

M-Estimators

Li Gan

Basic Asymptotic Theory

Convergence of deterministic sequences:

Definitions:

- (1) A sequence of nonrandom numbers $\{a_N: N=1,2,\dots\}$ converges to a if $a_N \rightarrow a$ if $\forall \varepsilon > 0, \exists N_\varepsilon \forall n$ such that $|a_N - a| < \varepsilon \quad \forall n > N_\varepsilon$. We write $a_N \rightarrow a$ as $N \rightarrow \infty$.
- (2) A sequence $\{a_N: N = 1,2,\dots\}$ is bounded iff $\exists b < \infty$ such that $|a_N| < b, \forall N=1,2,\dots$ otherwise, we say a_N is unbounded.

- Examples:
- (1) if $a_N = 2 + 1/N$ then $a_N \rightarrow 2$.
 - (2) if $a_N = (-1)^N$ then a_N doesn't have a limit but bounded.
 - (3) $a_N = N^{1/4}$, then a_N is not bounded.

Definitions:

- (1) A sequence $\{a_N\}$ is $O(N^\lambda)$ (at most of order N^λ) if $N^{-\lambda}a_N$ is bounded. We also write it as $a_N = O(N^\lambda)$. When $\lambda = 0$, $\{a_N\}$ is bounded, we also write it as $a_N = O(1)$ (big oh one).
- (2) $\{a_N\}$ is $o(N^\lambda)$ if $N^{-\lambda}a_N \rightarrow 0$. We also write it as $a_N = o(N^\lambda)$.

When $\lambda = 0$, a_N converge to zero, or $a_N = o(1)$ (little oh one).

- Example:
- (1) if $a_N = \log(N)$, then $a_N = o(N^\lambda)$ for any $\lambda > 0$.
 - (2) if $a_N = 10 + N^{1/2}$, then $a_N = O(N^{1/2})$ & $a_N = o(N^{1/2+\gamma})$ for $\gamma > 0$.

Convergence in probability

Definitions:

- (1) A sequence of random variables $\{x_N: N = 1, 2, \dots\}$ converges in probability to the constant a if for any $\varepsilon > 0 \quad \Pr(|x_N - a| \geq \varepsilon) < \varepsilon \rightarrow 0$ as $N \rightarrow \infty$. We write it as $x_N \xrightarrow{p} a$ or $\text{plim } x_N = a$.

- (2) For $\{x_N\}$, when $a = 0$ then $x_N \xrightarrow{p} 0$ or $x_N = o_p(1)$ (little oh p one)
 (3) $\{x_N\}$ is bounded in probability if for any $\varepsilon > 0$, there exists a $b_\varepsilon < \infty$ and an integer N_ε such that $\Pr(|x_N| \geq b_\varepsilon) < \varepsilon$ for all $N \geq N_\varepsilon$ or $x_N = O_p(1)$ (x_N is big oh p one)

Questions: (1) Is $x_N \rightarrow N(0, \sigma^2)$ bounded? $\Pr(|x_N| \geq b_\varepsilon) < \varepsilon$ is always true.
 (2) Are all random variables bounded in probability? Cauchy? Yes. A random sequence, however, is not bounded.

Discussions: If c_N is a nonrandom sequence:
 then $c_N = O_p(1)$ iff $c_N = O(1)$
 $c_N = o_p(1)$ iff $c_N = o(1)$

Lemma: if $x_N \xrightarrow{p} a$ then $x_N = O_p(1)$.

Definitions:

- (1) A random sequence $\{x_N: N = 1, 2, \dots\}$ is $o_p(a_N)$, where a_N is a nonrandom positive sequence and $x_N/a_N = o_p(1)$.
 (2) $\{x_N\}$ is $O_p(a_N)$ if $x_N/a_N = O_p(1)$.

Examples:

- (a) Define $\bar{x}_N = \frac{1}{N} \sum_i x_i$. We have $\bar{x}_N \xrightarrow{p} E(x_i)$. Then $\bar{x}_N = o_p(1)$.
 (b) Central limit theorem: $\sqrt{N}(\bar{x}_N - E(x_i)) \rightarrow N(0, \sigma_x^2)$, then: $\bar{x}_N = O_p(n^{-1/2})$
 (c) If z is a random variable, $x_N = \sqrt{N}z$, then $x_N = O_p(n^{1/2})$, and
 $x_N = o_p(n^{1/2+\varepsilon})$ for any $\varepsilon > 0$.

Lemma:

If $w_N = o_p(1)$ or $w_N \rightarrow 0$
 $x_N = o_p(1)$ or $x_N \rightarrow 0$
 $y_N = O_p(1)$ y_N is bounded in probability
 $z_N = O_p(1)$ z_N is bounded in probability.

Then: (1) $w_N + x_N = o_p(1)$, or $o_p(1) + o_p(1) = o_p(1)$
 (2) $y_N + z_N = O_p(1)$, or $O_p(1) + O_p(1) = O_p(1)$
 (3) $y_N z_N = O_p(1)$, or $O_p(1) O_p(1) = O_p(1)$

$$(4) x_N z_N = o_p(1), \text{ or } o_p(1) O_p(1) = o_p(1)$$

$$(5) w_N x_N = o_p(1), \text{ or } o_p(1) o_p(1) = o_p(1)$$

$$(6) x_N + z_N = O_p(1), \text{ or } o_p(1) + O_p(1) = o_p(1)$$

Lemma: $g: R^K \rightarrow R^J$ continuous at some point $c \in R^K$. Suppose a random sequence $\{x_N\}$, $x_N \xrightarrow{p} c$. Then $g(x_N) \xrightarrow{p} g(c)$, or $\text{plim } g(x_N) = g(\text{plim } x_N)$

This is the so-called Slutsky theorem, which is very useful in practice. It shows that the plim passes through nonlinear functions, provided they are continuous. The expectation operator does not have this feature.

Definition: Let (Ω, F, P) be a probability space. A sequence of events $\{\Omega_N: N = 1, 2, \dots\}$ F is said to occur with probability approaching one (w.p.a.1), if and only if $\Pr(\Omega_N) \rightarrow 1$ as $N \rightarrow \infty$

Let $\{z_N: N = 1, 2, \dots\}$ be a sequence of random $K \times K$ matrix, and let A be a nonrandom, invertible $K \times K$ matrix. If $z_N \xrightarrow{p} A$, then:

$$(1) z_N^{-1} \text{ exists with probability approaches to } 1;$$

$$(2) z_N^{-1} \xrightarrow{p} A^{-1} \text{ or } p \lim z_N^{-1} = A^{-1}.$$

Convergence in distribution

Definition: a sequence of random variables $\{x_N\}$ convergence in distribution to the continuous random variable x iff $F_N(x) \rightarrow F(x)$ as $N \rightarrow \infty$ for all $x \in R$, where $F_N(x)$ is the cdf of x_N , and $F(x)$ is the cdf of x .

Example: $\sqrt{N} \bar{x}_N \xrightarrow{d} N(\mu, \sigma^2)$. Note $F_N(x)$ does not have continuous for any N . A good example is where x_N is discrete for all but has an asymptotically normal distribution. A good example is Bernoulli.

Definition: A sequence of $K \times 1$ random vectors $\{x_N: N = 1, 2, \dots\}$ converges in distribution to the continuous vector x iff for $K \times 1$ nonrandom vector c , such that $c'c = 1$, and $c'x_N \xrightarrow{d} c'x$. We write $x_N \xrightarrow{d} x$.

When $x_N \xrightarrow{d} N(m, V)$ then $c'x_N \xrightarrow{d} N(c'm, c'Vc)$ for any $c \in R^K$ and $c'c = 1$.

Lemma: if $x_N \xrightarrow{d} x$ where x is any $K \times 1$ random vector, then $x_N = O_p(1)$.

This lemma is useful to establish if a sequence is bounded – often by simply figuring out if the sequence has a limiting distribution or not.

Lemma (continuous mapping theorem): let $\{x_N: N = 1, 2, \dots\}$ and $x_N \xrightarrow{d} x$. If $g: R^K \rightarrow R^J$ is a continuous function then $g(x_N) \xrightarrow{d} g(x)$.

The importance of the Continuous mapping theorem cannot be overstated. It says if we know the limiting distribution of x_N , we then know the limiting distribution of any function of x_N .

Corollary $\{z_n\}: z_N \xrightarrow{d} N(0, V)$ then

(1) For any $K \times M$ nonrandom matrix A , $A' z_N \xrightarrow{d} N(0, A'VA)$

(2) $z_N' V^{-1} z_N \xrightarrow{d} \chi_K^2$.

Lemma (asymptotic equivalence lemma):

Let $\{x_n\}$ and $\{z_n\}$. If $z_N \xrightarrow{d} z$, and $x_N - z_N \xrightarrow{p} 0$. Then $x_N \xrightarrow{d} z_N$.

Limit Theorem for Random Samples

Here we state two classical limit theorems for iid sequence.

Theorem 1: $\{w_i, i=1, 2, \dots\}$ are iid $G \times 1$ random vectors with $E(|w_{ig}|) < \infty$, then the sequence satisfies the WLLN (Weak Law of Large Numbers):

$$\frac{1}{N} \sum_{i=1}^N w_i \xrightarrow{p} \mu_w \equiv E(w_i)$$

Theorem 2 (Lindeberg-Levy CLT)

If $\{w_i, i=1, 2, \dots\}$ with $E(w_{ig}^2) < \infty$, and $E(w_{ig}) = 0$. Then

$$N^{-1/2} \sum_{i=1}^N w_i \xrightarrow{d} Normal(0, B) \text{ where } B = \text{Var}(w_i) = E(w_i w_i')$$

B is necessarily positive semidefinite.

Limiting Behavior of Estimators and Test Statistics

Asymptotic properties of estimators

Definition: $\{\hat{\theta}_N, N=1, 2, \dots\}$ be a sequence of estimators of $P \times 1$ vector $\theta \in \Theta$

If $\hat{\theta}_N \rightarrow \theta$ for any value of θ then we say $\hat{\theta}_N$ is a consistent estimator of θ .

Why for any value of θ ? Because we don't know θ .

Definition: if $\{\hat{\theta}_N, N=1, 2, \dots\}$ be a sequence, and if $\sqrt{N}(\hat{\theta}_N - \theta) \xrightarrow{d} N(0, V)$,

$\hat{\theta}_N$ is \sqrt{N} -normally distributed, and V is the asymptotic variance of $\sqrt{N}(\hat{\theta}_N - \theta)$.

Definition: Two estimators $\hat{\theta}_N$ and $\tilde{\theta}_N$ with $\sqrt{n}(\hat{\theta}_N - \theta) \xrightarrow{d} N(0, V)$ and $\sqrt{n}(\tilde{\theta}_N - \theta) \xrightarrow{d} N(0, D)$.

- (1) $\hat{\theta}_N$ is asymptotically efficient relative to $\tilde{\theta}_N$ if $D-V$ is positive or semi definite for all θ
- (2) $\hat{\theta}_N$ and $\tilde{\theta}_N$ are \sqrt{n} -equivalent if $\sqrt{n}(\tilde{\theta}_N - \hat{\theta}_N) = o_p(1)$.

Definition: $\hat{\theta}_{N1}$ and $\hat{\theta}_{N2}$ are asymptotically independent if

$$\text{Cov}(\hat{\theta}_{N1}, \hat{\theta}_{N2}) = \begin{pmatrix} V_1 & 0 \\ 0 & V_2 \end{pmatrix}.$$

Asymptotic properties of test statistics

Definition (1) asymptotic size: $\lim_{N \rightarrow \infty} P_N(\text{reject } H_0 | H_0)$. (2) A test is consistent against alternative H_1 if the null hypothesis is rejected with probability approaching one, or: $\lim_{N \rightarrow \infty} P_N(\text{reject } H_0 | H_1) = 1$.

Lemma: If $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, V)$, then:

- (1) $\sqrt{n}R(\hat{\theta} - \theta) \xrightarrow{d} N(0, RVR')$
- (2) $\sqrt{n}(\hat{\theta} - \theta)R'(RVR')^{-1}\sqrt{n}R(\hat{\theta} - \theta) \xrightarrow{d} \chi^2_{\dim(\theta)}$

For testing $H_0: R\theta=r$ where r is a $Q \times 1$ non-random vector. Wald statistic

Lemma: If $\sqrt{n}(\hat{\theta}_N - \theta) \xrightarrow{d} N(0, V)$. Let $c: \Theta \rightarrow R^Q$ be a continuously differentiable function on the parameter space $\Theta \subset R^P$ where $Q \leq P$. Assume that θ is in the interior of the parameter space. Define $C(\theta) \equiv \nabla_{\theta} c(\theta) = \frac{\partial c(\theta)}{\partial \theta}$. Then we have:

$$\sqrt{N}(c(\hat{\theta}_N) - c(\theta)) \xrightarrow{d} N(0, C(\theta)VC(\theta)')$$

How to get this? Delta method:

$$c(\hat{\theta}_N) = c(\theta) + C(\theta_N^*)(\hat{\theta}_N - \theta),$$

where θ_N^* is between $\hat{\theta}_N$ and θ . Therefore, $\hat{\theta}_N \xrightarrow{p} \theta$ would result in $\theta_N^* \xrightarrow{p} \theta$.

Therefore,

$$\begin{aligned}\sqrt{N}(c(\hat{\theta}_N) - c(\theta)) &= \sqrt{N}C(\theta_N^*)(\hat{\theta}_N - \theta) \\ &= C(\theta)\sqrt{N}(\hat{\theta}_N - \theta) + (C(\theta_N^*) - C(\theta))\sqrt{N}(\hat{\theta}_N - \theta) \\ &= C(\theta)\sqrt{N}(\hat{\theta}_N - \theta) + o_p(1)O_p(1) \\ &= C(\theta)\sqrt{N}(\hat{\theta}_N - \theta) + o_p(1) \\ &\xrightarrow{d} N(0, C(\theta)VC(\theta))\end{aligned}$$

M-estimator / nonlinear estimator

“M” indicates either the minimum or the maximum. Examples of M-estimators include MLE, least absolute deviation, GMM, non-linear least square, etc.

Nonlinear: we have a random variable y we would like to model $E(y|x)$:

If $E(y|x)=x\theta$, then linear models.

If $E(y|x) = m(x, \theta)$, a nonlinear function of θ , then nonlinear models.

Examples:

(1) If $y > 0$, and $m(x, \theta) = e^{x\theta}$, exponential

(2) If $0 < y < 1$, and $m(x, \theta) = \frac{e^{x\theta}}{1 + e^{x\theta}}$, logistic

How do we know if $m(x, \theta) = e^{x\theta}$? Maybe it is the case that: $m(x, \theta) = \frac{e^{x\theta}}{1 + e^{x\theta}}$?

Identification condition: A correctly specified model for the conditional mean, $E(y|x)$ if for some $\theta_0 \in \Theta$ such that:

$$E(y|x) = m(x, \theta_0)$$

Example: consider a model: $m(x, \theta) = \theta_1 x^{\theta_2}$.

$$E(y|x)=4x^{1.5}, \text{ then } \theta_{01}=4, \theta_{02}=1.5.$$

The model $m(x, \theta) = \theta_1 x^{\theta_2}$ would be estimated.

In linear models, we add an error term: if $E(y|x) = x\beta \rightarrow y = x\beta + u$ in which u can be identical (homoscedastic), and we assume $E(u|x)=0$

Question: In nonlinear models, are u and x still uncorrelated?

Example: if $y \geq 0$, then $m(x, \theta) + u \geq 0 \rightarrow u \geq -m(x, \theta)$.

In which case, u and x cannot be independent. Further, u cannot be homoskedasticity since $\text{var}(u|x) \neq \text{var}(u)$.

Assumption *NLSI*: For some $\theta_0 \in \Theta$, $E(y|x) = m(x, \theta_0)$, where θ_0 is the true parameter.

How to find θ_0 ? To find an estimate $\hat{\theta}$: such that $\hat{\theta} \xrightarrow{p} \theta_0$.

Proposition: Given that $E(y|x) = m(x, \theta_0)$. Show that the solution to the following minimization problem is θ_0 :

$$\min E[(y - m(x, \theta))^2]$$

Proof:

$$\begin{aligned} (y - m(x, \theta))^2 &= ((y - m(x, \theta_0)) + (m(x, \theta_0) - m(x, \theta)))^2 \\ &= (y - m(x, \theta_0))^2 + 2(y - m(x, \theta_0))(m(x, \theta_0) - m(x, \theta)) + (m(x, \theta_0) - m(x, \theta))^2 \end{aligned}$$

Given $E(y|x) = m(x, \theta_0)$, and take the conditional expectation of y :

$$\begin{aligned} E[(y - m(x, \theta))^2] &= E[(y - m(x, \theta_0))^2] + (m(x, \theta_0) - m(x, \theta))^2 \\ &\geq E[(y - m(x, \theta_0))^2] \end{aligned}$$

The inequality is strict when $\theta \neq \theta_0$.

Therefore: $\theta_0 = \arg \min E[(y - m(x, \theta))^2]$.

Denote: $\hat{\theta}_{NLS} = \arg \min \frac{1}{N} \sum_{i=1}^N (y_i - m(x_i, \theta))^2$

If $\frac{1}{N} \sum_{i=1}^N (y_i - m(x_i, \theta))^2 \xrightarrow{\text{uniform}} E(y_i - m(x_i, \theta))^2$ then $\hat{\theta}_{NLS} \xrightarrow{p} \theta_0$

Identification: uniform convergence and consistency:

Assumption *NLS2* (Identification): $E[(y - m(x, \theta))^2] > 0 \quad \forall \theta \in \Theta$ and $\theta \neq \theta_0$.

Example: $m(x, \theta) = \theta_1 + \theta_2 x_2 + \theta_3 x_3^{\theta_4}$

If true $\theta_3 = 0$ then $E[(y - m(x, \theta))^2]$ is minimized for any θ with $\theta_1 = \theta_{01}$, $\theta_2 = \theta_{02}$, $\theta_3 = 0$ and θ_4 at any values.

Assumption *NLS2* doesn't hold. The model is not identified.

If $\theta_3 \neq 0$, then *NLS2* holds. The model is identified and can be estimated.

Example: point wise convergence but not uniform convergence.

$$f_n(x) = x^n, \quad 0 \leq x \leq 1$$

$$\text{Let } f(x) = \begin{cases} 0 & \text{if } 0 \leq x < 1 \\ 1 & \text{if } x = 1 \end{cases}$$

In this case, $f_n(x) \rightarrow f(x)$ pointwise.

However, $f_n(x)$ does not converge to $f(x)$ uniformly.

$|f_n(x) - f(x)| = x^n$ cannot be bounded although $x^n \rightarrow 0$ but cannot be bounded

independent of x .

Because to have $x^n < \varepsilon$, we require $n \log x < \log \varepsilon$. Therefore, $n > \log \varepsilon / \log x$, which is dependent on x . Uniform convergence (the maximum converges):

$$\max \left| \frac{1}{N} \sum q(w_i, \theta) - E[q(w_i, \theta)] \right| \xrightarrow{p} 0$$

Note: uniform convergence clearly implies point wise convergence.

Point-wise convergence: for each $\theta \in \Theta$, $\frac{1}{N} \sum_i q(w_i, \theta) \xrightarrow{p} E(q(w, \theta))$.

Uniform weak law of large numbers

Theorem: w be random variable taking values $W \subset R^M$, Θ be a subset of R^P

$q: W \times \Theta \rightarrow R$ be real value function. Assume that: (a) Θ is compact (b) For each $\theta \in \Theta$, $q(\cdot, \theta)$ is bound measurable on W . (c) For each $w \in W$, $q(w, \cdot)$ is continuous on Θ . (d) $|q(w, \theta)| \leq b(w)$ for all $\theta \in \Theta$.

Theorem: If

- (i) $\frac{1}{N} \sum q(w_i, \theta)$ uniformly converges to $E[q(w_i, \theta)]$;
- (ii) $\theta_0 = \arg \min E(q(w_i, \theta))$, and
- (iii) $\hat{\theta}_N = \arg \min \frac{1}{N} \sum_{i=1}^N q(w_i, \theta)$.

Then we must have: $\hat{\theta}_N \xrightarrow{p} \theta_0$.

This theorem can be illustrated by the following graph:

$$\begin{array}{ccc} \frac{1}{N} \sum q(w_i, \theta) & \xrightarrow[\text{(Condition (i))}]{\text{uniformly}} & E[q(w_i, \theta)] \\ \min \uparrow \text{numerically} & & \min \uparrow \\ \hat{\theta}_N & \xrightarrow{p} & \theta_0 \end{array}$$

Theorem applies to median regression too.

LAD (least absolute deviation) estimator of θ_0 . If

(i) $\frac{1}{N} \sum |y_i - m(x_i, \theta)|$ uniformly converges to $E |y_i - m(x_i, \theta)|$;

(ii) $\theta_0 = \arg \min E |y_i - m(x_i, \theta)|$, and

(iii) $\hat{\theta}_{LAD} = \arg \min \frac{1}{N} \sum_{i=1}^N |y_i - m(x_i, \theta)|$.

Then we must have: $\hat{\theta}_{LAD} \xrightarrow{p} \theta_0$.

Asymptotic normality (working with 1st derivatives):

Define: $S(w_i, \theta) = \frac{\partial q(w_i, \theta)}{\partial \theta}$. The first order condition to the problem ensures that for any

set of solution $\hat{\theta}_N$, we must have: $\sum_{i=1}^N S(w_i, \hat{\theta}_N) = 0$. We will mostly work with this equation. By Taylor expansion:

$$\sum_{i=1}^N S(w_i, \hat{\theta}_N) = \sum_{i=1}^N S(w_i, \theta_0) + \sum_{i=1}^N \frac{\partial S(w_i, \theta_N^*)}{\partial \theta} (\hat{\theta}_N - \theta_0) = 0$$

where θ_N^* is between $\hat{\theta}_N$ and θ_0 .

Define $H(\theta) = \frac{\partial S(\theta)}{\partial \theta'} = \frac{\partial^2 q(\theta)}{\partial \theta \partial \theta'}$. Multiply both sides by $1/\sqrt{N}$, we have:

$$\frac{1}{\sqrt{N}} \sum_{i=1}^N S(w_i, \theta_0) = -\sqrt{N} (\hat{\theta}_N - \theta_0) \frac{1}{N} \sum_{i=1}^N \frac{\partial S(w_i, \theta_N^*)}{\partial \theta}$$

Applying CLT, we have:

$$\frac{1}{\sqrt{N}} \sum_{i=1}^N S(w_i, \theta_0) \xrightarrow{d} N(0, E[S(w_i, \theta_0)S(w_i, \theta_0)'])$$

Further let $\frac{1}{N} \sum_{i=1}^N \frac{\partial S(w_i, \theta_N^*)}{\partial \theta} \xrightarrow{p} A(\theta_0)$, and $E[S(w_i, \theta_0)S(w_i, \theta_0)'] \equiv B_0(\theta_0)$

Therefore, we have the limiting distribution of the estimate

$$\sqrt{N} (\hat{\theta}_N - \theta_0) \xrightarrow{d} N(0, A^{-1}(\theta_0)B_0(\theta_0)A^{-1}(\theta_0))$$

Examples: For NLS,

$$q(w_i, \theta) = \frac{1}{2}(y_i - m(x_i, \theta))^2 \text{ and } S(w_i, \theta) = -\nabla_{\theta} m(x_i, \theta)(y_i - m(x_i, \theta))$$

Note the conditional expectation is zero:

$$\begin{aligned} E(S(w_i, \theta) | x_i) &= -\nabla_{\theta} m(x_i, \theta)(E(y_i | x_i) - m(x_i, \theta)) \\ &= 0 \end{aligned}$$

Let $E(y_i | x_i) - m(x_i, \theta) = u_i$. We have:

$$\begin{aligned} B_0(\theta_0) &= E[S(w_i, \theta_0)S(w_i, \theta_0)'] \\ &= E[\nabla_{\theta} m(x_i, \theta_0)' u u' \nabla_{\theta} m(x_i, \theta_0)] \end{aligned}$$

If homoscedastic, i.e. $E(uu') = \sigma^2 I$, then $B_0(\theta_0) = \sigma^2 \nabla_{\theta} m(x_i, \theta_0)' \nabla_{\theta} m(x_i, \theta_0)$.

The Hessian is given by:

$$H(w_i, \theta) = \nabla_{\theta} m(x_i, \theta) \nabla_{\theta} m(x_i, \theta)' - \nabla_{\theta\theta} m(x_i, \theta)(y_i - m(x_i, \theta))$$

Take the conditional expectation:

$$A_0 = E(H(w_i, \theta_0) | x_i) = \nabla_{\theta} m(x_i, \theta_0) \nabla_{\theta} m(x_i, \theta_0)'$$

Note the 2nd term has zero expectation.

Therefore $B_0 = \sigma^2 A_0$.

Examples:

- (i) If linear, $m(x, \theta) = x\theta$. Then $A_0 = E(x'x)$.
- (ii) If $m(x, \theta) = \exp(x\theta)$. Then $A_0 = E(\exp(2x\theta_0)x'x)$
- (iii) If $m(x, \theta) = \theta_1 + \theta_2 x_2 + \theta_3 x_3^{\theta_4}$, the model is not identified.

For nonlinear regression, A_0 and B_0 are similar because they both depend on $\nabla_{\theta} m(x_i, \theta_0) \nabla_{\theta} m(x_i, \theta_0)'$. Generally, there is no single relationship between A_0 and B_0 , without homoscedasticity.

Estimating asymptotic variance:

$$\hat{A} = \frac{1}{N} \sum_{i=1}^N H(w_i, \hat{\theta}), \text{ and } \hat{B} = \frac{1}{N} \sum_{i=1}^N S(w_i, \hat{\theta})S(w_i, \hat{\theta})'$$

and the covariance matrix is given by:

$$A \text{ var}(\sqrt{N}\hat{\theta}) = \hat{A}^{-1} \hat{B} \hat{A}^{-1}, \text{ or } A \text{ var}(\hat{\theta}) = \frac{1}{N} \hat{A}^{-1} \hat{B} \hat{A}^{-1}$$

Two-step M-estimators

A two-step M-estimators $\hat{\theta}$ of θ_0 solves the problem

$$\min_{\theta \in \Theta} \sum_{i=1}^N q(w_i, \theta, \hat{\gamma})$$

Example: Weighted nonlinear least squares

$$\min_{\theta \in \Theta} \sum_{i=1}^N (y_i - m(x_i, \theta))^2 / h(x_i, \hat{\gamma})$$

The conditions for consistency are:

$$\text{If } \hat{\gamma} \rightarrow \gamma^*, \text{ and } E[q(w, \theta_0, \gamma^*)] < E[q(w, \theta, \gamma^*)] \quad \forall \theta \in \Theta, \theta \neq \theta_0$$

Asymptotic normality: The first order condition gives:

$$\sum_{i=1}^N S(w_i, \hat{\theta}_N; \hat{\gamma}) = \sum_{i=1}^N \frac{\partial q(w_i, \hat{\theta}_N; \hat{\gamma})}{\partial \theta} = 0.$$

1st-order Taylor expansion:

$$\sum_{i=1}^N S(w_i, \hat{\theta}_N; \hat{\gamma}) = \sum_{i=1}^N S(w_i, \theta_0; \hat{\gamma}) + \sum_{i=1}^N \frac{\partial S(w_i, \theta_N^*; \hat{\gamma})}{\partial \theta} (\hat{\theta}_N - \theta_0) = 0.$$

where θ_N^* is between $\hat{\theta}_N$ and θ_0 .

Define $H(\theta) = \frac{\partial S(\theta)}{\partial \theta'}$ $= \frac{\partial^2 q(\theta)}{\partial \theta \partial \theta'}$. Multiply both sides by $1/\sqrt{N}$, we have:

$$\frac{1}{\sqrt{N}} \sum_{i=1}^N S(w_i, \theta_0; \hat{\gamma}) = -\sqrt{N}(\hat{\theta}_N - \theta_0) \frac{1}{N} \sum_{i=1}^N \frac{\partial S(w_i, \theta_N^*; \hat{\gamma})}{\partial \theta}$$

Rearrange this equation:

$$\sqrt{N}(\hat{\theta}_N - \theta_0) = A_0^{-1}(\theta_0, \hat{\gamma}) \left(-\frac{1}{\sqrt{N}} \sum_{i=1}^N S(w_i, \theta_0; \hat{\gamma}) \right) + o_p(1)$$

Given we have: $\hat{\gamma} \rightarrow \gamma^*$. Taylor expansion again:

$$\begin{aligned} & -\frac{1}{\sqrt{N}} \sum_{i=1}^N S(w_i, \theta_0; \hat{\gamma}) \\ &= -\frac{1}{\sqrt{N}} \sum_{i=1}^N S(w_i, \theta_0; \gamma^*) + F_0 \sqrt{N}(\hat{\gamma} - \gamma^*) + o_p(1) \end{aligned}$$

where $F_0 = E(\nabla_{\gamma} S(w_i, \theta_0; \gamma^*))$

Therefore,

$$\begin{aligned}
\sqrt{N}(\hat{\theta}_N - \theta_0) &= A_0^{-1}(\theta_0, \hat{\gamma}) \left(-\frac{1}{\sqrt{N}} \sum_{i=1}^N S(w_i, \theta_0; \hat{\gamma}) \right) + o_p(1) \\
&= A_0^{-1}(\theta_0, \hat{\gamma}) \left[-\frac{1}{\sqrt{N}} \sum_{i=1}^N S(w_i, \theta_0; \gamma^*) + F_0 \sqrt{N}(\hat{\gamma} - \gamma^*) + o_p(1) \right] + o_p(1) \\
&= A_0^{-1}(\theta_0, \hat{\gamma}) \left(-\frac{1}{\sqrt{N}} \sum_{i=1}^N S(w_i, \theta_0; \gamma^*) \right) + A_0^{-1}(\theta_0, \hat{\gamma}) F_0 \sqrt{N}(\hat{\gamma} - \gamma^*) + o_p(1)
\end{aligned}$$

Therefore, if $F_0=0$, we can ignore the effect of $\hat{\gamma}$; There is no need to make adjustment. We just simply treat $\hat{\gamma}$ as a constant. Otherwise, it is necessary to make corrections. One case that $F_0=0$ is the weighted nonlinear least squares:

$$\min_{\theta \in \Theta} \sum_{i=1}^N (y_i - m(x_i, \theta))^2 / h(x_i, \hat{\gamma}).$$

In which

$$S(w_i, \theta_0; \hat{\gamma}) = -\nabla_{\theta} m(x_i, \theta_0) (y - m(x_i, \theta_0)) / h(x_i, \hat{\gamma})$$

Therefore,

$$\begin{aligned}
F_0 &= E(\nabla_{\gamma} S(w_i, \theta_0; \gamma^*)) \\
&= E(-\nabla_{\theta} m(x_i, \theta_0) (y - m(x_i, \theta_0))) \left(-\frac{1}{h^2(x_i, \gamma^*)} \cdot \frac{\partial h(x_i, \gamma^*)}{\partial \gamma} \right) \\
&= 0
\end{aligned}$$

How to make adjustments (when estimating variance?)

Suppose that $\sqrt{N}(\hat{\gamma} - \gamma^*)$ can be written, similarly, as the form:

$$\sqrt{N}(\hat{\gamma} - \gamma^*) = N^{-1/2} \sum_{i=1}^N r_i(\gamma^*) + o_p(1)$$

Almost all estimators we encounter have the representation like this. Given this, we could write:

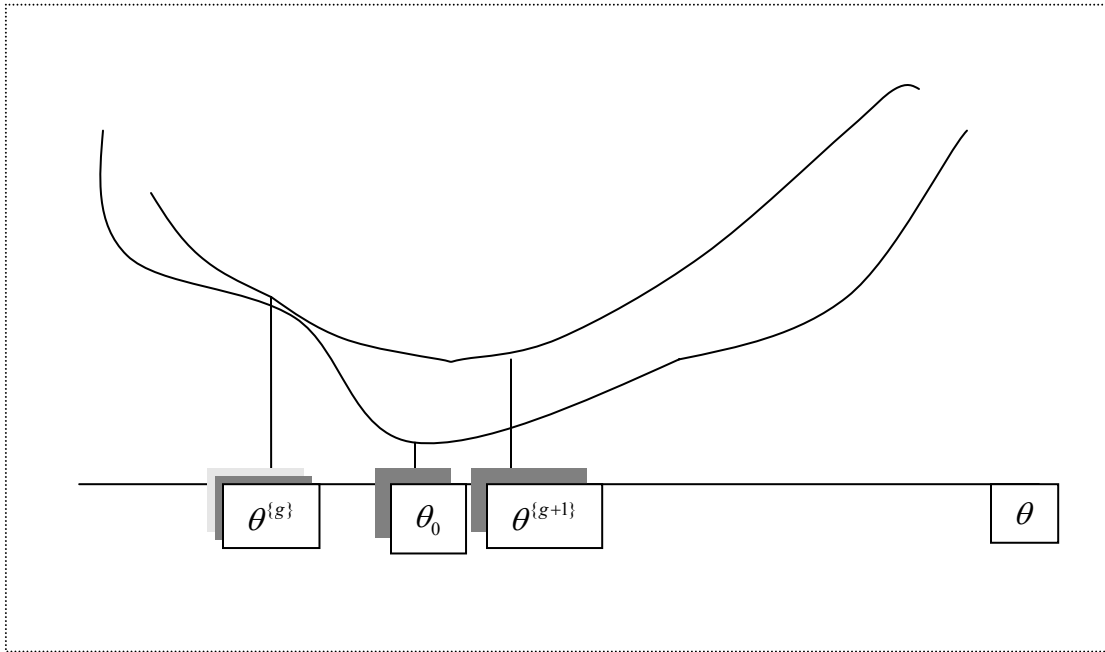
$$\begin{aligned}
\sqrt{N}(\hat{\theta}_N - \theta_0) &= A_0^{-1}(\theta_0, \hat{\gamma}) \left[-\frac{1}{\sqrt{N}} \sum_{i=1}^N (S(w_i, \theta_0; \gamma^*) + F_0 r_i(\gamma^*)) \right] + o_p(1) \\
&\equiv A_0^{-1}(\theta_0, \hat{\gamma}) \left[-\frac{1}{\sqrt{N}} \sum_{i=1}^N g(w_i, \theta_0; \gamma^*) \right] + o_p(1)
\end{aligned}$$

Let $D_0 \equiv E(g(w_i, \theta_0; \gamma^*) g(w_i, \theta_0; \gamma^*))$. Then we have:

$$A \text{ var } \sqrt{N}(\hat{\theta}_N - \theta_0) = A_0^{-1} D_0 A_0^{-1}$$

Optimization methods:

The most popular way of numerically searching for maximum or minimum is via line search.



Given we are now at point $\theta^{(g)}$ how do we move to the next point $\theta^{(g+1)}$? More specifically, we need to find out what direction to move, and what the step size is.

A common way is to apply the quadratic approximation: the objective function can be approximated by a second-order Taylor expansion.

Let the objective function be $l(\theta)$, the second order Taylor expansion is:

$$l(\theta) \approx l(\theta^{(g)}) + \left(\frac{\partial l(\theta^{(g)})}{\partial \theta} \right) (\theta - \theta^{(g)}) + \frac{1}{2} (\theta - \theta^{(g)}) \left(\frac{\partial^2 l(\theta^{(g)})}{\partial \theta \partial \theta'} \right) (\theta - \theta^{(g)})$$

Recognize that:

$$\frac{\partial l(\theta^{(g)})}{\partial \theta} = s(\theta^{(g)})$$

Assume that the optimum is reached at $\theta^{(g+1)}$. The first order condition is given by:

$$s(\theta^{(g)}) + \frac{\partial s(\theta^{(g)})}{\partial \theta} (\theta^{(g+1)} - \theta^{(g)}) = 0$$

This equation shows a possibility for iteration

$$\theta^{\{g+1\}} = \theta^{\{g\}} - \left(\frac{\partial S(\theta^{\{g\}})}{\partial \theta} \right)^{-1} S(\theta^{\{g\}})$$

So the direction is given by: $S(\theta^{\{g\}})$, and the step size is given by: $\left(\frac{\partial S(\theta^{\{g\}})}{\partial \theta} \right)^{-1} S(\theta^{\{g\}})$.

Now the optimization methods include how to specify $\left(\frac{\partial S(\theta^{\{g\}})}{\partial \theta} \right)$.

(1) Newton- Raphson method

$$\theta^{\{g+1\}} = \theta^{\{g\}} - \left(\sum_{i=1}^N H_i(\theta^{\{g\}}) \right)^{-1} \left(\sum_{i=1}^N S_i(\theta^{\{g\}}) \right)$$

Check if:

$$\sum_{i=1}^N q_i(\theta^{\{g+1\}}) - \sum_{i=1}^N q_i(\theta^{\{g\}}) < 0$$

If so, we take the step, otherwise, we reduce the step-size

$$\theta^{\{g+1\}} = \theta^{\{g\}} - r \left(\sum_{i=1}^N H_i(\theta^{\{g\}}) \right)^{-1} \left(\sum_{i=1}^N S_i(\theta^{\{g\}}) \right)$$

When do we stop?

$$\left(\sum_{i=1}^N S_i(\theta^{\{g\}}) \right)' \left(\sum_{i=1}^N H_i(\theta^{\{g\}}) \right)^{-1} \left(\sum_{i=1}^N S_i(\theta^{\{g\}}) \right) < \eta$$

(2) Berndt, Hall, Hall and Hausman (BHHH)

Replace $\left(\sum_{i=1}^N H_i(\theta^{\{g\}}) \right)$ by $\left(\sum_{i=1}^N S_i(\theta^{\{g\}}) S_i(\theta^{\{g\}})' \right)$

Note that BHHH requires computation of scores only and sum of outer product is always at least positive semi-definite.

Maximum Likelihood Estimate

Intuition:

Observe n observations, with one random draw from n random variables that are iid. The intuition is to find out the parameter values which make the joint distribution, or the observed values most likely occur.

This is a powerful intuition. It turns out to be extremely useful.

Plan:

- (1) A theorem shows that estimates from this $\hat{\theta}_{MLE}$ is indeed close to true θ (or converges to true θ)
- (2) The limiting distribution of $\hat{\theta}_{MLE}$
- (3) The efficiency of $\hat{\theta}_{MLE}$

Examples of MLE

$$\text{Bernoulli: } f_n(x | \theta) = \prod_{i=1}^n \theta^{x_i} (1 - \theta)^{1-x_i}$$

The likelihood function is:

$$l_n(\theta) = \sum_{i=1}^n (x_i \log \theta + (1 - x_i) \log(1 - \theta))$$

The first order condition is given by:

$$\frac{\partial l_n(\theta)}{\partial \theta} = \frac{\sum_{i=1}^n x_i}{\theta} - \frac{n - \sum_{i=1}^n x_i}{1 - \theta} = 0$$

$$\text{Therefore, } \hat{\theta}_{MLE} = \bar{x}_n$$

Properties of MLE

1. Consistency: suppose the true value of θ will be denoted as θ_0 :

$$\Pr_{\theta_0} \left\{ \prod_{i=1}^n f(x_i, \theta_0) > \prod_{i=1}^n f(x_i, \theta) \right\} \rightarrow 1 \text{ as } n \rightarrow \infty \text{ for any } \theta \neq \theta_0.$$

Proof: (homework)

This property says that the probability is the largest at the point of the true θ_0 , so we try to get the maximum. Note the point has to be the global one.

Intuitively, if $\hat{\theta}_n$ maximize likelihood function, and θ_0 maximizes the expectation of the likelihood function. Then $\hat{\theta}_n \rightarrow \theta_0$

2. Efficiency

Cramer-Rao information inequality:

Lemma: Let x_i be iid $x_i \rightarrow f(x_i, \theta)$. Let T be a function of x_i . $T = t(x_1, x_2, \dots, x_n)$, and $E(T) = u(\theta)$.

Show that $Var(T) \geq \frac{(u'(\theta))^2}{I(\theta)}$ where $I(\theta) = E\left[\left(\frac{\partial \ln f_n(x, \theta)}{\partial \theta}\right)^2\right]$.

Proof: Let $S_n = \frac{\partial \ln f_n(x)}{\partial \theta} = \frac{f_n'(x, \theta)}{f_n(x, \theta)}$ Then:

$$E(S_n) = \int \frac{f_n'(x, \theta)}{f_n(x, \theta)} f_n(x, \theta) dx = \int f_n'(x, \theta) dx = 0$$

$$u(\theta) = \int T(x) f_n(x, \theta) dx$$

$$u'(\theta) = \int T(x) \frac{\partial f_n(x, \theta)}{\partial \theta} dx$$

$$= \int T(x) \frac{\partial f_n(x, \theta)}{\partial \theta} \cdot \frac{1}{f_n(x, \theta)} \cdot f_n(x, \theta) dx$$

$$= \int T(x) \frac{\partial \ln f_n(x, \theta)}{\partial \theta} \cdot f_n(x, \theta) dx$$

$$= \int T(x) S_n(x, \theta) \cdot f_n(x, \theta) dx$$

$$= E(T(x) S_n(x, \theta)) = Cov(T(x), S_n(x, \theta))$$

The last equality is obtained because $E(S_n) = 0$. Given the fact that

$Cov(x, y) \leq \text{var}(x)^{1/2} \text{var}(y)^{1/2}$, we have:

$$u'(\theta) = Cov(T(x), S_n(x, \theta)) \rightarrow (u'(\theta))^2 \leq \text{var}(T(x)) \text{var}(S_n(x, \theta))$$

$$\text{Note } \text{var}(S_n) = E\left[\left(\frac{\partial \ln f_n(x, \theta)}{\partial \theta}\right)^2\right] \equiv I(\theta)$$

$$\text{We have: } \text{var}(T(x, \theta)) \geq \frac{(u'(\theta))^2}{I(\theta)}$$

Therefore, when $u'(\theta) = 1$, i.e., θ estimate is unbiased, we have the Cramer-Rao inequality:

$$\text{var}(T(x, \theta)) \geq \frac{1}{I(\theta)}$$

$\frac{1}{I(\theta)}$ is the minimum variance that an unbiased estimate may attain.

3. Asymptotic efficiency and normality

Next we show that MLE is asymptotic normal and efficient. First, we show the asymptotic normality.

The likelihood function is given by:

$$l_n(\theta) = \frac{1}{n} \sum_{i=1}^n \ln f(x_i, \theta) \xrightarrow{p} E_{\theta_0}(\ln f(x_i, \theta)) \equiv \int \ln f(x_i, \theta) \cdot f(x_i, \theta_0) dx_i$$

The first order condition is:

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

Note that: $\sqrt{n}S_n(x, \theta_0) \xrightarrow{d} N(0, I(\theta_0))$

Note the solution of equation implies that:

$$S_n(x, \hat{\theta}_{MLE}) = 0.$$

Taylor expansion at θ :

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

Therefore,

$$\sqrt{n}S_n(x, \theta_0) = -\sqrt{n}(\hat{\theta}_{MLE} - \theta_0) \frac{1}{n} \sum_{i=1}^n \left[\frac{f''(x_i, \theta_n^*)}{f(x_i, \theta_n^*)} - \left(\frac{\ln f'(x_i, \theta_n^*)}{f(x_i, \theta_n^*)} \right)^2 \right]$$

Note that we have $|\theta_n^* - \theta_0| \leq |\hat{\theta}_{MLE} - \theta_0|$. So,

$$\hat{\theta}_{MLE} \xrightarrow{p} \theta_0 \Rightarrow \theta_n^* \xrightarrow{p} \theta_0.$$

It is important to notice that a fact:

$$E_{\theta_0} \left(\frac{f''(x, \theta_0)}{f(x, \theta_0)} \right) = \int \frac{f''(x, \theta_0)}{f(x, \theta_0)} f(x, \theta_0) dx = \int f''(x, \theta_0) dx = 0.$$

$$\text{In fact, } E_{\theta_0} \left(\frac{f^{(k)}(x, \theta_0)}{f(x, \theta_0)} \right) = 0 \text{ for } k > 0.$$

Therefore, by the WLLN:

$$\frac{1}{n} \sum_{i=1}^n \frac{f''(x_i, \theta_n^*)}{f(x_i, \theta_n^*)} \xrightarrow{p} E_{\theta_0} \left(\frac{f''(x_i, \theta_0)}{f(x_i, \theta_0)} \right) = 0, \text{ and}$$

$$\frac{1}{n} \sum_{i=1}^n \left(\frac{\ln f'(x_i, \theta_n^*)}{f(x_i, \theta_n^*)} \right)^2 \xrightarrow{p} I(\theta_0)$$

Therefore,

$$\sqrt{n}(\hat{\theta}_{MLE} - \theta_0) \xrightarrow{d} N(0, I(\theta_0)^{-1})$$

Since the minimum variance is achieved for MLE, therefore, the MLE efficient estimator is efficient.

Example: let x have density $f(x) = (1 + \theta)x^\theta \quad \theta > -1 \quad 0 < x < 1$

Obtain the MLE estimator of the mean and compare it with sample average \bar{x} .

Note we can express likelihood in terms of μ , and found the value μ . We could, however, to express the likelihood function in term of $\theta \rightarrow$ more generally, if θ_1 and θ_2 are 1-1 mapping, and $\theta_1 = g(\theta_2)$, then we must have:

Example: $y = x\beta + \varepsilon$, ε is iid normal $N(0, \sigma^2)$. σ^2 is known MLE estimate of β ?

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

$$\ln f(\varepsilon_i, \beta) = c - \ln \sigma - \frac{(y_i - x_i\beta)^2}{2\sigma^2}$$

Therefore,

$$\frac{\partial \ln f(\varepsilon_i, \beta)}{\partial \beta} = \frac{x_i'(y_i - x_i\beta)}{2\sigma^2}$$

$$\frac{\partial^2 \ln f(\varepsilon_i, \beta)}{\partial \beta \partial \beta'} = -\frac{x_i' x_i}{2\sigma^2}$$

$$\rightarrow \sqrt{n}(\hat{\beta}_{MLE} - \beta) \xrightarrow{d} N(0, \sigma^2(x'x)^{-1})$$

Likelihood Function example

$x_i \sim N(e^{\alpha\beta}, 1)$, $\alpha \neq 0$, $y_i \sim N(e^\alpha, 1)$, and $\beta \neq 0$.

x_i and y_i are independent. The density (x_i, y_i) is:

$$f(x_i, y_i) = \frac{1}{2\pi} \exp\left(-\frac{(x_i - e^{\alpha\beta})^2 + (y_i - e^\alpha)^2}{2}\right)$$

The log-likelihood function is given by:

$$l = -\frac{1}{2} \sum \left((x_i - e^{\alpha\beta})^2 + (y_i - e^\alpha)^2 \right)$$

The first order condition is given by:

$$\frac{\partial l}{\partial \alpha} = -\frac{1}{2} \left(-2\beta e^{\alpha\beta} \sum x_i + 2\beta n e^{2\alpha\beta} - 2e^\alpha \sum y_i + 2n e^\alpha \right) = 0$$

$$\frac{\partial l}{\partial \beta} = \frac{1}{2} \left(-2\alpha e^{\alpha\beta} \sum x_i + 2\alpha n e^{2\alpha\beta} \right) = 0$$

Solving this two-equation system:

$$\text{From the second equation: } e^{\alpha\beta} = \frac{1}{n} \sum x_i = \bar{x}$$

$$\text{Plug this into the first equation: } \bar{y} = e^\alpha$$

$$\hat{\beta}_{MLE} = \frac{\ln \bar{x}}{\ln \bar{y}} \text{ is the MLE estimator}$$

How to find the variance of $\hat{\beta}_{MLE}$?

$$\frac{\partial^2 l}{\partial \alpha^2} = \beta^2 e^{\alpha\beta} \sum x_i - 2\beta^2 n e^{2\alpha\beta} + e^\alpha \sum y_i - n e^\alpha$$

$$\frac{\partial^2 l}{\partial \beta^2} = -\alpha^2 e^{\alpha\beta} \sum x_i - 2\alpha^2 n e^{2\alpha\beta}$$

$$\frac{\partial^2 l}{\partial \alpha \partial \beta} = -(1 + \alpha\beta) e^{\alpha\beta} \sum x_i + (1 + 2\alpha\beta) n e^{2\alpha\beta}$$

Plug $\alpha\beta = \ln \bar{x}$ and $\alpha = \ln \bar{y}$, we have:

$$\frac{\partial^2 l}{\partial \alpha^2} = \beta^2 e^{\alpha\beta} \sum x_i - 2\beta^2 n e^{2\alpha\beta} + e^\alpha \sum y_i - n e^\alpha$$

$$\frac{\partial^2 l}{\partial \beta^2} = -\alpha^2 n \bar{x}^2 = -n(\ln \bar{y})^2 \bar{x}^2$$

$$\frac{\partial^2 l}{\partial \alpha \partial \beta} = n \bar{x}^2 \ln \bar{x}$$

$$\rightarrow \text{var}(\sqrt{n} \hat{\beta}_{MLE}) = \frac{1}{(\ln \bar{y})^2 \bar{x}^2}$$

Therefore, $\sqrt{n}(\hat{\beta}_{MLE} - \beta) \xrightarrow{d} N\left(0, \frac{1}{(\ln \bar{y})^2 \bar{x}^2}\right)$

Example: $x_i \sim N(\theta, \theta^2)$:

The log-likelihood is given by:

$$l(x_1, \dots, x_n; \theta) = -\frac{1}{2} \ln 2\pi - n \ln \theta - \frac{\sum (x_i - \theta)^2}{2\theta^2}$$

The first order condition is:

$$\frac{\partial l}{\partial \theta} = -\frac{n}{\theta} + \frac{\sum (x_i - \theta)}{\theta^2} + \frac{\sum (x_i - \theta)^2}{\theta^3} = 0$$

We have two roots to this equality: \rightarrow

$$\begin{aligned} \hat{\theta}_{MLE} &= \frac{-\sum x_i \pm \sqrt{(\sum x_i)^2 + 4n \sum x_i^2}}{2n} \\ &= -\frac{\sum x_i}{2n} \pm \frac{1}{2} \sqrt{\left(\frac{\sum x_i}{n}\right)^2 + \frac{4}{n} \sum x_i^2} \end{aligned}$$

Assume the true θ (in which the data x_i were generated) value is θ_0 . By the Weak

Law of Large Numbers:

$$\begin{aligned} \frac{\sum x_i}{n} &\xrightarrow{p} E(x_i) = \theta_0 \\ \frac{\sum x_i^2}{n} &\xrightarrow{p} E(x_i^2) = \text{Var}(x_i) + (E(x_i))^2 = 2\theta_0^2 \end{aligned}$$

Therefore,

$$\hat{\theta}_{MLE} \xrightarrow{p} -\frac{\theta_0}{2} \pm \frac{3\theta_0}{2}$$

When there are multiple roots, only one of them is a consistent estimator that will converge to the true value θ_0 . The other root is not. Next we show how to distinguish if a root is the consistent estimator of the true parameter value.

Note that, for any θ , we have:

$$\frac{\partial^2 \ln f(x, \theta)}{\partial \theta^2} + \left(\frac{\partial \ln f(x, \theta)}{\partial \theta} \right)^2 = \frac{1}{f(x, \theta)} \frac{\partial^2 f(x, \theta)}{\partial \theta^2} \quad (*)$$

For the normal density:

$$\begin{aligned} f(x, \theta) &= \frac{1}{\theta\sqrt{2\pi}} \exp\left(-\frac{(x-\theta)^2}{2\theta^2}\right) \\ \frac{\partial f(x, \theta)}{\partial \theta} &= -\frac{1}{\theta} f(x, \theta) + f(x, \theta) \left[\frac{(x-\theta)^2}{\theta^3} + \frac{(x-\theta)}{\theta^2} \right] \\ \frac{\partial^2 f(x, \theta)}{\partial \theta^2} &= \frac{1}{\theta^2} f(x, \theta) - \frac{1}{\theta} \left\{ -\frac{1}{\theta} f(x, \theta) + f(x, \theta) \left[\frac{(x-\theta)^2}{\theta^3} + \frac{(x-\theta)}{\theta^2} \right] \right\} \\ &\quad + \left\{ -\frac{1}{\theta} f(x, \theta) + f(x, \theta) \left[\frac{(x-\theta)^2}{\theta^3} + \frac{(x-\theta)}{\theta^2} \right] \right\} \left[\frac{(x-\theta)^2}{\theta^3} + \frac{(x-\theta)}{\theta^2} \right] \\ &\quad + f(x, \theta) (-3\theta^{-4}(x-\theta)^2 - \theta^{-3}2(x-\theta) - 2\theta^{-3}(x-\theta) - \theta^{-2}) \\ &= f(x, \theta) \left[\frac{(x-\theta)^4}{\theta^6} + \frac{2(x-\theta)^3}{\theta^5} - \frac{4(x-\theta)^2}{\theta^4} - \frac{6(x-\theta)}{\theta^3} + \frac{1}{\theta^2} \right] \end{aligned}$$

Based on moment generating function: $E_{\theta_0}(e^{tx}) = \exp\left(\theta t + \frac{1}{2}\theta^2 t^2\right)$, we have:

$$E_{\theta_0}(x) = \theta_0, \quad E_{\theta_0}(x^2) = 2\theta_0^2, \quad E_{\theta_0}(x^3) = 4\theta_0^3, \quad E_{\theta_0}(x^4) = 10\theta_0^4$$

Therefore,

$$\begin{aligned} E_{\theta_0}(x - \theta) &= \int (x - \theta) f(x, \theta_0) dx = \theta_0 - \theta \\ E_{\theta_0}(x - \theta)^2 &= 2\theta_0^2 - 2\theta_0\theta + \theta^2 \\ E_{\theta_0}(x - \theta)^3 &= 4\theta_0^3 - 6\theta_0^2\theta + 3\theta_0\theta^2 - \theta^3 \\ E_{\theta_0}(x - \theta)^4 &= 10\theta_0^4 - 16\theta_0^3\theta + 12\theta_0^2\theta^2 - 43\theta_0\theta^3 + \theta^4 \end{aligned}$$

Take the expectation of equation (*) with respect to the true value θ_0 :

$$\begin{aligned} E_{\theta_0} \left[\frac{\partial^2 \ln f(x, \theta)}{\partial \theta^2} \right] + E_{\theta_0} \left[\left(\frac{\partial \ln f(x, \theta)}{\partial \theta} \right)^2 \right] &= E_{\theta_0} \left[\frac{1}{f(x, \theta)} \frac{\partial^2 f(x, \theta)}{\partial \theta^2} \right] \\ &= E_{\theta_0} \left[\frac{(x - \theta)^4}{\theta^6} + \frac{2(x - \theta)^3}{\theta^5} - \frac{4(x - \theta)^2}{\theta^4} - \frac{6(x - \theta)}{\theta^3} + \frac{1}{\theta^2} \right] \\ &= \frac{10\theta_0^4}{\theta^6} - \frac{8\theta_0^3}{\theta^5} - \frac{8\theta_0^2}{\theta^4} + \frac{4\theta_0}{\theta^3} + \frac{2}{\theta^2} \end{aligned}$$

At the first root, $\hat{\theta}_{MLE} \xrightarrow{p} \theta_0$,

$$\begin{aligned} \frac{1}{N} \sum_i \frac{\partial^2 \ln f(x, \hat{\theta}_{MLE})}{\partial \theta^2} + \left(\frac{\partial \ln f(x, \hat{\theta}_{MLE})}{\partial \theta} \right)^2 &\xrightarrow{p} \\ E_{\theta_0} \left[\frac{\partial^2 \ln f(x, \theta_0)}{\partial \theta^2} \right] + E_{\theta_0} \left[\left(\frac{\partial \ln f(x, \theta_0)}{\partial \theta} \right)^2 \right] & \\ = \frac{10}{\theta_0^2} - \frac{8}{\theta_0^2} - \frac{8}{\theta_0^2} + \frac{4}{\theta_0^2} + \frac{2}{\theta_0^2} &= 0 \end{aligned}$$

At the second root, $\hat{\theta}_{MLE} \xrightarrow{p} -2\theta_0$.

$$\begin{aligned} \frac{1}{N} \sum_i \frac{\partial^2 \ln f(x, \hat{\theta}_{MLE})}{\partial \theta^2} + \left(\frac{\partial \ln f(x, \hat{\theta}_{MLE})}{\partial \theta} \right)^2 &\xrightarrow{p} \\ E_{\theta_0} \left[\frac{\partial^2 \ln f(x, -2\theta_0)}{\partial \theta^2} \right] + E_{\theta_0} \left[\left(\frac{\partial \ln f(x, -2\theta_0)}{\partial \theta} \right)^2 \right] & \\ = -\frac{3}{32\theta_0^2} &\neq 0 \end{aligned}$$

This result shows that the $\frac{1}{N} \sum_i \frac{\partial^2 \ln f(x, \hat{\theta}_{MLE})}{\partial \theta^2} + \left(\frac{\partial \ln f(x, \hat{\theta}_{MLE})}{\partial \theta} \right)^2 = 0$ if the root $\hat{\theta}_{MLE}$ converges to the true value, and $\neq 0$ when $\hat{\theta}_{MLE}$ does not converge to the true value. Therefore, one may test if $\frac{1}{N} \sum_i \frac{\partial^2 \ln f(x, \hat{\theta}_{MLE})}{\partial \theta^2} + \left(\frac{\partial \ln f(x, \hat{\theta}_{MLE})}{\partial \theta} \right)^2 = 0$ as a test of the root that is the consistent estimate.

Generalized Method of Moments (GMM)

First consider a linear model:

$$y_i = x_i\beta + u_i, \quad \text{with } E(x_i' u_i) \neq 0$$

Assumption 1: $E(z_i' u_i) = 0$

Assumption 2: $E(z_i' x_i) \neq 0$

So the moment conditions are: $E(z_i' u_i) = E(z_i' (y_i - x_i\beta)) = 0$

Find β , such that:

$$\min_{\beta} \left(\sum_{i=1}^N z_i' (y_i - x_i\beta) \right)' W \left(\sum_{i=1}^N z_i' (y_i - x_i\beta) \right)$$

The solution to this problem:

$$\begin{aligned} \hat{\beta}_{MM} &= (X' ZWZ' X)^{-1} X' ZWZ' Y \\ &= \beta + (X' ZWZ' X)^{-1} X' ZWZ' u \end{aligned}$$

and the covariance of the estimator:

$$\text{Var}(\hat{\beta}_{MM}) = (X' ZWZ' X)^{-1} X' ZW\Lambda WZ' X (X' ZWZ' X)^{-1}$$

where $\Lambda = E(z_i' u_i u_i' z)$. This is a very long but intuitive covariance matrix.

The next step is to find the optimal weighting matrix W , which is the *GMM*.

When $W = A^{-1}$, then the optimal covariance matrix is reached. In this case,

$$\hat{\beta}_{GMM} = (X' Z\Lambda^{-1} Z' X)^{-1} X' Z\Lambda^{-1} Z' Y, \text{ and}$$

$$\begin{aligned} \text{Var}(\hat{\beta}_{GMM}) &= (X' Z\Lambda^{-1} Z' X)^{-1} X' Z\Lambda^{-1} \Lambda \Lambda^{-1} Z' X (X' Z\Lambda^{-1} Z' X)^{-1} \\ &= (X' Z\Lambda^{-1} Z' X)^{-1} X' Z\Lambda^{-1} Z' X (X' Z\Lambda^{-1} Z' X)^{-1} \\ &= (X' Z\Lambda^{-1} Z' X)^{-1} \end{aligned}$$

Under homoscedasticity, $\Lambda = E(z_i' u_i u_i' z) = \sigma^2 E(z_i' z)$. Then:

$$\begin{aligned} \hat{\beta}_{GMM} &= (X' Z(Z' Z)^{-1} Z' X)^{-1} X' Z(Z' Z)^{-1} Z' Y \\ &= (\hat{X}' X)^{-1} \hat{X}' y = \hat{\beta}_{2SLS} \end{aligned}$$

$$\begin{aligned}
\text{Var}(\hat{\beta}_{GMM}) &= (X' Z \Lambda^{-1} Z' X)^{-1} \\
&= \sigma^2 (X' Z (Z' Z)^{-1} Z' X)^{-1} \\
&= \sigma^2 (\hat{X}' X)^{-1}
\end{aligned}$$

Generalized Method of Moments (GMM)

Similar to the M-estimator setup:

$$E[g(w_i, \theta_0)] = 0, \text{ where } \theta \in R^P, \text{ and } g(w_i, \theta) \in R^L.$$

In other words, the number of parameters is P , and number of moments is L .

When $L < P$, the model is not identified; when $L = P$, the model is exactly identified; and when $L > P$, the model is over-identified.

Similar to the nonlinear least square case, if we define:

$$q(w_i, \theta) = g(w_i, \theta)' g(w_i, \theta)$$

Then a typical model is given by:

$$\min_{\theta} \sum_{i=1}^n q(w_i, \theta) = \sum_{i=1}^n g(w_i, \theta)' g(w_i, \theta) \quad (*)$$

All the properties from previous discussions apply.

Alternatively, consider a different model:

$$\min_{\theta} \left[\sum_{i=1}^n g(w_i, \theta) \right]' \hat{\Xi} \left[\sum_{i=1}^n g(w_i, \theta) \right] \quad (**)$$

when $\hat{\Xi} = I$, $(**) = (*)$ when cross terms are close to zero, which is true when two observations are independent.

However, compare $(*)$ and $(**)$, the advantage of $(**)$ is that a weighting matrix $\hat{\Xi}$ can be included in $(**)$ while it is difficult to include a weighting matrix to $(*)$.

Including the optimal weighting matrix would improve the asymptotic efficiency.

Define:

$$Q_N(\theta) \equiv \left[\frac{1}{n} \sum_{i=1}^n g(w_i, \theta) \right]' \hat{\Xi} \left[\frac{1}{n} \sum_{i=1}^n g(w_i, \theta) \right]$$

We can use the first-order condition and Taylor expansion to obtain the asymptotic distribution of this estimator.