Advanced Econometrics I

Jürgen Meinecke

Lecture 12 of 12

Research School of Economics, Australian National University

Roadmap

Many estimators share a common structure

They are members of a broader class

This commonality can be useful for deriving consistency and asymptotic normality for a broader class of estimators

Instead of deriving asymptotic results for individual estimators, we only need to establish it once for the broader class

The class of estimators that we are looking at today are those estimators that maximize an objective function that depends on data and sample size

This class of estimators is called *extremum estimators* or *M-estimators*

Today's lecture is a summary of Newey and McFadden, *Handbook of Econometrics*, chapter 36

Wooldridge's graduate textbook also has a good discussion

The nice thing about today's lecture is that it brings together all our estimators from earlier in the semester

I will show how they can be regarded as special cases of a more general framework

We will establish consistency and asymptotic normality results for this more general framework

As a result, consistency and asymptotic normality follow for the special cases

Another nice thing of today's lecture is that it builds you a bridge to EMET8008

Thomas will teach you GMM estimation which is closely related to the framework presented here

(Unfortunately, EMET8008 won't be offered next semester)

Definition (Extremum Estimator)

Given data $W_i, i = 1, \dots, N$, an estimator $\hat{\theta}^{\text{EE}}$ is called **extremum estimator** if there is an objective function $Q_N(W_i, \theta)$ such that $\hat{\theta}^{\text{EE}} := \underset{\theta \in \Theta}{\operatorname{argmax}} Q_N(W_i, \theta)$.

Often the objective function is a sample average:

$$Q_N(W_i, \theta) = \frac{1}{N} \sum_{i=1}^{N} q(W_i, \theta)$$

Huber (1981) calls such estimators "maximum likelihood type" estimators, which justifies the next definition

Definition (M-Estimator)

$\hat{\theta}^{\text{M}} := \underset{\theta \in \Theta}{\operatorname{argmax}} \frac{1}{N} \sum_{i=1}^{N} q(W_i, \theta)$.

Aside: Wooldridge defines $\hat{\theta}^{\text{M}}$ as a *minimizer* instead (irrelevant, can simply put a minus sign in front)

The class of extremum estimators contains the class of M-estimators

All our estimators from this semester satisfy this general structure

They are extremum estimators and M-estimators

$$\hat{\theta}^{\text{EE}} := \underset{\theta \in \Theta}{\operatorname{argmax}} \underbrace{\frac{1}{N} \sum_{i=1}^{N} q(W_i, \theta)}_{Q_N(W_i, \theta)}$$

with

OLS	$q(W_i, \theta) = -(Y_i - X_i'\theta)^2$
IV	$q(W_i, \theta) = -(Y_i - \hat{\pi}' Z_i \theta)^2$
ML	$q(W_i, \theta) = \ln f_Y(Y_i \mid \theta)$

The asymptotic properties of $\hat{\theta}^{\text{EE}}$ depend on the limit behavior of $Q_N(W_i, \theta)$

Using the WLLN, let's look at the probability limit of $Q_N(W_i, \theta)$

Denote it by $Q_0(\theta)$: $Q_N(W_i, \theta) \xrightarrow{p} Q_0(\theta)$

In our applications:

OLS	$Q_0(\theta) := -\mathsf{E}(Y_i - X_i'\theta)^2$
IV	$Q_0(\theta) := -\mathsf{E}(Y_i - \pi' Z_i \theta)^2$
ML	$Q_0(\theta) := \mathsf{E}\left(\ln f_Y(Y_i \vert \theta)\right)$

While we have pointwise convergence at all $\theta \in \Theta$ by the WLLN, for a generic consistency proof we need something stronger...

Roadmap

Extremum Estimators, M-Estimation

- Motivation
- Consistency
- Asymptotic Distribution
- Asymptotic Variance Estimation
- Last Slide

Definition (Uniform Convergence in Probability)

A sequence $Q_N(W_i, \theta)$ is said to **converge uniformly in probability** to $Q_0(\theta)$ if

$$\sup_{\theta \in \Theta} \left| Q_N(W_i, \theta) - Q_0(\theta) \right| = o_p(1).$$

Uniform implies pointwise convergence, not vice versa

Theorem (Consistency of Extremum Estimators)

If there is a function $Q_0(\theta)$ such that

(i) *$Q_0(\theta)$ is uniquely maximized at θ_0;*

(ii) *Θ is compact;*

(iii) *$Q_0(\theta)$ is continuous;*

(iv) *$Q_N(W_i, \theta)$ converges uniformly in probability to $Q_0(\theta)$;*

then $\hat{\theta}^{EE} = \theta_0 + o_p(1)$.

Sketch of proof: consistency
(to get a rough idea how these four conditions are used)

- uniform convergence, roughly, means that Q_N is very similar to Q_0 when N is large
- $\hat{\theta}$ as the maximizer of Q_N is then also a maximizer of Q_0
- but then, for every $\varepsilon > 0$, $Q_0(\hat{\theta}) > Q_0(\theta_0) - \varepsilon$ $(\star)$
- we still need to show that this implies that $\hat{\theta}$ is close to θ_0

- from previous slide: $Q_0(\hat{\theta}) > Q_0(\theta_0) - \varepsilon \ (\star)$
- pick any closed subset of Θ not containing θ_0
  (remove an open set containing θ_0 from Θ, the set that remains is compact because Θ is compact)
- by (iii) we can use Weierstrass' theorem:
  a continuous function over a compact set attains a maximum
- denote this maximum (over the compact subset) by $Q_0(\theta^*)$
- by (i): $Q_0(\theta_0) - Q_0(\theta^*) > 0$
- set $\varepsilon = Q_0(\theta_0) - Q_0(\theta^*)$ in $(\star)$, therefore $Q_0(\hat{\theta}) > Q_0(\theta^*)$
- then $\hat{\theta}$ must be from the open set containing θ_0
- it follows that $\hat{\theta} \xrightarrow{\text{p}} \theta_0$

(for details consult proof of Theorem 2.1 in Newey and McFadden)

The theorem gives four conditions

How should we interpret them?

Distinguish *substantive* conditions from *regularity* conditions

Conditions (i) and (ii) are substantive, meaning they could be restrictive in some applications

Conditions (iii) and (iv) are regularity conditions, meaning they are satisfied in many applications

In addition, we can think of conditions to be *primitive* or *non-primitive*

A condition is called primitive, if it is easy to interpret
(akin to high-level programming languages)

Only one of the four conditions is primitive: the compactness condition

In practice, condition (ii) is rarely checked

If you are estimating a probability, then compactness is easily verified

But if you run a regression, then you are implicitly thinking that your coefficients come from a closed and bounded set

Condition (i) is substantive and non-primitive

It has to be checked on a case by case basis

For the maximum likelihood model, we actually have checked condition (i) already

Recall from the week 10 lecture, that for $\theta \neq \theta_0$,

$$\mathrm{E}\left(\ln f_Y(Y|\theta_0)\right) > \mathrm{E}\left(\ln f_Y(Y|\theta)\right)$$

This means that the expected value of the log likelihood is maximized at the *true* value of the parameter

In other applications, condition (i) may not hold

What about conditions (iii) and (iv)?

They are non-primitive regularity conditions

While they are not easy to interpret, we believe that they may be easily satisfied nevertheless

The following Lemma helps when Q_N is based on a sample average, that is, when we consider M-estimators

Lemma (Uniform Law of Large Numbers for M-Estimators)

Let $Q_N(W_i, \theta) = \frac{1}{N}\sum_{i=1}^{N} q(W_i, \theta)$ *and* $Q_0(\theta) = E\left(q(W_i, \theta)\right)$. *If*

(i) *the sample data* W_i *are iid;*

(ii) Θ *is compact;*

(iii) $q(W_i, \cdot)$ *is continuous at each* $\theta \in \Theta$ *with probability one;*

(iv) *there exists a dominating function* $d(W_i)$ *with* $E\left(d(W_i)\right) < \infty$, *such that* $|q(W_i, \theta)| \leq d(W_i)$ *for all* $\theta \in \Theta$;

then

- $Q_N(W_i, \theta)$ *converges uniformly in probability to* $Q_0(\theta)$, *and*
- $Q_0(\theta)$ *is continuous.*

The conditions in the Lemma are all primitive and also quite weak

Notably, $q(W_i, \cdot)$ can be discontinuous on a set of measure zero

Conditions (iii) and (iv), both on q, are reasonably easy to check

Condition (iii) can be checked by inspection and condition (iv) boils down to existence of moments

OLS example (scalar X_i)

- let β be from the compact set $\left[\underline{b}, \bar{b}\right]$
- $q(W_i, \theta) = -(Y_i - X_i\beta)^2$, clearly continuous
- dominating function

$$\begin{aligned} |q(X_i, Y_i, \beta)| &:= (Y_i - X_i\beta)^2 = 2Y_i^2 + 2\beta^2 X_i^2 - (Y_i + X_i\beta)^2 \\ &\leq 2Y_i^2 + 2\beta^2 X_i^2 \\ &\leq 2Y_i^2 + 2\bar{b}^2 X_i^2 =: d(X_i, Y_i) \end{aligned}$$

- finite second moments of X_i and Y_i imply $\mathrm{E}(d(X_i, Y_i)) < \infty$

Connecting the dots:

To establish consistency of an extremum estimator, we need to check the four conditions of the Lemma plus the identification condition of the theorem

Roadmap

Extremum Estimators, M-Estimation

- Motivation
- Consistency
- Asymptotic Distribution
- Asymptotic Variance Estimation
- Last Slide

For the class of extremum estimators, there also exists a generic result on the asymptotic distribution

Conditions to establish asymptotic normality are a bit involved

However, if we restrict attention to extremum estimators with objective functions that are sufficiently smooth, then we can use an intuitive approximation via the mean value theorem

All we need is for Q_N to be twice continuously differentiable

Sketch of proof: asymptotic normality
(this is similar to the asymptotic distribution of MLE in week 10)

Notice, by definition, $\nabla_\theta Q_N(W_i, \hat{\theta}^{\text{EE}}) = 0$

Apply the mean value theorem around θ_0

$$0 = \nabla_\theta Q_N(W_i, \hat{\theta}^{\text{EE}}) = \nabla_\theta Q_N(W_i, \theta_0) + \left(\nabla_{\theta\theta} Q_N(W_i, \tilde{\theta})\right)\left(\hat{\theta}^{\text{EE}} - \theta_0\right),$$

where $\tilde{\theta}$ is between θ_0 and $\hat{\theta}^{\text{EE}}$

Multiplying through by $\sqrt{N}$ and solving for $\sqrt{N}\left(\hat{\theta}^{\text{EE}} - \theta_0\right)$,

$$\begin{aligned}
\sqrt{N}\left(\hat{\theta}^{\text{EE}} - \theta_0\right) &= -\left(\nabla_{\theta\theta} Q_N(W_i, \tilde{\theta})\right)^{-1} \cdot \sqrt{N}\, \nabla_\theta Q_N(W_i, \theta_0) \\
&\qquad\qquad \downarrow_p \text{ (WLLN)} \qquad\qquad \downarrow_d \text{ (CLT)} \\
&= \qquad H^{-1} \qquad\qquad \cdot \qquad \mathcal{N}(0, \Sigma) \\
&\xrightarrow{\text{d}} \mathcal{N}\left(0, H^{-1}\Sigma H^{-1}\right)
\end{aligned}$$

(also using Slutsky's theorem)

The formal result is

Theorem (Asymptotic Normality of Extremum Estimators)

Let $\hat{\theta}^{EE}$ be an extremum estimator such that $\hat{\theta}^{EE} = \theta + o_p(1)$. If

(i) $\theta_0 \in$ *interior*(Θ);

(ii) $Q_N(W_i, \theta)$ *is twice continuously differentiable in a neighborhood* T *of* θ_0;

(iii) $\sqrt{N}\nabla_\theta Q_N(\theta_0) \xrightarrow{d} \mathcal{N}(0, \Sigma)$;

(iv) *there exists* $H(\theta)$ *that is continuous at* θ_0 *and* $\sup_{\theta \in T} |\nabla_{\theta\theta} Q_N(W_i, \theta) - H(\theta)| = o_p(1)$

(v) $H := H(\theta_0)$ *is nonsingular;*

then $\sqrt{N}\left(\hat{\theta}^{EE} - \theta_0\right) \xrightarrow{d} \mathcal{N}\left(0, H^{-1}\Sigma H^{-1}\right)$.

Condition (i) is necessary because estimators may not be asymptotically normal when θ_0 sits at the boundary of Θ (this is not easy to show)

Condition (ii) is stronger than necessary, but convenient to obtain the Hessian matrix as the probability limit

Condition (iii) is easy to establish for M-estimators with differentiable q (such as OLS, IV, and MLE)

Condition (iv) is a U-WLLN for the Hessian

Condition (v) makes sure that the Hessian is invertible

Roadmap

Extremum Estimators, M-Estimation

- Motivation
- Consistency
- Asymptotic Distribution
- Asymptotic Variance Estimation
- Last Slide

For M-estimators in which q is differentiable, it is easy to estimate the asymptotic variance $H^{-1}\Sigma H^{-1}$

You can simply use plug-in estimators for H and Σ

Consistent estimators are

$$\hat{H} = \nabla_{\theta\theta} Q_N(W_i, \hat{\theta}^{\mathsf{EE}}) = (1/N) \sum_{i=1}^{N} \nabla_{\theta\theta} q(W_i, \hat{\theta}^{\mathsf{EE}})$$

$$\hat{\Sigma} = (1/N) \sum_{i=1}^{N} \nabla_{\theta} q(W_i, \hat{\theta}^{\mathsf{EE}}) \nabla_{\theta} q(W_i, \hat{\theta}^{\mathsf{EE}})'$$

Roadmap

Extremum Estimators, M-Estimation

- Motivation
- Consistency
- Asymptotic Distribution
- Asymptotic Variance Estimation
- Last Slide

Last Slide (yey!)

I hope you ...

- enjoyed it a bit
  (didn't hate it too much!?)
- learned something useful

Feel free to come by my office for a chat anytime!

Good luck with all your exams!