

Lecture Note 2: Extremum Estimators and GMM

(Based on Newey and McFadden 1994)

Extremum Estimator: $\hat{\theta}$ maximizes $\hat{Q}_n(\theta)$ over $\theta \in \Theta$.

Here Θ is the set of possible parameter values. We use a “hat” in \hat{Q}_n to indicate that the objective function can depend on data, and the n subscript indicates that it can depend on sample size.

It turns out that many estimators can be viewed as extremum estimators. We’ll look at a number of examples next.

Maximum Likelihood Estimator

Let z_1, z_2, \dots, z_n be IID with PDF $f(z|\theta_0)$ for $\theta_0 \in \Theta$. Then the MLE solves

$$\max_{\theta} \frac{1}{n} \sum_{i=1}^n \log f(z_i|\theta).$$

So this is an extremum estimator with

$$\hat{Q}_n(\theta) = \frac{1}{n} \sum_{i=1}^n \log f(z_i|\theta).$$

We can also handle conditional MLE. Suppose $z_i = (y_i, x_i)$, and let $f(z_i|\theta)$ denote the conditional density of y given x : $f(y_i|x_i, \theta)$.

Example: logit regression

Suppose y_i is binary with

$$P(y_i = 1|x_i) = \frac{\exp(x_i'\beta)}{1 + \exp(x_i'\beta)}.$$

Then the conditional likelihood function is

$$P(y_1, \dots, y_n|x_1, \dots, x_n, \beta) = \prod_{i=1}^n \left(\frac{\exp(x_i'\beta)}{1 + \exp(x_i'\beta)} \right)^{y_i} \left(\frac{1}{1 + \exp(x_i'\beta)} \right)^{1-y_i}.$$

The MLE $\hat{\beta}$ maximizes the likelihood. Equivalently, we can maximize the log likelihood:

$$\log P(y_1, \dots, y_n|x_1, \dots, x_n, \beta) = \sum_{i=1}^n y_i \log \frac{\exp(x_i'\beta)}{1 + \exp(x_i'\beta)} + (1 - y_i) \log \frac{1}{1 + \exp(x_i'\beta)}.$$

We can also multiply the log-likelihood by $1/n$ to get

$$\hat{Q}_n(\beta) = \frac{1}{n} \sum_{i=1}^n y_i \log \frac{\exp(x_i'\beta)}{1 + \exp(x_i'\beta)} + (1 - y_i) \log \frac{1}{1 + \exp(x_i'\beta)},$$

and

$$\hat{\beta}_{MLE} = \arg \max_{\beta} \hat{Q}_n(\beta).$$

Although in most cases we cannot solve for the value of $\hat{\beta}$ by hand, the log-likelihood function can be shown to be globally concave, so relatively simple numerical methods can be used to find the MLE.

Nonlinear Least Squares

Let $z_i = (y_i, x_i)$, and suppose that

$$E[y|x] = h(x, \theta_0).$$

For example, we might have

$$E[y|x] = \exp(x'\theta_0).$$

The nonlinear least squares (NLS) estimator $\hat{\theta}$ minimizes

$$\sum_{i=1}^n (y_i - h(x_i, \theta))^2.$$

This is equivalent to maximizing

$$\hat{Q}_n(\theta) = -\frac{1}{n} \sum_{i=1}^n (y_i - h(x_i, \theta))^2.$$

LAD estimator

Suppose that y is a random variable. The median of y , denoted $Med(y)$, is any number c such that

$$P(y \leq c) \geq 1/2, \quad P(y \geq c) \geq 1/2.$$

The idea is that half the probability mass of y is above c , and half the probability mass is below c . A useful result is that the median of y solves

$$\min_c E[|y - c|].$$

Suppose that the median of y given x has the form

$$Med(y|x) = m(x, \theta_0),$$

for a known function m . For example, if the conditional median is a linear function of x ,

$$Med(y|x) = x'\theta_0.$$

A natural estimator based on a random sample of (y_i, x_i) , $i = 1, 2, \dots, n$, is the least absolute

deviations estimator, defined as

$$\hat{\theta} = \arg \min_{\theta} \frac{1}{n} \sum_{i=1}^n |y_i - m(x_i, \theta)|.$$

(or maximizing the negative of the sum of absolute deviations).

GMM Estimator

Suppose we have a “moment function” $g(z, \theta)$ such that, at the “true” θ_0 ,

$$E[g(z, \theta_0)] = 0$$

Here $g(z, \theta)$ is a k -vector valued function so that the preceding display should be interpreted as k equality restrictions.

Let \hat{W} be a positive semidefinite matrix. So for a vector m ,

$$(m' \hat{W} m)^{1/2}$$

can be thought of as a “distance” of the vector m from 0. The GMM estimator maximizes

$$\hat{Q}_n(\theta) = - \left[\frac{1}{n} \sum_{i=1}^n g(z_i, \theta) \right]' \hat{W} \left[\frac{1}{n} \sum_{i=1}^n g(z_i, \theta) \right].$$

The interpretation is that $\hat{\theta}$ tries to minimize the distance of

$$\frac{1}{n} \sum_{i=1}^n g(z_i, \theta)$$

from 0; that is, the GMM estimator tries to set the sample version of $E[g(z, \theta)]$ as close as possible to 0.

Example of GMM: Linear IV

Suppose $z_i = (y_i, x_i, v_i)$, where

$$y_i = x_i' \theta_0 + \epsilon_i,$$

but x_i may be correlated with ϵ_i . The variable v_i is an instrumental variable:

$$E[v_i \epsilon_i] = 0.$$

Rewrite this as

$$E[v_i (y_i - x_i' \theta_0)] = 0.$$

So we can take

$$g(z, \theta) = v(y - x'\theta).$$

Suppose that both v and x are k -vectors. Then $g(z, \theta)$ is $k \times 1$, as is θ . So we can typically find a value $\hat{\theta}$ that exactly solves

$$\frac{1}{n} \sum_{i=1}^n g(z_i, \theta) = 0.$$

Then, this will be the solution to the GMM problem for any matrix \hat{W} . If $\dim(v) > \dim(x)$, however, it will typically not be possible to choose a θ to set $\frac{1}{n} \sum_{i=1}^n g(z_i, \theta)$ exactly equal to 0. Different choices for \hat{W} will lead to different solutions to the GMM problem.

A popular choice for \hat{W} is

$$\hat{W} = \left(\frac{1}{n} \sum_{i=1}^n v_i v_i' \right)^{-1}.$$

The solution for this weighting matrix turns out to be the two-stage least squares (TSLS) estimator.

Exercise: show that this choice for \hat{W} leads to TSLS.

Another example of GMM: Euler Equations (Hansen and Singleton, 1982)

Suppose a consumer is choosing a consumption stream c_1, \dots, c_T to maximize expected utility, under a constant relative risk aversion utility function

$$u(c) = \frac{c^\gamma - 1}{\gamma}.$$

The consumer's maximization problem is

$$\max_{c_1, \dots, c_T} E \left[\sum_{t=1}^T \beta^{t-1} u(c_t) \right]$$

subject to a dynamic budget constraint. Here β is the rate of time preference.

The first-order (Euler) conditions for a maximum are

$$E \left[\beta \left(\frac{c_{t+1}}{c_t} \right)^{\gamma-1} - 1 | \mathcal{I}_t \right] = 0,$$

for all t , where \mathcal{I}_t denotes the information available to the agent at time t .

Let $\theta = (\beta, \gamma)$ and

$$\rho(z_t, \theta) = \beta \left(\frac{c_{t+1}}{c_t} \right)^{\gamma-1} - 1.$$

So we have

$$E[\rho(z_t, \theta_0) | \mathcal{I}_t] = 0.$$

Now, suppose x_t are variables that are included in the information set at time t . For example, it could include lagged values of the consumption variable and other lagged variables such as income. Then

$$E[x_t \rho(z_t, \theta)] = 0.$$

So we can define

$$g(z_t, \theta) = x_t \rho(z_t, \theta),$$

and we have that

$$E[g(z_t, \theta_0)] = 0.$$

This is a nonlinear version of the previous instrumental variables problem.

Consistency of Extremum Estimators

Heuristic example: consider the maximum likelihood estimator,

$$\hat{\theta} = \arg \max_{\theta} \frac{1}{n} \sum_{i=1}^n \log f(z_i | \theta).$$

The criterion function has the form of a sample average, and by a law of large numbers converges in probability to

$$Q(\theta) = E[\log f(z_i | \theta)] = \int \log f(z | \theta) f(z | \theta_0) dz.$$

You may recall from Econ 520, that $Q(\theta) = E[\log f(z_i | \theta)]$ is maximized at the true value of θ . When the parameter is identified, θ_0 is the unique maximizer of $Q(\theta)$, i.e.

$$\theta_0 = \arg \max_{\theta} Q(\theta).$$

So it seems plausible that the maximizer of $\hat{Q}_n(\theta)$ would converge to the maximizer of $Q(\theta)$. For this to hold, we need some further conditions, in particular a “uniform” notion of convergence in probability.

Uniform convergence in probability: the function $\hat{Q}_n(\theta)$ converges uniformly in probability to $Q(\theta)$ if

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

Theorem: if there is a function $Q(\theta)$ such that

- (i) $Q(\theta)$ is uniquely maximized at θ_0 ,
 - (ii) Θ is compact,
 - (iii) $Q(\theta)$ is continuous,
 - (iv) $\hat{Q}_n(\theta)$ converges uniformly in probability to $Q(\theta)$,
- then $\hat{\theta} \xrightarrow{p} \theta_0$.

Proof: see Newey and McFadden.

Remark: in some cases, particularly simulation-based estimators, it is not always feasible to find the true maximizer of \hat{Q}_n . Suppose we instead have a “near-maximizer,” in the sense that $\hat{\theta}$ satisfies

$$\hat{Q}_n(\hat{\theta}) \geq \sup_{\theta \in \Theta} \hat{Q}_n(\theta) + o_p(1).$$

Then the previous theorem’s conclusion continues to hold.

Some of the conditions in the theorem above are straightforward to check (for example, compactness of the parameter space is usually assumed). Others require more work. To show conditions (iii) and (iv), the following “uniform law of large numbers” is very handy:

Uniform Law of Large Numbers: suppose that z_i are IID, Θ is compact, and we are given a function $a(z, \theta)$ such that

(i) for each $\theta \in \Theta$, $a(z, \theta)$ is continuous in z with probability one.

(ii) There is a function $d(z)$ with $\|a(z, \theta)\| \leq d(z)$ for all $\theta \in \Theta$, and $E[d(z)] < \infty$.

Then $E[a(z, \theta)]$ is continuous and

$$\sup_{\theta \in \Theta} \left\| \frac{1}{n} \sum_{i=1}^n a(z_i, \theta) - E[a(z, \theta)] \right\| \xrightarrow{p} 0.$$