2.Tutorium Generalisierte Regression

- Binäre Regression -

Nicole Schüller:

14.11.2016 und 24.11.2016

Minh Anh Le:

17.11.2016 und 21.11.2016

Institut für Statistik, LMU München

Gliederung

Gliederung

Wilkinson-Rogers-Notation I

- **+** Aneinanderreihung von Effekten
- **−** Ausschluss von Effekten
- **:** reiner Interaktionseffekt
- ***** Haupteffekte und Interaktionen zwischen diesen
- $\wedge$ Einschluss aller Interaktionen bis zur angegebenen Ordnung

[**Achtung**: Verwende **I(**Variable ^ 2**)** zur Einbindung des Quadrates einer Variable in das Modell! Analoges gilt für alle anderen Transformationssymbole, die auch in der Wilkinson-Rogers-Notation existieren.]

Wilkinson-Rogers-Notation II

Gegeben: x sei metrische Größe und a, b, c Faktoren

$y \sim a + x$	Modell ohne Interaktion: Einheitl. Steigungsparameter; von a abhängige Intercepts
$y \sim a * x =$ $a + x + a : x$	Modell mit Interaktion: Unterschied der Steigungsparameter im Vergleich zur Referenzkategorie durch a:x erzeugt
$y \sim (a + b + c)\,\hat{}\,2$ $= a * b * c - a : b : c$	Modell mit Interaktionen: Alle Zweifachinteraktionen werden durch $\hat{}\,2$ ins Modell aufgenommen

Gliederung

Datensituation

- Bisher:
  Betrachtung von Regressionsmodellen mit einer **skalaren Zielgröße** y.
  Implizit galt folgende Verteilungsannahme: $y_i|\mathbf{x}_i \sim N(\mu_i, \sigma^2)$
- Jetzt:
  Betrachtung von Regressionsmodellen mit einer **binären Zielgröße** y, d.h. $y_i \in \{0, 1\}$ bzw. $y_i|\pi_i \sim B(\pi_i)$:
- Dichtefunktion:

$$f(y_i; \pi_i) = \pi_i^{y_i}(1 - \pi_i)^{1-y_i}$$

Linearer Prädiktor

- Bisher:
  Modellierung des Zusammenhangs zwischen Erwartungswert und Kovariablen durch: $\mu_i = E(y_i|\mathbf{x}_i) = \mathbf{x}_i^T \boldsymbol{\beta}$
  (der erwartete Response entspricht dem linearen Prädiktor!)
- Jetzt:
  Für den Erwartungswert der binären Zielgröße gilt, dass $E(y_i|\mathbf{x}_i) = \mu_i = \pi_i$. Der erwartete Response ist eine Wahrscheinlichkeit zwischen 0 und 1 und kann **nicht dem linearen Prädiktor** entsprechen.
  $\Rightarrow$ Transformation nötig!

Link- und Responsefunktion

Für Regressionsmodelle mit binärem Response gilt:

$$E(y_i|\mathbf{x}_i) = \mu_i = \pi_i = h(\mathbf{x}_i^T \boldsymbol{\beta})$$

und umgekehrt:

$$g(\pi_i) = g(\mu_i) = h^{-1}(\mu_i) = h^{-1}[E(y_i|\mathbf{x}_i)] = \mathbf{x}_i^T \boldsymbol{\beta} = \eta_i$$

mit h(): Responsefunktion

und g(): Linkfunktion

Zusammenhang: $g() = h^{-1}()$

Charakterisierung des binären Regressionsmodells

1. Verteilungsannahme:

$$y_i|\pi_i \sim B(\pi_i)$$

2. Strukturannahme:

$$\eta_i = \mathbf{x}_i^T \boldsymbol{\beta} \quad \text{(linearer Prädiktor)}$$

3. Link-Funktion:

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

Gliederung

Logit-Link (natürlicher Link)

$$\mu_i = \pi_i = h(\eta_i) = \frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})} \Rightarrow g(\pi_i) = \log(\frac{\pi_i}{1 - \pi_i}) = \mathbf{x}_i^T \boldsymbol{\beta}$$

Interpretation

- $\log(\frac{\pi}{1-\pi}) = logit(\pi)$: logarithmierte Chancen, sog. Log Odds

  $\Rightarrow$ Linearer Einfluss der Prädiktoren auf die Log Odds, d.h. steigt x_j um eine Einheit, so ändert sich die logarithmierte Chance von Y um β_j

  $\Rightarrow$ Exponentieller Einfluss der Prädiktoren auf die Odds, d.h. steigt x_j um eine Einheit, so ändert sich die Chance von Y um $exp(\beta_j)$
- Falls $\pi = 1 - \pi = 0.5$: Chancengleichheit, d.h. kein Erklärungswert der Prädiktoren ($\hat{\boldsymbol{\beta}} = 0$).

Probit-Link

$$\mu_i = \pi_i = h(\eta_i) = \Phi(\mathbf{x}_i^T \boldsymbol{\beta}) \Rightarrow g(\pi_i) = \Phi^{-1}(\pi_i) = \mathbf{x}_i^T \boldsymbol{\beta}$$

Interpretation

- $\Phi(\mathbf{x}_i^T \boldsymbol{\beta})$: Wert der Verteilungsfunktion der Standardnormalverteilung an der Stelle $\mathbf{x}_i^T \boldsymbol{\beta}$
- $\Phi^{-1}(\pi)$: π-Quantil der Standardnormalverteilung
- Falls $\pi = 0.5 \Rightarrow \Phi^{-1}(\pi) = 0$, d.h. kein Erklärungswert der Prädiktoren ($\hat{\boldsymbol{\beta}} = 0$).
- Wenn man die Verteilungsfunktion $\Phi(\cdot)$ durch die logistische Verteilungsfunktion ersetzt, resultiert der Logit-Link.

c log log-Link (Extremwertmodell)

$$\mu_i = \pi_i = 1 - \exp\{-\exp(\mathbf{x}_i^T \boldsymbol{\beta})\}$$

$$\Rightarrow g(\pi_i) = \log\{-\log(1 - \pi_i)\} = \mathbf{x}_i^T \boldsymbol{\beta}$$

Interpretation

- $1 - \exp\{-\exp(\mathbf{x}_i^T \boldsymbol{\beta})\}$: Wert der Verteilungsfunktion der Gompertz-Verteilung an der Stelle $\mathbf{x}_i^T \boldsymbol{\beta}$
- $\log\{-\log(1 - \pi)\}$: c log log-Link
- für speziellere Anwendungen interessant (z.B. Analyse zeitdiskreter Verweildauern)

Gliederung

Modellbildung

`glm(formula=`*Modellformel*, `data=`*Quelldatensatz*, family = binomial(`link=`*Linkfunktion*), ...)

1. Angabe der Modellformel:
   Analog zu LMs! (ggf. mit Wilkinson-Rogers-Notation)
2. Spezifizierung der Verteilungsfamilie:
   Angabe der Verteilung der abhängigen Variable, in diesem Fall: “binomial“
3. Spezifikation der Linkfunktion:
   “logit“, “probit“, “cloglog“ usw...