

# **Statistische Modellierung III**

## **-Verallgemeinerte Schätzgleichungen-**

Dr. Martin Scharpenberg

MSc Medical Biometry/Biostatistics

WiSe 2017/2018

## Setup

- Im GLM haben wir gesehen, wie wir allgemein lineare Modelle auf Daten, deren Verteilung einer Exponentialfamilie angehört, erweitern konnten
- Eine zentrale Annahme des Kapitels war, dass wir unabhängige Beobachtungen haben
- Wenn die Beobachtungen nicht unabhängig sind, ist die Methodik aus dem GLM-Kapitel nicht anwendbar
- Gegenstand dieser Vorlesung soll es nun sein die Schätzggleichung aus dem GLM-Ansatz zu verallgemeinern
- Mögliche Anwendungen sind wieder Messwiederholungen, z.B. Longitudinaldaten

## Wiederholung GLM

## Wiederholung: Generalisierte lineare Modelle

- $(Y_i, \mathbf{x}_i)$ ,  $i = 1, \dots, n$ , wieder Daten von stochastisch unabhängigen Beobachtungseinheiten mit zufälliger Zielvariable  $Y_i$  und fixen (nicht-zufälligen) Kovariablen
- Verallgemeinerung auf nicht-unabhängige Messungen später
- Zielvariablen  $Y_i$  sollen einen gemeinsamen Träger  $\mathbb{T}$  haben
- Verknüpfen Verteilung von  $Y_i$  mit den Kovariablen
- Machen dazu eine Verteilungs- und eine Strukturannahme

## Wiederholung: Generalisierte lineare Modelle

- Verteilungsannahme: Zielvariablen  $Y_i$  haben Dichten bzgl. des Lebesgue- oder Zählmaßes haben, die einer Exponentialfamilie entstammen, d.h. für alle Werte  $y_i$  im gemeinsamen Träger  $\mathbb{T}$  sei die Dichte von  $Y_i$

$$f(y_i|\theta_i, \phi, \omega_i) = f(y_i|\theta_i) = \exp \left\{ \frac{y_i \theta_i - b(\theta_i)}{\phi} \omega_i - c(y_i, \phi, \omega_i) \right\}.$$

- Strukturannahme:

$$\mu_i = E(Y_i) = b'(\theta_i) = h(\eta_i) \quad \text{für } \eta_i = \mathbf{x}_i \beta,$$

$$\text{bzw. } \mathbf{x}_i \beta = \eta_i = g(\mu_i) = g[b'(\theta_i)] \text{ für } g = h^{-1}$$

## Wiederholung: Generalisierte lineare Modelle

- Wir haben gesehen, dass für stochastisch unabhängige Beobachtungen für die Score-Funktion

$$\mathbf{s}(\beta) = \frac{\partial}{\partial \beta} l(\beta) = \sum_{i=1}^n \mathbf{x}_i^T \frac{h'(\eta_i)}{v(\mu_i)} \{Y_i - \mu_i\} \omega_i / \phi$$

gilt

- Der MLE für  $\beta$  wurde dann als Lösung der Schätzgleichung

$$\sum_{i=1}^n \mathbf{x}_i^T \frac{h'(\eta_i)}{v(\mu_i)} \{Y_i - \mu_i\} \omega_i / \phi = \mathbf{0}$$

bestimmt

- Wollen diese Schätzgleichung nun verallgemeinern

## Setup

## Setup

- Die Zufallsvariable  $Y_{ij}$  ist die  $j$ -te Messung der Zielvariable für das  $i$ -te Individuum,  $i = 1, \dots, N$ ,  $j = 1, \dots, n_i$
- $\mathbf{x}_{ij}$  ist der zugehörige Kovariablenvektor
- Fassen die Beobachtungen für das  $i$ -te Individuum im  $n_i$ -dimensionalen Vektor  $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{in_i})$  zusammen
- Individuum steht (wie bei den Mixed Models) wieder für eine Messeinheit (z.B. Patient, Schule, Familie, ...)
- Wir nehmen an, dass die  $\mathbf{Y}_i$ ,  $i = 1, \dots, N$ , stochastisch unabhängig sind



## Setup

- Die Struktur- und Verteilungsannahmen aus dem GLM bleiben bestehen
- Verteilungsannahme: Zielvariablen  $Y_{ij}$  haben Dichten bzgl. des Lebesgue- oder Zählmaßes haben, die einer Exponentialfamilie entstammen, d.h. für alle Werte  $y_{ij}$  im gemeinsamen Träger  $\mathbb{T}$  sei die Dichte von  $Y_{ij}$

$$f(y_{ij}|\theta_{ij}, \phi, \omega_{ij}) = f(y_{ij}|\theta_{ij}) = \exp \left\{ \frac{y_{ij}\theta_{ij} - b(\theta_{ij})}{\phi} \omega_{ij} - c(y_{ij}, \phi, \omega_{ij}) \right\}.$$

- Strukturannahme:

$$\mu_{ij} = E(Y_{ij}) = b'(\theta_{ij}) = h(\eta_{ij}) \quad \text{für } \eta_{ij} = \mathbf{x}_{ij}\beta,$$

$$\text{bzw. } \mathbf{x}_{ij}\beta = \eta_{ij} = g(\mu_{ij}) = g[b'(\theta_{ij})] \text{ für } g = h^{-1}$$

## Schätzgleichung und ihre Verallgemeinerung

## Schätzgleichung bei Unabhängigkeit

- Im ersten Schritt nehmen wir an, dass die Messungen innerhalb eines Individuums unabhängig sind
- Damit erfüllen die  $Y_{ij}$ ,  $i = 1, \dots, N$ ,  $j = 1, \dots, n_i$  die Annahmen eines GLM
- Wir können also den Score bestimmen als:

$$\mathbf{s}(\beta) = \sum_{i=1}^N \sum_{j=1}^{n_i} \mathbf{x}_{ij}^T \frac{h'(\eta_{ij})}{v(\mu_{ij})} \{Y_{ij} - \mu_{ij}\} \omega_{ij} / \phi$$

- Der MLE für  $\beta$  kann dann als Lösung der Schätzgleichung

$$\sum_{i=1}^N \sum_{j=1}^{n_i} \mathbf{x}_{ij}^T \frac{h'(\eta_{ij})}{v(\mu_{ij})} \{Y_{ij} - \mu_{ij}\} \omega_{ij} / \phi = \mathbf{0}$$

bestimmt werden (z.B. numerisch via Fisher-Scoring)

## Schätzgleichung bei Unabhängigkeit – Matrixschreibweise

- Wir können diese Schätzgleichung in Matrixschreibweise wie folgt auf Individuenebene darstellen:

$$\sum_{i=1}^N \mathbf{x}_i^T \mathbf{D}_i \mathbf{V}_i^{-1} (\mathbf{Y}_i - \mu_i(\beta)) = 0$$

- Dabei sind

$$\mathbf{x}_i = \begin{pmatrix} \mathbf{x}_{i1} \\ \vdots \\ \mathbf{x}_{in_i} \end{pmatrix}, \quad \mathbf{D}_i = \begin{pmatrix} h'(\eta_{i1}) & & 0 \\ & \ddots & \\ 0 & & h'(\eta_{in_i}) \end{pmatrix} \quad \text{und} \quad \mathbf{V}_i = \phi \begin{pmatrix} \frac{v(\mu_{i1})}{\omega_{i1}} & & 0 \\ & \ddots & \\ 0 & & \frac{v(\mu_{in_i})}{\omega_{in_i}} \end{pmatrix}$$

## Schätzgleichung bei Unabhängigkeit – Matrixschreibweise

- Die Matrix  $\mathbf{V}_i$  in der Schätzgleichung beschreibt im Wesentlichen die Kovarianzstruktur der Daten des  $i$ -ten Individuums
- Sie lässt sich wie folgt darstellen:

$$\mathbf{V}_i = \phi \mathbf{A}_i^{1/2} \mathbf{W}_i^{-1/2} \mathbf{I}_{n_i} \mathbf{W}_i^{-1/2} \mathbf{A}_i^{1/2},$$

dabei sind

$$\mathbf{A}_i = \begin{pmatrix} v(\mu_{i1}) & & 0 \\ & \ddots & \\ 0 & & v(\mu_{in_i}) \end{pmatrix} \quad \text{und} \quad \mathbf{W}_i = \begin{pmatrix} \omega_{i1} & & 0 \\ & \ddots & \\ 0 & & \omega_{in_i} \end{pmatrix}$$

## Schätzgleichung bei Unabhängigkeit – Verallgemeinerung

- Aus der Darstellung

$$\mathbf{V}_i = \phi \mathbf{A}_i^{1/2} \mathbf{W}_i^{-1/2} \mathbf{I}_{n_i} \mathbf{W}_i^{-1/2} \mathbf{A}_i^{1/2},$$

wird klar, dass die Schätzgleichung die Beobachtungen innerhalb der Individuen als unabhängig behandelt

- Die Einheitsmatrix  $\mathbf{I}_{n_i}$  ist die Korrelationsmatrix innerhalb der Individuen für die marginale Verteilung des Outcomes
- Wir verallgemeinern die Schätzgleichung, indem wir die Matrix  $\mathbf{I}_{n_i}$  durch eine allgemeinere Korrelationsmatrix  $\mathbf{R}_i(\alpha)$  ersetzen
- $\mathbf{R}_i(\alpha)$  ist hierbei über einen Vektor  $\alpha$  parametrisiert und muss in der Regel geschätzt werden
- Werden später typische Strukturen für  $\mathbf{R}_i(\alpha)$  betrachten

## Verallgemeinerte Schätzgleichung

- Wir betrachten also die verallgemeinerte Schätzgleichung (engl. *generalized estimating equation – GEE*)

$$\sum_{i=1}^N \mathbf{x}_i^T \mathbf{D}_i \mathbf{V}_i^{-1} (\mathbf{Y}_i - \mu_i(\beta)) = \mathbf{0},$$

wobei

$$\mathbf{V}_i = \phi \mathbf{A}_i^{1/2} \mathbf{W}_i^{-1/2} \mathbf{R}_i(\alpha) \mathbf{W}_i^{-1/2} \mathbf{A}_i^{1/2}$$

- Damit nehmen wir nichtmehr an, dass die Beobachtungen innerhalb eines Individuums unabhängig sind
- Korrelationsstruktur wird über  $\mathbf{R}_i(\alpha)$  modelliert
- Es ist zu beachten, dass obige Schätzgleichung **nicht** die Scoregleichung einer Likelihood ist (Stichwort: Quasi-Likelihood)

## Berechnung der Schätzer



## Schätzen von R

- In der Regel wird  $\mathbf{R}_i(\alpha)$  über die Pearson-Residuen

$$e_{ij} = \frac{Y_{ij} - \mu_{ij}}{\sqrt{v(\mu_{ij})/\omega_{ij}}}$$

parametrisiert

- Hat man einen Schätzer  $\hat{\beta}$  für  $\beta$  können diese geschätzt werden durch

$$\hat{e}_{ij} = \frac{Y_{ij} - \hat{\mu}_{ij}}{\sqrt{v(\hat{\mu}_{ij})/\omega_{ij}}},$$

wobei  $\hat{\mu}_{ij} = h(\mathbf{x}_i \hat{\beta})$

## Schätzen des Dispersionsparameters

- Der Dispersionsparameter wird analog zum GLM via

$$\hat{\phi} = \frac{1}{N - k} \sum_{i=1}^n \sum_{j=1}^{n_i} \frac{(Y_{ij} - \hat{\mu}_{ij})^2}{v(\hat{\mu}_{ij})} \omega_{ij} = \frac{1}{N - k} \sum_{i=1}^n \sum_{j=1}^{n_i} \hat{e}_{ij}^2$$

geschätzt

## Berechnung von $\hat{\beta}$

Zur Berechnung von  $\hat{\beta}$  wird folgender Algorithmus benutzt:

1. Berechne einen ersten Schätzer für  $\beta$  aus dem herkömmlichen GLM unter Annahme von Unabhängigkeit (z.B. MLE)
2. Berechne daraus die geschätzten Pearson-Residuen und damit Schätzer für  $\mathbf{R}_i(\alpha)$  und  $\phi$
3. Berechne den Schätzer der Kovarianz

$$\hat{\mathbf{V}}_i = \hat{\phi} \hat{\mathbf{A}}_i^{1/2} \mathbf{W}_i^{-1/2} \mathbf{R}_i(\hat{\alpha}) \mathbf{W}_i^{-1/2} \hat{\mathbf{A}}_i^{1/2}$$

4. Update  $\hat{\beta}$  via

$$\hat{\beta}_{r+1} = \hat{\beta}_r + \left[ \sum_{i=1}^N \mathbf{x}_i^T \mathbf{D}_i \mathbf{V}_i^{-1} \mathbf{D}_i \mathbf{x}_i \right]^{-1} \sum_{i=1}^N \mathbf{x}_i^T \mathbf{D}_i \mathbf{V}_i^{-1} (\mathbf{Y}_i - \mu_i(\beta))$$

(analog zu Fisher-Scoring)

5. Wiederhole 2.-4. bis zur Konvergenz

## Eigenschaften der Schätzer

## Eigenschaften von $\hat{\beta}$

- Sei  $\hat{\beta}$  nun die Lösung der verallgemeinerten Schätzgleichung von Folie 15
- $\hat{\beta}$  ist kein MLE, daher können wir keine Likelihoodtheorie anwenden
- Es lässt sich aber zeigen (unter gewissen Regularitätsannahmen):

$$(\hat{\beta} - \beta) \stackrel{a}{\sim} N \left( \mathbf{0}, \left[ \sum_{i=1}^N \mathbf{X}_i^T \mathbf{D}_i \mathbf{V}_i^{-1} \mathbf{D}_i \mathbf{X}_i \right]^{-1} \right)$$

## Kovarianz von $\hat{\beta}$

- Aus der letzten Folie wird ersichtlich, dass wir die Kovarianz von  $\hat{\beta}$  schätzen können als

$$\widehat{\text{Cov}}(\hat{\beta}) = \left[ \sum_{i=1}^N \mathbf{x}_i^T \mathbf{D}_i \mathbf{V}_i^{-1} \mathbf{D}_i \mathbf{x}_i \right]^{-1}$$

- Dieser Kovarianzschätzer setzt voraus, dass die Korrelationsstruktur innerhalb eines Individuums mit  $\mathbf{V}_i$  bzw.  $\mathbf{R}_i(\alpha)$  korrekt spezifiziert wurde
- In SAS wird dieser Schätzer daher mit *model based* bezeichnet

## Misspezifikation der Varianzstruktur

- In der Regel entspricht  $\mathbf{R}_i(\alpha)$  nicht der wahren Korrelationsstruktur
- $\mathbf{R}_i(\alpha)$  wird als eine *working correlation matrix* betrachtet, die nicht notwendigerweise der Wahrheit entspricht
- Man konnte zeigen, dass  $\hat{\beta}$  dann trotzdem konsistent (und approximativ normalverteilt) ist
- Misspezifikation der Varianzstruktur muss aber in der Schätzung der Kovarianz berücksichtigt werden

## Misspezifikation der Varianzstruktur

- Auch wenn  $\mathbf{R}_i(\alpha)$  nicht der wahren Korrelationsstruktur entspricht, kann man zeigen:

$$(\hat{\beta} - \beta) \stackrel{a}{\sim} N(\mathbf{0}, \widehat{\text{Cov}}_S(\hat{\beta}))$$

- Die Kovarianz wird hierbei durch den sog. *sandwich-estimate* geschätzt:

$$\widehat{\text{Cov}}_S(\hat{\beta}) = \left[ \sum_{i=1}^N \mathbf{x}_i^T \mathbf{D}_i \mathbf{V}_i^{-1} \mathbf{D}_i \mathbf{x}_i \right]^{-1} \sum_{i=1}^N \mathbf{x}_i^T \mathbf{D}_i \mathbf{V}_i^{-1} \text{Cov}(\mathbf{Y}_i) \mathbf{V}_i^{-1} \mathbf{D}_i \mathbf{x}_i \left[ \sum_{i=1}^N \mathbf{x}_i^T \mathbf{D}_i \mathbf{V}_i^{-1} \mathbf{D}_i \mathbf{x}_i \right]^{-1}$$

- $\text{Cov}(\mathbf{Y}_i)$  ist dabei unbekannt und wird durch seinen Schätzer ersetzt:

$$\widehat{\text{Cov}}(\mathbf{Y}_i) = (\mathbf{Y}_i - \mu_i(\hat{\beta}))(\mathbf{Y}_i - \mu_i(\hat{\beta}))^T$$



## Misspezifikation der Varianzstruktur

- Bei korrekt spezifizierter Kovarianzstruktur ( $\text{Cov}(\mathbf{Y}_i) = \mathbf{V}_i$ ) ist offenbar

$$\widehat{\text{Cov}}(\hat{\beta}) = \widehat{\text{Cov}}_S(\hat{\beta})$$

- Da das Normalverteilungsergebnis der letzten Folie auch gilt, wenn die Kovarianzstruktur misspezifiziert wurde, spricht man bei  $\widehat{\text{Cov}}_S(\hat{\beta})$  von einem *robusten* Varianzschätzer
- In SAS heißt  $\widehat{\text{Cov}}_S(\hat{\beta})$  *empirical covariance*

## Testen linearer Hypothesen

## Testen linearer Hypothesen

- Sind nun wieder interessiert lineare Hypothesen der Form

$$H_0 : \mathbf{C}\beta = \mathbf{d} \quad \text{vs.} \quad H_1 : \mathbf{C}\beta \neq \mathbf{d}$$

zu Testen

- Aus der Normalverteilungseigenschaft des Schätzers  $\hat{\beta}(\alpha)$  folgt, dass

$$W = (\mathbf{C}\hat{\beta} - \mathbf{d})^T [\mathbf{C}\widehat{\text{Cov}}_S(\hat{\beta})\mathbf{C}^T]^{-1} (\mathbf{C}\hat{\beta} - \mathbf{d}) \stackrel{a}{\sim} \chi_r^2$$

mit  $r = \text{Rang}(\mathbf{C})$

- Wir können  $H_0$  verwerfen, falls  $W \geq Q_r^{\chi^2}(1 - \alpha)$
- In der Praxis ersetzt man, wie erwähnt,  $\mathbf{V}_i$ ,  $\mathbf{D}_i$  und  $\text{Cov}(\mathbf{Y}_i)$  durch ihre Schätzer

## Testen linearer Hypothesen

- Es gibt auch eine GEE-Version des Scoretests (nicht in dieser VL)
- Ein Likelihood-Ratio-Test ist nicht möglich
- Für einzelne Kontrasthypothese  $H_0 : \mathbf{c}\beta = d$ , mit  $\mathbf{c} \in \mathbb{R}^k$  und  $d \in \mathbb{R}$  kann ein Gauß-Test benutzt werden
- Verwerfe dafür  $H_0$  zum Niveau  $\alpha$ , falls

$$Z = \frac{\mathbf{c}\hat{\beta} - d}{\widehat{\mathbf{cCov}_S(\hat{\beta})\mathbf{c}^T}} \geq \Phi^{-1}(1 - \alpha)$$

- Dieser Test kann auch einseitig durchgeführt werden

## Beispiele für die working correlation

## Working Correlation

- Die Schätzung von  $\beta$  via GEE ist am effizientesten, wenn die Matrix  $\mathbf{R}_i(\alpha)$  korrekt spezifiziert ist
- Wir haben gesehen, dass diese Schätzung robust gegenüber einer Misspezifizierung der Korrelationsmatrix  $\mathbf{R}_i(\alpha)$  ist (Sandwich-Kovarianz)
- In der Praxis kennen wir  $\mathbf{R}_i(\alpha)$  nicht und machen hier möglichst plausible Annahmen (working correlation)
- Im Folgenden werden typische Annahmen vorgestellt

## Unabhängigkeit

- Der einfachste Fall ist als „Arbeitshypothese“ (working correlation) von unabhängigen Beobachtungen am gleichen Individuum auszugehen

- In diesem Fall ist

$$\mathbf{R}_i(\alpha) = \mathbf{I}_{n_i}$$

- Die Matrix  $\mathbf{R}_i(\alpha)$  muss also nicht geschätzt werden

## Fixe Korrelation

- In diesem Fall wird die gesamte Korrelationsmatrix  $\mathbf{R}_i(\alpha)$  als bekannt angesehen und vorgegeben
- Die Matrix  $\mathbf{R}_i(\alpha)$  muss also nicht geschätzt werden
- Dieser Fall ist eher unrealistisch



## Austauschbare Korrelationsmatrix

- Nehmen an, dass die Beobachtungen innerhalb eines Individuums alle die gleiche Korrelation haben
- In diesem Fall ist

$$\mathbf{R}_i(\alpha) = \begin{pmatrix} 1 & \alpha & \alpha & \dots & \alpha \\ \alpha & 1 & \alpha & \dots & \alpha \\ \alpha & \alpha & 1 & \dots & \alpha \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \alpha & \alpha & \alpha & \dots & 1 \end{pmatrix}$$

- Diese working correlation ist dann sinnvoll, wenn die Beobachtungen innerhalb eines Individuums keine Zeitabhängigkeit haben und jede Umordnung der Beobachtungen zulässig ist

## Austauschbare Korrelationsmatrix

- Ein Beispiel ist, wenn das „Individuum“ eine Klinik ist und die einzelnen Beobachtungen von Patienten innerhalb der Klinik stammen
- Diese Form von Korrelationsmatrix wird auch als *exchangeable correlation* oder *compound symmetry* bezeichnet
- Der Parameter  $\alpha$  wird mit Hilfe der Pearson-Residuen geschätzt  $\hat{e}_{ij}$ :

$$\hat{\alpha} = \frac{1}{\hat{\phi}} \sum_{i=1}^N \left\{ \frac{\sum_{u=1}^{n_i} \sum_{v=1}^{n_i} \hat{e}_{iu} \hat{e}_{iv} - \sum_{u=1}^{n_i} \hat{e}_{iu}^2}{n_i(n_i - 1)} \right\}$$

- SAS verwendet den alternativen Schätzer:

$$\hat{\alpha} = \frac{1}{\hat{\phi}(N^* - k)} \sum_{i=1}^N \sum_{u < v} \hat{e}_{iu} \hat{e}_{iv},$$

wobei  $N^* = 0.5 \sum_{i=1}^N n_i(1 - n_i)$  und  $k$  die Zahl der Kovariablen ist

## AR(1) - Struktur

- Gehen nun von einem zeitlichen Verlauf der Messungen aus
- Die Korrelation zwischen zwei Messungen soll mit ihrer zeitlichen Distanz kleiner werden
- Eine Mögliche Wahl der working correlation ist dann eine AR(1) Matrix:

$$\mathbf{R}_i(\alpha) = \begin{pmatrix} 1 & \alpha & \alpha^2 & \dots & \alpha^{n_i} \\ \alpha & 1 & \alpha & \dots & \alpha^{n_i-1} \\ \alpha^2 & \alpha & 1 & \dots & \alpha^{n_i-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \alpha^{n_i} & \alpha^{n_i-1} & \alpha^{n_i-2} & \dots & 1 \end{pmatrix}$$

## AR(1) - Struktur

- Der Parameter  $\alpha$  kann nun wieder über die Pearson-Residuen geschätzt werden
- Es existieren verschiedene Ansätze dafür
- In SAS wird die Schätzung wie folgt durchgeführt:

$$\hat{\alpha} = \frac{1}{\hat{\phi}(N - k)} \sum_{i=1}^N \sum_{j \leq n_i - 1} \hat{e}_{ij} \hat{e}_{i,j+1}$$

## ***m*-dependent**

- Nimmt man an, dass Beobachtungen die (zeitlich) nah beieinander liegen korreliert sind (und die Korrelation vom Abstand abhängt), aber die Korrelation ab einem gewissen (zeitlichen) Abstand 0 ist, so kann man dies über folgende Korrelationsstruktur ausdrücken:

$$\text{Corr}(Y_{ij}, Y_{i,j+t}) = \begin{cases} 1 & t = 0 \\ \alpha_t & 0 < t \leq m \\ 0 & t > m \end{cases}$$

- Die Einträge der Matrix  $\mathbf{R}_i(\alpha)$  sind also 1 auf der Diagonalen,  $\alpha_1$  auf der ersten Nebendiagonale usw.
- Diese Korrelationsstruktur wird auch *stationary correlation* genannt

## ***m*-dependent**

- Eine Möglichkeit (in SAS implementiert) die Parameter  $\alpha_t$  zu schätzen ist:

$$\hat{\alpha}_t = \frac{1}{\hat{\phi}(N_t - k)} \sum_{i=1}^N \sum_{j \leq n_i - t} \hat{e}_{ij} \hat{e}_{i,j+t},$$

wobei  $N_t = \sum_{i=1}^N (n_i - t)$

## Unstrukturiert

- Möchte man gar keine Annahmen an die Struktur der Korrelation machen, kann man eine unstrukturierte working correlation wählen:

$$\text{Corr}(Y_{iu}, Y_{iv}) = \begin{cases} 1 & u = v \\ \alpha_{uv} & u \neq v \end{cases}$$

- In diesem Fall werden also alle paarweisen Korrelationen zwischen den Beobachtungen geschätzt
- Die Parameter  $\alpha_{uv}$  können wie folgt geschätzt werden:

$$\hat{\alpha}_{uv} = \frac{1}{\hat{\phi}(N - k)} \sum_{i=1}^N \hat{e}_{iu} \hat{e}_{iv}$$

## Unstrukturiert – Schätzprobleme

- Es ist nicht garantiert, dass die so geschätzte Matrix  $\mathbf{R}_i(\alpha)$  invertierbar ist
- Besonders bei unbalancierten Datensätzen (stark unterschiedliche Anzahlen von Messungen an den Individuen) können numerische Probleme auftreten
- In diesem Fall geht in die Schätzung der unterschiedlichen Korrelationen  $\hat{\alpha}_{uv}$  unterschiedlich viel Information ein



## Unstrukturiert - Abwandlungen

- Eine Abwandlung der unstrukturierten Korrelationsmatrix ist, anzunehmen, dass bei einem bestimmten (zeitlichen) Abstand der Messungen keine Korrelation mehr vorliegt und ansonsten keine Annahmen zu treffen
- In diesem Fall währe die Korrelation spezifiziert über

$$\text{Corr}(Y_{iu}, Y_{iv}) = \begin{cases} 1 & u = v \\ \alpha_{uv} & 0 < |u - v| \leq m \\ 0 & |u - v| > m \end{cases}$$

- Die Parameter  $\alpha_{uv}$  können genauso geschätzt werden wie im unstrukturierten Fall, es gibt dieselben numerischen Probleme
- Diese Art der Korrelationsstruktur wird auch *nonstationary* genannt

## GEE vs. (Generalized) Linear Mixed Model

## Beispieldaten – Fußnägel

- In einer Studie zur Behandlung von Infektionen von Fußnägeln wurden 378 Patienten zu 2 Behandlungen randomisiert
- Behandlungsdauer 3 Monate
- Beobachtungszeitraum 12 Monate
- $Y_{ij}$  sei eine Binärvariable für schwere Entzündungen von Patient  $i$  zur Messung  $j$  (Zeitpunkt  $t_{ij}$ )
- Messungen nach 0, 1, 2, 3, 6, 9, 12 Monaten

## Gemischtes Modell in Exponentialfamilien (kurz)

- Bislang gemischte Modelle nur für normalverteilte Endpunkte betrachtet
- Verallgemeinerung auf Daten deren Verteilung eine Exponentialfamilie bildet ist möglich (Generalized Linear Mixed Model – GLMM)
- Sei  $\mathbf{b}_i \sim N(0, \mathbf{D})$  ein Vektor von zufälligen Effekten für das Individuum  $i$ . Beim verallgemeinerten gemischten linearen Modell gehen wir davon aus, dass

$$f(y_{ij}|\mathbf{b}_i, \theta_{ij}, \phi, \omega_{ij}) = f(y_{ij}|\theta_{ij}) = \exp \left\{ \frac{y_{ij}\theta_{ij} - b(\theta_{ij})}{\phi} \omega_{ij} - c(y_{ij}, \phi, \omega_{ij}) \right\},$$

wobei  $\mu_{ij} = E(Y_{ij}) = b'(\theta_{ij}) = h(\mathbf{x}_{ij}\beta + \mathbf{z}_{ij}\mathbf{b}_i)$ , für eine Responsefunktion  $h$  und Kovariablenvektoren  $\mathbf{x}_{ij}$  und  $\mathbf{z}_{ij}$

## Gemischtes Modell in Exponentialfamilien (kurz)

- Wie gemischten Modell für normalverteilte Daten wird also die Kovarianzstruktur über die zufälligen Effekte Modelliert
- Die Berechnung der marginalen Verteilung von  $Y_i$  und der Likelihood der Daten ist aufwendiger also zuvor
- In der Regel benötigt man numerische Approximation von Integralen (hier nicht behandelt)
- Es existieren Verfahren die Regressionskoeffizienten zu schätzen und Tests zu konstruieren
- Ein Spezialfall sind Binärdaten wie im Fußnagel datensatz

## Vergleich GEE vs. (Generalized) Linear Mixed Model

- Wollen nun die Ergebnisse einer GEE-Analyse und einer Mixed Model Analyse des Fußnageldatensatzes miteinander vergleichen
- Angepasstes GEE-Modell:

$$Y_{ij} \sim B(1, \pi_{ij}), \quad \log \left( \frac{\pi_{ij}}{1 - \pi_{ij}} \right) = \beta_0 + \beta_1 t_{ij},$$

unstrukturierte working correlation

- Angepasstes GLMM:

$$Y_{ij} | b_i \sim B(1, \pi_{ij}), \quad \log \left( \frac{\pi_{ij}}{1 - \pi_{ij}} \right) = \beta_0 + b_i + \beta_1 t_{ij} \quad (\text{random intercept})$$

## Vergleich GEE vs. (Generalized) Linear Mixed Model

- Die geschätzten Regressionskoeffizienten aus den beiden Modellen sind wie folgt:

Parameter	GEE	GLMM
	Estimate (s.e.)	Estimate (s.e.)
Intercept group A	-0.7219 (0.1656)	-1.6308 (0.4356)
Intercept group B	-0.6493 (0.1671)	-1.7454 (0.4478)
Slope group A	-0.1409 (0.0277)	-0.4043 (0.0460)
Slope group B	-0.2548 (0.0380)	-0.5657 (0.0601)

- Sehr große Unterschiede zwischen den Modellen. Erklärung?

## Interpretation der Koeffizienten im GEE

- Im GEE modellieren wir mit den Regressionskoeffizienten den Erwartungswert von  $Y_{ij}$ :

$$E(Y_{ij}) = \frac{\exp(\beta_0 + \beta_1 t_{ij})}{1 + \exp(\beta_0 + \beta_1 t_{ij})}$$

- Es wird also der Erwartungswert innerhalb der Population modelliert
- Man spricht daher auch von einem *population average (PA)* Modell



## Interpretation der Koeffizienten im GLMM

- Für das betrachtete GLMM haben wir

$$Y_{ij}|b_i \sim B(1, \pi_{ij}), \quad \log \left( \frac{\pi_{ij}}{1 - \pi_{ij}} \right) = \beta_0 + b_i + \beta_1 t_{ij} \quad (\text{random intercept})$$

- Wir modellieren also den Bedingten erwartungswert von  $Y_{ij}$  gegeben den Randomeffekt  $b_i$
- In Formeln heißt dies:

$$E(Y_{ij}|b_i) = \frac{\exp(\beta_0 + b_i + \beta_1 t_{ij})}{1 + \exp(\beta_0 + b_i + \beta_1 t_{ij})}$$

## Interpretation der Koeffizienten im GLMM

- Für das betrachtete GLMM haben wir

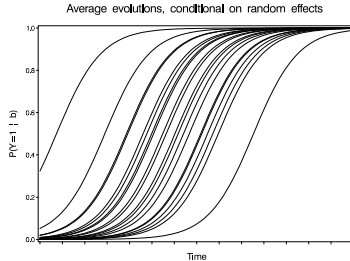
$$Y_{ij}|b_i \sim B(1, \pi_{ij}), \quad \log \left( \frac{\pi_{ij}}{1 - \pi_{ij}} \right) = \beta_0 + b_i + \beta_1 t_{ij} \quad (\text{random intercept})$$

- Wir modellieren also den Bedingten erwartungswert von  $Y_{ij}$  gegeben den Randomeffekt  $b_i$
- In Formeln heißt dies:

$$E(Y_{ij}|b_i) = \frac{\exp(\beta_0 + b_i + \beta_1 t_{ij})}{1 + \exp(\beta_0 + b_i + \beta_1 t_{ij})}$$

## Interpretation der Koeffizienten im GLMM

- Man spricht daher auch von einem *subject specific (SS)* Modell
- Im GLMM modellieren wir hier also den mittleren Verlauf des Erwartungswertes bei gegebenen Randomeffects:

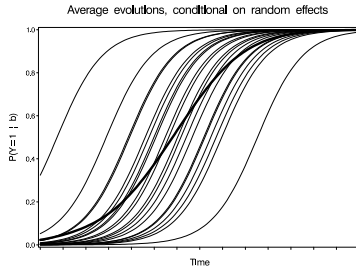


Mittlerer Verlauf gegeben die Randomeffects – aus Verbeke & Molenberghs (Vorlesungsfolien)

## Interpretation der Koeffizienten im GLMM

- Aus dem GLMM kann man das Modell für den marginalen (PA) Verlauf erhalten, indem man über die Randeffects mittelt:

$$E(Y_{ij}) = E[E(Y_{ij}|b_i)] = E \left[ \frac{\exp(\beta_0 + b_i + \beta_1 t_{ij})}{1 + \exp(\beta_0 + b_i + \beta_1 t_{ij})} \right] \neq \frac{\exp(\beta_0 + \beta_1 t_{ij})}{1 + \exp(\beta_0 + \beta_1 t_{ij})}$$



Mittlerer Verlauf gegeben die Randeffects – aus Verbeke & Molenberghs (Vorlesungsfolien)

## Interpretation der Koeffizienten (GEE vs. GLMM)

- Im die GEE-Parameter und die GLMM-Parameter müssen unterschiedlich interpretiert werden
  - GEE: Marginal (PA)
  - GLMM: Bedingt auf den Randomeffekt (SS)
- Im Allgemeinen ist das Modell für den marginalen Erwartungswert, welches aus dem GLMM folgt (durch Mitteln über die Randomeffects) nicht von der gleichen Form wie das Modell für den bedingten Erwartungswert im GLMM
- Es gibt Spezialfälle, in denen beide Modelle (fast) die gleiche Form haben (z.B. das betrachtete GLMM für Binärdaten mit normalverteilten random intercepts)

## Interpretation der Koeffizienten (GEE vs. GLMM)

- Die Tatsache, dass das Modell für den marginalen Erwartungswert, welches aus dem GLMM folgt nicht die gleiche Form hat, wie das GLMM selbst, kommt daher, dass im allgemeinen:

$$E[g(Y)] \neq g[E(Y)]$$

- Wenn also Randeffects nicht-linear in den bedingten Erwartungswert des GLMM eingehen, müssen die Parameter des marginalen Modells anders interpretiert werden, als die des GLMM
- Im Falle von linearen Modellen, stellt dies also kein Problem dar
- In der Praxis kann der marginale Erwartungswert aus dem GLMM abgeleitet werden, indem über die Randeffects gemittelt (integriert) wird (numerische Verfahren)





## Erweiterungen für GEEs

## Erweiterungen

- Es gibt auch ansätze für subject specific GEEs (arbeiten mit Randomeffects)
- Eine Erweiterung der GEE sind die *second order GEE (GEE2)* bei denen auch die Kovarianzstruktur mitmodelliert wird
- Für weitere Details siehe Literaturliste



## Literatur

-  K.-Y. Liang und S.L. Zeger.  
Longitudinal data analysis using generalized linear models.  
*Biometrika*, 73:13–22, 1986.
-  J.W. Hardin und J.M. Hilbe.  
Generalized Estimating Equations (Second Edition).  
*Chapman and Hall/CRC* , New York, 2012.
-  G. Verbeke und G. Molenberghs.  
Introduction to Logitudinal Data Analysis.  
*Vorlesungsfolien*, <https://gbiomed.kuleuven.be/english/research/50000687/50000696/geertverbeke/cursussen/longitudinal.pdf>.
-  A. Ziegler.  
Generalized Estimating Equations.  
*Springer*, New York Heidelberg Dodrecht London, 2011.