Statistics for Applications

Chapter 3: Parameter Estimation

Likelihood, Discrete case (1)

Let $\left(E, \mathcal{F}, (\mathbb{P}_\theta)_{\theta \in \Theta}\right)$ be a statistical model associated with a sample of i.i.d. r.v. $X_1, \ldots, X_n$. Assume that E is discrete (i.e., finite or countable).

Definition

The *likelihood* of the model is the map L_n (or just L) defined as:

$$\begin{array}{rccl} L_n \quad : & E^n \times \Theta & \to & \mathbb{R} \\ & (x_1, \ldots, x_n, \theta) & \mapsto & \mathbb{P}_\theta[X_1 = x_1, \ldots, X_n = x_n]. \end{array}$$

Likelihood, Discrete case (2)

Example 1 (Bernoulli trials): If $X_1, \ldots, X_n \overset{iid}{\sim} \text{Ber}(p)$ for some $p \in (0,1)$:

- $E = \{0,1\}$;
- $\Theta = (0,1)$;
- $\forall (x_1, \ldots, x_n) \in \{0,1\}^n, \quad \forall p \in (0,1),$

$$\begin{aligned} L(x_1, \ldots, x_n, p) &= \prod_{i=1}^{n} \mathbb{P}_p[X_i = x_i] \\ &= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} \\ &= p^{\sum_{i=1}^{n} x_i}(1-p)^{n-\sum_{i=1}^{n} x_i}. \end{aligned}$$

Likelihood, Discrete case (3)

Example 2 (Poisson model):

If $X_1, \ldots, X_n \overset{iid}{\sim} \text{Poiss}(\lambda)$ for some $\lambda > 0$:

- $E = \mathbb{N}$;
- $\Theta = (0, \infty)$;
- $\forall (x_1, \ldots, x_n) \in \{0, 1\}^n, \quad \forall \lambda > 0,$

$$\begin{aligned} L(x_1, \ldots, x_n, p) &= \prod_{i=1}^{n} \mathbb{P}_\lambda[X_i = x_i] \\ &= \prod_{i=1}^{n} e^{-\lambda} \frac{\lambda_i^x}{x_i!} \\ &= e^{-n\lambda} \frac{\lambda^{\sum_{i=1}^n x_i}}{x_1! \ldots x_n!}. \end{aligned}$$

Likelihood, Continuous case (1)

Let $\left(E, \mathcal{F}, (\mathbb{P}_\theta)_{\theta \in \Theta}\right)$ be a statistical model associated with a sample of i.i.d. r.v. $X_1, \dots, X_n$. Assume that all the $\mathbb{P}_\theta$ have a density f_θ w.r.t. the Lebesgue measure.

Definition

The *likelihood* of the model is the map L defined as:

$$\begin{array}{rccl} L & : & E^n \times \Theta & \to & \mathbb{R} \\ & & (x_1, \dots, x_n, \theta) & \mapsto & \prod_{i=1}^n f_\theta(x_i). \end{array}$$

Likelihood, Continuous case (2)

Example 1 (Gaussian model): If $X_1, \dots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, for some $\mu \in \mathbb{R}, \sigma^2 > 0$:

- $E = \mathbb{R}$;
- $\Theta = \mathbb{R} \times \mathbb{R}_+^*$;
- $\forall (x_1, \dots, x_n) \in \{0, 1\}^n, \quad \forall (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+^*$,

$$L(x_1, \dots, x_n, \mu, \sigma^2) = \frac{1}{(\sqrt{2\pi\sigma^2})^n} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right).$$

Maximum likelihood estimator (1)

Let $X_1, \dots, X_n$ be an i.i.d. sample associated with a statistical model $\big(E, \mathcal{F}, (\mathbb{P}_\theta)_{\theta\in\Theta}\big)$ and let L be the corresponding likelihood.

Definition

The *likelihood estimator* of θ is defined as:

$$\hat{\theta}_n^{MLE} = \underset{\theta\in\Theta}{\operatorname{argmax}}\ L(X_1, \dots, X_n, \theta),$$

provided it exists.

Remark (log-likelihood estimator): In practice, we use the fact that

$$\hat{\theta}_n^{MLE} = \underset{\theta\in\Theta}{\operatorname{argmax}}\ \ln L(X_1, \dots, X_n, \theta).$$

Maximum likelihood estimator (2)

Examples

- Bernoulli trials: $\hat{p}_n^{MLE} = \bar{X}_n$.
- Poisson model: $\hat{\lambda}_n^{MLE} = \bar{X}_n$.
- Gaussian model: $(\hat{\mu}_n, \hat{\sigma}_n^2) = \left(\bar{X}_n, \hat{S}_n\right)$.

Maximum likelihood estimator (3)

Definition: Fisher information

Define the log-likelihood for one observation as:

$$\ell(\theta) = \ln L_1(X, \theta), \qquad \theta \in \Theta.$$

Assume that ℓ is a.s. twice differentiable. Under some regularity conditions, the *Fisher information* of the statistical model is defined as:

$$I(\theta) = \mathbb{V}_\theta \left(\nabla_\theta \ell(\theta) \right) = -\mathbb{E}_\theta \left[\frac{\partial^2 \ell}{\partial \theta \partial \theta'}(\theta) \right].$$

Maximum likelihood estimator (4)

Theorem

Let $\theta^* \in \Theta$ (the *true* parameter). Assume the following:

1. The model is identified.
2. For all $\theta \in \Theta$, the support of $\mathbb{P}_\theta$ does not depend on θ;
3. θ^* is not on the boundary of Θ;
4. $I(\theta)$ is invertible in a neighborhood of θ^*;
5. A few more technical conditions.

Then, $\hat{\theta}_n^{MLE}$ satisfies:

- $\hat{\theta}_n^{MLE} \xrightarrow[n\to\infty]{\mathbb{P}} \theta^*$ w.r.t. $\mathbb{P}_{\theta^*}$;
- $\sqrt{n}\left(\hat{\theta}_n^{MLE} - \theta^*\right) \xrightarrow[n\to\infty]{(d)} \mathcal{N}\left(0, I(\theta^*)^{-1}\right)$ w.r.t. $\mathbb{P}_{\theta^*}$.

Method of moments (1)

Let $X_1, \ldots, X_n$ be an i.i.d. sample associated with a statistical model $\big(E, \mathcal{F}, (\mathbb{P}_\theta)_{\theta \in \Theta}\big)$. Assume that $\Theta \subseteq \mathbb{R}^d$, for some $d \geq 1$.

- *Population moments*: Let $m_k(\theta) = \mathbb{E}_\theta[X_1^k], \ 1 \leq k \leq d$.
- *Empirical moments*: Let $\hat{m}_k = \overline{X_n^k} = \frac{1}{n}\sum_{i=1}^{n} X_i^k, \quad 1 \leq k \leq d$.
- Let

$$\begin{array}{rccl} \psi & : & \Theta & \to & \mathbb{R}^d \\ & & \theta & \mapsto & (m_1(\theta), \ldots, m_d(\theta))\,. \end{array}$$

Method of moments (2)

Assume ψ is one to one:

$$\theta = \psi^{-1}(m_1(\theta), \ldots, m_d(\theta)).$$

Definition

Moments estimator of θ:

$$\hat{\theta}_n^{MM} = \psi^{-1}(\hat{m}_1, \ldots, \hat{m}_d),$$

provided it exists.

Method of moments (3)

Analysis of $\hat{\theta}_n^{MM}$

- Let $M(\theta) = (m_1(\theta), \ldots, m_d(\theta))$;
- Let $\hat{M} = (\hat{m}_1, \ldots, \hat{m}_d)$.
- Let $\Sigma(\theta) = \mathbb{V}_\theta(X_1, X_1^2, \ldots, X_1^d)$ be the covariance matrix of the random vector $(X_1, X_1^2, \ldots, X_1^d)$.
- Assume ψ^{-1} is continuously differentiable at $M(\theta)$.

Method of moments (4)

- LLN: $\hat{\theta}_n^{MM}$ is weakly/strongly consistent.
- CLT:

$$\sqrt{n}\left(\hat{M} - M(\theta)\right) \xrightarrow[n\to\infty]{(d)} \mathcal{N}\left(0, \Sigma(\theta)\right) \qquad \text{(w.r.t. } \mathbb{P}_\theta\text{)}.$$

Hence, by the Delta method (see next slide):

Theorem

$$\sqrt{n}\left(\hat{\theta}_n^{MM} - \theta\right) \xrightarrow[n\to\infty]{(d)} \mathcal{N}\left(0, \Gamma(\theta)\right) \qquad \text{(w.r.t. } \mathbb{P}_\theta\text{)},$$

where $\Gamma(\theta) = \nabla\left(\psi^{-1}\right)(M(\theta))\,\Sigma(\theta)\nabla\left(\psi^{-1}\right)(M(\theta))'$.

Multivariate Delta method

Let $(T_n)_{n\geq 1}$ sequence of random vectors in $\mathbb{R}^p$ ($p \geq 1$) that satisfies

$$\sqrt{n}(T_n - \vartheta) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, \Sigma),$$

for some $\vartheta \in \mathbb{R}^p$ and some symmetric positive semidefinite matrix $\Sigma \in \mathbb{R}^{p\times p}$.

Let $g : \mathbb{R}^p \to \mathbb{R}^k$ ($k \geq 1$) be continuously differentiable at ϑ. Then,

$$\sqrt{n}\,(g(T_n) - g(\vartheta)) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, \nabla g(\vartheta)' \Sigma \nabla g(\vartheta)),$$

where $\nabla g(\vartheta) = \left(\dfrac{\partial g_j}{\partial \theta_i}\right)_{1\leq i\leq d, 1\leq j\leq k} \in \mathbb{R}^{k\times d}$.

MLE vs. Moment estimator

- Comparison of the quadratic risks: In general, the MLE is more accurate.

- Computational issues: Sometimes, the MLE is intractable.

M-estimators (1)

Idea:

- Let $X_1, \dots, X_n$ be i.i.d with some unknown distribution P in some space E ($E \subseteq \mathbb{R}^d$ for some $d \geq 1$).
- No statistical model needs to be assumed.
- Goal: estimate a parameter θ^* associated with P, e.g. its mean, variance, median, other quantiles, the true parameter in some statistical model...
- Find a function $\rho : E \times \Theta \to \mathbb{R}$, where Θ is the set of all possible values for the unknown θ^*, such that:

$$\mathcal{Q}(\theta) := \mathbb{E}\left[\rho(X_1, \theta)\right]$$

  achieves its minimum only at $\theta = \theta^*$.

M-estimators (2)

- E.g., $\rho(x,\theta) = (x-\theta)^2$, $\rho(x,\theta) = |x-\theta|$, $\rho = -\ln L_1$, etc...
- Let $J(\theta) = -\mathbb{E}\left[\frac{\partial^2 \rho}{\partial\theta\partial\theta'}(X_1,\theta)\right] = -\frac{\partial^2 Q}{\partial\theta\partial\theta'}(\theta)$.
- Let $K(\theta) = \mathbb{V}\left[\frac{\partial \rho}{\partial\theta}(X_1,\theta)\right]$.
- Define $\hat{\theta}_n$ as a minimizer of:

$$\mathcal{Q}_n(\theta) := \frac{1}{n}\sum_{i=1}^{n} \rho(X_i,\theta).$$

M-estimators (3)

Theorem

Let $\theta^* \in \Theta$ (the *true* parameter). Assume the following:

1. θ^* is not on the boundary of Θ;
2. $J(\theta)$ is invertible in a neighborhood of θ^*;
3. A few more technical conditions.

Then, $\hat{\theta}_n$ satisfies:

- $\hat{\theta}_n \xrightarrow[n \to \infty]{\mathbb{P}} \theta^*$;
- $\sqrt{n}\left(\hat{\theta}_n - \theta^*\right) \xrightarrow[n \to \infty]{(d)} \mathcal{N}\left(0, J(\theta^*)^{-1} K(\theta^*) J(\theta^*)^{-1}\right)$.

M-estimators (4)

Example: Location parameter

If $X_1, \ldots, X_n$ are i.i.d. with density $f(\cdot - \theta)$, where:

- f is an unknown, positive, even function;
- θ is a real number of interest, a *location parameter*;

How to estimate θ ?

- M-estimators: empirical mean, empirical median, ...
- Compare their risks or asymptotic variances;
- The empirical median is more *robust*.