Prof. Vincent Tassion
Laurin Köhler-Schindler

Wahrscheinlichkeit & Statistik

Serie 10

Abgabe bis Mittwoch (11.05.2022) um 10:15 Uhr

Diese Serie beschäftigt sich mit der Maximum-Likelihood-Methode.

Weitere Informationen und Instruktionen zur Abgabe unter
https://metaphor.ethz.ch/x/2022/fs/401-0614-00L/

Aufgabe 10.1 [MLE I: Stetige Verteilung]
Wir betrachten eine stetige Verteilung mit Dichte

$$f(x) = \begin{cases} \frac{\theta}{x^{\theta+1}} & x \geq 1\,, \\ 0 & x < 1\,, \end{cases}$$

wobei $\theta > 0$ ein unbekannter Parameter ist. Wir wollen den Parameter θ mit Hilfe eines Datensatzes schätzen.

(a) Sei $X_1, \ldots, X_n$ eine Stichprobe von unabhängigen Zufallsvariablen, welche alle die Dichte f besitzen. Bestimme die Likelihood- und log-Likelihood-Funktion.

(b) Bestimme den zugehörigen Maximum-Likelihood-Schätzer für θ. Schreibe zuerst die allgemeine Formel für n Beobachtungen hin und berechne dann den Schätzwert für die folgende konkrete Stichprobe:

x_1	x_2	x_3	x_4	x_5
12.0	4.0	6.9	27.9	15.4

Aufgabe 10.2 [MLE II: Hochwasser im Zürichsee]
In Aufgabe 2 von Serie 9 haben wir folgendes Modell betrachtet: Die Zufallsvariable X messe die Wasserhöhe in cm über der kritischen Marke von 140 cm über Normalniveau im Zürichsee. Zur Modellierung von X können wir eine sogenannte verallgemeinerte Pareto-Verteilung mit Dichte

$$f_X(x;\theta) = \begin{cases} \frac{1}{\theta}(1+x)^{-(1+\frac{1}{\theta})} & \text{falls } x > 0, \\ 0 & \text{falls } x \leq 0 \end{cases}$$

verwenden. Dabei ist $\theta > 0$ ein unbekannter Parameter, der auf Basis von Daten $x_1, \ldots, x_n$ geschätzt werden soll; diese Daten werden wie üblich als Realisierungen von Zufallsvariablen $X_1, \ldots, X_n$ aufgefasst, die für jede Wahl des Parameters θ unter $\mathbb{P}_\theta$ i.i.d. sind mit Dichte $f_X(x;\theta)$. Zeige, dass der Schätzer

$$T^{(n)} = \sum_{i=1}^{n} \frac{\log(1+X_i)}{n}$$

der Maximum-Likelihood-Schätzer für θ ist.

Aufgabe 10.3 [Mittlerer quadratischer Schätzfehler: Normalverteilung]
Die Zufallsvariablen $X_1, \ldots, X_n$ seien unabhängig und je $\mathcal{N}(\mu, \sigma^2)$-verteilt unter $\mathbb{P}_\theta$, wobei $\theta = (\mu, \sigma^2) \in \mathbb{R} \times (0, \infty)$ ein 2-dimensionaler unbekannter Parameter ist. Als Schätzer für σ^2 betrachte man

$$T^{(n)}(c) := c \sum_{i=1}^{n} (X_i - \overline{X}_n)^2,$$

wobei $\overline{X}_n := \frac{1}{n} \sum_{i=1}^{n} X_i$ und $c > 0$ ist.

(a) Für welches $c^* > 0$ wird der mittlere quadratische Schätzfehler $\mathbb{E}_\theta\left[\left(T^{(n)}(c^*) - \sigma^2\right)^2\right]$ minimiert?
Hinweis: Verwende, dass $\frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2 \sim \chi^2_{n-1}$-verteilt ist unter $\mathbb{P}_\theta$.

(b) Entspricht der in (a) gefundene Schätzer $T^{(n)}(c^*)$ dem Maximum-Likelihood-Schätzer?

Aufgabe 10.4 [MLE III: Exponentialverteilung]
Eine Tankstelle veranschlagt für einen Ölwechsel mindestens $\theta_1 > 0$ Minuten. Die tatsächlich benötigte Zeit X variiert natürlich im Bereich $X \geq \theta_1$ und ist je Kunde/-in verschieden. Man kann jedoch annehmen, dass die zusätzliche Zeit durch eine exponentialverteilte Zufallsvariable gut modelliert wird.

(a) Sei $\theta_2 > 0$ und Z eine $\mathrm{Exp}(\theta_2)$-verteilte Zufallsvariable, Zeige, dass die Zufallsvariable $X = \theta_1 + Z$ die folgende Dichte hat:

$$f_{\theta_1,\theta_2}(x) = \begin{cases} \theta_2 e^{\theta_1\theta_2 - \theta_2 x} & \text{falls } x \geq \theta_1, \\ 0 & \text{sonst.} \end{cases}$$

Wir betrachten nun den Parameterraum $\Theta = \mathbb{R}_+ \times \mathbb{R}_+$ mit $\theta = (\theta_1, \theta_2)$ und die Modellfamilie $(\mathbb{P}_\theta)_{\theta \in \Theta}$, wobei die Zufallsvariablen $X_1, \ldots, X_n$ unter $\mathbb{P}_\theta$ unabhängig und identisch verteilt sind mit Dichte f_{θ_1,θ_2}.

(b) Bestimme den Maximum-Likelihood-Schätzer T_{ML} für $\theta = (\theta_1, \theta_2)$.

(c) Die benötigte Arbeitszeit in Minuten wurde für 10 zufällig ausgewählte Kunden/-innen notiert:

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
4.2	3.1	3.6	4.5	5.1	7.6	4.4	3.5	3.8	4.3

Der Stichprobenumfang ist hier also $n = 10$. Welcher Schätzwert ergibt sich basierend auf dem Maximum-Likelihood-Schätzer aus (b)?

Lösung 10.1

(a) Die Likelihood-Funktion ergibt sich aus dem Produkt der Dichten. Für $x_1, \dots, x_n \geq 1$ erhalten wir

$$L(x_1, \dots x_n; \theta) = \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} \frac{\theta}{x_i^{\theta+1}} = \theta^n \frac{1}{(\prod_{i=1}^{n} x_i)^{\theta+1}} \ .$$

Falls $x_i < 1$ für ein $1 \leq i \leq n$, ist die Likelihood-Funktion gleich 0, wir können uns also auf den Fall $x_1, \dots, x_n \geq 1$ beschränken.

Die log-Likelihood-Funktion erhalten wir durch Logarithmieren obiger Formel als

$$\ell(x_1, \dots x_n; \theta) = \log L(x_1, \dots x_n; \theta) = n \log \theta - (\theta + 1) \sum_{i=1}^{n} \log x_i.$$

(b) Ableiten und Nullsetzen der log-Likelihood-Funktion ergibt

$$\frac{\partial}{\partial \theta} \ell(x_1, \dots x_n; \theta) = \frac{n}{\theta} - \sum_{i=1}^{n} \log x_i = 0$$

für $\theta^* = \frac{n}{\sum_{i=1}^{n} \log x_i}$. Für $\theta < \theta^*$ ist die Ableitug strikt positiv, für $\theta > \theta^*$ strikt negativ, es handelt sich also um das Maximum. Also ist der Maximum-Likelihood-Schätzer

$$T_{ML} = \frac{n}{\sum_{i=1}^{n} \log X_i} = \frac{1}{\frac{1}{n} \sum_{i=1}^{n} \log X_i}.$$

Der realisierte Schätzwert für die gegebenen Daten ist dann

$$t_{ML} = \frac{5}{\sum_{i=1}^{5} \log x_i} = 0.4214.$$

Hinweis: Wir verwenden den natürlichen Logarithmus.

Lösung 10.2 Die log-Likelihood-Funktion ist gegeben durch

$$\begin{aligned} \log L(x_1, \dots, x_n; \theta) &= \log \left(\frac{1}{\theta^n} \prod_{i=1}^{n} (1 + x_i)^{-(1+\frac{1}{\theta})} \right) \\ &= -n \log \theta - (1 + \frac{1}{\theta}) \sum_{i=1}^{n} \log(1 + x_i). \end{aligned}$$

Ableiten nach θ und Nullsetzen ergibt

$$\frac{\partial}{\partial \theta} \log L(x_1, \dots, x_n; \theta) = -\frac{n}{\theta} + \frac{1}{\theta^2} \sum_{i=1}^{n} \log(1 + x_i) = 0$$

für

$$\theta^* = \frac{1}{n} \sum_{i=1}^{n} \log(1 + x_i).$$

Die zweite Ableitung ist

$$\frac{n}{\theta^2} - \frac{2}{\theta^3} \sum_{i=1}^{n} \log(1 + x_i),$$

also strikt kleiner als 0 an der Stelle θ^* und somit handelt es sich also um das Maximum. Der Maximum-Likelihood-Schätzer für θ ist also

$$T_{ML} = \frac{1}{n} \sum_{i=1}^{n} \log(1 + X_i) = T^{(n)}.$$

Lösung 10.3

(a) Wir fixieren zuerst ein $c > 0$ und setzen

$$T^* := \frac{1}{\sigma^2 c} T^{(n)}(c) = \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2.$$

Mit dem Hinweis folgt, dass

$$E_\theta[T^*] = n - 1 \quad \text{und} \quad \mathrm{Var}_\theta[T^*] = 2(n-1).$$

Also ist

$$\begin{aligned} f(c) := \mathrm{MSE}_\theta[T^{(n)}(c)] &= \mathrm{Var}_\theta[T^{(n)}(c) - \sigma^2] + \big(E_\theta\big[T^{(n)}(c) - \sigma^2\big]\big)^2 \\ &= \mathrm{Var}_\theta[\sigma^2 c T^*] + \big(E_\theta[\sigma^2 c T^* - \sigma^2]\big)^2 \\ &= \sigma^4 c^2 2(n-1) + \sigma^4 \big(c(n-1) - 1\big)^2 \\ &= \sigma^4 c^2 \big((n-1)^2 + 2(n-1)\big) - \sigma^4 c 2(n-1) + \sigma^4. \end{aligned}$$

Für beliebiges $c > 0$ ist also

$$f'(c) = 2c\left(2(n-1) + (n-1)^2\right)\sigma^4 - 2(n-1)\sigma^4$$

und

$$f''(c) = \big(4(n-1) + 2(n-1)^2\big)\sigma^4 > 0.$$

Insbesondere ist f strikt konvex und nimmt das globale Minimum bei $c^* = \frac{1}{n+1}$ an. Der gesuchte Schätzer ist also

$$T^{(n)}(c^*) = \frac{1}{n+1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2.$$

(b) Nein, der Maximum-Likelihood-Schätzer für σ^2 ist

$$T_{ML}^{(n)} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2,$$

wie in Beispiel 2 von Abschnitt 1.4 des Skripts hergeleitet wurde.

Lösung 10.4

(a) Wir verwenden die Charakterisierung der Dichte aus Proposition 4.16 und nehmen an, dass Z $\mathrm{Exp}(\theta_2)$-verteilt ist (unter $\mathbb{P}$). Z hat also die Dichte $f_Z(z) = \theta_2 e^{-\theta_2 z} \mathbb{1}_{z \geq 0}$. Für $\phi : \mathbb{R} \to \mathbb{R}$ stückweise stetig und beschränkt gilt

$$\begin{aligned} \mathbb{E}\left[\phi(X)\right] = \mathbb{E}\left[\phi(\theta_1 + Z)\right] &= \int_{-\infty}^{\infty} \phi(\theta_1 + z) f_Z(z) dz = \int_{0}^{\infty} \phi(\theta_1 + z)\theta_2 e^{-\theta_2 z} dz \\ &= \int_{\theta_1}^{\infty} \phi(x)\theta_2 e^{-\theta_2(x-\theta_1)} dx = \int_{-\infty}^{\infty} \phi(x)\theta_2 e^{\theta_1\theta_2 - \theta_2 x} \mathbb{1}_{x \geq \theta_1} dx \end{aligned}$$

und somit hat X die Dichte $f_X(x) = \theta_2 e^{\theta_1\theta_2 - \theta_2 x} \mathbb{1}_{x \geq \theta_1}$.

(b) Die Likelihood-Funktion ist

$$L(x_1, \dots, x_n; \theta_1, \theta_2) = \prod_{i=1}^n f(x_i; \theta_1, \theta_2) = \theta_2^n \exp\left(n\theta_1\theta_2 - \theta_2 \sum_{i=1}^n x_i\right) \prod_{i=1}^n \mathbb{1}_{x_i \in [\theta_1, \infty)}$$
$$= \theta_2^n \exp\left(n\theta_1\theta_2 - \theta_2 \sum_{i=1}^n x_i\right) \mathbb{1}_{\min_{1 \le i \le n} x_i \ge \theta_1}$$

und sie ist genau dann positiv, wenn alle x_i grösser oder gleich θ_1 sind. Unter diesen Nebenbedingungen müssen wir $n(\log \theta_2 + \theta_1\theta_2) - \theta_2 \sum_{i=1}^n x_i$ maximieren und erhalten

$$\theta_1 = \min_{1 \le i \le n} x_i \quad \text{und dann} \quad \theta_2 = \frac{n}{\sum_{i=1}^n x_i - n\theta_1}.$$

Daraus folgt sofort, dass

$$T_1 = \min_{1 \le i \le n} X_i \quad \text{und} \quad T_2 = \frac{n}{\sum_{i=1}^n X_i - nT_1} = \frac{1}{\overline{X}_n - T_1}$$

die Maximum-Likelihood-Schätzer von θ_1 und θ_2 sind.

(c) Für die gegebenen Daten erhalten wir die realisierten Schätzwerte

$$T_1(\omega) = t_1(x_1, \dots, x_{10}) = 3.1 \quad \text{und} \quad T_2(\omega) = t_2(x_1, \dots, x_{10}) \approx 0.7634.$$