ECON 6200: Section 10

4/9/25

Agenda

- Review Extremum Estimators
  - ↳ Consistency
- Asymptotic Distribution (proof)
- Identification
- MLE Asymptotic Distribution
- Hypothesis Testing

(Review)

Extremum Estimation

An extremum estimator is any estimator defined as

$$\hat{\theta} = \operatorname{argmin} Q_n(w_1, \ldots, w_n; \theta)$$

for some parameter θ in parameter space (Θ)
and $w_1, \ldots, w_n$ is the sample

- $Q_n(\cdot)$ = criterion / objective function (sample)
  - ↳ indexed by n because it must depend on your sample
  - ↳ $Q(\cdot)$ = population analog of $Q_n(\cdot)$

[Consistency Thm (A)]

Theorem

Assume:

1. The sample criterion uniformly consistently estimates the population criterion:

$$\sup_{\theta \in \Theta} |Q_n(\theta) - Q(\theta)| \xrightarrow{p} 0.$$

2. θ_0 is a unique and well-separated global minimum of $Q(\cdot)$:

$$\forall \epsilon > 0 \exists \delta > 0 : Q^{\epsilon} \equiv \inf_{\theta \in \Theta : \|\theta - \theta_0\| \geq \epsilon} Q(\theta) \geq Q(\theta_0) + \delta.$$

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

Consistency Thm (B)

Consolidated Theorem

Assume that:

1. $Q(\cdot)$ is continuous,
2. Θ is compact,
3. θ_0 uniquely minimizes $Q(\theta)$,
4. $\sup_{\theta \in \Theta} |Q_n(\theta) - Q(\theta)| \xrightarrow{p} 0$.

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

Consistency Thm (C)

Consistency for Convex $Q(\cdot)$

Assume that:

1. Θ is convex,
2. $\theta_0 \in \operatorname{int} \Theta$,
3. θ_0 uniquely minimizes $Q(\theta)$,
4. $Q_n(\cdot)$ is convex,
5. $|Q_n(\theta) - Q(\theta)| \xrightarrow{p} 0, \forall \theta \in \Theta$.

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

Asymptotic Distribution

Theorem: Asymptotic Distribution

Assume that:

1. $\hat{\theta} \xrightarrow{p} \theta_0$,
2. $\theta_0 \in \text{int}(\Theta)$,
3. $Q_n(\cdot)$ is twice continuously differentiable in an open neighborhood $\mathcal{N}$ of θ_0,
4. $\sqrt{n}\frac{dQ_n(\theta_0)}{d\theta} \xrightarrow{d} N(0, \Sigma)$,
5. $\sup_{\theta \in \mathcal{N}} \left\| \frac{dQ_n(\theta)^2}{d\theta d\theta'} - \frac{dQ(\theta)^2}{d\theta d\theta'} \right\| \xrightarrow{p} 0$,
6. $\boldsymbol{H} \equiv \frac{dQ(\theta_0)^2}{d\theta d\theta'}$ is nonsingular.

Then

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, \boldsymbol{H}^{-1}\Sigma\boldsymbol{H}^{-1}).$$

Note: Since we assume $\hat{\theta} \xrightarrow{p} \theta_0$, all the assumptions needed to assume consistency are are applicable for the asymptotic distribution as well

Also, since we have assumed consistency, we are mainly focused on $Q(\cdot)$ around θ_0.

Mean Value Thm

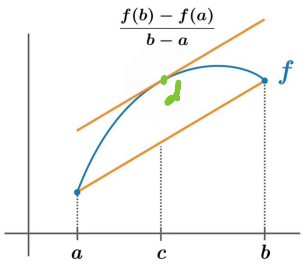
The Mean Value Theorem

Suppose f is continuous on $[a, b]$ and differentiable on (a, b). Then there exists some argument c such that

$$a < c < b$$

and

$$f'(c) = \frac{f(b) - f(a)}{b - a}$$



Proof

PART (1) :

4. $\sqrt{n}\frac{dQ_n(\theta_0)}{d\theta} \xrightarrow{d} N(0, \Sigma)$,

Issue: Don't know behavior of sample $Q_n(\cdot)$ w/ true θ_0

Since $\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmin}}\, Q_n(\theta)$, and $\hat{\theta} \xrightarrow{p} \theta_0$, $\theta_0 \in \operatorname{int} \Theta$,
$Q_n(\cdot)$ twice cont. diff. around θ

then w/ probability approaching 1,

$$\frac{d\,Q_n(\hat{\theta})}{d\theta} = 0$$

Recall MVT: $f'(c) = \frac{f(b)-f(a)}{b-a}$

$\Rightarrow f(b) = f(a) + f'(c)(b-a)$

By MVT and (A.3), w/ probability approaching 1

$$\frac{dQ_n(\hat{\theta})}{d\theta} = \frac{dQ_n(\theta_0)}{d\theta} + \frac{d^2 Q_n(\bar{\theta})}{d\theta d\theta'}(\hat{\theta}-\theta_0) = 0$$

for some $\bar{\theta}$ between θ_0 and $\hat{\theta}$.

Since $\bar{\theta}$ between $\hat{\theta}$, θ_0 and $\hat{\theta} \xrightarrow{p} \theta_0$

$\Rightarrow \bar{\theta} \xrightarrow{p} \theta_0$

Now, performing some algebraic manipulation to get to some CLT form:

$$\frac{dQ_n(\hat{\theta})}{d\theta} = \frac{dQ_n(\theta_0)}{d\theta} + \frac{dQ_n(\bar{\theta})^2}{d\theta d\theta'}(\hat{\theta} - \theta_0)$$

$$0 = \left(\frac{dQ_n(\theta_0)}{d\theta} + \frac{dQ_n(\bar{\theta})^2}{d\theta d\theta'}(\hat{\theta} - \theta_0) \right) \sqrt{n}$$

$$\sqrt{n}(\hat{\theta} - \theta_0) = -\left(\frac{dQ_n(\bar{\theta})^2}{d\theta d\theta'} \right)^{-1} \sqrt{n} \frac{dQ_n(\theta_0)}{d\theta}$$

$$\left[\text{PART ②}: \underbrace{\frac{dQ_n(\bar{\theta})^2}{d\theta\, d\theta'}}_{(H_n(\bar{\theta}))} \xrightarrow{P} H\right]$$

Clearly, this looks like (A.5) ❺ $\sup_{\theta\in\mathcal{N}} \left\| \frac{dQ_n(\theta)^2}{d\theta d\theta'} - \frac{dQ(\theta)^2}{d\theta d\theta'} \right\| \xrightarrow{P} 0,$

Let $H = \frac{dQ(\theta)^2}{d\theta\, d\theta'}$, $H_n(\theta) = \frac{dQ_n(\bar{\theta})^2}{d\theta\, d\theta'}$

$$\|H_n(\bar{\theta}) - H\| = \|H_n(\bar{\theta}) - H(\bar{\theta}) + H(\bar{\theta}) - H\|$$

$$\leq \|H_n(\bar{\theta}) - H(\bar{\theta})\| + \|H(\bar{\theta}) - H\|$$

$$\leq \sup_{\theta\in\mathcal{N}} \|H_n(\bar{\theta}) - H(\theta)\| + \|H(\bar{\theta}) - H\|$$

$$\xrightarrow{P} 0$$

Hence, $\frac{dQ_n(\bar{\theta})^2}{d\theta\, d\theta'} \xrightarrow{P} H$

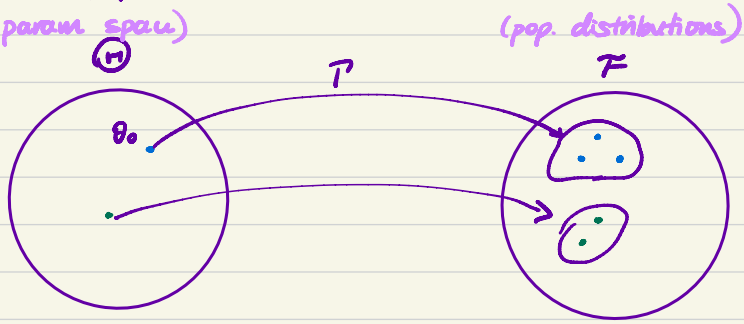
Thus, $\sqrt{n}(\hat{\theta} - \theta_0) = -\left(\frac{dQ_n(\bar{\theta})^2}{d\theta\, d\theta'}\right)^{-1} \sqrt{n}\, \frac{dQ_n(\theta_0)}{d\theta}$

$$\Rightarrow \sqrt{n}(\hat{\theta} - \theta_0) \longrightarrow N(0, H^{-1}\Sigma H^{-1})$$

check slides for other specifications

Identification

If we know the population distribution of the data, can we back out θ_0?

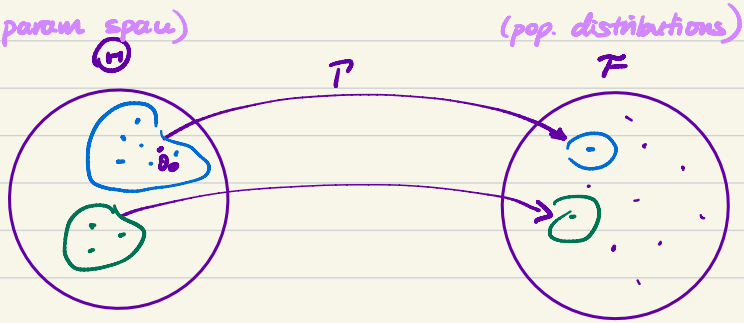


$T: \Theta \to \mathcal{F}$ (a correspondence that maps each parameter value to its corresponding distribution)

θ_0 is **identified** if the true population distribution $F \in T(\theta_0)$ implies $T^{-1}(F) = \{\theta_0\}$

↳ ie: Given a population distribution, you can find a unique value of θ_0

θ_0 is **partially identified** if $T^{-1}(F)$ is a set



ie: Given a population distribution, I can give you potential θ candidates, but I can't tell which value of θ is true

θ_0 is overidentified if there is an excess of information, which leaves us with testable assumption

↳ testable assumption are assumptions that can be observed and compared against data

- if data contradicts assumptions $\Rightarrow$ assumption wrong
- if data consistent $\Rightarrow$ didn't disprove assumption

Ex 1 Let $Y = h_0(X) + \varepsilon$ and assume ① $\varepsilon \perp X$
② $h_0 : \mathbb{R}^d \to \mathbb{R}$ cont function
③ X has support $\mathbb{R}^d$

Param of interest $\theta_0 = (h_0, F_{\varepsilon_0})$

(Ex: OLS w/ normal error where
h_0 = OLS coef
F_{ε_0} = dist w/ $E[\varepsilon]$, $\text{var}(\varepsilon)$)

Parameter space is $\mathcal{H} = \{$ set of pairs (h, F) s.t.
$h : \mathbb{R}^d \to \mathbb{R}$ cont,
F: cdf that doesn't depend on $x\}$

Question: Is θ_0 identified in $\mathcal{H}$?

↳ NO!

Claim: Let $h(x) = h_0(x) - \alpha$, $F_\varepsilon(t) = F_{\varepsilon_0}(t-\alpha)$

The pair $(h(x), F_\varepsilon(t))$ gives the same distribution of the data as $(h_0(x), F_{\varepsilon_0}(x))$

$$P(Y \le y, X \le x; h, F_\varepsilon) = \int_X P(h(w) + \varepsilon \le y \mid X = w; h, F_\varepsilon) f_x(w)\, dw$$

$$= \int_X P(\varepsilon \le y - h(w) \mid X = w; h, F_\varepsilon) f_x(w)\, dw$$

$$= \int_X F_\varepsilon(y - h(w))\, f_x(w)\, dw$$

$$= \int_X F_{\varepsilon_0}\Big(\big[y - \underbrace{(h_0(w) - \alpha)}_{= h(w)}\big] - \alpha\Big) f_x(w)\, dw$$

$$= \int_X F_{\varepsilon_0}(y - h_0(w))\, f_x(w)\, dw$$

$$= \int_X P(\varepsilon \le y - h_0(w) \mid X = w; h_0, F_{\varepsilon_0})\, f_x(w)\, dw$$

$$= P(Y \le y, X \le x; h_0, F_{\varepsilon_0})$$

This example fails the "location restriction"

Ex 2. Let $Y = h_0(X) + \varepsilon$ and assume ① $\varepsilon \perp X$
② $h_0 : \mathbb{R}^d \to \mathbb{R}$ cont function
③ X has support $\mathbb{R}^d$
④ $E[\varepsilon] = 0$

Param of interest $\theta_0 = (h_0, F_{\varepsilon_0})$

Parameter space is $\mathcal{M} = \{$ set of pairs (h, F) s.t.
$h : \mathbb{R}^d \to \mathbb{R}$ cont,
F : cdf that doesn't depend on x $\}$
and $E_F[\varepsilon] = 0$

Question: Is θ_0 identified in $\mathcal{M}$?

↳ YES!

Maximum Likelihood

Recall MLE is a special case of Extremum Estimators where

$$\begin{aligned} Q(\theta) &= \mathbb{E}\ell(W;\theta) \\ Q_n(\theta) &= \mathbb{E}_n\ell(W;\theta). \end{aligned}$$

Thm: Identification

Consistency of MI follows from the m-estimator consistency result above. Importantly, we can relate the identification assumption

$$\theta_0 \text{ uniquely maximizes } Q(\cdot)$$

to the likelihood identification condition

$$\theta \neq \theta_0 \implies \Pr(f(W;\theta) \neq f(W;\theta_0)) > 0.$$

Other Useful Things to Know:

- **Score Equation**

$$E\left[\frac{\partial \log f(w;\theta_0)}{\partial \theta}\right] = 0$$

  - basically an FOC condition under the log-likelihood

$$
\begin{aligned}
\int f(w;\theta_0)\,dw &= 1 \\
\implies \int \frac{\partial f(w;\theta_0)}{\partial \theta}\,dw &= 0 \\
\implies \int \frac{\partial \log f(w;\theta_0)}{\partial \theta} f(w;\theta_0)\,dw &= 0 \\
\implies \mathbb{E}\left(\frac{\partial \log f(w;\theta_0)}{\partial \theta}\right) &= 0.
\end{aligned}
$$

- **Information Matrix**

$$
\underbrace{E\left[\frac{\partial^2 \log f(w;\theta_0)}{\partial\theta\,\partial\theta'}\right]}_{\text{Fisher information } H} = -E\left[\frac{\partial \log f(w;\theta_0)}{\partial\theta}\,\frac{\partial \log f(w;\theta_0)}{\partial\theta'}\right]
$$

In class, we showed that MLE asymptotic distribution to be

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

through the Extremum Estimator asymptotic distribution theorem.

However, we can also derive the MLE asymptotic distribution through typical FOC procedure

Consider the linear regression model

$$Y = X'\beta + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2), \quad \varepsilon \perp X$$

Denote the true parameters as $\theta_0 = (\beta_0, \sigma_0^2)$

$$\Rightarrow f(y|X, \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{Y - X'\beta}{\sigma}\right)^2}$$

Then the conditional log likelihood is

$$\log f(Y|X, \theta) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(Y - X'\beta)^2$$

The ML estimator $\hat{\theta} = (\hat{\beta}, \hat{\sigma}^2)$ solves

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmax}}\ Q_n(\beta, \sigma^2)$$
$$= \underset{\theta \in \Theta}{\operatorname{argmax}} \sum_{i=1}^{n} \log f(y_i | x_i, \theta)$$

1) Solve for $\hat{\beta}$

$$\hat{\beta} = \underset{\beta}{\operatorname{argmax}} -\sum_{i=1}^{n} (y_i - x_i'\beta)^2$$
$$= \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - x_i'\beta)^2$$

minimizing MSE!

Note that this is just the OLS estimator,

$$\hat{\beta} = \left(\sum_i x_i x_i'\right)^{-1} \left(\sum_i x_i y_i\right)$$

2) Solve for $\hat{\sigma}^2$

$$\hat{\sigma}^2 = \underset{\sigma^2}{\operatorname{argmax}} \sum_{i}^{n} \left(-\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(y_i - x_i'\hat{\beta})^2\right)$$
$$= \underset{\sigma^2}{\operatorname{argmin}}\ n\log\sigma^2 + \frac{1}{\sigma^2}\sum_{i}^{n}(y_i - x_i'\hat{\beta})^2$$

FOC $\quad \frac{n}{\sigma^2} - \frac{1}{(\sigma^2)^2}\sum (y_i - x_i'\hat{\beta})^2 = 0$

$$\Rightarrow \hat{\sigma}^2 = \frac{1}{n}\sum_{i}^{n}(y_i - x_i'\beta)^2$$

Now, to show $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, -H^{-1})$:

3) Consider population criterion function:

$$Q(\theta) = E\left[-\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(Y - X'\beta)^2\right]$$

$$\frac{\partial Q(\theta)}{\partial \theta} = \begin{bmatrix} \frac{\partial Q(\theta)}{\partial \beta} \\ \frac{\partial Q(\theta)}{\partial \sigma^2} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma^2} E[X(Y - X'\beta)] \\ -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4} E[(Y - X'\beta)^2] \end{bmatrix}$$

$$\frac{\partial^2 Q(\theta)}{\partial \theta \partial \theta'} = \left[\begin{array}{c|c} \frac{\partial^2 Q(\theta)}{\partial \beta \partial \beta'} & \frac{\partial^2 Q(\theta)}{\partial \beta \partial \sigma^2} \\ \hline \frac{\partial^2 Q(\theta)}{\partial \sigma^2 \partial \beta} & \frac{\partial^2 Q(\theta)}{\partial \sigma^2 \partial \sigma^2} \end{array}\right] = \left[\begin{array}{c|c} -\frac{1}{\sigma^2} E[XX'] & -\frac{1}{\sigma^4} E[X(Y - X'\beta)] \\ \hline \frac{1}{\sigma^4} E[X(Y - X'\beta)] & \frac{1}{2\sigma^4} - \frac{1}{\sigma^6} E[(Y - X\beta)^2] \end{array}\right]$$

Note: Why is $\frac{\partial Q(\theta_0)}{\partial \theta} = 0$?

Intuitively, FOC = 0 at max and θ_0 is the argmax for true $Q(\cdot)$

(You can also do the math as well)

4) To find the asymptotic variance of ML estimator, evaluate $\frac{\partial Q(\cdot)}{\partial \theta}$ at θ_0.

$$\frac{\partial^2 Q(\theta_0)}{\partial \theta \partial \theta'} = \begin{bmatrix} -\frac{1}{\sigma^2} E[XX'] & -\frac{1}{\sigma^4} E[X(y - X'\beta)] \\ \frac{1}{\sigma^4} E[X(y - X'\beta)] & \frac{1}{2\sigma^4} - \frac{1}{\sigma^6} E[(y - X\beta)^2] \end{bmatrix}$$

($-\frac{1}{\sigma^4}E[X(y-X'\beta)] = 0$; $\frac{1}{2\sigma^4} - \frac{1}{\sigma^6}E[(y-X\beta)^2] = -\frac{1}{\sigma_0^4}$ since $E[(y-X\beta)^2] = \sigma_0^2$)

$$= \begin{bmatrix} -\frac{1}{\sigma_0^2} E[XX'] & 0 \\ 0 & -\frac{1}{2\sigma_0^4} \end{bmatrix} = H$$

$$\Rightarrow \quad -H^{-1} = \left(\frac{1}{\sigma_0^2} E[XX']\right)^{-1} \begin{bmatrix} \frac{1}{2\sigma_0^4} & 0 \\ 0 & \frac{1}{\sigma_0^2} E[XX'] \end{bmatrix} = \begin{bmatrix} \sigma^2 E[XX']^{-1} & 0 \\ 0 & 2\sigma^4 \end{bmatrix}$$

Recall

Information Matrix

$$\underbrace{E\left[\frac{\partial^2 \log f(w;\theta_0)}{\partial \theta \partial \theta'}\right]}_{\text{Fisher information}} = -E\left[\frac{\partial \log f(w;\theta_0)}{\partial \theta} \frac{\partial \log f(w;\theta_0)}{\partial \theta'}\right] = H$$

(Simplified)

Intuitive idea of where variance comes from:

Let $s_n(\hat{\theta})$ be the score function at $\hat{\theta}$.

If we expand $s_n(\hat{\theta})$ around θ_0 using MVT

$$\underbrace{s_n(\hat{\theta})}_{=0} \approx s_n(\theta_0) + H_n(\tilde{\theta})(\hat{\theta} - \theta_0),$$

$\tilde{\theta}$ between $\hat{\theta}, \theta_0$

Rearrange

$$\sqrt{n}(\hat{\theta} - \theta_0) \approx \underbrace{-H_n(\tilde{\theta})^{-1}}_{\xrightarrow{p} -H(\theta_0)^{-1} = I(\theta_0)^{-1}} \underbrace{\sqrt{n}\, s_n(\theta_0)}_{\xrightarrow{d} N(0, I(\theta_0)) \text{ by CLT}}$$

$$\xrightarrow{d} N(0, -H^{-1})$$

Hypothesis Testing

Hypothesis Testing

Suppose we want to test $H_0 : r(\theta) = 0$, where $r(\cdot)$ is a known function whose Jacobian $\boldsymbol{R}(\cdot)$ is both continuous and has full rank at θ_0.

The "trinity" of test statistics are:

- Wald:
  $W = nr(\hat{\theta})'(\boldsymbol{R}(\hat{\theta})\hat{\Sigma}^{-1}\boldsymbol{R}(\hat{\theta})')^{-1}r(\hat{\theta})$,
- Likelihood Ratio:
  $LR = 2n(Q_n(\hat{\theta}) - Q_n(\tilde{\theta}))$,
- Lagrange Multiplier:
  $LM = n\frac{\partial Q_n(\tilde{\theta})'}{\partial \theta}\tilde{\Sigma}^{-1}\frac{\partial Q_n(\tilde{\theta})}{\partial \theta}$,

$\hat{\theta}$ = unconstrained

$\tilde{\theta}$ = constrained

where $\tilde{\theta}$ is the constrained estimator

$$\tilde{\theta} \equiv \arg\min_{\theta \in \Theta} Q_n(\theta) \text{ s.t. } r(\theta) = 0$$

and where $(\hat{\Sigma}, \tilde{\Sigma})$ estimate the outer product of gradients at $(\hat{\theta}, \tilde{\theta})$.

Theorem

Assume that:

1. $\sqrt{n}(\hat{\theta} - \theta_0) = -\boldsymbol{H}^{-1}\sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial \theta} + o_P(1)$,
2. $\sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial \theta} \xrightarrow{d} N(0, \Sigma)$, Σ p.d.,
3. $\sqrt{n}(\tilde{\theta} - \theta_0) = O_P(1)$,
4. $\Sigma = -\boldsymbol{H}$.

Then all of (W, LR, LM) converge in distribution to $\chi^2_{\#r}$.

Furthermore (stated without proof), they are asymptotically equivalent: The difference between any two converges in probability to 0.

Key Points :

- Wald, LM, LR statistics are asymptotically equivalent
  - ↳ ie: share the same asymptotic distribution χ^2

- Trio of statistics in ML can be extended to GMM because of the relationship between ML and GMM.