Maximum Likelihood Estimation

Zhou Yahong

School of Economics
SHUFE

Part I

Maximum Likelihood Estimation

Estimation

- Popularity due to its asymptotic optimality.

Estimation

- Popularity due to its asymptotic optimality.
- To maximize $L = \Pi_i f(x_i, \theta)$—likelihood—given realized X, find value of θ makes it most likely.

Estimation

- Popularity due to its asymptotic optimality.
- To maximize $L = \Pi_i f(x_i, \theta)$—likelihood—given realized X, find value of θ makes it most likely.
- Equivalently, to maximize

$$\ln L = \sum_{i=1}^{n} \ln f(x_i, \theta)$$

Estimation

- Popularity due to its asymptotic optimality.
- To maximize $L = \Pi_i f(x_i, \theta)$—likelihood—given realized X, find value of θ makes it most likely.
- Equivalently, to maximize

$$\ln L = \sum_{i=1}^{n} \ln f(x_i, \theta)$$

- First order condition

$$\frac{\partial \ln L}{\partial \theta} = \sum_{i=1}^{n} \frac{\partial \ln f(x_i, \theta)}{\partial \theta} = \sum_{i=1}^{n} g_i = 0$$

Estimation

- By approximation

$$
\begin{aligned}
0 &= \sum_{i=1}^{n} \frac{\partial \ln f(x_i, \hat{\theta})}{\partial \theta} \\
&= \sum_{i=1}^{n} \frac{\partial \ln f(x_i, \theta_0)}{\partial \theta} + \left[\sum_{i=1}^{n} \frac{\partial^2 \ln f(x_i, \theta_0)}{\partial \theta \partial \theta'} \right] (\hat{\theta} - \theta_0)
\end{aligned}
$$

Estimation

- By approximation

$$0 = \sum_{i=1}^{n} \frac{\partial \ln f(x_i, \hat{\theta})}{\partial \theta}$$
$$= \sum_{i=1}^{n} \frac{\partial \ln f(x_i, \theta_0)}{\partial \theta} + \left[\sum_{i=1}^{n} \frac{\partial^2 \ln f(x_i, \theta_0)}{\partial \theta \partial \theta'} \right] (\hat{\theta} - \theta_0)$$

- Usefulness of the FOC/Linearization—in MLE and Nonlinear regression.

Estimation

- By approximation

$$0 = \sum_{i=1}^{n} \frac{\partial \ln f(x_i, \hat{\theta})}{\partial \theta}$$
$$= \sum_{i=1}^{n} \frac{\partial \ln f(x_i, \theta_0)}{\partial \theta} + \left[\sum_{i=1}^{n} \frac{\partial^2 \ln f(x_i, \theta_0)}{\partial \theta \partial \theta'} \right] (\hat{\theta} - \theta_0)$$

- Usefulness of the FOC/Linearization—in MLE and Nonlinear regression.
  - Justify the limiting distribution.

Estimation

- By approximation

$$0 = \sum_{i=1}^{n} \frac{\partial \ln f(x_i, \hat{\theta})}{\partial \theta}$$
$$= \sum_{i=1}^{n} \frac{\partial \ln f(x_i, \theta_0)}{\partial \theta} + \left[\sum_{i=1}^{n} \frac{\partial^2 \ln f(x_i, \theta_0)}{\partial \theta \partial \theta'} \right] (\hat{\theta} - \theta_0)$$

- Usefulness of the FOC/Linearization—in MLE and Nonlinear regression.
  - Justify the limiting distribution.
  - Facilitate computation—iteration method.

Estimation

- Consistency, asymptotic normality, and efficiency,

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, nI(\theta_0)^{-1}) = N(0, E\left[-\frac{\partial^2 \ln f(x_i, \theta_0)}{\partial\theta\partial\theta'}\right]^{-1})$$

where

$$I(\theta) = E\left[-\frac{\partial^2 \ln L}{\partial\theta\partial\theta'}\right] = nE\left[-\frac{\partial^2 \ln f(x_i, \theta)}{\partial\theta\partial\theta'}\right]$$
$$= nE\left[\frac{\partial \ln f(x_i, \theta)}{\partial\theta}\frac{\partial \ln f(x_i, \theta)}{\partial\theta'}\right]$$

$I(\theta)$ is the information matrix

Estimation

- Estimate the asymptotic Variance of the MLE

Estimation

- Estimate the asymptotic Variance of the MLE
  - approach one

$$\hat{I}(\hat{\theta}) = n * \frac{1}{n} \sum_i \left[-\frac{\partial^2 \ln f(x_i, \hat{\theta})}{\partial\theta\partial\theta'} \right] = \left[-\frac{\partial^2 L(\hat{\theta})}{\partial\hat{\theta}\partial\hat{\theta}'} \right]$$

Estimation

- Estimate the asymptotic Variance of the MLE
  - approach one

$$\hat{I}(\hat{\theta}) = n * \frac{1}{n} \sum_i \left[-\frac{\partial^2 \ln f(x_i, \hat{\theta})}{\partial \theta \partial \theta'} \right] = \left[-\frac{\partial^2 L(\hat{\theta})}{\partial \hat{\theta} \partial \hat{\theta}'} \right]$$

  - approach two

$$\hat{I}^*(\hat{\theta}) = n * \frac{1}{n} \left[\sum_{i=1}^{n} \hat{g}_i \hat{g}_i' \right] = \left[\sum_{i=1}^{n} \hat{g}_i \hat{g}_i' \right]$$

  where $\hat{g}_i = g(x_i, \hat{\theta})$, which is known as BHHH estimator and the outer product of gradients, or OPG.

Three Asymptotically equivalent test procedures

- The null hypothesis H_0: $c(\theta) = q$. Let $\hat{\theta}$ denote the MLE and $\bar{\theta}$ the constrained MLE.

Three Asymptotically equivalent test procedures

- The null hypothesis H_0: $c(\theta) = q$. Let $\hat{\theta}$ denote the MLE and $\bar{\theta}$ the constrained MLE.
- Likelihood ration test: LR$= 2(\ln L(\hat{\theta}) - \ln L(\bar{\theta}))$, which asymptotically χ^2_J under the null hypothesis

Three Asymptotically equivalent test procedures

- The null hypothesis H_0: $c(\theta) = q$. Let $\hat{\theta}$ denote the MLE and $\bar{\theta}$ the constrained MLE.
- Likelihood ration test: LR$= 2(\ln L(\hat{\theta}) - \ln L(\bar{\theta}))$, which asymptotically χ^2_J under the null hypothesis
- The Wald test:
$W = (c(\hat{\theta}) - q)'(\text{Var}(c(\hat{\theta}) - q))^{-1}(c(\hat{\theta}) - q) =$
$(c(\hat{\theta}) - q)'(\frac{\partial c(\hat{\theta})}{\partial \theta'}\text{Est. Var}(\hat{\theta})\frac{\partial' c(\hat{\theta})}{\partial \theta})^{-1}(c(\hat{\theta}) - q)$

Three Asymptotically equivalent test procedures

- The null hypothesis H_0: $c(\theta) = q$. Let $\hat{\theta}$ denote the MLE and $\bar{\theta}$ the constrained MLE.
- Likelihood ration test: LR$= 2(\ln L(\hat{\theta}) - \ln L(\bar{\theta}))$, which asymptotically χ^2_J under the null hypothesis
- The Wald test:
  $W = (c(\hat{\theta}) - q)'(\text{Var}(c(\hat{\theta}) - q))^{-1}(c(\hat{\theta}) - q) = (c(\hat{\theta}) - q)'(\frac{\partial c(\hat{\theta})}{\partial \theta'}\text{Est. Var}(\hat{\theta})\frac{\partial' c(\hat{\theta})}{\partial \theta})^{-1}(c(\hat{\theta}) - q)$
- The Lagrange Multiplier Test–Rao's Score test:
  LM$= \left(\frac{\partial L(\hat{\theta})}{\partial \theta'}\right)[I(\hat{\theta})]^{-1}\left(\frac{\partial' L(\hat{\theta})}{\partial \theta}\right)$.

The linear model under normality

- The model

$$y_i = x_i'\beta + \varepsilon_i$$

The linear model under normality

- The model

$$y_i = x_i'\beta + \varepsilon_i$$

- The density for ε_i——$f(\varepsilon_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\varepsilon_i^2/2\sigma^2\}$.

The linear model under normality

- The model

$$y_i = x_i'\beta + \varepsilon_i$$

- The density for ε_i——$f(\varepsilon_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\varepsilon_i^2/2\sigma^2\}$.
- log Likelihood

$$\begin{aligned} \ln L(\theta) &= -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - [1/2\sigma^2]\sum_{i=1}^{n}(y_i - x_i'\beta)^2 \\ &= -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - [1/2\sigma^2](y - X\beta)'(y - X\beta) \end{aligned}$$

The linear model under normality

- The first order conditions

$$\frac{\partial \ln L}{\partial \beta} = \frac{1}{\sigma^2} X'(y - X\beta) = 0$$

and

$$\frac{\partial \ln L}{\partial \sigma^2} = \frac{-n}{2\sigma^2} + \frac{1}{2\sigma^4}(y - X\beta)'(y - X\beta) = 0$$

The linear model under normality

- The first order conditions

$$\frac{\partial \ln L}{\partial \beta} = \frac{1}{\sigma^2} X'(y - X\beta) = 0$$

and

$$\frac{\partial \ln L}{\partial \sigma^2} = \frac{-n}{2\sigma^2} + \frac{1}{2\sigma^4}(y - X\beta)'(y - X\beta) = 0$$

- Thus

$$\hat{\beta}_{ML} = (X'X)^{-1}X'y = b$$

and

$$\hat{\sigma}^2_{ML} = \frac{e'e}{n}$$

The linear model under normality

- $$\left[\begin{array}{cc} \partial^2 \ln L/\partial\beta\partial\beta' & \partial^2 \ln L/\partial\beta\partial\sigma^2 \\ \partial^2 \ln L/\partial\sigma^2\partial\beta' & \partial^2 \ln L/\partial(\sigma^2)^2 \end{array}\right]\Bigg|_{\hat{\theta}}$$
$$= \left[\begin{array}{cc} -(1/\sigma^2)X'X & -(1/\sigma^4)X'\varepsilon \\ -(1/\sigma^4)\varepsilon'X & n/(2\sigma^4) - \varepsilon'\varepsilon/\sigma^6 \end{array}\right]$$
  hence
$$[I(\beta, \sigma^2)]^{-1} = \left[\begin{array}{cc} \sigma^2(X'X)^{-1} & 0 \\ 0 & 2\sigma^4/n \end{array}\right]$$