ECON 6200: Econometrics II

Extremum Estimation

Extremum Estimators

An extremum estimator is any estimator defiend as

$$\hat{\theta} = \arg\min_{\theta \in \Theta} Q_n(W_1, \ldots, W_n; \theta)$$

for some parameter θ in parameters apace Θ and where $W_1, \ldots, W_n$ is a sample.

Extremum Estimation

Extremum Estimators

An extremum estimator is any estimator defiend as

$$\hat{\theta} = \arg\min_{\theta \in \Theta} Q_n(W_1, \ldots, W_n; \theta)$$

for some parameter θ in parameters apace Θ and where $W_1, \ldots, W_n$ is a sample.

- The criterion function $Q_n(\cdot)$ must be indexed by n because its mathematical form necessaily depends on n.
- But it usually is is intuitively "the same" function at different n.
  For example, consider $Q_n(\cdot) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i'b)^2$.
- Similarly to GMM notation, we will often drop the data from the function's argument and just write $Q_n(\theta)$.

Extremum Estimation

Why would this estimate a true parameter value θ_0?

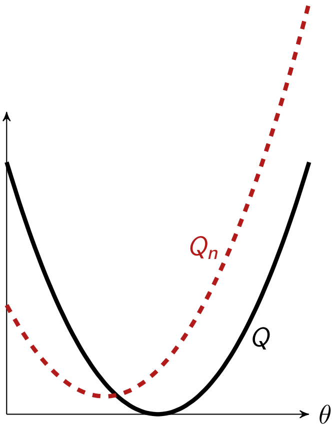
Invariably the intuition is as follows:

$$\begin{aligned}
\theta_0 &= \arg\min_{\theta\in\Theta} Q(\theta) \\
\hat{\theta} &= \arg\min_{\theta\in\Theta} Q_n(\theta) \\
Q_n(\cdot) &\to Q(\theta) \\
\overset{?}{\Longrightarrow} \hat{\theta} &\to \theta_0
\end{aligned}$$

That is, the sample criterion $Q_n(\cdot)$ estimates some population criterion $Q_n(\cdot)$ that is minimized at θ_0.

It is intuitively compelling that in "nice" cases, that implies $\hat{\theta} \to \theta_0$.

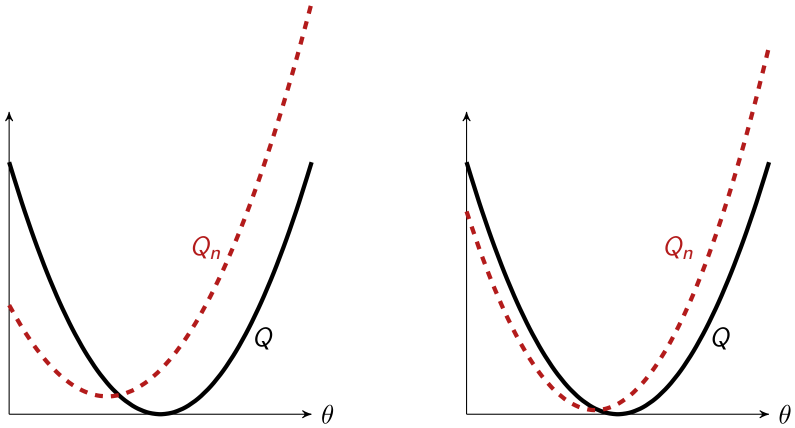
Extremum Estimation



Here's a visualization of the "nice" case:
We see two "typical" realizations with the larger n on the right.

Can you spot θ_0 and $\hat{\theta}$?

Extremum Estimation



Next steps:

- Some examples. M-estimation as special case.
- Working out the theory.

Examples of Extremum Estimators

GMM

$$\begin{aligned} Q(\theta) &= \mathbb{E}g(\theta)'\boldsymbol{W}\mathbb{E}g(\theta) \\ Q_n(\theta) &= \mathbb{E}_n g(\theta)'\hat{\boldsymbol{W}}\mathbb{E}_n g(\theta). \end{aligned}$$

Examples of Extremum Estimators

GMM

$$\begin{aligned} Q(\theta) &= \mathbb{E}g(\theta)'\boldsymbol{W}\mathbb{E}g(\theta) \\ Q_n(\theta) &= \mathbb{E}_n g(\theta)'\hat{\boldsymbol{W}}\mathbb{E}_n g(\theta). \end{aligned}$$

Method of Simulated Moments

$$\begin{aligned} Q(\theta) &= (\pi(\theta) - \pi_0)'\boldsymbol{W}(\pi(\theta) - \pi_0) \\ Q_n(\theta) &= (\tilde{\pi}(\theta) - \hat{\pi})'\hat{\boldsymbol{W}}(\tilde{\pi}(\theta) - \hat{\pi}), \end{aligned}$$

where

- the function $\pi(\cdot)$ maps parameter values onto implied moments of the data, e.g. means, variances, or entire time series of inflation, uinemployment,...
- π_0 are the true such moments and $\hat{\pi}$ an estimate,
- $\tilde{\pi}(\cdot)$ is a simulated analog of $\pi(\cdot)$.

This interestingly differs from GMM if simulation noise in $\tilde{\pi}$ cannot be ignored. Otherwise, it really is GMM but sometimes still called MSM.

Examples of Extremum Estimators

Nonlinear Least Squares

$$\begin{aligned} Q(\theta) &= \mathbb{E}(Y - m(X,\theta))^2 \\ Q_n(\theta) &= \mathbb{E}_n(Y - m(X,\theta))^2. \end{aligned}$$

You could argue this is just GMM (consider the FOC) but it was developed separately.

Examples of Extremum Estimators

Nonlinear Least Squares

$$
\begin{aligned}
Q(\theta) &= \mathbb{E}(Y - m(X,\theta))^2 \\
Q_n(\theta) &= \mathbb{E}_n(Y - m(X,\theta))^2.
\end{aligned}
$$

You could argue this is just GMM (consider the FOC) but it was developed separately.

Maximum Likelihood

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

The "conceptual" definition is at first glance different, but we will later derive the above from it.

M-Estimation

An important special case are m-estimators:

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

for some known, real-valued function $m(\cdot)$.

Examples:

- Maximum Likelihood: $m(W;\theta) = \ell(W;\theta)$,
- One-Step GMM: $m(W;\theta) = g(W;\theta)'\boldsymbol{W}g(W;\theta)$.
  (Why is efficient GMM not an m-estimator?)

This class is of interest because some building blocks of asymptotic theory are easily available at exactly this level of generality.

Warning: Some texts use m-estimation as synonym for extremum estimation.

Consistency

We next formalize the intuitive argument for consistency.

We start with high-level assumptions that we then verify in special cases.

Extremum Estimation

Consistency

We next formalize the intuitive argument for consistency.

We start with high-level assumptions that we then verify in special cases.

Note:

For simplicity, the following slides assume that $\arg\min_{\theta\in\Theta} Q_n(\theta)$ exists.

Can verify (homework) that everything goes through as long as

$$Q_n(\hat{\theta}) \leq \inf_{\theta\in\Theta} Q_n(\theta) + 1/n.$$

Thus, $\hat{\theta}$ can be an arbitrary choice fulfilling this constraint.

That settles existence and is also practically relevant because $\hat{\theta}$ may be numerically evaluated and then not exact.

Consistency

We next formalize the intuitive argument for consistency.

We start with high-level assumptions that we then verify in special cases.

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

Proof

Fix $\epsilon > 0$ and define $Q_n^\epsilon \equiv \inf_{\theta\in\Theta:\|\theta-\theta_0\|\geq\epsilon} Q_n(\theta)$, then

$$\begin{aligned}
& \Pr(\|\hat{\theta}-\theta_0\| > \epsilon) \\
\leq\ & \Pr\big(Q_n^\epsilon \leq Q_n(\theta_0)\big) \\
=\ & 1-\Pr\big(Q_n^\epsilon > Q_n(\theta_0)\big) \\
\leq\ & 1-\Pr\big(Q_n^\epsilon > Q_\epsilon - \delta/2, Q_n(\theta_0) < Q(\theta_0)+\delta/2\big) \\
\to\ & 0,
\end{aligned}$$

where all inequalities exploit logical implications; the last step uses that, by Assumption 1, $Q_n(\theta_0) \xrightarrow{p} Q(\theta_0)$ and $Q_n^\epsilon \xrightarrow{p} Q^\epsilon$.

(Where was the uniform convergence from Assumption 1 used?)

Extremum Estimation

Proof

Fix $\epsilon > 0$ and define $Q_n^\epsilon \equiv \inf_{\theta\in\Theta:\|\theta-\theta_0\|\geq\epsilon} Q_n(\theta)$, then

$$\begin{aligned} & \Pr(\|\hat{\theta}-\theta_0\| > \epsilon) \\ \leq\ & \Pr\big(Q_n^\epsilon \leq Q_n(\theta_0)\big) \\ =\ & 1-\Pr\big(Q_n^\epsilon > Q_n(\theta_0)\big) \\ \leq\ & 1-\Pr\big(Q_n^\epsilon > Q_\epsilon - \delta/2, Q_n(\theta_0) < Q(\theta_0)+\delta/2\big) \\ \to\ & 0, \end{aligned}$$

where all inequalities exploit logical implications; the last step uses that, by Assumption 1, $Q_n(\theta_0) \xrightarrow{p} Q(\theta_0)$ and $Q_n^\epsilon \xrightarrow{p} Q^\epsilon$.

(Where was the uniform convergence from Assumption 1 used?)

In the very last bit:

$$|Q_n^\epsilon - Q^\epsilon| = |Q_n(\theta_n^\epsilon) - Q(\theta^\epsilon)| \leq \sup_{\theta\in\Theta} |Q_n(\theta) - Q(\theta)| \xrightarrow{d} 0$$

but the last step used uniform convergence.

Extremum Estimation

The preceding result used uniform convergence and well-separated minimum.

We next provide lower-level conditions that imply these.

Extremum Estimation

The preceding result used uniform convergence and well-separated minimum.

We next provide lower-level conditions that imply these.

Theorem

Assume that:

1. $Q(\cdot)$ is continuous,
2. Θ is compact,
3. θ_0 uniquely minimizes $Q(\theta)$.

Then θ_0 is a well-separated minimum.

Proof

Fix $\epsilon > 0$. By the Weierstrass Theorem, Q^ϵ is attained by some θ^ϵ with $\|\theta^\epsilon - \theta_0\| \geq \epsilon$. Set $\delta = Q(\theta^\epsilon) - Q(\theta_0)$, which is not zero by Assumption 3.

Extremum Estimation

The preceding result used uniform convergence and well-separated minimum.

We next provide lower-level conditions that imply these.

Theorem

Assume that:

1. $\hat{\theta}$ is an m-estimator,
2. The data are i.i.d. realizations of W,
3. $m(W;\theta)$ is a.s. continuous in θ,
4. $|m(W;\theta)| \leq G(W)$ for some function G s.t. $\mathbb{E}G(W) < \infty$,
5. Θ is compact.

Then $Q_n(\cdot)$ converges to $Q(\cdot)$ uniformly.

Proof

This is the Uniform Law of Large Numbers.

Extremum Estimation

We can consolidate two of the above results into one handy theorem. (Compare Theorem 2.1 in Newey/McFadden 1994.)

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

This result covers many cases of interest.

We proved it already. The next slides illustrate necessity of the assumptions.

Consolidated Theorem: Necessity of Uniqueness

The example illustrates that unique minimization is an identification condition: Without it, even knowledge of $Q(\cdot)$ does not imply knowledge of θ_0.

The example can also be seen as illustrating partial identification. Estimation and inference theory for $\Theta_I \equiv \arg\min_{\theta\in\Theta} Q(\theta)$ (a possibly nonsingleton identified set) is an active literature.

Consolidated Theorem: Necessity of Continuity

Consolidated Theorem: Necessity of Compactness

Consolidated Theorem: Necessity of Uniform Convergence of $Q_n(\cdot)$

Extremum Estimation

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

Proof:

The proof of a simplified statement will be a homework.
If $Q_n(\cdot)$ (and by implication $Q(\cdot)$) is convex, we only need pointwise convergence.

The assumptions also ensure existence of $\hat{\theta}$ that exactly minimizes $Q_n(\cdot)$.

A Comment on Rate of Convergence

We are about to move on to $\sqrt{n}$-asymptotic normality.

Are there intermediate assumptions under which we can ensure a rate of convergence without ensuring asymptotic normality?

Yes: They relate the curvature of $Q(\cdot)$ at θ_0 to such a rate.

This rate is $\sqrt{n}$ if $Q(\cdot)$ locally dominates some quadratic function.

For the exact result, see van der Vaart and Wellner's "Argmax Theorem" (in *Weak Convergence and Empirical Processes*).

Extremum Estimation

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

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

Proof

By the definition of $\hat{\theta}$ and the first two assumptions, we have that with probability approaching 1,

$$\frac{dQ_n(\hat{\theta})}{d\theta} = 0.$$

Proof

By the definition of $\hat{\theta}$ and the first two assumptions, we have that with probability approaching 1,

$$\frac{dQ_n(\hat{\theta})}{d\theta} = 0.$$

Also using assumption 3 and the Mean Value Theorem, we can write (again with probability approaching 1)

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

Proof

By the definition of $\hat{\theta}$ and the first two assumptions, we have that with probability approaching 1,

$$\frac{dQ_n(\hat{\theta})}{d\theta} = 0.$$

Also using assumption 3 and the Mean Value Theorem, we can write (again with probability approaching 1)

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

where $\bar{\theta}$ is co-ordinatewise between θ_0 and $\hat{\theta}$; in particular, $\bar{\theta} \xrightarrow{p} \theta_0$.

Proof

By the definition of $\hat{\theta}$ and the first two assumptions, we have that with probability approaching 1,

$$\frac{dQ_n(\hat{\theta})}{d\theta} = 0.$$

Also using assumption 3 and the Mean Value Theorem, we can write (again with probability approaching 1)

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

where $\bar{\theta}$ is co-ordinatewise between θ_0 and $\hat{\theta}$; in particular, $\bar{\theta} \xrightarrow{p} \theta_0$.

From here, the idea is to combine and rearrange to find

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

Extremum Estimation

Proof

By the definition of $\hat{\theta}$ and the first two assumptions, we have that with probability approaching 1,

$$\frac{dQ_n(\hat{\theta})}{d\theta} = 0.$$

Also using assumption 3 and the Mean Value Theorem, we can write (again with probability approaching 1)

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

where $\bar{\theta}$ is co-ordinatewise between θ_0 and $\hat{\theta}$; in particular, $\bar{\theta} \xrightarrow{p} \theta_0$.

From here, the idea is to combine and rearrange to find

$$\sqrt{n}(\hat{\theta} - \theta_0) = -\underbrace{\left(\frac{dQ_n(\bar{\theta})^2}{d\theta d\theta'}\right)^{-1}}_{\xrightarrow{p} \boldsymbol{H}^{-1}} \underbrace{\sqrt{n}\frac{dQ_n(\theta_0)}{d\theta}}_{\xrightarrow{d} N(0,\Sigma)}.$$

Proof (ctd.)

We wrap up by clarifying the convergence to $\boldsymbol{H}^{-1}$.

To keep displays neat, define $H(\theta) = \frac{dQ_n(\bar{\theta})^2}{d\theta d\theta'}$ and $H_n(\cdot)$ analogously, then

$$\begin{aligned}
\left\| H_n(\bar{\theta}) - \boldsymbol{H} \right\| &= \left\| H_n(\bar{\theta}) - H(\bar{\theta}) + H(\bar{\theta}) - \boldsymbol{H} \right\| \\
&\leq \left\| H_n(\bar{\theta}) - H(\bar{\theta}) \right\| + \left\| H(\bar{\theta}) - \boldsymbol{H} \right\| \\
&\leq \sup_{\theta \in \mathcal{N}} \| H_n(\theta) - H(\theta) \| + \left\| H(\bar{\theta}) - \boldsymbol{H} \right\| \\
&\xrightarrow{p} 0,
\end{aligned}$$

where

- the first step is an add-and-subtract trick,
- the second one is the triangle inequality,
- we next use assumption 1 (strictly speaking, this step "only" holds with probability approaching 1),
- the last step uses assumptions 3 and 5.

The claim now follows by nonsingularity of $\boldsymbol{H}$ and the Continuous Mapping Theorem in close analogy to earlier proofs.

Specialization to GMM

We can slightly improve on the theorem if the application is nonlinear GMM. Recall that $\hat{\theta} = \arg\min_{\theta \in \Theta} \{\overline{g}_n(\theta)' \boldsymbol{W} \overline{g}_n(\theta)\}$.

Assume that:

1. $\hat{\theta} \xrightarrow{p} \theta_0$,
2. $\theta_0 \in \text{int}(\Theta)$,
3. $g(W; \theta)$ is a.s. continuously differentiable in an open neighborhood $\mathcal{N}$ of θ_0,
4. $\sqrt{n}\overline{g}_n(\theta_0) \xrightarrow{d} N(0, \boldsymbol{S})$, $\boldsymbol{S}$ p.d.
5. $\sup_{\theta \in \mathcal{N}} \left\| \frac{d\overline{g}_n(\theta)}{d\theta'} - \mathbb{E}\left(\frac{d\overline{g}(\theta_0)}{d\theta'}\right) \right\| \xrightarrow{p} 0$,
6. $\boldsymbol{G} \equiv \frac{dg(\theta_0)}{d\theta'}$ is of full column rank.

Then

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N\left(0, (\boldsymbol{G}'\boldsymbol{W}\boldsymbol{G})^{-1}\boldsymbol{G}'\boldsymbol{W}\boldsymbol{S}\boldsymbol{W}\boldsymbol{G}(\boldsymbol{G}'\boldsymbol{W}\boldsymbol{G})^{-1}\right).$$

Specialization to GMM

We can slightly improve on the theorem if the application is nonlinear GMM. Recall that $\hat{\theta} = \arg\min_{\theta \in \Theta} \{\overline{g}_n(\theta)' \boldsymbol{W} \overline{g}_n(\theta)\}$.

Assume that:

1. $\hat{\theta} \xrightarrow{p} \theta_0$,
2. $\theta_0 \in \text{int}(\Theta)$,
3. $g(W;\theta)$ is a.s. continuously differentiable in an open neighborhood $\mathcal{N}$ of θ_0,
4. $\sqrt{n}\overline{g}_n(\theta_0) \xrightarrow{d} N(0, \boldsymbol{S})$, $\boldsymbol{S}$ p.d.
5. $\sup_{\theta \in \mathcal{N}} \left\| \frac{d\overline{g}_n(\theta)}{d\theta'} - \mathbb{E}\left(\frac{d\overline{g}(\theta_0)}{d\theta'}\right) \right\| \xrightarrow{p} 0$,
6. $\boldsymbol{G} \equiv \frac{dg(\theta_0)}{d\theta'}$ is of full column rank.

Then

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N\left(0, (\boldsymbol{G}'\boldsymbol{W}\boldsymbol{G})^{-1}\boldsymbol{G}'\boldsymbol{W}\boldsymbol{S}\boldsymbol{W}\boldsymbol{G}(\boldsymbol{G}'\boldsymbol{W}\boldsymbol{G})^{-1}\right).$$

- Two-stage (efficient) GMM works just as before.
- The main improvement is that we need only once differentiability of $g(\cdot)$. Why?

Maximum Likelihood

Maximum Likelihood is an extremely important special case.

Say we are able to specify the distribution of data up to θ.

For example, the data are distributed with density

$$f(W_1, \ldots, W_n; \theta),$$

where the function $f(\cdot)$ is known.
(Assuming existence of a density is not essential.)

Then the Maximum Likelihood estimator is

$$\hat{\theta}_{ML} \equiv \arg\max_{\theta \in \Theta} f(w_1, \ldots, w_n; \theta).$$

Intuitively, this is the parameter value that maximizes the likelihood of observing the data that were in fact observed.

(For discussion of ML, we will think of extremum estimators as maximizing $Q(\cdot)$.)

Maximum Likelihood as M-Estimator

As we assume that data are i.i.d., we have the simplification

$$
\begin{aligned}
\hat{\theta}_{ML} &\equiv \arg\max_{\theta\in\Theta} f(w_1,\ldots,w_n;\theta) \\
&= \arg\max_{\theta\in\Theta} \prod_{i=1}^{n} f(w_i;\theta) \\
&= \arg\max_{\theta\in\Theta} \sum_{i=1}^{n} \log f(w_i;\theta) \\
&= \arg\max_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \log f(w_i;\theta).
\end{aligned}
$$

- This is much easier (often the only realistic objective) to compute.
- It is typically consistent even if the data are not i.i.d.
  (We might get to why. We will mostly assume that data are i.i.d.)
- The last step just reminds us that this is an m-estimator.

Remarks on Identification

You may have encountered different definitions of identification:

- In linear moment-based models, it is a rank condition.
- In extremum estimation, it is that θ_0 **uniquely** minimizes $Q(\cdot)$.
- In Maximum Likelihood, it is

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

  or equivalently,

$$\theta \neq \theta_0 \implies \exists A \subseteq \mathcal{W}, \Pr(A) > 0, f(w;\theta) \neq f(w;\theta_0) \forall w \in A,$$

  where $\mathcal{W}$ is the sample space or set of all possible realizations of W.

  Verbally, data that signal whether θ or θ_0 is true have positive probability. (The above probabilities are evaluated under the true distribution.)

What do these have in common?

Extremum Estimation

Remarks on Identification

All of the above operationalize the same concept:

If we knew the population distribution of the data, we could back out θ_0.

- In linear moment-based models, the rank condition implies that the population moment conditions can be solved for θ_0.
- In extremum estimation, uniqueness of the minimum at θ_0 means that knowledge of $Q(\cdot)$ implies knowledge of θ_0 (at least in principle).
- In Maximum Likelihood... well, we'll see.

Remarks on Identification

All of the above operationalize the same concept:

If we knew the population distribution of the data, we could back out θ_0.

- In linear moment-based models, the rank condition implies that the population moment conditions can be solved for θ_0.
- In extremum estimation, uniqueness of the minimum at θ_0 means that knowledge of $Q(\cdot)$ implies knowledge of θ_0 (at least in principle).
- In Maximum Likelihood... well, we'll see.

Warning: The term "identification" is loaded.

There are subtly different usages (see a JEL suvey by Lewbel).

There is a rather different usage in empirical work:
"Where does your identification come from?"

Our usage corresponds to identifiability in statistics.

Extremum Estimation

The following notation may be helpful.

- $\mathcal{F}$ is the set of all possible population distributions of data W,
- Θ is parameter space,
- The correspondence $\Gamma : \Theta \mapsto \mathcal{F}$ maps each parameter value on the set of distributions consistent with it.
  That set is a singleton if a likelihood is specified.
  For GMM, it would be $\Gamma(\theta) = \{F(W) \in \mathcal{F} : \mathbb{E}_F g(W;\theta) = 0\}$.
- Then θ_0 is identified if $\mathcal{F} \in \Gamma(\theta_0)$ implies $\Gamma^{-1}(F) = \{\theta_0\}$.
- We usually consider θ identified if the above holds for all possible true values.

Extremum Estimation

The following notation may be helpful.

- $\mathcal{F}$ is the set of all possible population distributions of data W,
- Θ is parameter space,
- The correspondence $\Gamma : \Theta \mapsto \mathcal{F}$ maps each parameter value on the set of distributions consistent with it.
  That set is a singleton if a likelihood is specified.
  For GMM, it would be $\Gamma(\theta) = \{F(W) \in \mathcal{F} : \mathbb{E}_F g(W;\theta) = 0\}$.
- Then θ_0 is identified if $\mathcal{F} \in \Gamma(\theta_0)$ implies $\Gamma^{-1}(F) = \{\theta_0\}$.
- We usually consider θ identified if the above holds for all possible true values.

This can be used to motivate some extensions (not pursued in this class):

- Partial identification: $\Gamma^{-1}(\theta_0)$ is a set, completely uninformative if it is Θ, point-identifying if it is $\{\theta_0\}$, but frequently in between.
- Irregular Identification or Ill-posed Inverse Problems: $\Gamma^{-1}(\cdot)$ is sufficiently ill-behaved so that identifiability formally obtains but, for example, convergence of the empirical distribution F_n to F may imply convergence of $\Gamma^{-1}(F_n)$ to $\Gamma^{-1}(F)$ at a slower, if any, rate.

Extremum Estimation

Extremum Estimation

Conditional Maximum Likelihood

In many cases, the distribution of regressors X is not informative about θ.

That is, we can write

$$f(Y, X; \theta) = f_y(Y|X; \theta) f_x(X).$$

In this case, we have simplification

$$\begin{aligned}
\hat{\theta}_{ML} &= \arg\max_{\theta \in \Theta} \sum_{i=1}^{n} \log f(Y, X; \theta) \\
&= \arg\max_{\theta \in \Theta} \sum_{i=1}^{n} \big(\log f_y(Y|X; \theta) + \log f_x(X)\big) \\
&= \arg\max_{\theta \in \Theta} \sum_{i=1}^{n} \log f_y(Y|X; \theta).
\end{aligned}$$

In practice, many ML estimators reflect this simplification.

For the purpose of theoretical analysis, we always write the estimator as maximizing the complete likelihood.

Consistency of Maximum Likelihood

Consistency of Ml 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

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

Extremum Estimation

Theorem

θ_0 uniquely maximizes $\mathbb{E}(\log f(W;\theta))$ if, and only if, $\theta \neq \theta_0$ implies $\Pr(f(W;\theta) \neq f(W;\theta_0)) > 0$.

Proof

Write

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

where the inequality is Jensen's inequality and is strict unless

$$\frac{f(W;\theta)}{f(W;\theta_0)} \text{ constant a.s.} \iff \Pr(f(W;\theta) \neq f(W;\theta_0)) = 0.$$

Asymptotic Distribution of Maximum Likelihood

The structure of ML allows us to both verify the "CLT assumption" and provide an important expression for the asymptotic variance.

Write

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

Asymptotic Distribution of Maximum Likelihood

The structure of ML allows us to both verify the "CLT assumption" and provide an important expression for the asymptotic variance.

Write

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

This result (the score equation) is important in its own right:
It implies that ML can be interpreted as method-of-moments estimator.

Extremum Estimation

Taking derivatives once more:

$$
\begin{aligned}
&\int \frac{\partial^2 \log f(w;\theta_0)}{\partial\theta\partial\theta'} f(w;\theta_0)dw + \int \frac{\partial \log f(w;\theta_0)}{\partial\theta}\frac{\partial \log f(w;\theta_0)}{\partial\theta'} f(w;\theta_0)dw = 0 \\
\Longrightarrow \quad & \mathbb{E}\left(\frac{\partial^2 \log f(w;\theta_0)}{\partial\theta\partial\theta'}\right) + \mathbb{E}\left(\frac{\partial \log f(w;\theta_0)}{\partial\theta}\frac{\partial \log f(w;\theta_0)}{\partial\theta'}\right) = 0 \\
\Longrightarrow \quad & \mathbb{E}\left(\frac{\partial^2 \log f(w;\theta_0)}{\partial\theta\partial\theta'}\right) = -\mathbb{E}\left(\frac{\partial \log f(w;\theta_0)}{\partial\theta}\frac{\partial \log f(w;\theta_0)}{\partial\theta'}\right).
\end{aligned}
$$

The last line is famous as information matrix equality.

Taking derivatives once more:

$$\int \frac{\partial^2 \log f(w;\theta_0)}{\partial\theta\partial\theta'} f(w;\theta_0)dw + \int \frac{\partial \log f(w;\theta_0)}{\partial\theta}\frac{\partial \log f(w;\theta_0)}{\partial\theta'} f(w;\theta_0)dw = 0$$

$$\implies \mathbb{E}\left(\frac{\partial^2 \log f(w;\theta_0)}{\partial\theta\partial\theta'}\right) + \mathbb{E}\left(\frac{\partial \log f(w;\theta_0)}{\partial\theta}\frac{\partial \log f(w;\theta_0)}{\partial\theta'}\right) = 0$$

$$\implies \mathbb{E}\left(\frac{\partial^2 \log f(w;\theta_0)}{\partial\theta\partial\theta'}\right) = -\mathbb{E}\left(\frac{\partial \log f(w;\theta_0)}{\partial\theta}\frac{\partial \log f(w;\theta_0)}{\partial\theta'}\right).$$

The last line is famous as information matrix equality.

Now write

$$Q_n(\theta_0) = \frac{1}{n}\sum_{i=1}^{n}\log f(w_i;\theta_0) \implies \frac{\partial Q_n(\theta_0)}{\partial\theta} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial \log f(w;\theta_0)}{\partial\theta}.$$

But we just showed that $\mathbb{E}\left(\frac{\partial \log f(W;\theta_0)}{\partial\theta}\right) = 0$. We thus have

$$\sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial\theta} \xrightarrow{d} N\left(0, \mathbb{E}\left(\frac{\partial \log f(w;\theta_0)}{\partial\theta}\frac{\partial \log f(w;\theta_0)}{\partial\theta'}\right)\right)$$

by the CLT. This establishes assumption 4.

Extremum Estimation

Substituting these findings into the theorem, we get that

$$
\begin{aligned}
&\sqrt{n}(\hat{\theta} - \theta_0) \\
&\xrightarrow{d} N\left(0, \left(\underbrace{\mathbb{E}\left(\frac{\partial^2 \log f(w;\theta_0)}{\partial\theta\partial\theta'}\right)}_{\boldsymbol{H}}\right)^{-1} \underbrace{\mathbb{E}\left(\frac{\partial \log f(w;\theta_0)}{\partial\theta}\frac{\partial \log f(w;\theta_0)}{\partial\theta'}\right)}_{\boldsymbol{S}} \left(\underbrace{\mathbb{E}(\cdot)}_{\boldsymbol{H}}^{-1}\right)\right) \\
&= N(0, \boldsymbol{H}^{-1})
\end{aligned}
$$

using the information matrix equality.

Now, under our i.i.d. assumption, $\boldsymbol{H}$ is the (Fisher) information matrix $\mathbb{I}(\theta_0)$.

Thus, ML asymptotically attains the Cramer-Rao lower bound.

Indeed, it is known (but we will not show formally) that ML is asymptotically efficient in the sense of having the smallest asymptotic variance in a broad class of "regular" estimators.

This creates a strong case for using ML – assuming you are willing to specify a likelihood and can compute the ML estimator.

Comparing GMM and ML

- Whenever we have a complete likelihood, we can perform ML.
- But we could also do GMM! Knowledge of tyhe likelihood implies knowledge of moment conditions, certainly the "score equations" but possibly others.
- Can GMM match of beat the performance of ML?

Extremum Estimation

Comparing GMM and ML

- Whenever we have a complete likelihood, we can perform ML.
- But we could also do GMM! Knowledge of tyhe likelihood implies knowledge of moment conditions, certainly the "score equations" but possibly others.
- Can GMM match of beat the performance of ML?

Fact:

$$\left(\boldsymbol{G}'\boldsymbol{S}^{-1}\boldsymbol{G}\right)^{-1} - \mathbb{I}(\theta_0)^{-1} \text{ is positive semidefinite.}$$

$$\left(\boldsymbol{G}'\boldsymbol{S}^{-1}\boldsymbol{G}\right)^{-1} = \mathbb{I}(\theta_0)^{-1} \text{ if } g(w,\theta) = \frac{\partial \log f(w;\theta_0)}{\partial \theta}.$$

Hence:

- GMM cannot (asymptotically) beat ML estimation: $\text{avar}(\hat{\theta}_{GMM}) \geq \text{avar}(\hat{\theta}_{ML})$.
- If the likelihood is known, GMM can trivially match ML by mimicking it.
- But, since those moment conditions would reflect likelihood information, we cannot in general get ML efficiency without knowing the likelihood.

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:
  $LM = n\frac{\partial Q_n(\tilde{\theta})'}{\partial\theta}\tilde{\Sigma}^{-1}\frac{\partial Q_n(\tilde{\theta})}{\partial\theta}$,

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 $(\hat{\theta}, \tilde{\theta})$.

Extremum Estimation

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.

Extremum Estimation

- When invoking the result, we take on faith that $\sqrt{n}(\tilde{\theta} - \theta_0) = O_P(1)$.
  This follows from the theory of restricted estimators, which is very similar to what we already did (with some additional linearization/matrix algebra); see Hansen or Hayashi (notably Table 7.1).
  Alternatively, under current assumptions it follows from the aforementioned Argmax Theorem.
- We only spell out the details for Maximum Likelihood. For other estimators, $\boldsymbol{H}$ must be redefined. See in particular Hayashi (ch. 7, notation $\boldsymbol{\Psi}$).
- Assumptions 1,3, and 4 restate things we know for ML.
  However, it is instructive to disentangle their role in the proof.
- We do need that $\boldsymbol{H} = -\Sigma$ and therefore that ML is well-specified.
  We will return to what happens in misspecified models.

Proof of Theorem

The argument for the Wald statistic is much as before and we omit it.

The first-order condition of the constrained estimation problem can be written as

$$\begin{aligned}
\sqrt{n}\frac{\partial Q_n(\tilde{\theta})}{\partial\theta} + \sqrt{n}\boldsymbol{R}(\tilde{\theta})'\gamma_n &= 0 \\
\sqrt{n}r(\tilde{\theta}) &= 0.
\end{aligned}$$

Proof of Theorem

The argument for the Wald statistic is much as before and we omit it.

The first-order condition of the constrained estimation problem can be written as

$$\begin{aligned}\sqrt{n}\frac{\partial Q_n(\tilde{\theta})}{\partial \theta} + \sqrt{n}\boldsymbol{R}(\tilde{\theta})'\gamma_n &= 0 \\ \sqrt{n}r(\tilde{\theta}) &= 0.\end{aligned}$$

Use the Mean Value Theorem to write

$$\begin{aligned} r(\tilde{\theta}) &= r(\theta_0) + \boldsymbol{R}(\bar{\theta})\left(\tilde{\theta} - \theta_0\right) \\ \Longrightarrow \sqrt{n}r(\tilde{\theta}) &= \sqrt{n}\boldsymbol{R}(\bar{\theta})\left(\tilde{\theta} - \theta_0\right) \\ &= \underbrace{\sqrt{n}\left(\boldsymbol{R}(\bar{\theta}) - \boldsymbol{R}(\theta_0)\right)\left(\tilde{\theta} - \theta_0\right)}_{\xrightarrow{p} 0} + \sqrt{n}\boldsymbol{R}(\theta_0)\left(\tilde{\theta} - \theta_0\right) \\ &= \boldsymbol{R}(\theta_0) \cdot \sqrt{n}\left(\tilde{\theta} - \theta_0\right) + o_P(1). \end{aligned}$$

Next, a Taylor expansion of $\frac{\partial Q_n(\theta)}{\partial \theta}$ about θ_0 yields

$$\sqrt{n}\frac{\partial Q_n(\tilde{\theta})}{\partial \theta} = \underbrace{\sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial \theta}}_{\xrightarrow{d} N(0,\Sigma)} + \sqrt{n}\underbrace{\frac{\partial^2 Q_n(\theta_0)}{\partial \theta \partial \theta'}}_{\xrightarrow{p} \boldsymbol{H}}\left(\tilde{\theta} - \theta_0\right) + o_P(1).$$

The second and third assumption now imply that $\sqrt{n}\frac{\partial Q_n(\tilde{\theta})}{\partial \theta}$, and hence $\sqrt{n}\gamma_n$, are of order $O_P(1)$. This, in turn, allows us to write

$$\boldsymbol{R}(\tilde{\theta})'\sqrt{n}\gamma_n = \boldsymbol{R}(\theta_0)'\sqrt{n}\gamma_n + \left(\boldsymbol{R}(\tilde{\theta}) - \boldsymbol{R}(\theta_0)\right)'\sqrt{n}\gamma_n = \boldsymbol{R}(\theta_0)'\sqrt{n}\gamma_n + o_P(1)$$

by similar arguments as before.

Extremum Estimation

Now some collecting of terms. We have

$$\begin{aligned}
&\sqrt{n} r(\tilde{\theta}) = 0 \\
&\sqrt{n} r(\tilde{\theta}) = \boldsymbol{R}(\theta_0)\sqrt{n}\big(\tilde{\theta} - \theta_0\big) + o_P(1) \\
\implies \quad &\boldsymbol{R}(\theta_0)\sqrt{n}\big(\tilde{\theta} - \theta_0\big) = o_P(1)
\end{aligned}$$

as well as

$$\begin{aligned}
&\sqrt{n}\frac{\partial Q_n(\tilde{\theta})}{\partial \theta} + \sqrt{n}\boldsymbol{R}(\tilde{\theta})'\gamma_n = 0 \\
&\sqrt{n}\frac{\partial Q_n(\tilde{\theta})}{\partial \theta} = \sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial \theta} + \boldsymbol{H}\sqrt{n}\big(\tilde{\theta} - \theta_0\big) + o_P(1) \\
&\boldsymbol{R}(\tilde{\theta})'\sqrt{n}\gamma_n = \boldsymbol{R}(\theta_0)'\sqrt{n}\gamma_n + o_P(1) \\
\implies \quad &\boldsymbol{H}\sqrt{n}\big(\tilde{\theta} - \theta_0\big) + \boldsymbol{R}(\theta_0)'\sqrt{n}\gamma_n = -\sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial \theta} + o_P(1)
\end{aligned}$$

Counting equations, this should characterize the joint distribution of $\sqrt{n}(\tilde{\theta} - \theta_0)$ and $\sqrt{n}\gamma_n$. However, the characterization is rather implicit.

Extremum Estimation

We consolidate into (for brevity, we drop the argument of $\boldsymbol{R}$)

$$\begin{bmatrix} \boldsymbol{H} & \boldsymbol{R}' \\ \boldsymbol{R} & 0 \end{bmatrix} \begin{bmatrix} \sqrt{n}(\tilde{\theta} - \theta_0) \\ \sqrt{n}\gamma_n \end{bmatrix} = \begin{bmatrix} -\sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial \theta} \\ 0 \end{bmatrix} + o_P(1).$$

implying (by mechanical application of partitioned matrix inversion) that

$$\sqrt{n}\begin{bmatrix} \tilde{\theta} - \theta_0 \\ \gamma_n \end{bmatrix} = \begin{bmatrix} -\boldsymbol{H}^{-1} + \boldsymbol{H}^{-1}\boldsymbol{R}'\left(\boldsymbol{R}\boldsymbol{H}^{-1}\boldsymbol{R}'\right)^{-1}\boldsymbol{R}\boldsymbol{H}^{-1} \\ -\left(\boldsymbol{R}\boldsymbol{H}^{-1}\boldsymbol{R}'\right)^{-1}\boldsymbol{R}\boldsymbol{H}^{-1} \end{bmatrix} \sqrt{n}\frac{\partial Q_n\left(\theta_0\right)}{\partial \theta} + o_P(1).$$

We consolidate into (for brevity, we drop the argument of $\boldsymbol{R}$)

$$\begin{bmatrix} \boldsymbol{H} & \boldsymbol{R}' \\ \boldsymbol{R} & 0 \end{bmatrix} \begin{bmatrix} \sqrt{n}(\tilde{\theta}-\theta_0) \\ \sqrt{n}\gamma_n \end{bmatrix} = \begin{bmatrix} -\sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial \theta} \\ 0 \end{bmatrix} + o_P(1).$$

implying (by mechanical application of partitioned matrix inversion) that

$$\sqrt{n}\begin{bmatrix} \tilde{\theta}-\theta_0 \\ \gamma_n \end{bmatrix} = \begin{bmatrix} -\boldsymbol{H}^{-1} + \boldsymbol{H}^{-1}\boldsymbol{R}'\left(\boldsymbol{R}\boldsymbol{H}^{-1}\boldsymbol{R}'\right)^{-1}\boldsymbol{R}\boldsymbol{H}^{-1} \\ -\left(\boldsymbol{R}\boldsymbol{H}^{-1}\boldsymbol{R}'\right)^{-1}\boldsymbol{R}\boldsymbol{H}^{-1} \end{bmatrix} \sqrt{n}\frac{\partial Q_n\left(\theta_0\right)}{\partial \theta} + o_P(1).$$

This gets us to the LM statistic pretty quickly:

$$\begin{aligned} \sqrt{n}\gamma_n &= -\left(\boldsymbol{R}\boldsymbol{H}^{-1}\boldsymbol{R}'\right)^{-1}\boldsymbol{R}\boldsymbol{H}^{-1}\sqrt{n}\frac{\partial Q_n\left(\theta_0\right)}{\partial \theta} + o_P(1) \\ &\xrightarrow{d} N\left(0, \left(\boldsymbol{R}\boldsymbol{H}^{-1}\boldsymbol{R}'\right)^{-1}\boldsymbol{R}\boldsymbol{H}^{-1}\Sigma\boldsymbol{H}^{-1}\boldsymbol{R}'\left(\boldsymbol{R}\boldsymbol{H}^{-1}\boldsymbol{R}'\right)^{-1}\right) \\ &= N\left(0, \left(\boldsymbol{R}\Sigma^{-1}\boldsymbol{R}'\right)^{-1}\right) \end{aligned}$$

Extremum Estimation

We consolidate into (for brevity, we drop the argument of $\boldsymbol{R}$)

$$\left[\begin{array}{cc} \boldsymbol{H} & \boldsymbol{R}' \\ \boldsymbol{R} & 0 \end{array}\right]\left[\begin{array}{c} \sqrt{n}(\tilde{\theta}-\theta_0) \\ \sqrt{n}\gamma_n \end{array}\right]=\left[\begin{array}{c} -\sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial\theta} \\ 0 \end{array}\right]+o_P(1).$$

implying (by mechanical application of partitioned matrix inversion) that

$$\sqrt{n}\left[\begin{array}{c} \tilde{\theta}-\theta_0 \\ \gamma_n \end{array}\right]=\left[\begin{array}{c} -\boldsymbol{H}^{-1}+\boldsymbol{H}^{-1}\boldsymbol{R}'\left(\boldsymbol{R}\boldsymbol{H}^{-1}\boldsymbol{R}'\right)^{-1}\boldsymbol{R}\boldsymbol{H}^{-1} \\ -\left(\boldsymbol{R}\boldsymbol{H}^{-1}\boldsymbol{R}'\right)^{-1}\boldsymbol{R}\boldsymbol{H}^{-1} \end{array}\right]\sqrt{n}\frac{\partial Q_n\left(\theta_0\right)}{\partial\theta}+o_P(1).$$

This gets us to the LM statistic pretty quickly:

$$\begin{aligned}
\sqrt{n}\gamma_n &= -\left(\boldsymbol{R}\boldsymbol{H}^{-1}\boldsymbol{R}'\right)^{-1}\boldsymbol{R}\boldsymbol{H}^{-1}\sqrt{n}\frac{\partial Q_n\left(\theta_0\right)}{\partial\theta}+o_P(1) \\
&\xrightarrow{d} N\left(0,\left(\boldsymbol{R}\boldsymbol{H}^{-1}\boldsymbol{R}'\right)^{-1}\boldsymbol{R}\boldsymbol{H}^{-1}\Sigma\boldsymbol{H}^{-1}\boldsymbol{R}'\left(\boldsymbol{R}\boldsymbol{H}^{-1}\boldsymbol{R}'\right)^{-1}\right) \\
&= N\left(0,\left(\boldsymbol{R}\Sigma^{-1}\boldsymbol{R}'\right)^{-1}\right) \\
\Longrightarrow &\quad \sqrt{n}\gamma_n'\boldsymbol{R}\Sigma^{-1}\boldsymbol{R}'\sqrt{n}\gamma_n\xrightarrow{d}\chi^2_{\#r}.
\end{aligned}$$

Extremum Estimation

We conclude with another Mean Value Theorem expansion:

$$Q_n(\tilde\theta) = Q_n(\hat\theta) + \frac{\partial Q_n(\hat\theta)}{\partial\theta}\left(\tilde\theta - \hat\theta\right) + \frac{1}{2}\left(\tilde\theta - \hat\theta\right)' \frac{\partial^2 Q_n(\bar\theta)}{\partial\theta\partial\theta'}\left(\tilde\theta - \hat\theta\right),$$

but $\frac{\partial Q_n(\hat\theta)}{\partial\theta} = 0$ (with high probability) and $\frac{\partial^2 Q_n(\bar\theta)}{\partial\theta\partial\theta'} \xrightarrow{p} \boldsymbol{H}$. Conclude

$$\begin{aligned} 2n\big(Q_n(\hat\theta) - Q_n(\tilde\theta)\big) &= -\sqrt{n}\big(\tilde\theta - \hat\theta\big)'\big(\boldsymbol{H} + o_P(1)\big)\sqrt{n}\big(\tilde\theta - \hat\theta\big) \\ &= -\sqrt{n}\big(\tilde\theta - \hat\theta\big)'\boldsymbol{H}\sqrt{n}\big(\tilde\theta - \hat\theta\big) + o_P(1). \end{aligned}$$

Extremum Estimation

We conclude with another Mean Value Theorem expansion:

$$Q_n(\tilde{\theta}) = Q_n(\hat{\theta}) + \frac{\partial Q_n(\hat{\theta})}{\partial \theta}(\tilde{\theta} - \hat{\theta}) + \frac{1}{2}(\tilde{\theta} - \hat{\theta})' \frac{\partial^2 Q_n(\bar{\theta})}{\partial \theta \partial \theta'}(\tilde{\theta} - \hat{\theta}),$$

but $\frac{\partial Q_n(\hat{\theta})}{\partial \theta} = 0$ (with high probability) and $\frac{\partial^2 Q_n(\bar{\theta})}{\partial \theta \partial \theta'} \xrightarrow{p} \boldsymbol{H}$. Conclude

$$\begin{aligned} 2n\big(Q_n(\hat{\theta}) - Q_n(\tilde{\theta})\big) &= -\sqrt{n}\big(\tilde{\theta} - \hat{\theta}\big)'\big(\boldsymbol{H} + o_P(1)\big)\sqrt{n}\big(\tilde{\theta} - \hat{\theta}\big) \\ &= -\sqrt{n}\big(\tilde{\theta} - \hat{\theta}\big)'\boldsymbol{H}\sqrt{n}\big(\tilde{\theta} - \hat{\theta}\big) + o_P(1). \end{aligned}$$

Substitute in from Assumption 1 respectively the big matrix equation to write

$$\begin{aligned} &\sqrt{n}\big(\tilde{\theta} - \hat{\theta}\big) \\ =&\sqrt{n}\big(\tilde{\theta} - \theta_0\big) - \sqrt{n}\big(\hat{\theta} - \theta_0\big) \\ =&-\big(\boldsymbol{H}^{-1} - \boldsymbol{H}^{-1}\boldsymbol{R}'(\boldsymbol{R}\boldsymbol{H}^{-1}\boldsymbol{R}')^{-1}\boldsymbol{R}\boldsymbol{H}^{-1}\big)\sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial \theta} + \boldsymbol{H}^{-1}\sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial \theta} + o_P(1) \\ =&\boldsymbol{H}^{-1}\boldsymbol{R}'(\boldsymbol{R}\boldsymbol{H}^{-1}\boldsymbol{R}')^{-1}\boldsymbol{R}\boldsymbol{H}^{-1}\sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial \theta} + o_P(1). \end{aligned}$$

This determines the distribution of LR. The rest is algebra.

Extremum Estimation

Combine the previous slide's displays to find (here $\approx$ absorbs $o_P(1)$)

$$
\begin{aligned}
& 2n\big(Q_n(\hat{\boldsymbol{\theta}}) - Q_n(\tilde{\boldsymbol{\theta}})\big) \\
\approx\ & -\sqrt{n}\frac{\partial Q_n(\theta_0}{\partial \theta'}\big(\boldsymbol{H}^{-1}\boldsymbol{R}'(\boldsymbol{R}\boldsymbol{H}^{-1}\boldsymbol{R}')^{-1}\boldsymbol{R}\boldsymbol{H}^{-1}\big)'\boldsymbol{H}\boldsymbol{H}^{-1}\boldsymbol{R}'(\boldsymbol{R}\boldsymbol{H}^{-1}\boldsymbol{R}')^{-1}\boldsymbol{R}\boldsymbol{H}^{-1}\sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial \theta} \\
=\ & -\sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial \theta'}\boldsymbol{H}^{-1}\boldsymbol{R}'(\boldsymbol{R}\boldsymbol{H}^{-1}\boldsymbol{R}')^{-1}\boldsymbol{R}\boldsymbol{H}^{-1}\sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial \theta'} \\
=\ & \sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial \theta'}\Sigma^{-1}\boldsymbol{R}'(\boldsymbol{R}\Sigma^{-1}\boldsymbol{R}')^{-1}\boldsymbol{R}\Sigma^{-1}\sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial \theta'}.
\end{aligned}
$$

Combine the previous slide's displays to find (here $\approx$ absorbs $o_P(1)$)

$$
\begin{aligned}
& 2n\big(Q_n(\hat{\boldsymbol{\theta}}) - Q_n(\tilde{\boldsymbol{\theta}})\big) \\
\approx\ & -\sqrt{n}\frac{\partial Q_n(\theta_0}{\partial \theta'}\big(\boldsymbol{H}^{-1}\boldsymbol{R}'(\boldsymbol{R}\boldsymbol{H}^{-1}\boldsymbol{R}')^{-1}\boldsymbol{R}\boldsymbol{H}^{-1}\big)'\boldsymbol{H}\boldsymbol{H}^{-1}\boldsymbol{R}'(\boldsymbol{R}\boldsymbol{H}^{-1}\boldsymbol{R}')^{-1}\boldsymbol{R}\boldsymbol{H}^{-1}\sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial \theta} \\
=\ & -\sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial \theta'}\boldsymbol{H}^{-1}\boldsymbol{R}'(\boldsymbol{R}\boldsymbol{H}^{-1}\boldsymbol{R}')^{-1}\boldsymbol{R}\boldsymbol{H}^{-1}\sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial \theta'} \\
=\ & \sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial \theta'}\Sigma^{-1}\boldsymbol{R}'(\boldsymbol{R}\Sigma^{-1}\boldsymbol{R}')^{-1}\boldsymbol{R}\Sigma^{-1}\sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial \theta'}.
\end{aligned}
$$

Recalling again that $\sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial \theta} \xrightarrow{d} N(0, \Sigma)$, we have

$$
\begin{aligned}
\boldsymbol{R}\Sigma^{-1}\sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial \theta} \quad & \xrightarrow{d} \quad N\big(0, \boldsymbol{R}\Sigma^{-1}\Sigma\Sigma^{-1}\boldsymbol{R}'\big) = N\big(0, \boldsymbol{R}\Sigma^{-1}\boldsymbol{R}'\big) \\
\Longrightarrow\ LR \quad & \xrightarrow{d} \quad \chi^2_{\#r}.
\end{aligned}
$$

Extremum Estimation

Why is it called Likelihood Ratio Statistic?

- If we take the likelihood literally, then

$$n(Q_n(\hat{\theta}) - Q_n(\tilde{\theta}) = \sum_{i=1}^{n} \ell(\hat{\theta}) - \sum_{i=1}^{n} \ell(\tilde{\theta}) = \frac{f(W_1, \ldots, W_n; \hat{\theta})}{f(W_1, \ldots, W_n; \tilde{\theta})}.$$

- The additional factor of 2 aligns the statistic with others.
- But we see that the interpretation of $Q_n(\cdot)$ as likelihood is not essential!

Visualization

Suppose $\Theta = \mathbb{R}^2$ and $\boldsymbol{H} = \boldsymbol{I}_2$. Then aspects of the result can be visualized as orthogonal decomposition in the linearized constrained estimation problem.