USC, Fall 2016, Economics 513

Lecture 14: Discrete Choice I: Maximum Likelihood

Likelihood function

Let $y_1, \ldots, y_n$ be a random sample from $N(\mu, \sigma^2)$. The pdf of y_i is

$$f(y_i; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{1}{2\sigma^2}(y_i - \mu)^2 \right\}$$

The joint pdf of $y_1, \ldots, y_n$ is

$$f(y_1, \ldots, y_n; \mu, \sigma^2) = \prod_{i=1}^{n} f(y_i; \mu, \sigma^2) = \frac{1}{\sigma^n (2\pi)^{\frac{n}{2}}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 \right\}$$

If we have the observations $y_1, \ldots, y_n$, this is a function of the parameters μ, σ^2. In sequel the vector of parameters is denoted by θ with in example

$$\theta = \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}$$

The joint pdf of the random variables of which the observed data are a realization, considered as a function of the parameters θ is called a *likelihood function.*

In general, we do not know the population value of the parameters which we denote by θ_0. The likelihood function is used to find an estimator/estimate of θ_0: we estimate the population parameter(s) by the value if θ that maximizes the likelihood function.

We denote the estimator/estimate by $\hat{\theta}$, so that

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

with L the likelihood function and Θ the parameter space.

In the example

$$L(\theta) = \frac{1}{\sigma^n (2\pi)^{\frac{n}{2}}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 \right\}$$

and

$$\Theta = \left\{ (\mu, \sigma^2) | -\infty < \mu < \infty, \sigma^2 > 0 \right\}$$

The estimate/estimator $\hat{\theta}$ is called the *maximum likelihood estimate/estimator (MLE).*

Equivalent definition

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \ln L(\theta)$$

with $\ln L(\theta)$ the *log likelihood function* and

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

Log likelihood preferred because

- First order conditions are easier.
- Log likelihood is a sum of (functions of) random variables.

In example

$$\ln L(\theta) = -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2$$

The MLE can be found by solving the first order conditions

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

$$\hat{\sigma^2} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \overline{y})^2$$

Example: CLR model with normal errors.

Because

$$y_i = x_i'\beta + \varepsilon_i \qquad , \varepsilon_i \sim N(0, \sigma^2)$$

we have that the conditional pdf of y_i give x_i, the i-th row of X is

$$f(y_i|x_i;\beta,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{1}{2\sigma^2}(y_i - x_i'\beta)^2\right\}$$

The data are $y_1, x_1, \ldots, y_n, x_n$ and to obtain the joint pdf $f(y_i, x_i; \mu, \sigma^2)$ we need to make an assumption on the marginal pdf of x_i. Possible assumptions

- x_i is a vector of non-stochastic constants
- x_i has pdf $g(x_i)$ and the $x_i, i = 1, \ldots$ are independent and identically distributed and g does not depend on θ.

We use the second option, so that

$$f(y_i, x_i; \mu, \sigma^2) = f(y_i|x_i;\mu,\sigma^2)g(x_i)$$

and

$$L(\theta) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{1}{2\sigma^2}(y_i - x_i'\beta)^2\right\} g(x_i)$$

and

$$\ln L(\theta) = \ln\left(\prod_{i=1}^{n} g(x_i)\right) - \frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i'\beta)^2$$

The first two terms on the rhs are indenpent of the parameters.

The MLE are

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

$$\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n} e_i^2 \quad , e_i = y_i - x_i'\hat{\beta}$$

Hence the MLE of β is the OLS estimator and the MLE of σ^2 is (K is the number of independent variables including the constant)

$$\hat{\sigma^2} = \frac{n-K}{n}\hat{\sigma}^2_{OLS}$$

so that in the CLR model

$$E(\hat{\sigma^2}) = \frac{n-K}{n}\sigma^2$$

The MLE of σ^2 is biased (downwards) but the bias goes to 0 if $n \to \infty$.

Note MLE does not depend on assumed distribution of x_i if that distribution does not depend on θ. The independence or identical distribution of $x_1, \ldots, x_n$ is not essential.

Example: Binary logit

Consider regression model with a 0-1 dependent variable, e.g. y_i is 1 if i owns a house and 0 if i rents.

Because y_i is 0-1 and an indicator of an event, it makes sense to consider

$$\Pr(y_i = 1|x_i) = p(x_i)$$

with x_i a vector of explanatory variables.

We choose the *logit* model for the conditional probability $p(x_i)$

$$p(x_i) = \frac{e^{x_i'\beta}}{1 + e^{x_i'\beta}}$$

Now we can write the conditional density of y_i given x_i as

$$f(y_i|x_i;\beta) = \left(\frac{e^{x_i'\beta}}{1 + e^{x_i'\beta}}\right)^{y_i} \left(\frac{1}{1 + e^{x_i'\beta}}\right)^{1-y_i}$$

and with same assumption on the distribution of x_i as in the CLR model we have

$$\ln L(\theta) = \sum_{i=1}^{n} \left\{ y_i \ln\left(\frac{e^{x_i'\beta}}{1 + e^{x_i'\beta}}\right) + (1 - y_i) \ln\left(\frac{1}{1 + e^{x_i'\beta}}\right) \right\} + \ln\left(\prod_{i=1}^{n} g(x_i)\right)$$

The first order conditions are

$$\sum_{i=1}^{n} \left(y_i - \frac{e^{x_i'\beta}}{1 + e^{x_i'\beta}}\right) x_i = 0$$

This is a nonlinear system of equations that does not have a closed-form solution, so that we have to find the MLE by numerical methods. The sampling distribution cannot be obtained as in the CLR model.
We now consider the sampling distribution of the MLE $\hat{\theta}$.

Questions

- What is the sampling distribution of $\hat{\theta}$ if $n \to \infty$?
- Can we use the asymptotic sampling distribution to find confidence intervals for θ_0?
- How does the asymptotic sampling distribution of the MLE compare to that of other estimators?

The theory applies to *all* MLE.

Maximum Likelihood theory

The MLE is in general consistent

$$\hat{\theta} \xrightarrow{p} \theta_0$$

The MLE is in general also asymptotically normal

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, I(\theta_0))$$

with

$$I(\theta_0) = -\mathrm{E}\left(\frac{\partial^2 \ln f(y|x;\theta)}{\partial\theta\partial\theta'}\right)\Bigg|_{\theta=\theta_0}$$

Example: The information matrix for the binary logit model is

$$I(\theta) = -\mathrm{E}\left[\frac{\partial^2 \ln f(y|x;\theta)}{\partial\theta\partial\theta'}\right] = \mathrm{E}\left[\frac{e^{\theta'x}}{(1+e^{\theta'x})^2}xx'\right]$$

The matrix $I(\theta)$ is called the *information matrix*. It can be estimated by

$$\hat{I}(\theta_0) = -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \ln f(y_i|x_i;\hat{\theta})}{\partial\theta\partial\theta'}$$

In the binary logit case

$$\hat{I}(\theta_0) = \frac{1}{n}\sum_{i=1}^{n}\frac{e^{\hat{\theta}'x_i}}{\left(1+e^{\hat{\theta}'x_i}\right)^2}x_i x_i'$$

In general the information matrix and its inverse will be computed iteratively when computing the MLE by numerical minimization (see below), so that there is no need to write a separate program for the estimation of the variance matrix of the MLE.
The diagonal elements of $\hat{I}(\theta_0)$ are estimates of the asymptotic variances of the MLE $\hat{\theta}$ and their square roots are the asymptotic standard deviations or standard errors $s(\hat{\theta}_k)$ so that a 95% confidence interval for θ_k is

$$[\hat{\theta}_k - 1.96s(\hat{\theta}_k), \hat{\theta}_k + 1.96s(\hat{\theta}_k)]$$

ML estimators have the smallest (asymptotic) variance of all consistent estimators for θ_0. We say that the MLE are *efficient*.
Of the asymptotic properties of ML we only look at consistency in some detail. The log likelihood is

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

If

$$\mathrm{E}[|\ln f(y,x,\theta)|] < \infty$$

then by the LLN

$$\frac{1}{n}\ln L(\theta) \xrightarrow{p} \mathrm{E}(\ln f(y,x,\theta))$$

We have

$$\mathrm{E}(\ln f(y,x,\theta)) = \int\int \ln f(y,x,\theta) f(y,x,\theta_0)\mathrm{d}y\mathrm{d}x \equiv h(\theta)$$

with θ_0 the population value of θ. The right hand side is a function of θ and we can ask at which θ this function is maximal. The first order condition is, if we can interchange differentiation and integration

$$\frac{\partial h}{\partial \theta}(\theta) = \int\int \frac{\partial f}{\partial \theta}(y,x,\theta)\frac{f(y,x,\theta_0)}{f(y,x,\theta)}\mathrm{d}y\mathrm{d}x = 0$$

If $\theta = \theta_0$ then

$$\frac{\partial h}{\partial \theta}(\theta_0) = \int\int \frac{\partial f}{\partial \theta}(y,x,\theta_0)\mathrm{d}y\mathrm{d}x = 0$$

because a density integrates to 1 for all θ so that (again we assume that we can interchange differentiation and integration) for all θ

$$\int\int \frac{\partial f}{\partial \theta}(y,x,\theta)\mathrm{d}y\mathrm{d}x = 0$$

If in addition we can show that the second order derivative at θ_0 is positive, $h(\theta)$ has a local maximum at θ_0. We need actually that the maximum is unique on Θ which holds if there is no $\theta \in \Theta$ such that

$$f(y,x,\theta) = f(y,x,\theta_0)$$

except for a set of values of y, x that has probability 0. If such a θ exists it is called observationally equivalent and θ is not identified.

If all this holds then

$$\hat{\theta} = \mathrm{argmax}_{\theta\in\Theta}\frac{1}{n}\sum_{i=1}^{n}\ln f(y_i,x_i,\theta) \xrightarrow{p} \theta_0 = \mathrm{argmax}_{\theta\in\Theta}\mathrm{E}(\ln f(y,x,\theta))$$

i.e. the MLE is consistent.

Computation of ML estimators

Remember the definitions of the ML estimators

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \ln L(\theta) = \arg\min_{\theta \in B} -\ln L(\theta)$$

These is usually a non-linear minimization problem (under constraints). This raises at least two questions:

(i) Does the minimum exist? Is it unique?

(ii) How do we compute this minimum?

In the sequel we use $Q(\theta)$ to denote the function that has to be minimized.

Does minimum exist? Is it unique?

We start with two well-known facts:

> A continuous function has a minimum on compact (closed and bounded) set, in our case the parameter space Θ. This establishes existence.
>
> A convex function has a *unique* minimum on a convex set, in our case the parameter space Θ. This establishes uniqueness.

Most log likelihoods are continuous on Θ. Sometimes you can show that $Q(\theta)$ is convex, but this is rare.

Example: binary logit model. The second derivative of the minus log likelihood is

$$\frac{\partial^2 Q}{\partial \theta \partial \theta'}(\theta) = \sum_{i=1}^{n} \frac{e^{\theta' x_i}}{(1 + e^{\theta' x_i})^2} x_i x_i'$$

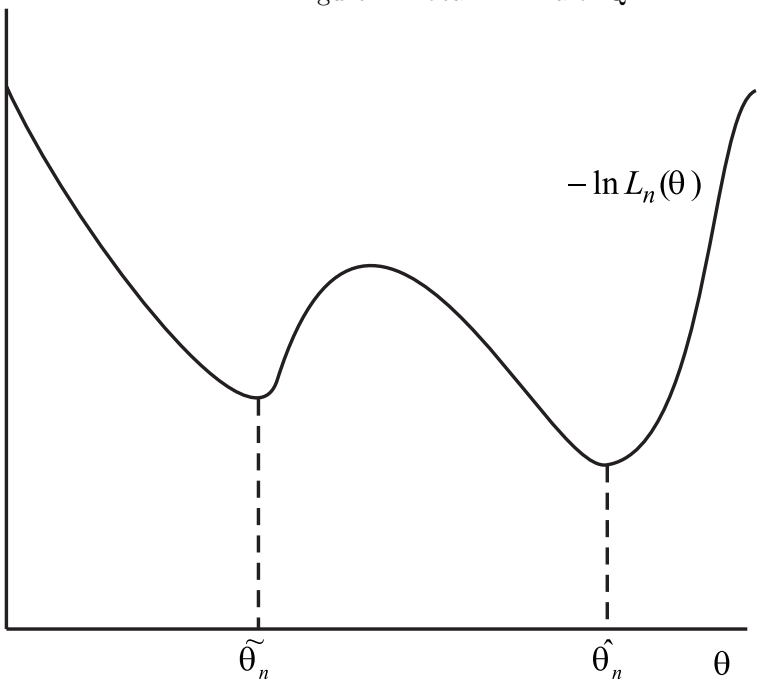
This is a positive definite matrix (because it is a sum of positive definite matrices). Hence the log likelihood is convex in θ.

Most minimization algorithms solve first-order conditions for a minimum

$$q\left(\hat{\theta}\right) = \frac{\partial Q}{\partial \theta}\left(\hat{\theta}\right) = 0$$

If Q is not convex, there is no guarantee that a solution to this equation is the ML estimator.

Figure 1: Local minima of Q



Note that in this figure $\tilde{\theta}$ satisfies the first-order condition for a minimum, but it is not the minimizer on Θ.

Most algorithms for minimization are based on a quadratic approximation of $Q(\theta)$. We start out with

$$0 = q\left(\hat{\theta}\right) \approx q(\theta) + \frac{\partial^2 Q}{\partial\theta\partial\theta'}(\theta)\left(\hat{\theta} - \theta\right)$$

We solve this to obtain

$$\hat{\theta} - \theta = -\left[\frac{\partial^2 Q}{\partial\theta\partial\theta'}(\theta)\right]^{-1} q(\theta)$$

If the approximation is exact, then we can find $\hat{\theta}$ from some starting value θ by

$$\hat{\theta} = \theta - \left[\frac{\partial^2 Q}{\partial\theta\partial\theta'}(\theta)\right]^{-1} q(\theta)$$

If the approximation is not exact, then this suggests an iterative algorithm:

(i) Start at a starting value θ_1

(ii) Iterate

$$\theta_{j+1} = \theta_j - \left[\frac{\partial^2 Q}{\partial\theta\partial\theta'}(\theta_j)\right]^{-1} q(\theta_j)$$

(iii) Stop if $q(\theta_j)$ is small, e.g. $q(\theta_j)'q(\theta_j) < \varepsilon$.

This is the Newton-Raphson algorithm. It requires computation of first and second derivatives of Q, either analytically or numerically by

$$q(\theta) \approx \frac{Q(\theta+\varepsilon) - Q(\theta)}{\varepsilon}$$

with ε a small number(here q is scalar).

Note that if the minimand is a minus log likelihood, then $Q\left(\hat{\theta}\right)^{-1}$ is a consistent estimator of the asymptotic variance matrix of $\hat{\theta}$.