Stochastik–Praktikum
Parametrische Schätztheorie

Thorsten Dickhaus

Humboldt-Universität zu Berlin

05.10.2010

Prolog

$X : \Omega^{-1} \to \Omega$ Zufallsgröße, die Experiment beschreibt.

Ein statistisches Modell (statistisches Experiment) ist ein Tripel $(\Omega, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta \in \Theta})$.

Ein Beispiel zur Illustration:

Binomial-Modell:

$Y = (Y_i)_{i=1,\dots,n}, \quad Y_i \sim \text{Bernoulli}(p)$,
Y_i stoch. unabhängige Zufallsvariablen mit Werten in $\{0, 1\}$,
$p \in (0, 1)$ Trefferw'keit, $n \in \mathbb{N}$, $X = \sum_{i=1,\dots,n} Y_i$

$\Omega = \{0, \dots, n\}$, $\mathcal{F} = 2^\Omega$,
$\vartheta = (n, p) \in \mathbb{N} \times (0, 1) = \Theta$

Schätzfunktionen und Eigenschaften

$T : \Omega \to \Theta$ heißt Schätzfunktion (Vorschrift).

Die Zufallsvariable $T(X)$ heißt Schätzer,
$T(x)$ Schätzwert (Realisierung).

Definition (Eigenschaften von $T(X)$)

- $T(X)$ unverzerrt (unbiased), falls $\mathbb{E}[T(X)] = \vartheta$.
- $T(X)$ effizient, falls $\mathbb{V}\mathrm{ar}(T(X))$ minimal (in einer Klasse von Schätzern).
- $T(X) = T_n(X)$ konsistent, falls $T(X) \xrightarrow{\mathbb{P}_\vartheta} \vartheta$.
- $T(X)$ suffizient, falls $T(X)$ alle Information der Stichprobe (über ϑ) ausnutzt.

Varianz-Bias Zerlegung

$\mathrm{MSE}_\vartheta(T(X)) = \mathbb{E}_\vartheta[(T(X) - \vartheta)^2]$ mittlerer quadratischer Fehler

Es gilt:

$$\begin{aligned}
\mathrm{MSE}_\vartheta(T(X)) &= \mathbb{E}_\vartheta[(T(X) - \mathbb{E}_\vartheta[T(X)] + \mathbb{E}_\vartheta[T(X)] - \vartheta)^2] \\
&= \mathbb{E}_\vartheta[(T(X) - \mathbb{E}_\vartheta[T(X)])^2] + (\mathbb{E}_\vartheta[T(X)] - \vartheta)^2 \\
&\quad + \underbrace{2(\mathbb{E}_\vartheta[T(X)] - \mathbb{E}_\vartheta[T(X)])(\mathbb{E}_\vartheta[T(X)] - \vartheta)}_{=0}
\end{aligned}$$

Also: $\mathrm{MSE}_\vartheta(T(X)) = \mathbb{V}\mathrm{ar}_\vartheta(T(X)) + (\mathbb{E}_\vartheta[T(X)] - \vartheta)^2$.

Übersicht

Übersicht

Momente

Definition

Das k-te Moment einer Zufallsvariable X ist definiert als

$$\mathbb{E}[X^k] = \begin{cases} \sum_{x \in \Omega} x^k f(x), & \text{falls } X \text{ diskret,} \\ \int_\Omega x^k f(x) dx, & \text{falls } X \text{ stetig.} \end{cases}$$

Dabei bezeichnet f die Verteilungsdichte bzw. die Wahrscheinlichkeitsfunktion von X.

Momentenschätzung (MOM)

Sei $\mu_k := \mathbb{E}[X^k]$.

Liege eine iid. Stichprobe $X_1, \ldots, X_n$ vom Umfang n für das interessierende statistische Experiment vor.

Dann ist der MOM-Schätzer $\hat{\mu}_k$ für μ_k gegeben durch

$$\hat{\mu}_k = \frac{\sum_{i=1}^{n} X_i^k}{n}.$$

Momentenschätzer: Beispiele

Habe X endliches zweites Moment.

Nach der Momentenmethode gilt:

1. $\hat{\mu} = \bar{X}_n$ (arithmetischer Mittelwert)
2. $\hat{\mu}_2 = \sum_{i=1}^{n} X_i^2 / n$
3. $\hat{\sigma}^2 = \hat{\mu}_2 - \hat{\mu}^2$, da $\mathbb{V}\mathrm{ar}(X) = \mathbb{E}[X^2] - \mathbb{E}^2[X]$

Unter 3. zeigt sich, dass aus Momenten ableitbare Parameter bei der Momentenmethode durch „plug-in“ geschätzt werden.

Momentenschätzer: Eigenschaften

Falls ein Momentenschätzer für einen Parameter ϑ von Interesse verfügbar ist, so gilt für $\hat{\vartheta}$:

- Einfach aus Stichprobenmomenten zu berechnen
- Asymptotisch erwartungstreu
- Für finite Stichprobenumfänge oftmals verzerrt
- Immer konsistent
- Oft weder effizient noch suffizient

Übersicht

Likelihood-Funktion

Definition

Statistisches Experiment $(\Omega, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta \in \Theta})$ sei gegeben.

Zufallsexperiment $(X_i)_{i=1,\ldots,n}$ sei durchgeführt worden.

Dann definieren wir

$$L(\vartheta) = \prod_{i=1}^{n} \mathbb{P}_\vartheta(X_i = x_i).$$

$L(\vartheta)$ approximiert die Wahrscheinlichkeit dafür, die konkret realisierte Stichprobe $(x_1, \ldots, x_n)$ unter ϑ zu beobachten.

Maximum Likelihood-Schätzer

Nach der MLE-Schätzmethode wird ϑ durch

$$\hat{\vartheta} = \arg\max_{\vartheta \in \Theta} L(\vartheta)$$

geschätzt.

In vielen Modellen ist die Maximalstelle eindeutig bestimmt (Konkavität von L!).

Maximum Likelihood-Schätzung: Beispiel

Wir betrachten das Binomialmodell aus der Einleitung mit $n = 10$ fest vorgegeben.

Es werden neun „Treffer“ beobachtet.

$$L(p) = \binom{10}{9} p^9 (1-p)$$

$$\ln L(p) = \ln \binom{10}{9} + 9 \ln(p) + \ln(1-p)$$

$$\frac{dL(p)}{dp} = \frac{9}{p} - \frac{1}{1-p}$$

Es ergibt sich $\hat{p} = 9/10$.

Maximum Likelihood-Schätzung: Eigenschaften

Vorteile:

- Universelles Schätzprinzip, vielseitig anwendbar
- Stets konsistent und suffizient
- Asymptotisch erwartungstreu und effizient
- Beste asymptotisch normalverteilte Schätzer

Nachteile:

- Likelihoodfunktion bzw. Optimierungsaufgabe unter Umständen nicht-trivial
- Zum Teil komplizierte Numerik erforderlich
- Für finites n zum Teil erheblich verzerrt!

Übersicht

Fehlerquadratsummenminimierung

Modellbildung:

Stichprobe $(x_1, \ldots, x_n)$ setzt sich additiv aus einer Funktion $f(\vartheta)$ des Parameters und einer Störung (Rauschen) zusammen.

Definition

Die Fehlerquadratsumme ist definiert durch

$$S(\vartheta) = \sum_{i=1}^{n} (x_i - f(\vartheta))^2.$$

OLS-Methode: Minimiere $S(\vartheta)$ bezüglich ϑ !

OLS-Methode: Beispiel

Betrachte einfache lineare Regression von Y auf X wie in Tag 1 thematisiert.

Gesucht wird also die „beste“ Approximation von Y durch eine lineare Funktion in X. Also: $f(\vartheta) = f(\alpha, \beta) = \alpha + \beta X$.

Seien Wertepaare $(x_1, y_1), \ldots, (x_n, y_n)$ beobachtet worden.

$$S(\alpha, \beta) = \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2.$$

OLS-Methode: Beispiel (II)

$$
\begin{aligned}
\frac{\partial S(\alpha, \beta)}{\partial \alpha} &= -2 \sum_{i=1}^{n} (y_i - \alpha - \beta x_i), \\
\frac{\partial S(\alpha, \beta)}{\partial \beta} &= -2 \sum_{i=1}^{n} (y_i - \alpha - \beta x_i) x_i, \\
\hat{\alpha} &= \bar{y}_n - \hat{\beta} \bar{x}_n, \\
\hat{\beta} &= \frac{\sum_{i=1}^{n} (x_i - \bar{x}_n)(y_i - \bar{y}_n)}{\sum_{i=1}^{n} (x_i - \bar{x}_n)^2} = \frac{s_{xy}}{s_x^2}.
\end{aligned}
$$

Übersicht

Bereichsschätzungen

Statistisches Experiment $(\Omega, \mathcal{F}, (\mathbb{P}_\vartheta)_{\vartheta \in \Theta \subset \mathbb{R}})$ sei gegeben.

Definition

Ein Konfidenzintervall $I = [\hat{\vartheta}_u, \hat{\vartheta}_o]$ besteht aus zwei messbaren Abbildungen $\hat{\vartheta}_{u/o} : \Omega \to \Theta$ mit der Eigenschaft

$$\forall x \in \Omega : \mathbb{P}_\vartheta(\hat{\vartheta}_u(x) \leq \vartheta \leq \hat{\vartheta}_o(x)) \geq 1 - \alpha$$

für ein vorgegebenes Konfidenzniveau $(1 - \alpha)$, $\alpha \in (0, 1)$.

In der Praxis werden die Schätzer $\hat{\vartheta}_u$ und $\hat{\vartheta}_o$ so ausgewählt, dass sich eine minimale Länge von I ergibt (Informativität!).

Konfidenzintervalle unter Normalverteilung

Sei $T(X)$ ein unverzerrter, normalverteilter Punktschätzer für ϑ.

Dann wird die Länge von I genau dann minimal, falls I symmetrisch um $T(x)$ ist.

$$
\begin{aligned}
\mathbb{P}_\vartheta(T(X) - c \leq \vartheta \leq T(X) + c) &= 1 - \alpha \\
\iff \mathbb{P}_\vartheta(|T(X) - \vartheta| \leq c) &= 1 - \alpha \\
\iff \mathbb{P}_\vartheta(T(X) - \vartheta \leq c) &= 1 - \alpha/2 \\
\iff \mathbb{P}_\vartheta\left(\frac{T(X) - \vartheta}{\sigma_T} \leq \frac{c}{\sigma_T}\right) &= 1 - \alpha/2 = \Phi(\frac{c}{\sigma_T}) \\
\iff c &= \sigma_T \Phi^{-1}(1 - \alpha/2).
\end{aligned}
$$

Ausblick: Höhere Dimensionen (von Θ)

Falls Θ hochdimensional ist, z.B. $\Theta \subseteq \mathbb{R}^k$, $k > 2$,
so sind klassische Schätzverfahren oftmals nicht optimal.

Moderne Schätzverfahren sind z. B.:

- Regularisierte Schätzverfahren (Strafterm für Komplexität)
- Shrinkage-Schätzer (James-Stein etc.)
- Statistische Lerntheorie (ERM, SRM, ...)