Statistik IV für Nebenfachstudierende

4.6. Inferenz für Mehrkategorielle Regression

Prof. Dr. Andreas Mayr

Institut für Statistik, LMU München

Sommersemester 2017

Situation

Daten: $(\underline{y}_i, \underline{x}_i)$

$\underline{y}_i \sim M(n_i, (\pi_{i1}, ..., \pi_{ik}))$

Multinomialverteilung

$\underline{y}^\top = (y_1, ..., y_k)$ ist multinomialverteilt

$\underline{y} \sim M(n, \underline{\pi}^\top = (\pi_1, ..., \pi_k))$ und

$$\mathbb{P}(\underline{y}^\top = (n_1, ..., n_k)) = \frac{n!}{n_1! \cdot ... \cdot n_k!} \pi_1^{n_1} \cdot ... \cdot \pi_k^{n_k}$$

mit $\sum_{r=1}^k n_r = n$ und $n_r \in \{0, 1, ..., n\}$.

Formulierung

- Jetzt nicht für $\underline{y}_i = (y_{i1}, ..., y_{ik})$

  mit $y_{ir} \in \{0, 1, ..., n_i\}$
- Sondern $\underline{\bar{y}}_i = (\bar{y}_{i1}, ..., \bar{y}_{ik}) = \underline{y}_i / n_i$

  mit $\underline{\bar{y}}_i = (p_{i1}, ..., p_{ik})$
- Also:

$$\mathbb{P}(\underline{\bar{y}}_i^\top = (p_{i1}, ..., p_{ik})) = \frac{n_i!}{(n_{i1}p_{i1})! \cdot ... \cdot (n_{ik}p_{ik})!} \pi_1^{(n_{i1}p_{i1})} \cdot ... \cdot \pi_k^{(n_{ik}p_{ik})}$$

Exponentialfamilie

Eine Verteilungsfamilie heißt k-parametrige Exponentialfamilie, wenn sich die Dichten auf folgende Gestalt bringen lassen:

$$f(x, \theta) = c(\theta) b(x) \exp\left(\sum_{j=1}^{k} \gamma_j(\theta) t_j(x)\right)$$

In unserem Fall:

$$f(p, \pi) = c(\pi) b(p) \exp\left(\sum_{j=1}^{k} \gamma_j(\pi) t_j(p)\right)$$

Multinomialverteilung als Exponentialfamilie

$$f(p_{i1}, ..., p_{ik}; \pi) =$$

Multinomialverteilung als Exponentialfamilie

... natürlicher r-ter Parameter

Multinomialverteilung als Exponentialfamilie

- Durch die Exponentialfamilie ist die klassische ML-Inferenz verfügbar (z.B. Score Funktion, Fisher Matrix, etc).
- Bzw. die Inferenzmethoden aus dem GLM-Framework (IWLS).

Das Maximum-Likelihood Prinzip

- Klassischerweise geht man davon aus, dass Parameter einer Verteilung fix sind.
- Die Realisierung einer Zufallsvariable ist zufällig (bei uns die Stichprobe) und folgt einer Verteilung mit Parameter θ.
- Die Wahrscheinlichkeitsverteilung $f(x, \theta)$ sagt etwas über das Auftreten von x gegeben θ aus.
- Wir haben aber x schon, und suchen θ.

► Suche nach dem θ welches am meisten für das Auftreten von x spricht.

Das Maximum-Likelihood Prinzip

Graphisch:

ML Schätzung

- log-Likelihood:
- Score-Funktion:
- Schätzgleichung:

ML Schätzung

- Fisher-Matrix:

- Asymptotische Normalität des ML Schätzers:

Goodness-of-fit

- Pearson:

- Devianz:

Hypothesentests...

.... siehe Übung.