# 5.Tutorium Generalisierte Regression

## - Multinomiales/Kumulatives Logit-Modell -

Nicole Schüller:

14.12.2015 und 11.01.2016

Hannah Busen:

17.12.2015 und 14.01.2016

Institut für Statistik, LMU München

# Gliederung

# Gliederung

# Multinomialverteilung mit Redundanzen:

$$Y \sim M(n, \pi^T = (\pi_1, ..., \pi_k))$$

$$P(\mathbf{y}^T = (m_1, ..., m_k)) = \frac{n!}{m_1! \cdot ... \cdot m_k!} \, \pi_1^{m_1} \cdot ... \cdot \pi_k^{m_k}$$

| | |
|---|---|
| $s = 1, ..., k$: | Index für Kategorien |
| $i = 1, ..., n$: | Index für Beobachtungen |
| $m_s$: | Anzahl der Beobachtungen in Kategorie s |
| $\pi_s$: | Wahrscheinlichkeit für Kategorie s |

# Multinomialverteilung ohne Redundanzen

Sei Kategorie $k$ Referenzkategorie:

$$\sum_{s=1}^{k} \pi_s = 1 \quad \Rightarrow \quad \pi_k = 1 - \pi_1 - ... - \pi_{k-1=q}$$

$$\sum_{s=1}^{k} m_s = n \quad \Rightarrow \quad m_k = n - m_1 - ... - m_{k-1=q}$$

$$\forall s \quad : \quad 0 \leq \pi_s \leq 1 \ \wedge \ m_s \in \{1, ..., n\}$$

$$P(\mathbf{y}^T = (m_1, ..., m_k)) = \frac{n!}{m_1! \cdot ... \cdot (n - m_1 - ... - m_q)!} \cdot \pi_1^{m_1} \cdot ... \cdot (1 - \pi_1 - ... - \pi_q)^{(n - m_1 - ... - m_q)}$$

# Gliederung

# Multinomiale Zielgrößen

- bisher:
  Betrachtung des Logit-Modells für binäre Zielgrößen:

$$log\left(\frac{P(y_i = 1|\mathbf{x}_i)}{P(y_i = k = 2|\mathbf{x}_i)}\right) = \mathbf{x}_i^T \boldsymbol{\beta}$$

- jetzt:
  Betrachtung des Logit-Modells für **multinomialverteilte** Zielgrößen, d.h. $y_i \in \{1, \ldots, k\}$

# Link-und Responsefunktion

Multinomiales Logit-Modell für $r = 1, ..., \overbrace{k-1}^{q}$:

$$log\left(\frac{P(y_i = r|\mathbf{x}_i)}{P(y_i = k|\mathbf{x}_i)}\right) = \mathbf{x}_i^T\boldsymbol{\beta}_r$$

bzw.

$$P(y_i = r|x) = \frac{exp\{\mathbf{x}_i^T\boldsymbol{\beta}_r\}}{1 + \sum_{s=1}^{k-1} exp\{\mathbf{x}_i^T\boldsymbol{\beta}_s\}}$$

# Link-und Responsefunktion

Außerdem gilt:

$$\mathbf{x}_i^T \boldsymbol{\beta}_k = log\left(\frac{P(y_i = k|\mathbf{x}_i)}{P(y_i = k|\mathbf{x}_i)}\right) = log(1) = 0$$

und somit

$$P(y_i = k|\mathbf{x}_i) = \frac{exp\{\overbrace{\mathbf{x}_i^T \boldsymbol{\beta}_k}^{0}\}}{1 + \sum_{s=1}^{k-1} exp\{\mathbf{x}_i^T \boldsymbol{\beta}_s\}} = \frac{1}{1 + \sum_{s=1}^{k-1} exp\{\mathbf{x}_i^T \boldsymbol{\beta}_s\}}$$

# Interpretation

- Beachte: Kategorie k ist Referenzkategorie!
- Interpretation über die logarithimierten Chancen wie beim binären Logit-Modell, jedoch jetzt immer im Bezug auf die Referenzkategorie k.
- Betrachte: $log\left(\frac{P(y=r|\mathbf{x}_i)}{P(y=k|\mathbf{x}_i)}\right) = log\left(\frac{\pi_r}{\pi_k}\right) = \mathbf{x}_i^T \boldsymbol{\beta}_r$

  $\Rightarrow$ Steigt $x_j$ um eine Einheit, so ändert sich die logarithmierte Chance von Y=r zu Y=k um $\beta_{rj}$ (Effekt von der $j$-ten Einflussgröße aus der $r$-ten Kategorie).

  $\Rightarrow$ Steigt $x_j$ um eine Einheit, so ändert sich die Chance von Y=r zu Y=k um $exp(\beta_{rj})$

# Gliederung

# Problemstellung

- Betrachte wieder eine Zielgröße $Y \in \{1, \dots, k\}$, jedoch nun ist Y ordinalskaliert, d.h. die Kategorien lassen sich ordnen.
- Bisheriges multinomiales Modell ist anwendbar, nutzt jedoch die ordinale Struktur der Daten nicht aus
  $\Rightarrow$ kumulatives Modell!
- Allgemeine Modellformulierung:

$$P(Y \leq r | \mathbf{x}_i) = F(\gamma_{0r} + \mathbf{x}_i^T \gamma)$$

# Logistische Verteilungsfunktion

Nimmt man für F die logistische Verteilungsfunktion an, so erhält man folgendes Modell:

$$P(y \leq r|\mathbf{x}_i) = \frac{exp\{\gamma_{0r} + \mathbf{x}_i^T \boldsymbol{\gamma}\}}{1 + exp\{\gamma_{0r} + \mathbf{x}_i^T \gamma\}}$$

bzw.

$$\underbrace{log\left(\frac{P(y \leq r|\mathbf{x}_i)}{P(y > r|\mathbf{x}_i)}\right)}_{\text{kumulierte Logits}} = \gamma_{0r} + \mathbf{x}_i^T \boldsymbol{\gamma}$$

# Besonderheit des Modells

- Betrachtet man zwei Populationen $x_1$ und $x_2$ (z.B. jung und alt), so gilt:

$$\frac{P(Y \leq r|\mathbf{x}_1)/P(Y > r|\mathbf{x}_1)}{P(Y \leq r|\mathbf{x}_2)/P(Y > r|\mathbf{x}_2)} = \frac{exp(\gamma_{0r} + \mathbf{x}_1^T \gamma)}{exp(\gamma_{0r} + \mathbf{x}_2^T \gamma)} = exp((\mathbf{x}_1^T - \mathbf{x}_2^T)\gamma)$$

$\rightarrow$ das Odds-Ratio ist unabhängig von Kategorie r

$\Rightarrow$ Proportional Odds Model

- Interpretation:
  Die Chancenverhältnisse sind unabhängig von der betrachteten Schwelle r und nur proportional zum Unterschied von $x_1$ und $x_2$

# Gliederung

# Nützliche R-Funktionen

- `multinom()` aus Paket `"nnet"`

  multinomiales Logit-Modell mit beobachtungsspezifischem Prädiktor $\eta_{ir} = \mathbf{x}_i^T \boldsymbol{\beta}_r$

- `polr()` aus Library `"MASS"`

  **p**roportional **o**dds **l**ogistic **r**egression

  $\rightarrow$ Fit ordinaler Logit-Modelle (per Default)

  $\rightarrow$ Fit ordinaler Probit-Modelle bei Angabe des Arguments `method = "probit"`