Kapitel 3

Likelihood-Inferenz

3.1 Parametrische Likelihood-Inferenz

Situation: $\mathcal{P}_\theta = \{f(\boldsymbol{x}|\theta) : \theta \in \Theta\}$, $\Theta \subseteq \mathbb{R}^p$, $p \ll n$, p konstant für $n \to \infty$. $f(\boldsymbol{x}|\theta)$ ist eine diskrete oder stetige oder allgemeiner eine Radon-Nikodym-Dichte.

Definition 3.1 (Likelihoodfunktion)**.** *Die* Likelihoodfunktion *von* $\theta \in \Theta$*,*

$$L(\theta) = f(\boldsymbol{x}|\theta),$$

ist definiert als die Dichte der beobachteten Daten $\boldsymbol{X} = (X_1, \ldots, X_n) = \boldsymbol{x} = (x_1, \ldots, x_n)$*, betrachtet als Funktion von* θ*. Mit* $L(\theta)$ *ist auch* $\widetilde{L}(\theta) = const \times L(\theta)$ *eine Likelihoodfunktion.*

Zu unterscheiden sind folgende Situationen:

1. $X_1, \ldots, X_n$ sind i.i.d. wie $X_i \sim f_1(x|\theta)$ (Statistik IV). Es gilt die Faktorisierung

$$L(\theta) = \prod_{i=1}^{n} f_1(x_i|\theta).$$

2. $X_1, \ldots, X_n$ — bzw. $Y_1|\boldsymbol{z}_1, \ldots, Y_n|\boldsymbol{z}_n$ im Regressionsfall bei einer Zielvariable $\boldsymbol{Y}$ und Kovariablenvektor $\boldsymbol{z}$ — sind unabhängig, aber nicht mehr identisch verteilt. Es gilt die Faktorisierung

$$L(\theta) = \prod_{i=1}^{n} f_i(x_i|\theta).$$

3. Die Paare $(X_1^d, X_1^s), \ldots, (X_i^d, X_i^s), \ldots, (X_n^d, X_n^s)$ sind unabhängig, die einzelnen Komponenten innerhalb eines Paares unter Umständen abhängig. Die Indizes s, d beziehen sich auf stetige bzw. diskrete Variablen. Eine derartige Datenlage ergibt sich beispielsweise bei Survivaldaten mit stetigen Überlebenszeiten und einem diskreten Zensierungsindikator $X_i^d = I(C_i \leq T_i)$, wobei C_i bzw. T_i den Zensierungs- bzw. Verweildauerprozess bezeichnen. Unter obige Situation fallen auch Mischverteilungsmodelle. X_i^d enspricht dann einer Klassenzugehörigkeit und X_i^s einem stetigen Merkmal(svektor).

4. Zeitlich korrelierte Daten / Stichprobenvariablen $X_1, \ldots, X_t, \ldots, X_n$ mit Dichtefunktion

$$f(x_1, \ldots, x_t, \ldots, x_n|\theta) = \\ f(x_n|x_{n-1}, \ldots, x_t, \ldots, x_1; \theta) \cdot \ldots \cdot f(x_{n-1}|x_{n-2}, \ldots, x_1; \theta) \cdot \ldots \cdot f(x_2|x_1; \theta) f(x_1|\theta).$$

Bei Markov-Ketten erster Ordnung mit der Eigenschaft

$$f(x_n|x_{n-1}, \ldots, x_1; \theta) = f(x_n|x_{n-1}; \theta)$$

vereinfacht sich die Likelihood zu

$$L(\theta) = \left(\prod_{i=2}^{n} f(x_i|x_{i-1}; \theta) \right) f(x_1|\theta).$$

Beispiel 3.1 (zu diesen vier Situationen)**.**

1. *Siehe Statistik IV bzw. Grundstudium.*
2. *Regressionssituationen (Querschnittsdaten) mit unabhängigen Zielvariablen $Y_1|\boldsymbol{z}_1, \ldots, Y_n|\boldsymbol{z}_n$ und festen Kovariablen $\boldsymbol{z}_i$:*
   - *klassisches lineares Modell: $Y_i|\boldsymbol{z}_i \sim N(\boldsymbol{z}_i^\top \boldsymbol{\beta}, \sigma^2)$,*
   - *Logit- oder Probitmodell: $Y_i|\boldsymbol{z}_i \sim Bin(1, \pi_i = h(\boldsymbol{z}_i^\top \boldsymbol{\beta}))$,*
   - *Poisson-Regression: $Y_i|\boldsymbol{z}_i \sim Po(\lambda_i = h(\boldsymbol{z}_i^\top \boldsymbol{\beta}))$.*
3. *Markov-Ketten, autoregressive Modelle für Zeitreihen/Longitudinaldaten.*
4. *Autoregressiver Prozess 1. Ordnung: Sei*

$$X_t = \alpha + \gamma X_{t-1} + \varepsilon_t$$

mit $\varepsilon_t \overset{i.i.d.}{\sim} N(0, \sigma^2)$ oder — mit zusätzlichem (zeitabhängigen) Kovariablenvektor $\boldsymbol{z}_t$ —

$$X_t = \alpha + \gamma X_{t-1} + \boldsymbol{z}_t^\top \boldsymbol{\beta} + \varepsilon_t.$$

In letzterem Fall hat die Likelihood die Form

$$L(\theta) = \left(\prod_{i=2}^{n} f_i(x_i|x_{i-1}; \theta) \right) f_1(x_1)$$

mit

$$f_i(x_i|x_{i-1}; \theta) = \phi(x_i|\alpha + \gamma x_{i-1} + \boldsymbol{z}_i^\top \boldsymbol{\beta}, \sigma^2),$$

wobei $\phi(x|\mu, \tau^2)$ den Wert der Normalverteilungsdichte mit Erwartungswert μ und Varianz τ^2 an der Stelle x bezeichnet.

Beispiel 3.2. *Wir betrachten unabhängige, aber (teils) unvollständige Ziehungen aus $N(\theta, 1)$.*

1. *Ziehung: Es sei $x_1 = 2.45$. Dann ist*

$$L_1(\theta) = \phi(x_1 - \theta) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(2.45 - \theta)^2 \right).$$

2. Ziehung: *Es sei nur* $0.9 < x_2 < 4$ *bekannt (unvollständige oder intervallzensierte Beobachtung). Die Likelihood lautet dann:*

$$L_2(\theta) = \mathbb{P}_\theta(0.9 < X_2 < 4) = \Phi(4-\theta) - \Phi(0.9-\theta).$$

Formal könnte man auch eine binäre Variable

$$X_2^d = \begin{cases} 1, \ 0.9 < X_2 < 4, \\ 0, \ \textit{sonst} \end{cases}$$

mit Dichtefunktion

$$f_2^d(1) = \mathbb{P}(X_2^d = 1) = \Phi(4-\theta) - \Phi(0.9-\theta)$$

definieren.

3. Ziehung: $z_1, \ldots, z_n$ *seien i.i.d. Realisierungen aus* $N(\theta, 1)$. *Bekannt sei aber nur*

$$x_3 = \max_{1 \le i \le n} z_i = z_{(n)}.$$

Der Rest sind fehlende Werte („missing values"). Die Verteilungsfunktion von $X_3 = Z_{(n)}$ *ist*

$$\begin{aligned} F_\theta(z_{(n)}) &= \mathbb{P}_\theta(Z_{(n)} \le z_{(n)}) = \mathbb{P}_\theta(Z_i \le z_{(n)} \ \forall \ i) \\ &= [\Phi(z_{(n)} - \theta)]^n. \end{aligned}$$

Die Dichte ergibt sich über Differentiation bezüglich θ*:*

$$f_\theta(z_{(n)}) = n[\Phi(z_{(n)} - \theta)]^{n-1} \phi(z_{(n)} - \theta),$$

d.h. für zum Beispiel $n = 5$ *und* $z_{(n)} = x_3 = 3.5$ *gilt*

$$L_3(\theta) = 5[\Phi(3.5 - \theta)]^4 \phi(3.5 - \theta).$$

Die gesamte Likelihood der drei Beobachtungen ist

$$L(\theta) = L_1(\theta) \cdot L_2(\theta) \cdot L_3(\theta),$$

also das Produkt der Likelihoodfunktionen L_1*,* L_2 *und* L_3*.*

Fazit: Die Likelihood ist sehr allgemein definiert.

Beziehung zur Bayes-Inferenz

- $p(\theta)$ sei die Prioriverteilung,
- $f(x|\theta) = L(\theta)$ die Likelihood.
- Dann ist

$$\begin{aligned} p(\theta|x) &\propto p(\theta) \cdot L(\theta) \\ \text{„Posteriori"} &\propto \text{„Priori"} \times \text{Likelihood}. \end{aligned}$$

Likelihood-Quotient

Frage: Wie vergleicht man die Likelihoods $L(\theta_1)$ und $L(\theta_2)$ für $\theta_1 \neq \theta_2$?

Antwort: Man betrachtet den Quotienten (nicht die Differenz), da dieser invariant gegenüber eineindeutigen Transformationen

$$x \mapsto y = y(x) \;\Leftrightarrow\; y \mapsto x(y)$$

ist. Für stetige x, y gilt mit dem Transformationssatz für Dichten:

$$f_Y(y|\theta) = f_X(x(y)|\theta) \left|\det\left(\frac{\partial x}{\partial y}\right)\right|$$

und somit

$$L(\theta; y) = L(\theta; x) \left|\det\left(\frac{\partial x}{\partial y}\right)\right| \;\Rightarrow\; \frac{L(\theta_2; y)}{L(\theta_1; y)} = \frac{L(\theta_2; x)}{L(\theta_1; x)}.$$

Satz 3.2.

1. *Sei $T = T(X)$ suffizient für θ. Dann gilt $L(\theta; x) = const \times L(\theta; t)$ mit $t = T(x)$, d.h. $L(\theta; x)$ und $L(\theta; t)$ sind äquivalent.*
2. *$L(\theta; x)$ ist minimalsuffzient.*

Beweis. Folgt unmittelbar aus den Resultaten aus Abschnitt 2. □

3.2 Maximum-Likelihood-Schätzung

Die Maximum-Likelihood-Schätzung ist die populärste Methode zur Konstruktion von Punktschätzern bei rein parametrischen Problemstellungen.

3.2.1 Schätzkonzept

Maximum-Likelihood-Prinzip: Finde Maximum-Likelihood-Schätzwert $\widehat{\theta}$, so dass

$$L(\widehat{\theta}; x) \geq L(\theta; x) \text{ für alle } \theta \in \Theta.$$

Dazu äquivalent ist

$$\ell(\widehat{\theta}; x) \geq \ell(\theta; x), \;\; \ell(\theta) = \log L(\theta)$$

mit der Log-Likelihood ℓ. Meist sucht man nach (lokalen) Maxima von $\ell(\theta)$ durch Nullsetzen der Score-Funktion

$$s(\theta) = \frac{\partial \ell(\theta)}{\partial \theta} = \left(\frac{\partial \ell(\theta)}{\partial \theta_1}, \ldots, \frac{\partial \ell(\theta)}{\partial \theta_p}\right)^\top$$

(soweit die 1. Ableitung der Log-Likelihood existiert!) als Lösung der sogenannten *ML-Gleichung*

$$s(\widehat{\theta}) = 0.$$