The Maximum Likelihood Estimator (MLE)

Empirical Analysis II, Econ 311: Topic 1

Prof. Harald Uhlig[1]

[1]University of Chicago
Department of Economics
huhlig@uchicago.edu

Winter 2019

Outline

Measure Spaces

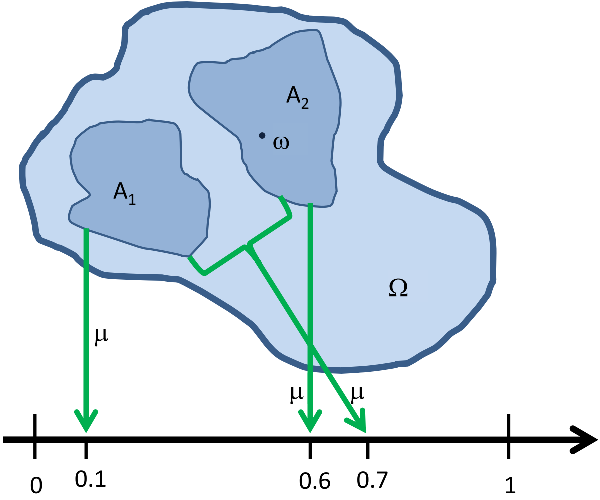
A measure space $(\Omega, \mathcal{F}, \mu)$ consists of

1. Ω: a set of points ("states of nature") ω.
2. $\mathcal{F}$: a set of subsets ("events") of Ω, which form a σ-algebra:
    1. $\Omega \in \mathcal{F}$.
    2. If $A \in \mathcal{F}$, then so is its complement $A^c = \Omega \backslash A \in \mathcal{F}$.
    3. $A_j \in \mathcal{F}, j = 1, 2, \ldots$ implies $\bigcup_{j=1}^{\infty} A_j \in \mathcal{F}$.
3. A measure μ, i.e. a mapping $\mu : \mathcal{F} \to \mathbb{R}_+ \cup \{\infty\}$ with
    1. Positivity: $\mu(A) \geq 0$.
    2. σ-additivity: if $A_j \in \mathcal{F}, j = 1, 2, \ldots$ are disjoint, then

    $$\mu\left(\bigcup_{j=1}^{\infty} A_j\right) = \sum_{j=1}^{\infty} \mu(A_j)$$

    3. $\mu(\emptyset) = 0$.
4. Probability space / probability measure: $\mu(\Omega) = 1$.

A measure space



Example 1: Rolling two dice

- $\Omega = \{\omega = (x, y) \mid x, y \in \{1, \ldots, 6\}\}$
- Three σ-algebras:
  - $\mathcal{F}_0 = \{\emptyset, \Omega\}$.
  - $\mathcal{F}_1 = \{A_x \times \{1, \ldots, 6\} \mid A_x \subseteq \{1, \ldots, 6\}\}$
  - $\mathcal{F}_2 = \{A \subseteq \Omega\}$
- $\mu(A) = \sum_{\omega \in A} \frac{1}{36}$: probability measure.
- A filtration: $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \mathcal{F}_2$.
- Dice roll at $t = 1, t = 2$. $\mathcal{F}_t$: events "known" at t. Information.

Example 1: Rolling two dice

- $\Omega = \{\omega = (x, y) \mid x, y \in \{1, \ldots, 6\}\}$
- Three σ-algebras:
    - $\mathcal{F}_0 = \{\emptyset, \Omega\}$.
    - $\mathcal{F}_1 = \{A_x \times \{1, \ldots, 6\} \mid A_x \subseteq \{1, \ldots, 6\}\}$
    - $\mathcal{F}_2 = \{A \subseteq \Omega\}$
- $\mu(A) = \sum_{\omega \in A} \frac{1}{36}$: probability measure.
- A filtration: $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \mathcal{F}_2$.
- Dice roll at $t = 1, t = 2$. $\mathcal{F}_t$: events “known” at t. Information.

Example 2: An infinite sequence

- $\Omega = \mathbb{N}$
- $\mathcal{F} = \{A \subseteq \Omega\}$.
- Let $\alpha_j = \mu(\{j\})$. Then $\mu(A) = \sum_{j \in A} \alpha_j$.

Example 3: the Lebesgue measure

- $\Omega = \mathbb{R}^m$.
- $\mathcal{F} = \mathcal{B}(\Omega)$: the Borel-$\sigma$-algebra, i.e. the smallest σ-algebra, which contains all open subsets of Ω.
- Let $I_j = [a_j, b_j], a_j \leq b_j \in \mathbb{R}$ be intervals. Define the box $B = I_1 \times \ldots \times I_n$. Define

$$\mu(B) = (b_1 - a_1)(b_2 - a_2) \ldots (b_m - a_m).$$

  Extend this to $\mathcal{F}$.
- For the mathematicians.
  - Extend to the Lebesgue-measurable sets $\bar{\mathcal{F}} = \{A \cup B \mid A \in \mathcal{F}, B \subseteq C \in \mathcal{F}, \mu(C) = 0\}$ per $\mu(A \cup B) = \mu(A)$.
  - Or this way. Outer measure:

$$\mu^*(A) = \inf\{\sum_j \mu(B_i) \mid A \subseteq \bigcup_i B_i, B_i \text{ is a box}\}$$

    A is Lebesgue measurable if $\mu^*(A) = \mu^*(A \cap C) + \mu^*(A \backslash C)$ for all $C \subseteq \mathbb{R}^n$. Define $\mu(A) = \mu^*(A)$.

Integration

Let $(\Omega, \mathcal{F}, \mu)$ be a measure space.

- A function $f : \Omega \to \mathbb{R}^k$ is called $\mathcal{F}$-measurable if $f^{-1}(B) \in \mathcal{F}$ for every Borel set $B \in \mathcal{B}(\mathbb{R}^k)$.
- Suppose $f = 1_A$ is an indicator function on a set $A \in \mathcal{F}$: $1_A(\omega) = 1$, if $\omega \in A$ and $1_A(\omega) = 0$, if $\omega \notin A$. Define the integral

$$\int f \, d\mu = \int f(\omega) \, \mu(d\omega) = \mu(A)$$

- Suppose f is a linear combination of indicator functions,

$$f(\omega) = \sum_{j=1}^{n} \psi_j 1_{A_j}, \; A_j \in \mathcal{F}$$

  Define the integral per linear extension,

$$\int f \, d\mu = \sum_{j=1}^{n} \psi_j \int 1_{A_j} d\mu = \sum_{j=1}^{n} \psi_j \, \mu(A_j)$$

Integration

- Extend to all positive measurable functions. Extend to all measurable functions f per

$$\int f d\mu = \int \max(f, 0) d\mu - \int \max(-f, 0) d\mu$$

provided at least one of the integrals is finite.

- For $A \in \mathcal{F}$, define

$$\int_A f d\mu = \int 1_A(\omega) f(\omega) \mu(d\omega).$$

- If μ is a probability measure, define the expectation $E[f] = \int f d\mu$.

The Radon-Nikodym Theorem

Theorem

Let $\mathcal{F}$ be a σ-algebra on Ω. Let μ and ν be two measures on $\mathcal{F}$. Suppose that $\mu(\Omega) < \infty$ and $\nu(\Omega) < \infty$. Suppose that ν is absolutely continuous with respect to μ, $\nu \ll \mu$, i.e. $\mu(A) = 0$ implies $\nu(A) = 0$ for $A \in \mathcal{F}$. Then there exists a positive measurable function g, called the Radon-Nikodym derivative,

$$g : \Omega \to \mathbb{R}_+ \text{ or } g = \frac{d\nu}{d\mu} \text{ with } \nu(A) = \int_A g\, d\mu = \int_A \frac{d\nu}{d\mu} d\mu$$

for all $A \in \mathcal{F}$.

Remark: This can be extended to signed measures $\nu : \Omega \to \mathbb{R}$: drop the requirement of "positiveness", but impose $-\infty < \nu(\Omega) < \infty$.

Example 1: Rolling two dice

- $\Omega = \{\omega = (x, y) \mid x, y \in \{1, \ldots, 6\}\}$.
- $\mathcal{F} = \mathcal{F}_j$, for one of $j = 0, 1, 2$.
- $\mu(A) = \sum_{\omega \in A} \frac{1}{36}$.
- Let $f : \Omega \to \mathbb{R}$ be measurable.

$$\int f \, d\mu = \sum_{\omega \in \Omega} \frac{f(\omega)}{36}$$

- Conditional expectation. Let $f : \Omega \to \mathbb{R}$ be $\mathcal{F}_2$-measurable. Find a $\mathcal{F}_1$-measurable function $g : \Omega \to \mathbb{R}$, $g = E[f \mid \mathcal{F}_1] = E_1[f]$ per the property

$$\int_A g \, d\mu = \int_A f \, d\mu, \quad \text{for all } A \in \mathcal{F}_1$$

- Existence of g: per the Radon-Nikodym-Theorem for signed measures. Here:

$$g(x, y) = E_1[f(x, y)] = E[f(x, y) \mid \mathcal{F}_1] = E[f(x, y) \mid x] = \sum_{j=1}^{6} \frac{1}{6} f(x, j)$$

Example 2: An infinite sequence

- $\Omega = \mathbb{N}$
- $\mathcal{F} = \{A \subseteq \Omega\}$.
- $\mu(A) = \sum_{j \in A} \alpha_j$.
- For $f : \Omega \to \mathbb{R}$,

$$\int f \, d\mu = \sum_{j=1}^{\infty} \alpha_j f(j)$$

  where $\alpha_j = \mu(\{j\})$.
- "Summation is integration".

Example 3: the Lebesgue measure

- Ω: an open subset of some $\mathbb{R}^n$.
- $\mathcal{F} = \mathcal{B}(\Omega)$: the Borel-$\sigma$-algebra.
- μ: the Lebesgue measure.
- For a measurable function $f : \Omega \to \mathbb{R}$, $\int f \, d\mu$ is what you expect it to be.
- Example: let $I_j = [a_j, b_j], a_j \leq b_j \in \mathbb{R}$ be intervals. Define the box $B = I_1 \times \ldots \times I_n$. Let $f(\omega) = \kappa 1_B$. Then,

$$\int f \, d\mu = \kappa(b_1 - a_1)(b_2 - a_2) \ldots (b_n - a_n)$$

Outline

The framework

- (Unknown) parameter $\theta \in \Theta$. Measure $\mu(d\theta)$. Often: $\Theta \subseteq \mathbb{R}^m$.
- Observation $y \in Y$. Measure $\nu(dy)$.
- Probability density $f(y \mid \theta)$ wrt ν. Thus,

$$\int f(y \mid \theta)\nu(dy) = 1, \text{ for all } \theta \in \Theta$$

- Likelihood function: $L(\theta \mid y) = f(y \mid \theta)$.
- Log-likelihood function: $\ell(\theta \mid y) = \ln L(\theta \mid y)$.
- Experiment on θ. Leads to an observation $y \sim f(y \mid \theta)$ for some known f, if it is carried out.
- Unconditional vs conditional likelihood. Suppose, draws of X do not depend on θ. Then

$$f(X, y \mid \theta) = f(y \mid \theta, X) f(X)$$

- Often: iid observations, $y = (y_1, \ldots, y_n)$,

$$f^{(n)}(y \mid \theta) = \prod^{n} f(y_j \mid \theta)$$

Example 1: linear regression

- $y \in \mathbb{R}^n, X \in \mathbb{R}^{n \times k}, \beta \in \mathbb{R}^k, \Sigma \in \mathbb{R}^{n \times n}$ pos.def.,

$$y = X\beta + \epsilon, \; \epsilon \sim \mathcal{N}(0, \Sigma)$$

- Solve for ϵ:

$$\epsilon = y - X\beta \sim \mathcal{N}(0, \Sigma)$$

- $\theta = \beta$. Or: $\theta = (\beta, \Sigma)$. Or: $\theta = \Sigma$. Or ...
- Assume: Distribution of X does not depend on θ.
- Conditional likelihood function: $L(\theta \mid y, X) = f(y \mid \theta, X)$,

$$L(\theta \mid y, X) = \frac{1}{(2\pi)^{n/2} \mid \Sigma \mid^{1/2}} \exp\left(-\frac{1}{2}(y - X\beta)'\Sigma^{-1}(y - X\beta)\right)$$

- Special case: iid assumption, $\Sigma = \sigma^2 I_n$.

$$L(\theta \mid y, X) = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left(-\frac{1}{2\sigma^2}\sum_{j=1}^{n}(y_j - X_j\beta)^2\right)$$

Example 2: binomial distribution

"Biased coin"

- $y \in \{0, 1\}$.
- Likelihood function:

$$\begin{aligned} L(\theta \mid y = 1) &= f(y = 1 \mid \theta) = P(y = 1 \mid \theta) &= \theta \\ L(\theta \mid y = 0) &= f(y = 0 \mid \theta) = P(y = 0 \mid \theta) &= 1 - \theta \end{aligned}$$

- Observation: $y = 1$.
- "How likely is it to get the observed data $y = 1$?". Answer: $L(\theta \mid y = 1) = \theta$.
- $y = 1$ is 9 times more likely for $\theta = 0.9$ than $\theta = 0.1$.
- Likelihood ratio: 9. Log-Likelihood-Ratio: 2.2.

Example 3: binary choice. Probit and Logit.

- Example: choose to accept ($y = 1$) or reject ($y = 0$) a job, if $X\beta$ is larger than some random outside option, i.e.

$$y = 1, \text{ iff } \epsilon \leq X\beta, \qquad y = 0, \text{ iff } \epsilon > X\beta$$

- Data $y \in \{0, 1\}$, X. Cannot observe ϵ. $\theta = \beta$.
- CDF (=cum.distr.function) G for ϵ. $\theta = \beta$. Likelihood function:

$$L(\theta \mid y = 1) = G(X\beta), \;\; L(\theta \mid y = 0) = 1 - G(X\beta)$$

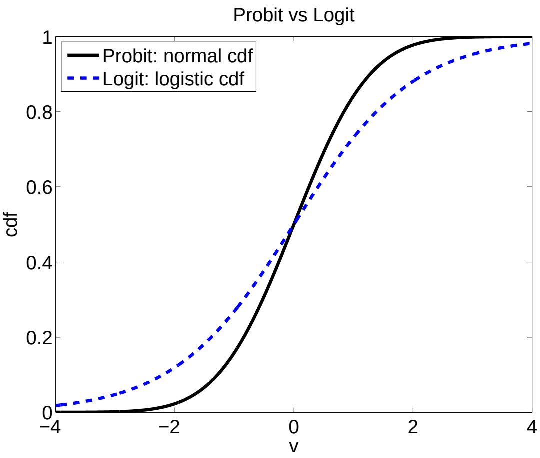
- Probit: standard normal,

$$G(v) = \int_{-\infty}^{v} \frac{1}{\sqrt{2\pi}} e^{-s^2/2} ds$$

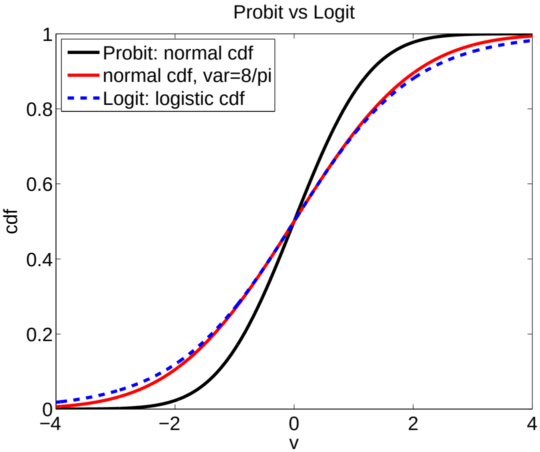
- Logit: logistic distribution,

$$G(v) = \frac{e^v}{1 + e^v}$$

Probit vs Logit



Probit vs Logit vs normal cdf with $\sigma = \sqrt{8/\pi} \approx 1.6$.



Example 4: censored regression. Tobit

- Example: hours worked depend on wage, provided the agent chooses to work at all. Data on hours available only then, on wage always.
- Tobit model:

$$y = \max\{y^*; 0\}, \text{ where } y^* = X\beta + \epsilon, \ \epsilon \sim \mathcal{N}(0, \sigma^2)$$

- Data: y, X. Not observed: y^*, ϵ.
- $\theta = (\beta, \sigma^2)$. Likelihood function:

$$L(\theta \mid y, X) = \begin{cases} \frac{1}{\sigma}\varphi\left(\frac{y - X\beta}{\sigma}\right) & \text{if } y > 0 \\ \Phi\left(\frac{-X\beta}{\sigma}\right) & \text{if } y = 0 \end{cases}$$

  where φ, Φ are the pdf and cdf of a standard normal.
- $L(\cdot \mid y, X)$: a Radon-Nikodym derivative wrt to which measure?

Outline

The Maximum Likelihood Estimator

$$\hat{\theta} = \text{argmax}_{\theta} L(\theta \mid y)$$

Remarks

- Maximizing the log-likelihood is the same:

$$\begin{aligned}\hat{\theta} &= \text{argmax}_\theta L(\theta \mid y) \\ &= \text{argmax}_\theta \ell(\theta \mid y)\end{aligned}$$

- Unconditional vs unconditional likelihood. Suppose, draws of X do not depend on θ, i.e.

$$f((X, y) \mid \theta) = f(y \mid \theta, X) f(X)$$

  Then, the MLE of the conditional and the unconditional likelihood function are the same,

$$\hat{\theta} = \text{argmax}_\theta f((X, y) \mid \theta) = \text{argmax}_\theta f(y \mid \theta, X)$$

The Score

From here on, $\theta \in \Theta \subseteq \mathbb{R}^m$, an open set.
The score is the first derivative of the log-likelihood function,

$$s(\theta) = s(\theta \mid y) = \frac{\partial \ell(\theta \mid y)}{\partial \theta}$$

Remark: the score is defined to be a column vector.

The Score

$$\boxed{E[s(\theta \mid y)] = 0}$$

Proof: Note: θ in arg. is also "truth". For emphasis: $E_\theta[s(\theta \mid y)] = 0$.

- For all θ, $\int f(y \mid \theta)\nu(dy) = 1$.
- Therefore $\int \frac{\partial f(y|\theta)}{\partial \theta}\nu(dy) = 0$.
- Rewrite:

$$\begin{aligned} 0 &= \int \frac{\partial f(y \mid \theta)}{\partial \theta}\nu(dy) \\ &= \int \frac{\frac{\partial f(y|\theta)}{\partial \theta}}{f(y \mid \theta)} f(y \mid \theta)\nu(dy) \\ &= \int \frac{\partial \ell(\theta \mid y)}{\partial \theta} f(y \mid \theta)\nu(dy) \\ &= E[s(\theta \mid y)] \end{aligned}$$

The Score

$$\boxed{E[s(\theta \mid y)] = 0}$$

Proof: Note: θ in arg. is also "truth". For emphasis: $E_\theta[s(\theta \mid y)] = 0$.

- For all θ, $\int f(y \mid \theta)\nu(dy) = 1$.
- Therefore $\int \frac{\partial f(y|\theta)}{\partial\theta}\nu(dy) = 0$.
- Rewrite:

$$\begin{aligned}
0 &= \int \frac{\partial f(y \mid \theta)}{\partial \theta}\nu(dy) \\
&= \int \frac{\frac{\partial f(y|\theta)}{\partial\theta}}{f(y \mid \theta)} f(y \mid \theta)\nu(dy) \\
&= \int \frac{\partial \ell(\theta \mid y)}{\partial \theta} f(y \mid \theta)\nu(dy) \\
&= E[s(\theta \mid y)]
\end{aligned}$$

The Score

$$\boxed{E[s(\theta \mid y)] = 0}$$

Proof: Note: θ in arg. is also "truth". For emphasis: $E_{\theta}[s(\theta \mid y)] = 0$.

- For all θ, $\int f(y \mid \theta)\nu(dy) = 1$.
- Therefore $\int \frac{\partial f(y|\theta)}{\partial \theta}\nu(dy) = 0$.
- Rewrite:

$$
\begin{aligned}
0 &= \int \frac{\partial f(y \mid \theta)}{\partial \theta}\nu(dy) \\
&= \int \frac{\frac{\partial f(y|\theta)}{\partial \theta}}{f(y \mid \theta)} f(y \mid \theta)\nu(dy) \\
&= \int \frac{\partial \ell(\theta \mid y)}{\partial \theta} f(y \mid \theta)\nu(dy) \\
&= E[s(\theta \mid y)]
\end{aligned}
$$

The Score

$$\boxed{E[s(\theta \mid y)] = 0}$$

Proof: Note: θ in arg. is also “truth”. For emphasis: $E_\theta[s(\theta \mid y)] = 0$.

- For all θ, $\int f(y \mid \theta)\nu(dy) = 1$.
- Therefore $\int \frac{\partial f(y|\theta)}{\partial \theta}\nu(dy) = 0$.
- Rewrite:

$$\begin{aligned} 0 &= \int \frac{\partial f(y \mid \theta)}{\partial \theta}\nu(dy) \\ &= \int \frac{\frac{\partial f(y|\theta)}{\partial \theta}}{f(y \mid \theta)} f(y \mid \theta)\nu(dy) \\ &= \int \frac{\partial \ell(\theta \mid y)}{\partial \theta} f(y \mid \theta)\nu(dy) \\ &= E[s(\theta \mid y)] \end{aligned}$$

The Information Matrix

The Fisher information matrix is defined as

$$\mathcal{I}(\theta) = E\left[s(\theta \mid y)s(\theta \mid y)'\right]$$

The Information Matrix Equality

Theorem

$$\mathcal{I}(\theta) = E\left[s(\theta \mid y)s(\theta \mid y)'\right] = -E\left[\frac{\partial^2 \ell(\theta \mid y)}{\partial\theta\,\partial\theta'}\right]$$

Proof: θ is "truth", i.e. $\mathcal{I}(\theta) = E_\theta\left[s(\theta \mid y)s(\theta \mid y)'\right] = -E_\theta\left[\frac{\partial^2 \ell(\theta|y)}{\partial\theta\,\partial\theta'}\right]$.

- Recall: $0 = E[s(\theta \mid y)] = \int \frac{\partial \ell(\theta|y)}{\partial\theta} f(y \mid \theta)\nu(dy)$.
- Differentiate wrt θ',

$$\begin{aligned} 0 &= \int \frac{\partial^2 \ell(\theta \mid y)}{\partial\theta\,\partial\theta'} f(y \mid \theta)\nu(dy) \\ &\quad + \int \frac{\partial \ell(\theta \mid y)}{\partial\theta} \frac{\frac{\partial f(y|\theta)}{\partial\theta'}}{f(y \mid \theta)} f(y \mid \theta)\nu(dy) \\ &= E\left[\frac{\partial^2 \ell(\theta \mid y)}{\partial\theta\,\partial\theta'}\right] + \mathcal{I}(\theta) \end{aligned}$$

The Information Matrix Equality

Theorem

$$\mathcal{I}(\theta) = E\left[s(\theta \mid y)s(\theta \mid y)'\right] = -E\left[\frac{\partial^2 \ell(\theta \mid y)}{\partial\theta\,\partial\theta'}\right]$$

Proof: θ is "truth", i.e. $\mathcal{I}(\theta) = E_{\theta}\left[s(\theta \mid y)s(\theta \mid y)'\right] = -E_{\theta}\left[\frac{\partial^2 \ell(\theta|y)}{\partial\theta\,\partial\theta'}\right]$.

- Recall: $0 = E[s(\theta \mid y)] = \int \frac{\partial \ell(\theta|y)}{\partial\theta} f(y \mid \theta)\nu(dy)$.
- Differentiate wrt θ',

$$\begin{aligned} 0 &= \int \frac{\partial^2 \ell(\theta \mid y)}{\partial\theta\,\partial\theta'} f(y \mid \theta)\nu(dy) \\ &\quad + \int \frac{\partial \ell(\theta \mid y)}{\partial\theta} \frac{\frac{\partial f(y|\theta)}{\partial\theta'}}{f(y \mid \theta)} f(y \mid \theta)\nu(dy) \\ &= E\left[\frac{\partial^2 \ell(\theta \mid y)}{\partial\theta\,\partial\theta'}\right] + \mathcal{I}(\theta) \end{aligned}$$

The Information Matrix Equality

Theorem

$$\mathcal{I}(\theta) = E\left[s(\theta \mid y)s(\theta \mid y)'\right] = -E\left[\frac{\partial^2 \ell(\theta \mid y)}{\partial\theta\,\partial\theta'}\right]$$

Proof: θ is "truth", i.e. $\mathcal{I}(\theta) = E_\theta\left[s(\theta \mid y)s(\theta \mid y)'\right] = -E_\theta\left[\frac{\partial^2 \ell(\theta|y)}{\partial\theta\,\partial\theta'}\right]$.

- Recall: $0 = E[s(\theta \mid y)] = \int \frac{\partial \ell(\theta|y)}{\partial\theta} f(y \mid \theta)\nu(dy)$.
- Differentiate wrt θ',

$$\begin{aligned}
0 &= \int \frac{\partial^2 \ell(\theta \mid y)}{\partial\theta\,\partial\theta'} f(y \mid \theta)\nu(dy) \\
&\quad + \int \frac{\partial \ell(\theta \mid y)}{\partial\theta} \frac{\frac{\partial f(y|\theta)}{\partial\theta'}}{f(y \mid \theta)} f(y \mid \theta)\nu(dy) \\
&= E\left[\frac{\partial^2 \ell(\theta \mid y)}{\partial\theta\,\partial\theta'}\right] + \mathcal{I}(\theta)
\end{aligned}$$

First- and second order expansions around some θ

$$
\begin{aligned}
\ell(\tilde{\theta}) &\approx \ell(\theta) + s(\theta)'(\tilde{\theta} - \theta) + \frac{1}{2}(\tilde{\theta} - \theta)' \frac{\partial^2 \ell(\theta)}{\partial\theta\,\partial\theta'}(\tilde{\theta} - \theta) \\
s(\tilde{\theta}) &\approx s(\theta) + \frac{\partial^2 \ell(\theta)}{\partial\theta\,\partial\theta'}(\tilde{\theta} - \theta)
\end{aligned}
$$

Likewise, with θ as "truth":

$$
\begin{aligned}
E_\theta[\ell(\tilde{\theta})] &\approx E_\theta[\ell(\theta)] + E_\theta[s(\theta)]'(\tilde{\theta} - \theta) - \tfrac{1}{2}(\tilde{\theta} - \theta)'\mathcal{I}(\theta)(\tilde{\theta} - \theta) \\
&\approx E_\theta[\ell(\theta)] - \tfrac{1}{2}(\tilde{\theta} - \theta)'\mathcal{I}(\theta)(\tilde{\theta} - \theta)
\end{aligned}
$$

$$
\begin{aligned}
E_\theta[s(\tilde{\theta})] &\approx E_\theta[s(\theta)] - \mathcal{I}(\theta)(\tilde{\theta} - \theta) \\
&\approx - \mathcal{I}(\theta)(\tilde{\theta} - \theta)
\end{aligned}
$$

Outline

Asymptotic Theory

- iid sample, $y = (y_1, \ldots, y_n)$. Truth: $\theta = \theta_0$.
- Correct is: $\ell, s, \mathcal{I}$ depend on entire sample. For example, $\ell(\theta \mid y) = \ell(\theta \mid (y_1, \ldots, y_n))$.
- Now: slight abuse of notation. $\ell, s, \mathcal{I}$ for one obs., e.g. $\ell(\theta \mid y_j)$.
- Let

$$
\begin{aligned}
\ell_n(\theta) &= \tfrac{1}{n} \textstyle\sum_{j=1}^{n} \ell(\theta \mid y_j) &&= \frac{1}{n} \ell(\theta \mid y_1, \ldots, y_n) \\
s_n(\theta) = \frac{\partial \ell_n(\theta)}{\partial \theta} &= \tfrac{1}{n} \textstyle\sum_{j=1}^{n} \frac{\partial \ell(\theta|y_j)}{\partial \theta} &&\xrightarrow{P} E_{\theta_0}[s(\theta)] = 0 \text{ at } \theta = \theta_0 \\
H_n(\theta) = \frac{\partial^2 \ell_n(\theta)}{\partial \theta \partial \theta'} &= \tfrac{1}{n} \textstyle\sum_{j=1}^{n} \frac{\partial^2 \ell(\theta|y_j)}{\partial \theta \partial \theta'} &&\xrightarrow{P} -\mathcal{I}(\theta_0) \text{ at } \theta = \theta_0
\end{aligned}
$$

- Central limit theorem:

$$
\sqrt{n}\, s_n(\theta_0) \xrightarrow{d} \mathcal{N}(0, E[s(\theta_0)s(\theta_0)']) = \mathcal{N}(0, \mathcal{I}(\theta_0))
$$

Asymptotic Normality for the MLE: Delta method

- The MLE $\hat{\theta}_n$ solves $s_n(\hat{\theta}_n) = 0$.
- First-order expansion around θ_0:

$$0 = s_n(\hat{\theta}_n) \approx s_n(\theta_0) + H_n(\theta_0)(\hat{\theta}_n - \theta_0)$$

- Assume $\mathcal{I}(\theta)$ is invertible (hence: positive definite).

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \approx -\sqrt{n}\,\mathcal{I}(\theta_0)^{-1} H_n(\theta_0)(\hat{\theta}_n - \theta_0) \approx \sqrt{n}\,\mathcal{I}(\theta_0)^{-1} s_n(\theta_0)$$

- Take the limit.

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}\left(0, \mathcal{I}(\theta_0)^{-1}\mathcal{I}(\theta_0)\,\mathcal{I}(\theta_0)^{-1}\right) = \mathcal{N}\left(0, \mathcal{I}(\theta_0)^{-1}\right)$$

Asymptotic Normality for the MLE

Theorem

If $\mathcal{I}(\theta)$ is invertible at the true θ, then

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}\left(0, \mathcal{I}(\theta)^{-1}\right)$$

The inverse of the information matrix is the asymptotic variance of the MLE

Feasible estimation of $\mathcal{I}(\theta)$

- Recall:

$$\mathcal{I}(\theta) = E\left[s(\theta \mid y)s(\theta \mid y)'\right] = -E\left[\frac{\partial^2 \ell(\theta \mid y)}{\partial\theta\, \partial\theta'}\right]$$

  Typically, not known and needs to be estimated.

- As average of score products

$$\hat{\mathcal{I}}_n^{(1)}(\theta_n) = \frac{1}{n}\sum_{j=1}^{n} s(\theta_n \mid y_j)s(\theta_n \mid y_j)'$$

- As average of second derivatives (see $H_n(\theta)$):

$$\hat{\mathcal{I}}_n^{(2)}(\theta_n) = -\frac{1}{n}\sum_{j=1}^{n} \frac{\partial^2 \ell(\theta_n \mid y_j)}{\partial\theta_n\, \partial\theta_n'}$$

- If $\theta_n \xrightarrow{P} \theta$, then $\hat{\mathcal{I}}_n^{(j)}(\theta_n) \xrightarrow{P} \mathcal{I}(\theta)$ is a consistent estimator.
- Often: $\theta_n = \hat{\theta}_n$.

Example 1: linear regression

$y_i \in \mathbb{R}$, $X_i \in \mathbb{R}^{1\times k}$, $\theta = [\beta, \sigma^2]$, $\beta \in \mathbb{R}^k, \sigma^2 > 0$,

$$y_i = X_i\beta + \epsilon_i,\ \epsilon_i \sim \mathcal{N}(0, \sigma^2),\ i = 1, \ldots, n \text{ iid}$$

Let X be $n \times k$, with X_i as i-th row. Similarly, let y be $n \times 1$.

$$\ell_n(\theta \mid y, X) = -\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln(\sigma^2) - \frac{1}{2n\sigma^2}\sum_{j=1}^{n}(y_j - X_j\beta)^2$$

$$s_n(\theta \mid y, X) = \begin{bmatrix} \frac{1}{n\sigma^2}(X'y - X'X\beta) \\ -\frac{1}{2\sigma^2} + \frac{1}{2n\sigma^4}\sum_{j=1}^{n}(y_j - X_j\beta)^2 \end{bmatrix}$$

$$\hat{\theta}_n = \begin{bmatrix} \hat{\beta} \\ \widehat{\sigma^2} \end{bmatrix} = \begin{bmatrix} (X'X)^{-1}X'y \\ \frac{1}{n}\sum_{j=1}^{n}(y_j - X_j\hat{\beta})^2 \end{bmatrix}$$

$$-E\left[H_n(\theta) \mid X\right] = \begin{bmatrix} \frac{X'X}{n\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{bmatrix} = \begin{bmatrix} \sigma^2\left(\frac{X'X}{n}\right)^{-1} & 0 \\ 0 & 2\sigma^4 \end{bmatrix}^{-1} = \mathcal{I}(\theta \mid X)$$

Example 1: linear regression

$y_i \in \mathbb{R}$, $X_i \in \mathbb{R}^{1\times k}$, $\theta = [\beta, \sigma^2]$, $\beta \in \mathbb{R}^k, \sigma^2 > 0$,

$$y_i = X_i\beta + \epsilon_i,\ \epsilon_i \sim \mathcal{N}(0, \sigma^2),\ i = 1, \ldots, n \ \text{ iid}$$

Let X be $n \times k$, with X_i as i-th row. Similarly, let y be $n \times 1$.

$$\ell_n(\theta \mid y, X) = -\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln(\sigma^2) - \frac{1}{2n\sigma^2}\sum_{j=1}^{n}(y_j - X_j\beta)^2$$

$$s_n(\theta \mid y, X) = \begin{bmatrix} \frac{1}{n\sigma^2}(X'y - X'X\beta) \\ -\frac{1}{2\sigma^2} + \frac{1}{2n\sigma^4}\sum_{j=1}^{n}(y_j - X_j\beta)^2 \end{bmatrix}$$

$$\hat{\theta}_n = \begin{bmatrix} \hat{\beta} \\ \widehat{\sigma^2} \end{bmatrix} = \begin{bmatrix} (X'X)^{-1}X'y \\ \frac{1}{n}\sum_{j=1}^{n}(y_j - X_j\hat{\beta})^2 \end{bmatrix}$$

$$-E\left[H_n(\theta) \mid X\right] = \begin{bmatrix} \frac{X'X}{n\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{bmatrix} = \begin{bmatrix} \sigma^2\left(\frac{X'X}{n}\right)^{-1} & 0 \\ 0 & 2\sigma^4 \end{bmatrix}^{-1} = \mathcal{I}(\theta \mid X)$$

Example 1: linear regression

$y_i \in \mathbb{R}$, $X_i \in \mathbb{R}^{1\times k}$, $\theta = [\beta, \sigma^2]$, $\beta \in \mathbb{R}^k, \sigma^2 > 0$,

$$y_i = X_i\beta + \epsilon_i,\ \epsilon_i \sim \mathcal{N}(0, \sigma^2),\ i = 1, \ldots, n \ \text{ iid}$$

Let X be $n \times k$, with X_i as i-th row. Similarly, let y be $n \times 1$.

$$\begin{aligned}
\ell_n(\theta \mid y, X) &= -\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln(\sigma^2) - \frac{1}{2n\sigma^2}\sum_{j=1}^{n}(y_j - X_j\beta)^2 \\
s_n(\theta \mid y, X) &= \begin{bmatrix} \frac{1}{n\sigma^2}(X'y - X'X\beta) \\ -\frac{1}{2\sigma^2} + \frac{1}{2n\sigma^4}\sum_{j=1}^{n}(y_j - X_j\beta)^2 \end{bmatrix} \\
\hat{\theta}_n &= \begin{bmatrix} \hat{\beta} \\ \widehat{\sigma^2} \end{bmatrix} = \begin{bmatrix} (X'X)^{-1}X'y \\ \frac{1}{n}\sum_{j=1}^{n}(y_j - X_j\hat{\beta})^2 \end{bmatrix} \\
-E\left[H_n(\theta) \mid X\right] &= \begin{bmatrix} \frac{X'X}{n\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{bmatrix} = \begin{bmatrix} \sigma^2\left(\frac{X'X}{n}\right)^{-1} & 0 \\ 0 & 2\sigma^4 \end{bmatrix}^{-1} = \mathcal{I}(\theta \mid X)
\end{aligned}$$

Example 2: binomial distribution

$y_i \in \{0,1\}, i = 1, \ldots, n$, with $P(y_i = 1 \mid \theta) = \theta$ iid. Observe: k "ones".

$$
\begin{aligned}
L(\theta) &= \theta^k (1-\theta)^{n-k} \\
\ell_n(\theta) &= \frac{1}{n} \sum_{i=1}^{n} \left(y_i \ln \theta + (1-y_i) \ln(1-\theta) \right) \\
&= \frac{k}{n} \ln \theta + \frac{n-k}{n} \ln(1-\theta) \\
s_n(\theta) &= \frac{k}{n} \frac{1}{\theta} - \frac{n-k}{n} \frac{1}{1-\theta} = \frac{\frac{k}{n} - \theta}{\theta(1-\theta)} \\
\hat{\theta}_n &= \frac{k}{n} \\
H_n(\theta) &= -\frac{k}{n} \frac{1}{\theta^2} - \frac{n-k}{n} \frac{1}{(1-\theta)^2} \\
-E[H_n(\theta)] &= \frac{1}{\theta} + \frac{1}{1-\theta} = (\theta(1-\theta))^{-1} = \mathcal{I}(\theta)
\end{aligned}
$$

Example 2: binomial distribution

$y_i \in \{0,1\}, i = 1,\ldots,n$, with $P(y_i = 1 \mid \theta) = \theta$ iid. Observe: k "ones".

$$
\begin{aligned}
L(\theta) &= \theta^k (1-\theta)^{n-k} \\
\ell_n(\theta) &= \frac{1}{n}\sum_{i=1}^{n} \left(y_i \ln\theta + (1-y_i)\ln(1-\theta)\right) \\
&= \frac{k}{n}\ln\theta + \frac{n-k}{n}\ln(1-\theta) \\
s_n(\theta) &= \frac{k}{n}\frac{1}{\theta} - \frac{n-k}{n}\frac{1}{1-\theta} = \frac{\frac{k}{n}-\theta}{\theta(1-\theta)} \\
\hat{\theta}_n &= \frac{k}{n} \\
H_n(\theta) &= -\frac{k}{n}\frac{1}{\theta^2} - \frac{n-k}{n}\frac{1}{(1-\theta)^2} \\
-E\left[H_n(\theta)\right] &= \frac{1}{\theta} + \frac{1}{1-\theta} = (\theta(1-\theta))^{-1} = \mathcal{I}(\theta)
\end{aligned}
$$

Example 3: binary choice.

Data $y_i \in \{0,1\}$, X_i. $P(y_i = 1 \mid X_i) = G(X\beta)$ iid. Density $g(v) = G'(v)$. $\theta = \beta$. Abbreviate $G_i = G(X_i\beta),\ g_i = g(X_i\beta)$. Define

Inverse Mills ratio or hazard: $\lambda(v) = \dfrac{g(v)}{1 - G(v)}$ (think: "$= \dfrac{P(V = v)}{P(V \geq v)}$")

$$
\begin{aligned}
\ell(\beta \mid y_i, X_i) &= y_i \ln G(X_i\beta) + (1 - y_i) \ln(1 - G(X_i\beta)) \\
&= y_i \ln G_i + (1 - y_i) \ln(1 - G(X_i\beta)) \\
s(\beta \mid X_i) &= \frac{y_i g_i X_i'}{G_i} - \frac{(1 - y_i) g_i X_i'}{1 - G_i} \\
&= \frac{(y_i - G_i) g_i X_i'}{G_i(1 - G_i)} \\
-E[H(\beta) \mid X_i] &= \frac{g_i^2 X_i' X_i}{G_i(1 - G_i)} = \left(\frac{G_i(1 - G_i)}{g_i^2 X_i' X_i} \right)^{-1} = \mathcal{I}(\beta \mid X_i) \\
&= \lambda(X_i\beta)\,\lambda(-X_i\beta)\,X_i' X_i \text{ ,if symmetry } g(v) = g(-v)
\end{aligned}
$$

Outline

Information Inequality

A statistic is a function $T : y \in Y \to \mathbb{R}^k$.

Theorem

Suppose $Var_\theta(T(y)) < \infty$ and $\mathcal{I}(\theta)$ is invertible. Let $\psi(\theta) = E_\theta[T(y)]$. Then,

$$Var_\theta(T(y)) \geq \left(\frac{\partial \psi(\theta)}{\partial \theta}\right) \mathcal{I}(\theta)^{-1} \left(\frac{\partial \psi(\theta)}{\partial \theta}\right)'$$

In particular:

Cramér-Rao lower bound:
Suppose that $E_\theta[T(y)] = \theta$,
i.e. $T(y)$ is an unbiased estimator of θ. Then
$Var_\theta(T(y)) \geq \mathcal{I}(\theta)^{-1}$

In words: MLE's are asymptotically efficient: they have minimal asymptotic variance among all asymptotically unbiased estimators.

Proof of the Information Inequality (scalar case)

Recall that

$$0 = E[s(\theta \mid y)] = \int \frac{\partial \ell(\theta \mid y)}{\partial \theta} f(y \mid \theta) \nu(dy) \tag{1}$$

Therefore

$$\begin{aligned}
\psi(\theta) &= \int T(y) f(y \mid \theta) \nu(dy) \\
\mid \psi'(\theta) \mid &= \mid \int T(y) \frac{\frac{\partial f(y \mid \theta)}{\partial \theta}}{f(y \mid \theta)} f(y \mid \theta) \nu(dy) \mid \\
per(1): &= \mid \int (T(y) - E_\theta[T(y)]) (s(\theta \mid y) - E[s(\theta \mid y)]) f(y \mid \theta) \nu(dy) \mid \\
&= \mid (Cov_\theta(T(y), s(\theta \mid y)) \mid \\
&\leq \sqrt{\text{Var}_\theta(T(y)) \, \text{Var}_\theta(s(\theta \mid y))}
\end{aligned}$$

or

$$\mid \psi'(\theta) \mid^2 \leq \text{Var}_\theta(T(y)) \mathcal{I}(\theta)$$

Proof of the Information Inequality (scalar case)

Recall that

$$0 = E[s(\theta \mid y)] = \int \frac{\partial \ell(\theta \mid y)}{\partial \theta} f(y \mid \theta)\nu(dy) \qquad (1)$$

Therefore

$$\begin{aligned}
\psi(\theta) &= \int T(y) f(y \mid \theta)\nu(dy) \\
\mid \psi'(\theta) \mid &= \mid \int T(y) \frac{\frac{\partial f(y \mid \theta)}{\partial \theta}}{f(y \mid \theta)} f(y \mid \theta)\nu(dy) \mid \\
per(1): &= \mid \int (T(y) - E_\theta[T(y)])\,(s(\theta \mid y) - E[s(\theta \mid y)])\, f(y \mid \theta)\nu(dy) \mid \\
&= \mid (Cov_\theta(T(y), s(\theta \mid y)) \mid \\
&\leq \sqrt{\mathrm{Var}_\theta(T(y))\,\mathrm{Var}_\theta(s(\theta \mid y))}
\end{aligned}$$

or

$$\mid \psi'(\theta) \mid^2 \leq \mathrm{Var}_\theta(T(y))\,\mathcal{I}(\theta)$$

Proof of the Information Inequality (scalar case)

Recall that

$$0 = E[s(\theta \mid y)] = \int \frac{\partial \ell(\theta \mid y)}{\partial \theta} f(y \mid \theta) \nu(dy) \qquad (1)$$

Therefore

$$\begin{aligned} \psi(\theta) &= \int T(y) f(y \mid \theta) \nu(dy) \\ \mid \psi'(\theta) \mid &= \mid \int T(y) \frac{\frac{\partial f(y \mid \theta)}{\partial \theta}}{f(y \mid \theta)} f(y \mid \theta) \nu(dy) \mid \\ per(1): &= \mid \int (T(y) - E_\theta[T(y)]) (s(\theta \mid y) - E[s(\theta \mid y)]) f(y \mid \theta) \nu(dy) \mid \\ &= \mid (Cov_\theta(T(y), s(\theta \mid y)) \mid \\ &\leq \sqrt{\mathrm{Var}_\theta(T(y)) \, \mathrm{Var}_\theta(s(\theta \mid y))} \end{aligned}$$

or

$$\mid \psi'(\theta) \mid^2 \leq \mathrm{Var}_\theta(T(y)) \mathcal{I}(\theta)$$

Proof of the Information Inequality (scalar case)

Recall that

$$0 = E[s(\theta \mid y)] = \int \frac{\partial \ell(\theta \mid y)}{\partial \theta} f(y \mid \theta)\nu(dy) \qquad (1)$$

Therefore

$$\begin{aligned}
\psi(\theta) &= \int T(y) f(y \mid \theta)\nu(dy) \\
\mid \psi'(\theta) \mid &= \mid \int T(y) \frac{\frac{\partial f(y|\theta)}{\partial \theta}}{f(y \mid \theta)} f(y \mid \theta)\nu(dy) \mid \\
per(1): &= \mid \int (T(y) - E_\theta[T(y)])\,(s(\theta \mid y) - E[s(\theta \mid y)])\, f(y \mid \theta)\nu(dy) \mid \\
&= \mid (Cov_\theta(T(y), s(\theta \mid y)) \mid \\
&\leq \sqrt{\text{Var}_\theta(T(y))\,\text{Var}_\theta(s(\theta \mid y))}
\end{aligned}$$

or

$$\mid \psi'(\theta) \mid^2 \leq \text{Var}_\theta(T(y))\,\mathcal{I}(\theta)$$

Outline

Identification

- With enough data, will there eventually be a unique θ or will several θ work as well, even asymptotically? Or: is θ identified?
- Definition: θ is identified, if there is no other $\tilde{\theta}$ with $L(\theta \mid y) = L(\tilde{\theta} \mid y)$ for all y.
- Example for lack of identification:
  - Suppose, some function $\vartheta : x \in (a, b) \rightarrow \theta = \vartheta(x)$ results in constant log-likelihood $\ell(\vartheta(x) \mid y)$, for all y.
  - ... therefore, $0 = s(\vartheta(x) \mid y)v(x)$, where $v(x) = \frac{d\vartheta(x)}{dx}$.
  - ... therefore

$$0 = v(x)' \frac{\partial s(\vartheta(x) \mid y)}{\partial \theta'} v(x) + s(\vartheta(x) \mid y) \frac{d^2 \vartheta(x)}{dx^2}$$

  - Take expectations:

$$0 = v(x)' \mathcal{I}(\vartheta(x)) v(x)$$

Lack of identification

If $\mathcal{I}(\theta)$ is not invertible,
θ may not be identified.

Outline

A framework for testing

- Hypothesis: $\theta \in \Theta_0$. Alternative: $\theta \in \Theta_1$.
- Observations y.
- Test: A decision $\delta(y) \in 0, 1$. $\delta(y) = 0$: "accept". $\delta(y) = 1$: "reject".
- Power function: $\beta(\theta, \delta) = P_\theta(\delta(y) = 1)$. Probability of rejection, if θ is true.
- Error of Type I: reject, even though $\theta \in \Theta_0$.
- Size of the test: $\sup_{\theta \in \Theta_0} \beta(\theta, \delta)$. Maximal probability of rejecting true hypothesis, i.e. max. prob. of type-I error. "5 % significance".
- Error of Type II: accept, even though $\theta \in \Theta_1$.

Likelihood-ratio test: a special case

- Hypothesis $\theta = \theta_0$. Alternative: $\theta = \theta_1$.
- Likelihood ratio:

$$LR^*(\theta_1, \theta_0 \mid y) = \frac{L(\theta_1 \mid y)}{L(\theta_0 \mid y)}$$

- Reject, if likelihood ratio exceeds some threshold Ψ,

$$\delta_\Psi^{(LR)}(y) = \begin{cases} 1 & \text{if } LR^*(\theta_1, \theta_0 \mid y) \geq \Psi \\ 0 & \text{if } LR^*(\theta_1, \theta_0 \mid y) < \Psi \end{cases}$$

The Neyman-Pearson-Lemma

Theorem

Consider testing $\theta = \theta_0$ vs $\theta = \theta_1$. Given Ψ, let $\delta(y)$ be some test with size equal or smaller than the size of the test $\delta_\Psi^{(LR)}$. Then, the power at the alternative is smaller too,

$$\beta(\delta, \theta_0) \leq \beta(\delta_\Psi^{(LR)}, \theta_0) \textit{ implies } \beta(\delta, \theta_1) \leq \beta(\delta_\Psi^{(LR)}, \theta_1)$$

In words: test δ that are as least as "careful" in avoiding type-I errors as a likelihood-ratio test will make at least as many type-II errors as a likelihood-ratio test.

Likelihood ratio tests are nice!

The Neyman-Pearson-Lemma

Theorem

Consider testing $\theta = \theta_0$ vs $\theta = \theta_1$. Given Ψ, let $\delta(y)$ be some test with size equal or smaller than the size of the test $\delta_\Psi^{(LR)}$. Then, the power at the alternative is smaller too,

$$\beta(\delta, \theta_0) \leq \beta(\delta_\Psi^{(LR)}, \theta_0) \textit{ implies } \beta(\delta, \theta_1) \leq \beta(\delta_\Psi^{(LR)}, \theta_1)$$

In words: test δ that are as least as “careful” in avoiding type-I errors as a likelihood-ratio test will make at least as many type-II errors as a likelihood-ratio test.

Likelihood ratio tests are nice!

The Neyman-Pearson-Lemma

Theorem

Consider testing $\theta = \theta_0$ vs $\theta = \theta_1$. Given Ψ, let $\delta(y)$ be some test with size equal or smaller than the size of the test $\delta_\Psi^{(LR)}$. Then, the power at the alternative is smaller too,

$$\beta(\delta, \theta_0) \leq \beta(\delta_\Psi^{(LR)}, \theta_0) \text{ implies } \beta(\delta, \theta_1) \leq \beta(\delta_\Psi^{(LR)}, \theta_1)$$

In words: test δ that are as least as "careful" in avoiding type-I errors as a likelihood-ratio test will make at least as many type-II errors as a likelihood-ratio test.

Likelihood ratio tests are nice!

Likelihood ratio tests: general case

- Hypothesis $\theta \in \Theta_0$. Alternative: $\theta \in \Theta_1$.
- Likelihood ratio: typically written in terms of 2*log-likelihood

$$LR(\Theta_1, \Theta_0 \mid y) = 2(\sup_{\theta_1 \in \Theta_1} \ell(\theta_1 \mid y) - \sup_{\theta_0 \in \Theta_0} \ell(\theta_0 \mid y)) = 2\log(LR^*(\Theta_1, \Theta_0 \mid y))$$

- Reject, if likelihood ratio exceeds some threshold $\psi = 2\log(\Psi)$,

$$\delta_\Psi^{(LR)}(y) = \begin{cases} 1 & \text{if } LR(\Theta_1, \Theta_0 \mid y) \geq \psi \\ 0 & \text{if } LR(\Theta_1, \Theta_0 \mid y) < \psi \end{cases}$$

- Suppose $\sup L = \max L$ on Θ_0. Suppose $\sup_{\theta_1 \in \Theta_1} L = \max_{\theta \in \Theta} L$. Let $\hat{\theta}_c$ be the constrained MLE on Θ_0. Let $\hat{\theta}$ be unconstrained MLE. Then

$$\begin{aligned} LR(\Theta_1, \Theta_0 \mid y) &= 2(\ell(\hat{\theta} \mid y) - \ell(\hat{\theta}_c \mid y)) \\ &= 2\ln\left(\frac{L(\hat{\theta} \mid y)}{L(\hat{\theta}_c \mid y)}\right) \end{aligned}$$

Outline

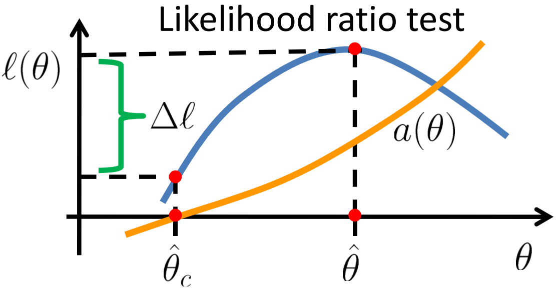
Testing a parametric constraint

- $\theta \in \mathbb{R}^m$ (or open subset)
- Constraint $a : \mathbb{R}^m \to \mathbb{R}^k$, differentiable. Derivative: rank k.
- Or: $a : \mathbb{R}^m \to \mathbb{R}^l$, differentiable, $l > k$, but derivative has rank k.
- Hypothesis: k constraints hold,

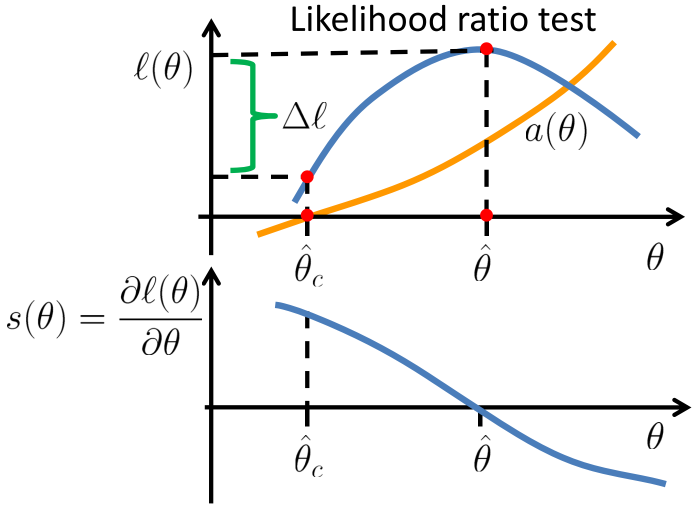
$$H_0 : a(\theta) = 0$$

- $\hat{\theta}$ or $\hat{\theta}_n$: unconstrained MLE.
- $\hat{\theta}_c$ or $\hat{\theta}_{c,n}$: constrained MLE, satisfies $a(\hat{\theta}_c) = 0$.
- Some material: Hayashi, Econometrics, Princeton Univ. Press (2000).

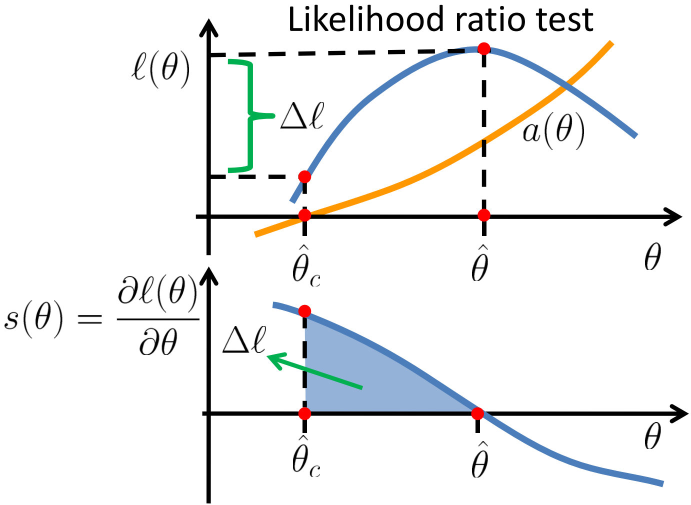
LR, Score and Wald



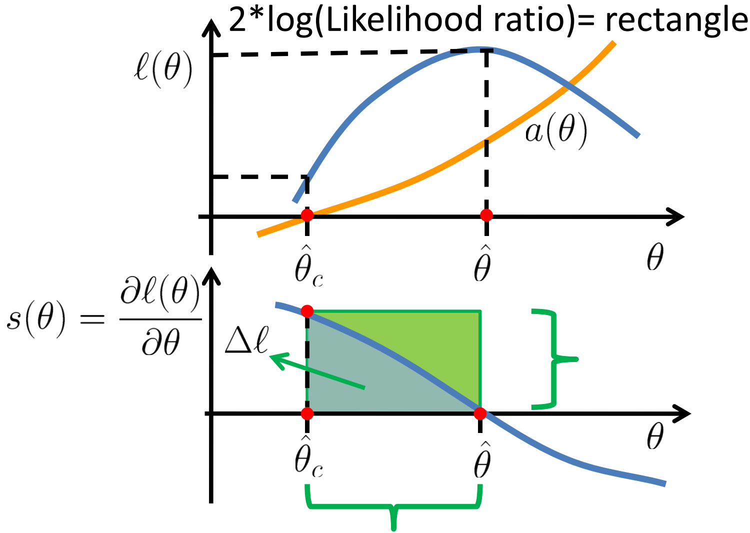
LR, Score and Wald



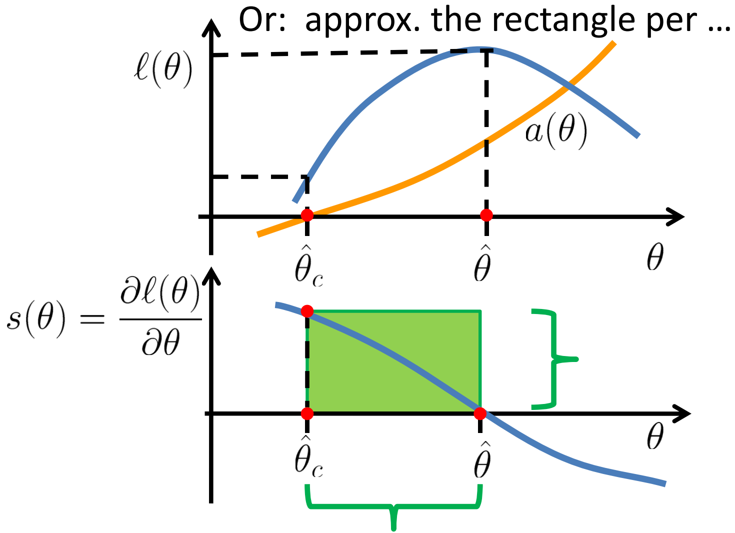
LR, Score and Wald



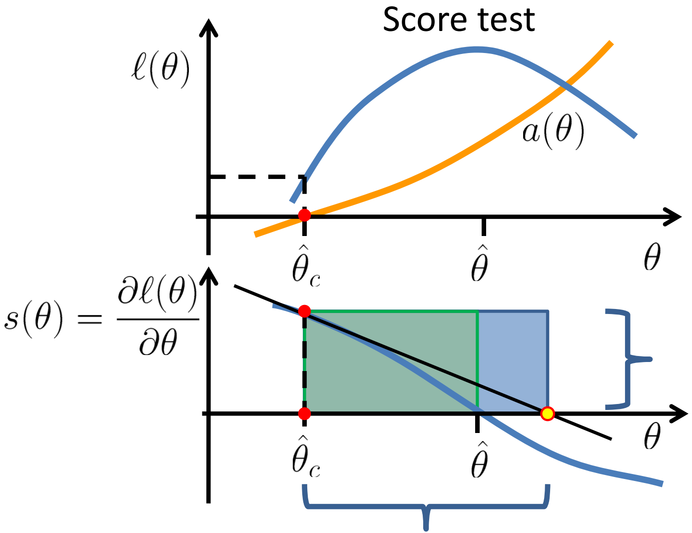
LR, Score and Wald



LR, Score and Wald

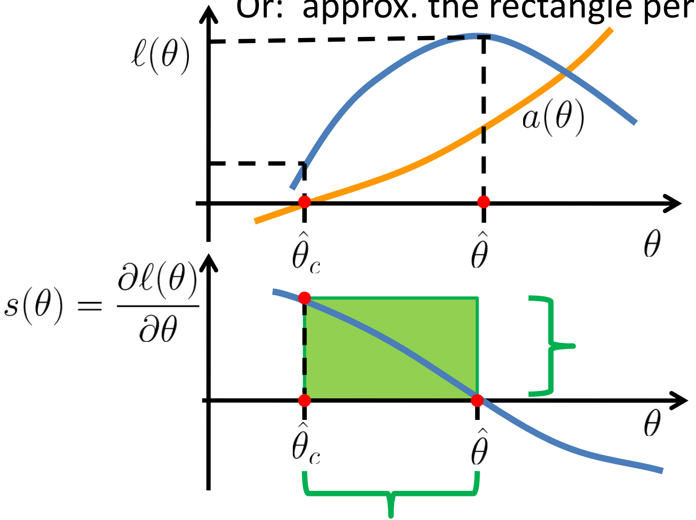


LR, Score and Wald

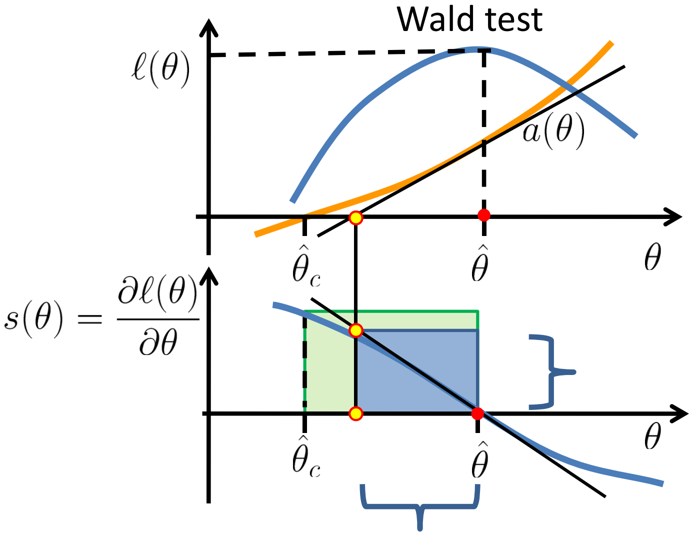


LR, Score and Wald

Or: approx. the rectangle per …



LR, Score and Wald



Three tests:

$$\textbf{rectangle} = s(\hat{\theta}_c)'(\hat{\theta} - \hat{\theta}_c)$$

1. Likelihood-ratio test:

$$\textbf{rectangle} \approx LR = 2 * (\ell(\hat{\theta}) - \ell(\hat{\theta}_c))$$

2. Score test or Lagrange multiplier test or Rao test:

$$\begin{aligned} \textbf{rectangle} &\approx s(\hat{\theta}_c)'\mathcal{I}(\hat{\theta}_c)^{-1}s(\hat{\theta}_c) \\ \text{per: } -s(\hat{\theta}_c) &= s(\hat{\theta}) - s(\hat{\theta}_c) \approx -\mathcal{I}(\hat{\theta}_c)(\hat{\theta} - \hat{\theta}_c) \end{aligned}$$

3. Wald test: [Remark: invertibility of $\partial a / \partial \theta$?]

$$\begin{aligned} \textbf{rectangle} &\approx a(\hat{\theta})' \left(\frac{\partial a(\hat{\theta})}{\partial \theta} \mathcal{I}(\hat{\theta})^{-1} \left(\frac{\partial a(\hat{\theta})}{\partial \theta} \right)' \right)^{-1} a(\hat{\theta}) \\ \text{per: } -a(\hat{\theta}) &= a(\hat{\theta}_c) - a(\hat{\theta}) \approx \frac{\partial a(\hat{\theta})}{\partial \theta}(\hat{\theta}_c - \hat{\theta}) \\ s(\hat{\theta}_c) &= s(\hat{\theta}_c) - s(\hat{\theta}) \approx -\mathcal{I}(\hat{\theta})(\hat{\theta}_c - \hat{\theta}) \end{aligned}$$

3. Wald statistic. Asymptotics.

- Truth: θ_0. Under H_0: $a(\theta_0) = 0$.
- Define

$$A(\theta) = \frac{\partial a(\theta)}{\partial \theta}, \ \ A = A(\theta_0), \ \ \mathcal{I} = \mathcal{I}(\theta_0)$$

- MLE asymptotics:

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

- Delta method:

$$\sqrt{n}(a(\hat{\theta}_n) - a(\theta_0)) \xrightarrow{d} \mathcal{N}\left(0, A\,\mathcal{I}^{-1} A'\right)$$

- Wald statistic:

$$W = n\, a(\hat{\theta}_n)' \left(A(\hat{\theta}_n)\,\hat{\mathcal{I}}_n^{-1} A(\hat{\theta}_n)'\right)^{-1} a(\hat{\theta}_n) \xrightarrow{d} \chi^2_k$$

2. Score test or Lagrange multiplier test. Asymptotics.

- Truth: θ_0. Maximize $\ell_n(\theta)$ s.t. $a(\theta) = 0$. Lagrangian. FONC:

$$s_n(\hat{\theta}_{c,n}) - A(\hat{\theta}_{c,n})' \lambda_n = 0$$

- Solve for $\lambda_n \in \mathbb{R}^k$: pre-multiply with some $B_n \xrightarrow{P} B$, size $k \times m$:

$$\begin{aligned} \sqrt{n}\,\lambda_n &= \sqrt{n}\left(B_n A(\hat{\theta}_{c,n})'\right)^{-1} B_n\, s_n(\hat{\theta}_{c,n}) \\ &\xrightarrow{d} \mathcal{N}\left(0, (B\,A')^{-1} B\,\mathcal{I}\,B'(A\,B')^{-1}\right) \end{aligned}$$

- A clever choice: $B = A\,\mathcal{I}^{-1}$ and $B_n = A(\hat{\theta}_{c,n})\hat{\mathcal{I}}_n^{-1}$. Then

$$\sqrt{n}\,\lambda_n \xrightarrow{d} \mathcal{N}\left(0, \left(A\,\mathcal{I}^{-1} A'\right)^{-1}\right)$$

- Lagrange multiplier statistic: [Remark: not $s_n' \hat{\mathcal{I}}_n^{-1} s_n \xrightarrow{d} \chi^2_m$]

$$LM = n\, s_n(\hat{\theta}_{c,n})' \,\hat{\mathcal{I}}_n^{-1}\, s_n(\hat{\theta}_{c,n}) = n\,\lambda_n' A(\hat{\theta}_{c,n})\,\hat{\mathcal{I}}_n^{-1} A(\hat{\theta}_{c,n})'\,\lambda_n \xrightarrow{d} \chi^2_k$$

2. Score test or Lagrange multiplier test. Asymptotics.

- Truth: θ_0. Maximize $\ell_n(\theta)$ s.t. $a(\theta) = 0$. Lagrangian. FONC:

$$s_n(\hat{\theta}_{c,n}) - A(\hat{\theta}_{c,n})'\lambda_n = 0$$

- Solve for $\lambda_n \in \mathbb{R}^k$: pre-multiply with some $B_n \xrightarrow{P} B$, size $k \times m$:

$$\begin{aligned} \sqrt{n}\,\lambda_n &= \sqrt{n}\left(B_n A(\hat{\theta}_{c,n})'\right)^{-1} B_n\, s_n(\hat{\theta}_{c,n}) \\ &\xrightarrow{d} \mathcal{N}\left(0, (B\,A')^{-1} B\,\mathcal{I}\,B'(A\,B')^{-1}\right) \end{aligned}$$

- A clever choice: $B = A\,\mathcal{I}^{-1}$ and $B_n = A(\hat{\theta}_{c,n})\hat{\mathcal{I}}_n^{-1}$. Then

$$\sqrt{n}\,\lambda_n \xrightarrow{d} \mathcal{N}\left(0, \left(A\,\mathcal{I}^{-1}A'\right)^{-1}\right)$$

- Lagrange multiplier statistic: [Remark: not $s_n'\hat{\mathcal{I}}_n^{-1}s_n \xrightarrow{d} \chi^2_m$]

$$LM = n\,s_n(\hat{\theta}_{c,n})'\,\hat{\mathcal{I}}_n^{-1}\,s_n(\hat{\theta}_{c,n}) = n\,\lambda_n'\,A(\hat{\theta}_{c,n})\,\hat{\mathcal{I}}_n^{-1}\,A(\hat{\theta}_{c,n})'\,\lambda_n \xrightarrow{d} \chi^2_k$$

1. Likelihood ratio test. Asymptotics.

- Truth: θ_0. Second-order approximation around $\hat{\theta}_n$:

$$\begin{aligned} LR &= 2n\,(\ell_n(\hat{\theta}_n) - \ell_n(\hat{\theta}_{c,n})) \\ &= n\,(\hat{\theta}_n - \hat{\theta}_{c,n})'\,\hat{\mathcal{I}}_n^{(2)}\,(\hat{\theta}_n - \hat{\theta}_{c,n}) + o_P \end{aligned}$$

- From calculation on score statistic plus first-order approx. of score:

$$A(\hat{\theta}_{c,n})'\,\lambda_n = s_n(\hat{\theta}_{c,n}) = \hat{\mathcal{I}}_n^{(2)}(\hat{\theta}_n - \hat{\theta}_{c,n}) + \frac{o_P}{n}$$

  Solve for $(\hat{\theta}_n - \hat{\theta}_{c,n})$.

- Therefore

$$LR = n\,\lambda_n' A(\hat{\theta}_{c,n}) \left(\hat{\mathcal{I}}_n^{(2)}\right)^{-1} A(\hat{\theta}_{c,n})'\,\lambda_n + o_P \xrightarrow{d} \chi_k^2$$

  per result about Lagrange multiplier statistic.