Statistik IV für Nebenfachstudierende

4.6. Inferenz für Mehrkategorielle Regression

Prof. Dr. Andreas Mayr

Institut für Statistik, LMU München

Sommersemester 2017

Situation

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

z.B. Für eine Messstation i gibt es n_i einzelne Beobachtungen (Urnenmodell)

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

Multinomialverteilung

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

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

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

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

Formulierung

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

  mit $y_{ir} \in \{0, 1, ..., n_i\}$ $\Leftarrow$ absolute Häufigkeiten

- Sondern $\bar{\underline{y}}_i = (\bar{y}_{i1}, ..., \bar{y}_{ik}) = \underline{y}_i / n_i$

  mit $\bar{\underline{y}}_i = (p_{i1}, ..., p_{ik})$ $\Leftarrow$ relative Häufigkeiten

- Also:

$(n_i p_{i1})$ $(n_i p_{ik})$

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

$(n_i p_{i1})!$ $(n_i p_{ik})$

weil $n_{i1} = n_i \cdot p_{i1} = y_{i1}$

Exponentialfamilie

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

k-parametrische Expofam

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

Statistik

natürliche Parametrisierung

In unserem Fall:

wahren Wert

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

rel Häufigkeiten "Stichprobe"

wahrer parameter

Stichprobe

Multinomialverteilung als Exponentialfamilie

$$f(p_{i1}, \dots, p_{ik}; \pi) = \frac{n_i!}{(n_i p_{i1})! \cdot \ldots \cdot (n_i p_{ik})!} \pi_{i1}^{n_i p_{i1}} \cdot \ldots \cdot \pi_{iq}^{n_i p_{iq}} (1 - \pi_{i1} - \ldots - \pi_{iq})^{n_i (1 - p_{i1} - \ldots - p_{iq})}$$

$$= \underbrace{\frac{n_i!}{(n_i p_{i1})! \cdot (n_i p_{iq})!}}_{c_i} \underbrace{\left(\frac{\pi_{i1}}{1 - \pi_{i1} - \ldots - \pi_{iq}}\right)^{n_i p_{i1}} \cdot \ldots \cdot \left(\frac{\pi_{iq}}{1 - \pi_{i1} - \ldots - \pi_{iq}}\right)^{n_i p_{iq}}} (1 - \pi_{i1} - \ldots - \pi_{iq})^{n_i}$$

$$\overset{\exp(\log(..))}{=} \exp\left(\log(c_i) + n_i \cdot \sum_{r=1}^{q} p_{iq} \log\left(\frac{\pi_{ir}}{1 - \pi_{i1} - \ldots - \pi_{iq}}\right) + n_i \cdot \log(1 - \pi_{i1} - \ldots - \pi_{iq})\right)$$

$$= \underbrace{c_i \cdot \exp(n_i)}_{b(p)} \cdot \underbrace{(1 - \pi_{i1} - \ldots - \pi_{iq})}_{c(\pi)} \cdot \exp\left(n_i \cdot \sum_{r=1}^{q} p_{iq} \log\left(\frac{\pi_{ir}}{1 - \pi_{i1} - \ldots - \pi_{iq}}\right)\right)$$

$\Uparrow$

q-parametrische ExpoFam

Multinomialverteilung als Exponentialfamilie

... natürlicher r-ter Parameter

$$\gamma_r(\pi) = \log\left(\frac{\pi_{ir}}{1-\pi_{i1}-\dots-\pi_{iq}}\right)$$

Spezialfall: $k=2$

$$\Rightarrow \log\left(\frac{\pi_i}{1-\pi_i}\right)$$

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).

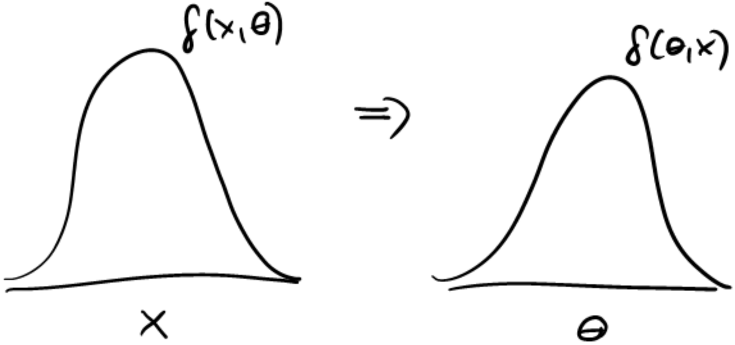
Das Maximum-Likelihood Prinzip

- 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:

$$\ell = \sum_{i=1}^{n} \log f(p_{i1}, \ldots, p_{in})$$

- Score-Funktion:

z.B. $$s(\beta) = \frac{\partial \ell(\beta)}{\partial \beta}$$ Ableitung der log-LH nach interessierendem Parameter

- Schätzgleichung:

$$s(\beta) \overset{!}{=} 0$$

$\Rightarrow$ Äquivalent zur max. der Likelihood ist max. der log-LH.
$\Rightarrow$ Ableitung Null setzen.

ML Schätzung

$\hat{=}$ 2. Ableitung der log-LH, bzw Ableitung der Scorefunktion "Krümmung"
auch Fisher-Information genannt.

- Fisher-Matrix:

$$F(\underline{\beta}) = E\left(-\frac{\partial\, \ell(\beta)}{\partial \beta\, \partial \beta^T}\right)$$

- Asymptotische Normalität des ML Schätzers:

$$\hat{\beta}_{ML} \xrightarrow{n \to \infty} N\left(\underline{\beta},\ F(\underline{\beta})^{-1}\right)$$

asympt. erwartungstreu

Σ = Fisher Matrix

Goodness-of-fit für mehrkategorielle Regression

- Pearson:

rel Häufigkeiten

Werte aus dem Modell geschätzt

$$\chi^2_P = \sum_{i=1}^{n} n_i \sum_{r=1}^{q} \frac{(p_{ir} - \hat{\pi}_{ir})^2}{\hat{\pi}_{ir}}$$

"quadratische Differenz"

- Devianz:

rel. Häufigkeit

$$\chi^2_D = 2 \sum_{i=1}^{n} n_i \sum_{r=1}^{q} p_{ir} \log\left(\frac{p_{ir}}{\hat{\pi}_{ir}}\right)$$

aus Modell

"log-Verhältnis"

Hypothesentests...

.... siehe Übung.