei auch  $\rho$  eine bekannte Funktion ist, welche die Abhängigkeitsstruktur quantifiziert, und  $\alpha \in \mathbb{R}$  eine Variable dieser Funktion ist, die gegebenenfalls geschätzt werden muss.

Im Folgenden wird es das Ziel sein, die obigen Charakterisierungen eines marginalen Modells für wiederholte Zählraten zu spezialisieren. Wie bereits erwähnt, wird für die Beschreibung dieser Daten im Allgemeinen die Poissonverteilung vorgeschlagen. Bei dieser Wahl muss allerdings berücksichtigt werden, dass die Poissonverteilung durch den Erwartungswert, welcher gleichzeitig der Varianz entspricht, vollständig bestimmt ist. Die meisten Datensätze werden der Voraussetzung der Gleichheit von Erwartungswert und Varianz nicht genügen, da eine sogenannte *Überdispersion* vorliegt, d.h. die ermittelte Varianz übertrifft die nach der Theorie vorliegende Varianz. Diesem Umstand muss daher im Modell für Zählraten Rechnung getragen werden. Darüber hinaus muss geklärt werden, welche Linkfunktion  $h$  im Fall poissonverteilter Daten in Gleichung (3.1) verwendet werden soll. Prinzipiell kommt jede Funktion in Frage, die das Intervall  $(0, \infty[$  auf  $\mathbb{R}$  abbildet. Aufgrund der besonderen Eigenschaften der kanonischen Linkfunktion, d.h. dem natürlichen Logarithmus, werden im Folgenden die Betrachtungen auf diese Linkfunktion reduziert. Insgesamt lässt sich also ein marginales Modell für poissonverteilte Daten mit Überdispersion durch die nachfolgenden Punkte charakterisieren:

- Für die marginale Erwartung von  $Y_{ij} \sim \mathcal{P}(\lambda_{ij})$  und die erklärenden Variablen gilt

$$\begin{aligned} & - E(Y_{ij}) = \lambda_{ij} \quad \forall i \in \{1, \dots, m\}, j \in \{1, \dots, n_i\} \text{ und} \\ & - \ln(\lambda_{ij}) = x_{ij}^\top \beta \quad \forall i \in \{1, \dots, m\}, j \in \{1, \dots, n_i\}. \end{aligned}$$

- Die marginale Varianz wird beschrieben durch die Gleichung

$$\text{Var}(Y_{ij}) = \lambda_{ij}\phi, \quad (3.3)$$

wobei  $\phi$  der Überdispersions-Parameter ist, der gegebenenfalls geschätzt werden muss. Für  $\phi = 1$  ergibt sich  $E(Y_{ij}) = \text{Var}(Y_{ij})$ .

- Die Korrelation zwischen  $Y_{ij}$  und  $Y_{ik}$  ist eine Funktion von  $\lambda_{ij}$ ,  $\lambda_{ik}$  und einem zusätzlichen Parameter  $\alpha \in \mathbb{R}$ , d.h.

$$\text{Corr}(Y_{ij}, Y_{ik}) = \rho(\lambda_{ij}, \lambda_{ik}, \alpha), \quad (3.4)$$

wobei  $\rho$  eine bekannte Funktion ist.

Wie bereits erwähnt, besteht das marginale Modell aus zwei Gleichungskomplexen, wobei der erste Komplex den Regressionszusammenhang liefert und der zweite die Abhängigkeitsstrukturen beschreibt.

Das betrachtete marginale Modell besitzt drei unbekannte Größen, die bestimmt bzw. geschätzt werden müssen. Diese sind der Regressionsvektor  $\beta$ , der den Regressionszusammenhang liefert, der Überdispersions-Parameter  $\phi$  und  $\alpha$ , eine der Variablen der Korrelationsfunktion in Gleichung (3.2). Welche unterschiedlichen Schätzmöglichkeiten angewendet werden können und welche sinnvoll sind, wird im nächsten Abschnitt betrachtet.

### 3.1.1 Schätzmethoden für das marginale Modell

Die zu bestimmenden Parameter  $\beta$ ,  $\phi$  und  $\alpha$  könnten wie üblich mit Hilfe der Maximum-Likelihood-Methode geschätzt werden. Dies setzt allerdings voraus, dass die gemeinsame Verteilung von  $Y = (Y_1^\top, \dots, Y_m^\top)^\top$  vorliegt. Aufgrund der Abhängigkeit der Daten lässt sich die gemeinsame Dichte von  $Y$  allerdings nicht als Produktdichte der Einzeldichten berechnen, so dass die Berechnung der gemeinsamen Dichte kompliziert oder gar unmöglich ist. Darüber hinaus ist die gesamte Verteilung von  $Y$  nur selten von Interesse. Vielmehr interessieren der Erwartungswertvektor und die Kovarianz bzw. die Korrelation.

Die Korrelationsstruktur der Daten darf auch keinesfalls ignoriert werden, da die Genauigkeit und die Signifikanz der Ergebnisse andernfalls oftmals überschätzt wird.<sup>2</sup> Darüber hinaus kann nach Cannon u. a. (2001) die Vernachlässigung der Korrelation zu falschen Schlussfolgerungen führen.

Aus diesem Grund waren und sind Alternativen gefragt, die das Schätzen der genannten Parameter erleichtern. Eine Möglichkeit, die nicht der Bestimmung der gesamten Verteilung bedarf, wurde 1986 von Liang und Zeger veröffentlicht. Ihr Artikel beschäftigt sich mit *Verallgemeinerten Schätzgleichungen* (*generalized estima-*

<sup>2</sup>Vgl. Sherman und Le Cessie (1997).

*ting equations*, GEE), die konsistente Schätzer für den Regressionsvektor  $\beta$  liefern. Diese Methode wurde dann sowohl in theoretischen als auch in angewandten Zeitschriften weiter entwickelt und erweitert, so dass heutzutage unterschieden wird zwischen GEE1, GEE2 und der *alternierenden logistischen Regression*. All diese Methoden basieren auf den *Verallgemeinerten Schätzgleichungen*, und sie werden im folgenden Abschnitt, wie von Liang und Zeger (1986), hergeleitet. Allerdings wird im Anschluss auch eine alternative Möglichkeit zur Herleitung der GEE vorgestellt, um basierend auf dieser den Beweis der Eigenschaften von den GEE-Schätzern führen zu können.

### 3.1.1.1 Verallgemeinerte Schätzgleichungen (GEE1)

Die Methode der *Verallgemeinerten Schätzgleichungen*, wie sie von Liang und Zeger (1986) hergeleitet wurde, basiert auf der Grundidee, den Regressionsparameter  $\beta$  mit Hilfe von Schätzungen für  $\alpha$  und  $\phi$  zu berechnen. Dazu wird zunächst angenommen, dass die Zufallsgrößen  $Y_{ij}$  für alle  $i \in \{1, \dots, m\}$  und  $j \in \{1, \dots, n_i\}$  unabhängig sind. Im Folgenden wird zunächst der allgemeine Fall untersucht, d.h. die Verteilung von  $Y_{ij}$  entstammt einer 1-parametrischen natürlichen Exponentialfamilie, und anschließend der Fall für poissonverteilte Daten.

Für die Loglikelihood-Funktion von  $Y = (Y_1^\top, \dots, Y_m^\top)^\top$  ergibt sich aufgrund der Verteilungsannahme folgendes:

$$\begin{aligned}
 l_y(\beta) &= \ln \left( \prod_{i=1}^m \prod_{j=1}^{n_i} \exp \left( [y_{ij} \theta_{ij} - \psi(\theta_{ij})] \zeta^{-1} + c(y_{ij}, \zeta) \right) \right) \\
 &= \sum_{i=1}^m \sum_{j=1}^{n_i} [y_{ij} \theta_{ij} - \psi(\theta_{ij})] \zeta^{-1} + c(y_{ij}, \zeta) \\
 &\stackrel{\text{Poisson}}{=} \sum_{i=1}^m \sum_{j=1}^{n_i} y_{ij} \ln(\lambda_{ij}) - \lambda_{ij} - \ln(y_{ij}!).
 \end{aligned} \tag{3.5}$$

Die Differentiation dieses Ausdrucks nach  $\beta$  und anschließendes Nullsetzen dieser Ableitung liefert unter der Annahme, dass  $\zeta$  unabhängig von  $\beta$  ist, die Gleichung

$$\tilde{S}(\beta) = \sum_{i=1}^m \sum_{j=1}^{n_i} \left[ y_{ij} \frac{\partial \theta_{ij}}{\partial \beta} - \psi'(\theta_{ij}) \frac{\partial \theta_{ij}}{\partial \beta} \right] \zeta^{-1} \tag{3.6}$$

$$\begin{aligned}
 &\stackrel{\text{Poisson}}{=} \sum_{i=1}^m \sum_{j=1}^{n_i} \left[ y_{ij} \frac{\partial \ln(\lambda_{ij})}{\partial \beta} - \frac{\partial \lambda_{ij}}{\partial \beta} \right] \\
 &= 0_{1 \times q},
 \end{aligned} \tag{3.7}$$

wobei  $\theta_{ij} = \psi^{-1}(\mu_{ij}) = \psi^{-1}(h^{-1}(x_{ij}^\top \beta))$  gilt. Darüber hinaus ist  $E(Y_{ij}) = \mu_{ij} = \psi'(\theta_{ij}) \stackrel{\text{Poisson}}{=} \lambda_{ij}$ , so dass mit der Kettenregel die Gleichung

$$\frac{\partial \mu_{ij}}{\partial \beta} = \frac{\partial \psi'(\theta_{ij})}{\partial \beta} = \psi''(\theta_{ij}) \frac{\partial \theta_{ij}}{\partial \beta} \quad (3.8)$$

folgt. Im Fall poissonverteilter Daten ergibt sich unter Berücksichtigung der Überdispersion (d.h.  $\psi''(\theta_{ij}) \hat{=} \phi \psi''(\theta_{ij})$ ) analog die Beziehung

$$\frac{\partial \lambda_{ij}}{\partial \beta} = \lambda_{ij} \frac{\partial \ln(\lambda_{ij})}{\partial \beta} = \phi \psi''(\ln(\lambda_{ij})) \frac{\partial \ln(\lambda_{ij})}{\partial \beta}.$$

Das Umstellen der Gleichung (3.8) liefert

$$\frac{\partial \theta_{ij}}{\partial \beta} = \frac{\partial \mu_{ij}}{\partial \beta} \left( \psi''(\theta_{ij}) \right)^{-1}$$

und für Zähldaten

$$\frac{\partial \ln(\lambda_{ij})}{\partial \beta} = \frac{\partial \lambda_{ij}}{\partial \beta} \left( \phi \psi''(\ln(\lambda_{ij})) \right)^{-1},$$

so dass durch Einsetzen in Gleichung (3.6) die veränderte Gleichung

$$\begin{aligned} \tilde{S}(\beta, \phi) &= \sum_{i=1}^m \sum_{j=1}^{n_i} y_{ij} \frac{\partial \theta_{ij}}{\partial \beta} - \psi'(\theta_{ij}) \frac{\partial \theta_{ij}}{\partial \beta} \\ &= \sum_{i=1}^m \sum_{j=1}^{n_i} \frac{\partial \theta_{ij}}{\partial \beta} (y_{ij} - \psi'(\theta_{ij})) \\ &= \sum_{i=1}^m \sum_{j=1}^{n_i} \frac{\partial \mu_{ij}}{\partial \beta} \left( \psi''(\theta_{ij}) \right)^{-1} (y_{ij} - \psi'(\theta_{ij})) \\ &\stackrel{\text{Poisson}}{=} \sum_{i=1}^m \sum_{j=1}^{n_i} \frac{\partial \lambda_{ij}}{\partial \beta} \phi^{-1} \left( \psi''(\ln(\lambda_{ij})) \right)^{-1} (y_{ij} - \psi'(\ln(\lambda_{ij}))) \\ &= 0_{1 \times q} \end{aligned} \quad (3.9)$$

entsteht. Da  $\beta = (\beta_1, \dots, \beta_q)^\top \in \mathbb{R}^q$  ist, entspricht die Ableitung  $\frac{\partial \mu_{ij}}{\partial \beta}$  dem Vektor

$$\left( \frac{\partial \mu_{ij}}{\partial \beta_1}, \dots, \frac{\partial \mu_{ij}}{\partial \beta_q} \right) \in \mathbb{R}^{1 \times q},$$

weshalb der gesamte Ausdruck in Gleichung (3.9) ebenfalls in  $\mathbb{R}^{1 \times q}$  liegt. Nun ist es unerheblich, ob ein  $q$ -dimensionaler Zeilenvektor oder ein Spaltenvektor gleich Null gesetzt wird. Aus diesem Grund wird im Folgenden die transponierte Gleichung

$$\begin{aligned} S(\beta, \phi) &= \sum_{i=1}^m \sum_{j=1}^{n_i} \left( \frac{\partial \mu_{ij}}{\partial \beta} \right)^\top \left( \psi''(\theta_{ij}) \right)^{-1} (y_{ij} - \psi'(\theta_{ij})) \\ &= 0_q, \end{aligned} \quad (3.10)$$

die sogenannte *Score-Gleichung*, benutzt. In den vorangegangenen Beobachtungen wird die Ableitung von  $\mu_{ij}$  nach  $\beta$  betrachtet und es gilt insbesondere

$$\begin{aligned}\mu_{ij} : \mathbb{R}^q &\longrightarrow \mathbb{R} \\ \beta &\mapsto \mu_{ij}(\beta) = h^{-1}(x_{ij}^\top \beta).\end{aligned}$$

Für die Herleitung von Gleichung (3.11) wird die Ableitung von  $\mu_i = (\mu_{i1}, \dots, \mu_{in_i})^\top$  nach  $\beta$  benötigt, so dass diese bestimmt werden muss. Auch  $\mu_i$  ist eine Funktion in Abhängigkeit von  $\beta$ , denn es gilt

$$\begin{aligned}\mu_i : \mathbb{R}^q &\longrightarrow \mathbb{R}^{n_i} \\ \beta &\mapsto (\mu_{i1}(\beta), \dots, \mu_{in_i}(\beta))^\top.\end{aligned}$$

Die Ableitung dieser Funktion nach  $\beta$  ist durch die  $(n_i \times q)$ -Jacobi-Matrix

$$\frac{\partial \mu_i}{\partial \beta} = \begin{pmatrix} \frac{\partial \mu_{i1}}{\partial \beta_1} & \dots & \frac{\partial \mu_{i1}}{\partial \beta_q} \\ \vdots & \ddots & \vdots \\ \frac{\partial \mu_{in_i}}{\partial \beta_1} & \dots & \frac{\partial \mu_{in_i}}{\partial \beta_q} \end{pmatrix}$$

gegeben. Somit folgt aus Gleichung (3.10) mit  $\psi'(\theta_{ij}) = \mu_{ij}$ :

$$\begin{aligned}S(\beta, \phi) &= \sum_{i=1}^m \sum_{j=1}^{n_i} \left( \frac{\partial \mu_{ij}}{\partial \beta} \right)^\top \left( \psi''(\theta_{ij}) \right)^{-1} (y_{ij} - \psi'(\theta_{ij})) \\ &= \sum_{i=1}^m \left( \frac{\partial \mu_i}{\partial \beta} \right)^\top \text{diag} \left( \psi''(\theta_{i1}), \psi''(\theta_{i2}), \dots, \psi''(\theta_{in_i}) \right)^{-1} \times (y_i - \mu_i) \\ &= \sum_{i=1}^m \left( \frac{\partial \mu_i}{\partial \beta} \right)^\top (A_i)^{-1} (y_i - \mu_i) \\ &= \sum_{i=1}^m \left( \frac{\partial \mu_i}{\partial \beta} \right)^\top \left( A_i^{1/2} I_{n_i} A_i^{1/2} \right)^{-1} (y_i - \mu_i)\end{aligned}\tag{3.11}$$

$$\begin{aligned}&\stackrel{\text{Poisson}}{=} \sum_{i=1}^m \left( \frac{\partial \lambda_i}{\partial \beta} \right)^\top \left( (A_i^\phi)^{1/2} I_{n_i} (A_i^\phi)^{1/2} \right)^{-1} (y_i - \lambda_i) \\ &= 0_q,\end{aligned}\tag{3.12}$$

wobei  $A_i = \text{diag}(\psi''(\theta_{i1}), \dots, \psi''(\theta_{in_i}))$  die Diagonalmatrix der marginalen Varianzen ist. Für poissonverteilte Daten gilt aufgrund der Überdispersion, dass  $A_i^\phi = \phi \text{diag}(\lambda_{i1}, \dots, \lambda_{in_i})$  ist. Darüber hinaus ist  $\mu_i = (\mu_{i1}, \dots, \mu_{in_i})^\top \stackrel{\text{Poisson}}{=} (\lambda_{i1}, \dots, \lambda_{in_i})^\top = \lambda_i$  und  $I_{n_i}$  die  $n_i \times n_i$  Einheitsmatrix, welche im Fall der Unabhängigkeit der Korrelationsmatrix von  $Y_i = (Y_{i1}, \dots, Y_{in_i})^\top$  entspricht. Diese Darstellung ist motiviert durch die Tatsache, dass  $A_i^{1/2} I_{n_i} A_i^{1/2}$  die Kovarianzmatrix im Fall unabhängiger Beobachtungen darstellt.

Der Ausdruck (3.11) basiert allerdings auf der Annahme, dass die Zufallsgrößen  $Y_{ij}$  unabhängig sind, so dass diese Vorgehensweise für wiederholte und damit korrelierte Daten unzulässig ist. Eine einfache und gleichzeitig naive Erweiterung für korrelierte Daten kann nach Liang und Zeger (1986) aufgestellt werden, indem die Einheitsmatrix in Gleichung (3.11) ersetzt wird durch eine quadratische,  $n_i$ -dimensionale Korrelationsmatrix  $R_i$ , die die Abhängigkeitsstruktur der Daten beschreibt. Da die Korrelation im marginalen Modell (vgl. Gleichung (3.2)) in Abhängigkeit vom Parameter  $\alpha$  beschrieben wurde, gilt  $R_i = R_i(\alpha)$ .

Somit ergibt sich

$$S(\beta, \phi, \alpha) = \sum_{i=1}^m \left( \frac{\partial \mu_i}{\partial \beta} \right)^\top \left( A_i^{1/2} R_i(\alpha) A_i^{1/2} \right)^{-1} (y_i - \mu_i) \quad (3.13)$$

$$\stackrel{\text{Poisson}}{=} \sum_{i=1}^m \left( \frac{\partial \lambda_i}{\partial \beta} \right)^\top \underbrace{\left( (A_i^\phi)^{1/2} R_i(\alpha)^{-1} (A_i^\phi)^{1/2} \right)}_{=: V_i^{\phi^{-1}}} (y_i - \lambda_i) \quad (3.14)$$

$$= 0_q,$$

wobei  $A_i = A_i(\beta)$  und bei Überdispersion  $A_i^\phi = A_i(\beta, \phi)$  gilt.

Nach Liang und Zeger (1986) ist es jetzt zulässig, die Korrelationsstruktur in Gleichung (3.2) willkürlich festzusetzen und den Parameter  $\alpha$  basierend auf dieser Strukturannahme zu schätzen. Die Matrix  $R_i(\alpha)$  wird somit entsprechend der gewählten Struktur durch die sogenannte *Arbeitskorrelationsmatrix*  $R_i(\hat{\alpha})$  ersetzt. Dabei können für die Abhängigkeitsstrukturen von Unabhängigkeit bis zur Autoregression unterschiedliche Vorgaben gemacht werden. Je nach Wahl der Abhängigkeitsstruktur kann  $\alpha$  durch einen konsistenten momentbasierten Schätzer  $\hat{\alpha}(\beta, \hat{\phi}(\beta))$  geschätzt werden, wobei  $\hat{\phi}(\beta)$  ein konsistenter Schätzer für  $\phi$  ist.<sup>3</sup> Die Gleichung (3.13) lässt sich somit schreiben als

$$S(\beta) = \sum_{i=1}^m \left( \frac{\partial \mu_i}{\partial \beta} \right)^\top \left( \hat{V}_i \right)^{-1} (y_i - \mu_i)$$

$$\stackrel{\text{Poisson}}{=} \sum_{i=1}^m \left( \frac{\partial \lambda_i}{\partial \beta} \right)^\top \left( \hat{V}_i^\phi \right)^{-1} (y_i - \lambda_i)$$

$$= 0_q, \quad (3.15)$$

mit

$$\hat{V}_i := A_i^{1/2}(\beta) R_i(\hat{\alpha}[\beta, \hat{\phi}(\beta)]) A_i^{1/2}(\beta),$$

bzw.

$$\hat{V}_i^\phi := A_i^{1/2}(\beta, \hat{\phi}(\beta)) R_i(\hat{\alpha}[\beta, \hat{\phi}(\beta)]) A_i^{1/2}(\beta, \hat{\phi}(\beta)).$$

<sup>3</sup>Konsistente Schätzer für  $\alpha$  und  $\phi$  werden in Abschnitt 3.1.1.3 vorgestellt.

Somit ergibt sich eine Gleichung, die nur noch vom Parameter  $\beta$  abhängt. Sie wird von Liang und Zeger (1986) *generalized estimating equation* (Verallgemeinerte Schätzgleichung, GEE) genannt und zur Berechnung des Schätzers von  $\beta$  verwendet. Eine Lösung dieser Gleichung und damit ein Schätzer  $\hat{\beta}$  kann mit Hilfe einer leicht veränderten Form der *Fisher-Scoring-Methode* bestimmt werden.

### 3.1.1.2 Fisher-Scoring-Methode

In diesem Abschnitt wird das Fisher-Scoring-Verfahren vorgestellt, da dieses in leicht veränderter Form genutzt wird, um den Regressionsvektor  $\beta$  zu schätzen. Die Legitimation dieser Vorgehensweise kann allerdings erst im Abschnitt 3.1.1.5 geliefert werden.

Das Fisher-Scoring-Verfahren ist eine iterative Möglichkeit zur Bestimmung des Regressionsvektors  $\beta$ . Das Verfahren ähnelt dem Newton-Raphson-Verfahren. Auch in diesem Fall müssen die erste und zweite Ableitung der zu maximierenden Loglikelihood-Funktion ermittelt werden. Zur Bestimmung des Startwerts  $\beta_0$  der Iteration wird vorausgesetzt, dass die Beobachtungen unabhängig sind. Die Loglikelihood-Funktion ist in diesem Fall nach Gleichung (3.5) gegeben durch

$$l_y(\beta) = \sum_{i=1}^m \sum_{j=1}^{n_i} [y_{ij} \theta_{ij} - \psi(\theta_{ij})] \zeta^{-1} + c(y_{ij}, \zeta),$$

wobei  $y = (y_{11}, \dots, y_{mn_m})^\top$  ist. Der Gradient von  $l_y(\beta)$  – und damit die Transponierte der ersten Ableitung – wurde bereits berechnet (vgl. Gleichung (3.11))

$$S_1(\beta) = \left( \frac{\partial l_y(\beta)}{\partial \beta} \right)^\top = \sum_{i=1}^m \left( \frac{\partial \mu_i}{\partial \beta} \right)^\top \left( (A_i^\phi)^{1/2} I_{n_i} (A_i^\phi)^{1/2} \right)^{-1} (y_i - \mu_i),$$

mit  $y_i = (y_{i1}, \dots, y_{in_i})^\top$  und  $\mu_i = (\mu_{i1}, \dots, \mu_{in_i})^\top$ . Der Fall für Daten ohne Überdispersion folgt mit  $\phi = 1$ .

Um die Score-Gleichung zu differenzieren, wird es hilfreich sein, folgende Eigenschaften auszunutzen:

- (i) Es gilt  $h(\mu_{ij}) = x_{ij}^\top \beta$  und damit  $\mu_{ij} = h^{-1}(x_{ij}^\top \beta)$  mit einer stetigen, streng monotonen und differenzierbaren Linkfunktion  $h: \mathbb{R} \rightarrow \mathbb{R}$ . Somit ist  $h'(\mu_{ij}) \neq 0$  für alle  $i \in \{1, \dots, m\}$  und  $j \in \{1, \dots, n_i\}$  und

$$(h^{-1})'(x_{ij}^\top \beta) = \frac{1}{h'(h^{-1}(x_{ij}^\top \beta))} = \frac{1}{h'(\mu_{ij})}.$$

- (ii) Es gilt  $E(Y_{ij}) = \psi'(\theta_{ij}) = \mu_{ij}$  und  $\text{Var}(Y_{ij}) = \psi''(\theta_{ij}) \phi = v(\mu_{ij}) \phi$ .

Für die Score-Gleichung und damit für den Gradienten der Loglikelihood-Funktion folgt mit Gleichung (3.11):

$$\begin{aligned}
 S_1(\beta) &= \sum_{i=1}^m \left( \frac{\partial \mu_i}{\partial \beta} \right)^\top \left( (A_i^\phi)^{1/2} I_{n_i} (A_i^\phi)^{1/2} \right)^{-1} (y_i - \mu_i) \\
 &= \sum_{i=1}^m \sum_{j=1}^{n_i} \left( \frac{\partial \mu_{ij}}{\partial \beta} \right)^\top \left( \phi \psi''(\theta_{ij}) \right)^{-1} (y_{ij} - \psi'(\theta_{ij})) \\
 &= \sum_{i=1}^m \sum_{j=1}^{n_i} \left( \frac{\partial h^{-1}(x_{ij}^\top \beta)}{\partial \beta} \right)^\top (v(\mu_{ij}) \phi)^{-1} (y_{ij} - \mu_{ij}) \\
 &\stackrel{\text{Kettenregel}}{=} \sum_{i=1}^m \sum_{j=1}^{n_i} \frac{x_{ij}}{\phi} (h^{-1})' (x_{ij}^\top \beta) (v(\mu_{ij}))^{-1} (y_{ij} - \mu_{ij}) \\
 &\stackrel{(i)}{=} \sum_{i=1}^m \sum_{j=1}^{n_i} \frac{x_{ij}}{\phi} \frac{(y_{ij} - \mu_{ij})}{h' (h^{-1}(x_{ij}^\top \beta)) v(\mu_{ij})} \\
 &= \sum_{i=1}^m \sum_{j=1}^{n_i} \frac{x_{ij}}{\phi} \frac{(y_{ij} - \mu_{ij})}{h'(\mu_{ij}) v(\mu_{ij})}. \tag{3.16}
 \end{aligned}$$

Für die zweite Ableitung nach  $\beta$  und somit für die Hessematrix  $H_y(\beta)$  ergibt sich unter Verwendung der Quotienten- und Kettenregel folgendes

$$\begin{aligned}
 \frac{\partial S_1(\beta)}{\beta} &= \sum_{i=1}^m \sum_{j=1}^{n_i} x_{ij} \phi^{-1} \frac{-(h^{-1})' (x_{ij}^\top \beta) x_{ij}^\top \cdot h'(\mu_{ij}) v(\mu_{ij})}{(h'(\mu_{ij}))^2 v^2(\mu_{ij})} - (y_{ij} - \mu_{ij}) \\
 &\quad \times \frac{h''(\mu_{ij}) (h^{-1})' (x_{ij}^\top \beta) x_{ij}^\top v(\mu_{ij}) + h'(\mu_{ij}) v'(\mu_{ij}) (h^{-1})' (x_{ij}^\top \beta) x_{ij}^\top}{(h'(\mu_{ij}))^2 v^2(\mu_{ij})} \\
 &\stackrel{(i)}{=} \sum_{i=1}^m \sum_{j=1}^{n_i} x_{ij} \phi^{-1} \frac{-x_{ij}^\top v(\mu_{ij})}{(h'(\mu_{ij}))^2 v^2(\mu_{ij})} - (y_{ij} - \mu_{ij}) \\
 &\quad \times \frac{h''(\mu_{ij}) (h'(\mu_{ij}))^{-1} x_{ij}^\top v(\mu_{ij}) + v'(\mu_{ij}) x_{ij}^\top}{(h'(\mu_{ij}))^2 v^2(\mu_{ij})} \\
 &= \sum_{i=1}^m \sum_{j=1}^{n_i} x_{ij} \phi^{-1} \frac{-x_{ij}^\top}{(h'(\mu_{ij}))^2 v(\mu_{ij})} - (y_{ij} - \mu_{ij}) \\
 &\quad \times \frac{h''(\mu_{ij}) x_{ij}^\top v(\mu_{ij}) + v'(\mu_{ij}) h'(\mu_{ij}) x_{ij}^\top}{(h'(\mu_{ij}))^3 v^2(\mu_{ij})} \\
 &=: H_y^1(\beta). \tag{3.17}
 \end{aligned}$$

Das Newton-Raphson-Verfahren zur iterativen Bestimmung des Startwerts  $\beta_0$  basiert sowohl auf dem Gradienten der Loglikelihood-Funktion als auch auf deren Hessematrix und ist definiert durch

$$\beta_0^{n+1} = \beta_0^n - (H_y^1)^{-1} (\beta_0^n) S_1(\beta_0^n).$$

Das Fisher-Scoring-Verfahren unterscheidet sich insofern von dem Newton-Raphson-Verfahren, als dass nicht die Hessematrix, sondern der Erwartungswert der Hessematrix in der Iteration verwendet wird, d.h.

$$\beta_0^{n+1} = \beta_0^n - (\mathbb{E} [H_Y^1(\beta_0^n)])^{-1} S_1(\beta_0^n).$$

In diesem speziellen Fall ergibt sich für den Erwartungswert der Hessematrix

$$\begin{aligned} \mathbb{E} (H_Y^1(\beta)) &= \mathbb{E} \left[ \sum_{i=1}^m \sum_{j=1}^{n_i} x_{ij} \phi^{-1} \frac{-x_{ij}^\top}{(h'(\mu_{ij}))^2 v(\mu_{ij})} - (Y_{ij} - \mu_{ij}) \right. \\ &\quad \left. \times \frac{h''(\mu_{ij}) x_{ij}^\top v(\mu_{ij}) + v'(\mu_{ij}) h'(\mu_{ij}) x_{ij}^\top}{(h'(\mu_{ij}))^3 v^2(\mu_{ij})} \right] \\ &= \sum_{i=1}^m \sum_{j=1}^{n_i} x_{ij} \phi^{-1} \frac{-x_{ij}^\top}{(h'(\mu_{ij}))^2 v(\mu_{ij})} - \underbrace{\mathbb{E} [Y_{ij} - \mu_{ij}]}_{=0} \\ &\quad \times \frac{h''(\mu_{ij}) x_{ij}^\top v(\mu_{ij}) + v'(\mu_{ij}) h'(\mu_{ij}) x_{ij}^\top}{(h'(\mu_{ij}))^3 v^2(\mu_{ij})} \\ &= \sum_{i=1}^m \sum_{j=1}^{n_i} x_{ij} \phi^{-1} \frac{-1}{(h'(\mu_{ij}))^2 v(\mu_{ij})} x_{ij}^\top \\ &= - \sum_{i=1}^m \sum_{j=1}^{n_i} x_{ij} \frac{1}{h'(\mu_{ij})} (v(\mu_{ij}) \phi)^{-1} \frac{1}{h'(\mu_{ij})} x_{ij}^\top \\ &\stackrel{(i)}{=} - \sum_{i=1}^m \sum_{j=1}^{n_i} \left( (h^{-1})' (x_{ij}^\top \beta) x_{ij}^\top \right)^\top (v(\mu_{ij}) \phi)^{-1} \left( (h^{-1})' (x_{ij}^\top \beta) x_{ij}^\top \right) \\ &\stackrel{(ii)}{=} - \sum_{i=1}^m \sum_{j=1}^{n_i} \left( \frac{\partial \mu_{ij}}{\partial \beta} \right)^\top (\psi''(\theta_{ij}) \phi)^{-1} \left( \frac{\partial \mu_{ij}}{\partial \beta} \right) \\ &= - \sum_{i=1}^m \left( \frac{\partial \mu_i}{\partial \beta} \right)^\top \left( (A_i^\phi)^{1/2} I_{n_i} (A_i^\phi)^{1/2} \right)^{-1} \left( \frac{\partial \mu_i}{\partial \beta} \right). \end{aligned}$$

Somit ist ein iteratives Verfahren zur Bestimmung des Startwerts  $\beta_0$  gegeben durch

$$\begin{aligned} \beta_0^{n+1} &= \beta_0^n + \left[ \sum_{i=1}^m \left( \frac{\partial \mu_i}{\partial \beta} \right)^\top \left( (A_i^\phi)^{1/2} I_{n_i} (A_i^\phi)^{1/2} \right)^{-1} \left( \frac{\partial \mu_i}{\partial \beta} \right) \right]^{-1} \\ &\quad \times \sum_{i=1}^m \left( \frac{\partial \mu_i}{\partial \beta} \right)^\top \left( (A_i^\phi)^{1/2} I_{n_i} (A_i^\phi)^{1/2} \right)^{-1} (y_i - \mu_i). \quad (3.18) \end{aligned}$$

Nun bleibt zu diskutieren, wie ein Schätzer für den Regressionsvektor  $\beta$  im Fall korrelierter Daten bestimmt werden kann. Zu diesem Zweck wird nach dem Verfahren von Liang und Zeger (1986) eine Score-Gleichung unter Einführung einer Arbeitskorrelationsmatrix  $R_i$  aufgestellt. Soll eine eventuell vorliegende Überdispersion

berücksichtigt werden, verändert sich die Score-Gleichung nach Gleichung (3.15) zu

$$S(\beta) = \sum_{i=1}^m \left( \frac{\partial \mu_i}{\partial \beta} \right)^\top (V_i^\phi)^{-1} (y_i - \mu_i).$$

Auch im Fall korrelierter Daten kann eine approximative Methode, basierend auf dem Fisher-Scoring-Verfahren, zur Bestimmung von  $\beta$  genutzt werden. Allerdings wird die Score-Gleichung so verändert, dass  $V_i^\phi$  unabhängig von  $\beta$  ist. Dies wird erreicht, indem  $\beta$  durch  $\beta_0$ , und  $\alpha$  und  $\phi$  durch konsistente Schätzer  $\alpha_0 := \alpha(\beta_0, \phi_0)$  bzw.  $\phi_0 := \phi(\beta_0)$  ersetzt werden, wobei  $\beta_0$  durch das Iterationsverfahren in Gleichung (3.18) bestimmt wird. Somit ergibt sich

$$\hat{V}_i^{\phi_0} = A_i^{1/2} \left( \beta_0, \hat{\phi}_0 \right) R_i(\hat{\alpha}_0) A_i^{1/2} \left( \beta_0, \hat{\phi}_0 \right). \quad (3.19)$$

Dabei sind  $\hat{\alpha}_0$  und  $\hat{\phi}_0$  als konsistente Schätzer für  $\alpha$  bzw.  $\phi$  unabhängig von  $\beta$ . Für die Ermittlung der Schätzer  $\hat{\alpha}_0$  und  $\hat{\phi}_0$  sei auf Abschnitt 3.1.1.3 verwiesen.

Die veränderte Score-Gleichung ist somit durch

$$S_2(\beta) = \sum_{i=1}^m \left( \frac{\partial \mu_i}{\partial \beta} \right)^\top \left( \hat{V}_i^{\phi_0} \right)^{-1} (y_i - \mu_i) \quad (3.20)$$

gegeben. Um ein Iterationsverfahren in Anlehnung an das Fisher-Scoring-Verfahren zu ermitteln, wird die erste Ableitung der Score-Funktion in Gleichung (3.20) bestimmt. Es muss dabei lediglich beachtet werden, dass  $\hat{V}_i^{\phi_0}$  unabhängig von  $\beta$  ist. Mit der verkürzten Schreibweise  $\hat{V}_i := \hat{V}_i^{\phi_0}$  folgt

$$\begin{aligned} \frac{\partial S_2(\beta)}{\partial \beta} &= \frac{\partial}{\partial \beta} \left[ \sum_{i=1}^m \left( \frac{\partial \mu_i}{\partial \beta} \right)^\top \hat{V}_i^{-1} y_i \right] - \frac{\partial}{\partial \beta} \left[ \sum_{i=1}^m \left( \frac{\partial \mu_i}{\partial \beta} \right)^\top \hat{V}_i^{-1} \mu_i \right] \\ &\stackrel{\text{Produktregel}}{=} \sum_{i=1}^m \frac{\partial}{\partial \beta} \left( \frac{\partial \mu_i}{\partial \beta} \right)^\top \hat{V}_i^{-1} y_i - \left[ \sum_{i=1}^m \frac{\partial}{\partial \beta} \left( \frac{\partial \mu_i}{\partial \beta} \right)^\top \hat{V}_i^{-1} \mu_i \right. \\ &\quad \left. + \sum_{i=1}^m \left( \frac{\partial \mu_i}{\partial \beta} \right)^\top \hat{V}_i^{-1} \left( \frac{\partial \mu_i}{\partial \beta} \right) \right] \\ &= \sum_{i=1}^m \frac{\partial}{\partial \beta} \left( \frac{\partial \mu_i}{\partial \beta} \right)^\top \hat{V}_i^{-1} (y_i - \mu_i) - \sum_{i=1}^m \left( \frac{\partial \mu_i}{\partial \beta} \right)^\top \hat{V}_i^{-1} \left( \frac{\partial \mu_i}{\partial \beta} \right) \\ &=: H_y^2(\beta). \end{aligned}$$

Analog zum Vorgehen bei unabhängigen Daten muss für die Iterationsvorschrift der Erwartungswert der zweiten Ableitung berechnet werden:

$$\begin{aligned} \mathbb{E} [H_Y^2(\beta)] &= \sum_{i=1}^m \frac{\partial}{\partial \beta} \left( \frac{\partial \mu_i}{\partial \beta} \right)^\top \hat{V}_i^{-1} \underbrace{\mathbb{E} [Y_i - \mu_i]}_0 - \sum_{i=1}^m \left( \frac{\partial \mu_i}{\partial \beta} \right)^\top \hat{V}_i^{-1} \left( \frac{\partial \mu_i}{\partial \beta} \right) \\ &= - \sum_{i=1}^m \left( \frac{\partial \mu_i}{\partial \beta} \right)^\top \hat{V}_i^{-1} \left( \frac{\partial \mu_i}{\partial \beta} \right). \end{aligned}$$

Basierend auf dieser Größe ergibt sich dann die Iterationsvorschrift

$$\begin{aligned}\beta^{n+1} &= \beta^n - (\mathbb{E}[H_Y^2(\beta^n)])^{-1} S_2(\beta^n) \\ &= \beta^n + \left[ \sum_{i=1}^m \left( \frac{\partial \mu_i}{\partial \beta} \right)^\top (\hat{V}_i^{\phi_0})^{-1} \left( \frac{\partial \mu_i}{\partial \beta} \right) \right]^{-1} \sum_{i=1}^m \left( \frac{\partial \mu_i}{\partial \beta} \right)^\top (\hat{V}_i^{\phi_0})^{-1} (y_i - \mu_i).\end{aligned}\quad (3.21)$$

Es sei darauf hingewiesen, dass diese Vorgehensweise nicht vollständig dem Fisher-Scoring-Verfahren entspricht, da nicht  $V_i$ , sondern die geschätzte Matrix  $\hat{V}_i^{\phi_0}$  für die Herleitung benutzt wurde.

Zusammenfassend lässt sich das Iterationsverfahren zur Bestimmung des Regressionsparameters  $\beta$  bei Wahl einer Arbeitskorrelationsstruktur somit folgendermaßen darstellen:

1. Bestimme einen Anfangswert  $\beta_0$  für  $\beta$  unter der Annahme, dass die einzelnen Beobachtungen unabhängig sind. Die Iterationsvorschrift ist dabei nach Gleichung (3.18) durch

$$\hat{\beta}_0^{n+1} = \hat{\beta}_0^n + \left[ \sum_{i=1}^m \left( \frac{\partial \mu_i}{\partial \beta} \right)^\top (A_i^\phi)^{-1} \left( \frac{\partial \mu_i}{\partial \beta} \right) \right]^{-1} \sum_{i=1}^m \left( \frac{\partial \mu_i}{\partial \beta} \right)^\top (A_i^\phi)^{-1} (y_i - \mu_i)$$

gegeben.

2. Ausgehend von dem ermittelten Anfangswert  $\beta_0$  kann ein konsistenter Schätzer  $\hat{\phi}_0$  für  $\phi$  berechnet werden (vgl. Abschnitt 3.1.1.3, Definition 3.1). Bedingt durch die gewählte Arbeitskorrelationsstruktur kann ebenfalls ein konsistenter Schätzer  $\hat{\alpha}_0$  für  $\alpha$  berechnet werden (vgl. Abschnitt 3.1.1.3, Tabelle 3.1).
3. Basierend auf  $\beta_0$ ,  $\hat{\alpha}_0$  und  $\hat{\phi}_0$  kann  $R_i(\hat{\alpha}_0)$  und mit Gleichung (3.19) auch  $\hat{V}_i^{\phi_0}$  durch

$$\hat{V}_i^{\phi_0} = A_i^{1/2} \left( \beta_0, \hat{\phi}_0 \right) R_i(\hat{\alpha}_0) A_i^{1/2} \left( \beta_0, \hat{\phi}_0 \right)$$

bestimmt werden.

4. Mit Hilfe der Schätzungen für  $V_i^\phi$  und dem Startwert  $\beta_0$  kann das Fisher-Scoring-Verfahren gemäß Vorschrift (3.21) durchgeführt werden.

Damit ergibt sich ein neuer Schätzer  $\beta_1$  für  $\beta$ . Ausgehend von diesem werden die Schritte zwei, drei und vier erneut durchgeführt. Dieses Vorgehen wird iterativ fortgesetzt, bis das Verfahren mit einer vorgegebenen Genauigkeit konvergiert.

### 3.1.1.3 Schätzer für $\alpha$ und $\phi$

Ein konsistenter Schätzer für  $\phi$  kann mit Hilfe der momentbasierten verallgemeinerten Pearson-Statistik ermittelt werden.

#### Definition 3.1.

Die Pearson-Statistik ist mit  $\sum_{i=1}^m n_i =: N$  und  $\beta \in \mathbb{R}^q$  definiert als

$$\hat{\phi}(\beta) = \frac{1}{N - q} \sum_{i=1}^m \sum_{j=1}^{n_i} (e_{ij})^2, \quad (3.22)$$

wobei  $e_{ij}$  das Pearson-Residuum ist, welches mit  $v(\mu_{ij}) = \psi''(\theta_{ij})$  und  $\mu_{ij} = \mu_{ij}(X_i, \beta)$  durch

$$e_{ij} = \frac{y_{ij} - \mu_{ij}}{\sqrt{v(\mu_{ij})}}$$

und im Fall von poissonverteilten Daten mit Überdispersion durch

$$e_{ij} = \frac{y_{ij} - \lambda_{ij}}{\sqrt{\phi \lambda_{ij}}}$$

gegeben ist.

Diese Statistik liefert unter gewissen Annahmen, die hier nicht weiter erläutert werden, stark konsistente Schätzer.

Für  $\alpha$  gibt es je nach genutzter Abhängigkeitsstruktur unterschiedliche konsistente Schätzer. Die gängigsten Strukturen, die auch von der Prozedur GENMOD verwendet werden, sind in der folgenden Tabelle 3.1 dargestellt.

Struktur	$Corr(Y_{ij}, Y_{ik}), j \neq k$	konsistente Schätzer
Unabhängigkeit	0	-
Austauschbarkeit	$\alpha$	$\hat{\alpha} = \frac{1}{m} \sum_{i=1}^m \frac{1}{n_i(n_i - 1)} \sum_{j \neq k} e_{ij} e_{ik}$
Autoregressive AR(1)	$\alpha^{ j-k }$	$\hat{\alpha} = \frac{1}{m} \sum_{i=1}^m \frac{1}{n_i - 1} \sum_{j \leq n_i - 1} e_{ij} e_{i,j+1}$
Unstrukturiert	$\alpha_{jk}$	$\hat{\alpha}_{jk} = \frac{1}{m} \sum_{i=1}^m e_{ij} e_{ik}$

Tabelle 3.1: Die gängigsten Schätzer für  $\alpha$  bei unterschiedlichen Abhängigkeitsstrukturen.

Es bleibt zu diskutieren, unter welchen Umständen welche Abhängigkeitsstruktur sinnvoll zu wählen ist und ob es Indikatoren für eine gute bzw. schlechte Wahl gibt. Es ist naheliegend, dass die Abhängigkeitsstruktur abhängig vom Datensatz und den Versuchsbedingungen gewählt werden sollte, da diese oftmals einen Aufschluss über die zugrundeliegende Struktur liefern.

Die einfachste Struktur ist durch die Unabhängigkeitsstruktur gegeben, doch sind die dazu ermittelten Schätzer im Allgemeinen ineffizient. Dies wurde von Liang und Zeger (1986) in Simulationsstudien gezeigt.

Liegt ein Datensatz vor, bei dem die wiederholten Daten eines Subjekts zeitunabhängig sind, und ist eine beliebige Permutation der wiederholten Daten zulässig, erscheint die Austauschbarkeitsstruktur (*exchangeable correlation structure*) sinnvoll.<sup>4</sup> Die Korrelationsmatrix ist in diesem Fall gegeben durch

$$[R_i(\alpha)_{jk}]_{\substack{k \in \{1, \dots, n_i\} \\ j \in \{1, \dots, n_i\}}} = \begin{cases} 1, & \text{falls } k = j, \\ \alpha, & \text{falls } k \neq j. \end{cases}$$

Ein Beispiel für diese Art von Daten ist eine Studie, in der die Subjekte bzw. Cluster Krankenhäuser, und die wiederholten Daten eines Clusters unterschiedliche Patienten innerhalb eines Krankenhauses sind.

In vielen Fällen ist die Annahme von zeitunabhängigen wiederholten Daten nicht erfüllt. Daher wird eine Abhängigkeitsstruktur benötigt, die der Zeitabhängigkeit der Daten Rechnung trägt. In einer Studie, bei der die wiederholten Daten in bestimmten Zeitabständen erhoben werden, ist die Annahme, dass näher zusammen liegende Beobachtungen stärker korreliert sind als weiter auseinander liegende, nachvollziehbar. Für einen solchen Fall empfehlen Hardin und Hilbe (2003) die autoregressive Abhängigkeitsstruktur AR(1), bei der die Korrelationsmatrix durch

$$[R_i(\alpha)_{jk}]_{\substack{k \in \{1, \dots, n_i\} \\ j \in \{1, \dots, n_i\}}} = \begin{cases} 1, & \text{falls } k = j, \\ \alpha^{|j-k|}, & \text{falls } k \neq j \end{cases}$$

gegeben ist.

Ein Gegenstück zu den bisher vorgestellten Abhängigkeitsstrukturen ist die unstrukturierte Korrelation, welche die allgemeinste und uneingeschränkste Struktur darstellt, da sie keine Vorgaben an die Beschaffenheit der Korrelationsmatrix stellt. Allerdings kann es bei dieser Art von Struktur numerische Probleme geben. Darüber hinaus ist die auf dieser Struktur basierende geschätzte Korrelationsmatrix in manchen Fällen singulär.<sup>5</sup> In diesem Zusammenhang wird daher die Nutzung der un-

---

<sup>4</sup>Vgl. Hardin und Hilbe (2003), S. 59.

<sup>5</sup>Dieses Problem wird im Abschnitt 5.3 auftreten.

strukturierten Korrelation nur für Datensätze ohne fehlende Daten und mit einer geringen Anzahl von Beobachtungen pro Subjekt bzw. Cluster empfohlen.<sup>6</sup>

Die letzte Korrelationsstruktur, der wir uns in diesem Abschnitt widmen werden, ist die feste Korrelationsstruktur. Diese kann dann angesetzt werden, wenn aufgrund von vorhergehenden Studien, Versuchen oder sonstigen Informationen die zugrundeliegende Korrelation bekannt ist. In diesem Fall entfällt im vorgestellten Iterationsverfahren das Schätzen der Korrelationsmatrix.

Selbstverständlich gibt es viele andere Korrelationsstrukturen, die hier nicht vorgestellt wurden. Je nach Datensatz und Studiendesign muss abgewägt werden, welche Korrelationsstruktur sinnvoll ist. Die gängigsten Strukturen sind in Prozeduren wie GENMOD vordefiniert, doch werden diese bei vielen Anwendungsproblemen nicht ausreichen, so dass eigene Korrelationsstrukturen definiert werden müssen.

Die Frage nach der Güte der Wahl kann in diesem Abschnitt noch nicht diskutiert, soll aber im nächsten beantwortet werden.

#### 3.1.1.4 Eigenschaften des GEE-Schätzers

Der durch die Gleichung (3.21) ermittelte Schätzer  $\hat{\beta}$  für den Regressionsvektor  $\beta$  ist nach Liang und Zeger (1986) unabhängig von der verwendeten Abhängigkeitsstruktur, und unter gewissen Regularitätsvoraussetzungen sowie einer korrekten Spezifikation der Mittelwertstruktur  $\mu_i = \mu(X_i, \beta)$  konsistent. Darüber hinaus gilt:

#### Satz 3.2.<sup>7</sup>

*Unter gewissen Regularitätsbedingungen und den folgenden Voraussetzungen*

- (i)  $\hat{\alpha}(\beta, \phi(\beta))$  ist für gegebenes  $\beta$  und  $\phi$  ein  $m^{1/2}$ -konsistenter Schätzer für  $\alpha$ ,
- (ii)  $\hat{\phi}(\beta)$  ist bei gegebenem  $\beta$  ein  $m^{1/2}$ -konsistenter Schätzer für  $\phi$  und
- (iii)  $\left| \frac{\partial \hat{\alpha}(\beta, \phi(\beta))}{\partial \phi} \right| \leq H(Y, \beta)$  mit  $H(Y, \beta) \in O_p(1)$ ,

*ist  $m^{1/2}(\hat{\beta} - \beta)$  asymptotisch normalverteilt mit Erwartungswert  $0_q$  und asymptotischer Kovarianzmatrix*

$$\text{Cov}(\beta) = I_0^{-1} I_1 I_0^{-1}, \tag{3.23}$$

<sup>6</sup>Vgl. Hardin und Hilbe (2003), S. 141.

<sup>7</sup>Vgl. Liang und Zeger (1986), S. 16.

mit

$$\begin{aligned}
 I_0 &= \sum_{i=1}^m \left( \frac{\partial \mu_i}{\partial \beta} \right)^\top V_i^{-1} \frac{\partial \mu_i}{\partial \beta} \quad \text{und} \\
 I_1 &= \sum_{i=1}^m \left( \frac{\partial \mu_i}{\partial \beta} \right)^\top V_i^{-1} \text{Cov}(Y_i) V_i^{-1} \frac{\partial \mu_i}{\partial \beta}.
 \end{aligned}$$

Ein konsistenter Schätzer für  $\text{Cov}(\hat{\beta})$  kann ermittelt werden, indem die Größen  $\beta$ ,  $\alpha$  und  $\phi$  durch die entsprechenden konsistenten Schätzer ersetzt werden.<sup>8</sup>

Darüber hinaus muss  $\text{Cov}(Y_i)$  geschätzt werden. Ein erwartungstreuer Schätzer ist im Fall der richtigen Spezifikation der Erwartungswertstruktur durch

$$(Y_i - \mu_i(X_i, \hat{\beta}))(Y_i - \mu_i(X_i, \hat{\beta}))^\top$$

gegeben. Allerdings besitzt dieser Schätzer den Rang 1, so dass diese Wahl bedenklich erscheint. Da die Berechnung einer Inversen im Laufe der Analyse nicht notwendig ist, wird nach Molenberghs und Verbeke (2005) dieser Schätzer trotz dieser Bedenken benutzt.

Wird die Korrelation von  $Y_i$  durch  $R(\alpha)$  richtig bestimmt, so gilt  $\text{Cov}(Y_i) = V_i$ ,  $\hat{I}_0$  entspricht der Matrix  $\hat{I}_1$  und demzufolge gilt  $\text{Cov}(\hat{\beta}) = \hat{I}_0^{-1}$ . Aus diesem Grund wird bei  $\hat{I}_0^{-1}$  von dem *modellbasierten* und bei  $\hat{I}_0^{-1} \hat{I}_1 \hat{I}_0^{-1}$  von dem *empirischen* Schätzer für die Kovarianzmatrix  $\text{Cov}(\hat{\beta})$  gesprochen. Als Schätzer für die Kovarianzmatrix  $\text{Cov}(\hat{\beta})$  sollte allerdings stets der empirische Schätzer gewählt werden, weil dieser ein genaueres Schätzergebnis liefert als der modellbasierte Schätzer.

Da die Distanz zwischen dem empirischen Schätzer und dem modellbasierten Schätzer in Abhängigkeit von der Distanz zwischen der wahren Korrelationsmatrix  $\text{Corr}(Y_i)$  und der Arbeitskorrelationsmatrix  $R(\alpha)$  größer bzw. kleiner wird, sollte der modellbasierte Schätzer lediglich als Indikator für die Güte der Wahl einer Korrelationsstruktur und damit der Matrix  $R(\alpha)$  betrachtet werden. Ein großer Unterschied zwischen der modellbasierten und empirischen Schätzung für  $\text{Cov}(\hat{\beta})$  kann daher auf einer schlechten Wahl der Korrelationsstruktur beruhen. Es ist allerdings zu beachten, dass man jede beliebige Abhängigkeitsstruktur definieren und schätzen kann, denn selbst eine unpassende Abhängigkeitsstruktur führt zu keinem inkonsistenten Ergebnis; sie kann lediglich eine Einschränkung der Effizienz bedingen.<sup>9</sup> Somit kann mit Hilfe der modellbasierten und empirischen Schätzer für  $\text{Cov}(\hat{\beta})$  die schuldig gebliebene Antwort auf die Frage nach einem Indikator für die Güte der Wahl einer

---

<sup>8</sup>Vgl. Molenberghs und Verbeke (2005), S. 155.

<sup>9</sup>Vgl. Molenberghs und Verbeke (2005), S. 155f.

Korrelationsstruktur nachgereicht werden (vgl. Abschnitt 3.1.1.3).

Beim Schätzen mit Hilfe der Verallgemeinerten Schätzgleichungen sind also Annahmen über die Abhängigkeitsstruktur erforderlich. Sollen solche Annahmen vermieden werden, oder ist die Abhängigkeitsstruktur selbst von Interesse, dann ist die GEE1-Methode nicht anwendbar. Alternativ muss dann die GEE2-Methode benutzt werden.<sup>10</sup>

Beispielsweise ist die Abhängigkeitsstruktur für eine Studie in der Arbeit von Diepgen und Bleittner (1996) von Interesse. Bei dieser Studie geht es darum, die Bedingtheit des Vorkommens einer Hautkrankheit durch Verwandtschaft und Umwelt zu untersuchen. Zu diesem Zweck wurde eine Fall-Kontroll-Studie durchgeführt, in der 426 erkrankte Patienten und eine Kontrollgruppe von 628 Patienten untersucht wurden. Ziel der Studie war es, zu überprüfen, ob ein signifikanter Zusammenhang zwischen Verwandtschaft und Erkrankung vorliegt. Daher wurden insgesamt etwa 5000 Familienmitglieder untersucht. Die abhängige Variable  $Y_i$  ist in diesem Fall ein binärer Vektor, der angibt ob die Hautkrankheit bei  $n_i$  Verwandten vorliegt, oder nicht und die interessierende Größe ist die Korrelation zwischen dem Verwandtschaftsgrad und der Erkrankung.

Die grundlegende Idee der Verallgemeinerten Schätzgleichungen (*generalized estimating equations*) besteht, wie gesagt, in dem Ersetzen der Einheitsmatrix in Gleichung (3.11) durch eine Arbeitskorrelationsmatrix. So naiv und einfach diese Methode auch wirken mag, ihre Resultate bezüglich der Konsistenz und der asymptotischen Normalverteilung sind es indes nicht. So finden Hardin und Hilbe (2003) die anerkennenden Worte:

*„However, that is not to say that the proof is simple, or that our appreciation of the results is any way lessened. Indeed, the Liang and Zeger paper is an impressive presentation of sophisticated asymptotic and distributional statistics.“<sup>11</sup>*

Jedoch musste die Autorin feststellen, dass der genannte Beweis zur Konsistenz des geschätzten Regressionsvektors weder in Büchern, die lediglich von Verallgemeinerten Schätzgleichungen handeln, noch in theoretischen Artikeln zu finden ist. Selbst im vielzitierten Artikel *„Longitudinal data analysis using generalized linear models“*

---

<sup>10</sup>Vgl. Molenberghs und Verbeke (2005), S. 164 ff und Hardin und Hilbe (2003), S. 104ff.

<sup>11</sup> Hardin und Hilbe (2003), S. 58

von Liang und Zeger (1986) ist der Beweis nicht enthalten.<sup>12</sup> Allerdings kann der Beweis der Eigenschaften der GEE-Schätzer auch basierend auf einer alternativen Methode zur Herleitung der Verallgemeinerten Schätzgleichungen geführt werden. Die Vorstellung dieser alternativen Methode wird daher Gegenstand des nächsten Abschnitts sein.

### 3.1.1.5 Pseudo-Maximum-Likelihood-Methode und Quasi-Generalisierte-Pseudo-Maximum-Likelihood-Methode

Wie in den vorangegangenen Abschnitten erläutert, ist die Bestimmung des zugrundeliegenden Regressionsvektors  $\beta \in \Theta$  im Fall korrelierter Daten komplizierter als im unabhängigen Fall. Die Dichte von  $Y_i = (Y_{i1}, \dots, Y_{in_i})^\top$  lässt sich aufgrund der Abhängigkeitsstruktur der Daten nicht ohne Weiteres berechnen. Eine falsche Spezifikation der Dichte oder der Varianz ist zudem eine grobe Verletzung der zentralen Voraussetzung der Maximum-Likelihood-Methode und führt zu inkonsistenten Schätzungen der Erwartungswertstruktur.

Daher wird in diesem Abschnitt ein Verfahren vorgestellt, bei dem ausschließlich die korrekte Spezifikation der Mittelwertstruktur benötigt wird, um konsistente und asymptotisch normalverteilte Schätzer für  $\beta$  zu erhalten. Dieses Verfahren basiert sowohl auf der *Pseudo-Maximum-Likelihood-Methode* (PML-Methode), als auch auf der *Quasi-Generalisierten-PML-Methode* (QGPML-Methode). Beide Methoden wurden von Gourieroux u. a. (1984) vorgestellt.

Es wird gezeigt, dass mit Hilfe der PML- und der QGPML-Methode die Schätzgleichungen von Liang und Zeger hergeleitet werden können, doch werden zunächst die beiden Methoden vorgestellt.

#### 3.1.1.5.1 PML-Methode

Bevor die PML-Methode betrachtet werden kann, muss die lineare Exponentialfamilie definiert und Eigenschaften dieser Wahrscheinlichkeitsdichten aufgezeigt werden. Nach Gourieroux u. a. (1984) wird die lineare Exponentialfamilie definiert durch:

**Definition 3.3.**

*Seien  $y \in \mathbb{R}^p$ ,  $\Psi \in \mathbb{R}^{p \times p}$  und  $a$  und  $b$  Abbildungen, die den Raum  $\mathbb{R}^p \times \mathbb{R}^{p \times p}$  in den Raum  $\mathbb{R}$  abbilden. Darüber hinaus sei  $c$  eine Abbildung, die den Raum  $\mathbb{R}^p \times \mathbb{R}^{p \times p}$  in den Raum  $\mathbb{R}^p$  abbildet und es gelte  $E(Y) = \mu$ .*

*Eine Familie von Wahrscheinlichkeitsverteilungen auf  $\mathbb{R}^p$  heißt  $p$ -parametrische einfache lineare Exponentialfamilie, falls jedes Element dieser Familie eine Dichte der*

---

<sup>12</sup>Darüber hinaus sind Versuche, mit den Autoren Kontakt aufzunehmen, fehlgeschlagen.

Form

$$f(y, \mu, \Psi) = \exp\{c(\mu, \Psi)^\top y + a(\mu, \Psi) + b(y, \Psi)\}$$

besitzt.

**Beispiel 3.4. Eine multivariate lineare Exponentialfamilie**

Für  $y \in \mathbb{R}^p$ ,  $\mu \in \mathbb{R}^p$  und  $\Psi$  aus der Menge der positiv definiten, symmetrischen  $p \times p$  Matrizen ist die Dichte der multivariaten Normalverteilung  $\mathcal{N}_p(\mu, \Psi)$  gegeben durch

$$\begin{aligned} f(y) &= (2\pi)^{-p/2} \det(\Psi)^{-1/2} \exp\left\{-\frac{1}{2}(y - \mu)^\top \Psi^{-1} (y - \mu)\right\} \\ &= \exp\{c(\mu, \Psi)^\top y + a(\mu, \Psi) + b(y, \Psi)\}, \end{aligned}$$

mit  $a(\mu, \Psi) = -\frac{1}{2}\mu^\top \Psi^{-1} \mu$ ,  $b(y, \Psi) = -\frac{p}{2} \ln(2\pi) - \frac{1}{2}y^\top \Psi^{-1} y - \frac{1}{2} \ln(\det(\Psi))$  und  $c(\mu, \Psi) = \Psi^{-1} \mu$ .

Für die weiteren Betrachtungen werden die folgenden Eigenschaften dieser Verteilungsfamilie benötigt.

**Satz 3.5.**

Ist  $f(y, \mu, \Psi)$  die Dichte eines Mitglieds einer linearen Exponentialfamilie und  $\Psi$  fest vorgegeben, dann gilt

$$\frac{\partial a(\mu, \Psi)}{\partial \mu} + \mu^\top \frac{\partial c(\mu, \Psi)^\top}{\partial \mu} = 0_{1 \times p}.$$

*Beweis:* Da  $f(y, \mu, \Psi)$  eine Dichtefunktion ist, ergibt sich

$$\int f(y, \mu, \Psi) dy = \int \exp\{c(\mu, \Psi)^\top y + a(\mu, \Psi) + b(y, \Psi)\} dy = 1,$$

und das Differenzieren dieser Gleichung liefert mit der Kettenregel folgendes:

$$\int \left( \frac{\partial a(\mu, \Psi)}{\partial \mu} + y^\top \frac{\partial c(\mu, \Psi)^\top}{\partial \mu} \right) f(y, \mu, \Psi) dy = 0_{1 \times p}.$$

Die Berechnung des Integrals liefert

$$\frac{\partial a(\mu, \Psi)}{\partial \mu} + \mathbb{E}(Y)^\top \frac{\partial c(\mu, \Psi)^\top}{\partial \mu} = 0_{1 \times p}$$

und damit die Behauptung. □

**Satz 3.6.**

Bei fest vorgegebenem  $\Psi$  gilt

$$\frac{\partial}{\partial \mu} \left( \frac{\partial a(\mu, \Psi)}{\partial \mu} \right) + \frac{\partial}{\partial \mu} \left( \frac{\partial c(\mu, \Psi)^\top}{\partial \mu} \right) \mu + \frac{\partial c(\mu, \Psi)^\top}{\partial \mu} = 0_{p \times p}. \quad (3.24)$$

*Beweis:* Die Behauptung folgt unmittelbar aus der Differentiation der Gleichung in Satz 3.5, dem Anwenden der Produktregel und dem Satz von Schwarz<sup>13</sup>.  $\square$

**Satz 3.7.**

*Besitzt die Zufallsvariable  $Y$  die Verteilung einer linearen Exponentialfamilie und die Dichte  $f(y, \mu, \Psi)$  mit fest vorgegebenem Parameter  $\Psi$ , so ist die Kovarianzmatrix  $\Sigma$  dieser Zufallsvariable gegeben durch*

$$\Sigma^{-1} = \frac{\partial c(\mu, \Psi)^\top}{\partial \mu}.$$

*Beweis:* Es gilt

$$\int y f(y, \mu, \Psi) dy = E(Y) = \mu,$$

so dass die Differentiation nach  $\mu$  unter Verwendung der Produktregel die Gleichung

$$\int y \left( \frac{\partial a(\mu, \Psi)}{\partial \mu} + y^\top \frac{\partial c(\mu, \Psi)^\top}{\partial \mu} \right) f(y, \mu, \Psi) dy = I_{p \times p} \quad (3.25)$$

liefert. Die Gleichheit im Satz 3.5 lässt sich erweitern zu

$$\frac{\partial a(\mu, \Psi)}{\partial \mu} + y^\top \frac{\partial c(\mu, \Psi)^\top}{\partial \mu} = -\mu^\top \frac{\partial c(\mu, \Psi)^\top}{\partial \mu} + y^\top \frac{\partial c(\mu, \Psi)^\top}{\partial \mu}, \quad (3.26)$$

und das Einsetzen dieser Beziehung in Gleichung (3.25) ergibt

$$\begin{aligned} I_{p \times p} &= \int y \left( \frac{\partial a(\mu, \Psi)}{\partial \mu} + y^\top \frac{\partial c(\mu, \Psi)^\top}{\partial \mu} \right) f(y, \mu, \Psi) dy \\ &= \int y \left( -\mu^\top \frac{\partial c(\mu, \Psi)^\top}{\partial \mu} + y^\top \frac{\partial c(\mu, \Psi)^\top}{\partial \mu} \right) f(y, \mu, \Psi) dy \\ &= \int y (y - \mu)^\top \frac{\partial c(\mu, \Psi)^\top}{\partial \mu} f(y, \mu, \Psi) dy \\ &= \frac{\partial c(\mu, \Psi)^\top}{\partial \mu} \int y (y - \mu)^\top f(y, \mu, \Psi) dy. \end{aligned} \quad (3.27)$$

Darüber hinaus gilt

$$\begin{aligned} -\frac{\partial c(\mu, \Psi)^\top}{\partial \mu} \int \mu (y - \mu)^\top f(y, \mu, \Psi) dy &= \frac{\partial c(\mu, \Psi)^\top}{\partial \mu} \left( -\mu \int y^\top f(y, \mu, \Psi) dy \right. \\ &\quad \left. + \mu \mu^\top \int f(y, \mu, \Psi) dy \right) \\ &= \frac{\partial c(\mu, \Psi)^\top}{\partial \mu} (-\mu \mu^\top + \mu \mu^\top) \quad (3.28) \\ &= 0_{p \times p}. \end{aligned}$$

---

<sup>13</sup>Vgl. Schmieder (1994).

Insgesamt ergibt sich damit aus den Gleichungen (3.27) und (3.28) folgendes:

$$\begin{aligned} & \frac{\partial c(\mu, \Psi)^\top}{\partial \mu} \cdot \left( \int y (y - \mu)^\top f(y, \mu, \Psi) dy - \int \mu (y - \mu)^\top f(y, \mu, \Psi) dy \right) = I_{p \times p} \\ & \Leftrightarrow \frac{\partial c(\mu, \Psi)^\top}{\partial \mu} \int (y - \mu) (y - \mu)^\top f(y, \mu, \Psi) dy = I_{p \times p} \\ & \Leftrightarrow \frac{\partial c(\mu, \Psi)^\top}{\partial \mu} \Sigma = I_{p \times p} \\ & \Leftrightarrow \frac{\partial c(\mu, \Psi)^\top}{\partial \mu} = \Sigma^{-1} \end{aligned}$$

und somit die Behauptung. □

Um den Satz 3.10 beweisen zu können, wird zunächst die Ungleichung von Kullback und der Begriff der *Identifizierbarkeit* vorgestellt.

**Definition 3.8. (Identifizierbarkeit)**

*Eine Funktion*

$$\begin{aligned} f : \mathbb{R}^p \times \mathbb{R}^q & \rightarrow \mathbb{R} \\ (y, \mu) & \mapsto f(y, \mu) \end{aligned}$$

heißt *identifizierbar bezüglich des Parameters*  $\mu \in \mathbb{R}^q$ , wenn für alle  $\mu_1, \mu_2 \in \mathbb{R}^q$  folgendes gilt:

$$f(y, \mu_1) = f(y, \mu_2) \text{ für fast alle } y \in \mathbb{R}^p \Rightarrow \mu_1 = \mu_2.$$

*Ist bekannt, welche Funktion betrachtet wird, so wird auch von einem identifizierbaren  $\mu$  gesprochen.*

**Lemma 3.9. (Ungleichung von Kullback)**

*Seien  $h$  und  $g$  zwei nicht negative und bezüglich des Maßes  $\nu$  integrierbare Funktionen. Zudem sei  $S$  der Integrationsbereich, in dem  $h > 0$  gilt, und es gelte  $\int_S (h(x) - g(x)) d\nu(x) \geq 0$ . Dann gilt*

$$\int_S h(x) \ln \frac{h(x)}{g(x)} d\nu(x) \geq 0$$

*und die Gleichheit gilt genau dann fast sicher, wenn  $h \equiv g$  ist.*

*Beweis:* Für den Beweis siehe Rao (1973), S.59. □

Mit Hilfe dieses Satzes kann die folgende Eigenschaft gezeigt werden.

**Satz 3.10.**

Für ein beliebiges identifizierbares  $\mu \in \mathbb{R}^q$ , ein festes  $\mu^* \in \mathbb{R}^q$  und ein festes  $\Psi$  gilt

$$a(\mu, \Psi) + c(\mu, \Psi)^\top \mu^* \leq a(\mu^*, \Psi) + c(\mu^*, \Psi)^\top \mu^*.$$

Dabei gilt die Gleichheit genau dann fast sicher, wenn  $\mu = \mu^*$  ist.

*Beweis:*

Sei  $Y$  eine Zufallsvariable mit einer Verteilung aus einer linearen Exponentialfamilie und es gelte  $E(Y) = \mu^*$ . Dann ergibt sich

$$\begin{aligned} \int f(y, \mu^*, \Psi) - f(y, \mu, \Psi) dy &= \int f(y, \mu^*, \Psi) dy - \int f(y, \mu, \Psi) dy \\ &= 1 - 1 \\ &= 0, \end{aligned} \tag{3.29}$$

und  $f(y, \mu^*, \Psi) \geq 0$  für alle  $y$ . Die Ungleichung von Kullback liefert dann mit  $h(y) = f(y, \mu^*, \Psi)$  und  $g(y) = f(y, \mu, \Psi)$

$$\int \ln \left( \frac{f(y, \mu^*, \Psi)}{f(y, \mu, \Psi)} \right) f(y, \mu^*, \Psi) dy \geq 0$$

und folglich

$$\begin{aligned} &\int \ln (\exp\{c(\mu^*, \Psi)^\top y + a(\mu^*, \Psi) - c(\mu, \Psi)^\top y - a(\mu, \Psi)\}) f(y, \mu^*, \Psi) dy \\ &= \int \{c(\mu^*, \Psi)^\top y + a(\mu^*, \Psi) - c(\mu, \Psi)^\top y - a(\mu, \Psi)\} f(y, \mu^*, \Psi) dy \\ &\geq 0. \end{aligned}$$

Die Berechnung des Integrals und Umstellen der Ungleichung liefern

$$c(\mu^*, \Psi)^\top \mu^* + a(\mu^*, \Psi) \geq c(\mu, \Psi)^\top \mu^* + a(\mu, \Psi).$$

Nach Satz 3.9 gilt die Gleichheit genau dann fast sicher, wenn  $h(y) \equiv g(y)$  ist. Diese Gleichheit besteht aufgrund der Identifizierbarkeit für  $\mu_i \equiv \mu_i^*$ , so dass die Behauptung folgt.  $\square$

Mit Hilfe der gezeigten Eigenschaften kann jetzt die Pseudo-Maximum-Likelihood-Methode vorgestellt werden. Dazu müssen folgende Voraussetzungen getroffen werden:

- Für alle  $i \in \{1, \dots, m\}$  sind  $Y_i$   $n_i$ -dimensionale Zufallsvektoren und  $X_i$  die  $(n_i \times q)$ -dimensionalen Prädiktormatrizen.

- $f^w(y_i, \mu_i^w, \Psi^w)$  ist die wahre, aber unbekannte Dichte des Zufallsvektors  $Y_i$ .
- Die Komponenten des wahren Erwartungswertvektors sind über den wahren, aber unbekanntem Regressionsvektor  $\beta^*$  und der Prädiktormatrix  $X_i$  parametrisiert. Es gilt also  $\mu_i^w = \mu^w(X_i, \beta^*)$ .
- Darüber hinaus ist  $V_i^w(\beta^*)$  die Kovarianzmatrix von  $Y_i$ , die in der PML-Methode nicht näher spezifiziert wird.

Im Rahmen der PML-Methode werden nun folgende Annahmen getroffen:

- Die Verteilung von  $Y_i$  entstammt der linearen Exponentialfamilie mit fest vorgegebenem Nuisance-Parameter  $\Psi$  und der dazugehörige Erwartungswert ist durch  $\mu^a(X_i, \beta^*)$  gegeben. Die Dichte von  $Y_i$  wird mit  $f^a(y_i, \mu^a(X_i, \beta^*), \Psi)$  bezeichnet.
- Der Erwartungswert  $\mu^a(X_i, \beta^*)$  stimmt mit dem wahren Erwartungswert überein, d.h. es gilt

$$\mu^a(X_i, \beta^*) = \mu^w(X_i, \beta^*).$$

Das zugrundeliegende stochastische Modell für  $Y_i$  mit gegebenem Prädiktor  $X_i$  ist für die wahre Dichte  $f^w(y_i, \mu^w(X_i, \beta^*))$  somit bestimmt durch

$$\begin{aligned} \mathbb{E}(Y_i) &= \mu^w(X_i, \beta^*) \\ \mu^w(X_i, \beta^*) &= \mu^a(X_i, \beta^*) =: \mu_i^* \\ \text{Cov}(Y_i) &= V_i^w(\beta^*). \end{aligned}$$

Für die angenommene Dichte ergibt sich das Modell

$$\begin{aligned} \mathbb{E}(Y_i) &= \mu_i^a(X_i, \beta^*) \\ \mu^w(X_i, \beta^*) &= \mu^a(X_i, \beta^*) = \mu_i^* \\ \text{Cov}(Y_i) &= V_i^a(\beta^*). \end{aligned}$$

Das Schätzen des Regressionsparameters kann nun mit Hilfe der Loglikelihood-Methode vollzogen werden. Für  $i \in \{1, \dots, m\}$  lautet die auf der angenommenen Dichte basierende Loglikelihood-Funktion von  $Y = (Y_1^\top, \dots, Y_m^\top)^\top$  aufgrund der Unabhängigkeit von  $Y_1, \dots, Y_m$  folgendermaßen

$$\tilde{\ell}(\beta) = \sum_{i=1}^m \{a(\mu(X_i, \beta), \Psi) + b(y_i, \Psi) + c(\mu(X_i, \beta), \Psi)^\top y_i\}. \quad (3.30)$$

**Definition 3.11. Pseudo-Loglikelihood-Funktion und PML-Schätzer**

Die normierte Loglikelihood-Funktion der angenommenen Dichte

$$\begin{aligned}\ell(\beta) &= \frac{1}{m} \sum_{i=1}^m \{a(\mu(X_i, \beta), \Psi) + b(y_i, \Psi) + c(\mu(X_i, \beta), \Psi)^\top y_i\} \\ &=: \frac{1}{m} \sum_{i=1}^m \ell_i(\beta)\end{aligned}$$

wird Pseudo-Loglikelihood-Funktion genannt. Darüber hinaus wird ein Vektor  $\hat{\beta}$  als Pseudo-Maximum-Likelihood-Schätzer (PML-Schätzer) bezeichnet, wenn dieser  $\ell(\beta)$  maximiert.

**Bemerkung 3.12.**

Es gilt

$$\begin{aligned}\operatorname{argmax}_{\beta \in \mathbb{R}^q} \frac{1}{m} \sum_{i=1}^m \{a(\mu(X_i, \beta), \Psi) + b(y_i, \Psi) + c(\mu(X_i, \beta), \Psi)^\top y_i\} \\ = \operatorname{argmax}_{\beta \in \mathbb{R}^q} \frac{1}{m} \sum_{i=1}^m \{a(\mu(X_i, \beta), \Psi) + c(\mu(X_i, \beta), \Psi)^\top y_i\}.\end{aligned}$$

**Bemerkung 3.13.**

Die PML-Methode basiert auf der Idee, den Nuisance-Parameter  $\Psi$  als fest vorgegeben anzusehen und einen Schätzer für den wahren Regressionsparameter  $\beta^*$ , basierend auf der angenommenen Verteilung, zu bestimmen. Aus diesem Grund wird für den Rest dieses Abschnittes  $f(y, \mu, \Psi)$  durch  $f(y, \mu)$  ersetzt.

Die feste Wahl von  $\Psi$  entspricht im Fall der multivariaten Normalverteilung einer festen Wahl der Kovarianzmatrix. Diese Tatsache wird für die QGPML-Methode in Abschnitt 3.1.1.5.2 von Bedeutung sein.

Es kann gezeigt werden, dass die Konsistenz des PML-Schätzers lediglich der korrekten Spezifikation der Erwartungswertstruktur, d.h.

$$\mu^w(X_i, \beta^*) = \mu^a(X_i, \beta^*),$$

und der von Gourieroux u. a. (1984) aufgeführten Regularitätsbedingungen bedarf. Diese sind gegeben durch:

- (i) Die  $n_i$ -dimensionalen Zufallsvektoren  $Y_i$  sind für alle  $i \in \{1, \dots, m\}$  unabhängig und identisch verteilt. Ohne Beschränkung der Allgemeinheit wird  $n_i = n$  angenommen.

- (ii) Die Funktionen  $f^a(y, \mu(X, \beta))$  und  $\ell_i(\beta)$  (siehe Definition 3.11) sind dreimal stetig differenzierbar.
- (iii)  $\Theta \subset \mathbb{R}^q$  ist kompakt und der wahre Parameter  $\beta^*$  ist ein innerer Punkt von  $\Theta$ .
- (iv)  $\ell(\beta)$ ,  $\frac{\partial \ell(\beta)}{\partial \beta}$  und  $\frac{\partial^2 \ell(\beta)}{\partial \beta^2}$  sind bezüglich der Verteilungsfunktion  $F^a(y)$  integrierbar.
- (v) Für alle  $i \in \{1, \dots, m\}$  und  $k \in \{1, 2, 3\}$  gilt

$$\frac{\partial^k}{\partial \beta^k} \int f^a(y, \mu(X, \beta)) dy = \int \frac{\partial^k}{\partial \beta^k} f^a(y, \mu(X, \beta)) dy,$$

d.h. die Integration nach  $y$  und die Differentiation nach  $\beta$  lassen sich vertauschen.

- (vi) Sei  $y = (y_1, \dots, y_p)^\top \in \mathbb{R}^p$ , dann existiert eine Funktion  $K : \mathbb{R} \rightarrow \mathbb{R}$ , so dass für alle  $j \in \{1, \dots, p\}$

- $E_{f^a}(K(Y_j)) = \int K(y_j) f^a(y_j, \mu(X, \beta^*)) dy_j < \infty$  ist und
- der Betrag von

$$\sum_{\nu=1}^q \sum_{\mu=1}^q \frac{\partial^3 \ell(x)}{\partial \beta_\mu \partial \beta_\nu \partial \beta_j}$$

für alle  $x$  in einer Umgebung  $U_{\beta^*}$  von  $\beta^*$  gleichmäßig durch  $K(y_j)$  beschränkt ist.

- (vii) Die Matrix  $W(\beta^*) = \frac{\partial^2 \ell(\beta^*)}{\partial \beta^2}$  ist positiv definit und insbesondere auch invertierbar.
- (viii) Der Parametervektor  $\beta$  und der Erwartungswertvektor  $\mu(X, \beta)$  sind identifizierbar, d. h.:

$$\begin{aligned} f(y, \mu(X, \beta_1)) &= f(y, \mu(X, \beta_2)) \text{ für fast alle } y \\ \implies \mu(X, \beta_1) &= \mu(X, \beta_2) \text{ für fast alle } X \end{aligned}$$

und

$$\mu(X, \beta_1) = \mu(X, \beta_2) \text{ für fast alle } X \implies \beta_1 = \beta_2.$$

- (ix) Es gelten darüber hinaus die in Burguete u. a. (1982) aufgeführten Voraussetzungen, da die Resultate aus dieser Arbeit später für den Beweis des Satzes 3.20 benötigt werden.

**Bemerkung 3.14.**

Die Voraussetzung bezüglich der identischen Verteilung der Zufallsvariablen  $Y_i$ ,  $i \in \{1, \dots, m\}$ , ist bei mehreren zu vergleichenden Gruppen im Allgemeinen nicht erfüllt, da diese die Gleichheit der Prädiktormatrizen impliziert.

Allerdings lässt sich diese Voraussetzung abschwächen, indem angenommen wird, dass die Prädiktormatrizen  $X_i$ ,  $i \in \{1, \dots, m\}$ , Realisierungen von Zufallsmatrizen  $\bar{X}_i$  und die Paare  $(Y_1, \bar{X}_1), \dots, (Y_m, \bar{X}_m)$  unabhängig und identisch verteilt sind. In randomisierten Studien ist diese Forderung stets erfüllt; insbesondere ist die Verteilung der erklärenden Variablen in diesem Fall eine diskrete.

Im Rahmen dieser Arbeit werden die Eigenschaften der PML-Schätzer für identisch verteilte  $Y_i$ ,  $i \in \{1, \dots, m\}$ , gezeigt. Allerdings lässt sich der Beweis des folgenden Satzes mit der Annahme zufälliger Prädiktormatrizen und identisch verteilter Paare  $(Y_i, X_i)$  analog führen. Hierbei muss lediglich die Randdichte von  $\bar{X}_i$  berücksichtigt werden. Alle im Beweis getroffenen Aussagen über Integrierbarkeit und verwendete Grenzwertsätze bleiben im Fall zufälliger diskret verteilter  $\bar{X}_i$  bestehen.

**Satz 3.15.**<sup>14</sup>

Die vorangegangenen Voraussetzungen und die bisher verwendeten Bezeichnungen mit  $\mu_i = \mu(X_i, \beta)$  liefern folgende Eigenschaften:

- (i) Es existiert ein stark konsistenter PML-Schätzer  $\hat{\beta}$  für  $\beta^*$ .
- (ii) Es gilt die asymptotische Verteilung

$$\sqrt{m}(\hat{\beta} - \beta^*) \sim \mathcal{N}_q(0_q, [-A(\beta^*)]^{-1}B(\beta^*)[-A(\beta^*)]^{-1}).$$

Dabei sind

$$A(\beta) = E(W(\beta)) = \frac{1}{m} \sum_{i=1}^m \left[ \left( \frac{\partial \mu_i}{\partial \beta} \right)^\top (V_i^a(\beta))^{-1} \frac{\partial \mu_i}{\partial \beta} \right]$$

der Erwartungswert der Matrix

$$W(\beta) = \frac{\partial^2 \ell(\beta)}{\partial^2 \beta}$$

und

$$B(\beta) = E(u(\beta)u(\beta)^\top) = \frac{1}{m} \sum_{i=1}^m \left( \frac{\partial \mu_i}{\partial \beta} \right)^\top (V_i^a(\beta))^{-1} (\hat{V}_i^w(\beta))^{-1} (V_i^a(\beta))^{-1} \frac{\partial \mu_i}{\partial \beta}.$$

<sup>14</sup>Vgl. Ziegler (1999), S.33, Satz 3.3.3.

(iii) *Stark konsistente Schätzer für  $A(\beta^*)$  und  $B(\beta^*)$  sind mit  $\hat{\beta}$  aus (i) gegeben durch*

$$\hat{A}(\hat{\beta}) = \frac{1}{m} \sum_{i=1}^m \left[ \left( \frac{\partial \hat{\mu}_i}{\partial \beta} \right)^\top (V_i^a(\hat{\beta}))^{-1} \frac{\partial \hat{\mu}_i}{\partial \beta} \right],$$

sowie

$$\hat{B}(\hat{\beta}) = \frac{1}{m} \sum_{i=1}^m \left[ \left( \frac{\partial \hat{\mu}_i}{\partial \beta} \right)^\top (V_i^a(\hat{\beta}))^{-1} (\hat{V}_i^w(\hat{\beta})) (V_i^a(\hat{\beta}))^{-1} \frac{\partial \hat{\mu}_i}{\partial \beta} \right],$$

mit  $\hat{\mu}_i = \mu(X_i, \hat{\beta})$  und  $\hat{V}_i^w(\hat{\beta}) = (y_i - \hat{\mu}_i)(y_i - \hat{\mu}_i)^\top$ .

Bevor dieser Satz bewiesen werden kann, müssen Sätze aus der multivariaten Stochastik, sowie ein zentraler Satz der Arbeit von White (1981) zitiert werden.

**Satz 3.16. (Satz von White)**

Sei eine stetige Funktion  $h$  gegeben durch

$$\begin{aligned} h : \mathbb{R}^p \times \mathbb{R}^q &\rightarrow \mathbb{R} \\ (y, \beta) &\mapsto h(y, \beta). \end{aligned}$$

Für alle  $(y, \beta)$  gelte die Ungleichung  $|h(y, \beta)| \leq m(y)$  mit einer Funktion  $m$ , die bezüglich einer Verteilungsfunktion  $F$  integrierbar ist. Darüber hinaus sei  $(\beta_k^*)_{k \in \mathbb{N}}$  eine Folge von  $q$ -dimensionalen Vektoren, die

$$Q_k(\beta) = \frac{1}{k} \sum_{i=1}^k h(X_i, \beta)$$

maximieren. Besitzt die Funktion

$$Q(\beta) = \int h(x, \beta) dF(x)$$

ein einziges Maximum  $\beta^* \in \mathbb{R}^q$ , so konvergiert die Folge  $(\beta_k^*)_{k \in \mathbb{N}}$  aus  $\mathbb{R}^q$  gegen  $\beta^*$  für fast jede Folge von gemäß  $F$  verteilten Zufallsvariablen  $(X_t)_{t \in \mathbb{N}}$  (d.h.  $(\beta_k^*)_{k \in \mathbb{N}} \xrightarrow{f.s.} \beta^*$ ).

*Beweis:* Diese Aussage folgt nach White (1981), Theorem 2.1. □

**Satz 3.17. (Multivariater Grenzwertsatz)**

Sei  $(Z_n)_{n \in \mathbb{N}}$  eine Folge von unabhängigen identisch verteilten  $p$ -dimensionalen Zufallsvektoren mit  $E(Z_i) = 0$  und endlicher Kovarianzmatrix  $B$ . Dann gilt

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n Z_i \xrightarrow{\mathcal{D}} Z,$$

wobei  $Z$  ein normalverteilter Zufallsvektor mit Erwartungswertvektor  $0_p$  und Kovarianzmatrix  $B$  ist.

*Beweis:* Der Beweis ist in Gänsler und Stute (1977), S. 358, Satz 8.8.1 gegeben.  $\square$

**Satz 3.18. (Satz von Slutsky)**

Sei  $(X_n)_{n \in \mathbb{N}}$  eine Folge von  $(q \times p)$ -dimensionalen Zufallsmatrizen und  $(Z_n)_{n \in \mathbb{N}}$  eine Folge von  $p$ -dimensionalen Zufallsvektoren. Darüber hinaus konvergiere  $Z_n$  in Verteilung gegen  $Z$  und  $X_n$  stochastisch gegen  $X$ . Dann konvergiert  $X_n Z_n$  in Verteilung gegen  $XZ$ , d.h. es gilt:

$$X_n \xrightarrow{P} X, \quad Z_n \xrightarrow{D} Z \implies X_n Z_n \xrightarrow{D} XZ.$$

Mit Hilfe der vorgestellten Sätze kann jetzt Satz 3.15 bewiesen werden.

*Beweis von Satz 3.15:*

(i) Die Existenz eines konsistenten PML-Schätzers kann mit Hilfe von Satz 3.16 gezeigt werden. Sei

$$h(y_i, \beta) := a(\mu_i, \Psi) + c(\mu_i, \Psi)^\top y_i$$

mit  $\mu_i = \mu(X_i, \beta)$ . Die Funktion  $h(y_i, \beta)$  ist nach Voraussetzung (ii) stetig und damit auch meßbar. Da der Parameterraum  $\Theta$  nach Voraussetzung (iii) kompakt ist, nehmen die Funktionen  $a(\mu_i, \Psi)$  und  $c(\mu_i, \Psi)$  auf  $\Theta$  ein Maximum an. Sei

$$m(y_i) := \max_{\beta \in \Theta} |a(\mu_i, \Psi)| + \max_{\beta \in \Theta} |c(\mu_i, \Psi)^\top y_i|$$

mit  $\mu_i = \mu(X_i, \beta)$ . Dann ist  $m$  bezüglich der angenommenen Verteilungsfunktion von  $Y_i = (Y_{i1}, \dots, Y_{in})^\top$  integrierbar, weil mit  $c(\mu_i, \Psi) = (c_1(\mu_i, \Psi), \dots, c_n(\mu_i, \Psi))^\top$  und der Existenz des Erwartungswertes von  $Y_i$ ,  $i \in \{1, \dots, m\}$ , folgendes gilt

$$\begin{aligned} \int m(y_i) dF^a(y_i) &= \int m(y_i) f^a(y_i, \mu(X_i, \beta^*)) dy_i \\ &= \int \{ \max_{\beta \in \Theta} |a(\mu_i, \Psi)| + \max_{\beta \in \Theta} |c(\mu_i, \Psi)^\top y_i| \} f^a(y_i, \mu(X_i, \beta^*)) dy_i \\ &= \max_{\beta \in \Theta} |a(\mu_i, \Psi)| + \mathbb{E}(\max_{\beta \in \Theta} |c(\mu_i, \Psi)^\top Y_i|) \\ &\leq \max_{\beta \in \Theta} |a(\mu_i, \Psi)| + \sum_{k=1}^n \mathbb{E}(\max_{\beta \in \Theta} |c_k(\mu_i, \Psi) Y_{ik}|) \\ &< \infty. \end{aligned} \tag{3.31}$$

Zudem ergibt sich für alle  $\beta \in \Theta$  und  $y_i \in \mathbb{R}^p$  folgende Abschätzung

$$|h(y_i, \beta)| \leq m(y_i).$$

Um den Satz von White anwenden zu können, muss noch gezeigt werden, dass

$$Q(\beta) = \int h(y_i, \beta) dF^a(y_i)$$

ein eindeutiges Maximum besitzt. Es gilt

$$\begin{aligned}
 Q(\beta) &= \int h(y_i, \beta) dF^a(y_i) \\
 &= \int h(y_i, \beta) f^a(Y_i, \mu(X_i, \beta^*)) dy_i \\
 &= \int [a(\mu_i, \Psi) + c(\mu_i, \Psi)^\top y_i] f^a(y_i, \mu(X_i, \beta^*)) dy_i \\
 &= a(\mu_i, \Psi) \int f^a(y_i, \mu(X_i, \beta^*)) dy_i + c(\mu_i, \Psi)^\top \int y_i f^a(y_i, \mu(X_i, \beta^*)) dy_i \\
 &= a(\mu_i, \Psi) + c(\mu_i, \Psi)^\top \underbrace{\mu^a(X_i, \beta^*)}_{=\mu_i^*}.
 \end{aligned}$$

Nach Satz 3.10 gilt für ein beliebiges  $\beta \in \mathbb{R}^q$  und ein festes  $\beta^*$

$$a(\mu_i, \Psi) + c(\mu_i, \Psi)^\top \mu_i^* \leq a(\mu_i^*, \Psi) + c(\mu_i^*, \Psi)^\top \mu_i^*.$$

Die Gleichheit gilt fast sicher, wenn  $\mu_i = \mu(X_i, \beta) = \mu(X_i, \beta^*) = \mu_i^*$  für fast alle  $X_i$  erfüllt wird. Mit der vorausgesetzten Identifizierbarkeit erster Ordnung folgt

$$\beta = \beta^*,$$

also liegt in  $\beta^*$  ein eindeutiges globales Maximum vor.

Sei  $(\hat{\beta}_k)_{k \in \mathbb{N}}$  eine Folge, und jedes Folgenglied maximiere die Summe

$$Q_k(\beta) = \frac{1}{k} \sum_{i=1}^k h(Y_i, \beta).$$

Dann ist  $(\hat{\beta}_k)_{k \in \mathbb{N}}$  nach Bemerkung 3.12 eine Folge von PML-Schätzfunktionen und mit Hilfe des Satzes von White wird die fast sichere Konvergenz von  $(\hat{\beta}_k)_{k \in \mathbb{N}}$ , d.h.

$$\hat{\beta}_k \longrightarrow \beta^*,$$

und damit die Behauptung geliefert.

(ii) Die zu maximierende Pseudo-Likelihood-Funktion ist nach Definition 3.11 gegeben durch

$$\ell(\beta) = \ln(f^a(y, \mu)) = \frac{1}{m} \sum_{i=1}^m \underbrace{\{a(\mu(X_i, \beta), \Psi) + b(y_i, \Psi) + c(\mu(X_i, \beta), \Psi)^\top y_i\}}_{=\ln(f^a(y_i, \mu(X_i, \beta)))}, \quad (3.32)$$

wobei  $y = (y_{11}, \dots, y_{mn_m})^\top$  ist. Die Bestimmung der lokalen Maxima bedarf der Berechnung der ersten und zweiten Ableitung von  $\ell(\beta)$  nach  $\beta$ . Die Transponierte der ersten Ableitung, und damit die Score-Gleichung, ist mit  $\mu_i = \mu(X_i, \beta)$  durch

$$u(\beta) := \left( \frac{\partial \ell(\beta)}{\partial \beta} \right)^\top = \frac{1}{m} \sum_{i=1}^m \underbrace{\left( \frac{\partial \{a(\mu_i, \Psi) + b(y_i, \Psi) + c(\mu_i, \Psi)^\top y_i\}}{\partial \beta} \right)^\top}_{=: u_i(\beta)}$$

gegeben. Dabei handelt es sich um einen  $q$ -dimensionalen Spaltenvektor, dessen Einträge für alle  $i \in \{1, \dots, m\}$  durch  $a_i(\beta)$  gegeben seien. Die Ableitung der Score-Gleichung sieht wie folgt aus:

$$W(\beta) := \frac{\partial u(\beta)}{\partial \beta} = \frac{1}{m} \sum_{i=1}^m \underbrace{\frac{\partial u_i(\beta)}{\partial \beta}}_{=: W_i(\beta)}.$$

Eine Taylor-Entwicklung erster Ordnung von  $u(\beta)$  um den Entwicklungspunkt  $\beta^*$  ist komponentenweise durchzuführen. Die  $i$ -te Komponente

$$a_i(\beta) = a_i(\beta^*) + \text{grad}(a_i(\beta^*))(\beta - \beta^*) + \frac{1}{2}(\beta - \beta^*)^\top H_{a_i}(\xi)(\beta - \beta^*)$$

enthält mit  $\text{grad}(a_i(\beta^*))$  den Gradienten und mit  $H_{a_i}(\xi)$  die Hessematrix von  $a_i(\beta)$ . Zusätzlich erfüllt  $\xi$  die Gleichung  $\xi = \beta^* + \vartheta(\beta - \beta^*)$  mit  $\vartheta \in [-1, 1]$  und somit  $\xi \in U_{\beta^*}$ . Fügt man diese einzelnen Komponenten wieder zusammen, ergibt sich

$$u(\beta) = u(\beta^*) + W(\beta^*)(\beta - \beta^*) + R(\beta)$$

mit dem Restgliedvektor

$$R(\beta) = \left( \frac{1}{2}(\beta - \beta^*)^\top H_{a_i}(\xi)(\beta - \beta^*) \right)_{i \in \{1, \dots, q\}}.$$

Die Hessematrix von  $a_i(\beta) = \frac{\partial \ell(\beta)}{\partial \beta_i}$  ist nach dem Satz von Schwarz symmetrisch und es gilt

$$H_{a_i}(\xi) = \left( \frac{\partial^3 \ell(\xi)}{\partial \beta_\mu \partial \beta_\nu \partial \beta_i} \right)_{\substack{\nu \in \{1, \dots, q\} \\ \mu \in \{1, \dots, q\}}}.$$

Diese Darstellung liefert gemeinsam mit Voraussetzung (vi) folgendes:

$$\begin{aligned} \frac{1}{2}(\beta - \beta^*)^\top H_{a_i}(\xi)(\beta - \beta^*) &= \frac{1}{2} \sum_{\nu=1}^q \sum_{\mu=1}^q \frac{\partial^3 \ell(\xi)}{\partial \beta_\mu \partial \beta_\nu \partial \beta_i} (\beta_\mu - \beta_\mu^*) (\beta_\nu - \beta_\nu^*) \\ &\leq \frac{1}{2} \left\| \sum_{\nu=1}^q \sum_{\mu=1}^q \frac{\partial^3 \ell(\xi)}{\partial \beta_\mu \partial \beta_\nu \partial \beta_i} (\beta_\mu - \beta_\mu^*) (\beta_\nu - \beta_\nu^*) \right\| \\ &\leq \frac{2q}{2} \|\beta - \beta^*\|^2 \left\| \sum_{\nu=1}^q \sum_{\mu=1}^q \frac{\partial^3 \ell(\xi)}{\partial \beta_\mu \partial \beta_\nu \partial \beta_i} \right\| \\ &\leq q \|\beta - \beta^*\|^2 \frac{1}{m} \sum_{i=1}^m \left\| \sum_{\nu=1}^q \sum_{\mu=1}^q \frac{\partial^3 \ell_i(\xi)}{\partial \beta_\mu \partial \beta_\nu \partial \beta_i} \right\| \\ &\leq q \|\beta - \beta^*\|^2 \frac{1}{m} \sum_{i=1}^m K(y_i). \end{aligned} \tag{3.33}$$

Darüber hinaus folgt aus dem schwachen Gesetz der großen Zahlen und Voraussetzung (i) und (vi) diese Konvergenz

$$\frac{1}{m} \sum_{i=1}^m K(Y_i) \xrightarrow{\mathcal{P}} \mathbb{E}_{f^a}(K(Y_i)) < \infty,$$

woraus sich zusammen mit den in Kapitel 2 definierten Landau-Symbolen die Gleichung

$$\frac{1}{2}(\beta - \beta^*)^\top H_{a_i}(\xi)(\beta - \beta^*) = \|\beta - \beta^*\|^2 O_p(1)$$

ergibt. Für den Restgliedvektor lässt sich daher

$$R(\beta) = \|\beta - \beta^*\|^2 O_p(1) \mathbf{1}_q$$

schreiben. Insgesamt gilt somit

$$u(\beta) = u(\beta^*) + W(\beta^*)(\beta - \beta^*) + \|\beta - \beta^*\|^2 O_p(1) \mathbf{1}_q.$$

Sei nun  $\hat{\beta}$  der nach (i) existierende konsistente PML-Schätzer. Dann ist  $u(\hat{\beta}) = 0$  nach Definition 3.11 und somit

$$0 = u(\beta^*) + W(\beta^*)(\hat{\beta} - \beta^*) + \|\hat{\beta} - \beta^*\|^2 O_p(1) \mathbf{1}_q. \quad (3.34)$$

Das Umstellen dieser Gleichung sowie anschließendes Multiplizieren mit  $\sqrt{m}$  liefert

$$\begin{aligned} -\sqrt{m}(W(\beta^*))^{-1}u(\beta^*) &= \sqrt{m}(\hat{\beta} - \beta^*) + \sqrt{m}(W(\beta^*))^{-1}\|\hat{\beta} - \beta^*\|^2 O_p(1) \mathbf{1}_q \\ &= \underbrace{\left( I_{q \times q} + (W(\beta^*))^{-1}(\hat{\beta} - \beta^*)^\top O_p(1) \mathbf{1}_q \right)}_{B_m} \sqrt{m}(\hat{\beta} - \beta^*). \end{aligned} \quad (3.35)$$

Mit Hilfe dieser Resultate und des Satzes von Slutsky soll jetzt die Verteilung von  $\sqrt{m}(\hat{\beta} - \beta^*)$  bestimmt werden.

Sei  $X := W(\beta^*)^{-1}$ , dann gilt aufgrund der Unabhängigkeit und der identischen Verteilung von  $Y_i$  für alle  $i \in \{1, \dots, m\}$  sowie aufgrund des starken Gesetzes der großen Zahlen die Konvergenz

$$X^{-1} = W(\beta^*) = \frac{1}{m} \sum_{i=1}^m W_i(\beta^*) \xrightarrow{f.s.} \mathbb{E}_{f^a}(W_i(\beta^*)) =: A(\beta^*). \quad (3.36)$$

Sei darüber hinaus

$$Z := \sqrt{m}u(\beta^*) = \frac{1}{\sqrt{m}} \sum_{i=1}^m u_i(\beta^*),$$

dann liefert die nach Voraussetzung (v) zugelassene Vertauschung von Differentiation nach  $\beta$  und Integration nach  $y_i$  den Wert  $0_q$  für  $E(u_i(\beta^*))$ :

$$\begin{aligned}
 E(u_i(\beta^*)) &= \int u_i(\beta^*) f^a(y_i, \mu(X_i, \beta^*)) dy_i \\
 &= \int \left( \frac{\partial \ln(f^a(y_i, \mu(X_i, \beta^*)))}{\partial \beta} \right)^\top f^a(y_i, \mu(X_i, \beta^*)) dy_i \\
 &= \int \frac{\left( \frac{\partial f^a(y_i, \mu(X_i, \beta^*))}{\partial \beta} \right)^\top}{f^a(y_i, \mu(X_i, \beta^*))} f^a(y_i, \mu(X_i, \beta^*)) dy_i \\
 &= \int \left( \frac{\partial f^a(y_i, \mu(X_i, \beta^*))}{\partial \beta} \right)^\top dy_i \\
 &= \left( \frac{\partial}{\partial \beta} \int f^a(y_i, \mu(X_i, \beta^*)) dy_i \right)^\top \\
 &= 0_q.
 \end{aligned}$$

Zudem gilt

$$\begin{aligned}
 \text{Cov}(u_i(\beta^*)) &= E(u_i(\beta^*) u_i(\beta^*)^\top) - E(u_i(\beta^*)) E(u_i(\beta^*))^\top \\
 &\stackrel{\text{s.o.}}{=} E(u_i(\beta^*) u_i(\beta^*)^\top) \\
 &=: B(\beta^*).
 \end{aligned}$$

Nach dem Multivariaten Grenzwertsatz folgt somit

$$Z \xrightarrow{\mathcal{D}} \mathcal{N}(0, B(\beta^*)). \quad (3.37)$$

Die Konvergenzen in (3.36) und (3.37) ermöglichen es den Satz von Slutsky anzuwenden. Es wird ausgenutzt, dass die Matrix  $A(\beta^*)$  aufgrund der zweifach stetig differenzierbaren Funktionen  $a, b$  und  $c$  symmetrisch ist. Darüber hinaus ist  $W(\beta^*)$  nach Voraussetzung (vii) invertierbar und daher auch  $A(\beta^*)$  aufgrund der Stetigkeit der Determinante. Für die linke Seite der Gleichung (3.35) ergibt sich somit

$$\begin{aligned}
 XZ &= \sqrt{m} W(\beta^*)^{-1} u(\beta^*) \xrightarrow{\mathcal{D}} A(\beta^*)^{-1} \mathcal{N}(0, B(\beta^*)) \\
 &= \mathcal{N}_q(0, A(\beta^*)^{-1} B(\beta^*) A(\beta^*)^{-1}). \quad (3.38)
 \end{aligned}$$

Unter der Voraussetzung, dass die Inverse von  $B_m$  aus Gleichung (3.35) existiert, und es gelingt  $B_m \xrightarrow{\mathcal{P}} I_{q \times q}$  zu zeigen, dann folgt aus den Gleichungen (3.35) und (3.38), sowie dem Satz von Slutsky, folgende Konvergenz:

$$\sqrt{m}(\hat{\beta} - \beta^*) = -B_m^{-1} \underbrace{\sqrt{m}(W(\beta^*))^{-1} u(\beta^*)}_{=XZ} \xrightarrow{\mathcal{D}} \mathcal{N}_q(0, A(\beta^*)^{-1} B(\beta^*) A(\beta^*)^{-1}). \quad (3.39)$$

Wegen

$$B_m = I_{q \times q} + (W(\beta^*))^{-1} (\hat{\beta} - \beta^*)^\top O_p(1) \mathbf{1}_q,$$

wobei  $\hat{\beta}$  ein stark konsistenter Schätzer für  $\beta^*$  und  $O_p(1)$  eine stochastisch beschränkte Folge ist, gilt

$$\underbrace{(W(\beta^*))^{-1}}_{\xrightarrow{P} A(\beta^*)^{-1}} \underbrace{(\hat{\beta} - \beta^*)^\top}_{\xrightarrow{P} \mathbf{0}} \underbrace{O_p(1) \mathbf{1}_q}_{\xrightarrow{P} c} \xrightarrow{P} 0_{q \times q}$$

mit einer Konstante  $c \in \mathbb{R}$ , womit auch  $B_m \xrightarrow{P} I_{q \times q}$  folgt.

Insgesamt ergibt sich daher mit den obigen Ausführungen, der Gleichung (3.39) und dem Satz von Slutsky die Konvergenz

$$\sqrt{m} (\hat{\beta} - \beta^*) \xrightarrow{D} \mathcal{N}(0, A(\beta^*)^{-1} B(\beta^*) A(\beta^*)^{-1})$$

und damit die behauptete Verteilung.

Nun bleibt noch die Aussage über die Gestalt der Matrizen  $A(\beta)$  und  $B(\beta)$  zu beweisen. Es gilt

$$\begin{aligned} W(\beta) &= \frac{\partial u(\beta)}{\partial \beta} \\ &= \frac{\partial}{\partial \beta} \left[ \frac{1}{m} \sum_{i=1}^m \frac{\partial a(\mu_i, \Psi) + b(y_i, \Psi) + c(\mu_i, \Psi)^\top y_i}{\partial \beta} \right]^\top \\ &\stackrel{\text{Kettenregel}}{=} \frac{1}{m} \sum_{i=1}^m \frac{\partial}{\partial \beta} \left[ \left( \frac{\partial a(\mu_i, \Psi)}{\partial \mu_i} + y_i^\top \frac{\partial c(\mu_i, \Psi)^\top}{\partial \mu_i} \right) \frac{\partial \mu_i}{\partial \beta} \right]^\top \\ &= \frac{1}{m} \sum_{i=1}^m \frac{\partial}{\partial \beta} \left[ \left( \frac{\partial \mu_i}{\partial \beta} \right)^\top \left[ \left( \frac{\partial a(\mu_i, \Psi)}{\partial \mu_i} \right)^\top + \left( \frac{\partial c(\mu_i, \Psi)^\top}{\partial \mu_i} y_i \right) \right] \right] \\ &= \frac{1}{m} \sum_{i=1}^m \frac{\partial}{\partial \beta} \left[ \sum_{\ell=1}^q \left( \frac{\partial \mu_{i\ell}}{\partial \beta} \right)^\top \left( \frac{\partial a(\mu_i, \Psi)}{\partial \mu_{i\ell}} + \frac{\partial c(\mu_i, \Psi)^\top y_i}{\partial \mu_{i\ell}} \right) \right] \\ &\stackrel{\text{Produktregel}}{=} \frac{1}{m} \sum_{i=1}^m \left[ \sum_{\ell=1}^q \frac{\partial}{\partial \beta} \left( \frac{\partial \mu_{i\ell}}{\partial \beta} \right)^\top \left( \frac{\partial a(\mu_i, \Psi)}{\partial \mu_{i\ell}} + \frac{\partial c(\mu_i, \Psi)^\top y_i}{\partial \mu_{i\ell}} \right) \right. \\ &\quad \left. + \left( \frac{\partial \mu_i}{\partial \beta} \right)^\top \left( \frac{\partial^2 a(\mu_i, \Psi)}{\partial \mu_i \partial \mu_i} + \frac{\partial^2 c(\mu_i, \Psi)^\top y_i}{\partial \mu_i \partial \mu_i} \right) \left( \frac{\partial \mu_i}{\partial \beta} \right) \right] \\ &\stackrel{\text{vgl. Satz 3.5}}{=} \frac{1}{m} \sum_{i=1}^m \left[ \sum_{\ell=1}^q \frac{\partial}{\partial \beta} \left( \frac{\partial \mu_{i\ell}}{\partial \beta} \right)^\top \left( -\frac{\partial c(\mu_i, \Psi)^\top}{\partial \mu_{i\ell}} \mu_i + \frac{\partial c(\mu_i, \Psi)^\top y_i}{\partial \mu_{i\ell}} \right) \right. \\ &\quad \left. + \left( \frac{\partial \mu_i}{\partial \beta} \right)^\top \left( \frac{\partial^2 a(\mu_i, \Psi)}{\partial \mu_i \partial \mu_i} + \frac{\partial^2 c(\mu_i, \Psi)^\top y_i}{\partial \mu_i \partial \mu_i} \right) \left( \frac{\partial \mu_i}{\partial \beta} \right) \right] \\ &\stackrel{\text{vgl. Satz 3.6}}{=} \frac{1}{m} \sum_{i=1}^m \left[ \sum_{\ell=1}^q \frac{\partial}{\partial \beta} \left( \frac{\partial \mu_{i\ell}}{\partial \beta} \right)^\top \left( \frac{\partial c(\mu_i, \Psi)^\top}{\partial \mu_{i\ell}} \right) (y_i - \mu_i) \right] \end{aligned}$$

$$\begin{aligned}
 & + \left( \frac{\partial \mu_i}{\partial \beta} \right)^\top \left( -\frac{\partial^2 c(\mu_i, \Psi)^\top}{\partial \mu_i \partial \mu_i} \mu_i - \frac{\partial c(\mu_i, \Psi)^\top}{\partial \mu_i} \right. \\
 & \left. + \frac{\partial^2 c(\mu_i, \Psi)^\top y_i}{\partial \mu_i \partial \mu_i} \right) \left( \frac{\partial \mu_i}{\partial \beta} \right) \Big] \\
 = & \frac{1}{m} \sum_{i=1}^m \left[ \sum_{\ell=1}^q \frac{\partial}{\partial \beta} \left( \frac{\partial \mu_{i\ell}}{\partial \beta} \right)^\top \left( \frac{\partial c(\mu_i, \Psi)^\top}{\partial \mu_{i\ell}} \right) (y_i - \mu_i) \right. \\
 & \left. + \left( \frac{\partial \mu_i}{\partial \beta} \right)^\top \left( \frac{\partial^2 c(\mu_i, \Psi)^\top}{\partial \mu_i \partial \mu_i} (y_i - \mu_i) - \frac{\partial c(\mu_i, \Psi)^\top}{\partial \mu_i} \right) \left( \frac{\partial \mu_i}{\partial \beta} \right) \right].
 \end{aligned}$$

Da  $E_{f^a}(Y - \mu) = 0$ , ergibt sich für den Erwartungswert von  $W(\beta)$

$$\begin{aligned}
 A(\beta) = E_{f^a}(W(\beta)) & = -\frac{1}{m} \sum_{i=1}^m \left( \frac{\partial \mu_i}{\partial \beta} \right)^\top \frac{\partial c(\mu_i, \Psi)^\top}{\partial \mu_i} \left( \frac{\partial \mu_i}{\partial \beta} \right) \\
 & \stackrel{\text{vgl. Satz 3.7}}{=} -\frac{1}{m} \sum_{i=1}^m \left( \frac{\partial \mu_i}{\partial \beta} \right)^\top (V_i^a(\beta))^{-1} \left( \frac{\partial \mu_i}{\partial \beta} \right).
 \end{aligned}$$

Außerdem gilt

$$\begin{aligned}
 u(\beta) & = \frac{1}{m} \sum_{i=1}^m \left( \frac{\partial a(\mu_i, \Psi) + b(y_i, \Psi) + c(\mu_i, \Psi)^\top y_i}{\partial \beta} \right)^\top \\
 & = \frac{1}{m} \sum_{i=1}^m \left( \frac{\partial \mu_i}{\partial \beta} \right)^\top \left[ \frac{\partial a(\mu_i, \Psi)}{\partial \mu_i} + y_i^\top \frac{\partial c(\mu_i, \Psi)^\top}{\partial \mu_i} \right]^\top \\
 & \stackrel{\text{vgl. Satz 3.5}}{=} \frac{1}{m} \sum_{i=1}^m \left( \frac{\partial \mu_i}{\partial \beta} \right)^\top \left[ (y_i - \mu_i)^\top \frac{\partial c(\mu_i, \Psi)^\top}{\partial \mu_i} \right]^\top \\
 & = \frac{1}{m} \sum_{i=1}^m \left( \frac{\partial \mu_i}{\partial \beta} \right)^\top \frac{\partial c(\mu_i, \Psi)^\top}{\partial \mu_i} (y_i - \mu_i)
 \end{aligned}$$

und damit auch

$$\begin{aligned}
 & B(\beta) \\
 = & E_{f^a}(u(\beta) u(\beta)^\top) \\
 = & \frac{1}{m} \sum_{i=1}^m E_{f^a} \left( \left( \frac{\partial \mu_i}{\partial \beta} \right)^\top \frac{\partial c(\mu_i, \Psi)^\top}{\partial \mu_i} (y_i - \mu_i) (y_i - \mu_i)^\top \frac{\partial c(\mu_i, \Psi)^\top}{\partial \mu_i} \left( \frac{\partial \mu_i}{\partial \beta} \right) \right) \\
 = & \frac{1}{m} \sum_{i=1}^m \left( \frac{\partial \mu_i}{\partial \beta} \right)^\top \frac{\partial c(\mu_i, \Psi)^\top}{\partial \mu_i} E_{f^a} \left( (y_i - \mu_i) (y_i - \mu_i)^\top \right) \frac{\partial c(\mu_i, \Psi)^\top}{\partial \mu_i} \left( \frac{\partial \mu_i}{\partial \beta} \right) \\
 = & \frac{1}{m} \sum_{i=1}^m \left( \frac{\partial \mu_i}{\partial \beta} \right)^\top \frac{\partial c(\mu_i, \Psi)^\top}{\partial \mu_i} \hat{V}_i^w(\beta) \frac{\partial c(\mu_i, \Psi)^\top}{\partial \mu_i} \left( \frac{\partial \mu_i}{\partial \beta} \right) \\
 \stackrel{\text{vgl. Satz 3.7}}{=} & \frac{1}{m} \sum_{i=1}^m \left( \frac{\partial \mu_i}{\partial \beta} \right)^\top (V_i^a(\beta))^{-1} \hat{V}_i^w(\beta) (V_i^a(\beta))^{-1} \left( \frac{\partial \mu_i}{\partial \beta} \right).
 \end{aligned}$$

Mit  $\beta = \beta^*$  folgt die Behauptung.

(iii) siehe White (1981), Theorem 3.3. □

Darüber hinaus zeigen Gourieroux u. a. (1984), dass die Zugehörigkeit der angenommenen Dichte zur linearen Exponentialfamilie notwendig für die starke Konsistenz des PML-Schätzers ist.<sup>15</sup>

Der Satz 3.15 liefert die starke Konsistenz des PML-Schätzers, jedoch wird keine Aussage über die Effizienz des Schätzers getroffen. Die von Liang und Zeger (1986) geführten Simulationsstudien belegen, dass eine Effizienzsteigerung erreicht werden kann, wenn sich die angenommene Korrelationsmatrix nur in geringem Maße von der wahren Korrelationsmatrix unterscheidet. Diese Ergebnisse liefern den Anreiz, die Kovarianzmatrix und damit auch die Korrelationsmatrix von  $Y_i$ ,  $i \in \{1, \dots, m\}$ , nicht im Vorfeld festzusetzen (vgl. Bemerkung 3.13), sondern diese gegebenenfalls zu schätzen. Da wir in unserem speziellen Fall in Abschnitt 3.1.1.5.3 die angenommene Verteilung als multivariate Normalverteilung wählen werden, entspricht der Bedarf eines Schätzers für die Kovarianzmatrix dem Bedarf eines Schätzers für den Parameter  $\Psi$  (vgl. Bemerkung 3.13). Dieses Vorhaben führt uns direkt zu der *Quasi-Generalisierten-Pseudo-Maximum-Likelihood-Methode* (QGPML-Methode).

### 3.1.1.5.2 QGPML-Methode

Im letzten Abschnitt wurde die PML-Methode mit einem fest vorgegebenen Parameter  $\Psi$  in der angenommenen Dichte vorgestellt. Jedoch ist im Allgemeinen nicht bekannt, wie man diese Größe festsetzen soll, so dass sie bei bestimmten Verteilungen gegebenenfalls geschätzt werden muss.

Darüber hinaus kann durch die Berücksichtigung von  $\Psi_i$ ,  $i \in \{1, \dots, m\}$  ein im Vergleich zu dem PML-Schätzer effizienterer Schätzer für  $\beta^*$  erhalten werden.<sup>16</sup> Eine Methode, die den Nuisance-Parameter in der Analyse berücksichtigt ist durch die QGPML-Methode gegeben.

Sei die angenommene Verteilung von  $Y_i$  aus einer linearen Exponentialfamilie und die Dichte bestimmt durch

$$f^a(y_i, \mu_i, \Psi_i) = \exp\{a(\mu_i, \Psi_i) + b(y_i, \Psi_i) + c(\mu_i, \Psi_i)^\top y_i\}$$

mit den Nuisance-Parametern  $\Psi_i$ ,  $i \in \{1, \dots, m\}$ . Für die Kovarianzmatrix  $\Sigma_i$  von  $Y_i$  ergibt sich nach Satz 3.7 folgendes:

$$\Sigma_i^{-1} := \Sigma_i^{-1}(X_i, \beta, \Psi) = \frac{\partial c(\mu_i, \Psi)^\top}{\partial \mu_i}.$$

<sup>15</sup>Vgl. Gourieroux u. a. (1984), S. 686, Theorem 2.

<sup>16</sup>Vgl. Ziegler (1999), S. 37.

Somit liefert der *Satz über implizit gegebene Funktionen*<sup>17</sup> die Gleichheit

$$\Psi_i = G(\mu_i, \Sigma_i)$$

mit einer stetig differenzierbaren Funktion  $G$ .

Des Weiteren hängt  $\Sigma_i$ ,  $i \in \{1, \dots, m\}$ , aufgrund der Modellannahmen eines marginalen Modells, von einem zusätzlichen Parameter  $\alpha$  ab. Somit lässt sich mit einer weiteren Funktion  $H$  auch

$$\Psi_i = G(\mu_i, H(X_i, \beta^*, \alpha))$$

schreiben.<sup>18</sup>

Die QGPML-Methode ist in zwei Schritte unterteilt: Im ersten Schritt werden die Parameter  $\Psi_i$ ,  $i \in \{1, \dots, m\}$ , geschätzt und somit fixiert, und erst im zweiten Schritt werden dann effizientere Schätzer für den Regressionsvektor  $\beta^*$  bestimmt.

Um einen Schätzer für  $\Psi_i = G(\mu_i, H(X_i, \beta^*, \alpha))$  zu ermitteln, werden konsistente Schätzer für  $\alpha$  und  $\beta^*$  benötigt. Ein konsistenter Schätzer für  $\beta^*$  kann mittels der PML-Methode mit Normalverteilungsannahme und der Einheitsmatrix als Kovarianzmatrix geschätzt werden.<sup>19</sup> Entsprechende Schätzer für  $\alpha$  können ebenfalls durch die PML-Methode berechnet oder anhand der Tabelle 3.1 gewählt werden.<sup>20</sup> Seien diese PML-Schätzer durch  $\tilde{\alpha}$  und  $\tilde{\beta}$  gegeben. Dann ist für alle  $i \in \{1, \dots, m\}$

$$\tilde{\Psi}_i := G\left(\mu(X_i, \tilde{\beta}), H(X_i, \tilde{\beta}, \tilde{\alpha})\right) \quad (3.40)$$

ein Schätzer für  $\Psi_i$  und infolgedessen unabhängig von  $\beta$ . Im zweiten Schritt kann dann ein im Vergleich zum PML-Schätzer effizienterer Schätzer für  $\beta^*$  unter Berücksichtigung des geschätzten Parameters  $\tilde{\Psi}_i$ ,  $i \in \{1, \dots, m\}$ , bestimmt werden. Dieser Schätzer wird auch *Quasi-Generalisierter-Pseudo-Maximum-Likelihood-Schätzer* (QGPML-Schätzer) genannt und ist definiert durch:

**Definition 3.19.**

Ein Vektor  $\hat{\beta} \in \mathbb{R}^q$ , der die normierte Summe

$$\ell(\beta, \alpha) := \frac{1}{m} \sum_{i=1}^m \ln \left( f(y_i, \mu(X_i, \beta)), G\left(\mu(X_i, \tilde{\beta}), H(X_i, \tilde{\beta}, \tilde{\alpha})\right) \right) \quad (3.41)$$

maximiert, heißt *Quasi-Generalisierter-Pseudo-Maximum-Likelihood-Schätzer* für  $\beta^*$ .

---

<sup>17</sup>Vgl. Schmieder (1994), S. 193, Satz 26.3.

<sup>18</sup>Vgl. Ziegler (1999), S. 37.

<sup>19</sup>Vgl. Ziegler (1999), S.56.

<sup>20</sup>Vgl. Ziegler (1999), S. 37.

Für die QGPML-Schätzer gelten nun folgende Eigenschaften:

**Satz 3.20.**<sup>21</sup>

Die Voraussetzungen in 3.1.1.5.1 und die bisher verwendeten Bezeichnungen mit  $\mu_i = \mu(X_i, \beta)$  liefern folgende Eigenschaften:

- (i) Es existiert ein stark konsistenter QGPML-Schätzer  $\hat{\beta}$  für  $\beta^*$ .
- (ii) Es gilt die asymptotische Verteilung

$$\sqrt{n} (\hat{\beta} - \beta^*) \sim \mathcal{N}_q (0_q, [-A(\beta^*, \alpha)]^{-1} B(\beta^*, \alpha) [-A(\beta^*, \alpha)]^{-1}).$$

Dabei sind  $A(\beta, \alpha)$  der Erwartungswert der zweiten Ableitung von  $\ell(\beta, \alpha)$  nach  $\beta$  und  $B(\beta, \alpha)$  der Erwartungswert von  $Z^\top Z$ , wobei  $Z := \frac{\partial \ell(\beta, \alpha)}{\partial \beta}$  ist.

- (iii) Stark konsistente Schätzer für  $A(\beta^*, \alpha)$  und  $B(\beta^*, \alpha)$  sind gegeben durch

$$\hat{A} = \frac{1}{m} \sum_{i=1}^m \left[ \left( \frac{\partial \hat{\mu}_i}{\partial \beta} \right)^\top \hat{\Sigma}_i^{-1} \frac{\partial \hat{\mu}_i}{\partial \beta} \right],$$

sowie

$$\hat{B} = \frac{1}{m} \sum_{i=1}^m \left[ \left( \frac{\partial \hat{\mu}_i}{\partial \beta} \right)^\top \hat{\Sigma}_i^{-1} \hat{V}_i^w(\hat{\beta}) \hat{\Sigma}_i^{-1} \frac{\partial \hat{\mu}_i}{\partial \beta} \right]$$

mit  $\hat{\mu}_i = \mu(X_i, \hat{\beta})$  und  $\hat{\Sigma}_i = \Sigma_i(X_i, \hat{\beta}, \tilde{\Psi})$ .

*Beweis:* Der Beweis dieser Aussage erfolgt mit Hilfe des Satzes von White und der Untersuchungen von Burguete u. a. (1982). Diese Resultate ermöglichen das Ersetzen des Parameters  $\Psi$  durch einen konsistenten Schätzer  $\tilde{\Psi}$  und liefern mit  $\tilde{\Psi}$  anstelle von  $\Psi$  die behaupteten Eigenschaften. □

### 3.1.1.5.3 PML-, QGPML-Methode und GEE

In diesem Abschnitt werden die PML- und die QGPML-Methode für die Analyse wiederholter und damit korrelierter Daten angewandt. Es wird erläutert, dass die Schätzgleichungen der QGPML-Methode mit den GEE von Liang und Zeger (1986) übereinstimmen. Allerdings beruht diese Gleichheit auf der Annahme, dass die erklärenden Variablen zufällig und die Paare  $(Y_1, X_1), \dots, (Y_m, X_m)$  unabhängig und identisch verteilt sind (vgl. dazu Bemerkung 3.14). Diese Forderung ist für randomisierte Studien stets erfüllt, so dass sich in diesem Fall die Äquivalenz der QGPML- und GEE-Schätzer ergibt.

Sollen solche Annahmen vermieden werden, liefert Satz 3.15 nur eine eingeschränkte

---

<sup>21</sup>Vgl. Ziegler (1999), S.38, Satz 3.4.2.

Vergleichbarkeit der QGPML- und GEE-Schätzer. Allerdings bringen die Resultate des letzten Abschnitts und die im Folgenden vorgestellte Herleitung der Verallgemeinerten Schätzgleichungen (GEE) zumindest die Gewissheit, dass die Analyse mittels der GEE innerhalb einer Gruppe stark konsistente Schätzer liefert. Es erscheint daher in diesem Fall ratsam, den Regressionsvektor  $\beta^*$  für jede Gruppe bzw. Cluster gesondert zu schätzen und die Ergebnisse im Anschluss auf Gleichheit zu testen.

Als angenommene Verteilung der  $n_i$ -dimensionalen Zufallsvektoren  $Y_i$  wird, wie bereits in den letzten Abschnitten angedeutet, die multivariate Normalverteilung benutzt. Nach Beispiel 3.4 ist diese Verteilung ein Mitglied einer linearen Exponentialfamilie. Ohne Beschränkung der Allgemeinheit sei  $n_i = n$  für alle  $i \in \{1, \dots, m\}$ .

Im Folgenden werden zur Herleitung der Verallgemeinerten Schätzgleichung die beiden Schritte der QGPML-Methode umgesetzt:

(1) Die angenommene Verteilung der Zufallsvariablen  $Y_i$  sei eine multivariate Normalverteilung mit dem Erwartungswert  $\mu_i = \mu(X_i, \beta)$  und der Kovarianzmatrix  $V_i$ , d.h.  $Y_i \sim \mathcal{N}_n(\mu(X_i, \beta^*), V_i)$  für alle  $i \in \{1, \dots, m\}$ .

Die Varianz der wahren Verteilung von  $Y_{ij}$ ,  $i \in \{1, \dots, m\}$  und  $j \in \{1, \dots, n\}$ , ist nach den Modellannahmen eines marginalen Modells und Satz 2.2 gegeben durch

$$\text{Var}(Y_{ij}) = \psi''(\theta_{ij})\phi.$$

Durch diese Tatsache motiviert, wird als angenommene Kovarianzmatrix die Matrix

$$V_i := A_i^{1/2} R(\alpha) A_i^{1/2},$$

mit einer Korrelationsmatrix  $R(\alpha)$  und der Matrix der marginalen Varianzen  $A_i = A_i(\beta) = \text{diag}(\psi''(\theta_{i1}), \dots, \psi''(\theta_{in_i}))$  bzw. im Fall von Überdispersion  $A_i^\phi = A_i^\phi(\beta, \phi) = \phi \text{diag}(\psi''(\theta_{i1}), \dots, \psi''(\theta_{in_i}))$  verwendet. Somit gilt  $V_i = V_i(\beta, \alpha)$  bzw. im Fall von Überdispersion  $V_i = V_i(\beta, \alpha, \phi)$ .

Nun muss nach den Ausführungen in Abschnitt 3.1.1.5.2 die Kovarianzmatrix geschätzt werden, d.h. es werden Schätzer für  $\beta$ ,  $\alpha$  und gegebenenfalls für  $\phi$  gesucht. Ein konsistenter Schätzer  $\tilde{\beta}$  für  $\beta$  kann durch die Anwendung der PML-Methode bestimmt werden.<sup>22</sup> Des Weiteren können anhand von Tabelle 3.1 ein konsistenter Schätzer  $\tilde{\alpha}$  für  $\alpha$  und mit Hilfe von Definition 3.1 ein konsistenter Schätzer  $\tilde{\phi}$  für  $\phi$  ermittelt werden. Eine Schätzung für die Kovarianzmatrix  $V_i$  ist daher gegeben durch  $\hat{V}_i := V_i(\tilde{\beta}, \tilde{\alpha})$  bzw. im Fall von Überdispersion durch  $\hat{V}_i^\phi := V_i(\tilde{\beta}, \tilde{\alpha}, \tilde{\phi})$ . Diese

---

<sup>22</sup>Dies kann beispielsweise mit der Verteilungsannahme  $Y_i \sim \mathcal{N}_n(\mu(X_i, \beta), I_{n \times n})$  erreicht werden. Vgl. hierzu auch Bemerkung 3.13.

geschätzten Kovarianzmatrizen werden in Ziegler (1999) auch *Arbeitsvarianzmatrizen* genannt. Somit ist der erste Schritt der QGPML-Methode abgeschlossen und es kann mit Schritt 2 fortgefahren werden.

(2) Im zweiten Schritt der QGPML-Methode wird mit der im ersten Schritt geschätzten Arbeitsvarianzmatrix folgende Annahme getroffen:<sup>23</sup>

$$Y_i \sim \mathcal{N}_n \left( \mu(X_i, \beta), \hat{V}_i^\phi \right).$$

Basierend auf dieser Verteilung kann dann nach Satz 3.20 ein stark konsistenter und effizienterer Schätzer für  $\beta$  bestimmt werden.

Die Loglikelihood-Funktion von  $Y$  ist mit  $\tilde{\Psi}_i = \hat{V}_i^\phi$ ,  $\mu_i = \mu(X_i, \beta)$  und nach Definition 3.11 sowie Beispiel 3.4 gegeben durch

$$\begin{aligned} & \ell_y(\beta) \\ &= \frac{1}{m} \sum_{i=1}^m -\frac{1}{2} \mu_i^\top \tilde{\Psi}_i^{-1} \mu_i - \frac{n}{2} \ln(2\pi) - \frac{1}{2} y_i^\top \tilde{\Psi}_i^{-1} y_i - \frac{1}{2} \ln(\det(\tilde{\Psi}_i)) + \mu_i^\top (\tilde{\Psi}_i^\top)^{-1} y_i \\ &= \frac{1}{m} \sum_{i=1}^m -\frac{1}{2} \left[ y_i^\top \tilde{\Psi}_i^{-1} y_i - y_i^\top \tilde{\Psi}_i^{-1} \mu_i - \mu_i^\top \tilde{\Psi}_i^{-1} y_i + \mu_i^\top \tilde{\Psi}_i^{-1} \mu_i \right] - w_i \\ &= \frac{1}{m} \sum_{i=1}^m -\frac{1}{2} \left[ (y_i - \mu_i)^\top \tilde{\Psi}_i^{-1} (y_i - \mu_i) \right] - w_i \\ &= \frac{1}{m} \sum_{i=1}^m -\frac{1}{2} \left[ (y_i - \mu_i)^\top \left( \hat{V}_i^\phi \right)^{-1} (y_i - \mu_i) \right] - w_i, \end{aligned}$$

wobei  $w_i = \frac{n}{2} \ln(2\pi) - \frac{1}{2} \ln(\det(\tilde{\Psi}_i))$  für alle  $i \in \{1, \dots, m\}$  ist.

Da die geschätzte Kovarianzmatrix  $\hat{V}_i^\phi$  unabhängig von  $\beta$  ist, ergibt sich mit der Produktregel folgende Score-Gleichung:

$$S(\beta) = \frac{1}{m} \sum_{i=1}^m \left( \frac{\partial \mu_i}{\partial \beta} \right)^\top \left( \hat{V}_i^\phi \right)^{-1} (y_i - \mu_i) = 0.$$

Diese Gleichung entspricht der in Abschnitt 3.1.1.1 vorgestellten Verallgemeinerten Schätzgleichung (*generalized estimating equation*, GEE) von Liang und Zeger (1986). Insbesondere folgen daher die behaupteten Eigenschaften der GEE-Schätzer in Satz 3.2 mit den Eigenschaften der QGPML-Schätzer in Satz 3.20.

Damit wurde in diesem Abschnitt für randomisierte Studien die Äquivalenz der QGPML-Schätzgleichungen und der Verallgemeinerten Schätzgleichungen skizziert. Darüber hinaus wird durch den ersten Schritt der QGPML-Methode deutlich, weshalb die Korrelationsstruktur beliebig wählbar ist und die Methode von Liang und

<sup>23</sup>Der Fall ohne Überdispersion folgt mit  $\phi = 1$  und wird daher im Folgenden vernachlässigt.

Zeger (1986) stets konsistente Schätzer liefert.

Abschließend wird die schuldig gebliebene Legitimation für die auf Seite 26 gegebene Iterationsvorschrift nachgereicht. Mit dem Wissen aus den letzten drei Abschnitten lässt sich die Iterationsvorschrift mit diesen Rechtfertigungen versehen:

1. (A. Erster Schritt der QGPML Methode) Bestimme einen stark konsistenten Anfangswert  $\beta_0$  für  $\beta$  unter der Annahme, dass die einzelnen Beobachtungen unabhängig sind. Die Iterationsvorschrift ist dabei nach Gleichung (3.18) durch

$$\hat{\beta}_0^{n+1} = \hat{\beta}_0^n + \left[ \sum_{i=1}^m \left( \frac{\partial \mu_i}{\partial \beta} \right)^\top A_i^{-1} \left( \frac{\partial \mu_i}{\partial \beta} \right) \right]^{-1} \sum_{i=1}^m \left( \frac{\partial \mu_i}{\partial \beta} \right)^\top A_i^{-1} (y_i - \mu_i).$$

gegeben.

2. (B. Erster Schritt der QGPML-Methode) Ausgehend von dem ermittelten Anfangswert  $\beta_0$  kann nach Definition 3.1 ein konsistenter Schätzer  $\hat{\phi}_0$  für  $\phi$  bestimmt werden. Entsprechend der gewählten Arbeitskorrelationsstruktur kann nach Tabelle (3.1) ebenfalls der konsistente Schätzer  $\hat{\alpha}_0$  für  $\alpha$  berechnet werden.
3. (C. Erster Schritt der QGPML-Methode) Basierend auf  $\beta_0$ ,  $\hat{\alpha}_0$  und  $\hat{\phi}_0$  kann  $R_i(\hat{\alpha}_0)$  und mit Gleichung (3.19) auch  $\hat{V}_i^{\phi_0}$  durch

$$\hat{V}_i^{\phi_0} = A_i^{1/2} \left( \beta_0, \hat{\phi}_0 \right) R_i(\hat{\alpha}_0) A_i^{1/2} \left( \beta_0, \hat{\phi}_0 \right)$$

geschätzt werden. Der Schätzer der Kovarianzmatrix  $\hat{V}_i^{\phi_0}$  entspricht dem Schätzer von  $\Psi$ , der im ersten Schritt der QGPML-Methode berechnet wird.

4. (Zweiter Schritt der QGPML-Methode) Mit Hilfe der Schätzungen für  $V_i^\phi$  und dem Startwert  $\beta_0$  kann das Fisher-Scoring-Verfahren gemäß Vorschrift (3.21) durchgeführt werden.

Damit erhält man einen neuen Schätzer für  $\beta$ , der durch  $\beta_1$  gegeben sei. Ausgehend von diesem werden die Schritte zwei, drei und vier erneut durchgeführt. Dieses Vorgehen wird iterativ fortgesetzt, bis das Verfahren eine vorgegebene Genauigkeit erreicht.

## 3.2 Subjekt-spezifische Modelle

Die subjekt-spezifischen Modelle unterscheiden sich von den marginalen Modellen durch die Einbeziehung von spezifischen subjekt-gebundenen Größen  $U_i$ . Diese spiegeln die natürliche Heterogenität der Subjekte wider, die auf nicht-messbare Faktoren, wie etwa der genetischen Veranlagung von Patienten oder deren Lebensumständen, beruhen kann.

Folglich unterscheidet sich die Interpretation der Regressionsparameter bei beiden Modellen: Während die Regressionsparameter im marginalen Modell eine Aussage über den Populationsdurchschnitt liefern, konzentriert sich die Aussagekraft des subjekt-spezifischen Modells auf einzelne Subjekte. Erst eine anschließende Marginalisierung des Modells ermöglicht Aussagen über den Populationsdurchschnitt, doch darauf soll erst in Kapitel 4 eingegangen werden.

In der Literatur wird für subjekt-spezifische Modelle oftmals auch der Begriff *random effects models*, d.h. Modelle mit zufälligen Effekten, verwendet.<sup>24</sup> Dies ist allerdings nur eine von drei unterschiedenen Teilfamilien der subjekt-spezifischen Modelle. Diese drei Teilfamilien unterscheiden sich durch verschiedene Handhabungen der subjekt-spezifischen Parameter  $U_i$ , da diese

1. als feste unbekannte Effekte,
2. als störende Effekte oder
3. als zufällige Effekte

betrachtet werden können.

Das allgemeine zugrundeliegende Modell der subjekt-spezifischen Modelle kann nach Diggle u. a. (1996) wie folgt formuliert werden:

Wie bereits im marginalen Modell bezeichne der  $n_i$ -dimensionale Vektor  $Y_i$  den zum Subjekt  $i$  gehörenden Zufallsvektor für alle  $i \in \{1, \dots, m\}$ .  $U_i$  sei der zum  $i$ -ten Subjekt gehörende subjekt-spezifische Parametervektor. Für das Modell werden jetzt folgende Annahmen getroffen:

- Bedingt auf den Größen  $U_i$  sind die Zufallsgrößen  $Y_{i1}, \dots, Y_{in_i}$  paarweise unabhängig und besitzen für alle  $j \in \{1, \dots, n_i\}$  die Verteilung einer 1-parametrischen natürlichen Exponentialfamilie. Für die Dichte von  $Y_{ij}$ ,  $i \in \{1, \dots, m\}$  und  $j \in \{1, \dots, n_i\}$ , gilt daher

$$f_{Y_{ij}|U_i=u_i}(y_{ij}) \stackrel{\text{Poisson}}{=} \exp[(y_{ij} \theta_{ij} - \psi(\theta_{ij})) \zeta^{-1} + c(y_{ij}, \zeta)] \quad (3.42)$$

$$\exp\{y_{ij} \ln(\lambda_{ij}) - \lambda_{ij} - \ln(y_{ij}!)\}.$$

<sup>24</sup>Vgl. zum Beispiel Diggle u. a. (1996).

- Die bedingten Momente  $\mu_{ij} = E(Y_{ij}|U_i) = \psi'(\theta_{ij}) \stackrel{Poisson}{=} \lambda_{ij}$  und  $v_{ij} = \text{Var}(Y_{ij}|U_i)$  erfüllen das verallgemeinerte lineare Modell

$$h(\mu_{ij}) = x_{ij}^\top \beta + d_{ij}^\top U_i,$$

wobei  $h$  eine bekannte Linkfunktion ist und im Fall poissonverteilter Daten beispielsweise durch den natürlichen Logarithmus ersetzt werden kann. Darüber hinaus gilt  $v_{ij} = v(\mu_{ij}) \phi = \psi''(\theta_{ij}) \phi \stackrel{Poisson}{=} \lambda_{ij} \phi$  mit einer bekannten Varianzfunktion  $v$  und einem unbekanntem Überdispersions-Parameter  $\phi$ . Die Größe  $d_{ij}$  sei ein  $p$ -dimensionaler Vektor mit Komponenten aus dem erklärenden  $q$ -dimensionalen Parametervektor  $x_{ij}$ .

- Die Parametervektoren  $U_i$  sind entweder fest oder zufällig und für alle  $i \in \{1, \dots, m\}$  paarweise unabhängig und identisch verteilt gemäß einer Verteilung  $G$ . Darüber hinaus seien die Größen  $U_i$  unabhängig von den erklärenden Variablen  $x_{ij}$ .

Die Einführung der Parameter  $U_i$  ermöglicht die Modellierung der subjekt-spezifischen Einflüsse, sowie die Modellierung der Abhängigkeit der Daten eines Subjekts, da alle Beobachtungen eines Subjekts die Größe  $U_i$  gemein haben. Damit ist das subjekt-spezifische Modell in gewissem Sinne eine einfache Erweiterung der verallgemeinerten linearen Modelle, da die Effekte auch zufällig sein können. Aus diesem Grund entspricht die Theorie der subjekt-spezifischen Modelle mit zufälligem  $U_i$  im Wesentlichen der Theorie sogenannter *verallgemeinerter linearer gemischter Modelle* (*generalized linear mixed models*, GLMM).

Im Folgenden werden die drei bereits erwähnten Betrachtungsweisen der Parameter  $U_i$  einzeln behandelt und die jeweilige Vorgehensweise kurz erläutert. Hierbei liegt der Schwerpunkt auf der Betrachtung von zufälligen Größen  $U_i$ . Anschließend werden Schätzmöglichkeiten für die Parameter  $U_i$  und  $\theta_{ij} = \theta_{ij}(\beta)$  vorgestellt.

### 3.2.1 Drei Möglichkeiten für die Handhabung der $U_i$

#### 3.2.1.1 Die subjekt-spezifischen Größen als feste unbekannte Parameter

Im Fall von festen subjekt-spezifischen Größen werden für die Bestimmung des Regressionsvektors  $\beta$  Schätzungen der Größen  $\theta_{ij}$  und  $U_i$  für alle  $i \in \{1, \dots, m\}$  und  $j \in \{1, \dots, n_i\}$  benötigt. Diese können mit Hilfe der klassischen Maximum-Likelihood-Schätzung für

$$\prod_{i=1}^m \prod_{j=1}^{n_i} f_{Y_{ij}|U_i=u_i}(y_{ij})$$

bestimmt werden. Nach Molenberghs und Verbeke liefert dieses Verfahren bei großem Datenumfang allerdings inkonsistente Schätzer, so dass alternative Verfahren, in denen beispielsweise die  $U_i$  entfernt werden, entwickelt wurden.<sup>25</sup>

### 3.2.1.2 Die subjekt-spezifischen Größen als Störparameter

Auch im Fall der subjekt-spezifischen Größen als Störparameter ist die klassische Maximum-Likelihood-Methode aufgrund des Aufwands keine sinnvolle Methode zur Ermittlung von Schätzern. Alternativ wird von Diggle u. a. (1996) vorgeschlagen, die bedingte Maximum-Likelihood-Methode anzuwenden. Bei diesem Verfahren werden die  $U_i$  als *überflüssige* Variablen behandelt und *suffiziente Schätzer* für die  $U_i$  ermittelt. Diese Schätzer können für alle  $i \in \{1, \dots, m\}$  gewählt werden als

$$\hat{U}_i = \sum_{j=1}^{n_i} d_{ij} y_{ij}.$$

Mittels dieser Schätzungen und der Maximierung der bedingten Verteilung von  $Y_i$  unter  $U_i$ , d.h.

$$f_{Y_i|U_i=\sum_{j=1}^{n_i} d_{ij} y_{ij}}(y_i) = \frac{f_{Y_i, U_i=\sum_{j=1}^{n_i} d_{ij} y_{ij}}(y_{ij})}{f_{U_i}(\sum_{j=1}^{n_i} d_{ij} y_{ij})},$$

können die Schätzer für  $\theta_{ij}$ ,  $i \in \{1, \dots, m\}$  und  $j \in \{1, \dots, n_i\}$ , bestimmt werden. Ausgehend von den Schätzern für  $U_i$  und  $\theta_{ij}$  kann dann ein Schätzer für den Regressionsparameter  $\beta$  berechnet werden, weil die Gleichheit

$$h(\mu_{ij}) = h(\psi'(\theta_{ij})) = x_{ij}^\top \beta + d_{ij}^\top U_i$$

mit bekannten Funktionen  $h$  und  $\psi$  gilt.

Dieses Vorgehen hat sowohl Vor- als auch Nachteile. Ein entscheidender Vorteil ist, dass keine weiteren Informationen über die subjekt-spezifischen Größen vorgegeben sein müssen. Allerdings werden suffiziente Statistiken für alle  $U_i$  benötigt, und es muss die bedingte Dichtefunktion maximiert werden. Dies kann bei einem Datensatz mit einer hohen Anzahl an Subjekten sehr aufwendig werden. Darüber hinaus geht jegliche Aussagekraft über die subjekt-spezifischen Parameter verloren, so dass diese Methode bei Interesse an den  $U_i$  als nicht sinnvoll erscheint.

### 3.2.1.3 Die subjekt-spezifischen Größen als zufällige Parameter

Der Ansatz bei subjekt-spezifischen Größen als zufällige Parameter ist dann anzuwenden, wenn der Effekt der subjekt-spezifischen Größen von Interesse ist oder die

<sup>25</sup>Vgl. Molenberghs und Verbeke (2005), S. 258.

Beseitigung dieser Größen zu einem zu hohen Informationsverlust für den Regressionsparameter  $\beta$  oder einzelnen Komponenten des Regressionsparameters führen würde. Die Vektoren  $U_i$  sind in diesem Fall unabhängige Zufallsvektoren und gemäß einer Verteilung  $G$  identisch verteilt. Bei wiederholten Zählraten können für  $G$  beispielsweise die Normalverteilung, die Gammaverteilung und als Spezialfall insbesondere die Exponentialverteilung in Betracht gezogen werden.

Interessiert hauptsächlich der Regressionsparameter  $\beta$ , so ist die Elimination der  $p$ -dimensionalen Zufallsgrößen  $U_i$  durch Integration der gemeinsamen Dichte von  $Y = (Y_1^\top, \dots, Y_m^\top)^\top$  und  $U = (U_1^\top, \dots, U_m^\top)^\top$  das vordergründige Ziel. Zu diesem Zweck ist es notwendig, die Randverteilung von  $Y$  zu berechnen:

Die gemeinsame Verteilung von  $Y$  und dem  $p$ -dimensionalen Zufallsvektor  $U_i$  ist nach dem Satz von Fubini<sup>26</sup> gegeben durch

$$F_{Y,U_i}(y, u_i) = \int_{-\infty}^{y_{11}} \cdots \int_{-\infty}^{y_{m n_m}} \int_{-\infty}^{u_{i1}} \cdots \int_{-\infty}^{u_{ip}} f_{Y,U_i}(s, t) dt_1 \dots dt_p ds_1 \dots ds_k$$

mit  $t = (t_1, \dots, t_p)$ ,  $s = (s_1, \dots, s_k)$  und  $k = \sum_{i=1}^m n_i$ . Zudem ergibt sich für den Zufallsvektor  $Y_i$  die Randverteilung

$$F_{Y_i}(y_i) = P(Y_i \leq y_i) = \int_{-\infty}^{y_{i1}} \cdots \int_{-\infty}^{y_{in_i}} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{Y_i,U_i}(s, t) dt_1 \dots dt_p ds_1 \dots ds_{n_i}$$

mit  $s = (s_1, \dots, s_{n_i})$ .

Die dazugehörige Dichte kann durch Differentiation nach  $y_i = (y_{i1}, \dots, y_{in_i})^\top$  ermittelt werden:

$$\begin{aligned} f_{Y_i}(y_i) &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{Y_i,U_i}(y_i, t) dt_1 \dots dt_p \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{Y_i|U_i=t}(y_i) f_{U_i}(t) dt_1 \dots dt_p \\ &= \int f_{Y_i|U_i=t}(y_i) f_{U_i}(t) dt \\ &= \int f_{Y_i|U_i=u_i}(y_i) f_{U_i}(u_i) du_i. \end{aligned} \tag{3.43}$$

Basierend auf dieser marginalen Dichte von  $Y_i$  und der vorausgesetzten Unabhängigkeit von  $Y_{i1}, \dots, Y_{in_i}$  bei gegebenen  $U_i$  folgt

$$f_{Y_i}(y_i) = \int \prod_{j=1}^{n_i} f_{Y_{ij}|U_i=u_i}(y_{ij}) f_{U_i}(u_i) du_i. \tag{3.44}$$

<sup>26</sup>Vgl. Schmieder (1994).

Zudem können jetzt die Parameter  $\theta_{ij} \stackrel{\text{Poisson}}{=} \ln(\lambda_{ij})$  mit Hilfe der Maximum-Likelihood-Methode geschätzt werden. Die Likelihood-Funktion für  $Y$  ist dabei erklärt durch

$$L_y(\delta) = \prod_{i=1}^m f_{Y_i}(y_i) = \prod_{i=1}^m \int \prod_{j=1}^{n_i} f_{Y_{ij}|U_i=u_i}(y_{ij}) f_{U_i}(u_i) du_i,$$

wobei  $\delta$  ein Parametervektor ist, der die Parameter der Verteilung  $G$  und die gesamten  $\theta_{ij} = \theta_{ij}(\beta)$  als Komponenten umfasst. Die Maximierung dieses Ausdrucks kann bei einem großen Datensatz recht aufwendig sein, da zunächst  $m$  Integrale über die mehrdimensionalen Größen  $U_i$ ,  $i \in \{1, \dots, m\}$ , gelöst werden müssen. Darüber hinaus existieren für die Berechnung der Integrale nur in wenigen Spezialfällen geschlossene Ausdrücke. Aus diesem Grund wurden zahlreiche Verfahren und Algorithmen entwickelt, um die Loglikelihood-Schätzung von  $\delta$  durch das Maximieren von

$$l_y(\delta) = \ln(L_y(\delta)) = \sum_{i=1}^m \ln \left( \int \prod_{j=1}^{n_i} f_{Y_{ij}|U_i=u_i}(y_{ij}) f_{U_i}(u_i) du_i \right) \quad (3.45)$$

zu ermitteln. Dies wird Gegenstand der Betrachtungen im folgenden Abschnitt sein.

Neben dem Regressionsvektor  $\beta$  können auch die subjekt-spezifischen Größen  $U_i$ ,  $i \in \{1, \dots, m\}$ , von Interesse sein. Dies ist beispielsweise der Fall, wenn Vorhersagen über subjekt-spezifische Entwicklungen getroffen werden sollen, oder wenn die Varianz innerhalb der Daten eines Subjekts von Interesse ist, da diese einen Indikator für die Existenz von Ausreißern darstellt. Zu diesem Zweck kann nach Molenberghs und Verbeke (2005) die *posteriori*-Verteilung der subjekt-spezifischen Größen, d.h.

$$f_{U_i|Y_i=y_i}(u_i) = \frac{\prod_{i=1}^{n_i} f_{Y_{ij}|U_i=u_i}(y_{ij}) f_{U_i}(u_i)}{\int \prod_{i=1}^{n_i} f_{Y_{ij}|U_i=u_i}(y_{ij}) f_{U_i}(u_i) du_i}, \quad (3.46)$$

betrachtet werden. Anschließend können dann mittels der Maximum-Likelihood-Methode die sogenannten *empirischen Bayes-Schätzer* bestimmt werden.

### 3.2.2 Schätzmethode im Fall zufälliger subjekt-spezifischer Parameter

Wie bereits im letzten Abschnitt eingeführt, wird ein subjekt-spezifisches Modell u.a. durch die Beziehung

$$h(\mu_{ij}) = x_{ij}^\top \beta + d_{ij}^\top U_i \quad (3.47)$$

mit bekannter Linkfunktion  $h$  und den Größen  $\mu_{ij} = E(Y_{ij}|U_i) = \psi'(\theta_{ij}) \stackrel{\text{Poisson}}{=} \lambda_{ij}$  und  $v_{ij} = \text{Var}(Y_{ij}|U_i) = \psi''(\theta_{ij}) \phi \stackrel{\text{Poisson}}{=} \lambda_{ij} \phi$  charakterisiert. Die subjekt-spezifischen Größen sind zufällig und identisch verteilte Zufallsgrößen, so dass die

Gleichung (3.47) ein sogenanntes *generalized linear mixed model* (GLMM) darstellt. Im letzten Abschnitt wurde bereits die Loglikelihood-Funktion

$$l_y(\delta) = \sum_{i=1}^m \ln \left( \int \prod_{j=1}^{n_i} f_{Y_{ij}|U_i=u_i}(y_{ij}) f_{U_i}(u_i) du_i \right) \quad (3.48)$$

eingeführt. Um  $\delta$  schätzen zu können, wird diese Funktion abgeleitet und die so erhaltene Score-Funktion gleich Null gesetzt. Zu diesem Zweck müssen zunächst die Integrale im Ausdruck der Loglikelihood-Funktion bestimmt werden. Dies ist im Allgemeinen sehr aufwendig, da es  $m$  mehrdimensionale Integrale zu lösen gilt. In Spezialfällen, wie bei der Betrachtung normalverteilter  $Y_{ij}$ , besitzen diese Integrale eine geschlossene analytische Form; bei nicht-normalverteilten Fällen ist die Berechnung der Integrale allerdings komplexer.

Im Fall normalverteilter subjekt-spezifischer Größen  $U_i$  können die Integrale mit Hilfe von numerischen Approximationsmethoden bestimmt werden.<sup>27</sup> Diese Methoden lassen sich in zwei Gruppen unterteilen, da entweder die Daten, oder die Integrale angenähert werden können.

Sind die subjekt-spezifischen Größen nicht normalverteilt, fehlt es leider an passenden Approximationsmethoden und Implementierungen in Software-Paketen. Allerdings vereinfacht sich im Fall von gammaverteilten  $U_i$  die marginale Dichte von  $Y_{ij}$ ,  $i \in \{1, \dots, m\}$  und  $j \in \{1, \dots, n_i\}$ , zu der Dichte einer Negativ-Binomialverteilung, so dass sich zumindest die Berechnung der mehrdimensionalen Integrale erübrigt.

### 3.2.2.1 Normalverteilte subjekt-spezifische Größen

Wie bereits erwähnt, sind im Fall normalverteilter  $U_i$  zwei Gruppen von Approximationsmethoden für die Berechnung der Integrale in Gleichung (3.44) anwendbar. Die eine Methode beschäftigt sich mit der Approximation der Daten und die andere mit der Approximation der Integrale durch die *Hermite-Gauß-Quadratur*.

Bei der Approximation der Daten werden diese durch normalverteilte sogenannte *Pseudodaten* ersetzt. Für die *Pseudodaten* kann dann ein lineares Modell formuliert werden, und wie oben bereits erwähnt, lassen sich die Integrale für normalverteilte Daten analytisch lösen. Molenberghs und Verbeke gehen in ihrer Arbeit auf zwei Möglichkeiten zur Erzeugung der *Pseudodaten* ein.<sup>28</sup> Die Umsetzung dieser Methoden ist im Statistik-Programm **SAS** durch die Prozedur **GLIMMIX** gegeben. Allerdings muss diese Prozedur gesondert eingebunden werden und war der Autorin nicht zugänglich. Daher wird im Rahmen dieser Arbeit die alternative Methode zur

---

<sup>27</sup>Vgl. Molenberghs und Verbeke (2005), S. 267ff.

<sup>28</sup>Vgl. Molenberghs und Verbeke (2005), S. 269ff.

Approximation der Integrale vorgestellt, da diese in der vorinstallierten Prozedur NLMIXED enthalten ist.

In der Numerik sind zahlreiche Methoden zur Approximation von Integralen bekannt und viele dieser Verfahren sind in Statistik-Programmen implementiert. Angelehnt an die Vorgehensweise von Pinheiro und Bates werden in diesem Abschnitt die *Hermite-Gauß-Quadratur* und die *adaptive Hermite-Gauß-Quadratur* zur Approximation der Integrale in Gleichung (3.48) vorgestellt.<sup>29</sup> Diese werden auch in der SAS-Prozedur NLMIXED verwendet.

Die Gauß'sche-Quadratur wurde zur numerischen Berechnung gewichteter Integrale

$$I(f) = \int_a^b f(z) \phi(z) dz \quad (3.49)$$

entwickelt, wobei  $f : [a, b] \rightarrow \mathbb{R}$  die zu integrierende Funktion mit  $a, b \in \mathbb{R} \cup \{-\infty, \infty\}$  und die Gewichtsfunktion  $\phi$  eine gegebene positive, stückweise stetige Funktion ist. Dieses Integral wird durch die gewichtete Summe

$$I_Q(f) = \sum_{q=1}^Q \omega_q f(z_q) \quad (3.50)$$

approximiert. Dabei bezeichnen die  $\omega_q$ ,  $q \in \{1, \dots, Q\}$ , die zu wählenden Gewichte und  $z_q$  Stützstellen, die eine Zerlegung des Integrationsbereichs liefern. Die Gauß-Quadratur basiert auf einer guten Wahl der Stützstellen und einer ausreichend großen Anzahl von Quadraturpunkten  $Q$ , welche die Genauigkeit der Approximation determinieren. Bei der herkömmlichen Gauß-Quadratur werden die Stützstellen lediglich abhängig von der Gewichtsfunktion  $\phi(z)$  berechnet. Alternativ ermöglicht die adaptive Gauß-Quadratur eine höhere Genauigkeit mit einer geringeren Anzahl an Quadraturpunkten, indem zur Berechnung der Stützstellen auch Informationen über die Funktion  $f$  genutzt werden. Aus der Numerik ist bekannt, dass die Wahl der Stützstellen als Nullstellen gewisser orthogonaler Polynome die beste ist.

Im Fall der Gleichung (3.48)

$$l_y(\delta) = \sum_{i=1}^m \ln \left( \int \prod_{j=1}^{n_i} f_{Y_{ij}|U_i=u_i}(y_{ij}) f_{U_i}(u_i) du_i \right)$$

geht es zunächst darum, das Integral

$$\int \prod_{j=1}^{n_i} f_{Y_{ij}|U_i=u_i}(y_{ij}) f_{U_i}(u_i) du_i = \int f_{Y_i|U_i=u_i}(y_i) f_{U_i}(u_i) du_i$$

---

<sup>29</sup>Vgl. Pinheiro und Bates (1995).

zu lösen und durch eine anschließende Maximum-Likelihood-Schätzung den Parametervektor  $\delta$  zu bestimmen. Dieses Integral lässt sich für den Fall eindimensionaler subjekt-spezifischer Größen, d.h.  $p = 1$ , auch schreiben als

$$\int_{-\infty}^{\infty} f_{Y_i|U_i=u_i}(y_i) f_{U_i}(u_i) du_i,$$

wobei die Funktion  $f_{U_i}(u_i)$  der Dichte einer Normalverteilung entspricht. Unter der Annahme, dass die  $U_i, i \in \{1, \dots, m\}$ , normalverteilt sind, also  $U_i \sim \mathcal{N}(\mu, \sigma^2)$  ergibt sich durch die Substitution von

$$u := \frac{u_i - \mu}{\sigma}$$

die Gleichung

$$\int_{-\infty}^{\infty} f_{Y_i|U_i=u_i}(y_i) f_{U_i}(u_i) du_i = \int_{-\infty}^{\infty} g(u) \underbrace{\exp(-u^2)}_{\phi(u)} du \quad (3.51)$$

mit der Funktion

$$\begin{aligned} g: \mathbb{R} &\rightarrow \mathbb{R} \\ u &\mapsto f_{Y_i|U_i=u\sigma+\mu}(y_i) \frac{1}{\sqrt{2\pi}} \exp\left(\frac{u^2}{2}\right). \end{aligned}$$

Die Gewichtsfunktion in Gleichung (3.49) entspricht somit der Funktion

$$\phi(z) = \exp(-z^2).$$

Die Gauß-Quadratur wird in diesem Fall auch Hermite-Gauß-Quadratur genannt. Um das Integral in Gleichung (3.51) durch eine gewichtete Summe zu approximieren, müssen u.a. die Stützstellen bestimmt werden. Daher stellt sich die Frage, welche Polynome bezüglich des gewichteten Skalarprodukts  $\langle \cdot, \cdot \rangle_\phi$  orthogonal zueinander sind. Dabei ist das gewichtete Skalarprodukt wie folgt definiert:

**Definition 3.21.**

Sei  $\phi : \mathbb{R} \rightarrow \mathbb{R}$  eine positive, stückweise stetige Funktion. Auf dem Vektorraum  $\mathcal{P} = \{p(x) | p(x) = a_0 + a_1x + a_2x^2 + \dots, \text{ mit } a_i \in \mathbb{R} \forall i \in \mathbb{N}_0\}$  aller Polynome über  $\mathbb{R}$  wird für  $g, h \in \mathcal{P}$  und  $a, b \in \mathbb{R}$  mit  $a < b$  das gewichtete Skalarprodukt  $\langle g, h \rangle_\phi$  durch

$$\langle g, h \rangle_\phi := \int_a^b g(x) f(x) \phi(x) dx$$

definiert.

**Definition 3.22.**

Sei  $n \in \mathbb{N}_0$ , dann ist das  $n$ -te Hermite-Polynom definiert durch

$$H_n(x) := (-1)^n e^{x^2} \frac{\partial^n e^{-x^2}}{\partial^n x}.$$

Im Anhang A wird gezeigt, dass die *Hermite-Polynome* bezüglich des gewichteten Skalarprodukts orthogonal zur Funktion  $\phi(x) = \exp(-x^2)$  sind. Mit Hilfe dieser Polynome kann in den nächsten Abschnitten sowohl die herkömmliche *Hermite-Gauß-Quadratur*, als auch die *adaptive Hermite-Gauß-Quadratur* erläutert werden.

**3.2.2.1.1 Hermite-Gauß-Quadratur**

Die folgende Definition beschreibt die Hermite-Gauß'sche Quadraturformel an  $n$  Stützstellen mit positiven Gewichten und Genauigkeitsgrad  $2n - 1$ .

**Definition 3.23.**

Sei  $n \in \mathbb{N}$  und seien  $z_1, \dots, z_n$  die Nullstellen des  $n$ -ten Hermite-Polynoms. Die Gewichte für alle  $j \in \{1, \dots, n\}$  seien durch

$$\omega_j := \langle L_j, 1 \rangle_\phi$$

definiert, wobei  $L_j$  das durch

$$L_j(x) := \prod_{\substack{i=1 \\ i \neq j}}^n \frac{x - z_i}{z_j - z_i}, \quad j \in \{1, \dots, n\}$$

definierte Lagrange'sche Basispolynom bezeichne. Zudem sei  $\phi(x) = \exp(-x^2)$  die positive, stückweise stetige Gewichtsfunktion.

Für das Integral

$$I(f) = \int_a^b f(x) \phi(x) dx$$

wird die numerische Integrationsformel

$$I_n(f) = \sum_{j=1}^n \omega_j f(z_j)$$

die *Hermite-Gauß-Quadratur* der  $n$ -ten Ordnung bezüglich der Gewichtsfunktion  $\phi$  genannt.

Es kann bewiesen werden, dass diese Integrationsformel für alle beliebig oft differenzierbaren Funktionen asymptotisch exakt ist. Zu diesem Zweck reicht es zu zeigen, dass die Integrationsformel für jedes Polynom exakt ist, denn eine Funktion  $g \in C^\infty$  lässt sich mittels Taylor-Entwicklung beliebig genau durch ein Polynom approximieren. Bevor der entsprechende Satz formuliert und bewiesen wird, muss allerdings noch die *Lagrange'sche Interpolationsformel* vorgestellt werden.

**Satz 3.24.** (Lagrange'sche Interpolationsformel)

Sei  $\mathcal{P}_n$  der Vektorraum aller Polynome, deren Grad kleiner oder gleich  $n$  ist und  $f$  eine stetige Funktion. Zudem seien  $(x_0, f(x_0)), \dots, (x_n, f(x_n))$  paarweise verschiedene Stützpunkte. Dann existiert genau ein interpolierendes Polynom  $p \in \mathcal{P}_n$  mit

$$p(x_j) = f(x_j) \quad \text{für alle } j \in \{1, \dots, n\}.$$

Dieses Polynom wird mit  $p_n(x)$  bezeichnet und ist durch die sogenannte Lagrange'sche Interpolationsformel

$$p_n(x) = \sum_{j=0}^n f(x_j) L_j(x)$$

mit

$$L_j(x) = \prod_{\substack{i=0 \\ i \neq j}}^n \frac{x - x_i}{x_j - x_i}, \quad j \in \{0, \dots, n\}$$

explizit gegeben.

Nun kann die Exaktheit der Approximation der Hermite-Gauß-Quadratur gezeigt werden.

**Satz 3.25.**

Die Hermite-Gauß-Quadratur der Ordnung  $n$  ist für jedes Polynom  $p$  vom Grad kleiner oder gleich  $2n - 1$  exakt, d.h. es gilt

$$\langle p, 1 \rangle_\phi = \sum_{j=1}^n \omega_j p(z_j)$$

mit  $\phi = e^{-x^2}$ .

*Beweis:* Sei  $p$  ein beliebiges Polynom vom Grad kleiner oder gleich  $2n - 1$  und  $H_n$  das  $n$ -te Hermite-Polynom. Dann existieren nach dem Euklidischen Algorithmus Polynome  $q, r \in \mathcal{P}_{n-1}$  mit

$$p(x) = q(x) H_n(x) + r(x).$$

Es gilt  $H_n(z_j) = 0$  für alle  $j \in \{1, \dots, n\}$  und somit auch

$$p(z_j) = r(z_j).$$

Mit Hilfe der Lagrange'schen Interpolationsformel, angewendet auf  $r$ , ergibt sich

$$\begin{aligned} r(x) &= \sum_{j=1}^n r(z_j) L_j(x) \\ &= \sum_{j=1}^n p(z_j) L_j(x). \end{aligned}$$

Zudem ist offensichtlich, dass für den Vektorraum  $\mathcal{P}_n$  gilt:

$$\mathcal{P}_n = \text{span}\{H_0, H_1, \dots, H_n\}.$$

Sei  $\mathcal{P}$  die Menge aller Polynome und  $\mathcal{P}_{n-1}^\perp := \{f \in \mathcal{P} \mid \langle f, g \rangle_\phi = 0 \quad \forall g \in \mathcal{P}_{n-1}\}$ . Dann gilt insbesondere  $H_n(x) \in \mathcal{P}_{n-1}^\perp$  und mit  $q \in \mathcal{P}_{n-1}$  folgt daher

$$\begin{aligned} \langle p, 1 \rangle_\phi &= \langle q H_n + r, 1 \rangle_\phi \\ &= \langle q H_n, 1 \rangle_\phi + \langle r, 1 \rangle_\phi \\ &= \underbrace{\langle q, H_n \rangle_\phi}_{=0} + \left\langle \sum_{j=1}^n p(z_j) L_j(x), 1 \right\rangle_\phi \\ &= \sum_{j=1}^n p(z_j) \langle L_j, 1 \rangle_\phi \\ &= \sum_{j=1}^n \omega_j p(z_j) \end{aligned}$$

und damit die Behauptung. □

Das Korollar 1 ist, wie bereits angedeutet, eine direkte Folgerung aus Satz 3.25.

**Korollar 3.26.**

*Die Hermite-Gauß-Quadraturformel ist für alle  $f \in C^\infty$  asymptotisch exakt, daher gilt*

$$\lim_{n \rightarrow \infty} I_n(f) = \int_a^b f(x) \phi(x) dx.$$

In diesem Abschnitt wurde somit eine Möglichkeit zur Berechnung eines Integrals, wie es in Gleichung (3.51) durch

$$I(g) = \int_{-\infty}^{\infty} g(u) \exp(-u^2) du$$

gegeben ist, vorgestellt. Die Stützstellen  $z_j$  entsprechen dabei den Nullstellen der Hermite-Polynome, die bezüglich des gewichteten Skalarprodukts orthogonal zur Gewichtsfunktion  $\phi$  sind. Damit basiert die Berechnung der Stützstellen und auch der Gewichte  $\omega_j$  lediglich auf der Gewichtsfunktion  $\phi$ .

Die Hermite-Polynome erfüllen die Symmetrie-Relation

$$H_n(-x) = (-1)^n H_n(x).$$

Um dies erkennen zu können, muss die erzeugende Funktion der Hermite-Polynome angegeben werden.

**Satz 3.27.**

Die Hermite-Polynome entsprechen den Koeffizienten der Reihenentwicklung der Funktion

$$w(x, \alpha) := e^{2x\alpha - \alpha^2} = \sum_{n=0}^{\infty} \frac{\alpha^n}{n!} H_n(x), \quad x \in \mathbb{R}.$$

Die Funktion  $w$  wird in diesem Zusammenhang auch *erzeugende Funktion* der Hermite-Polynome genannt.

*Beweis:* Der Beweis ist im Anhang A, S. 161, Satz A.5 gegeben. □

Für die erzeugende Funktion  $w(x, \alpha)$  gilt nun einerseits

$$w(-x, \alpha) = e^{2(-x)\alpha - \alpha^2} = \sum_{n=0}^{\infty} \frac{\alpha^n}{n!} H_n(-x)$$

und andererseits

$$\begin{aligned} w(-x, \alpha) &= e^{2x(-\alpha) - (-\alpha)^2} = \sum_{n=0}^{\infty} \frac{(-\alpha)^n}{n!} H_n(x) \\ &= e^{-2x\alpha - \alpha^2} = \sum_{n=0}^{\infty} \frac{(-1)^n \alpha^n}{n!} H_n(x), \end{aligned}$$

so dass die behauptete Symmetrie-Relation durch Koeffizientenvergleich folgt. Für die Nullstellen der Hermite-Polynome und damit für die Stützstellen der Quadraturformel ergibt sich eine zum Ursprung symmetrische Verteilung. Was geschieht aber, wenn die Funktion  $g(u) \cdot \exp(-u^2)$  nicht symmetrisch ist? Die Genauigkeit der Integral-Approximation lässt sich in diesem Fall zwar durch die Ordnung  $n$  des Hermite-Polynoms steuern, allerdings wären nicht alle Informationen optimal genutzt. In Abbildung 3.1 werden zur Veranschaulichung dieses Problems das Ergebnis einer Integralberechnung durch eine herkömmliche Hermite-Gauß-Quadratur mit dem einer adaptiven Hermite-Gauß-Quadratur verglichen.<sup>30</sup>

Die bessere Approximation einer adaptiven Hermite-Gauß-Quadratur ist dabei unübersehbar und diese wird infolgedessen Gegenstand des nächsten Abschnitts sein.

**3.2.2.1.2 Adaptive Hermite-Gauß-Quadratur**

Das Integral, das approximiert werden soll, ist für  $p = 1$  von der Gestalt

$$\int_{-\infty}^{\infty} f_{Y_i|U_i=u_i}(y_i) f_{U_i}(u_i) du_i,$$

---

<sup>30</sup>Diese Abbildungen wurden mit Hilfe eines Maple-Makros erstellt. Vgl. dazu [http://math.cmaisonneuve.qc.ca/plantagne/Maple/Integrale\\_Riemann\\_Maple9/Integrale\\_Riemann5.html](http://math.cmaisonneuve.qc.ca/plantagne/Maple/Integrale_Riemann_Maple9/Integrale_Riemann5.html).

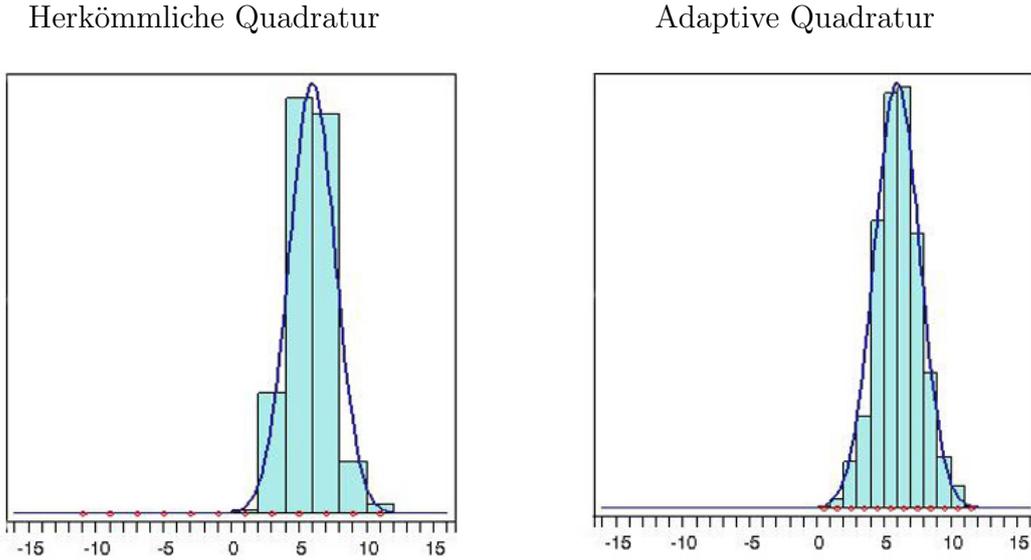


Abbildung 3.1: Darstellung der Ergebnisse einer adaptiven Hermite-Gauß-Quadratur (rechts) und einer herkömmlichen Hermite-Gauß-Quadratur (links) mit 12 Quadraturpunkten. Die roten Kreise stellen dabei die Stützstellen der Integralberechnung dar.

wobei  $U_i \sim \mathcal{N}(0, D)$  gilt. Durch die Substitution von  $u D^{1/2} := u_i$  ergibt sich

$$\int_{-\infty}^{\infty} f_{Y_i|U_i=u_i}(y_i) f_{U_i}(u_i) du_i = \int_{-\infty}^{\infty} f_{Y_i|U_i=u D^{1/2}}(y_i) f_U(u) du$$

mit  $U \sim \mathcal{N}(0, 1)$ .

Die adaptive Hermite-Gauß-Quadratur basiert nun auf der Idee, das Gitter der Stützstellen nicht um den Ursprung, sondern vielmehr um den Bayes-Schätzer von  $U$  zu zentrieren. Zudem wird von der Annahme ausgegangen, dass

$$f_{Y_i|U_i=u D^{1/2}}(y_i) f_U(u)$$

die Dichte einer Normalverteilung ist.

Der empirische Bayes-Schätzer für  $U$  wird nach Gleichung (3.46) durch Maximierung von

$$\begin{aligned} f_{U|Y_i=y_i}(u) &= \frac{\prod_{i=1}^{n_i} f_{Y_{ij}|U_i=u D^{1/2}}(y_{ij}) f_U(u)}{\int \prod_{i=1}^{n_i} f_{Y_{ij}|U_i=u D^{1/2}}(y_{ij}) f_U(u) du} \\ &= \frac{f_{Y_i|U_i=u D^{1/2}}(y_i) f_U(u)}{\int f_{Y_i|U_i=u D^{1/2}}(y_i) f_U(u) du} \end{aligned}$$

ermittelt. Nun gilt

$$\operatorname{argmax}_{u \in \mathbb{R}} (f_{U|Y_i=y_i}(u)) = \operatorname{argmax}_{u \in \mathbb{R}} (\ln(f_{U|Y_i=y_i}(u)))$$

$$\begin{aligned}
 &= \operatorname{argmin}_{u \in \mathbb{R}} \left( -\ln \left( f_{U|Y_i=y_i}(u) \right) \right) \\
 &= \operatorname{argmin}_{u \in \mathbb{R}} \left[ -\ln \left( f_{Y_i|U_i=u} D^{1/2}(y_i) f_U(u) \right) \right. \\
 &\quad \left. + \ln \left( \int f_{Y_i|U_i=u} D^{1/2}(y_i) f_U(u) du \right) \right] \\
 &= \operatorname{argmin}_{u \in \mathbb{R}} -\ln \left( \underbrace{f_{Y_i|U_i=u} D^{1/2}(y_i) f_U(u)}_{=:g(u)} \right). \quad (3.52)
 \end{aligned}$$

Zur Vermeidung von Unklarheiten wird jede im Folgenden genutzte Funktion von derjenigen Variablen, von der sie abhängt, indiziert, d.h. insbesondere  $g(u) =: g_u(u)$ . Aufgrund der Annahme, dass  $g_u(u)$  der Dichte einer Normalverteilung entspricht, ist der Erwartungswert dieser Verteilung durch

$$\hat{u} := \operatorname{argmin}_{u \in \mathbb{R}} \left( -\ln(g_u(u)) \right)$$

gegeben. Die Varianz  $\sigma^2$  dieser Normalverteilung wird mit Hilfe des folgenden Satzes berechnet.

**Satz 3.28.**

*Sei  $X$  eine normalverteilte Zufallsvariable mit dem Erwartungswert  $\hat{u}$  und der Dichte  $g_u(u)$ . Dann ist die Varianz dieser Zufallsvariable durch*

$$\sigma^2 = \left[ -\frac{\partial^2}{\partial^2 u} \ln(g_u(u)) \right]^{-1}$$

*gegeben.*

*Beweis:* Sei  $X$  normalverteilt,  $X \sim \mathcal{N}(\hat{u}, \sigma^2)$ , mit bekanntem Erwartungswert  $\hat{u}$  und unbekannter Varianz  $\sigma^2$ . Dann gilt

$$g_u(u) = \frac{1}{\sigma \sqrt{2\pi}} \exp \left( -\frac{1}{2} \left( \frac{u - \hat{u}}{\sigma} \right)^2 \right)$$

und somit auch

$$\ln(g_u(u)) = -\ln(\sigma \sqrt{2\pi}) - \frac{1}{2} \left( \frac{u - \hat{u}}{\sigma} \right)^2.$$

Die Differentiation liefert

$$\frac{\partial}{\partial u} \ln(g_u(u)) = -\frac{1}{\sigma} \frac{u - \hat{u}}{\sigma} \quad \text{und} \quad \frac{\partial^2}{\partial^2 u} \ln(g_u(u)) = -\frac{1}{\sigma^2}.$$

Somit folgt

$$\sigma^2 = \left[ -\frac{\partial^2}{\partial^2 u} \ln(g_u(u)) \right]^{-1}$$

und damit die Behauptung. □

**Bemerkung 3.29.**

Der Satz 3.28 gilt auch für die multivariate Normalverteilung, d.h. für  $p > 1$ . Der Beweis kann in diesem Fall analog geführt werden.

Die Dichte  $g_u(u)$  ist damit die Dichte einer Zufallsvariablen  $X \sim \mathcal{N}(\hat{u}, \sigma^2)$ . Diese Dichte lässt sich durch Substitution von

$$t := \frac{u - \hat{u}}{\sigma}$$

umformen, so dass mit  $U_i \sim \mathcal{N}(0, D)$  und  $U \sim \mathcal{N}(0, 1)$  die Gleichung

$$\begin{aligned} \int_{-\infty}^{\infty} f_{Y_i|U_i=u_i}(y_i) f_{U_i}(u_i) du_i &= \int_{-\infty}^{\infty} f_{Y_i|U_i=u D^{1/2}}(y_i) f_U(u) du \\ &\stackrel{\text{vgl. (3.52)}}{=} \int_{-\infty}^{\infty} g_u(u) du \\ &= \int_{-\infty}^{\infty} \sigma g_t(t) dt \end{aligned} \tag{3.53}$$

gilt. Im Folgenden wird nun mit Hilfe der Hermite-Gauß-Quadratur, wie sie im letzten Abschnitt vorgestellt wurde, das Integral

$$\int_{-\infty}^{\infty} \sigma g_t(t) dt$$

approximiert. Zu diesem Zweck wird die Gewichtsfunktion  $\phi(t) = e^{-t^2}$  eingefügt und das Integral

$$\int_{-\infty}^{\infty} \underbrace{\sigma g_t(t) e^{t^2}}_{=: f_t(t)} e^{-t^2} dt \tag{3.54}$$

bestimmt. Nach Definition 3.23 und Korollar 3.26 kann dieses Integral durch die numerische Integrationsformel

$$I_n(f) = \sum_{j=1}^n \omega_j f_t(z_j)$$

approximiert werden. Für alle  $j \in \{1, \dots, n\}$  sind dabei die  $z_j$  die Nullstellen des  $n$ -ten Hermite-Polynoms,

$$\omega_j = \langle L_j, 1 \rangle_{\phi}$$

die benötigten Gewichte und  $L_j$ , die in Definition 3.23 definierten Lagrange'schen Basispolynome. Es folgt mit Gleichung (3.53)

$$\begin{aligned}
 \int_{-\infty}^{\infty} f_{Y_i|U_i=u} D^{1/2}(y_i) f_U(u) du &= \int_{-\infty}^{\infty} \sigma g_t(t) e^{t^2} e^{-t^2} dt \\
 &= \int_{-\infty}^{\infty} f_t(t) \phi(t) dt \\
 &\approx \sum_{j=1}^n \omega_j f_t(z_j). \tag{3.55}
 \end{aligned}$$

Zudem gilt

- $t = (u - \hat{u}) \sigma^{-1}$ , und somit insbesondere  $u = \sigma t + \hat{u}$ ,
- $g_u(u) = f_{Y_i|U_i=u} D^{1/2}(y_i) f_U(u)$  und
- $f_t(t) = \sigma g_t(t) e^{t^2}$ ,

so dass sich in Abhängigkeit von  $u$

$$\begin{aligned}
 \int_{-\infty}^{\infty} f_{Y_i|U_i=u} D^{1/2}(y_i) f_U(u) du &\stackrel{vgl.(3.55)}{\approx} \sum_{j=1}^n \omega_j f_t(z_j) \\
 &\stackrel{vgl.(3.54)}{=} \sum_{j=1}^n \omega_j \sigma g_t(z_j) e^{z_j^2} \\
 &= \sum_{j=1}^n \omega_j \hat{\sigma} g_u(\sigma z_j + \hat{u}) e^{z_j^2} \\
 &\stackrel{vgl.(3.52)}{=} \sum_{j=1}^n \omega_j \sigma f_{Y_i|U_i=(\sigma z_j + \hat{u})} D^{1/2}(y_i) \\
 &\quad \times f_U(\sigma z_j + \hat{u}) \exp(z_j^2)
 \end{aligned}$$

ergibt. Diese Integrationsformel wird von Pinheiro und Bates (1995) die *adaptive Gauß-Quadratur* genannt und wurde hier für die Gewichtsfunktion  $\phi(x) = e^{-x^2}$  hergeleitet. Allerdings basierte diese Herleitung auf der Annahme, dass eindimensionale subjekt-spezifische Größen vorliegen.

Im Anhang B wird die Hermite-Gauß-Quadratur und die adaptive Hermite-Gauß-Quadratur für mehrdimensionale subjekt-spezifische Größen hergeleitet und somit eine Möglichkeit erarbeitet, die Integrale der Loglikelihood-Funktion

$$\begin{aligned}
 l_y(\delta) &= \sum_{i=1}^m \ln \left( \int \prod_{j=1}^{n_i} f_{Y_{ij}|U_i=u_i}(y_{ij}) f_{U_i}(u_i) du_i \right) \\
 &= \sum_{i=1}^m \ln \left( \int f_{Y_i|U_i=u_i}(y_i) f_{U_i}(u_i) du_i \right)
 \end{aligned}$$

approximativ zu lösen. Dort wird gezeigt, dass die Approximation

$$\begin{aligned}
 l_y(\delta) &\approx \sum_{i=1}^m \ln \left[ \sum_{j_1=1}^r \dots \sum_{j_p=1}^r \prod_{k=1}^p \omega_{j_k} f_{Y_i|U_i=\hat{a}_i}(y_i) \right. \\
 &\quad \left. \times f_{U_i}(\hat{a}_i) 2^{p/2} |\det(\Sigma)|^{1/2} \exp\left(\sum_{k=1}^p z_{j_k}^2\right) \right] \\
 &=: \tilde{l}_y(\delta)
 \end{aligned}$$

gilt.

Nun ist es das Ziel, den Parametervektor  $\delta$  und somit sowohl den Regressionsvektor  $\beta$  als auch die Parameter der Verteilung von  $U_i$  zu bestimmen. Es wird daher der Loglikelihood-Schätzer von  $\delta$  berechnet, indem nicht  $l_y(\delta)$ , sondern  $\tilde{l}_y(\delta)$  maximiert wird. Das Maximum kann z.B durch Anwendung des Newton-Raphson-Verfahrens ermittelt werden.

Die Methode der Hermite-Gauß- bzw. adaptiven Hermite-Gauß-Quadratur mit anschließender Maximierung der Loglikelihood-Funktion wurde auch in der Prozedur NLMIXED des Statistik-Programms SAS implementiert. Diese Prozedur wird im Kapitel 5 vorgestellt.



# Kapitel 4

## Ein Vergleich der Modelle

In Kapitel 3 wurden die am häufigsten genutzten Modelle für wiederholte Daten, die marginalen und die subjekt-spezifischen Modelle, vorgestellt. Beide Modelle basieren auf den verallgemeinerten linearen Modellen und ermöglichen eine Aufstellung des Regressionsmodells unter Berücksichtigung der Abhängigkeitsstruktur der Daten. Trotz einiger Ähnlichkeiten zwischen den zwei Modellen erzielen sie oftmals sehr unterschiedliche Resultate, so dass Statistiker, die wenig Erfahrung mit diesen Modellen haben, gelegentlich verunsichert sind.<sup>1</sup> Ziel dieses Kapitels ist es daher, die Unterschiede zwischen den Ergebnissen beider Modellfamilien aufzuzeigen und zu diskutieren, wodurch diese Unterschiede zustande kommen.

Wie bereits in Kapitel 3 erwähnt, können sowohl die marginalen Modelle als auch die subjekt-spezifischen Modelle genutzt werden, um Aussagen über die durchschnittliche Entwicklung einer Population zu treffen. Allerdings ermöglichen lediglich die Resultate eines marginalen Modells unmittelbare Rückschlüsse auf den Populationsdurchschnitt. Die geschätzten Parameter der Analyse eines subjekt-spezifischen Modells beschreiben dagegen eine Entwicklung der abhängigen Variablen, basierend auf einem festen Niveau der subjekt-spezifischen Größe  $U_i$ . Sie liefern daher keinen direkten Aufschluss über die durchschnittliche Entwicklung der abhängigen Variablen. Ist letztere jedoch von Interesse, müssen weitere Berechnungen durchgeführt werden.

Zur Veranschaulichung werden im Folgenden die Unterschiede der Modellfamilien und die notwendigen zusätzlichen Berechnungen bei einem subjekt-spezifischen Modell anhand einer Beispielstudie erläutert.

---

<sup>1</sup>Vgl. Molenberghs und Verbeke (2005), S. 297.

## 4.1 Epilepsie-Studie

Bei der betrachteten Epilepsie-Studie handelt es sich um eine multizentrische<sup>2</sup>, zwei-armige<sup>3</sup>, randomisierte<sup>4</sup>, doppelt-verblindete<sup>5</sup> Studie zum Nachweis der Wirksamkeit eines Antiepileptikums (*anti-epileptic drug*, AED). Dieses Antiepileptikum wurde zur Reduzierung der Anzahl von Epilepsieanfällen betroffener Patienten entwickelt.

Zur Überprüfung der Wirkung werden bereits in einem Zeitraum von zwölf Wochen vor Behandlungsbeginn die Epilepsieanfälle der Studienteilnehmer gezählt und dokumentiert, so dass vor Beginn der Behandlung Baseline-Werte vorliegen.

Mit Abschluss der zwölften Woche werden 45 Patienten der Placebo-Gruppe, 44 Patienten der AED-Gruppe zugeordnet und mit der Behandlung begonnen. Während der Durchführung der Studie wird die Anzahl der wöchentlichen Epilepsieanfälle über einen Zeitraum von mindestens sechzehn Wochen dokumentiert. Mit dem Ende der sechzehnten Woche werden die Patienten in eine Langzeitstudie aufgenommen und im Rahmen dieser Studie maximal bis zu 27 Wochen behandelt. Allerdings lässt sich anhand der Daten die Tendenz erkennen, dass Patienten aus der Placebo-Gruppe die Studie früher verlassen haben als Patienten aus der AED-Gruppe. Für das detaillierte Studiendesign wird auf die Arbeit von Faught u. a. (1996) verwiesen.

Die interessierende Größe dieser Studie ist die Entwicklung der Anzahl an Epilepsieanfällen während der Behandlung. Insbesondere stellt sich die Frage nach der Existenz eines signifikanten Behandlungseffekts. Dazu sollen für diese Studie sowohl ein marginales als auch ein subjekt-spezifisches Modell angepasst und die Resultate im Anschluss verglichen werden.

Für beide Modelle sei  $y_{ij}$  die Anzahl der Epilepsieanfälle von Patient  $i$  in der Woche  $j$  für alle  $i \in \{1, \dots, 99\}$ ,  $j \in \{1, \dots, n_i\}$  und  $n_i \in \{1, \dots, 27\}$ . Darüber hinaus sei  $t_{ij}$  der Zeitpunkt, an dem diese Beobachtung erhoben wurde, d.h.  $t_{ij} \in \{1, \dots, n_i\}$ . Sei zudem  $y_{ij}$  eine Realisierung der Zufallsvariable  $Y_{ij}$ .

---

<sup>2</sup>Die Daten wurden in mehreren Kliniken oder Praxen erhoben.

<sup>3</sup>Es handelt sich um zwei Medikationen, die sich in der Darreichungsform, der Dosierung oder den Wirkstoffen unterscheiden. Hier handelt es sich um eine Placebo-kontrollierte Studie, d.h. einer Gruppe wird ein Placebo verabreicht und der anderen Gruppe das zu untersuchende Medikament.

<sup>4</sup>Die Patienten werden den beiden Behandlungsgruppen per Zufall zugeteilt. Diese Prozedur ermöglicht ein theoretisch gleichmäßiges Auftreten bekannter und unbekannter Effekte in beiden Behandlungsgruppen (vgl. Beller u. a. (2002)).

<sup>5</sup>Weder der behandelnde Arzt, noch der Patient wissen, ob es sich bei dem eingenommenen Medikament um ein Placebo oder der AED handelt.

## 4.2 Marginales Modell

Für die Epilepsie-Studie aus Faught u. a. (1996) wird in diesem Abschnitt ein marginales Modell aufgestellt und mit Hilfe der Verallgemeinerten Schätzgleichungen (GEE) und der Prozedur `GENMOD` ausgewertet werden.

Da es sich bei der Anzahl von Epilepsieanfällen um Zähldaten handelt, wird angenommen, dass diese Beobachtungen poissonverteilt sind. Zur Analyse des Zusammenhangs zwischen der Behandlung und der Anzahl der Epilepsieanfälle wird nun ein verallgemeinertes lineares Modell mit kanonischer Linkfunktion (d.h.  $\ln(\cdot)$ ) und behandlungsspezifischen linearen Zeiteffekten aufgestellt:

$$\ln(\lambda_{ij}) = \begin{cases} \beta_0 + \beta_1 t_{ij}, & \text{falls Placebo,} \\ \beta_0 + \beta_2 t_{ij}, & \text{falls AED,} \end{cases} \quad (4.1)$$

wobei  $Y_{ij} \sim \mathcal{P}(\lambda_{ij})$ . Aufgrund der Randomisierung wird für beide Behandlungsgruppen der gleiche Intercept  $\beta_0$  vorausgesetzt. Ein marginales Modell wird allerdings nicht allein durch ein verallgemeinertes lineares Modell beschrieben. Nach Kapitel 3 muss eine Abhängigkeitsstruktur zur Berechnung der Arbeitskorrelationsmatrix bestimmt werden. Nachdem es sich bei diesem Beispiel um eine Studie handelt, die über einen längeren Zeitraum durchgeführt wurde, muss einer eventuell existierenden Zeitabhängigkeit Rechnung getragen werden. Da die Beobachtungen darüber hinaus in gleichen Zeitabständen erhoben wurden, erscheint es nach Abschnitt 3.1.1.3 sinnvoll, die autoregressive Abhängigkeitsstruktur  $\text{AR}(1)$  für die Modellierung der Abhängigkeitsstruktur zu verwenden.

Eine Auswertung dieses Modells mit Hilfe der Prozedur `GENMOD` liefert nach Molenberghs und Verbeke (2005) folgende Schätzer, Standardabweichungen (SD) und p-Werte:<sup>6</sup>

Parameter	Schätzer	empirische SD	modellbasierte SD	p-Wert
$\beta_0$	1.3140	0.1435	0.1601	Keine Angabe
$\beta_1$	-0.0142	0.0234	0.0185	0.5429
$\beta_2$	-0.0192	0.0178	0.0174	0.2795
$\beta_1 - \beta_2$	Keine Angabe	Keine Angabe	Keine Angabe	0.8721

Zudem ergibt sich für den Autokorrelationsparameter  $\alpha$  die Schätzung  $\hat{\alpha} = 0.5963$ . Aus den in der Tabelle 4.2 aufgelisteten Schätzern und p-Werten lässt sich folgern, dass die Nullhypothesen  $H_0 : \beta_i = 0$  für  $i \in \{1, 2\}$  zum Signifikanzniveau  $\alpha^* = 0.05$

<sup>6</sup>Für detaillierte Informationen zur Implementierung und Durchführung der Analyse mit Hilfe der `SAS` Prozedur `GENMOD` wird auf das Kapitel 5 verwiesen.

nicht verworfen werden können. Ein signifikanter Unterschied zwischen den beiden Behandlungsgruppen lässt sich aufgrund des p-Wertes von 0.8721 für den Test  $H_0 : \beta_1 - \beta_2 = 0$  gegen  $H_1 : \beta_1 - \beta_2 \neq 0$  ebenfalls nicht nachweisen.

### 4.3 Subjekt-spezifisches Modell

Eine alternative Analyse der Epilepsie-Studie unter Berücksichtigung der Abhängigkeitsstruktur kann durch ein subjekt-spezifisches Modell erhalten werden. Zu diesem Zweck wird angenommen, dass die Zufallsvariablen  $Y_{ij}$  bedingt unter den zufälligen Größen  $U_i$  unabhängig und poissonverteilt sind. Wie bei der Analyse eines marginalen Modells wird die kanonische Linkfunktion genutzt und ein Modell mit behandlungsspezifischen linearen Zeiteffekten angepasst. Ferner wird erneut angenommen, dass beide Behandlungsgruppen den gleichen Intercept besitzen, so dass das Modell

$$\ln(\lambda_{ij}) = \begin{cases} \beta_0 + U_{i1} + (\beta_1 + U_{i2})t_{ij}, & \text{falls Placebo,} \\ \beta_0 + U_{i1} + (\beta_2 + U_{i2})t_{ij}, & \text{falls AED,} \end{cases}$$

mit  $U_i = (U_{i1}, U_{i2})$  als normalverteilten Zufallsvektor,  $U_i \sim \mathcal{N}_2(0, D)$  mit

$$D = \begin{pmatrix} d_{11} & d_{12} \\ d_{12} & d_{22} \end{pmatrix}$$

und  $Y_{ij}|U_i \sim \mathcal{P}(\lambda_{ij})$  ausgewertet wird. Die Auswertung kann mit Hilfe der Prozedur NLMIXED durchgeführt werden. Diese schätzt den gesuchten Parametervektor  $\delta = (\beta_0, \beta_1, \beta_2, d_{11}, d_{12}, d_{22})^\top$ . Nach Molenberghs und Verbeke (2005) ergeben sich die folgenden Schätzer, Standardabweichungen (SD) und p-Werte:<sup>7</sup>

Parameter	Schätzer	SD	p-Wert
$\beta_0$	0.7739	0.1293	Keine Angabe
$\beta_1$	-0.0245	0.0096	< 0.05
$\beta_2$	-0.0193	0.0097	< 0.05
$\beta_1 - \beta_2$	Keine Angabe	Keine Angabe	> 0.6
$d_{11}$	1.2859	0.2231	Keine Angabe
$d_{22}$	0.0024	0.0006	Keine Angabe
$d_{12}$	-0.3349	0.1318	Keine Angabe

<sup>7</sup>Diese Auswertung wurde mit Hilfe der Hermite-Gauss-Quadratur mit  $Q = 10$  Quadraturpunkten durchgeführt. Das Maximum der Loglikelihood-Funktion  $\ell_y(\delta)$  wurde mittels des *Quasi-Newton-Algorithmus* berechnet. Für detaillierte Informationen zur Implementierung und Durchführung der Analyse mit Hilfe der SAS-Prozedur NLMIXED wird auf das Kapitel 5 verwiesen.

Der Unterschied zwischen den beiden Behandlungsgruppen ist wie im marginalen Modell nicht signifikant, d.h.  $\beta_1 - \beta_2 = 0$  wird zum Signifikanzniveau  $\alpha^* = 0.05$  nicht abgelehnt. Im Gegensatz zu der Analyse, die auf den marginalen Modellen basiert, ergibt sich allerdings in diesem Fall, dass  $\beta_1$  und  $\beta_2$  jeweils signifikant ungleich Null sind.

Inwiefern lassen sich die resultierenden Regressionsparameter  $\hat{\beta}_0$  und  $\hat{\beta}_1$  beider Analysemethoden allerdings vergleichen? Des Weiteren interessiert die Frage, ob sich bei beiden Modellen die gleiche Entwicklung der Anzahl an Epilepsieanfällen ergibt. Diese Fragen sollen im nächsten Abschnitt behandelt werden.

## 4.4 Vergleich der Resultate

Die Analyse eines marginalen Modells liefert für den Regressionsparameter Schätzer, die unmittelbar Aussagen über die Entwicklung des Populationsdurchschnitts ermöglichen. Dieser Sachverhalt gilt für das subjekt-spezifische Modell allerdings nicht. Um hier die durchschnittliche Entwicklung in jeder Behandlungsgruppe zu erhalten, muss der marginalisierte Erwartungswert, d.h. der Erwartungswert der bedingten Erwartungswerte  $E(Y_{ij}|U_i)$  über die subjekt-spezifischen Größen  $U_i$ , ermittelt werden. In der Placebo-Gruppe entspricht dies der Berechnung von

$$\begin{aligned} E(Y_{ij}) &= E[E(Y_{ij}|U_i)] \\ &= E[\exp(\beta_0 + U_{i1} + (\beta_1 + U_{i2})t_{ij})] \\ &\neq \exp(\beta_0 + \beta_1 t_{ij}) \end{aligned} \quad (4.2)$$

und entsprechend muss in der AED-Gruppe

$$\begin{aligned} E(Y_{ij}) &= E[E(Y_{ij}|U_i)] \\ &= E[\exp(\beta_0 + U_{i1} + (\beta_2 + U_{i2})t_{ij})] \\ &\neq \exp(\beta_0 + \beta_2 t_{ij}) \end{aligned} \quad (4.3)$$

bestimmt werden. Es wird deutlich, dass sich die marginalen Erwartungen beider Analysemethoden unterscheiden. Darüber hinaus bedarf die Berechnung der marginalisierten Erwartung in einem festen Punkt  $t = t_{ij}$  numerischer Integrationsmethoden. In vielen Fällen ist es allerdings einfacher, den Erwartungswert zu approximieren, indem eine große Anzahl  $M$  an Realisierungen von  $U_i$  generiert wird. Dann ergibt sich für die Placebo-Gruppe der Erwartungswert

$$E[Y(t)] = \frac{1}{M} \sum_{i=1}^M \exp(0.7739 + U_{i1} + (-0.0245 + U_{i2})t_{ij}) \quad (4.4)$$

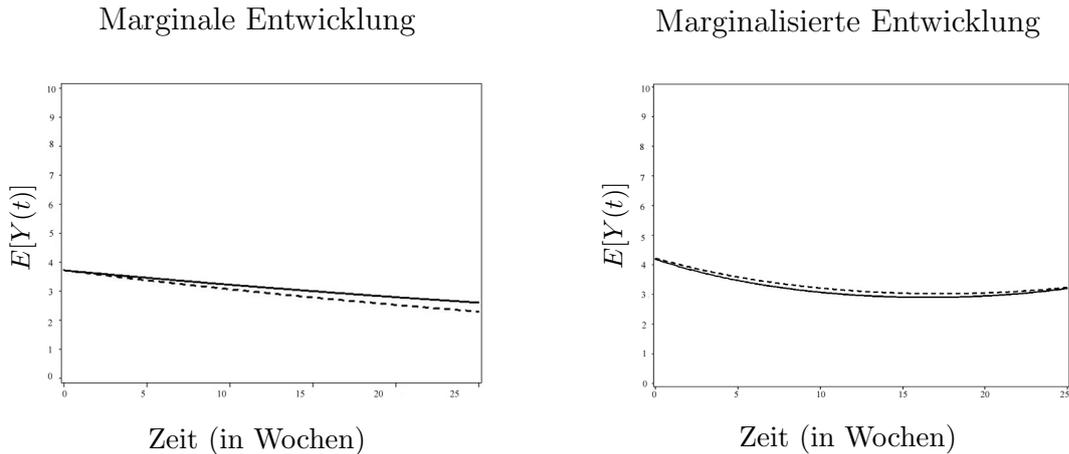


Abbildung 4.1: Die gestrichelte Linie stellt die Entwicklung in der Verum-Gruppe und die durchgezogene Linie die in der Placebo-Gruppe dar. Die marginalisierte Entwicklung wurde hierbei für  $M = 1000$  berechnet.

sowie für die AED-Gruppe der Erwartungswert

$$E[Y(t)] = \frac{1}{M} \sum_{i=1}^M \exp(0.7739 + U_{i1} + (-0.0193 + U_{i2})t_{ij}). \quad (4.5)$$

Eine grafische Veranschaulichung dieser durchschnittlichen Entwicklung in beiden Behandlungsgruppen kann ermittelt werden, indem der Schätzer  $E[Y(t)]$  für ausreichend viele Zeitpunkte  $t$  berechnet und anschließend  $t$  gegen  $E[Y(t)]$  aufgetragen wird. Eine solche Grafik ist für  $M = 1000$  in Abbildung 4.1 gegeben.<sup>8</sup> Beim Vergleich dieser marginalisierten Entwicklung mit den Resultaten eines marginalen Modells (vgl. ebenfalls Abbildung 4.1), fallen folgende Unterschiede auf:

Die subjekt-spezifische marginalisierte Entwicklung weist keine Monotonie auf, da die durchschnittliche Anzahl der Anfälle gegen Ende der Studienzeit wieder ansteigt. Im Gegensatz dazu ist die marginale Entwicklung in beiden Behandlungsgruppen, entsprechend der Modellgleichung in (4.1), monoton fallend. Zudem wird anhand Abbildung 4.1 deutlich, dass die Analyse mittels der subjekt-spezifischen Modelle eine geringfügig bessere Wirksamkeit des Placebos angibt.<sup>9</sup>

Nun sind die Unterschiede zwischen den Erwartungswerten eines marginalen Modells einerseits und den marginalisierten Erwartungswerten eines subjekt-spezifischen Modells andererseits mit der Variabilität der subjekt-spezifischen Größen zu begründen:

<sup>8</sup>Eine entsprechende Abbildung ist in Molenberghs und Verbeke (2005) auf Seite 339 zu finden. Die SAS-Anweisungen zur Erstellung der beiden Teilabbildungen sind in Anhang C gegeben und wurden mit Hilfe der Anweisung in Molenberghs und Verbeke (2005), S. 344, erstellt.

<sup>9</sup>Wie in den letzten beiden Abschnitten gezeigt, ist der Behandlungsunterschied allerdings nicht signifikant.

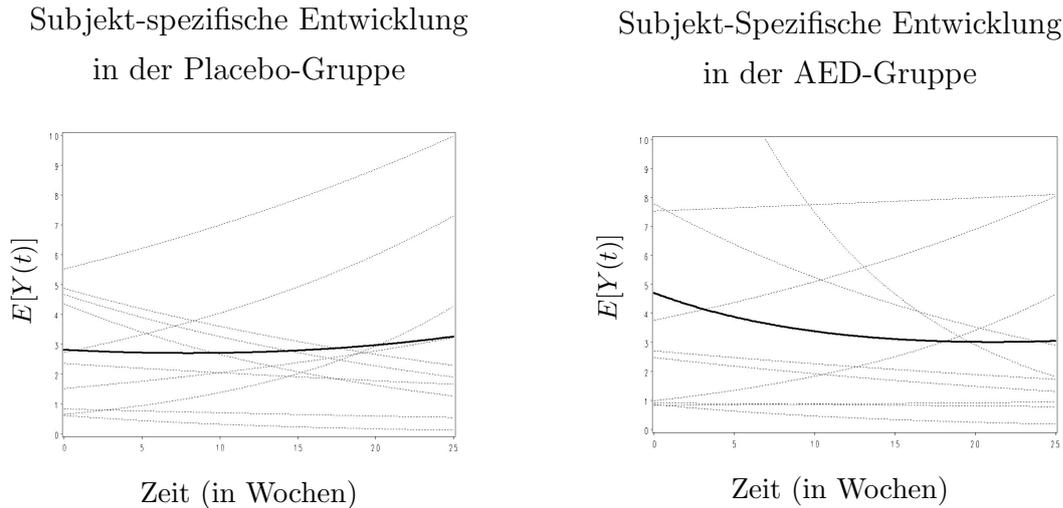


Abbildung 4.2: Für beide Behandlungsgruppen wird die subjekt-spezifische Entwicklung für zehn Patienten in der Epilepsiestudie (gestrichelte Linien) dargestellt. Zudem ist die durch Gleichung (4.4) bzw. (4.5) und  $M = 10$  berechnete marginalisierte Entwicklung für beide Behandlungsgruppen gegeben (durchgezogene Linie).

*„Intuitively, it is to be expected that this effect strongly depends on the amount of between-subject variability: In case the random-intercepts variability is large, parameters from fitting marginal models and random-effects models will be very different, while equal parameter values hold if the variance of the random-effects equals zero.“<sup>10</sup>*

Um diese „between-subject variability“ und die Auswirkungen auf den marginalisierten Erwartungswert deutlich zu machen, sind in Abbildung 4.2 die subjekt-spezifischen Entwicklungen von zehn Patienten und deren marginalisierte Entwicklung in der Placebo-Gruppe (berechnet gemäß Gleichung (4.4)) bzw. in der AED-Gruppe (berechnet gemäß Gleichung (4.5)) gegeben.<sup>11</sup> Folglich unterscheiden sich die Entwicklungen für lediglich zehn Patienten bereits im hohen Maße.

Zusammenfassend ergibt sich:

*„The severe differences in results obtained from marginal and random-effects models follow from the fact that the parameters in both models have completely different interpretations.[...] Whenever a marginal model is fitted, one directly obtains estimates and inferences for the components in  $\beta^M$ , the regression vector that models the average trend in the population. In case a random-effects model is fitted, one should realize that, even when estimation and inference is based on likelihood principles*

<sup>10</sup>Molenberghs und Verbeke (2005), S. 299.

<sup>11</sup>Die SAS-Anweisung zur Erstellung der Abbildung ist in Anhang C gegeben.

*for the marginal likelihood where the random effects have been integrated out, the parameters keep their original random-effects interpretation, such that estimates, as well as inferences are obtained for the components in  $\beta^{RE}$  rather than  $\beta^M$ .*<sup>12</sup>

Dabei ist  $\beta^M$  der Regressionsvektor, der aus der Analyse des marginalen Modells hervorgegangen ist, und  $\beta^{RE}$  entsprechend der Regressionsvektor eines subjekt-spezifischen Modells.

In einigen Spezialfällen gibt es einen direkten Zusammenhang zwischen den Regressionsvektoren  $\beta^M$  und  $\beta^{RE}$ . Eine solche Beziehung liegt beispielsweise bei linearen Modellen, d.h. bei normalverteilten Daten  $Y_{ij}$ , vor. Sei das subjekt-spezifische Modell durch

- $Y_i|U_i \sim \mathcal{N}(X_i\beta^{RE} + Z_iU_i, \Sigma)$  mit einer regulären Planungsmatrix  $X_i$  für alle  $i \in \{1, \dots, m\}$  und
- dem linearen gemischten Modell  $E(Y_{ij}|U_i) = X_i\beta^{RE} + Z_iU_i$

bestimmt. Dann ergibt sich der marginalisierte Erwartungswert

$$E(Y_{ij}) = E[E(Y_{ij}|U_i)] = X_i\beta^{RE}.$$

Im Rahmen der Analyse eines marginalen Modells wird das lineare Modell  $E(Y_i) = X_i\beta^M$  betrachtet und die Abhängigkeitsstruktur gesondert modelliert. Da die Planungsmatrix regulär ist, folgt somit die Gleichheit der Regressionsvektoren

$$\beta^M = \beta^{RE},$$

und damit auch die Gleichheit des marginalen und des marginalisierten Erwartungswertes. Mit diesem Beispiel wird deutlich, dass sich die Regressionsparameter im nicht-normalverteilten Fall insbesondere aufgrund der benötigten Linkfunktion unterscheiden.

Ein weiteres Beispiel, bei dem die Regressionsparameter  $\beta^M$  und  $\beta^{RE}$  zusammenhängen, ist durch binäre Daten und der Modellierung mittels eines *Logistischen Modells* gegeben.<sup>13</sup> Dieses Beispiel soll hier allerdings nicht weiter ausgeführt werden.

---

<sup>12</sup>Molenberghs und Verbeke (2005), S. 298.

<sup>13</sup>Vgl. Molenberghs und Verbeke (2005), S. 300.

# Kapitel 5

## Implementierung und Beispiele

### 5.1 Implementierung für das marginale Modell

Die Methode des Schätzens mittels Verallgemeinerter Schätzgleichungen (*generalized estimating equations*, GEE) wurde von Liang und Zeger im Jahr 1986 veröffentlicht. Bereits zwei Jahre später wurde ein SAS-Makro vorgestellt, das zur Auswertung marginaler Modelle herangezogen werden kann.<sup>1</sup> Inzwischen ist die Methode der GEE in einer ganzen Serie von Softwarepaketen implementiert, so z.B. in STATA, SUDAAN und SPIDA.<sup>2</sup> Auch im Rahmen der Software-Programme S-PLUS und SPSS ist die Analyse mittels der GEE möglich, allerdings müssen die zugehörigen Pakete und Makros gesondert eingebunden werden.<sup>3</sup>

Für die Analyse der Beispiele in Abschnitt 5.3 wird die Statistik-Software SAS benutzt. Diese findet besonders in Bereichen der Biometrie breite Anwendung und wird von vielen Nutzern als Standard angesehen.<sup>4</sup>

Die Prozedur zur Auswertung marginaler Modelle wird in SAS unter dem Namen GENMOD geführt. Im Wesentlichen wertet diese Prozedur sowohl verallgemeinerte lineare Modelle für korrelierte Daten als auch für unkorrelierte Daten aus.

Die Funktionsweise dieses Programms entspricht den Ausführungen in Abschnitt 3.1.1.1. Im Folgenden werden zunächst die wichtigsten Punkte der Syntax von GENMOD vorgestellt. Zu detaillierteren Informationen wird auf das Handbuch SAS/STAT (1999) verwiesen.

Ein typisches Programm zur Auswertung eines marginalen Modells ist gegeben durch:

---

<sup>1</sup>Vgl. Karim und Zeger (1988).

<sup>2</sup>Vgl. Ziegler und Grömping (1998).

<sup>3</sup>Vgl. Ziegler u. a. (1998).

<sup>4</sup>Vgl. Molenberghs und Verbeke (2005), S. 203.

```
proc genmod data=Datensatz ;  
class qualitative Daten ;  
model abhängige Variable=erklärende Variablen / dist=Verteilung /  
link=Linkfunktion ;  
repeated subject=Größe, zu der wiederholt Daten erhoben werden /  
corrw covb  
type=Abhängigkeitsstruktur within=Ordnung der Wiederholung ;  
run ;
```

Dabei gilt folgendes:

- Die **class**-Anweisung gibt die qualitativen Variablen der Analyse an;
- die **model**-Anweisung legt das Modell fest, indem die Beobachtungsvariable (abhängige Variable, z.B. Anzahl der Hitzewallungen) und die erklärenden Variablen (unabhängige Variablen, z.B. Behandlung, Alter, Geschlecht) bestimmt werden;
- die **dist**-Anweisung gibt die angenommene Verteilung der Daten an (in diesem Fall die Poisson-Verteilung)
- die **link**-Anweisung bestimmt die Linkfunktion;
- die **repeated**-Anweisung besitzt mehrere Optionen:
  - **subject** legt fest, zu welcher Variable wiederholte Beobachtungen erhoben wurden (z.B. Patient);
  - **type** bestimmt die Struktur der Arbeitskorrelationsmatrix z.B. **EXCH** für Austauschbarkeit, **AR(1)** für Autoregression erster Ordnung, **IND** für Unabhängigkeit, **UNSTR** für eine unstrukturierte Arbeitskorrelationsmatrix und **USER** für eine benutzerdefinierte;
  - **corr**w liefert die geschätzte Arbeitskorrelationsmatrix;
  - **cov**b berechnet die Korrelationsmatrix des geschätzten Regressionsparameters, wobei sowohl die modellbasierte Schätzung ( $I_0^{-1}$ ) als auch die empirische Schätzung ( $I_0^{-1} I_1 I_0^{-1}$ ) ausgegeben werden;
  - **within** gibt diejenige Variable an, nach der die Beobachtungen innerhalb eines Subjekts geordnet sind (z.B. Zeit, Tag, Ort).

Die Ausführung eines solchen Programms resultiert in Tabellen und Angaben zu den genutzten Variablen und Größen, die im Wesentlichen folgendes beinhalten:

- Modellinformationen:
  - Name des Datensatzes;
  - angenommene Verteilung;
  - Linkfunktion (es wird die kanonische Linkfunktion verwendet, wenn keine `link`-Anweisung angegeben wird);
  - Name der abhängigen Variable;
  - Anzahl der vorliegenden Beobachtungen;
  - Anzahl der Beobachtungen, die für die Analyse genutzt werden sowie
  - Anzahl der fehlenden Daten;
- Levelinformationen der qualitativen Daten;
- Maße für die Anpassungsgüte des geschätzten Modells, wie die Deviance und die Pearson- $\chi^2$ -Statistik. Da diese Größen aber auf der Unabhängigkeit der Beobachtungen basieren (vgl. McCullagh und Nelder (1989), S.33f), sind sie im Fall korrelierter Daten nicht aussagekräftig (vgl. Molenberghs und Verbeke (2005), S.206f);
- eine Tabelle der Anfangswerte für die anschließende GEE-Analyse mittels der Fisher-Scoring-Methode, d.h.:
  - Berechnungen finden unter der Annahme der Unabhängigkeit statt;
  - Schätzer für die Regressionsparameter, Standardabweichungen und 95%-Konfidenzintervalle für die Regressionsparameter werden angegeben sowie
  - p-Werte des  $\chi^2$ -Tests zum Testen der Hypothesen  $H_0 : \beta_i = 0$  gegen  $H_0 : \beta_i \neq 0$  werden für alle  $i \in \{1, \dots, q\}$  angegeben;
- GEE-Modellinformationen:
  - Korrelationsstruktur;
  - Ordnung der wiederholten Beobachtungen;
  - maximale und minimale Anzahl der Beobachtungen pro Subjekt sowie
  - Anzahl der Subjekte mit fehlenden Daten;
- ein modellbasierter Schätzer ( $I_0^{-1}$ ) für die Kovarianzmatrix von  $\beta$ ;

- ein empirischer Schätzer  $(I_0^{-1}I_1I_0^{-1})$  für die Kovarianzmatrix von  $\beta$ ;
- ein Schätzer für die Arbeitskorrelationsmatrix und
- GEE-Schätzer:
  - Schätzer für die Regressionsparameter, die mit der Methode von Liang und Zeger (1986) ermittelt wurden;
  - Standardabweichungen der Schätzer, die basierend auf  $I_0^{-1}$  oder  $I_0^{-1}I_1I_0^{-1}$  berechnet werden können;
  - 95%-Konfidenzintervalle sowie
  - p-Werte des zweiseitigen Z-Tests zum Testen der Hypothesen  $H_0 : \beta_i = 0$  gegen  $H_0 : \beta_i \neq 0$  für alle  $i \in \{1, \dots, q\}$ .

Anhand dieser Informationen können signifikante Aussagen über Zusammenhänge zwischen den abhängigen Größen  $Y_{ij}$  und den unabhängigen erklärenden Variablen getätigt werden. Zudem wird die Ausgabe der Prozedur GENMOD in Abschnitt 5.3.4 durch ein Beispiel erläutert.

## 5.2 Implementierung für das subjekt-spezifische Modell

Für das subjekt-spezifische Modell wurde eine Methode zur Schätzung des gesuchten Parametervektors  $\delta$  bereits im Abschnitt 3.2.2 diskutiert.<sup>5</sup> In diesem Zusammenhang hatte die Loglikelihood-Funktion die Form

$$l_y(\delta) = \sum_{i=1}^m \ln \left( \int \prod_{j=1}^{n_i} f_{Y_{ij}|U_i=u_i}(y_{ij}) f_{U_i}(u_i) du_i \right).$$

Die Maximum-Likelihood-Schätzung für  $\delta$  basiert allerdings nicht auf  $l_y(\delta)$ , sondern auf der approximierten Funktion<sup>6</sup>

$$\begin{aligned} \tilde{l}_y(\delta) = \sum_{i=1}^m \ln \left[ \sum_{j_1=1}^r \dots \sum_{j_p=1}^r \prod_{k=1}^p \omega_{j_k} f_{Y_i|U_i=\hat{a}_i}(y_i) \right. \\ \left. \times f_{U_i}(\hat{a}_i) 2^{p/2} |\det(\Sigma)|^{1/2} \exp \left( \sum_{k=1}^p z_{j_k}^2 \right) \right]. \end{aligned}$$

---

<sup>5</sup>Der Parametervektor  $\delta$  beinhaltet alle unbekanntes Größen der Modellierung. Da  $\theta_{ij} \stackrel{\text{Poisson}}{=} \ln(\lambda_{ij}) = x_{ij}^\top \beta + z_{ij}^\top U_i$  gilt, sind die unbekanntes Größen durch den Regressionsvektor  $\beta$  und die Parameter der Verteilung von  $U_i$  gegeben.

<sup>6</sup>Vgl. Herleitung und Notation auf Seite 71.

Diese Approximation wurde mit Hilfe der adaptiven Hermite-Gauß-Quadratur berechnet und das Maximum kann durch verschiedene numerische Optimierungsverfahren, wie beispielsweise dem Newton-Raphson-Verfahren, ermittelt werden. Die Umsetzung dieser Methode wird in SAS durch die Prozedur NLMIXED zur Verfügung gestellt, daher soll diese Prozedur im Rahmen dieser Arbeit erläutert werden.

Die NLMIXED-Prozedur wertet *generalized linear mixed models* (GLMM), aber auch nicht-lineare Modelle aus, indem sie eine Approximation der Loglikelihood-Funktion maximiert. Diese Loglikelihood-Funktion wird durch Integration über die subjekt-spezifischen Größen ermittelt.<sup>7</sup> Für die Berechnung der Integrale sind verschiedene Approximationsmethoden in NLMIXED verfügbar, so auch die in Kapitel 3 vorgestellte Hermite-Gauß-Quadratur und die adaptive Hermite-Gauß-Quadratur. Für das Lösen des Maximierungsproblems stehen ebenfalls eine Auswahl von verschiedenen numerischen Optimierungsverfahren, wie etwa das Newton-Raphson-Verfahren oder der duale Quasi-Newton-Algorithmus, zur Verfügung.

Ein typisches Programm zur Auswertung eines subjekt-spezifischen Modells ist gegeben durch:

```
proc nlmixed data=Datensatz technique = Methode zur Berechnung des
  Maximums method = Methode zur Approximation der Integrale
  qpoints=Anzahl der Quadraturpunkte noad;
parms Anfangswerte für die einzelnen Komponenten von delta;
model Legt die bedingte Verteilung der Daten,
  gegeben den subjekt-spezifischen Größen fest;
random subjekt-spezifische Größe und deren Verteilung
subject = Variable, zu der wiederholt Daten erhoben werden;
run;
```

Dabei sind die Aufgaben der Anweisungen wie folgt:

- Die *technique*-Anweisung legt fest, welche Methode zur Berechnung des Maximums von  $\tilde{\ell}_y(\delta)$  genutzt wird. Es können beispielsweise folgende Optimierungsmethoden gewählt werden:
  - CONGRA: konjugierte Gradientenoptimierung;
  - NMSIMP: Nelder-Mead Simplex-Optimierung;
  - NEWRAP: Newton-Raphson-Methode und

---

<sup>7</sup>Vgl. Abschnitt 3.2.1.3.

- QUANEW: Quasi-Newton-Verfahren (fehlt die `technique` Anweisung, so wird diese Methode angewandt);
- die `method`-Anweisung gibt an, welche Methode zur Approximation der Integrale in der Loglikelihood-Funktion  $\ell_y(\delta)$  benutzt werden soll. Auch für diese Option sind verschiedene Vorgaben möglich:
  - FIRO setzt die so genannte *first order* Methode von Beal und Sheiner (1982) um;
  - GAUSS gibt die in Abschnitt 3.2.2.1.2 vorgestellte adaptive Hermite-Gauß-Quadratur an. Die Adaption kann durch Angabe der Option `noad` verhindert werden;
  - HARDY approximiert mit Hilfe der *Hardy-Quadratur*, die auf einer adaptiven Trapezregel basiert;
- die `qpoints`-Anweisung ermöglicht die Steuerung der Genauigkeit durch Angabe der Anzahl an Quadraturpunkten. Sollte diese Angabe nicht gemacht werden, berechnet NLMIXED die Anzahl der Quadraturpunkte adaptiv;
- die `parms`-Anweisung wird genutzt, um die Startwerte für alle unbekannt Parameter der Analyse anzugeben;
- die `model`-Anweisung spezifiziert die Verteilung der Daten, bedingt unter den subjekt-spezifischen Größen. Für die bedingte Verteilung können alle gängigen Verteilungen angegeben werden;
- die `random`-Anweisung definiert die subjekt-spezifischen Größen und deren Verteilung;
- die `subject`-Anweisung gibt diejenige Variable an, zu der die wiederholten Daten erhoben wurden.

Durch diese Ausführungen wird unmittelbar deutlich, welche Schwachstellen die Prozedur NLMIXED aufweist. Einerseits generiert NLMIXED die Startwerte nicht selbst, obwohl die Güte der Konvergenz des in der `technique`-Anweisung gewählten Optimierungsverfahrens vor allem bei komplexen Modellen deutlich von der Wahl der Startwerte abhängt.<sup>8</sup> Andererseits beruht die Approximation der gesuchten Parameter auf der Anzahl der gewählten Quadraturpunkte. Ferner kann es in Fällen, in denen beispielsweise die Verteilung der subjekt-spezifischen Parameter  $U_i$  durch

---

<sup>8</sup>Vgl. Molenberghs und Verbeke (2005), S. 291.

$U_i \sim \mathcal{N}_p(0_p, D)$  gegeben ist, zu einer nicht positiv definiten Schätzung  $\hat{D}$  für die Kovarianzmatrix  $D$  kommen.

Im Verlauf der Analyse mittels der NLMIXED-Prozedur kann es zu weiteren Problemen kommen, allerdings wird für detaillierte Informationen auf den Abschnitt *Computational Problems* in SAS/STAT (1999) verwiesen.

Trotz der genannten Nachteile stellt die hohe Flexibilität bei der Modellbeschreibung einen entscheidenden Vorteil dar. So kann das gemischte Modell

$$\begin{aligned} \ln(\lambda_{ij}) &= x_{ij}^\top \beta + d_{ij}^\top U_i \\ &= \begin{cases} \beta_0 + U_i + \beta_1, & \text{falls Placebo,} \\ \beta_0 + U_i, & \text{falls Verum} \end{cases} \end{aligned}$$

charakterisiert werden durch die Anweisung

```
if(treat=Placebo) then eta = beta_0 + u + beta_1;
else if
(treat=Verum) then eta = beta_0 + u;
lambda = exp(eta);
```

und das Modell

$$\ln(\lambda_{ij}) = \begin{cases} \beta_0 + U_i + \beta_1 t_{ij}, & \text{falls Placebo,} \\ \beta_0 + U_i + \beta_2 t_{ij}, & \text{falls Verum,} \end{cases} \quad (5.1)$$

durch die Anweisung

```
if(treat=Placebo) then eta = beta_0 + u + beta_1*time ;
else if
(treat=Verum) then eta =beta_0 + u + beta_2*time;
lambda=exp(eta);
```

beschrieben werden. Für die Analyse muss dann nur noch der Zusammenhang zwischen den abhängigen Variablen und der Größe `lambda` angegeben werden. Dies kann z.B. durch

```
model abhängige Variable ~ Poisson(lambda)
```

geschehen.

Die Ausführung eines solchen Programms liefert, wie die Prozedur `GENMOD`, Tabellen und Angaben, die im Wesentlichen folgendes beinhalten:

- Modellinformationen:
  - Namen des Datensatzes;
  - Namen der abhängigen Variable;
  - Verteilung der abhängigen Variable, bedingt unter den Größen  $U_i$ ;
  - Bezeichnung der subjekt-spezifischen Größe;
  - Verteilung der subjekt-spezifischen Größe;
  - Namen der Variable, von der die wiederholten Daten stammen;
  - Namen des genutzten Optimierungsverfahrens;
  - Namen der Approximationsmethode zur Berechnung der Integrale;
  - Anzahl der Beobachtungen, die vorliegen;
  - Anzahl der Beobachtungen, die genutzt werden;
  - Anzahl der Subjekte;
  - maximale Anzahl an Beobachtungen pro Subjekt;
  - Anzahl der unbekannt Parameter sowie
  - Anzahl der Quadraturpunkte, die zur Analyse genutzt werden;
- Startwerte, die bei der Optimierung angewandt werden und der Funktionswert der negativen Loglikelihood-Funktion;
- den Iterationsverlauf bei der Minimierung der negativen Loglikelihood-Funktion. Für jeden Iterationsschritt des gewählten Optimierungsverfahrens werden dabei der Funktionswert der Loglikelihood-Funktion, die Differenz zum vorhergehenden Iterationswert und der größte absolute Wert der Komponenten des Gradienten bestimmt;
- Maße für die Anpassungsgüte des geschätzten Modells, z.B.
  - Maximalwert der Loglikelihood-Funktion;
  - *Akaike's Informationskriterium*<sup>9</sup>, sowie das
  - *Bayes'sche Informationskriterium*<sup>10</sup>;

---

<sup>9</sup>Vgl. Akaike (1974).

<sup>10</sup>Vgl. Schwarz (1978).

- eine Tabelle mit den Schätzern für die unbekannt Parameter:
  - Maximum-Likelihood-Schätzer für die einzelnen Komponenten des Regressionsvektors  $\beta$  und die Parameter der Verteilung von  $U_i$ ;
  - Standardabweichungen, die basierend auf der Inversen der Fisher-Informationsmatrix berechnet werden;
  - 95%-Konfidenzintervalle sowie
  - p-Werte und Freiheitsgrad des genutzten t-Tests zum Testen der Hypothesen  $H_0 : \delta_i = 0$  gegen  $H_1 : \delta_i \neq 0$  für alle  $i \in \{1, \dots, k\}$  und  $\delta \in \mathbb{R}^k$ .

Mit Hilfe der Prozedur `NLMIXED` können somit subjekt-spezifische Modelle für Studien ausgewertet werden. Darüber hinaus können anhand der genannten Ausgaben des Programms Aussagen über den Zusammenhang zwischen abhängigen und unabhängigen Variablen getroffen werden.

Sowohl die in Abschnitt 5.1 vorgestellte Prozedur `GENMOD` als auch die soeben vorgestellte Prozedur `NLMIXED` sollen in diesem Kapitel genutzt werden, um Studien der Schering AG zu analysieren.

## 5.3 Beispiele

In diesem Abschnitt werden zwei von der Schering AG durchgeführte Phase-3-Studien vorgestellt. Die Datenerhebung liefert in beiden Studien wiederholte Zähl-daten und das Ziel beider Studien ist es jeweils nachzuprüfen, ob ein Medikament (Medikament A bzw. Medikament B) nachweisbar das Auftreten von Hitzewallungen während des Klimakteriums<sup>11</sup> reduziert.

Das Klimakterium wird im Volksmund auch Wechseljahre genannt und bezeichnet die Zeit der hormonellen Umstellung vor, während und nach der Menopause.<sup>12</sup> Während des Klimakteriums, das meistens zwischen dem 48. und 52. Lebensjahr auftritt, kann es zu Beschwerden wie Hitzewallungen, Schweißausbrüchen bis hin zu Depressionen kommen.<sup>13</sup> Allgemein ist das Klimakterium ein natürlicher Abschnitt im Leben einer Frau und keine Krankheit. Allerdings ist in vielen Fällen der Wunsch nach Medikamenten zur Linderung der Beschwerden groß, um so u.a. Einschränkung im Beruf und Alltag zu unterbinden. Daher werden bei ein bis zwei Drittel aller Frauen die Beschwerden laut Pschyrembel (1994) behandelt.

<sup>11</sup>griechisch: *κλιμακτηριον*, kritischer Punkt im menschlichen Leben, Schwelle

<sup>12</sup>Die Menopause entspricht dem Zeitpunkt des physiologischen Aufhörens der Menstruation.

<sup>13</sup>Vgl. Pschyrembel (1994), S. 784.

Die Medikament-A-Studie wird sowohl mit Hilfe marginaler als auch subjekt-spezifischer Modelle ausgewertet, und die Ergebnisse anschließend hinsichtlich der Wirksamkeit des Medikaments interpretiert. Im Fall der Medikament-B-Studie wird lediglich das marginale Modell in verschiedenen Ausführungen ausgewertet. Zunächst einmal werden aber die Studiendesigns für die Medikament-A- und Medikament-B-Studie vorgestellt.

### 5.3.1 Medikament-A-Studie

Bei der Medikament-A-Studie handelt es sich um eine multizentrische<sup>14</sup>, doppelverblindete<sup>15</sup>, vierarmige<sup>16</sup>, Placebo-kontrollierte Studie, in deren Rahmen 212 postmenopausale Frauen in vier Behandlungsgruppen eingeteilt wurden. Die Studienteilnehmerinnen wurden aufgefordert, während des Behandlungszeitraumes von 12 Wochen ein Tagebuch zu führen, um darin die Anzahl der täglichen Hitzewallungen zu dokumentieren. Nach Ablauf der zwölf Wochen lagen somit im optimalen Fall für jede Frau 84 Daten – wiederholte Zählzeiten – vor.<sup>17</sup>

Im Rahmen dieser Diplomarbeit wird allerdings nicht der komplette Datensatz verwendet. Erstens werden nur zwei der vier Behandlungsgruppen betrachtet, und zweitens werden lediglich die Daten der vierten Behandlungswoche für die Analyse herangezogen.<sup>18</sup> Darüber hinaus werden die Beobachtungen von Patientinnen mit fehlenden Daten nicht in der Analyse berücksichtigt.<sup>19</sup> Demzufolge werden die Daten von nur 100 postmenopausalen Frauen für den Zeitraum der vierten Woche betrachtet, wobei, ohne weiter auf die Dosierung einzugehen,

- 53 Frauen der Placebo-Gruppe und
- 47 Frauen der Medikament-A-Gruppe

---

<sup>14</sup>Die Patienten wurden in mehreren Kliniken oder Praxen behandelt.

<sup>15</sup>Weder der behandelnde Arzt noch der Patient wissen, ob und welche Wirkstoffe im eingenommenen Medikament enthalten sind.

<sup>16</sup>Es handelt sich um vier Medikationen, die sich in der Darreichungsform, der Dosierung oder den Wirkstoffen unterscheiden.

<sup>17</sup>Es wurden neben der Anzahl an Hitzewallungen auch andere Faktoren gemessen, jedoch reduzieren wir hier die Betrachtungen lediglich auf die Anzahl der Hitzewallungen.

<sup>18</sup>Die Restriktion auf die vierte Woche ist dadurch motiviert, dass neue gesetzliche Richtlinien den Nachweis der Wirksamkeit des Medikaments nach vier Wochen fordern.

<sup>19</sup>Die Beobachtungen dieser Personen werden nicht berücksichtigt, da im Rahmen dieser Arbeit nicht diskutiert wird, wie mit fehlenden Daten umzugehen ist. Literatur zu diesem Thema ist in Molenberghs und Verbeke (2005), Teil VI zu finden.

zugeteilt waren. Ziel dieser Studie war es, die Wirksamkeit des Medikaments (unabhängig von der Zeitvariablen) zu überprüfen.

### 5.3.2 Medikament-B-Studie

Das Studiendesign der Medikament-B-Studie ähnelt dem der Medikament-A-Studie. Auch hier handelte es sich um eine multizentrische, doppelt-verblindete, Placebo-kontrollierte Studie. Anstatt von vier Armen wurden im Fall der Medikament-B-Studie nur drei Behandlungsarme betrachtet. Zudem war diese Studie mit 425 Frauen größer angelegt als die Medikament-A-Studie. Die Hauptzielvariable ist wieder die Anzahl der Hitzewallungen, deren tägliches Auftreten während eines Behandlungszeitraumes von 12 Wochen per Tagebuch festgehalten werden sollte. Auch die Stärke der Hitzewallungen wurde festhalten, wobei zwischen schwachen, mittelstarken und starken Hitzewallungen unterschieden wurde. Nach Ablauf der zwölf Wochen lagen somit im optimalen Fall für jede Frau  $84 \times 3 = 252$  Zählzeiten vor.

Ähnlich zu dem Vorgehen bei der Medikament-A-Studie wird auch hier nicht der gesamte Datensatz für die Auswertungen benutzt. Neben dem Placebo wird nur ein weiterer Behandlungsarm betrachtet. Außerdem werden die verschiedenen Stärken der Hitzewallungen zu einer neuen Größe zusammengefasst, indem die Summe der mittleren und starken Hitzewallungen gebildet, und die Analyse mit dieser neuen Größe durchgeführt wird.

Ziel der Untersuchung ist es, die Existenz eines behandlungsspezifischen Zeiteffekts zu überprüfen. Zu diesem Zweck werden sechs verschiedene Designs betrachtet:

#### **Design 1**

Es werden pro Subjekt alle Werte zur Analyse herangezogen.

#### **Design 2**

Es werden lediglich alle Daten der vierten, achten und zwölften Woche betrachtet.

#### **Design 3**

Es werden zwölf Beobachtungen, die jeweils sieben Tage auseinander liegen, für die Analyse benutzt.

#### **Design 4**

Nur die Daten der vierten Woche sind von Interesse.

#### **Design 5**

Es werden nur die Daten der achten Woche betrachtet.

#### **Design 6**

Es werden nur die Daten der zwölften Woche benutzt.

### 5.3.3 Modellierung im Fall eines marginalen Modells

#### Medikament-A-Studie

Um die Wirksamkeit der Behandlung hinsichtlich der Reduzierung der Hitzewallungsanzahl zu untersuchen, wird im Fall der Medikament-A-Studie ein einfaches Regressionsmodell angepasst. Sowohl die Baseline-Werte als auch eventuell existierende Zeiteffekte werden nicht berücksichtigt. Lediglich der Behandlungseffekt in Abhängigkeit von der Behandlung ist von Interesse. Demzufolge kann das marginale Modell für die Medikament-A-Studie wie folgt formuliert werden:

- Das verallgemeinerte lineare Modell ist gegeben durch

$$\begin{aligned} \ln(\mathbb{E}(Y_{ij})) &= \ln(\lambda_{ij}) \\ &= \begin{cases} \beta_0 + \beta_1, & \text{falls Placebo,} \\ \beta_0, & \text{falls Medikament A.} \end{cases} \end{aligned} \quad (5.2)$$

- Die Varianz

$$\text{Var}(Y_{ij}) = \lambda_{ij}\phi \quad (5.3)$$

enthält den Überdispersions-Parameter  $\phi$ .

- Die Korrelation wird beschrieben durch

$$\text{Corr}(Y_{ij}, Y_{ik}) = \rho(\lambda_{ij}, \lambda_{ik}, \alpha),$$

wobei  $\text{Corr}(Y_{ij}, Y_{ik})$  fest sein (Austauschbarkeit) oder beispielsweise  $\text{Corr}(Y_{ij}, Y_{ik}) = \alpha^{|j-k|}$  (Autoregression erster Ordnung) gelten kann.

Der Behandlungsunterschied ist gemäß der Modellierung in Gleichung (5.2) multiplikativ und gegeben durch den Faktor  $\exp(\beta_1)$ , d.h. wenn  $\beta_1 \neq 0$  ist, ergibt sich eine Veränderung der Anzahl an Hitzewallungen um den Faktor  $\exp(\beta_1)$ .

#### Medikament-B-Studie

In der Medikament-B-Studie soll der Behandlungseffekt von Medikament B in Abhängigkeit von der Zeit untersucht werden. Daher wird analog zur vorherigen Studie ein verallgemeinertes lineares Modell mit natürlicher Linkfunktion angepasst. Im Gegensatz zu der Medikament-A-Studie spielt aber nicht nur die Behandlung, sondern auch die Zeit eine Rolle. Zu diesem Zweck wird das Modell in Gleichung (5.2) zu dem Modell

$$\ln(\lambda_{ij}) = \begin{cases} \beta_0 + \beta_1 + \beta_2 t_{ij}, & \text{falls Placebo,} \\ \beta_0 + \beta_3 t_{ij}, & \text{falls Medikament B} \end{cases} \quad (5.4)$$

erweitert, das die behandlungsspezifischen Zeiteffekte miteinbezieht. Wird die Nullhypothese  $H_0^1 : \beta_1 = 0$  zum Signifikanzniveau  $\alpha^* = 0.05$  nicht verworfen, folgt eine Anpassung des Modells

$$\ln(\lambda_{ij}) = \begin{cases} \beta_0 + \beta_2 t_{ij}, & \text{falls Placebo,} \\ \beta_0 + \beta_3 t_{ij}, & \text{falls Medikament B.} \end{cases} \quad (5.5)$$

Die Existenz eines Behandlungseffekts kann nun untersucht werden, indem getestet wird, ob die Differenz von  $\beta_2$  und  $\beta_3$  ungleich Null ist. Der Behandlungsunterschied lässt sich dabei pro Zeiteinheit  $t_{ij}$  durch den Faktor  $\exp((\beta_2 - \beta_3) t_{ij})$  berechnen. Das Modell in Gleichung (5.4) und gegebenenfalls das Modell in Gleichung (5.5) werden im Fall der Medikament-B-Studie für alle sechs Studiendesigns mit Hilfe der Prozedur `GENMOD` ausgewertet.

### 5.3.4 Auswertung der Beispiele im Fall marginaler Modelle

In diesem Abschnitt werden die im vorherigen Abschnitt vorgestellten Beispiele und Modelle mit Hilfe der Prozedur `GENMOD` ausgewertet. Die Funktionsweise der Prozedur wurde im Abschnitt 5.1 erläutert.

#### 5.3.4.1 Auswertung der Medikament-A-Studie

Die Prozedur zur Auswertung des Modells in (5.2) ist gegeben durch:

```
proc genmod data=hot_flushes_a;
class subject treat day;
model number_of_hf=treat/ dist=poisson;
repeated subject=subject/ corrw covb
type=exch within=day;
run;
```

wobei `hot_flushes_a` gerade der Name des verwendeten Datensatzes ist. Als qualitative Daten werden die Größen

- `subjekt`  $\hat{=}$  Nummer des Patienten,
- `treat`  $\hat{=}$  Code für die Behandlung (0 für Placebo, 1 für Medikament A) und
- `day`  $\hat{=}$  Tagesnummer (von Tag 1 bis Tag 7)

angegeben. Durch die Anweisung `model` wird die Modellgleichung (5.2) vorgegeben. Für diesen Fall heißt das folglich, dass die Anzahl der Hitzewallungen in Abhängigkeit von der Behandlung modelliert wird.

Darüber hinaus wird die zugrundeliegende Verteilung festgelegt. Da die Anzahl der Hitzewallungen Zähldaten liefert, wird die Poissonverteilung benutzt. Als Linkfunktion wird der natürliche Logarithmus, d.h. die natürliche Linkfunktion in Gleichung (5.2) verwendet. Eine Angabe der `link`-Anweisung ist in diesem Fall somit nicht erforderlich.

Wie im Abschnitt 5.1 erläutert, setzt die Prozedur `GENMOD` die Methode der Verallgemeinerten Schätzgleichungen (GEE) von Liang und Zeger (1986) um. Daher bedarf die Auswertung des Datensatzes der Vorgabe einer Arbeitskorrelationsstruktur. Das durch Gleichung (5.2) aufgestellte Modell berücksichtigt keinerlei Zeiteffekte, daher können die Beobachtungen für ein Subjekt beliebig permutiert werden, ohne dass dies Einfluss auf die geschätzten Größen hat. Aus diesem Grund erscheint es nach Abschnitt 3.1.1.3 sinnvoll, für die Analyse die *exchangeable*-Korrelationsstruktur zu wählen, da diese Struktur es ermöglicht, eine Gleichheit der Korrelationen zwischen unterschiedlichen Beobachtungen zu modellieren. Dementsprechend wird in der obigen Prozedur für die `type`-Anweisung diese Struktur gewählt, deren Kurzform in SAS durch `exch` gegeben ist.

Schließlich muss nur noch festgelegt werden, nach welcher Größe die wiederholten Beobachtungen sortiert werden sollen.

Die Auswertung der Prozedur, wie in Abschnitt 5.1 beschrieben, liefert die folgenden Informationen. Dabei wird die Tabelle mit den Größen zur Anpassungsgüte des Modells ebenfalls aufgeführt; allerdings verlieren diese Größen, wie bereits erläutert, für korrelierte Daten ihre Aussagekraft.

**Die Prozedur GENMOD**

**Modellinformationen**

Datei	WORK.HOT_FLUSHES_A	
Verteilung	Poisson	
Verknüpfungsfunktion	Log	
Abhängige Variable	number_of_hf	number of hf

Number of Observations Read	700
Number of Observations Used	700

**Klassifizierungsausprägungsinformationen**

Klasse	Ausprägungen	Werte
SUBJECT	100	1 17 23 24 25 28 30 31 33 35 37 38 41 42 50 51 53
		56 57 59 62 63 65 75 79 83 84 97 98 103 104 106
		107 110 118 119 126 127 129 132 134 135 137 138
		143 144 145 155 156 157 165 192 193 194 199 203
		204 205 206 209 210 213 216 218 219 221 224 225
		227 230 ...
TREAT	2	0 1
Day	7	1 2 3 4 5 6 7

**Parameterinformationen**

Parameter	Effekt	TREAT
Prm1	Intercept	
Prm2	TREAT	0
Prm3	TREAT	1

**Kriterien für Bewertung der Anpassungsgüte**

Kriterium	Freiheits- grade	Wert	Wert/DF
Devianz	698	3107.0407	4.4513
Skalierte Devia	698	3107.0407	4.4513
Pearsonsches Chi-Q	698	3524.6051	5.0496

**Kriterien für Bewertung der Anpassungsgüte**

Kriterium	Freiheits- grade	Wert	Wert/DF
Skaliertes Pearso	698	3524.6051	5.0496
Log-Likelihood		1342.8345	

HINWEIS: Algorithmus konvergiert.

**Analyse der ursprgl. Parameterschätzer**

Parameter		Freiheits- grade		Standardfehler	Waldsch 95% Konfidenzgrenzen		Chi- Quadrat
		Schätzwert					
Intercept		1	0.7348	0.0382	0.6600	0.8097	370.41
TREAT	0	1	0.9944	0.0440	0.9081	1.0806	510.74
TREAT	1	0	0.0000	0.0000	0.0000	0.0000	.
Skala		0	1.0000	0.0000	1.0000	1.0000	

**Analyse der ursprgl.  
Parameterschätzer**

Parameter		Pr > ChiSq
Intercept		<.0001
TREAT	0	<.0001
TREAT	1	.
Skala		

HINWEIS: The scale parameter was held fixed.

**GEE-Modellinformationen**

Korrelationsstruktur	Austauschbar
Innerhalb-Subjekt Effekt	Day (7 Auspräg.)
Subjekteffekt	SUBJECT (100 Auspräg.)
Anzahl der Cluster	100
Korrelationsmatrix-Dimension	7
Maximum Clustergröße	7
Minimum Clustergröße	7

**Kovarianzmatrix (Modellbasiert)**

	Prm1	Prm2
Prm1	0.04627	-0.04627
Prm2	-0.04627	0.06145

**Kovarianzmatrix (Empirisch)**

	Prm1	Prm2
Prm1	0.04580	-0.04580
Prm2	-0.04580	0.06108

HINWEIS: Algorithmus konvergiert.

**Working-Korrelationsmatrix**

	Col1	Col2	Col3	Col4	Col5	Col6	Col7
Row1	1.0000	0.8810	0.8810	0.8810	0.8810	0.8810	0.8810
Row2	0.8810	1.0000	0.8810	0.8810	0.8810	0.8810	0.8810
Row3	0.8810	0.8810	1.0000	0.8810	0.8810	0.8810	0.8810
Row4	0.8810	0.8810	0.8810	1.0000	0.8810	0.8810	0.8810
Row5	0.8810	0.8810	0.8810	0.8810	1.0000	0.8810	0.8810
Row6	0.8810	0.8810	0.8810	0.8810	0.8810	1.0000	0.8810
Row7	0.8810	0.8810	0.8810	0.8810	0.8810	0.8810	1.0000

**Austauschbare  
Working-Korrelation**

Korrelation    0.8810283599

**GEE-Parameterschätzeranalyse  
Empirische Standardfehlerschätzer**

Parameter	Schätzwert	Standardfehler	95%		Z	Pr >  Z
			Konfidenzgrenzen			
Intercept	0.7348	0.2140	0.3153	1.1543	3.43	0.0006
TREAT    0	0.9944	0.2472	0.5100	1.4788	4.02	<.0001
TREAT    1	0.0000	0.0000	0.0000	0.0000	.	.

Die Tabelle der *GEE-Parameterschätzeranalyse / Empirische Standardfehlerschätzer* ist von besonderem Interesse, da hier die Schätzer für die Regressionskoeffizienten  $\beta_0$  (vgl. erste Zeile) und  $\beta_1$  (vgl. zweite Zeile) angegeben werden. Auf den ersten Blick fällt auf, dass der Regressionskoeffizient für TREAT = 1 (vgl. dritte Zeile) auf Null gesetzt wird. Dies liegt an der internen Parametrisierung, die durch die Prozedur vorgenommen wird. So versucht SAS, das Modell

$$\begin{aligned} \ln(\lambda_{ij}) &= x_{ij}^\top \beta \\ &= \beta_0 + \beta_1 x_1(y_{ij}) + \beta_2 x_2(y_{ij}) \\ &= \beta_0 + \beta_1 \mathbb{1}_{\{x_{ij}=\text{Placebo}\}}(y_{ij}) + \beta_2 \mathbb{1}_{\{x_{ij}=\text{MedikamentA}\}}(y_{ij}) \end{aligned}$$

anzupassen. Dieses Modell ist allerdings überparametrisiert, so dass die Planungsmatrix, gegeben durch

$$\begin{pmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix},$$

singulär ist. Aus diesem Grund wird  $\beta_2$  auf Null gesetzt und daher das Modell

$$\begin{aligned} \ln(\lambda_{ij}) &= \beta_0 + \beta_1 \mathbb{1}_{\{x_{ij}=\text{Placebo}\}}(y_{ij}) \\ &= \begin{cases} \beta_0 + \beta_1, & \text{falls Placebo,} \\ \beta_0, & \text{falls Medikament A,} \end{cases} \end{aligned}$$

ausgewertet. Dieses Modell entspricht gerade dem in Gleichung (5.2) vorgeschlagenen Modell.

Neben den Regressionskoeffizienten werden an gleicher Stelle die Standardabweichungen und die 95%-Konfidenzintervalle angegeben. Mit Hilfe des Z-Tests wird

getestet, ob die jeweiligen Regressionskoeffizienten gleich Null sind. Wird ein Signifikanzniveau von  $\alpha^* = 0.05$  festgelegt, können die Nullhypothesen verworfen und  $\beta_0$  und  $\beta_1$  dementsprechend als ungleich Null angenommen werden. Ist  $\beta_1 \neq 0$ , ist dies gleichbedeutend mit dem Vorliegen eines signifikanten Behandlungsunterschieds, welcher gerade durch den Faktor

$$\begin{aligned} \exp(\beta_1) &= \exp(0.9944) \\ &\approx 2.7031 \end{aligned}$$

gegeben ist. Das heißt, dass die durchschnittliche Anzahl der Hitzewallungen in der Placebo-Gruppe um den Faktor 2.7031 größer ist als die Anzahl in der Medikament-A-Gruppe:

$$E(Y_{\text{Med. A}}) \cdot 2.7031 = E(Y_{\text{Placebo}}).$$

Für die Placebo-Gruppe erhält man

$$\begin{aligned} \ln(E(Y_{ij})) &= \ln(\lambda_{ij}) \\ &= \beta_0 + \beta_1, \end{aligned}$$

und für den Erwartungswert von  $Y_{ij}$ , d.h. für die täglich zu erwartende Hitzewallungsanzahl in der Placebo-Gruppe, ergibt sich entsprechend

$$\begin{aligned} \hat{E}(Y_{ij}) = \exp(\hat{\beta}_0 + \hat{\beta}_1) &= \exp(0.7348 + 0.9944) \\ &\approx 5.6361. \end{aligned}$$

Für den Erwartungswert in der Medikament-A-Gruppe erhält man dagegen eine durchschnittliche Hitzewallungsanzahl von

$$\begin{aligned} \hat{E}(Y_{ij}) = \lambda_{ij} = \exp(\hat{\beta}_0) &= \exp(0.7348) \\ &\approx 2.0851, \end{aligned}$$

so dass der signifikante Behandlungsunterschied noch einmal deutlich wird.

Die ermittelten GEE-Schätzer basieren auf der Annahme der *exchangeable*-Korrelationsstruktur, allerdings unterscheiden sich die Schätzer für alternative Arbeitsstrukturen nur minimal. Darüber hinaus ergibt sich auch für andere Korrelationsstrukturen ein signifikanter Behandlungsunterschied, wie Tabelle 5.1 zeigt.

Aufgrund der hohen Anzahl zu schätzender Größen konnten mit Hilfe der Prozedur GENMOD keine Ergebnisse für die *unstructured*-Korrelationsmatrix ermittelt werden. Es ist unklar, ob dieses Problem lediglich auf numerischen Schwierigkeiten beruht.<sup>20</sup>

---

<sup>20</sup>Vgl. Abhandlung in Abschnitt 3.1.1.3.

	IND	EXCH	AR(1)
$\hat{\beta}_0$	0.7472	0.7348	0.7578
$\hat{\beta}_1$	0.9866	0.9944	0.9897
$\hat{\beta}_2$	0.0000	0.0000	0.0000
SD $\hat{\beta}_0$	0.2163	0.2140	0.2158
SD $\hat{\beta}_1$	0.2489	0.2472	0.2502
SD $\hat{\beta}_2$	0.0000	0.0000	0.0000
p-Wert $\hat{\beta}_0$	0.0006	0.0006	0.0004
p-Wert $\hat{\beta}_1$	<.0001	<.0001	<.0001
p-Wert $\hat{\beta}_2$	.	.	.

Tabelle 5.1: Dargestellt sind die geschätzten Regressionsparameter, Standardabweichungen (SD) und p-Werte für unterschiedliche Korrelationsstrukturen. Dabei steht IND für Unabhängigkeit, EXCH für Austauschbarkeit und AR(1) für den autoregressiven Prozess erster Ordnung.

#### 5.3.4.2 Auswertung der Medikament-B-Studie

Die Medikament-B-Studie wird in zwei Schritten analysiert. Zunächst wird das Modell in Gleichung (5.4) betrachtet und im Anschluss wird die Auswertung des Modells in Gleichung (5.5) nur dann vorgenommen, wenn die Hypothese  $H_0^1 : \beta_1 = 0$  zum Signifikanzniveau  $\alpha^* = 0.05$  nicht verworfen werden kann. Die Prozedur zur Analyse des Modells (5.4) entspricht dem Vorgehen bei der Medikament-A-Studie und ist für Design  $i \in \{1, 2, 4, 5, 6\}$  gegeben durch:

```
proc genmod data=hot_flushes_b_design_i;
class PID Week Day Treat;
model Summe=treat treat*time/ dist=poisson;
repeated subject=PID/ corrw covb
type=AR(1) within=Day;
run;
```

und für das Design 3 gegeben durch:

```
proc genmod data=hot_flushes_b_design_3;
class PID Week Treat;
model Summe=treat treat*time/ dist=poisson;
repeated subject=PID/ corrw covb
type=AR(1) within=Week;
run;
```

Diese beiden Aufrufe der Prozedur `GENMOD` unterscheiden sich lediglich in der `within`-Anweisung. In der Design-3-Studie wird pro Woche nur eine Beobachtung benutzt, so dass die Daten eines Subjekts bezüglich der Wochenanzahl angeordnet sind.

Der Name des betroffenen Datensatzes ist durch `hot_flushes_b_design_i` gegeben und folgende qualitative Größen werden angewandt:

- `PID`  $\hat{=}$  Nummer des Patienten;
- `Week`  $\hat{=}$  Wochennummer (Woche 1 bis Woche 12);
- `Day`  $\hat{=}$  Tagesnummer ( Tag 1 bis Tag 84) sowie
- `Treat`  $\hat{=}$  Code für das Medikament (0 für Placebo und 1 für Medikament B).

Es ist zu beachten, dass sowohl die Variable `time` als auch die Variable `Day` genutzt werden. Beide Variablen enthalten die Tagesnummer (Tag 1 bis Tag 84), allerdings wird die Variable `Day` bzw. `Week` als qualitative Variable gekennzeichnet, damit diese dann in der `within`-Anweisung stehen kann. Die `time`-Variable ist eine stetige Variable, wie im Regressionsmodell mit Zeiteffekten vorgesehen.

Ziel der Medikament-B-Studie ist es, den Behandlungseffekt in Abhängigkeit von der Behandlung und der Zeit zu untersuchen. In der `model`-Anweisung wird zunächst vorgegeben, dass die Summe der mittleren und starken Hitzewallungen, zusammengefasst in der Variable `summe`, gerade der abhängigen Variablen entspricht. Diese soll dann gemäß der Modellierung in Gleichung (5.4) in Abhängigkeit von der Behandlung (`treat`) und den behandlungsspezifischen Zeiteffekten (`treat*time`) modelliert werden.

Aufgrund der Datenerhebung über einen längeren Zeitraum ist es realistisch anzunehmen, dass die Korrelation der Daten abhängig von der Zeit ist. So wird vorausgesetzt, dass zeitlich weit auseinander liegende Beobachtungen schwächer korreliert sind als nah zusammen liegende Beobachtungen. Diese Korrelationsstruktur entspricht der eines autoregressiven Prozesses erster Ordnung (`AR(1)`), so dass in der `type`-Anweisung diese Struktur mit der Kurzform `AR(1)` festgelegt wird.<sup>21</sup>

Die Ausführung der Prozedur für die sechs Designs liefert die in Tabelle 5.2 aufgelisteten Schätzer und p-Werte für den Regressionsparameter  $\beta_1$ . Die vollständigen Tabellen mit den GEE-Parameterschätzungen sind im Anhang D.1 zu finden.

In jedem der Designs ergibt sich für  $\beta_1$  ein p-Wert, der größer ist als  $\alpha^* = 0.05$ , somit kann die Nullhypothese  $H_0 : \beta_1 = 0$  zum Niveau  $\alpha$  nicht verworfen werden. Dementsprechend kann für jedes der sechs Designs mit dem Modell in Gleichung

---

<sup>21</sup>Vgl. Brockwell und Davis (1991).

Design	$\hat{\beta}_1$	SD	95%-Konfidenzintervalle	p-Wert
Design 1	0.0252	0.0732	[-0.1182, 0.1686]	0.7303
Design 2	-0.0138	0.1131	[-0.2356, 0.2079]	0.9027
Design 3	0.0531	0.0864	[-0.1163, 0.2225]	0.5388
Design 4	0.2546	0.2378	[-0.2115, 0.7206]	0.2844
Design 5	0.0028	0.48	[-0.9379, 0.9434]	0.9954
Design 6	-0.4696	0.784	[-2.0062, 1.0669]	0.5492

Tabelle 5.2: Schätzer, p-Werte und Standardabweichungen (SD) für  $\beta_1$  in der Analyse von Modell (5.4).

(5.5) fortgefahen werden. In diesem Modell wird die Differenz von  $\beta_2$  und  $\beta_3$  mitgeschätzt, um anschließend zu testen, ob diese Differenz gleich Null ist. Wird diese Hypothese zum Signifikanzniveau  $\alpha^*$  verworfen, ist ein signifikanter Behandlungsunterschied nachgewiesen.

Zu diesem Zweck wird zusätzlich die Option `estimate` in der `GENMOD`-Prozedur benutzt, um einen Vektor  $L$  angeben und die Hypothese  $H_0 : L\beta = 0$  testen zu können. Um die Hypothese

$$\begin{aligned}
 H_0 : \beta_2 - \beta_3 &= 0 \\
 \Leftrightarrow H_0 : \begin{pmatrix} 1 & -1 \end{pmatrix} \begin{pmatrix} \beta_2 \\ \beta_3 \end{pmatrix} &= 0
 \end{aligned}$$

zu testen, verändert sich die Prozedur-Anweisung für Design  $i \in \{1, 2, 4, 5, 6\}$  zu

```

proc genmod data=hot_flushes_b_design_i;
class PID Week Day Treat;
model Summe= treat*time/ dist=poisson;
repeated subject=PID/ corrw covb
type=AR(1) within=Day;
estimate 'diff slopes' treat*time 1 -1
run;

```

und für das dritte Design zu

```

proc genmod data=hot_flushes_b_design_3;
class PID Week Treat;
model Summe=treat treat*time/ dist=poisson;
repeated subject=PID/ corrw covb

```

```

type=AR(1) within=week;
estimate 'diff slopes' treat*time 1 -1
run;
    
```

Es ergeben sich die in Tabelle 5.3 aufgelisteten Schätzer, Standardabweichungen (SD) und p-Werte für die Differenz  $\beta_2 - \beta_3$ .

Design	$\widehat{\beta_2 - \beta_3}$	SD	95%-Konfidenzintervalle	p-Wert
Design 1	-0.0045	0.0019	[-0.0083, -0.0007]	0.0195
Design 2	-0.0037	0.0017	[-0.0070, -0.0004]	0.0279
Design 3	-0.0325	0.0113	[-0.0546, -0.0104]	0.0040
Design 4	-0.0048	0.0035	[-0.0117, 0.0021]	0.1737
Design 5	-0.0054	0.0024	[-0.0100, -0.0008]	0.0224
Design 6	-0.0044	0.0016	[-0.0076, -0.0012]	0.0076

Tabelle 5.3: Schätzer und p-Werte für die Analyse des Modells in Gleichung (5.5).

In allen Designs bis auf Design 4 wird die Hypothese,  $H_0 : \beta_2 - \beta_3 = 0$  somit zum Signifikanzniveau  $\alpha^* = 0.05$  verworfen. Dies bedeutet folglich, dass in diesen Designs ein signifikanter Behandlungsunterschied festzustellen ist. Insgesamt ergeben sich für den Behandlungsfaktor  $\exp((\beta_2 - \beta_3) t_{ij})$ , um den sich die Anzahl der Hitzewallungen in der Medikament-B-Gruppe gegenüber der Anzahl in der Placebo-Gruppe reduziert, die Größen in Tabelle 5.4.

Design	$\exp(\beta_2 - \beta_3)$
Design 1	0.9955
Design 2	0.9963
Design 3	0.9680
Design 4	k.A.
Design 5	0.9946
Design 6	0.9956

Tabelle 5.4: Die signifikanten multiplikativen Behandlungseffekte für die Analyse des alternativen Modells in (5.5) und  $t_{ij} = 1$ .

Für das Design 3 ergibt sich beispielsweise die in Abbildung 5.1 gezeigte grafische Veranschaulichung der Entwicklung in beiden Behandlungsgruppen. Der Rückgang der Anzahl an Hitzewallungen in der Medikament-B-Gruppe ist sehr gut zu erken-

## Entwicklung der Hitzewallungsanzahl im Design 3

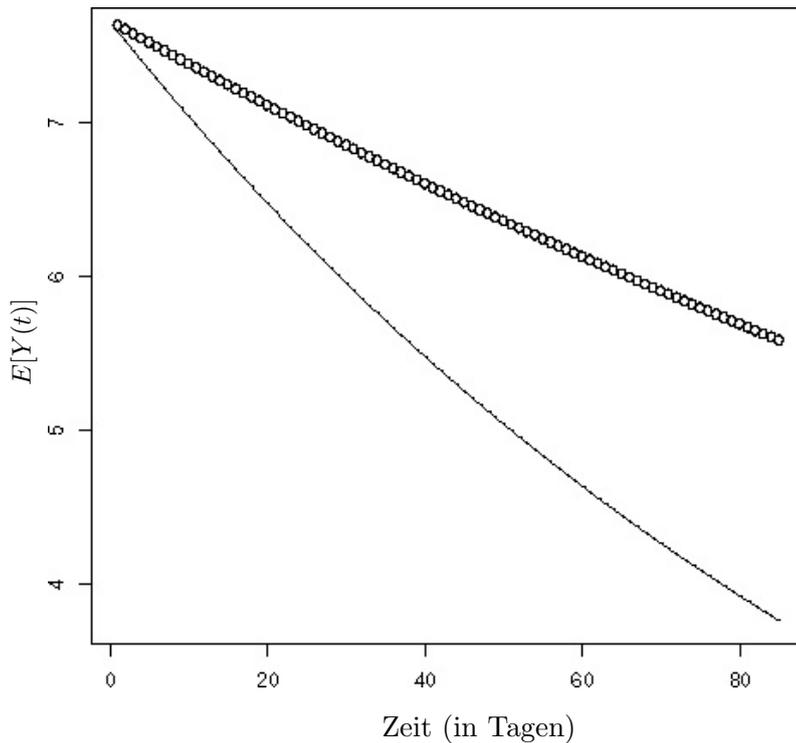


Abbildung 5.1: Die gepunktete Linie spiegelt die Entwicklung in der Placebo-Gruppe und die durchgezogene Linie die Entwicklung in der Medikament-B-Gruppe wieder.

nen. Allerdings ist nicht klar, wodurch die ebenfalls stete Reduktion in der Placebo-Gruppe zu begründen ist. Ob es sich bei diesem Effekt lediglich um den sogenannten Placeboeffekt handelt, oder ob die Verblindung fehlgeschlagen ist, kann im Rahmen dieser Analyse nicht beantwortet werden.

### 5.3.5 Modellierung im Fall eines subjekt-spezifischen Modells

Im Fall subjekt-spezifischer Modelle wird lediglich die Medikament-A-Studie ausgewertet. Für diese Studie soll ein einfaches Regressionsmodell angepasst werden. Analog zum Vorgehen bei der marginalen Modellierung werden Baseline-Werte und eventuell existierende Zeiteffekte in der Analyse nicht berücksichtigt.

Als Modellgleichung ergibt sich daher

$$Y_{ij}|U_i \sim \mathcal{P}(\lambda_{ij}),$$

$$\ln(\lambda_{ij}) = x_{ij}^\top \beta + d_{ij}^\top U_i$$

$$= \begin{cases} \beta_0 + U_i + \beta_1, & \text{falls Placebo,} \\ \beta_0 + U_i, & \text{falls Medikament A,} \end{cases} \quad (5.6)$$

wobei  $U_i \sim \mathcal{N}(0, D)$  die subjekt-spezifische Zufallsvariable für alle  $i \in \{1, \dots, m\}$  ist. Im Folgenden müssen die Frage nach einem existierenden Behandlungseffekt beantwortet und Schätzer der Regressionsparameter geliefert werden.

### 5.3.6 Auswertung mittels subjekt-spezifischer Modelle

Ein Programm zur Auswertung des Modells (5.6) ist gegeben durch

```
proc nlmixed data=hot_flushes_a;
parms beta_0=3 beta_1=0.5 D=0.25;
if(treat=0) then eta=beta_0 + u + beta_1;
else if
(treat=1) then eta=beta_0 + u; lambda=exp(eta);
model number_of_hf~poisson(lambda);
random u~normal(0,D)
subject= subject;
run;
```

wobei `hot_flushes_a` der Name des Datensatzes ist und `subject` die Nummer eines Patienten angibt. Durch die `parms`-Anweisung werden die Startwerte der Iteration auf  $\beta_0^0 = 3$ ,  $\beta_1^0 = 0.5$  und  $D^0 = 0.25$  festgesetzt. Zudem wird durch die `if`-Bedingung das Modell (5.6) beschrieben, d.h.

$$\eta = \begin{cases} \beta_0 + U + \beta_1, & \text{falls Placebo,} \\ \beta_0 + U, & \text{falls Medikament A,} \end{cases} \quad (5.7)$$

und entsprechend dem Modell (5.6) gilt somit  $\eta = \ln(\lambda)$ . In der `model`-Anweisung wird angegeben, dass die Anzahl der Hitzewallungen in Abhängigkeit von der Behandlung modelliert werden und dass die Verteilung von der Anzahl der Hitzewallungen bedingt durch die subjekt-spezifischen Größen  $U_i$ ,  $i \in \{1, \dots, m\}$ , einer Poissonverteilung mit Parameter `lambda=exp(eta)` entsprechen soll.

Des Weiteren wird festgesetzt, dass die subjekt-spezifischen Größen durch die Zufallsvariablen  $U_i$ ,  $i \in \{1, \dots, m\}$ , gegeben sind und dass die wiederholten Beobachtungen zu der Variablen `subject` erhoben wurden.

Die Auswertung dieser Prozedur liefert die folgenden Informationen und Tabellen:

## Die Prozedur NLMIXED

## Spezifikationen

Data Set	WORK.HOT_FLUSHES_A
Abhängige Variable	number_of_hf
Verteilung für abhängige Variable	Poisson
Zufällige Effekte	u
Verteilung für zufällige Effekte	Normal
Subjektvariable	SUBJECT
Optimierungstechnik	Quasi-Newton, dual
Integrationsmethode	Adaptive Gaussische Quadratur

## Dimensionen

Verwendete Beobachtungen	700
Nicht verwendete Beobachtungen	0
Gesamtbeobachtungen	700
Subjekte	100
Max Beob. je Subjekt	7
Parameter	3
Quadraturpunkte	1

## Parameter

beta_0	beta_1	D	NegLogLike
3	0.5	0.25	2241.67515

## Iterationsverlauf

Iteration	Aufrufe	NegLogLike	Diff	MaxGrad	Steigung
1	3	1310.23179	931.4434	102.4689	-111335
2	5	1212.66329	97.5685	23.71153	-97.1026
3	7	1206.46624	6.197051	23.7028	-6.81824
4	9	1190.7167	15.74953	7.840594	-7.43171
5	10	1190.28209	0.434613	6.963515	-0.76918
6	12	1190.21348	0.068608	6.595554	-0.04877
7	15	1188.20633	2.007156	1.840439	-0.08286
8	17	1187.95255	0.253781	1.133296	-0.34721
9	19	1187.89836	0.054188	0.210197	-0.09017
10	21	1187.89678	0.001575	0.016962	-0.00322
11	23	1187.89675	0.000035	0.003236	-0.00004
12	25	1187.89675	3.616E-7	0.000064	-7.07E-7

HINWEIS: GCONV convergence criterion satisfied.

Anpassungsstatistiken

-2 Log Likelihood	2375.8
AIC (kleiner ist besser)	2381.8
AICC (kleiner ist besser)	2381.8
BIC (kleiner ist besser)	2389.6

Parameterschätzwerte

Parameter	Schätzwert	Standardfehler	Freiheits- grade	t-Wert	Pr >  t	Alpha	Untere
beta_0	-0.3592	0.2619	99	-1.37	0.1734	0.05	-0.8788
beta_1	1.4098	0.3439	99	4.10	<.0001	0.05	0.7275
D	2.6170	0.5016	99	5.22	<.0001	0.05	1.6218

Parameterschätzwerte

Parameter	Obere	Gradient
beta_0	0.1605	0.000064
beta_1	2.0921	0.000027
D	3.6122	-6.62E-7

Da die Anweisung `qpoints` beim Aufruf der Prozedur fehlte, wird die Anzahl der Quadraturpunkte adaptiv berechnet. Bei dem aufgestellten Modell und dem gegebenen Datensatz ergibt sich hier lediglich ein benötigter Quadraturpunkt.

Die Ausführung der Prozedur liefert das Verwerfen der Nullhypothese  $H_0 : \beta_1 = 0$  zum Signifikanzniveau  $\alpha^* = 0.05$ . Ferner ergibt sich für die subjekt-spezifischen Größen die Verteilung  $U_i \sim \mathcal{N}(0, 2.617)$  für alle  $i \in \{1, \dots, m\}$ .

Wie bereits in Kapitel 4 beschrieben, unterscheidet sich die Interpretation des Regressionsvektors  $\beta$  im subjekt-spezifischen Modell von jener des Regressionsvektors im marginalen Modell. Die durch `NLMIXED` ermittelten Schätzer für  $\beta_k$ ,  $k \in \{0, 1\}$ , ermöglichen im Gegensatz zu denjenigen des marginalen Modells keine direkte Aussage über den Populationsdurchschnitt.

Um trotzdem Aussagen über den Populationsdurchschnitt tätigen zu können, muss

$$\begin{aligned}
 E(Y_{ij}) &= E[E(Y_{ij}|U_i)] \\
 &= \begin{cases} E\left[\exp\left(\hat{\beta}_0 + U_i + \hat{\beta}_1\right)\right], & \text{falls Placebo,} \\ E\left[\exp\left(\hat{\beta}_0 + U_i\right)\right], & \text{falls Medikament A} \end{cases}
 \end{aligned}$$

berechnet bzw. geschätzt werden.

Im Rahmen dieser Studie wird der Erwartungswert nicht mit Hilfe von numerischen Integrationsmethoden geschätzt. Vielmehr werden  $M = 2.000.000$  Realisierungen  $u_k$

der Zufallsgrößen  $U_i$  erzeugt und im Anschluss

$$E(Y_{ij}) = \begin{cases} \frac{1}{M} \sum_{k=1}^M \exp(-0.3592 + u_k + 1.4098), & \text{falls Placebo,} \\ \frac{1}{M} \sum_{k=1}^M \exp(-0.3592 + u_k), & \text{falls Medikament A} \end{cases}$$

berechnet.<sup>22</sup> Auf diese Weise ergibt sich eine durchschnittliche tägliche Hitzewal-  
lungsanzahl von

$$E(Y_{ij}) = \begin{cases} 10.62, & \text{falls Placebo,} \\ 2.58, & \text{falls Medikament A.} \end{cases} \quad (5.8)$$

Wird die Analyse für unterschiedliche Anzahlen an Quadraturpunkten  $Q$  durch-  
geführt, ergeben sich bei der Anwendung der adaptiven Hermite-Gauß-Quadratur  
und Maximierung mittels der dualen Quasi-Newton-Methode nur geringe Unter-  
schiede zwischen den Parameterschätzungen (vgl. Tabelle 5.5).

Quadraturpunkte	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{D}$
$Q = 1$	-0.3592	1.4098	2.6170
$Q = 2$	-0.3522	1.4055	2.5836
$Q = 5$	-0.3544	1.4052	2.6324
$Q = 10$	-0.3560	1.4060	2.6464
$Q = 20$	-0.3560	1.4061	2.6471

Tabelle 5.5: Ermittelte Schätzer für das Modell in Gleichung (5.6) mit unterschied-  
lichen Anzahlen an Quadraturpunkten  $Q$ .

Dieses Resultat steht im Gegensatz zu den Resultaten verschiedener, von Molenberg-  
hs und Verbeke (2005) analysierten Studien, bei denen sich in Abhängigkeit von der  
Quadraturpunkteanzahl große Unterschiede zwischen den geschätzten Parametern  
ergaben.<sup>23</sup>

Bei der Analyse durch die nicht-adaptiven Hermite-Gauß-Quadratur werden, ohne  
Angabe der `qpoints`-Anweisung, allerdings  $Q = 41$  Quadraturpunkte benötigt. Die  
Analyse liefert folgende Parameterschätzer:

$$\hat{\beta}_0 = 1.1689, \hat{\beta}_1 = -0.4418 \text{ und } \hat{D} = 1.8510.$$

<sup>22</sup>Die SAS-Anweisung zur Berechnung des marginalisierten Erwartungswertes ist in Anhang E  
gegeben.

<sup>23</sup>Vgl. beispielsweise *Toenail Studie* in Molenberghs und Verbeke (2005), S.278ff.

Die Schätzer, die mit Hilfe der adaptiven Quadratur ermittelt wurden, unterscheiden sich somit wesentlich von den Schätzern der nicht-adaptiven Quadratur. Gemäß dem Vorgehen auf Seite 107 gilt für die durchschnittliche Anzahl an Hitzewallungen

$$E(Y_{ij}) = \begin{cases} \frac{1}{M} \sum_{k=1}^M \exp(1.1689 + u_k - 0.4418), & \text{falls Placebo,} \\ \frac{1}{M} \sum_{k=1}^M \exp(1.1689 + u_k), & \text{falls Medikament A} \end{cases}$$

und  $u_k$  sind Realisierungen der Zufallsvariable  $U \sim \mathcal{N}(0, 1.851)$ . Mit  $M = 2.000.000$  ergibt sich

$$E(Y_{ij}) = \begin{cases} 5.23, & \text{falls Placebo,} \\ 8.14, & \text{falls Medikament A} \end{cases}$$

und somit ein Ergebnis, dass den Resultaten der adaptiven Hermite-Gauß-Quadratur, bei der nur ein Quadraturpunkt verwendet wurde, widerspricht. Weshalb die nicht-adaptive Quadratur ein solch unterschiedliches Ergebnis liefert, konnte im Rahmen dieser Arbeit nicht ermittelt werden. Für alternative Startwerte ( $\beta_0 = 2$ ,  $\beta_1 = 1$  und  $D = 1$ ) ergeben sich die folgenden Schätzer:

$$\hat{\beta}_0 = 0.94, \hat{\beta}_1 = 0.2298 \text{ und } \hat{D} = 1.6474$$

und für die marginalisierte Erwartung von  $Y_{ij}$  folgt

$$E(Y_{ij}) = \begin{cases} 7.19, & \text{falls Placebo,} \\ 5.83, & \text{falls Medikament A.} \end{cases}$$

Für ein und denselben Datensatz gibt es somit im Fall der nicht-adaptiven Quadratur sehr unterschiedliche Schätzungen und marginale Erwartungen. Darüber hinaus konvergiert die Prozedur in diesem Fall nicht für alle Startwerte, so dass der Nutzer aufgefordert werden kann, neue Startwerte zu wählen. Selbst wenn die von NLMIXED durchgeführte Iteration konvergiert ist dies, wie oben verdeutlicht, somit keineswegs ein Gütekriterium.

## 5.4 Vergleich der Ergebnisse

Im Kapitel 4 wurde erläutert, dass sich die Interpretationen der Regressionsparameter in den marginalen und subjekt-spezifischen Modellen unterscheiden. In den letzten Abschnitten wurde die durchschnittliche Hitzewallungsanzahl bestimmt. Für

das marginale Modell ergab sich eine durchschnittliche tägliche Hitzewallungsanzahl von

$$E(Y_{ij}) = \begin{cases} 2.09, & \text{falls Medikament A,} \\ 5.64, & \text{falls Placebo.} \end{cases}$$

Die Analyse des subjekt-spezifischen Modells lieferte bei der Analyse mittels adaptiver Hermite-Gauß-Quadratur eine durchschnittliche tägliche Hitzewallungsanzahl von

$$E(Y_{ij}) = \begin{cases} 2.58, & \text{falls Medikament A,} \\ 10.62, & \text{falls Placebo.} \end{cases}$$

Die durchschnittlichen Entwicklungen in der Medikament-A-Gruppe unterscheiden sich somit nur gering. Im Gegensatz dazu liefert die Analyse des subjekt-spezifischen Modells in der Placebo-Gruppe eine fast doppelt so hohe durchschnittliche Anzahl an Hitzewallungen als bei der Analyse des marginalen Modells.

Im Kapitel 4 wurde bereits darauf hingewiesen, dass sich die Entwicklung des Populationsdurchschnitts im marginalen und im subjekt-spezifischen Modell in Abhängigkeit von der Varianz der subjekt-spezifischen Größen unterscheidet. Diese Varianz wurde im letzten Abschnitt durch den Wert 1.851 geschätzt. Um aber sicherzustellen, dass der große Unterschied der Erwartungswerte lediglich durch die Variabilität der subjekt-spezifischen Größen hervorgeht, soll im nächsten Abschnitt eine weitere Analyse durchgeführt werden, um damit eventuell vorhandene Fehler bei der subjekt-spezifischen Analyse ausschließen zu können. Darüber hinaus kann durch die alternative Analyse möglicherweise geklärt werden, ob die Resultate der nicht-adaptiven Analyse im letzten Abschnitt tatsächlich falsch und lediglich mit den Schwächen der Prozedur NLMIXED zu begründen sind.

### 5.4.1 Alternatives subjekt-spezifisches Modell

Das subjekt-spezifische Modell im letzten Abschnitt hatte die folgenden Eigenschaften:

- Bedingt unter den unabhängigen und identisch normalverteilten Größen  $U_i \sim \mathcal{N}(0, D)$  sind für alle  $i \in \{1, \dots, 100\}$  die Zufallsgrößen  $Y_{i1}, \dots, Y_{i7}$  paarweise unabhängig und besitzen die bedingte Dichte

$$f_{Y_{ij}|U_i=u_i}(y_{ij}) = \exp\{y_{ij} \ln(\lambda_{ij}) - \lambda_{ij} - \ln(y_{ij}!)\},$$

d.h.  $Y_{ij}|U_i \sim \mathcal{P}(\lambda_{ij})$ .

- Der bedingte Erwartungswert  $\lambda_{ij} = E(Y_{ij}|U_i)$  erfüllt das verallgemeinerte gemischte lineare Modell

$$\begin{aligned} \ln(\lambda_{ij}) &= x_{ij}^\top \beta + d_{ij}^\top U_i \\ &= \begin{cases} \beta_0 + U_i + \beta_1, & \text{falls Placebo,} \\ \beta_0 + U_i, & \text{falls Medikament A.} \end{cases} \end{aligned} \quad (5.9)$$

Nun ist bekannt, dass die Summe unabhängiger poissonverteilter Zufallsvariablen  $X_k \sim \mathcal{P}(\lambda_k)$  wieder poissonverteilt ist und

$$\sum_{k=1}^n X_k \sim \mathcal{P}\left(\sum_{k=1}^n \lambda_k\right)$$

gilt. Somit lässt sich das obige Modell äquivalent formulieren als:

- Bedingt unter den unabhängigen und identisch normalverteilten Größen  $U_i \sim \mathcal{N}(0, D)$  sind für alle  $i \in \{1, \dots, 100\}$  die Zufallsgrößen  $Y_i = \sum_{j=1}^7 Y_{ij}$  paarweise unabhängig und besitzen die bedingte Dichte

$$f_{Y_i|U_i=u_i}(y_i) = \exp\{y_i \ln(\lambda_i^*) - \lambda_i^* - \ln(y_i!)\}$$

mit  $E(Y_i) = \sum_{j=1}^7 \lambda_{ij} =: \lambda_i^*$ .

- Der bedingte Erwartungswert  $\lambda_i^* = E(Y_i|U_i)$  erfüllt das verallgemeinerte gemischte lineare Modell

$$\begin{aligned} \ln(\lambda_i^*) &= x_i^\top \beta + d_i^\top U_i \\ &= \begin{cases} \beta_0 + U_i + \beta_1, & \text{falls Placebo,} \\ \beta_0 + U_i, & \text{falls Medikament A.} \end{cases} \end{aligned} \quad (5.10)$$

Dabei unterscheiden sich die Regressionsparameter in den Modellgleichungen (5.9) und (5.10), da die Parameter in Modell (5.9) die tägliche Anzahl an Hitzewallungen und die Parameter in Modell (5.10) dagegen die wöchentliche Anzahl charakterisieren.

Sind die Resultate im Abschnitt 5.3.6 korrekt, so ist zu erwarten, dass

$$E(\lambda_i^*)/7 = E(\lambda_{ij}) \stackrel{vgl.(5.8)}{=} 10.62$$

gilt. Um dies nachzuprüfen, wird die Summe der Hitzewallungen jeder Person berechnet und dann das verallgemeinerte gemischte lineare Modell mit Hilfe der Prozedur `NLMIXED` ausgewertet. Ist die Summe der Hitzewallungen durch die Variable `Summe` gegeben, kann folgender Aufruf genutzt werden, um das Modell in (5.10) anzupassen:

```

proc nlmixed data=hot_flushes_a;
parms beta_0=3 beta_1=0.5 D=0.25;
if(treat=0) then eta=beta_0 + u + beta_1;
else if
(treat=1) then eta=beta_0 + u; lambda=exp(eta);
model Summe~poisson(lambda);
random u~normal(0,D)
subject= subject;
run;

```

Die resultierenden Modellinformationen und Tabellen sind im Anhang F gegeben. Für die Parameterschätzer folgt:

Parameter	Schätzer	p-Wert
$\beta_0$	1.5915	< .0001
$\beta_1$	1.4052	< .0001
$D$	2.6324	< .0001

Somit liegt ein signifikanter Behandlungsunterschied zwischen beiden Behandlungsgruppen vor. Die Berechnung der zu erwartenden durchschnittlichen wöchentlichen Anzahl an Hitzewallungen mit einer großen Zahl  $M$  (hier:  $M = 2.000.000$ ) liefert:

$$\begin{aligned}
E(Y_{ij}) &= E[E(Y_{ij}|U_i)] \\
&= \begin{cases} \frac{1}{M} \sum_{k=1}^M \exp(1.5915 + u_k + 1.4042), & \text{falls Placebo,} \\ \frac{1}{M} \sum_{k=1}^M \exp(1.5915 + u_k), & \text{falls Medikament A} \end{cases} \\
&= \begin{cases} 74.57, & \text{falls Placebo,} \\ 18.36, & \text{falls Medikament A.} \end{cases}
\end{aligned}$$

Dabei sind  $u_k$  für alle  $k \in \{1, \dots, M\}$  Realisierungen von  $U \sim \mathcal{N}(0, 2.6324)$ . Die Analyse berechnet die durchschnittliche tägliche Hitzewallungsanzahl als

$$E(Y_{ij}) = \begin{cases} 10.65, & \text{falls Placebo,} \\ 2.62, & \text{falls Medikament A.} \end{cases}$$

Die Analyse in Abschnitt 5.3.6 ergab

$$E(Y_{ij}) = \begin{cases} 10.62, & \text{falls Placebo,} \\ 2.58, & \text{falls Medikament A,} \end{cases}$$

daher stimmen die Resultate beider Analysen fast überein.

Der große Unterschied zwischen den Ergebnissen der Analyse des marginalen und des subjekt-spezifischen Modells beruht somit tatsächlich auf der Varianz der identisch verteilten subjekt-spezifischen Größen  $U_i = U$ ,  $i \in \{1, \dots, m\}$ . Zudem ist dieses Resultat ein weiterer Hinweis dafür, dass die gewonnenen Ergebnisse der nicht-adaptiven Quadratur verfälscht sind.

# Kapitel 6

## Versuchsplanung

Die im Kapitel 3 vorgestellten Analysen konzentrieren sich auf das Schätzen des unbekanntes Regressionsvektors  $\beta$  mit Hilfe von Messwerten bzw. Beobachtungen  $y_{11}, \dots, y_{mn_m}$ . Diese Messwerte wurden im Rahmen einer wissenschaftlichen Studie oder eines Versuchs erhoben.

Das Ziel einer Studie kann eine Entscheidung zwischen zwei sich widersprechenden Aussagen sein. Diese Aussagen werden in der Statistik in Form von Hypothesen formuliert. Bei einer klinischen Studie, die für die Zulassung von Medikamenten notwendig ist, lassen sich diese beispielsweise wie folgt formulieren: „das neue Medikament ist wirksam“ und andererseits „das neue Medikament ist nicht wirksam“. Das Ziel dieser Studie ist es, durch geeignete Datenerhebung zu untersuchen, welche der Hypothesen nicht abgelehnt wird. Eine Entscheidungsregel zwischen zwei sich gegenseitig ausschließenden Hypothesen  $H_0$  und  $H_1$  auf Basis erhobener Daten heißt Test.

Die Entscheidungsregel basiert auf Teststatistiken. Aufgrund der begrenzten Anzahl von Daten der Grundgesamtheit kann es bei der Entscheidung für eine Hypothese zu zwei Arten von Fehlern kommen: Eine Fehlentscheidung liegt vor, wenn die Nullhypothese  $H_0$  verworfen wird, obwohl sie wahr ist. In diesem Zusammenhang wird vom *Fehler 1. Art* gesprochen. Wird die Nullhypothese allerdings beibehalten, obwohl die Alternativhypothese  $H_1$  wahr ist, so wird der Fehler mit *Fehler 2. Art* bezeichnet. Ist die Wahrscheinlichkeit für den Fehler 1. Art durch  $\alpha^*$  gegeben, d.h. gilt

$$P(\text{Fehler 1. Art}) = P(H_0 \text{ ablehnen} | H_0 \text{ ist wahr}) =: \alpha^*$$

mit einem Wahrscheinlichkeitsmaß  $P$ , so wird der Test auch *Test zum Niveau  $\alpha^*$*  genannt. Die Wahrscheinlichkeit den Fehler 2. Art zu begehen sei gegeben durch

$$P(\text{Fehler 2. Art}) = P(H_0 \text{ nicht ablehnen} | H_1 \text{ ist wahr}) =: \tilde{\beta}.$$

Beide Fehlerarten sind bei festem Stichprobenumfang gegenläufig, so dass sie nicht gleichzeitig minimiert werden können. Es kann lediglich entweder  $\alpha^*$  oder  $\tilde{\beta}$  fest gewählt und basierend auf dieser Größe die andere Wahrscheinlichkeit berechnet werden. In der Statistik ist es im Allgemeinen üblich, die Wahrscheinlichkeit für den Fehler 1. Art ( $\alpha^*$ -Fehler) festzusetzen und die Wahrscheinlichkeit für den Fehler 2. Art ( $\tilde{\beta}$ -Fehler) durch eine Vergrößerung des Stichprobenumfangs zu verkleinern. Daher werden die Hypothesen so formuliert, dass der Fehler 1. Art die schwerwiegendere Fehlentscheidung darstellt. In einer klinischen Studie werden die Hypothesen daher üblicherweise wie folgt gewählt:

- $H_0$ : Das Medikament ist nicht wirksam.
- $H_1$ : Das Medikament ist wirksam.

Mit dieser Wahl wird sicher gestellt, dass mit einer geringen Wahrscheinlichkeit wirkungslose Medikamente auf den Markt kommen. Für die Wahrscheinlichkeit  $\alpha^*$  wird gewöhnlich der Wert 0,05 angesetzt.

Obwohl die Wahrscheinlichkeit für den Fehler 1. Art fest gewählt wird, sollte ein Test zum Niveau  $\alpha^*$  so ausgesucht werden, dass die Wahrscheinlichkeit für den Fehler 2. Art möglichst gering ist. Diese Wahrscheinlichkeit des Fehlers 2. Art kann bei Hypothesen der Form  $H_0 : \theta = \theta_0$  versus  $H_1 : \theta \neq \theta_0$  mit Hilfe der sogenannten Gütefunktion  $G(\theta)$  bestimmt werden, denn für  $\theta \in \Theta_1 = \{\theta \in \Theta | \theta \neq \theta_0\}$  gilt  $1 - G(\theta) \geq \tilde{\beta}$ . Ein Gütekriterium statistischer Tests ist dessen geringe Wahrscheinlichkeit für den Fehler 2. Art. Dies lässt sich nach Rüger (1985) folgendermaßen erreichen:<sup>1</sup>

- Bei fester Anzahl von Subjekten  $m$  gilt: Je kleiner das vorgegebene  $\alpha^*$  ist, desto flacher verläuft die Gütefunktion, d.h. um so geringer ist die Wahrscheinlichkeit, bei einem  $\theta \in \Theta_1$  zur Ablehnung von  $H_0$  zu gelangen.
- Bei festem  $\alpha^*$  gilt: Je größer der Stichprobenumfang  $m$  ist, desto steiler verläuft die Gütefunktion, d.h. um so eher kommt der Test bei einem  $\theta \in \Theta_1$  zur Ablehnung von  $H_0$ . Der Test wird *trennschärfer* (in dem Sinne, dass er zwischen richtigen und falschen Hypothesen besser trennt) genannt.

Daher ist eine ausreichend große Anzahl von Subjekten  $m$  notwendig, um einen trennscharfen und damit genaueren Test zu erhalten. Diese Güte des Test wird in der englischen Literatur auch *statistical Power* genannt.

---

<sup>1</sup>Vgl. Rüger (1985), S. 250.

---

Die Berechnung der benötigten Anzahl von Subjekten zur Durchführung einer Studie ist eine wichtige Komponente in der Versuchsplanung: Einerseits reduziert eine zu geringe Subjektanzahl  $m$  die Güte des Tests; signifikante Effekte können nur unzureichend erkannt werden. Andererseits führt eine zu große Subjektanzahl zu einer unnötigen Erhöhung der Studienkosten. Des Weiteren kann ein zu großer Stichprobenumfang zu einer unnötigen Verzögerung der Studiendauer führen. Dies ist beispielsweise im Rahmen von AIDS-Studien ethisch nicht vertretbar.

Für poissonverteilte Daten, die für jedes Subjekt lediglich einmal erhoben werden, gibt es in der wissenschaftlichen Literatur einige Prozeduren und Möglichkeiten zur Berechnung des benötigten Stichprobenumfangs.<sup>2</sup> Für wiederholte poissonverteilte Daten – im Gegensatz zu binären oder normalverteilten<sup>3</sup> wiederholten Daten –en<sup>4</sup> wiederholten Daten – sind entsprechende Berechnungen nur unzureichend ausgearbeitet. Für marginale Modelle ist in Liu und Liang (1997), basierend auf der Arbeit von Self u. a. (1988), eine Herleitung für eine Formel zur Berechnung der benötigten Anzahl  $m$  angegeben. Diese Methode soll in diesem Kapitel vorgestellt werden. Bedauerlicherweise ist es der Autorin nicht gelungen, analoge Herleitungen für die subjekt-spezifischen Modelle zu finden.

Bevor die Herleitung von Liu und Liang (1997) behandelt wird, müssen die Einflussgrößen zur Berechnung des Stichprobenumfangs  $m$  charakterisiert werden. Nach Diggle u. a. (1996) sind diese in jeder Studie mit wiederholten Daten vor Beginn der Studie festzusetzen und durch die folgenden Größen gegeben:<sup>5</sup>

- (i) Der  $\alpha^*$ -Fehler, der im Allgemeinen auf den Wert 0.05 festgesetzt wird.
  
- (ii) Die *kleinste bedeutungsvolle Differenz*  $d$  soll so gewählt werden, dass die Nullhypothese signifikant abgelehnt wird, wenn sich der betrachtete Parameter von dem Parameterwert unter der Nullhypothese um mindestens  $d$  unterscheidet. Aus diesem Grund lassen sich die Hypothesen  $H_0 : \theta = \theta_0$  gegen  $H_1 : \theta \neq \theta_0$  mit  $d = |\theta_0 - \theta_1|$  und  $\theta_1 \in \Theta_1$  auch umschreiben in  $H_0 : \theta = \theta_0$  gegen  $H_1 : \theta = \theta_1$ . Der Parameter  $d$  muss in diesem Zusammenhang anhand von früheren Versuchen oder ähnlichen Informationen so gewählt werden, dass es zu einer hohen Aussagekraft des Tests kommt.

---

<sup>2</sup>Siehe dazu Cook (1995) und Signorini (1991).

<sup>3</sup>Vgl. Bloch (1986) und Lui und Cumberland (1992).

<sup>4</sup>Vgl. Bloch (1986) und Lui und Cumberland (1992).

<sup>5</sup>Vgl. Diggle u. a. (1996), S. 26ff.

(iii) Die *Power* (Güte) eines statistischen Tests ist durch die Wahrscheinlichkeit

$$P(H_0 \text{ ablehnen} | H_1 \text{ ist wahr}) = 1 - \tilde{\beta}$$

gegeben. Eine typische Anforderung bei statistischen Tests ist eine Power von 0.9 für alle  $\theta \in \Theta$ , deren Differenz sich von  $\theta_0$  um mindestens  $d$  unterscheidet.

(iv) Die Anzahl der Beobachtungen  $n_i$  pro Subjekt  $i$  muss festgesetzt werden.<sup>6</sup> Zur Vereinfachung werden die Analysen auf den Fall  $n_i = n$  für alle  $i \in \{1, \dots, m\}$  reduziert.

(v) Die Korrelationsstruktur und die einzelnen Korrelationen der wiederholten Beobachtungen müssen bekannt sein. Da dies im Allgemeinen nicht der Fall ist, können Erfahrungswerte aus vorherigen Studien oder Simulationen genutzt werden.

Die Methode von Liu und Liang (1997) zur Berechnung des Stichprobenumfangs basiert auf den marginalen Modellen und den Verallgemeinerten Schätzgleichungen (GEE), wie sie im Kapitel 3 vorgestellt wurden.

Sei für alle  $i \in \{1, \dots, m\}$  der Beobachtungsvektor  $(y_{i1}, \dots, y_{in})^\top$  eine Realisierung des  $n$ -dimensionalen Zufallsvektors  $Y_i$ . Die marginale Erwartung  $\mu_{ij}$  von  $Y_{ij}$  und die erklärenden Variablen  $x_{ij}$  hängen für alle  $i \in \{1, \dots, m\}$  und  $j \in \{1, \dots, n\}$  durch ein verallgemeinertes lineares Modell

$$\begin{aligned} h(\mu_{ij}) &= x_{ij}^\top \beta \\ &= w_{ij}^\top \psi + z_{ij}^\top \tau \end{aligned} \tag{6.1}$$

zusammen. Dabei sind  $h$  eine bekannte Linkfunktion,  $\psi$  der interessierende  $q_1$ -dimensionale Parametervektor,  $\tau$  der nicht interessierende  $q_2$ -dimensionale Parametervektor und  $w_{ij}, z_{ij}$  die  $q_1$ - bzw.  $q_2$ -dimensionalen, erklärenden Variablenvektoren. Insbesondere gilt  $q = q_1 + q_2$ . Bei dem Modell

$$\ln(\mu_{ij}) = \begin{cases} \beta_0 + \beta_1, & \text{falls Placebo,} \\ \beta_0, & \text{falls Verum} \end{cases}$$

ist der Behandlungsunterschied durch  $\exp(\beta_1)$  gegeben, d.h. bei der Überprüfung der Wirksamkeit des Medikaments interessiert lediglich der Parameter  $\beta_1$ . Demzufolge ist in diesem Fall  $\psi = \beta_1, \tau = \beta_0, z_{ij} \equiv 1$  und  $x_{ij} = 1$ , falls das  $i$ -te Subjekt ein

---

<sup>6</sup>Alternativ könnte die Anzahl der Beobachtungen  $n$  pro Person auch in Abhängigkeit eines festen Stichprobenumfangs optimiert werden.

Placebo erhält, und  $x_{ij} = 0$ , falls das  $i$ -te Subjekt das Verum erhält.

Die zu untersuchenden Hypothesen sind nun

$$H_0 : \psi = \psi_0 \quad \text{gegen} \quad H_1 : \psi \neq \psi_0,$$

so dass im obigen Beispiel  $\psi_0 = 0$  zu wählen ist. Wenn die *kleinste bedeutungsvolle Differenz*  $d$  durch  $d = |\psi_0 - \psi_1|$  für ein  $\psi_1 \neq \psi_0$  gegeben ist, können diese Hypothesen ohne Beschränkung der Allgemeinheit umformuliert werden zu

$$H_0 : \psi = \psi_0 \quad \text{gegen} \quad H_1 : \psi = \psi_1.$$

Nach Liu und Liang (1997) kann dieses Testproblem mit Hilfe der Teststatistik

$$T := S_\psi(\psi_0, \hat{\tau}_0, \hat{\alpha}, \hat{\phi}) \Sigma_0^{-1} S_\psi(\psi_0, \hat{\lambda}_0, \hat{\alpha}, \hat{\phi}) \quad (6.2)$$

entschieden werden. Dabei ist mit den Bezeichnungen aus Kapitel 3

- $\hat{\tau}_0$  ein Schätzer für  $\tau$ , der unter der Nullhypothese (d.h.  $\psi = \psi_0$  und  $\mu_{ij} = h^{-1}(w_{ij}^\top \psi_0 + z_{ij}^\top \tau)$ ) als Lösung der Gleichung

$$S_\tau(\psi_0, \tau, \alpha, \phi) = \sum_{i=1}^m \left( \frac{\partial \mu_i}{\partial \tau} \right)^\top (V_i^\phi)^{-1} (Y_i - \mu_i) = 0$$

ermittelt wird;

- $S_\psi(\psi_0, \hat{\tau}_0, \hat{\alpha}, \hat{\phi}) = \sum_{i=1}^m \left( \frac{\partial \mu_i}{\partial \psi} \right)^\top (V_i^\phi)^{-1} (Y_i - \mu_i)$  mit einem Schätzer  $\hat{\alpha}$  für  $\alpha$  und einem Schätzer  $\hat{\phi}$  für den Überdispersions-Parameter  $\phi$ ;<sup>7</sup>
- $\Sigma_0 = \text{Cov}_{H_0} [S_\psi(\psi_0, \hat{\tau}_0, \alpha, \phi)]$  und
- $V_i^\phi = (A_i^\phi)^{1/2} R(\alpha) (A_i^\phi)^{1/2}$  mit  $A_i^\phi = \phi \text{diag}(\mu_{i1}, \dots, \mu_{in_i})$ .

Für diese Teststatistik  $T$  lässt sich nach Liu und Liang (1997) folgender Satz formulieren. Dieser wird, wie die weiteren Sätze in diesem Abschnitt, ohne Beweis bleiben.

---

<sup>7</sup>Da die Herleitung des benötigten Stichprobenumfangs  $m$  von Liu und Liang (1997) auf den GEE basiert, ist die Existenz eines konsistenten Schätzers für  $\psi$  unabhängig von der Abhängigkeitsstruktur der Daten. Bei gewählter Abhängigkeitsstruktur kann dann anhand der Tabelle 3.1 ein Schätzer für  $\alpha$  bestimmt werden. Entsprechendes gilt für den Dispersions-Parameter  $\phi$ . Daher werden, analog zum Vorgehen von Liu und Liang (1997), im Folgenden  $\alpha$  und  $\phi$  als bekannt vorausgesetzt.

**Satz 6.1.**

Für  $m \rightarrow \infty$  konvergiert  $T$  unter der Nullhypothese  $H_0$  gegen eine zentrale  $\chi_{q_1}^2$ -Verteilung. Unter  $H_1$ , d.h. mit  $\psi = \psi_1$  und  $\tau = \tau_1$  folgt  $T$  asymptotisch einer nichtzentralen  $\chi_{q_1}^2$ -Verteilung mit Nichtzentralitätsparameter

$$\nu = \xi^\top \Sigma_1^{-1} \xi,$$

wobei  $\xi = E_{H_1}(S_\psi(\psi_0, \hat{\tau}_0, \alpha, \phi))$  und  $\Sigma_1 = \text{Cov}_{H_1}(S_\psi(\psi_0, \hat{\tau}_0, \alpha, \phi))$  ist.

Für einen Test zum Niveau  $\alpha^*$  und einer Power von  $1 - \tilde{\beta}$  gilt  $\nu = \nu(\alpha^*, \tilde{\beta})$ . Mit vorgegebenen Werten für  $\alpha^*$ ,  $\tilde{\beta}$ ,  $d$ , des Nichtzentralitätsparameters  $\nu$  sowie der Größen  $\xi$  und  $\Sigma_1$  lässt sich nun der Stichprobenumfang  $m$  berechnen. Bevor dieser Zusammenhang durch eine Gleichung geliefert wird, werden Approximationen für die Größen  $\xi$  und  $\Sigma_1$  vorgestellt:

**Satz 6.2.**

Die erklärenden Variablen des  $i$ -ten Subjekts  $\{(w_{ij}, z_{ij}), j \in \{1, \dots, n\}\}$  seien für jedes  $i \in \{1, \dots, m\}$  diskret verteilt mit

$$P(w_{ij\ell}, z_{ij\ell}; j \in \{1, \dots, n\}) = \pi_\ell,$$

wobei  $\ell \in \{1, \dots, L\}$  und  $\{(w_{ij\ell}, z_{ij\ell}), j \in \{1, \dots, n\}\}, \ell = \{1, \dots, L\}$  die  $L$  verschiedenen Werte sind, die  $\{(w_{ij}, z_{ij}), j \in \{1, \dots, n\}\}$  annehmen kann. Die Größen  $\xi$  und  $\Sigma_i$  können dann approximiert werden durch

$$\xi \approx m \tilde{\xi} = m \sum_{\ell=1}^L \pi_\ell P_\ell^*(V_\ell^{\phi,*})^{-1} (\mu_\ell^1 - \mu_\ell^*) \quad (6.3)$$

und

$$\Sigma_1 \approx m \tilde{\Sigma}_1 = m \sum_{\ell=1}^L \pi_\ell P_\ell^*(V_\ell^{\phi,*})^{-1} \text{Cov}_{H_1}(Y_\ell) (V_\ell^{\phi,*})^{-1} P_\ell^{*\top} \quad (6.4)$$

mit

- $\mu_\ell^1 = h^{-1}(w_{ij\ell}^\top \psi_1 + z_{ij\ell}^\top \tau_1)$  und  $\mu_\ell^* = h^{-1}(w_{ij\ell}^\top \psi_0 + z_{ij\ell}^\top \tau_0^*)$ ;
- $\tau_0^*$  als Lösung der Gleichung

$$\sum_{\ell=1}^L \pi_\ell \left( \frac{\partial \mu_\ell^*}{\partial \tau} \right)^\top (V_\ell^{\phi,*})^{-1} (\mu_\ell^1 - \mu_\ell^*) = 0; \quad (6.5)$$

- $P_\ell^* = P_\ell(\psi_0, \tau_0^*) = \left( \frac{\partial \mu_\ell^*}{\partial \psi} \right)^\top - I_{\psi\tau\phi}^* (I_{\tau\tau\phi}^*)^{-1} \left( \frac{\partial \mu_\ell^*}{\partial \tau} \right)$ ;

- 
- $I_{\psi\tau\phi}^* = I_{\psi\tau\phi}^*(\psi_0, \tau_0^*) = \sum_{\ell=1}^L \pi_\ell \left( \frac{\partial \mu_\ell^*}{\partial \psi} \right)^\top (V_\ell^{\phi,*})^{-1} \left( \frac{\partial \mu_\ell^*}{\partial \tau} \right);$
  - $I_{\tau\tau\phi}^* = I_{\tau\tau\phi}^*(\psi_0, \tau_0^*) = \sum_{\ell=1}^L \pi_\ell \left( \frac{\partial \mu_\ell^*}{\partial \tau} \right)^\top (V_\ell^{\phi,*})^{-1} \left( \frac{\partial \mu_\ell^*}{\partial \tau} \right).$

Für die Berechnung des Stichprobenumfangs gilt dann:

**Satz 6.3.**

Mit  $\tilde{\xi}$  und  $\tilde{\Sigma}_1$  aus Satz 6.2 und  $\nu$  aus Satz 6.1 ergibt sich für einen Test zum Niveau  $\alpha^*$ , einer Power von  $1 - \tilde{\beta}$  und einer kleinsten bedeutungsvollen Differenz  $d$  ein benötigter Stichprobenumfang von

$$m = \frac{\nu}{\tilde{\xi}^\top \tilde{\Sigma}_1^{-1} \tilde{\xi}}.$$

Zusammenfassend lässt sich somit basierend auf der Teststatistik  $T$  die Berechnung des benötigten Stichprobenumfangs bei wiederholten Daten durch folgende Schritte durchführen:

- (1) Aufstellen eines Regressionsmodells und Formulierung der Hypothesen  $H_0$  und  $H_1$ ;
- (2) Festsetzen des Signifikanzniveaus  $\alpha^*$ , der Power  $1 - \tilde{\beta}$  und der kleinsten bedeutungsvollen Differenz  $d$ ;
- (3) Auswählen einer Korrelationsstruktur, die zum Design der Studie passt (vgl. Abschnitt 3.1.1.3);
- (4) Schätzen der Parameter  $\alpha$  und  $\phi$  gemäß der Tabelle 3.1 bzw. Definition 3.1;
- (5) Spezifizieren der diskreten Verteilung der erklärenden Variablen  $\{(w_{ij}, z_{ij}), j \in \{1, \dots, n\}\}, i \in \{1, \dots, m\}$ ;
- (6) Berechnung von  $\tau_0^*$  mit Hilfe von Gleichung (6.5);
- (7) Berechnung aller Größen im Satz 6.2;
- (8) Berechnung des benötigten Stichprobenumfangs mittels Satz 6.3.

## 6.1 Der Stichprobenumfang im Fall von zwei Stichproben

Der folgende Satz liefert Aufschluss über den benötigten Stichprobenumfang für das Modell

$$h(\mu_{ij}) = \begin{cases} \beta_0 + \beta_1, & \text{falls Placebo,} \\ \beta_0, & \text{falls Verum.} \end{cases}$$

### Satz 6.4.

Sei ein Zweistichprobenproblem mit poissonverteilten Beobachtungen mit Überdispersion, den erklärenden Variablen

$$x_{ij} = x_i = \begin{cases} 0, & \text{falls Placebo,} \\ 1, & \text{falls Verum} \end{cases}$$

für alle  $j \in \{1, \dots, n\}$  und der Korrelationsmatrix  $R(\alpha)$  gegeben. Ferner wird angenommen, dass die Verum- und die Placebo-Gruppe gleich viele Subjekte enthalten. Basierend auf der Methode von Liang und Liu ergibt sich als minimal benötigter Stichprobenumfang  $m$  für einen Test zum Niveau  $\alpha^*$ , einer Power von  $1 - \tilde{\beta}$  und einer kleinsten bedeutungsvollen Differenz  $d$  der Wert

$$m = \frac{2(z_{1-\alpha^*/2} + z_{1-\tilde{\beta}})^2 (\exp(d) + \exp(\beta_0 + d))}{\phi^{-1}(\exp(\beta_0 + d) - \exp(\beta_0))^2 \mathbf{1}_n^\top R(\alpha)^{-1} \mathbf{1}_n}.$$

Dabei ist  $z_x$  das  $x$ -Quantil der Standardnormalverteilung und  $\phi$  der Überdispersionsparameter.

*Beweis:* Der folgende Beweis bedient sich der Notation des letzten Abschnitts. Zur Herleitung des behaupteten Stichprobenumfangs werden die Schritte auf Seite 119 umgesetzt.

Im Fall von poissonverteilten Daten mit Überdispersion gilt

$$E(Y_{ij}) = \lambda_{ij} \quad \text{und} \quad \text{Var}(Y_{ij}) = \phi \lambda_{ij}$$

für alle  $i \in \{1, \dots, m\}$  und  $j \in \{1, \dots, n\}$ . Das verallgemeinerte lineare Modell im betrachteten Zweistichprobenproblem ist nach den Voraussetzungen gegeben durch

$$\ln(\lambda_{ij}) = \ln(\lambda_i) = \begin{cases} \beta_0 + \beta_1, & \text{falls Placebo,} \\ \beta_0, & \text{falls Verum.} \end{cases}$$

Die Behandlungen unterscheiden sich somit durch den Faktor  $\exp(\beta_1)$ . Dieses Modell lässt sich mit den Bezeichnungen aus dem letzten Abschnitt auch schreiben als

$$\ln(\lambda_{ij}) = \ln(\lambda_i) = \begin{cases} z_i\tau + w_i\psi, & \text{falls Placebo,} \\ z_i\tau, & \text{falls Verum,} \end{cases}$$

mit  $z_i \equiv 1$ ,  $\tau = \beta_0$ ,  $w_i = x_i$  und  $\psi = \beta_1$ .

Die Hypothesen, die im Zusammenhang mit klinischen Studien von Interesse sind, können dann folgendermaßen formuliert werden:

$$H_0 : \exp(\tau) = \exp(\tau + \psi) \quad \text{gegen} \quad H_1 : \exp(\tau) \neq \exp(\tau + \psi),$$

d.h. insbesondere  $H_0 : \psi = 0$  gegen  $H_1 : \psi \neq 0$ .

Diese Hypothesen sollen mit Hilfe der Teststatistik (6.2) zum Signifikanzniveau  $\alpha^*$ , mit der Power  $1 - \tilde{\beta}$  und mit der *kleinsten bedeutungsvollen Differenz*  $d$  getestet werden. Basierend auf  $d$  lassen sich diese Hypothesen umformulieren zu

$$H_0 : \psi = \psi_0 = 0 \quad \text{gegen} \quad H_1 : \psi = \psi_1 = d.$$

Die behauptete Fallzahl  $m$  kann jetzt mit Hilfe von Satz 6.3 bestimmt werden. Die Anwendung dieses Satzes erfordert die Berechnung von  $\nu$ ,  $\tilde{\xi}$  und  $\tilde{\Sigma}_1$  und damit insbesondere die Berechnung der Terme in Satz 6.2.

Aufgrund der gleich großen Behandlungsgruppen wird angenommen, dass für alle  $i \in \{1, \dots, m\}$  und  $j \in \{1, \dots, n\}$  die diskreten erklärenden Variablen wie folgt verteilt sind:

$$P(W_{ij} = 0, Z_{ij} = 1) =: \pi_1 = 1/2 \quad \text{und} \quad P(W_{ij} = 1, Z_{ij} = 1) =: \pi_2 = 1/2,$$

wodurch  $L=2$  folgt.<sup>8</sup>

Zudem besitze jedes Subjekt  $i$ ,  $i \in \{1, \dots, m\}$  die gleiche Korrelationsstruktur  $R(\alpha)$ , so dass die Kovarianzmatrix für  $\ell \in \{1, 2\}$  gegeben ist durch

$$V_\ell^\phi = (A_\ell^\phi)^{1/2} R(\alpha) (A_\ell^\phi)^{1/2}$$

mit  $A_\ell^\phi = \phi \text{diag}(\lambda_\ell, \dots, \lambda_\ell)$ ,  $\lambda_1 := \exp(\tau)$  und  $\lambda_2 := \exp(\psi + \tau)$ .

Seien unter  $H_1$ , d.h. mit  $\psi = \psi_1$  und  $\tau = \tau_1$ ,

$$\lambda_1^1 := \exp(\tau_1) \mathbf{1}_n \quad \text{und} \quad \lambda_2^1 := \exp(\tau_1 + \psi_1) \mathbf{1}_n$$

und unter  $H_0$  gilt approximativ

$$\lambda_1^* := \exp(\tau_0^*) \mathbf{1}_n \quad \text{und} \quad \lambda_2^* := \exp(\tau_0^* + \psi_0) \mathbf{1}_n = \exp(\tau_0^*) \mathbf{1}_n,$$

---

<sup>8</sup>Dieses Vorgehen entspricht der Randomisierung beider Behandlungsgruppen.

wobei  $\tau_0^*$  nach Satz 6.2 eine Lösung der Gleichung

$$\sum_{\ell=1}^L \pi_{\ell} \left( \frac{\partial \lambda_{\ell}^*}{\partial \tau} \right)^{\top} (V_{\ell}^{\phi,*})^{-1} (\lambda_{\ell}^1 - \lambda_{\ell}^*) = 0 \quad (6.6)$$

ist. Hierbei ist  $V_{\ell}^{\phi,*} = (A_{\ell}^{\phi,*})^{1/2} R(\alpha) (A_{\ell}^{\phi,*})^{1/2}$ ,  $A_{\ell}^{\phi,*} = \phi \text{diag}(\lambda_{\ell}^*, \dots, \lambda_{\ell}^*)$  und  $\lambda_{\ell}^* = \exp(\tau_0^*)$  mit  $\ell \in \{1, 2\}$ , so dass

$$V_1^{\phi,*} = V_2^{\phi,*} = \phi \exp(\tau_0^*) R(\alpha)$$

gilt. Nun wird die Gleichung (6.6) gelöst:

$$\begin{aligned} & \sum_{\ell=1}^2 \pi_{\ell} \left( \frac{\partial \lambda_{\ell}^*}{\partial \tau} \right)^{\top} (V_{\ell}^{\phi,*})^{-1} (\lambda_{\ell}^1 - \lambda_{\ell}^*) = 0 \\ \Leftrightarrow^{\psi_0=0} & \pi_1 \exp(\tau_0^*) \mathbf{1}_n^{\top} \phi^{-1} \exp(-\tau_0^*) R(\alpha)^{-1} (\exp(\tau_1) - \exp(\tau_0^*)) \mathbf{1}_n \\ & + \pi_2 \exp(\tau_0^*) \mathbf{1}_n^{\top} \phi^{-1} \exp(-\tau_0^*) R(\alpha)^{-1} (\exp(\tau_1 + \psi_1) - \exp(\tau_0^*)) \mathbf{1}_n = 0 \\ \Leftrightarrow & [\pi_1 (\exp(\tau_1) - \exp(\tau_0^*)) + \pi_2 (\exp(\tau_1 + \psi_1) - \exp(\tau_0^*))] \mathbf{1}_n^{\top} R(\alpha)^{-1} \mathbf{1}_n = 0 \\ \Leftrightarrow & \pi_1 (\exp(\tau_1) - \exp(\tau_0^*)) + \pi_2 (\exp(\tau_1 + \psi_1) - \exp(\tau_0^*)) = 0 \\ \Leftrightarrow & \pi_1 \exp(\tau_1) + \pi_2 \exp(\tau_1 + \psi_1) - \underbrace{(\pi_1 + \pi_2)}_{=1} \exp(\tau_0^*) = 0 \end{aligned}$$

und damit

$$\tau_0^* = \ln(\pi_1 \exp(\tau_1) + \pi_2 \exp(\tau_1 + \psi_1)).$$

Basierend auf der Größe  $\tau_0^*$  können jetzt die benötigten Größen  $\nu$ ,  $\tilde{\xi}$  und  $\tilde{\Sigma}_1$  berechnet werden:

(1) Berechnung des Nichtzentralitätsparameters  $\nu$ :

Nach Satz 6.1 konvergiert die von Liu und Liang betrachtete Teststatistik  $T$  unter  $H_0$  gegen eine zentrale  $\chi^2$ -Verteilung mit  $\dim \psi = 1$  Freiheitsgraden, d.h.  $T \approx X^2$  mit einer standardnormalverteilten Zufallsvariablen  $X$ .

Unter der Alternative  $H_1$  ist die Statistik  $T$  nach Satz 6.1  $\chi^2$ -verteilt mit dem Nichtzentralitätsparameter  $\nu$ , so dass  $T \approx Z^2$  mit einer Zufallsvariablen  $Z \sim \mathcal{N}(\mu, 1)$  gilt. Der Erwartungswert von  $T$  kann mit Hilfe der momenterzeugenden Funktion  $M_Z(t)$  von  $Z$  berechnet werden; dabei ist  $M_Z(t) = \exp(\mu t + t^2/2)$ . Somit folgt

$$\begin{aligned} E(T) & \approx E(Z^2) \\ & = \left. \frac{\partial^2 M_Z(t)}{\partial t \partial t} \right|_{t=0} \\ & = \left. \frac{\partial}{\partial t} \left( (\mu + t) \exp(\mu t + \frac{t^2}{2}) \right) \right|_{t=0} \end{aligned}$$

$$\begin{aligned}
 &= \left( \exp\left(\mu t + \frac{t^2}{2}\right) + (\mu + t)^2 \exp\left(\mu t + \frac{t^2}{2}\right) \right) \Big|_{t=0} \\
 &= 1 + \mu^2
 \end{aligned}$$

und  $\mu^2$  ist der Nichtzentralitätsparameter  $\nu$ , der ermittelt werden muss. Insbesondere ist aufgrund der Konstruktion  $\sqrt{T} - \mu$  asymptotisch standardnormalverteilt.

Da  $T$  die Teststatistik eines Tests zum Niveau  $\alpha^*$  und der Power  $1 - \tilde{\beta}$  bei einer *kleinsten bedeutungsvollen Differenz* von  $d$  ist, gilt mit der Verteilungsfunktion  $\Phi(x)$  der Standardnormalverteilung folgendes

$$\begin{aligned}
 &1 - \tilde{\beta} &= P(\sqrt{T} > z_{1-\alpha^*/2}) \\
 \Leftrightarrow &1 - \tilde{\beta} &= P(\sqrt{T} - \mu > z_{1-\alpha^*/2} - \mu) \\
 \Leftrightarrow &1 - \tilde{\beta} &= 1 - \Phi(z_{1-\alpha^*/2} - \mu) \\
 \Leftrightarrow &1 - \tilde{\beta} &= \Phi(-z_{1-\alpha^*/2} + \mu) \\
 \Leftrightarrow &\Phi(z_{1-\tilde{\beta}}) &= \Phi(-z_{1-\alpha^*/2} + \mu) \\
 \Leftrightarrow &z_{1-\tilde{\beta}} &= -z_{1-\alpha^*/2} + \mu \\
 \Leftrightarrow &z_{1-\tilde{\beta}} + z_{1-\alpha^*/2} &= \mu \\
 \Rightarrow &(z_{1-\tilde{\beta}} + z_{1-\alpha^*/2})^2 &= \mu^2 = \nu.
 \end{aligned} \tag{6.7}$$

(2) Berechnung der Größe  $\tilde{\xi}$  mit Hilfe von Gleichung (6.3) und Satz 6.2:

Es gilt

$$\begin{aligned}
 \tilde{\xi} &= \sum_{\ell=1}^L \pi_{\ell} P_{\ell}^* (V_{\ell}^{\phi,*})^{-1} (\mu_{\ell}^1 - \mu_{\ell}^*) \\
 &\stackrel{\text{Poisson}}{=} \sum_{\ell=1}^2 \pi_{\ell} P_{\ell}^* (V_{\ell}^{\phi,*})^{-1} (\lambda_{\ell}^1 - \lambda_{\ell}^*).
 \end{aligned}$$

Für die Berechnung von  $\tilde{\xi}$  muss also zunächst  $P_{\ell}^* = P_{\ell}(\psi_0, \tau_0^*)$  bestimmt werden. Es ist

$$P_{\ell}^* = \left( \frac{\partial \lambda_{\ell}^*}{\partial \psi} \right)^{\top} - I_{\psi\tau\phi}^* (I_{\tau\tau\phi}^*)^{-1} \left( \frac{\partial \lambda_{\ell}^*}{\partial \tau} \right)$$

mit  $\psi_0 = 0$ ,

$$\begin{aligned}
 I_{\psi\tau\phi}^* = I_{\psi\tau\phi}^*(\psi_0, \tau_0^*) &= \sum_{\ell=1}^2 \pi_{\ell} \left( \frac{\partial \lambda_{\ell}^*}{\partial \psi} \right)^{\top} (V_{\ell}^{\phi,*})^{-1} \left( \frac{\partial \lambda_{\ell}^*}{\partial \tau} \right) \\
 &= 0_n + \pi_2 \exp(\tau_0^*) \mathbf{1}_n^{\top} \phi^{-1} \exp(-\tau_0^*) R(\alpha)^{-1} \exp(\tau_0^*) \mathbf{1}_n \\
 &= \pi_2 \phi^{-1} \exp(\tau_0^*) \mathbf{1}_n^{\top} R(\alpha)^{-1} \mathbf{1}_n
 \end{aligned}$$

und

$$\begin{aligned} I_{\tau\tau\phi}^* &= I_{\tau\tau\phi}^*(\psi_0, \tau_0^*) = \sum_{\ell=1}^2 \pi_\ell \left( \frac{\partial \lambda_\ell^*}{\partial \tau} \right)^\top (V_\ell^{\phi,*})^{-1} \left( \frac{\partial \lambda_\ell^*}{\partial \tau} \right) \\ &= (\pi_1 + \pi_2) \phi^{-1} \exp(\tau_0^*) \mathbf{1}_n^\top R(\alpha)^{-1} \mathbf{1}_n. \end{aligned}$$

Damit ist

$$\begin{aligned} P_1^* &= \mathbf{0}_n^\top - \frac{\pi_2 \phi^{-1} \exp(\tau_0^*) \mathbf{1}_n^\top R(\alpha)^{-1} \mathbf{1}_n}{(\pi_1 + \pi_2) \phi^{-1} \exp(\tau_0^*) \mathbf{1}_n^\top R(\alpha)^{-1} \mathbf{1}_n} \exp(\tau_0^*) \mathbf{1}_n^\top \\ &= -\pi_2 \exp(\tau_0^*) \mathbf{1}_n^\top, \end{aligned}$$

sowie

$$\begin{aligned} P_2^* &= \exp(\tau_0^*) \mathbf{1}_n^\top - \pi_2 \exp(\tau_0^*) \mathbf{1}_n^\top \\ &= (1 - \pi_2) \exp(\tau_0^*) \mathbf{1}_n^\top \\ &= \pi_1 \exp(\tau_0^*) \mathbf{1}_n^\top. \end{aligned}$$

Schließlich ergibt sich

$$\begin{aligned} \tilde{\xi} &= \sum_{\ell=1}^2 \pi_\ell P_\ell^* (V_\ell^{\phi,*})^{-1} (\lambda_\ell^1 - \lambda_\ell^*) \\ &= \pi_1 P_1^* \phi^{-1} \exp(-\tau_0^*) R(\alpha)^{-1} [\exp(\tau_1) - \exp(\tau_0^*)] \mathbf{1}_n \\ &\quad + \pi_2 P_2^* \phi^{-1} \exp(-\tau_0^*) R(\alpha)^{-1} [\exp(\tau_1 + \psi_1) - \exp(\tau_0^*)] \mathbf{1}_n \\ &= \pi_1 \pi_2 \phi^{-1} [\exp(\tau_1 + \psi_1) - \exp(\tau_0^*) - \exp(\tau_1) + \exp(\tau_0^*)] \mathbf{1}_n^\top R(\alpha)^{-1} \mathbf{1}_n \\ &= \pi_1 \pi_2 \phi^{-1} [\exp(\tau_1 + \psi_1) - \exp(\tau_1)] \mathbf{1}_n^\top R(\alpha)^{-1} \mathbf{1}_n. \end{aligned} \tag{6.8}$$

• Berechnung von  $\tilde{\Sigma}_1$  mit Hilfe von Gleichung (6.4) und Satz 6.2:

Es gilt

$$\tilde{\Sigma}_1 = \sum_{\ell=1}^2 \pi_\ell P_\ell^* (V_\ell^{\phi,*})^{-1} \text{Cov}_{H_1}(Y_\ell) (V_\ell^{\phi,*})^{-1} P_\ell^{*\top}.$$

Zudem ist  $\text{Cov}_{H_1}(Y_1) = \phi \exp(\tau_1) R(\alpha)$  und  $\text{Cov}_{H_1}(Y_2) = \phi \exp(\tau_1 + \psi_1) R(\alpha)$ , so dass sich durch Einsetzen und Zusammenfassen

$$\begin{aligned} \tilde{\Sigma}_1 &= \pi_1 \pi_2^2 \phi^{-1} \exp(\tau_1) \mathbf{1}_n^\top R(\alpha)^{-1} \mathbf{1}_n + \pi_1^2 \pi_2 \phi^{-1} \exp(\tau_1 + \psi_1) \mathbf{1}_n^\top R(\alpha)^{-1} \mathbf{1}_n \\ &= \pi_1 \pi_2 \phi^{-1} \mathbf{1}_n^\top R(\alpha)^{-1} \mathbf{1}_n [\pi_1 \exp(\tau_1 + \psi_1) + \pi_2 \exp(\tau_1)] \end{aligned} \tag{6.9}$$

ergibt. Für den benötigten Stichprobenumfang  $m$  folgt dann mit Satz 6.3, den Glei-

chungen (6.7), (6.8), (6.9) und mit  $\psi_1 = d$ ,  $\pi_1 = \pi_2 = 1/2$ ,  $\tau_1 = \beta_0$ :

$$\begin{aligned}
 m &= \frac{\nu}{\tilde{\xi}^\top \tilde{\Sigma}_1^{-1} \tilde{\xi}} \\
 &= \frac{(z_{1-\alpha^*/2} + z_{1-\tilde{\beta}})^2 \pi_1 \pi_2 \phi^{-1} \mathbf{1}_n^\top R(\alpha)^{-1} \mathbf{1}_n [\pi_1 \exp(\psi_1 + \tau_1) + \pi_2 \exp(\tau_1)]}{(\pi_1 \pi_2 \phi^{-1} [\exp(\psi_1 + \tau_1) - \exp(\tau_1)] \mathbf{1}_n^\top R(\alpha)^{-1} \mathbf{1}_n)^2} \\
 &= \frac{(z_{1-\alpha^*/2} + z_{1-\tilde{\beta}})^2 [\pi_1 \exp(\psi_1 + \tau_1) + \pi_2 \exp(\tau_1)]}{\pi_1 \pi_2 \phi^{-1} [\exp(\psi_1 + \tau_1) - \exp(\tau_1)]^2 \mathbf{1}_n^\top R(\alpha)^{-1} \mathbf{1}_n} \\
 &= \frac{2(z_{1-\alpha^*/2} + z_{1-\tilde{\beta}})^2 [\exp(d + \beta_0) + \exp(\beta_0)]}{\phi^{-1} [\exp(d + \beta_0) - \exp(\beta_0)]^2 \mathbf{1}_n^\top R(\alpha)^{-1} \mathbf{1}_n}
 \end{aligned}$$

und damit die Behauptung.  $\square$

### Bemerkung 6.5.

Für den Fall einer *exchangeable*-Korrelationsstruktur, d.h.

$$[R(\alpha)_{jk}]_{\substack{k \in \{1, \dots, n\} \\ j \in \{1, \dots, n\}}} = \begin{cases} 1, & \text{falls } k = j, \\ \alpha, & \text{falls } k \neq j, \end{cases}$$

gilt

$$[R(\alpha)_{jk}]_{\substack{k \in \{1, \dots, n\} \\ j \in \{1, \dots, n\}}}^{-1} = \begin{cases} \frac{-n+2}{(1+(n-1)\alpha)}, & \text{falls } k = j, \\ \frac{\alpha}{(\alpha-1)(1+(n-1)\alpha)}, & \text{falls } k \neq j \end{cases}$$

und somit

$$\mathbf{1}_n^\top R(\alpha)^{-1} \mathbf{1}_n = \frac{n}{1 + (n-1)\alpha}.$$

Für den benötigten Stichprobenumfang  $m$  ergibt sich dann nach dem letzten Satz

$$m = \frac{2(z_{1-\alpha^*/2} + z_{1-\tilde{\beta}})^2 (\exp(d) + \exp(\beta_0 + d))(1 + (n-1)\alpha)}{n\phi^{-1} (\exp(\beta_0 + d) - \exp(\beta_0))^2}. \quad (6.10)$$

Dabei sind die Quantile für  $\alpha^* = 0.05$  bzw.  $\tilde{\beta} = 0.9$  durch  $z_{1-\alpha^*/2} \approx 1.96$  bzw.  $z_{1-\tilde{\beta}} \approx 1.282$  gegeben.

Abschließend sei darauf hingewiesen, dass es eine weitere Methode zur Berechnung von Fallzahlen gibt. Diese basiert ebenfalls auf der Theorie der GEE und wurde von Dahmen u. a. (2004) vorgestellt. Des Weiteren ist diese Methode in einem SAS-Makro namens GEESIZE implementiert.<sup>9</sup>

<sup>9</sup>Vgl. Rochon (1998).

## 6.2 Der Stichprobenumfang und seine Einflussgrößen

In diesem Abschnitt wird ein marginales Modell für ein Zweistichprobenproblem betrachtet und die Fallzahlformel (6.10) in Abhängigkeit ihrer Einflussgrößen untersucht. Das marginale Modell ist gegeben durch:

- Für  $i \in \{1, \dots, m\}$  und  $j \in \{1, \dots, n\}$  sind  $Y_{ij}$  poissonverteilte Zufallsvariablen, und die Realisierung  $y_{ij}$  entspricht beispielsweise der Anzahl der Hitzewallungen von Patientin  $i$  während der  $j$ -ten Beobachtungseinheit (z.B. Tag, Woche, Monat).
- Die marginale Erwartung  $E(Y_{ij}) = \lambda_{ij}$  und die erklärenden Variablen hängen wie folgt miteinander zusammen:

$$\ln(E(Y_{ij})) = \ln(\lambda_{ij}) = \begin{cases} \beta_0 + \beta_1, & \text{falls Placebo,} \\ \beta_0, & \text{falls Verum.} \end{cases}$$

- Als Abhängigkeitsstruktur wird die *exchangeable*-Struktur gewählt. Deshalb ist die Korrelationsmatrix des Zufallsvektors  $Y_i = (Y_{i1}, \dots, Y_{in})$  für alle  $i \in \{1, \dots, m\}$  gegeben durch

$$[R(\alpha)_{jk}]_{\substack{k \in \{1, \dots, n\} \\ j \in \{1, \dots, n\}}} = \begin{cases} 1, & \text{falls } k = j, \\ \alpha, & \text{falls } k \neq j. \end{cases}$$

Der benötigte Stichprobenumfang  $m$  kann dann durch das Festsetzen des Signifikanzniveaus  $\alpha^*$ , der Power  $\tilde{\beta}$ , der Anzahl an Wiederholungen  $n$  pro Person, des Korrelationsparameters  $\alpha$ , der kleinsten bedeutungsvollen Differenz  $d = \beta_1$ , des Überdispersions-Parameters  $\phi$ , sowie der Angabe von  $\beta_0$  mit Hilfe der vorgestellten Fallzahlformel berechnet werden.

Um die Entwicklung des Stichprobenumfangs zu analysieren, sind in Anhang G Funktionen definiert, die in Abhängigkeit von den oben genannten Einflussgrößen die Fallzahl  $m$  ermitteln. Die Funktion `fallzahl1` berechnet den benötigten Stichprobenumfang exakt nach der Formel (6.10). Dagegen rundet die Funktion `fallzahl2` den exakten Stichprobenumfang auf die nächstgrößere gerade Zahl. Die Berechnung der Stichprobengröße mit Hilfe dieser Funktion erlaubt es, beiden Behandlungsgruppen gleich viele Subjekte zuzuordnen. Darüber hinaus sind Funktionen der Form `Grafik_fallzahl_Parameter` gegeben.<sup>10</sup> Diese verdeutlichen die Entwicklung der

<sup>10</sup>Die Größe `Parameter` kann dabei durch `n`, `phi`, `alpha`, `d` und `beta0` ersetzt werden.

Wiederholungsanzahl  $n$  versus Stichprobengröße  $m$

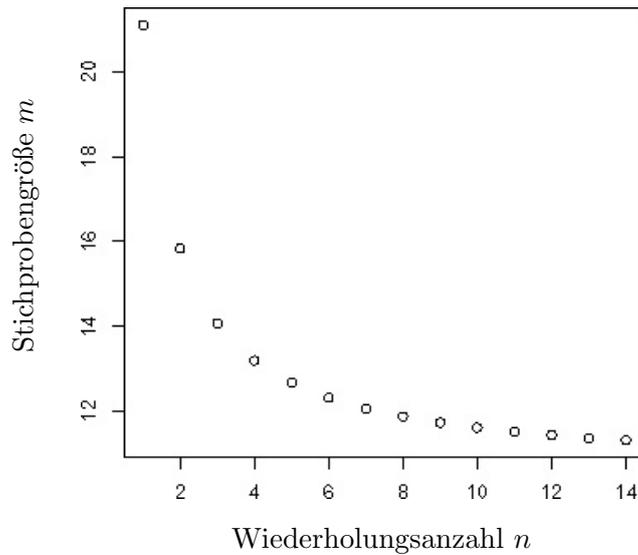


Abbildung 6.1: Die Entwicklung der Stichprobengröße  $m$  in Abhängigkeit der Anzahl  $n$  an wiederholten Beobachtungen pro Patientin für  $\beta_1 = \ln(2)$ .

Stichprobengröße  $m$  in Abhängigkeit des Parameters `Parameter` durch eine Grafik. Zudem werden die benötigten exakten und gerundeten Fallzahlen für unterschiedliche Werte des Parameters `Parameter` in einer Tabelle ausgegeben.

**Wiederholungen  $n$  versus Stichprobenumfang  $m$ :**

Wird angenommen, dass sich die Anzahl der Hitzewallungen in der Verum-Gruppe um die Hälfte gegenüber der Anzahl in der Placebo-Gruppe reduziert, gilt  $d = \beta_1 = \ln(2)$ . Des Weiteren seien die restlichen Parameter wie folgt festgesetzt:  $\beta_0 = 1.5$ ,  $\phi = 1.5$ ,  $\alpha = 0.5$ ,  $\alpha^* = 0.05$  und  $\tilde{\beta} = 0.9$ . Die Abbildung 6.1 veranschaulicht die Entwicklung der Stichprobengröße in Abhängigkeit von der Anzahl der Wiederholungen pro Patientin. Darüber hinaus wird bei der Auswertung der Funktion `Grafik_fallzahl_n(14,0.5,1.5,1.5,log(2))`<sup>11</sup> folgende Tabelle mit exakten und gerundeten Fallzahlen ausgegeben.

```
Grafik_fallzahl_n(14,0.5,1.5,1.5,log(2))
      n exaktes m gerundetes m
[1,]  1 21.10701          22
[2,]  2 15.83026          16
[3,]  3 14.07134          16
```

<sup>11</sup>Durch die Angabe von  $n = 14$  wird vorgegeben, dass die benötigte Stichprobengröße für  $n \in \{1, \dots, 14\}$  berechnet und dargestellt werden soll.

Wiederholungsanzahl  $n$  versus Stichprobengröße  $m$

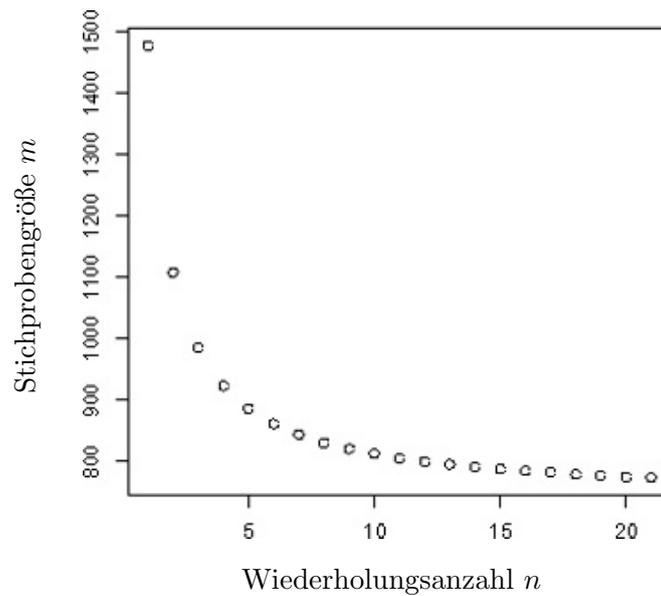


Abbildung 6.2: Die Entwicklung der Stichprobengröße  $m$  in Abhängigkeit der Anzahl  $n$  an wiederholten Beobachtungen pro Patientin für  $\beta_1 = \ln(1.1)$ .

[4,]	4	13.19188	14
[5,]	5	12.66421	14
[6,]	6	12.31243	14
[7,]	7	12.06115	14
[8,]	8	11.87270	12
[9,]	9	11.72612	12
[10,]	10	11.60886	12
[11,]	11	11.51292	12
[12,]	12	11.43297	12
[13,]	13	11.36532	12
[14,]	14	11.30733	12

Anhand der Grafik 6.1 und der Tabelle lässt sich erkennen, dass der benötigte Stichprobenumfang mit Zunahme der Wiederholungsanzahlen sinkt. Allerdings wird auch deutlich, dass es zu keiner nennenswerten Reduktion des Stichprobenumfangs für  $n \geq 4$  kommt.

Unterscheiden sich beide Behandlungsgruppen lediglich durch den Faktor 1.1, d.h.  $\beta_1 = \ln(1.1)$ , ergibt sich eine ähnliche Entwicklung. Diese ist in Abbildung 6.2 dargestellt. Zudem treten folgende Fallzahlen auf:<sup>12</sup>

<sup>12</sup>In diesem Fall wurde  $n = 21$  in der Funktion `Grafik.fallzahl_n` gewählt. Somit wird der Stichprobenumfang zur Erreichung der Power von 0.9 und eines Signifikanzniveaus von 0.05 für

Grafik\_fallzahl\_n(21,0.5,1.5,1.5,log(1.1))

	n	exaxtes m	gerundetes m
[1,]	1	1477.4910	1478
[2,]	2	1108.1183	1110
[3,]	3	984.9940	986
[4,]	4	923.4319	924
[5,]	5	886.4946	888
[6,]	6	861.8698	862
[7,]	7	844.2806	846
[8,]	8	831.0887	832
[9,]	9	820.8283	822
[10,]	10	812.6201	814
[11,]	11	805.9042	806
[12,]	12	800.3076	802
[13,]	13	795.5721	796
[14,]	14	791.5130	792
[15,]	15	787.9952	788
[16,]	16	784.9171	786
[17,]	17	782.2011	784
[18,]	18	779.7869	780
[19,]	19	777.6268	778
[20,]	20	775.6828	776
[21,]	21	773.9239	774

In diesem Fall reduziert sich die Stichprobengröße mit Anwachsen der Wiederholungen pro Subjekt ebenfalls. Allerdings ist auch hier ein weniger steiles Absinken für  $n \geq 6$  erkennbar.

### Korrelation versus Stichprobenumfang $m$ :

Jetzt soll analysiert werden, wie sich die Anzahl der benötigten Subjekte in Abhängigkeit zu der Korrelation verändert. In diesem Teil wird lediglich die `exchangeable`-Korrelationsstruktur betrachtet. Werden pro Subjekt drei Beobachtungen erhoben und gilt  $d = \ln(2)$ ,  $\beta_0 = 1.5$ , sowie  $\phi = 1.5$ , ergibt sich die in Abbildung 6.3 dargestellte Veränderung des benötigten Stichprobenumfangs und die folgende Tabelle:

---

$n \in \{1, \dots, 21\}$  berechnet.

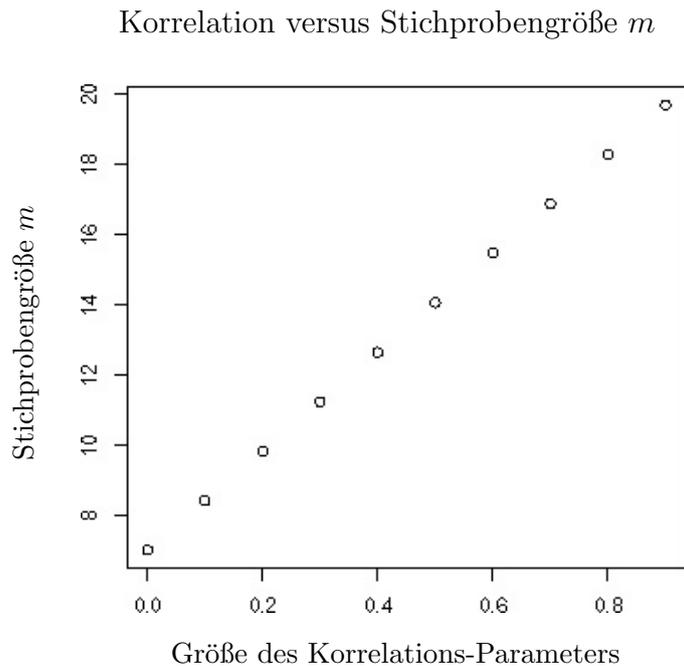


Abbildung 6.3: Die Entwicklung der Stichprobengröße  $m$  in Abhängigkeit zu der Korrelation  $\alpha \in \{0.1, \dots, 0.9\}$ .

Grafik\_fallzahl\_alpha(3,0.9,1.5,1.5,log(2))

	alpha	exaktes m	gerundetes m
[1,]	0.0	7.035671	8
[2,]	0.1	8.442806	10
[3,]	0.2	9.849940	10
[4,]	0.3	11.257074	12
[5,]	0.4	12.664209	14
[6,]	0.5	14.071343	16
[7,]	0.6	15.478477	16
[8,]	0.7	16.885612	18
[9,]	0.8	18.292746	20
[10,]	0.9	19.699880	20

Es lässt sich somit vermuten, dass sich der Stichprobenumfang  $m$  linear zu der Korrelation  $\alpha$  entwickelt. Es sei darauf hingewiesen, dass diese Analyse lediglich auf einer positiven Korrelation  $\alpha \in \{0.1, \dots, 0.9\}$  beruht.

### Überdispersion versus Stichprobenumfang $m$ :

Nun wird die Entwicklung der Stichprobengröße in Abhängigkeit der Überdispersion betrachtet. Es seien  $d = \ln(2)$ ,  $\beta_0 = 1.5$ ,  $\alpha = 0.5$ ,  $n = 3$  und  $\phi \in \{1, 1.2, \dots, 3\}$ . Analog zum vorigen Fall lässt sich in Abbildung 6.4 ein linearer Zusammenhang zwischen

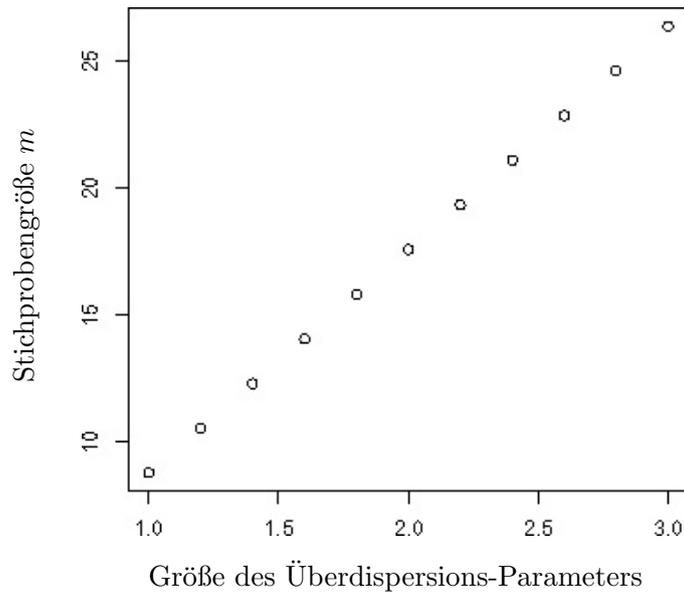
Überdispersion versus Stichprobengröße  $m$ 

Abbildung 6.4: Darstellung des Zusammenhangs zwischen der Stichprobengröße  $m$  und des Überdispersion-Parameters  $\phi \in \{1, 1.2, \dots, 3\}$ .

der Überdispersion und der Stichprobengröße erkennen. Auf eine tabellarische Ausgabe wurde verzichtet.

**Größe des Behandlungsunterschieds  $d$  versus Stichprobenumfang  $m$ :**

In früheren Betrachtungen wurde bereits deutlich, dass bei einem geringen Behandlungsunterschied mehr Studienteilnehmer benötigt werden als bei einem großen Behandlungsunterschied. Dieser Sachverhalt wird für  $n = 3$ ,  $\phi = 1.5$ ,  $\alpha = 0.5$  und  $\beta_0 = 1.5$  durch Abbildung 6.5 und die folgende Tabelle veranschaulicht.

```
Grafik_fallzahl_d(3,0.5,1.5,1.5,2)
      d  exaktes m  gerundetes m
[1,] 0.1 892.7098476      894
[2,] 0.2 212.5570661      214
[3,] 0.3  90.0472468       92
[4,] 0.4  48.3182407       50
[5,] 0.5  29.5212183       30
[6,] 0.6  19.5848580       20
[7,] 0.7  13.7549132       14
[8,] 0.8  10.0730551       12
[9,] 0.9   7.6167730        8
[10,] 1.0   5.9070113        6
```

Behandlungsunterschied versus Stichprobengröße  $m$

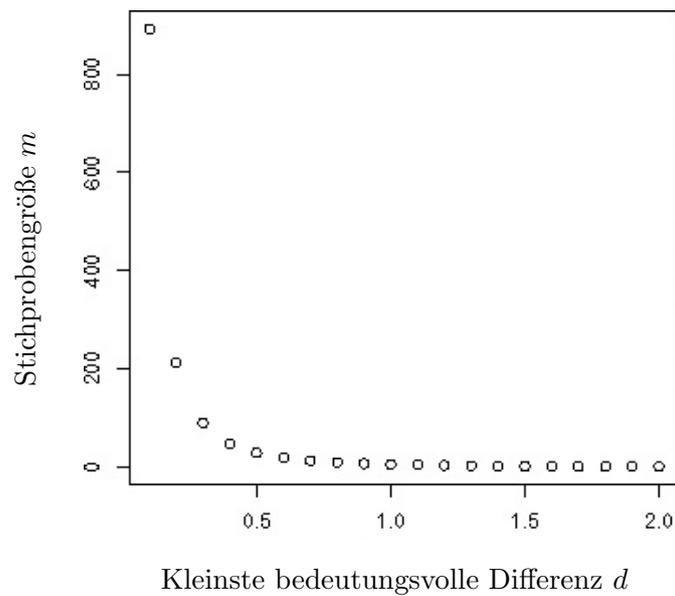


Abbildung 6.5: Die Entwicklung der Stichprobengröße  $m$  in Abhängigkeit zu dem Behandlungsunterschied, der multiplikativ durch den Faktor  $\exp(d)$  mit  $d \in \{\ln(1.1), \dots, \ln(3)\}$  gegeben ist.

[11,]	1.1	4.6758329	6
[12,]	1.2	3.7643505	4
[13,]	1.3	3.0737750	4
[14,]	1.4	2.5402318	4
[15,]	1.5	2.1210393	4
[16,]	1.6	1.7868652	2
[17,]	1.7	1.5170564	2
[18,]	1.8	1.2967599	2
[19,]	1.9	1.1150931	2
[20,]	2.0	0.9639488	2

**Logarithmus der marginalen Erwartung in der Verum-Gruppe versus Stichprobenumfang  $m$ :**

Schließlich soll jetzt noch die Entwicklung der Stichprobengröße in Abhängigkeit der Größe  $\beta_0$  betrachtet werden. Dabei spiegelt  $\beta_0$  gerade den Logarithmus der marginalen Erwartung in der Verum-Gruppe wieder.

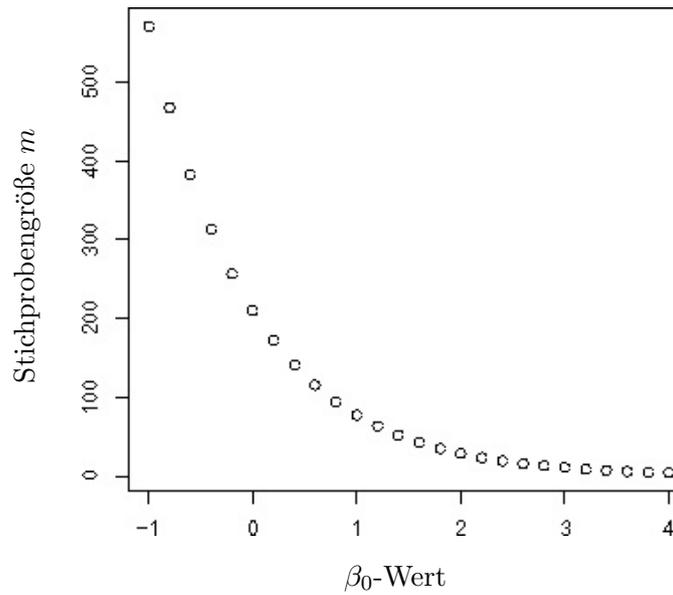
$\beta_0$  versus Stichprobengröße  $m$ 

Abbildung 6.6: Grafische Darstellung des Zusammenhangs zwischen der Stichprobengröße  $m$  und dem Regressionsparameter  $\beta_0$ .

Für  $n = 3$ ,  $d = \ln(1.5)$ ,  $\alpha = 0.5$ ,  $\phi = 1.5$  und  $\beta_0 \in \{-1, \dots, 4\}$  ergibt sich:

Grafik\_fallzahl\_beta0(3,0.5,1.5,4,log(1.5))

	beta0	exaktes m	gerundetes m
[1,]	-1.0	571.413503	572
[2,]	-0.8	467.833807	468
[3,]	-0.6	383.029925	384
[4,]	-0.4	313.598379	314
[5,]	-0.2	256.752637	258
[6,]	0.0	210.211280	212
[7,]	0.2	172.106440	174
[8,]	0.4	140.908835	142
[9,]	0.6	115.366397	116
[10,]	0.8	94.454017	96
[11,]	1.0	77.332408	78
[12,]	1.2	63.314421	64
[13,]	1.4	51.837463	52
[14,]	1.6	42.440925	44
[15,]	1.8	34.747691	36
[16,]	2.0	28.449003	30
[17,]	2.2	23.292074	24

[18,]	2.4	19.069937	20
[19,]	2.6	15.613144	16
[20,]	2.8	12.782961	14
[21,]	3.0	10.465803	12
[22,]	3.2	8.568675	10
[23,]	3.4	7.015438	8
[24,]	3.6	5.743755	6
[25,]	3.8	4.702589	6
[26,]	4.0	3.850154	4

Damit ist die Analyse der Entwicklung der Stichprobengröße in Abhängigkeit der unterschiedlichen Einflussgrößen abgeschlossen. Es sei allerdings darauf hingewiesen, dass bei der Entwicklung eines Studiendesigns neben der Abwägung zwischen Anzahl der Wiederholungen pro Studienteilnehmer und der Stichprobengröße natürlich auch die Kosten berücksichtigt werden müssen.

# Kapitel 7

## Simulation korrelierter Zählraten

Im Rahmen dieser Diplomarbeit werden die in Kapitel 3 vorgestellten Modelle und die dazugehörigen Tests auf ihre Güte untersucht. Einerseits muss die in Kapitel 6 berechnete Stichprobengröße bei gegebenen Einflussgrößen, wie beispielsweise einer Power von  $1 - \tilde{\beta} = 0.9$  und einem Signifikanzniveau von  $\alpha^* = 0.05$ , überprüft werden. Andererseits ist es auch notwendig, die bestehenden Implementierungen auf ihre Funktionsweise zu untersuchen. Da in Kapitel 6 lediglich für marginale Modelle eine Formel zur Berechnung des Stichprobenumfangs vorgestellt werden konnte, werden die Simulationen in diesem Kapitel entsprechend nur für diese Modelle durchgeführt. Zur Überprüfung der genannten Punkte, ist es unerlässlich, Daten mit einer vorgegebenen Korrelationsstruktur zu simulieren. Als Parameter sind hierbei nicht nur der Korrelationsparameter  $\alpha$ , sondern auch der Behandlungseffekt, die Anzahl der Subjekte und die Anzahl der Beobachtungen pro Subjekt notwendig.<sup>1</sup> Basierend auf den simulierten Daten zweier Behandlungsgruppen mit fest gewählten Parametern kann dann untersucht werden, ob die zuvor festgesetzte Power bei der Berechnung des Stichprobenumfangs erreicht wird.

### 7.1 Simulation von Daten für das marginale Modell

Um Zählraten zu simulieren, die mit dem marginalen Modell ausgewertet werden, müssen poissonverteilte Daten mit einer bekannten Korrelationsstruktur erzeugt werden. Für unabhängige, poissonverteilte Daten existieren sowohl in der Statistik-Software **SAS** als auch in **R** erzeugende Funktionen, gegeben durch `ranpoi` bzw.

---

<sup>1</sup>Die Bestimmung von  $\alpha$  setzt die Auswahl einer Korrelationsstruktur voraus (vgl. Abschnitt 3.1.1.3).

rpois. Allerdings gibt es für die Erzeugung korrelierter Daten keine bereits implementierten Programme in den genannten Softwarepaketen.

Mit dem wachsenden Bekanntheitsgrad der Verallgemeinerten Schätzgleichungen (GEE) stieg auch die Anzahl der Veröffentlichungen zur Gewinnung von korrelierten nicht-normalverteilten Daten. So beschäftigten sich Emrich und Piedemonte (1991), Lee (1993), Park u. a. (1996) und Gange (1995) mit der Generierung von korrelierten binären Daten. Darüber hinaus wurde eine Möglichkeit zur Erzeugung von korrelierten gammaverteilten Daten von Ronning (1977) und eine Methode zur Erzeugung von korrelierten rechteckverteilten Daten von Willemain und Desautels (1993) veröffentlicht.

Erfreulicherweise kann in unserem Fall zurückgegriffen werden auf eine Arbeit von Park und Shin (1998), die sich mit der Erzeugung von Daten beschäftigt, deren Verteilung bezüglich der Summation abgeschlossen ist. Später wird gezeigt, dass poissonverteilte Daten diese Eigenschaft erfüllen.

**Definition 7.1.**

Eine Klasse von Verteilungsfunktionen  $D = \{F_\alpha; \alpha \geq 0\}$  heißt abgeschlossen bezüglich der Summation, falls für  $F_\alpha \in D$  und  $F_\beta \in D$  auch

$$F_\alpha * F_\beta \in D$$

gilt, wobei  $F_\alpha * F_\beta$  die Faltung von  $F_\alpha$  und  $F_\beta$  bezeichnet und durch

$$(F_\alpha * F_\beta)(x) := \int_{-\infty}^{\infty} F_\alpha(x - u) dF_\beta(u)$$

definiert ist.

Diese Eigenschaft von Verteilungen ist äquivalent zu einer weiteren, die sich auf die zugehörigen Zufallsvariablen bezieht. Diese Beziehung wird hier lediglich für poissonverteilte Zufallsvariablen vorgestellt und nicht bewiesen:

**Satz 7.2.**<sup>2</sup>

Sei  $P_\lambda(k) = \frac{\lambda^k}{k!} e^{-\lambda}$  die Zähldichte einer poissonverteilten Zufallsvariablen, dann ist

$$D = \{F_\lambda(n) | \lambda \geq 0\} = \{F_\lambda(n) = \sum_{k=-\infty}^n P_\lambda(k) = \sum_{k=0}^n \frac{\lambda^k}{k!} e^{-\lambda} | \lambda \geq 0\}$$

eine bezüglich der Summation abgeschlossene Klasse von Verteilungen.<sup>3</sup> Zudem ist diese Eigenschaft äquivalent zur Aussage, dass für unabhängige Zufallsvariablen  $Y$  und  $Z$  und deren zugehörigen Verteilungsfunktionen  $F_\alpha \in D$  bzw.  $F_\beta \in D$  die Summe  $Y + Z$  die Verteilungsfunktion  $F_{\alpha+\beta} \in D$  besitzt.

---

<sup>2</sup>Vgl. Park und Shin (1998), S.129.

<sup>3</sup>Hierbei ist zu beachten, dass für alle  $\lambda \in \mathbb{N}_0$  und für alle  $k < 0$   $P_\lambda(k) = 0$  gilt.

Für die Verteilungen, die abgeschlossen bezüglich der Summation sind, wird jetzt in Anlehnung an Park und Shin (1998) ein Algorithmus zur Erzeugung korrelierter Daten vorgestellt werden. Nach Satz 7.2 wird damit auch eine Methode zur Erzeugung korrelierter Zähldaten geliefert. Zur Vereinfachung wird in diesem Kapitel angenommen, dass  $n_i = n$  für alle  $i \in \{1, \dots, m\}$  gilt.

Ziel ist es jetzt, zu einem vorgegebenen Erwartungswertvektor  $\lambda = (\lambda_1, \dots, \lambda_n)^\top$ , einer vorgegebenen Korrelationsmatrix  $R_i = (\rho_{jk})$  mit  $k, j \in \{1, \dots, n\}$  und einer vorgegebenen Klasse von bezüglich der Summation abgeschlossenen Verteilungen (hier: Poissonverteilung) einen Algorithmus zu erhalten, der nichtnegative Zufallsvariablen  $Y_{i1}, \dots, Y_{in}$  mit  $E(Y_{ij}) = \lambda_j$  und  $\text{Corr}(Y_{ij}, Y_{ik}) = \rho_{jk} \geq 0, k \neq j$ , für alle  $i \in \{1, \dots, m\}$  und  $k, j \in \{1, \dots, n\}$  liefert.

Da unterschiedliche subjekt-gebundene Korrelationsmatrizen in den marginalen Modellen und deren Analyse mittels GEE nicht vorgesehen sind, ist es sinnvoll,  $R_i =: R$  für alle  $i \in \{1, \dots, m\}$  anzunehmen. Darüber hinaus sollen, wie oben bereits angedeutet, die Erwartungswertvektoren der unterschiedlichen Subjekte gleich sein.

Die von Park und Shin (1998) vorgestellte Methode erzeugt einen Vektor  $X = (X_1, \dots, X_{\tau+n})^\top$ , mit  $\tau \in \mathbb{N}$ , von unabhängigen, poissonverteilten Zufallsvariablen und eine Matrix  $T$ , deren Einträge aus Nullen und Einsen bestehen, so dass für  $Y_i = (Y_{i1}, \dots, Y_{in})^\top$  und alle  $i \in \{1, \dots, m\}$  folgendes gilt:

$$\begin{aligned} Y_i &= TX \quad \text{mit} \quad T \in \mathbb{R}^{n \times \tau+n}, \\ E(Y_i) &= \lambda_i \otimes 1_n, \\ \text{und} \quad \text{Corr}(Y_{ij}, Y_{ik}) &= \rho_{jk} \quad \text{für} \quad j \neq k. \end{aligned}$$

Die Erzeugung der korrelierten Zähldaten wird somit erreicht, indem die einzelnen Zufallsvariablen  $Y_{i1}, \dots, Y_{in}$  als Linearkombination von unabhängigen, poissonverteilten Zufallsvariablen  $X_1, \dots, X_{\tau+n}$  dargestellt werden. Dabei kommt die geforderte Korrelation  $\text{Corr}(Y_{ij}, Y_{ik}) = \rho_{jk}$  u.a. dadurch zu Stande, dass alle Zufallsvariablen  $Y_{ij}$  mindestens die Zufallsvariable  $X_1$  gemeinsam haben. Dies wird im Folgenden genauer beschrieben.

Der vorgestellte Algorithmus wird aus Gründen der Übersichtlichkeit mit Hilfe von Kanten und Knoten beschrieben.

### 7.1.1 Der Algorithmus

Der Algorithmus zur Bestimmung der Matrix  $T$  und des Zufallsvektors  $X$  enthält nach Park und Shin (1998) folgende Schritte:

**Schritt 1**

Setze  $\ell = 1$  für den ersten Schritt der Iteration und berechne den Wert

$$\lambda_{jk}^\ell = \rho_{jk} \sqrt{\lambda_j \lambda_k} \text{ für alle } j, k \in \{1, \dots, n\}.$$

Konstruiere dann einen Graphen mit  $n$  Knoten, welche die Zufallsvariablen darstellen, und verbinde diese jeweils miteinander. Dabei bezeichne  $i$ ,  $i \in \{1, \dots, m\}$ , den  $i$ -ten Knoten und  $(j, k)$  die Kante zwischen den Knoten  $j$  und  $k$ ,  $j, k \in \{1, \dots, n\}$ . Dem  $i$ -ten Knoten wird für  $\ell = 1$  der Wert  $\lambda_i = \lambda_{ii}^\ell$  zugeordnet und der Kante zwischen den Knoten  $k$  und  $j$  der Wert  $\lambda_{jk}^\ell$ .

**Schritt 2**

Finde den kleinsten Wert einer Kante  $(r, s)$ , wobei  $r$  und  $s$  Knoten sind und setze  $\beta_\ell = \lambda_{rs}^k$ .

**Schritt 3**

Definiere die Menge  $S_\ell$  als Menge derjenigen Knoten  $i$ , die sowohl mit Knoten  $r$  als auch mit Knoten  $s$  verbunden sind ( $r$  und  $s$  aus Schritt 2).

**Schritt 4**

Aktualisiere die Werte der Kanten und Knoten gemäß

$$\lambda_{ij}^{\ell+1} = \begin{cases} \lambda_{ij}^\ell - \beta_\ell, & \text{falls } i \in S_\ell \text{ und } j \in S_\ell, \\ \lambda_{ij}^\ell, & \text{falls } i \notin S_\ell \text{ oder } j \notin S_\ell \end{cases}$$

und entferne alle Kanten  $(i, j)$  mit  $\lambda_{ij}^{\ell+1} = 0$ . Wenn alle Kantenwerte kleiner oder gleich Null sind muss mit Schritt 5 fortgefahren werden. Ansonsten setze  $\ell = \ell + 1$  und gehe zu Schritt 2.

**Schritt 5**

Wenn ein Knoten  $i$  existiert mit  $\lambda_{ii} < 0$ , dann bricht der Algorithmus ohne Lösung ab. Mögliche Gründe werden im Laufe dieses Abschnitts diskutiert. Andernfalls setze  $\ell = \tau$ ,  $\beta_{\tau+1} = \lambda_1^\tau, \dots, \beta_{\tau+n} = \lambda_n^\tau$  und  $S_{\tau+1} = \{1\}, \dots, S_{\tau+n} = \{n\}$ . Mit Hilfe dieser Mengen muss jetzt eine  $n \times (\tau + n)$  Matrix  $T = (T_{ij})$  mit  $i \in \{1, \dots, n\}$  und  $j \in \{1, \dots, \tau + n\}$  konstruiert werden. Dabei sind die Einträge der Matrix wie folgt definiert:

$$T_{ij} = \begin{cases} 1, & \text{falls } i \in S_j, \\ 0, & \text{falls } i \notin S_j \end{cases}$$

für alle  $j \in \{1, \dots, \tau + n\}$ . Für den Zufallsvektor  $X$  gilt

$$X = (X_1, \dots, X_{\tau+n})^\top \text{ und } X_i \sim \mathcal{P}(\beta_i)$$

für alle  $i \in \{1, \dots, \tau + n\}$ .

Eine ausführlichere Beschreibung der notwendigen Implementierungen und ein R-Programm, mit dessen Hilfe nach diesem Schema korrelierte Zähldaten erzeugt werden können, ist im Anhang H gegeben.

Führt dieses Verfahren zu einer Lösungsmatrix  $T$ , so gelten für  $T$  und den Zufallsvektor  $X = (X_1, \dots, X_{\tau+n})^\top$  mit  $X_i \sim \mathcal{P}(\beta_i)$  nach Satz 7.3 die geforderten Eigenschaften

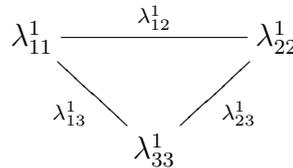
$$\begin{aligned} Y_i &= TX, \\ E(Y_i) &= \lambda_i \otimes 1_n, \\ \text{und } \text{Corr}(Y_{ij}, Y_{ik}) &= \rho_{jk} \text{ für } j \neq k \end{aligned}$$

für alle  $i \in \{1, \dots, m\}$  und  $j, k \in \{1, \dots, n\}$ . Der vorgestellte Algorithmus soll im Folgenden anhand eines Beispiels veranschaulicht werden. Gesucht sind poissonverteilte Zufallsvariablen  $Y_1, Y_2$  und  $Y_3$  mit  $Y = (Y_1, Y_2, Y_3)^\top$ ,

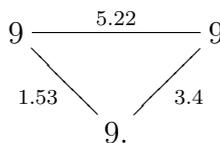
$$E(Y) = \begin{pmatrix} 9 \\ 9 \\ 9 \end{pmatrix} \text{ und } \text{Corr}(Y) = \begin{pmatrix} 1 & 0.58 & 0.17 \\ 0.58 & 1 & 0.37 \\ 0.17 & 0.37 & 1 \end{pmatrix}. \quad (7.1)$$

Nach dem Algorithmus gilt:

**Schritt 1** Setze  $\ell = 1$  und berechne die Werte  $\lambda_{jk}^1 = \rho_{jk} \sqrt{\lambda_j \lambda_k}$  für alle  $j, k \in \{1, 2, 3\}$ . Dann ergibt sich der Graph

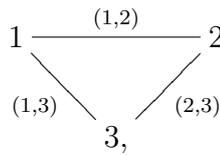


und mit eingesetzten Werten also



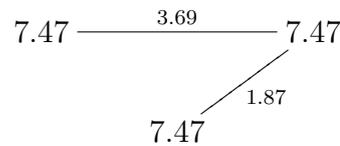
**Schritt 2a** Der kleinste Wert der Kante ist durch  $\lambda_{13}^1 = 1.53 =: \beta_1$  gegeben.

**Schritt 3a** Werden die Knoten und Kanten folgendermaßen benannt:



dann gilt, dass alle Knoten sowohl mit dem Knoten 1 als auch mit dem Knoten 3 über eine Kante verbunden sind, womit  $S_1 = \{1, 2, 3\}$  gilt.

**Schritt 4a** Aktualisiere die Werte der Kanten und Knoten gemäß den Vorgaben in Schritt 4, so dass sich der neue Graph

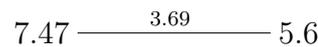


ergibt. Die Kante (1,3) wurde entfernt, weil der aktualisierte Wert hier zu Null wird. Da noch nicht alle Kantenwerte kleiner oder gleich Null sind, muss erneut bei Schritt 2 fortfahren und  $\ell = 2$  gesetzt werden.

**Schritt 2b** Der kleinste Wert des neuen Graphen ist gegeben durch  $\lambda_{23}^2 = 1.87 =: \beta_2$ .

**Schritt 3b** In diesem Fall ist der Knoten 1 nicht mit den Knoten 2 und 3 verbunden, so dass  $S_2 = \{2, 3\}$  gilt.

**Schritt 4b** Gemäß den Anforderungen in Schritt 4 ist der aktualisierte Graph gegeben durch:



5.6.

Da nicht alle Kantenwerte kleiner oder gleich Null sind, wird  $\ell = 3$  in Schritt 2 gesetzt.

**Schritt 2c** Der kleinste Kantenwert ist durch  $\lambda_{12}^3 = 3.69 =: \beta_3$  gegeben.

**Schritt 3c** Die Kante (1,2) ist lediglich mit den Knoten 1 und 2 verbunden, so dass  $S_3 = \{1, 2\}$  gilt.

**Schritt 4c** Die aktualisierten Werte sind gegeben durch



5.6.

Alle Kantenwerte sind in diesem Fall Null, so dass mit Schritt 5 fortgefahren wird.

**Schritt 5** Da alle Knoten positiv sind, kann jetzt die Matrix  $T$  aufgestellt werden.

Es ergibt sich  $\beta_4 := \lambda_1^3 = 3.78$ ,  $\beta_5 := \lambda_2^3 = 1.91$  und  $\beta_6 := \lambda_3^3 = 5.6$ . Darüber hinaus gilt  $S_4 = \{1\}$ ,  $S_5 = \{2\}$  und  $S_6 = \{3\}$ .

Die erste Spalte der Matrix  $T$  berechnet sich nun mit Hilfe der Menge  $S_1 = \{1, 2, 3\}$ . Die einzelnen Einträge der ersten Spalte sind hierbei gegeben durch

$$T_{i1} = \begin{cases} 1, & \text{falls } i \in S_1, \\ 0, & \text{falls } i \notin S_1, \end{cases}$$

so dass die erste Spalte die Form  $(1, 1, 1)^\top$  hat. Entsprechend erhält man für die zweite Spalte den Vektor  $(0, 1, 1)^\top$  und so weiter. Insgesamt ergibt sich damit die  $3 \times 6$  Matrix  $T$  mit

$$T = \begin{pmatrix} 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \end{pmatrix}. \quad (7.2)$$

Der Zufallsvektor  $X$  ist dabei gegeben durch den Vektor  $(X_1(\beta_1), \dots, X_6(\beta_6))^\top$ , wobei gilt

$$X_i \sim \mathcal{P}(\beta_i).$$

Im beschriebenen Beispiel folgt somit für den Erwartungswert

$$\mathbb{E}(Y) = \mathbb{E}(TX) = T \cdot \mathbb{E}(X) = \begin{pmatrix} 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1.53 \\ 1.87 \\ 3.69 \\ 3.78 \\ 1.91 \\ 5.6 \end{pmatrix} = \begin{pmatrix} 9 \\ 9 \\ 9 \end{pmatrix}.$$

Für die Korrelationsmatrix ergibt sich ferner

$$\begin{aligned} \text{Corr}(Y) &= \frac{1}{9} \text{Cov}(Y) = \frac{1}{9} \text{Cov}(TX) \\ &= \frac{1}{9} T \text{Cov}(X) T^\top \\ &= \frac{1}{9} T \text{diag}(1.53, 1.87, 3.69, 3.78, 1.91, 5.6) T^\top \\ &= \frac{1}{9} \begin{pmatrix} 9 & 5.22 & 1.53 \\ 5.22 & 9 & 3.4 \\ 1.53 & 3.4 & 9 \end{pmatrix} \end{aligned}$$

$$= \begin{pmatrix} 1 & 0.58 & 0.17 \\ 0.58 & 1 & 0.37 \\ 0.17 & 0.37 & 1 \end{pmatrix}.$$

Nun bleibt noch zu klären, unter welchen Bedingungen der Algorithmus eine Lösung  $Y = (Y_1, \dots, Y_n)^\top$  liefert, so dass

$$\begin{aligned} Y &= TX \\ E(Y_i) &= \lambda_i \\ \text{und } \text{Corr}(Y_j, Y_k) &= \rho_{jk} \text{ für } j \neq k \end{aligned}$$

für alle  $i \in \{1, \dots, m\}$  und  $j, k \in \{1, \dots, n\}$  gilt.

**Satz 7.3.**

Seien  $T$  die Matrix und  $X = (X_1(\beta_1), \dots, X_{\tau+n}(\beta_{\tau+n}))^\top$  der Zufallsvektor, die der beschriebene Algorithmus liefert. Liegen alle Verteilungen der Zufallsvariablen  $X_\ell(\beta_\ell)$  in einer Klasse  $D$  von bezüglich der Summation abgeschlossenen Verteilungen und gilt

- (i)  $X_\ell(\beta_\ell)$  sind für alle  $\ell \in \{1, \dots, \tau + n\}$  unabhängig,
- (ii)  $E(X_\ell(\beta_\ell)) = \beta_\ell$  für alle  $\ell \in \{1, \dots, \tau + n\}$  sowie
- (iii)  $\text{Var}(X_\ell(\beta_\ell)) = \eta \beta_\ell$  mit  $\eta > 0$ ,

dann ist die Verteilung jeder Zufallsvariablen des Zufallsvektors  $Y = (Y_1, \dots, Y_n)^\top = TX$  auch in  $D$  enthalten und es folgt

$$E(Y_i) = \lambda_i \text{ und } \text{Corr}(Y_i, Y_j) = \rho_{ij}$$

für alle  $i, j \in \{1, \dots, n\}$ .

*Beweis:*

Mit Hilfe des zuvor beschriebenen Algorithmus wird eine Matrix  $T = [T_{i\ell}]_{\substack{i=1, \dots, n \\ \ell=1, \dots, \tau+n}}$  mit Einträgen  $T_{i\ell} \in \{0, 1\}$  bestimmt. Es gilt  $Y = (Y_1, \dots, Y_n)^\top = TX$  und somit für alle  $i \in \{1, \dots, n\}$

$$Y_i = \sum_{\ell=1}^{\tau+n} T_{i\ell} X_\ell(\beta_\ell).$$

Da die Zufallsvariablen  $X_\ell(\beta_\ell)$  für alle  $\ell \in \{1, \dots, \tau + n\}$  unabhängig sind und deren Verteilungsfunktionen in einer bezüglich der Summation abgeschlossenen Klasse  $D$

liegen, ergibt sich nach Satz 7.2 auch  $Y_i \in D$  für alle  $i \in \{1, \dots, n\}$ .

Nun bleibt noch zu zeigen, dass für alle  $i, j \in \{1, \dots, n\}$

$$E(Y_i) = \lambda_i \quad \text{und} \quad \text{Corr}(Y_i, Y_j) = \rho_{ij}$$

gilt.

Im Laufe der Durchführung des Algorithmus wird im Schritt 3 für jeden Iterationsschritt  $\ell \in \{1, \dots, \tau+n\}$  die Menge  $S_\ell$  bestimmt und anschließend werden im Schritt 4 die Werte der Kanten und Knoten aktualisiert. Hierbei wird für alle  $i, j \in \{1, \dots, n\}$ ,  $\ell \in \{1, \dots, \tau+n\}$  und  $(i, j) \in S_\ell$  vom Kanten- bzw. Knotenwert  $\lambda_{ij}^\ell$  der Wert  $\beta_\ell$  abgezogen, bis der Kanten- bzw. der Knotenwert Null wird. Deshalb ergibt sich bei Existenz einer Lösung (vgl. Schritt 5):

$$\lambda_{ij}^1 = \sum_{\ell=1}^{\tau+n} \beta_\ell \mathbb{1}_{S_\ell}(i) \mathbb{1}_{S_\ell}(j). \quad (7.3)$$

Zudem gilt aufgrund der Konstruktion der Matrix  $T$  folgendes

$$Y_i = \sum_{\ell=1}^{\tau+n} T_{i\ell} X_\ell(\beta_\ell) = \sum_{\ell=1}^{\tau+n} \mathbb{1}_{S_\ell}(i) X_\ell(\beta_\ell),$$

so dass sich der Erwartungswert von  $Y_i$  berechnet als

$$\begin{aligned} E(Y_i) &= E\left(\sum_{\ell=1}^{\tau+n} \mathbb{1}_{S_\ell}(i) X_\ell(\beta_\ell)\right) \\ &= \sum_{\ell=1}^{\tau+n} \mathbb{1}_{S_\ell}(i) E(X_\ell(\beta_\ell)) \\ &\stackrel{\text{Vor. (ii)}}{=} \sum_{\ell=1}^{\tau+n} \mathbb{1}_{S_\ell}(i) \beta_\ell \\ &\stackrel{(7.3)}{=} \lambda_{ii}^1 \\ &= \lambda_i. \end{aligned}$$

Entsprechend ergibt sich für die Kovarianz von  $Y_i$  und  $Y_j$  und damit für die  $(i, j)$ -te Koordinate  $\text{Cov}(Y_i, Y_j)$  der Kovarianzmatrix  $\text{Cov}(Y)$  für alle  $i, j \in \{1, \dots, n\}$ :

$$\begin{aligned} \text{Cov}(Y_i, Y_j) &= (\text{Cov}(Y))_{ij} \\ &= (\text{Cov}(TX))_{ij} \\ &= (T \text{Cov}(X) T^\top)_{ij} \\ &\stackrel{\text{Vor. (iii)}}{=} T_{i\bullet} \text{diag}(\eta\beta_1, \dots, \eta\beta_{\tau+n}) (T^\top)_{\bullet j} \end{aligned}$$

$$\begin{aligned}
 & \stackrel{(*)}{=} \left( \mathbb{1}_{S_1}(i) \eta \beta_1, \dots, \mathbb{1}_{S_{\tau+n}}(i) \eta \beta_{\tau+n} \right) \begin{pmatrix} \mathbb{1}_{S_1}(j) \\ \mathbb{1}_{S_2}(j) \\ \vdots \\ \mathbb{1}_{S_{\tau+n}}(j) \end{pmatrix} \\
 & = \sum_{\ell=1}^{\tau+n} \eta \beta_{\ell} \mathbb{1}_{S_{\ell}}(i) \mathbb{1}_{S_{\ell}}(j) \\
 & = \eta \sum_{\ell=1}^{\tau+n} \beta_{\ell} \mathbb{1}_{S_{\ell}}(i) \mathbb{1}_{S_{\ell}}(j) \\
 & \stackrel{(7.3)}{=} \eta \lambda_{ij}^1. \tag{7.4}
 \end{aligned}$$

Dabei bezeichnet  $T_{i\bullet}$  die  $i$ -te Zeile sowie  $T_{\bullet j}$  die  $j$ -te Spalte der Matrix  $T$  und die Gleichheit in  $(*)$  folgt aufgrund der Konstruktion der Matrix  $T$ .

Nach den Anweisungen in Schritt 1 berechnet sich  $\lambda_{ij}^1$  als

$$\lambda_{ij}^1 = \rho_{ij} \sqrt{\lambda_i \lambda_j}.$$

Zudem gilt

$$\begin{aligned}
 \text{Var}(Y_i) \text{Var}(Y_j) & = \text{Var} \left( \sum_{\ell=1}^{\tau+n} \mathbb{1}_{S_{\ell}}(i) X_{\ell}(\beta_{\ell}) \right) \text{Var} \left( \sum_{k=1}^{\tau+n} \mathbb{1}_{S_k}(j) X_k(\beta_k) \right) \\
 & = \sum_{\ell=1}^{\tau+n} \mathbb{1}_{S_{\ell}}(i) \text{Var}(X_{\ell}(\beta_{\ell})) \sum_{k=1}^{\tau+n} \mathbb{1}_{S_k}(j) \text{Var}(X_k(\beta_k)) \\
 & \stackrel{\text{Vor. (iii)}}{=} \sum_{\ell=1}^{\tau+n} \mathbb{1}_{S_{\ell}}(i) \eta \beta_{\ell} \sum_{k=1}^{\tau+n} \mathbb{1}_{S_k}(j) \eta \beta_k \\
 & \stackrel{(7.3)}{=} \eta^2 \lambda_{ii}^1 \lambda_{jj}^1 \\
 & = \eta^2 \lambda_i \lambda_j, \tag{7.5}
 \end{aligned}$$

so dass die Korrelation von  $Y_i$  und  $Y_j$  für alle  $i, j \in \{1, \dots, n\}$  durch

$$\begin{aligned}
 \text{Corr}(Y_i, Y_j) & = \frac{1}{\sqrt{\text{Var}(Y_i) \text{Var}(Y_j)}} \text{Cov}(Y_i, Y_j) \\
 & \stackrel{(7.5), (7.4)}{=} \frac{1}{\sqrt{\eta^2 \lambda_i \lambda_j}} \eta \lambda_{ij}^1 \\
 & \stackrel{(7.5)}{=} \frac{1}{\sqrt{\eta^2 \lambda_i \lambda_j}} \eta \rho_{ij} \sqrt{\lambda_i \lambda_j} \\
 & = \rho_{ij}
 \end{aligned}$$

gegeben ist. □

Nun stellt sich die Frage, unter welchen Bedingungen der Algorithmus in Schritt 5 ohne Lösung abbricht. In Park und Shin (1998) steht dazu:

„When the algorithm fails, there are two possibilities. One is that the given correlation structure is inconsistent such as non-positive definite structure or infeasible structure [...]. The other lies in the algorithm itself. We should admit that the algorithm can fail even for feasible correlation structures in the extreme cases of strongly correlated structures with different means.“<sup>4</sup>

Mit Hilfe des vorgestellten Algorithmus können für alle anderen Fälle korrelierte Zählraten mit dem geforderten Erwartungswert und der geforderten Korrelation erzeugt werden. Dies wird erreicht, indem Realisierungen des Zufallsvektors  $X$  erzeugt und anschließend mit der Matrix  $T$  multipliziert werden, so dass  $Y = TX$  gilt. Wie in der Einleitung dieses Kapitels bereits beschrieben, werden jetzt Daten zweier Behandlungsgruppen betrachtet. Dabei wird von folgendem marginalen Modell ausgegangen:

- Für  $i \in \{1, \dots, m\}$  und  $j \in \{1, \dots, n\}$  sind  $Y_{ij}$  poissonverteilte Zufallsvariablen.
- Die marginale Erwartung  $E(Y_{ij}) = \lambda_{ij}$  und die erklärenden Variablen hängen wie folgt voneinander ab:

$$\ln(E(Y_{ij})) = \ln(\lambda_{ij}) = \begin{cases} \beta_0 + \beta_1, & \text{falls Placebo,} \\ \beta_0, & \text{falls Verum.} \end{cases}$$

- Die Korrelationsmatrix des Zufallsvektors  $Y_i = (Y_{i1}, \dots, Y_{in})^\top$  ist für alle  $i \in \{1, \dots, m\}$  gegeben durch

$$[R(\alpha)_{jk}]_{\substack{k \in \{1, \dots, n\} \\ j \in \{1, \dots, n\}}} = \begin{cases} 1, & \text{falls } k = j, \\ \alpha, & \text{falls } k \neq j. \end{cases}$$

Der benötigte Stichprobenumfang  $m$  kann durch das Festsetzen des Signifikanzniveaus  $\alpha^*$ , der Power  $1 - \tilde{\beta}$ , der Anzahl an Wiederholungen  $n$  pro Subjekt, des Korrelationsparameters  $\alpha$ , der kleinsten bedeutungsvollen Differenz  $d = \beta_1$  sowie der Angabe von  $\beta_0$  mit Hilfe der Formel in Gleichung (6.10) berechnet werden.<sup>5</sup> Basierend auf diesen Größen sind für die Simulation von Daten zweier Behandlungsgruppen zwei R-Programme im Anhang I gegeben. Dabei unterscheiden sich die beiden Programme nur darin, dass `simdatn1` lediglich für  $n = 1$  und `simdat` für  $n > 1$  genutzt werden kann.<sup>6</sup> Die erzeugten Datensätze enthalten dann vier Variablen:

<sup>4</sup>Park und Shin (1998), S. 135.

<sup>5</sup>Dabei werden jeder Behandlungsgruppe  $\lceil \frac{m}{2} \rceil$  Subjekte zugeordnet.

<sup>6</sup>Für weitere Details wird auf den Anhang I verwiesen.

- die Subjektnummer zwischen 1 und  $m$
- die Codierung der Behandlung, z.B. 0 für Placebo und 1 für Verum,
- die Nummer der Wiederholung zwischen 1 und  $n$
- die Realisierung der Zufallsvariablen  $Y_{ij}$  mit  $E(Y_{ij}) = \beta_0$  falls Verum bzw.  $E(Y_{ij}) = \beta_0 + \beta_1$  falls das Placebo eingesetzt wurde, und  $\text{Corr}(Y_i) = R(\alpha)$ .

Diese Variablen werden benötigt, um eine Auswertung mit Hilfe der SAS-Prozedur GENMOD vornehmen zu können.

## 7.2 Auswertung der simulierten Studien

Zur Auswertung der simulierten Datensätze wurde ein SAS-Makro geschrieben, das es ermöglicht, die geschätzten Parameter und die Power für unterschiedliche Parameterkonstellationen mit Hilfe der Prozedur GENMOD zu berechnen. Dieses Makro ist in Anhang J gegeben.

Mit Hilfe des Makros und der Programme zur Simulation der Daten<sup>7</sup> werden für jede betrachtete Parameterkonstellation 1000 Datensätze erzeugt und im Anschluss ausgewertet. Für jede Kombination der Parameter werden die Schätzer für die Regressionsparameter ( $\beta_0$  und  $d = \beta_1$ ), die Standardabweichungen ( $sd\beta_0$  und  $sd\beta_1$ ) und auch die Konfidenzintervallgrenzen über die 1000 simulierten Studien gemittelt. Darüber hinaus wird die Power ausgegeben, indem berechnet wird, in wie vielen der 1000 Studien die Nullhypothese  $H_0 : \beta_1 = 0$  zum Signifikanzniveau  $\alpha^* = 0.05$  abgelehnt wird. Die Zahl wird im Anschluss durch 1000 geteilt.

In der folgenden Tabelle sind nun die Parameterkombinationen, die geschätzten und gemittelten Parameter und insbesondere die Power für den Fall einer Beobachtung pro Subjekt aufgelistet. Dabei gilt bei allen Simulationen  $\phi = 1$ .<sup>8</sup>

Nr.	$n$	$\alpha$	$\beta_0$	$d = \beta_1$	$m$	$\hat{\beta}_0$	$sd\beta_0$	$\hat{\beta}_1$	$sd\beta_1$	Power
1	1	-	1	$\ln(2)$	24	0.984	0.168	0.701	0.207	0.915
2	1	-	1	$\ln(1.5)$	78	0.992	0.096	0.415	0.124	0.903
3	1	-	1	$\ln(1.1)$	1624	0.999	0.021	0.095	0.029	0.884
4	1	-	0.5	$\ln(1.5)$	128	0.499	0.097	0.401	0.125	0.901
Fortsetzung auf der nächsten Seite.										

---

<sup>7</sup>Vgl. Anhang I.

<sup>8</sup>Es war im Rahmen dieser Arbeit nicht möglich, korrelierte poissonverteilte Daten mit Überdispersion zu erzeugen, so dass für alle simulierten Datensätze  $\phi = 1$  gilt.

Nr.	$n$	$\alpha$	$\beta_0$	$d = \beta_1$	$m$	$\hat{\beta}_0$	$sd\beta_0$	$\hat{\beta}_1$	$sd\beta_1$	Power
5	1	-	1.5	$\ln(4)$	4	1.464	0.151	1.419	0.184	0.999
6	1	-	1.5	$\ln(3)$	6	1.456	0.208	1.142	0.248	0.975
7	1	-	1.5	$\ln(1.5)$	48	1.493	0.088	0.409	0.122	0.918
8	1	-	2	$\ln(1.5)$	30	1.999	0.090	0.403	0.117	0.920

Tabelle 7.1: Ergebnisse der simulierten Studien für  $n = 1$ .

Bei diesen Ergebnissen wird deutlich, dass die Prozedur GENMOD die simulierten Daten mit einer hohen Genauigkeit auswertet. Alle geschätzten und gemittelten Parameter liegen sehr nahe am wahren benutzten Wert. Darüber hinaus ergibt sich bei der Auswertung in allen Studien, mit Ausnahme der dritten Studie, eine Power von  $\geq 0.9$ .

Die Daten der dritten Studie wurden so simuliert, dass der zu erkennende multiplikative Behandlungseffekt durch den Faktor  $\exp(\beta_1) = \exp(d) = \exp(\ln(1.1)) = 1.1$  gegeben war. Um bei einer Beobachtung pro Person und  $\beta_0 = 1$  eine Power von  $1 - \tilde{\beta} = 0.9$  zu erhalten, ergab sich mit der Formel (6.10) ein benötigter Stichprobenumfang von  $m = 1624$  Subjekten. Nun kommt es bei der Auswertung der Daten lediglich zu einer Power von 0.884. Dies kann einerseits an einer zu geringen Anzahl von Simulationen (1000) liegen. Andererseits ist zu beachten, dass die Fallzahl basierend auf der Teststatistik (6.2) berechnet wurde, GENMOD allerdings die Hypothesen mit Hilfe des *Z-Tests* testet.

Für mehr als eine Beobachtung pro Person ergeben sich bei festem  $\beta_0 = 1,5$  und  $\phi = 1$  entsprechend die folgenden Größen:

Nr.	$n$	$\alpha$	$d = \beta_1$	$m$	$\hat{\beta}_0$	$sd\beta_0$	$\hat{\beta}_1$	$sd\beta_1$	Power
9	2	0.2	$\ln(2)$	10	1.488	0.139	0.701	0.175	0.958
10	2	0.4	$\ln(2)$	10	1.475	0.149	0.708	0.187	0.930
11	2	0.6	$\ln(2)$	12	1.485	0.152	0.701	0.189	0.939
12	2	0.8	$\ln(2)$	14	1.480	0.153	0.703	0.190	0.934
13	2	0.2	$\ln(1.1)$	592	1.4997	0.021	0.095	0.029	0.914
14	2	0.4	$\ln(1.1)$	690	1.4996	0.021	0.096	0.029	0.907
15	2	0.6	$\ln(1.1)$	788	1.499	0.022	0.096	0.029	0.914
16	2	0.8	$\ln(1.1)$	888	1.4995	0.212	0.096	0.029	0.914
17	2	0.2	$\ln(1.5)$	30	1.499	0.089	0.403	0.116	0.925
18	2	0.4	$\ln(1.5)$	34	1.492	0.092	0.414	0.119	0.926

Fortsetzung auf der nächsten Seite.

Nr.	$n$	$\alpha$	$d = \beta_1$	$m$	$\hat{\beta}_0$	$sd\beta_0$	$\hat{\beta}_1$	$sd\beta_1$	Power
19	2	0.6	ln(1.5)	38	1.499	0.093	0.406	0.121	0.909
20	2	0.8	ln(1.5)	44	1.495	0.093	0.412	0.120	0.915
21	5	0.2	ln(2)	6	1.479	0.121	0.706	1.556	0.959
22	5	0.4	ln(2)	8	1.482	0.137	0.707	0.173	0.952
23	5	0.6	ln(2)	10	1.491	0.145	0.697	0.183	0.938
24	5	0.8	ln(2)	12	1.489	0.155	0.697	0.193	0.925
25	5	0.2	ln(1.1)	356	1.499	0.022	0.096	0.029	0.917
26	5	0.4	ln(1.1)	514	1.501	0.021	0.095	0.029	0.892
27	5	0.6	ln(1.1)	670	1.499	0.021	0.095	0.029	0.908
28	5	0.8	ln(1.1)	828	1.501	0.021	0.094	0.029	0.897
29	5	0.2	ln(1.5)	18	1.493	0.087	0.412	0.113	0.921
30	5	0.4	ln(1.5)	26	1.497	0.089	0.405	0.117	0.916
31	5	0.6	ln(1.5)	32	1.489	0.093	0.414	0.121	0.916
32	5	0.8	ln(1.5)	40	1.494	0.094	0.409	0.121	0.910
33	7	0.2	ln(2)	6	1.482	0.113	0.705	0.146	0.975
34	7	0.4	ln(2)	8	1.494	0.129	0.691	0.164	0.955
35	7	0.6	ln(2)	10	1.484	0.143	0.704	0.181	0.936
36	7	0.8	ln(2)	12	1.483	0.154	0.694	0.193	0.901
37	7	0.2	ln(1.1)	310	1.501	0.021	0.095	0.029	0.907
38	7	0.4	ln(1.1)	480	1.501	0.021	0.094	0.029	0.906
39	7	0.6	ln(1.1)	648	1.499	0.021	0.094	0.029	0.896
40	7	0.8	ln(1.1)	818	1.500	0.021	0.095	0.029	0.900
41	7	0.2	ln(1.5)	16	1.498	0.084	0.402	0.111	0.931
42	7	0.4	ln(1.5)	24	1.493	0.088	0.411	0.116	0.923
43	7	0.6	ln(1.5)	32	1.497	0.091	0.406	0.119	0.909
44	7	0.8	ln(1.5)	40	1.497	0.092	0.404	0.120	0.904
45	14	0.2	ln(2)	4	1.492	0.095	0.689	0.127	0.954
46	14	0.4	ln(2)	8	1.491	0.127	0.697	0.161	0.956
47	14	0.6	ln(2)	10	1.486	0.141	0.702	0.178	0.944
48	14	0.8	ln(2)	12	1.487	0.155	0.698	0.193	0.916
49	14	0.2	ln(1.1)	254	1.4999	0.021	0.096	0.029	0.902
50	14	0.4	ln(1.1)	438	1.4998	0.021	0.094	0.029	0.896
51	14	0.6	ln(1.1)	620	1.5003	0.021	0.095	0.029	0.890
Fortsetzung auf der nächsten Seite.									

Nr.	$n$	$\alpha$	$d = \beta_1$	$m$	$\hat{\beta}_0$	$sd\beta_0$	$\hat{\beta}_1$	$sd\beta_1$	Power
52	14	0.8	ln(1.1)	804	1.4999	0.021	0.095	0.029	0.896
53	14	0.2	ln(1.5)	14	1.495	0.079	0.408	0.105	0.948
54	14	0.4	ln(1.5)	22	1.498	0.087	0.404	0.115	0.930
55	14	0.6	ln(1.5)	30	1.496	0.092	0.407	0.119	0.910
56	14	0.8	ln(1.5)	40	1.492	0.092	0.409	0.120	0.918
57	21	0.2	ln(2)	4	1.479	0.087	0.711	0.119	0.980
58	21	0.4	ln(2)	8	1.488	0.122	0.702	0.155	0.972
59	21	0.6	ln(2)	10	1.486	0.141	0.707	0.177	0.961
60	21	0.8	ln(2)	12	1.485	0.151	0.697	0.189	0.927
61	21	0.2	ln(1.1)	236	1.501	0.021	0.094	0.029	0.902
62	21	0.4	ln(1.1)	424	1.500	0.021	0.094	0.029	0.892
63	21	0.6	ln(1.1)	610	1.498	0.021	0.097	0.029	0.910
64	21	0.8	ln(1.1)	798	1.500	0.021	0.095	0.029	0.895
65	21	0.2	ln(1.5)	12	1.496	0.081	0.408	0.107	0.929
66	21	0.4	ln(1.5)	22	1.497	0.087	0.404	0.113	0.924
67	21	0.6	ln(1.5)	30	1.497	0.091	0.409	0.118	0.934
68	21	0.8	ln(1.5)	38	1.497	0.094	0.411	0.122	0.909
69	28	0.2	ln(2)	4	1.486	0.087	0.706	0.118	0.979
70	28	0.4	ln(2)	6	1.489	0.124	0.699	0.161	0.927
71	28	0.6	ln(2)	10	1.489	0.139	0.687	0.175	0.946
72	28	0.8	ln(2)	12	1.488	0.151	0.693	0.189	0.926
73	28	0.2	ln(1.1)	226	1.4998	0.021	0.095	0.029	0.906
74	28	0.4	ln(1.1)	416	1.499	0.021	0.095	0.020	0.897
75	28	0.6	ln(1.1)	606	1.499	0.021	0.096	0.029	0.899
76	28	0.8	ln(1.1)	796	1.500	0.021	0.095	0.029	0.893
77	28	0.2	ln(1.5)	12	1.498	0.077	0.405	0.102	0.939
78	28	0.4	ln(1.5)	20	1.495	0.089	0.409	0.118	0.910
79	28	0.6	ln(1.5)	30	1.496	0.091	0.411	0.119	0.911
80	28	0.8	ln(1.5)	38	1.496	0.094	0.409	0.122	0.898

Tabelle 7.2: Ergebnisse der simulierten Studien für  $n > 1$ 

Bei zwölf der 72 durchgeführten Simulationen mit mehr als einer Beobachtung pro Person ergibt sich eine Power  $1 - \tilde{\beta}$  von  $0.89 \leq 1 - \tilde{\beta} < 0.9$ . In allen anderen Fällen wird die geforderte Power von 0.9 erreicht, und auch die gemittelten Schätzer der

Regressionsparameter unterscheiden sich unwesentlich von den wahren genutzten Parametern. Des Weiteren liegt die Streuung der Schätzer in allen Simulationen im üblichen Rahmen.

Werden die Fälle verglichen, in denen eine Power von 0.9 nicht erreicht wird, fällt auf, dass elf dieser zwölf Fälle bei  $\beta_1 = \ln(1.1)$  aufgetreten sind. Daher ist anscheinend die unzureichende Power bei einem geringen Behandlungsunterschied nicht nur auf numerische Probleme zurückzuführen. Vielmehr ist es naheliegend, dass die Fallzahlformel für geringe Behandlungsunterschiede Schwächen aufweist.

Der zwölfte Fall ist bei der Auswertung der Datensätze der Simulation Nr. 80 aufgetreten. Hier ergibt sich eine Power von 0.898, so dass die geforderte Power vermutlich bei einer höheren Anzahl an Simulationsstudien (z.B. 2000) erreicht wird. Das Gleiche könnte natürlich auch für die restlichen elf Fälle gelten.

Insgesamt sind die Ergebnisse zufriedenstellend. Die Prozedur schätzt mit einer hohen Genauigkeit die wahren Parameter der erzeugten Datensätze. Darüber hinaus kommt es bei 83.75% aller durchgeführten Simulationen zu der gewünschten Power. Die restlichen 16.25% erfüllen immerhin auch eine Power von  $0.89 \leq 1 - \tilde{\beta} < 0.9$ . Daher erscheint die in Kapitel 6 berechnete Fallzahlformel richtig und kompatibel mit dem Z-Test.<sup>9</sup>

---

<sup>9</sup>Wie bereits erwähnt, wurde die Fallzahlformel (6.10) basierend auf der T-Statistik (6.2) berechnet. Allerdings ermittelt GENMOD die p-Werte basierend auf dem Z-Test.

# Kapitel 8

## Zusammenfassung und Ausblick

Das Ziel der vorliegenden Arbeit war es, Modellierungsmöglichkeiten und Aspekte der Versuchsplanung für wiederholte Zählraten zu erläutern. In den vorangegangenen sieben Kapiteln wurde die Modellierung von wiederholten Daten zunächst durch Beispiele motiviert und die Eigenschaften *marginaler* und *subjekt-spezifischer Modelle* erörtert.

In diesem Zusammenhang wurden in Kapitel 3 die *marginalen Modelle* charakterisiert. Der Regressionszusammenhang zwischen dem Beobachtungsvektor und den erklärenden Variablen wurde durch ein verallgemeinertes lineares Modell beschrieben. Die Abhängigkeitsstruktur wurde separat durch die Korrelation modelliert.

Die Analyse marginaler Modelle führte zu den von Liang und Zeger (1986) vorgestellten Verallgemeinerten Schätzgleichungen (*generalized estimating equations*, GEE), die das Schätzen des gesuchten Regressionsvektors ermöglichen. In Abschnitt 3.1.1.1 wurde erläutert, wie diese Schätzgleichungen durch die Einführung einer sogenannten *Arbeitskorrelationsmatrix* entstehen. Es wurde dargestellt, dass ein Schätzer für den gesuchten Regressionsvektor mit Hilfe einer leicht veränderten Form des *Fisher-Scoring-Verfahrens* ermittelt werden kann.

Aufgrund fehlender Hinweise für den Beweis der Eigenschaften dieses GEE-Schätzers wurden im Anschluss mit Hilfe der Arbeit von Ziegler (1999) die *Pseudo-Maximum-Likelihood-Methode* (PML-Methode) und die *Quasi-Generalisierte-Pseudo-Maximum-Likelihood-Methode* (QGPML-Methode) betrachtet. Es konnte für randomisierte Studien die Äquivalenz der QGPML- und GEE-Schätzer dargelegt werden. Des Weiteren wurde gezeigt, dass die GEE-Schätzer konsistent und asymptotisch normalverteilt sind. Außerdem wurde diskutiert, inwiefern diese Schätzer von der Wahl der Arbeitskorrelationsmatrix abhängen. Es wurden die modellbasierte und die em-

pirische Schätzung der Kovarianzmatrix des Regressionsparameters betrachtet und anhand dieser Schätzungen Gütekriterien für die Wahl einer Arbeitskorrelationsstruktur aufgestellt.

Außerdem wurden zwei marginale Modelle mit der Prozedur `GENMOD` des Statistik-Programms `SAS` ausgewertet. Die Analyse wurde für unterschiedliche Abhängigkeitsstrukturen (Unabhängigkeit, Austauschbarkeit, Autoregression erster Ordnung) vorgenommen und gezeigt, dass sich die Resultate nur minimal voneinander unterscheiden. Auch wurde auf Besonderheiten im Umgang mit dieser Prozedur hingewiesen und die Qualität der Auswertung anhand von Simulationsstudien in Kapitel 7 untersucht.

In Kapitel 6 wurde dann auf das Themenfeld der Versuchsplanung für marginale Modelle eingegangen. Zunächst wurden die Einflussgrößen zur Berechnung des Stichprobenumfangs im Fall wiederholter Daten aufgestellt, um im Anschluss, angelehnt an die Vorgehensweise von Liu und Liang (1997), eine Fallzahlformel für den Fall wiederholter Zählraten zu bestimmen. Diese Fallzahlformel wurde dann in Abhängigkeit verschiedener Parameter, wie etwa der Überdispersion oder der Korrelation, untersucht. In diesem Zusammenhang wurde deutlich, dass sich der Stichprobenumfang im Fall einer *exchangeable*-Korrelationsstruktur linear zu der Überdispersion und der Korrelation entwickelt. Darüber hinaus fiel auf, dass es ab einer gewissen Anzahl an wiederholten Beobachtungen pro Subjekt zu keiner nennenswerten Reduzierung des benötigten Stichprobenumfangs, d.h. der benötigten Anzahl an Subjekten, kommt.

Die erhaltene Formel zur Berechnung des benötigten Stichprobenumfangs wurde dann in Kapitel 7 auf ihre Güte untersucht. Anhand von 80 durchgeführten Simulationen mit jeweils 1000 simulierten Datensätzen wurde gezeigt, dass die geforderte Power von 0.9 in 83.75% der Simulationen eingehalten wurde. Für die restlichen 16.25% der Simulationen wurde eine Power zwischen 0.89 und 0.9 erreicht. Dabei wurde die Power oft bei Studien mit einem sehr geringen Behandlungsunterschied nicht eingehalten. Daher liegt die Vermutung nahe, dass die berechnete Fallzahlformel für geringe Behandlungsunterschiede Schwächen aufweist.

Neben den marginalen Modellen wurden in Kapitel 3 die subjekt-spezifischen Modelle eingeführt. Bei der Charakterisierung der *subjekt-spezifischen Modelle* wurde erneut ein verallgemeinertes lineares Modell aufgestellt. Im Gegensatz zum Vorgehen bei marginalen Modellen wurden die erklärenden Variablen jedoch durch zusätzliche subjekt-spezifische Größen ergänzt. Diese spiegeln die natürliche Heterogenität der

---

Subjekten wider, die auf nicht-messbaren Faktoren beruht. Die Berücksichtigung der subjekt-spezifischen Größen ermöglicht zudem die Modellierung der Abhängigkeiten zwischen den wiederholten Daten eines Subjekts, da alle Beobachtungen eines Subjekts diese Größe gemein haben.

Für die Analyse des subjekt-spezifischen Modells wurden drei Handhabungen der subjekt-spezifischen Größen erläutert. Diese können als feste unbekannte Parameter, als Störparameter oder als zufällige Parameter betrachtet werden. In dieser Arbeit konzentrierten sich die Untersuchungen auf zufällige subjekt-spezifische Größen, so dass die Theorie der subjekt-spezifischen Modelle im Wesentlichen der Theorie verallgemeinerter linearer Modelle mit zufälligen Effekten entspricht. Solch ein Modell ist aber gerade ein *verallgemeinertes lineares gemischtes Modell* (*generalized linear mixed models*, GLMM). Basierend auf der Verteilungsannahme, dass die Beobachtungen bedingt auf den subjekt-spezifischen Größen poissonverteilt sind, wurde die Maximum-Likelihood-Funktion aufgestellt.

Um ein Maximum dieser Funktion zu ermitteln, mussten zunächst Integrale über mehrdimensionale Größen und im Anschluss das Maximum mit Hilfe von Optimierungsverfahren, wie etwa dem *Newton-Raphson-Verfahren*, berechnet werden. Zur Bestimmung der Integrale wurden zwei Möglichkeiten genannt. Die eine basiert auf der Approximation der Integrale und die andere auf der Approximation der Daten. In dieser Arbeit wurden lediglich zwei Methoden zur Approximation der Integrale vorgestellt, nämlich die *Hermite-Gauß-Quadratur* und die *adaptive Hermite-Gauß-Quadratur* (vgl. Abschnitt 3.2.2.1) und anschließend hinsichtlich der Genauigkeit der Approximation verglichen. Im Rahmen dieser Untersuchungen wurde deutlich, dass für nicht-symmetrische Funktionen die adaptive Hermite-Gauß-Quadratur im Allgemeinen bessere Resultate liefert und daher für diese Fälle genutzt werden sollte.

Darüber hinaus wurde die Prozedur `NLMIXED` des Statistik-Programms `SAS` zur Auswertung subjekt-spezifischer Modelle vorgestellt, um mit dieser Prozedur anschließend in Kapitel 5 eine Studie auszuwerten. In diesem Kontext wurden die Auswirkungen von verschiedenen Anzahlen an Quadraturpunkten auf die Resultate der Analyse betrachtet. Im Gegensatz zu Untersuchungen von Molenberghs und Verbeke (2005) unterschieden sich die Ergebnisse nur in geringem Maße. Zudem wurden die Ergebnisse der Analyse mittels herkömmlicher Hermite-Gauß-Quadratur und adaptiver Hermite-Gauß-Quadratur verglichen: Dabei wurden gravierende Unterschiede in den Resultaten deutlich. Diese scheinen insbesondere auf Schwächen der Prozedur `NLMIXED` zu beruhen.

Neben den separaten Untersuchungen an den beiden Modellfamilien wurden diese in Kapitel 4 und z.T. in Kapitel 5 auch verglichen. Anhand einer Beispielstudie wurde erläutert, dass sich die Interpretationen der Regressionsparameter in beiden Modellen grundsätzlich voneinander unterscheiden. Zwar ermöglicht die Analyse beider Modelle Aussagen über die durchschnittliche Entwicklung einer Population, doch lediglich die Resultate eines marginalen Modells erlauben eine unmittelbare Interpretation. Es wurde dargelegt, dass im Fall subjekt-spezifischer Modelle zusätzliche Berechnungen notwendig sind. Diese wurden daher für die Beispielstudie und das Beispiel in Kapitel 5 durchgeführt. Im Anschluss konnten die resultierenden Entwicklungen beider Modellfamilien gegenübergestellt werden. In beiden Studien unterschieden sich die durchschnittlichen Entwicklungen voneinander und es wurde deutlich, dass sich die Resultate in Abhängigkeit der Variabilität der subjekt-spezifischen Größen unterscheiden.

Aus den vorangegangenen Kapiteln ist deutlich geworden, dass es – im Gegensatz zu Studien mit binären oder normalverteilten Daten – keine speziellen Modelle für die Auswertung wiederholter Zähldaten gibt. Es wurde daher der allgemeine Ansatz der GEE im Fall marginaler Modelle und die Theorie der GLMM im Fall subjekt-spezifischer Modelle vorgestellt. Welches der Modelle in welchem Fall angewandt wird, hängt nun im Wesentlichen von der Fragestellung der Studie ab. Interessiert der Populationsdurchschnitt, wie bei der Untersuchung der durchschnittlichen Hitzewallungsanzahl in einem bestimmten Zeitraum, sind beide Modelle geeignet. Folgende Aspekte könnten die Auswahl zwischen den Modellfamilien erleichtern:

(1) Die Modellierung marginaler Modelle ist recht unkompliziert und bedarf lediglich der Festlegung einer Arbeitskorrelationsstruktur. Unabhängig von dieser Wahl liefert die Auswertung mittels Verallgemeinerter Schätzgleichungen einen konsistenten Schätzer für den Regressionsvektor. Allerdings muss beachtet werden, dass eine unpassende Abhängigkeitsstruktur eine Einschränkung der Effizienz bedingen kann. Mit unpassend ist z.B. gemeint, dass bei einer Studie, in deren Rahmen über einen längeren Zeitraum die Hitzewallungsanzahl beobachtet wurde, die Annahme der Austauschbarkeit sicherlich nicht erfüllt sein wird. Im Gegensatz dazu ist diese Struktur realistisch, wenn die Cluster Krankenhäuser sind und die Beobachtungen innerhalb eines Clusters von unterschiedlichen Patienten stammen.

Darüber hinaus ist durch die Prozedur `GENMOD` eine leicht zu bedienende Implementierung für die Auswertung marginaler Modelle gegeben und mit der Arbeit von Liu

---

und Liang (1997) liegt eine Möglichkeit vor, den benötigten Stichprobenumfang zu berechnen.

(2) Soll nun aber die Heterogenität der Subjekte berücksichtigt werden, ist ein marginales Modell für die Modellierung nicht-messbarer Faktoren unzureichend. In diesem Fall sollte daher ein subjekt-spezifisches Modell angepasst werden. Interpretationen über den Populationsdurchschnitt sind in diesem Fall zwar nicht direkt aufgrund der geschätzten Regressionsparameter möglich, doch können im Gegensatz zu marginalen Modellen auch Aussagen über die Entwicklung einzelner Individuen getroffen werden. Dies kann insbesondere für das Aufspüren von Ausreißern von Bedeutung sein.

Die für die Auswertung subjekt-spezifischer Modelle benutzte SAS-Prozedur NLMIXED wies im Rahmen der Auswertungen in Kapitel 5 schwerwiegende Schwächen auf. Die geschätzten Parameter hingen von der Anzahl der Quadraturpunkte, von den gewählten Startwerten sowie von der Wahl einer adaptiven bzw. nicht-adaptiven Quadratur ab. Inwiefern die hier nicht vorgestellte SAS-Prozedur GLIMMIX zur Auswertung subjekt-spezifischer Modelle bessere Resultate liefert, bedarf Untersuchungen, die im Rahmen dieser Arbeit nicht geleistet werden konnten. Des Weiteren konnte keine Fallzahlformel für die subjekt-spezifischen Modelle bestimmt werden.

Abschließend soll nun erörtert werden, welche Fragen im Rahmen dieser Arbeit nicht beantwortet wurden.

Wie bereits erwähnt, konnte aufgrund fehlender Literatur zu diesem Thema keine Fallzahlformel für eine Auswertung mittels subjekt-spezifischer Modelle präsentiert werden. Da eine solche Formel allerdings für das Aufstellen eines Studiendesigns unerlässlich ist, sollte eine entsprechende Formel hergeleitet oder in der Literatur auffindig gemacht werden.

Des Weiteren wurden die zufälligen subjekt-spezifischen Größen als normalverteilt vorausgesetzt. In diesem Zusammenhang ist es aber auch interessant, die Gamma-Verteilung und als deren Spezialfall die Exponentialverteilung zu betrachten. Versuche der Autorin dazu, die in der Arbeit nicht aufgeführt sind, zeigten zwar, dass die marginale Verteilung von einzelnen Beobachtungen gegeben einer gammaverteilten subjekt-spezifischen Größe einer Negativ-Binomialverteilung folgen, doch traten Probleme bei der Berechnung der marginalen Verteilung der mehrdimensionalen Beobachtungsvektoren auf. Somit sind in diesem Fall erneut numerische Approximationsmethoden gefordert. Es sei darauf hingewiesen, dass die Hermite-Gauß-

Quadratur aufgrund der veränderten Gewichtsfunktion hier nicht angewandt werden konnte.

Im Fall marginaler Modelle stellt sich darüber hinaus die Frage, ob die Einschränkung, dass die Korrelationsmatrizen für alle Subjekte gleich sind, sinnvoll ist (vgl. dazu Kapitel 6 und Kapitel 7). Inwiefern eine Fallzahlformel für unterschiedliche Korrelationsmatrizen aufgestellt werden kann, bedarf weiterer Erforschung. Auch muss geklärt werden, ob eine solche Betrachtung zu aufwendig ist und ob ein nennenswerter Beitrag zur Bestimmung besserer Resultate geliefert wird.

Selbstverständlich könnten auch Aspekte zur Robustheit beider Modellfamilien diskutiert werden.

Außerdem wäre es von großem Interesse, die Resultate einer marginalen und einer subjekt-spezifischen Analyse mit den herkömmlichen Methoden zu vergleichen. Diese Methoden bilden den Mittelwert der Beobachtungen eines Subjekts und wenden im Anschluss den *t-Test* oder den *Wilcoxon-Rangsummen-Test* an.

Die Frage nach dem Vergleich der Analysemethoden stellt sich insbesondere, da in Bereichen der Biometrie Methoden zur Auswertung wiederholter Daten noch kaum Einzug gefunden haben, und wenn doch, dann nur der sekundären Analyse dienen.<sup>1</sup> Dies erscheint verwunderlich, zumal das *Committee for Proprietary Medicinal Products* (CPMP) der *European Agency for the Evaluation of Medicinal Products* (EMA) die Analyse mittels Modelle für wiederholte Daten ausdrücklich zulässt.<sup>2</sup>

Wie diese ausstehenden Fragen belegen, ist der Themenbereich dieser Arbeit längst nicht ausgeschöpft.

---

<sup>1</sup>Vgl. Dahmen und Ziegler (2004).

<sup>2</sup>Vgl. CPMP (1998).

# Anhang



# Anhang A

## Orthogonale Polynome

In diesem Teil des Anhangs sollen sogenannte orthogonale Polynome definiert und deren Eigenschaften aufgezeigt werden.<sup>1</sup> Ist eine Gewichtsfunktion  $\phi$  auf  $[a, b] \subset \mathbb{R}$  vorgegeben, dann lässt sich auf dem Vektorraum

$$\mathcal{P} := \{p(x) | p(x) = a_0 + a_1x + a_2x^2 + \dots \text{ mit } a_i \in \mathbb{R} \forall i \in \mathbb{N}_0\}$$

aller Polynome mit Koeffizienten in  $\mathbb{R}$  das gewichtete Skalarprodukt  $\langle \cdot, \cdot \rangle_\phi$  definieren.

### Definition A.1.

Sei  $\phi : \mathbb{R} \rightarrow \mathbb{R}$  eine positive, stückweise stetige Funktion. Auf dem Vektorraum  $\mathcal{P} = \{p(x) | p(x) = a_0 + a_1x + a_2x^2 + \dots, \text{ mit } a_i \in \mathbb{R} \forall i \in \mathbb{N}_0\}$  aller Polynome über  $\mathbb{R}$  wird für  $g, h \in \mathcal{P}$  und  $a, b \in \mathbb{R}$  mit  $a < b$  das gewichtete Skalarprodukt  $\langle g, h \rangle_\phi$  durch

$$\langle g, h \rangle_\phi := \int_a^b g(x) f(x) \phi(x) dx$$

definiert.

Der folgende Satz lässt sich dann leicht beweisen.

### Satz A.2.

Durch  $\langle \cdot, \cdot \rangle_\phi$  ist ein Skalarprodukt auf dem Vektorraum aller Polynome definiert.

*Beweis:* Um nachzuprüfen, ob  $\langle \cdot, \cdot \rangle_\phi : \mathcal{P} \times \mathcal{P} \rightarrow \mathbb{R}$  ein Skalarprodukt ist, muss gezeigt werden, dass es sich um eine symmetrische, positiv definite Bilinearform handelt:

---

<sup>1</sup>Dieser Teil der Diplomarbeit ist in Anlehnung an die Ausarbeitungen auf der Internetseite <http://www.matha.rwth-aachen.de/lehre/ss06/anaprosem/Alexandra.Goeke.Ausarbeitung.pdf> entstanden.

(i) Positive Definitheit: Für  $g \in \mathcal{P} \setminus \{0\}$  folgt

$$\begin{aligned} \langle g, g \rangle_\phi &= \int_a^b g(x) g(x) \phi(x) dx \\ &= \int_a^b g^2(x) \phi(x) dx \\ &\geq 0, \end{aligned}$$

weil  $\phi$  eine positive Funktion ist und  $g \in \mathcal{P} \setminus \{0\}$ . Für  $g \equiv 0$  gilt offensichtlich  $\langle g, g \rangle_\phi = 0$ . Nun die Rückrichtung:

Sei  $g \in \mathcal{P}[a, b] := \{p(x) | p(x) = a_0 + a_1x + a_2x^2 + \dots, \text{ mit } a_i \in \mathbb{R} \forall i \in \mathbb{N}_0, x \in [a, b]\}$  mit  $g \neq 0$ . Dann gibt es ein  $x \in [a, b]$  mit  $g(x) \neq 0$ .

Da  $g$  stetig ist, gibt es ein  $\varepsilon > 0$  mit  $|g(x')| \geq \frac{1}{2} |g(x)|$  für alle  $x' \in ]x - \varepsilon, x + \varepsilon[ \cap [a, b] = \{x'' \in [a, b] \mid |x'' - x| < \varepsilon\}$ . Dann gilt jedoch

$$\begin{aligned} \langle g, g \rangle_\phi &= \int_a^b g(x)^2 \phi(x) dx \\ &= \int_{]x-\varepsilon, x+\varepsilon[ \cap [a, b]} g(t)^2 \phi(t) dt + \int_{[a, b] \setminus ]x-\varepsilon, x+\varepsilon[} g(t)^2 \phi(t) dt \\ &\geq \int_{]x-\varepsilon, x+\varepsilon[ \cap [a, b]} \frac{1}{2} |g(x)| \phi(x) dt + \int_{[a, b] \setminus ]x-\varepsilon, x+\varepsilon[} 0 dt \\ &\geq \varepsilon \cdot \frac{1}{2} |g(x)| \phi(x) \\ &> 0. \end{aligned}$$

(ii) Symmetrie: Die Symmetrie gilt aufgrund der Kommutativität der Multiplikation in  $\mathcal{P}$ .

(iii) Bilinearform :

- Seien  $f, g, h \in \mathcal{P}$ , dann berechnet sich das Skalarprodukt wie folgt

$$\begin{aligned} \langle f, g + h \rangle_\phi &= \int_a^b f(x) (g + h)(x) \phi(x) dx \\ &= \int_a^b f(x) (g(x) + h(x)) \phi(x) dx \\ &= \int_a^b f(x) g(x) \phi(x) dx + \int_a^b f(x) h(x) \phi(x) dx \\ &= \langle f, g \rangle_\phi + \langle f, h \rangle_\phi. \end{aligned}$$

- Andererseits gilt aufgrund der Symmetrie

$$\langle f + g, h \rangle_\phi = \langle f, h \rangle_\phi + \langle g, h \rangle_\phi.$$

- Zudem folgt für  $\alpha \in \mathbb{R}$

$$\begin{aligned}
\langle \alpha g, h \rangle_\phi &= \int_a^b (\alpha g)(x) h(x) \phi(x) dx \\
&= \int_a^b \alpha g(x) h(x) \phi(x) dx \\
&= \alpha \int_a^b g(x) h(x) \phi(x) dx \\
&= \alpha \langle g, h \rangle_\phi.
\end{aligned}$$

□

Die folgende Definition zeigt auf, wann zwei Polynome bezüglich des in Definition A.1 definierten Skalarprodukts orthogonal sind.

**Definition A.3.**

Die Polynome  $g, h \in \mathcal{P}$  mit  $g \neq h$  heißen *orthogonal*, wenn

$$\langle g, h \rangle_\phi = 0$$

*gilt.*

Im Rahmen der in Abschnitt 3.2.2.1 vorgestellten Hermite-Gauß-Quadratur ist die Gewichtsfunktion für eindimensionale subjekt-spezifische Größen durch

$$\phi(z) = \exp(-z^2)$$

gegeben. Es stellt sich jetzt die Frage, welche Polynome bezüglich des gewichteten Skalarprodukts  $\langle \cdot, \cdot \rangle_\phi$  orthogonal zueinander sind. Es wird sich im Folgenden zeigen, dass die *Hermite-Polynome* dieser Anforderung genügen.

**Definition A.4.**

Sei  $n \in \mathbb{N}_0$ , dann ist das  $n$ -te Hermite-Polynom  $H_n$  definiert durch

$$H_n(x) := (-1)^n e^{x^2} \frac{\partial^n e^{-x^2}}{\partial^n x}.$$

**Satz A.5.**

Die Hermite-Polynome entsprechen den Koeffizienten der Reihenentwicklung der Funktion

$$w(x, \alpha) := e^{2x\alpha - \alpha^2} = \sum_{n=0}^{\infty} \frac{\alpha^n}{n!} H_n(x), \quad x \in \mathbb{R}. \quad (\text{A.1})$$

Die Funktion  $w$  wird in diesem Zusammenhang auch *erzeugende Funktion der Hermite-Polynome* genannt.

*Beweis:* Sei  $f(x) = \exp(-x^2)$ , dann gilt

$$\begin{aligned}
 f'(x) &= -2x e^{-x^2} \\
 f''(x) &= (4x^2 - 2) e^{-x^2} \\
 &\vdots \\
 f^{(n)}(x) &\stackrel{Def.}{=} (-1)^n H_n(x) e^{-x^2}.
 \end{aligned} \tag{A.2}$$

Wenn  $f(x + \alpha) =: g(\alpha)$  als Funktion aufgefasst wird, die nur von  $\alpha$  abhängt, ist sie stetig differenzierbar nach  $\alpha$ . Mit Hilfe der Taylor-Entwicklung um den Entwicklungspunkt 0 gilt dann:

$$\begin{aligned}
 f(x - \alpha) &= g(-\alpha) \\
 &\stackrel{Taylor-Ent.}{=} \sum_{n=0}^{\infty} \frac{g^{(n)}(0)}{n!} (-\alpha - 0)^n \\
 &\stackrel{(*)}{=} \sum_{n=0}^{\infty} \frac{f^{(n)}(x)}{n!} (-\alpha)^n \\
 &= \sum_{n=0}^{\infty} (-1)^n \frac{\alpha^n}{n!} f^{(n)}(x) \\
 &\stackrel{vgl.(A.2)}{=} \sum_{n=0}^{\infty} \frac{\alpha^n}{n!} H_n(x) e^{-x^2},
 \end{aligned}$$

so dass sich

$$\begin{aligned}
 w(x, \alpha) &= e^{2x\alpha - \alpha^2} \\
 &= f(x - \alpha) e^{x^2} \\
 &= \sum_{n=0}^{\infty} \frac{\alpha^n}{n!} H_n(x) e^{-x^2} e^{x^2} \\
 &= \sum_{n=0}^{\infty} \frac{\alpha^n}{n!} H_n(x)
 \end{aligned}$$

ergibt. Dabei lässt sich die Gleichheit in (\*) erklären durch

$$\begin{aligned}
 g(\alpha) &= e^{-(x+\alpha)^2} \\
 g'(\alpha) &= -2(x + \alpha) e^{-(x+\alpha)^2} \\
 g''(\alpha) &= -2e^{-(x+\alpha)^2} + 4(x + \alpha)^2 e^{-(x+\alpha)^2} \\
 &= (4(x + \alpha)^2 - 2) e^{-(x+\alpha)^2} \\
 &\vdots \\
 g^{(n)}(\alpha) &= (-1)^n H_n(x + \alpha) e^{-(x+\alpha)^2}
 \end{aligned}$$

und

$$\begin{aligned} g^{(n)}(0) &= g^{(n)}(\alpha)|_{\alpha=0} \\ &= (-1)^n H_n(x) e^{-x^2} \\ &\stackrel{\text{vgl. (A.2)}}{=} f^{(n)}(x). \end{aligned}$$

□

Mit Hilfe der Funktion  $w$  wird es möglich sein, Sätze zu verifizieren, die später für den Beweis der Orthogonalität der Hermite-Polynome gebraucht werden.

**Satz A.6.**

Das  $n$ -te Hermite-Polynom lässt sich durch seine beiden Vorgänger rekursiv darstellen, denn für alle  $n \in \mathbb{N}$  gilt

$$H_{n+1}(x) = 2xH_n(x) - 2nH_{n-1}(x).$$

*Beweis:* Es gilt

$$\begin{aligned} \frac{\partial w(x, \alpha)}{\partial \alpha} &= (2x - 2\alpha) e^{2x\alpha - \alpha^2} \\ &= (2x - 2\alpha) w(x, \alpha) \end{aligned}$$

und somit ergibt sich

$$\frac{\partial w(x, \alpha)}{\partial \alpha} - (2x - 2\alpha) w(x, \alpha) = 0.$$

Nach Satz A.5 folgt also:

$$\frac{\partial}{\partial \alpha} \sum_{n=0}^{\infty} \frac{H_n(x)}{n!} \alpha^n - 2x \sum_{n=0}^{\infty} \frac{H_n(x)}{n!} \alpha^n + 2\alpha \sum_{n=0}^{\infty} \frac{H_n(x)}{n!} \alpha^n = 0.$$

Durch das Differenzieren der ersten Summe lässt sich der vorherige Ausdruck umformen zu

$$\sum_{n=1}^{\infty} \frac{H_n(x)}{(n-1)!} \alpha^{n-1} - 2x \sum_{n=0}^{\infty} \frac{H_n(x)}{n!} \alpha^n + 2 \sum_{n=0}^{\infty} \frac{H_n(x)}{n!} \alpha^{n+1} = 0$$

und außerdem ergibt sich mittels einer Indexverschiebung:

$$\sum_{n=0}^{\infty} \frac{H_{n+1}(x)}{n!} \alpha^n - 2x \sum_{n=0}^{\infty} \frac{H_n(x)}{n!} \alpha^n + 2 \sum_{n=1}^{\infty} \frac{H_{n-1}(x)}{(n-1)!} \alpha^n = 0.$$

Aufgrund des Aufbaus der Hermite-Polynome gilt  $H_0(x) = 1$  und  $H_1(x) = 2x$ .

Damit folgt aus der letzten Gleichung:

$$\underbrace{H_1(x) - 2xH_0(x)}_{=0} + \sum_{n=1}^{\infty} \frac{H_{n+1}(x)}{n!} \alpha^n - 2x \sum_{n=1}^{\infty} \frac{H_n(x)}{n!} \alpha^n + 2 \sum_{n=1}^{\infty} \frac{H_{n-1}(x)}{(n-1)!} \alpha^n = 0.$$

Dieser Ausdruck entspricht einem Polynom in der Variable  $\alpha$ , so dass sich durch Koeffizientenvergleich

$$\begin{aligned} \frac{H_{n+1}(x)}{n!} - 2x \frac{H_n(x)}{n!} + 2 \frac{H_{n-1}(x)}{(n-1)!} &= 0 \\ \Leftrightarrow H_{n+1}(x) - 2xH_n(x) + 2nH_{n-1}(x) &= 0 \end{aligned}$$

und somit die Behauptung ergibt. □

Eine zweite Rekursionsformel ermöglicht die Darstellung der ersten Ableitung des  $n$ -ten Hermite-Polynoms in Abhängigkeit von dem  $(n-1)$ -ten Polynom.

**Satz A.7.**

Für alle  $n \in \mathbb{N}$  gilt

$$H'_n(x) = 2nH_{n-1}(x).$$

*Beweis:* Durch das Differenzieren der erzeugenden Funktion nach  $x$  ergibt sich

$$\begin{aligned} \frac{\partial w(x, \alpha)}{\partial x} &= 2\alpha e^{2x\alpha - \alpha^2} = 2\alpha w(x, \alpha) \\ \Leftrightarrow \frac{\partial w(x, \alpha)}{\partial x} - 2\alpha w(x, \alpha) &= 0 \\ \Leftrightarrow \sum_{n=1}^{\infty} \frac{H'_n(x)}{n!} \alpha^n - 2 \sum_{n=0}^{\infty} \frac{H_n(x)}{n!} \alpha^{n+1} &= 0 \quad (\text{mit } H'_0(x) = 0) \\ \Leftrightarrow \sum_{n=1}^{\infty} \frac{H'_n(x)}{n!} \alpha^n - 2 \sum_{n=1}^{\infty} \frac{H_{n-1}(x)}{(n-1)!} \alpha^n &= 0. \end{aligned}$$

Ein Koeffizientenvergleich liefert für alle  $n \in \mathbb{N}$ :

$$\begin{aligned} \frac{H'_n(x)}{n!} - 2 \frac{H_{n-1}(x)}{(n-1)!} &= 0 \\ \Leftrightarrow H'_n(x) - 2nH_{n-1}(x) &= 0 \end{aligned}$$

und damit die Behauptung. □

**Satz A.8.**

Die Funktion

$$u(x) = e^{-x^2/2} H_n(x)$$

löst die Differentialgleichung

$$u''(x) + (2n + 1 - x^2) u(x) = 0.$$

*Beweis:* Einsetzen der Funktion  $u(x) = e^{-x^2/2}H_n(x)$  in die gegebene Differentialgleichung liefert

$$\begin{aligned}
& u''(x) + (2n + 1 + x^2) u(x) \\
= & \left[ e^{-x^2/2} H_n(x) \right]'' + (2n + 1 - x^2) e^{-x^2/2} H_n(x) \\
= & \left[ -e^{-x^2/2} + x^2 e^{-x^2/2} \right] H_n(x) - 2x e^{-x^2/2} H_n'(x) \\
& + e^{-x^2/2} H_n''(x) + (2n + 1 - x^2) e^{-x^2/2} H_n(x) \\
= & 2n e^{-x^2/2} H_n(x) - 2x e^{-x^2/2} H_n'(x) + e^{-x^2/2} H_n''(x) \\
\stackrel{\text{Satz A.7}}{=} & e^{-x^2/2} [2n H_n(x) - 4nx H_{n-1}(x) + 4n^2 H_{n-2}(x)] \\
= & 2n e^{-x^2/2} [H_n(x) - 2x H_{n-1}(x) + 2n H_{n-2}(x)] \\
\stackrel{\text{Satz A.6}}{=} & 0
\end{aligned}$$

und damit die Behauptung. □

Schließlich kann jetzt mit Hilfe des vorherigen Satzes die Orthogonalität der Hermite-Polynome nachgewiesen werden.

**Satz A.9.**

*Die Hermite-Polynome sind bezüglich des gewichteten Skalarprodukts mit der Gewichtsfunktion  $\phi(x) = e^{-x^2}$  orthogonal zueinander, d.h. für  $n, m \in \mathbb{N}_0$  und  $n \neq m$  gilt*

$$\langle H_n(x), H_m(x) \rangle_\phi = \int_{-\infty}^{\infty} H_n(x) H_m(x) e^{-x^2} dx = 0.$$

*Beweis:* Seien  $u_n(x) = e^{-x^2/2}H_n(x)$  und  $u_m(x) = e^{-x^2/2}H_m(x)$ . Dann liefert Satz A.8

$$u_n''(x) + (2n + 1 - x^2) u_n(x) = 0 \quad \text{und} \quad u_m''(x) + (2m + 1 - x^2) u_m(x) = 0.$$

Durch Multiplikation der ersten Gleichung mit  $u_m$ , der zweiten mit  $u_n$  und durch anschließende Subtraktion der zweiten von der ersten Gleichung gelangt man zu:

$$\begin{aligned}
& u_n'' u_m - u_m'' u_n + (2n - 2m) u_n u_m = 0 \\
\Leftrightarrow & \frac{\partial}{\partial x} \left( u_n' u_m - u_n u_m' \right) + (2n - 2m) u_n u_m = 0.
\end{aligned}$$

Die anschließende Integration über das Intervall  $[-\infty, \infty]$  liefert

$$\begin{aligned}
 & \lim_{t \rightarrow \infty} \left[ u_n' u_m - u_n u_m' \right]_{-t}^t + \int_{-\infty}^{\infty} (2n - 2m) u_n(x) u_m(x) dx = 0 \\
 \Leftrightarrow & \lim_{t \rightarrow \infty} \left[ \left( -x e^{-x^2/2} H_n(x) + e^{-x^2/2} H_n'(x) \right) e^{-x^2/2} H_m(x) \right. \\
 & \quad \left. - \left( -x e^{-x^2/2} H_m(x) + e^{-x^2/2} H_m'(x) \right) e^{-x^2/2} H_n(x) \right]_{-t}^t \\
 & \quad + 2(n - m) \int_{-\infty}^{\infty} u_n(x) u_m(x) dx = 0 \\
 \Leftrightarrow & \lim_{t \rightarrow \infty} \underbrace{\left[ e^{-x^2} \left( H_n'(x) H_m(x) - H_n(x) H_m'(x) \right) \right]_{-t}^t}_{\rightarrow 0} \\
 & \quad + 2(n - m) \int_{-\infty}^{\infty} u_n(x) u_m(x) dx = 0
 \end{aligned}$$

und somit die gewünschte Gleichung

$$\begin{aligned}
 & 2(n - m) \int_{-\infty}^{\infty} u_n(x) u_m(x) dx = 0 \\
 \Leftrightarrow & \int_{-\infty}^{\infty} u_n(x) u_m(x) dx = 0 \quad (n \neq m) \\
 \Leftrightarrow & \int_{-\infty}^{\infty} H_n(x) H_m(x) e^{-x^2} dx = 0.
 \end{aligned}$$

□

Die Hermite-Gauß-Quadratur bedarf der Nullstellen der Hermite-Polynome. Die folgenden zwei Sätze geben Aufschluss darüber, wie viele Nullstellen es gibt und wie diese berechnet werden können. Beide Sätze werden hier nicht bewiesen.

**Satz A.10.**

Sei  $\phi$  eine Gewichtsfunktion auf  $[a, b]$  und  $n \in \mathbb{N}$  fest vorgegeben. Dann besitzt das  $n$ -te Hermite-Polynom die einfachen, reellen Nullstellen  $z_1, \dots, z_n$  mit  $z_i \in [a, b]$  für alle  $i \in \{1, \dots, n\}$ . Seien  $L_1, \dots, L_n$  die diesen Nullstellen zugeordneten Lagrange'schen Basispolynome, definiert durch

$$L_j(x) := \prod_{\substack{k=1 \\ k \neq j}}^n \frac{x - z_k}{z_j - z_k}, \quad j \in \{1, \dots, n\}.$$





# Anhang B

## Die Multivariate Hermite-Gauß-Quadratur

In Abschnitt 3.2.2.1 wurden die *Gauß-Quadratur* und die *adaptive Gauß-Quadratur* zur Approximation der Integrale in Gleichung (3.48) für eindimensionale subjekt-spezifische Größen vorgestellt. Das zu approximierende Integral im Fall  $p$ -dimensionaler subjekt-spezifischer Größen  $U_i = (U_{i1}, \dots, U_{ip})$  mit  $p > 1$  ist von der Form

$$\begin{aligned} \int \prod_{j=1}^{n_i} f_{Y_{ij}|U_i=u_i}(y_{ij}) f_{U_i}(u_i) du_i &= \int f_{Y_i|U_i=u_i}(y_i) f_{U_i}(u_i) du_i \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{Y_i|U_i=u_i}(y_i) f_{U_i}(u_i) du_{i1} \dots du_{ip}. \end{aligned}$$

Zur besseren Übersicht wird der Index  $i$  für das  $i$ -te Subjekt weggelassen und die Funktion

$$h(u) := f_{Y_i|U=u}(y_i) \tag{B.1}$$

als eine Funktion in Abhängigkeit von der Variablen  $u$  definiert. Zudem werden zur Vermeidung von Unklarheiten die Funktion  $h$  und auch weitere Funktionen, die in diesem Zusammenhang genutzt werden, mit derjenigen Variablen, von der sie abhängen, indiziert. Somit ergibt sich das Integral

$$\int h_u(u) f_U(u) du = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h_u(u_1, \dots, u_p) f_U(u_1, \dots, u_p) du_1 \dots du_p, \tag{B.2}$$

wobei  $U \sim \mathcal{N}_p(0, D)$  mit  $p$ -dimensionaler Kovarianzmatrix  $D$  gilt. Ohne Beschränkung der Allgemeinheit wird angenommen, dass  $D = I_{p \times p}$  ist, da dies andernfalls durch die Transformation  $D^{-1/2}U = V$  und Anwendung der Transformationsformel erreicht werden kann.

Der folgende Satz gibt nun Aufschluss über die Möglichkeiten, dieses Integral für mehrdimensionale subjekt-spezifische Größen approximativ zu lösen, und liefert dadurch das multivariate Analogon zur in Kapitel 3 vorgestellten Hermite-Gauß- bzw. adaptiven Hermite-Gauß-Quadratur.

**Satz B.1.**

Sei  $r \in \mathbb{N}$  und für alle  $j_\ell \in \{1, \dots, r\}$  und  $\ell \in \{1, \dots, p\}$  sei  $z_{j_\ell} \in \{z_1, \dots, z_r\}$  eine Nullstelle des  $r$ -ten Hermite-Polynoms. Zudem seien die Gewichte für alle  $j \in \{1, \dots, r\}$  und  $k \in \{1, \dots, p\}$  durch

$$\omega_{j_k} := \langle L_j, 1 \rangle_{\phi_k}$$

gegeben, wobei  $L_j$  das in Definition 3.23 definierte Lagrange'sche Basispolynom bezeichnet. Zudem gelte  $\phi_k(u_k) = e^{-u_k^2}$  und  $u = (u_1^\top, \dots, u_p^\top)^\top$ .

Für  $\phi(u) = \exp(-u^\top u)$  und das Integral

$$I_p(h) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h_u(u) \phi(u) du_1 \cdots du_p$$

ist dann die numerische Integrationsformel

$$I_{p,r}(h) = \sum_{j_1=1}^r \cdots \sum_{j_p=1}^r h_u(z_{j_1}, \dots, z_{j_p}) \prod_{k=1}^p \omega_{j_k}$$

das multivariate Analogon zur eindimensionalen Hermite-Gauß-Quadratur.

*Beweis:* Es gilt

$$\phi(u) = \exp(-u^\top u) = \exp\left(-\sum_{\ell=1}^p u_\ell^2\right) = \prod_{\ell=1}^p \exp(-u_\ell^2) = \prod_{\ell=1}^p \phi_\ell(u_\ell)$$

und damit auch

$$\begin{aligned} I_p(h) &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h_u(u_1, \dots, u_p) \phi(u) du_1 \cdots du_p \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h_u(u_1, \dots, u_p) \prod_{\ell=1}^p \phi_\ell(u_\ell) du_1 \cdots du_p \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \prod_{\ell=1}^{p-1} \phi_\ell(u_\ell) \left( \int_{-\infty}^{\infty} \phi_p(u_p) h_u(u_1, \dots, u_p) du_p \right) du_{p-1} \cdots du_1 \\ &\stackrel{(*)}{\approx} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \prod_{\ell=1}^{p-1} \phi_\ell(u_\ell) \left( \sum_{j_p=1}^r \omega_{j_p} h_u(u_1, \dots, u_{p-1}, z_{j_p}) \right) du_{p-1} \cdots du_1 \end{aligned}$$

$$\begin{aligned}
&= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \prod_{\ell=1}^{p-2} \phi_{\ell}(u_{\ell}) \sum_{j_p=1}^r \omega_{j_p} \left( \int_{-\infty}^{\infty} \phi_{p-1}(u_{p-1}) h_u(u_1, \dots, u_{p-1}, z_{j_p}) du_{p-1} \right) \\
&\quad du_{p-2} \cdots du_1 \\
&\stackrel{(*)}{\approx} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \prod_{\ell=1}^{p-3} \phi_{\ell}(u_{\ell}) \sum_{j_p=1}^r \omega_{j_p} \sum_{j_{p-1}=1}^r \omega_{j_{p-1}} h_u(u_1, \dots, z_{j_{p-1}}, z_{j_p}) du_{p-3} \cdots du_1 \\
&\quad \vdots \\
&\approx \sum_{j_1=1}^r \cdots \sum_{j_p=1}^r \prod_{k=1}^p \omega_{j_k} h_u(z_{j_1}, \dots, z_{j_p}), \tag{B.3}
\end{aligned}$$

wobei (\*) aus der eindimensionalen Gauß-Quadratur gemäß Definition 3.23 hervorgeht.  $\square$

Abschließend soll die multivariate Version der adaptiven Gauß-Quadratur zur Berechnung des Integrals

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h_u(u_1, \dots, u_p) f_U(u_1, \dots, u_p) du_1 \dots du_p \tag{B.4}$$

genutzt werden. Analog zum Vorgehen bei der eindimensionalen adaptiven Gauß-Quadratur wird angenommen, dass

$$h_u(u_1, \dots, u_p) f_U(u_1, \dots, u_p)$$

der Dichte einer  $p$ -dimensionalen normalverteilten Zufallsvariablen  $X$  entspricht. Dann ist nach Bemerkung 3.29 der Erwartungswert dieser Zufallsvariable durch

$$\hat{u} := \operatorname{argmax}_{u \in \mathbb{R}^p} \ln \underbrace{\left[ h_u(u) f_U(u) \right]}_{=: g_u(u)},$$

und die Kovarianzmatrix durch

$$\Sigma := \left[ -\frac{\partial}{\partial u} \left( \frac{\partial}{\partial u} \ln(g_u(u)) \right)^{\top} \right]^{-1}$$

gegeben. Die Substitution von

$$t := (u - \hat{u}) \Sigma^{-1/2} 2^{-1/2}$$

liefert mit Hilfe der Transformationsformel folgendes:

$$\begin{aligned}
&du = |\det(2^{-1/2} \Sigma^{-1/2})| dt \\
&\stackrel{(*)}{\Leftrightarrow} du = |2^{-p/2} (\det(\Sigma))^{-1/2}| dt \\
&\Leftrightarrow dt = 2^{p/2} |(\det(\Sigma))^{1/2}| du,
\end{aligned}$$

dabei gilt (\*) wegen

$$\begin{aligned} \sqrt{\det(\Sigma)} \sqrt{\det(\Sigma)} &= \det(\Sigma) \\ &= \det(\Sigma^{1/2} \Sigma^{1/2}) \\ &= \det(\Sigma^{1/2}) \det(\Sigma^{1/2}). \end{aligned}$$

Somit ergibt sich für das Integral in (B.4)

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g_u(u) \, du = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} 2^{p/2} |\det(\Sigma)|^{1/2} g_t(t) \, dt.$$

Um die mehrdimensionale Gauß-Quadratur auf dieses Integral anwenden zu können, wird der Integrand erweitert zu

$$\underbrace{2^{p/2} |\det(\Sigma)|^{1/2} g_t(t) \exp(t^\top t)}_{:=f_t(t)} \exp(-t^\top t).$$

Aus Satz B.1 folgt

$$\begin{aligned} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g_u(u) \, du &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} 2^{p/2} |\det(\Sigma)|^{1/2} g_t(t) \, dt \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_t(t) \exp(-t^\top t) \, dt \\ &\stackrel{\text{Satz B.1}}{\approx} \sum_{j_1=1}^r \cdots \sum_{j_p=1}^r \prod_{k=1}^p \omega_{j_k} f_t(z_{j_1}, \dots, z_{j_p}) \\ &= \sum_{j_1=1}^r \cdots \sum_{j_p=1}^r \prod_{k=1}^p \omega_{j_k} 2^{p/2} |\det(\Sigma)|^{1/2} g_t(z_{j_1}, \dots, z_{j_p}) \\ &\quad \times \exp\left(\sum_{k=1}^p z_{j_k}^2\right). \end{aligned} \tag{B.5}$$

Mit

- $t = (u - \hat{u})\Sigma^{-1/2} 2^{-1/2} \iff u = t \Sigma^{1/2} 2^{1/2} + \hat{u}$ ,
- $h_u(u) := f_{Y_i|U=u}(y_i)$  und
- $g_u(u) = h(u) f_U(u)$

gilt somit

$$\begin{aligned}
\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h_u(u) f_U(u) \, du &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g_u(u) \, du \\
&\stackrel{\text{vgl. B.5}}{\approx} \sum_{j_1=1}^r \cdots \sum_{j_p=1}^r \prod_{k=1}^p \omega_{j_k} 2^{p/2} |\det(\Sigma)|^{1/2} g_t(z_{j_1}, \dots, z_{j_p}) \\
&\quad \times \exp\left(\sum_{k=1}^p z_{j_k}^2\right) \\
&= \sum_{j_1=1}^r \cdots \sum_{j_p=1}^r \prod_{k=1}^p \omega_{j_k} 2^{p/2} |\det(\Sigma)|^{1/2} \\
&\quad g_u(z_{j_1} \Sigma^{1/2} \sqrt{2} + \hat{u}, \dots, z_{j_p} \Sigma^{1/2} \sqrt{2} + \hat{u}) \\
&\quad \times \exp\left(\sum_{k=1}^p z_{j_k}^2\right) \\
&= \sum_{j_1=1}^r \cdots \sum_{j_p=1}^r \prod_{k=1}^p \omega_{j_k} f_{Y_i|U=\hat{a}}(y_i) \\
&\quad \times f_U(\hat{a}) 2^{p/2} |\det(\Sigma)|^{1/2} \exp\left(\sum_{k=1}^p z_{j_k}^2\right),
\end{aligned}$$

wobei  $\hat{a}$  definiert ist als  $\hat{a} := (a_{j_1}, \dots, a_{j_p})$  mit  $a_{j_k} := z_{j_k} \Sigma^{1/2} \sqrt{2} + \hat{u}$  für alle  $k \in \{1, \dots, p\}$ .

Damit wurde eine Möglichkeit erarbeitet, die Integrale der Loglikelihood-Funktion

$$\begin{aligned}
l_\delta(y) &= \sum_{i=1}^m \ln \left( \int \prod_{j=1}^{n_i} f_{Y_{ij}|U_i=u_i}(y_{ij}) f_{U_i}(u_i) \, du_i \right) \\
&= \sum_{i=1}^m \ln \left( \int f_{Y_i|U_i=u_i}(y_i) f_{U_i}(u_i) \, du_i \right)
\end{aligned}$$

für mehrdimensionale subjekt-spezifische Größen  $U_i$  approximativ zu lösen. Es gilt

$$\begin{aligned}
l_\delta(y) &\approx \sum_{i=1}^m \ln \left[ \sum_{j_1=1}^r \cdots \sum_{j_p=1}^r \prod_{k=1}^p \omega_{j_k} f_{Y_i|U_i=\hat{a}_i}(y_i) \right. \\
&\quad \left. \times f_{U_i}(\hat{a}_i) 2^{p/2} |\det(\Sigma)|^{1/2} \exp\left(\sum_{k=1}^p z_{j_k}^2\right) \right] \\
&=: \tilde{l}_\delta(y).
\end{aligned} \tag{B.6}$$



# Anhang C

## Die SAS - Anweisungen zum Zeichnen der Abbildungen in Kapitel 4

### Erstellung der Abbildung 4.1

Die SAS-Anweisungen zum Zeichnen der marginalisierten subjekt-spezifischen Entwicklung sind wie folgt:

```
1 data op;
2   do treat=0 to 1 by 1;
3     do subject=1 to 1000 by 1;
4       b1=rannor(-1); b2=rannor(-1);
5       u1=1.1340*b1; u2=-0.0163*b1 + 0.0462*b2;
6       do t=0 to 27 by 0.1;
7         if treat=0 then y=exp(0.7739 + u1+ (-0.0245 + u2)*t);
8         else y=exp(0.7739 + u1 + (-0.0193 + u2)*t);
9         output;
10      end;
11    end;
12  end;
13
14 proc sort data=op;
15 by t treat;
16 run;
17
18 proc means data=op;
19 var y;
20 by t treat;
21 output out=out;
```

```
22 run;
23
24 data out1;
25 set out;
26 if _stat_='MEAN' then;
27 else delete;
28 run;
29
30 proc gplot data=out1;
31 plot y*t=treat/ haxis=axis1 vaxis=axis2 legend =legend1;
32 axis1 label =(h=2.5 'Zeit□(in□Wochen)') value=(h=1.5)
33 order=(0 to 25 by 5) minor =none;
34 axis2 label =(h=2.5 A=90 'E(Y(t))') value=(h=1.5)
35 order=(0 to 10 by 1) minor =none;
36 legend1 label=(h=2 'Behandlung:')
37 value=(h=2 'Placebo' 'Verum');
38 symbol1 c=black i=join w=3 l=1 mode=include;
39 symbol2 c=black i=join w=2 l=2 mode=include;
40 run;
```

Der SAS-Code für die marginale Entwicklung ist dann entsprechend:

```
1 data op1;
2 do treat=0 to 1 by 1;
3 do t=0 to 27 by 0.1;
4 if treat=0 then y=exp(1.314 + (-0.0142)*t);
5 else y=exp(1.314 + (-0.0192)*t);
6 output;
7 end;
8 end;
9
10
11 proc gplot data=op1;
12 plot y*t=treat/ haxis=axis1 vaxis=axis2 legend =legend1;
13 axis1 label =(h=2.5 'Zeit□(in□Wochen)') value=(h=1.5)
14 order=(0 to 25 by 5) minor =none;
15 axis2 label =(h=2.5 A=90 'E(Y(t))') value=(h=1.5)
16 order=(0 to 10 by 1) minor =none;
17 legend1 label=(h=2 'Behandlung:')
18 value=(h=2 'Placebo' 'Verum');
19 symbol1 c=black i=join w=3 l=1 mode=include;
20 symbol2 c=black i=join w=2 l=2 mode=include;
21 run;
```

---

## Erstellung der Abbildung 4.2

Auch die Abbildung 4.2 enthält zwei Sub-Grafiken:

Die Placebo-Gruppe:

```
1 data op;
2   do treat=0 to 1 by 1;
3     do subject=1 to 10 by 1;
4       b1=rannor(-1); b2=rannor(-1);
5       u1=1.1340*b1; u2=-0.0163*b1 + 0.0462*b2;
6       do t=0 to 27 by 0.1;
7         if treat=0 then y=exp(0.7739 + u1+ (-0.0245 + u2)*t);
8         else y=exp(0.7739 + u1 + (-0.0193 + u2)*t);
9         output;
10        end;
11      end;
12    end;
13
14 data placebo;
15   set op;
16   if treat=1 then delete;
17   run;
18
19 proc sort data =placebo;
20   by t;
21   run;
22
23 proc means data=placebo;
24   var y;
25   by t;
26   output out = outp;
27   run;
28
29 data outp2;
30   set outp;
31   if _stat_='MEAN' then;
32   else delete;
33   run;
34
35 data tmp1;
36   set outp2;
37   subject=11;
38   run;
39
```

```
40 data placebo1;
41   set placebo tmp1;
42 run;
43
44 proc sort data=placebo1;
45 by t;
46 run;
47
48 proc gplot data=placebo1;
49   plot y*t=subject/ haxis=axis1 vaxis=axis2 nolegend;
50 axis1 label =(h=2.5 'Zeit□(in□Wochen)') value=(h=1.5)
51       order=(0 to 25 by 5) minor =none;
52 axis2 label =(h=2.5 A=90 'E(Y(t))') value=(h=1.5)
53       order=(0 to 10 by 1) minor =none;
54 symbol1 c=black i=join w=1 l=2 mode=include;
55 symbol2 c=black i=join w=1 l=2 mode=include;
56 symbol3 c=black i=join w=1 l=2 mode=include;
57 symbol4 c=black i=join w=1 l=2 mode=include;
58 symbol5 c=black i=join w=1 l=2 mode=include;
59 symbol6 c=black i=join w=1 l=2 mode=include;
60 symbol7 c=black i=join w=1 l=2 mode=include;
61 symbol8 c=black i=join w=1 l=2 mode=include;
62 symbol9 c=black i=join w=1 l=2 mode=include;
63 symbol10 c=black i=join w=1 l=2 mode=include;
64 symbol11 c=black i=join w=4 l=1 mode=include;
65 run;
```

Die Verum-Gruppe:

```
1 data op;
2   do treat=0 to 1 by 1;
3     do subject=1 to 10 by 1;
4       b1=rannor(-1); b2=rannor(-1);
5       u1=1.1340*b1; u2=-0.0163*b1 + 0.0462*b2;
6       do t=0 to 27 by 0.1;
7         if treat=0 then y=exp(0.7739 + u1+ (-0.0245 + u2)*t);
8         else y=exp(0.7739 + u1 + (-0.0193 + u2)*t);
9         output;
10        end;
11      end;
12    end;
13
14 data treatment;
15 set op;
16 if treat=0 then delete;
```

---

```

17  run;
18
19  proc sort data =treatment;
20  by t;
21  run;
22
23  proc means data=treatment;
24  var y;
25  by t;
26  output out = outt;
27  run;
28
29  data outt2;
30  set outt;
31  if _stat_='MEAN' then;
32  else delete;
33  run;
34
35  data tmp1;
36  set outt2;
37  subject=11;
38  run;
39
40  data treat1;
41  set treatment tmp1;
42  run;
43
44  proc sort data=treat1;
45  by t;
46  run;
47
48  proc gplot data=treat1;
49  plot y*t=subject/ haxis=axis1 vaxis=axis2 ;
50  axis1 label =(h=2.5 'Zeit□(in□Wochen)') value=(h=1.5)
51  order=(0 to 25 by 5) minor =none;
52  axis2 label =(h=2.5 A=90 'E(Y(t))') value=(h=1.5)
53  order=(0 to 10 by 1) minor =none;
54  symbol1 c=black i=join w=1 l=2 mode=include;
55  symbol2 c=black i=join w=1 l=2 mode=include;
56  symbol3 c=black i=join w=1 l=2 mode=include;
57  symbol4 c=black i=join w=1 l=2 mode=include;
58  symbol5 c=black i=join w=1 l=2 mode=include;
59  symbol6 c=black i=join w=1 l=2 mode=include;
60  symbol7 c=black i=join w=1 l=2 mode=include;

```

```
61 symbol8 c=black i=join w=1 l=2 mode=include;  
62 symbol9 c=black i=join w=1 l=2 mode=include;  
63 symbol10 c=black i=join w=1 l=2 mode=include;  
64 symbol11 c=black i=join w=4 l=1 mode=include;  
65 run;
```

# Anhang D

## Die Endtableaus für die Analyse von Modell (5.4) und Modell (5.5)

### D.1 Endtableaus für die Analyse des Modells (5.4)

#### Design 1

The GENMOD Procedure

Analysis Of GEE Parameter Estimates  
Empirical Standard Error Estimates

Parameter		Estimate	Standard	95% Confidence		Z	Pr >  Z
			Error	Limits			
Intercept		2.0622	0.0469	1.9703	2.1541	43.97	<.0001
TREAT	1	0.0252	0.0732	-0.1182	0.1686	0.34	0.7303
TREAT	0	0.0000	0.0000	0.0000	0.0000	.	.
time*TREAT	1	-0.0098	0.0014	-0.0124	-0.0071	-7.24	<.0001
time*TREAT	0	-0.0048	0.0009	-0.0066	-0.0031	-5.43	<.0001

#### Design 2

Analysis Of GEE Parameter Estimates  
Empirical Standard Error Estimates

Parameter		Estimate	Standard	95% Confidence		Z	Pr >  Z
			Error	Limits			
Intercept		1.9476	0.0716	1.8072	2.0879	27.20	<.0001
TREAT	1	-0.0138	0.1131	-0.2356	0.2079	-0.12	0.9027
TREAT	0	0.0000	0.0000	0.0000	0.0000	.	.
time*TREAT	1	-0.0066	0.0015	-0.0096	-0.0035	-4.26	<.0001
time*TREAT	0	-0.0030	0.0012	-0.0053	-0.0006	-2.50	0.0125

### Design 3

Analysis Of GEE Parameter Estimates  
Empirical Standard Error Estimates

Parameter		Estimate	Standard	95% Confidence		Z Pr >  Z	
			Error	Limits			
Intercept		2.0095	0.0574	1.8970	2.1219	35.02	<.0001
TREAT	1	0.0531	0.0864	-0.1163	0.2225	0.61	0.5388
TREAT	0	0.0000	0.0000	0.0000	0.0000	.	.
time*TREAT	1	-0.0608	0.0095	-0.0794	-0.0422	-6.41	<.0001
time*TREAT	0	-0.0240	0.0062	-0.0362	-0.0118	-3.86	0.0001

### Design 4

Analysis Of GEE Parameter Estimates  
Empirical Standard Error Estimates

Parameter		Estimate	Standard	95% Confidence		Z Pr >  Z	
			Error	Limits			
Intercept		1.7179	0.1486	1.4284	2.0110	11.57	<.0001
TREAT	1	0.2546	0.2378	-0.2115	0.7206	1.07	0.2844
TREAT	0	0.0000	0.0000	0.0000	0.0000	.	.
time*TREAT	1	-0.0086	0.0066	-0.0215	0.0043	-1.30	0.1932
time*TREAT	0	0.0057	0.0057	-0.0056	0.0169	0.99	0.3225

### Design 5

Analysis Of GEE Parameter Estimates  
Empirical Standard Error Estimates

Parameter		Estimate	Standard	95% Confidence		Z Pr >  Z	
			Error	Limits			
Intercept		1.7730	0.3107	1.1640	2.3820	5.71	<.0001
TREAT	1	0.0028	0.4800	-0.9379	0.9434	0.01	0.9954
TREAT	0	0.0000	0.0000	0.0000	0.0000	.	.
time*TREAT	1	-0.0056	0.0069	-0.0191	0.0078	-0.82	0.4135
time*TREAT	0	-0.0002	0.0058	-0.0115	0.0111	-0.03	0.9739

Design 6

Analysis Of **GEE Parameter Estimates**  
Empirical Standard Error Estimates

Parameter		Estimate	Standard Error	95% Confidence		Z	Pr >  Z
				Limits			
Intercept		1.1423	0.4956	0.1709	2.1137	2.30	0.0212
TREAT	1	-0.4696	0.7840	-2.0062	1.0669	-0.60	0.5492
TREAT	0	0.0000	0.0000	0.0000	0.0000	.	.
time*TREAT	1	0.0088	0.0075	-0.0059	0.0235	1.17	0.2401
time*TREAT	0	0.0074	0.0060	-0.0043	0.0192	1.24	0.2138

Das Modell in (5.4) wird von der Prozedur GENMOD, wie bei der Medikament-A-Studie (vgl. Abschnitt 5.3.4), überparametrisiert, so dass die Parameter der dritten Zeile jeweils gleich Null gesetzt werden.

## D.2 Endtableaus für die Analyse des Modells (5.5)

Design 1

**The GENMOD Procedure**

Analysis Of **GEE Parameter Estimates**  
Empirical Standard Error Estimates

Parameter		Estimate	Standard Error	95% Confidence		Z	Pr >  Z
				Limits			
Intercept		2.0750	0.0367	2.0030	2.1469	56.51	<.0001
time*TREAT	1	-0.0096	0.0014	-0.0124	-0.0067	-6.62	<.0001
time*TREAT	0	-0.0051	0.0010	-0.0071	-0.0030	-4.86	<.0001

**Contrast Estimate Results**

Label	Estimate	Standard Error	Alpha	Confidence Limits		Chi-Square	Pr > ChiSq
diff slopes	<b>-0.0045</b>	0.0019	0.05	-0.0083	-0.0007	5.45	<b>0.0195</b>

### Design 2

Analysis Of GEE Parameter Estimates							
Empirical Standard Error Estimates							
Parameter		Estimate	Standard	95% Confidence		Z	Pr >  Z
			Error	Limits			
Intercept		1.9407	0.0564	1.8302	2.0513	34.39	<.0001
time*TREAT	1	-0.0066	0.0014	-0.0094	-0.0038	-4.58	<.0001
time*TREAT	0	-0.0029	0.0011	-0.0051	-0.0008	-2.65	0.0080

Contrast Estimate Results							
Label	Estimate	Standard Error	Alpha	Confidence Limits		Chi-Square	Pr > ChiSq
diff slopes	<b>-0.0037</b>	0.0017	0.05	-0.0070	-0.0004	4.83	<b>0.0279</b>

### Design 3

Analysis Of GEE Parameter Estimates							
Empirical Standard Error Estimates							
Parameter		Estimate	Standard	95% Confidence		Z	Pr >  Z
			Error	Limits			
Intercept		2.0365	0.0434	1.9513	2.1216	46.89	<.0001
time*TREAT	1	-0.0586	0.0094	-0.0771	-0.0402	-6.23	<.0001
time*TREAT	0	-0.0261	0.0063	-0.0385	-0.0138	-4.16	<.0001

Contrast Estimate Results							
Label	Estimate	Standard Error	Alpha	Confidence Limits		Chi-Square	Pr > ChiSq
diff slopes	<b>-0.0325</b>	0.0113	0.05	-0.0546	-0.0104	8.30	<b>0.0040</b>

### Design 4

Analysis Of GEE Parameter Estimates							
Empirical Standard Error Estimates							
Parameter		Estimate	Standard Error	95% Confidence Limits		Z	Pr >  Z
Intercept		1.8439	0.1198	1.6092	2.0786	15.40	<.0001
time*TREAT	1	-0.0037	0.0046	-0.0127	0.0052	-0.81	0.4160
time*TREAT	0	0.0011	0.0049	-0.0085	0.0106	0.22	0.8276

Contrast Estimate Results							
Label	Estimate	Standard Error	Alpha	Confidence Limits		Chi-Square	Pr > ChiSq
diff slopes	-0.0048	0.0035	0.05	-0.0117	0.0021	1.85	0.1737

### Design 5

Analysis Of GEE Parameter Estimates							
Empirical Standard Error Estimates							
Parameter		Estimate	Standard Error	95% Confidence Limits		Z	Pr >  Z
Intercept		1.7742	0.2368	1.3100	2.2383	7.49	<.0001
time*TREAT	1	-0.0056	0.0047	-0.0147	0.0036	-1.20	0.2316
time*TREAT	0	-0.0002	0.0045	-0.0090	0.0086	-0.05	0.0080

Contrast Estimate Results							
Label	Estimate	Standard Error	Alpha	Confidence Limits		Chi-Square	Pr > ChiSq
diff slopes	-0.0054	0.0024	0.05	-0.0100	-0.0008	5.21	0.0224

## Design 6

Analysis Of GEE Parameter Estimates  
Empirical Standard Error Estimates

Parameter		Standard		95% Confidence		Z Pr >  Z	
		Estimate	Error	Limits			
Intercept		0.9560	0.3847	0.2021	1.7100	2.49	0.0129
time*TREAT	1	0.0053	0.0049	-0.0043	0.0150	1.09	0.2762
time*TREAT	0	0.0097	0.0047	0.0006	0.0188	2.08	0.0371

Contrast Estimate Results

Label	Estimate	Standard		Confidence Limits		Chi-Square	Pr > ChiSq
		Error	Alpha				
diff slopes	<b>-0.0044</b>	0.0016	0.05	-0.0076	-0.0012	7.13	<b>0.0076</b>

# Anhang E

## Die SAS-Anweisung für die Berechnung des marginalisierten Erwartungswertes im Abschnitt 5.3.6

```
1 data op;
2 do treat=0 to 1 by 1;
3   do subject=1 to 2000000 by 1;
4     b1=rannor(-1);
5     u1=1.617714*b1;
6     if treat=0 then y=exp(-0.3592+u1+ 1.4098);
7     else y=exp(-0.3592+u1);
8     output;
9   end;
10 end;
11
12 data placebo;
13 set op;
14 if treat=1 then delete;
15 run;
16
17 data treatment;
18 set op;
19 if treat=0 then delete;
20 run;
21
22 proc means data=placebo;
23 var y;
```

```
24 run;  
25  
26 proc means data=treatment;  
27 var y;  
28 run;
```

# Anhang F

## Die Endtableaus für die alternative Analyse des subjekt- spezifischen Modells in Abschnitt 5.4.1

Die Analyse des alternativen subjekt-spezifischen Modells in Abschnitt 5.4.1 liefert folgende Tabellen:

### Die Prozedur NLMIXED

#### Spezifikationen

Data Set	WORK.HOT_FLUSHES_B
Abhängige Variable	Summe
Verteilung für abhängige Variable	Poisson
Zufällige Effekte	u
Verteilung für zufällige Effekte	Normal
Subjektvariable	SUBJECT
Optimierungstechnik	Quasi-Newton, dual
Integrationsmethode	Adaptive Gaussische Quadratur

#### Dimensionen

Verwendete Beobachtungen	100
Nicht verwendete Beobachtungen	0
Gesamtbeobachtungen	100
Subjekte	100
Max Beob. je Subjekt	1
Parameter	3
Quadraturpunkte	5

Parameter

beta_0	beta_1	D	NegLogLike
3	0.5	0.25	594.772875

Iterationsverlauf

Iteration	Aufrufe	NegLogLike	Diff	MaxGrad	Steigung
1	2	475.874082	118.8988	69.27542	-3084.6
2	3	462.504936	13.36915	64.74517	-26.394
3	5	438.918366	23.58657	22.18434	-11.5523
4	7	433.980197	4.938169	12.69723	-4.09032
5	9	431.777685	2.202512	6.124284	-0.74054
6	10	429.072749	2.704937	4.276912	-1.19803
7	12	428.602664	0.470085	1.816403	-0.37291
8	14	428.541353	0.061311	0.474687	-0.05799
9	16	428.537082	0.004271	0.025233	-0.00775
10	18	428.536999	0.000083	0.012724	-0.00011
11	20	428.536979	0.000002	0.000207	-0.00002
12	12	428.536979	7.525E-9	1.496E-6	-1.5E-8

HINWEIS: GCONV convergence criterion satisfied.

Anpassungsstatistiken

-2 Log Likelihood	857.1
AIC (kleiner ist besser)	863.1
AICC (kleiner ist besser)	863.3
BIC (kleiner ist besser)	870.9

Parameterschätzwerte

Parameter	Schätzwert	Standardfehler	Freiheits- grade	t-Wert	Pr >  t	Alpha	Untere
beta_0	1.5915	0.2618	99	6.08	<.0001	0.05	1.0721
beta_1	1.4052	0.3443	99	4.08	<.0001	0.05	0.7220
D	2.6324	0.5047	99	5.22	<.0001	0.05	1.6310

Parameterschätzwerte

Parameter	Obere	Gradient
beta_0	2.1109	1.496E-6
beta_1	2.0884	8.117E-7
d_11	3.6339	-6.62E-7

# Anhang G

## Die R-Funktionen zur Verdeutlichung der Einflussgrößen des benötigten Stichprobenumfangs

Die Berechnung des exakten benötigten Stichprobenumfangs mit dem Signifikanzniveau  $\alpha^* = 0.05$  und einer geforderten Power von  $\tilde{\beta} = 0.9$  wird wie folgt durchgeführt:

```
1 fallzahl1←function(n, alpha, phi, beta0, d)
2     2*((1.96 + 1.282)**2*(exp(beta0) + exp(beta0 + d))*(1 + (n-1)*alpha))/
3     (n*phi**(-1)*(exp(beta0 + d)-exp(beta0))**2)
```

Der benötigte Stichprobenumfang für gleich große Behandlungsgruppen wird berechnet durch:

```
1 fallzahl2←function(n, alpha, phi, beta0, d)
2     2* ceiling(((1.96 + 1.282)**2*(exp(beta0) + exp(beta0 + d))*(1 +
3     (n-1)*alpha))/(n*phi**(-1)*(exp(beta0 + d)-exp(beta0))**2))
```

Die Funktion Grafik\_fallzahl\_n hat die Form:

```
1 Grafik_fallzahl_n←function(n, alpha, phi, beta0, d){
2   fallzahl1←NULL
3   for(i in (1:n)){
4     fallzahl1←c(fallzahl1,fallzahl1(i, alpha, phi, beta0, d))
5   }
6   fallzahl2←NULL
7   for(i in (1:n)){
8     fallzahl2←c(fallzahl2,fallzahl2(i, alpha, phi, beta0, d))
9   }
10  plot(1:n, fallzahl1, xlab="Anzahl der Wiederholungen", ylab=
```

```

11 "Anzahl_ der_ Subjekte", main="Wiederholung_ versus
12 _ _ Stichprobenumfang_ m")
13 fallzahltabelle←matrix(c(1:n,fallzahl1,fallzahl2),ncol=3)
14 dimnames(fallzahltabelle)←list(NULL,c("n","exaktes_ m",
15 "gerundetes_ m"))
16 fallzahltabelle
17 };

```

Die Funktion Grafik\_fallzahl\_phi ist gegeben durch:

```

1 Grafik_fallzahl_phi←function(n, alpha, phi, beta0, d){
2   fallzahl1←NULL
3   for(i in seq(1,phi,0.2))\{
4     fallzahl1←c(fallzahl1,fallzahl1(n, alpha, i, beta0, d))
5   \}
6   fallzahl2←NULL
7   for(i in seq(1,phi,0.2)){
8     fallzahl2←c(fallzahl2,fallzahl2(n, alpha, i, beta0, d))
9   \}
10  plot(seq(1,phi,0.2), fallzahl1, ylab="Anzahl_ der_ Subjekte",
11       xlab="Größe_ des_ Überdispersions-Parameters",
12       main="Überdispersion_ versus_ Stichprobenumfang_ m")
13  fallzahltabelle←matrix(c(seq(1,phi,0.2),fallzahl1,fallzahl2),ncol=3)
14  dimnames(fallzahltabelle)←list(NULL,c("phi","exaktes_ m",
15 "gerundetes_ m"))
16  fallzahltabelle
17  \};

```

Die Funktion Grafik\_fallzahl\_alpha hat die Form:

```

1 Grafik_fallzahl_alpha←function(n, alpha, phi, beta0, d){
2   fallzahl1←NULL
3   for(i in seq(0,alpha,0.1)){
4     fallzahl1←c(fallzahl1,fallzahl1(n, i, phi, beta0, d))
5   \}
6   fallzahl2←NULL
7   for(i in seq(0,alpha,0.1)){
8     fallzahl2←c(fallzahl2,fallzahl2(n, i, phi, beta0, d))
9   \}
10  plot(seq(0,alpha,0.1), fallzahl1, xlab="Korrelationsparameter_ der
11  _ _ Exchangeable-Struktur", ylab="Anzahl_ der_ Subjekte",
12  main="Korrelation_ versus_ Stichprobenumfang_ m")
13  fallzahltabelle←matrix(c(seq(0,alpha,0.1),fallzahl1,fallzahl2),
14  ncol=3)
15  dimnames(fallzahltabelle)←list(NULL,c("alpha","exaktes_ m",
16  "gerundetes_ m"))

```

---

```

17 fallzahltabelle
18 \};

```

Die Funktion `Grafik_fallzahl_d` wird bestimmt durch:

```

1 Grafik_fallzahl_d←function(n, alpha, phi, beta0, d){
2   fallzahl1←NULL
3   for(i in seq(0.1,d,0.1)){
4     fallzahl1←c(fallzahl1,fallzahl1(n, alpha, phi, beta0, i))
5     \}
6   fallzahl2←NULL
7   for(i in seq(0.1,d,0.1)){
8     fallzahl2←c(fallzahl2,fallzahl2(n, alpha, phi, beta0, i))
9     \}
10  plot(seq(0.1,d,0.1), fallzahl1, xlab="Kleinste_bedeutungsvolle
11  _Differenz_d", ylab="Anzahl_der_Subjekte",
12  main="Behandlungsunterschied_versus_Stichprobenumfang_m")
13  fallzahltabelle←matrix(c(seq(0.1,d,0.1),fallzahl1,fallzahl2),
14  ncol=3)
15  dimnames(fallzahltabelle)←list(NULL,c("d","exaktes_m",
16  "gerundetes_m"))
17  fallzahltabelle
18  \};

```

Die Funktion `Grafik_fallzahl_beta0` ist gegeben durch:

```

1 Grafik_fallzahl_beta0←function(n, alpha, phi, beta0, d){
2   fallzahl1←NULL
3   for(i in seq(-1,beta0,0.2)){
4     fallzahl1←c(fallzahl1,fallzahl1(n, alpha, phi, i, d))
5     \}
6   fallzahl2←NULL
7   for(i in seq(-1,beta0,0.2)){
8     fallzahl2←c(fallzahl2,fallzahl2(n, alpha, phi, i, d))
9     \}
10  plot(seq(-1,beta0,0.2), fallzahl1, xlab="Logarithmus_der_marginalen
11  _Erwartung_in_der_Verum-Gruppe", ylab="Anzahl_der_Subjekte",
12  main="Logarithmus_der_marginalen_Erwartung_in_der_Verum-Gruppe
13  _versus_Stichprobenumfang_m")
14  fallzahltabelle←matrix(c(seq(-1,beta0,0.2),fallzahl1,fallzahl2),
15  ncol=3)
16  dimnames(fallzahltabelle)←list(NULL,c("beta0","exaktes_m",
17  "gerundetes_m"))
18  fallzahltabelle
19  \};

```



# Anhang H

## Das R-Programm für den Algorithmus in Abschnitt 7.1.1

Gegenstand dieses Anhangs ist eine detaillierte Beschreibung der Implementierung des Algorithmus in Abschnitt 7.1.1. Darüber hinaus wird ein R-Programm als Umsetzung des Algorithmus vorgestellt.

### H.1 Detaillierte Implementierung

#### Schritt 0

Bereite eine Liste  $A$  vor, in der die Größen

- $\ell$  als Laufindex,
- $(r, s)$  als Kante mit dem minimalen Wert,
- $S_\ell$  als Menge derjenigen Knoten  $i$ , die über eine Kante sowohl mit Knoten  $r$  als auch mit Knoten  $s$  verbunden sind, und
- die Kanten- und Knotenwerte  $\alpha_{ij}^\ell$

wenn nötig gespeichert werden.

#### Schritt 1

Sei  $\ell = 1$  und berechne  $\alpha_{ij}^\ell = \rho_{ij} (\lambda_i \cdot \lambda_j)^{1/2}$  für alle  $i, j \in \{1, \dots, k\}$ <sup>1</sup>. Die Werte  $\alpha_{ij}^\ell$  entsprechen für  $i \neq j$  den Kantenwerten und die Zahlen  $\alpha_{ii}$  entsprechen den Knotenwerten.

---

<sup>1</sup>In Abschnitt 7.1.1 gilt  $k = n$ .

**Schritt 2**

Finde den kleinsten Wert einer Kante  $(r, s)$ , wobei  $r$  und  $s$  Knoten sind und, setze  $\beta_\ell = \lambda_{rs}^\ell$ .

**Schritt 3**

**Schritt 3A** Definiere eine Menge  $B$  mit  $B := \emptyset$  und  $S_\ell^0 = \{r, s\}$ .

**Schritt 3B** Für  $i \in \{1, \dots, k\}$  sei

$$S_\ell^i = \begin{cases} S_\ell^{i-1} \cup \{i\}, & \text{falls } \alpha_{ij}^\ell > 0 \text{ für alle } j \in S_\ell^{i-1} \\ S_\ell^{i-1}, & \text{sonst} \end{cases}$$

und  $S_\ell = S_\ell^k$

**Schritt 3C** Setze  $B = B \cup S_\ell$ .

**Wenn** es ein  $j \notin B$  mit  $\alpha_{rj}\alpha_{sj} > 0$  gibt, werden die aktuellen Werte für  $\ell, r, s, S_\ell$  und  $\alpha_{ij}^\ell$  für alle  $i, j$  in der Liste  $A$  gespeichert,  $S_\ell^0 = \{r, s, j\}$  gesetzt und erneut mit Schritt 3B fortgefahren.

**Sonst** gehe zu Schritt 4.

**Schritt 4**

Aktualisiere die Werte der Kanten und Knoten gemäß

$$\lambda_{ij}^{\ell+1} = \begin{cases} \lambda_{ij}^\ell - \beta_\ell, & \text{falls } i \in S_\ell \text{ und } j \in S_\ell \\ \lambda_{ij}^\ell, & \text{falls } i \notin S_\ell \text{ oder } j \notin S_\ell \end{cases}$$

und entferne alle Kanten  $(i, j)$  mit  $\lambda_{ij}^{\ell+1} = 0$ .

**Wenn** alle Kantenwerte kleiner oder gleich Null sind, muss mit Schritt 5 fortgefahren werden.

**Ansonsten** setze  $\ell = \ell + 1$  und gehe zu Schritt 2.

**Schritt 5**

**Schritt 5A** **Wenn** ein Knotenwert  $\alpha_{ii}$  negativ ist, dann

- entscheide, dass der Algorithmus ohne Lösung abbricht, **falls** die Liste  $A$  leer ist.

- Entferne **sonst** die letzten Elemente  $\ell, r, s, S_\ell, \alpha_{ij}$ , die im Schritt 3C in der Liste  $A$  gespeichert wurden und führe mit diesen Größen den Schritt 4 erneut durch;

**sonst** gehe zu Schritt 5B.

**Schritt 5B** Setze  $\ell = \tau$ ,  $\beta_{\tau+1} = \lambda_1^\tau$ , ...,  $\beta_{\tau+n} = \lambda_n^\tau$  und  $S_{\tau+1} = \{1\}$ , ...,  $S_{\tau+n} = \{n\}$ . Definiere die  $k \times \tau + n$  Matrix  $T = [T_{i\ell}]_{\substack{i=1,\dots,k \\ \ell=1,\dots,\tau+n}}$  als

$$T_{i\ell} := \begin{cases} 1, & \text{falls } i \in S_\ell, \\ 0, & \text{falls } i \notin S_\ell \end{cases}$$

für alle  $i \in \{1, \dots, k\}$  und  $\ell = 1, 2, \dots, \tau + n$ .

Liefert der Algorithmus in Schritt 5 einer Lösung, so enthält diese eine Matrix  $T$  und einen Zufallsvektor  $X$ . Auf deren Basis können dann korrelierte poissonverteilte Daten erzeugt werden (vgl. Kapitel 7).

## H.2 Das R -Programm

```

1 corrpoi ←function(R,lambda) {
2   abbruch ←FALSE
3   k ←length(lambda)
4   if(!is.vector(lambda)) {
5     cat("lambda_is_not_a_vector!_Algorithm_did_not_start.\n")
6     abbruch ←TRUE
7   }
8   if(!is.matrix(R)) {
9     cat("R_is_not_a_matrix!_Algorithm_did_not_start.\n")
10    abbruch ←TRUE
11  } else {
12    if(ncol(R) != nrow(R)) {
13      cat("R_is_not_quadratic!_Algorithm_did_not_start.\n")
14      abbruch ←TRUE
15    }
16    if(ncol(R) != k | nrow(R) != k) {
17      cat("Dimension_of_R_and_lambda_do_not_match!");
18      cat("Algorithm_did_not_start.\n")
19      abbruch ←TRUE
20    }
21    if(sum(diag(R) != 1) > 0 | sum(R > 1) > 0 | sum(R < -1) > 0) {
22      cat("R_is_not_a_valid_correlation_matrix!")

```

```

23     cat("Algorithm_did_not_start.\n")
24     abbruch ←TRUE
25   } else {
26     if(ncol(R)==nrow(R) & prod(R==t(R))==0) {
27       cat("R_is_not_a_valid_correlation_matrix!\n")
28       cat("Algorithm_did_not_start.\n")
29       abbruch ←TRUE
30     }
31   }
32 }
33
34 if(abbruch==FALSE) {
35   beta ←numeric(0)
36   #Step 0
37   l ←NULL
38   r ←NULL
39   s ←NULL
40   S ←list(NULL)
41   ecken ←lambda
42   kanten ←matrix(rep(0,k*k), ncol=k)
43   A ←list(l=list(l),r=list(r),s=list(s),S=list(S),e=list(ecken),k=list(kanten))
44
45   #Step 1
46   l ←1
47   kanten ←R * sqrt(lambda%*%t(lambda))
48   kanten[lower.tri(kanten, diag=TRUE)] ←Inf
49
50   #Step 2
51   step2 ←TRUE
52   while(step2 == TRUE) {
53     r ←which(kanten == min(kanten),arr.ind = TRUE)[1,1]
54     #Falls das Minimum an mehreren Kanten steht nehmen wir
55     #einfach den zuerst ermittelten
56     s ←which(kanten == min(kanten),arr.ind = TRUE)[1,2]
57     beta[l] ←kanten[r,s]
58
59     #Step 3
60     #a
61     B ←numeric(0)
62     S[[l]] ←c(r,s)
63     #b
64     step3b ←TRUE
65     while(step3b == TRUE) {
66       for(i in 1:k) {

```

```

67     h ←rep(0,length(S[[1]]))
68     index ←0
69     for(j in S[[1]]) {
70         index ←index+1
71         if(i==j & ecken[i] > 0) {h[index] ←1}
72         if(i!=j & kanten[min(i,j),max(i,j)] > 0 &
73         kanten[min(i,j),max(i,j)] < Inf) {h[index] ←1}
74     }
75     if(prod(h)==1) {
76         S[[1]] ←unique(c(S[[1]],i))
77     }
78 }
79 #c
80 step3b ←FALSE
81 B ←c(B,S[[1]])
82 for(j in 1:k) {
83     if(prod(B!=j)==1) {
84         if(j==r & ecken[j]*kanten[min(s,j),max(s,j)]>0 &
85         ecken[j]*kanten[min(s,j),max(s,j)]<Inf) {
86             A$l[[1]] ←1
87             A$r[[1]] ←r
88             A$s[[1]] ←s
89             A$S[[1]] ←S
90             A$e[[1]] ←ecken
91             A$k[[1]] ←kanten
92             S[[1]] ←unique(c(r,s,j))
93             step3b ←TRUE
94         }
95         else {A$l[[1]] ←0}
96         if(j==s & ecken[j]*kanten[min(r,j),max(r,j)]>0 &
97         ecken[j]*kanten[min(r,j),max(r,j)]<Inf) {
98             A$l[[1]] ←1
99             A$r[[1]] ←r
100            A$s[[1]] ←s
101            A$S[[1]] ←S
102            A$e[[1]] ←ecken
103            A$k[[1]] ←kanten
104            S[[1]] ←unique(c(r,s,j))
105            step3b ←TRUE
106        }
107        else {A$l[[1]] ←0}
108        if(j!=r & j!=s & kanten[min(r,j),max(r,j)]*
109        kanten[min(s,j),max(s,j)]>0 & kanten[min(r,j),max(r,j)]*
110        kanten[min(s,j),max(s,j)]<Inf) {

```

```

111         A$l[[1]] ← 1
112         A$r[[1]] ← r
113         A$s[[1]] ← s
114         A$S[[1]] ← S
115         A$e[[1]] ← ecken
116         A$k[[1]] ← kanten
117         S[[1]] ← unique(c(r,s,j))
118         step3b ← TRUE
119     }
120     else {A$l[[1]] ← 0}
121 }
122 else {A$l[[1]] ← 0}
123 }
124 }
125
126 #Step4
127 step4 ← TRUE
128 while(step4 == TRUE) {
129     for(i in S[[1]]) {ecken[i] ← ecken[i]-beta[1]}
130     for(i in S[[1]]) {
131         for(j in S[[1]]) {
132             if(j>i) {
133                 kanten[i,j] ← kanten[i,j]-beta[1]
134                 if(kanten[i,j]==0) {kanten[i,j] ← Inf}
135             }
136
137         }
138     }
139     step4 ← FALSE
140     if(sum(kanten!=Inf)>=1 & sum(kanten>0)>=1) {
141         step5 ← FALSE
142         step2 ← TRUE
143         l ← l+1
144     } else {
145         step5 ← TRUE
146         step2 ← FALSE
147     }
148
149     #Step5
150     if(step5 == TRUE) {
151         if(sum(ecken<0)>=1) {
152             h ← numeric(0)
153             for(i in 1:l) {
154                 h[i] ← A$l[[i]]

```

```

155     }
156     if(sum(h)==0) {
157         cat("Algorithm_failed!\n")
158         T ←NULL
159         A ←NULL
160         beta ←NULL
161         step4 ←FALSE
162         break
163     } else {
164         l ←max(which(h!=0))
165         r ←A$r[[1]]
166         s ←A$s[[1]]
167         S ←A$S[[1]]
168         ecken ←A$e[[1]]
169         kanten ←A$k[[1]]
170         step4 ←TRUE
171     }
172 } else {
173     T ←matrix(rep(0,k*1), ncol=1)
174     for(i in 1:k) {
175         for(j in 1:l) {
176             if(sum(S[[j]]==i)>=1) {T[i,j] ←1}
177         }
178     }
179     T ←cbind(T,diag(rep(1,k)))
180     beta ←c(beta, ecken)
181 }
182
183 }
184 }
185 }
186 }
187 if(prod(round(T %*% diag(beta) %*% t(T),2)==round(R *
188 sqrt(lambda%*%t(lambda),2))==1) {
189     return(list(T=T,beta=beta,A=A, R=R, lambda=lambda))
190 } else {
191     cat("Something_went_wrong.")
192     cat("Please_try_again_with_new_R_and/or_lambda!\n")
193 }
194 }

```



# Anhang I

## Die R-Programme zur Daten-Erzeugung

### I.1 Das R -Programm für den Fall $n = 1$

Im Vergleich zur Simulation korrelierter poissonverteilter Daten ist die Simulation im Fall einer einzigen Beobachtung unproblematisch. Wird der benötigte Stichprobenumfang bei gegebenen Parametern  $\alpha^*$ ,  $\tilde{\beta}$ ,  $d = \beta_1$ ,  $\alpha$  und  $\beta_0$  mittels Formel (6.10) berechnet, können nämlich durch die folgenden zwei Programme `generpoin1` und `simdatn1` Daten mit den entsprechenden Eigenschaften erzeugt werden.

```
1 generpoin1←function(R,lambda,m,treat1,treat2,treateffect,dateiname){
2   n ←m/2
3   G ←rpois(n,lambda)
4   H ←rpois(n,lambda*exp(treateffect))
5   daten1←data.frame(ID=rep(1:n,each=length(lambda)),treat=rep(treat1,n*
6     length(lambda)), day=rep(1:length(lambda),n),flushes=as.vector(G))
7   daten2←data.frame(ID=rep((n+1):(2*n),each=length(lambda)),treat=rep(treat2,
8     n*length(lambda)), day=rep(1:length(lambda),n),flushes=as.vector(H))
9   daten ←rbind(daten1,daten2)
10  #return(daten)
11  write.table(daten,dateiname,quote=FALSE,col.names=FALSE,row.names=FALSE)
12 }
13
14 simdatn1 ←function(R,lambda,m,treat1,treat2,treateffect,simanz){
15   for(i in 1:simanz) {
16     simdat←paste("L:/My_Documents/Simdaten/Simdaten7/daten",i,".txt",sep="")
17     generpoin1(R,lambda,m,treat1,treat2,treateffect,simdat)
18   }
19 }
```

Dabei gibt die Variable `simanz` an, wie viele Datensätze erzeugt werden sollen.

## I.2 Das R -Programm für den Fall von $n > 1$

Das Programm zur Erzeugung korrelierter poissonverteilter Daten im Fall von  $n > 1$  besteht aus zwei Teilen, dem Programm `generpoi` und dem Programm `simdat`.

### Das Programm `generpoi`

Dieses Programm setzt für gegebene Parameter  $\alpha^*$ ,  $\tilde{\beta}$ ,  $d = \beta_1$ ,  $\alpha$  und  $\beta_0$  (vgl. die Notation im Kapitel 7, Seite 145) folgende Schritte um:

**Schritt 1** Für die gegebene Korrelationsmatrix  $R(\alpha)$  und den marginalen Erwartungswert in der Verum-Gruppe, d.h.  $(\exp(\beta_0), \dots, \exp(\beta_0))^T \in \mathbb{R}^n$ , werden mit Hilfe des R -Programms `corrpoi` (vgl. Anhang H), die Matrix  $T_{Verum}$  und der poissonverteilte Zufallsvektor  $X_{Verum}$  berechnet.

**Schritt 2** Analog werden für die gegebene Korrelationsmatrix  $R(\alpha)$  und den marginalen Erwartungswert in der Placebo-Gruppe, d.h.  $(\exp(\beta_0 + \beta_1), \dots, \exp(\beta_0 + \beta_1))^T \in \mathbb{R}^n$ , die Matrix  $T_{Placebo}$  und der poissonverteilte Zufallsvektor  $X_{Placebo}$  berechnet.

**Schritt 3** Mit Hilfe des R-Befehls `rpois` werden Realisierungen von  $X_{Verum}$  bzw.  $X_{Placebo}$  erzeugt, und im Anschluss mit der Matrix  $T_{Verum}$  bzw.  $T_{Placebo}$  multipliziert.

**Schritt 4** Die in Schritt 3 ermittelten Daten werden in einen Datensatz der gewünschten Form gebracht. Der Datensatz enthält vier Variablen, die Subjektnummer (ID, d.h. Zahlen von 1 bis  $m$ ), die Behandlung (`treat`, dabei ist `treat1` Verum und `treat0` Placebo), die Anzahl der Wiederholung (`day`, d.h. 1 bis  $n$ ) und die korrelierten poissonverteilten Daten (`flushes`).

Das Programm ist somit gegeben durch

```
1 generpoi←function(R,lambda,n,treat1,treat0,d,simdaten){
2   a1 ←corrpoi(R,lambda)
3   beta1 ←a1$beta
4   T1 ←a1$T
5   m1 ←length(beta1)
6   X1 ←matrix(rep(0,m1*n),ncol=n)
7   for(i in 1:m1) {
```

```

8       X1[i,] ← rpois(n,beta1[i])
9     }
10    G←T1%*%X1
11    a2 ←corrpoi(R,lambda*exp(d))
12    beta2 ←a2$beta
13    T2 ←a2$T
14    m2 ←length(beta2)
15    X2 ←matrix(rep(0,m2*n),ncol=n)
16    for(i in 1:m2) {
17      X2[i,] ←rpois(n,beta2[i])
18    }
19    H←T2%*%X2
20    daten1←data.frame(ID=rep(1:n,each=length(lambda)),treat=rep(treat1,n*
21      length(lambda)), day=rep(1:length(lambda),n),flushes=as.vector(G))
22    daten2←data.frame(ID=rep((n+1):(2*n),each=length(lambda)),treat=rep(treat0,
23      n*length(lambda)), day=rep(1:length(lambda),n),flushes=as.vector(H))
24    daten ←rbind(daten1,daten2)
25    #return(daten)
26    write.table(daten,simdaten,quote=FALSE,col.names=FALSE,row.names=FALSE)
27  }

```

wobei  $\lambda = (\exp(\beta_0), \dots, \exp(\beta_0))^T$  und  $d = d = \beta_1$ .

## Das Programm simdat

Das Programm `simdat` ermöglicht nun die gleichzeitige Erzeugung mehrerer Datensätze mit Hilfe des Programms `generpoi`. Diese Datensätze werden im Anschluss im Ordner *Simdaten* unter dem Namen `dateni` gespeichert. Dabei gilt  $i \in \{1, \dots, \text{simanz}\}$ .

```

1  simdat ←function(R,lambda,n,treat1,treat0,d,simanz){
2    for(i in 1:simanz) {
3      simdat ←paste("h:/my_documents/simdaten/daten",i,".txt",sep="")
4      generpoi(R,lambda,n,treat1,treat0,d,simdat)
5    }
6  }

```



# Anhang J

## Das SAS-Programm zur Auswertung der simulierten Datensätze

Das Makro `auswertung(k, corrstr)` hängt von der Anzahl der simulierten Datensätze `k` und der verwandten Korrelationsstruktur `corrstr` ab. Die im Rahmen dieser Arbeit durchgeführten Analysen wurden alle mit `k = 1000` und `corrstr = EXCH` ausgewertet.

```
1 %Macro auswertung(k, corrstr);
2
3 /* Erzeuge einen Datensatz "Ergebnisse", der später um die geschätzten
4    GEE-Parameter ergänzt werden soll.*/
5
6    data Ergebnisse;
7        Estimate=1; Stderr=1; ProbZ=1; OUTPUT;
8    run;
9
10 %DO i=1 %to &k;
11
12 /*Einlesen der Datensätze "dateni.txt" für die Simulationsanzahl "k".*/
13
14    data simdat;
15        Infile "L:/My Documents/Simdaten/Simdaten3/daten&i..txt";
16        Input ID treat day flushes;
17    run;
18
19 /*Durchführen von der Prozedur GENMOD für jeden Datensatz "dateni.txt"
20    und Benennung der Output-Tafel GEEEmpPEst, welche die geschätzten
21    Regressionsparameter enthält, mit "tmp" (temporäre Datei).*/
22
```

```
23  ods output GEEEmpPEst=tmp;
24  proc genmod data=simdat;
25  class ID treat day;
26  model flushes=treat/dist=poisson;
27  repeated subject=ID/ corrw type=&corrstr within=day;
28  run;
29
30
31  /*Da die Z-Werte für die weiteren Betrachtungen keine Rolle spielen,
32  werden diese im Folgenden aus der Datei "tmp" eliminiert.*/
33
34  data tmp;
35  set tmp;
36  keep parm Level1 Estimate Stderr LowerCL UpperCL ProbZ;
37  run;
38
39  /*Es wird eine weitere Spalte "Durchlauf" definiert, die angibt, welcher
40  Datensatz "dateni.txt" die entsprechenden Regressionsparameter geliefert
41  hat.*/
42
43  data tmp;
44  set tmp;
45  durchlauf=&i;
46  output;
47  run;
48
49  /*Nun werden die Datensätze "Ergebnisse" und "tmp" zusammengefügt, indem
50  nach jedem Iterationsschritt die "tmp"-Daten zum bereits bestehenden
51  Datensatz "Ergebnisse" hinzugefügt wird.*/
52
53  data Ergebnisse;
54  set Ergebnisse tmp;
55  run;
56
57  /*Der Datensatz "Ergebnisse" wird nun nach der Größe "Durchlauf"
58  (Simulationsschritt) und den Regressionsparametern \beta_0, \beta_1 und
59  \beta_2 sortiert.*/
60
61  proc sort data=Ergebnisse out=Ergebnisse;
62  by durchlauf parm Level1;
63  run;
64
65  %End;
66
```

---

```

67  /*Der Datensatz "Ergebnisse" wird im Folgenden so transformiert, dass in den
68  Spalten die Größen \beta_0, \beta_1, die dazugehörigen Standardabweichungen,
69  Konfidenzintervallgrenzen und p-Werte stehen. Die Schätzer für den Regres-
70  sionsparameter \beta_2 sind aufgrund der Überparametrisierung alle Null,
71  so dass die entsprechenden Werte aus dem Datensatz eliminiert werden.*/
72
73  data Ergebnisse;
74  set Ergebnisse;
75  retain beta0 beta1 sd0 sd1 LCL1 UCL1 LCLO UCLO p0 p1;
76  if parm="Intercept" then do
77      beta0=Estimate;
78      sd0=Stderr;
79      LCLO=LowerCL;
80      UCLO=UpperCL;
81      p0=ProbZ;
82  end;
83  if level1="0" then do
84      beta1=Estimate;
85      sd1=Stderr;
86      LCL1=LowerCL;
87      UCL1=UpperCL;
88      p1=ProbZ;
89  end;
90  if level1="1" then delete; output;
91  run;
92
93  data Ergebnisse;
94  set Ergebnisse;
95  by durchlauf parm Level1;
96  if Last.durchlauf;
97  drop Estimate Stderr Level1 Parm LowerCL UpperCL ProbZ ;
98  run;
99
100 /*Im Folgenden werden Lageschätzungen der einzelnen Spalten des Datensatzes
101 "Ergebnisse" berechnet. Darüber hinaus wird ein Datensatz "pwerte" erzeugt,
102 der lediglich die p-Werte der \beta_1-Schätzung enthält. Im Anschluss wird
103 gezählt, wie oft der p-Wert kleiner als \alpha* =0.05 ist.*/
104
105 data Ergebnisse;
106 set Ergebnisse (firstobs=2);
107 run;
108
109 proc univariate data=Ergebnisse;
110 var beta1 sd1 LCL1 UCL1 beta0 sd0 LCLO UCLO;

```

```
111    run;
112
113    data pwerte;
114        set Ergebnisse;
115        keep p1;
116    run;
117
118    data power;
119        set pwerte;
120        retain counter;
121        if _N_ = 1 then counter=0;
122        if p1 < 0.05 then counter = counter + 1;
123    run;
124    %Mend;
125
126    %auswertung(1000,Exch);
```

# Literaturverzeichnis

- [SAS/STAT(1999) ] SAS Institute Inc. (1999) *SAS/STAT Users Guide Version 8*. Cary, NC: SAS Institute Inc.
- [Akaike 1974] AKAIKE, Hirotosugu: A new look at the statistical model identification. In: *IEEE Transactions on Automatic Control* 19 (1974), S. 716–723
- [Arfken und Weber 1985] ARFKEN, George B. ; WEBER, Hans J.: *Mathematical methods for Physicists*. Orlando, FL: Academic Press, 1985
- [Beal und Sheiner 1982] BEAL, S.L. ; SHEINER, L.B.: Estimating population kinetics. In: *CRC Critical Reviews in Biomedical Engineering* 8 (1982), S. 195–222
- [Beller u. a. 2002] BELLER, Elaine M. ; GEBSKI, Val ; KEECH, Anthony C.: Randomisation in clinical trials. In: *The Medical Journal of Australia* 177 (2002), S. 565–567
- [Bloch 1986] BLOCH, D.A.: Sample size requirements and the cost of a randomized clinical trial with repeated measurements. In: *Statistics in Medicine* 5 (1986), S. 663–667
- [Bröcker 1995] BRÖCKER, Theodor: *Analysis II*. Spektrum Akademischer Verlag GmbH Heidelberg-Berlin-Oxford, 1995
- [Breslow und Clayton 1993] BRESLOW, N.E. ; CLAYTON, D.G.: Approximate inference in generalized linear mixed models. In: *Journal of the American Statistical Association* 88-421 (1993), S. 9–25
- [Brockwell und Davis 1991] BROCKWELL, J. ; DAVIS, R.: *Time Series: Theory and Methods*. Springer Verlag, 1991
- [Burguete u. a. 1982] BURGUETE, J. ; GALLANT, R. ; SOUZA, G.: On unification of the asymptotic theory of nonlinear econometric models. In: *Econometric Reviews* 1 (1982), S. 151–190

- [Cannon u. a. 2001] CANNON, M.J. ; WARNER, L. ; TADDEI, J.A. ; KLEINBAUM, D.G.: What can go wrong when you assume that correlated data are independent: An Illustration from the Evaluation of a Childhood Health Intervention in Brazil. In: *Statistics in Medicine* 20 (2001), S. 1461–1467
- [Cook 1995] COOK, R. J.: The design and analysis of randomized trial with recurrent events. In: *Statistics in Medicine* 14 (1995), S. 2081–2098
- [CPMP 1998] CPMP: *Note for guidance on statistical principles for clinical trials*. EMEA, CPMP/ICH/363/396, 1998
- [Dahmen u. a. 2004] DAHMEN, G. ; ROCHON, J. ; KÖNIG, I.R. ; ZIEGLER, A.: Sample size calculations for controlled clinical trials using generalized estimating equations (GEE). In: *Methods of Information in Medicine* 43 (2004), S. 451–456
- [Dahmen und Ziegler 2004] DAHMEN, G. ; ZIEGLER, A.: Generalized estimating equations in controlled clinical trials: Hypotheses testing. In: *Biometrical Journal* 46-3 (2004), S. 214–232
- [Diepgen und Bleittner 1996] DIEPGEN, T.L. ; BLEITTNER, M.: Analysis of familial aggregation of atopic exzema and other atopic diseases by odds ratio regression model. In: *Journal of Investigative Dermatology* 106 (1996), S. 977–981
- [Diggle u. a. 1996] DIGGLE, Peter J. ; LIANG, Kung-Yee ; ZEGER, Scott L.: *Analysis of longitudinal data*. Oxford University Press, 1996
- [Emrich und Piedemonte 1991] EMRICH, L.J. ; PIEDEMONTE, M.R.: A method for generating high-dimensional multivariate binary variables. In: *The American Statistician* 45 (1991), S. 302–304
- [Faught u. a. 1996] FAUGHT, E. ; WILDER, B.J. ; RAMSAY, R.E. ; REIFE, R.A. ; KRAMER, L.D. ; PLEDGER, G.W. ; KARIM, R.M.: Topiramate placebo-controlled dose-ranging trial in refractory partial epilepsy using 200-, 400- and 600-mg daily dosages. In: *Neurology* 46 (1996), S. 1684–1690
- [Ferguson 1996] FERGUSON, Thomas S.: *A course in large sample theory*. Chapman & Hall, 1996
- [Gange 1995] GANGE, S.J.: Generating multivariate categorical variates using the iterative proportional fitting algorithm. In: *The American Statistician* 49 (1995), S. 134–138

- [Gänssler und Stute 1977] GÄNSSLER, P. ; STUTE, W.: *Wahrscheinlichkeitstheorie*. Springer Verlag, 1977
- [Gourieroux u. a. 1984] GOURIEROUX, C. ; MONFORT, A. ; TROGNON, A.: Pseudo Maximum Likelihood Methods: Theory. In: *Econometrica* 52, No. 3 (1984), S. 681–700
- [Hardin und Hilbe 2003] HARDIN, James W. ; HILBE, Joseph M.: *Generalized Estimating Equations*. Chapman & Hall/CRC, 2003
- [Karim und Zeger 1988] KARIM, M.R. ; ZEGER, S.L.: GEE: A SAS Macro for Longitudinal Data Analysis / Department of Biostatistics, The Johns Hopkins University, Baltimore, MD. 1988. – Forschungsbericht
- [Klebauer 1998] KLEBAUER, Fima C.: *Introduction to stochastic calculus with applications*. Imperial College Press, 1998
- [Lee 1993] LEE, A.J.: Generating random binary deviates having fixed marginal distributions and specified degrees of association. In: *The American Statistician* 47 (1993), S. 209–215
- [Liang und Zeger 1986] LIANG, Kung-Yee ; ZEGER, Scott L.: Longitudinal data analysis using generalized linear models. In: *Biometrika* 73-1 (1986), S. 13–22
- [Lindsey 1999] LINDSEY, J.K.: *Models for repeated measurements*. Oxford University Press, 1999
- [Liu und Liang 1997] LIU, G. ; LIANG, K.Y.: Sample size calculations for studies with correlated observations. In: *Biometrics* 53 (1997), S. 937–947
- [Lui und Cumberland 1992] LUI, K.J. ; CUMBERLAND, W. G.: Sample size requirements for repeated measurements in continuous data. In: *Statistics in Medicine* 11 (1992), S. 633–641
- [McCullagh und Nelder 1989] MCCULLAGH, P. ; NELDER, J.A.: *Generalized Linear Models*. Chapman & Hall, 1989
- [Molenberghs und Verbeke 2005] MOLENBERGHS, Geert ; VERBEKE, Geert: *Models for discrete longitudinal data*. Springer, 2005
- [Park u. a. 1996] PARK, C.G. ; PARK, T. ; SHIN, D.W.: A simple method for generating correlated binary variables. In: *The American Statistician* 50 (1996), S. 306–310

- [Park und Shin 1998] PARK, C.G. ; SHIN, D.W.: An algorithm for generating correlated random variables in a class of infinitely divisible distributions. In: *Journal of Statistical Computation and Simulation* 61 (1998), S. 127–139
- [Pinheiro und Bates 1995] PINHEIRO, José C. ; BATES, Douglas M.: Approximations to the log-likelihood function in the nonlinear mixed-effect model. In: *Journal of Computational and Graphical Statistics* 4 (1995), S. 12–35
- [Pschyrembel 1994] PSCHYREMBEL (Hrsg.): *Klinisches Wörterbuch*. Walter de Gruyter, 1994
- [Rao 1973] RAO, C. R.: *Linear statistical inference and its applications*. New York: Wiley, 1973
- [Rüger 1985] RÜGER, Bernhard: *Induktive Statistik*. Oldenbourg Verlag, 1985
- [Rochon 1998] ROCHON, J.: Application of GEE procedures for sample size calculations in repeated measures experiments. In: *Statistics in Medicine* 17 (1998), S. 1643–1658
- [Ronning 1977] RONNING, G.: A simple scheme for generating multivariate gamma distributions with non-negative covariance matrix. In: *Technometrics* 19 (1977), S. 179–183
- [Schmieder 1994] SCHMIEDER, Gerald: *Analysis - Eine Einführung für Mathematiker und Informatiker*. Vieweg, 1994
- [Schwarz 1978] SCHWARZ, G.: Estimating the dimension of a model. In: *The Annals of Statistics* 6 (1978), S. 461–464
- [Self u. a. 1988] SELF, S.G. ; MAURITSEN, R.H. ; OHARA, J.: Power calculations for likelihood ratio tests in generalized linear models. In: *Biometrics* 44 (1988), S. 79–86
- [Sherman und Le Cessie 1997] SHERMAN, M. ; LE CESSIE, S.: A comparison between bootstrap methods and generalized estimating equations for correlated outcomes in generalized linear models. In: *Communications in Statistics - Simulation and Computation* 26 (1997), S. 901–925
- [Signorini 1991] SIGNORINI, D.F.: Sample size for Poisson regression. In: *Biometrika* 78 (1991), S. 446–450

- [White 1981] WHITE, H.: Consequences and detection of misspecified nonlinear regression models. In: *Journal of the American Statistical Association* 76 (1981), S. 419–433
- [Willemain und Desautels 1993] WILLEMAIN, T.R. ; DESAUTELS, P.A.: A method to generate autocorrelated uniform random numbers. In: *Journal of Statistical Computation and Simulation* 45 (1993), S. 23–31
- [Wolfinger und O’Connell 1993] WOLFINGER, Russ ; O’CONNELL, Michael: Generalized linear mixed models: A pseudo-likelihood approach. In: *Journal of Statistical Computation and Simulation* 48 (1993), S. 233–243
- [Zeger und Liang 1986] ZEGER, Scott L. ; LIANG, Kung-Yee: Longitudinal data analysis for discrete and continuous outcomes. In: *Biometrics* 42-1 (1986), S. 121–130
- [Ziegler 1999] ZIEGLER, Andreas: *Pseudo Maximum Likelihood Methode und Generalised Estimating Equations zur Analyse korrelierter Daten*. Peter Lang, 1999
- [Ziegler und Grömping 1998] ZIEGLER, Andreas ; GRÖMPING, Ulrike: The Generalised Estimating Equations: A comparison of procedures available in commercial statistical software packages. In: *Biometrical Journal* 40 (1998), S. 245–260
- [Ziegler u. a. 1998] ZIEGLER, Andreas ; KASTNER, Christian ; BLEITNER, Maria: The Generalised Estimating Equations: An annotated bibliography. In: *Biometrical Journal* 40-2 (1998), S. 115–139

## **Erklärung**

Hiermit versichere ich, dass ich diese Arbeit selbständig verfasst und keine anderen als die angegebenen Hilfsmittel und Quellen benutzt habe.