Carl von Ossietzky
Universität Oldenburg
Diplomstudiengang Mathematik

DIPLOMARBEIT

Modellwahl und Versuchsplanung in Studien mit wiederholten binären Beobachtungen

vorgelegt von: Mona Hagedorn
vorgelegt am: 01.02.2007

Betreuende Gutachterin: Prof. Dr. Christine Müller
Zweiter Gutachter: Prof. Dr. Dietmar Pfeifer
Auswärtiger Berater bei
Themenstellung und Betreuung: Dr. Norbert Benda

Inhaltsverzeichnis

1 Einleitung **1**

2 Grundlagen **3**

3 Modelle für wiederholte binäre Beobachtungen **8**

3.1 Marginale Modelle 9

3.1.1 Das George-Bowman-Modell 12

3.1.1.1 Modelldarstellung 12

3.1.1.2 Modellwahl für die Parameter $\lambda_{i,k}$ 17

3.1.1.3 Maximum-Likelihood-Schätzungen 19

3.1.2 Das Bahadur-Modell 21

3.1.2.1 Modelldarstellung 21

3.1.2.2 Modellwahl für die Parameter $\rho_i(2)$ und π_i 28

3.1.2.3 Maximum-Likelihood-Schätzungen 29

3.1.3 Schätzmethoden 31

3.1.3.1 Generalized Estimating Equations (GEEs) 32

3.1.3.1.1 Allgemeine Theorie der GEEs 32

3.1.3.1.2 Schätzen des Parameters β 36

3.2 Subjektspezifische Modelle 38

3.2.1 Das Beta-Binomial-Modell 41

3.2.1.1 Modelldarstellung 41

3.2.1.2 Modellwahl für die Parameter π_i und ρ_i 45

3.2.1.3 Maximum-Likelihood-Schätzungen 46

3.2.2 Verallgemeinerte lineare gemischte Modelle 47

3.2.2.1 Die logistische Regression mit gemischten Effekten 48

3.2.2.2 Allgemeine Darstellung 48

3.2.2.3 Maximum-Likelihood-Schätzungen 49

4 Vergleich der Modellklassen **51**

4.1 Vergleich der Modellklassen für binäre Daten mit dem Logit-Link 51

4.1.1 Indonesian Children's Health Study 54

4.2 Vergleich des Bahadur- und des Beta-Binomial-Modells 56

5 Versuchsplanung **61**

5.1 Berechnung des benötigten Stichprobenumfangs 62

5.1.1 Eine Beobachtung pro Person 64

5.1.2 Wiederholte Beobachtungen 70

5.1.2.1 Das Verfahren zur Fallzahlberechnung von Liu und Liang 71

5.2 Stichprobengröße versus Wiederholungen 81

5.2.1 Beispiel 82

6 Simulationen **90**

6.1 Erzeugung wiederholter binärer Daten 90

6.2 Auswertung der Daten 91

6.3 Ergebnisse der Simulationen 94

7 Schlussbemerkungen **97**

Anhang **98**

A Erstellung von Abbildung 1 **99**

B Monotonie **102**

C Funktionen von Beispiel 5.2.1 **103**

D Erzeugung simulierter binärer Daten **108**

D.1 Eine Beobachtung pro Subjekt 108

D.2 Zwei Beobachtungen pro Subjekt 111

D.3 Drei Beobachtungen pro Subjekt 116

E Auswertung der Daten **122**

Abbildungsverzeichnis

1 Graphische Darstellung von bedingten Erwartungen eines Logit-Modells mit verschiedenen Realisierungen des Random Effects zusammen mit der marginalisierten Erwartung. 52
2 Verschobene Normalverteilungen 66
3 Gütefunktion 70
4 Ausgabe der Funktion *zeichnenpA* 84
5 Ausgabe der Funktion *zeichnenpAcl* 85
6 Ausgabe der Funktion *kostenfunktionpAcl* 86
7 „Sprungstellen“ in der Kostenfunktion 88

Tabellenverzeichnis

1 Schranken für $\rho_i(2)$. 59
2 Beispiel für Datei *schaetzer* 94
3 Beispiel für Datei *parameterbeta1* 94
4 Auswertung der Simulationen für den Parameter β_1. 96

1 Einleitung

In dieser Arbeit werden wiederholte binäre Beobachtungen untersucht. Binäre Beobachtungen sind Beobachtungen mit nur zwei möglichen Ergebnissen, zum Beispiel, ob ein untersuchtes Ereignis aufgetreten ist oder nicht, Erfolg und Misserfolg oder An- und Auszustand. In der Praxis treten binäre Beobachtungen z.B. bei toxikologischen Experimenten mit schwangeren Mäusen auf. Das beobachtete unerwünschte Ereignis ist dann das Vorkommen von Missbildungen bei den Föten. Auch stetige Daten können in binäre Daten transformiert werden, indem ein bestimmter Wert als Grenze für beide Zustände benutzt wird, z.B. die Beobachtung des Blutdrucks in pharmazeutischen Experimenten. Gemessen wird, ob der Blutdruck größer oder kleiner als 140 mmHg ist.[1] Allerdings ist dies eher ein untypisches Beispiel; im Allgemeinen werden andere Laborparameter, die einen Grenzwert überschreiten, als Auftreten eines unerwünschten Ereignisses gewertet.

Von wiederholten binären Beobachtungen wird gesprochen, wenn mehrere Untersuchungen des gleichen Merkmals bei einem Subjekt bzw. bei einer Untersuchungseinheit gemacht werden. Sobald mehrere Subjekte untersucht werden, können Abhängigkeiten zwischen den Beobachtungen eines Subjekts auftreten, da die verschiedenen Subjekte jeweils ein unterschiedliches Basisniveau aufweisen. Zum Beispiel können bei einer Gruppe von Personen simultan beide Augen untersucht werden. Die Daten der Augen einer Person sind dann voneinander abhängig, da sie dieselben genetischen Grundvoraussetzungen haben. Bei anderen Personen sind diese Voraussetzungen wieder anders. Ein weiteres Beispiel sind klinische Studien, die Daten aus unterschiedlichen Kliniken verwenden. Die Daten aus einer Klinik weisen dann aufgrund der gleichen örtlichen Verhältnisse eine Abhängigkeitsstruktur, einen „Klinikeffekt“ auf. Im Allgemeinen werden solche Daten *geclustert* oder *korreliert* genannt.

[1] Ein Blutdruck größer als 140 mmHg bedeutet Hypertonie.

Im Folgenden werden für die Auswertung wiederholter binärer Daten verschiedene Modelle vorgestellt. Jedes Modell enthält Parameter, die mit Hilfe der Daten eines Experiments geschätzt werden können. Diese Parameter können dann zur Prognose über den möglichen Ausgang einer Wiederholung des Experiments benutzt werden. Dies geschieht zum Beispiel in klinischen Tests zur Zulassung eines Medikaments.

Generell werden drei Klassen von Modellen unterschieden: die marginalen Modelle, die subjekt-spezifischen (cluster-spezifischen) Modelle und Modelle, die den Ausgang der nächsten Beobachtung bedingt auf den vorherigen Beobachtungen modellieren. Im Rahmen dieser Arbeit wird allerdings nur auf die marginalen Modelle und auf eine spezielle Unterklasse der subjekt-spezifisches Modelle, der random-effects-Modelle, eingegangen. Im Folgenden werden die jeweiligen Anwendungsgebiete erläutert und zu jeder Klasse ein bis zwei spezielle Modelle vorgestellt und analysiert. Insbesondere wird auf die Anzahl der nötigen Parameter eingegangen. Nach diesen Modellvorstellungen erfolgt in Kapitel 5 ein Abschnitt über Versuchsplanung, in dem ein Verfahren von Liu u. Liang (1997) zur Herleitung einer Fallzahlformel für marginale Modelle mit wiederholten Beobachtungen vorgestellt wird. Diese Formel gibt die mindestens benötigte Anzahl von zu untersuchenden Subjekten an, damit ein statistischer Test mit hoher Wahrscheinlichkeit einen signifikanten Unterschied in den Daten erkennt. Für ein konkretes Beispiel wird mit Hilfe dieser Formel in Abhängigkeit der Wiederholungen pro Subjekt der jeweils benötigte Stichprobenumfang ermittelt. Mit diesen Ergebnissen kann die Anzahl von Wiederholungen bestimmt werden, bei der die gesamten Kosten einer Studie minimal wären. Im letzten Kapitel werden Simulationen durchgeführt, die die Fallzahlformel für ein Beispiel testen sollen. Diese Simulationen und Auswertungen erfolgen mit den Statistik-Programmen R und SAS.

2 Grundlagen

Zunächst werden einige Grundlagen erläutert, die in dieser Arbeit benutzt werden.

Bemerkung 1. Es werden Experimente betrachtet, die wiederholte binäre Daten bzw. Beobachtungen liefern. Solange nichts Gegenteiliges behauptet wird, wird einfachheitshalber davon ausgegangen, dass die Beobachtungen von Personen bzw. Subjekten stammen. Natürlich sind auch andere Untersuchungseinheiten denkbar. Insgesamt werden N Subjekte beobachtet und pro Subjekt i werden n_i Beobachtungen gemacht.

Bemerkung 2. Die Beobachtung bei Subjekt i und der j-ten Wiederholung wird mit y_{ij} bezeichnet. Diese Beobachtung wird als Realisierung einer Zufallsvariable Y_{ij} angesehen. Verkürzt wird auch die Zufallsvariable schon als Beobachtung bezeichnet. Die Zufallsvariablen Y_{ij} werden zu einem Zufallsvektor Y_i zusammengefasst.

Bemerkung 3. Die n_i Zufallsvariablen Y_{ij} von Subjekt i sind untereinander abhängig bzw. korreliert. Dies wird durch die Aussage, dass die Beobachtungen korreliert seien, vereinfacht ausgedrückt. Diese korrelierten Beobachtungen bilden ein sogenanntes Cluster.

Bemerkung 4. Die Zufallsvariablen Y_{ij} werden durch ein Modell mit den dazugehörigen erklärenden Variablen $x_{ij} \in \mathbb{R}^p$, also den bekannten oder auch vermuteten Faktoren, denen ein Einfluss auf die mögliche Beobachtung des Experiments nachgesagt wird, in Verbindung gebracht. Der Parametervektor β beschreibt diesen Einfluss und wird mit Hilfe der Beobachtungen geschätzt, um den möglichen Ausgang einer Wiederholung des Experiments vorherzusagen. Die Korrelation der Beobachtungen bei einem Subjekt wird von dem benutzten Modell berücksichtigt.

Bemerkung 5. Viele Modelle beschreiben mit ihrer Annahme über die Verteilung der Zufallsvariable gut den beobachteten Erwartungswert. Allerdings ist die

Varianz der tatsächlich beobachteten Daten meist größer als die im Modell angenommene. Dieser Zustand wird *overdispersion* genannt. In den meisten Modellen ist ein Parameter für *overdispersion* enthalten.

Bemerkung 6. Um die zwei im Folgenden betrachteten verschiedenen Ausprägungen der Modelle für wiederholte binäre Beobachtungen gegenüberzustellen, wird zur Illustration eine Studie benutzt, nämlich die Indonesian Children's Health Study (ICHS), die in Diggle u. a. (2002) kurz erläutert wird. In dieser Studie wurden die Gründe und der Effekt von Vitamin-A-Mangel bei Vorschulkindern untersucht. In einem vierteljährlichen Abstand wurden bis zu sechs Untersuchungen bei über 3000 Kindern gemacht. Dabei wurde überprüft, ob die Kinder unter Atemwegsinfektionen, Diarrhoe oder Xerophthalmie, einer Augenkrankheit, die auf Vitamin-A-Mangel zurückzuführen ist, leiden. Ebenso wurden Gewicht und Größe gemessen. Es wird nun die Frage betrachtet, ob Kinder mit Vitamin-A-Mangel ein erhöhtes Risiko für Atemwegsinfektionen haben.

Für unabhängige Daten kann das *verallgemeinerte lineare Modell* benutzt werden (z.B. $n_i = 1$).

Definition 7. Das *verallgemeinerte lineare Modell (GLM = generalized linear model)* besteht aus drei Komponenten:

1. den beobachteten Zufallsvektor $Y = (Y_1, \ldots, Y_N)^\top \in \mathbb{R}^N$ und dessen Verteilung (Zufallskomponente),

2. dem linearen Prädiktor $x_1, \ldots x_n \in \mathbb{R}^p$ (systematische Komponente, erklärende Variablen),

3. der Linkfunktion h.

Der Zufallsvektor $Y = (Y_1, \ldots, Y_N)\top \in \mathbb{R}^N$ besteht aus N unabhängigen Beobachtungen, die mit einer Verteilung aus einer 1-parametrischen natürlichen Exponential-Familie mit Dispersionsparameter (siehe Def. 8) verteilt sind. Die

Abhängigkeitsstruktur des Zufallsvektors Y von den erklärenden Variablen wird dann durch folgende Gleichung beschrieben:

$$h(\mathrm{E}(Y_i)) = x_i^\top \beta.$$

Dabei ist $\beta \in \mathbb{R}^p$ der zu schätzende Parametervektor.
Mit $X = (x_1, \ldots, x_N) \in \mathbb{R}^{p \times N}$ (Planungsmatrix) lässt sich somit schreiben

$$h(\mathrm{E}(Y)) = X^\top \beta,$$

wobei $h(\mathrm{E}(Y)) \in \mathbb{R}^N$ der Vektor mit den Einträgen $h(\mathrm{E}(Y_i))$ ist.

Definition 8. Besitzt die Zufallsvariable $Y_i \in \mathbb{R}$ eine Dichte der Form

$$f_{\theta,\phi}(y_i) = \exp\left\{ [y_i\theta - b(\theta)] \frac{w}{\phi} + c(y_i, \phi) \right\},$$

wobei $\theta, \phi \in \mathbb{R}$ unbekannt und $w \in \mathbb{R},\ b : \mathbb{R} \to \mathbb{R},\ c : \mathbb{R}^2 \to \mathbb{R}$ bekannt sind, dann besitzt Y_i eine Verteilung einer *1-parametrischen natürlichen Exponentialfamilie mit Dispersionsparameter ϕ.*[2]

Beispiel 9. Die Zufallsvariable $Y_i = \frac{\tilde{Y}_i}{M}$ mit $\tilde{Y}_i \sim B(M, p)$ (Binomialverteilung mit Parametern M und p) besitzt eine Verteilung einer 1-parametrischen natürlichen Exponentialfamilie mit Dispersionsparameter ϕ, denn es gilt:

$$\begin{aligned}
f_{M,p}(y_i) &= \mathrm{P}_{B(M,p)}(My_i) = \binom{M}{My_i} p^{My_i}(1-p)^{M-My_i} \\
&= \exp\left\{ \ln\left(\binom{M}{My_i}\right) + My_i \ln(p) + (M - My_i)\ln(1-p) \right\} \\
&= \exp\left\{ My_i(\ln(p) - \ln(1-p)) + M\ln(1-p) + \ln\left(\binom{M}{My_i}\right) \right\} \\
&= \exp\left\{ \left[y_i \ln\left(\frac{p}{1-p}\right) - (-\ln(1-p)) \right] M + \ln\left(\binom{M}{My_i}\right) \right\} \\
&= \exp\left\{ [y_i\theta - b(\theta)]\frac{w}{\phi} + c(y_i, \phi) \right\}
\end{aligned}$$

[2] vgl. McCullagh u. Nelder (1989)

mit $\theta = \ln\left(\frac{p}{1-p}\right)$, also $p = \frac{e^\theta}{1+e^\theta}$ und $b(\theta) = -\ln(1-p) = \ln(1+e^\theta)$, $w = M$, $\phi = 1$ und $c(y_i, \phi) = \ln\left(\binom{M}{My_i}\right)$ für $y_i \in \{0, \frac{1}{M}, \frac{2}{M}, \ldots, 1\}$.
Insbesondere gilt die Darstellung für $M = 1$, d.h. für die Bernoulli-Verteilung $Y_i = \tilde{Y}_i \sim B(1, p)$.

Bemerkung 10. Wenn Y_i eine Verteilung einer 1-parametrischen natürlichen Exponentialfamilie mit Dispersionsparameter ϕ, kurz Exponentialfamilie, besitzt, dann ist

$$\mathrm{E}(Y_i) = b'(\theta) \text{ und } \mathrm{Var}(Y_i) = b''(\theta)\frac{\phi}{w}.$$

Zum Beweis dieser Aussage siehe McCullagh u. Nelder (1989), S. 29.
Für Bernoulli-verteilte Zufallsvariablen ist dann

$$\begin{aligned} \mathrm{E}(Y_i) &= b'(\theta) = \frac{\exp(\theta)}{1+\exp(\theta)} = p \quad \text{und} \\ \mathrm{Var}(Y_i) &= b''(\theta)\frac{\phi}{w} = \frac{\exp(\theta)}{(1+\exp(\theta))^2} = p(1-p). \end{aligned}$$

Definition 11. Seien die Zufallsvariablen $Y_1, \ldots, Y_n$ gegeben. Dann ist eine *Korrelation k-ter Ordnung* ρ_k der Zufallsvariablen $Y_{i_1}, \ldots, Y_{i_k}$ $(k \leq n)$ definiert durch

$$\rho_k = \frac{\mathrm{E}[(Y_{i_1} - \mathrm{E}(Y_{i_1}))(Y_{i_2} - \mathrm{E}(Y_{i_2})) \cdot \ldots \cdot (Y_{i_k} - \mathrm{E}(Y_{i_k}))]}{(\mathrm{Var}(Y_{i_1})\,\mathrm{Var}(Y_{i_2}) \cdot \ldots \cdot \mathrm{Var}(Y_{i_k}))^{1/2}}.$$

Definition 12. Die *Stirlingnummer $S(k, j)$ der zweiten Art* ist die Anzahl der Möglichkeiten, eine k-elementige Menge in j nichtleere Menge zu teilen. Nach Vereinbarung ist $S(0, 0) = 1$.

Satz 13. *Die Stirlingnummer der zweiten Art lässt sich nach der Formel*

$$S(k, j) = \frac{1}{j!}\sum_{i=0}^{j}(-1)^{j-i}\binom{j}{i}i^k$$

berechnen. Außerdem gilt die Gleichung

$$\sum_{k \geq j} S(k, j)\frac{t^k}{k!} = \frac{(e^t - 1)^j}{j!}, \quad j \geq 0.$$

Beweis. Der Beweis ist in Stanley (1997), S. 34 zu finden. □

Bemerkung 14. Für eine Funktion $f : \mathbb{R}^p \to \mathbb{R}^N,\ x \mapsto f(x)$ ist die Ableitung nach dem Vektor x gegeben durch

$$\frac{\partial f(x)}{\partial x} = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_p} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_N}{\partial x_1} & \cdots & \frac{\partial f_N}{\partial x_p} \end{pmatrix} \in \mathbb{R}^{N \times p}.$$[3]

Insbesondere gilt für $X \in \mathbb{R}^{p \times N},\ \beta \in \mathbb{R}^p$:

$$\frac{\partial X^\top \beta}{\partial \beta} = X^\top.$$

Der für eine Funktion $f : \mathbb{R}^p \to \mathbb{R}^1$ durch Ableitung entstehende Vektor wird *Gradient* genannt. Wird die Transponierte des Gradienten erneut nach x abgeleitet, ergibt sich gerade die Transponierte der Hesse-Matrix. Bei stetigen zweiten Ableitungen ist wegen der Vertauschbarkeit der Differentiationsreihenfolge (Satz von Schwarz) diese Matrix symmetrisch, so dass das Transponieren der Matrix keine Änderung bewirkt.

[3] Erwe (1962)

3 Modelle für wiederholte binäre Beobachtungen

Für die Auswertung wiederholter binärer Daten gibt es nach Aerts u. a. (2002) verschiedene Ansätze. Die einfachste Möglichkeit besteht darin, alle Daten als unabhängig anzunehmen und mit Modellen für unabhängige Beobachtungen auszuwerten. Allerdings wird in diesem Fall die verfügbare Information der Korrelation vernachlässigt und deshalb ist diese Methode für die Auswertung nicht gut geeignet.

Eine weitere Möglichkeit ist, dass die Auswertung auf eine Beobachtung pro Subjekt beschränkt bleibt, z.B. dass von den Beobachtungen beider Augen nur die Beobachtung des schlechteren Auges in die Auswertung eingeht. Diese Modelle werden "*response feature models*" genannt. Aber auch bei dieser Möglichkeit gehen Informationen verloren.

Bei stetigen und asymptotisch normalverteilten Daten gibt es eine große Klasse von linearen Modellen, die die Korrelation berücksichtigen. Bei kategoriellen, insbesondere binären Daten, gibt es nicht so viele ausgereifte Techniken. Es existiert für dieses Problem keine Standardlösung und die vielen unterschiedlichen Ansätze fallen nicht wie bei der Normalverteilung zusammen. Für binäre Daten fehlt das multivariate Analogon zur Normalverteilung.

Eine Übersicht über die verschiedenen Auswertungstechniken von wiederholten binären Beobachtungen geben die Artikel von Pendergast u. a. (1996), Molenberghs u. Verbeke (2004), Hu u. a. (1998) und Neuhaus u. a. (1991), sowie die Bücher von Aerts u. a. (2002), Diggle u. a. (2002) und Molenberghs u. Verbeke (2005).

Im Großen und Ganzen lassen sich drei Klassen von Modellen für wiederholte binäre Beobachtungen unterscheiden: die marginalen Modelle, die subjekt- bzw. cluster-spezifischen Modelle und die bedingten Modelle. Im Rahmen dieser Arbeit

wird nur auf die ersten beiden Klassen eingegangen. Nach Diggle u. a. (2002) sind marginale Modelle für korrelierte Daten das natürliche Analogon zu verallgemeinerten linearen Modellen für unabhängige Daten. Marginale Modelle enthalten Parameter, die Aussagen über die durchschnittlich zu erwartende Beobachtung machen (Aerts u. a. (2002)). Sie werden auch population-averaged-Modelle genannt. Ihnen stehen subjektspezifische Modelle gegenüber, die Parameter enthalten, die sich direkt auf die Subjekte bzw. Cluster beziehen. Eine spezielle Klasse von subjektspezifischen Modellen sind die random-effects-Modelle. Fälscherlicherweise werden in mancher Literatur die subjektspezifischen Modelle schon direkt als random-effects-Modelle bezeichnet.
Welches Modell benutzt werden sollte, hängt von den untersuchten Fragestellungen ab. Bei klinischen Test zur Zulassung von Medikamenten ist zum Beispiel von Interesse, ob sich ein Medikament in seiner Wirkungsweise von einem anderen unterscheidet. Zur Beantwortung kann ein marginales Modell benutzt werden, das den durchschnittlich zu erwartenden Behandlungsunterschied beschreibt. Sollte die Wirkungsweise eines Medikaments direkt auf einen Probanden untersucht werden, z.B. ein Medikament für HIV-positive Patienten, kommen subjekt- bzw. clusterspezifische Parameter hinzu. In diesem Fall wird ein subjektspezifisches Modell benutzt.

3.1 Marginale Modelle

Bei marginalen Modellen wird laut Diggle u. a. (2002) der Einfluss der erklärenden Variablen auf die Beobachtung getrennt von der Korrelation der Beobachtungen bei einem Subjekt modelliert. Dabei wird die marginale Erwartung als eine Funktion der erklärenden Variablen modelliert. Erklärende Variablen könnten z.B. die untersuchten Medikamente, Alter und Geschlecht der Probanden oder die Giftkonzentration in toxikologischen Experimenten sein. Mit marginaler Erwartung wird die durchschnittliche Antwort einer Subpopulation mit den glei-

chen erklärenden Variablen bezeichnet.

Als Beispiel sei ein Experiment gegeben, in dem bei einer festgelegten Anzahl von Probanden wiederholte Beobachtungen gemacht werden. Die Probanden erhalten z.B. zwei unterschiedliche Behandlungen. Von Interesse ist der Behandlungseffekt, der marginal gemessen wird. Die Unterschiede in den Reaktionen der einzelnen Subjekte innerhalb der gleichen Behandlungsgruppe werden dabei nicht berücksichtigt.

Das allgemeine marginale Modell lässt sich nun durch folgende drei Punkte charakterisieren:[4]

1. Es ist $\mathrm{E}(Y_{ij}) = \mu_{ij}$ mit $h(\mu_{ij}) = x_{ij}^T\beta$, wobei h eine Linkfunktion, $x_{ij} \in \mathbb{R}^p$ die erklärenden Variablen von Person i bei der j-ten Beobachtung und $\beta \in \mathbb{R}^p$ der zu schätzende Parametervektor ist.

2. Die marginale Varianz hängt von der marginalen Erwartung ab:

$$\mathrm{Var}(Y_{ij}) = v(\mu_{ij})\phi,$$

   wobei v eine bekannte Varianzfunktion und ϕ ein Dispersionsparameter ist, der eventuell geschätzt werden muss.

3. Die Korrelation zwischen Y_{ij} und Y_{ik} ist eine Funktion der marginalen Erwartungswerte und eventuell eines zusätzlichen Parameters α:

$$\mathrm{Corr}(Y_{ij}, Y_{ik}) = \rho(\mu_{ij}, \mu_{ik}, \alpha),$$

   wobei $\rho(\cdot)$ eine bekannte Funktion ist.

Beispiel 15. In Diggle u. a. (2002) wird ein Beispiel für ein marginales logistisches Modell mit Intercept gegeben. Dort wird die Indonesian Children's Health Study (ICHS) zu Grunde gelegt (s.a. Bemerkung 6). In dieser Studie wird das Auftreten von Atemwegsinfektionen im Zusammenhang mit Vitamin-A-Mangel

[4] vgl. Diggle u. a. (2002)

untersucht. Sei z_{ij} die erklärende binäre Variable, die anzeigt, ob das Kind i bei der j-ten Untersuchung einen Vitamin-A-Mangel aufwies oder nicht ($z_{ij} = 1$ für ja, 0 für nein). Dann ist $x_{ij}^\top = (1, z_{ij})$. Sei Y_{ij} die binäre Beobachtung der Atemwegsinfektion ($Y_{ij} = 1$ für ja, 0 für nein). Dann kann folgendes logistisches Modell aufgestellt werden:

$$\text{logit } (\mu_{ij}) \;=\; \log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) \;=\; \log\left(\frac{\mathrm{P}(Y_{ij}=1)}{\mathrm{P}(Y_{ij}=0)}\right) \;=\; \beta_0 + \beta_1 z_{ij}.$$

Für die Varianz lässt sich die der Binomialverteilung benutzen; die Korrelation wird als konstant angenommen:

$$\text{Var } (Y_{ij}) = \mu_{ij}(1-\mu_{ij}) \quad \text{und} \quad \text{Corr } (Y_{ij}, Y_{ik}) = \alpha.$$

Der Interceptparameter β_0 ist der Logarithmus des Bruchs der Häufigkeiten von infizierten zu nicht infizierten Kindern in der Subpopulation der Kinder, die nicht unter Vitamin-A-Mangel leiden. Der Parameter β_1 ist der Logarithmus des Bruchs der Odds der Infektionen bei den Kindern mit Vitamin-A-Mangel und den Kindern ohne Mangel. Damit ist $\exp(\beta_1)$ ein Bruch von Populationshäufigkeiten, also ein population-averaged Parameter. Die Populationshäufigkeit ist der Durchschnitt der individuellen Risiken der Personen mit den gleichen erklärenden Variablen. Mit diesem marginalen Modell kann nun z.B. folgende Frage betrachtet werden:

- Ist die Verbreitung von Atemwegsinfektionen größer in der Subpopulation der Kinder mit Vitamin-A-Mangel?

Dies ist eine marginale Fragestellung, denn es werden Populationen von Kindern verglichen.
Diesem Beispiel wird später ein Beispiel eines subjektspezifischen Modells für die Indonesian Children's Health Study gegenübergestellt.

Im Allgemeinen sind z.B. das George-Bowman-Modell und das Bahadur-Modell zur Auswertung von wiederholten binären Beobachtungen geeignet. Auf diese zwei Modelle wird im Folgenden weiter eingegangen.

3.1.1 Das George-Bowman-Modell

Das George-Bowman Modell ist in der Literatur z.B. in Aerts u. a. (2002) und bei George u. Bowman (1995) zu finden. Es ist ein marginales Modell und kann zur Analyse von austauschbaren (exchangeable) binären Daten mit einer Likelihoodprozedur verwendet werden.

3.1.1.1 Modelldarstellung

Definition 16. Die Zufallsvariablen $X_1, X_2, ...$ sind *austauschbar*, falls für jedes $n \in \mathbb{N}$, für jeden Vektor $(x_1, x_2, \ldots, x_n)$ und für jede Permutation $\pi(1), \ldots, \pi(n)$ gilt:

$$\mathrm{P}(X_{\pi(1)} = x_1, \ldots, X_{\pi(n)} = x_n) = \mathrm{P}(X_1 = x_1, \ldots, X_n = x_n).^{5}$$

Die Eigenschaft der Austauschbarkeit ist gut geeignet, um die Abhängigkeit von Daten/ bzw. Beobachtungen von einem Subjekt (in einem Cluster) zu beschreiben. Sie wird zum Beispiel in toxikologischen Experimenten mit schwangeren Mäusen angenommen. Dabei gibt es Kontrollgruppen und Gruppen von schwangeren Tieren, die dem Toxin ausgesetzt werden. Mehrere Gruppen erhalten die gleiche Gift-Konzentration. Im Anschluss werden dann die Föten auf Missbildungen untersucht. Es interessiert nur die Anzahl der Missbildungen, nicht die Abfolge der Beobachtungen.[6]

Seien $(Y_{i1}, Y_{i2}, \ldots, Y_{in_i})$ die n_i binären Beobachtungen der Gruppe (Cluster) i, wobei für $k = 1, \ldots, n_i$ der Ausdruck $Y_{ik} = 1$ bedeutet, dass der k-te Fötus missgebildet ist und $Y_{ik} = 0$, dass der Fötus normal gewachsen ist. Dann ist die gemeinsame Verteilung von $(Y_{i1}, \ldots, Y_{in_i})$ invariant unter Permutationen der Indizes $(1, \ldots, n_i)$[7].

Für die Wahrscheinlichkeitsfunktion der Anzahl $Z_i = \sum_j Y_{ij}$ der Missbildungen bei Gruppe i mit n_i Beobachtungen gibt es einen geschlossenen Ausdruck:

[5] vgl. George u. Bowman (1995)

[6] vgl. George u. Bowman (1995)

[7] George u. Bowman (1995)

Satz 17. [8] *Seien $Y_{i1}, Y_{i2}, \ldots, Y_{in_i}$ austauschbare binäre Zufallsvariablen und sei*

$$\lambda_{i,k} := \begin{cases} \mathrm{P}(Y_{i1} = 1, Y_{i2} = 1, \ldots, Y_{ik} = 1) & \textit{für} \quad k = 1, \ldots, n_i, \\ 1 & \textit{für} \quad k = 0. \end{cases}$$

($\lambda_{i,k}$ ist also die Wahrscheinlichkeit, dass bei Gruppe (Cluster) i alle Beobachtungen in einer Menge von k Beobachtungen Erfolge (Missbildungen) sind.)

Dann gilt mit $Z_i = \sum_{j=1}^{n_i} Y_{ij}$ und $z_i = \sum_{j=1}^{n_i} y_{ij}$:

$$\mathrm{P}(Y_{i1} = y_{i1}, \ldots, Y_{in_i} = y_{in_i}) = \sum_{k=z_i}^{n_i} (-1)^{k-z_i} \binom{n_i - z_i}{k - z_i} \lambda_{i,k} \tag{1}$$

und

$$\begin{aligned} \mathrm{P}(Z_i = z_i) &= f(z_i; \lambda_{i,z_i}, \lambda_{i,z_i+1}, \ldots, \lambda_{i,n_i}, n_i) && (2) \\ &= \binom{n_i}{z_i} \sum_{k=z_i}^{n_i} (-1)^{k-z_i} \binom{n_i - z_i}{k - z_i} \lambda_{i,k}. && (3) \end{aligned}$$

Beweis. Der Beweis benutzt ein Inklusions-Exklusions-Prinzip.
Da die Zufallsvariablen $Y_{i1}, \ldots, Y_{in_i}$ austauschbar sind, genügt es zu zeigen, dass $\mathrm{P}(Y_{i1} = 1, \ldots, Y_{iz_i} = 1, Y_{iz_i+1} = 0, \ldots, Y_{in_i} = 0)$ durch Ausdruck (1) gegeben ist. Laut Definition ist

$$\begin{aligned} \lambda_{i,z_i} &= \mathrm{P}(Y_{i1} = 1, \ldots, Y_{iz_i} = 1) \\ &= \sum_{y_{iz_i+1}, \ldots, y_{in_i}} \mathrm{P}(Y_{i1} = 1, \ldots, Y_{iz_i} = 1, Y_{iz_i+1} = y_{iz_i+1}, \ldots, Y_{in_i} = y_{in_i}) \\ &= \mathrm{P}(Y_{i1} = 1, \ldots, Y_{iz_i} = 1, Y_{iz_i+1} = 0, \ldots, Y_{in_i} = 0) \\ &\quad + \binom{n_i - z_i}{1} \sum_{y_{iz_i+2}, \ldots, y_{in_i}} \mathrm{P}(Y_{i1} = 1, \ldots, Y_{iz_i+1} = 1, Y_{iz_i+2} = y_{iz_i+2}, \ldots, Y_{in_i} = y_{in_i}) \\ &\quad - \binom{n_i - z_i}{2} \sum_{y_{iz_i+3}, \ldots, y_{in_i}} \mathrm{P}(Y_{i1} = 1, \ldots, Y_{iz_i+2} = 1, Y_{iz_i+3} = y_{iz_i+3}, \ldots, Y_{in_i} = y_{in_i}) \\ &\quad + \cdots + (-1)^{n_i - z_i + 1} \mathrm{P}(Y_{i1} = 1, \ldots, Y_{in_i} = 1) \end{aligned}$$

[8] Satz 17 und Beweis stammen aus George u. Bowman (1995).

$$
\begin{aligned}
&= \mathrm{P}(Y_{i1} = 1, \ldots, Y_{iz_i} = 1, Y_{iz_i+1} = 0, \ldots, Y_{in_i} = 0) \\
&\quad + \sum_{k=z_i+1}^{n_i} (-1)^{k-z_i+1} \binom{n_i - z_i}{k - z_i} \lambda_{i,k}.
\end{aligned}
$$

Daraus folgt:

$$
\mathrm{P}(Y_{i1} = 1, \ldots, Y_{iz_i} = 1, Y_{iz_i+1} = 0, \ldots, Y_{in_i} = 0) = \sum_{k=z_i}^{n_i} (-1)^{k-z_i} \binom{n_i - z_i}{k - z_i} \lambda_{i,k}.
$$

Gleichung (3) folgt direkt aus (1), da $\mathrm{P}(Y_{i\pi(1)} = y_{i1}, \ldots, Y_{i\pi(n_i)} = y_{in_i})$ das Gleiche für alle Permutationen $\pi(1), \ldots, \pi(n_i)$ ist und es $\binom{n_i}{z_i}$ Möglichkeiten gibt z_i Einsen auf n_i Stellen zu verteilen. □

Korollar 18. Die Ausdrücke (1) und (3) verallgemeinern die binomiale Verteilung; wenn die Bernoulli-verteilten Zufallsvariablen Y_{ij} unabhängig sind, dann wird $\lambda_{i,k}$ zu $\lambda_{i,1}^k$ für $k = 1, \ldots, n_i$ und Z_i zu einer binomialverteilten Zufallsvariable mit Parametern n_i und $\lambda_{i,1}$.

Beweis. Es ist

$\lambda_{i,k} = \mathrm{P}(Y_{i1} = 1, \ldots, Y_{ik} = 1) = \prod_{j=1}^{k} \mathrm{P}(Y_{ij} = 1) = \prod_{j=1}^{k} \lambda_{i,1} = \lambda_{i,1}^k$

und

$$
\begin{aligned}
\mathrm{P}(Z_i = z_i) &= \binom{n_i}{z_i} \sum_{k=z_i}^{n_i} (-1)^{k-z_i} \binom{n_i - z_i}{k - z_i} \lambda_{i,k} \\
&= \binom{n_i}{z_i} \sum_{k=z_i}^{n_i} (-1)^{k-z_i} \binom{n_i - z_i}{k - z_i} \lambda_{i,1}^k \\
&= \binom{n_i}{z_i} \sum_{k=0}^{n_i - z_i} (-1)^{k} \binom{n_i - z_i}{k} \lambda_{i,1}^{k+z_i} \\
&= \binom{n_i}{z_i} \lambda_{i,1}^{z_i} \sum_{k=0}^{n_i - z_i} \binom{n_i - z_i}{k} (-\lambda_{i,1})^k 1^{n_i - z_i - k} \\
&= \binom{n_i}{z_i} \lambda_{i,1}^{z_i} (1 - \lambda_{i,1})^{n_i - z_i}.
\end{aligned}
$$

□

Der Erwartungswert, die Varianz und die Korrelation können durch den Parameter $\lambda_{i,k}$ ausgedrückt werden:

$$\begin{aligned}
\mathrm{E}(Z_i) &= \sum_{j=1}^{n_i} \mathrm{E}(Y_{ij}) \;=\; n_i \,\mathrm{P}(Y_{ij}=1) \;=\; n_i \lambda_{i,1}, \\
\mathrm{Var}(Y_{ij}) &= \mathrm{E}(Y_{ij}^2) - \mathrm{E}(Y_{ij})^2 \;=\; \lambda_{i,1}(1-\lambda_{i,1}), \\
\mathrm{Corr}(Y_{ij}, Y_{ik}) &= \frac{\mathrm{E}(Y_{ij}Y_{ik}) - \mathrm{E}(Y_{ij})\,\mathrm{E}(Y_{ik})}{\mathrm{E}({Y_{ij}}^2) - \mathrm{E}(Y_{ij})^2} \\
&= \frac{\mathrm{P}(Y_{ij}=1, Y_{ik}=1) - \mathrm{P}(Y_{ij}=1)\,\mathrm{P}(Y_{ik}=1)}{\mathrm{P}(Y_{ij}=1) - \mathrm{P}(Y_{ij}=1)^2} \\
&= \frac{\lambda_{i,2} - \lambda_{i,1}^2}{\lambda_{i,1}(1-\lambda_{i,1})} = \alpha.
\end{aligned}$$

Diese Aussage ergibt sich ebenso direkt aus dem nächsten Satz, der Aussagen über Momente und Korrelationen von höherer Ordnung macht. Um die Lesbarkeit zu erhöhen, wird vorerst der Gruppenindex i weggelassen.

Satz 19. *Sei $Y_1, Y_2, \ldots, Y_n$ eine Menge von austauschbaren binären Zufallsvariablen und sei ρ_k die Korrelation k-ter Ordnung sowie $Z = \sum_{j=1}^{n} Y_j$.*
Dann ist

$$\rho_k = \frac{\sum_{j=0}^{k} (-1)^{k-j} \binom{k}{j} \lambda_1^{k-j} \lambda_j}{(\lambda_1(1-\lambda_1))^{k/2}}$$

für alle natürlichen Zahlen k und

$$\mathrm{E}(Z^k) = \sum_{j=1}^{min(n,k)} \binom{n}{j} j! S(k,j) \lambda_j$$

wobei $S(k,j)$ die Stirlingnummer der zweiten Art (siehe Def. 12) ist.[9]

Beweis. Da $Y_1, Y_2, \ldots, Y_n$ austauschbar sind, sind alle Randverteilungen und gemeinsamen Verteilungen gleich. Also gilt für jede Teilmenge $Y_{i_1}, Y_{i_2}, \ldots, Y_{i_k}$:

$$[\mathrm{Var}(Y_1)]^{k/2} \rho_k = \mathrm{E}((Y_{i_1} - \mu) \cdots (Y_{i_k} - \mu)) = \mathrm{E}((Y_1 - \mu) \cdots (Y_k - \mu)).$$

[9] Satz 19 und Beweis stammen aus George u. Bowman (1995).

Da $\mathrm{Var}(Y_1) = \mu(1-\mu) = \lambda_1(1-\lambda_1)$ und

$$\mathrm{E}((Y_1-\mu)\cdots(Y_k-\mu)) = \mathrm{E}[\sum_{j=0}^{k}\binom{k}{j}\prod_{i=1}^{j}Y_i\mu^{k-j}(-1)^{k-j}] = \sum_{j=0}^{k}(-1)^{k-j}\binom{k}{j}\lambda_1^{k-j}\lambda_j,$$

folgt der obige Ausdruck für ρ_k.

Für den Beweis des Ausdrucks für $\mathrm{E}(Z^k)$ wird die momenterzeugende Funktion $M(t)$ von Z benutzt, da die k-te Ableitung der momenterzeugenden Funktion an der Stelle 0 gleich dem k-ten Moment $\mathrm{E}(Z^k)$ ist. Es gilt:

$$\begin{aligned}
M(t) &= \mathrm{E}(e^{tZ}) \\
&= \sum_{z=0}^{n} e^{tz}\,\mathrm{P}(Z=z) \\
&= \sum_{z=0}^{n} e^{tz}\binom{n}{z}\sum_{j=0}^{n-z}(-1)^j\binom{n-z}{j}\lambda_{z+j} \\
&= \sum_{z=0}^{n} e^{tz}\binom{n}{z}\sum_{j=z}^{n}(-1)^{j-z}\binom{n-z}{j-z}\lambda_j \\
&= \sum_{j=0}^{n}\sum_{z=0}^{j}(-1)^{j-z}e^{tz}\binom{n}{j}\binom{j}{z}\lambda_j \\
&= \sum_{j=0}^{n}\binom{n}{j}\lambda_j\sum_{z=0}^{j}(-1)^{j-z}e^{tz}\binom{j}{z} \\
&= \sum_{j=0}^{n} j!\binom{n}{j}\lambda_j\frac{(e^t-1)^j}{j!} \\
&= \sum_{j=0}^{n} j!\binom{n}{j}\lambda_j\sum_{l\geq j}S(l,j)\frac{t^l}{l!} \\
&= \sum_{l=0}^{\infty}\frac{t^l}{l!}\sum_{j=0}^{min(n,l)}\binom{n}{j}j!S_{l,j}\lambda_j.
\end{aligned}$$

□

3.1.1.2 Modellwahl für die Parameter $\lambda_{i,k}$ [10]

Insgesamt hat dieses Modell für Subjekt i n_i Parameter. Im Allgemeinen wird angenommen, dass sich diese Parameter $\lambda_{i,k}$ mit Hilfe einer Linkfunktion F_k durch die erklärenden Variablen $x_i \in \mathbb{R}^p$ bei Gruppe i und einen Parametervektor $\beta \in \mathbb{R}^p$ beschreiben lassen, wodurch dann Maximum-Likelihood-Schätzungen erleichtert werden. Somit folgt aus Satz 17:

$$P(Z_i = z_i) \quad = \quad \binom{n_i}{z_i} \sum_{k=z_i}^{n_i} (-1)^{k-z_i} \binom{n_i - z_i}{k - z_i} F_k(x_i^\top \beta). \tag{4}$$

Allerdings sind einige Voraussetzungen an die F_k notwendig. Laut der Gleichung (3) muss

$$\sum_{k=z_i}^{n_i} (-1)^{k-z_i} \binom{n_i - z_i}{k - z_i} \lambda_{i,k} \quad \geq \quad 0 \tag{5}$$

gelten und außerdem

$$1 = \lambda_{i,0} \geq \lambda_{i,1} \geq \lambda_{i,2} \geq \cdots \geq \lambda_{i,n} \geq 0. \tag{6}$$

Das bedeutet, dass die Funktionen F_k für $\lambda_{i,k}$ die Bedingungen in (5) und (6) erfüllen müssen.Ein Funktionentyp, der diese Anforderungen erfüllt, ist die *gefaltete logistische Funktion*

$$\begin{aligned} \lambda_k(\beta^*) \quad &= \quad \frac{2}{1 + (k+1)^{\beta^*}} \\ &= \quad \frac{2}{1 + \exp[\beta^* \ \log(k+1)]} \end{aligned}$$

mit $\beta^* > 0$ und $k = 0, 1, \ldots, n_i$.

Für diesen Funktionentyp lässt sich die Aussage der Gleichung (6) leicht zeigen. Es ist $\lambda_0(\beta^*) = 2/\{1 + \exp(\beta^* \ \log(1))\} = 2/\{1 + 1\} = 1$ und offenbar sind alle $\lambda_k(\beta^*) \geq 0$ für $k \geq 1$. Sei $k \in \{0, 1, \ldots, n_i - 1\}$ und $\beta^* > 0$. Die Monotonie ergibt sich nun aus der wahren Aussage $k + 2 \geq k + 1$ durch folgende äquivalente

[10] Dieses Modell stammt aus George u. Bowman (1995).

Umformungen:

$$\begin{array}{rcccc} & k+2 & \geq & k+1 \\ \Leftrightarrow & \log(k+2) & \geq & \log(k+1) \\ \Leftrightarrow & \beta^* \ \log(k+2) & \geq & \beta^* \ \log(k+1) \\ \Leftrightarrow & 1+\exp(\beta^* \ \log(k+2)) & \geq & 1+\exp(\beta^* \ \log(k+1)) \\ \Leftrightarrow & \dfrac{2}{1+\exp(\beta^* \ \log(k+1))} & \geq & \dfrac{2}{1+\exp(\beta^* \ \log(k+2))} \\ \Leftrightarrow & \lambda_k(\beta^*) & \geq & \lambda_{k+1}(\beta^*). \end{array}$$

Nach George u. Bowman (1995) erfüllt die gefaltete logistische Funktion auch Bedingung (5). Allerdings erfordert der Beweis umfangreiche Schritte aus der Analysis und wird hier nicht erbracht.

Für das vorangegangene Beispiel des toxikologischen Experiments kann solch eine logistische Funktion zur Modellierung der $\lambda_{i,k}$ entsprechend angepasst werden. Sei dazu d_i die Gift-Konzentration, mit der die i-te Gruppe von Tieren behandelt wird. Dann ergibt sich in diesem Fall $x_i = (1, d_i)^\top$ und der Parametervektor $\beta = (\beta_1, \beta_2)$. Die logistische Funktion verändert sich damit wie folgt:

$$\begin{aligned} \lambda_{i,k}(\beta_1, \beta_2) &= 2/\{1+\exp[x_i^\top \beta \ \log(k+1)]\} \\ &= 2/\{1+\exp[(\beta_1+\beta_2 d_i) \ \log(k+1)]\}. \end{aligned}$$

Demnach muss $\beta_1 + \beta_2 d_i \geq 0$ für jedes i sein, damit die obigen Bedingungen an die Modellfunktion für λ_k erfüllt sind. Diese Voraussetzung ist wichtig und nicht immer gegeben und muss daher beachtet werden.

Für die marginale Erwartung $E(Y_{ij}) = \lambda_{i,1}$ ergibt sich dann das Modell

$$\begin{aligned} E(Y_{ij}) &= \lambda_{i,1}(\beta_1, \beta_2) \\ &= 2/\{1+\exp[(\beta_1+\beta_2 d_i) \ \log(2)]\} \end{aligned}$$

bzw. mittels der Linkfunktion h geschrieben

$$h(\lambda_{i,1}) = \log_2\left(\frac{2}{\lambda_{i,1}} - 1\right) = \beta_1 + \beta_2 d_i.$$

3.1.1.3 Maximum-Likelihood-Schätzungen [11]

Seien $(Y_{i1}, Y_{i2}, \ldots, Y_{in_i})$ für $i = 1, \ldots, N$ unabhängige Vektoren. Jeder Vektor $(Y_{i1}, Y_{i2}, \ldots, Y_{in_i})$ habe austauschbare binäre Zufallsvariablen. Zur Bestimmmung der Maximum-Likelihood-Schätzung wird die Log-Likelihood-Funktion $l(\beta)$ benötigt. Diese ergibt sich wie folgt aus Satz 17:

$$\begin{aligned} l(\beta) &= \log \prod_{i=1}^{N} \mathrm{P}(Y_{i1} = y_{i1}, Y_{i2} = y_{i2}, \ldots, Y_{in_i} = y_{in_i}) \\ &= \sum_{i=1}^{N} \log \mathrm{P}(Y_{i1} = y_{i1}, Y_{i2} = y_{i2}, \ldots, Y_{in_i} = y_{in_i}) \\ &= \sum_{i=1}^{N} \log \left(\sum_{k=z_i}^{n_i} (-1)^{k-z_i} \binom{n_i - z_i}{k - z_i} \lambda_{i,k} \right). \end{aligned}$$

Mit einem Modell F_k für $\lambda_{i,k}$, das die obigen Voraussetzungen erfüllt, ergibt sich nun als zu maximierende Funktion

$$l(\beta) = \sum_{i=1}^{N} \log \left(\sum_{k=z_i}^{n_i} (-1)^{k-z_i} \binom{n_i - z_i}{k - z_i} F_k(x_i^\top \beta) \right). \tag{7}$$

Mit Hilfe des Newton-Raphson-Verfahrens kann nun iterativ der Maximum-Likelihood-Schätzer von β bestimmt werden. Die Iterationsvorschrift lautet damit:

$$\beta^{(j+1)} = \beta^{(j)} - H^{-1}(\beta^{(j)}) S(\beta^{(j)}), \quad \text{für } j = 1, 2, \ldots \ .$$

Dabei ist $S(\beta)$ die Transponierte der ersten Ableitung von $l(\beta)$ nach β und wird auch als Score-Funktion bezeichnet. H ist die Hesse-Matrix (es seien die hier benutzten Funktionen F_k mindestens zweimal stetig differenzierbar).
In diesem Fall ergibt sich für die Score-Funktion:

$$S(\beta) = \sum_{i=1}^{N} \sum_{k=z_i}^{n_i} x_i \Delta_k(x_i^\top \beta), \tag{8}$$

wobei

$$\Delta_k(x_i^\top \beta) = \frac{(-1)^{k-z_i} \binom{n_i - z_i}{k - z_i} f_k(x_i^\top \beta)}{\sum_{l=z_i}^{n_i} (-1)^{l-z_i} \binom{n_i - z_i}{l - z_i} F_l(x_i^\top \beta)} \tag{9}$$

[11] vgl. auch George u. Bowman (1995)

und f_k die erste Ableitung von F_k ist.

Zur Bestimmung der Hesse-Matrix muss die Score-Funktion $S(\beta)$ erneut nach β abgeleitet werden:

$$H(\beta) = \frac{\partial S(\beta)}{\partial \beta} = \sum_{i=1}^{N} \sum_{k=z_i}^{n_i} x_i \frac{\partial \Delta_k(x_i^\top \beta)}{\partial \beta}.$$

Mit Hilfe der Quotientenregel ergibt sich nun für $\frac{\partial \Delta_k(x_i^T \beta)}{\partial \beta}$:

$$\begin{aligned}
\frac{\partial \Delta_k(x_i^T \beta)}{\partial \beta} &= \frac{(-1)^{k-z_i}\binom{n_i-z_i}{k-z_i} f_k'(x_i^\top\beta) \sum_{l=z_i}^{n_i}(-1)^{l-z_i}\binom{n_i-z_i}{l-z_i} F_l(x_i^\top\beta)}{(\sum_{l=z_i}^{n_i}(-1)^{l-z_i}\binom{n_i-z_i}{l-z_i} F_l(x_i^\top\beta))^2} x_i^\top \\
&\quad - \frac{(-1)^{k-z_i}\binom{n_i-z_i}{k-z_i} f_k(x_i^\top\beta) \sum_{l=z_i}^{n_i}(-1)^{l-z_i}\binom{n_i-z_i}{l-z_i} f_l(x_i^\top\beta)}{(\sum_{l=z_i}^{n_i}(-1)^{l-z_i}\binom{n_i-z_i}{l-z_i} F_l(x_i^\top\beta))^2} x_i^\top \\
&= \frac{(-1)^{k-z_i}\binom{n_i-z_i}{k-z_i} f_k(x_i^\top\beta)}{\sum_{l=z_i}^{n_i}(-1)^{l-z_i}\binom{n_i-z_i}{l-z_i} F_l(x_i^\top\beta)} \frac{f_k'(x_i^\top\beta)}{f_k(x_i^\top\beta)} x_i^\top \\
&\quad - \frac{(-1)^{k-z_i}\binom{n_i-z_i}{k-z_i} f_k(x_i^\top\beta)}{\sum_{l=z_i}^{n_i}(-1)^{l-z_i}\binom{n_i-z_i}{l-z_i} F_l(x_i^\top\beta)} \frac{\sum_{l=z_i}^{n_i}(-1)^{l-z_i}\binom{n_i-z_i}{l-z_i} f_l(x_i^\top\beta)}{\sum_{l=z_i}^{n_i}(-1)^{l-z_i}\binom{n_i-z_i}{l-z_i} F_l(x_i^\top\beta)} x_i^\top \\
&= \Delta_k(x_i^\top\beta)\frac{f_k'(x_i^\top\beta)}{f_k(x_i^\top\beta)} x_i^\top - \Delta_k(x_i^\top\beta) \sum_{l=z_i}^{n_i} \Delta_l(x_i^\top\beta) x_i^\top \\
&= \Delta_k(x_i^\top\beta)\left[\frac{f_k'(x_i^\top\beta)}{f_k(x_i^\top\beta)} x_i^\top - \sum_{l=z_i}^{n_i} \Delta_l(x_i^\top\beta) x_i^\top\right].
\end{aligned}$$

Damit ergibt sich als Hesse-Matrix:

$$H(\beta) = \sum_{i=1}^{N}\sum_{k=z_i}^{n_i} \Delta_k(x_i^\top\beta)\frac{f_k'(x_i^\top\beta)}{f_k(x_i^\top\beta)} x_i x_i^\top - \sum_{i=1}^{N}\left[\sum_{k=z_i}^{n_i}\Delta_k(x_i^\top\beta)x_i\right]\left[\sum_{l=z_i}^{n_i}\Delta_l(x_i^\top\beta)x_i\right]^\top.$$

Die Newton-Raphson-Iteration muss nicht per Hand berechnet werden, da es Computerprogramme gibt, die dazu benutzt werden können. Z.B kann SAS die Berechnung übernehmen.

Damit soll die Betrachtung des George-Bowman-Modells abgeschlossen sein. Ein weiteres marginales Modell zur Auswertung wiederholter binärer Beobachtungen ist das Bahadur-Modell.

3.1.2 Das Bahadur-Modell

Das Bahadur-Modell ist ein marginales Modell für geclusterte binäre Daten (z.B. wiederholte Beobachtungen bei einem Subjekt) und wurde von Bahadur (1961) eingeführt. Es ist in der Literatur unter anderem in den Artikeln von Declerck u. a. (1998), Lipsitz u. a. (1995) und Pendergast u. a. (1996) zu finden, sowie in den Büchern von Molenberghs u. Verbeke (2005), Aerts u. a. (2002) und Diggle u. a. (2002).

3.1.2.1 Modelldarstellung

Das Bahadur-Modell kann z.B. bei klinischen Test benutzt werden, welche die Wirksamkeit von Medikamenten überprüfen. Dabei wird ein Medikament wiederholt den Probanden verabreicht. Das beobachtete Ereignis ist das Nichtwirken des Medikaments und wird mit 1 kodiert. Die Beobachtungen eines Subjekts sind voneinander abhängig, d.h. sie sind korreliert. Diese Korrelation wird vom Bahadur-Modell berücksichtigt.
Sei Y_{ij} die j-te binäre Beobachtung beim i-ten Subjekt, die anzeigt, ob das beobachtete Ereignis aufgetreten ist oder nicht. Die Verteilung von Y_{ij} ist dann eine Bernoulli-Verteilung mit

$$E(Y_{ij}) \;=\; P(Y_{ij}=1) \;=\; \pi_{ij}.$$

Um die Abhängigkeit der Daten bei einem Subjekt (Cluster) zu beschreiben, gibt es verschiedene Assoziationsparameter: marginal Odds-Ratio, den Korrelations-Koeffizienten und den Kappa-Koeffizienten.[12] Bei dem Bahadur-Modell wird der Korrelations-Koeffizient zwischen der j-ten und k-ten Beobachtung benutzt:

$$\mathrm{Corr}(Y_{ij}, Y_{ik}) = \rho_{ijk} = \frac{\pi_{ijk} - \pi_{ij}\pi_{ik}}{(\pi_{ij}(1-\pi_{ij})\pi_{ik}(1-\pi_{ik}))^{1/2}},$$

wobei $\mathrm{P}(Y_{ij}=1, Y_{ik}=1) = \mathrm{E}(Y_{ij}Y_{ik}) = \pi_{ijk}$ ist.
Damit kann die gemeinsame Wahrscheinlichkeit π_{ijk} geschrieben werden als:

[12] siehe Aerts u. a. (2002)

$$\pi_{ijk} = \pi_{ij}\pi_{ik} + \rho_{ijk}[\pi_{ij}(1-\pi_{ij})\pi_{ik}(1-\pi_{ik})]^{(1/2)}.$$

Für die übrigen gemeinsamen Wahrscheinlichkeiten von zwei Beobachtungen bei Subjekt i ergeben sich dann:

$$\begin{aligned}
\mathrm{P}(Y_{ij}=1, Y_{ik}=0) &= \pi_{ij} - \pi_{ijk} \\
\mathrm{P}(Y_{ij}=0, Y_{ik}=1) &= \pi_{ik} - \pi_{ijk} \\
\mathrm{P}(Y_{ij}=0, Y_{ik}=0) &= 1 - \pi_{ik} - \pi_{ij} + \pi_{ijk}
\end{aligned}$$

und entsprechend für die gemeinsamen Wahrscheinlichkeiten von drei Beobachtungen:

$$\begin{aligned}
\mathrm{P}(Y_{ij}=1, Y_{ik}=1, Y_{il}=1) &= \pi_{ijkl} \\
\mathrm{P}(Y_{ij}=1, Y_{ik}=1, Y_{il}=0) &= \pi_{ijk} - \pi_{ijkl} \\
\mathrm{P}(Y_{ij}=1, Y_{ik}=0, Y_{il}=1) &= \pi_{ijl} - \pi_{ijkl} \\
\mathrm{P}(Y_{ij}=1, Y_{ik}=0, Y_{il}=0) &= \pi_{ij} - \pi_{ijk} - \pi_{ijl} + \pi_{ijkl} \\
\mathrm{P}(Y_{ij}=0, Y_{ik}=1, Y_{il}=1) &= \pi_{ikl} - \pi_{ijkl} \\
\mathrm{P}(Y_{ij}=0, Y_{ik}=1, Y_{il}=0) &= \pi_{ik} - \pi_{ijk} - \pi_{ikl} + \pi_{ijkl} \\
\mathrm{P}(Y_{ij}=0, Y_{ik}=0, Y_{il}=1) &= \pi_{il} - \pi_{ikl} - \pi_{ijl} + \pi_{ijkl} \\
\mathrm{P}(Y_{ij}=0, Y_{ik}=0, Y_{il}=0) &= 1 - \pi_{ik} - \pi_{il} - \pi_{ij} + \pi_{ijk} + \pi_{ijk} + \pi_{ijl} + \pi_{ikl} \\
&\quad - \pi_{ijkl}.
\end{aligned}$$

Für die Korrelation von drei Beobachtungen ergibt sich:

$$\begin{aligned}
\mathrm{Corr}(Y_{ij}, Y_{ik}, Y_{il}) &= \rho_{ijkl} \\
&= \frac{\mathrm{E}((Y_{ij}-\pi_{ij})(Y_{ik}-\pi_{ik})(Y_{il}-\pi_{il}))}{[\pi_{ij}(1-\pi_{ij})\pi_{ik}(1-\pi_{ik})\pi_{il}(1-\pi_{il})]^{1/2}} \\
&= \frac{\mathrm{E}(Y_{ij}Y_{ik}Y_{il} - Y_{ij}\pi_{ik}Y_{il} - \pi_{ij}Y_{ik}Y_{il} + \pi_{ij}\pi_{ik}Y_{il})}{[\pi_{ij}(1-\pi_{ij})\pi_{ik}(1-\pi_{ik})\pi_{il}(1-\pi_{il})]^{1/2}} \\
&\quad + \frac{\mathrm{E}(-Y_{ij}Y_{ik}\pi_{il} + Y_{ij}\pi_{ik}\pi_{il} + \pi_{ij}Y_{ik}\pi_{il} - \pi_{ij}\pi_{ik}\pi_{il})}{[\pi_{ij}(1-\pi_{ij})\pi_{ik}(1-\pi_{ik})\pi_{il}(1-\pi_{il})]^{1/2}} \\
&= \frac{\pi_{ijkl} - \pi_{ijl}\pi_{ik} - \pi_{ikl}\pi_{ij} - \pi_{ijk}\pi_{il} + 2\pi_{ij}\pi_{ik}\pi_{il}}{[\pi_{ij}(1-\pi_{ij})\pi_{ik}(1-\pi_{ik})\pi_{il}(1-\pi_{il})]^{1/2}}
\end{aligned}$$

und damit ist die gemeinsame Wahrscheinlichkeit für drei Beobachtungen:

$$\begin{aligned}
\pi_{ijkl} &= \pi_{ijl}\pi_{ik} + \pi_{ikl}\pi_{ij} \\
&\quad +\pi_{ijk}\pi_{il} - 2\pi_{ij}\pi_{ik}\pi_{il} + \rho_{ijkl}[\pi_{ij}(1-\pi_{ij})\pi_{ik}(1-\pi_{ik})\pi_{il}(1-\pi_{il})]^{1/2} \\
&= \pi_{ij}\pi_{ik}\pi_{il} + \pi_{ij}\rho_{ikl}[\pi_{ik}(1-\pi_{ik})\pi_{il}(1-\pi_{il})]^{1/2} \\
&\quad + \pi_{ik}\rho_{ijl}[\pi_{ij}(1-\pi_{ij})\pi_{il}(1-\pi_{il})]^{1/2} \\
&\quad + \pi_{il}\rho_{ijk}[\pi_{ij}(1-\pi_{ij})\pi_{ik}(1-\pi_{ik})]^{1/2} \\
&\quad + \rho_{ijkl}[\pi_{ij}(1-\pi_{ij})\pi_{ik}(1-\pi_{ik})\pi_{il}(1-\pi_{il})]^{1/2}.
\end{aligned}$$

Für eine likelihood-basierte Annäherung wird die vollständige Angabe der gemeinsamen Wahrscheinlichkeit des binären Antwortvektors von jedem Subjekt benötigt. Sei $f(y)$ die gemeinsame Verteilung von $Y_i = (Y_{i1}, \cdots, Y_{in_i})^T$. Das Bahadur-Modell gibt einen geschlossenen Ausdruck für $f(y)$ an. Die Verbindung der binären Antworten wird dabei in Termen der Randwahrscheinlichkeiten und des Korrelations-Koeffizienten von zweiter, dritter und höherer Ordnung ausgedrückt.[13]

Satz 20. *Sei*

$$\epsilon_{ij} = \frac{Y_{ij} - \pi_{ij}}{\sqrt{\pi_{ij}(1-\pi_{ij})}} \text{ und } e_{ij} = \frac{y_{ij} - \pi_{ij}}{\sqrt{\pi_{ij}(1-\pi_{ij})}},$$

wobei y_{ij} eine Realisierung von Y_{ij} ist. Weiter sei

$$\rho_{ijk} = \mathrm{E}(\epsilon_{ij}\epsilon_{ik}),\ \rho_{ijkl} = \mathrm{E}(\epsilon_{ij}\epsilon_{ik}\epsilon_{il})\ ,\ \ldots\ ,\ \rho_{i12\ldots n_i} = E(\epsilon_{i1}\epsilon_{i2}\cdots\epsilon_{in_i}).$$

Dann kann das allgemeine Bahadur-Modell durch folgenden Ausdruck repräsentiert werden:

$$f(y_i) = f_1(y_i)c(y_i)$$

$$\text{mit } f_1(y_i) = \prod_{j=1}^{n_i} \pi_{ij}^{y_{ij}}(1-\pi_{ij})^{1-y_{ij}}$$

[13] vgl. Molenberghs u. Verbeke (2005)

und $c(y_i) = 1 + \sum_{j<k} \rho_{ijk} e_{ij} e_{ik} + \sum_{j<k<l} \rho_{ijkl} e_{ij} e_{ik} e_{il} + \cdots + \rho_{i12\cdots n_i} e_{i1} e_{i2} \cdots e_{in_i}$.[14]

Beweis. Hier wird der Beweis nur für den Fall $n_i = 2$ geliefert. Ein allgemeiner Beweis ist in Bahadur (1961) zu finden.

Es ist

$$\begin{aligned} f(y_i) &= f_1(y_i)c(y_i) \\ &= \prod_{j=1}^{2} \pi_{ij}^{y_{ij}} (1-\pi_{ij})^{1-y_{ij}} \times (1 + \sum_{j<k} \rho_{ijk} e_{ij} e_{ik}) \\ &= \pi_{i1}^{y_{i1}} (1-\pi_{i1})^{1-y_{i1}} \pi_{i2}^{y_{i2}} (1-\pi_{i2})^{1-y_{i2}} \times (1 + \rho_{i12} e_{i1} e_{i2}). \end{aligned}$$

Nun können alle möglichen Ausprägungen des 2-dimensionalen binären Beobachtungsvektors betrachtet werden:

$$\begin{aligned} f((1,1)) &= \pi_{i1}\pi_{i2} \times \left[1 + \frac{(\pi_{i12} - \pi_{i1}\pi_{i2})(1-\pi_{i1})(1-\pi_{i2})}{\pi_{i1}(1-\pi_{i1})\pi_{i2}(1-\pi_{i2})}\right] \\ &= \pi_{i1}\pi_{i2} \times \left[1 + \frac{\pi_{i12} - \pi_{i1}\pi_{i2}}{\pi_{i1}\pi_{i2}}\right] \\ &= \pi_{i12} \\ &= P(Y_{i1} = 1, Y_{i2} = 1), \\ f((1,0)) &= \pi_{i1}(1-\pi_{i2}) \times \left[1 + \frac{(\pi_{i12} - \pi_{i1}\pi_{i2})(1-\pi_{i1})(-\pi_{i2})}{\pi_{i1}(1-\pi_{i1})\pi_{i2}(1-\pi_{i2})}\right] \\ &= \pi_{i1}(1-\pi_{i2}) \times \left[1 + \frac{-\pi_{i12} + \pi_{i1}\pi_{i2}}{\pi_{i1}(1-\pi_{i2})}\right] \\ &= \pi_{i1} - \pi_{i12} \\ &= P(Y_{i1} = 1, Y_{i2} = 0), \\ f((0,1)) &= (1-\pi_{i1})\pi_{i2} \times \left[1 + \frac{(\pi_{i12} - \pi_{i1}\pi_{i2})(-\pi_{i1})(1-\pi_{i2})}{\pi_{i1}(1-\pi_{i1})\pi_{i2}(1-\pi_{i2})}\right] \\ &= (1-\pi_{i1})\pi_{i2} \times \left[1 + \frac{-\pi_{i12} + \pi_{i1}\pi_{i2}}{(1-\pi_{i1})\pi_{i2}}\right] \\ &= \pi_{i2} - \pi_{i12} \\ &= P(Y_{i1} = 0, Y_{i2} = 1), \\ f((0,0)) &= (1-\pi_{i1})(1-\pi_{i2}) \times \left[1 + \frac{(\pi_{i12} - \pi_{i1}\pi_{i2})(-\pi_{i1})(-\pi_{i2})}{\pi_{i1}(1-\pi_{i1})\pi_{i2}(1-\pi_{i2})}\right] \end{aligned}$$

[14] Satz 20 ist unter anderem in Molenberghs u. Verbeke (2005) zu finden.

$$
\begin{aligned}
&= (1-\pi_{i1})(1-\pi_{i2}) \times \left[1 + \frac{(\pi_{i12} - \pi_{i1}\pi_{i2})}{(1-\pi_{i1})(1-\pi_{i2})}\right] \\
&= 1 - \pi_{i1} - \pi_{i2} + \pi_{i12} \\
&= P(Y_{i1} = 0, Y_{i2} = 0).
\end{aligned}
$$

□

Damit ist die Wahrscheinlichkeitsdichte ein Produkt der Dichte $f_1(y_i)$ des unabhängigen Modells und des Korrekturfaktors $c(y_i)$. Dieser Faktor kann als ein Modell für overdispersion angesehen werden.[15]

Korollar 21. *Das allgemeine Bahadur-Modell hat für jedes Subjekt $2^{n_i} - 1$ Parameter.*

Beweis. Für die Wahrscheinlichkeiten π_{ij} ergeben sich $n_i = \binom{n_i}{1}$ Parameter und für die Korrelationen zweiter Ordnung $\binom{n_i}{2}$ Parameter. Allgemein gibt es für die Korrelationen k-ter Ordnung $\binom{n_i}{k}$ Parameter für $k = 2, \ldots, n_i$. Das ergibt insgesamt $\sum_{k=1}^{n_i} \binom{n_i}{k} = 2^{n_i} - 1$ Parameter. □

Nun wird der spezielle Fall mit austauschbaren Beobachtungen betrachtet. Dabei gilt $\pi_{ij} = \pi_i$ für $j = 1, \ldots, n_i$ und $i = 1, \ldots, N$ und dass die Korrelationen einer Ordnung konstant sind, d.h. $\rho_{ijk} = \rho_i(2)$ für $j < k$, $\rho_{ijkl} = \rho_i(3)$ für $j < k < l$, usw. bis $\rho_{i12\ldots n_i} = \rho_i(n_i)$ für $i = 1, \ldots, N$.[16]

Satz 22. *Seien die Beobachtungen im Bahadur-Modell austauschbar. Dann reduziert sich das Bahadur-Modell auf*

$$f_1(y_i) = \pi_i^{z_i}(1-\pi_i)^{n_i - z_i} \text{ und}$$

$$c(y_i) = 1 + \sum_{r=2}^{n_i} \rho_i(r) \sum_{s=0}^{r} \binom{z_i}{s}\binom{n_i - z_i}{r-s}(-1)^{s+r}\lambda_i^{r-2s}$$

$$\text{mit } \lambda_i = \sqrt{\pi_i/(1-\pi_i)} \text{ und } z_i = \sum_{j=1}^{n_i} y_{ij}.$$[17]

[15] Molenberghs u. Verbeke (2005)

[16] vgl. Molenberghs u. Verbeke (2005)

[17] Satz 22 ist unter anderem in Molenberghs u. Verbeke (2005)zu finden.

Beweis. Die Formel für f_1 ergibt sich nach den Potenzgesetzen. Interessanter ist die Formel für c. $c(y_i)$ vereinfacht sich mit austauschbaren Beobachtungen zu

$$c(y_i) = 1 + \rho_i(2) \sum_{j<k} e_{ij}e_{ik} + \rho_i(3) \sum_{j<k<l} e_{ij}e_{ik}e_{il} + \cdots + \rho_i(n_i)\ e_{i1}e_{i2}\cdots e_{in_i}.$$

Es ist nun noch zu zeigen, dass sich die Summe über die Produkte mit $r \geq 2$ Faktoren e_{ij} durch die Summe $\sum_{s=0}^{r} \binom{z_i}{s}\binom{n_i - z_i}{r-s}(-1)^{r-s}\lambda_i^{r-2s}$ darstellen lässt. Dies wird im folgenden Lemma gezeigt. □

Lemma 23. *Seien e_{ij}, n_i, z_i wie oben definiert. Außerdem sei $r \geq 2$.*
Dann gilt:

$$\sum_{j<k<\cdots<l} \overbrace{e_{ij}e_{ik}\cdot\ldots\cdot e_{il}}^{r\ \textit{Faktoren}} = \sum_{s=0}^{r} \binom{z_i}{s}\binom{n_i - z_i}{r-s}(-1)^{r-s}\lambda_i^{r-2s}.$$

Beweis. Im Beweis wird zuerst e_{ij} durch $\dfrac{y_{ij} - \pi_i}{\sqrt{\pi_i(1-\pi_i)}}$ ersetzt und dann ermittelt, welche möglichen Summanden wie oft vorkommen können. Dies geschieht durch Betrachtung der y_{ij}. Je nachdem wie viele gleich eins sind, ergeben sich unterschiedliche Produkte. Es gilt:

$$\begin{aligned}
\sum_{j<\ldots<l} \overbrace{e_{ij}\cdot\ldots\cdot e_{il}}^{\text{r Faktoren}} &= \frac{1}{(\sqrt{\pi_i(1-\pi_i)})^r} \sum_{j<k<\ldots<l} (y_{ij} - \pi_i)(y_{ik} - \pi_i)\cdot\ldots\cdot(y_{il} - \pi_i) \\
&= \frac{1}{(\sqrt{\pi_i(1-\pi_i)})^r} \left[\binom{z_i}{0}\binom{n_i - z_i}{r-0}(-1)^{r-0}\pi_i^{r-0}(1-\pi_i)^0 \right. \\
&\quad + \binom{z_i}{1}\binom{n_i - z_i}{r-1}(-1)^{r-1}\pi_i^{r-1}(1-\pi_i)^1 \\
&\quad \vdots \\
&\quad \left. + \binom{z_i}{r}\binom{n_i - z_i}{r-r}(-1)^{r-r}\pi_i^{r-r}(1-\pi_i)^r \right] \\
&= \frac{1}{(\sqrt{\pi_i(1-\pi_i)})^r} \sum_{s=0}^{r} \binom{z_i}{s}\binom{n_i - z_i}{r-s}(-1)^{r-s}\pi_i^{r-s}(1-\pi_i)^s \\
&= \sum_{s=0}^{r} \binom{z_i}{s}\binom{n_i - z_i}{r-s}(-1)^{r-s} \frac{\pi_i^{r-s}(1-\pi_i)^s}{(\sqrt{\pi_i(1-\pi_i)})^r}
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{s=0}^{r} \binom{z_i}{s} \binom{n_i - z_i}{r-s} (-1)^{r-s} \left(\frac{\sqrt{\pi_i}}{\sqrt{1-\pi_i}} \right)^r \left(\frac{\sqrt{1-\pi_i}}{\sqrt{\pi_i}} \right)^{2s} \\
&= \sum_{s=0}^{r} \binom{z_i}{s} \binom{n_i - z_i}{r-s} (-1)^{r-s} \lambda_i^{r-2s}.
\end{aligned}
$$

□

Bemerkung 24. Die Wahrscheinlichkeitsdichte von $Z_i = \sum_{j=1}^{n_i} Y_{ij}$ ist dann gegeben durch

$$f(z_i) = \binom{n_i}{z_i} f(y_i),$$

die Anzahl der Parameter pro Subjekt reduziert sich auf $1 + (n_i - 1) = n_i$.

Bemerkung 25. Wenn nun noch alle 3-fach und höheren Korrelationen gleich Null gesetzt werden, ergibt sich:

$$
\begin{aligned}
f(z_i) &= f(z_i; \pi_i, \rho_i(2), n_i) \\
&= \binom{n_i}{z_i} \pi_i^{z_i} (1-\pi_i)^{n_i - z_i} \\
&\quad \times \left[1 + \rho_i(2) \left\{ \binom{n_i - z_i}{2} \frac{\pi_i}{1-\pi_i} - z_i(n_i - z_i) + \binom{z_i}{2} \frac{1-\pi_i}{\pi_i} \right\} \right] \quad (10)
\end{aligned}
$$

Somit bleiben pro Subjekt i nur noch die zwei Parameter $\rho_i(2)$ und π_i übrig.

Diese Darstellung der Bahadur-Wahrscheinlichkeitsdichte ist vorteilhaft gegenüber anderen Darstellungen wie z.B. der Odds-Ratio-Repräsentation, da es für diese keine geschlossene Form der gemeinsamen Verteilung gibt.[18] Allerdings ist sie nur sinnvoll, wenn es sich um eine Wahrscheinlichkeitsdichte handelt. Bahadur (1961) hat gezeigt, dass die Summe über die Wahrscheinlichkeiten aller möglichen Beobachtungen gleich Eins ist. Jedoch ist es möglich, dass in Abhängigkeit von $\rho_i(2)$ und π_i der Ausdruck (10) nicht die zweite Eigenschaft einer Wahrscheinlichkeitsdichte aufweist; (10) ist nicht immer größer oder gleich Null.[19] Zum Beispiel ergibt sich für den Fall $n_i = 4$, $\rho_i(2) = 0.4$ und $\pi_i = 0.05$, dass $f(1) = -0.023465$

[18] Aerts u. a. (2002)

[19] vgl. Molenberghs u. Verbeke (2005)

ist; damit ist f keine Wahrscheinlichkeitsdichte mehr. Diese Tatsache führt zu Einschränkungen des Parameterraums.[20] Für den Fall, dass die höheren Korrelationen entfernt sind, wurden die Einschränkungen von Bahadur (1961) untersucht (siehe auch Kapitel 4). Aber auch wenn sie enthalten sind, ist der Parameterraum von einer eigenartigen Form.[21]

3.1.2.2 Modellwahl für die Parameter $\rho_i(2)$ und π_i

Die marginalen Parameter π_i und ρ_i werden durch eine Linkfunktion mit den erklärenden Variablen verbunden. Da Y_{ij} eine binäre Zufallsvariable ist, wird üblicherweise für die Modellierung des Parameters π_i die logistische Linkfunktion benutzt. Für den Parameter $\rho_i(2)$ ist Fischers Z-Transformation eine gebräuchliche Linkfunktion.[22] Aus diesen zwei Linkfunktionen kann eine zusammengesetzte Linkfunktion gebildet werden. Dies führt zu der folgenden verallgemeinerten linearen Regressionsbeziehung:[23]

$$h((\pi_i, \rho_i(2))^\top) = \begin{pmatrix} \ln\left(\frac{\pi_i}{1-\pi_i}\right) \\ \ln\left(\frac{1+\rho_i(2)}{1-\rho_i(2)}\right) \end{pmatrix} = \eta_i(\beta) = X_i^T\beta \tag{11}$$

mit $X_i \in \mathbb{R}^{2\times p}$ als Planungsmatrix des i-ten Subjekts und $\beta \in \mathbb{R}^p$ als unbekanntem Parametervektor.

Ein Beispiel dazu findet sich in Molenberghs u. Verbeke (2005). Dort ist ein toxikologisches Experiment gegeben, in dem die marginale Erwartung von der Giftkonzentration d_i abhängt. Zur Modellierung lässt sich ein lineares marginales Logitmodell benutzen. Dabei ist die Korrelation konstant, d.h. $\rho_i(2) = \rho(2)$. Und weiter ist

$$X_i = \begin{pmatrix} 1 & 0 \\ d_i & 0 \\ 0 & 1 \end{pmatrix} \quad \text{und} \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_d \\ \beta_2 \end{pmatrix}. \tag{12}$$

[20] vgl. Aerts u. a. (2002)
[21] Aerts u. a. (2002)
[22] siehe Aerts u. a. (2002)
[23] vgl. Aerts u. a. (2002)

Hierbei ist β_0 ein Intercept, β_d beschreibt den Dosiseffekt und β_2 die Korrelation.

3.1.2.3 Maximum-Likelihood-Schätzungen

Für eine Maximum-Likelihood-Schätzung des Parametervektors $\theta_i = (\pi_i, \rho_i(2))^T$ $= (\pi_i(\beta), \rho_i(2)(\beta))^T$ benötigt man die Log-Likelihood-Funktion vom i-ten Subjekt:

$$l_i(\beta) = \ln(f(z_i; \pi_i(\beta), \rho_i(2)(\beta), n_i)). \tag{13}$$

Da die Beobachtungen der einzelnen N Subjekte als unabhängig vorausgesetzt werden, ergibt sich für die Log-Likelihood-Funktion:

$$l(\beta) = \sum_{i=1}^{N} l_i(\beta).$$

Die Maximum-Likelihood-Schätzung $\hat{\beta}$ für β ist definiert als die Nullstelle der Ableitung von $l(\beta)$ nach β bzw. der Score-Funktion. Mit dem Newton-Raphson-Verfahren lässt sich die Maximum-Likelihood-Schätzung bestimmen.
Es soll hier dennoch die Score-Funktion aufgestellt werden. Mit Hilfe der Kettenregel lässt sich die Ableitung von $l_i(\beta)$ nach β bestimmen. Zunächst ist nach (11)

$$h(\theta_i) = \eta_i(\beta) = X_i^\top \beta$$

und damit

$$\theta_i = h^{-1}(\eta_i(\beta)),$$

wobei $h^{-1}(\mu_i(\beta))$ bedeutet, dass in jedem Eintrag des Vektors die Umkehrfunktion gebildet wird. Mit der Dichte als Funktion von θ_i und dem Vorangegangenen ist nun

$$\begin{aligned}
\frac{\partial l_i(\beta)}{\partial \beta} &= \frac{\partial \ln(f(\theta_i))}{\partial \beta} \\
&= \frac{\partial \ln(f(h^{-1}(\eta_i(\beta))))}{\partial \beta} \\
&= \left.\frac{\partial \ln(f(x))}{\partial x}\right|_{x=h^{-1}(\eta_i(\beta))=\theta_i} \left.\frac{\partial h^{-1}(y)}{\partial y}\right|_{y=\eta_i(\beta)} \frac{\partial \eta_i(\beta)}{\partial \beta}.
\end{aligned}$$

Dabei ist $\frac{\partial \eta_i(\beta)}{\partial \beta} = X_i^\top \in R^{2\times 3}$ und $\left.\frac{\partial \ln(f(x))}{\partial x}\right|_{x=h^{-1}(\eta_i(\beta))=\theta_i} \in R^{1\times 2}$.
Weiterhin ist h als Funktion von θ_i stetig differenzierbar und die Jacobimatrix ($\in R^{2\times 2}$) ist für alle Werte im Definitionsbereich $M = (0,1)\times(-1,1)$ invertierbar. Also gilt mit dem Satz über Umkehrfunktionen, dass die Jacobimatrizen von h und h^{-1} invers zueinander sind. Somit gilt

$$\begin{aligned}
\left.\frac{\partial h^{-1}(y)}{\partial y}\right|_{y=\eta_i(\beta)} &= \left(\left.\frac{\partial h(z)}{\partial z}\right|_{z=h^{-1}(\eta_i(\beta))=\theta_i}\right)^{-1} \\
&= \begin{pmatrix} \frac{1}{\pi_i(1-\pi)} & 0 \\ 0 & \frac{2}{(1-\rho_i(2))(1+\rho_i(2))} \end{pmatrix}^{-1} \\
&= \begin{pmatrix} \pi_i(1-\pi) & 0 \\ 0 & \frac{(1-\rho_i(2))(1+\rho_i(2))}{2} \end{pmatrix}.
\end{aligned}$$

Weiter ist es egal, ob ein 1×3-Vektor oder ein 3×1-Vektor zu Bestimmung der Nullstellen betrachtet wird. In der Literatur wird meist die Form $p\times 1$ der Score-Funktion bevorzugt. Durch das Transponieren und mit $T_i = \left.\frac{\partial h(z)}{\partial z}\right|_{z=\theta_i}$ und $L_i = \left.\frac{\partial \ln(f(x))}{\partial x}\right|_{x=\theta_i}$ lässt sich die Score-Funktion jetzt folgendermaßen schreiben:[24]

$$S(\beta) = \sum_{i=1}^{N} X_i (T_i^\top)^{-1} L_i^\top. \tag{14}$$

Wenn höhere Korrelationsordnungen enthalten sind, wird die Implementierung der Score-Funktion unhandlich. Fischers Z-Transformation kann für alle Korrelationsparameter $\rho_i(r)$ benutzt werden. Die Planungsmatrix X_i wird in diesem Fall entsprechend verändert. Allerdings ist das Schätzen eines Bahadur-Modells mit höheren Korrelationsordnungen nicht ohne weiteres möglich, da ansteigend komplexe Einschränkungen des Parameterraums auftreten.[25]

Deshalb werden alternativ in vielen Studien nur die marginale Erwartung und die

[24] vgl. Molenberghs u. Verbeke (2005)

[25] vgl. Aerts u. a. (2002)

paarweisen Korrelationen betrachtet. Dabei wird die gesamte Likelihood-Annäherung durch Schätzgleichungen (estimating-equations) ersetzt, wobei nur die ersten zwei Momente modelliert werden.[26] Diese Methode wird *verallgemeinerte lineare Schätzgleichungen (generalized estimating equations (GEEs))* genannt und wird in Abschnitt 3.1.3.1 näher betrachtet.

Für das Logit-Modell $h(\pi_{ij}) = \text{logit}(\pi_{ij}) = x_{ij}^{\top}\beta$ im Bahadur-Modell mit unabhängigen Beobachtungen ergibt sich als Score-Funktion $\sum_{i=1}^{N} X_i(y_i - \pi_i)$ mit $X_i = (x_{i1}, \ldots, x_{in_i})$. Diese Score-Funktion ergibt sich analog zu Obigem durch Ableitung der Log-Likelihood-Funktion

$$\ln\left(\prod_{i=1}^{N}\prod_{j=1}^{n_i}\pi_{ij}^{y_{ij}}(1-\pi_{ij})^{1-y_{ij}}\right).$$

Diese Score-Funktion ist gerade die Schätzgleichung der GEEs für unabhängige Beobachtungen mit Bernoulli-verteilten Zufallsvariablen (vgl. 3.1.3.1.1). Mehr zu estimating equations und generalized estimating equations (GEE) ist in Abschnitt 3.1.3.1 zu finden.

3.1.3 Schätzmethoden

Zum Schätzen des unbekannten Parametervektors β gibt es verschiedene Möglichkeiten. Die Bekannteste ist die Maximum-Likelihood-Schätzung. Dazu muss, wie oben schon angedeutet, die komplette gemeinsame Wahrscheinlichkeitsfunktion angegeben werden. Oftmals ist dies nicht möglich oder die Bestimmung des Maximum-Likelihood-Schätzers numerisch zu aufwändig. Dann können alternativ andere Methoden zur Parameterbestimmung verwendet werden, z.B. Pseudo-Likelihood-Methoden, Alternierende Logistische Regression (ALR) oder verallgemeinerte Schätzgleichungen (generalized estimating equations, GEE).[27] Auf letztere wird im Folgenden näher eingegangen.

[26] vgl.Molenberghs u. Verbeke (2005)

[27] vgl. Molenberghs u. Verbeke (2005)

3.1.3.1 Generalized Estimating Equations (GEEs)

Wenn das Interesse hauptsächlich auf Parametern der marginalen Erwartung und paarweiser Korrelation liegt, kann eine Log-Likelihoodgleichung für korrelierte bzw. geclusterte Daten durch *Verallgemeinerte Schätzgleichungen (generalized estimating equations, GEE)* ersetzt werden.[28] Diese führten Liang u. Zeger (1986) ein. Sie fordern die korrekte Spezifizierung der univariaten marginalen Verteilung und stellen nur eine Annahme über die Korrelationsstruktur auf. Sie schätzen die Parameter, die mit dem Erwartungswert eines individuellen binären Antwortvektors verbunden sind und drücken die Voraussetzungen über die Assoziation der Beobachtungen in Termen der marginalen Korrelation aus.[29]

Diese Methode von Liang u. Zeger (1986) wird heute auch GEE1 genannt und ist in vielen Software-Paketen, unter anderem in SAS in der Prozedur GENMOD enthalten. Nach Liang u. Zeger (1986) liefert diese Methode auch dann konsistente Schätzungen für die Parameter der Haupteffekte, wenn die Korrelationsstruktur falsch angenommen wurde, solange der Erwartungswert korrekt modelliert ist.[30] Allerdings ist diese Methode weniger geeignet, wenn das Interesse auf den Korrelationssparametern selbst liegt. Dafür sollte dann eher eine andere Methode, z.B. GEE2 benutzt werden. In dieser Arbeit wird hier nur die Theorie der GEE1 betrachtet, die im Folgenden vorgestellt wird.

3.1.3.1.1 Allgemeine Theorie der GEEs

Für das Schätzen des Parametervektors β, der mit der marginalen Erwartung in Modellen mit diskreten wiederholten Beobachtungen zusammenhängt, können GEE's benutzt werden. Um die Score-Gleichung $S(\beta)$ der GEE-Theorie für diskrete korrelierte Beobachtungen zu entwickeln, wird zunächst eine Score-Gleichung für unabhängige Beobachtungen hergeleitet. Dabei sei ein marginales verallge-

[28] vgl.Molenberghs u. Verbeke (2005)

[29] Molenberghs u. Verbeke (2005)

[30] Diese Aussage aus Liang u. Zeger (1986) wird im Rahmen dieser Arbeit ohne Beweis benutzt.

meinertes lineares Modell gegeben. Seien Y_{ij} die Zufallsvariable, die die j-te Beobachtung beim i-ten Subjekt beschreibt und besitze Y_{ij} eine Verteilung einer Exponential-Familie, bei der der Parameter ϕ unabhängig von β sei. Dann können mit der Theorie der *verallgemeinerten linearen Modelle* die Maximum-Likelihood-Schätzer bestimmt werden. Zunächst gilt für diese Modelle der Zusammenhang (siehe Grundlagen)

$$h(\mathrm{E}(Y_{ij})) = h(\mu_{ij}) = h(b'(\theta_{ij})) = x_{ij}^\top \beta.$$

Daraus folgt

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

Da es sich hier um unabhängige Beobachtungen handelt, ist die gemeinsame Dichte gerade das Produkt der Dichten der Y_{ij}. Die Maximum-Likelihood-Schätzungen sind damit Lösungen der Gleichung

$$\frac{\partial \sum_{i=1}^{N} \sum_{j=1}^{n_i} \ln f_{y_{ij}}(\theta_{ij}, \phi)}{\partial \beta} = \sum_{i=1}^{N} \sum_{j=1}^{n_i} \frac{\partial \ln f_{y_{ij}}((b')^{-1}(h^{-1}(x_{ij}^\top \beta)), \phi)}{\partial \beta} = 0. \tag{15}$$

Mit Hilfe der Kettenregel lässt sich $\sum_{i=1}^{N} \sum_{j=1}^{n_i} \frac{\partial \ln f_{y_{ij}}((b')^{-1}(h^{-1}(x_{ij}^\top \beta)), \phi)}{\partial \beta}$ nun weiter umformen. Es ergibt sich

$$\begin{aligned}
& \sum_{i=1}^{N} \sum_{j=1}^{n_i} \left. \frac{\partial \ln f_{y_{ij}}(x, y)}{\partial (x, y)} \right|_{(x,y)=((b')^{-1}(h^{-1}(x_{ij}^\top \beta)), \phi)} \left(\frac{\partial (b')^{-1}(h^{-1}(x_{ij}^\top \beta))}{\partial \beta}, \frac{\partial \phi}{\partial \beta} \right)^\top \\
= & \sum_{i=1}^{N} \sum_{j=1}^{n_i} \left. \frac{\partial \ln f_{y_{ij}}(x, y)}{\partial x} \right|_{(x,y)=(\theta_{ij}, \phi)} \frac{\partial (b')^{-1}(h^{-1}(x_{ij}^\top \beta))}{\partial \beta} \\
= & \sum_{i=1}^{N} \sum_{j=1}^{n_i} \left. \frac{\partial \ln f_{y_{ij}}(x, y)}{\partial x} \right|_{(x,y)=(\theta_{ij}, \phi)} \left. \frac{\partial (b')^{-1}(z)}{\partial z} \right|_{z=h^{-1}(x_{ij}^\top \beta)} \frac{\partial h^{-1}(x_{ij}^\top \beta)}{\partial \beta}.
\end{aligned}$$

Wird nun die Dichte abgeleitet (Dichte einer Verteilung aus der Exponentialfamilie, siehe Grundlagen), ergibt sich

$$\begin{aligned}
&\sum_{i=1}^{N}\sum_{j=1}^{n_i}(y_{ij}-b'(\theta_{ij}))\frac{w}{\phi}\left(\frac{\partial(b')(v)}{\partial v}\Big|_{v=(b')^{-1}(h^{-1}(x_{ij}^\top\beta))}\right)^{-1}\frac{\partial\mu_{ij}}{\partial\beta}\\
=&\sum_{i=1}^{N}\sum_{j=1}^{n_i}(y_{ij}-b'(\theta_{ij}))\frac{w}{\phi}(b''(\theta_{ij}))^{-1}\frac{\partial\mu_{ij}}{\partial\beta}\\
=&\sum_{i=1}^{N}(y_i-b'(\theta_i))^\top diag(b''(\theta_{ij})\frac{\phi}{w})^{-1}\frac{\partial\mu_i}{\partial\beta}\\
=&\sum_{i=1}^{N}(y_i-b'(\theta_i))^\top A_i^{-1}\frac{\partial\mu_i}{\partial\beta}.
\end{aligned}$$

Dabei ist A_i gerade die Kovarianzmatrix, eine Diagonalmatrix mit den marginalen Varianzen (siehe Grundlagen: $b''(\theta_{ij})\frac{\phi}{w}=\mathrm{Var}(Y_{ij}),\ b'(\theta_{ij})=\mu_{ij}$).
In der Literatur wird meist die Transponierte dieser Ableitung gleich Null gesetzt, so dass sich als Score-Gleichung ergibt:

$$\sum_{i=1}^{N}\left(\frac{\partial\mu_i}{\partial\beta}\right)^\top A_i^{-1}(y_i-\mu_i)=0 \tag{16}$$

bzw.

$$\sum_{i=1}^{N}\left(\frac{\partial\mu_i}{\partial\beta}\right)^\top (A_i^{1/2}I_{n_i}A_i^{1/2})^{-1}(y_i-\mu_i)=0. \tag{17}$$

Hier ist I_{n_i} die $n_i\times n_i$ Einheitsmatrix, die als Korrelationsmatrix betrachtet werden kann, die die Unabhängigkeit der Beobachtungen widerspiegelt.
Für binäre Beobachtungen ($\mu_{ij}=p_{ij}$) und das Modell

$$\mathrm{logit}(p_{ij})=\log\left(\frac{p_{ij}}{1-p_{ij}}\right)=x_{ij}^\top\beta$$

ergibt sich für unabhängige Beobachtungen als Score-Funktion

$$\begin{aligned}
&\sum_{i=1}^{N}\left(\frac{\partial p_i}{\partial\beta}\right)^\top A_i^{-1}(y_i-p_i)\\
=&\sum_{i=1}^{N}\sum_{j=1}^{n_i}\frac{\partial p_{ij}}{\partial\beta}(p_{ij}(1-p_{ij}))^{-1}(y_{ij}-p_{ij})
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=1}^{N}\sum_{j=1}^{n_i} \frac{\partial \frac{\exp(x_{ij}^\top\beta)}{1+\exp(x_{ij}^\top\beta)}}{\partial\beta}(p_{ij}(1-p_{ij}))^{-1}(y_{ij}-p_{ij}) \\
&= \sum_{i=1}^{N}\sum_{j=1}^{n_i} x_{ij}^\top p_{ij}(1-p_{ij})(p_{ij}(1-p_{ij}))^{-1}(y_{ij}-p_{ij}) \\
&= \sum_{i=1}^{N} X_i(y_i-p_i).
\end{aligned}
$$

Mit korrelierten Beobachtungen wird nun die Score-Gleichung der GEE-Theorie definiert durch

$$S(\beta) = \sum_{i=1}^{N}\left(\frac{\partial\mu_i}{\partial\beta}\right)^\top (A_i^{1/2}R_iA_i^{1/2})^{-1}(y_i-\mu_i) = 0, \tag{18}$$

wobei durch Ersetzen der Einheitsmatrix I_{n_i} durch R_i die Korrelation berücksichtigt wird. $A_i^{1/2}R_iA_i^{1/2}$ spiegelt auch hier die Kovarianzmatrix wider. A_i wird erneut durch β bestimmt, aber dieser Parameter enthält keine Information über R_i. Deshalb muss $R_i = R_i(\alpha)$ durch einen zusätzlichen Parameter bestimmt werden. Liang u. Zeger (1986) vermieden das Hinzufügen von zusätzlichen Modellkomponenten für diesen Parameter, indem sie dem Modellierer erlaubten, eine *inkorrekte Struktur*, eine sogenannte Arbeitskorrelationsmatrix, anzugeben. Ohne Beweis werden hier ihre Ergebnisse benutzt: Solange der Erwartungswert μ_i korrekt durch $h(\mu_i) = X_i\beta$ modelliert wurde und unter Hinzunahme von schwachen Regularitätsbedingungen ist der Schätzer $\hat{\beta}$, der durch Lösen der Gleichung (18) erhalten wird, konsistent und asymptotisch normalverteilt mit Erwartungswert β und asymptotischer Kovarianzmatrix

$$\mathrm{Var}(\hat{\beta}) = I_0^{-1}I_1I_0^{-1}.$$

Dabei ist

$$I_0 = \sum_{i=0}^{N}\left(\frac{\partial\mu_i}{\partial\beta}\right)^\top V_i^{-1}\frac{\partial\mu_i}{\partial\beta} \tag{19}$$

und

$$I_1 = \sum_{i=0}^{N}\left(\frac{\partial\mu_i}{\partial\beta}\right)^\top V_i^{-1}\,\mathrm{Var}(Y_i)V_i^{-1}\frac{\partial\mu_i}{\partial\beta}\ . \tag{20}$$

Ein zusätzlicher Parameter ϕ für die *overdispersion* kann folgendermaßen hinzugefügt werden:

$$V_i = V(\beta, \alpha, \phi) = \phi A_i(\beta)^{1/2} R_i(\alpha) A_i(\beta)^{1/2}. \tag{21}$$

3.1.3.1.2 Schätzen des Parameters β

Die Schätzung des Parameters β erfolgt in einem iterativen Prozess und wird aus Molenberghs u. Verbeke (2005) ohne Beweis übernommen. Dazu werden zunächst Fehlerterme

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

definiert und V_i wie oben benutzt. Für binäre Beobachtungen und mit dem Logit-Modell ist

$$e_{ij} = \frac{y_{ij} - p_{ij}}{\sqrt{p_{ij}(1 - p_{ij})}}.$$

und $A_i(\beta)^{1/2}$ aus V_i ist eine $n_i \times n_i$ Diagonalmatrix mit den Einträgen $\sqrt{p_{ij}(1 - p_{ij})}$ für $j = 1, \ldots, n_i$.

Außerdem muss die Korrelationsstruktur spezifiziert und der Parameter α geschätzt werden. Dazu gibt es verschiedene Möglichkeiten, z.B.:

- Unabhängig:

  $$\text{Corr}(Y_{ij}, Y_{ik}) = 0, \quad (j \neq k).$$

  Bei dieser Wahl müssen keine Korrelationsparameter geschätzt werden.

- Austauschbar:

  $$\begin{aligned} \text{Corr}(Y_{ij}, Y_{ik}) &= \alpha, \quad (j \neq k), \\ \hat{\alpha} &= \frac{1}{N} \sum_{i=1}^{N} \frac{1}{n_i(n_i - 1)} \sum_{j \neq k} e_{ij} e_{ik}. \end{aligned}$$

- AR(1):

  $$\begin{aligned} \text{Corr}(Y_{ij}, Y_{i,j+t}) &= \alpha^t, \quad (t = 0, 1, \ldots, n_i - j), \\ \hat{\alpha} &= \frac{1}{N} \sum_{i=1}^{N} \frac{1}{n_i - 1} \sum_{j \leq n_i - 1} e_{ij} e_{i,j+1}. \end{aligned}$$

- Unstrukturiert:

$$\begin{aligned} \text{Corr}(Y_{ij}, Y_{ik}) &= \alpha_{jk}, \quad (j \neq k), \\ \hat{\alpha_{jk}} &= \frac{1}{N} \sum_{i=1}^{N} e_{ij} e_{ik}. \end{aligned}$$

Der Dispersionsparameter ϕ kann dann durch

$$\hat{\phi} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{n_i} \sum_{j=1}^{n_i} e_{ij}^2$$

geschätzt werden.

Die Iteration zum Schätzen von β nach Liang u. Zeger (1986) läuft in folgenden Schritten ab:

1. Erzeuge eine initialen Schätzer β^0 für β, indem alle Beobachtungen als unabhängig vorausgesetzt werden.
2. Erzeuge die Größen $e_{ij}, \hat{\alpha}$ und $\hat{\phi}$.
3. Erzeuge mit diesen Werten $R_i(\hat{\alpha})$ und V_i.
4. Erzeuge mit dem letzten β-Wert den folgenden:

$$\begin{aligned} \beta^{k+1} = \beta^k - \left[\sum_{i=1}^{N} \left(\frac{\partial \mu_i}{\partial \beta} \right)^\top V_i^{-1} \sum_{i=1}^{N} \left(\frac{\partial \mu_i}{\partial \beta} \right) \right]^{-1} \\ \times \left[\sum_{i=1}^{N} \left(\frac{\partial \mu_i}{\partial \beta} \right)^\top V_i^{-1} (y_i - \mu_i) \right]. \end{aligned} \tag{22}$$

Nach diesem Verfahren berechnet auch das Statistik-Programm SAS mit der Prozedur GENMOD die Parameterschätzer in einem marginalen Modell mit wiederholten Beobachtungen.

Im weiteren Verlauf wird nun die zweite Klasse von Modellen für wiederholte Beobachtungen betrachtet: die subjektspezifischen Modelle.

3.2 Subjektspezifische Modelle

Die subjektspezifischen Modelle lassen sich auch als clusterspezifische Modelle bezeichnen, je nachdem, ob die wiederholten Beobachtungen bei Personen oder Gruppen bzw. Einheiten gemacht werden. Ohne Einschränkung wird im Folgenden von subjektspezifischen Modellen gesprochen.

Bei subjektspezifischen Modellen sind im Gegensatz zu den marginalen Modellen zusätzlich zu den erklärenden Variablen subjektspezifische Parameter vorhanden. Diese subjektspezifischen Parameter variieren, wie der Name schon sagt, von einer Person zur Nächsten. Sie spiegeln die natürliche Heterogenität innerhalb einer Population wider, die auf nicht-messbare Faktoren zurückzuführen ist. Die Antwortwahrscheinlichkeit wird als Funktion von den erklärenden Variablen und den subjektspezifischen Parametern modelliert. Die subjektspezifischen Modelle können benutzt werden, wenn auch subjektspezifische Entwicklungen betrachtet werden sollen. Die Interpretation der Parameter für die festen Effekte erfolgt bedingt auf einem konstanten Level der subjektspezifischen Parameter.[31]

Nach Molenberghs u. Verbeke (2005) gibt es grundlegend drei verschiedene Arten die subjektspezifischen Parameter zu behandeln. Die Erste ist, die subjektspezifischen Parameter wie feste Effekte zu behandeln, also für jede Person einen separaten Parameter zu benutzen und Rückschlüsse durch Maximum-Likelihood-Methoden zu erhalten. Allerdings ist diese Behandlung problematisch, da die Anzahl der Parameter proportional zu der Stichprobengröße ansteigt. Eine weitere Möglichkeit ist, Rückschlüsse durch Maximierung der Likelihoodfunktion, bedingt auf einer suffizienten Statistik für den subjektspezifischen Parameter, zu erhalten. Diese Statistik muss allerdings erst gefunden werden. Die dritte Möglichkeit, bei der die subjektspezifischen Parameter als zufällige Effekte angesehen werden, wird häufiger benutzt. Dies bedeutet, dass der subjektspezifische Parameter als Zufallsvariable angesetzt wird. Diese Betrachtungsweise wird auch im

[31] vgl. Aerts u. a. (2002) und Diggle u. a. (2002)

Rahmen dieser Arbeit weiter verfolgt, so fällt beispielsweise das unten betrachtete Beta-Binomial-Modell in diese Klasse. Doch zunächst wird ein Beispiel für ein subjektspezifisches Modell mit der Indonesian Children's Health Study gegeben (vgl. dieses Beispiel mit Beispiel 15).[32]

Beispiel 26. Es wird ein logistisches Modell mit subjektspezifischen Parametern für die Wahrscheinlichkeit einer Atemwegsinfektion in der Indonesian Children's Health Study (Bemerkung 6) betrachtet. Im Gegensatz zu Beispiel 15 wird diesmal die unterschiedliche Neigung der Kinder zur Erkrankung an einer Atemwegsinfektion berücksichtigt (z.B. hervorgerufen durch unterschiedliche Lebensumstände und genetische Veranlagung).
Das einfachste Modell dieser Art ist ein Modell, das für jedes Kind die Neigung zur Infektion berücksichtigt, aber den Effekt des Vitamin-A-Mangels auf die Wahrscheinlichkeit einer Infektion für jedes Kind gleich ansetzt. Sei Y_{ij} die binäre Beobachtung mit $Y_{ij} = 1$ für beobachtete eine Atemwegsinfektion und $Y_{ij} = 0$ sonst. Dann kann folgendes Modell für die bedingte Erwartung aufgestellt werden:

$$\operatorname{logit} \mathrm{P}(Y_{ij} = 1|U_i) = \beta_0 + U_i + \beta_1 x_{ij},$$

wobei x_{ij} die binäre erklärende Variable ist, die angibt, ob das Kind i bei der j-ten Beobachtung unter Vitamin-A-Mangel litt ($x_{ij} = 1$) oder nicht ($x_{ij} = 0$). U_i ist nun eine Zufallsvariable, die die Neigung des i-ten Kindes zur Atemwegsinfektion beschreibt. Die Verteilung der Zufallsvariable ist mit der Modellaufstellung festzulegen – in diesem Fall soll U_i eine normalverteilte Zufallsvariable mit Erwartungswert 0 und unbekannter Varianz ν^2 sein. Die Varianz ν^2 spiegelt den Grad der Heterogenität in der Neigung der Kinder zur Erkrankung wider, die nicht durch die erklärenden Variablen beeinflusst wird. Bedingt auf U_i sind im subjektspezifischen Modell die einzelnen Beobachtungen eines Kindes unabhängig.
Der Parameter β_0 gibt den Logarithmus des Odds einer Atemwegsinfektion eines

[32] vgl. Diggle u. a. (2002)

typischen Kindes mit $U_i = 0$ und ohne Vitamin-A-Mangel an. Der Parameter β_1 ist der Logarithmus des Bruchs des Odds einer Atemwegsinfektion für ein Kind mit Mangel und des Odds für dasselbe(!) Kind ohne Mangel.[33] Zu weiteren Interpretationen der Parameter und einem Vergleich von marginalen und subjektspezifischen Parametern siehe Abschnitt 4.

Im Folgenden wird eine allgemeine Formulierung der subjektspezifischen Modelle gegeben, in denen die subjektspezifischen Parameter wie zufällige Effekte behandelt werden.
Sei Y_i der n_i-dimensionale Vektor mit den wiederholten Beobachtungen bei Subjekt (Cluster) i für $i = 1, \ldots, N$. Hier wird die Modellformulierung wie in Molenberghs u. Verbeke (2005) benutzt, d.h. dass Y_i, bedingt auf U_i, einer vorher festgelegten Verteilung F_i folgt. F_i hängt eventuell von erklärenden Variablen ab und wird durch einen Vektor θ mit unbekannten Parametern beschrieben. Weiter hängt F_i von einem zufälligen Vektor U_i ab, der subjektspezifisch ist:

$$Y_i|U_i \ \sim \ F^{\theta}_{Y_i|U_i}(y_i).$$

Die Verteilung F_i kann dabei jede bekannte n_i-dimensionale Verteilung sein. Wenn die Komponenten Y_{ij} von Y_i bedingt auf U_i unabhängig sind, genügt es, die univariate Verteilung von allen Y_{ij} zu beschreiben. Somit wird die Verteilungsfunktion F_i zu einem Produkt über die n_i univariaten Verteilungsfunktionen von Y_{ij}.[34] Die zugehörige Dichte zu F_i sei

$$f^{\theta}_{Y_i|U_i}(y_i).$$

Dieser allgemeine Ansatz wird nun für die Behandlung der subjektspezifischen Parameter als zufällige Effekte angepasst. Dabei wird angenommen, dass mit dem

[33] Diese Interpretation des Parameters scheint widersinnig, da ein Kind nicht einen Mangel aufweisen und gleichzeitig keinen haben kann. Besser ist die Vorstellung von zwei verschiedenen Kindern mit den gleichen subjektspezifischen Veranlagungen.

[34] vgl. Molenberghs u. Verbeke (2005)

Ziehen der Subjekte aus einer Population auch die Parameter U_i aus einer Population von subjektspezifischen Parametern gezogen werden. D.h. die U_i's können als Zufallsvektoren angesehen werden, die unabhängig von einer Verteilungsfunktion $Q(U_i)$ gezogen werden. Die Verteilung $Q(U_i)$ wird auch *mixing distribution* genannt.[35]

Im Allgemeinen werden zwei Möglichkeiten für die *mixing distribution* benutzt: Stiratelli u. a. (1984) setzen den Parametervektor U_i als normalverteilt voraus; ebenso in der Arbeit mit verallgemeinerten linearen gemischten Modellen. Alternativ kann der Parametervektor auch von einer Beta-Verteilung stammen.[36] Beide Ansätze werden im Folgenden genauer betrachtet.

Schätzungen für den Parameter θ können durch Maximum-Likelihood-Schätzungen mit Hilfe der marginalen Dichte von Y_i erhalten werden. Es wird dazu über den Parameter U_i intergriert:

$$f_i^{\theta,Q}(y_i) \;=\; \int f^{\theta}_{Y_i|U_i}(y_i) dQ_i(U_i) \tag{23}$$

Auf diese Art und Weise können multivariate marginale Likelihoods erhalten werden; deshalb steht diese Annäherung zu anderen marginalen Modellen, wie z.B. dem Bahadur-Modell, in Konkurrenz.

Im Anschluss wird nun das Beta-Binomial-Modell betrachtet.

3.2.1 Das Beta-Binomial-Modell

Das Beta-Binomial-Modell ist ein subjektspezifisches Modell, in dem die subjektspezifischen Parameter wie zufällige Effekte behandelt werden. Dieses Modell kann für binäre Beobachtungen genutzt werden.

3.2.1.1 Modelldarstellung

Das Beta-Binomial-Modell kann auch wieder in toxikologischen Experimenten genutzt werden. Jedes Subjekt besitzt dabei einen Zufallsparameter(vektor) Π_i, der

[35] vgl. Molenberghs u. Verbeke (2005)

[36] vgl. Aerts u. a. (2002)

in diesem Fall die subjektspezifische Erfolgswahrscheinlichkeit für das Auftreten des unerwünschten Ereignisses bei den n_i Bernoulli-verteilten Beobachtungen Y_{ij} bei Subjekt i angibt. Sei Y_i der n_i-dimensionale Vektor der Beobachtungen und seien die Elemente Y_{ij} bedingt auf Π_i unabhängig. Dann ist die bedingte Dichte von Y_i gegeben Π_i proportional zu der Dichte $f_{Z_i|\Pi_i}(z_i)$ von $Z_i = \sum_j Y_{ij}$. Diese ist, bedingt auf Π_i, binomialverteilt mit Parametern n_i und Π_i. Die Erfolgswahrscheinlichkeit Π_i wird nun als Beta-verteilt mit den Parametern a_i und b_i mit Erwartungswert $\pi_i = \frac{a_i}{a_i+b_i}$ vorausgesetzt.[37] Die Dichte der Beta-Verteilung mit Parameter (a_i, b_i), $a_i, b_i > 0$ ist gegeben durch:

$$g_{\Pi_i}(x) = \frac{x^{a_i-1}(1-x)^{b_i-1}}{B(a_i, b_i)},\ 0 \leq x \leq 1.$$

Dabei ist $B(\cdot,\cdot)$ die Beta-Funktion: $B(a_i, b_i) = \int_0^1 u^{a_i-1}(1-u)^{b_i-1}\ du$.
Die Varianz der Beta-Verteilung ist $\frac{a_i b_i}{(a_i+b_i+1)(a_i+b_i)^2}$.[38]

Satz und Definition 27. *Sei Z_i wie oben definiert. Dann ist die marginale Dichte von Z_i gegeben durch*

$$f_{Z_i}(z_i) = \binom{n_i}{z_i} \frac{B(z_i + a_i, n_i - z_i + b_i)}{B(a_i, b_i)}.^{39} \tag{24}$$

Diese Dichte nennt man auch Beta-binomiale Dichte.

Beweis. Es ist

$$\begin{aligned}
f_{Z_i}(z_i) &= \int f_{Z_i|\Pi_i=x}(z_i)\ dQ(\Pi_i) \\
&= \int f_{Z_i|\Pi_i=x}(z_i) g_{\Pi_i}(x)\ dx \\
&= \int \binom{n_i}{z_i} x^{z_i}(1-x)^{n_i-z_i} \frac{x^{a_i-1}(1-x)^{b_i-1}}{B(a_i, b_i)}\ dx \\
&= \binom{n_i}{z_i} \int \frac{x^{z_i+a_i-1}(1-x)^{n_i-z_i+b_i-1}}{B(a_i, b_i)}\ dx
\end{aligned}$$

[37] vgl. Molenberghs u. Verbeke (2005)

[38] siehe DeGroot (1975) Seite 244

[39] Die Aussage dieses Satzes stammt aus Molenberghs u. Verbeke (2005).

$$= \binom{n_i}{z_i} \frac{B(z_i + a_i, n_i - z_i + b_i)}{B(a_i, b_i)}.$$

□

Korollar 28. *Für die Dichte von Y_{ij} ergibt sich entsprechend:*

$$f_{Y_{ij}}(y_{ij}) = \frac{B(y_{ij} + a_i, 1 - y_{ij} + b_i)}{B(a_i, b_i)}.$$

Beweis. Der Beweis folgt als Spezialfall aus Satz 27 mit $n_i = 1$. □

Mit Hilfe der Dichtefunktionen lassen sich nun leicht die Erwartungswerte von Y_{ij} und Z_i bestimmen.

Satz 29. *Es ist $E(Y_{ij}) = \frac{a_i}{a_i+b_i}$ und $E(Z_i) = n_i \frac{a_i}{a_i+b_i}$.*

Beweis. Es ist
$E(Y_{ij}) = P(Y_{ij} = 1) = f_{Y_{ij}}(1) = \int x g_{\Pi_i}(x)\, dx = E(\Pi_i) = \pi_i = \dfrac{a_i}{a_i + b_i}$ und damit
$\mu_i := E(Z_i) = E(\sum Y_{ij}) = \sum E(Y_{ij}) = n_i \pi_i = n_i \dfrac{a_i}{a_i + b_i}$. □

Auch die Varianzen lassen sich nun herleiten.

Satz 30. *Für die Varianzen von Zi und Y_{ij} ergibt sich:*

$$\begin{aligned}
\operatorname{Var}(Z_i) &= n_i \left[\frac{a_i}{a_i + b_i}\left(1 - \frac{a_i}{a_i + b_i}\right)\right]\left[1 + (n_i - 1)\frac{1}{a_i + b_i + 1}\right] \\
\operatorname{Var}(Y_{ij}) &= \frac{a_i}{a_i + b_i}\left(1 - \frac{a_i}{a_i + b_i}\right)
\end{aligned}$$

Beweis. Für die Varianz von Z_i gilt:

$$\begin{aligned}
\operatorname{Var}(Z_i) &= E(\operatorname{Var}(Z_i|\Pi_i)) + \operatorname{Var}(E(Z_i|\Pi_i)) \\
&= E(n_i\Pi_i(1 - \Pi_i)) + \operatorname{Var}(n_i\Pi_i) \\
&= n_i(E(\Pi_i) - E(\Pi_i^2)) + n_i^2 \operatorname{Var}(\Pi_i) \\
&= n_i(E(\Pi_i) - E(\Pi_i)^2) - n_i \operatorname{Var}(\Pi_i) + n_i^2 \operatorname{Var}(\Pi_i) \\
&= n_i[E(\Pi_i)(1 - E(\Pi_i)) + (n_i - 1)\operatorname{Var}(\Pi_i)] \\
&= n_i\left[\frac{a_i}{a_i + b_i}\left(1 - \frac{a_i}{a_i + b_i}\right) + (n_i - 1)\frac{a_i b_i}{(a_i + b_i + 1)(a_i + b_i)^2}\right] \\
&= n_i\left[\frac{a_i}{a_i + b_i}\left(1 - \frac{a_i}{a_i + b_i}\right)\right]\left[1 + (n_i - 1)\frac{1}{a_i + b_i + 1}\right]
\end{aligned}$$

und die Behauptung für $\mathrm{Var}(Y_{ij})$ folgt mit $n_i = 1$. □

Um in diesem Modell die Assoziation der Beobachtungen zu beschreiben, wird die Korrelation zwischen Y_{ij} und Y_{ik} verwendet. Dafür wird zuerst die Kovarianz berechnet.

Satz 31. *Für die Kovarianz zwischen zwei Beobachtungen Y_{ij} und Y_{ik} gilt:*

$$Cov(Y_{ij}, Y_{ik}) = \frac{a_i b_i}{(a_i + b_i + 1)(a_i + b_i)^2}.$$

Beweis. Es gilt mit Satz 27 (angewendet mit $n_i = 2$ und $z_i = 2$):

$$\begin{aligned}
\mathrm{Cov}(Y_{ij}, Y_{ik}) &= \mathrm{E}(Y_{ij}Y_{ik}) - \mathrm{E}(Y_{ij})\,\mathrm{E}(Y_{ik}) \\
&= \mathrm{P}(Y_{ij} = 1, Y_{ik} = 1) - \mathrm{E}(\Pi_i)^2 \\
&= \binom{2}{2} \frac{B(2 + a_i, b_i)}{B(a_i, b_i)} - \mathrm{E}(\Pi_i)^2 \\
&= \int \binom{2}{2} x^2 (1 - x)^0 g_{\Pi_i}(x)\, dx - \mathrm{E}(\Pi_i)^2 \\
&= \int x^2 g_{\Pi_i}(x)\, dx - \mathrm{E}(\Pi_i)^2 \\
&= \mathrm{E}(\Pi_i^2) - \mathrm{E}(\Pi_i)^2 \\
&= \mathrm{Var}(\Pi_i) \\
&= \frac{a_i b_i}{(a_i + b_i + 1)(a_i + b_i)^2}.
\end{aligned}$$

□

Korollar 32. *Für die Korrelation ergibt sich:*

$$\rho_i := \mathrm{Corr}(Y_{ij}, Y_{ik}) = \frac{1}{a_i + b_i + 1}.$$

Beweis. Mit Hilfe der Sätze 30 und 31 gilt:

$$\begin{aligned}
\mathrm{Corr}(Y_{ij}, Y_{ik}) &= \frac{\mathrm{Cov}(Y_{ij}, Y_{ik})}{\mathrm{Var}(Y_{ij})} \\
&= \frac{a_i b_i}{(a_i + b_i + 1)(a_i + b_i)^2} \frac{1}{\frac{a_i}{a_i+b_i}\left(1 - \frac{a_i}{a_i+b_i}\right)} \\
&= \frac{a_i b_i}{(a_i + b_i + 1)(a_i + b_i)^2} \frac{(a_i + b_i)^2}{a_i b_i} \\
&= \frac{1}{a_i + b_i + 1}.
\end{aligned}$$

□

Bemerkung 33. Um die Wahrscheinlichkeitsdichte $f_{Z_i}(z_i)$ in Termen des durchschnittlichen Erwartungswerts π_i und der Korrelation ρ_i auszudrücken, kann eine Umparametrisierung durchgeführt werden. Durch elementare Umformungen ergibt sich $a_i = \pi_i(\rho_i^{-1} - 1)$ und $b_i = (1 - \pi_i)(\rho_i^{-1} - 1)$. Damit kann die Dichte nun geschrieben werden als

$$f_{Z_i}^{(\pi_i, \rho_i)}(z_i) = \binom{n_i}{z_i} \frac{B[\pi_i(\rho_i^{-1} - 1) + z_i, (1 - \pi_i)(\rho_i^{-1} - 1) + n_i - z_i]}{B[\pi_i(\rho_i^{-1} - 1), (1 - \pi_i)(\rho_i^{-1} - 1)]}.^{40} \qquad (25)$$

Bemerkung 34. Für die Zufallsvariable Π_i des subjektspezifischen Effekts gilt mit dieser Parametrisierung

$$\mathrm{E}(\Pi_i) = \frac{\pi_i(\rho_i^{-1} - 1)}{\rho_i^{-1} - 1} = \pi_i \quad \text{und} \quad \mathrm{Var}(\Pi_i) = \frac{\pi_i(1 - \pi_i)(\rho_i^{-1} - 1)}{\rho_i^{-1}(\rho_i^{-1} - 1)^2} = \pi_i(1 - \pi_i)\rho_i.$$

Dieses Modell hat demnach pro Subjekt wieder 2 Parameter (π_i und ρ_i), die geschätzt werden müssen.

3.2.1.2 Modellwahl für die Parameter π_i und ρ_i

Die Parameter π_i und ρ_i können – wie beim Bahadur-Modell – durch die Logistische Linkfunktion und Fischers Z-Transformation modelliert werden. Deshalb wird hier nicht weiter darauf eingegangen.[41]

[40] vgl.Molenberghs u. Verbeke (2005)

[41] vgl. Molenberghs u. Verbeke (2005)

3.2.1.3 Maximum-Likelihood-Schätzungen

Für die Maximum-Likelihood-Schätzung wird erneut die Log-Likelihood-Funktion benötigt. Diese wird im folgenden Satz betrachtet.

Satz 35. *Gegeben sei das oben beschriebene Beta-Binomial-Modell. Dann gilt für die Log-Likelihood-Funktion des i-ten Subjekts:*

$$l_i = \sum_{r=0}^{z_i-1} \ln\left(\pi_i + \frac{r\rho_i}{1-\rho_i}\right) + \sum_{r=0}^{n_i-z_i-1} \ln\left(1-\pi+\frac{r\rho_i}{1-\rho_i}\right) - \sum_{r=0}^{n_i-1} \ln\left(1+\frac{r\rho_i}{1-\rho_i}\right).^{42}$$

Beweis. Der Beweis bedient sich der Gleichung (25). Der positive Faktor $\binom{n_i}{z_i}$ kann bei der Maximierung ignoriert werden und wird somit weggelassen. Für die Beta-Funktion gilt $B(a,b) = \dfrac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$, wobei Γ die Gamma-Funktion ist. Damit wird Gleichung (25) (ohne den Term $\binom{n_i}{z_i}$) zu

$$\frac{\Gamma[\pi_i(\rho_i^{-1}-1)+z_i]\cdot\Gamma[(1-\pi_i)(\rho_i^{-1}-1)+n_i-z_i]}{\Gamma[\pi_i(\rho_i^{-1}-1)+z_i+(1-\pi_i)(\rho_i^{-1}-1)+n_i-z_i]} \cdot \frac{\Gamma[\pi_i(\rho_i^{-1}-1)+(1-\pi_i)(\rho_i^{-1}-1)]}{\Gamma[\pi_i(\rho_i^{-1}-1)]\cdot\Gamma[(1-\pi_i)(\rho_i^{-1}-1)]}. \quad (26)$$

Vereinfacht ergibt sich daraus:

$$\frac{\Gamma[\pi_i(\rho_i^{-1}-1)+z_i]}{\Gamma[\pi_i(\rho_i^{-1}-1)]} \cdot \frac{\Gamma[(1-\pi_i)(\rho_i^{-1}-1)+n_i-z_i]}{\Gamma[(1-\pi_i)(\rho_i^{-1}-1)]} \cdot \frac{\Gamma[(\rho_i^{-1}-1)]}{\Gamma[(\rho_i^{-1}-1)+n_i]}. \quad (27)$$

Für die Gamma-Funktion Γ gilt: $\Gamma(x+1) = x\Gamma(x)$ für $x > 0$. Durch mehrmalige Anwendung dieser Funktionalgleichung wird (27) zu

$$\frac{\prod_{r=0}^{z_i-1}(\pi_i(\rho_i^{-1}-1)+r)\cdot\Gamma[\pi_i(\rho_i^{-1}-1)]}{\Gamma[\pi_i(\rho_i^{-1}-1)]} \cdot \frac{\prod_{r=0}^{n_i-z_i-1}((1-\pi_i)(\rho_i^{-1}-1)+r)\Gamma[(1-\pi_i)(\rho_i^{-1}-1)]}{\Gamma[(1-\pi_i)(\rho_i^{-1}-1)]} \cdot \frac{\Gamma[(\rho_i^{-1}-1)]}{\prod_{r=0}^{n_i-1}((\rho_i^{-1}-1)+r)\Gamma[(\rho_i^{-1}-1)]} \quad (28)$$

[42] Die Aussage dieses Satzes stammt aus Aerts u. a. (2002).

bzw. zu

$$\frac{\prod_{r=0}^{z_i-1}(\pi_i(\rho_i^{-1}-1)+r)\prod_{r=0}^{n_i-z_i-1}((1-\pi_i)(\rho_i^{-1}-1)+r)}{\prod_{r=0}^{n_i-1}((\rho_i^{-1}-1)+r)}. \tag{29}$$

Dies entspricht

$$\frac{\prod_{r=0}^{z_i-1}\left(\pi_i+\frac{r}{(\rho_i^{-1}-1)}\right)\prod_{r=0}^{n_i-z_i-1}\left((1-\pi_i)+\frac{r}{(\rho_i^{-1}-1)}\right)}{\prod_{r=0}^{n_i-1}\left(1+\frac{r}{(\rho_i^{-1}-1)}\right)}$$

bzw.

$$\frac{\prod_{r=0}^{z_i-1}\left(\pi_i+\frac{r\rho_i}{1-\rho_i}\right)\prod_{r=0}^{n_i-z_i-1}\left((1-\pi_i)+\frac{r\rho_i}{1-\rho_i}\right)}{\prod_{r=0}^{n_i-1}\left(1+\frac{r\rho_i}{1-\rho_i}\right)}. \tag{30}$$

Logarithmieren des Ausdrucks liefert dann die Behauptung. □

Wenn die gleichen verallgemeinerten linearen Regressionsbeziehungen wie in Gleichung (11) und (12) für π_i und ρ_i vorausgesetzt werden, ergibt sich der Maximum-Likelihod-Schätzer $\hat{\beta}$ als Lösung von $S(\beta) = 0$ mit der Score-Funktion für β wie in Gleichung (14) definiert.[43]

3.2.2 Verallgemeinerte lineare gemischte Modelle

Nach Molenberghs u. Verbeke (2005) sind verallgemeinerte lineare gemischte Modelle die am häufigsten benutzten Modelle mit zufälligen Effekten in der Arbeit mit diskreten wiederholten Beobachtungen. Bevor eine allgemeine Modelldarstellung der verallgemeinerten linearen gemischten Modelle vorgestellt wird, wird erst das meist benutzte verallgemeinerte lineare gemischte Modell in Verbindung mit binären Daten vorgestellt: die logistische Regression mit gemischten Effekten.

[43] vgl. Molenberghs u. Verbeke (2005)

3.2.2.1 Die logistische Regression mit gemischten Effekten

Das einfachste logistische Modell mit festen und zufälligen Effekten ist ein Modell mit einem zufälligen skalaren Intercept und einer skalaren erklärenden Variablen (vgl. Beispiel 26). Dazu seien Y_{ij} binäre Beobachtungen des Subjekts i mit erklärender Variable $x_{ij} \in \mathbb{R}$. Dann lässt sich das Modell folgendermaßen schreiben:

$$\text{logit}[\text{E}(Y_{ij}|V_{i,1})] = V_{i,1} + \beta_2 x_{ij}.^{44} \tag{31}$$

Dabei ist β_2 ein fester Effekt und $V_{i,1}$ ein zufälliger Effekt mit $V_{i,1} \sim N(\beta_1, \nu^2)$. Durch Setzen von

$$U_i = V_{i,1} - \beta_1$$

wird (31) zu

$$\text{logit}[\text{E}(Y_{ij}|U_i)] = (\beta_1 + U_i) + \beta_2 x_{ij} \tag{32}$$

mit $U_i \sim N(0, \nu^2)$. Nun sind β_1 und β_2 feste Effekte und der zufällige Effekt U_i gibt die Abweichung von dem Parameter β_1 an.

Dieses Modell lässt sich nach Hedeker (2005) zu einem Modell mit multiplen festen und zufälligen Effekten erweitern. Sei dazu $x_{ij} \in \mathbb{R}^p$ ein Vektor mit den Variablen mit festen Effekten und $z_{ij} \in \mathbb{R}^q$ ein Vektor der erklärenden Variablen, die zufällige Effekte besitzen. Dann wird das logistische Modell mit gemischten Effekten zu

$$\text{logit}[\text{E}(Y_{ij}|U_i)] = x_{ij}^\top \beta + z_{ij}^\top U_i, \tag{33}$$

wobei U_i nun multivariat normalverteilt mit Erwartungswertvektor 0 und Kovarianzmatrix D ist.

3.2.2.2 Allgemeine Darstellung

Eine allgemeine Darstellung der verallgemeinerten linearen gemischten Modelle ist in Molenberghs u. Verbeke (2005) und in Diggle u. a. (2002) gegeben und wird hier erläutert.

[44] siehe Pendergast u. a. (1996)

Sei Y_{ij} die j-te Beobachtung beim i-ten Subjekt für $i = 1, \ldots, N$, $j = 1, \ldots, n_i$ und Y_i der n_i-dimensionale Beobachtungsvektor. Das verallgemeinerte lineare gemischte Modell lässt sich mit einer Linkfunktion h darstellen durch

$$h(\mu_{ij}) = h[E(Y_{ij}|U_i)] = x_{ij}^\top \beta + z_{ij}^\top U_i. \tag{34}$$

Der bedingte Erwartungswert wird somit durch die Linkfunktion h mit den erklärenden Variablen $x_{ij} \in \mathbb{R}^p$ und $z_{ij} \in \mathbb{R}^q$ verbunden. Weiter ist β der unbekannte Parametervektor der festen Effekte und U_i ein Zufallsvektor, der den zufälligen Effekt beschreibt. Die U_i sind unabhängig identisch verteilt. Sei hier U_i normalverteilt mit Erwartungswertvektor 0 und Kovarianzmatrix D und sei $f_{U_i}(u_i)$ die zugehörige Dichte.
Bedingt auf den zufälligen Effektvektor U_i seien die Beobachtungen Y_{ij} unabhängig mit einer Dichte der Exponential-Familie, d.h.

$$f^{\beta,\phi}_{Y_{ij}|U_i}(y_{ij}) = \exp\{[y_{ij}\theta_{ij} - b(\theta_{ij})]\frac{w}{\phi} + c(y_{ij}, \phi)\}. \tag{35}$$

3.2.2.3 Maximum-Likelihood-Schätzungen

Nach Molenberghs u. Verbeke (2005) lassen sich die Parameter eines subjektspezifischen Modells, hier des verallgemeinerten linearen gemischten Modells, durch Maximierung der marginalen Likelihood-Funktion erhalten. Dies geschieht durch Integration über die zufälligen Effekte. Die marginale Dichte eines Subjekts i aus Gleichung (23) wird somit zu

$$f^{\beta,\phi,D}_{Y_i}(y_i) \;= \int \prod_{j=1}^{n_i} f^{\beta,\phi}_{Y_{ij}|U_i=u_i}(y_{ij}) f_{U_i}(u_i) du_i, \tag{36}$$

und so ergibt sich die Likelihood-Funktion

$$\begin{aligned} L^{\beta,\phi,D}(y) &= \prod_{i=1}^{N} f^{\beta,\phi,D}_{Y_i}(y_i) \\ &= \prod_{i=1}^{N} \int \prod_{j=1}^{n_i} f^{\beta,\phi}_{Y_{ij}|U_i=u_i}(y_{ij}) f_{U_i}(u_i) du_i. \end{aligned} \tag{37}$$

Entspricht $f^{\beta,\phi}_{Y_{ij}|U_i}(y_{ij})$ einer Bernoulli-Dichte und h dem Logit-Link

$$\text{logit}(\text{E}(Y_{ij}|\beta, U_i)) = \text{logit}(p_{ij}(\beta, U_i)) = x_{ij}^\top \beta + z_{ij}^\top U_i$$

wird die Likelihood-Funktion zu

$$\begin{aligned}
L^{\beta,D}(y) &= \prod_{i=1}^{N} \int \prod_{j=1}^{n_i} (p_{ij}(\beta, u_i))^{y_{ij}} (1 - p_{ij}(\beta, u_i))^{1-y_{ij}} f_{U_i}(u_i) du_i \\
&= \prod_{i=1}^{N} \int \prod_{j=1}^{n_i} \left(\frac{\exp(x_{ij}^\top \beta + z_{ij}^\top u_i)}{1 + \exp(x_{ij}^\top \beta + z_{ij}^\top u_i)} \right)^{y_{ij}} \left(\frac{1}{1 + \exp(x_{ij}^\top \beta + z_{ij}^\top u_i)} \right)^{1-y_{ij}} \\
&\quad \times f_{U_i}(u_i) du_i \\
&= \prod_{i=1}^{N} \int \prod_{j=1}^{n_i} \exp((x_{ij}^\top \beta + z_{ij}^\top u_i) y_{ij}) (1 + \exp(x_{ij}^\top \beta + z_{ij}^\top u_i))^{-1} f_{U_i}(u_i) du_i \\
&= \prod_{i=1}^{N} \int \exp(\beta^\top \sum_j x_{ij} y_{ij} + u_i^\top \sum_j z_{ij} y_{ij} - \sum_j \log(1 + x_{ij}^\top \beta + z_{ij}^\top u_i) \\
&\quad \times (2\pi)^{-q/2} |D|^{-1/2} \exp(-u_i D^{-1} u_i / 2) du_i.
\end{aligned}$$

4 Vergleich der Modellklassen

Bei Modellen mit nichtlinearen Linkfunktionen unterscheiden sich die marginalen und die subjektspezifischen Parameter. In marginalen Modellen beschreiben die Parameter einen durchschnittlich zu erwartenden Unterschied z.B. in den Behandlungen. In subjektspezifischen Modellen beschreiben die Parameter diesen Unterschied auf einem bestimmten Level der subjektspezifischen Parameter, d.h. die Variation in den Subjekten wird berücksichtigt. Im Folgenden werden diese Unterschiede der Parameterinterpretation an einem Beispiel mit einem Logit-Link betrachtet und anschließend die Parameter der Indonesian Children's Health Study (ICHS) (siehe Bemerkung 6) untersucht.

4.1 Vergleich der Modellklassen für binäre Daten mit dem Logit-Link

Um die Parameter eines marginalen und eines subjektspezifischen Modells für binäre Daten zu vergleichen, wird ein Modell mit einem Logit-Link und einer stetigen erklärenden Variablen t, z.B. Zeit, betrachtet (vgl. Molenberghs u. Verbeke (2005)). Im marginalen Modell sind also zwei Parameter zu schätzen: ein Intercept und ein Parameter für den Zeiteinfluss. Das subjektspezifische Modell enthält einen zufälligen Intercept, also $\text{logit}[\text{P}(Y_{ij}|U_i)] = \beta_0 + U_i + \beta_1 t$ mit $U_i \sim N(0, \sigma^2)$. Somit ist der bedingte Erwartungswert

$$\text{E}(Y_{ij}|U_i) = \frac{\exp(\beta_0 + U_i + \beta_1 t)}{1 + \exp(\beta_0 + U_i + \beta_1 t)}. \tag{38}$$

Um die durchschnittliche Erwartung zu erhalten, wird über die bedingte Erwartung integriert:

$$\begin{aligned} \text{E}(Y_{ij}) &= \text{E}[\text{E}(Y_{ij}|U_i)] \\ &= \text{E}\left[\frac{\exp(\beta_0 + U_i + \beta_1 t)}{1 + \exp(\beta_0 + U_i + \beta_1 t)}\right] \qquad (39) \\ &\neq \frac{\exp(\beta_0 + \beta_1 t)}{1 + \exp(\beta_0 + \beta_1 t)}. \qquad (40) \end{aligned}$$

Damit ist die marginale Erwartung, die sich aus dem subjektspezifischen Ansatz ergibt, nicht gleich der der marginialen Erwartung aus einem marginalen Modell. Trotzdem können mit dieser marginalisierten Erwartung aus dem subjektspezifischen Modell einige Rückschlüsse gezogen werden. In Abbildung 1 wurden

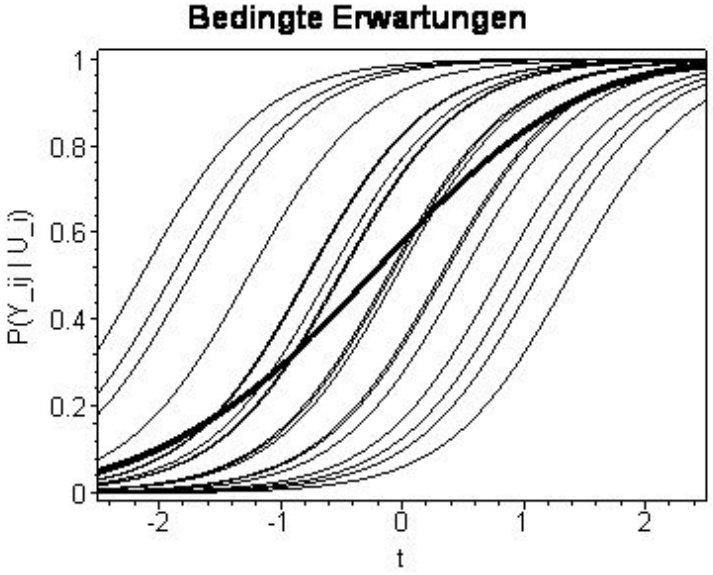

Abbildung 1: Graphische Darstellung von bedingten Erwartungen eines Logit-Modells mit verschiedenen Realisierungen des Random Effects zusammen mit der marginalisierten Erwartung.

einige bedingte Erwartungen und die dazugehörige marginalisierte Erwartung in Abhängigkeit von t gezeichnet.[45] Es ist deutlich zu erkennen, dass der marginale Zeittrend weniger steil ist als jeder subjektbezogene Zeittrend. Molenberghs u. Verbeke (2005), S. 299, führt das zu folgendem Schluss:

"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."

[45]Eine ähnliche Abbildung ist in Molenberghs u. Verbeke (2005) zu finden. Die Erstellung dieser Abbildung ist im Anhang A erläutert.

Somit sollten also Parameter aus dem marginalen und dem subjektspezifischen Modell auch in der Notation unterschieden werden. Seien β^{RE} der Parametervektor eines subjektspezifischen Modells und β^{M} der eines marginalen Modells. In diesem Beispiel beschreibt β^{RE} die Erfolgswahrscheinlichkeit eines jeden Subjekts, wobei die Variation in den Subjekten berücksichtigt wird und β^{M} gibt an, wie sich die durchschnittliche Erfolgswahrscheinlichkeit der Population entwickelt.

In Neuhaus u. a. (1991) wird das obige Beispiel zur Motivation verwendet, um die Parameter β^{RE} und β^{M} zu vergleichen. Deren Ergebnisse sind aber auch allgemein gültig. Sie erhielten folgenden Zusammenhang zwischen β^{RE} und β^{M}: Die marginalen Parameter sind betragsmäßig kleiner als die subjektspezifischen Parameter. Insbesondere bedeutet das, dass wenn $\beta^{\mathrm{RE}} = 0$ ist, auch $\beta^{\mathrm{M}} = 0$ ist. Außerdem stellten sie fest, dass, je größer die Varianz $\mathrm{Var}(U_i)$ ist, sich die Parameter um so mehr unterscheiden. Wenn die Varianz $\mathrm{Var}(U_i) = 0$ ist, es also keinen zufälligen Effekt gibt, stimmen marginale und subjektspezifische Parameter überein.

Zeger u. a. (1988) betrachteten den speziellen Fall, dass U_i normalverteilt ist mit Erwartungswertvektor 0 und Kovarianzmatrix D. Das obige Beispiel mit Erwartungswert 0 und Varianz σ^2 fällt in diese Klasse und Zeger u. a. (1988) zeigen, dass dann

$$\text{logit } \mathrm{E}(Y_{ij}) \approx (c^2\sigma^2 + 1)^{-1/2} x_{ij}\beta^{\mathrm{RE}} \tag{41}$$

ist, wobei $c^2 = 16\sqrt{3}/15\pi$.[46] Somit gilt

$$\beta^{\mathrm{M}} \approx (c^2\sigma^2 + 1)^{-1/2}\beta^{\mathrm{RE}} \tag{42}$$

bzw.

$$\left|\frac{\beta^{\mathrm{RE}}}{\beta^{\mathrm{M}}}\right| \approx \sqrt{c^2\sigma^2 + 1} > 1. \tag{43}$$

[46] Der Beweis wird hier aufgrund seines Umfangs nicht erbracht, ist aber in Zeger u. a. (1988) nachzulesen.

Auch hier ergeben sich gleiche Parameter, wenn die Varianz σ^2 des subjektspezifischen Intercepts gleich Null ist.

Da die Parameter also eine unterschiedliche Interpretation besitzen, sollte bei der Wahl des Modells darauf geachtet werden, welche Fragestellung untersucht werden soll. In einem marginalen Modell ergeben sich sofort marginale Parameter, entweder direkt durch Maximum-Likelihood-Schätzungen oder durch die GEE. Bei den subjektspezifischen Modellen, von denen hier die random-effects-Modelle untersucht werden, lassen sich Rückschlüsse entweder durch marginale oder hierarchische Methoden erhalten. Diese Parameter sind, wie oben beschrieben, nicht mit den marginalen Parametern vergleichbar. Aber es können mit ihnen Parameter erhalten werden, die auch einen marginalen Trend im subjektspezifischen Modell beschreiben. Dieser marginale Trend kann durch numerisches Mitteln graphisch dargestellt werden. Dazu wird eine große Anzahl von Realisierungen von U_i bestimmt und die Kurve folgendermaßen berechnet:

$$\hat{\mathrm{E}}(Y_{ij}) = \frac{1}{M}\sum_{i=1}^{M}\frac{\exp(\hat{\beta}_0^{\mathrm{RE}} + u_i + \hat{\beta}_1^{\mathrm{RE}}t)}{1+\exp(\hat{\beta}_0^{\mathrm{RE}} + u_i + \hat{\beta}_1^{\mathrm{RE}}t)}.\text{[47]} \tag{44}$$

4.1.1 Indonesian Children's Health Study

Im Folgenden wird der oben beschriebene Unterschied der Parameter in einem marginalen und einem subjektspezifischen Modell mit Hilfe der Indonesian Children's Health Study dargestellt.[48]

Das subjektspezifische Modell war in diesem Fall (siehe Beispiel 26)

$$\operatorname{logit}\mathrm{P}(Y_{ij}=1|U_i) = \beta_0^{\mathrm{RE}} + U_i + \beta_1^{\mathrm{RE}}x_{ij}, \tag{45}$$

wobei die U_i normalverteilt sind und $x_{ij} = 1$ für Kinder mit Vitamin-A-Mangel und sonst $x_{ij} = 0$ gesetzt wird. Ein Kind ohne Vitamin-A-Mangel hat also die Wahrscheinlichkeit einer Infektion von

$$\mathrm{P}(Y_{ij}=1|U_i) = \exp(\beta_0^{\mathrm{RE}} + U_i)/(1+\exp(\beta_0^{\mathrm{RE}} + U_i))$$

[47] vgl. Molenberghs u. Verbeke (2005)

[48] siehe Diggle u. a. (2002)

und ein Kind mit Vitamin-A-Mangel von

$$\exp(\beta_0^{\mathrm{RE}} + U_i + \beta_1^{\mathrm{RE}})/(1 + \exp(\beta_0^{\mathrm{RE}} + U_i + \beta_1^{\mathrm{RE}})).$$

Um nun den Unterschied in den Parametern deutlich zu machen, werden hypothetische Werte für die subjektspezifischen Parameter benutzt und zwar für $\beta_0^{\mathrm{RE}} = -2.0$, $\beta_1^{\mathrm{RE}} = 0.4$ und $\mathrm{Var}(U_i) = 2$ (s.a. Diggle u. a. (2002)). Für ein Kind mit $U_i = 0$ und ohne Mangel ergibt sich dann eine Infektionswahrscheinlichkeit von 0.12 und mit Mangel eine von 0.17.
Um den Vergleich mit einem marginalen Modell zu ermöglichen, wird in diesem subjektspezifischen Modell die durchschnittliche Erwartung berechnet, die annähernd gleich der marginalen Erwartung in einem marginalen Modell ist. Dies erfolgt wieder durch Integration:

$$\begin{aligned} \mathrm{P}(Y_{ij} = 1) &= \int \mathrm{P}(Y_{ij} = 1|U_i)\ \mathrm{d}Q(U_i) && (46) \\ &= \int \frac{\beta_0^{\mathrm{RE}} + u_i + \beta_1^{\mathrm{RE}} x_{ij}}{1 + \beta_0^{\mathrm{RE}} + u_i + \beta_1^{\mathrm{RE}} x_{ij}} f_{U_i}(u_i)\ \mathrm{d}u_i, && (47) \end{aligned}$$

wobei $f_{U_i}(u_i)$ die Dichte der Normalverteilung mit Erwartungswert 0 und Varianz 2 ist. Dieses Integral lässt sich schwer berechnen und wird deshalb mit dem oben angegebenen Ausdruck

$$\mathrm{E}(Y_{ij}) = \frac{1}{M} \sum_{i=1}^{M} \frac{\exp(\beta_0^{\mathrm{RE}} + u_i + \beta_1^{\mathrm{RE}} x_{ij})}{1 + \exp(\beta_0^{\mathrm{RE}} + u_i + \beta_1^{\mathrm{RE}} x_{ij})} \qquad (48)$$

mit M Realisierungen u_i von U_i angenähert. Mit den oben genannten Parameterwerten ergibt sich mit $M = 100000$ eine durchschnittliche Infektionswahrscheinlichkeit von 0.23 für Kinder mit Vitamin-A-Mangel und 0.18 für Kinder ohne Mangel. Für ein marginales Modell, das diese Wahrscheinlichkeiten von 0.23 und 0.18 besitzt, werden nun die nötigen Parameter berechnet. Es soll also gelten: $\exp(\beta_0^{\mathrm{M}})/(1 + \exp(\beta_0^{\mathrm{M}})) = 0.18$, und somit ist $\beta_0^{\mathrm{M}} = -1.52$ zu wählen. Für den Parameter β_1^{M}, der den Logarithmus des Odds Ratio beschreibt, ergibt sich somit $\exp(\beta_1^{\mathrm{M}}) = (0.23/(1 - 0.23))/(0.18/(1 - 0.18)) = 1.36$ und damit $\beta_1^{\mathrm{M}} = 0.31$.

Somit sind auch in diesem Beispiel die Parameter unterschiedlich. In dem einem Modell werden die Populations-Odds beschrieben und in dem anderen die individuellen Odds. Auch hier besteht ein Unterschied zwischen den Parametern. Es ist zum einem $\left|\frac{-2}{-1.52}\right| = 1.32$ und zum anderen $\left|\frac{0.4}{0.31}\right| = 1.29$.

4.2 Vergleich des Bahadur- und des Beta-Binomial-Modells

Im Folgenden wird nun das marginale Bahadur-Modell mit austauschbaren Beobachtungen mit dem subjektspezifischen Beta-Binomial-Modell verglichen. Im Bahadur-Modell mit austauschbaren Beobachtungen gibt es bis zu n_i Parameter pro Subjekt - den Parameter π_i für die marginale Erwartung bei einer Beobachtung und noch $n_i - 1$ mögliche Parameter für die Korrelation zweiter, dritter und höherer Ordnung. Im Beta-Binomial-Modell gibt es hingegen nur 2 Parameter - wieder den Parameter π_i für den Erwartungswert für eine Beobachtung und einen Parameter ρ_i, der die Korrelation zwischen zwei Beobachtungen beschreibt. Das Bahadur-Modell ist durch die vielen Parameter für die höheren Korrelationen sehr komplex. Meistens werden die Parameter für die Korrelationen dritter und höherer Ordnung deshalb gleich Null gesetzt. Es entsteht ein Modell mit nur noch zwei Parametern, das nun mit dem Beta-Binomial-Modell vergleichbar ist.
In dem Bahadur-Modell mit austauschbaren Beobachtungen, in dem die höheren Korrelationen gleich Null gesetzt wurden, entstehen Einschränkungen an den Parameterraum. Denn: Gleichung (10), die die Dichte beschreiben soll, ist, wie auf S. 27 gezeigt, nicht für alle möglichen Parameterkombinationen größer gleich Null. In dem folgenden Satz werden von Bahadur (1961) entwickelte Schranken für die Korrelation $\rho_i(2)$ angegeben, für die die in Gleichung (10) angegebene Funktion genau dann größer gleich Null ist, wenn $\rho_i(2)$ diese Schranken einhält.

Satz 36. *Die in Gleichung* (10) *angegebene Funktion ist genau dann größer gleich Null, wenn für* $\rho_i(2)$ *gilt:*

$$-\frac{2}{n_i(n_i-1)}\min\left\{\frac{\pi_i}{1-\pi_i},\frac{1-\pi_i}{\pi_i}\right\}\leq\rho_i(2)\leq\frac{2\pi_i(1-\pi_i)}{(n_i-1)\pi_i(1-\pi_i)+0.25-\gamma_0} \tag{49}$$

mit $\gamma_0=\min\limits_{z_i\in\{0,\ldots,n_i\}}\{[z_i-(n_i-1)\pi_i-0.5]^2\}$.[49]

Beweis. Die Dichte aus (10) ist

$$\begin{aligned} f(z_i) &= \binom{n_i}{z_i}\pi_i^{z_i}(1-\pi_i)^{n_i-z_i} \\ &\quad\times\left[1+\rho_i(2)\left\{\binom{n_i-z_i}{2}\frac{\pi_i}{1-\pi_i}-z_i(n_i-z_i)+\binom{z_i}{2}\frac{1-\pi_i}{\pi_i}\right\}\right]. \end{aligned}$$

Demnach ist $f(z_i)$ genau dann größer gleich Null, wenn

$$\rho_i(2)\left\{\binom{n_i-z_i}{2}\frac{\pi_i}{1-\pi_i}-z_i(n_i-z_i)+\binom{z_i}{2}\frac{1-\pi_i}{\pi_i}\right\}\geq-1$$

gilt. Dies lässt sich umformen zu

$$\frac{\rho_i(2)}{2\pi_i(1-\pi_i)}[(n_i-z_i)(n_i-z_i-1)\pi_i^2-2z_i(n_i-z_i)\pi_i(1-\pi_i)+z_i(z_i-1)(1-\pi_i)^2]\geq-1.$$

Durch Ausmultiplizieren und Zusammenfassen ergibt sich

$$\frac{\rho_i(2)}{2\pi_i(1-\pi_i)}[z_i^2-z_i+2\pi_iz_i-2z_in_i\pi_i+n_i^2\pi_i^2-n_i\pi_i^2]\geq-1.$$

Weitere Umformungen ergeben

$$\begin{aligned} \frac{\rho_i(2)}{2\pi_i(1-\pi_i)}[&z_i^2-2z_i\pi_i(n_i-1)+(n_i-1)^2\pi_i^2-(n_i-1)^2\pi_i^2 \\ &-z_i+(n_i-1)\pi_i-(n_i-1)\pi_i+\frac{1}{4}-\frac{1}{4}+n_i^2\pi_i^2-n_i\pi_i^2]\geq-1, \end{aligned}$$

und

$$\frac{\rho_i(2)}{2\pi_i(1-\pi_i)}[(z_i-(n_i-1)\pi_i)^2-(z_i-(n_i-1)\pi_i)+\frac{1}{4}-\frac{1}{4}+n_i\pi_i^2-\pi_i^2-(n_i-1)\pi_i]\geq-1,$$

[49] Diese Ungleichung und die Beweisidee stammen aus Bahadur (1961).

bzw.

$$\frac{\rho_i(2)}{2\pi_i(1-\pi_i)} \underbrace{\left[\left(z_i-(n_i-1)\pi_i-\frac{1}{2}\right)^2-\frac{1}{4}-(n_i-1)\pi_i(1-\pi_i)\right]}_{:=g(z_i,n_i,\pi_i)} \geq -1.$$

Letztendlich ist also folgende Ungleichung zu betrachten:

$$\frac{\rho_i(2)}{2\pi_i(1-\pi_i)} g(z_i,n_i,\pi_i) \geq -1.$$

Um nun die Schranken für $\rho_i(2)$ zu erhalten, müssen das Maximum und das Minimum von $g(z_i,n_i,\pi_i)$ in Abhängigkeit von z_i bestimmt werden. $g(z_i,n_i,\pi_i)$ wird maximal bzw. minimal, wenn $\left(z_i-(n_i-1)\pi_i-\frac{1}{2}\right)^2$ maximal bzw. minimal wird. $\left(z_i-(n_i-1)\pi_i-\frac{1}{2}\right)^2$ ist in Abhängigkeit von z_i eine verschobene Normalparabel mit ihrem Scheitelpunkt bei $0 \leq (n_i-1)\pi_i+1/2 \leq n_i$. Da z_i die ganzen Zahlen von 1 bis n_i durchläuft, gilt sicher:

$$g_0 := \min_{z_i}\left\{\left(z_i-(n_i-1)\pi_i-\frac{1}{2}\right)^2\right\} \leq \frac{1}{4}.$$

Und somit ist

$$\min_{z_i} g(z_i,n_i,\pi_i) = -\left[\frac{1}{4}+(n_i-1)\pi_i(1-\pi_i)-g_0\right].$$

Die größten Werte von $\left(z_i-(n_i-1)\pi_i-\frac{1}{2}\right)^2$ sind bei 0 und n_i möglich. Es ist $g(0,n_i,\pi_i) = n_i(n_i-1)\pi_i^2$ und $g(n_i,n_i,\pi_i) = n_i(n_i-1)(1-\pi)^2$. Somit ist

$$\max_{z_i} g(z_i,n_i,\pi_i) = n_i(n_i-1)\max\{\pi_i^2,(1-\pi_i)^2\}.$$

Damit folgen die behaupteten Schranken für $\rho_i(2)$. □

Tabelle 1 gibt einige mit Gleichung (49) berechnete mögliche Schranken für $\rho_i(2)$ in Abhängigkeit von n_i und π_i an.

Auch in dem Beta-Binomial-Modell gibt es Einschränkungen an den Parameter ρ_i. Nach Definition ist $\rho_i > 0$! Also sollte es nicht verwendet werden, wenn eine negative Korrelation erwartet wird.

Mit Gleichung (49) berechnete Schranken für $\rho_i(2)$ in Abhängigkeit von n_i und p_i.

	Wahrscheinlichkeit p_i			
n_i	0.05	0.1	0.3	0.5
2	(-0.053, 1.000)	(-0.111, 1.000)	(-0.429, 1.000)	(-1.000, 1.000)
3	(-0.018, 0.513)	(-0.037, 0.529)	(-0.143, 0.636)	(-0.333, 1.000)
4	(-0.009, 0.352)	(-0.019, 0.375)	(-0.071, 0.583)	(-0.167, 0.500)
5	(-0.005, 0.271)	(-0.011, 0.300)	(-0.043, 0.420)	(-0.100, 0.500)
7	(-0.003, 0.192)	(-0.005, 0.231)	(-0.020, 0.296)	(-0.048, 0.333)
10	(-0.001, 0.141)	(-0.002, 0.200)	(-0.010, 0.200)	(-0.022, 0.200)
15	(-0.0005, 0.109)	(-0.001, 0.120)	(-0.004, 0.135)	(-0.010, 0.143)

Tabelle 1: Schranken für $\rho_i(2)$.

Für den Fall $n_i = 2$ verglichen Kupper u. Haseman (1978) das Bahadur- und das Beta-Binomial-Modell: Der Parameter $\pi_i = \pi_i^{Ba}$ aus dem Bahadur-Modell entspricht dem Parameter $\pi_i = \pi_i^{Be}$ aus dem Beta-Binomial-Modell. Denn für eine Zufallsvariable Z_i, die sich durch beide Modelle darstellen lässt, gilt:

$$2\pi_i^{Ba} = E_{Ba}(Z_i) = E_{Be}(Z_i) = 2\pi_i^{Be}.$$

Durch einen Vergleich der Wahrscheinlichkeiten des Bahadur-Modells mit denen des Beta-Binomial-Modells ergibt sich die Vergleichbarkeit der Parameter $\rho_i(2)$ und ρ_i. Zur Berechnung der Wahrscheinlichkeiten des Bahadur-Modells kann Gleichung (10) benutzt werden und es ergibt sich direkt:

$$\begin{aligned}
\mathrm{P}(Z_i = 0) &= (1-\pi_i)^2 \times \left[1 + \rho_i(2)\left(\frac{\pi_i}{1-\pi_i}\right)\right], \\
\mathrm{P}(Z_i = 1) &= 2\pi_i(1-\pi_i) \times \left[1 + \rho_i(2)(-1)\right], \\
\mathrm{P}(Z_i = 2) &= \pi_i^2 \times \left[1 + \rho_i(2)\left(\frac{(1-\pi_i)}{\pi_i}\right)\right].
\end{aligned}$$

Für die Berechnung der Wahrscheinlichkeiten des Beta-Binomial-Modells werden Gleichung (25) und Bemerkung 34 benutzt. Zunächst wird Gleichung (25) zu

$$f_{Z_i}^{(\pi_i,\rho_i)}(z_i) = \binom{2}{z_i} \; \mathrm{E}[\Pi_i^{z_i}(1-\Pi_i)^{2-z_i}].$$

umgeformt. Für $z_i = 0, 1, 2$ ergeben sich dann

$$\begin{aligned}
f_{Z_i}^{(\pi_i,\rho_i)}(0) = &\quad 1 - 2\,\mathrm{E}(\Pi_i) + \mathrm{E}(\Pi_i^2) &&= 1 - 2\,\mathrm{E}(\Pi_i) + \mathrm{Var}(\Pi_i) + E(\Pi_i)^2,\\
f_{Z_i}^{(\pi_i,\rho_i)}(1) = &\quad 2(\mathrm{E}(\Pi_i) - \mathrm{E}(\Pi_i^2)) &&= 2(\mathrm{E}(\Pi_i) - \mathrm{Var}(\Pi_i) - E(\Pi_i)^2),\\
f_{Z_i}^{(\pi_i,\rho_i)}(2) = &\quad \mathrm{E}(\Pi_i^2) &&= \mathrm{Var}(\Pi_i) + E(\Pi_i)^2.
\end{aligned}$$

Mit Bemerkung 34 wird dies zu

$$\begin{aligned}
\mathrm{P}(Z_i = 0) &= (1 - \pi_i)^2 + \rho_i(\pi_i(1 - \pi_i)),\\
\mathrm{P}(Z_i = 1) &= 2\pi_i(1 - \pi_i) - 2\rho_i(\pi_i(1 - \pi_i)),\\
\mathrm{P}(Z_i = 2) &= \pi_i^2 + \rho_i(\pi_i(1 - \pi_i)).
\end{aligned}$$

Durch Gegenüberstellen der Wahrscheinlichkeiten des Bahadur-Modells mit denen des Beta-Binomial-Modells ergibt sich die Vergleichbarkeit der Parameter $\rho_i(2)$ und ρ_i. Die Parameter weisen allerdings den oben erwähnten Unterschied in ihren Wertebereichen auf. Bis auf diese Einschränkung sind in diesem speziellen Fall die beiden Modelle identisch.

5 Versuchsplanung

Zur Entscheidung von statistischen Fragestellungen werden Versuche durchgeführt. Dazu werden statistische Hypothesen aufgestellt. Das „unerwünschte“ Ereignis steht dabei in der Hypothese H_0 und das „erhoffte“ Ereignis in der Alternative H_1. Ziel ist es, mit Hilfe der durch den Versuch erhaltenden Daten, die Hypothese abzulehnen. Dies geschieht durch die Aufstellung einer Teststatistik mit der dann ein Test durchgeführt wird. Dabei treten meist unerwünschte Fehler auf. Der *Fehler 1. Art* (α*-Fehler*), d.h. die Wahrscheinlichkeit, dass die Hypothese H_0 abgelehnt wird, obwohl sie wahr ist, ist dabei der schlimmste Fehler. Der andere Fehler, der dabei auftreten kann, ist der *Fehler 2. Art* (β*-Fehler*), d.h. die Wahrscheinlichkeit, die Nullhypothese H_0 nicht abzulehnen, obwohl H_0 falsch ist. Die gleichzeitige Minimierung beider Fehlerwahrscheinlichkeiten ist nicht möglich. Allerdings kann der Fehler 1. Art konstant gehalten und der Fehler 2. Art verringert werden, wenn der Stichprobenumfang vergrößert wird.[50] Also wird ein genügend großer Stichprobenumfang benötigt, damit relevante Unterschiede erkannt werden. Allerdings kosten Versuche Zeit und Geld und darum sollte die Anzahl der Einzelversuche (Stichprobenumfang) möglichst klein sein. Das bedeutet, dass das Hauptproblem in der Bestimmung des richtigen Stichprobenumfangs liegt.

Mit einer Beobachtung pro Person müssen zur Überprüfung einer Hypothese also die folgenden, oben angesprochenen Kriterien festgelegt werden, um später einen aussagekräftigen Test zu erhalten:[51]

1. Zunächst wird der *Fehler 1. Art* festgelegt. Eine typische Wahl für diesen Fehler ist $\alpha = 0.05$.

2. Anschließend wird die *kleinste bedeutungsvolle Differenz d, die erkannt werden soll*, bestimmt. Die Nullhypothese soll mit einer hohen Wahrscheinlich-

[50] vgl. Genschel u. Becker (2005)

[51] Die folgenden Kriterien stammen aus Diggle u. a. (2002).

keit abgelehnt werden, wenn der betrachtete Parameter von dem Parameterwert unter der Nullhypothese um d oder mehr abweicht. Dieser Wert d wird nach praktischen Gesichtspunkten bestimmt.

3. Außerdem muss die *Power* oder *Güte*, die der Test bei der kleinsten bedeutungsvollen Differenz d haben soll, festgelegt werden. Die *Power (Güte)* ist die Wahrscheinlichkeit, dass der Test die Nullhypothese ablehnt, wenn sie falsch ist. Typischerweise werden Tests mit einer Power von 0.8 oder 0.9 bei der kleinsten bedeutungsvollen Differenz d konstruiert.

Für wiederholte Daten müssen zusätzlich noch weitere Entscheidungen getroffen werden:

1. Die *Korrelation* ρ der wiederholten Beobachtungen muss mit in Betracht gezogen werden. Entweder werden dafür Erfahrungswerte verwendet oder es müssen begründete Vermutungen benutzt werden.

2. Letztendlich muss die Anzahl n_i der Beobachtungen pro Person festgesetzt werden.

Mit diesen Werten lässt sich dann der benötigte Stichprobenumfang bestimmen.

5.1 Berechnung des benötigten Stichprobenumfangs

Die Berechnung des benötigten Stichprobenumfangs wird hier für ein Zwei-Stichproben-Problem durchgeführt. Zwei gleich große Probanden-Gruppen erhalten die Behandlungen A bzw. B. Untersucht wird die Wirkung der Behandlungen. Die beobachtete Variable ist demnach binär (0 für Symptom nicht aufgetreten bzw. Behandlung wirksam, 1 nicht wirksam). Seien p_A und p_B die entsprechenden Erfolgswahrscheinlichkeiten für das Versagen der Behandlung. Die Hypothese lautet damit: $H_0 : p_A = p_B$ gegen $H_1 : p_A \neq p_B$. Ziel ist es, aus dieser Hypothese aus einem Test eine Formel für den benötigten Umfang herzuleiten. Sei dazu

der Fehler 1. Art durch die Variable α beschrieben, damit die Formel später für unterschiedliche Niveaus anwendbar ist. In klinischen Studien ist meist eine der Behandlungen die Kontrolle, von der die Erfolgswahrscheinlichkeit schon bekannt ist. Sei in diesem Beispiel Behandlung A die Kontrolle, womit p_A bekannt sei. Nun wird die kleinste bedeutungsvolle Differenz d festgesetzt, indem ein Odds Ratio OR festgelegt wird.[52] Wird vermutet, dass die Behandlung B besser wirkt als die Kontrolle A, also $p_B < p_A$, wird ein Odds Ratio kleiner 1 festgesetzt (oder größer 1, wenn die Behandlung B mit einer schlechteren Wirkung vermutet wird). Dieser Odds Ratio kann mit p_A in eine Erfolgswahrscheinlichkeit p_B^d für Behandlung B umgerechnet werden. Somit ergibt sich eine kleinste bedeutungsvolle Differenz $d = |p_A - p_B^d|$, bei der die Nullhypothese mit einer Wahrscheinlichkeit von $1 - \beta$ (β = Fehler 2. Art) abgelehnt werden sollte. Eine weitere Eigenschaft der Gütefunktion sollte sein (halte den Stichprobenumfang vorerst konstant), dass die Güte um so größer wird, je weiter die Parameterwerte von denen der Nullhypothese abweichen. In diesem Fall sollte also die Güte bzw. Power umso größer werden, je mehr der Odds Ratio von 1 abweicht bzw. je größer die Differenz $|p_A - p_B|$ wird. Wenn also die Power von $1 - \beta$ schon bei d erreicht wird, dann sollte sie erst recht für Werte größer als d erreicht werden. Deshalb wird zunächst die Suche nach dem benötigten Stichprobenumfang N bei der Differenz d (durch den Odds Ratio bestimmt!) durchgeführt. Anschließend wird die Gütefunktion betrachtet, ob sie auch die oben angesprochenen Eigenschaften aufweist.

Zur Suche des benötigten Stichprobenumfangs wird der Umstand genutzt, dass sich die Power an jeder Stelle der Alternative verbessert, wenn der Stichprobenumfang N vergrößert wird. Gesucht ist nun der mindestens benötige Stichprobenumfang, mit dem es mit dem gewählten Test zum Niveau α eine Power von $1 - \beta$ bei einer Differenz d zum Entscheiden der Hypothese $H_0 : p_A = p_B$ gegen $H_1 : p_A \neq p_B$ (bzw: $H_0 : OR = 1$ gegen $H_1 : OR \neq 1$) geben kann.

[52] Odds Ratio $OR = \frac{\frac{p_B}{1-p_B}}{\frac{p_A}{1-p_A}}$

5.1.1 Eine Beobachtung pro Person

Für das oben angesprochene Problem lässt sich der benötigte Stichprobenumfang mit einer Variante des approximativen Gauß-Tests bestimmen. Gesucht wird die Anzahl N_G der benötigten Beobachtungen pro Gruppe, die mindestens gemacht werden müssen, damit der Test eine Power von $1-\beta$ bei einer Differenz $d = |p_A - p_B^d|$ hat.

Satz 37. *Für das oben angesprochene Zwei-Stichproben-Problem ist für einen Test, der sich vom approximativen Gaußtest ableitet und ein Niveau α sowie eine Power von $1-\beta$ bei einer kleinsten bedeutungsvollen Differenz $d = |p_A - p_B^d|$ besitzen soll,*

$$N_G \geq \frac{(z_{1-\alpha/2} + z_{1-\beta})^2 (p_A(1-p_A) + p_B^d(1-p_B^d))}{(p_A - p_B^d)^2} \tag{50}$$

zu wählen. Dabei sind $z_{1-\alpha/2}$ und $z_{1-\beta}$ die $(1-\alpha/2)$- bzw. $(1-\beta)$-Quantile der Standardnormalverteilung.

Im Beweis dieses Satzes wird folgender Hilfssatz benutzt:

Lemma 38. *Seien $X, X_n, Y_n, n \geq 1$, reellwertige Zufallsvariablen, wobei X_n und Y_n bei festem n auf demselben Wahrscheinlichkeitsraum erklärt seien. Unter der Voraussetzung $X_n \xrightarrow{D} X$ und $Y_n \xrightarrow{P} c$ für ein festes $c \in \mathbb{R}$ gilt*

$$X_n Y_n \xrightarrow{D} Xc. \tag{51}$$

Beweis. Der Beweis dieses Lemmas ist in Behnen u. Neuhaus (2003) auf S. 111 zu finden. □

Beweis von Satz 37. Es wird die Teststatistik

$$T = \frac{Y_A - Y_B}{\sqrt{\frac{p_A(1-p_A) + Y_B(1-Y_B)}{N_G}}} \tag{52}$$

für das Testproblem $H_0 : p_A = p_B$ gegen $H_1 : p_A \neq p_B$ (bzw. $H_0 : OR = 1$ gegen $H_1 : OR \neq 1$) benutzt. Sei p_A aus Erfahrung bekannt und sei p_B^d durch p_A und

einem Odds Ratio unter der Alternative festgelegt (Durch diese Festlegung wird also auch d festgelegt.). Weiter ist $Y_A = \frac{1}{N_G}\sum_{i=1}^{N_G} Y_{Ai}$ das arithmetische Mittel der Beobachtungen mit Behandlung A, Y_B entsprechend. Wenn N_G genügend groß ist ($N_G > 30$), sind Y_A und Y_B approximativ normalverteilt mit Erwartungswert p_A bzw. p_B und Varianz $\frac{1}{N_G}p_A(1-p_A)$ und $\frac{1}{N_G}p_B(1-p_B)$. Außerdem ist

$$\begin{aligned} T &= \frac{Y_A - Y_B}{\sqrt{\frac{p_A(1-p_A)+Y_B(1-Y_B)}{N_G}}} \frac{\sqrt{p_A(1-p_A)+p_B(1-p_B)}}{\sqrt{p_A(1-p_A)+p_B(1-p_B)}} \\ &= \frac{Y_A - Y_B}{\sqrt{\frac{p_A(1-p_A)+p_B(1-p_B)}{N_G}}} \frac{\sqrt{p_A(1-p_A)+p_B(1-p_B)}}{\sqrt{p_A(1-p_A)+Y_B(1-Y_B)}}. \end{aligned}$$

Unter H_0 konvergiert der erste Bruch in Verteilung für $N_G \to \infty$ gegen $\mathcal{N}(0,1)$ (Zentraler Grenzwertsatz). Der zweite Bruch konvergiert in Wahrscheinlichkeit gegen 1 (wegen des schwachen Gesetzes der großen Zahlen und den Rechenregeln für Konvergenzen in Wahrscheinlichkeit). Damit ist mit Lemma 38 T unter H_0 asymptotisch standardnormalverteilt. Für ein genügend großes N_G ist T unter H_1 approximativ normalverteilt mit Erwartungswert $\dfrac{p_A - p_B}{\sqrt{\frac{p_A(1-p_A)+p_B(1-p_B)}{N_G}}}$ und Varianz 1.

Für große N_G sollen folgende Gleichungen für einen Test zum Niveau α und mit einer Power $1-\beta$ bei einer Differenz d annähernd gelten:

$$\begin{aligned} \mathrm{P}_d(|T| \geq z_{1-\alpha/2}) &\leq \alpha \qquad \text{unter } H_0, \\ \mathrm{P}_d(|T| \geq z_{1-\alpha/2}) &\geq 1-\beta \qquad \text{unter } H_1. \end{aligned}$$

Aus der unteren Gleichung lässt sich nun die Bedingung für N_G herleiten:

$$\begin{aligned} & \mathrm{P}_d(|T| > z_{1-\alpha/2}) && \geq 1-\beta \\ \Leftrightarrow \quad & 1 - \mathrm{P}_d(|T| \leq z_{1-\alpha/2}) && \geq 1-\beta \\ \Leftrightarrow \quad & 1 - \mathrm{P}_d(-z_{1-\alpha/2} \leq T \leq z_{1-\alpha/2}) && \geq 1-\beta \\ \Leftrightarrow \quad & 1 - (\mathrm{P}_d(T \leq z_{1-\alpha/2}) - \mathrm{P}_d(T \leq -z_{1-\alpha/2})) && \geq 1-\beta \\ \Leftrightarrow \quad & 1 - \underbrace{\mathrm{P}_d(T \leq z_{1-\alpha/2})}_{*} + \underbrace{\mathrm{P}_d(T \leq -z_{1-\alpha/2})}_{**} && \geq 1-\beta \end{aligned} \tag{53}$$

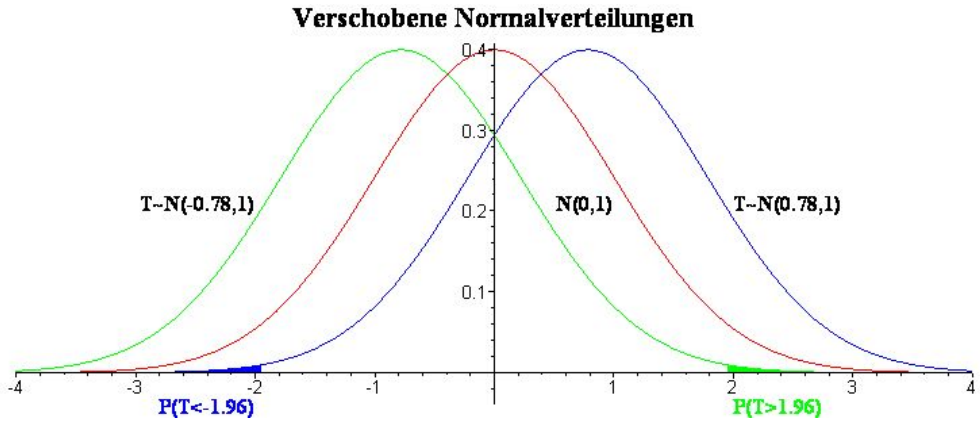

Abbildung 2: Verschobene Normalverteilungen

Unter H_1 an der Stelle d ist $T = \dfrac{Y_A - Y_B}{\sqrt{\frac{p_A(1-p_A)+Y_B(1-Y_B)}{N_G}}}$ eine verschobene Normalverteilung mit Erwartungswert $\dfrac{p_A - p_B^d}{\sqrt{\frac{p_A(1-p_A)+p_B^d(1-p_B^d)}{N_G}}}$ und Varianz 1.

Für $d = p_A - p_B^d > 0$ wird $**$ in Gleichung (53) ungefähr Null. (Siehe Abb. 2, $\alpha = 0.05$. Schon bei einem Behandlungsunterschied von nur $d = 0.05$ wird für $N_G = 30$ diese Wahrscheinlichkeit kleiner als 0.01. Bei ansteigendem Behandlungsunterschied oder größerem Stichprobenumfang N_G wird die Wahrscheinlichkeit rasch noch kleiner. Die Idee, diese „kleine“ Wahrscheinlichkeit zu vernachlässigen, stammt aus Henning (2002).) Somit wird Gleichung (53) in diesem Fall zu:

$$1 - \mathrm{P}_d \left(\frac{Y_A - Y_B}{\sqrt{\frac{p_A(1-p_A)+Y_B(1-Y_B)}{N_G}}} \leq z_{1-\alpha/2} \right) \geq 1 - \beta. \tag{54}$$

Es ist

$$\begin{aligned} & 1 - \mathrm{P}_d \left(\frac{Y_A - Y_B}{\sqrt{\frac{p_A(1-p_A)+Y_B(1-Y_B)}{N_G}}} \leq z_{1-\alpha/2} \right) \\ = \; & 1 - \mathrm{P}_d \left(T - \frac{p_A - p_B^d}{\sqrt{\frac{p_A(1-p_A)+p_B^d(1-p_B^d)}{N_G}}} \leq z_{1-\alpha/2} - \frac{p_A - p_B^d}{\sqrt{\frac{p_A(1-p_A)+p_B^d(1-p_B^d)}{N_G}}} \right) \end{aligned}$$

$$\approx \quad 1 - \Phi\left(z_{1-\alpha/2} - \frac{p_A - p_B^d}{\sqrt{\frac{p_A(1-p_A)+p_B^d(1-p_B^d)}{N_G}}} \right),$$

wobei die Approximation für genügend große N_G gilt. Damit wird Gleichung (54) zu

$$\begin{aligned}
& 1 - \Phi\left(z_{1-\alpha/2} - \frac{p_A - p_B^d}{\sqrt{\frac{p_A(1-p_A)+p_B^d(1-p_B^d)}{N_G}}} \right) && \geq 1-\beta \\
\Leftrightarrow \quad & \Phi\left(z_{1-\alpha/2} - \frac{p_A - p_B^d}{\sqrt{\frac{p_A(1-p_A)+p_B^d(1-p_B^d)}{N_G}}} \right) && \leq \beta \\
\Leftrightarrow \quad & z_{1-\alpha/2} - \frac{p_A - p_B^d}{\sqrt{\frac{p_A(1-p_A)+p_B^d(1-p_B^d)}{N_G}}} && \leq z_\beta = -z_{1-\beta} \\
\Leftrightarrow \quad & \frac{p_A - p_B^d}{\sqrt{\frac{p_A(1-p_A)+p_B^d(1-p_B^d)}{N_G}}} && \geq z_{1-\alpha/2} + z_{1-\beta} \\
\overset{(\oplus)}{\Leftrightarrow} \quad & \frac{(p_A - p_B^d)^2}{\frac{p_A(1-p_A)+p_B^d(1-p_B^d)}{N_G}} && \geq (z_{1-\alpha/2} + z_{1-\beta})^2 \\
\Leftrightarrow \quad & \frac{(z_{1-\alpha/2} + z_{1-\beta})^2 (p_A(1-p_A) + p_B^d(1-p_B^d))}{(p_A - p_B^d)^2} && \leq N_G.
\end{aligned}$$

Für $d = p_A - p_B^d < 0$ ist $*$ in Gleichung (53) ungefähr 1. Somit wird Gleichung (53) zu:

$$\mathrm{P}_d\left(\frac{Y_A - Y_B}{\sqrt{\frac{p_A(1-p_A)+Y_B(1-Y_B)}{N_G}}} \leq -z_{1-\alpha/2} \right) \geq 1-\beta. \tag{55}$$

Es ist

$$\begin{aligned}
& \mathrm{P}_d\left(\frac{Y_A - Y_B}{\sqrt{\frac{p_A(1-p_A)+Y_B(1-Y_B)}{N_G}}} \leq -z_{1-\alpha/2} \right) \\
= \quad & \mathrm{P}_d\left(T - \frac{p_A - p_B^d}{\sqrt{\frac{p_A(1-p_A)+p_B^d(1-p_B^d)}{N_G}}} \leq -z_{1-\alpha/2} - \frac{p_A - p_B^d}{\sqrt{\frac{p_A(1-p_A)+p_B^d(1-p_B^d)}{N_G}}} \right) \\
\approx \quad & \Phi\left(-z_{1-\alpha/2} - \frac{p_A - p_B^d}{\sqrt{\frac{p_A(1-p_A)+p_B^d(1-p_B^d)}{N_G}}} \right),
\end{aligned}$$

wobei die Approximation wieder für genügend große N_G gilt. Damit wird Gleichung (55) zu

$$\begin{aligned}
& \Phi\left(-z_{1-\alpha/2} - \frac{p_A - p_B^d}{\sqrt{\frac{p_A(1-p_A)+p_B^d(1-p_B^d)}{N_G}}}\right) && \geq 1-\beta \\
\Leftrightarrow \quad & -z_{1-\alpha/2} - \frac{p_A - p_B^d}{\sqrt{\frac{p_A(1-p_A)+p_B^d(1-p_B^d)}{N_G}}} && \geq z_{1-\beta} \\
\Leftrightarrow \quad & -\frac{p_A - p_B}{\sqrt{\frac{p_A(1-p_A)+p_B^d(1-p_B^d)}{N_G}}} && \geq z_{1-\alpha/2} + z_{1-\beta} \\
\overset{(\oplus)}{\Leftrightarrow} \quad & \frac{(p_A - p_B^d)^2}{\frac{p_A(1-p_A)+p_B^d(1-p_B^d)}{N_G}} && \geq (z_{1-\alpha/2} + z_{1-\beta})^2 \\
\Leftrightarrow \quad & \frac{(z_{1-\alpha/2} + z_{1-\beta})^2(p_A(1-p_A) + p_B^d(1-p_B^d))}{(p_A - p_B^d)^2} && \leq N_G.
\end{aligned}$$

Zu ($\oplus$): Für kleine α- und β-Werte sind die Quantile größer als Null. Dann gilt die Umformung. Die typischen Wahlen für $\alpha = 0.05$ und $\beta = 0.9$ oder 0.8 erfüllen diese Bedingung.

N_G ist also mindestens so groß wie der obige Bruch zu wählen, damit der Test zum Niveau α eine Power von $1-\beta$ bei einer kleinsten bedeutungsvollen Differenz d besitzen kann. □

Bemerkung 39. Als gesamter Stichprobenumfang N ist demnach

$$N \geq \frac{2(z_{1-\alpha/2} + z_{1-\beta})^2(p_A(1-p_A) + p_B^d(1-p_B^d))}{(p_A - p_B^d)^2} \tag{56}$$

zu wählen.

Bemerkung 40. Bei einem größeren Abstand $d^* > d$ mit $d^* = |p_A - p_b^{d*}|$ und der Anzahl N_G, die für d berechnet wurde, ist die Power auch größer als $1-\beta$, denn es gilt

$$\begin{aligned}
& 1 - \beta(d^*) \\
= \quad & 1 - P_{d^*}(T_{N_G} \leq z_{1-\alpha/2}) + P_{d^*}(T_{N_G} \leq -z_{1-\alpha/2})) \\
\approx \quad & 1 - \Phi\left(z_{1-\alpha/2} - \frac{p_A - p_B^{d*}}{\sqrt{\frac{p_A(1-p_A)+p_B^{d*}(1-p_B^{d*})}{N_G}}}\right) + \Phi\left(-z_{1-\alpha/2} - \frac{p_A - p_B^{d*}}{\sqrt{\frac{p_A(1-p_A)+p_B^{d*}(1-p_B^{d*})}{N_G}}}\right).
\end{aligned}$$

Nun ist $\frac{p_A - p_B^{d*}}{\sqrt{\frac{p_A(1-p_A)+p_B^{d*}(1-p_B^{d*})}{N_G}}}$ mit $p_B^{d*} = \frac{OR\frac{p_A}{1-p_A}}{1+OR\frac{p_A}{1-p_A}}$ für konstantes p_A eine in OR monoton fallende Funktion, die eine Nullstelle bei einem $OR = 1$ besitzt (Beweis siehe Anhang). Das bedeutet für $p_B^{d*} > p_B^d > p_A$ (also $OR_{p_B^{d*}} > OR_{p_B^d} > 1$):

$$
\begin{aligned}
& 1 - \Phi\left(z_{1-\alpha/2} - \frac{p_A - p_B^{d*}}{\sqrt{\frac{p_A(1-p_A)+p_B^{d*}(1-p_B^{d*})}{N_G}}}\right) + \Phi\left(-z_{1-\alpha/2} - \frac{p_A - p_B^{d*}}{\sqrt{\frac{p_A(1-p_A)+p_B^{d*}(1-p_B^{d*})}{N_G}}}\right) \\
> \quad & \Phi\left(-z_{1-\alpha/2} - \frac{p_A - p_B^{d*}}{\sqrt{\frac{p_A(1-p_A)+p_B^{d*}(1-p_B^{d*})}{N_G}}}\right) \\
\geq \quad & \Phi\left(-z_{1-\alpha/2} - \frac{p_A - p_B^{d}}{\sqrt{\frac{p_A(1-p_A)+p_B^{d}(1-p_B^{d})}{N_G}}}\right) \\
\geq \quad & 1 - \beta.
\end{aligned}
$$

Für $p_B^{d*} < p_B^d < p_A$ (also $OR_{p_B^{d*}} < OR_{p_B^d} < 1$) bedeutet das entsprechend:

$$
\begin{aligned}
& 1 - \Phi\left(z_{1-\alpha/2} - \frac{p_A - p_B^{d*}}{\sqrt{\frac{p_A(1-p_A)+p_B^{d*}(1-p_B^{d*})}{N_G}}}\right) + \Phi\left(-z_{1-\alpha/2} - \frac{p_A - p_B^{d*}}{\sqrt{\frac{p_A(1-p_A)+p_B^{d*}(1-p_B^{d*})}{N_G}}}\right) \\
> \quad & 1 - \Phi\left(z_{1-\alpha/2} - \frac{p_A - p_B^{d*}}{\sqrt{\frac{p_A(1-p_A)+p_B^{d*}(1-p_B^{d*})}{N_G}}}\right) \\
\geq \quad & 1 - \Phi\left(z_{1-\alpha/2} - \frac{p_A - p_B^{d}}{\sqrt{\frac{p_A(1-p_A)+p_B^{d}(1-p_B^{d})}{N_G}}}\right) \\
\geq \quad & 1 - \beta.
\end{aligned}
$$

Zum Beispiel ist in Abbildung 3 die Gütefunktion für $p_A = 0.5$ und einem N_G, das eine Power von 0.9 bei einem Odds Ratio von 2 liefert, zu sehen.

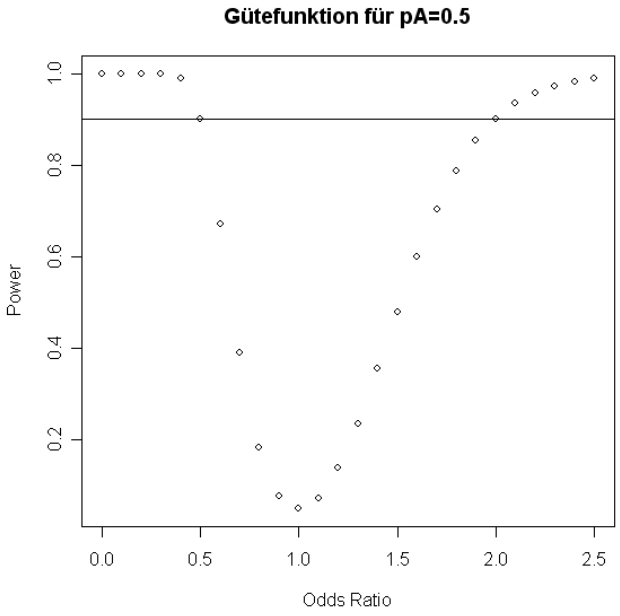

Abbildung 3: Gütefunktion

5.1.2 Wiederholte Beobachtungen

Für marginale Modelle mit wiederholte Beobachtungen lässt sich der benötigte Stichprobenumfang mit einem Verfahren berechnen, das Liu u. Liang (1997) entwickelt haben. Unter anderem kann der Stichprobenumfang N für das oben schon mehrmals erwähnte Modell mit zwei Gruppen mit binären Beobachtungen

$$\text{logit}(\mu_{ij}) = \beta_0 + \beta_1 x_{ij}$$

ermittelt werden. Dabei wird wieder davon ausgegangen, dass die Erfolgswahrscheinlichkeit p_A der Kontrolle bekannt ist und die kleinste bedeutungsvolle Differenz d durch die Festsetzung von p_B bzw. durch den OR festgelegt wird.

Satz 41. *Für ein Zwei-Stichprobenvergleich mit austauschbaren binären Beobachtungen und erklärenden Variablen $x_{ij} = 0$ für die Behandlung A und $x_{ij} = 1$ für die Behandlung B mit dem Modell*

$$\text{logit}(\mu_{ij}) = \beta_0 + \beta_1 x_{ij}$$

für $j = 1, \ldots, n_i = n$ und einer Korrelation ρ ergibt sich als mindestens benötigter Stichprobenumfang N für einen von Liu u. Liang (1997) benutzten Test zum Niveau α und einer Power $1 - \beta$ bei einer Differenz $d = |p_A - p_B|$:

$$N \geq \frac{2(z_{1-\alpha/2} + z_{1-\beta})^2 (p_A(1-p_A) + p_B(1-p_B))(1 + (n-1)\rho)}{n(p_B - p_A)^2}. \qquad (57)$$

Dabei ist z_x das x-Quantil der Standardnormalverteilung.

Zum Beweis dieses Satzes wird zunächst das Verfahren zur Fallzahlberechnung aus Liu u. Liang (1997) erläutert. Der obige Satz folgt dann als Anwendungsbeispiel.

5.1.2.1 Das Verfahren zur Fallzahlberechnung von Liu und Liang

Sei $Y_i = (y_{i1}, y_{i2}, \ldots, y_{in_i})^\top$ der Beobachtungsvektor des i-ten Subjekts (bzw. Clusters). Außerdem sei ein marginales Modell

$$g(\mu_{ij}) = x_{ij}^\top \psi + z_{ij}^\top \lambda, \quad j = 1, \ldots, n_i, \; i = 1, \ldots, N, \qquad (58)$$

gegeben, wobei $\mu_{ij} = \mathrm{E}(y_{ij})$ und g eine Linkfunktion ist, die den Erwartungswert von y_{ij} mit den erklärenden Variablen x_{ij} und z_{ij} verbindet. Dabei sei x_{ij} ein Vektor vom Format $p \times 1$ und z_{ij} von $q \times 1$. Weiter sei ψ der interessante Parameter des Formats $p \times 1$ und λ ein $(q \times 1)$-Vektor mit Parametern, die weniger von Interesse sind. Die zu untersuchende Hypothese laute: $H_0 : \psi = \psi_0$ gegen $H_1 : \psi \neq \psi_0$. Allerdings erfolgt die Bestimmung der Fallzahl wie oben an einer festen Stelle der Alternative, so dass hier die Alternative $H_1 : \psi = \psi_1$ benutzt wird. Durch die Festlegung von ψ_1 wird indirekt der kleinste bedeutungsvolle Unterschied d, der erkannt werden soll, festgesetzt.

Zur Entscheidung dieses Testproblems wird folgende Teststatistik benutzt, die auf den GEEs (siehe Abschnitt 3.1.3.1) basiert:

$$T = S_\psi(\psi_0, \hat{\lambda}_0, \hat{\alpha})^\top \hat{\Sigma}^{-1} S_\psi(\psi_0, \hat{\lambda}_0, \hat{\alpha}), \qquad (59)$$

wobei

$$S_\psi(\psi_0, \hat{\lambda}_0, \hat{\alpha}) = \left(\sum_{i=1}^{N} \left(\frac{\partial \mu_i}{\partial \psi} \right)^\top V_i^{-1} (Y_i - \mu_i) \right) (\psi_0, \hat{\lambda}_0, \hat{\alpha})$$

ist. Dabei sei $\mu_i = \mathrm{E}(Y_i)$. Diese Gleichung wird an den Stellen ψ_0, $\hat{\lambda}_0$ und α betrachtet, so dass für die Einträge in μ_i gilt: $\mu_{ij}(\psi_0, \hat{\lambda}_0, \alpha) = g^{-1}(x_{ij}^\top \psi_0 + z_{ij}^\top \hat{\lambda}_0)$. Es ist

$$\Sigma = \mathrm{Cov}(S_\psi(\psi_0, \hat{\lambda}_0, \alpha)),$$

und $\hat{\Sigma} \to \Sigma$ für großes N. Weiter ist $\hat{\lambda}_0$ ein Schätzer für λ unter H_0, der durch Lösen der Gleichung

$$S_\lambda(\psi_0, \lambda, \alpha) = \left(\sum_{i=1}^{N} \left(\frac{\partial \mu_i}{\partial \lambda} \right)^\top V_i^{-1} (Y_i - \mu_i) \right) (\psi_0, \lambda, \alpha) = 0 \tag{60}$$

erhalten wird. Hier ist $\mu_{ij}(\psi_0, \lambda, \alpha) = g^{-1}(x_{ij}^\top \psi_0 + z_{ij}^\top \lambda)$ und V_i die Kovarianz-Matrix von Y_i, die durch den Parameter α und durch ϕ und λ bestimmt und damit unabhängig von Y_i modelliert wird. (Siehe Abschnitt 3.1.3.1: $V_i = A_i^{1/2} R_i(\alpha) A_i^{1/2}$. Dort sind auch die verschiedenen Möglichkeiten für $R_i(\alpha)$ aufgeführt. Da die GEEs auch dann noch konsistente Schätzungen für die Hauptparameter liefern, wenn die Korrelationssstruktur falsch angenommen wurde[53], wird der Parameter α im folgenden als bekannnt vorausgesetzt und vernachlässigt.)

Satz 42. *Für $N \to \infty$ konvergiert T unter H_0 gegen eine χ^2_p-Verteilung. Unter H_1 mit $\psi = \psi_1$ und $\lambda = \lambda_1$ folgt T asymptotisch einer nichtzentralen χ^2_p-Verteilung mit Nichtzentralitätsparameter*

$$\nu = \xi^\top \Sigma_1^{-1} \xi, \tag{61}$$

wobei $\xi = \mathrm{E}_{H_1}(S_\psi(\psi_0, \hat{\lambda}_0))$ und $\Sigma_1 = \mathrm{Cov}_{H_1}(S_\psi(\psi_0, \hat{\lambda}_0))$.

Beweis. Unter H_0 ist $\hat{\Sigma} \approx \Sigma_0 = \mathrm{Cov}_{H_0}(S_\psi(\psi_0, \hat{\lambda}_0))$. Damit ist

$$\begin{aligned} T &= S_\psi(\psi_0, \hat{\lambda}_0)^\top \Sigma_0^{-1/2} \Sigma_0^{1/2} \hat{\Sigma}^{-1} \Sigma_0^{1/2} \Sigma_0^{-1/2} S_\psi(\psi_0, \hat{\lambda}_0) \\ &\approx S_\psi(\psi_0, \hat{\lambda}_0)^\top \Sigma_0^{-1/2} \Sigma_0^{-1/2} S_\psi(\psi_0, \hat{\lambda}_0). \end{aligned}$$

[53] siehe Liang u. Zeger (1986)

Für $N \to \infty$ ist $\Sigma_0^{-1/2} S_\psi(\psi_0, \hat{\lambda}_0) = \Sigma_0^{-1/2} \left(\sum_{i=1}^{N} \left(\frac{\partial \mu_i}{\partial \psi} \right)^\top V_i^{-1} (Y_i - \mu_i) \right) (\psi_0, \hat{\lambda}_0)$ asymptotisch normalverteilt mit Erwartungswert

$$\begin{aligned} & \Sigma_0^{-1/2} \left(\sum_{i=1}^{N} \left(\frac{\partial \mu_i}{\partial \psi} \right)^\top V_i^{-1} \underbrace{(\mathrm{E}_{H_0}(Y_i) - \mu_i)}_{=0} \right) (\psi_0, \hat{\lambda}_0) \\ = \; & 0_{p \times 1} \end{aligned}$$

und Kovarianzmatrix

$$\begin{aligned} & \Sigma_0^{-1/2} \operatorname{Cov}_{H_0}(S_\psi(\psi_0, \hat{\lambda}_0))(\Sigma_0^{-1/2})^\top \\ = \; & \Sigma_0^{-1/2} \Sigma_0 \Sigma_0^{-1/2} \\ = \; & E_{p \times p}. \end{aligned}$$

Damit ist $\Sigma_0^{-1/2} S_\psi(\psi_0, \hat{\lambda}_0) \sim \mathcal{N}_p(0_p, E_{p \times p})$ und somit besitzt jeder Eintrag in diesem $(p \times 1)$-Vektor eine $\mathcal{N}(0,1)$-Verteilung. Dann ist T als Summe über diese p quadrierten normalverteilten Zufallsvariablen χ^2_p-verteilt.
Unter H_1 wird Σ zu $\Sigma_1 = \operatorname{Cov}_{H_1}(S_\psi(\psi_0, \hat{\lambda}_0)$. T bleibt χ^2_p-verteilt, allerdings nicht zentral. Der Nichtzentralitätsparameter ν ist die Summe der quadrierten Erwartungswerte μ_i der p normalverteilten Zufallsvariablen aus $\Sigma_1^{-1/2} S_\psi(\psi_0, \hat{\lambda}_0))$ unter H_1:

$$\begin{aligned} \nu &= \sum_{i=1}^{p} \mu_i^2 \\ &= (\mu_1, \mu_2, \ldots, \mu_p)(\mu_1, \mu_2, \ldots, \mu_p)^\top \\ &= \mathrm{E}_{H_1}(S_\psi(\psi_0, \hat{\lambda}_0))^\top \Sigma_1^{-1/2} \Sigma_1^{-1/2} \mathrm{E}_{H_1}(S_\psi(\psi_0, \hat{\lambda}_0)) \\ &= \xi^\top \Sigma_1^{-1} \xi. \end{aligned}$$

□

Die folgenden Approximationen werden aus Liu u. Liang (1997) ohne Beweis übernommen, da der Beweis im Rahmen dieser Arbeit nicht mehr erbracht werden kann.

Bemerkung 43. Der Erwartungswert von $S_\psi(\psi_0, \hat{\lambda}_0)$ unter H_1 wird folgendermaßen approximiert:

$$\xi = \mathrm{E}_{H_1}(S_\psi(\psi_0, \hat{\lambda}_0)) \approx \sum_{i=1}^{N} P_i^* V_i^{-1} (\mu_i^1 - \mu_i^*), \tag{62}$$

wobei $\mu_i^1 = g^{-1}(x_{ij}^\top \psi_1 + z_{ij}^\top \lambda_1)$ und $\mu_{ij}^* = g^{-1}(x_{ij}^\top \psi_0 + z_{ij}^\top \lambda_0^*)$,

$$\begin{aligned} P_i^* &= \left(\left(\frac{\partial \mu_i}{\partial \psi} \right)^\top - I_{\psi\lambda}^* I_{\lambda\lambda}^{*-1} \left(\frac{\partial \mu_i}{\partial \lambda} \right) \right) (\psi_0, \lambda_0^*), \\ I_{\psi\lambda}^* &= \left(\sum_{i=1}^{N} \left(\frac{\partial \mu_i}{\partial \psi} \right)^\top V_i^{-1} \left(\frac{\partial \mu_i}{\partial \lambda} \right) \right) (\psi_0, \lambda_0^*), \end{aligned}$$

und

$$I_{\lambda\lambda}^* = \left(\sum_{i=1}^{N} \left(\frac{\partial \mu_i}{\partial \lambda} \right)^\top V_i^{-1} \left(\frac{\partial \mu_i}{\partial \lambda} \right) \right) (\psi_0, \lambda_0^*)$$

ist. Dabei ist λ_0^* der Grenzwert von $\hat{\lambda}_0$ für $N \to \infty$ mit gegebenen ψ_1 und λ_1. λ_0^* ist die Lösung von

$$\lim_{N \to \infty} N^{-1} \mathrm{E}_{H_1}(S_\lambda(\psi_0, \lambda_0^*); \psi_1, \lambda_1) = 0. \tag{63}$$

Bemerkung 44. Die Kovarianzmatrix Σ_1 von $S_\psi(\psi_0, \hat{\lambda}_0)$ wird approximiert durch

$$\Sigma_1 = \mathrm{Cov}_{H_1}(S_\psi(\psi_0, \hat{\lambda}_0)) \approx \sum_{i=1}^{N} P_i^* V_i^{-1} \mathrm{Cov}_{H_1}(Y_i) V_i^{-1} P_i^{*\top}. \tag{64}$$

Damit kann nun die Power zum Testen von H_0 gegen H_1 durch die nichtzentrale χ^2-Verteilung approximiert werden. Andererseits kann mit festgelegten α- und β-Fehler der erforderliche Nichtzentralitätsparameter ν erhalten werden. Dazu muss der erforderliche Stichprobenumfang bestimmt werden.

Zunächst wird vorausgesetzt, dass pro Subjekt gleich viele Beobachtungen gemacht werden, d.h. $n_i = n$. Weiterhin wird im Folgenden der Index i weggelassen. Um nun die benötigte Stichprobengröße zu bestimmen, wird angenommen, dass die erklärenden Variablen $\{(x_j, z_j), j = 1, \ldots, n\}$ diskret sind und eine Verteilung der Form

$$P(x_j = u_{jl}, z_j = v_{jl}; j = 1, \ldots, n) = \tau_l, \quad l = 1, \ldots, L \tag{65}$$

besitzen, wobei $\{(u_{jl}, v_{jl}; j = 1, \ldots, n), l = 1, \ldots, L\}$ die L möglichen verschiedenen Werte für $\{(x_j, z_j), j = 1, \ldots, n\}$ sind.

Bemerkung 45. Mit diesen zusätzlichen Voraussetzungen wird ξ aus Gleichung (62) weiter vereinfacht. Die Werte der einzelnen Summanden $P^* V^{-1}(\mu^1 - \mu^*)$ (Index i weggelassen) aus (62) hängen unter anderem von den Werten von x_j und z_j ab. Diese erklärenden Variablen besitzen nach Obigen eine diskrete Verteilung (siehe (65)). Damit sind auch die Summanden $P^* V^{-1}(\mu^1 - \mu^*)$ Zufallsvektoren. Zur Vereinfachung werden diese Summanden nun durch ihren Erwartungswert bzgl. der Gleichung (65) ersetzt. Somit wird Gleichung (62) zu

$$\xi = N \,\mathrm{E}(P^* V^{-1}(\mu^1 - \mu^*)) = N \sum_{l=1}^{L} \tau_l P_l^* V_l^{-1}(\mu_l^1 - \mu_l^*), \tag{66}$$

wobei P_l^*, V_l, μ_l^1 und μ_l^* wie in (62) definiert sind, aber mit $x_{ij} = u_{jl}$ und $z_{ij} = v_{jl}$.

Bemerkung 46. Mit den gleichen Argumenten wie oben vereinfacht sich die Kovarianz-Matrix Σ_1 zu

$$\Sigma_1 = N \,\mathrm{E}(P^* V^{-1} \,\mathrm{Cov}_{H_1}(Y) V^{-1} P^{*\top}) = N \sum_{l=1}^{L} \tau_l P_l^* V_l^{-1} \,\mathrm{Cov}_{H_1}(Y_l) V_l^{-1} P_l^{*\top}. \tag{67}$$

Bemerkung 47. Mit $\tilde{\xi} = \mathrm{E}(P^* V^{-1}(\mu^1 - \mu^*))$ und
$\tilde{\Sigma}_1 = \mathrm{E}(P^* V^{-1} \,\mathrm{Cov}_{H_1}(Y) V^{-1} P^{*\top})$ wird der Nichtzentralitätsparameter aus (62) zu

$$\xi = N \tilde{\xi}^\top \tilde{\Sigma}_1^{-1} \tilde{\xi}, \tag{68}$$

womit sich der benötigte Stichprobenumfang darstellen lässt als

$$N = \frac{\nu}{(\tilde{\xi}^\top \tilde{\Sigma}_1^{-1} \tilde{\xi})}. \tag{69}$$

Dabei wird ν bestimmt durch eine nichtzentrale χ^2-Verteilung und die vorgegebenen Werte des α- und des β-Fehlers.

Bemerkung 48. Im Folgenden wird Gleichung (63) vereinfacht. Zunächst wird $\mathrm{E}_{H_1}(S_\lambda(\psi_0, \lambda_0^*), \psi_1, \lambda_1)$ gebildet. Dadurch wird μ_i zu μ_i^* und $\mathrm{E}_{H_1}(Y_i)$ zu μ_i^1 und Gleichung (63) lässt sich schreiben als

$$\lim_{N\to\infty} N^{-1} \sum_{i=1}^{N} \left(\frac{\partial \mu_i^*}{\partial \lambda_0^*}\right)^\top V_i^{-1}(\mu_i^1 - \mu_i^*) = 0. \tag{70}$$

Mit den obigen Voraussetzungen wird die Summe wieder durch den Erwartungswert bzgl. (65) ersetzt und es ergibt sich $N \sum_{l=1}^{L} \tau_l \left(\frac{\partial \mu_l^*}{\partial \lambda_0^*}\right)^\top V_l^{-1}(\mu_l^1 - \mu_l^*)$, so dass nun Gleichung (63) zu

$$\sum_{l=1}^{L} \tau_l \left(\frac{\partial \mu_l^*}{\partial \lambda_0^*}\right)^\top V_l^{-1}(\mu_l^1 - \mu_l^*) = 0 \tag{71}$$

wird. λ_0^* ist die Lösung dieser Gleichung.

Damit wird nun die benötigte Stichprobengröße mit wiederholten Beobachtungen mit der Teststatistik T in folgenden Schritten ermittelt:

1. Erstelle ein Regressionsmodell für die marginale Erwartung und lege Parameterwerte für die Null- und die Alternativ-Hypothese fest.
2. Lege die Korrelationsstruktur, d.h. $R_i(\alpha)$ fest.
3. Bilde eine Verteilung für die diskreten erklärenden Variablen in der Form von Gleichung (65).
4. Berechne λ_0^* mit Gleichung (71).
5. Ermittle den Nichtzentralitätsparameter.
6. Berechne alle benötigten Größen für (69) und erhalte somit den benötigen Stichprobenumfang.

Nun zum Beweis von Satz 41:

Beweis von Satz 41. Das betrachtete Modell $\text{logit}(\mu_{ij}) = \beta_0 + \beta_1 x_{ij}$ gleicht mit $\lambda = \beta_0$ und $\psi = \beta_1$ dem Modell

$$\text{logit}(\mu_{ij}) = \lambda + x_{ij}\psi, \quad j = 1, \ldots, n$$

mit $x_{ij} = 0$ oder 1 und $z_{ij} \equiv 1$ für$j = 1, \ldots, n$, d.h.

$$p_A = \frac{\exp(\lambda)}{1 + \exp(\lambda)} \quad \text{und} \quad p_B = \frac{\exp(\lambda + \psi)}{1 + \exp(\lambda + \psi)}.$$

Die Nullhypothese H_0 lautet $H_0 : p_A = p_B \Leftrightarrow H_0 : \psi = \psi_0 = 0$ gegen $H_1 : \psi = \psi_1 \neq 0$. Jedes Cluster besitzt die gleiche Korrelationsstruktur und die Beobachtungen sind austauschbar, d.h. die Korrelationsmatrix R ist eine Matrix mit Einsen auf der Diagonalen und ρ sonst.
Die diskreten erklärenden Variablen seien folgendermaßen verteilt:

$$\begin{aligned} \text{P}(X_{ij} = 0, Z_{ij} = 1) &= \tau_1 = 1 - \tau_2, \\ \text{P}(X_{ij} = 1, Z_{ij} = 1) &= \tau_2. \end{aligned}$$

Demnach ist $L = 2$. Als nächstes muss λ_0^* berechnet werden. Dazu löse die Gleichung

$$\sum_{l=1}^{2} \tau_l \left(\frac{\partial \mu_l^*}{\partial \lambda_0^*}\right)^\top V_l^{-1}(\mu_l^1 - \mu_l^*) = 0 \tag{72}$$

unter H_1. Dazu ist zunächst

$$\mu_l^* = \frac{\exp(x_{ij}\psi_0 + z_{ij}\lambda_0^*)}{1 + \exp(x_{ij}\psi_0 + z_{ij}\lambda_0^*)}\mathbf{1_n} = \frac{\exp(x_{ij}\psi_0 + \lambda_0^*)}{1 + \exp(x_{ij}\psi_0 + \lambda_0^*)}\mathbf{1_n},$$

wobei $\mathbf{1_n}$ ein Vektor der Dimension $n \times 1$ ist, der nur aus Einsen besteht. Außerdem ist $\mu_l^* = \mu_1^*$ für $x_{ij} = 0$ und $\mu_l^* = \mu_2^*$ für $x_{ij} = 1$. Somit ist also

$$\begin{aligned} \left(\frac{\partial \mu_l^*}{\partial \lambda_0^*}\right) &= \frac{\exp(x_{ij}\psi_0 + \lambda_0^*)}{(1 + \exp(x_{ij}\psi_0 + \lambda_0^*))^2}\mathbf{1_n} \\ &\overset{\psi_0 = 0}{=} \frac{\exp(\lambda_0^*)}{(1 + \exp(\lambda_0^*))^2}\mathbf{1_n} \end{aligned}$$

für $l = 1, 2$. Außerdem ist $V_1 = V_2 = \dfrac{\exp(\lambda_0^*)}{(1+\exp(\lambda_0^*))^2} R$ und

$$\begin{aligned} \mu_l^1 &= \frac{\exp(x_{ij}\psi_1 + \lambda_1)}{1+\exp(x_{ij}\psi_1+\lambda_1)}\mathbf{1_n}, \\ \mu_l^* &= \frac{\exp(\lambda_0^*)}{1+\exp(\lambda_0^*)}\mathbf{1_n} \end{aligned}$$

für $l = 1, 2$. Weiterhin lässt sich Gleichung (72) unter H_1 vereinfachen:

$$\begin{aligned} & \sum_{l=1}^{2} \tau_l \left(\frac{\partial \mu_l^*}{\partial \lambda_0^*}\right)^\top V_l^{-1}(\mu_l^1 - \mu_l^*) && = 0 \\ \Leftrightarrow \quad & \sum_{l=1}^{2} \tau_l \frac{\exp(\lambda_0^*)}{(1+\exp(\lambda_0^*))^2} \mathbf{1_n}^\top \left(\frac{\exp(\lambda_0^*)}{(1+\exp(\lambda_0^*))^2}\right)^{-1} R^{-1} \\ & \quad \times \left(\frac{\exp(x_{ij}\psi_1+\lambda_1)}{1+\exp(x_{ij}\psi_1+\lambda_1)}\mathbf{1_n} - \frac{\exp(\lambda_0^*)}{1+\exp(\lambda_0^*)}\mathbf{1_n}\right) && = 0 \\ \Leftrightarrow \quad & \sum_{l=1}^{2} \tau_l \left(\frac{\exp(x_{ij}\psi_1+\lambda_1)}{1+\exp(x_{ij}\psi_1+\lambda_1)} - \frac{\exp(\lambda_0^*)}{1+\exp(\lambda_0^*)}\right) \mathbf{1_n}^\top R^{-1}\mathbf{1_n} && = 0 \\ \overset{(*)}{\Leftrightarrow} \quad & \sum_{l=1}^{2} \tau_l \left(\frac{\exp(x_{ij}\psi_1+\lambda_1)}{1+\exp(x_{ij}\psi_1+\lambda_1)} - \frac{\exp(\lambda_0^*)}{1+\exp(\lambda_0^*)}\right) && = 0 \\ \Leftrightarrow \quad & \tau_1 \frac{\exp(\lambda_1)}{1+\exp(\lambda_1)} + \tau_2 \frac{\exp(\psi_1+\lambda_1)}{1+\exp(\psi_1+\lambda_1)} - (\tau_1+\tau_2)\frac{\exp(\lambda_0^*)}{1+\exp(\lambda_0^*)} && = 0 \\ \Leftrightarrow \quad & \tau_1 \frac{\exp(\lambda_1)}{1+\exp(\lambda_1)} + \tau_2 \frac{\exp(\psi_1+\lambda_1)}{1+\exp(\psi_1+\lambda_1)} && = \frac{\exp(\lambda_0^*)}{1+\exp(\lambda_0^*)} \\ \Leftrightarrow \quad & \tau_1 p_A + \tau_2 p_B && = \frac{\exp(\lambda_0^*)}{1+\exp(\lambda_0^*)}. \end{aligned}$$

Die Äquivalenz $(*)$ gilt, da $\mathbf{1_n}^\top R^{-1}\mathbf{1_n} \neq 0$ ist. Insgesamt ergibt sich somit $\lambda_0^* = \text{logit}(\tau_1 p_A + \tau_2 p_B)$.

Als nächstes wird der Nichtzentralitätsparameter ermittelt. Unter H_0 konvergiert T gegen eine χ^2-Verteilung mit einem Freiheitsgrad, d.h. $T \approx Z^2$ mit $Z \sim \mathcal{N}(0,1)$. Damit ist $\sqrt{T}$ asymptotisch standardnormalverteilt.

Unter H_1 ist T asymptotisch nichtzentral χ^2-verteilt. Es sei hier $T \approx Y^2$ mit $Y \sim \mathcal{N}(\mu, 1)$. Dann ist der Erwartungswert $\mathrm{E}(T) \approx 1 + \mu^2$, da T einen Freiheitsgrad besitzt. μ^2 ist der Nichtzentralitätsparameter, der bestimmt werden muss.

$\sqrt{T} - \mu$ ist dann asymptotisch standardnormalverteilt und es gilt:

$$\begin{aligned} \mathrm{P}_d(\sqrt{T}) > z_{1-\alpha/2}) &= \mathrm{P}_d(\sqrt{T} - \mu > z_{1-\alpha/2} - \mu) \\ &\approx 1 - \Phi(z_{1-\alpha/2} - \mu). \end{aligned}$$

Für einen Test mit dem Niveau α und der Power $1 - \beta$ an der Stelle d gilt nun:

$$\begin{aligned} & 1 - \beta && = 1 - \Phi(z_{1-\alpha/2} - \mu) \\ \Leftrightarrow \quad & 1 - \beta && = \Phi(-z_{1-\alpha/2} + \mu) \\ \Leftrightarrow \quad & \Phi(z_{1-\beta}) && = \Phi(-z_{1-\alpha/2} + \mu) \\ \Leftrightarrow \quad & z_{1-\beta} && = -z_{1-\alpha/2} + \mu \\ \Leftrightarrow \quad & z_{1-\beta} + z_{1-\alpha/2} && = \mu \\ \Rightarrow \quad & (z_{1-\beta} + z_{1-\alpha/2})^2 && = \mu^2 = \nu. \end{aligned}$$

Nun sind noch $\tilde{\xi}$ und Σ_1 für Gleichung (69) zu bestimmen. Es ist

$$P_l^* = \left(\left(\frac{\partial \mu_l}{\partial \psi} \right)^\top - I_{\psi\lambda}^* I_{\lambda\lambda}^{*-1} \left(\frac{\partial \mu_l}{\partial \lambda} \right)^\top \right) (\psi_0, \lambda_0^*).$$

Es ergibt sich an den Stellen ψ_0 und λ_0^* für

$$\frac{\partial \mu_1}{\partial \psi} = \mathbf{0_n} \text{ und für } \frac{\partial \mu_2}{\partial \psi} = \frac{\exp(\lambda_0^*)}{(1 + \exp(\lambda_0^*))^2} \mathbf{1_n}.$$

Außerdem ist an den Stellen ψ_0 und λ_0^*

$$\frac{\partial \mu_1}{\partial \lambda} = \frac{\partial \mu_2}{\partial \lambda} = \frac{\exp(\lambda_0^*)}{(1 + \exp(\lambda_0^*))^2} \mathbf{1_n}.$$

Weiterhin ist

$$\begin{aligned} I_{\psi\lambda}^* &= \sum_{l=1}^{2} \tau_l \left(\frac{\partial \mu_l}{\partial \psi} \right)^\top V_l^{-1} \left(\frac{\partial \mu_l}{\partial \lambda} \right) \\ &= \tau_2 \frac{\exp(\lambda_0^*)}{(1 + \exp(\lambda_0^*))^2} \mathbf{1_n}^\top R^{-1} \mathbf{1_n}. \end{aligned}$$

Mit $p^* := \tau_1 p_A + \tau_2 p_B = \dfrac{\exp(\lambda_0^*)}{1 + \exp(\lambda_0^*)}$ wird dies zu

$$I_{\psi\lambda}^* = \tau_2 p^* (1 - p^*) \mathbf{1_n}^\top R^{-1} \mathbf{1_n}.$$

Ebenso ist

$$\begin{aligned} I^*_{\lambda\lambda} &= \sum_{l=1}^{2} \tau_l \left(\frac{\partial \mu_l}{\partial \lambda}\right)^\top V_l^{-1} \left(\frac{\partial \mu_l}{\partial \lambda}\right) \\ &= (\tau_1+\tau_2)\frac{\exp(\lambda_0^*)}{(1+\exp(\lambda_0^*))^2}\mathbf{1_n}^\top R^{-1}\mathbf{1_n} \\ &= p^*(1-p^*)\mathbf{1_n}^\top R^{-1}\mathbf{1_n}. \end{aligned}$$

Damit ist nun

$$\begin{aligned} P_1^* &= -\tau_2 p^*(1-p^*)\mathbf{1_n}^\top, \\ P_2^* &= (1-\tau_2)p^*(1-p^*)\mathbf{1_n}^\top. \end{aligned}$$

Also ist unter H_1

$$\begin{aligned} \tilde{\xi} &= \sum_{l=1}^{2} \tau_l P_l^* V_l^{-1}(\mu_l^1 - \mu_l^*) \\ &= \tau_1\tau_2 \mathbf{1_n}^\top R^{-1}\mathbf{1_n}\left(-\frac{\exp(\lambda_1)}{1+\exp(\lambda_1)} + \frac{\exp(\lambda_1+\psi_1)}{1+\exp(\lambda_1+\psi_1)}\right) \\ &= \tau_1\tau_2 \mathbf{1_n}^\top R^{-1}\mathbf{1_n}(p_B - p_A). \end{aligned}$$

Nun muss noch $\tilde{\Sigma}_1$ berechnet werden. Es ist

$$\tilde{\Sigma}_1 = \sum_{l=1}^{2} \tau_l P_l^* V_l^{-1} \operatorname{Cov}_{H_1}(Y_l) V_l^{-1} P_l^{*\top}.$$

Mit

$$\operatorname{Cov}_{H_1}(Y_0) = \frac{\exp(\lambda_1)}{(1+\exp(\lambda_1))^2} R$$

und

$$\operatorname{Cov}_{H_1}(Y_1) = \frac{\exp(\lambda_1+\psi_1)}{(1+\exp(\lambda_1+\psi_1))^2} R$$

und $V_l = p^*(1-p^*)R$ wird dies unter H_1 zu

$$\begin{aligned} \tilde{\Sigma}_1 &= \tau_1\tau_2\left(\tau_2\frac{\exp(\lambda_1)}{(1+\exp(\lambda_1))^2}\right)\mathbf{1_n}^\top R^{-1}\mathbf{1_n} \\ &\quad + \tau_1\tau_2\left(\tau_1\frac{\exp(\lambda_1+\psi_1)}{(1+\exp(\lambda_1+\psi_1))^2}\right)\mathbf{1_n}^\top R^{-1}\mathbf{1_n} \\ &= \tau_1\tau_2(\tau_2 p_A(1-p_A) + \tau_1 p_B(1-p_B))\mathbf{1_n}^\top R^{-1}\mathbf{1_n} \end{aligned}$$

und somit ergibt sich für den benötigten Stichprobenumfang:

$$\begin{aligned} N &= \frac{\nu}{\tilde{\xi}^\top \tilde{\Sigma_1}^{-1} \tilde{\xi}} \\ &= \frac{(z_{1-\alpha/2} + z_{1-\beta})^2 (\tau_2 p_A(1-p_A) + \tau_1 p_B(1-p_B))}{\tau_1 \tau_2 (p_B - p_A)^2 \mathbf{1_n}^\top R^{-1} \mathbf{1_n}}. \end{aligned}$$

Da es sich um austauschbare Beobachtungen handelt, ist

$$R = \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \rho \\ \rho & \cdots & \rho & 1 \end{pmatrix} \in \mathbb{R}^{n \times n}$$

und damit ist

$$R^{-1} = \frac{1}{(\rho-1)(1+(n-1)\rho)} \begin{pmatrix} -(n-2)\rho - 1 & \rho & \cdots & \rho \\ \rho & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \rho \\ \rho & \cdots & \rho & -(n-2)\rho - 1 \end{pmatrix}.$$

Es folgt $\mathbf{1_n}^\top R^{-1} \mathbf{1_n} = \dfrac{n}{1+(n-1)\rho}$ und mit $\tau_1 = \tau_2 = 1/2$ wird der benötigte Stichprobenumfang zu

$$N = \frac{2(z_{1-\alpha/2} + z_{1-\beta})^2 (p_A(1-p_A) + p_B(1-p_B))(1+(n-1)\rho)}{n(p_B - p_A)^2}.$$

□

Bemerkung 49. Für $n = 1$ ergibt sich aus (57) gerade die Formel (56), die für eine Beobachtung pro Person hergeleitet wurde.

5.2 Stichprobengröße versus Wiederholungen

Mit dem Verfahren von Liu u. Liang (1997) lässt sich nun für ein gegebenes Modell für eine feste Anzahl n von Wiederholungen pro Person der benötigte Stichprobenumfang (benötigte Personenzahl) berechnen, damit ein Test zum Niveau α

und einer Power $1 - \beta$ bei einer kleinsten bedeutungsvollen Differenz d erhalten wird. Das gleiche Ergebnis lässt sich allerdings auch erreichen, wenn die Anzahl n verändert und der Stichprobenumfang N entsprechend angepasst wird. Da pro Person in einer Studie Fixkosten (Rekrutierungskosten) sowie pro Untersuchung pro Person variable Kosten entstehen, sollte die Anzahl der Beobachtungen pro Person und die Anzahl der beobachteten Personen (Stichprobengröße) so gegeneinander abgewogen werden, dass die gesamten Kosten minimal sind.

Bemerkung 50. Seien in einer Studie pro Person Rekrutierungskosten r und pro Beobachtung pro Person Untersuchungskosten u gegeben. Sei die Anzahl der beobachteten Subjekte mit N bezeichnet und die Anzahl der Wiederholungen mit n. Dann kann als einfacher Ansatz für die gesamten Kosten K der Studie der Ansatz

$$K(N, r, u, n) = N * r + N * u * n \tag{73}$$

gewählt werden.

Bemerkung 51. Die Anzahl N der zu beobachtenden Subjekte wird nun in Abhängigkeit von n mit Hilfe des Verfahrens von Liu u. Liang (1997) berechnet. Für verschiedene Anzahlen n von Wiederholungen können so die Kosten verglichen werden. Daraus lässt sich diejenige Anzahl von Wiederholungen bestimmen, bei der die geringsten Kosten zu erwarten sind. Mit diesem n kann dann die benötigte Personenzahl bestimmt werden. Im Folgenden wird nun für ein Beispiel die benötigte Anzahl von Wiederholungen und Personen ermittelt, mit denen die Kosten minimal sind.

5.2.1 Beispiel

Als Beispiel wird eine Studie betrachtet, in der zwei Behandlungen miteinander verglichen werden sollen. Dabei sei Behandlung A die Kontrolle und Behandlung B ein Medikament, das auf Wirksamkeit untersucht werden soll. Dazu wird das oben schon erwähnte Modell $\text{logit}(\mu_{ij}) = \beta_0 + \beta_1 x_{ij}$ mit $x_{ij} = 0$ für Behandlung A

und $x_{ij} = 1$ für Behandlung B benutzt. Zwei gleich große Gruppen erhalten je eine Behandlung. Die Studie soll kostengünstig durchgeführt werden, so dass vorab die günstigste Kombination von wiederholten Beobachtungen n und Personen N bestimmt werden soll. Für verschiedene Werte von n werden die Kosten mit Gleichung (73) ermittelt, wobei die jeweils benötigte Anzahl N mit Gleichung (57) berechnet wird. Dazu werden allerdings noch einige Informationen benötigt. Zum einen die Erfolgswahrscheinlichkeit p_A bei Behandlung A (Kontrolle). Diese sollte durch Erfahrung bekannt sein oder muss geschätzt werden. Zum anderen muss eine Vorstellung über den zu Grunde liegenden Odds Ratio[54], der mindestens erkannt werden soll, vorliegen. Durch diesen kann mit p_A dann eine Wahrscheinlichkeit p_B^d errechnet werden, die wie p_A in die Fallzahlformel eingeht. Außerdem sollte eine Schätzung über die in den wiederholten Daten vorhandene Korrelation ρ vorliegen. Mit den gewünschten α- und β-Werten werden die benötigten Quantile berechnet. Für $\alpha = 0.05$ und $1 - \beta = 0.9$ ergeben sich $z_{1-\alpha/2} = 1.960$ und $z_{1-\beta} = 1.282$. Nun kann in Abhängigkeit von der Anzahl n der Wiederholungen die benötigte Subjektanzahl berechnet werden.

Für dieses Beispiel sei zunächst $p_A = 0.5$, der Odds Ratio $OR = 0.5$, die Korrelation $\rho = 0.3$, die Rekrutierungskosten betragen 1000 und die Untersuchungskosten 100 Einheiten. Die nötigen Berechnungen werden mit der Programmiersprache ***R*** durchgeführt. Die entsprechenden Funktionen sind im Anhang zu finden. Zunächst wird eine Funktion *anzahlpA* benötigt, die für übergebene Werte von p_A, OR, ρ und n mit Gleichung (57) die benötigte Personenanzahl ausrechnet. Die Funktion *zeichnenpA* berechnet in einer Schleife von 1 bis n die benötigte Personenzahl und gibt diese in einer Tabelle und einer Graphik aus. Für $n = 10$ liefert diese Funktion eine Tabelle, in der in der ersten Spalte die Anzahl der Wiederholungen n und in der zweiten Spalte die jeweilige benötigte Anzahl von Personen N steht:

[54] $OR = OR^d = \frac{\frac{p_B^d}{1-p_B^d}}{\frac{p_A}{1-p_A}}$

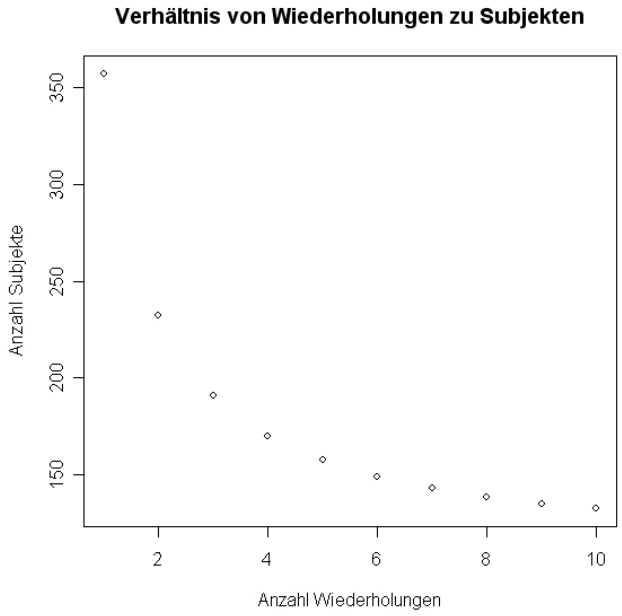

Abbildung 4: Ausgabe der Funktion *zeichnenpA*

```
> zeichnenpA(0.5,0.5,0.3,10)
      Wiederholungen Subjekte
 [1,]              1 357.3592
 [2,]              2 232.2835
 [3,]              3 190.5916
 [4,]              4 169.7456
 [5,]              5 157.2380
 [6,]              6 148.8997
 [7,]              7 142.9437
 [8,]              8 138.4767
 [9,]              9 135.0024
[10,]             10 132.2229
```

Diese Funktion liefert offensichtlich keine ganzen Zahlen und auch ein Aufrunden liefert nicht immer gerade Zahlen. Allerdings sollen zwei gleich große Gruppen

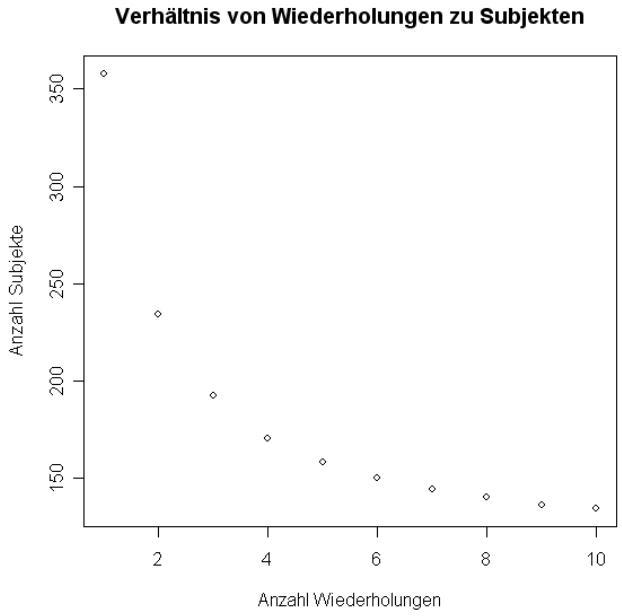

Abbildung 5: Ausgabe der Funktion *zeichnenpAcl*

betrachtet werden. Also muss die Fallzahl, die mit Gleichung (57) erhalten wird, noch zur nächsten ganzen geraden Zahl aufgerundet werden. Dies ist in der Funktion *zeichnenpAcl* gleich implementiert:

```
> zeichnenpAcl(0.5,0.5,0.3,10)
      Wiederholungen Subjekte
 [1,]              1      358
 [2,]              2      234
 [3,]              3      192
 [4,]              4      170
 [5,]              5      158
 [6,]              6      150
 [7,]              7      144
 [8,]              8      140
 [9,]              9      136
```

```
[10,]             10       134
```

Bemerkung 52. Durch dieses Anpassen der Fallzahlen kann es dazu kommen, dass für unterschiedliche Anzahlen von Wiederholungen n eine gleiche Anzahl von benötigten Personen errechnet wird.

Mit Hilfe einer Funktion *kostenfunktionpAcl* können nun die Kosten für Studien mit unterschiedlich vielen Wiederholungen pro Subjekt berechnet werden. Dazu wird Formel (73) benutzt. Die Funktion *kostenfunktionpAcl* gibt in diesem Fall eine Tabelle sowie eine Graphik mit den Kosten für 1 bis 10 Wiederholungen aus:

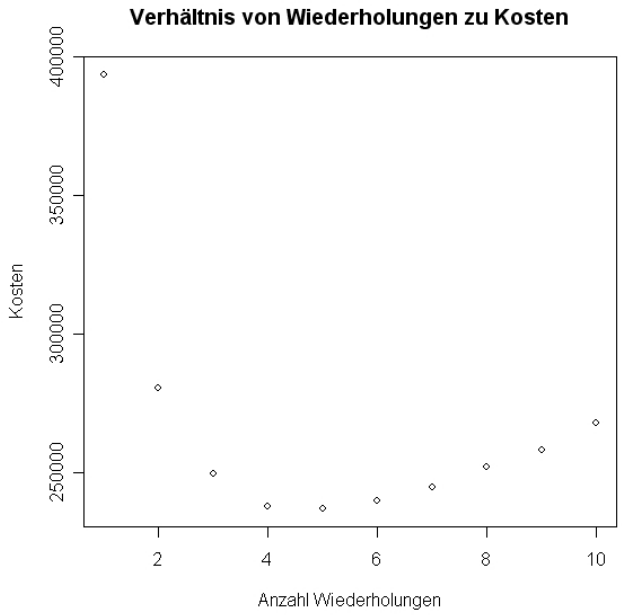

Abbildung 6: Ausgabe der Funktion *kostenfunktionpAcl*

```
> kostenfunktionpAcl(10,0.5,0.5,0.3,1000,100)
      Wiederholungen Kosten
 [1,]              1 393800
 [2,]              2 280800
```

```
 [3,]             3 249600
 [4,]             4 238000
 [5,]             5 237000
 [6,]             6 240000
 [7,]             7 244800
 [8,]             8 252000
 [9,]             9 258400
[10,]            10 268000
```

In diesem Beispiel sind die Kosten bei 5 Wiederholungen minimal, so dass die Studie mit 158 Subjekten mit je 5 Beobachtungen durchgeführt werden sollte.

Bemerkung 53. Bei einigen Variablenkombinationen kann es zu „Sprüngen“ in der Kostenfunktion kommen, d.h., dass die Kosten bei steigender Wiederholungsanzahl erst sinken, dann steigen, wieder sinken und schließlich wieder ansteigen (siehe Abb.7). Zum Beispiel ergibt sich mit einem OR von 0.02 und sonst den gleichen Variablen wie oben folgende Kostensituation:

```
> kostenfunktionpAcl(10,0.5,0.02,0.3,1000,100)
      Wiederholungen Kosten
 [1,]              1  28600
 [2,]              2  19200
 [3,]              3  18200
 [4,]              4  16800
 [5,]              5  18000
 [6,]              6  19200
 [7,]              7  17000
 [8,]              8  18000
 [9,]              9  19000
[10,]             10  20000
```

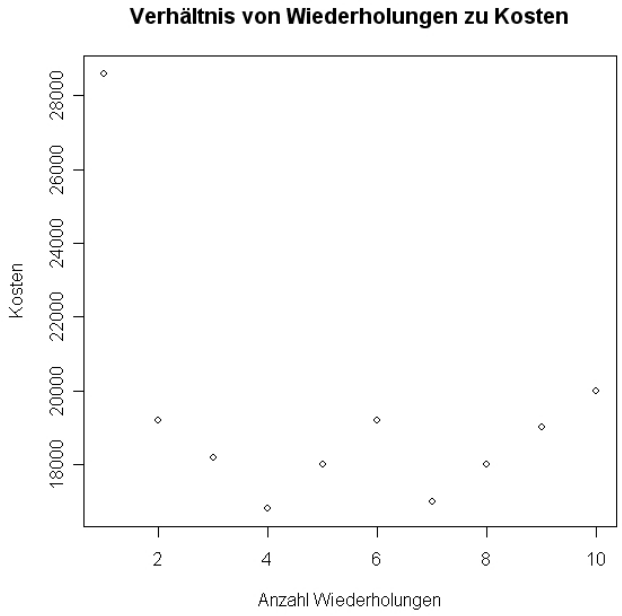

Abbildung 7: „Sprungstellen“ in der Kostenfunktion

Diese Situation lässt sich mit Bemerkung 52 erklären. Bei unterschiedlich vielen Wiederholungen ergibt die Funktion *zeichnenpAcl* gleich viele Subjekte:

```
> zeichnenpAcl(0.5,0.02,0.3,10)
      Wiederholungen Subjekte
 [1,]              1       26
 [2,]              2       16
 [3,]              3       14
 [4,]              4       12
 [5,]              5       12
 [6,]              6       12
 [7,]              7       10
 [8,]              8       10
 [9,]              9       10
[10,]             10       10
```

Dies wirkt sich auf die Funktion (73) aus: wenn die Anzahl der Wiederholungen n steigt und die Anzahl der Subjekte gleich bleibt, steigen die Kosten.

6 Simulationen

Zur Überprüfung der Formel für die benötigte Stichprobengröße aus Kapitel 5 werden wiederholte binäre Daten erzeugt. Diese Daten weisen einen Behandlungsunterschied auf, der durch den Odds Ratio angegeben wird. Für verschiedene Werte für die Erfolgswahrscheinlichkeit p_A der Kontrolle werden die jeweiligen Erfolgswahrscheinlichkeiten p_B der Behandlung mit diesem Odds Ratio berechnet. Die Daten werden dann mit Hilfe des Bahadur-Modells erzeugt und anschließend mit dem Statistik-Programm SAS ausgewertet. Dabei wird die in SAS enthaltene Prozedur `genmod` benutzt. In 200 Simulationen werden Schätzer für den Behandlungseffekt erzeugt und anschließend mit der SAS-Prozedur `univariate` statistisch ausgewertet. Weiterhin wird die Signifikanz des Behandlungsunterschieds betrachtet, um eine Aussage über die zu Grunde liegende Power bei diesem Behandlungsunterschied zu erhalten.

6.1 Erzeugung wiederholter binärer Daten

Es werden wiederholte binäre Daten für $n = n_i = 1, 2$ und 3 erzeugt. Dies geschieht mit Hilfe des Bahadur-Modells. Mit Formel (10) können mit vorgegebener Wahrscheinlichkeit p und Korrelation ρ die Wahrscheinlichkeiten für $0, 1, \ldots, n$ beobachtete Einsen berechnet werden. Die weitere Erzeugung der Daten wird hier beispielhaft für den Fall $n = 2$ erläutert und lässt sich leicht übertragen.
Die Erzeugung der Daten erfolgt mit SAS. Die jeweiligen Programme sind im Anhang zu finden. Zunächst werden mit Gleichung (10) die Wahrscheinlichkeiten für keine und eine beobachtete Eins bestimmt. Sei $f_0 = \mathrm{P}(Y = 0)$ und $f_1 = \mathrm{P}(Y = 1)$. Um nun damit wiederholte binäre Daten zu erhalten, wird mit einer Gleichverteilung auf dem Intervall [0,1] eine Zufallszahl x erzeugt. Wenn $x < f_0$ ist, wird die Variable *anz* gleich 0 gesetzt. Sie gibt die Anzahl der Einsen an, die der binäre Beobachtungsvektor später besitzen soll. Wenn $f_0 < x < f_0 + f_1$ ist, wird diese Variable *anz* gleich 1 gesetzt, sonst gleich 2. Dieses Vorgehen ist

in dem SAS-Makro *%beob1* implementiert.
Mit Hilfe dieses Makros wird anschließend mit dem SAS-Makro *%tabelle* eine Datei erzeugt, in der für eine übergebene Anzahl von Subjekten wiederholte Beobachtungen erzeugt werden. Dabei werden die Subjekte gleichmäßig auf zwei Gruppen aufgeteilt. Für die erste Gruppe wird für jedes Subjekt mit der übergebenen Wahrscheinlichkeit und Korrelation die Variable *anz* berechnet. Für die zweite Gruppe wird zunächst mit dem übergebenen Odds Ratio die Erfolgswahrscheinlichkeit berechnet und dann für jedes Subjekt wieder die Variable *anz* bestimmt. Aus diesen Werten wird nun eine Tabelle erstellt, die die Subjektnummer, die Nummer der wiederholten Beobachtungen (hier 1, 2 und 3) und die binären Beobachtungen selbst enthält.

6.2 Auswertung der Daten

Die Daten können nun mit der Prozedur `genmod` ausgewertet werden. Diese erlaubt die Auswertung von korrelierten Daten mit Hilfe der GEEs. Die Eingabe in SAS ist folgende:

```
proc genmod data = tabelle desc;
  class subject treat time;
  model y = treat / dist = bin;
  repeated subject = subject / type exch corrb corrw modelse;
run;
```

Die Prozedur `genmod` wertet die Datei *tabelle* aus. *tabelle* ist die Datei, die mit dem Makro *%tabelle* erzeugt wurde und die Daten enthält. Mit der Angabe `desc` modelliert **proc genmod** die Wahrscheinlichkeit für $Y_{ij} = 1$. Die erklärenden Variablen werden in dem *class*-Statement angegeben, sowie alle Variablen, die nicht stetig sind. *subject* beinhaltet hier die Subjektnummer, *time* die Nummer der Beobachtung und *treat* die Behandlungsart. Mit dem *model*-Statement wird das marginale Modell festgelegt. In diesem Fall hängt die beobachtete binäre Variable

y von der Behandlung *treat* ab. Durch die Angabe *dist=bin* wird eine binomiale Verteilung festgelegt und automatisch ein Logit-Modell benutzt. Durch das *repeated*-Statement wird angegeben, dass es sich um wiederholte Daten handelt, so dass die Korrelation durch Anwenden der GEE-Theorie berücksichtigt wird. Die Daten eines Subjekts sind korreliert. Die Korrelationsstruktur wird durch *type exch* angegeben. Die hier benutzte Korrelationsstruktur ist *exchangeable* (Es sind auch andere Korrelationsstrukturen möglich, wie z.B. *unstructured*, *identity* usw., vgl. 3.1.3.1.1). Die *working correlation matrix* wird durch *corrw* ausgegeben. Weiterhin gibt *corrb* die Korrelationsmatrix der geschätzten Parameter aus. Mit *genmod* können nun für das Modell $\text{logit}(p_{ij}) = \beta_0 + \beta_1 x_{ij},\ i = K, B,$ $x_K j = 0, x_B j = 1$ die Parameter β_0 und β_1 berechnet werden.[55] Dazu benutzt *genmod* die Theorie der GEEs und das Schätzen der Parameter läuft wie in Abschnitt 3.1.3.1.2 beschrieben ab. Für simulierte Daten mit einem Odds Ratio von 0.5 ergibt sich als Ausgabe von `genmod` dann beispielsweise:

GEE-Parameterschätzeranalyse

Empirische Standardfehlerschätzer

Parameter		Schätzwert	Standardfehler	95% Konfidenzgrenzen		Z
Intercept		1.2809	0.1294	1.0273	1.5346	9.90
treat	B	-0.4371	0.1799	-0.7896	-0.0845	-2.43
treat	K	0.0000	0.0000	0.0000	0.0000	.

[55] Die Behandlung A erhält im Folgenden die Bezeichnung K für Kontrolle.

GEE-Parameterschätzeranalyse

Empirische Standardfehlerschätzer

Pr > |Z|
<.0001
0.0151
.

Der Parameter, der in der Ausgabe bei *Intercept* steht, ist gerade β_0; der Parameter, der bei *treat B* steht, ist der Parameter β_1 aus dem oben benutzten Modell. Dieser Parameter β_1 entspricht, wie schon erwähnt, dem Logarithmus des Odds Ratio. Also ist $\exp(\beta_1) = OR$. Damit ergibt sich in diesem Fall für eine Schätzung des Odds Ratio $\exp(-0.4371) = 0.6459068$.
Bei der Erzeugung der Daten wurden so viele Subjekte benutzt, um den nach Kapitel 5 zu Grunde liegenden Unterschied in den Behandlungen signifikant mit einer Power von 0.9 zu erkennen. Die Signifikanz der Parameter wird mit der Z-Statistik geprüft. Die Hypothese lautet also jeweils $H_0 : \beta_i = 0$ für $i = 1, 2$. Für den Parameter β_1 beispielsweise errechnet sich die Prüfgröße durch $\frac{-0.4371-0}{0.1799} = -2.43$. Der p-Wert wird dann mit Hilfe der Standardnormalverteilung errechnet, indem die Wahrscheinlichkeiten für Werte kleiner gleich -2.34 und größer gleich 2.34 addiert werden. Hier ergibt sich für den Parameter β_1, der den Behandlungsunterschied anzeigt, ein p-Wert von 0.0151 (< 0.05) und dieser ist damit signifikant.[56]
Um zu überprüfen, wie gut `genmod` die wahren, in den Daten enthaltenen Parameter erkennt und welche Power dabei erreicht wird, werden Simulationen durchgeführt. Mit dem Makro *%wiederholen* werden für eine feste Erfolgswahrscheinlichkeit p_A, einen festen Odds Ratio und einer festen Korrelation je 200

[56]Da die Parameter β nach Liang u. Zeger (1986) asymptotisch normalverteilt sind, ist die Z-Statistik, die auf einer Normalverteilung beruht, geeignet um die p-Werte zu bestimmen.

Schätzwert	95% Untere Konfidenz-intervallgrenze LCL	95% Obere Konfidenz-intervallgrenze UCL	Pr> |Z|
-0.4865	-0.8437	-0.1292	0.0076
-0.3523	-0.6930	-0.0116	0.0427

Tabelle 2: Beispiel für Datei *schaetzer*

Mittelwert	Varianz	Stand.abweichung	LCL	UCL	Weite
-0.4194	0.0090	0.0949	-0.7684	-0.0704	0.6980

Tabelle 3: Beispiel für Datei *parameterbeta1*

verschiedene Datensätze erzeugt und anschließend mit `genmod` ausgewertet. Die Ergebnisse werden dabei in eine Datei *schaetzer* geschrieben, die den errechneten Wert für β_1, die 95%-tige untere und obere Konfidenzgrenzen und den p-Wert bzgl. der Z-Statistik enthält. Für nur zwei Simulationen ergibt sich beispielsweise die Datei *schaetzer* in Tabelle 2. Die Datei *schaetzer* wird anschließend mit dem Makro *%auswertung* ausgewertet, d.h. es wird eine Datei *parameterbeta1* erzeugt, die den Mittelwert, die Varianz, die Standardabweichung, den Mittelwert der unteren und oberen Konfidenzintervallgrenze (LCL und UCL) und die Weite dieses Intervalls enthält. Für das obige Beispiel ergeben sich somit die Werte in Tabelle 3. Desweiteren wird noch die Power bestimmt, indem die Anzahl der p-Werte < 0.05 aus der Datei *schaetzer* bestimmt und durch 200 geteilt wird.
Alle SAS-Makros sind im Anhang einschließlich der Dokumentation zu finden.

6.3 Ergebnisse der Simulationen

Zur Überprüfung der Formel (57) für die benötigte Stichprobengröße für das Logit-Modell wurden Simulationen für $n = 1, 2, 3$ durchgeführt. Dabei wurde stets ein Odds Ratio von 0.5 angenommen. Für die Erfolgswahrscheinlichkeit p_A der Kontrolle wurden die Werte 0.2, 0.4., 0.6 und 0.8 sowie für die Korrelation

die Werte 0.3 und 0.5 benutzt. Die benötigte Stichprobengröße wurde für die jeweiligen Kombinationen mit Gleichung (57) berechnet. Anschließend wurden für jede der möglichen Kombinationen 200 Datensätze nach obigem Schema erzeugt und wie beschrieben ausgewertet.

Die Ergebnisse sind in Tabelle 4 zusammengestellt. Angegeben sind jeweils die für die Anzahl der Wiederholungen (n), der Korrelation (ρ) und der Erfolgswahrscheinlichkeit p_A benötigten Subjekte (N) sowie die jeweils erhaltenen *Parameterbeta1*-Dateien und die dazugehörige Power.

Für $n = 1$ und $n = 2$ liefert die Formel eine sehr gute Power, die nahe an der gewünschten von 0.9 liegt. Die Schwankungen sind durch die geringe Anzahl von Simulationen (200) erklärbar. Für $n = 3$ sinkt die Power allerdings, besonders bei einer Korrelation von 0.5 und steigenden Werten für p_A. Dies lässt die Vermutung zu, dass diese Fallzahlformel besser für Daten mit wenigen wiederholten Beobachtungen pro Subjekt und geringer Korrelation geeignet ist. Um genauere Aussagen treffen zu können, müssten allerdings noch mehrere Kombinationen von Wiederholungen, Korrelationen und Wahrscheinlichkeiten betrachtet werden. Ingesamt sollten für ein klares Ergebnis auch mehr als 200 Simulationen durchgeführt werden, mindestens 1000, eher noch mehr.

n	ρ	p_A	N	Mittelw.	Var.	Std.	LCL	UCL	Weite	Power
1		0.2	690	-0.7064	0.0530	0.2302	-1.1363	-0.2766	0.8597	0.88
		0.4	400	-0.7038	0.0452	0.2126	-1.1336	-0.2740	0.8596	0.915
		0.6	348	-0.6909	0.0540	0.2324	-1.1190	-0.2628	0.8563	0.88
		0.8	452	-0.6692	0.0482	0.2196	-1.0978	-0.2406	0.8571	0.86
2	0.3	0.2	448	-0.7103	0.0567	0.2382	-1.1420	-0.2786	0.8634	0.90
		0.4	260	-0.7092	0.0548	0.2340	-1.1375	-0.2809	0.8566	0.885
		0.6	226	-0.6770	0.0463	0.2152	-1.1056	-0.2484	0.8572	0.875
		0.8	294	-0.6902	0.0404	0.2010	-1.1192	-0.2612	0.8580	0.90
	0.5	0.2	518	-0.6960	0.0477	0.2185	-1.1258	-0.2663	0.8595	0.885
		0.4	300	-0.6942	0.0410	0.2024	-1.1220	-0.2664	0.8556	0.905
		0.6	262	-0.7060	0.0434	0.2082	-1.1324	-0.2796	0.8528	0.935
		0.8	340	-0.6935	0.0389	0.1972	-1.1235	-0.2636	0.8599	0.915
3	0.3	0.2	368	-0.6377	0.0528	0.2298	-1.0729	-0.2026	0.8703	0.85
		0.4	214	-0.6077	0.0513	0.2266	-1.0463	-0.1691	0.8772	0.78
		0.6	186	-0.6037	0.0477	0.2183	-1.0439	-0.1636	0.8803	0.75
		0.8	242	-0.5912	0.0498	0.2230	-1.0304	-0.1520	0.8784	0.825
	0.5	0.2	460	-0.6163	0.0482	0.2195	-1.0528	-0.1797	0.8731	0.805
		0.4	268	-0.5400	0.0496	0.2227	-0.9788	-0.1013	0.8775	0.68
		0.6	232	-0.5055	0.0438	0.2093	-0.9510	-0.0599	0.8911	0.62
		0.8	302	-0.5583	0.0576	0.2399	-1.0039	-0.1128	0.8911	0.66

Tabelle 4: Auswertung der Simulationen für den Parameter β_1.

7 Schlussbemerkungen

In dieser Arbeit wird deutlich, dass es kein festes Modell zur Auswertung von wiederholten binären Daten gibt, sondern viele verschiedene Ansätze. Die schwierige Entscheidung, welche Modellklasse (marginal oder subjektspezifisch) benutzt werden sollte, muss der Anwender auf Grund seiner zu untersuchenden Fragestellung und Erfahrung selbst treffen. Marginale Modelle besitzen den Vorteil, dass sie relativ einfach auszuwerten sind. Allerdings berücksichtigen sie nicht die Variabilität, die in den untersuchten Subjekten vorhanden ist. Dafür sind subjektspezifische Modelle besser geeignet, bei denen allerdings die Schätzungen durch die nötigen Integrationen erschwert werden.

Ist die Wahl auf eine Modellklasse gefallen, bleibt die Entscheidung, welches Modell aus dieser Klasse benutzt werden sollte. In dieser Arbeit wurden pro Klasse nur ein oder zwei verschiedene vorgestellt, doch gibt es in der Literatur noch viele andere. Der bekannteste Ansatz in den marginalen Modellen ist die Theorie der verallgemeinerten linearen Modelle und bei den subjektspezifischen Modellen die verallgemeinerten linearen gemischten Modelle. Für ein spezielles Problem können aber genauso gut andere Modelle geeignet sein.

Nach der Entscheidung für ein Modell bleibt das Problem der Versuchsplanung. In der Literatur sind viele Ansätze dazu zu finden, allerdings selten eine Formel zur Bestimmung einer guten Fallzahl. Für marginale Modelle kann das Verfahren von Liu u. Liang (1997) benutzt werden, das allerdings relativ aufwändig ist. In dieser Arbeit wurde die Tauglichkeit dieser Formel nur ansatzweise untersucht. Eine umfassendere Prüfung wäre sinnvoll.

Für subjektspezifische Modelle ist im Rahmen dieser Arbeit kein entsprechender Artikel gefunden worden. Allerdings wäre es interessant, eine Formel für subjektspezifische Modelle zu finden und diese für Probleme, die mit beiden Modelltypen behandelt werden können, mit der entsprechenden Formel für marginale Modelle zu vergleichen.

Anhang

A Erstellung von Abbildung 1

Erstellung einer Graphik, die den unterschiedlichen Verlauf von subjektspezifischen Erwartungen und der daraus resultierenden marginalisierten Erwartung deutlich macht. Eine Vorlage dieser Graphik ist in Molenberghs und Verbeke (2005)zu finden. Hier wird die Graphik allerdings mit eigenen Werten selbst erstellt. Dazu wurde das Programm Maple benutzt. Zunächst wird das Programm neu gestartet und das Statistikprogramm geladen.

```
> restart:
> with(stats):
```

Jetzt werden 20 normalverteilte Zufallsvariablen mit Erwartungswert 0 und Varianz 2 erzeugt und in der Variable x gespeichert. Diese Zufallsvariablen spiegeln den subjektspezifischen Effekt wider.

```
> x:= stats[random, normald[0,2]](20):
```

In einer Schleife werden nun 20 subjektspezifische Funktionen f(i) = exp(0.3+x[i]+2*t)/(1+exp(0.3+x[i]+2*t)) erzeugt. Diese spiegeln die Erfolgswahrscheinlichkeit von i=1,...,20 Subjekten in einem logit-Modell logit(p(i)) = 0.3+ x[i] + 2* t wider. Hier ist x[i] der i-te Eintrag aus dem Vektor x mit den Realisierungen der normalverteilten Zufallsvariablen.

```
> for i from 1 by 1 to 20 do
> p(i):=exp(0.3+x[i]+2*t)/
```

```
> (1+exp(0.3+x[i]+2*t))  end do:
```

Diese Funktionen können nun geplottet werden. Dazu werden in einer Sequenz die verschiedenen plots erzeugt und in der Menge L gespeichert. Der nachfolgende Befehl erzeugt dann die Graphik. (Zur Ausgabe der Graphik muss der Doppelpunkt hinter dem letzten Befehl in ein Semikolon umgeändert werden.)

```
> L := {seq(plot(p(i),t=-2.5..2.5,
> color=black,labels= [t,"P(Y_ij | U_i)"],
> labeldirections=[HORIZONTAL,VERTICAL],
> axes=BOXED,title="Bedingte Erwartungen",
> titlefont=[ HELVETICA, BOLD, 12]),
> i=1..20)}:
> plots[display](L):
```

Nun muss noch die marginalisierte Erwartung berechnet werden. Dazu werden alle subjektspezifischen Kurven addiert und das Ergebnis durch ihre Anzahl geteilt. Die erhaltene Kurve wird in p(m) gespeichert.

```
> p(a):=0:
> for i from 1 by 1 to 20 do
> p(a) := p(a) + p(i) end do:
> p(m):= 1/20*p(a)
```

Die Kurve p(m) wird mit den subjektspezifischen Kurven gemeinsam geplottet. Sie ist mit einem stärkeren Strich gezeichnet.(Zur Ausgabe der Graphik muss wieder der letzte

```
Doppelpunkt zu einem Semikolon abgeändert werden.)

> G:= L union {plot(p(m), t=-2.5 ..2.5,
> color=black,thickness=4,axes=BOXED)}:
> plots[display](G):
```

B Monotonie

Im Folgenden werden die Aussagen über die Funktion

$$\frac{p_A - p_B^{d*}}{\sqrt{\frac{p_A(1-p_A)+p_B^{d*}(1-p_B^{d*})}{N_G}}} =: \sqrt{N_G} f(OR)$$

mit $p_B^{d*} = \frac{OR\frac{p_A}{1-p_A}}{1+OR\frac{p_A}{1-p_A}}$ und konstantem p_A bewiesen. Zunächst besitzt $\sqrt{N_G}f$ bei 1 eine Nullstelle, denn mit einem $OR = 1$ ist die Wahrscheinlichkeit $p_B^{d*} = \frac{\frac{p_A}{1-p_A}}{1+\frac{p_A}{1-p_A}} = p_A$ und damit $\sqrt{N_G}f(1) = 0$.
Um zu zeigen, dass $\sqrt{N_G}f$ monoton fallend ist, muss die erste Ableitung für alle OR kleiner 0 sein. Der Faktor $\sqrt{N_G}$ kann dabei vernachlässigt werden. Es wird im Weiteren nur noch die Funktion f betrachtet.
Sei zunächst $x = \frac{p_A}{1-p_A}$ und $y = p_A(1-p_A)+p_B^{d*}(1-p_B^{d*})$. Dann ergibt sich folgende Bedingung, die sich äquivalent zu einer wahren Aussage umformen lässt:

$$f'(OR) = \frac{-\frac{x}{(1+ORx)^2}\sqrt{y} - \left(p_A - \frac{ORx}{1+ORx}\right)\frac{1}{2}y^{-1/2}\left(\frac{x}{(1+ORx)^2} - 2\frac{ORx}{1+ORx}\frac{x}{(1+ORx)^2}\right)}{(\sqrt{y})^2} < 0$$

$$\begin{aligned}
\Leftrightarrow \quad & -\sqrt{y} - \left(p_A - \tfrac{ORx}{1+ORx}\right)\tfrac{1}{2}y^{-1/2}\left(1 - 2\tfrac{ORx}{1+ORx}\right) && < 0 \\
\Leftrightarrow \quad & -\left(p_A(1-p_A) + \tfrac{ORx}{1+ORx}\left(1 - \tfrac{ORx}{1+ORx}\right)\right) && \\
& \quad -\left(p_A - \tfrac{ORx}{1+ORx}\right)\tfrac{1}{2}\left(1 - 2\tfrac{ORx}{1+ORx}\right) && < 0 \\
\Leftrightarrow \quad & -p_A(1-p_A) - \tfrac{1}{2}\tfrac{ORx}{1+ORx} - \tfrac{1}{2}p_A + p_A\tfrac{ORx}{1+ORx} && < 0 \\
\Leftrightarrow \quad & \left(p_A - \tfrac{1}{2}\right)\tfrac{ORx}{1+ORx} && < p_A\left(1 - p_A + \frac{1}{2}\right) \\
\Leftrightarrow \quad & \left(p_A - \tfrac{1}{2}\right)\frac{OR\frac{p_A}{1-p_A}}{1+OR\frac{p_A}{1-p_A}} && < p_A\left(-p_A + \frac{3}{2}\right) \\
\Leftrightarrow \quad & \left(p_A - \tfrac{1}{2}\right)\left(OR\tfrac{p_A}{1-p_A}\right) && < p_A\left(-p_A + \frac{3}{2}\right)\left(1 + OR\frac{p_A}{1-p_A}\right) \\
\Leftrightarrow \quad & \underbrace{\left(-\frac{1}{2}(p_A+1) + p_A^2\right)}_{<0}\underbrace{(OR\frac{p_A}{1-p_A})}_{>0} && < \underbrace{p_A\left(-p_A + \frac{3}{2}\right)}_{>0}.
\end{aligned}$$

Dabei sind die Ungleichungen in der letzten Zeile wahr, da $p_A \in (0,1)$ ist.

C Funktionen von Beispiel 5.2.1

```
# Die folgenden Funktionen werden in der Versuchsplanung von
# Beispiel 5.2.1 zur Ermittlung der benötigten Stichprobengröße
# benutzt.

# Es wird das Modell logit(mu_ij) = beta0+beta1x_ij mit x_ij = 0
# für Behandlung A und x_ij = 1 für Behandlung B betrachtet. Sei
# i=1,...,N die Anzahl der Personen und j=1,...n die wiederholten
# Beobachtungen pro Person. Außerdem seien die Beobachtungen
# austauschbar. Demnach gilt für die eine Hälfte der Beobachtungen
# logit(p_A) = beta0 und logit(pB)=beta0 + beta1. Für dieses
# Modell soll nun die benötigte Fallzahl bestimmt werden,
# für die es einen Test mit H0: pA=pB gegen H1:pA ungleich pB
# zum Niveau alpha und Power 1-beta bei einer kleinsten
# bedeutungsvollen Differenz zwischen pA und pB geben kann.
# Diese Differenz wird durch den Odds Ratio bestimmt.
#
# Die Funktion anzahlpA berechnet nun aus den übergegebenen Werten
# pA (Erfolgwahrscheinlichkeit der Kontrollgruppe), dem Odds Ratio
# OR, der Korrelation rho der wiederholten Beobachtungen und der
# Anzahl n der wiederholten Beobachtungen pro Person die benötigte
# Stichprobengröße. Dazu wird die Formel von Liu und Liang be-
# nutzt, wobei alpha = 0.05 und beta = 0.10 angenommen wird.
# Diese Funktion gibt reelle Fließpunktzahlen aus.

anzahlpA<-function(pA,OR,rho,n)
(1.960+1.282)**2*0.5*(pA*(1-pA)+(OR*(pA/(1-pA))/(1+OR*pA/(1-pA)))
*(1-(OR*(pA/(1-pA))/(1+OR*pA/(1-pA)))))*(1+(n-1)*rho)/
```

```
(n*0.25*((OR*(pA/(1-pA))/(1+OR*pA/(1-pA)))-pA)**2)

# Die Funktion anzahlpAcl errechnet ebenso wie anzahlpA die
# benötigte Stichprobengröße, rundet sie aber zusätzlich auf
# die nächste ganze gerade Zahl auf, damit die Studie mit
# gleich großen Gruppen durchgeführt werden können.

anzahlpAcl<-function(pA,OR,rho,n)
2*ceiling(0.5*(1.960+1.282)**2*0.5*(pA*(1-pA)+(OR*(pA/(1-pA))/
(1+OR*pA/(1-pA)))*(1-(OR*(pA/(1-pA))/(1+OR*pA/(1-pA)))))
*(1+(n-1)*rho)/(n*0.25*((OR*(pA/(1-pA))/(1+OR*pA/(1-pA)))-pA)**2))

# Die Funktion zeichnenpA benötigt als übergebene Parameter die
# Erfolgswahrscheinlichkeit der Kontrolle pA, den OR, die
# Korrelation rho und die Anzahl der Wiederholungen n. Die
# Funktion gibt dann eine Tabelle aus, in der für 1 bis n die mit
# der Funktion anzahlpA berechneten Stichprobengrößen stehen.
# Außerdem werden die Werte dieser Tabelle graphisch dargestellt.

zeichnenpA<-function(pA,OR,rho,n){
 anz<-NULL
 for(i in (1:n)){
   anz<-c(anz,anzahlpA(pA,OR,rho,i))
   }
 plot(1:n,anz,xlab="Anzahl Wiederholungen",ylab="Anzahl Subjekte",
main="Verhältnis von Wiederholungen zu Subjekten")
 anzahltabelle<-matrix(c(1:n,anz),ncol=2)
 dimnames(anzahltabelle)<-list(NULL,c("Wiederholungen","Subjekte"))
 anzahltabelle
```

```
 }

# Die Funktion zeichnenpAcl ist mit der Funktion zeichnenpA
# identisch, nur dass sie die Funktion anzahlpAcl statt anzahlpA
# benutzt.

zeichnenpAcl<-function(pA,OR,rho,n){
 anz<-NULL
 for(i in (1:n)){
   anz<-c(anz,anzahlpAcl(pA,OR,rho,i))
   }
 plot(1:n,anz,xlab="Anzahl Wiederholungen",ylab="Anzahl Subjekte",
main="Verhältnis von Wiederholungen zu Subjekten")
 anzahltabelle<-matrix(c(1:n,anz),ncol=2)
 dimnames(anzahltabelle)<-list(NULL,c("Wiederholungen","Subjekte"))
 anzahltabelle
 }

# Die Funktion kosten errechnet die Kosten einer Studie mit
# Gleichung (72). Übergeben werden müssen die Anzahl der
# Subjekte anzahlsub, die Rekrutierungskosten
# rekrutierungskosten pro Person, die Untersuchungskosten
# untersuchungskosten pro Person und Beobachtung und die Anzahl
# der Wiederholungen wiederholungen.

kosten<-function(anzahlsub,rekrutierungskosten,
untersuchungskosten,wiederholungen)
rekrutierungskosten*anzahlsub+untersuchungskosten*wiederholungen
*anzahlsub
```

```
# Die Funktion kostenfunktionpA errechnet für 1 bis n
# Wiederholungen die Kosten einer Studie mit Hilfe der Funktion
# kosten. Übergeben werden müssen die Anzahl der Wiederholungen n,
# die Erfolgswahrscheinlichkeit pA bei der Kontrolle, der Odds
# Ratio OR, die Rekrutierungskosten rekrutierungskosten, und die
# Untersuchungskosten untersuchungskosten pro wiederholter
# Beobachtung. Die benötigte Anzahl von Personen für die jeweilige
# Anzahl von Wiederholungen wird mit der Funktion anzahlpA
# berechnet. Ausgegeben wird eine Tabelle für die Kosten von
# Studien mit 1 bis n Beobachtungen pro Person. Außerdem
# werden die Werte der Tabelle graphisch dargestellt.

kostenfunktionpA<-function(n,pA,OR,rho,rekrutierungskosten,
untersuchungskosten)
   {
   kost<-NULL
   for(i in (1:n)){
     kost<-c(kost,kosten(anzahlpA(pA,OR,rho,i),rekrutierungskosten,
untersuchungskosten,i))
     }
   plot(1:n,kost,xlab="Anzahl Wiederholungen",ylab="Kosten",
main="Verhältnis von Wiederholungen zu Kosten")
   kostentabelle<-matrix(c(1:n,kost),ncol=2)
   dimnames(kostentabelle)<-list(NULL,c("Wiederholungen","Kosten"))
   kostentabelle
  }

# Die Funktion kostenfunktionpAcl ist identisch mit der Funktion
```

```
# kostenfunktionpA, nur dass sie statt der Funktion anzahlpA die
# Funktion anzahlpAcl benutzt.

kostenfunktionpAcl<-function(n,pA,OR,rho,rekrutierungskosten,
untersuchungskosten)
   {
   kost<-NULL
   for(i in (1:n)){
     kost<-c(kost,kosten(anzahlpAcl(pA,OR,rho,i),
rekrutierungskosten,untersuchungskosten,i))
     }
   plot(1:n,kost,xlab="Anzahl Wiederholungen",ylab="Kosten",
main="Verhältnis von Wiederholungen zu Kosten")
   kostentabelle<-matrix(c(1:n,kost),ncol=2)
   dimnames(kostentabelle)<-list(NULL,c("Wiederholungen","Kosten"))
   kostentabelle
  }
```

D Erzeugung simulierter binärer Daten

D.1 Eine Beobachtung pro Subjekt

```
/* Das nachfolgende SAS-Makro beob1 erzeugt für eine zu übergeben-
de Subjektenummer sub mit Erfolgswahrscheinlichkeit p und der Be-
handlung treat eine einfache binäre Beobachtung. Die Beobachtung
wird in der Datei anzahl unter dem Namen anz gespeichert.*/

%macro beob1(sub, p, treat);
  data anzahl;
    subject=&sub;
    treat=&treat;
    f0=(1-&p);
    x=rand('uniform');
    if x < f0 then do;
      anz = 0;
    end;
    else do;
      anz = 1;
    end;
    keep subject treat anz;
  run;
%mend beob1;


/* Das Makro tabelle1 erzeugt für n Subjekte einfache binäre
Beobachtungen. Dazu wird das Makro beob1 benutzt. Für die erste
Hälfte der Subjekte werden Daten mit einer Erfolgswahrscheinlich-
```

```
keit pA erzeugt, für die zweite Hälfte wird die Erfolgswahrschein-
lichkeit mit Hilfe von pA und dem angegeben Odds Ratio OR
berechnet. Die erste Behandlung erhält die Bezeichnung K, die
zweite B. Die Daten werden in einer Datei tabelle gespeichert,
die die jeweilige Subjektnummer und die dazugehörige binäre
Beobachtung enthält.*/

%macro tabelle1(n, pA, OR);
  data tabelle;
  %do lauf=1 %to (&n/2);
    %beob1(&lauf, &pA, 'K');
    data tabelle;
      set tabelle anzahl;
    run;
  %end;
  %do lauf = &n/2+1 %to &n;
    %beob1(&lauf,((&OR*&pA)/(1-&pA))/(1+(&OR*&pA)/(1-&pA)),'B');
    data tabelle;
      set tabelle anzahl;
    run;
  %end;
  data tabelle;
    set tabelle(firstobs=2);
  run;
%mend tabelle1;


/* Das Makro wiederholen1 erzeugt nun mehrere Datensätze mit
n Personen, einer Erfolgswahrscheinlichkeit pA für die erste
```

```
Behandlung K und einem Odds Ratio OR. Die Anzahl der erzeugten
Datensätze wird durch die Variable menge übergeben. Die einzelnen
Datensätze werden mit dem Makro tabelle1 erzeugt. Jeder Datensatz
wird mit der SAS-Prozedur genmod ausgewertet. Die erhaltenen Para-
meterschätzer für den Behandlungseffekt sowie die untere und
obere Konfidenzgrenze und der p-Wert für die Signifikanz dieses
Parameters werden in eine Datei schaetzer geschrieben.*/

%macro wiederholen1(menge, n, pA, OR);
  data schaetzer;
  %do i=1 %to &menge;
    %tabelle1(&n,&pA,&OR);
    data myGEE;
    data contrast;
/* GEEEmpEst ist die outputtafel, von der die Schätzungen
genommen werden*/
    ods output GEEEmpPEst = myGEE;
    ods output Contrasts = contrast;
/*proc genmod wertet die binären Daten aus.*/
    proc genmod data = tabelle desc;
      class subject treat;
      model anz = treat /dist = bin;
/*Da hier keine wiederholten Daten vorliegen, ist das repeated-
Statement eigentlich überflüssig.*/
      repeated subject = subject /type= exch corrb corrw;
      contrast 'beta1' treat 1 -1 /E;
    run;
/*Auswahl der benötigten Parameter, hier die Treatment-
Parameter.*/
```

```
    data myGEE;
      set myGEE(firstobs=2 obs=2);
      keep Estimate LowerCL UpperCL ProbZ;
    run;
    data contrast;
      set contrast;
      keep ProbChiSq;
    run;
    data myGEE;
      merge myGEE contrast;
    run;
    data schaetzer;
      set schaetzer myGEE;
    run;
  %end;
  data schaetzer;
    set schaetzer(firstobs=2);
  run;
%mend wiederholen1;
```

D.2 Zwei Beobachtungen pro Subjekt

```
*/ Das Makro beob2 erzeugt wie das Makro beob1 binäre Daten für
ein Subjekt mit der Nummer sub. Diesmal werden allerdings zwei
Beobachtungen pro Subjekt erzeugt. Dazu wird das Bahadur-Modell
benutzt. Mit Hilfe der Wahrscheinlichkeit p für das Auftreten
einer Eins und der Korrelation rho wird bei diesem Modell die
Wahrscheinlichkeit für das Auftreten keiner und einer Eins bei
zwei Beobachtungen bestimmt. Mit Hilfe einer gleichverteilten
```

```
Zufallsvariablen wird damit die Anzahl der Einsen in dem Beobach-
tungsvektor bestimmt. Diese Anzahl wird in der Datei anzahl
unter der Variable anz gespeichert.*/

%macro beob2(sub, rho, p, treat);
  data anzahl;
    subject=&sub;
    treat=&treat;
    f0=((1-&p)**2)*(1+&rho*(&p/(1-&p)));
    f1=2*&p*(1-&p)*(1+&rho*(-1));
    x=rand('uniform');
    if x < f0 then do;
      anz=0;
    end;
    else do;
      if x< f0+f1 then do;
        anz=1;
      end;
      else do;
        anz=2;
      end;
    end;
    keep subject treat anz;
  run;
%mend beob2;


/* Das Makro tabelle2 erzeugt wie das Makro tabelle1 für n
Subjekte binäre Beobachtungen, diesmal allerdings zwei Beobachtun-
```

```
gen pro Subjekt. Dazu wird das Makro beob2 benutzt. Zur Erzeugung
der Daten wird die mit dem Makro beob2 erhaltene Anzahl von
Einsen in binäre Beobachtungen umgewandelt. Für anz=0 ergibt sich
0,0; für anz=1 1,0 und sonst 1,1. Die Reihenfolge der Nullen und
Einsen kann so festgelegt werden, da es sich um austauschbare
Daten handeln soll.
Für die erste Hälfte der Subjekte werden Daten mit einer Erfolgs-
wahrscheinlichkeit von pA für eine Eins erzeugt, für die zweite
Hälfte wird die Erfolgswahrscheinlichkeit mit Hilfe von pA und
dem angegebenen Odds Ratio OR berechnet. Die erste Behandlung
erhält die Bezeichnung K, die zweite B.
Die Daten werden in einer Datei tabelle gespeichert,
die die jeweilige Subjektnummer, die Nummer der Wiederholung
time und die dazugehörige wiederholte binäre Beobachtung
enthält.*/

%macro tabelle2(n, rho, pA, OR);
  data alle;
/*Erzeugung der Variablen anz*/
    %do lauf=1 %to (&n/2);
      %beob2(&lauf,&rho,&pA,'K');
      data alle;
        set alle anzahl;
      run;
    %end;
  data alle;
    set alle(firstobs=2);
  run;
  %do lauf=&n/2+1 %to &n;
```

```
    %beob2(&lauf,&rho,((&OR*&pA)/(1-&pA))/(1+(&OR*&pA)/(1-&pA)),'B');
    data alle;
      set alle anzahl;
    run;
  %end;
/*Einfügen der Wiederholungsnummer*/
  data probe;
    do time=1 to 2;
      output;
    end;
  run;
/*Erstellen der Datei tabelle*/
  data zwischen;
  data wiederholung;
  %do oft=1 %to &n;
    %do weg=1 %to 2;
      data zwischen;
        set zwischen alle(firstobs=&oft obs=&oft);
      run;
    %end;
    data wiederholung;
      set wiederholung probe;
    run;
  %end;
  data tabelle;
    merge wiederholung zwischen;
  run;
/*Umwandeln der Variable anz in binäre Daten*/
  data tabelle;
```

```
    set tabelle(firstobs=2);
    y= anz ge time;
  run;
  proc datasets;
    delete probe zwischen anzahl wiederholung;
  run;
%mend tabelle2;


/*Das Makro wiederholen2 arbeitet analog wie das Makro
wiederholen1, nur dass es das Makro tabelle2 statt tabelle1
benutzt.*/

%macro wiederholen2(menge,personenzahl,korr,pA, OR);
  data schaetzer;
  %do i=1 %to &menge;
    %tabelle2(&personenzahl,&korr,&pA,&OR);
    data myGEE;
    data contrast;
/* GEEEmpPEst ist die outputtafel, von der die Schätzungen
  genommen werden*/
    ods output GEEEmpPEst= myGEE;
    ods output Contrasts = contrast;
    proc genmod data= tabelle desc;
      class subject treat time;
      model y = treat /dist=bin;
/*Mit dem repeated-Statement kann genmod die Korrelation
  berücksichtigen.*/
      repeated subject = subject /type= exch corrb corrw modelse;
```

```
      contrast 'beta1' treat 1 -1 /E;
    run;
/*Auswahl der benötigten Parameter, hier die Treatment-
Parameter.*/
    data myGEE;
      set myGEE(firstobs=2 obs=2);
      keep Estimate LowerCL UpperCL ProbZ;
    run;
    data contrast;
      set contrast;
      keep ProbChiSq;
    run;
    data myGEE;
      merge myGEE contrast;
    run;
    data schaetzer;
      set schaetzer myGEE;
    run;
  %end;
  data schaetzer;
    set schaetzer(firstobs=2);
  run;
%mend wiederholen2;
```

D.3 Drei Beobachtungen pro Subjekt

```
/* Das Makro beob3 arbeitet analog wie das Makro beob2, nur dass
diesmal mit dem Bahadur-Modell die Wahrscheinlichkeiten für keine,
eine und zwei beobachtete Einsen bestimmt werden.*/
```

```
%macro beob3(sub, rho, p, treat);
  data anzahl;
    subject=&sub;
    treat=&treat;
    f0=((1-&p)**3)*(1+&rho*3*(&p/(1-&p)));
    f1=3*&p*((1-&p)**2)*(1+&rho*((&p/(1-&p)) -2));
    f2=3*((&p)**2)*(1-&p)*(1+&rho*(-2+(1-&p)/&p));
    x=rand('uniform');
    if x < f0 then do;
      anz=0;
    end;
    else do;
      if x< f0+f1 then do;
        anz=1;
      end;
      else do;
        if x< f0+f1+f2 then do;
          anz=2;
        end;
        else do;
          anz=3;
        end;
      end;
    end;
    keep subject treat anz;
  run;
%mend beob3;
```

```
/*Das Makro tabelle3 arbeitet analog wie das Makro tabelle2, nur
dass es das Makro tabelle3 statt tabelle2 benutzt. Die Variable
anz wird für anz = 0 in 0,0,0; für anz = 1 in 1,0,0; für anz = 2
in 1,1,0 und sonst in 1,1,1 umgewandelt. */

%macro tabelle3(n, rho, pA, OR);
  data alle;
    %do lauf=1 %to (&n/2);
      %beob3(&lauf,&rho,&pA,'K');
      data alle;
        set alle anzahl;
      run;
    %end;
  data alle;
    set alle(firstobs=2);
  run;
  %do lauf=&n/2+1 %to &n;
    %beob3(&lauf,&rho,((&OR*&pA)/(1-&pA))/(1+(&OR*&pA)/(1-&pA)),'B');
    data alle;
      set alle anzahl;
    run;
  %end;
  data probe;
    do time=1 to 3;
      output;
    end;
  run;
  data zwischen;
```

```
  data wiederholung;
  %do oft=1 %to &n;
    %do weg=1 %to 3;
      data zwischen;
        set zwischen alle(firstobs=&oft obs=&oft);
      run;
    %end;
    data wiederholung;
      set wiederholung probe;
    run;
  %end;
  data tabelle;
    merge wiederholung zwischen;
  run;
  data tabelle;
    set tabelle(firstobs=2);
    y= anz ge time;
  run;
  proc datasets;
    delete probe zwischen anzahl wiederholung;
  run;
%mend tabelle3;


/*Das Makro wiederholen3 arbeitet analog wie das Makro
wiederholen2, nur dass es statt dem Makro tabelle2 das
Makro tabelle3 benutzt.*/

%macro wiederholen3(menge,personenzahl,korr,pA, OR);
```

```
data schaetzer;
  %do i=1 %to &menge;
    %tabelle3(&personenzahl,&korr,&pA,&OR);
    data myGEE;
    data contrast;
/* GEEEmpPEst ist die outputtafel, von der die Schätzungen
genommen werden*/
    ods output GEEEmpPEst= myGEE;
    ods output Contrasts = contrast;
    proc genmod data= tabelle desc;
      class subject treat time;
      model y = treat /dist=bin;
/*Mit dem repeated-Statement kann genmod die Korrelation
  berücksichtigen.*/
      repeated subject = subject /type= exch corrb corrw modelse;
      contrast 'beta1' treat 1 -1 /E;
    run;
/*Auswahl der benötigten Parameter, hier die Treatment-
Parameter.*/
    data myGEE;
      set myGEE(firstobs=2 obs=2);
      keep Estimate LowerCL UpperCL ProbZ;
    run;
    data contrast;
      set contrast;
      keep ProbChiSq;
    run;
    data myGEE;
      merge myGEE contrast;
```

```
    run;
    data schaetzer;
      set schaetzer myGEE;
    run;
  %end;
  data schaetzer;
    set schaetzer(firstobs=2);
  run;
%mend wiederholen3;
```

E Auswertung der Daten

```
/*Das Makro auswertung übernimmt die Auswertung der mit den
Makros wiederholen1, wiederholen2 und wiederholen3 erhaltenen
beta1-Daten des Behandlungseffektes. Dazu wird die SAS-Prozedur
proc univariate benutzt. Proc univariate berechnet unter
anderem den Mittelwert, Varianz, Standardabweichung und
Konfidenzintervalle von übergebenen Daten. Hier werden die
Ergebnisse in die Datei parameterbeta1 geschrieben.*/

%macro auswertung(schaetzer);
/*führe Auswertung für den Parameter beta1 durch*/
data Kennzahlen;
  ods output BasicMeasures=Kennzahlen;
proc univariate data=&schaetzer;
  var estimate;
run;
/*lese die Varianz ab*/
data varianz;
  set Kennzahlen(firstobs=2 obs=2);
  keep VarValue;
  rename VarValue=Varianz;
run;
/*lese die Standardabweichung ab*/
data standardabweichung;
  set Kennzahlen(firstobs=1 obs=1);
  keep VarValue;
  rename VarValue=Standardabweichung;
run;
```

```
/*lese den Mittelwert ab*/
data mean;
  set Kennzahlen(firstobs=1 obs=1);
  keep LocValue;
  rename LocValue=mean;
run;
/*führe Auswertung für die untere Konfidenzintervall-
schranke durch*/
data LCL;
  ods output BasicMeasures=LCL;
proc univariate data=&schaetzer;
  var LowerCL;
run;
data LCL;
  set LCL(firstobs=1 obs=1);
  keep LocValue;
  rename LocValue=LowerCL;
run;
/*führe Auswertung für die obere Konfidenzintervall-
schranke durch*/
data UCL;
  ods output BasicMeasures=UCL;
proc univariate data=&schaetzer;
  var UpperCL;
run;
data UCL;
  set UCL(firstobs=1 obs=1);
  keep LocValue;
  rename LocValue=UpperCL;
```

```
run;
/*stelle die Ergebnisse zusammen*/
data parameterbeta1;
  merge mean Varianz Standardabweichung LCL UCL;
run;

data parameterbeta1;
  set parameterbeta1;
  Weite=UpperCL-LowerCL;
run;
/*lösche nicht benötigte Datein*/
proc datasets;
  delete Kennzahlen varianz standardabweichung mean UCL LCL;
run;
%mend auswertung;


/*Errechne die Power, indem in der Datei schaetzer die
p-Werte kleiner als 0.05 bzgl. der Z-Statistik gezählt werden.*/
data power;
  set schaetzer;
  keep ProbZ;
run;

data sign;
  set power;
  retain counter;
  if _N_ = 1 then counter = 0;
  if ProbZ < 0.05 then counter = counter + 1;
```

```
run;


/*Errechne die Power, indem in der Datei schaetzer die
p-Werte kleiner als 0.05 bzgl. der Chi-Quadrat-Statistik
gezählt werden.*/
data power1;
  set schaetzer;
  keep ProbChiSq;
run;

data sign1;
  set power1;
  retain counter;
  if _N_ = 1 then counter = 0;
  if ProbChiSq < 0.05 then counter = counter + 1;
run;
```

Literatur

Aerts u. a. 2002

Aerts, Marc ; Geys, Helena ; Molenberghs, Geert ; Ryan, Louise M.: *Monographs on Statistics and Applied Probability*. Bd. 96: *Topics in Modelling of Clustered Data*. Chapman and Hall, 2002

Bahadur 1961

Bahadur, Raghu R.: A representation of the joint distribution of responses to n dichotomous items. In: Solomon, H. (Hrsg.): *Studies in Item Analysis and Prediction*, Standford Mathematical Studies in the Social Science VI. Standford, CA: Standford University Press, 1961, S. 158–168

Behnen u. Neuhaus 2003

Behnen, Konrad ; Neuhaus, Georg: *Grundkurs Stochastik*. 4. Auflage. PD-Verlag, 2003

Declerck u. a. 1998

Declerck, Lieven ; Aerts, Marc ; Molenberghs, Geert: Behavior of the likelihood ratio test statistic under a Bahadur model for exchangeable binary data. In: *Journal of Statistical Computation and Simulation* 61 (1998), S. 15–38

DeGroot 1975

DeGroot, Morris H. ; Mosteller, Frederick (Hrsg.): *Probability and Statistics*. Addison-Wesley Publishing Company, 1975 (Addison-Wesley Series in Behavioral Science: Quantative Methods)

Diggle u. a. 2002

Diggle, Peter J. ; Heagerty, Patrick ; Liang, Kung-Yee ; Zeger, Scott: *Analysis of Longitudinal Data*. Oxford University Press, 2002

Erwe 1962

Erwe, Friedhelm: *Differential- und Integralrechnung.* Bd. 1: *Elemente der Infinitesimalrechnung, Differentialrechnung*. BI Hochschultaschenbücher-Verlag, 1962

Genschel u. Becker 2005

Genschel, Ulrike ; Becker, Claudia: *Schließende Statistik, Grundlegende Methoden.* Springer, 2005 (EMILeA-stat)

George u. Bowman 1995

George, E. O. ; Bowman, Dale: A Full Likelihood Procedur for Analysing Exchangable Binary Data. In: *Biometrics* 51 (1995), S. 512–523

Hedeker 2005

Hedeker, Donald: Generalized Linear Mixed Models. In: Everitt, Brian S. (Hrsg.) ; Howell, David C. (Hrsg.): *Encyclopedia of Statistics in Behavioral Science*, John Wiley and Sons, 2005

Henning 2002

Henning, Michael: *Fallzahladjustierung im Rahmen von Klinischen Studien*, Universität Dortmund, Diss., 2002

Hu u. a. 1998

Hu, Frank B. ; Goldberg, Jack ; Hedeker, Donald ; Flay, Brian R. ; Pentz, Mary A.: Comparison of Population-Averaged and Subject-Specific Approaches for Analysing Repeated Binary Outcomes. In: *American Journal of Epidermiology* 147 (1998), S. 694–703

Kupper u. Haseman 1978

Kupper, L. L. ; Haseman, J. K.: The Use of a Correlated Binomial Model for the Analysis of Certain Toxicological Experiments. In: *Biometrics* 34 (1978), S. 69–76

Liang u. Zeger 1986

Liang, Kung-Yee ; Zeger, Scott L.: Longitudinal data analysis using generalized linear models. In: *Biometrika* 73 (1986), S. 13–22

Lipsitz u. a. 1995

Lipsitz, Stuart R. ; Fitzmaurice, Garret M. ; Sleeper, Lynn ; Zhao, L.P.: Estimation Methods for the Joint Distribution of Repeated Binary Observations. In: *Biometrics* 51 (1995), S. 562–570

Liu u. Liang 1997

Liu, Guanghan ; Liang, Kung-Yee: Sample Size Calculations for Studies with Correlated Observations. In: *Biometrics* 53 (1997), S. 937–947

McCullagh u. Nelder 1989

McCullagh, P. ; Nelder, J. A. ; Cox, D. R. (Hrsg.) ; Hinkley, D. V. (Hrsg.) ; Reid, N. (Hrsg.) ; Rubin, D. R. (Hrsg.) ; Silverman, B. W. (Hrsg.): *Monographs on Statistics and Applied Probability*. Bd. 37: *Generalized Linear Models*. zweite Ausgabe. Chapman and Hall, 1989, Nachdruck 1997

Molenberghs u. Verbeke 2004

Molenberghs, Geert ; Verbeke, Geert: Meaningful statistical model formulations for repeated measures. In: *Statistica Sinica* 14 (2004), S. 989–1020

Molenberghs u. Verbeke 2005

Molenberghs, Geert ; Verbeke, Geert: *Models for Discrete Longitudinal Data*. Springer, 2005

Neuhaus u. a. 1991

Neuhaus, J.M. ; Kalbfleisch, J.D. ; Hauck, W.W.: A Comparison of Cluster-Specific and Population-Averaged Approaches for Analyzing Correlated Binary Data. In: *International Statistical Review* 59 (1991), S. 25–35

Pendergast u. a. 1996

Pendergast, Jane F. ; Gange, Stephen J. ; Newton, Michael A. ; Lindstrom, Mary J. ; Palta, Mari ; Fisher, Marian R.: A Survey of Methods for Analyzing Clustered Binary Response Data. In: *International Statistical Review* 64 (1996), S. 89–118

Stanley 1997

Stanley, Richard P. ; Fulton, W. (Hrsg.) ; Garling, D.J.H. (Hrsg.) ; Ribet, K. (Hrsg.) ; Walters, P. (Hrsg.): *Cambridge Studies in Advanced Mathematics 49*. Bd. 1: *Enumerative Combinatorics*. Cambridge University Press, 1997

Stiratelli u. a. 1984

Stiratelli, Robert ; Laird, Nan ; Ware, James H.: Random-Effects Models for Serial Observations with Binary Response. In: *Biometrics* 40 (1984), S. 961–971

Zeger u. a. 1988

Zeger, Scott L. ; Liang, Kung-Yee ; Albert, Paul S.: Models for Longitudinal Data: A Generalized Estimating Equation Approach. In: *Biometrics* 44 (1988), S. 1049–1060

Erklärung

Hiermit versichere ich, dass ich diese Arbeit selbständig verfasst und keine anderen als die angegebenen Hilfsmittel und Quellen benutzt habe.

Oldenburg, den 01.02.2007 .....