# Advanced Econometrics

## Chapter 12: Maximum Likelihood Estimation

In Choi

Sogang University

# References


- Newey, W. and D. McFadden (1994). "Large sample estimation and hypothesis testing," Chapter 36 in R. Engle and D. McFadden (eds.) Handbook of Econometrics, Vol. IV, North-Holland.
- Wooldridge, J. (2002). *Econometric Analysis of Cross Section and Panel Data,* The MIT Press.

# Introduction

- Maximum likelihood estimation provide a unified approach to estimation.
- MLE is generally the most efficient estimation procedure in the class of estimators that use information on the distribution of the data.
- Maximum likelihood estimators are generally inconsistent if some part of the specified distribution is misspecified.

# Introduction

- There are cases in which MLE is robust to failure of certain assumptions. But these must be examined on a case-by-case basis.
- MLE for the linear model under a normality assumption is robust to the assumption.

# Introduction

- In almost all economic applications, we are interested in estimating parameters in conditional distributions.
- We assume that each random draw is partitioned as $(x_i, y_i)$ where $x_i \in R^K$ and $y_i \in R^G$. We are interested in estimating a model for the conditional distribution of $y_i$ given $x_i$.
- Thus, conditional maximum likelihood estimation (CMLE) is the main focus.

## Example

(Probit) Suppose that the (unobserved) latent variable $y_i^*$ follows

$$y_i^* = x_i'\theta + e_i,$$

where $e_i$ is independent of $x_i$ and $e_i \sim iidN(0,1)$. We observe only the sign of $y_i$, i.e.,

$$y_i = \begin{cases} 1 & \text{if } y_i^* > 0 \\ 0 & \text{if } y_i^* \leq 0 \end{cases}$$

Then,

$$\begin{aligned} P(y_i &= 1 \mid x_i) = P(y_i^* > 0 \mid x_i) \\ &= P(x_i'\theta + e_i > 0 \mid x_i) \\ &= P(e_i > -x_i'\theta \mid x_i) \\ &= 1 - \Phi(-x_i'\theta) = \Phi(x_i'\theta). \end{aligned}$$

## Example

and

$$P(y_i = 0 \mid x_i) = 1 - \Phi(x_i'\theta).$$

The density function of $y_i$ given $x_i$ is

$$f(y_i \mid x_i) = \left[\Phi(x_i'\theta)\right]^{y_i} \left[1 - \Phi(x_i'\theta)\right]^{1-y_i},$$

which gives the log-likelihood function

$$L(\theta) = \sum_{i=1}^{N} \left(\log\left[\Phi(x_i'\theta)\right] + (1 - y_i)\log\left[1 - \Phi(x_i'\theta)\right]\right).$$

# General framework for CMLE

- Let $f(y \mid x, \theta_0)$ denote the true conditional density of $y_i$ given $x = x_i$.
- We have

$$K(f; x, \theta) = \int \log \left[ f(y \mid x, \theta_0) / f(y \mid x, \theta) \right] f(y \mid x, \theta_0) dy \geq 0.$$

This is called the conditional Kullback-Leibler information inequality. This is minimized when $\theta = \theta_0$.

# General framework for CMLE

- The Kullback-Leibler information inequality is rewritten as

$$E\left(\log f\left(y_i \mid x_i, \theta_0\right) \mid x_i = x\right) \geq E\left(\log f\left(y_i \mid x_i, \theta\right) \mid x_i = x\right)$$

or

$$E\left[l_i(\theta_0) \mid x_i\right] \geq E\left[l_i(\theta) \mid x_i\right],$$

where $l_i(\theta)$ is the conditional log likelihood for observation i.

- We see that $\theta_0$ solves

$$\max_{\theta \in \Theta} E\left[l_i(\theta)\right].$$

- The sample analogue of the optimization problem is

$$\max_{\theta \in \Theta} N^{-1} \sum_{i=1}^{N} \log f\left(y_i \mid x_i, \theta\right).$$

A solution of this problem is the CMLE of $\theta_0$.

# Consistency of CMLE

**Theorem 1** (Consistency of CMLE): Let $\{(x_i, y_i) : i = 1, 2, ...\}$ be a random sample. Let $\Theta$ be a parameter set. Assume
(i) $f(y \mid x, \theta)$ is the true density.
(ii) $\Theta$ is compact parameter set.
(iii) $l(\theta)$ is a continuous function on $\Theta$.
(iv) $|l(w, \theta)| \leq b(W)$ for all $\theta \in \Theta$ and $E(b(W)) < \infty$.
Then, the CMLE $\hat{\theta}$ exists and $\hat{\theta} \xrightarrow{p} \theta_0$.

# Asymptotic normality of CMLE

- The score of the log likelihood for observation $i$ is defined as

$$s_i(\theta) = \nabla_\theta l_i(\theta).$$

In most cases $Es_i(\theta) = 0$.

Theorem 2 (Asymptotic Normality of CMLE): Let the conditions of Theorem 1 hold, and let $H_i(\theta) = \nabla^2_\theta l_i(\theta)$. In addition, assume that
(i) $\theta_0 \in int(\Theta)$.
(ii) $l(y, x, \cdot)$ is twice continuously differentiable on $int(\Theta)$.
(iii) $Es_i(\theta) = 0$ for all $\theta \in int(\Theta)$.
(iv) the elements of $\nabla^2_\theta l_i(\theta)$ are bounded in absolute value by a function $b(y, x)$ which is integrable.
(v) $A_0 = -E\left[H_i(\theta)_{|\theta=\theta^0}\right]$ is positive definite.
Then,

$$\sqrt{N}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, A_0^{-1}).$$

# Estimating the asymptotic variance

- The asymptotic variance can be estimated by

$$\left[ N^{-1} \sum_{i=1}^{N} A(x_i, \hat{\theta}) \right]^{-1},$$

  where $A(x_i, \theta) = -E\left[ \nabla_\theta^2 l_i(\theta) \mid x_i \right]$.
- The estimator is positive definite when it exists, and tends to have better finite sample properties than the outer product of the score estimator.