Lecture 4: Maximum Likelihood Estimation

(Text Section 1.6)

Maximum likelihood estimation (ML estimation) is another estimation method. In the case of the linear model with errors distributed as $N(0, \sigma^2)$, the ML and least-squares estimators are the same. However, in the general case, this is not true. Often the variance of the ML estimator (MLE) is less than the variance of other estimators (including the least-squares estimator), and hence is the preferable estimation method. MLEs also have many other nice properties, such as the fact that $\hat{\boldsymbol{\beta}}$ gets close to $\boldsymbol{\beta}$ (the true parameter) with high probability as the sample size gets large. For these reasons, we will focus on ML estimation in this course.

The *likelihood function* is algebraically the same as the probability distribution of the observed data. The joint distribution of $Y_1, \ldots, Y_n$, $f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n; \boldsymbol{\theta})$, is seen as a function of $y_1, \ldots, y_n$ with fixed parameters $\boldsymbol{\theta}$. In contrast, the likelihood, $\mathcal{L}(\boldsymbol{\theta}; y_1, \ldots, y_n)$ is seen as a function of $\boldsymbol{\theta}$ for a given set of data points.

The idea behind ML estimation is that we try to find the values of the parameters $\boldsymbol{\theta}$ that seem most likely, given our observed data. To do this, we locate $\hat{\boldsymbol{\theta}}$, the value which maximizes $\mathcal{L}(\boldsymbol{\theta}; \mathbf{y})$. The value $\hat{\boldsymbol{\theta}}$ also maximizes the function $\log \mathcal{L}(\boldsymbol{\theta}; \mathbf{y})$, since log is a monotonically increasing function. Often, it is easier to maximize the log-likelihood than the likelihood itself.

Example: $Y_1, \ldots, Y_n$ independent, $Y_i \sim \mathrm{N}(\mu, \sigma^2)$ with σ^2 known. The likelihood is

$$\begin{aligned}
\mathcal{L}(\mu; \mathbf{y}) &= \prod_{i=1}^{n} f_{Y_i}(y_i) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y_i - \mu)^2}{2\sigma^2}\right\} \\
&= (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right\}
\end{aligned}$$

and the log-likelihood is

$$\log \mathcal{L}(\mu; \mathbf{y}) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2.$$

Differentiating with respect to μ, we obtain

$$\begin{aligned}
\frac{d}{d\mu}\log \mathcal{L}(\mu; \mathbf{y}) &= -\frac{1}{2\sigma^2}\frac{d}{d\mu}\sum_{i=1}^{n}(y_i - \mu)^2 \\
&= \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \mu)
\end{aligned}$$

The derivative of the log-likelihood is known as the *score function*. To find the MLE, we set the score function equal to 0 and solve:

$$\begin{aligned} 0 &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \hat{\mu}) \\ \hat{\mu} &= \frac{1}{n}\sum_{i=1}^{n} y_i \equiv \bar{y}. \end{aligned}$$

To prove that an estimator is a *maximum* of the likelihood function (not a minimum or saddle point), we take the second derivatives of $\log \mathcal{L}(\boldsymbol{\theta}; \mathbf{y})$ with respect to the unknown parameters, i.e. we calculate $\frac{\partial^2}{\partial\theta_j \partial\theta_k} \log \mathcal{L}(\boldsymbol{\theta}; \mathbf{y})$ for all j, k. The matrix with $(j,k)^{th}$ entry $\frac{\partial^2}{\partial\theta_j \partial\theta_k} \log \mathcal{L}(\boldsymbol{\theta}; \mathbf{y})$ is called the *Hessian*.

Assuming that $\hat{\boldsymbol{\theta}}$ is the only root of the score function, a sufficient condition for $\hat{\boldsymbol{\theta}}$ to be a maximum is that this matrix of second derivatives be negative definite when evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$. When there is only one parameter (i.e. $\boldsymbol{\theta}$ is 1-dimensional, and we use the notation θ), it is sufficient to check that

$$\frac{\partial^2}{\partial\theta^2} \log \mathcal{L}(\theta; \mathbf{y})\Big|_{\theta=\hat{\theta}} < 0.$$

Alternatively, it is sufficient to show that the second derivative is negative for all θ (including $\hat{\theta}$).

Strictly speaking, we should also check that the maximum of the log-likelihood does not occur at one of the boundaries of the parameter space. However, I will not require you to do this.

For example, in the above case where $Y_1, \ldots, Y_n$ are iid $\mathrm{N}(\mu, \sigma^2)$ random variables with σ^2 known, we can calculate

$$\begin{aligned} \frac{d^2}{d\mu^2} \log \mathcal{L}(\mu; \mathbf{y}) &= \frac{d}{d\mu}\left\{\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \mu)\right\} \\ &= -\frac{n}{\sigma^2} \\ &< 0, \end{aligned}$$

which proves that $\hat{\mu} = \bar{y}$ is indeed the MLE of μ.

In the case where $Y_1, \ldots, Y_n$ are independent with $Y_i \sim \mathrm{N}(\mathbf{x}_i'\boldsymbol{\beta}, \sigma^2)$, σ^2 known, the Hessian is

$$-\frac{1}{\sigma^2}\mathbf{X}'\mathbf{X}.$$

This matrix can be shown to be negative definite.

Example (normal distribution, cont.)

If σ^2 is also unknown, we can use the ML method to estimate it. Differentiating the log-likelihood with respect to σ^2:

$$\begin{aligned}
\frac{d}{d\sigma^2}\log\mathcal{L}(\boldsymbol{\theta};\mathbf{y}) &= -\frac{n}{2\sigma^2}+\frac{\sum_{i=1}^n(y_i-\mathbf{x}_i'\boldsymbol{\beta})^2}{2\sigma^4}\\
0 &= -\frac{n}{2\hat{\sigma}^2}+\frac{\sum_{i=1}^n(y_i-\mathbf{x}_i'\hat{\boldsymbol{\beta}})^2}{2\hat{\sigma}^4}\\
\hat{\sigma^2} &= \frac{\sum_{i=1}^n(y_i-\mathbf{x}_i'\hat{\boldsymbol{\beta}})^2}{n}
\end{aligned}$$

Note that the MLE of σ^2 is biased (i.e., $\mathrm{E}[\hat{\sigma^2}]\neq\sigma^2$)! That is why, in practice, we usually use the unbiased estimator $\frac{\sum_{i=1}^n(y_i-\mathbf{x}_i'\hat{\boldsymbol{\beta}})^2}{n-p}$.

ML estimation is a useful method for estimating parameters of many different models. Sometimes this method leads to closed forms for the parameter estimates (e.g. normal case), and other times we need to use numerical methods to locate the MLEs.

Example (Poisson distribution)

Some laboratory test are run on samples of river water in order to determine whether the water is safe for swimming. Of particular interest is the concentration of coliform bacteria in the water. A collection of n independent unit-volume samples are taken from the river, and the number of bacteria in each are counted. Let $Y_1,\ldots,Y_n$ represent these counts. Assume that the counts are Poisson distributed with mean μv, where v is the volume (in this case 1) and μ is an unknown parameter. Problem: Estimate the mean number of coliform per unit-volume in this river using the ML method.

$$\begin{aligned}
\mathcal{L}(\mu;\mathbf{y}) &= \prod_{i=1}^n f_{Y_i}(y_i)\\
&= \prod_{i=1}^n \frac{e^{-\mu}\mu^{y_i}}{y_i!}
\end{aligned}$$

So

$$\begin{aligned}
\log\mathcal{L}(\mu;\mathbf{y}) &= -n\mu+\log\mu\sum_{i=1}^n y_i-\sum_{i=1}^n\log y_i!\\
\frac{d}{d\mu}\log\mathcal{L}(\mu;\mathbf{y}) &= -n+\frac{\sum_{i=1}^n y_i}{\mu}\\
0 &= -n+\frac{\sum_{i=1}^n y_i}{\hat{\mu}}\\
\hat{\mu} &= \frac{\sum_{i=1}^n y_i}{n}\equiv\bar{y}.
\end{aligned}$$

In this case, the model has only one unknown parameter (μ), so it is easy to verify that

$$\left.\frac{d^2}{d\mu^2}\log\mathcal{L}(\mu;\mathbf{y})\right|_{\mu=\hat{\mu}} = -\frac{\sum_{i=1}^n y_i}{\hat{\mu}^2} \leq 0,$$

i.e. that $\hat{\mu}$ is indeed a maximum.

Example (Bernoulli distribution)

With reference to the last example, say that we can't count the *number* of bacteria in a sample of river water but rather can determine only whether or not *any* bacteria are present. If y out of n test tubes containing volume v of water test negative for the presence of bacteria, what is the MLE of μ?

Let the number of bacteria in a test tube be N. Then

$$\begin{aligned}
\mathrm{P}(\text{test is negative}) &\equiv p \\
&= \mathrm{P}(N=0) \\
&= \frac{(\mu v)^0 e^{-\mu v}}{0!} \\
&= e^{-\mu v}
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{P}(\text{test is positive}) &= \mathrm{P}(N>0) \\
&= 1-\mathrm{P}(N=0) \\
&= 1-p
\end{aligned}$$

Since disjoint volumes are independent, the n test tubes constitute independent samples. The likelihood is then

$$\mathcal{L}(p;y) = \left(\begin{array}{c} n \\ y \end{array}\right) p^y (1-p)^{n-y}$$

and

$$\log\mathcal{L}(p;y) = \log\left(\begin{array}{c} n \\ y \end{array}\right) + y\log p + (n-y)\log(1-p).$$

Differentiating, we get

$$\begin{aligned}
\frac{d}{dp}\log\mathcal{L}(p;y) &= \frac{y}{p} - \frac{n-y}{1-p} \\
0 &= \frac{y}{\hat{p}} - \frac{n-y}{1-\hat{p}} \\
\hat{p} &= \frac{y}{n}
\end{aligned}$$

To confirm that $\hat{p}$ is indeed the MLE, we compute

$$\begin{aligned}
\frac{d^2}{dp^2}\log\mathcal{L}(p;y) &= -\frac{y}{p^2} - \frac{n-y}{(1-p)^2} \\
&< 0.
\end{aligned}$$

We can now use the *invariance property* of the MLEs to compute $\hat{\mu}$. This property states that if we reparameterize the model, we can simply "plug in" the MLEs of the reparameterized model to get the MLEs of the original model, i.e.

$$\begin{aligned}\hat{\mu} &= -\frac{1}{v}\log\hat{p} \\ &= \frac{\log n - \log y}{v}.\end{aligned}$$

Example (logarithmic distribution)

Let $Y_1, \ldots, Y_n$ be the number of distinct animal species observed in independent quadrants $1, \ldots, n$. Assume that Y_i has a logarithmic distribution with parameter θ, i.e.

$$f_{Y_i}(y_i) = \frac{(1-\theta)^{y_i}}{-y_i \log\theta},\ y_i = 1, 2, \ldots,\ \text{and } 0 < \theta < 1.$$

Problem: Find the MLE of θ.

$$\begin{aligned}\mathcal{L}(\theta; \mathbf{y}) &= \prod_{i=1}^{n} \frac{(1-\theta)^{y_i}}{-y_i \log\theta} \\ &= \frac{(1-\theta)^t}{(-\log\theta)^n \prod_{i=1}^{n} y_i}\end{aligned}$$

where $t = \sum_{i=1}^{n} y_i$. So,

$$\begin{aligned}\log\mathcal{L}(\theta; \mathbf{y}) &= t\log(1-\theta) - n\log(-\log(\theta)) - \sum_{i=1}^{n}\log y_i \\ \frac{d}{d\theta}\log\mathcal{L}(\theta; \mathbf{y}) &= -\frac{t}{1-\theta} - \frac{n}{\theta\log(\theta)}\end{aligned}$$

The solution to $\frac{d}{d\theta}\log\mathcal{L}(\theta; \mathbf{y}) = 0$ cannot be found analytically. One way of finding the MLE numerically is the *Newton-Raphson* or *scoring* method.