14.385

Nonlinear Econometrics

Lecture 4.

Theory: Asymptotic Distribution of GMM/Nonlinear IV

Application: Revisit probits and logits. Multinomial choice.

Topics to be covered in TA Session:

Testing (Parallels OLS)

Variance Estimation (Parallels OLS)

Asymptotic Normality of GMM and Nonlinear IV.

Recall Idea: Estimate parameters by setting sample moments to be close to population counterpart.

Definitions:

β : $p \times 1$ parameter vector, with true value β_0.
$g_i(\beta)$ = $g(z_i, \beta) : m \times 1$ vector of functions
of i^{th} data observation z_i and parameter.

Model (or moment restriction):

$$E[g_i(\beta_0)] = 0.$$

Definitions:

$\widehat{g}(\beta) := E_n[g_i(\beta)]$: Sample averages.
$\widehat{A}$: $m \times m$ positive definite matrix.

Lead Examples:
IV: $g_i(\beta) = (Y_i - X_i\beta)Z_i,\ \widehat{A} = \widehat{V}ar[g_i(\beta_0)]^{-1}$
NIV: $g_i(\beta) = f(Y_i, X_i, \beta)Z_i,\ \widehat{A} = \widehat{V}ar[g_i(\beta_0)]^{-1}$
MLE: $g_i(\beta) = \nabla \ln f(Z_i, \beta),\ \widehat{A} = I$
M: $g_i(\beta) = \nabla m(Z_i, \beta),\ \widehat{A} = I.$

GMM ESTIMATOR:

$$\widehat{\beta} = \arg\min_{\beta} \widehat{g}(\beta)'\widehat{A}\widehat{g}(\beta).$$

This is a special case of extremum estimator, so the arguments of the previous type can be applied to get the following result.

ASYMPTOTIC NORMALITY OF GMM: *If the data are i.i.d. or stationary strongly mixing with rate greater than two,* $\widehat{\beta} \xrightarrow{p} \beta_0$ *and i)* β_0 *is in the interior of the parameter set over which minimization occurs; ii)* $g_i(\beta)$ *is continuously differentiable on a neighborhood* $\mathcal{N}$ *of* β_0*; iii)* $E[\sup_{\beta \in \mathcal{N}} \|\nabla g_i(\beta)\|]$ *is finite; iv)* $\widehat{A} \xrightarrow{p} A$ positive definite *and* $G'AG$ *is nonsingular, for* $G = E[\nabla g_i(\beta_0)]$*; v)* for i.i.d. data, $\Omega = E[g_i(\beta_0) g_i(\beta_0)']$ is finite and for mixing data $\Omega = \lim_n Var[\widehat{g}(\theta_0)]$ exists and is finite, then

$$\begin{aligned} \sqrt{n} \quad & (\widehat{\beta} - \beta_0) \xrightarrow{d} N(0, V), \\ V \quad = \quad & (G'AG)^{-1} G'A\Omega AG (G'AG)^{-1}. \end{aligned}$$

Proof: For $\widehat{G} = \nabla \widehat{g}(\widehat{\beta})$, we have the FOC,

$$0 = \widehat{G}' \widehat{A} \widehat{g}(\widehat{\beta}).$$

We can expand them as

$$0 = \widehat{G}' \widehat{A} \{\widehat{g}(\beta_0) + \nabla \widehat{g}(\bar{\beta})[\widehat{\beta} - \beta_0]\},$$

where $\nabla \widehat{g}(\bar{\beta})$ stands for the matrix whose each row evaluated at (a row-dependent) $\bar{\beta}$ located on the line joining θ_0 and $\widehat{\theta}$, and solve for

$$\sqrt{n}(\widehat{\beta} - \beta_0) = -[\widehat{G}' \widehat{A} \nabla \widehat{g}(\bar{\beta})]^{-1} \widehat{G}' \widehat{A} \sqrt{n} \widehat{g}(\beta_0)$$

By the ULLN and the CMT, we have that

$$-[\widehat{G}' \widehat{A} \nabla \widehat{g}(\bar{\beta})]^{-1} \widehat{G}' \widehat{A} \xrightarrow{p} -(G'AG)^{-1} G'A.$$

Application of the CLT gives us

$$\sqrt{n}\widehat{g}(\beta_0) \xrightarrow{d} N(0, \Omega).$$

Applying Slutsky, we get

$$\sqrt{n}(\widehat{\beta} - \beta_0) \xrightarrow{d} -(G'AG)^{-1}G'A \cdot N(0, \Omega).$$

Notes: (1) $\widehat{A}$ affects V only through $plim(\widehat{A})$.

(2) If $m = p$, then

$$V = G^{-1}\Omega G^{-1\prime},$$

and A drops out. Thus, the choice of the matrix A has no effect on asymptotic variance in this case. We have $m = p$ for MLE, M-estimators, and "exactly identified" GMM. We have that $m > p$ for "overidentified" GMM.

(3) The optimal choice of A is given by $A \propto \Omega^{-1}$, in which case

$$V = (G'\Omega^{-1}G)^{-1}.$$

(4) **MLE.** Assume i.i.d. data. For MLE, under correct specification:

$$g(z, \theta) = \nabla \ln f(z, \theta), \; G = E\nabla^2 \ln f(z, \theta),$$

$$\Omega = var[\nabla \ln f(z, \theta_0)], \text{ and } -G = \Omega$$

so that

$$V = -G^{-1} = \Omega^{-1}.$$

The fact that $G = \Omega$ is known as information matrix equality, which holds under some regularity conditions.

(5) **M-estimators, including MLE under incorrect specification.** We have

$$g(z,\theta) = \nabla m(z,\theta),\ G = E\nabla^2 m(z,\theta_0),$$

$$\Omega = \mathsf{Var}[\nabla m(z,\theta_0)],$$

so that

$$V = G^{-1}\Omega G^{-1}.$$

This is known as Huber's sandwich formula or, more simply, robust variance-covariance matrix.

(6) **Linear IV.** Illustrate plausibility of conditions with Linear IV:

interior parameter condition (i) holds by assumption; continuous differentiability (ii) holds by linearity of

$$g_i(\beta) = Z_i(y_i - X_i'\beta)$$

in β; dominance condition (iii) holds as long as second moments exist, by

$$\nabla g_i(\beta) = -Z_i X_i' \quad G = -EZ_i X_i'$$

; (iv) holds as long as A is nonsingular and $G = -E[Z_i X_i']$ has full column rank; and (v) holds as long as

$$\Omega = E(Y_i - X_i'\beta_0)^2 Z_i Z_i'$$

is finite. Then, with $A \propto \Omega^{-1}$,

$$V = (G'\Omega^{-1}G)^{-1}.$$

(7) As stated in the theorem, i.i.d. sampling can be replaced by strict stationarity and strong mixing with rate larger than 2, that is provided mixing coefficients $\alpha(j)$ go to zero at rate $j^{-\alpha}$ as $j \to \infty$, for $\alpha > 2$, in which case the limit variance takes the form

$$\Omega = \lim_{n\to\infty} Var[\sqrt{n}\widehat{g}(\theta_0)].$$

(8) However, in many cases we can use the CLT for martingale difference sequences, which is much more relevant in dynamic economic applications. For instance, in Hansen and Singleton $g_i(\beta_0)$ being a martingale difference sequence is implied by economic assumptions. In this case, the limit variance is the same as in the i.i.d. case:

$$\Omega = Var[g_i(\theta_0)] = E[g_i(\beta_0)g_i(\beta_0)'].$$

Revisit Maximum Likelihood Estimation for Binary Choice:

$$\ln f_i(\beta) = y_i \log F(x_i'\beta) + (1 - y_i) \log(1 - F(x_i'\beta))$$

$$\nabla \ln f_i(\beta) = \Big(\frac{y_i - F(x_i'\beta)}{F(x_i'\beta) \cdot (1 - F(x_i'\beta))}\Big) f(x_i'\beta) x_i$$

CLT for the score

$$\sqrt{n} E_n \nabla \ln f_i(\beta_0) \to_d N(0, E \frac{f_i^2}{F_i(1 - F_i)} x_i x_i')$$

since $Var(y_i - F_i | x_i) = F_i(1 - F_i)$. Therefore:

$$\sqrt{n}(\widehat{\beta} - \beta) \to_d N(0, [E \frac{f_i^2}{F_i(1 - F_i)} x_i x_i']^{-1})$$

These variance formula is valid only under correct specification. Under incorrect specification, treat as an M-estimator, and use the Huber's robust sandwich formula.

Probit and Logit Examples:

For probit: $\sqrt{n}(\widehat{\beta} - \beta) \xrightarrow{d} N(0, E \frac{\phi_i^2}{\Phi_i(1-\Phi_i)} x_i x_i']^{-1})$.

For logit:

$$\sqrt{n}(\widehat{\beta} - \beta) \to_d N(0, [E \Lambda(x_i'\beta)(1 - \Lambda(x_i'\beta)) x_i x_i']^{-1}).$$

You can also do NLS by minimizing

$$\tilde{\beta} = \arg\min E_n (y_i - F(x_i'\beta))^2$$

by using the relations:

$$y_i = F(x_i'\beta) + \epsilon_i, E(\epsilon_i|x_i) = 0, V(\epsilon_i|x_i) = F(x_i'\beta)(1 - F(x_i'\beta)).$$

The NLS $\tilde{\beta}$ is not efficient because of heteroscedasticity, so you can do GLS:

$$\widehat{\beta} = \arg\min E_n\Big[\frac{1}{F(x_i'\tilde{\beta})(1 - F(x_i'\tilde{\beta}))}(y_i - F(x_i'\beta))^2\Big],$$

which attains the same efficiency as MLE.

The usual choice for probit and logit is MLE due to computational reasons: the least squares problem is non-convex, whereas log-likelihood for probits and logits is concave.

Multinomial Choice Models

The multinomial choice or qualitative response models provide a framework for analysis of choice among finite sets of alternatives, with each alternative characterized as a bundle of attributes.

We consider the following example that captures the basics of these models. Consider predicting urban demand for various modes of transportation. The utility of a commuter from taking choices car, train, and bus are given by:

1. Bus($y = 1$): $U_{1i} = \mu_{1i} + \epsilon_{1i}$

2. Train($y = 2$): $U_{2i} = \mu_{2i} + \epsilon_{2i}$

3. Car($y = 3$): $U_{3i} = \mu_{3i} + \epsilon_{3i}$

Here μ_{ki} characterizes the systematic part of utility, called the "mean utility," attributable to observed characteristics x_i of modes of transportation and of the commuter, and ϵ_{ki} is an unobserved disturbance. Often we use

$$\mu_{ki} = x_i'\beta_k$$

or more general functional forms. In what follows, we use $P_i(y_i = 1)$ to denote $P(y_i = 1|x_i)$.

Then

$$P_i(y_i = 1) = P_i(U_{1i} \geq U_{2i}, U_{1i} \geq U_{3i}),$$

$P_i(y_i = 2) = P_i(U_{2i} \geq U_{1i}, U_{3i} \geq U_{1i})$, and $P_i(y_i = 3) = P_i(U_{3i} \geq U_{1i}, U_{3i} \geq U_{2i})$. This allows use to write down the likelihood function, using the fact that the conditional log-likelihood function of a single observation (y_i, x_i) is

$$\sum_k 1\{y_i = k\} \ln P_i(y_i = k)$$

Theory: The usual MLE analysis applies to get CAN and efficient estimators, subject to regularity conditions. We should use M-estimation approach to account for possible misspecification.

Multinomial Logit Model: This model postulates that

$$P_i(y_i = k) = \frac{e^{\mu_{ik}}}{\sum_k e^{\mu_{ik}}}.$$

McFadden has shown that this model can be derived from optimizing behavior: Suppose disturbances ϵ_{1i}, ϵ_{2i}, ϵ_{3i} are i.i.d. and each one distributed as Type-I extreme Value, with Gumbel distribution function,

$$F(\epsilon) = e^{-e^{-\epsilon}}$$

that has density $f(\epsilon) = e^{-\epsilon}e^{-e^{-\epsilon}}$. Then the choice probabilities are described by the formula above. The assumption on the errors is also a necessary condition for the formula above to hold.

We illustrate the calculation using the travel example:

$$P_i(y_i = 2) = P_i(\epsilon_1 + \mu_1 \leq \epsilon_2 + \mu_2, \epsilon_3 + \mu_3 \leq \epsilon_2 + \mu_2)$$
$$= \int_{-\infty}^{\infty} f(\epsilon_2) P_i(\epsilon_1 \leq \epsilon_2 + \mu_2 - \mu_1 | \epsilon_2) P_i(\epsilon_3 \leq \epsilon_2 + \mu_2 - \mu_3 | \epsilon_2) d\epsilon_2$$
$$= \int_{-\infty}^{\infty} e^{-\epsilon_2} [e^{-e^{-\epsilon_2}[1 + e^{\mu_1 - \mu_2} + e^{\mu_3 - \mu_2}]}] d\epsilon_2$$
$$= \int_{0}^{\infty} e^{-[1 + e^{\mu_1 - \mu_2} + e^{\mu_3 - \mu_2}]x} dx \quad [\text{ using } x = e^{-\epsilon_2}]$$
$$= \frac{1}{1 + e^{\mu_1 - \mu_2} + e^{\mu_3 - \mu_2}}$$
$$= \frac{e^{\mu_2}}{e^{\mu_1} + e^{\mu_2} + e^{\mu_3}}.$$

Similarly, $P_i(y_i = 1) = \frac{e^{\mu_1}}{e^{\mu_1} + e^{\mu_2} + e^{\mu_3}}$, and $P_i(y_i = 3) = \frac{e^{\mu_1}}{e^{\mu_1} + e^{\mu_2} + e^{\mu_3}}$.

It is also easy to get that

$$P_i(y = 1 | y = 1 \ or \ y = 2) = \frac{e^{\mu_1}}{e^{\mu_1} + e^{\mu_2}}$$
$$P_i(y = 3 | y = 1 \ or \ y = 3) = \frac{e^{\mu_3}}{e^{\mu_1} + e^{\mu_3}}$$

and so on.

IIA property: Relative choice probabilities depend only on pairwise comparisons. This is the independence from the "irrelevant" alternatives. E.g., in the last equation, the relative frequency of choosing between 1(bus)

and 3(car) does not depend on characteristics of choice 2(train). This might be unreasonable since bus(1) and train(2) are both public transportation and the unobservable components of utility from them might be correlated. (A more extreme example would be to look at buses painted in different colors.) Another way to look at this is as follows: Suppose $y = 2$ is not available, then

$$\begin{aligned} P_i(y = 3 \text{ without choice 2}) &= P_i(U_3 \geq U_1) \\ &= \frac{e^{\mu_3}}{e^{\mu_1} + e^{\mu_3}} \end{aligned}$$

Compare this to the probability of choosing $y = 3$, having chosen either 1 or 3,

$$P_i(y = 3 | y = 1 \ or \ y = 3) = \frac{e^{\mu_3}}{e^{\mu_1} + e^{\mu_3}},$$

which is the same as above. Here, the relative frequency of choosing between 1(bus) and 3(car) does not depend whether the choice 2(train) is present. Again, this might be unreasonable since bus(1) and train(2) are both public transportation and utilities from them might be correlated.

Discussion: The logit model is highly tractable and computable due to the concave log-likelihood function. Moreover, it is derivable from the optimizing choice behavior. The drawback of this model is the IIA property.

Nested Logit: To allow for correlation between choice 1 and 2, introduce some dependence between ϵ_1 and ϵ_2 as follows:

$$F(\epsilon_1, \epsilon_2, \epsilon_3) = G(\epsilon_1, \epsilon_2)F(\epsilon_3) = G(\epsilon_1, \epsilon_2)e^{-e^{-\epsilon_3}}$$

and

$$G(\epsilon_1, \epsilon_2) = e^{-[e^{-\epsilon_1/\rho}+e^{-\epsilon_2/\rho}]^\rho},$$

for $0 \leq \rho \leq 1$. This is a bivariate type I extreme-value (Gumbel) distribution. When $\rho = 1$ we are back to the logit case, since then $G(\epsilon_1, \epsilon_2) = F(\epsilon_1) \cdot F(\epsilon_2)$. When $\rho = 0$, ϵ_1 and ϵ_2 become perfectly correlated.

Then

$$\begin{aligned} P_i(y_i = 1 | y_i \neq 3) &= P_i(U_1 \geq U_2 | U_1 \geq U_3 \text{ or } U_2 \geq U_3) \\ &= \frac{e^{\mu_1/\rho}}{e^{\mu_1/\rho} + e^{\mu_2/\rho}}, \end{aligned}$$

using calculations similar to the previous ones. Also,

$$P_i(y = 3) = \frac{e^{\mu_3}}{e^{\mu_3} + [e^{\mu_1/\rho} + e^{\mu_2/\rho}]^\rho}$$

These two quantities enable us to fully specify the likelihood. The IIA property no longer holds.

Discussion: Nested logit allows us to avoid some of the limitations of logit. It also preserves the logit's computational tractability. There are further generalizations, in particular, McFadden's generalized extreme value model, of which nested logit is a special case.

(Lecture 4 contd.)

Multinomial Probit: Here we specify,

$$\epsilon_i = (\epsilon_{ki}, k = 1, .., K) \sim N(0, \Sigma).$$

This allows for very flexible correlation structure, but the choice probabilities

$$P_i(y = j|x_i) = P_i[U_j \geq U_k, \text{ for all } k]$$

are usually not available in closed form, so one has to evaluate choice probabilities numerically, using Monte-Carlo approach. In the MC approach, we simulate many draws of

$$\epsilon_s = \Sigma^{1/2} Z_s, \ Z_s \sim_{\text{iid}} N(0, I), \ s = 1, ..., S,$$

then evaluate

$$U_{ks} = x_i' \beta_k + \epsilon_s, s = 1, ..., S,$$

and approximate

$$P_i(y = j|x_i) \approx \frac{1}{S} \sum_{s=1}^{S} 1\{U_{js} \geq U_{ks}, \text{ for all } k\}.$$

The draws of $\{Z_s, s = 1, ..., S\}$ are generated only once and reused in the evaluation of $P_i(y = j|x_i)$ at different parameter values. Note that in the formula above, the utility terms U_k depend on parameters β and Σ. The approach was pioneered by McFadden.

Notes: There are many useful extensions, including **generalized extreme value models** and **sequential**

choice models (that exploit the sequentiality of decisions). Both of these approaches help facilitate the implementation.

The qualitative response models and their derivation from the optimizing behavior were largely pioneered by McFadden. McFadden also introduced the simulated likelihood methods to econometrics, which was a major innovation. Others who introduced simulated likelihood in statistics and econometrics include Čencov (60s), Geyer, and Polard and Pakes.

References:

Amemiya, Advanced Econometrics, Chapter 9, and others on the reading list.