14.385

Nonlinear Econometrics

Lecture 1. Basic Overview of Some Principal Methods.

Main Methods to be Covered in the Course:

1. Moment Methods (Generalized Method of Moments, Nonlinear IV, Z-estimation), including **moment inequalities** (later in the course).

2. Extremum Methods (MLE, M-estimators, quantile regression, minimum distance, GMM).

3. Bayesian and Quasi-Bayesian Methods.

All methods are interrelated.

Example 1: Optimizing behavior of economic agents.

A representative agent maximizes expected utility from consumption, Hansen and Singleton (1982). Euler (first-order) conditions for optimizing behavior imply a **conditional moment restriction**

$$E[\rho(z_t, \theta_0)|x_t] = 0,$$

where x_t represents information available at time t or before, and

$$\rho(z_t, \theta) = \theta_1 (c_{t+1}/c_t)^{\theta_2} R_{t+1} - 1.$$

where c_t is consumption in t and R_{t+1} is the return on an asset between t and $t+1$. Here θ_1 is time discount factor and per-period utility function is $(c_t)^{\theta_2+1}$.

Generalized Method of Moments (GMM) is the main method for dealing with the moment restriction.

In many cases GMM is known as **Nonlinear Instrumental Variable estimator**. In statistics, the analogs of GMM are known as **Z-estimators** or **Estimating Function estimators**.

GMM Notation:

$$\begin{aligned} \theta &: p \times 1 \text{ parameter vector.} \\ z_i &: \text{data observation.} \\ g(z,\theta) &: \text{and } r \times 1 \text{ vector-function, } r \geq p. \end{aligned}$$

Main Assumption: $z_1, ..., z_n$ are independently and identically distributed (i.i.d.) with

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

The problem is to estimate θ_0 from the available data.

Generalized Method of Moments (GMM):

For

$$\widehat{g}(\theta) := E_n[g(z_i, \theta)].$$

$$\widehat{\theta} \quad \text{minimizes} \quad \widehat{g}(\theta)' \widehat{A} \widehat{g}(\theta).$$

where $\widehat{A}$ is a positive semi-definite (p.s.d.) matrix. E_n is the empirical expectation operator:

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

The estimator sets the empirical moments as close as possible to their population counterparts, which are known to be 0.

Properties: Consistent Asymptotically Normal (CAN).

Choice of $\widehat{A}$ that minimizes asymptotic variance is a consistent estimator of the asymptotic variance of $\sqrt{n}\widehat{g}(\theta_0)$.

$$A = \left(\lim_n Var[\sqrt{n}\widehat{g}(\theta_0)] \right)^{-1}.$$

For $E[g(z_i, \theta_0)g(z_j, \theta_0)'] = 0$ for $i \neq j$ (no autocorrelation in the moment function) an optimal choice of weighting matrix is

$$A = \left(E[g(z_i, \tilde{\theta})g(z_i, \tilde{\theta})']\right)^{-1},$$

This can be estimated by

$$\hat{A} = \left(E_n[g(z_i, \tilde{\theta})g(z_i, \tilde{\theta})']\right)^{-1},$$

where $\tilde{\theta}$ is a preliminary (GMM) estimator. One can iterate on the estimate.

Conditional Moment Restrictions: Often have residual vector $\rho(z, \theta)$ satisfying a conditional moment restriction

$$E[\rho(z, \theta_0)|x] = 0.$$

Then by iterated expectations, for any conformable matrix $B(x)$ of functions of x, we have

$$g(z, \theta) = B(x)\rho(z, \theta)$$

satisfy unconditional moment restrictions. Can think of $B(x)$ as instrumental variables, in which case the estimator is known as the **Nonlinear Instrumental Variable Estimator**.

Maximum Likelihood Estimation

Notation:

θ : $p \times 1$ parameter vector.

z_i : data observation; e.g. $z_i = (y_i, x_i)$.

$f(z|\theta)$: family of density or probability functions.

Data Assumption: $z_1, ..., z_n$ are i.i.d. with

$$z_i \text{ having pdf } f(z|\theta_0).$$

The key fact is that

$$\theta_0 \text{ maximizes } E[\ln f(z_i|\theta)],$$

which is a consequence of **information inequality**.

Maximum Likelihood Estimator (MLE):

$$\widehat{\theta} \text{ maximizes } E_n[\ln f(z_i|\theta)].$$

Properties: Consistent and asymptotically normal (CAN) and asymptotically efficient.

Thus, MLE is asymptotically the most efficient estimator if $f(z|\theta)$ is correct. But, estimator may not be consistent if $f(z|\theta)$ is misspecified.

Conditional and Marginal MLE: Often we do not want to specify the entire distribution of the data. For example, we generally do not want to specify distribution of regressors. Recall the classical normal regression model as the primary example.

Suppose that the data take the form $z = (y, x)$ and the density can be factored as

$$f(z|\theta) = f(y|x, \beta)h(x|\gamma),$$

where β and γ are subvectors of θ, $f(y|x, \beta)$ is a conditional density of y given x and $h(x|\gamma)$ is a marginal density for x.

Then we can consider the estimator that maximizes the log-likelihood from either the conditional or the marginal part.

Conditional MLE (CMLE):

$$\widehat{\beta} \quad \text{maximizes} \quad E_n[\ln f(y_i|x_i,\beta)]$$

This estimator is CAN and asymptotically efficient when the information matrix for (β,γ) is block-diagonal, and when $h(x|\gamma)$ has unknown functional form. Here x can often be thought of as regressors, where we leave the marginal distribution unspecified, so that the CMLE is efficient.

Example (Censored Regression Model of Type I): The following example often arises when wages are top-coded or censored in other ways. The equation of interest is

$$Y_i^* = x_i'\beta_0 + u_i,$$

but we only observe

$$Y_i = \min\{Y_i^*, L\}.$$

Simple least squares where we regress Y_i on $x_i'\beta$ on the resulting data will give an inconsistent estimate, because the error does not satisfy the usual orthogonality condition $E[Y_i - x_i'\beta_0|x_i] \neq 0$. A way to deal with the problem is to use likelihood methods that explicitly account for how the data is generated.

The log-likelihood for a single observation (y, x) conditional on x is

$$\begin{aligned} \ln f(y|x,\theta) &= 1(y < L) \ln[f(y|x,\theta)] \\ &+ 1(y = L) \ln Pr[y = L|x,\theta]. \end{aligned}$$

Note that likelihood is a mixture of discrete and continuous distribution.

Let's specify each of the pieces. Suppose u_i is independent of x_i and is distributed $N(0, \sigma^2)$. Then for the standard normal pdf $\phi(u)$ with a cdf $\Phi(u)$ we have

$$\begin{aligned} \Pr(y = L|x, \beta, \sigma^2) &= \Pr(u_i \geq L - x'\beta|x, \beta, \sigma^2) \\ &= 1 - \Phi((L - x'\beta)/\sigma) \\ &= \Phi((x'\beta - L)/\sigma). \end{aligned}$$

Also, for $y < L$ the conditional pdf of y given x is

$$f(y|x, \beta, \sigma) = \sigma^{-1}\phi((y - x'\beta)/\sigma), y < L.$$

Then the log-likelihood for a single observation (y, x) conditional on x is

$$\begin{aligned} \ln f(y|x, \beta, \sigma^2) &= 1(y < L) \ln[\sigma^{-1}\phi((y - x'\beta)/\sigma)] \\ &+ 1(y = L) \ln \Phi((x'\beta - L)/\sigma). \end{aligned}$$

The consistency of the MLE for estimating θ_0 will **depend** crucially on the correctness of the normality assumption, in particular upon the distributional shape of the upper tail. Censored quantile regression methods introduced later do not require strong distributional assumptions.

Another way to relax the distributional assumptions is to make use of more flexible models of distributions. For example, take a t-family for the error term instead of the normal family.

Example 2 Contd. (Type-II Censored Regression Model or Selection Model): The equation of interest is

$$Y_{i2}^* = x_i'\beta_{02} + u_{i2},$$

and we observe $Y_{i2} = Y_{i2}^*$ if

$$Y_{i1}^* = x_i'\beta_{01} + u_{i1} \geq 0.$$

For unobserved values of Y_{i2}^*, we set $Y_{i2} = 0$. We also observe $Y_{i1} = 1(Y_{i1}^* > 0)$.

In the context of labor force participation, this models a two-part decision process: first a person decides whether to work or not, on the basis of some index Y_{i1}, and then decides how many hours Y_{i2} to work. This is all given the circumstances modeled by x_i and disturbances u_{i1} and u_{i2}.

The conditional log-likelihood function for a single observation (y_2, y_1) takes the form

$$\begin{aligned} \ln f(y_2, y_1|x, \theta) &= 1(y_1 = 0) \ln Pr[y_1 = 0|x] \\ &+ 1(y_1 = 1)\Big[\ln\Big(f[y_2|x, y_1 = 1] \\ &\cdot \quad Pr[y_1 = 1|x]\Big)\Big]. \end{aligned}$$

We can parametrically specify each piece, using the joint normality of disturbances and their independence from regressors. Then write the conditional likelihood function and proceed with

maximum likelihood estimation. You will work through many details of this in HW.

Marginal MLE:

$$\widehat{\gamma} \quad \text{maximizes} \quad E_n[\ln f(x_i|\gamma)].$$

This estimator is CAN and asymptotically efficient when $f(y|x,\beta)$ has unknown functional form or when β and when information matrix for β and γ is block-diagonal. This can be thought of as throwing away some data.

Need an example here.

M-estimators are a generalization of MLE, and the name stands for "maximum likelihood like estimators." Another name, more often used in econometrics, is **Quasi Maximum Likelihood Estimator (QMLE)**.
The parameter θ_0 is known to solve

$$\theta_0 \text{ minimizes } Em(z, \theta)$$

and the estimator takes the form

$$\widehat{\theta} \text{ minimizes } E_n m(z, \theta).$$

Example (Quantile Regression). Other than linear and nonlinear least squares, an important example is the median regression, or least absolute deviation estimator, where m-function takes the form $m(z, \theta) = |Y - X'\theta|$, so that the estimator minimizes

$$E_n[|Y - X'\theta|].$$

Median regression aims to estimate the **conditional median** of Y given X, modeled by $X'\theta_0$.

Quantile regression is a generalization of the median regression that aims to estimate the $\tau-$**conditional quantile** of Y given X. Quantile regression minimizes the asymmetric absolute deviation with m-function $m(z,\theta)$ equal to

$$\rho_\tau(Y - X'\theta) := \tau|Y - X'\theta|_+ + (1-\tau)|Y - X'\theta|_-,$$

where $\tau \in (0,1)$.

Minimum Distance Estimation

Notation:

θ : $p \times 1$ parameter vector of interest

$\widehat{\pi}$: preliminary "reduced form estimator"

$h(\pi, \theta)$: $r \times 1$ vector function of the parameters,

where $r \geq p$.

Data Assumption: $\widehat{\pi}$ is a consistent estimator of a parameter vector π_0. The parameter of interest θ_0 is an (implicit) function of θ_0 defined by:

$$h(\pi_0, \theta_0) = 0.$$

Thus, here the data enters only through the parameters estimator $\widehat{\pi}$.

Minimum Distance Estimator (MDE): The estimator is obtained from

$$\widehat{\theta} \text{ minimizes } h(\widehat{\pi}, \theta)' \widehat{A} h(\widehat{\pi}, \theta).$$

where $\hat{A}$ is a positive semi-definite (p.s.d.) matrix. Idea is that estimator sets $h(\hat{\pi}, \theta)$ close to population counterpart of 0.

Properties: CAN. Choice of $\hat{A}$ that minimizes asymptotic variance is

$$\hat{A} = \hat{\Omega}^{-1}, \text{ for } \sqrt{n}h(\hat{\pi}, \theta_0) \xrightarrow{d} N(0, \Omega).$$

Form of $\hat{\Omega}$ depends on $\hat{\pi}$ and h.

Example. Combining data from different sources for instrumental variables estimation. Suppose that we are interested in estimating the slope δ_0 of an equation

$$Y_i = \alpha_0 + \delta_0 D_i + u_i = X_i'\beta_0 + u_i, \; X_i' = (1, D_i), \beta' = (\alpha, \delta$$

where D_i is a dummy variable taking on 0 or 1. Suppose also that there is an instrument z_i available, with $Cov(z_i, u_i) = 0 \;\; Cov(z_i, D_i) \neq 0$. The problem is that individual data is not available. Instead, for $Z_i' = (1, z_i)$ one has

an estimator $\widehat{\pi}_{ZX}$ of $\pi_{ZX} = E[Z_i X_i']$ from one data set and another estimator $\widehat{\pi}_{ZY}$ of $\pi_{ZY} = E[Z_i Y_i]$ from another data set. The instrument orthogonality condition gives $0 = E[Z_i u_i] = \pi_{ZY} - \pi_{ZX}\beta_0$. This condition can be exploited to form a MDE as $\widehat{\beta}$ minimizes

$$(\widehat{\pi}_{ZY} - \widehat{\pi}_{ZX}\beta)' \widehat{A} (\widehat{\pi}_{ZY} - \widehat{\pi}_{ZX}\beta).$$

An empirical example is given in Angrist (1990, "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records"). There Y_i is earnings, D_i indicates whether they served in the military, and z_i indicates whether their birthday was after a certain date. In that application, $\widehat{\pi}_{ZX}$ comes from military records and $\widehat{\pi}_{ZY}$ from social security data.

Extremum Estimator

Notation:

θ : $p \times 1$ parameter vector of interest

$\widehat{Q}(\theta)$: $r \times 1$ function of data and parameters.

Data Assumption: Here we assume that

$$\sup_{\theta \in \Theta} |\widehat{Q}(\theta) - Q(\theta)| \to_p 0,$$

and

$$\theta_0 \text{ minimizes } Q(\theta).$$

Here the data enters only through the function $\widehat{Q}(\theta)$.

Extremum Estimator: The estimator is obtained as follows:

$$\widehat{\theta} \text{ minimizes } \widehat{Q}(\theta).$$

The idea is that $\hat{Q}(\theta)$ is close to a function that is minimized at θ_0, so minimizer of $\hat{Q}(\theta)$ should be close to θ_0.

Special Cases:

$$\begin{aligned} MLE &: \hat{Q}(\theta) = -E_n \ln f(z_i, \theta), \\ M &: \hat{Q}(\theta) = E_n m(z_i, \theta), \\ GMM &: \hat{Q}(\theta) = \hat{g}(\theta)' \hat{A} \hat{g}(\theta), \\ MD &: \hat{Q}(\theta) = h(\hat{\pi}, \theta)' \hat{A} h(\hat{\pi}, \theta). \end{aligned}$$

Discussion: The extremum approach provides one unified approach to the asymptotic properties of estimators. The GMM approach can also be viewed as a fairly general approach. Recent literature on **empirical likelihood** can be viewed as an important refinement of the GMM approach.

Computational considerations often also lean toward extremum estimators. For good computing, we would like $\hat{Q}(\theta)$ to be convex in

the parameters. An example is quantile regression that constructs a convex M-function – as asymmetric least absolute deviation – that can be easily minimized. In contrast, the GMM approach to the quantile problem is basically intractable. **Computability** is **extremely important** for applicability.

Some remarks on the history of thought:

1. MLE was pioneered by Laplace and Gauss, and Fisher. Cramer gave the first rigorous treatment of asymptotic normality. Ibragimov and Hasminskii and LeCam's treatment were other milestones in asymptotics for MLE, including both regular and non-regular cases. Huber gave the first rigorous study of M-estimators.

2. Method of moments was introduced by Karl Pearson. Nonlinear IV was introduced by Takeshi Amemiya who also derived efficient instruments. In statistics, Godambe pioneered estimating equations and studied their efficient form. GMM, especially with a view towards economic time series, was introduced by Lars Hansen. Hansen used ergodicity and structure of the economic optimizing model to justify the weighting matrix and CAN properties. Extremum approach was developed in depth in

Amemiya's Advanced Econometrics text. Keith Knight's and Charles Geyer's treatment of extremum estimators and M-estimators using epi-convergence appears to be the best current treatments of asymptotics.

3. Hansen and Singleton's Example 1 is one of the paradigms that define econometrics as a field. Tobin introduced Type I censored regression models, another paradigm of econometrics, and Amemiya gave the first rigorous treatment. Selection models or Type II models were developed by Gronau and Heckman.

4. M-estimators were studied by Huber. Median regression goes back to Boskovich in 17th century and Laplace (1818). Quantile regression was introduced by Koenker and Bassett (1978) and developed fully in an impressive body of work led by Roger Koenker. The maximum likelihood, least squares, and quantile regression are probably the most prominent representatives of M-estimators.