Chapter 1

Maximum Likelihood Estimation

1.1 Parametric Model

A parametric model is a complete specification of the distribution. Once the parameter is given, the distribution function is determined. Instead, a semiparametric model only gives a few features rather than a complete description of the distribution.

Example 1.1. Semiparametric model: If we know $X \sim i.i.d. \left(\mu, \sigma^2\right)$, we can estimate μ, σ^2 by method of moments.

Parametric model: If we assume $X \sim N\left(\mu, \sigma^2\right)$, we can estimate μ, σ^2 by MLE.

Example 1.2. Conditional model: the conditioning variable can be viewed as if it is fixed and the randomness comes from the error term only.

$$y = X'\beta + \varepsilon$$

x is the conditional variable. The condition $E(\varepsilon|X) = 0$ together with a full rank $E[XX']$can help to identify β. This is semiparametric model. However, if we assume $f(\varepsilon \mid X) \sim N\left(0, \sigma^2\right)$, then conditional parametric model as it completely describes $f(y \mid X)$ and it becomes a conditional parametric model.

Definition 1.1. Parametric model. The distribution of the data $(x_1, ..., x_n)$ is known up to a finite dimensional parameter.

Let Θ be the parameter space a researcher specifies.

Definition 1.2. A model is **correctly specified**, if the true DGP is $f(X \mid \theta_0)$ for some $\theta_0 \in \Theta$. Otherwise, the model is **misspecified**.

1.2 Likelihood

In this chapter we will mostly talk about unconditional models. The results can be carried over to conditional models. To keep the setting simple, let $(X_1, \ldots, X_n)$ be i.i.d. The **likelihood** of the sample is $\prod_{i=1}^n f(X_i \mid \theta_0)$. The **log-likelihood** is

$$\ell_n(\theta) = \frac{1}{n}\sum_{i=1}^n \log f(X_i \mid \theta).$$

Here, we put $1/n$ to average the log-likelihood. This scaling factor does not change the estimation at all.

In practice, we work with the log-likelihood, which is more convenient. the MLE estimator is defined as

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \ell_n(\theta).$$

To justify the likelihood principle, consider the population version of the

$$\ell(\theta) = E[\log f(X \mid \theta)]$$

Theorem 1.1. *When model is correctly specified, θ_0 is the maximizer.*

Proof. Kullback-Leibler distence

$$\begin{aligned} E[\log p(\theta_0)] - E[\log p(\theta)] &= E[\log(p(\theta_0)/p(\theta))] \\ &= -E[\log(p(\theta)/p(\theta_0))] \\ &\geq -\log E[p(\theta)/p(\theta_0)] = 0 \end{aligned}$$

where the inequality holds by Jensen's inequality for the convex function $-\log(\cdot)$. □

1.3 Score, Hessian, and Information

Score:

$$\psi_n(\theta) = \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log f(X_i \mid \theta)$$

Hessian:

$$\mathscr{H}_n(\theta) = -\sum_{i=1}^{n} \frac{\partial^2}{\partial\theta\partial\theta'} \log f(X_i \mid \theta)$$

Efficient score:

$$\psi_0 = \frac{\partial}{\partial\theta} \log f(X_i \mid \theta_0)$$

Theorem 1.2. *If the model is correctly specified, the support of X does not depend on θ, and θ_0 is in the interior of Θ, then $E(\psi_0) = 0$.*

Proof. By the Leibniz rule,

$$E(\psi_0) = E\left[\frac{\partial}{\partial\theta} \log f(X_i \mid \theta_0)\right] = \frac{\partial}{\partial\theta} E[\log f(X_i \mid \theta_0)] = 0$$

as θ_0 is the maximizer in an interior. □

Definition 1.3. Fisher information matrix:

$$\mathscr{I}_0 = E[\psi_0 \psi_0']$$

Definition 1.4. Expected Hessian:

$$\mathscr{H}_0 = -E\left[\frac{\partial^2}{\partial\theta\partial\theta'} \log f(X \mid \theta_0)\right]$$

Theorem 1.3. *If the model is correctly specified, we have the* ***information matrix equality:*** $\mathscr{I}_0 = \mathscr{H}_0$.

Proof. Start with Hessian,

$$\begin{aligned} E\left[\frac{\partial^2}{\partial\theta\partial\theta'}\log f(\theta_0)\right] &= E\left[\frac{\partial}{\partial\theta}\frac{\partial}{\partial\theta'}\log f(\theta_0)\right] \\ &= E\left[\frac{\partial}{\partial\theta}\frac{\frac{\partial}{\partial\theta'}f(\theta)}{f(\theta)}\Big|_{\theta=\theta_0}\right] \\ &= E\left[\frac{\frac{\partial^2}{\partial\theta\partial\theta'}f(\theta)}{f(\theta_0)}\right] + E\left[\frac{\frac{\partial}{\partial\theta}f(\theta)\frac{\partial}{\partial\theta'}f(\theta)}{f^2(\theta_0)}\right]. \end{aligned}$$

The first term:

$$E\left[\frac{\frac{\partial^2}{\partial\theta\partial\theta'}f(\theta)}{f(\theta_0)}\right] = \int\frac{\frac{\partial^2}{\partial\theta\partial\theta'}f(\theta)}{f(\theta_0)}f(\theta_0)\,dx = \int\frac{\partial^2}{\partial\theta\partial\theta'}f(\theta)\,dx = \frac{\partial^2}{\partial\theta\partial\theta'}\int f(\theta)\,dx = \frac{\partial^2}{\partial\theta\partial\theta'}1 = 0.$$

The second term:

$$E\left[\frac{\frac{\partial}{\partial\theta}f(\theta)\frac{\partial}{\partial\theta'}f(\theta)}{f^2(\theta_0)}\right] = E\left[\frac{\partial}{\partial\theta}\log f(\theta_0)\frac{\partial}{\partial\theta'}\log f(\theta_0)\right] = E\left[\psi_0\psi_0'\right].$$

□

Notice that the information matrix equality holds only when the model is correctly specified. It fails when the model is misspecified.

1.4 Cramér-Rao Lower Bound

Theorem 1.4. *Suppose the model is correctly specified, the support of X does not depend on θ, and θ_0 is in the interior of Θ. If $\widetilde{\theta}$ is unbiased estimator, then var $\left(\widetilde{\theta}\right) \geq (n\mathscr{I}_0)^{-1}$.*

Proof. Because of unbiasedness,

$$\theta = E_\theta\left[\widetilde{\theta}\right] = \int\widetilde{\theta}f(\mathbf{X}\mid\theta)\,d\mathbf{x}$$

for any $\theta\in\Theta$. $\mathbf{X}$ here is for the entire sample, $f(\mathbf{X}\mid\theta) = f(X_1,...,X_n\mid\theta) = \prod_{i=1}^n f(X_i\mid\theta)$. Take derivative at the two sides. The LHS is

$$\frac{\partial\theta}{\partial\theta'} = \mathbf{I}_p$$

. The RHS:

$$\begin{aligned} \frac{\partial}{\partial\theta'}\int\widetilde{\theta}f(\mathbf{X}\mid\theta)\,d\mathbf{x} &= \int\widetilde{\theta}\frac{\partial}{\partial\theta'}f(\mathbf{X}\mid\theta)\,d\mathbf{x} \\ &= \int\widetilde{\theta}\frac{\frac{\partial}{\partial\theta'}f(\mathbf{X}\mid\theta)}{f(\mathbf{X}\mid\theta)}f(\mathbf{X}\mid\theta)\,d\mathbf{x} \\ &= \int\widetilde{\theta}\frac{\partial}{\partial\theta'}\log f(\mathbf{X}\mid\theta)f(\mathbf{X}\mid\theta)\,d\mathbf{x} \\ &= \int\widetilde{\theta}\psi_n(\theta)f(\mathbf{X}\mid\theta)\,d\mathbf{x} \end{aligned}$$

Evaluate at the true θ_0, and due to i.i.d. data

$$\mathbf{I}_p = \int \widetilde{\theta}\psi_n(\theta_0) f(\mathbf{X} \mid \theta_0)\, d\mathbf{x} = E\left[\widetilde{\theta}\psi_n(\theta_0)\right] = E\left[\left(\widetilde{\theta} - \theta_0\right)\psi_n(\theta_0)\right]$$

where the last equality holds by $E\left[\theta_0\psi_n(\theta_0)\right] = \theta_0 E\left[\psi_n(\theta_0)\right] = \theta_0 E\left[n\psi_0\right] = 0$. We thus have

$$var\begin{pmatrix} \widetilde{\theta} - \theta_0 \\ \psi_n(\theta_0) \end{pmatrix} = \begin{bmatrix} \boldsymbol{V} & \mathbf{I}_p \\ \mathbf{I}_p & n\mathscr{I}_0 \end{bmatrix}.$$

Pre- and post-multiply $\begin{bmatrix} \boldsymbol{I}_p & -(n\mathscr{I}_0)^{-1} \end{bmatrix}$, we have

$$\begin{bmatrix} \boldsymbol{I}_p & -(n\mathscr{I}_0)^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{V} & \boldsymbol{I}_p \\ \boldsymbol{I}_p & n\mathscr{I}_0 \end{bmatrix} \begin{bmatrix} \boldsymbol{I}_p \\ -(n\mathscr{I}_0)^{-1} \end{bmatrix} = \boldsymbol{V} - (n\mathscr{I}_0)^{-1} \geq 0.$$

□

The Cramér-Rao Lower Bound is a lower bound. It may not reachable. When it is reached, an estimator is **Cramér-Rao efficient** if it is unbiased and the variance is $(n\mathscr{I}_0)^{-1}$.

Example 1.3. Normal distribution: Let $\gamma = \sigma^2$

$$\log \ell_n\left(X \mid \mu, \sigma^2\right) = -\frac{n}{2}\log\gamma - \frac{n}{2}\log\pi - \frac{1}{2\gamma}\sum_{i=1}^{n}(X_i - \mu)^2$$

$$\psi_n\left(\mu, \sigma^2\right) = \begin{cases} \frac{1}{\gamma}\sum_{i=1}^{n}(X_i - \mu) \\ -\frac{n}{2\gamma} + \frac{1}{2\gamma^2}\sum_{i=1}^{n}(X_i - \mu)^2 \end{cases}$$

$$\mathscr{H}_n\left(\mu, \sigma^2\right) = \begin{bmatrix} \frac{n}{\gamma} & \frac{1}{2\gamma^2}\sum_{i=1}^{n}(X_i - \mu) \\ \frac{1}{2\gamma^2}\sum_{i=1}^{n}(X_i - \mu) & -\frac{n}{2\gamma^2} + \frac{1}{\gamma^3}\sum_{i=1}^{n}(X_i - \mu)^2 \end{bmatrix}$$

Expected Hessian:

$$E\left[\mathscr{H}_n\left(\mu, \sigma^2\right)\right] = \begin{bmatrix} \frac{n}{\gamma} & 0 \\ 0 & \frac{n}{2\gamma^2} \end{bmatrix}$$

Take inverse:

$$\begin{bmatrix} \frac{\gamma}{n} & 0 \\ 0 & 2\frac{\gamma^2}{n} \end{bmatrix}$$

This is the lower bound.
Check:
the sample mean:

$$var\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right) = \frac{\sigma^2}{n}$$

The sample mean is Cramér-Rao efficient.

$$s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 = \frac{1}{n-1}X'\left(I - \frac{1}{n}1_n 1_n'\right)X$$

$E\left(S_n^2\right)=\sigma^2$ is unbiased

$$(n-1)\frac{s_n^2}{\sigma^2}=\left(\frac{X}{\sigma}\right)'\left(I-\frac{1}{n}1_n1_n'\right)\left(\frac{X}{\sigma}\right)\sim\chi^2(n-1)$$

So,

$$s_n^2=\frac{\chi^2(n-1)}{n-1}\sigma^2$$

$$var\left(s_n^2\right)=\frac{\sigma^4}{(n-1)^2}2(n-1)=\frac{2\sigma^4}{n-1}>\frac{2\sigma^4}{n}$$

Does not satisfy Cramér-Rao efficient.

1.5 Asymptotic Normality

MLE is a special case of m-estimator. Under regularity conditions, $\hat{\theta}\xrightarrow{p}\theta_0$, and asymptotically normal:

$$\sqrt{n}\left(\hat{\theta}-\theta_0\right)\xrightarrow{d}N\left(0,\mathscr{H}_0^{-1}\mathscr{I}_0\mathscr{H}_0^{-1}\right)$$

When the information equality hods, the asymptotic variance is simplified as $\mathscr{I}_0^{-1}\mathscr{I}_0\mathscr{I}_0^{-1}=\mathscr{I}_0^{-1}$, and thus

$$\sqrt{n}\left(\hat{\theta}-\theta_0\right)\xrightarrow{d}N\left(0,\mathscr{I}_0^{-1}\right).$$

In other words, it achieves asymptotic efficiency.

Caveat:

1. need correct specification
2. the comparison is restricted to asymptotically unbiased estimator. There are biased estimators with better overall performance.

1.6 Kullback-Leibler Divergence

$$KLIC(f,g)=\int f(x)\log\frac{f(x)}{g(x)}dx$$

Properties:

1. $KLIC(f,f)=0$
2. $KLIC(f,g)\geq 0$
3. $f=\arg\min_g KLIC(f,g)$

If $f(x)=f(x\mid\theta)$ is a parametric family

$$\theta_0=\arg\min_{\theta\in\Theta}KLIC(f,f_\theta)$$

which is correctly specified model.

Pseudo-true parameter:

$$\theta_0=\arg\min_{\theta\in\Theta}KLIC(f,f_\theta)$$

which is misspecified model.

KLIC is the distance measure of any two distributions.

$$\begin{aligned} KLIC\left(f, f_{\theta}\right) &= \int f\left(x\right)\log f\left(x\right)dx - \int f\left(x\right)\log f\left(x \mid \theta\right)dx \\ &= \int f\left(x\right)\log f\left(x\right)dx - E\left[\log f\left(x \mid \theta\right)\right] \\ &= \int f\left(x\right)\log f\left(x\right)dx - \ell\left(\theta\right) \end{aligned}$$

the pseudo-true value

$$\theta^{*} = \arg\max_{\theta \in \Theta} \ell\left(\theta\right)$$

The information equality was proved under correct specification. When the model is misspecified,

$$E\left[S\left(\theta^{*}\right)S\left(\theta^{*}\right)'\right] \neq E\left[\frac{\partial^{2}}{\partial\theta\partial\theta'}\log f\left(\theta^{*}\right)\right].$$

As a result, we will have a sandwich-form asymptotic variance in

$$\sqrt{n}\left(\hat{\theta} - \theta^{*}\right) \xrightarrow{d} N\left(0, \mathscr{H}_{*}^{-1}\mathscr{I}_{*}\mathscr{H}_{*}^{-1}\right)$$

understand that $\mathscr{I}_{*}$ and $\mathscr{H}_{*}$ are evaluated at the pseudo-true value.

Zhentao Shi. Feb 14, 2023. Transcribed by Shu Shen.