Lec 6. Point Estimation

Zhang W.P.

Contents

6.1 Estimation principle 1
6.1.1 Method of Moments, MOM 8
6.1.2 Maximum Likelihood Estimators (MLEs) 11

Learning objectives

After this section, you should be able to ...

- understand the criteria of point estimation
- understand the popular estimation methods
- find the estimator and its properties

6.1 Estimation principle

A point estimator uses the information available in a sample to obtain a single number that estimates a population parameter.

Let $X = (X_1, \ldots, X_n) \in \mathcal{X}$ be a sample from the population $F_\theta, \theta \in \Theta \subset R^d$, then any statistic $\hat{g}_n(X) = \hat{g}(X_1, \ldots, X_n)$ is said to be an estimator of parameter of interest $g(\theta)$. Def.

Suppose we want to estimate θ, how do we find a function $\hat{\theta}_n(X)$ of the sample X that in some sense "is close" to the unknown θ?

The fundamental heuristic

Minimal contrast estimation: We consider a function that we shall call a contrast function:

$$\rho : \mathcal{X} \times \Theta \to R$$

and define $\mathcal{D}(\theta_0, \theta)$ to measure the discrepancy between θ and θ_0(true value):

$$D(\theta_0, \theta) \equiv E_{\theta_0} \rho(X, \theta)$$

- We require that $D(\theta_0, \theta)$ is uniquely minimized for $\theta = \theta_0$.
- Since $\rho(X, \theta)$ is an estimate of $D(\theta_0, \theta)$, it is natural to consider $\widehat{\theta}(X)$ minimizing $\rho(X, \theta)$.
- Now suppose the true θ_0 is an interior point of Θ, and $\theta \to D(\theta_0, \theta)$ is smooth. Then we expect

$$\nabla_\theta D(\theta_0, \theta) = 0$$

where ∇ denotes the gradient $\nabla_\theta = \left(\frac{\partial}{\partial \theta_1}, \ldots, \frac{\partial}{\partial \theta_d} \right)'$

- Arguing heuristically again we are led to estimates $\widehat{\theta}$ that solve

$$\nabla_\theta \rho(X, \widehat{\theta}) = 0$$

which is a special form of estimating equations.

Estimation equation: More generally, suppose we are given a function $\Psi : \mathcal{X} \times R^d \to R^d, \Psi \equiv (\psi_1, \ldots, \psi_d)^T$ and define

$$S(\theta_0, \theta) = E_{\theta_0} \Psi(X, \theta)$$

Suppose $S(\theta_0, \theta) = 0$ has $\theta = \theta_0$ as its unique solution for all $\theta_0 \in \Theta$. Then we say $\widehat{\theta}$ solving

$$\Psi(X, \hat{\theta}) = 0$$

is an estimating equation estimate.

Small sample properties

- The criterion of unbiasedness

Def.

An estimator $\hat{g}_n(X)$ is an unbiased estimator of $g(\theta)$, if

$$E_\theta \hat{g}_n(X) = g(\theta), \quad \forall \theta \in \Theta$$

The bias of estimator $\hat{g}_n(X)$ is defined as $\hat{g}_n(X) - g(\theta)$. $\hat{g}_n(X)$ is said to be asymptotically unbiased if its bias goes to zero as the sample size $n \to \infty$.

Small sample properties

- Relative efficiency

Given two different unbiased estimators $\hat{g}_n(X)$ and $\tilde{g}_n(X)$ of $g(\theta)$, $\hat{g}_n(X)$ is said to be relatively more efficient than $\tilde{g}_n(X)$ if

$$Var_\theta(\tilde{g}_n(X)) \geq Var_\theta(\hat{g}_n(X), \ \ \forall \theta \in \Theta$$

and '>' holds for at least one value of θ. The efficiency of $\hat{g}_n(X)$ relative to $\tilde{g}_n(X)$, denoted $eff(\hat{g}_n, \tilde{g}_n)$, is given by

$$eff(\hat{g}_n, \tilde{g}_n) = \frac{Var_\theta(\tilde{g}_n(X))}{Var_\theta(\hat{g}_n(X))}$$

Def.

- **Mean square error (MSE):** for an estimator $\hat{g}_n(X)$ of $g(\theta)$:

$$MSE(\hat{g}_n(X)) = E(\hat{g}_n(X) - g(\theta))^2 = Var(\hat{g}_n(X)) + (E\hat{g}_n(X) - g(\theta))^2$$

Large sample properties

- The criterion of consistency

An estimator $\hat{g}_n(X)$ is a consistent estimator of $g(\theta)$, if $\hat{g}_n(X) \xrightarrow{p} g(\theta)$, i.e., if $\hat{g}_n(X)$ converges in probability to $g(\theta)$. Otherwise $\hat{g}_n(X)$ is said to be inconsistent. Def.

- Strongly consistent estimator: $\hat{g}_n(X) \overset{a.s.}{\to} g(\theta)$
- Consistent estimator in r-mean: $\lim_{n\to\infty} E|\hat{g}_n(X) - g(\theta)|^r = 0$.

Large sample properties

- Asymptotical normality

We say an estimator $\hat{g}_n(X)$ of $g(\theta)$ has asymptotical normality property, if

$$B_n^{-1}(\theta)[\hat{g}_n(X) - g(\theta)] \xrightarrow{\mathcal{L}} N(0, I)$$

Def.

for some positive series $B_n(\theta)$ as $n \to \infty$.

6.1.1 Method of Moments, MOM

The method of moments is a very simple (plug-in or substitution) procedure for finding an estimator for one or more parameters of a statistical model. It is one of the oldest methods for deriving point estimators.

If $g(\theta) = G(\alpha_1, \ldots, \alpha_k; \mu_2, \ldots, \mu_s)$, i.e., a function of population raw and central moments, then the method of moments estimator for $g(\theta)$ is defined as

$$\hat{g}_n(X) = G(a_{n1}, \ldots, a_{nk}; m_{n2}, \ldots, m_{ns}) \qquad \text{Def.}$$

where $a_{n1}, \ldots, a_{nk}; m_{n2}, \ldots, m_{ns}$ are sample moments of corresponding population moments.

Plug-in principle: $\hat{\theta}(X) = \theta(F_n)$ is a plug-in estimate of $\theta(F)$ as F_n is an estimate of F.

Properties

Theorem 1. *$\hat{g}_n(X) = G(a_{n1}, \ldots, a_{nk}; m_{n2}, \ldots, m_{ns})$ is a consistent estimator for $g(\theta) = G(\alpha_1, \ldots, \alpha_k; \mu_2, \ldots, \mu_s)$ if G is continuous function.*

Theorem 2. *$\hat{g}_n(X) = G(a_{n1}, \ldots, a_{nk}; m_{n2}, \ldots, m_{ns})$ is an asymptotically normal estimator for $g(\theta) = G(\alpha_1, \ldots, \alpha_k; \mu_2, \ldots, \mu_s)$ if G is continuously differentiable.*

6.1.2 Maximum Likelihood Estimators (MLEs)

Let $f(x|\theta)$ be the "probability function" of a sample X with parameter space Θ. If $X = x$ is observed, then

$$L(\theta) \equiv L(\theta \mid x) = f(x \mid \theta)$$

is called the likelihood function. For each sample point $x = (x_1, \ldots, x_n)$, let

$$\hat{\theta}(x) = \arg\min_{\theta \in \Theta} L(\theta)$$

Then $\hat{\theta}(X)$ is called the maximum likelihood estimator (MLE) of the parameter θ.

Def.

The following remarks are important

- The likelihood function is a function of θ (and that $x_1, \ldots, x_n$ act as fixed constants).
- If $X_1, \ldots, X_n$ is an independent sample from a population with pdf or pmf $f(x \mid \theta)$, then the likelihood function

$$L(\theta) \equiv L(\theta \mid x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i \mid \theta).$$

- If the MLE $\hat{\theta}$ exists, then $\hat{\theta} \in \Theta$. Hence if $\hat{\theta}$ is not in the parameter space Θ, then $\hat{\theta}$ is not the MLE of θ.
- If T is a sufficient statistic for θ and a unique MLE $\hat{\theta}$ of θ exists, $\hat{\theta}$ must be a function of T, and thus is minimal sufficient.
- Let $D(\theta_0, \theta) = -E_{\theta_0} \log f(X|\theta)$ and to say that $D(\theta_0, \theta)$ is minimized uniquely when $P_\theta = P_{\theta_0}$ is equivalent to

$$\begin{aligned} D(\theta_0, \theta) - D(\theta_0, \theta_0) &= -\left(E_{\theta_0} \log f(X|\theta) - E_{\theta_0} \log f(X|\theta_0)\right) \\ &= -E_{\theta_0} \log \frac{f(X|\theta)}{f(X|\theta_0)} > 0 \end{aligned}$$

unless $\theta = \theta_0$. Here $D(\theta_0, \theta_0)$ is called the entropy of X. In the case of $X_1, \ldots, X_n$ i.i.d.

$$\rho(X, \theta) = -\frac{1}{n} \sum \log f(X_i, \theta)$$

satisfies the condition of being a contrast function, and the MLE is a minimum contrast estimate.

- Define the mutual entropy or Kullback-Leibler information divergence between $f(X|\theta_0)$ and $f(X|\theta)$ by

$$K(f(X|\theta_0), f(X|\theta)) = -E_{\theta_0} \log \frac{f(X|\theta)}{f(X|\theta_0)}$$

Then maximizing likelihood is equivalent to minimizing Kullback-Leibler divergence, i.e., maximizing the likelihood of data is equal to minimizing the difference between the estimate and the real data distribution.

Theorem 3 (**Invariance Principle**). *If $\hat{\theta}$ is the MLE of θ, then $h(\hat{\theta})$ is the MLE of $h(\theta)$ where h is a function with domain Θ.*

Proof. When h is one to one, let $\eta = h(\theta)$. Then the inverse function h^{-1} exists and $\theta = h^{-1}(\eta)$. Hence

$$f(x \mid \theta) = f\left(x \mid h^{-1}(\eta)\right)$$

is the joint pdf or pmf of x. So the likelihood function of $h(\theta) = \eta$ is

$$L^*(\eta) \equiv L\left(h^{-1}(\eta)\right).$$

Also note that

$$\sup_{\eta} L^*(\eta) = \sup_{\eta} L\left(h^{-1}(\eta)\right) = L(\hat{\theta})$$

Thus

$$\hat{\eta} = h(\hat{\theta})$$

is the MLE of $\eta = h(\theta)$ when h is one to one.

Berk (1967) said that if h is not one to one, define

$$w(\theta) = (h(\theta), u(\theta)) = (\eta, \gamma) = \xi$$

such that $w(\theta)$ is one to one. Note that the choice

$$w(\theta) = (h(\theta), \theta)$$

works. In other words, we can always take u to be the identity function. The choice of w is not unique, but the inverse function

$$w^{-1}(\xi) = \theta$$

is unique. Hence the likelihood is well defined, and $w(\hat{\theta})$ is the MLE of ξ. $\square$

Properties

Theorem 4. *Let X be a sample from a population with corresponding densities $f(x|\theta), \theta \in \Theta$ open $\subset R^d$. If further $l_x(\theta) \equiv \log f(x|\theta)$ is strictly concave and $l_x(\theta) \to -\infty$ as $\theta \to \partial\Theta$ (the boundary of Θ), then the MLE $\hat{\theta}(x)$ exists and is unique.*

Theorem 5. *Suppose that* $Q\left(\theta; \boldsymbol{X}_n\right)$ *is continuous in* θ *and there exists a function* $Q_0(\theta)$ *such that*

1. $Q_0(\theta)$ *is uniquely maximized at* θ_0
2. Θ *is compact (i.e., closed and bounded)*
3. $Q_0(\theta)$ *is continuous in* θ
4. $Q\left(\theta; \boldsymbol{X}_n\right)$ *converges uniformly in probability to* $Q_0(\theta)$*, i.e., we have that for all* $\epsilon > 0$*,*

$$P_{\theta_0}\left[\sup_{\theta \in \Theta} \left|Q\left(\theta; \boldsymbol{X}_n\right) - Q_0(\theta)\right| > \epsilon\right] \to 0, n \to \infty$$

then $\hat{\theta}\left(\boldsymbol{X}_n\right)$ *defined as the value of* $\theta \in \Theta$ *which for each* $\boldsymbol{X}_n = \boldsymbol{x}_n$ *maximizes the objective function* $Q\left(\theta; \boldsymbol{X}_n\right)$ *satisfies* $\hat{\theta}\left(\boldsymbol{X}_n\right) \xrightarrow{P} \theta_0$

Theorem 6. *Let $X_1, \ldots, X_n$ be a simple sample from a population with probability function $f(x \mid \theta), \theta \in \Theta \subset R\}$, and $\hat{\theta}$ be the MLE of true parameter θ_0. Suppose certain regularity conditions hold, including:*

- *All probability function $f(x \mid \theta)$ in the model have the same support,*
- *θ_0 is an interior point (i.e., not on the boundary) of Θ,*
- *The log-likelihood $\ell(\theta) = logL(\theta)$ is differentiable in θ, and*
- *$\hat{\theta}$ is the unique value of $\theta \in \Theta$ that solves the equation $0 = \ell'(\theta)$.*

Then $\hat{\theta}$ is consistent and asymptotically normal, with

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \to \mathcal{N}\left(0, \frac{1}{I\left(\theta_0\right)}\right)$$

in distribution. Here, $I(\theta)$ is defined by the two equivalent expressions

$$I(\theta) := \mathrm{Var}_\theta[S(X, \theta)] = -\mathbb{E}_\theta\left[S'(X, \theta)\right]$$

where Var_θ and $\mathbb{E}_\theta$ denote variance and expectation with respect to $X \sim f(x \mid \theta)$, and

$$S(x, \theta) = \frac{\partial}{\partial \theta} \log f(x \mid \theta), \quad S'(x, \theta) = \frac{\partial^2}{\partial \theta^2} \log f(x \mid \theta)$$

$S(x, \theta)$ is called the score function, and $I(\theta)$ is called the Fisher information.

Different likelihoods

Let the parameter be (β, θ) and θ is not of interest, then the standard likelihood by $L(\beta, \theta \mid x)$.

- **Maximum Likelihood:** Find β and θ that maximizes $L(\beta, \theta \mid x)$.
- **Profile Likelihood:** If we can express θ as a function of β then we replace θ with the corresponding function. Say, $\theta = g(\beta)$. Then, we maximize:

  $$L(\beta, g(\beta) \mid x)$$

  Usually, the profile likelihood estimate $\hat{\beta}$ can be done in the following two steps:

  - fixed β, find

    $$g(\beta) = \arg\max_{\theta} L(\beta, \theta | x)$$

  - find $\hat{\beta}$ such that

    $$\hat{\beta} = \arg\max L(\beta, g(\beta) \mid x)$$

- **Induced likelihood** If $\eta = \tau(\theta)$, then

$$L^*(\eta) \equiv \sup_{\theta:\tau(\theta)=\eta} L(\theta)$$