

Ludwig-Maximilians-Universität München  
Fakultät für Mathematik, Informatik und Statistik  
Institut für Statistik  
Prof. Dr. Gerhard Tutz

## Bachelorarbeit

zum Thema

Ansätze zur Modellierung von Messwiederholungen mit  
kategorialen Responsevariablen

10.09.2010

Claudia Stuckart  
Statistik, Fachsemester 6

# Inhaltsverzeichnis

<b>1</b>	<b>Einführung</b>	<b>1</b>
1.1	Motivation . . . . .	1
1.2	Generalisierte lineare Modelle . . . . .	2
1.3	Generalisierte Schätzgleichungen . . . . .	3
<b>2</b>	<b>Marginale Modelle</b>	<b>4</b>
2.1	Notation . . . . .	4
2.2	Modellierung . . . . .	5
2.2.1	Spezifizierung der Arbeitskovarianz . . . . .	6
2.2.2	Das binäre logistische marginale Modell . . . . .	11
2.2.3	Das marginale Modell mit ordinalem Response . . . . .	12
2.3	Inferenz . . . . .	13
2.3.1	GEE für $\beta$ und Momentenmethode . . . . .	13
2.3.2	GEE für $\beta$ und GEE für $\alpha$ . . . . .	18
2.3.3	Odds Ratio Parametrisierung . . . . .	20
2.4	Anwendungsbeispiel . . . . .	21
<b>3</b>	<b>Modelle mit zufälligen Effekten</b>	<b>26</b>
3.1	Konzept . . . . .	26
3.2	Notation . . . . .	27
3.3	Modellierung . . . . .	27
3.3.1	Das binäre logistische Modell mit zufälligen Effekten . . . . .	29
3.3.2	Das Modell mit zufälligen Effekten bei ordinalem Response . . . . .	30
3.4	Inferenz . . . . .	31
3.4.1	Posteriori-Modus-Schätzer . . . . .	32
3.4.2	Schätzung basierend auf Integrationstechniken . . . . .	37
3.4.2.1	Maximum-Likelihood-Schätzung für $\beta$ und $Q$ . . . . .	37
3.4.2.2	Posteriori-Erwartungswert-Schätzung für $b_i$ . . . . .	43
<b>4</b>	<b>Anwendung von marginalen Modellen und Modellen mit zufälligen Effekten</b>	<b>45</b>
4.1	Allgemeine Bevölkerungsumfrage der Sozialwissenschaften . . . . .	45
4.2	Marginales Modell mit ordinalem Response . . . . .	47
4.3	Ordinales Modell mit zufälligem Intercept . . . . .	51
4.4	Vergleich der Modelle . . . . .	51
<b>5</b>	<b>Zusammenfassung</b>	<b>54</b>

<b>Literaturverzeichnis</b>	<b>55</b>
<b>Eigenständigkeitserklärung</b>	<b>57</b>

# 1 Einführung

## 1.1 Motivation

Statistische Methoden werden in vielen Studien eingesetzt um Zusammenhänge von Variablen zu erklären. Die zu erklärende Variable bezeichnet man als Zielvariable oder Response. Sie wird im Allgemeinen mit  $\mathbf{y}$  benannt. Die Variablen, deren Einfluss geprüft wird, werden als Kovariablen bezeichnet und haben die Notation  $\mathbf{x}_1, \dots, \mathbf{x}_p$ .

Ist die Zielvariable metrisch und normalverteilt wird ein linearer Zusammenhang zwischen dem Response und den Kovariablen unterstellt. Metrische Variablen können alle Werte innerhalb eines Intervalls annehmen. Regressionsanalysen mit metrischen normalverteilten Response Variablen basieren auf linearen Modellen. Liegen Daten der Form  $(y_i, x_{i1}, \dots, x_{ip})$ ,  $i=1,2,\dots,n$  vor, so hat das klassische (multiple) lineare Regressionmodell die Form:

$$E(y_i|\mathbf{x}_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \epsilon_i,$$

wobei  $\beta_0, \dots, \beta_p$  die unbekanntenen Regressionskoeffizienten bezeichnet. Die Fehler  $\epsilon_1, \dots, \epsilon_n$  sind unabhängig und identisch normalverteilt mit  $E(\epsilon_i) = 0$  und  $var(\epsilon_i) = \sigma^2$ .

Neben den metrischen Variablen existieren auch diskrete Variablen, welche nur endlich viele oder abzählbar viele Ausprägungen annehmen können. Bei endlich vielen Ausprägungen wird die Variable als kategorial bezeichnet. Für die Regressionsanalyse bei kategorialer Response werden generalisierte lineare Modelle verwendet. In diesem Modell wird der Einfluss der Kovariablen auf den Response über eine Responsefunktion  $h$  spezifiziert:

$$E(y_i|\mathbf{x}_i) = h(\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}).$$

Die detailliertere Betrachtung dieses Modells findet sich in Abschnitt 1.2 dargestellt.

Das lineare Modell und das generalisierte lineare Modell decken zwar eine große Anzahl an möglichen Modellen und Fragestellungen ab, sind jedoch nicht für alle Daten geeignet.

Messwiederholungsdaten weisen eine spezielle Struktur auf, die bei der Modellierung berücksichtigt werden muss und können sehr allgemein als wiederholte Messungen rund um ein Thema aufgefasst werden. Eine speziellere Art von Messwiederholungen sind Wiederholungen über die Zeit. Diese werden als Längsschnittdaten bezeichnet. Betrachtet man nicht nur eine Einheit wiederholt über die Zeit, sondern mehrere Objekte, so wird von Paneldaten gesprochen.

Bei Regressionsanalysen werden die Beobachtungen meist als unabhängig angenommen. Diese Annahme ist für die Messwiederholungsdaten nur bedingt sinnvoll. Die Annahme, dass die Objekte voneinander unabhängig sind ist naheliegend, jedoch ist für die wiederholten Messungen an einer Einheit die Annahme der Unabhängigkeit oft nicht korrekt.

Die Analyse von Paneldaten oder allgemeiner Messwiederholungsdaten bedarf spezieller Verfahren und Modelle.

In dieser Arbeit werden zwei Modelle für Messwiederholungen bei kategorialer Resonse präsentiert.

In Kapitel 2 werden die marginalen Modelle und in Kapitel 3 die Modelle mit zufälligen Effekten vorgestellt. In diesen Kapiteln wird zuerst die Notation der Daten erklärt, anschließend die Modellierung, wobei speziell auf binären und ordinalen Response eingegangen wird. Der dritte Abschnitt des jeweiligen Kapitels erläutert die Inferenz im Modell.

Eine Anwendung des marginalen Modells an den Daten des “Sozio-ökonomischen Panels“ (SOEP) wird in Abschnitt 2.4 vorgestellt. Ein Vergleich von Modellen mit zufälligen Effekten mit marginalen Modellen anhand der Daten der “Allgemeinen Bevölkerungsumfrage der Sozialwissenschaften“ (ALLBUS) befindet sich in Kapitel 4. Die Auswertung der Daten erfolgt mit der statistischen Software R.

## 1.2 Generalisierte lineare Modelle

Im folgenden Abschnitt wird das generalisierte lineare Modell (GLM) für Daten ohne Messwiederholung eingeführt.

Bei kategorialer Zielvariable ist eine direkte lineare Modellierung nicht sinnvoll, da der Response nicht-normal verteilt ist. Das generalisierte lineare Modell spezifiziert die Modellierung der Kovariablen über eine Funktion.

Die Spezifizierung des Modells unterteilt sich in eine Verteilungsannahme und ein Strukturannahme.

Die Verteilungsannahme besagt, dass die Zielvariablen für gegebene Kovariablen  $\mathbf{x}_i = (1, x_{i1}, \dots, x_{ip})$  voneinander unabhängig sind und dass die bedingte Dichte von  $y_i$  einer Exponentialfamilie folgt:

$$f(y_i|\theta_i) = \exp \left[ \frac{y_i \theta_i - \psi(\theta_i)}{\phi} + c(y_i, \phi) \right].$$

Der Parameter  $\theta_i$  wird als natürlicher Parameter und der Parameter  $\phi$  als Dispersionsparameter bezeichnet. Erwartungswert und Varianz sind somit:

$$E(y_i|\mathbf{x}_i) = \mu_i = \psi'(\theta_i) \text{ und } \text{var}(y_i|\mathbf{x}_i) = \sigma_i^2 = \phi \psi''(\theta_i).$$

Die Strukturannahme bestimmt, dass der Erwartungswert  $\mu_i$  mit dem linearen Prädiktor  $\eta_i = \mathbf{x}_i\boldsymbol{\beta} = \beta_0 + \beta_1x_{i1} + \dots + \beta_px_{ip}$  durch

$$\mu_i = h(\eta_i) = h(\mathbf{x}_i\boldsymbol{\beta}) \text{ bzw. } \eta_i = g(\mu_i)$$

verknüpft ist, wobei  $h$  eine (eindeutige und zweimal differenzierbare) Responsefunktion und  $g$  die Linkfunktion, d.h. die Umkehrfunktion  $g = h^{-1}$  von  $h$  ist (Fahrmeir *et al.* (2007)).

Beispiele für Verteilungen der Exponentialfamilie sind die Normal-, Binomial- und Poissonverteilung. Das lineare Modell ist somit also ein Spezialfall des generalisierten linearen Modells.

Im generalisierten linearen Modell ist der Erwartungswert und die Varianz durch die Exponentialfamilie bestimmt. Diese beschränkte Varianz muss jedoch nicht zwangsläufig der wahren Varianz entsprechen, deshalb werden die Modellannahmen bei der Parameterschätzung abgeschwächt.

### 1.3 Generalisierte Schätzgleichungen

Eine Möglichkeit die Annahmen des GLM in der Inferenz abzuschwächen ist die Quasi-Likelihood-Schätzung. Bei dieser wird davon ausgegangen, dass  $\mu_i$  der wahre Erwartungswert und  $\sigma_i^2 = \phi\psi''(\theta_i)$  die wahre Varianz ist. Der Parameter  $\phi$  sei fest und  $\psi''(\theta_i)$  eine beliebig spezifizierte Funktion. Das bedeutet, dass nur der Erwartungswert und die Varianz, aber keine Verteilungsannahme spezifiziert wird. Für die Schätzung resultiert eine Schätzgleichung, die formal identisch mit der Maximum-Likelihood-Schätzung ist, jedoch nicht aus der Log-Likelihood abgeleitet wird.

Noch schwächere Annahmen werden bei der generalisierten Schätzgleichung (generalized estimating equation, GEE) festgelegt. Bei dieser Schätzung wird lediglich der Erwartungswert als korrekt spezifiziert angenommen. Die Varianz wird durch die Arbeitsvarianz ersetzt. Man geht nicht davon aus, dass die Arbeitsvarianz der wahren Varianz entspricht.

## 2 Marginale Modelle

In folgendem Kapitel werden die marginalen Modelle erklärt. Der Abschnitt 2.1 befasst sich mit der Struktur der zugrundeliegenden Daten und deren Notation. Die Modellierung des Modells erfolgt in Abschnitt 2.2, welcher sich in die Annahmen des Modells, die Spezifizierung der Arbeitskovarianz in Abschnitt 2.2.1, die explizite Formulierung für binären Response in Abschnitt 2.2.2 und für ordinalen Response in Abschnitt 2.2.3 unterteilt. Die Inferenz des marginalen Modells lässt sich grundsätzlich mit spezifizierter Korrelation oder mit spezifiziertem Odds Ratio vornehmen. Methoden mit Korrelation werden in Abschnitt 2.3.1 und Abschnitt 2.3.2 vorgestellt. Abschnitt 2.3.3 beschreibt die Inferenz bei Modellierung mittels Odds Ratio. Auf die Anwendung des marginalen Modells an den SOEP Daten wird in Abschnitt 2.4 eingegangen.

Liang und Zeger (1986) und Zeger und Liang (1986) stellen die ersten marginalen Modelle vor. Weitere wichtige Literatur zu den marginalen Modellen ist Diggle *et al.* (2002) und Fahrmeir und Tutz (1994).

Marginale Modelle werden bei Messwiederholungsdaten bevorzugt verwendet, wenn das primäre Interesse in der Regression liegt und nicht bei dem Zusammenhang der wiederholten Messungen. Diese Modellklasse ist deshalb im Besonderen geeignet um Effekte der Grundgesamtheit im Durchschnitt zu analysieren.

### 2.1 Notation

Bei Messwiederholungsdaten werden mehrere Einheiten wiederholt gemessen bzw. befragt. Diese Einheiten können Personen oder wie im SOEP Haushalte sein. Diese Einheiten werden im weiteren als Cluster bezeichnet. Für jedes Cluster  $i$ ,  $i = 1, \dots, n$  und jede Messung  $t$ ,  $t = 1, \dots, T_i$  liegen Beobachtungen  $(y_{it}, \mathbf{x}_{it})$  vor. Die Beobachtungen bestehen aus dem Response  $y_{it}$  und dem Vektor der Kovariablen  $\mathbf{x}_{it}$ . Der Response für das  $i$ -te Cluster wird in dem  $T_i \times 1$  Responsevektor

$$\mathbf{y}_i' = (y_{i1}, \dots, y_{iT_i})$$

dargestellt. Der Vektor der Kovariablen  $\mathbf{x}_{it}$  der  $i$ -ten Einheit zur Messung  $t$  ist ein  $p \times 1$  Vektor. Die Matrix  $X_i$  der Kovariablen für die  $i$ -te Einheit ist folglich die  $T_i \times p$  Matrix:

$$X_i = (\mathbf{x}_{i1}, \dots, \mathbf{x}_{iT_i})' = \begin{pmatrix} x_{i11} & x_{i12} & \dots & x_{i1p} \\ x_{i21} & x_{i22} & & \\ \vdots & & & \\ x_{iT_i1} & x_{iT_i2} & & x_{iT_ip} \end{pmatrix}, i = 1, \dots, n.$$

Es wird angenommen, dass die Messwiederholungen, gegeben den Kovariablen voneinander unabhängig sind:

$$y_1, \dots, y_n | \mathbf{x}_1, \dots, \mathbf{x}_n \text{ unabhängig,}$$

wobei  $\mathbf{x}_i' = (\mathbf{x}_{i1}', \dots, \mathbf{x}_{iT_i}')$ .

Fehlende Werte lassen sich in Studien oft nicht vermeiden. Für diese Werte wird angenommen, dass sie zufällig fehlen (Rubin (1976)), d.h. die fehlenden Antworten stehen weder mit der Ausprägung des Merkmals, noch mit einer anderen Variable in Verbindung.

## 2.2 Modellierung

Der folgende Abschnitt beschreibt die Modellierung des marginalen Erwartungswerts des Responses als eine Funktion der Kovariablen. Für die Modellierung wird angenommen, dass Unabhängigkeit zwischen den Clustern besteht. Die wiederholten Messungen innerhalb eines Clusters können unabhängig oder korreliert sein. In marginalen Modellen wird der Einfluss der Kovariablen auf den Response separat von den Korrelationen zwischen den Responses modelliert. Die wahre Kovarianz ist nicht von vordergründiger Bedeutung. Deshalb wird die Kovarianz lediglich als eine Arbeitskovarianzmatrix definiert. Das Wesentliche in marginalen Modellen ist die korrekte Spezifikation des marginalen Erwartungswerts. Ist der marginale Erwartungswert korrekt spezifiziert, so werden die Regressionskoeffizienten  $\boldsymbol{\beta}$  konsistent geschätzt, auch wenn die angenommene Kovarianz nicht der wahren Kovarianz entspricht.

Die Schätzung des Parameter-Vektors  $\boldsymbol{\beta}$  erfolgt durch die Verwendung der generalisierten Schätzgleichung. Die Verfahren zur Inferenz werden in Abschnitt 2.3 detailliert erläutert.

Das marginale Modell hat folgende Annahmen:

1. Der marginale Erwartungswert ist korrekt spezifiziert durch

$$\mu_{it} = E(y_{it} | \mathbf{x}_{it}) = h(\mathbf{x}_{it}' \boldsymbol{\beta}),$$

mit bekannter Responsefunktion  $h$ . Der marginale Erwartungswert des Responses hängt von den Kovariablen  $\mathbf{x}_{it}$  durch die Link Funktion ( $g = h^{-1}$ ) ab, wie z.B den Logit-Link für binären Response. Der Vektor  $\boldsymbol{\beta}$  ist ein  $p \times 1$  Vektor von unbekanntem Regressionskoeffizienten. Dieser Koeffizientenvektor ist konstant über alle Einheiten.

2. Die marginale Varianz ist eine Funktion des marginalen Erwartungswerts:

$$\sigma_{it}^2 = \text{var}(y_{it} | \mathbf{x}_{it}) = \phi v(\mu_{it}),$$

mit bekannter Varianzfunktion  $v$  und Skalierungsparameter  $\phi$ .

3. Die Kovarianz zwischen  $y_{is}$  und  $y_{it}$  ist:

$$\text{cov}(y_{is}, y_{it}) = c(\mu_{is}, \mu_{it}; \boldsymbol{\alpha}),$$

wobei  $c$  eine bekannte Funktion ist. Die Kovarianz ist folglich über eine Funktion  $c$  abhängig von den marginalen Erwartungswerten  $\mu_{is}$ ,  $\mu_{it}$  und eventuell von dem Vektor der Assoziationsparameter  $\boldsymbol{\alpha}$ .

Da die Schätzung der Regressionskoeffizienten  $\boldsymbol{\beta}$  auch für die falsch spezifizierte Kovarianz konsistent ist, wird im weiteren die Arbeitskovarianz verwendet.

Die Arbeitskovarianzmatrix für das  $i$ -te Cluster ist definiert durch:

$$\text{cov}(\mathbf{y}_i) = \Sigma_i(\boldsymbol{\beta}, \boldsymbol{\alpha})$$

und ist abhängig von  $\boldsymbol{\beta}$  und eventuell von  $\boldsymbol{\alpha}$ .

### 2.2.1 Spezifizierung der Arbeitskovarianz

Es gibt zwei Wege die Arbeitskovarianz zu spezifizieren. Die erste Möglichkeit ist über Korrelation und die zweite ist über das Odds Ratio für binäre Daten, beziehungsweise das Globale Odds Ratio für ordinale Daten.

Für die Spezifizierung der Arbeitskovarianz über die Korrelation wird

$$A_i(\boldsymbol{\beta}) = \text{diag}(y_{it} | \mathbf{x}_{it}) = \text{diag}(\sigma_{i1}^2, \dots, \sigma_{iT_i}^2) = \begin{pmatrix} \sigma_{i1}^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_{iT_i}^2 \end{pmatrix}$$

definiert.

Die Arbeitskovarianzmatrix für  $\mathbf{y}_i$  ist:

$$\Sigma_i(\boldsymbol{\beta}, \boldsymbol{\alpha}) = A_i^{1/2}(\boldsymbol{\beta}) R_i(\boldsymbol{\alpha}) A_i^{1/2}(\boldsymbol{\beta}), \quad (1)$$

wobei  $R_i(\boldsymbol{\alpha})$  die  $T_i \times T_i$  Arbeitskorrelationsmatrix für  $\mathbf{y}_i$  ist. Daraus folgt, dass auch  $\Sigma_i(\boldsymbol{\beta}, \boldsymbol{\alpha})$  eine  $T_i \times T_i$  Matrix ist. In der Praxis wird  $R_i(\boldsymbol{\alpha})$  oft nicht richtig spezifiziert, deshalb wird sie nicht als die wahre Korrelationsmatrix angenommen.  $R_i(\boldsymbol{\alpha})$  ist also die gewählte Korrelation, welche von Cluster zu Cluster variieren kann. Der unbekannte Parametervektor  $\boldsymbol{\alpha}$  beschreibt die Korrelation  $R_i(\boldsymbol{\alpha})$  vollständig und ist für alle Cluster  $i$  gleich. Sollte  $R_i(\boldsymbol{\alpha})$  der wahren Korrelation entsprechen, entspricht  $\Sigma_i(\boldsymbol{\beta}, \boldsymbol{\alpha})$  der wahren Kovarianzmatrix. Die Schätzung von  $\boldsymbol{\beta}$  ist konsistent, auch falls  $R_i(\boldsymbol{\alpha})$  nicht korrekt spezifiziert wurde.

In Liang und Zeger (1986) werden einige Beispiele für die Spezifizierung der Arbeitskorrelationsmatrix  $R_i(\boldsymbol{\alpha})$  gegeben. Imai *et al.* (2008) erläutern wie diese in die

statistische Software R implementiert sind.

Die einfachste Möglichkeit ist von unkorrelierten wiederholten Messungen auszugehen:

$$\begin{aligned} \text{corr}(y_{is}, y_{it}) &= 0, \forall s \neq t \\ R_i(\boldsymbol{\alpha}) &= I = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & & \\ \vdots & & \ddots & \\ 0 & & & 1 \end{pmatrix}. \end{aligned}$$

Die Kovarianzmatrix entspricht der Identitätsmatrix. Der Parameter  $\boldsymbol{\alpha}$  ist hier überflüssig.

Geht man von korrelierten Messungen aus, ist eine Möglichkeit das Modell mit stationärer Abhängigkeit m-ten Grades zu spezifizieren. Bei diesem Modell wird die Korrelation definiert durch:

$$\text{corr}(y_{is}, y_{it}) = \begin{cases} \alpha_{|s-t|} & \text{für } |s-t| \leq m \\ 0 & \text{für } |s-t| > m \end{cases},$$

wobei  $m$  die Anzahl der abhängigen Perioden  $t$  ist.

Für  $m = 1$  ist die Korrelationsmatrix:

$$R_i(\boldsymbol{\alpha}) = \begin{pmatrix} 1 & \alpha_1 & 0 & 0 & 0 \\ \alpha_1 & 1 & \alpha_1 & 0 & 0 \\ 0 & \alpha_1 & 1 & \alpha_1 & 0 \\ 0 & 0 & \alpha_1 & 1 & \alpha_1 \\ \vdots & & & & \ddots \end{pmatrix}. \quad (2)$$

Die Matrix für die stationäre Abhängigkeit 2-ten Grades lautet folglich:

$$R_i(\boldsymbol{\alpha}) = \begin{pmatrix} 1 & \alpha_1 & \alpha_2 & 0 & 0 \\ \alpha_1 & 1 & \alpha_1 & \alpha_2 & 0 \\ \alpha_2 & \alpha_1 & 1 & \alpha_1 & \alpha_2 \\ 0 & \alpha_2 & \alpha_1 & 1 & \alpha_1 \\ \vdots & & & & \ddots \end{pmatrix}.$$

Neben der stationären Abhängigkeit m-ten Grades, gibt es auch die nicht-stationäre m-Abhängigkeit. Diese ist bestimmt durch:

$$\text{corr}(y_{is}, y_{it}) = \begin{cases} \alpha_{st} & \text{für } |s-t| \leq m \\ 0 & \text{für } |s-t| > m \end{cases}.$$

Bei nicht-stationärer Abhängigkeit 1-ten Grades ist  $R_i(\boldsymbol{\alpha})$ :

$$R_i(\boldsymbol{\alpha}) = \begin{pmatrix} 1 & \alpha_{12} & 0 & 0 & 0 \\ \alpha_{12} & 1 & \alpha_{23} & 0 & 0 \\ 0 & \alpha_{23} & 1 & \alpha_{34} & 0 \\ 0 & 0 & \alpha_{34} & 1 & \alpha_{45} \\ 0 & 0 & 0 & \alpha_{45} & 1 \\ \vdots & & & & \ddots \end{pmatrix}. \quad (3)$$

Die Korrelationsmatrix der nicht-stationären Abhängigkeit 2-ten Grades ist:

$$R_i(\boldsymbol{\alpha}) = \begin{pmatrix} 1 & \alpha_{12} & \alpha_{13} & 0 & 0 \\ \alpha_{12} & 1 & \alpha_{23} & \alpha_{24} & 0 \\ \alpha_{13} & \alpha_{23} & 1 & \alpha_{34} & \alpha_{35} \\ 0 & \alpha_{24} & \alpha_{34} & 1 & \alpha_{45} \\ 0 & 0 & \alpha_{35} & \alpha_{45} & 1 \\ \vdots & & & & \ddots \end{pmatrix}.$$

Ein Alternative ist die Annahme der Äquikorrelation. Dies bedeutet, dass je zwei Beobachtungen des Clusters  $i$  dieselbe Korrelation  $\alpha$  haben:

$$\text{corr}(y_{is}, y_{it}) = (R_i)_{st} = \alpha (\forall s \neq t)$$

$$R_i(\boldsymbol{\alpha}) = \begin{pmatrix} 1 & \alpha & \alpha & \alpha & \dots \\ \alpha & 1 & \alpha & \alpha & \\ \alpha & \alpha & 1 & \alpha & \\ \vdots & & & & \end{pmatrix}.$$

Eine weitere Möglichkeit ist die Autokorrelationsfunktion eines Gauß AR(1)-Prozesses:

$$\text{corr}(y_{is}, y_{it}) = (R_i)_{st} = \alpha^{|t-s|} = \alpha^{|t-s|} (\forall s \neq t)$$

$$R_i(\boldsymbol{\alpha}) = \begin{pmatrix} 1 & \alpha & \alpha^2 & \alpha^3 & \dots \\ \alpha & 1 & \alpha & \alpha^2 & \\ \alpha^2 & \alpha & 1 & \alpha & \\ \alpha^3 & \alpha^2 & \alpha & 1 & \\ \vdots & & & & \ddots \end{pmatrix}.$$

Diese Arbeitskovarianz ist bei Messwiederholungen oft sinnvoll. Denn die hier spezifizierte Korrelation legt fest, dass ein größerer zeitlicher Abstand eine schwächere Korrelation bedingt.

Es besteht aber auch die Option  $R_i(\boldsymbol{\alpha}) = R(\boldsymbol{\alpha})$  mit  $T_i = T$  unspezifiziert zu lassen und empirisch zu schätzen. Die Korrelationsmatrix ist dann gegeben durch:

$$R(\boldsymbol{\alpha}) = \begin{pmatrix} 1 & \alpha_{12} & \alpha_{13} & \alpha_{14} & \cdots \\ & 1 & \alpha_{23} & \alpha_{24} & \\ & & 1 & \alpha_{34} & \\ & & & & \ddots \end{pmatrix}.$$

Wie bereits angemerkt gibt es zwei Wege die Arbeitskovarianz zu spezifizieren. Für die Spezifikation über die Korrelation wurden einige Möglichkeiten aufgezeigt. Der zweite Weg ist für binären Response über das Odds Ratio (Kreuzproduktverhältnis, Chancenverhältnis) für die Variablen  $y_{is}$  und  $y_{it}$  (Lipsitz *et al.* (1991)). Dies hat einen entscheidenden Vorteil, welcher im Anschluss an die Definition erläutert wird.

Das Odds Ratio von  $y_{it}$  und  $y_{is}$  ist folgendermaßen definiert:

$$\gamma_{ist} = \frac{P(y_{is} = y_{it} = 1)P(y_{is} = y_{it} = 0)}{P(y_{is} = 1, y_{it} = 0)P(y_{is} = 0, y_{it} = 1)}, \quad s \neq t. \quad (4)$$

Im Allgemeinen kann der Vektor der paarweisen Odds Ratio durch

$$\boldsymbol{\gamma} = \boldsymbol{\gamma}(\boldsymbol{\alpha})$$

parametrisiert werden. Die Kovarianz  $cov(y_{it}, y_{is}) = c(\mu_{it}, \mu_{is}, \boldsymbol{\alpha})$  (Abschnitt 2.2, Annahme 3) bezieht das Odds Ratio durch den Parameter  $\boldsymbol{\alpha}$  ein.

Für jedes Cluster  $i$  gibt es  $\binom{T_i}{2}$  Paare von Odds Ratios. Bei großem  $T_i$  ist die Anzahl der Odds Ratio dementsprechend hoch. Die einfachste Möglichkeit ist alle Paare von Odds Ratios als gleich anzunehmen:

$$\gamma_{ist} = \gamma.$$

Dies wird als Äquikorrelation gezeichnet.

Eine weitere Option ist es die Messwiederholungen als unkorreliert anzunehmen. Das Odds Ratio ist infolgedessen:

$$\gamma_{ist} = 1.$$

Die Spezifizierung des Erwartungswerts, der Varianz und des Odds Ratio bestimmen vollständig die Kovarianzen von wiederholten Beobachtungen für ein Cluster. Wie bereits erwähnt hat das Odds Ratio gegenüber der Korrelation bei binären Daten einen entscheidenden Vorteil. Dieser wird im Folgenden erläutert.

Bei binärem Response ist die Kovarianz von zwei Beobachtungen gegeben durch:

$$cov(y_{is}, y_{it}) = P(y_{is} = 1, y_{it} = 1) - \mu_{is}\mu_{it},$$

wobei  $\mu_{is} = \pi_{is} = P(y_{is} = 1) = h(\mathbf{x}_{is}'\boldsymbol{\beta})$ .

Daraus folgt die Korrelation:

$$\text{corr}(y_{is}, y_{it}) = \frac{P(y_{is} = 1, y_{it} = 1) - \mu_{is}\mu_{it}}{(\mu_{is}(1 - \mu_{is})\mu_{it}(1 - \mu_{it}))^{1/2}} = \delta_{ist}.$$

Für  $P(y_{is} = 1, y_{it} = 1)$  gilt:

$$P(y_{is} = 1, y_{it} = 1) = \begin{cases} \frac{1 - (\mu_{is} + \mu_{it})(1 - \gamma_{ist}) - s(\mu_{is}, \mu_{it}, \gamma_{ist})}{2(\gamma_{ist} - 1)} & \text{für } \gamma_{ist} \neq 1 \\ \mu_{is}\mu_{it} & \text{für } \gamma_{ist} = 1, \end{cases} \quad (5)$$

wobei  $s(\mu_{is}, \mu_{it}, \gamma_{ist}) = [(1 - (\mu_{is} + \mu_{it})(1 - \gamma_{ist}))^2 - 4(\gamma_{ist} - 1)\gamma_{ist}\mu_{is}\mu_{it}]^{1/2}$ .

Für die Wahrscheinlichkeit  $P(y_{is} = 1, y_{it} = 1)$  gilt  $P(y_{is} = 1, y_{it} = 1)(\boldsymbol{\beta}, \boldsymbol{\alpha})$ , da  $\mu_{it} = \mu_{it}(\boldsymbol{\beta})$ ,  $\mu_{is} = \mu_{is}(\boldsymbol{\beta})$  und  $\gamma_{ist} = \gamma_{ist}(\boldsymbol{\alpha})$ .

Die gemeinsame Wahrscheinlichkeit  $P(y_{is} = 1, y_{it} = 1)$  ist restringiert durch:

$$\max(0; \mu_{is} + \mu_{it} - 1) \leq P(y_{is} = y_{it} = 1) \leq \min(\mu_{is}, \mu_{it}).$$

Daraus folgt, dass auch die Korrelation  $\rho_{ist} = \text{corr}(y_{is}, y_{it})$  restringiert ist. Um dies zu zeigen wird:

$$\gamma_{ist}^m = \frac{\pi_{is}/(1 - \pi_{is})}{\pi_{it}/(1 - \pi_{it})}$$

definiert. Die Korrelation ist damit beschränkt durch:

$$\max(-\sqrt{\gamma_{ist}^m}, -\sqrt{1/\gamma_{ist}^m}) \leq \rho_{ist} \leq \min(\sqrt{\gamma_{ist}^m}, \sqrt{1/\gamma_{ist}^m}). \quad (6)$$

Der rechte Teil der Ungleichung (6) wird im Folgenden bewiesen:

$$\begin{aligned} \frac{P(y_{is}=1, y_{it}=1) - \mu_{is}\mu_{it}}{(\mu_{is}(1 - \mu_{is})\mu_{it}(1 - \mu_{it}))^{1/2}} &\leq \min\left(\frac{\pi_{is} - \pi_{is}\pi_{it}}{\sqrt{\pi_{is}(1 - \pi_{is})\pi_{it}(1 - \pi_{it})}}, \frac{\pi_{it} - \pi_{is}\pi_{it}}{\sqrt{\pi_{is}(1 - \pi_{is})\pi_{it}(1 - \pi_{it})}}\right) \\ \rho_{ist} &\leq \min\left(\sqrt{\frac{\pi_{is}(1 - \pi_{it})}{(1 - \pi_{is})\pi_{it}}}, \sqrt{\frac{\pi_{it}(1 - \pi_{is})}{\pi_{is}(1 - \pi_{it})}}\right) \\ \rho_{ist} &\leq \min\left(\sqrt{\frac{\pi_{is}/(1 - \pi_{is})}{\pi_{it}/(1 - \pi_{it})}}, \sqrt{\frac{\pi_{it}/(1 - \pi_{it})}{\pi_{is}/(1 - \pi_{is})}}\right). \end{aligned}$$

Aus (6) folgt:

$$\|\rho_{ist}\| \leq \min(\sqrt{\gamma_{ist}}, \sqrt{1/\gamma_{ist}}).$$

Die Korrelation ist von den marginalen Wahrscheinlichkeiten  $\mu_{is}$  und  $\mu_{it}$  abhängig und durch sie beschränkt. Der ursprüngliche Definitionsbereich der Korrelation,  $\rho \in [-1, 1]$  ist bei binären Daten im Allgemeinen größer als das tatsächlich mögliche

Intervall für die Korrelation.

Deshalb wird häufig das Odds Ratio bevorzugt, da es ist nicht durch die marginalen Erwartungswerte beschränkt.

Das Chancenverhältnis darf nicht nur zur Spezifizierung der Kovarianzmatrix bei binärem Response, sondern auch bei ordinalem Response verwendet werden. Heagerty und Zeger (1996), Fahrmeir und Pritscher (1996) und Williamson *et al.* (1995) verwenden das sogenannte Globale Odds Ratio bei ordinalem Response. Das Globale Odds Ratio ist definiert durch:

$$\gamma_{ist}(r, m) = \frac{P(y_{is} \leq r, y_{it} \leq m)P(y_{is} > r, y_{it} > m)}{P(y_{is} \leq r, y_{it} > m)P(y_{is} > r, y_{it} \leq m)}.$$

### 2.2.2 Das binäre logistische marginale Modell

Das Modell mit logistischer Responsefunktion wird bei Regression mit binärem Response verwendet.

Das logistische marginale Modell hat die folgenden Annahmen:

1. Spezifikation der marginale Wahrscheinlichkeit:

$$\pi_{it} = \mu_{it} = E(y_{it} = 1|x_{it}) = h(\mathbf{x}_{it}'\boldsymbol{\beta}). \quad (7)$$

Die Linkfunktion ist der Logit-Link:

$$\text{logit}(\mu_{it}) = \log \frac{\mu_{it}}{1 - \mu_{it}} = \log \frac{P(y_{it} = 1)}{P(y_{it} = 0)} = \mathbf{x}_{it}'\boldsymbol{\beta}.$$

2. Die marginale Varianz der Komponenten ist eine Funktion des marginalen Erwartungswerts:

$$\text{var}(y_{it}) = \phi v(\mu_{it}) = \mu_{it}(1 - \mu_{it}) = \pi_{it}(1 - \pi_{it}) \text{ und } \phi = 1.$$

3. Die Kovarianz ist eine Funktion der marginalen Wahrscheinlichkeiten und gegebenenfalls eines Assoziationsparameters  $\alpha$ :

$$\text{cov}(y_{is}, y_{it}) = c(\mu_{is}, \mu_{it}; \alpha) = c(\pi_{is}, \pi_{it}; \boldsymbol{\alpha}), \text{ für } s \neq t.$$

Wie bereits in Abschnitt 2.2 erwähnt, ist das Wesentliche die richtige Spezifikation der marginalen Wahrscheinlichkeit. Bei richtiger Spezifikation der marginalen Wahrscheinlichkeit lässt sich der Vektor der unbekanntenen Regressionskoeffizienten  $\boldsymbol{\beta}$  auch dann konsistent schätzen, wenn die Kovarianz nicht korrekt spezifiziert ist. Die Festlegung der Arbeitskovarianz ist deshalb ausreichend.

Da bei dem logistischen marginalen Modell der Response binär ist, gibt es zwei

Methoden den Zusammenhang zwischen den Messungen innerhalb eines Clusters zu spezifizieren.

Bei der ersten Methode ist die Kovarianzmatrix:

$$\Sigma_i(\boldsymbol{\beta}, \boldsymbol{\alpha}) = A_i(\boldsymbol{\beta})^{1/2} R_i(\boldsymbol{\alpha}) A_i(\boldsymbol{\beta})^{1/2}, \quad (8)$$

wobei  $A = \text{diag}(\pi_{it}(1 - \pi_{it}))$  und  $R_i(\boldsymbol{\alpha})$  ist die Arbeitskorrelationsmatrix.  $R_i(\boldsymbol{\alpha})$  kann wie in Abschnitt 2.2.1 spezifiziert werden. Die zweite Methode ist über das Odds Ratio  $\gamma_{ist}$  (4).

### 2.2.3 Das marginale Modell mit ordinalem Response

Neben dem binären Response gibt es bei kategorialen Zielvariablen auch den ordinalen Response. Die Zielvariable hat dabei  $k$  Kategorien,  $y_{it} \in \{1, \dots, k\}$ , wobei diese einer Ordnung unterliegen.

Die Messung  $t$  zur Person  $i$ ,  $\mathbf{y}_{it}$  ist auch als ein Vektor darstellbar. Dieser ist definiert durch:

$$\mathbf{y}_{it} = (y_{it1}, y_{it2}, \dots, y_{itq}), \quad q = k - 1,$$

wobei  $y_{itj} = 1$ , falls die Beobachtung in Kategorie  $j$  ist und  $y_{itj} = 0$  sonst.

In welcher Kategorie  $r$  zugeordnet wird, lässt sich motivieren durch die Lage einer latenten Variable bezüglich der Schwellenparameter  $\theta_1, \dots, \theta_q$ . Diese Schwellenparameter müssen einer Ordnung unterliegen:  $-\infty < \theta_1 < \dots < \theta_q < \infty$ .

Das kumulative marginale Modell ist bestimmt durch (Fahrmeir und Pritscher (1996)):

$$g(P(y_{it} \leq r | \mathbf{x}_{it})) = \theta_r + \mathbf{x}_{it}' \boldsymbol{\gamma},$$

wobei  $g$  die Linkfunktion und  $\boldsymbol{\gamma}$  der Vektor der Kovariablen Effekte ist. Die Verwendung des Logit-Links liefert das Proportional Odds Modell:

$$\log \frac{P(y_{it} \leq r | \mathbf{x}_{it})}{P(y_{it} > r | \mathbf{x}_{it})} = \theta_r + \mathbf{x}_{it}' \boldsymbol{\gamma}.$$

Ebenso wie  $y_{it}$  als Vektor  $\mathbf{y}_{it}$  umstrukturierbar ist, kann auch der lineare Prädiktor  $\eta_{it}$  als Vektor  $\boldsymbol{\eta}_{it} = (\eta_{it1}, \dots, \eta_{itq})'$  umstrukturiert werden. Der linearen Prädiktor des marginalen Modells ist, bei der Schreibweise in Matrixform, bestimmt durch  $\boldsymbol{\eta}_{it} = X'_{it} \boldsymbol{\beta}$ , wobei

$$X'_{it} = \begin{pmatrix} 1 & & & \mathbf{x}_{it}' \\ & \ddots & & \vdots \\ & & 1 & \mathbf{x}_{it}' \end{pmatrix} \quad \text{und} \quad \boldsymbol{\beta} = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_q \\ \boldsymbol{\gamma} \end{pmatrix}.$$

Eine alternative Formulierung des kumulativen Modells ist (Heagerty und Zeger (1996)):

$$\log \frac{P(y_{it} > r | \mathbf{x}_{it})}{P(y_{it} \leq r | \mathbf{x}_{it})} = \theta_r + \mathbf{x}_{it}' \boldsymbol{\gamma}.$$

## 2.3 Inferenz

Statistische Inferenz durch die Maximum-Likelihood-Methode ist bei marginalen Modellen nicht möglich, da bei diesen nur die ersten zwei Momente spezifiziert sind. Generalisierte Schätzgleichungen hingegen lassen eine separate Spezifikation von Erwartungswert, Varianz und Kovarianz zu.

Die Schätzung für marginale Modelle ist über den Ansatz von Liang und Zeger (1986) möglich. Dabei wird eine generalisierte Schätzgleichung (Abschnitt 1.3) für  $\boldsymbol{\beta}$  und die Momentenmethode für  $\boldsymbol{\alpha}$  und  $\phi$  verwendet. Ein weiterer Ansatz wird von Prentice (1988) für binäre Daten vorgestellt. Wie im Ansatz von Liang und Zeger (1986) wird eine generalisierte Schätzgleichung für  $\boldsymbol{\beta}$  verwendet. Anstatt der Momentenmethode wird jedoch eine zweite generalisierte Schätzgleichung für  $\boldsymbol{\alpha}$  eingeführt.

Bei diesen Ansätzen wird die Arbeitskovarianz über die Korrelation spezifiziert. Die Inferenz mit der Alternative, dem Odds Ratio wird in Abschnitt 2.3.3 erläutert.

### 2.3.1 GEE für $\boldsymbol{\beta}$ und Momentenmethode

Die generalisierte Schätzgleichung für  $\boldsymbol{\beta}$ , bei festen  $\boldsymbol{\alpha}$  ist:

$$\mathbf{S}_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^n \left( \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} \right)' \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\beta}, \boldsymbol{\alpha}) (\mathbf{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})).$$

Wegen (Toutenburg, 2003, Seite 430):

$$\begin{aligned} \left( \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} \right)' &= \left( \frac{\partial \mathbf{h}(\boldsymbol{\eta}_i)}{\partial \boldsymbol{\beta}} \right)' = \left( \frac{\partial \boldsymbol{\eta}_i}{\partial \boldsymbol{\beta}} \right)' \left( \frac{\partial \mathbf{h}(\boldsymbol{\eta}_i)}{\partial \boldsymbol{\eta}_i} \right)' \\ &= \begin{pmatrix} \mathbf{x}_{i1}' \\ \mathbf{x}_{i2}' \\ \vdots \end{pmatrix}' \begin{pmatrix} \frac{\partial h(\eta_{i1})}{\partial \eta_{i1}} & 0 & 0 \\ 0 & \frac{\partial h(\eta_{i2})}{\partial \eta_{i2}} & \\ 0 & 0 & \ddots \end{pmatrix}, \end{aligned}$$

ist die folgende Schreibweise äquivalent:

$$\mathbf{S}_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^n S_{i,\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^n X_i' D_i(\boldsymbol{\beta}) \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\beta}, \boldsymbol{\alpha}) (\mathbf{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})) = \mathbf{0}, \quad (9)$$

wobei  $\boldsymbol{\mu}_i(\boldsymbol{\beta}) = E(\mathbf{y}_i|X_i) = h(X_i\boldsymbol{\beta})$  der korrekte Erwartungswert von  $\mathbf{y}_i$ , die Kovarianz  $\Sigma_i(\boldsymbol{\beta}, \boldsymbol{\alpha})$  (1) nicht zwingend die wahre Kovarianz und

$$D_i(\boldsymbol{\beta}) = \text{diag}(\partial h/\partial \eta) = \begin{pmatrix} \frac{\partial h(\eta_{i1})}{\partial \eta_{i1}} & & 0 \\ & \ddots & \\ 0 & & \frac{\partial h(\eta_{iT_i})}{\partial \eta_{iT_i}} \end{pmatrix}$$

ist.

Für die Lösung der generalisierten Schätzgleichung ist es wünschenswert, dass sie nur von einem Parameter abhängt. Deshalb wird in Gleichung (9)  $\boldsymbol{\alpha}$  durch  $\hat{\boldsymbol{\alpha}}(\mathbf{y}_1, \dots, \mathbf{y}_n, \boldsymbol{\beta}, \phi)$ , einen  $n^{1/2}$ -konsistenten Schätzer für  $\boldsymbol{\alpha}$ , falls  $\boldsymbol{\beta}$  und  $\phi$  bekannt sind, ersetzt. Folglich ist die generalisierte Schätzgleichung nur noch eine Funktion von  $\boldsymbol{\beta}$ . Zur Vollständigkeit wird auch  $\phi$  durch  $\hat{\phi}(\mathbf{y}_1, \dots, \mathbf{y}_n, \boldsymbol{\beta})$ , einen  $n^{1/2}$ -konsistenten Schätzer für  $\phi$ , falls  $\boldsymbol{\beta}$  bekannt ist, ersetzt.

Die generalisierte Schätzgleichung für  $\boldsymbol{\beta}$  ist folglich:

$$\mathbf{S}_\beta(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^n \mathbf{S}_{i,\beta}[\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}\{\boldsymbol{\beta}, \hat{\phi}(\boldsymbol{\beta})\}] = \mathbf{0}$$

Der Schätzer  $\hat{\boldsymbol{\beta}}$  ist als die Nullstelle der generalisierten Schätzgleichung bestimmt. Des Weiteren kann gezeigt werden, dass  $\hat{\boldsymbol{\beta}}$  unter bestimmten Bedingungen konsistent und asymptotisch normalverteilt ist. Diese Regularitätsbedingungen sind, dass  $T_i$  fest,  $n \rightarrow \infty$  und die Annahmen von Theorem 2 aus Liang und Zeger (1986) gelten.

Die Bedingungen des Theorems 2 sind:

- $\hat{\boldsymbol{\alpha}}$  ist  $n^{1/2}$ - konsistent gegeben  $\boldsymbol{\beta}, \phi$
- $\hat{\phi}$  ist  $n^{1/2}$ - konsistent gegeben  $\boldsymbol{\beta}$ .

Der Schätzer  $\hat{\boldsymbol{\beta}}$  ist dann  $n^{1/2}$ -konsistent und asymptotisch multivariat normalverteilt:

$$\hat{\boldsymbol{\beta}} \stackrel{a}{\sim} N(\boldsymbol{\beta}, F^{-1}VF^{-1}),$$

mit  $F = \sum_{i=1}^n X_i' D_i \Sigma_i^{-1} D_i X_i$  und  $V = \sum_{i=1}^n X_i' D_i \Sigma_i^{-1} \text{cov}(\mathbf{y}_i) \Sigma_i^{-1} D_i X_i$ .

Sei  $A = F^{-1}VF^{-1}$  die Kovarianzmatrix. Ersetzt man  $\boldsymbol{\beta}, \boldsymbol{\alpha}$  und  $\phi$  durch ihre Schätzungen und  $\text{cov}(\mathbf{y}_i)$  durch  $(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_i)(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_i)'$ , so kann die asymptotische Kovarianzmatrix  $\hat{A}$  konsistent geschätzt werden durch:

$$\hat{A} = \text{cov}(\hat{\boldsymbol{\beta}}) \stackrel{a}{\sim} \hat{F}^{-1} \left\{ \sum_{i=1}^n X_i' \hat{D}_i \hat{\Sigma}_i^{-1} (\mathbf{y}_i - \hat{\boldsymbol{\mu}}_i)(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_i)' \hat{\Sigma}_i^{-1} \hat{D}_i X_i \right\} \hat{F}^{-1},$$

mit  $\hat{\boldsymbol{\mu}}_i = \boldsymbol{\mu}_i(\hat{\boldsymbol{\beta}})$ ,  $D_i, \Sigma_i$  ermittelt an der Stelle  $\hat{\boldsymbol{\alpha}}$  bzw.  $\hat{\boldsymbol{\beta}}$  und  $\hat{F}$  :

$$\hat{F} = \sum_{i=1}^n X_i' \hat{D}_i \hat{\Sigma}_i^{-1} \hat{D}_i X_i.$$

### Schätzung von $\boldsymbol{\beta}$

Die Schätzung von  $\boldsymbol{\beta}$  wird errechnet durch eine Iteration zwischen

- dem veränderten Fisher-Scoring-Algorithmus, bei festen  $\hat{\boldsymbol{\alpha}}$
- und der Momentenschätzung von  $\boldsymbol{\alpha}$  und  $\phi$ , bei festen  $\hat{\boldsymbol{\beta}}$ .

Die generalisierte Schätzgleichung wird für gegebene Schätzungen  $\hat{\boldsymbol{\alpha}}$  und  $\hat{\phi}$  durch den Fisher-Scoring-Algorithmus gelöst. Der Vektor  $\hat{\boldsymbol{\beta}}$  sei der nach Abbruch des Fisher-Scoring-Algorithmus gewonnene Schätzer.

Das Fisher-Scoring-Verfahren geht von einem Startwert  $\hat{\boldsymbol{\beta}}^{(0)}$  aus.

Die Iterationen des Algorithmus sind definiert durch:

$$\hat{\boldsymbol{\beta}}^{(k+1)} = \hat{\boldsymbol{\beta}}^{(k)} + (\hat{F}^{(k)})^{-1} \hat{\boldsymbol{S}}^{(k)}, \quad (10)$$

mit der Arbeits-Fisher Matrix

$$\hat{F}^{(k)} = \sum_{i=1}^n X_i' D_i(\hat{\boldsymbol{\beta}}^{(k)}) \tilde{\Sigma}_i^{-1}(\hat{\boldsymbol{\beta}}^{(k)}) D_i(\hat{\boldsymbol{\beta}}^{(k)}) X_i,$$

und der generalisierten Schätzgleichung

$$\hat{\boldsymbol{S}}^{(k)} = \left\{ \sum_{i=1}^n X_i' D_i(\hat{\boldsymbol{\beta}}^{(k)}) \tilde{\Sigma}_i^{-1}(\hat{\boldsymbol{\beta}}^{(k)}) (\mathbf{y}_i - \boldsymbol{\mu}_i(\hat{\boldsymbol{\beta}}^{(k)})) \right\},$$

wobei  $\tilde{\Sigma}_i(\boldsymbol{\beta}) = \Sigma_i[\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}\{\boldsymbol{\beta}, \hat{\phi}(\boldsymbol{\beta})\}]$ .

Gestoppt wird der Algorithmus, falls ein Abbruchkriterium, zum Beispiel

$$\|\hat{\boldsymbol{\beta}}^{(k+1)} - \hat{\boldsymbol{\beta}}^{(k)}\| / \|\hat{\boldsymbol{\beta}}^{(k)}\| \leq \epsilon,$$

erfüllt ist (Fahrmeir *et al.* (2007)). Demzufolge ist  $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{(k)}$ .

Das Fisher-Scoring löst die generalisiert Schätzgleichung und liefert den konsistenten und asymptotisch normalverteilten Schätzer  $\hat{\boldsymbol{\beta}}$ , falls  $\hat{\boldsymbol{\alpha}}$  ein konsistenter Schätzer ist.

### Schätzung von $\boldsymbol{\alpha}$ und $\phi$

Die Parameter  $\boldsymbol{\alpha}$  und  $\phi$  lassen sich schätzen, falls  $\hat{\boldsymbol{\beta}}$  durch den Algorithmus gegeben

ist. Zur Schätzung werden einfache Funktionen von Pearson Residuen verwendet, welche definiert sind durch:

$$\hat{r}_{it} = \frac{y_{it} - \hat{\mu}_{it}}{(v(\hat{\mu}_{it}))^{1/2}} = \frac{y_{it} - \mu_{it}(\hat{\boldsymbol{\beta}})}{(v(\mu_{it}(\hat{\boldsymbol{\beta}})))^{1/2}}.$$

Die Schätzung des Skalierungsparameter  $\phi$  ist:

$$\hat{\phi} = \frac{1}{N - p} \sum_{i=1}^n \sum_{t=1}^{T_i} \hat{r}_{it}^2,$$

wobei  $N = \sum_{i=1}^n T_i$ .

Wie in Abschnitt 2.2.1 beschrieben, gibt es verschiedene Möglichkeiten die Korrelation  $R_i(\boldsymbol{\alpha})$  zu spezifizieren, welche vollständig durch  $\boldsymbol{\alpha}$  beschrieben werden.

Sind die Beobachtungen innerhalb eines Clusters unabhängig, also  $R_i(\boldsymbol{\alpha}) = I$ , ist kein Parameter  $\boldsymbol{\alpha}$  zu schätzen. Der Parameter  $\phi$  wird zur Schätzung von  $\boldsymbol{\beta}$  und  $\text{var}(\boldsymbol{\beta})$  nicht benötigt.

Sei  $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_{T_i-1})'$  und  $\alpha_t = \text{corr}(y_{it}, y_{i,t+1})$ ,  $t = 1, \dots, T_i - 1$ . Dies entspricht der nicht-stationären Abhängigkeit 1-ten Grades (3). Der Doppelindex in (3) ist nicht zwingend nötig und wird deshalb hier durch einen einfachen Index ersetzt. Der Schätzer für  $\boldsymbol{\alpha}$ , gegeben  $\boldsymbol{\beta}$  und  $\phi$  ist:

$$\hat{\alpha}_t = \frac{1}{\phi(N - p)} \sum_{i=1}^n \hat{r}_{it} \hat{r}_{i,t+1}.$$

Der Spezialfall  $\alpha_t = \alpha$ ,  $t = 1, \dots, T_i - 1$  entspricht der stationären Abhängigkeit 1-ten Grades (2) und hat folgende Schätzung:

$$\hat{\alpha} = \frac{1}{T_i - 1} \sum_{t=1}^{T_i-1} \hat{\alpha}_t.$$

In diesem Fall wird die Schätzung von  $\phi$  nicht benötigt.

Ist die Korrelationsstruktur die Äquikorrelation, also  $\text{corr}(y_{is}, y_{it}) = \alpha$ ,  $\forall s \neq t$ , kann  $\alpha$ , für gegebenes  $\phi$  geschätzt werden durch:

$$\hat{\alpha} = \frac{1}{\phi(\sum_{i=1}^n \frac{1}{2} T_i (T_i - 1) - p)} \sum_{i=1}^n \sum_{t>s} \hat{r}_{it} \hat{r}_{is}.$$

Für die Schätzung bei Korrelation  $\alpha^{|t-s|}$ ,  $\forall s \neq t$ , siehe Liang und Zeger (1986)).

Die un spezifizierte Korrelationsstruktur  $R(\boldsymbol{\alpha})$  wird folgendermaßen geschätzt:

$$\frac{1}{\phi n} \sum_{i=1}^n A_i^{-1/2} (\mathbf{y}_i - \hat{\boldsymbol{\mu}}_i) (\mathbf{y}_i - \hat{\boldsymbol{\mu}}_i)' A_i^{-1/2}.$$

Diese Berechnung ist nur nützlich, falls  $T$  im Vergleich zu  $n$  klein ist. Unter dieser Voraussetzung ist der Schätzer von allen Vorgestellten der Beste (Liang und Zeger (1986)).

### Binärer Response

Ein marginales Modell mit binärem Response wurde bereits in Abschnitt 2.2.2 vorgestellt. Für die Inferenz bei binärem Response ist die generalisierte Schätzgleichung für  $\boldsymbol{\beta}$ :

$$\mathbf{S}_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^n \mathbf{S}_{i,\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^n X_i' D_i(\boldsymbol{\beta}) \Sigma_i^{-1}(\boldsymbol{\beta}, \boldsymbol{\alpha}) (\mathbf{y}_i - \boldsymbol{\pi}_i(\boldsymbol{\beta})) = \mathbf{0},$$

wobei  $\Sigma_i^{-1}(\boldsymbol{\beta}, \boldsymbol{\alpha})$  gleich (8).

Ist die marginale Wahrscheinlichkeit  $\boldsymbol{\pi}_i$  (7) korrekt spezifiziert, so ist  $\hat{\boldsymbol{\beta}}$  asymptotisch normalverteilt mit Kovarianz:

$$\hat{F}^{-1} \sum_{i=1}^n X_i' \hat{D}_i \hat{\Sigma}_i^{-1} (\mathbf{y}_i - \hat{\boldsymbol{\pi}}_i) (\mathbf{y}_i - \hat{\boldsymbol{\pi}}_i)' \hat{\Sigma}_i^{-1} \hat{D}_i X_i \hat{F}^{-1},$$

wobei  $\hat{\pi}_{it} = \pi_{it}(\hat{\boldsymbol{\beta}})$ . Die Schätzung des Assoziationsparameters  $\boldsymbol{\alpha}$  ist eine Funktion von Pearson Residuen. Für binären Response ist das Pearson Residuum:

$$\hat{r}_{it} = \frac{y_{it} - \hat{\pi}_{it}}{(\hat{\pi}_{it}(1 - \hat{\pi}_{it}))^{1/2}}.$$

Die Vorschriften zur Schätzung von  $\boldsymbol{\alpha}$  bleiben als Funktionen der Residuen identisch.

Bei un spezifizierten  $R$  ist die Schätzung keine Funktion der Residuen, sondern definiert durch:

$$\frac{1}{n} \sum_{i=1}^n A_i^{-1/2} (\mathbf{y}_i - \hat{\boldsymbol{\pi}}_i) (\mathbf{y}_i - \hat{\boldsymbol{\pi}}_i)' A_i^{-1/2}.$$

### Effizienz

Liang und Zeger (1986) befassen sich nicht nur mit der Schätzung der Parameter,

sondern auch mit der Frage um wie viel  $\beta$  effizienter ist, wenn von korrelierten ( $\hat{\beta}_G$ ) statt von unkorrelierten ( $\hat{\beta}_I$ ) Beobachtungen ausgegangen wird.

Anhand einer Simulation wird gezeigt, dass der Unterschied zwischen  $\hat{\beta}_G$  und  $\hat{\beta}_I$  klein ist, falls die wahre Korrelation kleiner als 0.3 ist. Sowohl  $\hat{\beta}_G$  als auch  $\hat{\beta}_I$  sind sehr gut bei einer kleinen Korrelation  $\alpha$ . Ist  $\alpha$  jedoch groß, so kann eine wesentliche Verbesserung der Schätzer durch die Verwendung der wahren Korrelationsmatrix erreicht werden. Je größer  $\alpha$  ist desto besser ist  $\hat{\beta}_G$  und desto schlechter ist  $\hat{\beta}_I$ . Der Unterschied zwischen den Schätzern ist am größten, wenn die Stichprobengröße unterschiedlich ist.

Hervorzuheben ist, dass der Schätzer und seine Varianz konsistent geschätzt werden können, falls  $\alpha$  und  $\phi$  konsistent geschätzt werden können. Bei falsch spezifizierter Korrelation ist der Schätzer trotzdem konsistent und asymptotisch normalverteilt.

### 2.3.2 GEE für $\beta$ und GEE für $\alpha$

Neben dem Ansatz von Liang und Zeger (1986) zur Inferenz, stellte Prentice (1988) einen weiteren Ansatz für korrelierte binäre Daten vor.

Die generalisierte Schätzgleichung für  $\beta$  ist, wie bisher definiert durch (9). Zur Schätzung von  $\alpha$  wird jedoch keine Funktion von Residuen verwendet. Stattdessen wird die generalisierte Schätzgleichung für  $\alpha$  eingeführt.

Dazu wird die Stichprobenvarianz definiert:

$$Z_{its} = Z_{its}(\beta) = \frac{(y_{it} - \pi_{it})(y_{is} - \pi_{is})}{(\pi_{it}(1 - \pi_{it})\pi_{is}(1 - \pi_{is}))^{1/2}} = r_{it}r_{is}. \quad (11)$$

Daraus folgt:

$$\mathbf{Z}_i' = (r_{i1}r_{i2}, r_{i1}r_{i3}, \dots, r_{i,T_i-1}r_{iT_i}).$$

Der Erwartungswert der Stichprobenvarianz liefert die Elemente von  $R_i(\alpha)$ :

$$E(Z_{its}) = \text{corr}(y_{it}, y_{is}) = \delta_{ist}(\alpha)$$

Die Varianz der Stichprobenvarianz ist:

$$\begin{aligned} \text{var}(Z_{its}) &= w_{ist} = \text{var}(r_{it}r_{is}) \\ &= 1 + (1 - 2\pi_{it})(1 - 2\pi_{is})(\pi_{it}(1 - \pi_{it})\pi_{is}(1 - \pi_{is}))^{-1/2}\delta_{ist} - \delta_{ist}^2, \end{aligned}$$

wobei  $t < s \leq T_i$ .

Demzufolge ist die Matrix  $W_i$  gegeben durch:

$$W_i = \text{diag}(w_{ist}) = \begin{pmatrix} \text{var}(r_{i1}r_{i2}) & & & \\ & \ddots & & \\ & & \ddots & \\ & & & \text{var}(r_{i,T_i-1}r_{iT_i}) \end{pmatrix}.$$

Des Weiteren ist der Vektor der paarweisen Korrelationen  $\boldsymbol{\delta}_i$  definiert durch:

$$\boldsymbol{\delta}_i' = (\delta_{i12}, \dots, \delta_{i1T_i}, \delta_{i23}, \dots).$$

Die generalisierte Schätzgleichung für  $\boldsymbol{\alpha}$  ist:

$$\mathbf{S}_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{i=1}^n E_i' W_i^{-1} (\mathbf{Z}_i - \boldsymbol{\delta}_i(\boldsymbol{\alpha})) = \mathbf{0},$$

wobei  $E_i = \left( \frac{\partial \boldsymbol{\delta}_i}{\partial \boldsymbol{\alpha}} \right)$ .

Der Schätzalgorithmus um  $\boldsymbol{\beta}$  und  $\boldsymbol{\alpha}$  zu berechnen startet mit  $\boldsymbol{\beta}_0$  und  $\boldsymbol{\alpha}_0$  und iteriert nach der Regel:

$$\boldsymbol{\alpha}^{(k+1)} = \boldsymbol{\alpha}^{(k)} + \left( \sum_i E_i' W_i^{-1} E_i \right)^{-1} \sum_i E_i' W_i^{-1} (\mathbf{Z}_i - \boldsymbol{\delta}_i)$$

Die Schätzgleichung für  $\boldsymbol{\beta}$  ist (9). Die Iteration erfolgt nach der Regel (10). Die Lösung  $\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}$  ist asymptotisch normalverteilt mit Varianz:

$$\begin{pmatrix} \hat{F}^{-1} & 0 \\ B & \left( \sum_i \hat{E}_i' \hat{W}_i^{-1} \hat{E}_i \right)^{-1} \end{pmatrix} \begin{pmatrix} \hat{V} & \mathbf{C} \\ \mathbf{C}' & \sum_i \hat{E}_i' \hat{W}_i^{-1} \text{cov}(\hat{\mathbf{Z}}_i) \hat{W}_i^{-1} \hat{E}_i \end{pmatrix} \begin{pmatrix} \hat{F}^{-1} & B' \\ 0 & \left( \sum_i \hat{E}_i' \hat{W}_i^{-1} \hat{E}_i \right)^{-1} \end{pmatrix},$$

wobei

$$\begin{aligned} \mathbf{C} &= \sum_i X_i' \hat{D}_i \hat{\Sigma}_i \text{cov}(\mathbf{y}_i, \hat{\mathbf{Z}}_i) \hat{W}_i^{-1} \hat{E}_i, \\ B &= \left( \sum_i \hat{E}_i' \hat{W}_i^{-1} \hat{E}_i \right)^{-1} \left( \sum_i \hat{E}_i' \hat{W}_i^{-1} \frac{\partial \hat{\mathbf{Z}}_i}{\partial \boldsymbol{\beta}} \right) \hat{F}^{-1}, \\ \text{cov}(\mathbf{y}_i, \mathbf{Z}_i) &= (\mathbf{y}_i - \hat{\boldsymbol{\pi}}_i) (\hat{\mathbf{Z}}_i - \hat{\boldsymbol{\delta}}_i)', \\ \text{cov}(\mathbf{Z}_i) &= (\hat{\mathbf{Z}}_i - \hat{\boldsymbol{\delta}}_i) (\hat{\mathbf{Z}}_i - \hat{\boldsymbol{\delta}}_i)'. \end{aligned}$$

In dem Ansatz von Prentice (1988) wurden Korrelationen als Parameter in der generalisierten Schätzgleichung von  $\boldsymbol{\alpha}$  verwendet.

Einen leicht veränderten Ansatz stellten Fahrmeir und Tutz (1994) vor. Die generalisierte Schätzgleichung in diesem Ansatz ist definiert durch:

$$\sum_i \left( \frac{\partial \boldsymbol{\theta}_i}{\partial \boldsymbol{\alpha}} \right)' C_i^{-1} (\mathbf{w}_i - \boldsymbol{\theta}_i),$$

wobei  $\mathbf{w}_i = (y_{i1}y_{i2}, y_{i1}y_{i3}, \dots, y_{iT_i-1}y_{iT_i})$ ,  $\boldsymbol{\theta}_i = E(\mathbf{w}_i)$  und  $C_i = \text{diag}(\text{var}(\mathbf{w}_i))$ .

Der Ansatz von Liang und Zeger (1986) und Prentice (1988) wird als GEE1 (first-order generalized estimating equation) bezeichnet. Miller *et al.* (1993) beschreiben diesen Ansatz nochmal explizit für nominalen und ordinalen Response.

### 2.3.3 Odds Ratio Parametrisierung

In Abschnitt 2.2.1 wurde die Korrelation und das Odds Ratio als Möglichkeit zur Spezifizierung der Abhängigkeit der Messwiederholungen angeführt. Die Inferenzmethoden von Liang und Zeger (1986) im Abschnitt 2.3.1 und von Prentice (1988) im Abschnitt 2.3.2 verwenden die Korrelationen als Assoziationsmaß.

Im folgenden Abschnitt wird die Inferenz bei Verwendung des Odds Ratio erklärt. Lipsitz *et al.* (1991) spezifizieren den Zusammenhang über das Odds Ratio für binäre Daten und erweitern den Ansatz von Prentice (1988) entsprechend.

Dafür wird der Vektor  $\mathbf{U}_i = \{y_{ist}\}$  mit

$$y_{its} = I\{y_{it} = y_{is}\} = y_{it}y_{is}$$

definiert. Der Erwartungswert der Komponenten ist:

$$E(y_{its}) = P(y_{it} = y_{is} = 1)$$

und die Varianz des Vektor  $\mathbf{U}_i$  ist:

$$H_i = \text{var}(\mathbf{U}_i) = \text{diag}\{\text{var}(y_{its})\} = \text{diag}\{P(y_{it} = y_{is} = 1)(1 - P(y_{it} = y_{is} = 1))\}.$$

Die Wahrscheinlichkeit  $P(y_{it} = y_{is} = 1)$  ist bestimmt durch die Wahrscheinlichkeiten  $\pi_{is}(\boldsymbol{\beta}), \pi_{it}(\boldsymbol{\beta})$  und das Odds Ratio  $\gamma_{ist}(\boldsymbol{\alpha})$ , siehe (5).

Die Generalisierte Schätzgleichung für  $\boldsymbol{\beta}$  bleibt (9).

Die Schätzgleichung für  $\boldsymbol{\alpha}$  ist definiert durch:

$$\mathbf{S}_\alpha(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{i=1}^n G_i' H_i^{-1} (\mathbf{U}_i - \boldsymbol{\tau}_i(\boldsymbol{\beta}, \boldsymbol{\alpha})) = \mathbf{0},$$

wobei  $\boldsymbol{\tau}_i(\boldsymbol{\beta}, \boldsymbol{\alpha}) = E(\mathbf{U}_i)$  und  $G_i = \partial \boldsymbol{\tau}_i / \partial \boldsymbol{\alpha}$ . Die Iteration für  $\boldsymbol{\beta}$  ist (10). Der Fisher-Scoring-Algorithmus für  $\boldsymbol{\alpha}$  iteriert folgendermaßen:

$$\hat{\boldsymbol{\alpha}}^{(k+1)} = \hat{\boldsymbol{\alpha}}^{(k)} - \left( \sum_{i=1}^n \hat{G}_i' \hat{H}_i^{-1} \hat{G}_i \right)^{-1} \left( \sum_{i=1}^n \hat{G}_i' \hat{H}_i^{-1} \{ \mathbf{U}_i - \boldsymbol{\tau}_i(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}) \} \right).$$

Die Schätzung  $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$  ist konsistent für  $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ .

Die Parametrisierung über das Odds Ratio ist auch für ordinalen Response möglich. Williamson *et al.* (1995) erweiterten diese Methode auf ordinalen Response, jedoch nur für bivariate Daten.

Heagerty und Zeger (1996) und Fahrmeir und Pritscher (1996) erläutern die GEE1 Inferenzmethode für korrelierten ordinalen Response. Der Response geht in der Form  $\mathbf{y}_i = (\mathbf{y}_{i1}', \dots, \mathbf{y}_{iT_i}')'$  und die Daten in der Form  $X_i = (X'_{i1}, \dots, X'_{iT_i})'$  in die generalisierte Schätzgleichung ein.

Die Inferenz marginaler Modelle ist auch mit anderen Ansätzen entwickelt worden. Liang *et al.* (1992) entwickelten einen Ansatz für binäre Daten und Odds Ratio Parametrisierung, den sogenannte GEE2 Ansatz (second-order generalized estimating equation method). Bei diesem wird ein Parametervektor, der alle unbekannt Parameter enthält, durch eine einzige generalisierte Schätzgleichung geschätzt. Ein alternativer Ansatz ist die ALR Methode (alternating logistic regression) von Carey *et al.* (1993) für binäre Daten und Parametrisierung über das Odds Ratio. Die Schätzung basiert auf Iterationen zwischen GEE1 für die Regressionskoeffizienten und logistischen Regressionen für die Odds Ratios. Heagerty und Zeger (1996) erläutern die GEE2 und die ALR Methode für ordinalen Response.

## 2.4 Anwendungsbeispiel

In diesem Abschnitt wird die Anwendung der marginalen Modellen auf Messwiederholungsdaten vorgestellt.

Das Deutsche Institut für Wirtschaftsforschung (DIW) befragt jährlich tausende Haushalte in ganz Deutschland zu mehreren hundert sozioökonomischen Merkmalen (Fahrmeir *et al.* (2003)). Die ausgewählten Haushalte bleiben über mehrere Jahre hinweg in der Studie. Für die Haushalte existieren folglich Messwiederholungen über die Zeit. Da viele Haushalte parallel und über die Zeit befragt werden, liegen Paneldaten vor, die als das “Sozio-ökonomische Panel“ (SOEP) bezeichnet werden.

Es werden darin Merkmale zu den Personen im Haushalt und der Haushaltsausstattung erhoben.

Im Folgenden werden jedoch keine longitudinalen Daten betrachtet, sondern die Daten des Jahres 2007. Die Auswertung bezieht sich auf Messwiederholungen rund um ein Thema.

Die Variablen auf die sich die Auswertung bezieht sind:

- Farb-TV
- Telefon
- Internet
- Geburtsjahr
- Geschlecht
- Nettoverdienst
- Deutsche Staatsangehörigkeit

Die Struktur des originalen Datensatzes ist wie folgt:

TV	Tel.	Internet	Geb.jahr	Geschlecht	Verdienst	Staatsangehörigkeit
1	1	0	1950	2	2000	1
1	0	1	1988	1	500	0
⋮						

Der Medienbesitz ist die Zielgröße, welche untersucht werden soll. Diese setzt sich aus den Variablen Farb-TV, Telefon und Internet zusammen. Zur Auswertung müssen diese drei Variablen ein Responsevektor sein. Der Datensatz wird deshalb umstrukturiert zu:

Medien	Geburtsjahr	Geschlecht	Verdienst	Staatsangehörigkeit
1	1950	2	2000	1
1	1950	2	2000	1
0	1950	2	2000	1
1	1988	1	500	0
0	1988	1	500	0
1	1988	1	500	0
⋮				

Die Antworten zu einer Person stehen jeweils untereinander. Dies lässt sich durch eine selbst geschriebene Schleife oder mit dem Befehl `reshape()` aus dem Paket `stats` vornehmen. Da sich die Kovariablen während der drei Befragungen zum Medienbesitz nicht ändern, stehen sie drei mal wiederholt im Datensatz. Der Befehl `reshape()` erzeugt von alleine eine Variable `id`, welche die Personen durchnummeriert und eine Variable `time`, die angibt nach welchem Medium gefragt wurde. Die Variablen Geschlecht, Deutsche Staatsangehörigkeit, Medien sind 1-2 kodiert, wobei 2 für weiblich oder nein steht. Besser zur Auswertung und Interpretation ist eine 0-1 Kodierung, wobei 0 für nein stehen soll.

Dementsprechend sind die Ausprägungen und Kodierungen der im folgenden verwendeten Regression:

- Medienbesitz: nein (0), ja (1)
- time: Farb- TV (1), Telefon (2), Internet (3)
- Geschlecht: weiblich (0), männlich (1)
- Deutsche Nationalität: nein (0), ja (1)

- Alter: in Jahren angegeben
- Nettoverdienst: in Euro angeben

An der Umfrage haben sich 7896 Männer und 5401 Frauen beteiligt, deren Alters-Median bei 51 Jahren lag. Der Nettoverdienst wurde zwischen 20 Euro und 25000 Euro abgegeben, wobei der Median bei 1600 Euro war. Im Besitz eines Farb-TV waren 11428 Personen, eines Telefons 11282 Personen und eines Internetanschlusses 7097 Personen. Dahingegen besaßen 237 Personen keinen Farb-TV, 384 kein Telefon und 4501 Personen kein Internet.

Die Anzahl der Cluster, die in die Regression eingehen, sind 6008 und die maximale Clustergröße ist 3. Die Zielvariable Medienbesitz ist binär. Das binäre logistische marginale Modell ist im Paket `geepack` in der Funktion `geeglm()` implementiert.

Für die erste Regression wird Unabhängigkeit zwischen den Messungen angenommen. Dementsprechend ist die Korrelationsmatrix  $R_i(\boldsymbol{\alpha})$  gleich der Einheitsmatrix. Aber auch die Modellierung von abhängigen Messwiederholungen ist ebenfalls zulässig. Die Korrelationsmatrix  $R_i(\boldsymbol{\alpha})$  kann mit der Äquikorrelation (`corstr=exchangeable`), Autokorrelation (`corstr=ar1`) oder als unspezifiziert (`corstr=unstructured`) in die Regression eingehen. Die geschätzten Regressionskoeffizienten, die geschätzten Dispersionsparameter und die Schätzer für den Korrelationsparameter finden sich in Tabelle 1.

Der Tabelle 1 kann entnommen werden, dass die Regressionskoeffizienten, Standardabweichungen und die Signifikanzen in allen vier Modellen fast gleich sind. Die Schätzer für  $\boldsymbol{\alpha}$  sind sehr klein. Die Beobachtungen sind also annähernd unkorreliert.

In einer Simulation von Liang und Zeger (1986) wurde festgestellt, dass bei einer wahren Korrelation kleiner als 0.3 die Regressionskoeffizienten bei Annahme von Korreliertheit und Unkorreliertheit nur wenig Unterschied aufweisen (Abschnitt 2.3.1). Dies bestätigt sich bei dieser Auswertung.

Die Interpretation der Regressionskoeffizienten  $\boldsymbol{\beta}$  ist diesselbe wie in Querschnittstudien und unabhängig von der Anzahl der Messwiederholungen (Liang *et al.* (1992)).

Im folgenden Abschnitt werden die Regressionskoeffizienten interpretiert. Die Zahlenwerte beziehen sich auf das Unabhängigkeitsmodell. Die Werte für die Modelle mit Korrelation sind von diesen nur minimal verschieden.

Der transformierte Koeffizient  $\exp(\beta_0) = 12.34251920$  gibt das Verhältnis der Häufigkeiten von Medienbesitz zu kein Medienbesitz an, wenn alle anderen Variablen 0 sind. Der Parameter  $\beta_{time2} = -3.3771414$  gibt das logarithmierte Chancenverhältnis von Telefonbesitzern zu Farbfernsehbesitzern an, unter den Personen nicht deutscher Staatsangehörigkeit, im Alter von 0 Jahren und 0 Euro Einkommen an. Mit  $\exp(\beta_{time2}) = 0.03414492$  erhält man das Chancenverhältnis an

Tabelle 1: Regressionskoeffizienten, geschätzte Dispersionsparameter  $\phi$  und geschätzte Korrelationsparameter  $\alpha$

	independence				exchangeable				ar1				unstructured			
	Estimate	Std.err	Pr(>  W )	Pr(>  W )	Estimate	Std.err	Pr(>  W )	Pr(>  W )	Estimate	Std.err	Pr(>  W )	Pr(>  W )	Estimate	Std.err	Pr(>  W )	
(Intercept)	2.5130501	0.5904193	2.08e-05	2.08e-05	2.605641	0.595624	1.2e-05	1.2e-05	2.551189	0.598231	2.0e-05	2.0e-05	2.543010	0.590947	1.7e-05	
time2	-3.3771414	0.6792990	6.64e-07	6.64e-07	-3.354947	0.682682	8.9e-07	8.9e-07	-3.254357	0.682506	1.9e-06	1.9e-06	-3.098621	0.673828	4.3e-06	
time3	-2.0618508	0.6161209	0.000818	0.000818	-2.139864	0.621455	0.00057	0.00057	-2.082552	0.623556	0.00084	0.00084	-2.042943	0.616735	0.00092	
Geschlecht1	-0.3747206	0.2042464	0.066558	0.066558	-0.373894	0.206255	0.06987	0.06987	-0.372015	0.205300	0.06998	0.06998	-0.373513	0.204681	0.06802	
Alter	0.0495662	0.0081507	1.19e-09	1.19e-09	0.046423	0.008137	1.2e-08	1.2e-08	0.048385	0.008177	3.3e-09	3.3e-09	0.048574	0.008135	2.4e-09	
Nationalität1	-0.8012728	0.5125048	0.117948	0.117948	-0.803583	0.516314	0.11962	0.11962	-0.817571	0.520385	0.11616	0.11616	-0.799558	0.512496	0.11873	
Einkommen	0.0001160	0.0001078	0.281811	0.281811	0.000139	0.000116	0.23052	0.23052	0.000131	0.000115	0.25464	0.25464	0.000121	0.000109	0.26760	
time2:Geschlecht1	0.0601482	0.2586621	0.816122	0.816122	0.091004	0.260549	0.72688	0.72688	0.104718	0.258706	0.68564	0.68564	0.122309	0.257678	0.63503	
time3:Geschlecht1	0.3565023	0.2150505	0.097366	0.097366	0.355335	0.217340	0.10206	0.10206	0.356306	0.216073	0.09915	0.09915	0.362089	0.215561	0.09301	
time2:Alter	0.0248553	0.0116028	0.032179	0.032179	0.024663	0.011474	0.03160	0.03160	0.021421	0.011385	0.05989	0.05989	0.016887	0.011199	0.13159	
time3:Alter	-0.0662721	0.0086783	2.23e-14	2.23e-14	-0.063315	0.008684	3.1e-13	3.1e-13	-0.065269	0.008705	6.5e-14	6.5e-14	-0.066089	0.008667	2.4e-14	
time2:Nationalität1	1.5450488	0.5508006	0.005030	0.005030	1.560414	0.554641	0.00490	0.00490	1.578523	0.557354	0.00462	0.00462	1.570095	0.547861	0.00416	
time3:Nationalität1	1.5126146	0.5275051	0.004137	0.004137	1.510252	0.530321	0.00440	0.00440	1.520148	0.535072	0.00450	0.00450	1.498350	0.526892	0.00446	
time2:Einkommen	0.0004845	0.0001730	0.005094	0.005094	0.000446	0.000182	0.01405	0.01405	0.000446	0.000180	0.01330	0.01330	0.000453	0.000170	0.00749	
time3:Einkommen	0.0004390	0.0001160	0.000154	0.000154	0.000415	0.000124	0.00082	0.00082	0.000421	0.000124	0.00066	0.00066	0.000429	0.000117	0.00025	
$\phi$	1.106	2.880	1.09	1.09	1.09	2.41	2.41	2.41	1.08	2.20	1.07	1.07	1.07	2.15	2.15	
$\alpha$					0.116	0.262			0.158	0.321						
$\alpha_{12}$													0.0178	0.0385		
$\alpha_{13}$													0.0358	0.0733		
$\alpha_{23}$													0.2980	0.6139		

sich. Bei deutscher Staatsangehörigkeit oder Einkommen bzw. Alter größer 0 ist zusätzlich die jeweilige Zweier-Interaktion mit *time2* zu beachten. Die Zweier-Interaktion besagt, dass der Effekt einer Kovariable von dem Wert einer anderen Kovariable abhängt. Das Chancenverhältnis von Telefonbesitzern zu Farbfernsehbesitzern bei 20 jährigen Personen deutscher Staatsangehörigkeit ohne Einkommen ist  $\exp(\beta_{time2} + 20 * \beta_{time2:Alter}) = 0.05613278$ .

Die Variablen *time2*, *time3*, *Alter* und die Interaktionen *time3:Alter*, *time:Nationalität*, *time:Einkommen* sind hoch signifikant.

Steigt das Alter um 1 Jahr, so steigt die Chance Medien zu besitzen um  $\exp(\beta_{Alter}) = 1.05081516$  bei den Fernsehbesitzern, um  $\exp(\beta_{Alter}) + \exp(\beta_{time2:Alter}) = 2.075982$  bei den Telefonbesitzer und um  $\exp(\beta_{Alter}) + \exp(\beta_{time3:Alter}) = 1.986691$  bei den Internetbesitzern. Der Einfluss der Variable Nationalität ist an sich nicht signifikant (p-Wert=0.117948 > 0.5), hingegen ihre Interaktion mit Telefon (p-Wert=0.005030 < 0.5) bzw Internet (p-Wert=0.004137 < 0.5) schon. Ebenso verhält es sich bei der Variable Einkommen. Steigt das Einkommen um 1 Euro, so steigt die Chance Medien zu besitzen um  $\exp(\beta_{Einkommen}) = 1.00011603$  bei den Fernsehbesitzern, um  $\exp(\beta_{Einkommen}) + \exp(\beta_{time2:Einkommen}) = 2.000601$  bei den Telefonbesitzer und um  $\exp(\beta_{Einkommen}) + \exp(\beta_{time3:Einkommen}) = 2.000555$  bei den Internetbesitzern. Der Schätzer  $\exp(\beta_{Nationalität}) = 0.44875741$  gibt das Chancenverhältnis von Personen deutscher Nationalität zu Personen nicht deutscher Nationalität bei Fernsehbesitzern an. Bei Telefon- und Internetbesitzern muss zusätzlich die jeweilige Interaktion beachtet werden.

Die Variable Geschlecht (p-Wert=0.066558 > 0.5) und ihre Interaktionen (p-Wert=0.816122 > 0.5, p-Wert=0.097366 > 0.5) haben keinen signifikanten Einfluss und könnten deshalb aus dem Regressionsmodell gestrichen werden.

Der Dispersionsparameter  $\phi$  ist in allen vier Modellen im Intervall von 1,106 bis 1,09. Bei Annahme einer Bernoulli-Verteilung für den Response wäre  $\phi = 1$ . Somit weisen die SOEP-Daten eine Überdispersion (Fahrmeir *et al.* (2007)) auf. Die empirische Varianz ist größer, als die, durch die Bernoulli-Verteilung bestimmte Varianz.

## 3 Modelle mit zufälligen Effekten

In diesem Kapitel werden die Modelle mit zufälligen Effekten erklärt. Abschnitt 3.1 befasst sich mit dem Konzept dieser Modelle. Die Struktur und Notation der Daten wird in Abschnitt 3.2 beschrieben. Im Anschluss wird in Abschnitt 3.3 die Modellierung der Modelle mit zufälligen Effekten zuerst allgemein, in Abschnitt 3.3.1 explizit für binären Response und in Abschnitt 3.3.2 für ordinalen Response angegeben. Die Inferenz ist in Abschnitt 3.4 erläutert.

Random effects Modelle für kategorialen Response wurden in dem Büchern von Fahrmeir und Tutz (1994), Diggle *et al.* (2002) und Fahrmeir *et al.* (2007) vorgestellt. Neben diesen Büchern sind auch die Artikel von Zeger *et al.* (1988) und Stiratelli *et al.* (1984) zu erwähnen.

### 3.1 Konzept

Bei der Planung einer Studie wird festgelegt welche Merkmale von Interesse sind und welche Fragestellungen beantwortet werden sollen. Die Fragestellung, die der Studie zugrunde liegt entscheidet welche statistischen Verfahren bzw. Modelle zur Analyse verwendet werden.

Im Abschnitt 2 wurden die marginalen Modelle vorgestellt. Diese schätzen Effekte der Grundgesamtheit im Durchschnitt und die Korrelation wird als nuisance Parameter gehandhabt.

Oft liegt jedoch der Schwerpunkt des Interesses mehr bei den Individuen (Clustern, Subjekten) selbst. Die Regressionskoeffizienten sollen den individuellen Response und nicht den populationsdurchschnittlichen Response erklären. Die Abhängigkeit, die durch Messwiederholungen entsteht, wird nicht mehr durch die Kovarianz beschrieben, sondern es sollen die Ursachen der Kovarianz verwendet werden (Zeger *et al.* (1988)). Diese Ursachen sind zufällige subjektspezifische Effekte, welche die Heterogenität der Individuen widerspiegeln.

Daraus resultiert eine Alternative zu den marginalen Modellen, nämlich die Modelle mit zufälligen Effekten (random effects models). Diese Modelle werden auch als gemischte Modelle (mixed models) bezeichnet. Bei diesem Modell werden dem linearen Prädiktor neben den festen Effekten auch noch zusätzlich zufällige Effekte beigefügt.

Diese zufälligen Effekte sind cluster- bzw. subjektspezifisch und deshalb besonders geeignet bei Fragestellungen, bei denen cluster- bzw. subjektspezifisches Verhalten von primärem Interesse ist.

Werden die ALLBUS-Daten anhand von Modellen mit zufälligen Effekten analysiert, so stehen die Befragten an sich im Interessensmittelpunkt und der linearen Prädiktor wird um personenspezifische zufällige Effekte ergänzt. Für eine genauere

Analyse der ALLBUS-Daten siehe Abschnitt 4.

## 3.2 Notation

Wie die marginalen Modelle werden auch die Modelle mit zufälligen Effekten zur Auswertung von Messwiederholungsdaten bzw. longitudinalen Daten verwendet. Solche Daten sind natürlicherweise charakterisiert durch (zeitlich) wiederholte Messungen an einem Cluster (Einheit, Subjekt). Die Messung  $t$ ,  $t = 1, \dots, T_i$  am Cluster  $i$ ,  $i = 1, \dots, n$  führt zu der Beobachtung  $(y_{it}, \mathbf{x}_{it})$ , wobei  $y_{it}$  der Response und  $\mathbf{x}_{it}$  der Vektor der Kovariablen ist.

## 3.3 Modellierung

Wie bereits erwähnt beschäftigt sich diese Arbeit ausschließlich mit kategorialen Zielvariablen. Die Verwendung eines linearen Modells ist bei dieser Datenstruktur nicht sinnvoll. Deshalb wird ein generalisiertes lineares Modell mit zufälligen Effekten spezifiziert. Eine äquivalente Bezeichnung ist generalisiertes lineares gemischtes Modell (generalized linear mixed model, GLMM). Wie im GLM wird unter anderem eine Verteilungsannahme für die Dichte  $f(y_{it}|\mathbf{b}_i)$  und eine Strukturannahme für den Erwartungswert  $E(y_{it}|\mathbf{b}_i)$  modelliert.

Die Annahmen des GLMM sind:

1. Die Messungen zwischen den Clustern sind voneinander unabhängig. Die Responses eines Clusters sind wegen der Messwiederholungen im Allgemeinen nicht voneinander unabhängig. Gegeben den zufälligen Effekten sind die Messungen innerhalb eines Cluster voneinander unabhängig:

$$y_{i1}, \dots, y_{iT_i} | \mathbf{b}_i \text{ unabhängig.}$$

2. Die bedingte Dichte  $f(y_{it}|\mathbf{b}_i)$  folgt einer Exponentialfamilie:

$$f(y_{it}|\mathbf{b}_i) = \exp\left[\frac{y_{it}\theta_{it} - \psi(\theta_{it})}{\phi} + c(y_{it}, \phi)\right].$$

Der Parameter  $\theta_{it}$  wird als natürlicher oder kanonischer Parameter bezeichnet. Die Funktionen  $\psi()$  und  $c()$  sind bekannt. Der Dispersionsparameter ist  $\phi$ .

3. Der lineare Prädiktor  $\eta_{it} = \mathbf{x}_{it}'\boldsymbol{\beta}$  wird erweitert zu:

$$\eta_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{w}_{it}'\mathbf{b}_i,$$

wobei  $\boldsymbol{\beta}$  der Vektor der festen Effekte und  $\mathbf{b}_i$  der Vektor der zufälligen, clusterspezifischen Effekte ist. Der Vektor  $\mathbf{w}_{it}$  ist ein Teilvektor von dem Kovariablenvektor  $\mathbf{x}_{it}$ . Der Vektor  $\mathbf{x}_{it}$  hat die Dimension  $p \times 1$  und  $\mathbf{w}_{it}$  hat die Dimension  $q \times 1$ . Es kann dementsprechend sowohl der Intercept als auch die Steigungsparameter zufällig sein.

4. Der Erwartungswert ist bestimmt durch:

$$\mu_{it} = E(y_{it}|\mathbf{b}_i) = h(\eta_{it}) = h(\mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{w}_{it}'\mathbf{b}_i)$$

und für die Varianz gilt:

$$v_{it} = \text{var}(y_{it}|\mathbf{b}_i) = v(\mu_{it})\phi,$$

wobei  $h$  eine bekannte Responsefunktion und  $v$  eine bekannte Varianzfunktion ist.

5. Die zufälligen Effekte  $\mathbf{b}_i$ ,  $i = 1, \dots, n$  sind voneinander unabhängig und identisch verteilt entsprechend einer gemeinsamen zugrundeliegenden multivariaten Verteilung  $F$ , beziehungsweise Dichte  $p$ . Die zufälligen Effekte spiegeln die natürliche Heterogenität in der Population wieder. Wie hier spezifiziert, folgt die natürliche Heterogenität einer Wahrscheinlichkeitsverteilung.

Aus der Annahme, dass  $f(y_{it}|\mathbf{b}_i)$  einer Exponentialfamilie folgt (zweite Annahme), sind die bedingten Momente definiert durch:

$$\mu_{it} = E(y_{it}|\mathbf{b}_i) = \psi'(\theta_{it})$$

$$v_{it} = \text{var}(y_{it}|\mathbf{b}_i) = \psi''(\theta_{it})\phi.$$

Zusammengefasst bedeuten die Annahmen, dass in einem generalisierten linearen gemischten Modell die Daten für ein Cluster einem generalisierten linearen Modell folgen, mit der zusätzlichen Annahme, dass die Regressionskoeffizienten von Cluster zu Cluster variieren entsprechend einer Verteilung  $F$ .

Im multinomialen Fall ist der lineare Prädiktor definiert durch:

$$\boldsymbol{\eta}_{it} = X_{it}\boldsymbol{\beta} + W_{it}\mathbf{b}_i,$$

wobei  $X_{it}$  und  $W_{it}$  Designmatrix sind. Die Matrix  $W_{it}$  ist eine Teilmatrix von  $X_{it}$ .

Die multivariate Verteilung  $F$  (Annahme 5) lässt sich beispielsweise durch die Normalverteilung spezifizieren. Die clusterspezifischen Effekte  $\mathbf{b}_i$  sind dementsprechend

unabhängig und normalverteilt.

Der Erwartungswert und die Kovarianz der zufälligen Effekte sind:

$$E(\mathbf{b}_i) = \mathbf{0} , cov(\mathbf{b}_i) = Q, \quad (12)$$

wobei  $Q$  eine positiv definite, meist unbekannte Matrix ist.

Für eine konsistente Schätzung der Parameter muss sowohl die Verteilung der zufälligen Effekte, als auch die Linkfunktion korrekt spezifiziert sein (Zeger *et al.* (1988)).

### 3.3.1 Das binäre logistische Modell mit zufälligen Effekten

Im folgenden Abschnitt wird das Modell mit zufälligen Effekten bei binärer Zielvariable vorgestellt. Voraussetzung ist, dass die Responses zu einem Cluster, gegeben den zufälligen Effekt, voneinander unabhängig sind. Die entsprechende bedingte Dichte folgt einer Bernoulli-Verteilung, welche zur Exponentialfamilie gehört. Der Erwartungswert ist über eine Responsefunktion vom linearen Prädiktor abhängig, welcher durch feste und zufällige Effekte bestimmt ist.

Die allgemeine Formel für das Logit-Modell mit zufälligen Effekten ist gegeben durch:

$$\log \frac{P(y_{it} = 1 | \mathbf{b}_i)}{P(y_{it} = 0 | \mathbf{b}_i)} = g(P(y_{it} = 1 | \mathbf{b}_i)) = \eta_{it}$$

beziehungsweise

$$\pi_{it} = h(\eta_{it}) = P(y_{it} = 1 | \mathbf{b}_i) = \frac{\exp(\eta_{it})}{1 + \exp(\eta_{it})},$$

wobei  $\eta_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{w}_{it}'\mathbf{b}_i$ ,  $\mathbf{b}_i \sim F()$ ,  $h()$  die Linkfunktion und  $g()$  die Responsefunktion ist.

Da die zufälligen Effekte die Heterogenität der Cluster darstellen und  $\mathbf{w}_{it}$  eine Teilmenge von  $\mathbf{x}_{it}$  ist, folgt, dass sich die natürliche Heterogenität in einer Teilmenge der Regressionskoeffizienten wieder findet. Der einfachste Fall ist es die Heterogenität auf den Intercept zu beschränken. Ein solches Modell wird als Modell mit zufälligem Intercept bezeichnet. Der Intercept kann als normalverteilt angenommen werden. Bei diesem Modell ist der lineare Prädiktor definiert durch:

$$\eta_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + b_i$$

und es gilt:

$$b_i \sim N(0, \sigma^2).$$

Der zufällige Effekt  $b_i$  ist spezifisch für die Messungen am  $i$ -ten Cluster. Es kann zum Beispiel das Grundlevel an einer speziellen Person repräsentieren.

Im Allgemeinen ist es zulässig, dass  $\mathbf{x}_{it}$  die 1 enthält und somit  $\eta_{it} = \beta_0 + b_i + \tilde{\mathbf{x}}_{it}\tilde{\boldsymbol{\beta}}$ , wobei  $\tilde{\mathbf{x}}_{it}$  die 1 nicht enthält und  $\tilde{\boldsymbol{\beta}}$  der Vektor ohne  $\beta_0$  ist.

Anhand des binären logistischen Modells mit zufälligen Effekten lassen sich individuelle Chancen und Risiken berechnen.

### 3.3.2 Das Modell mit zufälligen Effekten bei ordinalem Response

In einigen Studien ist die Zielvariable nicht binär sondern mehrkategorial, genauer gesagt ordinal. Der Response des Cluster  $i$  zur Messung  $t$  ist  $y_{it}$ . Dies wird im Vergleich zum univariaten Fall als Vektor  $\mathbf{y}_{it} = (y_{it1}, \dots, y_{itk})$  dargestellt, wobei  $k$  die Anzahl der Kategorien ist. Ist die Beobachtung in der Kategorie  $j$ , so ist  $y_{itj} = 1$  und 0 sonst. Die Wahrscheinlichkeit  $\boldsymbol{\pi}_{it}$  ist dementsprechend  $(\pi_{it1}, \dots, \pi_{itq})$ ,  $q = k - 1$ . Da die Summe über alle  $k$  Wahrscheinlichkeiten gleich eins ist, ist es ausreichend  $k - 1$  Wahrscheinlichkeiten zu berechnen.

Das Modell lässt sich motivieren durch eine latente Variable  $U$  und Schwellen  $\theta$ . Die beobachtete Kategorie  $r$  ist Resultat der Lage der latenten Variable bezüglich der Schwellen:

$$y_{it} = r \Leftrightarrow \theta_{i,r-1} < U_{it} < \theta_{i,r}, \quad r = 1, \dots, q.$$

Die Wahrscheinlichkeiten im kumulativen Modell sind demnach definiert durch:

$$\pi_{it1} = F(\theta_{i1} + \mathbf{x}_{it}'\boldsymbol{\gamma})$$

$$\pi_{itr} = F(\theta_{ir} + \mathbf{x}_{it}'\boldsymbol{\gamma}) - \pi_{it,r-1},$$

wobei  $F$  eine Verteilungsfunktion und  $\boldsymbol{\gamma}$  der Vektor der Kovariableneffekte ist. Die Schwellen müssen geordnet sein:

$$-\infty = \theta_{i0} < \theta_{i1} < \dots < \theta_{iq} < \theta_{ik} = \infty.$$

Um ein Modell mit zufälligen Effekten zu erhalten, muss der lineare Prädiktor um zufällige Effekte ergänzt werden. Genau wie beim binären logistischen Modell mit zufälligen Effekten ist es die einfachste Möglichkeit nur den Intercept als zufällig zu spezifizieren.

Es resultiert wiederum das sogenannte Modell mit zufälligen Intercept.

Der lineare Prädiktor ist dann definiert durch:

$$\eta_{itr} = \underbrace{\theta_r + b_i}_{\theta_{ir}} + \mathbf{x}_{it}'\boldsymbol{\gamma}$$

mit

$$b_i \sim N(0, \sigma^2).$$

Die Schwellen werden also clusterspezifisch verändert:  $\theta_{ir} = \theta_r + b_i$ .  
Der lineare Prädiktor des Modells mit zufälligem Intercept ist bei der Schreibweise in Matrixform bestimmt durch  $\boldsymbol{\eta}_{it} = X_{it}\boldsymbol{\beta} + \mathbf{W}_{it}b_i$  wobei

$$X_{it} = \begin{pmatrix} 1 & & & \mathbf{x}_{it}' \\ & \ddots & & \vdots \\ & & 1 & \mathbf{x}_{it}' \end{pmatrix}, \mathbf{W}_{it} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \text{ und } \boldsymbol{\beta} = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_q \\ \boldsymbol{\gamma} \end{pmatrix}.$$

Die latente Variable selbst ist nicht mehr im Modell enthalten.  
Ist die Ordnung der Schwellen verletzt, so können numerische Probleme bei der Schätzung auftreten. Um diese zu verhindern werden die Schwellen reparametrisiert:

$$\alpha_{i1} = \theta_{i1}$$

$$\alpha_{ij} = \log(\theta_{ij} - \theta_{i,j-1}), j = 2, \dots, q$$

und somit  $\boldsymbol{\alpha}_i = (\alpha_{i1}, \dots, \alpha_{iq})$ . Der Erwartungswert  $E(\boldsymbol{\alpha}_i)$  sei  $\tilde{\boldsymbol{\alpha}}$  und die Kovarianzmatrix  $cov(\boldsymbol{\alpha}_i)$  sei  $Q$ . Die Wahrscheinlichkeiten im kumulativen Modell mit Verteilungsfunktion  $F$  sind dann definiert durch:

$$\pi_{it1} = F(\alpha_{i1} + \mathbf{x}_{it}'\boldsymbol{\gamma}),$$

$$\pi_{itj} = F(\alpha_{i1} + \sum_{m=2}^j \exp(\alpha_{im}) + \mathbf{x}_{it}'\boldsymbol{\gamma}) - \pi_{it,j-1}, j = 2, \dots, q.$$

Für die reparametrisierten Schwellen gilt:

$$\boldsymbol{\alpha}_i = \tilde{\boldsymbol{\alpha}} + \mathbf{b}_i, \mathbf{b}_i \sim N(\mathbf{0}, Q) \text{ und } \boldsymbol{\beta} = (\tilde{\boldsymbol{\alpha}}, \boldsymbol{\gamma})'.$$

Bei einem Modell mit zufälligem Intercept und Verwendung der reparametrisierten Schwellen sind die Design-Matrizen des kumulativen logistischen Modells:

$$X_{it} = \begin{pmatrix} 1 & & & 0 & \mathbf{x}_{it}' \\ & \ddots & & & 0 \\ & & \ddots & & \vdots \\ 0 & & & 1 & 0 \end{pmatrix} \text{ und } W_{it} = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix}.$$

### 3.4 Inferenz

In dem Modell mit zufälligen Effekten sind die festen und zufälligen Effekte unbekannt. Infolgedessen müssen diese geschätzt werden. In Fahrmeir und Tutz (1994) werden zwei Ansätze zur Inferenz in Modellen mit zufälligen Effekten vorgestellt.

Zum einen eine Schätzung, die auf der Maximierung der Posteriori Verteilung ohne Integration basiert und zum anderen ein zweistufiger Schätzalgorithmus bei dem Integrationsverfahren verwendet werden.

Die Methode, basierend auf dem Posteriori-Modus, kann auch als penalisierte Likelihood Schätzung betrachtet werden. Diese wird ebenfalls kurz in Fahrmeir *et al.* (2003) und Diggle *et al.* (2002) vorgestellt.

Stiratelli *et al.* (1984) erläutern den Posteriori-Modus-Schätzer für binären Response.

### 3.4.1 Posteriori-Modus-Schätzer

Zu schätzen ist der Parametervektor  $\beta$ , die zufälligen Effekte  $\mathbf{b}_i$ ,  $i = 1, \dots, n$  und eventuell die Kovarianz  $Q$  der zufälligen Effekte. Es wird ein Vektor eingeführt, der die festen und zufälligen Effekte enthält:

$$\boldsymbol{\delta}' = (\beta', \mathbf{b}_1', \dots, \mathbf{b}_n').$$

Die Schätzung der unbekanntenen Größen soll gleichzeitig durch die Maximierung der Posteriori erfolgen.

Die Varianz-Kovarianz-Komponenten  $Q$  können bekannt oder unbekannt sein. Die Inferenz ist dementsprechend unterschiedlich. Als Erstes wird die Inferenz bei bekanntem  $Q$  geschildert und im Anschluss daran die Inferenz bei unbekanntem  $Q$ .

#### Bekannte Varianz-Kovarianz-Komponenten

In Abschnitt 3.3, Annahme 1 würde die Unabhängigkeit zwischen den Clustern und die bedingte Unabhängigkeit innerhalb der Cluster festgelegt.

Die bedingte Dichte  $f(Y|B; \beta)$ , mit  $Y = (\mathbf{y}_1, \dots, \mathbf{y}_n)$  und  $B = (\mathbf{b}_1, \dots, \mathbf{b}_n)$  kann somit als das Produkt über die einzelnen Dichten  $f(y_{it}|\mathbf{b}_i; \beta)$  formuliert werden:

$$f(Y|B; \beta) = \prod_{i=1}^n f(\mathbf{y}_i|\mathbf{b}_i; \beta) = \prod_{i=1}^n \prod_{t=1}^{T_i} f(y_{it}|\mathbf{b}_i; \beta). \quad (13)$$

Die uninformative Priori Dichte von  $\boldsymbol{\delta}$  ist proportional zum Produkt über die Priori Dichten von  $\mathbf{b}_i$ :

$$p(\boldsymbol{\delta}; Q) \propto \prod_{i=1}^n p(\mathbf{b}_i; Q) \quad (14)$$

Unter Verwendung des Satz von Bayes, der Proportionalität (14) und der Dichte (13) ist die Posteriori gegeben durch:

$$f(\boldsymbol{\delta}|Y; Q) = \frac{\prod_{i=1}^n f(\mathbf{y}_i|\mathbf{b}_i, \beta) \prod_{i=1}^n p(\mathbf{b}_i; Q)}{\int \prod_{i=1}^n f(\mathbf{y}_i|\mathbf{b}_i, \beta) p(\mathbf{b}_i; Q) d\mathbf{b}_1, \dots, d\mathbf{b}_n d\beta}. \quad (15)$$

Anhand der Posteriori gegeben den Daten, soll ein Schätzer für  $\delta$  gefunden werden.

Ein möglicher Punktschätzer ist der Posteriori-Mittelwert. Dieser kann approximiert werden durch den Posteriori-Modus. Durch die Nullstelle der ersten Ableitung der Posteriori erhält man den Posteriori-Modus, durch die zweite Ableitung die Posteriori-Krümmung.

Wird die Integration im Nenner der Posteriori bezüglich der unbekanntenen Größen, durchgeführt, so ist dieser nicht mehr von  $\beta$  und den zufälligen Effekten  $\mathbf{b}_i$  abhängig. Der resultierende Nenner ist also fest und kann als ein Faktor betrachtet werden. Es gilt somit folgende Proportionalität für die Posteriori:

$$f(\delta|Y; Q) \propto \prod_{i=1}^n f(\mathbf{y}_i|\mathbf{b}_i, \beta) \prod_{i=1}^n p(\mathbf{b}_i; Q).$$

Für die Inferenz ist es äquivalent statt der Posteriori die logarithmierte Posteriori zu verwenden. Diese ist:

$$\sum_{i=1}^n \log f(\mathbf{y}_i|\mathbf{b}_i, \beta) + \sum_{i=1}^n \log p(\mathbf{b}_i; Q).$$

Wird die Dichte der zufälligen Effekte  $p(\mathbf{b}_i; Q)$  als Normalverteilung angenommen so folgt:

$$p(\mathbf{b}_i; Q) = (2\pi)^{-q/2} |Q|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{b}_i - E(\mathbf{b}_i))' Q^{-1} (\mathbf{b}_i - E(\mathbf{b}_i))\right\}, \mathbf{b}_i \in \mathbb{R}^q$$

Da der Erwartungswert der zufälligen Effekte gleich Null ist (12) und bei einer Ableitung nach  $\mathbf{b}_i$ , die von  $\mathbf{b}_i$  unabhängigen Terme gleich Null werden, können diese bei der Formulierung der Dichte der normalverteilten zufälligen Effekte ignoriert werden. Für die Log-Posteriori mit normalverteilten zufälligen Effekten folgt dann:

$$l(\delta) = \sum_{i=1}^n \log f(\mathbf{y}_i|\mathbf{b}_i, \beta) - \frac{1}{2} \sum_{i=1}^n \mathbf{b}_i' Q^{-1} \mathbf{b}_i.$$

Die Dichte  $f(\mathbf{y}_i|\mathbf{b}_i, \beta)$  entspricht der Likelihood und folglich  $\log f(\mathbf{y}_i|\mathbf{b}_i, \beta)$  der Log-Likelihood. Der Term  $-\frac{1}{2} \sum_{i=1}^n \mathbf{b}_i' Q^{-1} \mathbf{b}_i$  kann als Penalisierung auf die Log-Likelihood interpretiert werden. Die Posteriori-Modus-Schätzung wird deshalb in verschiedener Literatur auch als penalisierte Likelihood-Schätzung bezeichnet.

Zur Maximierung der penalisierten Log-Likelihood  $l(\delta)$  werden die Score-Funktionen bezüglich dem festen Effekt  $\beta$  und den zufälligen Effekten  $\mathbf{b}_i$  bestimmt:

$$\mathbf{s}_\beta = \frac{\partial l(\delta)}{\partial \beta} = \sum_{i=1}^n \sum_{t=1}^{T_i} X'_{it} D_{it}(\delta) \Sigma_{it}^{-1}(\delta) (\mathbf{y}_{it} - \boldsymbol{\mu}_{it}(\delta)),$$

$$\mathbf{s}_i = \frac{\partial l(\boldsymbol{\delta})}{\partial \mathbf{b}_i} = \sum_{t=1}^{T_i} W'_{it} D_{it}(\boldsymbol{\delta}) \Sigma_{it}^{-1}(\boldsymbol{\delta}) (\mathbf{y}_{it} - \boldsymbol{\mu}_{it}(\boldsymbol{\delta})) - Q^{-1} \mathbf{b}_i,$$

wobei  $D_{it}(\boldsymbol{\delta}) = \partial h(\boldsymbol{\eta}_{it}) / \partial \boldsymbol{\eta}$ ,  $\Sigma_{it}(\boldsymbol{\delta}) = \text{cov}(\mathbf{y}_{it} | \boldsymbol{\beta}, \mathbf{b}_i)$ ,  $\boldsymbol{\mu}_{it}(\boldsymbol{\delta}) = h(\boldsymbol{\eta}_{it})$  und  $i = 1 \dots, n$ . Bei binärem Response wird der Designvektor  $\mathbf{x}_{it}$  und der Skalar  $y_{it}$  anstelle der Designmatrix  $X_{it}$  und dem Vektor  $\mathbf{y}_{it}$  verwendet.

Die Komponenten  $\mathbf{s}_\beta$  und  $\mathbf{s}_i$  ergeben die Score-Funktion von  $\boldsymbol{\delta}$ :

$$\mathbf{s}(\boldsymbol{\delta}) = \frac{\partial l(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}} = (\mathbf{s}_\beta, \mathbf{s}_1, \dots, \mathbf{s}_n).$$

Der penalisierte Maximum-Likelihood- bzw. Posteriori-Modus-Schätzer  $\hat{\boldsymbol{\delta}}$  wird bestimmt durch:

$$\mathbf{s}(\boldsymbol{\delta}) = \mathbf{0}.$$

Die Berechnung erfolgt iterativ über den Fisher-Scoring-Algorithmus. Für dieses Verfahren wird die erwartete bedingte Fisher-Informationsmatrix  $F(\boldsymbol{\delta}) = \text{cov} \mathbf{s}(\boldsymbol{\delta})$  benötigt. Diese wird errechnet aus der erwarteten negativen zweiten Ableitung von  $l(\boldsymbol{\delta})$ . Diese Matrix  $F(\boldsymbol{\delta})$  ist folgendermaßen partitioniert:

$$F(\boldsymbol{\delta}) = \begin{pmatrix} F_{\beta\beta} & F_{\beta 1} & F_{\beta 2} & \dots & F_{\beta n} \\ F_{1\beta} & F_{11} & & & 0 \\ F_{2\beta} & & F_{22} & & \\ \vdots & & & \ddots & \\ F_{n\beta} & 0 & & & F_{nn} \end{pmatrix},$$

wobei

$$F_{\beta\beta} = -E \left( \frac{\partial^2 l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} \right) = \sum_{i=1}^n \sum_{t=1}^{T_i} X'_{it} D_{it}(\boldsymbol{\delta}) \Sigma_{it}^{-1}(\boldsymbol{\delta}) D'_{it}(\boldsymbol{\delta}) X_{it},$$

$$F_{\beta i} = F'_{i\beta} = -E \left( \frac{\partial^2 l(\boldsymbol{\delta})}{\partial \boldsymbol{\beta} \partial \mathbf{b}_i'} \right) = \sum_{t=1}^{T_i} X'_{it} D_{it}(\boldsymbol{\delta}) \Sigma_{it}^{-1}(\boldsymbol{\delta}) D'_{it}(\boldsymbol{\delta}) W_{it},$$

$$F_{ii} = -E \left( \frac{\partial^2 l(\boldsymbol{\delta})}{\partial \mathbf{b}_i \partial \mathbf{b}_i'} \right) = \sum_{t=1}^{T_i} W'_{it} D_{it}(\boldsymbol{\delta}) \Sigma_{it}^{-1}(\boldsymbol{\delta}) D'_{it}(\boldsymbol{\delta}) W_{it} + Q^{-1}.$$

Auffällig ist die Struktur der erwarteten Fisher-Information. Der rechte untere Teil der Matrix  $F(\boldsymbol{\delta})$  ist eine Block-Diagonal-Matrix.

Die Iterationsvorschrift des Fisher-Scoring-Algorithmus lautet:

$$\boldsymbol{\delta}^{(k+1)} = \boldsymbol{\delta}^{(k)} + F^{-1}(\boldsymbol{\delta}^{(k)}) \mathbf{s}(\boldsymbol{\delta}^{(k)}), \quad (16)$$

wobei  $k$  den Iterationsschritt angibt.

Die Dimension von  $F(\boldsymbol{\delta})$  kann zu Problemen bei der Berechnung führen. Der Algorithmus wird deshalb unter Nutzung der Blockstruktur und der partitionierten Struktur folgendermaßen umstrukturiert:

$$F_{\beta\beta}^{(k)} \Delta\boldsymbol{\beta}^{(k)} + \sum_{i=1}^n F_{\beta i}^{(k)} \Delta\mathbf{b}_i^{(k)} = \mathbf{s}_\beta^{(k)}, \quad (17)$$

$$F_{i\beta}^{(k)} \Delta\boldsymbol{\beta}^{(k)} + F_{ii}^{(k)} \Delta\mathbf{b}_i^{(k)} = \mathbf{s}_i^{(k)}, \quad (18)$$

wobei  $i = 1, \dots, n$ ,  $\Delta\boldsymbol{\beta}^{(k)} = \boldsymbol{\beta}^{(k+1)} - \boldsymbol{\beta}^{(k)}$  und  $\Delta\mathbf{b}_i^{(k)} = \mathbf{b}_i^{(k+1)} - \mathbf{b}_i^{(k)}$ . Es ist wünschenswert, den Algorithmus als Gleichung von  $\Delta\boldsymbol{\beta}^{(k)}$  und  $\Delta\mathbf{b}_i^{(k)}$  zu formulieren. Deshalb wird (18) nach  $\Delta\mathbf{b}_i^{(k)}$  aufgelöst:

$$\begin{aligned} F_{ii}^{(k)} \Delta\mathbf{b}_i^{(k)} &= \mathbf{s}_i^{(k)} - F_{i\beta}^{(k)} \Delta\boldsymbol{\beta}^{(k)} \\ \Leftrightarrow \Delta\mathbf{b}_i^{(k)} &= (F_{ii}^{(k)})^{-1} \{ \mathbf{s}_i^{(k)} - F_{i\beta}^{(k)} \Delta\boldsymbol{\beta}^{(k)} \}, \end{aligned} \quad (19)$$

und (17) nach  $\Delta\boldsymbol{\beta}^{(k)}$  auflösen:

$$\begin{aligned} \Delta\boldsymbol{\beta}^{(k)} &= (F_{\beta\beta}^{(k)})^{-1} \{ \mathbf{s}_\beta^{(k)} - \sum_{i=1}^n F_{\beta i}^{(k)} \Delta\mathbf{b}_i^{(k)} \} \\ \Leftrightarrow \Delta\boldsymbol{\beta}^{(k)} &= (F_{\beta\beta}^{(k)})^{-1} \{ \mathbf{s}_\beta^{(k)} - \sum_{i=1}^n F_{\beta i}^{(k)} [(F_{ii}^{(k)})^{-1} \{ \mathbf{s}_i^{(k)} - F_{i\beta}^{(k)} \Delta\boldsymbol{\beta}^{(k)} \}] \} \Leftrightarrow \\ \Delta\boldsymbol{\beta}^{(k)} - (F_{\beta\beta}^{(k)})^{-1} \sum_{i=1}^n F_{\beta i}^{(k)} (F_{ii}^{(k)})^{-1} F_{i\beta}^{(k)} \Delta\boldsymbol{\beta}^{(k)} &= (F_{\beta\beta}^{(k)})^{-1} \{ \mathbf{s}_\beta^{(k)} - \sum_{i=1}^n F_{\beta i}^{(k)} (F_{ii}^{(k)})^{-1} \mathbf{s}_i^{(k)} \} \\ \Leftrightarrow \Delta\boldsymbol{\beta}^{(k)} &= \{ F_{\beta\beta}^{(k)} - \sum_{i=1}^n F_{\beta i}^{(k)} (F_{ii}^{(k)})^{-1} F_{i\beta}^{(k)} \}^{-1} \{ \mathbf{s}_\beta^{(k)} - \sum_{i=1}^n F_{\beta i}^{(k)} (F_{ii}^{(k)})^{-1} \mathbf{s}_i^{(k)} \} \end{aligned} \quad (20)$$

Nach diesen Umformungen besteht der Algorithmus aus den Iterationsschritten (20) und (19).

Der durch den Algorithmus ermittelte Schätzer  $\hat{\boldsymbol{\delta}}' = (\hat{\boldsymbol{\beta}}', \hat{\mathbf{b}}_1', \dots, \hat{\mathbf{b}}_n')$  ist für eine genügend große Anzahl an Messwiederholungen  $T_i$  approximativ normalverteilt:

$$\hat{\boldsymbol{\delta}} \sim N(\boldsymbol{\delta}, F^{-1}(\boldsymbol{\delta})).$$

Der Posteriori-Modus und  $F^{-1}(\boldsymbol{\delta})$  sind gute Approximationen für den Posteriori-Mittelwert und Kovarianzmatrix. Die approximative Kovarianzmatrix ist partitioniert in:

$$F^{-1}(\boldsymbol{\delta}) = \begin{pmatrix} V_{\beta\beta} & V_{\beta 1} & V_{\beta 2} & \dots & V_{\beta n} \\ V_{1\beta} & V_{11} & V_{12} & \dots & V_{1n} \\ V_{2\beta} & V_{21} & V_{22} & \dots & V_{2n} \\ \vdots & \vdots & & & \vdots \\ V_{n\beta} & V_{n1} & V_{n2} & \dots & V_{nn} \end{pmatrix},$$

wobei

$$\begin{aligned}
V_{\beta\beta} &= (F_{\beta\beta} - \sum_{i=1}^n F_{\beta i} F_{ii}^{-1} F_{i\beta})^{-1}, \\
V_{ii} &= F_{ii}^{-1} + F_{ii}^{-1} F_{i\beta} V_{\beta\beta} F_{\beta i} F_{ii}^{-1}, \\
V_{\beta i} &= V_{i\beta}' = -V_{\beta\beta} F_{\beta i} F_{ii}^{-1}, \\
V_{ij} &= V_{ji}' = F_{ii}^{-1} F_{i\beta} V_{\beta\beta} F_{\beta j} F_{jj}^{-1}, i \neq j.
\end{aligned}$$

### Unbekannte Varianz-Kovarianz-Komponenten

Bisher wurde angenommen, dass die Komponenten von  $Q$  bekannt sind. In diesem zweiten Abschnitt zum Posteriori-Modus-Schätzer wird von unbekanntem  $Q$  ausgegangen.

Die Basis zur Bestimmung von  $Q$  ist die marginale Log-Likelihood:

$$l(Q) = \log \int \prod_{i=1}^n f(\mathbf{y}_i | \beta, \mathbf{b}_i) p(\mathbf{b}_i; Q) d\mathbf{b}_1, \dots, d\mathbf{b}_n d\beta.$$

Den Schätzer für  $Q$  erhält man durch die Maximierung der marginalen Log-Likelihood. Auf Grund der verschachtelten Integralstruktur benötigt man dazu ein spezielles Verfahren. Das Verfahren zur Maximierung von  $l(Q)$  ist der sogenannte EM-type Algorithmus. Dieser leitet sich vom EM-Algorithmus (McLachlan und Krishnan (1997)) ab. EM steht für die zwei Schritte, die innerhalb eines Iterationsschritts ausgeführt werden: Expectation und Maximizing.

Im Erwartungswertschritt wird  $M(Q|Q^{(p)})$  berechnet und im Maximierungsschritt wird durch die Maximierung von  $M(Q|Q^{(p)})$  das aktuelle  $Q^{(p+1)}$  berechnet.

Dabei ist  $M(Q|Q^{(p)})$  definiert durch:

$$M(Q|Q^{(p)}) = E\{\log p(\boldsymbol{\delta}; Q) | Y; Q^{(p)}\} = \int \log p(\boldsymbol{\delta}; Q) f(\boldsymbol{\delta} | Y; Q^{(p)}) d\boldsymbol{\delta}.$$

Die Posteriori Dichte  $f(\boldsymbol{\delta} | Y; Q^{(p)})$  (15) wird an  $Q^{(p)}$  berechnet. Die Dichte der zufälligen Effekte  $p(\mathbf{b}_i; Q)$  kann beispielsweise als Normalverteilung angenommen werden.

Die Bestimmung von  $Q^{(p+1)}$  im Maximierungsschritt, vereinfacht sich mit der Normalverteilungsannahme für  $p(\mathbf{b}_i; Q)$  zu:

$$Q^{(p+1)} = \frac{1}{n} \sum_{i=1}^n (\text{cov}(\mathbf{b}_i | \mathbf{y}_i; Q^{(p)}) + E(\mathbf{b}_i | \mathbf{y}_i; Q^{(p)}) E(\mathbf{b}_i | \mathbf{y}_i; Q^{(p)})'). \quad (21)$$

Ist  $Q^{(p+1)}$  durch (21) geschätzt, so kann wie im Fall mit bekannten Varianz-Kovarianz-Komponenten der Posteriori-Modus und die Posteriori-Krümmung durch

den Fisher-Scoring-Algorithmus (16) bestimmt werden. Der EM-type Algorithmus integriert den Fisher-Scoring-Algorithmus im Erwartungswertschritt. Dieser Algorithmus iteriert wie folgt:

1. Startwert  $Q^{(0)}$  wählen.

Für  $p=0,1,2,\dots$

2. Posteriori-Modus-Schätzer  $\hat{\boldsymbol{\delta}}^{(p)}$  und Posteriori-Krümmung  $\hat{V}_{ii}^{(p)}$  werden durch den Fisher-Scoring-Algorithmus (16) bestimmt, wobei die Kovarianz-Komponenten durch die aktuelle Schätzung  $Q^{(p)}$  ersetzt werden.
3. EM-step: Berechne  $Q^{(p+1)}$ :

$$Q^{(p+1)} = \frac{1}{n} \sum_{i=1}^n (\hat{V}_{ii}^{(p)} + \hat{\mathbf{b}}_i^{(p)} (\hat{\mathbf{b}}_i^{(p)})').$$

### 3.4.2 Schätzung basierend auf Integrationstechniken

Der Posteriori-Modus-Schätzer (Abschnitt 3.4.1) kommt ohne Integration aus. Deshalb ist dieser Schätzer besonders von Bedeutung bei hochdimensionalen zufälligen Effekten, da die Integration über alle diese Effekte zu Problemen bei der Berechnung führen kann.

Eine Alternative zum Posteriori-Modus-Schätzer bietet ein zweistufiges Schätzverfahren bei dem Integration verwendet wird.

Die erste Stufe ist die Berechnung der festen Effekte  $\boldsymbol{\beta}$  und der Varianz-Kovarianz-Komponenten  $Q$  mittels Maximum-Likelihood-Schätzung und in der zweiten Stufe wird die Posteriori-Erwartungswert-Schätzung für die zufälligen Effekte  $\mathbf{b}_1, \dots, \mathbf{b}_n$  durchgeführt.

Dieses Verfahren wird in Fahrmeir und Tutz (1994) detailliert beschrieben.

Des Weiteren wird der Dispersionsparameter  $\phi$  als bekannt angenommen.

#### 3.4.2.1 Maximum-Likelihood-Schätzung für $\boldsymbol{\beta}$ und $Q$

Die marginale Log-Likelihood der festen Effekte  $\boldsymbol{\beta}$  und  $Q$  ist definiert durch:

$$l(\boldsymbol{\beta}, Q) = \sum_{i=1}^n \log \int f(\mathbf{y}_i | \mathbf{b}_i; \boldsymbol{\beta}) p(\mathbf{b}_i; Q) d\mathbf{b}_i.$$

Die unbekannt Parameter werden durch die Maximierung der marginalen Log-Likelihood geschätzt.

Ist das Integral hochdimensional, so ist eine analytische Berechnung in den meisten

Fällen nicht möglich. Abhilfe schafft die Gauß-Hermite Integration oder die Monte Carlo Integration.

Für diese Techniken wird die reparametrisierte Log-Likelihood verwendet. Diese basiert auf der Reparametrisierung der zufälligen Effekte unter Verwendung der Cholesy-Zerlegung.

Die Cholesky-Zerlegung besagt, dass sich jede symmetrische positiv definite Matrix  $Q$  eindeutig darstellen lässt als:

$$Q = Q^{1/2}Q^{T/2}.$$

Dabei besitzt  $Q^{1/2}$  die Gestalt einer unteren Dreiecksmatrix und hat positive Diagonalelemente.

Die Dichte der zufälligen Effekte  $p(\mathbf{b}_i; Q)$  sei symmetrisch um den Mittelwert. Die Reparametrisierung der zufälligen Effekte ist gegeben durch:

$$\mathbf{b}_i = Q^{1/2}\mathbf{a}_i,$$

wobei  $\mathbf{a}_i$  ein standardisierter Zufallsvektor mit Mittelwert Null und der Identitätsmatrix als Kovarianzmatrix ist. Der lineare Prädiktor in  $\boldsymbol{\beta}$  und in  $Q^{1/2}$  ist:

$$\boldsymbol{\eta}_{it} = X_{it}\boldsymbol{\beta} + W_{it}Q^{1/2}\mathbf{a}_i$$

und kann mit dem Kronecker Produkt geschrieben werden als:

$$\boldsymbol{\eta}_{it} = [X_{it}, \mathbf{a}_i' \otimes W_{it}] \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\theta} \end{bmatrix}, \quad (22)$$

wobei  $\boldsymbol{\theta}$  die Vektorisierung der Matrix  $Q^{1/2}$  ist.

Ist der zufällige Effekt  $\mathbf{a}_i$  kein Vektor sondern nur ein Skalar, so vereinfacht sich das Kronecker Produkt zu  $a_i W_{it}$ .

Der lineare Prädiktor (22) legt die Definition eines neuen Parametervektors  $\boldsymbol{\alpha}$  nahe:

$$\boldsymbol{\alpha}' = (\boldsymbol{\beta}', \boldsymbol{\theta}'). \quad (23)$$

Die reparametrisierte marginale Log-Likelihood in Abhängigkeit des Parameters  $\boldsymbol{\alpha}$  ist gegeben durch:

$$l(\boldsymbol{\alpha}) = \sum_{i=1}^n \log L_i(\boldsymbol{\alpha}) = \sum_{i=1}^n \log \int f(\mathbf{y}_i | \mathbf{a}_i; \boldsymbol{\alpha}) g(\mathbf{a}_i) d\mathbf{a}_i, \quad (24)$$

wobei die Dichte  $g$  den Erwartungswert Null und die Identitätsmatrix als Kovarianzmatrix hat. Wird die Dichte  $g$  spezifiziert, so führt die Maximierung von  $l(\boldsymbol{\alpha})$  bezüglich  $\boldsymbol{\alpha}$  zu den Maximum-Likelihood-Schätzern für  $\boldsymbol{\beta}$  und  $Q^{1/2}$ .

Die Maximierung der Log-Likelihood bzw. der reparametrisierten Log-Likelihood kann durch direkte oder indirekte Methoden vorgenommen werden. Direkte Methoden sind die Gauß-Hermite Integration (Hildebrand (1956)) und die Monte Carlo Integration (Terejanu (2009)). Eine indirekte Methode ist der EM-Algorithmus. Die beiden Integrationsstechniken arbeiten in Kombination mit dem Fisher-Scoring-Algorithmus.

### Direkte Maximierung

Die direkte Maximierung soll im Folgenden beschrieben werden. Der Maximum-Likelihood-Schätzer wird im Allgemeinen durch die Nullstelle der Scorefunktion ermittelt. Hier soll für die ML-Schätzung die reparametrisierte Log-Likelihood  $l(\boldsymbol{\alpha})$  (24) maximiert werden. Die Scorefunktion ist die Ableitung dieser Log-Likelihood bezüglich dem Parametervektor  $\boldsymbol{\alpha}$ :

$$\mathbf{s}(\boldsymbol{\alpha}) = \frac{\partial l(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} = \sum_{i=1}^n \frac{\partial L_i(\boldsymbol{\alpha}) / \partial \boldsymbol{\alpha}}{L_i(\boldsymbol{\alpha})} = \mathbf{0}.$$

Da das Integral in der Likelihood  $L_i$  analytisch nicht mehr lösbar sein muss, bedarf es Integrationsverfahren.

Wird die Dichte  $g$  als Normalverteilung spezifiziert, so kann die Integration in der Likelihood  $L_i$  durch die Gauß-Hermite-Quadratur durchgeführt werden. Dieses Integrationsverfahren ist nur für niedrigdimensionale Integrale durchführbar. Für die Erläuterung des Gauß-Hermite-Verfahrens wird sich deshalb auf ein eindimensionales Integral beschränkt. Der zufällige Effekt ist dementsprechend ein Skalar und  $\boldsymbol{\theta}$  eine einzige Varianz Komponente.

Die Gauß-Hermite Approximation der Likelihood  $L_i$  ist:

$$L_i(\boldsymbol{\alpha}) \approx L_i^{gh}(\boldsymbol{\alpha}) = \sum_{j=1}^m \mathbf{v}_j f(\mathbf{y}_i | \mathbf{d}_j; \boldsymbol{\alpha}).$$

Die Koeffizienten  $\mathbf{v}_j$  sind Gewichte, in welche  $g(\mathbf{a}_i)$  eingeht (Hildebrand (1956)).

Die Punkte  $\mathbf{d}_j$ ,  $j = 1 \dots, m$  werden als Quadraturpunkte bezeichnet.

Die approximierte Log-Likelihood ergibt sich dementsprechend:

$$l^{gh}(\boldsymbol{\alpha}) = \sum_{i=1}^n \log \sum_{j=1}^m \mathbf{v}_j f(\mathbf{y}_i | \mathbf{d}_j; \boldsymbol{\alpha}). \quad (25)$$

Die Ableitung von (25) ist die durch Gauß-Hermite approximierte Scorefunktion:

$$\mathbf{s}(\boldsymbol{\alpha}) \approx \mathbf{s}^{gh}(\boldsymbol{\alpha}) = \sum_{i=1}^n \sum_{j=1}^m \frac{1}{\underbrace{\sum_{k=1}^m \mathbf{v}_k f(\mathbf{y}_i | \mathbf{d}_k; \boldsymbol{\alpha})}_{c_{ij}^{gh}(\boldsymbol{\alpha})}} \mathbf{v}_j f(\mathbf{y}_i | \mathbf{d}_j; \boldsymbol{\alpha}) \frac{\partial \log f(\mathbf{y}_i | \mathbf{d}_j; \boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}, \quad (26)$$

wobei  $c_{ij}^{gh}(\boldsymbol{\alpha})$  die Gewichtung ist, für die gelten muss:

$$\sum_{j=1}^m c_{ij}^{gh}(\boldsymbol{\alpha}) = 1.$$

Die Gewichtung ist von dem unbekanntem Parameter  $\boldsymbol{\alpha}$  abhängig. Da  $\boldsymbol{\alpha}$  ein Vektor zweier unbekannter Parameter (23) ist, folgt, dass die Ableitung  $\partial \log f(\mathbf{y}_i | d_j; \boldsymbol{\alpha}) / \partial \boldsymbol{\alpha}'$  in der approximierten Scorefunktion (26) aus zwei Komponenten besteht. Zum einen die Ableitung nach dem Parameter  $\boldsymbol{\beta}$  und zum anderen die Ableitung nach dem Parameter  $\boldsymbol{\theta}$ :

$$\begin{aligned} \frac{\partial \log f(\mathbf{y}_i | d_j; \boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}'} &= \left( \frac{\partial \log f(\mathbf{y}_i | d_j; \boldsymbol{\alpha})}{\partial \boldsymbol{\beta}'}, \frac{\partial \log f(\mathbf{y}_i | d_j; \boldsymbol{\alpha})}{\partial \boldsymbol{\theta}} \right), \\ \frac{\partial \log f(\mathbf{y}_i | d_j; \boldsymbol{\alpha})}{\partial \boldsymbol{\beta}'} &= \sum_{t=1}^{T_i} X_{it}' D_{it}(\boldsymbol{\alpha}, \mathbf{d}_j) \Sigma_{it}^{-1}(\boldsymbol{\alpha}, \mathbf{d}_j) (\mathbf{y}_{it} - \boldsymbol{\mu}_{it}(\boldsymbol{\alpha}, \mathbf{d}_j)), \\ \frac{\partial \log f(\mathbf{y}_i | d_j; \boldsymbol{\alpha})}{\partial \boldsymbol{\theta}} &= \sum_{t=1}^{T_i} \mathbf{d}_j W_{it}' D_{it}(\boldsymbol{\alpha}, \mathbf{d}_j) \Sigma_{it}^{-1}(\boldsymbol{\alpha}, \mathbf{d}_j) (\mathbf{y}_{it} - \boldsymbol{\mu}_{it}(\boldsymbol{\alpha}, \mathbf{d}_j)), \end{aligned}$$

wobei  $D_{it}(\boldsymbol{\alpha}, \mathbf{d}_j) = \partial h(\boldsymbol{\eta}_{itj}) / \eta$ ,  $\Sigma_{it}(\boldsymbol{\alpha}, \mathbf{d}_j) = \text{cov}(\mathbf{y}_{it} | d_j)$ ,  $\boldsymbol{\mu}_{it}(\boldsymbol{\alpha}, \mathbf{d}_j) = h(\boldsymbol{\eta}_{itj})$  und  $\boldsymbol{\eta}_{itj} = X_{it} \boldsymbol{\beta} + \mathbf{d}_j * W_{it} \boldsymbol{\theta}$ .

Eine kleine Anzahl an Knoten  $m$  hat den Vorteil, dass der numerische Aufwand gering gehalten wird. Für  $n$  und  $m$  gegen unendlich ist der Maximum-Likelihood-Schätzer für  $\boldsymbol{\alpha}$  konsistent und asymptotisch normalverteilt.

Die Gauß-Hermite-Quadratur wird in der Anwendung für univariate zufällige Effekte bzw. niedrigdimensionale zufällige Effekte verwendet. Bei multivariaten bzw. hochdimensionalen zufälligen Effekten wird das Monte-Carlo-Integrationsverfahren angewandt.

Die Monte-Carlo Approximation der Likelihood  $L_i$  ist:

$$L_i(\boldsymbol{\alpha}) \approx L_i^{mc}(\boldsymbol{\alpha}) = \frac{1}{m} \sum_{j=1}^m f(\mathbf{y}_i | d_{ij}; \boldsymbol{\alpha}),$$

wobei  $d_{i1}, \dots, d_{im}$  unabhängig und identisch verteilte Zufallsvariablen der Verteilung  $g$  sind.

Die durch das Monte-Carlo-Integrationsverfahren approximierten Scorefunktion ist:

$$\mathbf{s}(\boldsymbol{\alpha}) \approx \mathbf{s}^{mc}(\boldsymbol{\alpha}) = \sum_{i=1}^n \sum_{j=1}^m \frac{1}{\underbrace{\sum_{k=1}^m f(\mathbf{y}_i | d_{ik}; \boldsymbol{\alpha})}_{c_{ij}^{mc}(\boldsymbol{\alpha})}} f(\mathbf{y}_i | d_{ij}; \boldsymbol{\alpha}) \frac{\partial \log f(\mathbf{y}_i | d_{ij}; \boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}},$$

wobei die Gewichte  $c_{ij}^{mc}$  zu Eins addieren:  $\sum_{j=1}^m c_{ij}^{mc}(\boldsymbol{\alpha}) = 1$ .

Das Weitere ist analog zur Gauß-Hermite Approximation.

Der Maximum-Likelihood-Schätzer für  $\boldsymbol{\alpha}$ , welcher sich aus der Nullstelle der Scorefunktion  $\boldsymbol{s}^{gh}$  bzw.  $\boldsymbol{s}^{mc}$  ergibt, kann durch verschiedene iterative Verfahren ermittelt werden. Geeignete iterative Verfahren sind zum Beispiel der Fisher-Scoring Algorithmus oder der Newton-Raphson Algorithmus.

Ein Nachteil solcher Verfahren ist, dass der ML-Schätzer stark von den Startwerten abhängt oder das Verfahren nicht konvergiert.

### Indirekte Maximierung

Neben den direkten Methoden die Log-Likelihood  $l(\boldsymbol{\alpha})$  zu maximieren, gibt es auch die Möglichkeit das Maximum indirekt durch den Expectation-Maximizing (EM-) Algorithmus zu bestimmen. Dieser besteht iterativ aus dem E-Schritt und dem M-Schritt.

Der E-Schritt basiert auf der Berechnung des bedingten Erwartungswerts dieser Dichte:

$$\log f(Y, A; \boldsymbol{\alpha}) = \sum_{i=1}^n \log f(\mathbf{y}_i | \mathbf{a}_i; \boldsymbol{\alpha}) + \sum_{i=1}^n \log g(\mathbf{a}_i), \quad (27)$$

wobei  $A = (\mathbf{a}_1, \dots, \mathbf{a}_n)$ .

Der bedingte Erwartungswert der Dichte  $\log f(Y, A; \boldsymbol{\alpha})$ , gegeben den Daten und der aktuellen Schätzung von  $\boldsymbol{\alpha}$ , ist definiert durch:

$$M(\boldsymbol{\alpha} | \boldsymbol{\alpha}^{(p)}) = E\{\log f(Y, A; \boldsymbol{\alpha}) | Y; \boldsymbol{\alpha}^{(p)}\} = \int \log(f(Y, A; \boldsymbol{\alpha})) f(A | Y; \boldsymbol{\alpha}^{(p)}) dA, \quad (28)$$

wobei  $f(A | Y; \boldsymbol{\alpha}^{(p)})$  die Posteriori ist.

Wegen dem Satz von Bayes und (13) gilt folgende Proportionalität:

$$f(A | Y; \boldsymbol{\alpha}^{(p)}) \propto \prod_{i=1}^n f(\mathbf{y}_i | \mathbf{a}_i; \boldsymbol{\alpha}^{(p)}) \prod_{i=1}^n g(\mathbf{a}_i) \quad (29)$$

Durch Einsetzen von (27) und (29) in (28) vereinfacht sich die Berechnung im E-Schritt zu:

$$M(\boldsymbol{\alpha} | \boldsymbol{\alpha}^{(p)}) = \sum_{i=1}^n \frac{\int (\log f(\mathbf{y}_i | \mathbf{a}_i; \boldsymbol{\alpha}) + \log g(\mathbf{a}_i)) f(\mathbf{y}_i | \mathbf{a}_i; \boldsymbol{\alpha}^{(p)}) g(\mathbf{a}_i) d\mathbf{a}_i}{\int f(\mathbf{y}_i | \mathbf{a}_i; \boldsymbol{\alpha}^{(p)}) g(\mathbf{a}_i) d\mathbf{a}_i}$$

Zur Integration kann Monte-Carlo Integration oder bei Normalverteilungsannahme die Gauß-Hermite Integration verwendet werden.

Bei der Monte-Carlo Integration ist der bedingte Erwartungswert  $M(\boldsymbol{\alpha}|\boldsymbol{\alpha}^{(p)})$  definiert durch:

$$M(\boldsymbol{\alpha}|\boldsymbol{\alpha}^{(p)}) \approx M^{mc}(\boldsymbol{\alpha}|\boldsymbol{\alpha}^{(p)}) = \sum_{i=1}^n \sum_{j=1}^m c_{ij}^{mc} (\log f(\mathbf{y}_i|\mathbf{d}_{ij}; \boldsymbol{\alpha}) + \log g(\mathbf{d}_{ij})),$$

$$c_{ij}^{mc} = \frac{f(\mathbf{y}_i|\mathbf{d}_{ij}; \boldsymbol{\alpha}^{(p)})}{\sum_{k=1}^m f(\mathbf{y}_i|\mathbf{d}_{ik}; \boldsymbol{\alpha}^{(p)})} \quad \text{und} \quad \sum_{j=1}^m c_{ij}^{mc} = 1.$$

Ist die Dichte  $g$ , aus welcher die  $\mathbf{d}_{ij}$  unabhängig und identisch verteilte Zufallszahlen sind, normalverteilt, so kann auch analog das Gauß-Hermite Verfahren verwendet werden.

Der EM-Algorithmus basiert iterativ auf zwei Schritten. Zum einen auf dem bereits beschriebenen E-Schritt, indem der Erwartungswert  $M(\boldsymbol{\alpha}|\boldsymbol{\alpha}^{(p)})$  berechnet wird und zum anderen dem M-Schritt, indem  $M(\boldsymbol{\alpha}|\boldsymbol{\alpha}^{(p)})$  bezüglich  $\boldsymbol{\alpha}$  maximiert wird. Bei Verwendung des Monte-Carlo Integrationsverfahrens muss dementsprechend das Maximum von  $M^{mc}(\boldsymbol{\alpha}|\boldsymbol{\alpha}^{(p)})$  berechnet werden. Die Maximierung erfolgt durch Verwendung der ersten Ableitung von  $M^{mc}(\boldsymbol{\alpha}|\boldsymbol{\alpha}^{(p)})$ :

$$u(\boldsymbol{\alpha}|\boldsymbol{\alpha}^{(p)}) = \frac{\partial M^{mc}(\boldsymbol{\alpha}|\boldsymbol{\alpha}^{(p)})}{\partial \boldsymbol{\alpha}} = \sum_{i=1}^n \sum_{j=1}^m c_{ij}^{mc} \frac{\partial \log f(\mathbf{y}_i|\mathbf{d}_{ij}; \boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} = \mathbf{0}.$$

Die erwartete bedingte Informationsmatrix ist der negative Erwartungswert der zweiten Ableitung von  $E^{mc}(\boldsymbol{\alpha}|\boldsymbol{\alpha}^{(p)})$  und somit:

$$U(\boldsymbol{\alpha}|\boldsymbol{\alpha}^{(p)}) = -E \left( \frac{\partial^2 E^{mc}(\boldsymbol{\alpha}|\boldsymbol{\alpha}^{(p)})}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}'} \right) = \sum_{i=1}^n \sum_{j=1}^m c_{ij}^{mc} F_{ij}(\boldsymbol{\alpha}),$$

wobei

$$F_{ij}(\boldsymbol{\alpha}) = \begin{bmatrix} F_{ij}^{\beta\beta} & F_{ij}^{\beta\theta} \\ F_{ij}^{\theta\beta} & F_{ij}^{\theta\theta} \end{bmatrix}$$

mit

$$F_{ij}^{\beta\beta} = \sum_{t=1}^{T_i} X'_{it} D_{it} \Sigma_{it}^{-1} D'_{it} X_{it},$$

$$F_{ij}^{\beta\theta} = (F_{ij}^{\theta\beta})' = \sum_{t=1}^{T_i} X'_{it} D_{it} \Sigma_{it}^{-1} D'_{it} (\mathbf{d}_{ij}' \otimes W_{it}),$$

$$F_{ij}^{\theta\theta} = \sum_{t=1}^{T_i} (\mathbf{d}_{ij}' \otimes W_{it})' D_{it} \Sigma_{it}^{-1} D'_{it} (\mathbf{d}_{ij}' \otimes W_{it}),$$

und  $D_{it} = D_{it}(\boldsymbol{\alpha}, \mathbf{d}_{ij}) = \partial h(\boldsymbol{\eta}_{itj}) / \partial \boldsymbol{\eta}$ ,  $\Sigma_{it} = \Sigma_{it}(\boldsymbol{\alpha}, \mathbf{d}_{ij}) = \text{cov}(\mathbf{y}_{it} | \mathbf{d}_{ij})$  und  $\boldsymbol{\eta}_{itj} = X_{it}\boldsymbol{\beta} + \mathbf{d}_{ij}' \otimes W_{it}\boldsymbol{\theta}$ . Der M-Schritt der Iteration ist dann definiert durch:

$$\boldsymbol{\alpha}_{k+1} = \mathbf{a}_k + U^{-1}(\boldsymbol{\alpha}_k | \boldsymbol{\alpha}^{(p)})u(\mathbf{a}_k | \boldsymbol{\alpha}^{(p)}), \quad (30)$$

wobei  $k$  der Iterationsindex ist. Die Iteration startet bei  $\boldsymbol{\alpha}_0 = \boldsymbol{\alpha}^{(p)}$ .

Das Konzept EM-Algorithmus mit Monte-Carlo Integration ist folgendes:

- Berechne den Startwert  $\boldsymbol{\beta}^{(0)}$  durch Verwendung eines GLM ohne zufällige Effekte. Lege einen beliebigen Startwert für  $\boldsymbol{\theta}^{(0)}$  fest, zum Beispiel  $\boldsymbol{\theta}^{(0)} = \mathbf{0}$ . Definierte  $\boldsymbol{\alpha}^{(0)} = (\boldsymbol{\beta}^{(0)}, \boldsymbol{\theta}^{(0)})$ .

Für  $p=0,1,2,\dots$

- Approximation von  $M(\boldsymbol{\alpha} | \boldsymbol{\alpha}^{(p)})$  durch  $M^{mc}(\boldsymbol{\alpha} | \boldsymbol{\alpha}^{(p)})$ .
- Berechnung von  $\boldsymbol{\alpha}^{(p+1)}$  durch (30).

### 3.4.2.2 Posteriori-Erwartungswert-Schätzung für $b_i$

Durch die Maximum-Likelihood-Methode wurden die festen Effekte  $\boldsymbol{\beta}$  und die Kovarianz der zufälligen Effekte  $Q$  geschätzt. Die Schätzung der zufälligen Effekte  $\mathbf{b}_1, \dots, \mathbf{b}_n$  erfolgt durch die Posteriori-Erwartungswert-Schätzung.

Die Berechnungen basieren auf der Posteriori von  $\mathbf{b}_i$ , gegeben  $\mathbf{y}_i$ :

$$f(\mathbf{b}_i | \mathbf{y}_i; \boldsymbol{\beta}, Q) = \frac{f(\mathbf{y}_i | \mathbf{b}_i; \boldsymbol{\beta})p(\mathbf{b}_i; Q)}{\int f(\mathbf{y}_i | \mathbf{b}_i; \boldsymbol{\beta})p(\mathbf{b}_i; Q)d\mathbf{b}_i},$$

und der Proportionalität:

$$f(\mathbf{b}_i | \mathbf{y}_i; \boldsymbol{\beta}, Q) \propto f(\mathbf{y}_i | \mathbf{b}_i; \boldsymbol{\beta})p(\mathbf{b}_i; Q). \quad (31)$$

Die festen Effekte  $\boldsymbol{\beta}$  und  $Q$  werden durch ihre, in Abschnitt 3.4.2.1 ermittelten Schätzer  $\hat{\boldsymbol{\beta}}$  und  $\hat{Q}$  ersetzt. Der beste empirische bayesianische Punktschätzer im Sinne des MSE (mean squared error) ist der Posteriori-Erwartungswert:

$$\mathbf{b}_i^m = E(\mathbf{b}_i | \mathbf{y}_i) = \int \mathbf{b}_i f(\mathbf{b}_i | \mathbf{y}_i; \hat{\boldsymbol{\beta}}, \hat{Q})d\mathbf{b}_i.$$

Die Posteriori Kovarianz Matrix  $V_i^m$  ist definiert durch:

$$V_i^m = \text{cov}(\mathbf{b}_i | \mathbf{y}_i) = E(\mathbf{b}_i \mathbf{b}_i') - \mathbf{b}_i(\mathbf{b}_i)'$$

Zur Berechnung des Posteriori-Erwartungswerts und der Kovarianz bedarf also der Berechnung von Integralen. Der Einfachheit halber werden diese benötigten

Integrale in einer Gleichung  $S(q(b_i))$  formuliert, wobei  $q(\mathbf{b}_i) = 1$ ,  $q(\mathbf{b}_i) = b_i$  oder  $q(b_i) = b_i b_i'$  und unter Verwendung von (31) folgt:

$$S(q(\mathbf{b}_i)) = \int q(b_i) f(\mathbf{y}_i | \mathbf{b}_i; \hat{\boldsymbol{\beta}}) p(\mathbf{b}_i; \hat{Q}) d\mathbf{b}_i.$$

Zur Lösung kann zum Beispiel das Gauß-Hermite oder Monte-Carlo Verfahren verwendet werden. Weitere Einzelheiten findet man bei Fahrmeir und Tutz (1994).

## 4 Anwendung von marginalen Modellen und Modellen mit zufälligen Effekten

In diesem Kapitel werden die marginalen Modelle und die Modelle mit zufälligen Effekte auf die Daten der Allgemeinen Bevölkerungsumfrage der Sozialwissenschaften angewendet. Abschnitt 4.1 beschreibt diese Umfrage und führt die wichtigsten Variablen für die weiteren Abschnitte auf. Ein marginales Modell mit ordinalem Response wird in Abschnitt 4.2 und ein ordinales Modell mit zufälligem Intercept in Abschnitt 4.3 gefittet. Ein Vergleich dieser Modelle ist in Abschnitt 4.4 gegeben.

### 4.1 Allgemeine Bevölkerungsumfrage der Sozialwissenschaften

Die Daten der Allgemeinen Bevölkerungsumfrage der Sozialwissenschaften (ALLBUS) werden von der Gesellschaft Sozialwissenschaftlicher Infrastruktur (GESIS) in Zusammenarbeit mit dem ALLBUS-Ausschuss erhoben.

Das Untersuchungsgebiet ist die Bundesrepublik Deutschland. ALLBUS stellt eine Trenderhebung zur gesellschaftlichen Dauerbeobachtung von Einstellungen, Verhalten und sozialem Wandel in Deutschland dar.

Unter anderem wird das Vertrauen in öffentliche Einrichtungen und Organisationen erhoben. Dies geschieht über Fragen zum Vertrauen in das Gesundheitswesen, das Bundesverfassungsgericht, den Bundestag, die Gemeindeverwaltung, die Bundeswehr, die Kirche, die Justiz, das Fernsehen, das Zeitungswesen, die Hochschulen, die Bundesregierung, die Gewerkschaften, die Polizei, die Arbeitsämter, die Rentenversicherung, die Arbeitsverbände, die Kommission der Europäischen Union, das Europäische Parlament und den Europäischen Gerichtshof.

Des Weiteren werden Merkmale über die befragten Personen in die Daten aufgenommen. Diese sind zum Beispiel: Alter, Geschlecht, Einkommen und Bundesland der Befragung.

Die Daten des ALLBUS (1980-2006) liegen im `*.sav` Format vor, also im Format für die statistische Software SPSS. Dieses Datei Format lässt sich in die Software R mit Hilfe der Pakets `foreign` und der darin implementierten Funktion `read.spss()` einlesen. Alternativ kann der Datensatz als `*.csv` Datei aus SPSS exportiert und mit `read.csv(*.csv, sep= ;)` aus dem Paket `utils` in R importiert werden.

Die Variablen, welche im Folgenden untersucht werden sind:

- Vertrauen in öffentliche Einrichtungen und Organisationen (Variablen: v164-v183)

- Nettoeinkommen des Befragten
- Alter des Befragten
- Geschlecht des Befragten
- Bundesland der Befragung

Bei ALLBUS werden nicht jährlich alle Variablen erhoben. Im Jahr 2002 wurden 11 der 20 Vertrauensvariablen erhoben. Dies ist der größte Wert in den Jahren 1980-2006. Deshalb wird im Weiteren ausschließlich das Jahr 2002 betrachtet. Im Jahr 2002 wurden 2820 Personen befragt.

Die Struktur des Originaldatensatzes ist wie folgt:

v164	v165	v166	...	Alter	Geschlecht	Einkommen
5	6	3		65	2	2000
3	2	2		20	1	500
⋮						

Die Zielgröße muss als Vektor formuliert werden. Der Datensatz wird deshalb umstrukturiert zu:

Vertrauen	Alter	Geschlecht	Einkommen
5	65	2	2000
6	65	2	2000
3	65	2	2000
⋮			
3	20	1	500
2	20	1	500
2	20	1	500
⋮			

Die Antworten einer Person stehen jeweils untereinander. Dies lässt sich durch eine selbst geschriebene Schleife oder mit dem Befehl `reshape()` vornehmen. Da sich die Kovariablen während der elf Befragungen zum Vertrauen nicht ändern, stehen sie elf Mal wiederholt im Datensatz.

Der Befehl `reshape()` erzeugt automatisch die Variable `id`, die angibt welche Person befragt wurde und die Variable `time`, die angibt welche Vertrauensfrage gestellt wurde.

Die Ausprägungen bzw. Kodierungen der in den Regressionen verwendeten Variablen sind:

**Vertrauen** in öffentliche Einrichtungen und Organisationen: Überhaupt kein Vertrauen (1),..., sehr großes Vertrauen (7)

**time**: Gesundheitswesen (1), Bundesverfassungsgericht (2), Bundestag (3), Katholische Kirche (4), Evangelische Kirche (5), Justiz (6), Fernsehen (7), Zeitungswesen (8), Hochschulen und Universitäten (9), Bundesregierung (10), Polizei (11)

**Nettoeinkommen** des Befragten: in Euro angegeben

**Alter** des Befragten: in Jahren angegeben

**Geschlecht** des Befragten: weiblich (0), männlich (1)

**Bundesland** der Befragung: Baden-Württemberg (1), Bayern (2), ehemaliges Berlin-West (3), Bremen (4), Hamburg (5), Hessen (6), Niedersachsen (7), Nordrhein-Westfalen (8), Rheinland-Pfalz (9), Saarland (10), Schleswig-Holstein (11), ehemaliges Berlin-Ost (12), Brandenburg (13), Mecklenburg-Vorpommern (14), Sachsen (15), Sachsen-Anhalt (16), Thüringen (17).

An der Umfrage im Jahr 2002 beteiligten sich 1423 Frauen und 1397 Männer, deren Alters-Median bei 45 Jahren lag. Das Einkommen wurde zwischen 50 Euro und 15200 Euro angegeben. Der Median lag bei 1358 Euro. Pro Bundesland wurden zwischen 11 Personen und 490 Personen befragt. Die Verteilung der Vertrauensstärken war:

1	2	3	4	5	6	7
2768	3013	4823	7321	6739	4194	1431

## 4.2 Marginales Modell mit ordinalem Response

Marginale Modelle mit ordinalem Response sind im Paket `geepack` durch die Funktion `ordgee()` berechenbar.

Die Auswertung schließt 2017 Cluster ein. Die maximale Clustergröße ist 11. Verwendet wird der Logit-Link. Bei unabhängigen Beobachtungen innerhalb eines Clusters, also Odds Ratio  $\gamma = 1$ , ist die Regression:

```
ordgee(formula = Vertrauen ~ time + Bundesland + Geschlecht *
time + Alter * time + log(Einkommen) * time, id=id, data=allbus,
corstr = "independence", int.const = TRUE).
```

Für korrelierte Messwiederholungen, zum Beispiel Äquikorrelation  $\gamma = \alpha$  lautet die Regression:

```
ordgee(formula = Vertrauen ~ time + Bundesland + Geschlecht *
time + Alter * time + log(Einkommen) * time, id=id, data=allbus,
```

```
corstr = "exchangeable", control = geese.control(maxit = 900,
epsilon = 0.9), int.const = TRUE)
```

Aus Konvergenzgründen muss das Konvergenzkriterium von  $\epsilon = 0.0001$  auf  $\epsilon = 0.9$  geändert und die Anzahl der Iterationen erhöht werden. Bei unspezifizierter Korrelation konvergiert der Algorithmus nicht. Die Schätzer der Regressionskoeffizienten  $\beta$  und der Schätzer für  $\alpha$  befinden sich in Tabelle 2.

Die Funktion `ordgee` verwendet das Logit-Modell von Heagerty und Zeger (1996):

$$\text{logit}(P(y_{it} > r|\mathbf{x}_{it})) = \log \frac{P(y_{it} > r|\mathbf{x}_{it})}{P(y_{it} \leq r|\mathbf{x}_{it})} = \theta_r + \mathbf{x}_{it}'\beta.$$

Um eine vergleichbare Interpretation mit dem Modell mit zufälligen Effekten zu erhalten, muss man sich folgender Äquivalenz bewusst sein:

$$\begin{aligned} \Leftrightarrow -\log \frac{P(y_{it} > r|\mathbf{x}_{it})}{P(y_{it} \leq r|\mathbf{x}_{it})} &= -(\theta_r + \mathbf{x}_{it}'\beta) \\ \Leftrightarrow \log \frac{P(y_{it} \leq r|\mathbf{x}_{it})}{P(y_{it} > r|\mathbf{x}_{it})} &= -\theta_r - \mathbf{x}_{it}'\beta. \end{aligned}$$

Bei Annahme von unabhängigen Messwiederholungen sind mehr Schätzer signifikant und die Standardabweichungen kleiner als bei dem Modell mit Äquikorrelation. Die Schätzer der beiden Modelle sind leicht unterschiedlich. Zum Beispiel ist der Geschlechtseffekt im Unabhängigkeitsmodell 0.455130076 und im Äquikorrelationsmodell 0.4352151714. Bei Schätzern nahe um Null sind manchmal auch die Vorzeichen verschieden. Die Interaktion `time3:Alter` hat bei Unabhängigkeit den Wert  $-0.002544152$  und bei Äquikorrelation den Wert 0.0068748827.

Die Intercepts sind die Schwellenparameter, welche die Vertrauensstufen unterteilen. So liegt die Chance auf Vertrauen in Stufe drei oder niedriger im Verhältnis zu Vertrauen in Stufe vier oder höher bei  $\exp(-2.564413142) = 0.07696434$ . Die Interpretation der restlichen Intercepts ist analog.

Die Kovariable `time` gibt an, bei welcher Vertrauens-Frage das Vertrauen hoch oder niedrig ist. Bei dem marginalen Modell mit Unabhängigkeitsannahme ist das Vertrauen in allen betrachteten Fragen eher niedrig als bei der Frage nach dem Vertrauen in das Gesundheitswesen (1). Bei Annahme der Äquikorrelation ist es, bis auf die Frage nach der Justiz (6) ebenso. Bei der Frage nach der Justiz (6) ist das Vertrauen höher als bei der Frage nach dem Gesundheitswesen (1). Diese Interpretationen sind für den Fall, dass die Interaktionen Null und alle anderen Variablen konstant sind. Wird für die Regression das Modell mit Äquikorrelation spezifiziert, so ist die Frage nach der Justiz die einzige signifikante Frage. Bei Unabhängigkeit sind alle Vertrauensfragen, bis auf Fernsehen (7), Hochschulen und

Tabelle 2: Regressionskoeffizienten und geschätztes Odds Ratio

	independence			exchangeable		
	estimate	san.se	p	estimate	san.se	p
Inter:1	4.472091776	0.502888582	0.000000e+00	4.5995809890	0.936990773	9.159515e-07
Inter:2	3.491969243	0.501075428	3.193112e-12	3.5709194730	1.089441422	1.046371e-03
Inter:3	2.564413142	0.500739828	3.035192e-07	2.6139026663	1.149499518	2.296909e-02
Inter:4	1.422817048	0.500101101	4.440322e-03	1.4392623383	1.202583208	2.313810e-01
Inter:5	0.184332745	0.499842196	7.122903e-01	0.1657736854	1.248962826	8.944076e-01
Inter:6	-1.294711321	0.502063811	9.915036e-03	-1.3720433498	1.237380948	2.675041e-01
time2	-4.264023548	0.681959470	4.036609e-10	-3.4971135635	6.080567490	5.652037e-01
time3	-2.070858651	0.603467872	6.000381e-04	-2.2494355495	6.133379407	7.138032e-01
time4	-3.569075906	0.679240334	1.484182e-07	-3.1262114437	5.359277276	5.596845e-01
time5	-2.122789957	0.650349560	1.098220e-03	-1.9371523668	1.485347987	1.921742e-01
time6	-1.933827333	0.613756146	1.628199e-03	-1.6667416722	0.714127580	1.959808e-02
time7	-1.111245627	0.684697103	1.045950e-01	-1.1624547691	1.317924980	3.777584e-01
time8	-1.449923112	0.638367942	2.312901e-02	-1.6830501225	0.949606048	7.633371e-02
time9	-0.270613697	0.637575928	6.712438e-01	0.1625656579	3.115117975	9.583805e-01
time10	-2.349344436	0.611728452	1.227829e-04	-2.1041541440	1.641618707	1.999284e-01
time11	-0.316055839	0.597545759	5.968587e-01	-0.1284808764	4.171638924	9.754301e-01
Bundesland2	-0.005792695	0.113012000	9.591205e-01	0.1738001104	0.205453160	0.975888e-01
Bundesland3	-0.005206398	0.167284053	9.751714e-01	0.0797908951	0.311571519	3.978800e-01
Bundesland4	0.179457952	0.335466400	5.926840e-01	0.2422068937	0.415372283	5.598206e-01
Bundesland5	-0.401661758	0.225617353	7.503037e-02	-0.4667348081	0.371505697	2.089953e-01
Bundesland6	-0.003428392	0.132148516	9.793024e-01	0.2510480793	0.190444077	1.874285e-01
Bundesland7	0.139993011	0.131114591	2.856491e-01	0.2189898006	0.143831141	1.278719e-01
Bundesland8	-0.020972300	0.106961666	8.445530e-01	0.1325938596	0.143463263	3.553636e-01
Bundesland9	-0.076202562	0.175403199	6.639673e-01	0.0249659300	0.210482032	9.055819e-01
Bundesland10	0.058800460	0.253017052	8.162295e-01	0.0175397271	0.427740147	9.672915e-01
Bundesland11	0.008501148	0.172148942	9.606145e-01	0.0405341523	0.211361217	8.479171e-01
Bundesland12	-0.417063096	0.205920769	4.283071e-02	-0.3069233052	0.235004399	1.915416e-01
Bundesland13	-0.431582499	0.129647496	8.719224e-04	-0.2804525510	0.150187615	6.185300e-02
Bundesland14	0.055733385	0.153705346	7.169046e-01	0.2326690378	8.463808301	9.978090e-01
Bundesland15	-0.186637623	0.119307785	1.177384e-01	0.0120950686	0.149172745	9.353776e-01
Bundesland16	-0.132215238	0.130993132	3.128166e-01	-0.0539281643	0.147149434	7.140027e-01
Bundesland17	-0.266620331	0.139034653	5.515492e-02	-0.3480876070	0.215668645	1.065290e-01
Geschlecht1	0.455130076	0.105485462	1.598750e-05	0.4352151714	0.109821985	7.404021e-05
Alter	0.008580840	0.002966453	3.820360e-03	0.0081163528	0.003255728	1.266882e-02
log(Einkommen)	-0.296074168	0.072831738	4.799529e-05	-0.3135315738	0.142072041	2.732459e-02
time2:Geschlecht1	-0.048573514	0.145593591	7.386633e-01	-0.1189761710	0.816284361	8.841161e-01
time3:Geschlecht1	-0.333402376	0.133068807	1.222823e-02	-0.4944441322	1.950362400	7.998711e-01
time4:Geschlecht1	-0.452570891	0.150610291	2.656564e-03	-0.4729728782	0.909657592	6.031012e-01
time5:Geschlecht1	-0.564309860	0.145836424	1.090716e-04	-0.4661526320	1.369436899	7.335573e-01
time6:Geschlecht1	-0.243862542	0.134049915	6.888213e-02	-0.2643812035	0.161003564	1.005728e-01
time7:Geschlecht1	-0.113925426	0.149682179	4.465881e-01	-0.0943427436	0.208546505	6.509936e-01
time8:Geschlecht1	-0.227989024	0.141123242	1.061957e-01	-0.2508140507	0.155030704	1.056982e-01
time9:Geschlecht1	-0.184821971	0.142516070	1.946828e-01	-0.2205646998	0.641613572	7.310225e-01
time10:Geschlecht1	-0.482356340	0.134192120	3.249894e-04	-0.4588776107	0.253316549	7.006674e-02
time11:Geschlecht1	-0.233300476	0.136420073	8.723585e-02	-0.2713618473	0.450487779	5.469256e-01
time2:Alter	-0.008414819	0.004064872	3.844005e-02	-0.0002856003	0.128957498	9.982329e-01
time3:Alter	-0.002544152	0.003664877	4.875577e-01	0.0068748827	0.091598228	9.401711e-01
time4:Alter	0.013167244	0.004180779	1.635663e-03	0.0100113230	0.130739587	9.389620e-01
time5:Alter	0.018309102	0.004134566	9.497830e-06	0.0137122806	0.032867747	6.765353e-01
time6:Alter	-0.013926111	0.003799765	2.473452e-04	-0.0139344533	0.004920504	4.627002e-03
time7:Alter	0.006280873	0.004209690	1.356982e-01	0.0034069862	0.010490448	7.453549e-01
time8:Alter	-0.006970887	0.003974975	7.948304e-02	-0.0094016734	0.004055316	2.042987e-02
time9:Alter	-0.007829882	0.004022998	5.162108e-02	-0.0025348731	0.080800203	9.749728e-01
time10:Alter	-0.005610012	0.003841183	1.441552e-01	-0.0082336984	0.010925770	4.510873e-01
time11:Alter	0.006836745	0.003772091	6.991588e-02	0.0129052344	0.045494590	7.766669e-01
time2:log(Einkommen)	0.758940163	0.101782663	8.881784e-14	0.6154258923	0.264124371	1.980293e-02
time3:log(Einkommen)	0.231881109	0.089643683	9.690056e-03	0.2210693787	0.542435992	6.836046e-01
time4:log(Einkommen)	0.272285379	0.100578604	6.785543e-03	0.2392527749	0.162914116	1.419461e-01
time5:log(Einkommen)	0.120943949	0.096060877	2.080179e-01	0.1251259451	0.470958925	7.904831e-01
time6:log(Einkommen)	0.372349385	0.090769951	4.093789e-05	0.3398483030	0.110579599	2.116797e-03
time7:log(Einkommen)	-0.060639719	0.101588284	5.505632e-01	-0.0391917173	0.159622273	8.060477e-01
time8:log(Einkommen)	0.164295651	0.094537151	8.222930e-02	0.2149690107	0.138312582	1.201305e-01
time9:log(Einkommen)	0.232396667	0.094128042	1.355132e-02	0.1486497413	0.116937464	2.036610e-01
time10:log(Einkommen)	0.297273149	0.090983692	1.085696e-03	0.2896860955	0.215341886	1.785484e-01
time11:log(Einkommen)	0.110785849	0.089162933	2.140484e-01	0.0561411371	0.360914083	8.763857e-01
$\alpha$				0.9675752	0.1346265	6.61804e-13

Universitäten (9) und Polizei (11) signifikant.

Die exakte Interpretation des Regressionskoeffizienten `time` für das Unabhängigkeitsmodell wird im folgenden Absatz durchgeführt.

Die Chance auf Vertrauen in der Kategorie  $r$  oder niedriger im Verhältnis zu einer höheren Vertrauenskatgorie ist bei der Frage nach dem Vertrauen in das Bun-

desverfassungsgericht (2) um den Faktor  $\exp(4.264023548) = 71.09546$  höher als bei der Frage nach dem Vertrauen in das Gesundheitswesen (1). Die Chance auf Vertrauen in der Kategorie  $r$  oder niedriger im Verhältnis zu einer höheren Vertrauenskategorie ist bei der Frage nach dem Vertrauen in den Bundestag (3) um den Faktor  $\exp(2.070858651) = 7.93163$  höher als bei der Frage nach dem Vertrauen in das Gesundheitswesen (1). Die Interpretation der anderen Regressionskoeffizienten ist analog.

Die Regressionskoeffizienten der Einflussvariable **Bundesland** im Unabhängigkeitsmodell geben an, dass in Bayern (2), im ehemaligen Berlin-West (3), Hamburg (5), Hessen (6), Nordrhein-Westfalen (8), Rheinland-Pfalz (9), im ehemaligen Berlin-Ost (12), Brandenburg (13), Sachsen (15), Sachsen-Anhalt (16) und in Thüringen (17) das Vertrauen in öffentliche Einrichtungen und Organisationen geringer ist als in Baden-Württemberg (1). Genauer betrachtet ist die Chance auf Vertrauen in der Kategorie  $r$  oder niedriger im Verhältnis zu einer höheren Vertrauenskategorie bei einer Befragung im ehemaligen Berlin-Ost (12) um den Faktor  $\exp(0.417063096) = 1.517498$  höher als bei einer Befragung in Baden-Württemberg (1). Die Chance auf Vertrauen in der Kategorie  $r$  oder niedriger im Verhältnis zu einer höheren Vertrauenskategorie ist bei der Frage nach dem Vertrauen in Brandenburg (13) um den Faktor  $\exp(0.431582499) = 1.539692$  höher als bei der Frage nach dem Vertrauen in Baden-Württemberg (1).

Einen signifikanten Einfluss haben die Regionen ehemaliges Berlin-Ost (12) und Brandenburg (13).

In Bremen (4), Niedersachsen (7), Saarland (10), Schleswig-Holstein (11) und Mecklenburg-Vorpommern (14) ist das Vertrauen höher als in Baden-Württemberg (1).

Die Chance auf Vertrauen in der Kategorie  $r$  oder niedriger im Verhältnis zu einer höheren Vertrauenskategorie ist bei einer Befragung in Bremen (4) um den Faktor  $\exp(-0.179457952) = 0.8357231$  niedriger als in Baden-Württemberg (1).

Bei Annahme der Äquikorrelation ist keines der Bundesländer signifikant.

Bei der Unabhängigkeitsannahme und bei der Äquikorrelationsannahme hat die Kovariable **Geschlecht** einen signifikanten Einfluss. Bei der Unabhängigkeitsannahme ist die Chance auf Vertrauen in der Kategorie  $r$  oder niedriger im Verhältnis zu einer höheren Vertrauenskategorie bei Männern (1) um den Faktor  $\exp(-0.455130076) = 0.6343654$  niedriger als bei Frauen (0).

In beiden Modellen ist die Einflussvariable **Alter** signifikant. Bei Annahme von Unabhängigkeit senkt die Chance auf Vertrauen in der Kategorie  $r$  oder niedriger im Verhältnis zu einer höheren Vertrauenskategorie, mit dem Zuwachs um ein Jahr, um den Faktor  $\exp(-0.008580840) = 0.9914559$ . Je älter also die Personen sind, desto mehr Vertrauen haben sie in die öffentlichen Einrichtungen und Organisationen.

Sowohl bei der Annahme von Unabhängigkeit, als auch bei der Annahme von Äquikorrelation ist das logarithmierte Einkommen signifikant. Bei der Annahme von Unabhängigkeit steigt die Chance auf Vertrauen in der Kategorie  $r$  oder niedriger im Verhältnis zu einer höheren Vertrauenskategorie, mit dem Zuwachs um eine Einheit des logarithmierten Einkommens, um den Faktor  $\exp(0.296074168) = 1.34457$ . Je höher das Einkommen, desto niedriger ist das Vertrauen in öffentliche Einrichtungen und Organisationen.

Alle Interpretationen waren unter der Voraussetzung, dass alle anderen Variablen konstant sind und dass die Interaktionen gleich Null sind.

Der Assoziationsparameter  $\alpha$  ist im Äquikorrelationsmodell signifikant.

### 4.3 Ordinales Modell mit zufälligem Intercept

Eine Implementierung ordinaler Modelle mit zufälligen Effekten ist bei R im Paket `ordinal` zu finden. Die Funktion `clmm()` berechnet die zufälligen und festen Effekte bei ordinalem Response. Die Implementierung des Modells bei Verwendung des Logit-Link ist folgende:

$$\text{logit}(P(y_{ij} \leq r | \mathbf{x}_{it})) = \log \frac{P(y_{it} \leq r | \mathbf{x}_{it})}{P(y_{it} > r | \mathbf{x}_{it})} = \theta_r - \mathbf{x}_{ij}'\boldsymbol{\beta} - b_i.$$

Die Algorithmus der Regression

```
clmm(Vertrauen ~ 1+time+Bundesland+Geschlecht*time+Alter*time+
  log(Einkommen)*time, random=id, data=allbus_5,threshold="flexible",
  link="logistic",Hess=..,control=clmm.control(gradTol=..,maxIter=..))
```

konvergiert nicht, egal welches Schätzverfahren, wieviele Iterationen und welches Konvergenzkriterium angegeben wird.

Ein Vergleich dieses Modell mit zufälligen Intercept mit dem marginalen Modell von Abschnitt 4.2 ist deshalb nicht möglich.

### 4.4 Vergleich der Modelle

Der Vergleich der Modelle soll mit einem Modell mit zufälligen Intercept und einem marginalen Modell bei Unabhängigkeit, welche jeweils nur `time` als Kovariable haben, geschehen. Es werden 2820 Cluster mit einer maximalen Cluster Größe von 11 verwendet. Die Schätzer sind in Tabelle 3. Für die zufälligen Effekte gilt:

	Var	Std.Dev
id	1.400713	1.183517

Tabelle 3: Regressionskoeffizienten

	marginales Modell			Modell mit zufälligen Intercept			
	estimate	san.se	p	Estimate	Std. Error	z value	$Pr(>  z )$
Inter:1	2.89785319	0.05269650	0.00000000	-3.5427	0.0474	-74.6671	
Inter:2	1.93247125	0.04529007	0.00000000	-2.4578	0.0442	-55.6268	
Inter:3	1.01075695	0.04245199	0.00000000	-1.3420	0.0424	-31.6760	
Inter:4	-0.10124244	0.04106159	0.01367749	0.0452	0.0416	1.0868	
Inter:5	-1.30562390	0.04406386	0.00000000	1.5467	0.0429	36.0393	
Inter:6	-2.78225065	0.06299495	0.00000000	3.3962	0.0500	67.9649	
time2	0.59054129	0.05827693	0.00000000	0.6692	0.0498	13.4300	< 2.22e-16
time3	-0.72092520	0.05261930	0.00000000	-0.9117	0.0477	-19.1141	< 2.22e-16
time4	-1.16045412	0.05902002	0.00000000	-1.9404	0.0522	-37.1834	< 2.22e-16
time5	-0.62200973	0.05680442	0.00000000	-1.2240	0.0511	-23.9426	< 2.22e-16
time6	-0.07034848	0.05255760	0.18073272	-0.2364	0.0480	-4.9255	8.4156e-07
time7	-1.18768094	0.05847686	0.00000000	-1.4047	0.0488	-28.8116	< 2.22e-16
time8	-0.69695619	0.05546388	0.00000000	-0.8256	0.0480	-17.1921	< 2.22e-16
time9	0.90441563	0.05730381	0.00000000	0.8470	0.0488	17.3604	< 2.22e-16
time10	-0.81788295	0.05380582	0.00000000	-1.0327	0.0478	-21.5929	< 2.22e-16
time11	0.65540839	0.05377258	0.00000000	0.6206	0.0480	12.9315	< 2.22e-16

Das marginale Modell mit Äquikorrelation wird hier nicht aufgeführt, da es für ein genügend kleines  $\epsilon$  im Konvergenzkriterium nicht konvergiert. Für großes  $\epsilon$  konvergiert der Algorithmus. Die Standardabweichungen sind jedoch nicht mehr sinnvoll.

Die Schätzer des marginalen Modells sind näher an der Null als die Schätzer des Modells mit zufälligem Intercept. Die Standardabweichung des Modells mit zufälligen Intercept sind kleiner als die des marginalen Modells. Die einzig nicht-signifikante Kovariable ist im marginalen Modell `time6`, also die Frage nach dem Vertrauen in die Justiz. Für beide Modelle resultiert, mit dieser einen Ausnahme, dass es signifikant für das Vertrauen in öffentliche Einrichtungen und Organisationen ist, nach welcher Einrichtung bzw. Organisation gefragt wird.

Die Interpretation der Effekte der Kovariablen `time` des marginalen Modells und des Modells mit zufälligem Intercept findet sich im folgenden Absatz.

Wird nach dem Vertrauen in den Bundestag (3), Katholische Kirche (4), Evangelische Kirche (5), Justiz (6), Fernsehen (7), Zeitungswesen (8) und Bundesregierung (10) gefragt so ist das Vertrauen niedriger als bei der Frage nach dem Vertrauen in das Gesundheitswesen (1).

Die Chance auf Vertrauen in der Kategorie  $r$  oder niedriger im Verhältnis zu einer höheren Vertrauenskategorie ist bei der Frage nach dem Vertrauen in den Bundestag (3) um den Faktor  $\exp(0.72092520) = 2.056335$  (marginales Modell) bzw.  $\exp(0.9117) = 2.488549$  (Modell mit zufälligen Effekten) höher als bei der Frage nach dem Vertrauen in das Gesundheitswesen (1).

Bei den Fragen nach dem Bundesverfassungsgericht (2), Hochschulen und Universitäten (9) und Polizei (11) ist das Vertrauen höher als bei der Frage nach dem Vertrauen in das Gesundheitswesen (1).

Die Chance auf Vertrauen in der Kategorie  $r$  oder niedriger im Verhältnis zu einer höheren Vertrauenskategorie ist bei der Frage nach dem Vertrauen in das Bundesverfassungsgericht (2) um den Faktor  $\exp(-0.59054129) = 0.5540273$  (marginales Modell) bzw.  $\exp(-0.6692) = 0.5121181$  (Modell mit zufälligen Effekten) niedriger als bei der Frage nach dem Vertrauen in das Gesundheitswesen (1).

Die Intercepts sind die Schwellenparameter, welche die Trennung zwischen den Vertrauensstufen ziehen.

Bei marginalen Modellen sind diese Schwellenparameter für alle Personen gleich. Die Chance auf Vertrauen der Stufe 2 oder niedriger im Verhältnis zu dem Vertrauen in Stufe 3 oder höher liegt bei  $\exp(-1.93247125) = 0.1447899$ .

Bei Modellen mit zufälligen Effekten variieren diese für die Personen in der Form, dass von jedem Schwellenparameter der individuelle zufällige Effekt subtrahiert wird. Dementsprechend liegt die Chance auf Vertrauen der Stufe 2 oder niedriger im Verhältnis zu dem Vertrauen in Stufe 3 oder höher bei  $\exp(-2.4578)\exp(-b_i) = 0.08562311 \exp(-b_i)$ .

## 5 Zusammenfassung

Modelle mit zufälligen Effekten und marginale Modelle werden bei Messwiederholungsdaten verwendet. Marginale Modelle modellieren den Zusammenhang über die Korrelation oder das Odds Ratio und schätzen die Effekte für die Grundgesamtheit im Durchschnitt. Im Vergleich dazu beziehen Modelle mit zufälligen Effekten die Heterogenität der Cluster in die Modellierung ein und schätzen clusterspezifische Effekte. Die Interpretation der Regressionskoeffizienten ist also zum einen, wie sich die durchschnittliche Grundgesamtheit und zum anderen wie sich Cluster der Grundgesamtheit bezüglich Veränderungen der Kovariablen ändern. Beide Modelle können auf kategoriale Zielvariablen angewendet werden.

Für eine konsistente Schätzung muss bei marginalen Modellen der Erwartungswert und bei Modellen mit zufälligen Effekten der Link und die Verteilung der zufälligen Effekte korrekt spezifiziert sein.

Die Auswertung der Daten des Sozio-ökonomischen Panels von 2007 mit dem marginalen Modell liefert Schätzer für verschiedene Korrelationen, die für die verschiedenen Modelle nur sehr kleine Differenzen aufweisen. Die verschiedenen geschätzten Korrelationen sind klein.

Für die Daten der Allgemeinen Bevölkerungsumfrage der Sozialwissenschaften 2002 wurden marginale Modelle und Modelle mit zufälligen Effekten gefittet. Bei dem Modell mit vielen Kovariablen konvergierte der Algorithmus des Modells mit zufälligem Intercept nicht und der Algorithmus für das marginale Modell mit Korrelation nur für ein gelockertes Konvergenzkriterium. Ein Vergleich der Modelle ist somit nur über ein reduziertes Modell möglich. Im Modell mit nur einer Kovariablen waren die Regressionskoeffizienten leicht unterschiedlich und der zufällige Intercept deutete auf eine geringe Heterogenität der befragten Personen hin.

## Literatur

- Carey V, Zeger S, Diggle P (1993). “Modelling multivariate binary data with alternating logistic regressions.” *Biometrika*, **80**(3), 517–526.
- Christensen R (2010). *ordinal: Regression Models for Ordinal Data*. R package version 2010.06-12, URL <http://CRAN.R-project.org>.
- Diggle P, Heagerty P, Liang K, Zeger S (2002). *Analysis of longitudinal data*. Oxford University Press, USA.
- Fahrmeir L, Kneib T, Lang S (2007). *Regression: Modelle, Methoden und Anwendungen*. Springer.
- Fahrmeir L, Künstler R, Pigeot I, Tutz G (2003). *Statistik: Der Weg zur Datenanalyse*. Springer.
- Fahrmeir L, Pritscher L (1996). “Regression analysis of forest damage by marginal models for correlated ordinal responses.” *Environmental and Ecological Statistics*, **3**(3), 257–268.
- Fahrmeir L, Tutz G (1994). *Multivariate statistical modelling based on generalized linear models*. Springer Verlag.
- Heagerty P, Zeger S (1996). “Marginal Regression Models for Clustered Ordinal Measurements.” *Journal of the American Statistical Association*, **91**(435), 1024–1036.
- Hildebrand F (1956). *Introduction to numerical analysis*. Dover Pubns.
- Imai K, King G, Lau O (2008). “Zelig: Everyone’s statistical software.” *R package version*, pp. 474–483. ”<http://gking.harvard.edu/zelig/docs/probit.gee.pdf>”; besucht am 4.Mai 2010.
- Liang K, Zeger S (1986). “Longitudinal data analysis using generalized linear models.” *Biometrika*, **73**(1), 13–22.
- Liang K, Zeger S, Qaqish B (1992). “Multivariate regression analyses for categorical data.” *Journal of the Royal Statistical Society. Series B (Methodological)*, **54**(1), 3–40.
- Lipsitz S, Laird N, Harrington D (1991). “Generalized estimating equations for correlated binary data: using the odds ratio as a measure of association.” *Biometrika*, **78**(1), 153–160.

- McLachlan G, Krishnan T (1997). *The EM algorithm and extensions*. Wiley New York.
- Miller M, Davis C, Landis J (1993). “The analysis of longitudinal polytomous data: Generalized estimating equations and connections with weighted least squares.” *Biometrics*, **49**(4), 1033–1044.
- Prentice R (1988). “Correlated binary regression with covariates specific to each binary observation.” *Biometrics*, **44**(4), 1033–1048.
- Rubin D (1976). “Inference and missing data.” *Biometrika*, **63**(3), 581–592.
- Stiratelli R, Laird N, Ware J (1984). “Random-effects models for serial observations with binary response.” *Biometrics*, **40**(4), 961–971.
- Terejanu G (2009). “Tutorial on Monte Carlo Techniques.” <http://www.acsu.buffalo.edu/~terejanu/files/tutorialMC.pdf>; besucht am 11. August 2010.
- Toutenburg H (2003). *Lineare Modelle: Theorie und Anwendungen*. Springer.
- Williamson J, Kim K, Lipsitz S (1995). “Analyzing Bivariate Ordinal Data Using a Global Odds Ratio.” *Journal of the American Statistical Association*, **90**(432), 1432–1437.
- Yan J, Hojsgaard S, Halekoh U (2010). *geepack: Generalized Estimating Equation Package*. R package version 1.0-17, URL <http://CRAN.R-project.org>.
- Zeger S, Liang K (1986). “Longitudinal data analysis for discrete and continuous outcomes.” *Biometrics*, **42**(1), 121–130.
- Zeger S, Liang K, Albert P (1988). “Models for longitudinal data: a generalized estimating equation approach.” *Biometrics*, **44**(4), 1049–1060.

## **Eigenständigkeitserklärung**

Hiermit erkläre ich, Claudia Stuckart, dass es sich bei der vorliegenden Bachelorarbeit um eine selbständig verfasste Arbeit handelt und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt wurden.

München, den 10. September 2010

Claudia Stuckart