Tutorium 4

Analyse longitudinaler Daten

Prof. Dr. Sonja Greven, Almond Stöcker,
Johanna Völkl

Material: Alexander Bauer

04. Juli 2017

Übersicht

Serielle Korrelation

Serielle Korrelation

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

- **Conditional independence:** $\mathbf{\Sigma}_i = \begin{pmatrix} \sigma^2 & 0 & 0 \\ 0 & \sigma^2 & 0 \\ 0 & 0 & \sigma^2 \end{pmatrix}$
- **serielle Korrelation:** $\boldsymbol{\epsilon}_i = \boldsymbol{\epsilon}_{(1)i} + \boldsymbol{\epsilon}_{(2)i}, \qquad \epsilon_{(1)i} \perp \boldsymbol{\epsilon}_{(2)i}$

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

  Spezialfälle:

  - **Compound symmetry:** $\mathbf{\Sigma}_i = \begin{pmatrix} \sigma_1^2 & \sigma_2 & \sigma_2 \\ \sigma_2 & \sigma_1^2 & \sigma_2 \\ \sigma_2 & \sigma_2 & \sigma_1^2 \end{pmatrix}$
  - **Unstructured:** $\mathbf{\Sigma}_i = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 \end{pmatrix}$
  - **AR(p)-Prozess:** $\boldsymbol{\epsilon}_i = \sum_{k=1}^{p} \rho_k \boldsymbol{\epsilon}_{i-k} + \phi_t$ mit weißem Rauschen ϕ_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 $\mathbf{\Sigma}_i = \sigma^2 \mathbf{I}_{n_i}$ 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 $\{Y(t), t \in \mathbb{R}\}$)

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

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

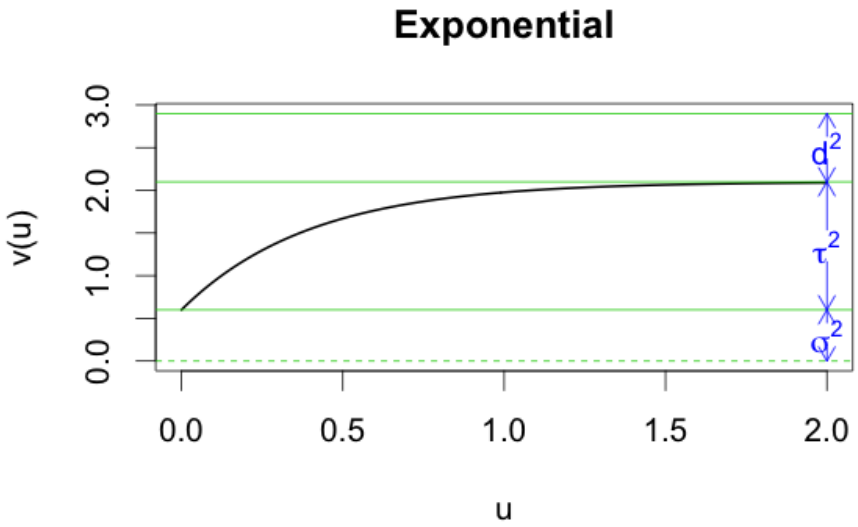
$$v(u) = \sigma^2 + \tau^2(1 - g(u)).$$

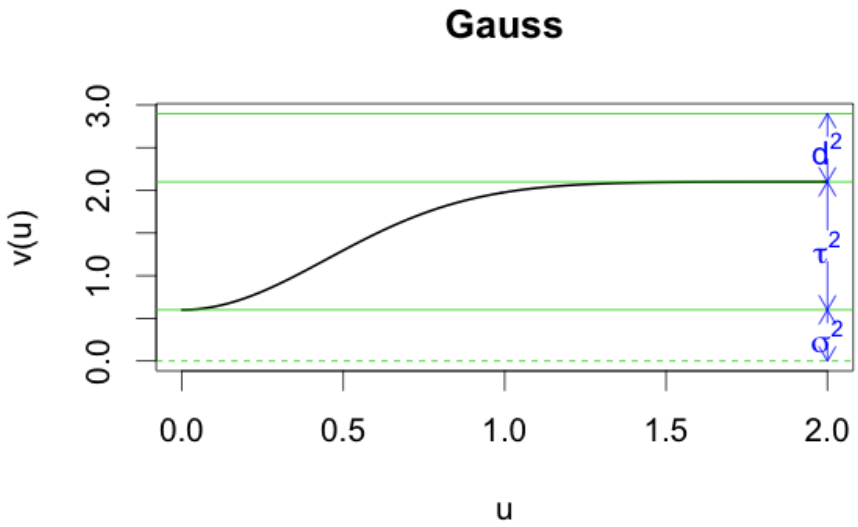
Eigenschaften:

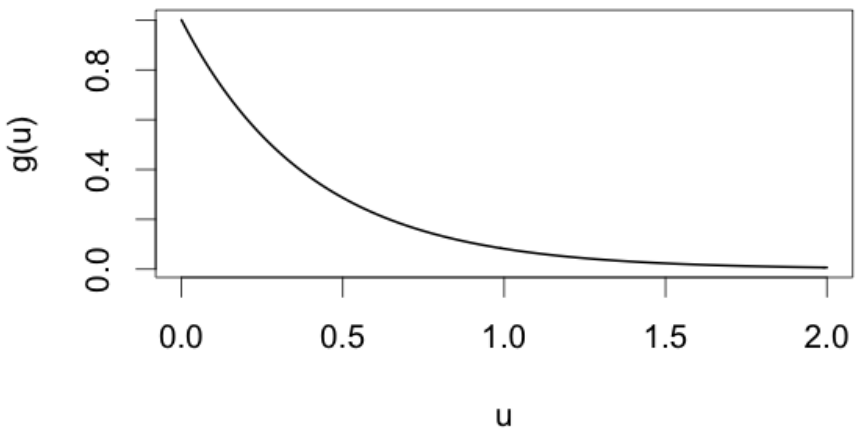
- Abnehmendes $g \Leftrightarrow$ steigendes Semi-Variogramm
- Nugget-Effekt: $v(0) = \sigma^2$
- $\lim_{u \to \infty} v(u) = \sigma^2 + \tau^2 \leq Var(Y_{ij}^c) = d^2 + \sigma^2 + \tau^2$

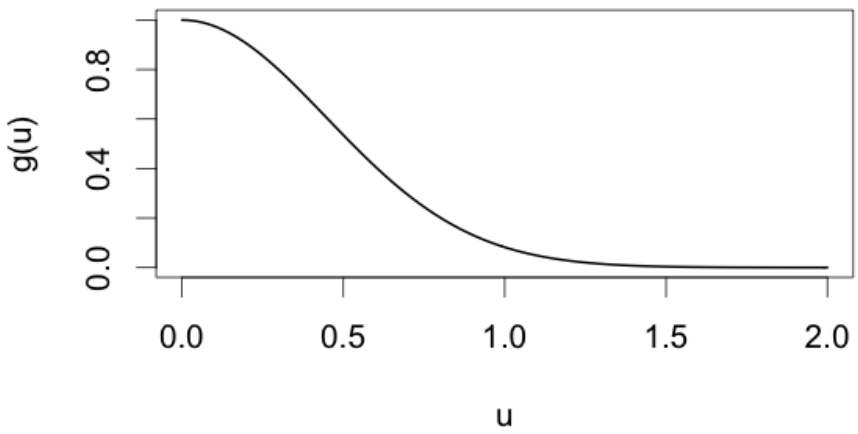
- Modell mit Random Slopes: Y_{ij}^c nicht mehr stationär
  $\Rightarrow$ Semi-Variogramm nicht anwendbar!

Serielle Korrelation









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) Plotte 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 $\hat{v}(\cdot)$

Modelldiagnostik

Modelldiagnostik

Mögliche Residuen:

- Populationsspezifische Residuen: $\mathbf{y}_i - \mathbf{X}_i\hat{\beta}$

  Problem: Residuen sind korreliert und heteroskedastisch
- Subjektspezifische Residuen: $\mathbf{y}_i - \mathbf{X}_i\hat{\beta} - \mathbf{Z}_i\hat{\mathbf{b}}_i$

  Problem: $\hat{\mathbf{b}}_i$ abhängig von Annahmen (Normalität, Varianzstruktur)

  $\Rightarrow$ Deshalb: Betrachtung von **transformierten Residuen**

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

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

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

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

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 $\mathbf{L}$?

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

*: da $Cov(\mathbf{X}) = \mathbb{E}(\mathbf{X}\mathbf{X}^T) - \mathbb{E}(\mathbf{X})\mathbb{E}(\mathbf{X}^T)$ 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: $d_i = \mathbf{r}_i^{*T} \mathbf{r}_i^* \overset{a}{\sim} \chi^2_{n_i}$

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

Überprüfung der Kovarianzstruktur:

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

$\Rightarrow$ Semi-Variogramm sollte bei korrekt spezifizierter Kovarianz zufällig um 1 streuen

Modelldiagnostik

Random Effects Assumption

Random effects:

- Idee: Auffangen von Effekten nicht gemessener / messbarer subjekt-spezifischer Kovariablen

  $\Rightarrow$ Beispiel: Random Intercept zur Berücksichtigung individueller CD4-Niveaus

- Zentrale Annahme: **Random effects assumption**

$$\forall i, j: \ \mathbb{E}(b_i | x_{ij}) = 0$$

$\Rightarrow$ Unabhängigkeit der b_i von den Kovariablen

Auswirkungen:

- Annahme erfüllt: Schätzer sind effizienter als im fixed effect model
- Annahmeverletzung: Schätzer nicht mehr konsistent!

Random Effects Assumption

- Beispiel: *Fall-Kontroll-Studie zum Testen eines Medikaments*

$$CD4_{ij} = \beta_0 + \beta_1 Medikament_{ij} + \beta_2 Alter_{ij} + b_{i0} + \epsilon_{ij}$$

  - Annahme ist verletzt, wenn Personen mit überdurchschnittlichem CD4-Level länger leben
  - Annahme ist verletzt, wenn es Confounder-Variable gibt, welche sowohl auf x als auch auf y wirkt:

    Hypothetisches Bsp.: *Confounder Rauchen*
    Nichtraucher leben länger und haben höheres CD4-Niveau
    $\Rightarrow$ Wenn Rauchen nicht im Modell ist wandert Effekt teilweise in b_{i0}
    $\Rightarrow$ Dadurch korrelieren $Alter_{ij}$ und b_{i0}
  - Annahme ist bzgl. Medikamenteneffekt immer erfüllt, wenn Aufteilung in Fall- und Behandlungsgruppe randomisiert wurde

Hausman-Test

Hausman-Test: zur Überprüfung der Annahme

$$T = (\hat{\beta}_{FE} - \hat{\beta}_{RE})^T \left(\hat{Var}(\hat{\beta}_{FE}) - \hat{Var}(\hat{\beta}_{RE})\right)^{-1} (\hat{\beta}_{FE} - \hat{\beta}_{RE}) \overset{H_0}{\sim} \chi^2_p$$

- Testidee:
  - $\hat{\beta}_{FE}$ ist konsistent
  - $\hat{\beta}_{RE}$ ist unter REA konsistent (und effizient)

  $\Rightarrow$ Schätzer sollten sich unter REA nicht systematisch unterscheiden
- Testentscheidung:
  - „H_1“: REA ist verletzt
  - „H_0“: Keine gesicherte Aussage bzgl. REA möglich
- **Umsetzung in R:** Keine Implementierung für `lme`/`lmer`/`gam`

  $\Rightarrow$ Deshalb: Test per Hand rechnen

Random Effects Assumption

Lösungsansätze bei verletzter Annahme:

1) **Fixed effects model**

- Schätzung der personenspezifischen Effekte als fixed effects
- Vorteil: β-Schätzer haben keinen Bias (trotz Multikollinearität)
- Nachteil: Keine Schätzung von zeitkonstanten Kovariablen-Effekten

2) **Hybrid model**

$$Y_{ij} = \beta_0 + (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)^T \beta_{W1} + \bar{\mathbf{x}}_i^T \beta_{G1} + b_i + \epsilon_{ij}$$

- $\mathbf{x}_{ij} - \bar{\mathbf{x}}_i$ nicht mit b_i korreliert; wenn dann nur $\bar{\mathbf{x}}_i$!
- Vorteil: Auch Schätzung zeitkonstanter Effekte möglich
- Test $\beta_{W1} = \beta_{G1}$ ähnlich zu Hausman-Test

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
$\Rightarrow$ 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 θ, Skalen- bzw. Dispersionsparameter ϕ, Funktionen $\psi(\cdot)$, $c(\cdot,\cdot)$.

Eigenschaften der ersten beiden Momente:

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

mit $v(\mu)$ der Varianzfunktion.

Wiederholung: GLM

Notation:

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

Annahmen:

- $Y_i \overset{u.}{\sim}$ Expo-Fam.(θ_i, ϕ)
- 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:**

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

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 - $Y_i \overset{u.}{\sim} \mathcal{N}(\mu_i, \sigma^2)$

$$f(y_i) = exp\left\{-\frac{1}{2}\left(\frac{y_i - \mu_i}{\sigma}\right)^2 - \log(\sqrt{2\pi\sigma^2})\right\}$$

$$= \exp\left\{\frac{y_i\mu_i - \mu_i^2/2}{\sigma^2} - \frac{y_i^2}{2\sigma^2} - \log(\sqrt{2\pi\sigma^2})\right\}.$$

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 - $Y_i \overset{u.}{\sim} \mathcal{B}(\pi_i)$

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

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:

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

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

Wiederholung: GLM

GLM mit Poissonverteilung - $Y_i \overset{u.}{\sim} Po(\lambda_i)$

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

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:

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

Wiederholung: GLM

Übersicht der betrachteten Spezialfälle:

Verteilung	ϕ	Varianzfunktion	natürliche Linkfunktion
Normal	σ^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 θ_i abhängig von $\boldsymbol{\beta}$.

2) Log-Likeliood:

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

Wiederholung: GLM

ML-Schätzung:

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

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

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 $\hat{\boldsymbol{\beta}}_{ML}$

Wiederholung: GLM

Alternative Formulierung der Score-Gleichungen:
wird später bei GEEs (Generalized Estimating Equations) benötigt

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

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

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

und somit

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

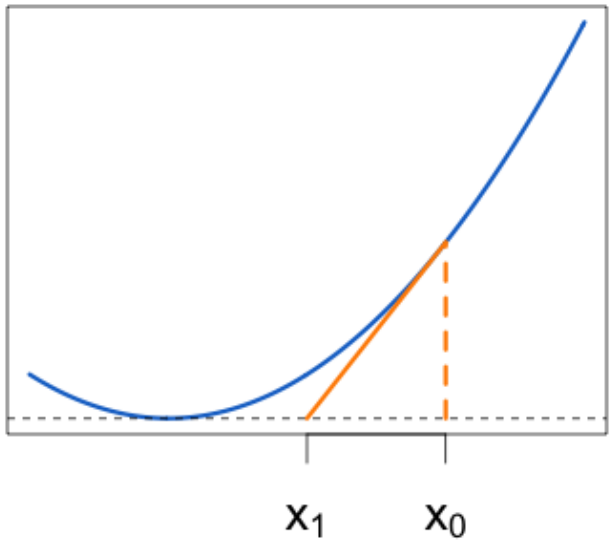
Wiederholung: GLM

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)

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



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.