

# Tutorium 4

## Analyse longitudinaler Daten

Prof. Dr. Sonja Greven, Dipl. Stat. Jona Cederbaum,  
Alexander Bauer

14. Juni 2016

# Übersicht

# Serielle Korrelation

# Serielle Korrelation

**Korrelationsstrukturen** im LLMM (Bsp. mit  $n_i = 3$ ):

- **Conditional independence:**  $\Sigma_i = \begin{pmatrix} \sigma^2 & 0 & 0 \\ 0 & \sigma^2 & 0 \\ 0 & 0 & \sigma^2 \end{pmatrix}$
- **serielle Korrelation:**  $\epsilon_i = \epsilon_{(1)i} + \epsilon_{(2)i}, \quad \epsilon_{(1)i} \perp \epsilon_{(2)i}$

mit  $\text{Cov}(\epsilon_{(2)i}) = \tau^2 \mathbf{H}_i$  und somit  $\Sigma_i = \tau^2 \mathbf{H}_i + \sigma^2 \mathbf{I}_{n_i}$ .

Spezialfälle:

- **Compound symmetry:**  $\Sigma_i = \begin{pmatrix} \sigma_1^2 & \sigma_2 & \sigma_2 \\ \sigma_2 & \sigma_1^2 & \sigma_2 \\ \sigma_2 & \sigma_2 & \sigma_1^2 \end{pmatrix}$
- **Unstructured:**  $\Sigma_i = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 \end{pmatrix}$
- **AR(p)-Prozess:**  $\epsilon_i = \sum_{k=1}^p \rho_k \epsilon_{i-k} + \phi_t$  mit weißem Rauschen  $\phi_t$
- **Monoton abnehmende Funktion**  $g(\cdot)$  als Korrelationsstruktur

# Serielle Korrelation

Modell mit **serieller Korrelation** und Funktion  $g(\cdot)$ :

Wähle  $h_{ijk} = g(|t_{ij} - t_{ik}|)$  mit  $g(\cdot)$  monoton sinkend mit  $g(0) = 1$ .

**Wahl von  $g(\cdot)$ :**

- Gängige Strukturen:

$$\text{Exponential : } \tau^2 g(u) = \tau^2 \exp(-\phi u)$$

$$\text{Gauß : } \tau^2 g(u) = \tau^2 \exp(-\phi u^2)$$

aus der Power exponential family

- Alternativen: Matérn-Familie, Fraktionelle Polynome

# Serielle Korrelation

## Das Semi-Variogramm:

- Motivation:
  - Ist eine allgemeinere Korrelationsstruktur als  $\Sigma_i = \sigma^2 \mathbf{1}_{n_i}$  nötig?
  - Wie sollte die Korrelationsstruktur aussehen (vor Modellschätzung)?
  - Ist die spezifizierte Korrelationsstruktur sinnvoll (nach Schätzung)?
- Semi-Variogramm zur Charakterisierung der Kovarianz:  
(i.A. nur für stationäre  $\{Y(t), t \in \mathbb{R}\}$ )

$$\nu(u) = \frac{1}{2} \mathbb{E}[\{Y(t) - Y(t-u)\}^2]$$

mit  $u$  dem absoluten zeitlichen Abstand zweier Beobachtungen.

# Serielle Korrelation

## Das Semi-Variogramm:

- Modell nur mit einem Random Intercept:

$$Y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + b_i + \epsilon_{(1)ij} + \epsilon_{(2)ij}$$

Betrachtung von  $Y_{ij}^c = Y_{ij} - \mathbf{x}_{ij}^T \boldsymbol{\beta}$  als stationärem Prozess liefert

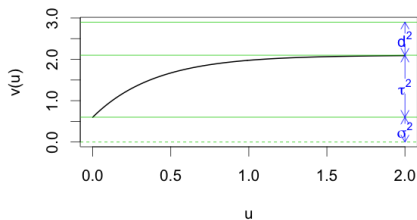
$$v(u) = \sigma^2 + \tau^2(1 - g(u)).$$

## Eigenschaften:

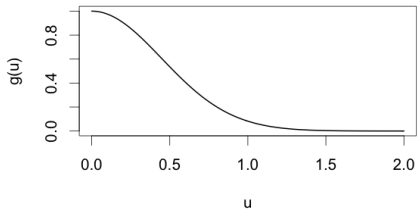
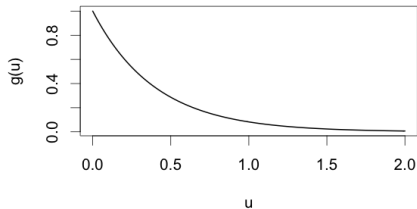
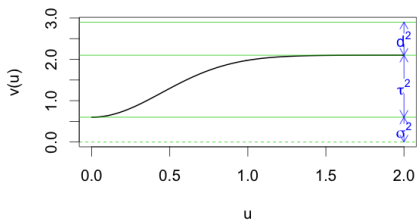
- Abnehmendes  $g \Leftrightarrow$  steigendes Semi-Variogramm
- Nugget-Effekt:  $v(0) = \sigma^2$
- $\lim_{u \rightarrow \infty} v(u) = \sigma^2 + \tau^2 \leq \text{Var}(Y_{ij}^c) = d^2 + \sigma^2 + \tau^2$
- Modell mit Random Slopes:  $Y_{ij}^c$  nicht mehr stationär  
 $\Rightarrow$  Semi-Variogramm nicht anwendbar!

## Serielle Korrelation

Exponential



Gauss





## Serielle Korrelation

### Das empirische Semi-Variogramm:

Datenbasierte Schätzung des Semi-Variogramms anhand der Residuen

Vorgehen für Random Intercept-Modell:

- 1) Schätze das lineare Modell  $Y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + \epsilon_{ij}$  mittels KQ-Schätzung
- 2) Berechne  $v_{ijk} = \frac{1}{2}(r_{ij} - r_{ik})^2$   
zwischen Residuenpaaren  $(r_{ij}, r_{ik})$  am gleichen Subjekt  
und mit  $r_{ij} = y_{ij} - \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}}$
- 3) Plote  $v_{ijk}$  gegen zeitliche Abstände  $u = |t_{ij} - t_{ik}|$
- 4) Glätte  $(u, v_{ijk})$  (z.B. per LOESS) und erhalte dadurch das empirische Semi-Variogramm  $\hat{v}(\cdot)$

Vorgehen für Modell mit Random Slope:

⇒ Betrachtung der subjekt-spezifischen Residuen

# Modelldiagnostik

# Modelldiagnostik

## Mögliche Residuen:

- Populationsspezifische Residuen:  $\mathbf{y}_i - \mathbf{X}_i \hat{\boldsymbol{\beta}}$   
Problem: Residuen sind korreliert und heteroskedastisch
- Subjektspezifische Residuen:  $\mathbf{y}_i - \mathbf{X}_i \hat{\boldsymbol{\beta}} - \mathbf{Z}_i \hat{\mathbf{b}}_i$   
Problem:  $\hat{\mathbf{b}}_i$  abhängig von Annahmen (Normalität, Varianzstruktur)  
⇒ Deshalb: Betrachtung von **transformierten Residuen**

## Transformierte Residuen: $\mathbf{r}_i^* = \mathbf{L}_i^{-1} \mathbf{r}_i$

mit  $\mathbf{r}_i = \mathbf{y}_i - \mathbf{X}_i \hat{\boldsymbol{\beta}}$  und der Cholesky-Zerlegung  $\hat{\mathbf{V}}_i = \mathbf{L}_i \mathbf{L}_i^T$ .

⇒  $\mathbf{r}_i^*$  approximativ unkorreliert mit Mittelwert 0 und Varianz 1

⇒ Plotten der  $\mathbf{r}_i^*$  gegen den transformierten Prädiktor  $\hat{\boldsymbol{\mu}}_{ij}^* = \mathbf{L}_i^{-1} \mathbf{X}_i \hat{\boldsymbol{\beta}}$

# Modelldiagnostik

Wiederholung: **Cholesky-Zerlegung**

Eindeutige Zerlegung der Matrix  $\mathbf{A}$  durch eine untere Dreiecksmatrix  $\mathbf{L}$ :

$$\mathbf{A} = \mathbf{L}\mathbf{L}^T$$

Möglich für jede symmetrische, positiv definite  $n \times n$ -Matrix  $\mathbf{A}$ .

Warum funktioniert das Dekorrelieren durch  $\mathbf{L}$ ?

$$\begin{aligned} \hat{\mathbf{V}}_i &= \mathbf{L}_i \mathbf{L}_i^T, & \Rightarrow \mathbf{L}_i^{-1} \hat{\mathbf{V}}_i \mathbf{L}_i^{-1,T} &= \mathbf{I}_{n_i} \\ \text{Cov}(\mathbf{r}_i^*) &= \text{Cov}(\mathbf{L}_i^{-1} \mathbf{r}_i) \\ &\stackrel{*}{=} \mathbb{E}((\mathbf{L}_i^{-1} \mathbf{r}_i) \cdot (\mathbf{L}_i^{-1} \mathbf{r}_i)^T) \\ &= \mathbb{E}(\mathbf{L}_i^{-1} \mathbf{r}_i \mathbf{r}_i^T \mathbf{L}_i^{-1,T}) = \mathbf{L}_i^{-1} \mathbb{E}(\mathbf{r}_i \mathbf{r}_i^T) \mathbf{L}_i^{-1,T} \\ &= \mathbf{L}_i^{-1} \text{Cov}(\mathbf{r}_i) \mathbf{L}_i^{-1,T} = \mathbf{L}_i^{-1} \hat{\mathbf{V}}_i \mathbf{L}_i^{-1,T} = \mathbf{I}_{n_i}. \end{aligned}$$

\*: da  $\text{Cov}(\mathbf{X}) = \mathbb{E}(\mathbf{X}\mathbf{X}^T) - \mathbb{E}(\mathbf{X})\mathbb{E}(\mathbf{X}^T)$  und  $\mathbb{E}(\mathbf{L}_i^{-1} \mathbf{r}_i) = \mathbf{L}_i^{-1} \mathbb{E}(\mathbf{r}_i) = \mathbf{0}$ .

# Modelldiagnostik

## Ermitteln auffälliger Subjekte:

Durchführung je eines Tests pro Subjekt:  $d_i = \mathbf{r}_i^{*T} \mathbf{r}_i^* \stackrel{a}{\sim} \chi_{n_i}^2$

Beachten: p-Werte  $< \alpha$  werden in  $\alpha N$  Fällen erwartet

## Überprüfung der Kovarianzstruktur:

$\mathbf{r}_i^*$  sind unkorreliert mit Mittelwert 0 und Varianz 1

⇒ Semi-Variogramm sollte bei korrekt spezifizierter Kovarianz zufällig um 1 streuen

# Wiederholung: GLM

# Wiederholung: GLM

## Wieso Generalisierte Modelle?

Da Response teilweise nicht metrisch oder NV-Annahme nicht optimal

### Beispiel I: Binärer (0/1) Response

$Y_i$ : Person  $i$  kauft bestimmtes Produkt ja/nein

### Beispiel II: Zähldaten als Response

$Y_i$ : Anzahl an epileptischen Anfällen von Person  $i$  in bestimmtem Zeitraum

**Umsetzung:** Betrachtung von Verteilungen aus der Exponentialfamilie

⇒ einheitliches Framework (Schätzung, Inferenz, etc.)

# Wiederholung: GLM

Betrachtung von Verteilungen aus der **Exponentialfamilie**:

$$f(y|\theta, \phi) = \exp \left\{ \frac{y\theta - \psi(\theta)}{\phi} + c(y, \phi) \right\}$$

mit natürlichem Parameter  $\theta$ , Skalen- bzw. Dispersionsparameter  $\phi$ , Funktionen  $\psi(\cdot)$ ,  $c(\cdot, \cdot)$ .

Eigenschaften der ersten beiden Momente:

$$\begin{aligned} \mathbb{E}(Y) &= \mu = \psi'(\theta) \\ \text{Var}(Y) &= \sigma^2 = \phi\psi''(\theta) \\ &\quad \phi\psi''(\psi'^{-1}(\mu)) =: \phi v(\mu), \end{aligned}$$

mit  $v(\mu)$  der Varianzfunktion.



# Wiederholung: GLM

## Notation:

- Stichprobe mit  $i = 1, \dots, N$  Beobachtungen
- $Y_i$ : unabhängige zufällige Zielgrößen
- $\mathbf{x}_i$ : Vektoren der  $p$  Kovariablen

## Annahmen:

- $Y_i \stackrel{u}{\sim} \text{Expo-Fam.}(\theta_i, \phi)$
- Darstellung von  $\mathbb{E}(Y_i) = \mu_i$  durch **Responsefunktion**  $h(\cdot)$ :

$$\mu_i = h(\eta_i) = h(\mathbf{x}_i^T \boldsymbol{\beta})$$

(bzw. Darstellung durch zugehörige **Linkfunktion**  $g(\cdot) = h^{-1}(\cdot)$ ).

**Anmerkung:** Benutzung des natürlichen (kanonischen) Links  $h(\cdot) = \psi'(\cdot)$  führt zu linearem Zusammenhang  $\theta_i = \mathbf{x}_i^T \boldsymbol{\beta}$ .

# Wiederholung: GLM

Notwendige Komponenten für **Definition eines GLM:**

1) **Verteilungsannahme:**

$$Y_i \stackrel{u.}{\sim} \text{Expo-Fam.}(\theta_i, \phi)$$

2) **Systematische Komponente:** Spezifizierung des linearen Prädiktors

$$\eta_i = \mathbf{x}_i^T \boldsymbol{\beta}.$$

3) **Linkfunktion:** Wahl der Linkfunktion  $g(\cdot)$

$$g(\mu_i) = \eta_i$$

# Wiederholung: GLM

**GLM mit Normalverteilung** -  $Y_i \stackrel{u.}{\sim} \mathcal{N}(\mu_i, \sigma^2)$

$$\begin{aligned}
 f(y_i) &= \exp \left\{ -\frac{1}{2} \left( \frac{y_i - \mu_i}{\sigma} \right)^2 - \log(\sqrt{2\pi\sigma^2}) \right\} \\
 &= \exp \left\{ \frac{y_i\mu_i - \mu_i^2/2}{\sigma^2} - \frac{y_i^2}{2\sigma^2} - \log(\sqrt{2\pi\sigma^2}) \right\}.
 \end{aligned}$$

NV gehört zur Expo-Fam. mit  $\theta_i = \mu_i$ ,  $\psi(\theta_i) = \theta_i^2/2$  und  $\phi = \sigma^2$ .

- Verwendung der Varianzfunktion  $v(\mu_i) = 1$
- Verwendung der Identität als natürliche Linkfunktion:

$$g(\mu_i) = \mu_i = \eta_i = \mathbf{x}_i^T \boldsymbol{\beta}$$

# Wiederholung: GLM

**GLM mit Bernoulliverteilung** -  $Y_i \stackrel{u.}{\sim} \mathcal{B}(\pi_i)$

$$f(y_i) = \exp \left\{ y_i \log \left( \frac{\mu_i}{1 - \mu_i} \right) + \log(1 - \mu_i) \right\},$$

mit  $\mu_i = \pi_i$ .  $\mathcal{B}(\pi_i)$  gehört zur Expo-Fam. mit  $\theta_i = \log \left( \frac{\mu_i}{1 - \mu_i} \right)$ ,  
 $\psi(\theta_i) = \log(1 + \exp(\theta_i))$  und  $\phi = 1$ .

- Varianzfunktion:  $v(\mu_i) = \mu_i(1 - \mu_i)$
- Logit-Link als natürlicher Link:

$$g(\mu_i) = \log \left( \frac{\mu_i}{1 - \mu_i} \right) = \log \left( \frac{\pi_i}{1 - \pi_i} \right) = \mathbf{x}_i^T \boldsymbol{\beta}$$

$$\Rightarrow P(Y = 1 | \mathbf{x}_i) = \pi_i = \frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})}.$$

# Wiederholung: GLM

**GLM mit Poissonverteilung** -  $Y_i \stackrel{u}{\sim} Po(\lambda_i)$

$$f(y_i) = \exp \{y_i \log \mu_i - \mu_i - \log y_i!\},$$

mit  $\mu_i = \lambda_i$ .  $Po(\lambda_i)$  gehört zur Expo.-Fam. mit  $\theta_i = \log(\mu_i)$ ,  
 $\psi(\theta_i) = \exp(\theta_i)$  und  $\phi = 1$ .

- Varianzfunktion:  $v(\mu_i) = \mu_i$
- Log-Link als natürlicher Link:

$$g(\mu_i) = \log(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta}$$
$$\Rightarrow \mu_i = \exp(\mathbf{x}_i^T \boldsymbol{\beta}).$$

# Wiederholung: GLM

## Übersicht der betrachteten Spezialfälle:

Verteilung	$\phi$	Varianzfunktion	natürliche Linkfunktion
Normal	$\sigma^2$	$v(\mu) = 1$	$\mu = \eta$ (Identität)
Bernoulli	1	$v(\mu) = \mu(1 - \mu)$	$\log\left(\frac{\mu}{1-\mu}\right) = \eta$ (Logit)
Poisson	1	$v(\mu) = \mu$	$\log(\mu) = \eta$ (Log)

# Wiederholung: GLM

## ML-Schätzung:

1) Likelihood für unabhängige Beobachtungen:

$$L(\boldsymbol{\beta}, \phi) = \prod_{i=1}^N \exp \left[ \frac{y_i \theta_i - \psi(\theta_i)}{\phi} + c(y_i, \phi) \right],$$

mit  $\theta_i$  abhängig von  $\boldsymbol{\beta}$ .

2) Log-Likelihood:

$$l(\boldsymbol{\beta}, \phi) = \frac{1}{\phi} \sum_{i=1}^N [y_i \theta_i - \psi(\theta_i)] + \sum_{i=1}^N c(y_i, \phi).$$

# Wiederholung: GLM

## ML-Schätzung:

3) Ableiten nach  $\beta$  und Nullsetzen ( $\frac{\partial l}{\partial \beta} = 0$ )  $\rightarrow$  Score-Gleichungen:

$$S(\beta) = \sum_{i=1}^N \frac{\partial \theta_i}{\partial \beta} [y_i - \psi'(\theta_i)] = \mathbf{0}.$$

4) Unter Benutzung des natürlichen Links ergibt sich schließlich:

$$\sum_{i=1}^N \mathbf{x}_i (y_i - \mu_i) = \mathbf{0}.$$

(gilt für alle Verteilungen aus der Exponentialfamilie!)

$\Rightarrow$  Lösen der Gleichungen führt zu  $\hat{\beta}_{ML}$



# Wiederholung: GLM

**Alternative Formulierung** der Score-Gleichungen:

wird später bei GEEs (Generalized Estimating Equations) benötigt

$$S(\beta) = \sum_{i=1}^N \frac{\partial \theta_i}{\partial \beta} [y_i - \psi'(\theta_i)] = \mathbf{0}.$$

Aus  $\mu_i = \psi'(\theta_i)$  und  $v_i := v(\mu_i) = \psi''(\theta_i)$  folgt

$$\frac{\partial \mu_i}{\partial \beta} = \psi''(\theta_i) \frac{\partial \theta_i}{\partial \beta} = v_i \frac{\partial \theta_i}{\partial \beta}$$

und somit

$$S(\beta) = \sum_{i=1}^N \frac{\partial \mu_i}{\partial \beta} v_i^{-1} [y_i - \mu_i] = \mathbf{0}.$$

# Wiederholung: GLM

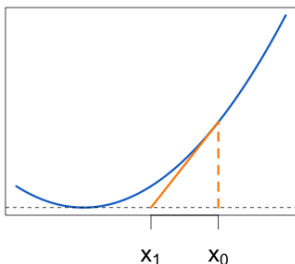
## ML-Schätzung:

Meist Optimierung durch iterative Algorithmen notwendig.

## Newton-Raphson-Algorithmus:

- Univariater Fall: Minimum von  $f(x)$  bestimmen (mit Startwert  $x_0$ )  
( $\Leftrightarrow$  Nullstelle von  $f'(x)$  bestimmen)

$$x_{n+1} = x_n - \frac{f'(x_n)}{f''(x_n)}, \quad n \geq 0$$



# Wiederholung: GLM

## Newton-Raphson-Algorithmus:

- Multivariater Fall: Minimum von  $f(\mathbf{x})$  bestimmen

$$\mathbf{x}_{n+1} = \mathbf{x}_n - \mathbf{H}^{-1} \nabla f(\mathbf{x}_n)$$

mit  $\mathbf{H} = \nabla^2 f(\mathbf{x}_n)$  der Hesse-Matrix mit Elementen  $\mathbf{H}_{i,j} = \frac{\partial^2 f}{\partial x_i \partial x_j}$

- **Fisher-Scoring:** Spezialfall von Newton-Raphson  
Verwendung der Fisher-Information (EW der Hesse-Matrix) anstelle der Hesse-Matrix. Führt u.a. zu schnellerer Berechnung.