Lecture 13 — Maximum likelihood estimation

Last lecture, we introduced the method of moments for estimating one or more parameters θ in a parametric model. This lecture, we discuss a different method called maximum likelihood estimation. The focus of this lecture will be on how to compute this estimate; subsequent lectures will study its statistical properties.

13.1 Maximum likelihood estimation

Consider data $X_1, \ldots, X_n \overset{IID}{\sim} f(x|\theta)$, for a parametric model $\{f(x|\theta) : \theta \in \Omega\}$. Given the observed values $X_1, \ldots, X_n$ of the data, the function

$$\text{lik}(\theta) = f(X_1|\theta) \times \ldots \times f(X_n|\theta)$$

of the parameter θ is called the **likelihood function**. If $f(x|\theta)$ is the PMF of a discrete distribution, then $\text{lik}(\theta)$ is simply the probability of observing the values $X_1, \ldots, X_n$ if the true parameter were θ. The **maximum likelihood estimator (MLE)** of θ is the value of $\theta \in \Omega$ that maximizes $\text{lik}(\theta)$. Intuitively, it is the value of θ that makes the observed data "most probable" or "most likely".

The idea of maximum likelihood is related to the use of the likelihood ratio statistic in the Neyman-Pearson lemma. Recall that for testing

$$H_0 : (X_1, \ldots, X_n) \sim g$$
$$H_1 : (X_1, \ldots, X_n) \sim h$$

where g and h are joint PDFs or PMFs for n random variables, the most powerful test rejects for small values of the likelihood ratio

$$L(X_1, \ldots, X_n) = \frac{g(X_1, \ldots, X_n)}{h(X_1, \ldots, X_n)}.$$

In the context of a parametric model, we may consider testing $H_0 : X_1, \ldots, X_n \overset{IID}{\sim} f(x|\theta_0)$ versus $H_1 : X_1, \ldots, X_n \overset{IID}{\sim} f(x|\theta_1)$, for two different parameter values $\theta_0, \theta_1 \in \Omega$. Then

$$g(X_1, \ldots, X_n) = f(X_1|\theta_0) \times \ldots \times f(X_n|\theta_0),$$
$$h(X_1, \ldots, X_n) = f(X_1|\theta_1) \times \ldots \times f(X_n|\theta_1),$$

so the likelihood ratio is exactly $\text{lik}(\theta_0)/\text{lik}(\theta_1)$. The MLE (if it exists and is unique) is the value of $\theta \in \Omega$ for which $\text{lik}(\theta)/\text{lik}(\theta') > 1$ for any other value $\theta' \in \Omega$.

13.2 Examples

Computing the MLE is an optimization problem. Maximizing $\text{lik}(\theta)$ is equivalent to maximizing its (natural) logarithm

$$l(\theta) = \log\left(\text{lik}(\theta)\right) = \sum_{i=1}^n \log f(X_i|\theta),$$

which in many examples is easier to work with as it involves a sum rather than a product. Let's work through several examples:

Example 13.1. Let $X_1, \ldots, X_n \overset{IID}{\sim} \text{Poisson}(\lambda)$. Then

$$\begin{aligned} l(\lambda) &= \sum_{i=1}^n \log \frac{\lambda^{X_i} e^{-\lambda}}{X_i!} \\ &= \sum_{i=1}^n (X_i \log \lambda - \lambda - \log(X_i!)) \\ &= (\log \lambda) \sum_{i=1}^n X_i - n\lambda - \sum_{i=1}^n \log(X_i!). \end{aligned}$$

This is differentiable in λ, so we maximize $l(\lambda)$ by setting its first derivative equal to 0:

$$0 = l'(\lambda) = \frac{1}{\lambda} \sum_{i=1}^n X_i - n.$$

Solving for λ yields the estimate $\hat{\lambda} = \bar{X}$. Since $l(\lambda) \to -\infty$ as $\lambda \to 0$ or $\lambda \to \infty$, and since $\hat{\lambda} = \bar{X}$ is the unique value for which $0 = l'(\lambda)$, this must be the maximum of l. In this example, $\hat{\lambda}$ is the same as the method-of-moments estimate.

Example 13.2. Let $X_1, \ldots, X_n \overset{IID}{\sim} \mathcal{N}(\mu, \sigma^2)$. Then

$$\begin{aligned} l(\mu, \sigma^2) &= \sum_{i=1}^n \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(X_i-\mu)^2}{2\sigma^2}}\right) \\ &= \sum_{i=1}^n \left(-\frac{1}{2}\log(2\pi\sigma^2) - \frac{(X_i-\mu)^2}{2\sigma^2}\right) \\ &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^n (X_i-\mu)^2. \end{aligned}$$

Considering σ^2 (rather than σ) as the parameter, we maximize $l(\lambda)$ by settings its partial derivatives with respect to μ and σ^2 equal to 0:

$$\begin{aligned} 0 &= \frac{\partial l}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^n (X_i - \mu), \\ 0 &= \frac{\partial l}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^n (X_i-\mu)^2. \end{aligned}$$

Solving the first equation yields $\hat{\mu} = \bar{X}$, and substituting this into the second equation yields $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$. Since $l(\mu, \sigma^2) \to -\infty$ as $\mu \to -\infty$, $\mu \to \infty$, $\sigma^2 \to 0$, or $\sigma^2 \to \infty$, and as $(\hat{\mu}, \hat{\sigma}^2)$ is the unique value for which $0 = \frac{\partial l}{\partial \mu}$ and $0 = \frac{\partial l}{\partial \sigma^2}$, this must be the maximum of l. Again, the MLEs are the same as the method-of-moments estimates.

Example 13.3. Let $X_1, \ldots, X_n \overset{IID}{\sim} \text{Gamma}(\alpha, \beta)$. Then

$$\begin{aligned} l(\alpha, \beta) &= \sum_{i=1}^n \log\left(\frac{\beta^\alpha}{\Gamma(\alpha)} X_i^{\alpha-1} e^{-\beta X_i}\right) \\ &= \sum_{i=1}^n (\alpha \log \beta - \log \Gamma(\alpha) + (\alpha - 1)\log X_i - \beta X_i) \\ &= n\alpha \log \beta - n \log \Gamma(\alpha) + (\alpha - 1)\sum_{i=1}^n \log X_i - \beta \sum_{i=1}^n X_i. \end{aligned}$$

To maximize $l(\alpha, \beta)$, we set its partial derivatives equal to 0:

$$\begin{aligned} 0 &= \frac{\partial l}{\partial \alpha} = n \log \beta - \frac{n\Gamma'(\alpha)}{\Gamma(\alpha)} + \sum_{i=1}^n \log X_i, \\ 0 &= \frac{\partial l}{\partial \beta} = \frac{n\alpha}{\beta} - \sum_{i=1}^n X_i. \end{aligned}$$

The second equation implies that the MLEs $\hat{\alpha}$ and $\hat{\beta}$ satisfy $\hat{\beta} = \hat{\alpha}/\bar{X}$. Substituting into the first equation and dividing by n, $\hat{\alpha}$ satisfies

$$0 = \log \hat{\alpha} - \frac{\Gamma'(\hat{\alpha})}{\Gamma(\hat{\alpha})} - \log \bar{X} + \frac{1}{n}\sum_{i=1}^n \log X_i. \tag{13.1}$$

The function $f(\alpha) = \log \alpha - \frac{\Gamma'(\alpha)}{\Gamma(\alpha)}$ decreases from ∞ to 0 as α increases from 0 to ∞, and the value $-\log \bar{X} + \frac{1}{n}\sum_{i=1}^n \log X_i$ is always negative (by Jensen's inequality)—hence (13.1) always has a single unique root $\hat{\alpha}$, which is the MLE for α. The MLE for β is then $\hat{\beta} = \hat{\alpha}/\bar{X}$.

Unfortunately there is no closed-form expression for this root $\hat{\alpha}$. (In particular, the MLE $\hat{\alpha}$ is *not* the method-of-moments estimator for α.) We may compute the root numerically using the **Newton-Raphson method**: We start with an initial guess $\alpha^{(0)}$, which (for example) may be the method-of-moments estimator

$$\alpha^{(0)} = \frac{\bar{X}^2}{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2}.$$

Having computed $\alpha^{(t)}$ for any $t = 0, 1, 2, \ldots$, we compute the next iteration $\alpha^{(t+1)}$ by approximating the equation (13.1) with a linear equation using a first-order Taylor expansion around $\hat{\alpha} = \alpha^{(t)}$, and set $\alpha^{(t+1)}$ as the value of $\hat{\alpha}$ that solves this linear equation. In detail, let $f(\alpha) = \log \alpha - \frac{\Gamma'(\alpha)}{\Gamma(\alpha)}$. A first-order Taylor expansion around $\hat{\alpha} = \alpha^{(t)}$ in (13.1) yields the linear approximation

$$0 \approx f(\alpha^{(t)}) + (\hat{\alpha} - \alpha^{(t)}) f'(\alpha^{(t)}) - \log \bar{X} + \frac{1}{n}\sum_{i=1}^n \log X_i,$$

and we set $\alpha^{(t+1)}$ to be the value of $\hat{\alpha}$ solving this linear equation, i.e.[1]

$$\alpha^{(t+1)} = \alpha^{(t)} + \frac{-f(\alpha^{(t)}) + \log \bar{X} - \frac{1}{n}\sum_{i=1}^n \log X_i}{f'(\alpha^{(t)})}.$$

The iterations $\alpha^{(0)}, \alpha^{(1)}, \alpha^{(2)}, \ldots$ converge to the MLE $\hat{\alpha}$.

Example 13.4. Let $(X_1, \ldots, X_k) \sim \text{Multinomial}(n, (p_1, \ldots, p_k))$. (This is not quite the setting of n IID observations from a parametric model, as we have been considering, although you can think of $(X_1, \ldots, X_k)$ as a summary of n such observations $Y_1, \ldots, Y_n$ from the parametric model Multinomial$(1, (p_1, \ldots, p_k))$, where Y_i indicates which of k possible outcomes occurred for the ith observation.) The log-likelihood is given by

$$l(p_1, \ldots, p_k) = \log\left(\binom{n}{X_1, \ldots, X_k} p_1^{X_1} \ldots p_k^{X_k}\right) = \log\binom{n}{X_1, \ldots, X_k} + \sum_{i=1}^k X_i \log p_i,$$

and the parameter space is

$$\Omega = \{(p_1, \ldots, p_k) : 0 \leq p_i \leq 1 \text{ for all } i \text{ and } p_1 + \ldots + p_k = 1\}.$$

To maximize $l(p_1, \ldots, p_k)$ subject to the linear constraint $p_1 + \ldots + p_k = 1$, we may use the method of **Lagrange multipliers**: Consider the Lagrangian

$$L(p_1, \ldots, p_k, \lambda) = \log\binom{n}{X_1, \ldots, X_k} + \sum_{i=1}^k X_i \log p_i + \lambda(p_1 + \ldots + p_k - 1),$$

for a constant λ to be chosen later. Clearly, subject to $p_1 + \ldots + p_k = 1$, maximizing $l(p_1, \ldots, p_k)$ is the same as maximizing $L(p_1, \ldots, p_k, \lambda)$. Ignoring momentarily the constraint $p_1 + \ldots + p_k = 1$, the unconstrained maximizer of L is obtained by setting for each $i = 1, \ldots, k$

$$0 = \frac{\partial L}{\partial p_i} = \frac{X_i}{p_i} + \lambda,$$

which yields $\hat{p}_i = -X_i/\lambda$. For the specific choice of constant $\lambda = -n$, we obtain $\hat{p}_i = X_i/n$ and $\sum_{i=1}^n \hat{p}_i = \sum_{i=1}^n X_i/n = 1$, so the constraint is satisfied. As $\hat{p}_i = X_i/n$ is the unconstrained maximizer of $L(p_1, \ldots, p_k, -n)$, this implies that it must also be the constrained maximizer of $L(p_1, \ldots, p_k, -n)$, so it is the constrained maximizer of $l(p_1, \ldots, p_k)$. So the MLE is given by $\hat{p}_i = X_i/n$ for $i = 1, \ldots, k$.

[1] If this update yields $\alpha^{(t+1)} \leq 0$, we may reset $\alpha^{(t+1)}$ to be a very small positive value.