Toulouse School of Economics

M1 INTERMEDIATE ECONOMETRICS

Asymptotics for nonlinear estimators

Koen Jochmans   François Poinas

2024 — 2025

Overview

This deck of slides goes over asymptotics for extremum estimators (NLLS and MLE)

The relevant chapter in Hansen is 22, but we give some additional detail and examples.

Global and local identification (H22.3)

Consider an estimator of θ_0

$$\hat{\theta} = \arg\max_{\theta \in \Theta} n^{-1} \sum_{i=1}^{n} \rho(Y_i, X_i, \theta).$$

This is based on the understanding that

$$\mathbb{E}\left(\rho(Y, X, \theta_0)\right) > \mathbb{E}\left(\rho(Y, X, \theta)\right)$$

for any $\theta \in \Theta$ different from θ_0.

This condition is a global identification condition.

With NLLS we maximize

$$-n^{-1}\sum_{i=1}^{n}\left(Y_i - m(X_i,\theta)\right)^2.$$

For MLE we maximize

$$n^{-1}\ell_n(\theta) = n^{-1}\sum_{i=1}^{n}\log f(Y_i|X_i,\theta).$$

In each of these cases we know that θ_0 is a global maximizer of the limit problem.

It need not be the only maximizer.

Local identification is the weaker requirement that the Hessian of the limit problem is negative definite at θ_0, and so the maximum is well isolated.

In the NLLS case, local identification is the requirement that

$$\operatorname{rank} \mathbb{E}\left(\frac{\partial m(X,\theta_0)}{\partial \theta}\,\frac{\partial m(X,\theta_0)}{\partial \theta'}\right) = k.$$

For the linear model this boils down to the usual no-multicolinearity condition.

Global identification is that

$$\begin{aligned}\mathbb{E}\left((Y - m(X,\theta))^2\right) &= \mathbb{E}\left((Y - m(X,\theta_0))^2\right) + \mathbb{E}\left((m(X,\theta_0) - m(X,\theta))^2\right) \\ &> \mathbb{E}\left((Y - m(X,\theta_0))^2\right)\end{aligned}$$

and so that $\mathbb{E}\left((m(X,\theta_0) - m(X,\theta))^2\right) > 0$. This happens if and only if

$$\mathbb{P}(m(X,\theta) \neq m(X,\theta_0)) > 0$$

for all $\theta \neq \theta_0$ in Θ. In the linear model this is the no-multicolinearity condition because $\mathbb{E}\left((m(X,\theta_0) - m(X,\theta))^2\right) = \mathbb{E}((\theta-\theta_0)'XX'(\theta-\theta_0))$ but we know that $\alpha'\mathbb{E}(XX')\alpha > 0$ for any $\alpha \neq 0$ when $\mathbb{E}(XX')$ is positive definite.

Uniform law of large numbers (H22.5)

Function $\rho(x, \theta)$ for $\theta \in \Theta$ (continuous on Θ compact) with $\mathbb{E}(\sup_{\theta \in \Theta} \rho(X, \theta)) < +\infty$.

A pointwise convergence result (i.e., for any fixed $\theta \in \Theta$) is

$$\mathbb{P}\left(\left|n^{-1} \sum_i \rho(X_i, \theta) - \mathbb{E}(\rho(X, \theta))\right| > \epsilon\right) < \delta, \qquad \text{for all} \qquad n > \underline{n}_\theta,$$

A uniform result is that, for all $\theta \in \Theta$,

$$\mathbb{P}\left(\left|n^{-1} \sum_i \rho(X_i, \theta) - \mathbb{E}(\rho(X, \theta))\right| > \epsilon\right) < \delta, \qquad \text{for all} \qquad n > \underline{n},$$

with $\underline{n}$ independent of θ.
We write

$$\sup_{\theta \in \Theta} \left|n^{-1} \sum_i \rho(X_i, \theta) - \mathbb{E}(\rho(X, \theta))\right| \underset{p}{\rightarrow} 0$$

as $n \rightarrow \infty$.

To appreciate the difference take a non-stochastic example:

$$n\theta e^{-n\theta}$$

for $\theta \in \Theta = [0, 1]$. This function is continuous in θ.

For any fixed θ,

$$n\theta e^{-n\theta} \to 0$$

as $n \to \infty$. (because the exponential term vanishes more quickly than the linear term grows.)

However, at $\theta = n^{-1}$ the function equals e^{-1} for any n. Hence,

$$\sup_{\theta \in \Theta} |n\theta e^{-n\theta}| \nrightarrow 0$$

as $n \to \infty$.

Uniform convergence implies pointwise convergence.

Argmax theorem and consistency (H22.4)

Let θ_0 be globally identified as the solution to

$$\max_{\theta \in \Theta} S(\theta), \qquad S(\theta) = \mathbb{E}(\rho(Y, X, \theta))$$

and let $\hat{\theta}$ be the solution to

$$\max_{\theta \in \Theta} S_n(\theta), \qquad S_n(\theta) = n^{-1} \sum_{i=1}^{n} \rho(Y_i, X_i, \theta).$$

By a uniform law of large numbers,

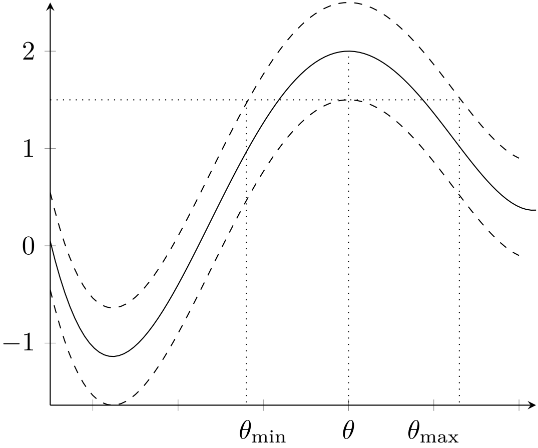
$$S_n(\theta) \underset{p}{\rightarrow} S(\theta)$$

uniformly in $\theta \in \Theta$.

Then

$$\arg\max_{\theta \in \Theta} S_n(\theta) = \hat{\theta} \underset{p}{\rightarrow} \theta_0 = \arg\max_{\theta \in \Theta} S(\theta).$$

Below is a uniform ε-band around $S(\theta)$ in which $S_n(\theta)$ must lie with high probability, and the corresponding interval $[\theta_{\min}, \theta_{\max}]$ in which $\hat{\theta}$ must then also lie with high probability.



As $n \to \infty$, the ε-band tightens and so the interval $[\theta_{\min}, \theta_{\max}]$ shrinks to a point. By identification this point must be θ_0. As $\hat{\theta} \in [\theta_{\min}, \theta_{\max}]$ it must be that $\hat{\theta}$ converges to θ_0.

If ρ is twice continuously-differentiable in θ and $S(\theta)$ is not maximized at the boundary of Θ we have that

$$\frac{\partial S_n(\hat{\theta})}{\partial \theta} = \frac{\partial S_n(\theta_0)}{\partial \theta} + \frac{\partial^2 S_n(\theta_*)}{\partial \theta \partial \theta'} (\hat{\theta} - \theta_0) = 0$$

by the first-order condition and a mean-value expansion. We may then solve for $\hat{\theta} - \theta_0$ to obtain

$$\hat{\theta} - \theta_0 = -\left(\frac{\partial^2 S_n(\theta_*)}{\partial \theta \partial \theta'} \right)^{-1} \frac{\partial S_n(\theta_0)}{\partial \theta}.$$

We handle each of the right-hand side terms separately next.

First, we would like to show that

$$\frac{\partial^2 S_n(\theta_*)}{\partial\theta\partial\theta'} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \rho(Y_i, X_i, \theta_*)}{\partial\theta\partial\theta'} \underset{p}{\rightarrow} \mathbb{E}\left(\frac{\partial^2 \rho(Y, X, \theta_0)}{\partial\theta\partial\theta'}\right)$$

where $\theta_* \underset{p}{\rightarrow} \theta_0$.
First,

$$\left\|\frac{\partial^2 S_n(\theta_*)}{\partial\theta\partial\theta'} - \frac{\partial^2 S(\theta_0)}{\partial\theta\partial\theta'}\right\| \leq \left\|\frac{\partial^2 S_n(\theta_*)}{\partial\theta\partial\theta'} - \frac{\partial^2 S(\theta_*)}{\partial\theta\partial\theta'}\right\| + \left\|\frac{\partial^2 S(\theta_*)}{\partial\theta\partial\theta'} - \frac{\partial^2 S(\theta_0)}{\partial\theta\partial\theta'}\right\|$$

Continuity of the second derivative of $S(\theta)$ together with consistency of θ_* implies that

$$\left\|\frac{\partial^2 S(\theta_*)}{\partial\theta\partial\theta'} - \frac{\partial^2 S(\theta_0)}{\partial\theta\partial\theta'}\right\| \underset{p}{\rightarrow} 0$$

by the continuous mapping theorem.

Next,

$$\left\|\frac{\partial^2 S_n(\theta_*)}{\partial\theta\partial\theta'} - \frac{\partial^2 S(\theta_*)}{\partial\theta\partial\theta'}\right\| = \left\|n^{-1}\sum_{i=1}^{n}\frac{\partial^2\rho(Y_i,X_i,\theta_*)}{\partial\theta\partial\theta'} - \mathbb{E}\left(\frac{\partial^2\rho(Y,X,\theta_*)}{\partial\theta\partial\theta'}\right)\right\|$$
$$\leq \sup_{\theta\in\Theta}\left\|n^{-1}\sum_{i=1}^{n}\frac{\partial^2\rho(Y_i,X_i,\theta)}{\partial\theta\partial\theta'} - \mathbb{E}\left(\frac{\partial^2\rho(Y,X,\theta)}{\partial\theta\partial\theta'}\right)\right\|$$

so that we can apply a uniform law of large numbers, provided that $\mathbb{E}\left(\sup_{\theta\in\Theta}\frac{\partial^2\rho(Y,X,\theta)}{\partial\theta\partial\theta'}\right) < \infty$, to obtain that

$$\left\|\frac{\partial^2 S_n(\theta_*)}{\partial\theta\partial\theta'} - \frac{\partial^2 S(\theta_*)}{\partial\theta\partial\theta'}\right\| \underset{p}{\to} 0.$$

Taken together we have shown that

$$\frac{\partial^2 S_n(\theta_*)}{\partial\theta\partial\theta'} \underset{p}{\to} \mathbb{E}\left(\frac{\partial^2\rho(Y,X,\theta_0)}{\partial\theta\partial\theta'}\right) = \boldsymbol{Q}$$

If $\boldsymbol{Q}$ is invertible then we apply the continuous mapping theorem and find that

$$\sqrt{n}(\hat{\theta} - \theta_0) = \left(-\boldsymbol{Q}^{-1} + o_p(1)\right) \sqrt{n}\, \frac{\partial S_n(\theta_0)}{\partial \theta}.$$

Next,

$$\sqrt{n}\, \frac{\partial S_n(\theta_0)}{\partial \theta} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\rho(Y_i, X_i, \theta_0)}{\partial \theta} \underset{d}{\rightarrow} N(0, \Omega).$$

provided that

$$\Omega = \mathbb{E}\left(\frac{\rho(Y_i, X_i, \theta_0)}{\partial \theta} \frac{\rho(Y_i, X_i, \theta_0)}{\partial \theta'} \right)$$

exists.

Hence,

$$\sqrt{n}(\hat{\theta} - \theta_0) \underset{d}{\rightarrow} N(0, \boldsymbol{Q}^{-1} \Omega \boldsymbol{Q}^{-1}).$$

This applies to both NLLS and MLE (and more generally still).

NLLS (H23.4)

For

$$S(\theta) = -\frac{1}{2}\mathbb{E}((Y - m(X,\theta))^2),$$

we have

$$\frac{\partial S(\theta)}{\partial \theta} = \mathbb{E}\left(\frac{\partial m(X,\theta)}{\partial \theta}(Y - m(X,\theta))\right),$$

so that

$$\Omega = \text{var}\left(\frac{\partial m(X,\theta_0)}{\partial \theta}\, e\right) = \mathbb{E}\left(\frac{\partial m(X,\theta_0)}{\partial \theta}\,\frac{\partial m(X,\theta_0)}{\partial \theta'}\, e^2\right),$$

and also

$$\boldsymbol{Q} = -\mathbb{E}\left(\frac{\partial m(X,\theta_0)}{\partial \theta}\,\frac{\partial m(X,\theta_0)}{\partial \theta'}\right).$$

Under conditional homoskedasticity, $\mathbb{E}(e^2|X) = \sigma^2$, $\Omega = -\boldsymbol{Q}\,\sigma^2$ and the asymptotic variance simplifies to

$$\sigma^2\,\boldsymbol{Q}^{-1}.$$

Compare all this to OLS and GLS.

We estimate the asymptotic variance $\boldsymbol{Q}^{-1}\Omega\boldsymbol{Q}^{-1}$ by the obvious plug-in estimator that uses

$$\hat{\boldsymbol{Q}} = -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial m(X_i,\hat{\theta})}{\partial\theta}\,\frac{\partial m(X_i,\hat{\theta})}{\partial\theta'}$$

and

$$\hat{\Omega} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial m(X_i,\hat{\theta})}{\partial\theta}\,\frac{\partial m(X_i,\hat{\theta})}{\partial\theta'}\,\hat{e}_i^2$$

for $\hat{e}_i = Y_i - m(X_i,\hat{\theta})$.

This is the analog of the usual robust variance-covariance matrix in the linear model.

The likelihood case has

$$S(\theta) = \mathbb{E}\left(\log f(Y|X,\theta)\right).$$

The relevant derivatives are

$$\frac{\partial S(\theta)}{\partial \theta} = \mathbb{E}\left(\frac{\partial \log f(Y|X,\theta)}{\partial \theta}\right), \quad \boldsymbol{Q} = \mathbb{E}\left(\frac{\partial^2 \log f(Y|X,\theta)}{\partial \theta \partial \theta'}\right),$$

and

$$\Omega = \operatorname{var}\left(\frac{\partial \log f(Y|X,\theta_0)}{\partial \theta}\right) = \mathbb{E}\left(\frac{\partial \log f(Y|X,\theta_0)}{\partial \theta}\frac{\partial \log f(Y|X,\theta_0)}{\partial \theta'}\right).$$

In this case we can use the information equality to achieve an important simplification.

Information inequality

The likelihood problem is fully parametric. For any function φ let

$$\mathbb{E}_{\theta}(\varphi(Y, X, \theta)|X = x) = \int \varphi(Y, X, \theta)\, f(y|x, \theta)\, dy.$$

Note that

$$\frac{\log f(y|x,\theta)}{\partial\theta} = \frac{1}{f(y|x,\theta)}\,\frac{\partial f(y|x,\theta)}{\partial\theta}. \tag{1}$$

Therefore,

$$\begin{aligned}\mathbb{E}_{\theta}\left(\frac{\partial \log f(Y|X,\theta)}{\partial\theta}\,\middle|\, X = x\right) &= \int \left(\frac{1}{f(y|x,\theta)}\,\frac{\partial f(y|x,\theta)}{\partial\theta}\right) f(y|x,\theta)\, dy \\ &= \int \frac{\partial f(y|x,\theta)}{\partial\theta}\, dy \\ &= \frac{\partial}{\partial\theta}\int f(y|x,\theta)\, dy = 0\end{aligned}$$

(provided that the support of f does not change with θ, which we will use as a regularity condition.)

Now we can differentiate the condition

$$\int \frac{\partial \log f(y|x,\theta)}{\partial \theta} f(y|x,\theta)\,dy = 0$$

with respect to θ to obtain

$$\int \frac{\partial^2 \log f(y|x,\theta)}{\partial \theta \partial \theta'} f(y|x,\theta) + \frac{\partial \log f(y|x,\theta)}{\partial \theta} \frac{\partial f(y|x,\theta)}{\partial \theta'}\,dy = 0.$$

Using Eq. (1) this gives the identify

$$-\int \frac{\partial^2 \log f(y|x,\theta)}{\partial \theta \partial \theta'} f(y|x,\theta)\,dy = \int \frac{\partial \log f(y|x,\theta)}{\partial \theta} \frac{\partial \log f(y|x,\theta)}{\partial \theta'} f(y|x,\theta)\,dy.$$

Both sided of this equation are expectations. Moreover,

$$-\boldsymbol{Q} = \Omega.$$

This is the information equality.

We call $\partial \log f(y|x, \theta)/\partial\theta$ the score.

The variance of the score, Ω, is called the information.

The asymptotic variance of the MLE is thus equal to the inverse of the information matrix.

Recall the linear model

$$Y|X = x \sim N(x'\beta, \sigma^2).$$

Here, the score for β was (at true values)

$$\frac{X(Y - X'\beta_0)}{\sigma_0^2} = \frac{Xe}{\sigma_0^2}$$

and so

$$\Omega = \frac{\mathbb{E}(XX')}{\sigma_0^2}.$$

Furthermore, the Hessian matrix for β was

$$-\frac{XX'}{\sigma^2}$$

so that, clearly,

$$\boldsymbol{Q} = -\frac{\mathbb{E}(XX')}{\sigma_0^2}.$$

For the Poisson model with conditional mean and variance $\exp(X'\beta)$ the score was

$$X(Y - \exp(X'\beta))$$

which, has variance

$$\Omega = \mathbb{E}(XX'\text{var}(Y|X)) = \mathbb{E}(XX'\exp(X'\beta_0)) = -\boldsymbol{Q}.$$

Here the mean-variance equality embedded in the poisson distribution is important.

Relaxing this restriction to allow for over/under dispersion leads to negative binomial models.

Logit (H25.7)

The binary-choice logit model has a simpler form than the probit model because, with

$$F(u) = \frac{1}{1+\exp(-u)},$$

it is easy to see that the associated density is

$$F'(u) = \tfrac{\exp(-u)}{(1+\exp(-u))^2} = \tfrac{1}{1+\exp(-u)}\,\tfrac{\exp(-u)}{1+\exp(-u)} = F(u)(1-F(u)).$$

Then, recall that the log-pmf is

$$Y\log(F(X'\theta)) + (1-Y)\log(1-F(X'\theta)).$$

The score, which is,

$$X\,F'(X'\theta)\,\left(\tfrac{Y}{F(X'\theta)} - \tfrac{(1-Y)}{1-F(X'\theta)}\right) = X F'(X'\theta)\,\tfrac{Y-F(X'\theta)}{F(X'\theta)\,(1-F(X'\theta))},$$

thus simplifies to just

$$X(Y - F(X'\theta)),$$

which clearly has mean zero and variance

$$\Omega = \mathbb{E}(XX'F'(X'\theta_0)) = -\boldsymbol{Q}$$

at the truth.

Like before we can estimate the asymptotic variance $\boldsymbol{Q}^{-1}\Omega\boldsymbol{Q}^{-1}$ by the obvious plug-in estimator that uses

$$\hat{\boldsymbol{Q}} = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2 \log f(Y_i|X_i,\hat{\theta})}{\partial\theta\partial\theta'}$$

and

$$\hat{\Omega} = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \log f(Y_i|X_i,\hat{\theta})}{\partial\theta}\frac{\partial \log f(Y_i|X_i,\hat{\theta})}{\partial\theta'}.$$

The information identify also justifies and estimator based only on one of the two. Some software programs use $-\hat{\boldsymbol{Q}}^{-1}$ as the default variance estimator because this quantity has usually already been calculated in the optimization of the likelihood (recall Newton's algorithm), while $\hat{\Omega}$ requires an additional step.

We can follow the Wald principle in exactly the same way as before.

We have

$$\sqrt{n}(\hat{\theta} - \theta_0) \underset{d}{\to} N(0, \boldsymbol{V}_\theta)$$

and so

$$n\,(\hat{\theta} - \theta_0)' \hat{\boldsymbol{V}}_\theta^{-1}\,(\hat{\theta} - \theta_0) \underset{d}{\to} \chi^2_k.$$

Tests about nonlinear transformation of θ_0 follow in the same way by a delta-method argument.

Consider a parametric problem with parameter θ and a random sample $Z_1, \dots, Z_n$.

Suppose that $\hat{\theta}$ is unbiased for θ. (Bias can be accommodated at the expense of some additional notation.)

Then

$$\mathbb{E}_\theta(\hat{\theta} - \theta) = 0.$$

Moreover,

$$\iint \cdots \int (\hat{\theta}(z_1, z_2, \dots, z_n) - \theta) \prod_{i=1}^{n} f(z_i, \theta)\, dz_1 dz_2 \dots dz_n = 0$$

holds for every θ.

The derivative of the expression under the integral is

$$(\hat{\theta}(z_1, z_2, \dots, z_n) - \theta)\frac{\partial \prod_{i=1}^n f(z_i, \theta)}{\partial \theta} - \prod_{i=1}^n f(z_i, \theta).$$

This gives

$$\int \cdots \int (\hat{\theta}(z_1, z_2, \dots, z_n) - \theta)\frac{\partial \prod_{i=1}^n f(z_i, \theta)}{\partial \theta} \, dz_1 \dots dz_n = 1$$

because $\int \cdots \int \frac{\partial \prod_{i=1}^n f(z_i, \theta)}{\partial \theta} \, dz_1 \dots dz_n = 1$.

Further, observe that

$$\sum_{i=1}^{n} \frac{\partial \log f(z_i,\theta)}{\partial \theta} = \sum_{i=1}^{n} \frac{1}{f_\theta(z_i,\theta)} \frac{\partial f(z_i,\theta)}{\partial \theta}$$

$$= \sum_{i=1}^{n} \left(\frac{\prod_{j \neq i} f(z_j,\theta)}{\prod_j f(z_j,\theta)} \right) \frac{\partial f(z_i,\theta)}{\partial \theta}$$

$$= \frac{\sum_{i=1}^{n} \prod_{j \neq i} f(z_j,\theta) \frac{\partial f(z_i,\theta)}{\partial \theta}}{\prod_j f(z_j,\theta)} = \frac{\frac{\partial \prod_i f(z_i,\theta)}{\partial \theta}}{\prod_j f_\theta(x_j)}$$

so that

$$\frac{\partial \prod_{i=1}^{n} f(z_i,\theta)}{\partial \theta} = \left(\sum_{i=1}^{n} \frac{\partial \log f(z_i,\theta)}{\partial \theta} \right) \left(\prod_{j=1}^{n} f(z_j,\theta) \right).$$

Therefore, our unbiasedness condition implies that the integral

$$\int .. \int (\hat{\theta}(z_1, \ldots, z_n) - \theta) \left(\sum_{i=1}^{n} \frac{\partial \log f(z_i, \theta)}{\partial \theta} \right) \left(\prod_{j=1}^{n} f(z_j, \theta) \right) dz_1..dz_n$$

must equal one. But the integral is equal to

$$\mathbb{E}\left((\hat{\theta} - \theta) \left(\sum_{i=1}^{n} \frac{\partial \log f(Z_i, \theta)}{\partial \theta} \right) \right) = \text{cov}\left(\hat{\theta}, \sum_{i=1}^{n} \frac{\partial \log f(Z_i, \theta)}{\partial \theta} \right)$$

and so, by Cauchy-Schwarz,

$$1^2 = \text{cov}\left(\hat{\theta}, \sum_{i=1}^{n} \frac{\partial \log f(Z, \theta)}{\partial \theta} \right)^2 \leq \text{var}\left(\hat{\theta} \right) n \, \text{var}\left(\frac{\partial \log f(Z, \theta)}{\partial \theta} \right).$$

Hence,

$$\text{var}\left(\hat{\theta} \right) \geq \frac{\Omega^{-1}}{n}.$$

Achieving the bound is not possible for a given n in general.

However, the MLE achieves it as $n \to \infty$.

Henc, MLE is asymptotically efficient.

This result uses the information equality, which requires correct specification of the likelihood function.