Lecture 7: Methods of Estimation

MATH 667-01
Statistical Inference
University of Louisville

September 19, 2017

Introduction

- We consider methods for finding estimators of the unknown parameter(s) in a model which are discussed in Sections 7.1 and 7.2 of Casella and Berger (2001)[1].
- Specifically, in class we will cover three widely used methods:
    1. method of moments
    2. maximum likelihood estimation
    3. Bayes estimation
- When discussing the likelihood principle, one motivation is mentioned from Section 6.3.
- When discussing Bayes estimation, we will review Bayes' Rule from Section 1.3 and conditional probabilities from Section 4.2.

[1] Casella, G. and Berger, R. (2001). Statistical Inference, second edition. Duxbury Press.

Introduction

- We consider point estimation of the unknown parameter θ (or function of the unknown parameter) in a parametric model $\boldsymbol{X} \sim f_{\boldsymbol{X}}(\boldsymbol{x}|\theta)$.
- Usually, we assume $X_1, \ldots, X_n$ is a random sample from a population with pdf/pmf $f(x|\theta)$. Estimates $\hat{\theta}$ of θ based on observed data $x_1, \ldots, x_n$ gives us an estimated model from the parametric family.
- *Definition L7.1* (Def 7.1.1 on p.311): A *point estimator* is any function $W(X_1, \ldots, X_n)$ of a sample. That is, any statistic is a point estimator.
- Note that an estimator is a function of the sample $X_1, \ldots, X_n$ so it is random.
- Alternately, we refer to the observed value of a point estimator based on a realized data values $x_1, \ldots, x_n$ as a *point estimate*. So, the point estimate $W(x_1, \ldots, x_n)$ is not random.

Method of Moments

- This is a simple approach based on matching the sample and poulation moments.
- Let $X_1, \ldots, X_n$ be a sample from a population with pdf/pmf $f(x|\theta_1, \ldots, \theta_k)$.
- The method of moments estimator of the parameters is denoted by $(\tilde{\theta}_1, \ldots, \tilde{\theta}_k)$ and is obtained by solving the equations

$$\begin{aligned} m_1 &= \mu_1'(\theta_1, \ldots, \theta_k) \\ &\vdots \\ m_k &= \mu_k'(\theta_1, \ldots, \theta_k) \end{aligned}$$

for $(\theta_1, \ldots, \theta_k)$ where $m_j = \frac{1}{n}\sum_{i=1}^n X_i^j$ and $\mu_j' = \mathsf{E}[X^j]$ for $j = 1, \ldots, k$.

- *Example L7.1*: Let $X_1, \ldots, X_n$ be a random sample from a Bernoulli(p) distribution which has probability mass function

$$P(X = x) = p^x (1-p)^{1-x} I_{\{0,1\}}(x)$$

where $p \in [0, 1]$. Find the method of moments estimator of p.

- *Answer to Example L7.1*: Setting $m_1 = \mu_1'$ where $m_1 = \bar{X}$ and $\mu_1' = \mathsf{E}[X_1] = p$, the method of moments estimator is $\tilde{p} = \bar{X}$.

- *Example L7.2*: Suppose 10 voters are randomly selected in an exit poll and 4 voters say that they voted for the incumbent. What is the method of moments estimate of p?

- *Answer to Example L7.2*: The method of moments estimate of the proportion of all voters who voted for the incumbent is $\tilde{p} = \dfrac{\sum_{i=1}^n x_i}{n} = \dfrac{4}{10} = .4.$

Method of Moments

- *Example L7.3*: Suppose $X_1, \ldots, X_n$ are iid Normal(μ, σ^2) random variables. Find the method of moments estimator of (μ, σ^2).
- *Answer to Example L7.3*: Here $\mu_1' = \mathsf{E}[X] = \mu$ and $\mu_2' = \mathsf{E}[X^2] = \mathsf{Var}[X] + (\mathsf{E}[X])^2 = \sigma^2 + \mu^2$.
  So, we have $\tilde{\mu} = m_1 = \bar{X}$ and $\widetilde{\sigma^2} + \tilde{\mu}^2 = m_2$. Solving for $\widetilde{\sigma^2}$, we obtain

$$\begin{aligned} \widetilde{\sigma^2} &= m_2 - \tilde{\mu}^2 \\ &= \frac{\sum_{i=1}^n X_i^2}{n} - \bar{X}^2 \\ &= \frac{\sum_{i=1}^n X_i^2 - n\bar{X}^2}{n} \\ &= \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}. \end{aligned}$$

  Thus, $(\tilde{\mu}, \widetilde{\sigma^2}) = \left(\bar{X}, \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2\right)$.

- *Example L7.4*: Suppose $X_1, \ldots, X_n$ are iid Uniform$(-\theta, \theta)$ random variables which have probability density function

$$f(x|\theta) = \frac{1}{2\theta} I_{(-\theta,\theta)}(x)$$

where $\theta > 0$. Find a method of moments estimator of θ based on the second moment.

- Note that since $\mathsf{E}[X] = 0$, it cannot be used to estimate θ.

- *Answer to Example L7.4*: Since

$$\mu_2' = \int_{-\theta}^{\theta} x^2 \frac{1}{2\theta}\, dx = \frac{1}{2\theta}\left[\frac{1}{3}x^3\right]_{-\theta}^{\theta} = \frac{1}{2\theta}\left(\frac{\theta^3}{3} - \left(-\frac{\theta^3}{3}\right)\right) = \theta^2/3,$$

  we solve the equation

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 = \theta^2/3$$

  for θ and obtain

$$\tilde{\theta} = \pm\sqrt{\frac{3}{n}\sum_{i=1}^{n} X_i^2}.$$

- In Lecture 1, the likelihood function $L(\boldsymbol{\theta}; \boldsymbol{x}) = f_{\boldsymbol{\theta}}(\boldsymbol{x})$ was introduced.
- (Section 6.3, p.290): Intuitively, the rationale for this principle is as follows. If $L(\boldsymbol{\theta}_1|\boldsymbol{x}) > L(\boldsymbol{\theta}_2|\boldsymbol{x})$, then the sample we actually observed is more likely to have occurred if $\boldsymbol{\theta} = \boldsymbol{\theta}_1$ than if $\boldsymbol{\theta} = \boldsymbol{\theta}_2$.
    - If $\boldsymbol{X}$ is discrete, then $L(\boldsymbol{\theta}_1|\boldsymbol{x}) > L(\boldsymbol{\theta}_2|\boldsymbol{x})$ directly implies that $P_{\boldsymbol{\theta}_1}(\boldsymbol{X} = \boldsymbol{x}) > P_{\boldsymbol{\theta}_2}(\boldsymbol{X} = \boldsymbol{x})$.
    - If $X_1, \ldots, X_n$ is continuous and independent and ε is a small positive number, then $L(\boldsymbol{\theta}_1|\boldsymbol{x}) > L(\boldsymbol{\theta}_2|\boldsymbol{x})$ implies that

$$1 < \frac{L(\boldsymbol{\theta}_1|\boldsymbol{x})}{L(\boldsymbol{\theta}_2|\boldsymbol{x})} \approx \frac{\prod_{i=1}^{n} P_{\boldsymbol{\theta}_1}(x_i - \frac{\varepsilon}{2} < X < x_i + \frac{\varepsilon}{2})}{\prod_{i=1}^{n} P_{\boldsymbol{\theta}_2}(x_i - \frac{\varepsilon}{2} < X < x_i + \frac{\varepsilon}{2})}.$$

Maximum Likelihood Estimation

- *Definition L7.2* (Def 7.2.4 on p.316): For each sample point $\boldsymbol{x}$, let $\hat{\boldsymbol{\theta}}(\boldsymbol{x})$ be the parameter value at which $L(\boldsymbol{\theta}; \boldsymbol{x})$ attains its maximum as a function of $\boldsymbol{\theta}$, with $\boldsymbol{x}$ held fixed. The *maximum likelihood estimator* (MLE) of the parameter $\boldsymbol{\theta}$ based on a sample $\boldsymbol{X}$ is $\hat{\boldsymbol{\theta}}(\boldsymbol{X})$.
- By construction, we maximize $\boldsymbol{\theta}$ over its parameter space.
- There is no guarantee that the MLE will be unique in general.
- The MLE has some nice large sample properties (see Chapter 10).
- The likelihood might be difficult to maximize directly, so numerical methods such as the EM algorithm are often needed.
- The MLE can be sensitive to small changes in $\boldsymbol{x}$.

Maximum Likelihood Estimation

- Often the parameter space is an interval instead of a discrete set of values.
- If in addition the likelihood function is differentiable with respect to the parameters, then possible candidates for the MLE are (1) the solutions to the *score equations*

$$\frac{\partial}{\partial \theta_i} L(\theta_1, \ldots, \theta_k | \boldsymbol{x}) = 0, i = 1, \ldots, k$$

  and (2) the boundaries of the parameter space.

- *Example L7.5*: Let $X_1, \ldots, X_n$ be a random sample from a Bernoulli(p) distribution which has probability mass function

$$P(X = x) = p^x (1-p)^{1-x} I_{\{0,1\}}(x)$$

where $p \in (0,1)$. Find the maximum likelihood estimator of p. and show that it is a maximizer.

- *Answer to Example L7.5*: The log-likelihood function for p is

$$\begin{aligned}
\ell(p|x_1, \ldots, x_n) &= \ln L(p|x_1, \ldots, x_n) \\
&= \ln \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} \\
&= \sum_{i=1}^{n} \ln \left\{ p^{x_i}(1-p)^{1-x_i} \right\} \\
&= \sum_{i=1}^{n} \{ x_i \ln p + (1-x_i)\ln(1-p) \}
\end{aligned}$$

Maximum Likelihood Estimation

- *Answer to Example L7.5 continued*:

$$\begin{aligned}
\ell(p|x_1,\ldots,x_n) &= \sum_{i=1}^{n}\{x_i \ln p + (1-x_i)\ln(1-p)\} \\
&= \left(\sum_{i=1}^{n} x_i\right)\ln p + \left(n - \sum_{i=1}^{n} x_i\right)\ln(1-p) \\
&= n\left\{\bar{x}\ln p + (1-\bar{x})\ln(1-p)\right\}.
\end{aligned}$$

- Differentiating ℓ, we obtain

$$\frac{d\ell}{dp} = n\left(\frac{\bar{x}}{p} - \frac{1-\bar{x}}{1-p}\right) = n\left(\frac{\bar{x}(1-p)-(1-\bar{x})p}{p(1-p)}\right) = \frac{n(\bar{x}-p)}{p(1-p)}.$$

- $\hat{p} = \bar{x}$ maximizes $\ell(p|x_1,\ldots,x_n)$ since $\frac{d\ell}{dp} = 0$ if and only if $p = \bar{x}$ and $\frac{d^2\ell}{dp^2} = -n\left\{\bar{x}/p^2 + (1-\bar{x})/(1-p)^2\right\} < 0$.

- Suppose we want to estimate $\tau(\boldsymbol{\theta})$ where $\tau : \Theta \to \mathbb{R}^k$ is a function of the parameter and Θ is the parameter space (domain of $L(\boldsymbol{\theta}; \boldsymbol{x})$).
- If $L(\boldsymbol{\theta}|\boldsymbol{x})$ is the likelihood function for $\boldsymbol{\theta}$ based on $\boldsymbol{x}$, then define the induced likelihood function for $\tau(\boldsymbol{\theta})$ as

$$L^*(\boldsymbol{\eta}; \boldsymbol{x}) = \sup_{\{\boldsymbol{\theta}:\tau(\boldsymbol{\theta})=\boldsymbol{\eta}\}} L(\boldsymbol{\theta}; \boldsymbol{x})$$

  and the value $\hat{\boldsymbol{\eta}}$ which minimizes L^* is the MLE of $\boldsymbol{\eta} = \tau(\boldsymbol{\theta})$.
- The following theorem states the *invariance property of maximum likelihood estimators*.
- *Theorem L7.1* (Thm 7.2.10 on p.330): If $\hat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}$, then for any function $\tau(\boldsymbol{\theta})$, the MLE of $\tau(\boldsymbol{\theta})$ is $\tau(\hat{\boldsymbol{\theta}})$.

- *Proof of Theorem L7.1*:

$$
\begin{aligned}
L^*(\hat{\boldsymbol{\eta}};\boldsymbol{x}) &= \sup_{\boldsymbol{\eta}} L^*(\boldsymbol{\eta};\boldsymbol{x}) \\
&= \sup_{\boldsymbol{\eta}} \sup_{\{\boldsymbol{\theta}:\tau(\boldsymbol{\theta})=\boldsymbol{\eta}\}} L(\boldsymbol{\theta};\boldsymbol{x}) \\
&= \sup_{\boldsymbol{\theta}} L(\boldsymbol{\theta};\boldsymbol{x}) \\
&= L(\hat{\boldsymbol{\theta}};\boldsymbol{x}) \\
&= \sup_{\left\{\boldsymbol{\theta}:\tau(\boldsymbol{\theta})=\tau(\hat{\boldsymbol{\theta}})\right\}} L(\boldsymbol{\theta};\boldsymbol{x}) \\
&= L^*(\tau(\hat{\boldsymbol{\theta}});\boldsymbol{x})
\end{aligned}
$$

- *Example L7.6*: Suppose $X_1, \ldots, X_n$ are iid Uniform$(-\theta, \theta)$ random variables which have probability density function

$$f(x|\theta) = \frac{1}{2\theta} I_{(-\theta,\theta)}(x)$$

where $\theta > 0$.
(a) Find the maximum likelihood estimator of θ.
(b) Find the maximum likelihood estimator of $e^{-\theta}$.
(c) Find the maximum likelihood estimator of $\sqrt{\theta - 1}$.

Maximum Likelihood Estimation

- *Answer to Example L7.6*: (a) The likelihood function is

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} \frac{1}{2\theta} I_{(-\theta,\theta)}(x_i) \\ &= \frac{1}{2^n\theta^n} \prod_{i=1}^{n} I_{(-\theta,\theta)}(x_i) \\ &= \frac{1}{2^n\theta^n} I_{(0,\theta)}\left(\max_{i=1,\ldots,n} |x_i|\right) \\ &= \begin{cases} 0 & \text{if } \theta < \max\limits_{i=1,\ldots,n} |x_i| \\ \frac{1}{2^n\theta^n} & \text{if } \theta \geq \max\limits_{i=1,\ldots,n} |x_i| \end{cases}. \end{aligned}$$

- Since $L(\theta)$ is decreasing when $\theta \geq \max\limits_{i=1,\ldots,n} |x_i|$, the maximum likelihood estimator is $\hat{\theta} = \max\limits_{i=1,\ldots,n} |X_i|$.

- *Answer to Example L7.6 continued*: (b) The invariance property of the MLE (*Theorem L7.6*) implies that the MLE of $e^{-\theta}$ is

$$e^{-\hat{\theta}} = \exp\left(-\max_{i=1,\ldots,n} |X_i|\right).$$

- (c) Note that the domain of $\tau(\theta) = \sqrt{\theta - 1}$ is $[1, \infty)$. The maximizer of $L(\theta)$ if θ is restricted to $\Theta = [1, \infty)$ is

$$\hat{\theta} = \max\left\{1, \max_{i=1,\ldots,n} |X_i|\right\}.$$

  Then the invariance property of the MLE implies that the MLE of $\sqrt{\theta - 1}$ is

$$\sqrt{\hat{\theta} - 1} = \sqrt{\max\left\{1, \max_{i=1,\ldots,n} |X_i|\right\} - 1}.$$

- *Definition L7.3* (Def 1.3.2 on p.20): If $A, B \in S$ and $P(B) > 0$, then

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

- Bayes' Rule
  Theorem L7.2 (Thm 1.3.5 on p.23): Let $A_1, A_2, \ldots$ be a partition of the sample space S and $B \subset S$. If $P(B) > 0$ and $P(A_i) > 0$, then

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{j:P(A_j)>0} P(B|A_j)P(A_j)}.$$

- *Definition L7.4* (Def 4.2.1 on p.148): Let (X, Y) be a discrete bivariate random vector with joint pmf $f(x, y)$ and marginal pmfs $f_X(x)$ and $f_Y(y)$. For any x such that $P(X = x) = f_X(x) > 0$, the *conditional pmf of Y given that $X = x$* is the function of y defined by

$$f(y|x) = P(Y = y|X = x) = \frac{f(x, y)}{f_X(x)}.$$

  For any y such that $P(Y = y) = f_Y(y) > 0$, the *conditional pmf of X given that $Y = y$* is the function of x defined by

$$f(x|y) = P(X = x|Y = y) = \frac{f(x, y)}{f_Y(y)}.$$

- If $g(Y)$ is a function of a discrete random variable Y, then the *conditional expected value of $g(Y)$ given that $X = x$* is

$$\mathsf{E}(g(Y)|x) = \sum_y g(y) f(y|x).$$

- *Definition L7.5* (Def 4.2.3 on p.150): Let (X, Y) be a continuous bivariate random vector with joint pdf $f(x, y)$ and marginal pdfs $f_X(x)$ and $f_Y(y)$. For any x such that $f_X(x) > 0$, the *conditional pdf of* Y *given that* $X = x$ is the function of y defined by

$$f(y|x) = \frac{f(x, y)}{f_X(x)}.$$

For any y such that $f_Y(y) > 0$, the *conditional pdf of* X *given that* $Y = y$ is the function of x defined by

$$f(x|y) = \frac{f(x, y)}{f_Y(y)}.$$

- If $g(Y)$ is a function of a continuous random variable Y, then the *conditional expected value of* $g(Y)$ *given that* $X = x$ is

$$\mathsf{E}(g(Y)|x) = \int_{-\infty}^{\infty} g(y) f(y|x)\, dy.$$

Bayesian Estimation

- The Bayesian approach differs greatly from the classical approach that we have been discussing.
- In the Bayesian approach, the parameter $\boldsymbol{\theta}$ is assumed to be a random variable/vector with *prior distribution* $\pi(\boldsymbol{\theta})$.
- Then we can find update the pdf/pmf of the distribution of $\boldsymbol{\theta}$ given data $\boldsymbol{X} = \boldsymbol{x}$ using Bayes' Rule

$$\pi(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{f(\boldsymbol{x}, \boldsymbol{\theta})}{m(\boldsymbol{x})} = \frac{f(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{m(\boldsymbol{x})}$$

  where $m(\boldsymbol{x})$ is the pdf/pmf of the marginal distribution of $\boldsymbol{X}$. The updated prior is referred to as the *posterior distribution*.
- The Bayes estimator of $\boldsymbol{\theta}$ is obtained by finding the mean of the posterior distribution; that is, $\hat{\boldsymbol{\theta}}_B = \mathsf{E}[\boldsymbol{\theta}|\boldsymbol{X}]$.

- *Example L7.7*: Let $X_1, \ldots, X_n$ be a random sample from a Bernoulli(p) distribution. Find the Bayes estimator of p, assuming that the prior distribution on p is beta(α,β).
- *Answer to Example L7.7*: Since $X_1, \ldots, X_n$ are iid Bernoulli(p) random variables, $\sum_{i=1}^n X_i$ is binomial(n,p). The posterior distribution of $p|\sum_{i=1}^n X_i = x$ is

$$
\begin{aligned}
\pi(p|x) &= \frac{f(x|p)\pi(p)}{m(x)} \\
&= \frac{\binom{n}{x} p^x (1-p)^{n-x} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}}{\int_0^1 \binom{n}{x} p^x (1-p)^{n-x} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}\ dp} \\
&= \frac{p^{x+\alpha-1}(1-p)^{n-x+\beta-1}}{\int_0^1 p^{x+\alpha-1}(1-p)^{n-x+\beta-1}\ dp} \\
&= \frac{\Gamma(n+\alpha+\beta)}{\Gamma(x+\alpha)\Gamma(n-x+\beta)} p^{x+\alpha-1}(1-p)^{n-x+\beta-1} I_{(0,1)}(p).
\end{aligned}
$$

Bayesian Estimation

- *Answer to Example L7.7 continued*: Thus, $p|\sum_{i=1}^{n} X_i = x$ follows a beta($\sum_{i=1}^{n} x_i + \alpha$,$n - \sum_{i=1}^{n} x_i + \beta$) distribution. The Bayes estimator (posterior mean) is

$$
\begin{aligned}
\hat{p}_B &= \frac{\sum_{i=1}^{n} X_i + \alpha}{\alpha + \beta + n} \\
&= \left(\frac{n}{\alpha + \beta + n}\right) \frac{\sum_{i=1}^{n} X_i}{n} + \left(\frac{\alpha + \beta}{\alpha + \beta + n}\right) \frac{\alpha}{\alpha + \beta}.
\end{aligned}
$$

  The Bayes estimator is a weighted average of $\bar{X}$ (the sample mean based on the data) and $\mathsf{E}[p] = \frac{\alpha}{\alpha+\beta}$ (the mean of the prior distribution).

Bayesian Estimation

- *Definition L7.6* (Def 7.2.15 on p.325): Let $\mathcal{F}$ denote the class of pdfs or pmfs $f(x|\theta)$ (indexed by θ). A class Π of prior distributions is a *conjugate family* for $\mathcal{F}$ if the posterior distribution is in the class Π for all $f \in \mathcal{F}$, all priors in Π, and all $x \in \mathcal{X}$.
- As seen in Example L7.7, the beta family is conjugate for the binomial family.