3.Tutorium Generalisierte Regression

- Allgemeine GLM-Theorie -

Minh Anh Le:

28.11.2016 und 05.12.2016

Nicole Schüller:

01.12.2016 und 08.12.2016

Institut für Statistik, LMU München

Gliederung

Gliederung

Verteilungsannahme

Die bedingte Verteilung $y_i|\mathbf{x}_i$ folgt einer Exponentialfamilie, d.h. die Dichte hat die Form

$$f(y_i|\theta_i, \phi_i) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{\phi_i} + c(y_i; \phi_i)\right\}$$

wobei:

θ_i : natürlicher bzw. kanonischer Parameter

ϕ_i : Dispersionsparameter

b(), c(): Funktionen

Strukturannahme

Der (bedingte) Erwartungswert μ_i ist mit dem linearen Prädiktor $\eta_i = \mathbf{x}_i^T \boldsymbol{\beta}$ durch

$$\mu_i = h(\eta_i) \quad \text{bzw.} \quad \eta_i = g(\mu_i)$$

verknüpft, wobei

h(): Responsefunktion und

g(): Linkfunktion

Vergleich von LM und GLM

LMs	GLMs
Verteilungsannahme	
$y_i \vert \mathbf{x}_i \sim N(\mu_i, \sigma^2)$ Normalverteilung	$f(y_i \vert \theta_i, \phi_i) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{\phi_i} + c(y_i;\phi_i)\right\}$ Exponentialfamilie
Strukturannahme	
Der Erwartungswert μ_i ist mit dem linearen Prädiktor $\eta_i = \mathbf{x}_i^T\boldsymbol{\beta}$ verknüpft.	
$\mu_i = \eta_i$	$\mu_i = h(\eta_i)$ bzw. $\eta_i = g(\mu_i)$ Dabei ist $h()$ die *Responsefunktion* und $g()$ die *Linkfunktion* und es gilt $g() = h()^{-1}$. Kanonischer bzw. natürlicher Link: $\theta_i = \eta_i = \mathbf{x}_i^T\boldsymbol{\beta}$

Gliederung

Poisson-Verteilung

- Dichtefunktion von $y_i|\lambda_i \sim Po(\lambda_i)$:

$$f(y_i;\lambda_i) = \frac{\lambda_i^{y_i}}{y_i!} exp\{-\lambda_i\}, \quad \lambda_i > 0$$

- Log-Link (Kanonischer bzw. natürlicher Link)

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

Interpretation

- $\log(\lambda_i)$: sog. Log-Link für Zähldatenmodelle
  $\Rightarrow$ linearer Einfluss der Prädiktoren auf den logarithmierten Erwartungswert
  $\Rightarrow$ exponentieller Einfluss der Prädiktoren auf den Erwartungswert

Exponential-Verteilung

- Dichtefunktion von $y_i|\lambda_i \sim Exp(\lambda_i)$:

$$f(y_i;\lambda_i) = \lambda_i \exp\{-\lambda_i y_i\}, \quad \lambda_i > 0$$

- Kanonischer bzw. natürlicher Link

$$\theta = -\lambda_i \text{ und } \mu_i = \frac{1}{\lambda_i} \quad \Rightarrow \quad \theta_i = -\frac{1}{\mu_i}$$

$$\Rightarrow \quad g(\mu_i) = -\frac{1}{\mu_i} = \mathbf{x}_i^T \boldsymbol{\beta}$$

Beachte

- Kanonischer (bzw. natürlicher) Link ist in diesem Fall problematisch, da $y_i > 0$ gelten muss, daher besser: $\mu_i = exp\{\mathbf{x}_i^T \boldsymbol{\beta}\}$ (Log-Link wie bei Poisson-Verteilung)!

Gliederung

Log-Likelihood

- Unter Ausschluss der von θ_i unabhängigen Komponente $c(y_i, \phi)$, $\forall i = 1, ..., n$

$$\Rightarrow \quad \ell_i(\theta_i) = \log(f(y_i|\theta_i, \phi)) = \frac{y_i\theta_i - b(\theta_i)}{\phi}$$

- Relation $\theta_i = \theta(\mu_i)$ zw. nat. Parameter und Erwartungswert

$$\Rightarrow \quad \ell_i(\mu_i) = \frac{y_i\theta(\mu_i) - b(\theta(\mu_i))}{\phi}$$

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

$$\Rightarrow \quad \ell_i(\mu_i(\boldsymbol{\beta})) = \ell_i(h(\mathbf{x}_i^T\boldsymbol{\beta})) = \frac{y_i\theta(h(\mathbf{x}_i^T\boldsymbol{\beta})) - b(\theta(h(\mathbf{x}_i^T\boldsymbol{\beta})))}{\phi}$$

Scorefunktion

- Beobachtungsspezifische Scorefunktion (vektoriell!):

$$s_i(\beta) = \mathbf{x}_i D_i(\beta) \sigma_i^{-2}(\beta)[y_i - \mu_i(\beta)],$$

mit

$$\mu_i(\beta) = h(\mathbf{x}_i^T \beta); \quad \sigma_i^2(\beta) = v(h(\mathbf{x}_i^T \beta))\phi; \quad D_i(\beta) = \frac{\partial h(\mathbf{x}_i^T \beta)}{\partial \eta}$$

- Fall der kanonischen (bzw. natürlichen) Linkfunktion:
  $s_i(\beta) = \mathbf{x}_i \frac{[y_i - \mu_i(\beta)]}{\phi}$
- ML-Schätzer auf Basis der i-ten Beobachtung:

$$s_i(\beta) = 0$$

Problem: Oft nicht eindeutig oder nur mit sehr hohem Aufwand lösbar! $\Rightarrow$ Fisher-Scoring!

Gliederung

Modellbildung

`glm(formula=`*Modellformel*, family = *Verteilungsfamilie der abh. Variable*`(link="`*Link*`")`, `data=`*Quelldatensatz*, ...)

1. Angabe der Modellformel:
   Analog zu LMs!
2. Spezifizierung der Verteilungsfamilie:
   Angabe der Verteilung der abhängigen Variable;
   Voraussetzung: Verteilung ist Exponentialfamilie
3. Spezifikation der Linkfunktion:
   Bei Verwendung des kanonischen bzw. natürlichen Links vernachlässigbar