Topics in Econometrics

Generalized Method of Moments: Extensions

Instructor: Ma, Jun

Renmin University of China

May 12, 2022

Moment equation models

- Let $g_i(\beta)$ be a known $l \times 1$ function of the i-th observation W_i ($g_i(\beta) = g(W_i, \beta)$) and the parameter $\beta \in \mathbb{R}^k$. A moment equation model is

$$\mathrm{E}[g_i(\beta)] = 0.$$

  We know that the true parameter β satisfies the system of equations.
- For example, in the instrumental variables model $g_i(\beta) = Z_i \left(Y_i - X_i^\top \beta\right)$ ($W_i = (Y_i, X_i, Z_i)$).
- We say the parameter is identified if there is unique β solves the equations. A necessary condition for identification is $l \geq k$.
- $l = k$: just identified;
- $l > k$: over-identified.

Method of moments

- We consider the just identified case: $l = k$
- The sample analogue of $\mathrm{E}\left[g_i\left(\beta\right)\right]$:

$$\overline{g}_n\left(\beta\right) = \frac{1}{n}\sum_{i=1}^{n} g_i\left(\beta\right).$$

- The method of moments estimator (MME) $\widehat{\beta}_{\mathrm{mm}}$ for β is the solution to

$$\frac{1}{n}\sum_{i=1}^{n} g_i\left(\widehat{\beta}_{\mathrm{mm}}\right) = 0.$$

Overidentified moment equations

- Define
$$\overline{g}_n\left(b\right) = \frac{1}{n}\sum_{i=1}^{n} g_i\left(b\right).$$
- We defined the MME $\widehat{\beta}$ for β to be the solution to $\overline{g}_n\left(\widehat{\beta}\right) = 0$. However, if the model is over-identified, there are more equations than parameters. The MME is not defined.
- We cannot find an estimator $\widehat{\beta}$ which sets $\overline{g}_n\left(\widehat{\beta}\right) = 0$ but we can try to find an estimator $\widehat{\beta}$ which makes $\overline{g}_n\left(\widehat{\beta}\right)$ as close to zero as possible.

- Let W be an $l \times l$ positive definite weight matrix. The GMM criterion function is

$$J\left(b\right) = n \cdot \overline{g}_n\left(b\right)^\top W \overline{g}_n\left(b\right).$$

- When $W = \mathrm{I}_l$ (l-dimensional identity matrix), $J\left(b\right) = n \cdot \overline{g}_n\left(b\right)^\top \overline{g}_n\left(b\right) = n \cdot \left\|\overline{g}_n\left(b\right)\right\|^2$.
- The Generalized method of moments (GMM) estimator is $\widehat{\beta}_{\mathrm{gmm}} = \mathrm{argmin}_b J_n\left(b\right)$.

Asymptotic distribution

- Asymptotic distribution of the GMM estimator

$$\sqrt{n}\left(\widehat{\beta}_{\mathrm{gmm}} - \beta\right) \to_d \mathrm{N}\left(0, V_W\right).$$

where

$$V_W = \left(\mathrm{Q}^\top W \mathrm{Q}\right)^{-1}\left(\mathrm{Q}^\top W \Omega W \mathrm{Q}\right)\left(\mathrm{Q}^\top W \mathrm{Q}\right)^{-1}$$

with

$$\Omega = \mathrm{E}\left[g_i\left(\beta\right) g_i\left(\beta\right)^\top\right] \text{ and } \mathrm{Q} = \mathrm{E}\left[\left.\frac{\partial}{\partial b^\top} g_i\left(b\right)\right|_{b=\beta}\right]$$

- If the efficient weight matrix $W = \Omega^{-1}$ is used then

$$V_\beta = \left(\mathrm{Q}^\top \Omega^{-1} \mathrm{Q}\right)^{-1}.$$

Efficient GMM

- The efficient GMM estimator can be constructed by using

$$\widehat{\Omega} = \frac{1}{n}\sum_{i=1}^{n} g_i\left(\widetilde{\beta}\right) g_i\left(\widetilde{\beta}\right)^\top - \overline{g}_n\left(\widetilde{\beta}\right) \overline{g}_n\left(\widetilde{\beta}\right)^\top,$$

  with a preliminary consistent estimator $\widetilde{\beta}$.
- The asymptotic covariance matrix can be estimated by sample counterparts of the population matrices.

Continuously-updated GMM

- An alternative to the two-step GMM estimator can be constructed by letting the weight matrix be an explicit function of b:

$$J(b) = n \cdot \overline{g}_n(b)^\top \left(\frac{1}{n} \sum_{i=1}^{n} g_i(b) g_i(b)^\top \right)^{-1} \overline{g}_n(b)$$

or

$$J(b) = \overline{g}_n(b)^\top \left(\frac{1}{n} \sum_{i=1}^{n} g_i(b) g_i(b)^\top - \overline{g}_n(b) \overline{g}_n(b)^\top \right)^{-1} \overline{g}_n(b).$$

- The $\widehat{\beta}$ which minimizes this function is the CU-GMM estimator. The minimization requires numerical methods.
- We have:

$$\sqrt{n} \left(\widehat{\beta}_{\mathrm{cu-gmm}} - \beta \right) \to_d \mathrm{N}(0, V_\beta).$$

Wald statistic

- The parameter of interest θ is a function of the coefficients, $\theta = r(\beta)$ for some function $r : \mathbb{R}^k \to \mathbb{R}^q$. The estimator of θ is given by $\widehat{\theta} = r\left(\widehat{\beta}\right)$.
- If $r(\cdot)$ is continuous at the true value of β, then $\widehat{\theta} \to_p \theta$. Suppose that $r : \mathbb{R}^k \to \mathbb{R}^q$ is continuously differentiable at the true value of β and $R = \partial r(b)^\top / \partial b\Big|_{b=\beta}$ has rank q. Then, $\sqrt{n}\left(\widehat{\theta} - \theta\right) \to_d \mathrm{N}(0, V_\theta)$ where $V_\theta = R^\top V_\beta R$.
- Consider the Wald statistic

$$W(\theta) = n\left(\widehat{\theta} - \theta\right)^\top \widehat{V}_\theta^{-1}\left(\widehat{\theta} - \theta\right),$$

where $\widehat{V}_\theta$ is a consistent estimator of V_θ. Then, $W(\theta) \to_d \chi^2_q$.

Confidence set

- A confidence region $\widehat{C}$ is a set estimator for $\theta \in \mathbb{R}^q$. A natural confidence region is

$$\widehat{C} = \left\{\theta \in \mathbb{R}^q : W\left(\theta\right) \leq c_{1-\alpha}\right\},$$

with $c_{1-\alpha}$ being the $1-\alpha$ quantile of the χ^2_q distribution: $F_{\chi^2_q}\left(c_{1-\alpha}\right) = 1-\alpha$.
- Then,

$$\Pr\left[\theta \in \widehat{C}\right] \to \Pr\left[\chi^2_q \leq c_{1-\alpha}\right] = 1-\alpha.$$

- Note that the shape of the confidence set $\widehat{C}$ is predetermined (i.e., ellipse).

OverIdentification test

- Consider the linear IV model:

$$\begin{aligned} Y_i &= X_i^\top \beta + e_i \\ \mathrm{E}\left[e_i Z_i\right] &= 0, \end{aligned}$$

  where $X_i \in \mathbb{R}^k$ and $Z_i \in \mathbb{R}^l$. The model is over-identified: $l > k$.
- The model specifies

$$\mathrm{E}\left[e_i Z_i\right] = 0 \Longleftrightarrow \mathrm{E}\left[Z_i Y_i\right] = \mathrm{E}\left[Z_i X_i^\top\right] \beta.$$

- This is equivalent to saying that $\mathrm{E}\left[Z_i Y_i\right]$ is in the column space of $\mathrm{E}\left[Z_i X_i^\top\right]$. The model imposes a restriction on the distribution of the oberved variables (Y_i, X_i, Z_i).
- Since β is of dimension $k < l$, it is not certain if such a vector exists. In such a case, we say that the model is misspecified.

- Suppose that $X_i \in \mathbb{R}^1$ and $Z_i = \left(Z_i^{(1)}, Z_i^{(2)}\right)^\top \in \mathbb{R}^2$. Then the model specifies

$$\begin{aligned} \mathrm{E}\left[Z_i^{(1)} Y_i\right] &= \mathrm{E}\left[Z_i^{(1)} X_i\right] \beta \\ \mathrm{E}\left[Z_i^{(2)} Y_i\right] &= \mathrm{E}\left[Z_i^{(2)} X_i\right] \beta, \end{aligned}$$

which requires

$$\frac{\mathrm{E}\left[Z_i^{(1)} Y_i\right]}{\mathrm{E}\left[Z_i^{(1)} X_i\right]} = \frac{\mathrm{E}\left[Z_i^{(2)} Y_i\right]}{\mathrm{E}\left[Z_i^{(2)} X_i\right]}.$$

- The true distribution of (Y_i, X_i, Z_i) may violate this condition.
- We can do a hypothesis test of the model specification. This is known as the overidentification test:

$$\mathrm{H}_0 : \text{There exists } \beta \in \mathbb{R}^k \text{ such that } \mathrm{E}\left[Z_i \left(Y_i - X_i^\top \beta\right)\right] = 0.$$

- For the more general model, the null hypothesis of correct model specification is

$$\mathrm{H}_0 : \text{There exists } \beta \in \mathbb{R}^k \text{ such that } \mathrm{E}\left[g_i\left(\beta\right)\right] = 0.$$

- H_0 is true if and only if

$$\min_b n \cdot \mathrm{E}\left[g_i\left(b\right)\right]^\top \Omega^{-1} \mathrm{E}\left[g_i\left(b\right)\right] = 0.$$

- We estimate $\min_b n \cdot \mathrm{E}\left[g_i\left(b\right)\right]^\top \Omega^{-1} \mathrm{E}\left[g_i\left(b\right)\right]$ by

$$\min_b n \cdot \overline{g}_n\left(b\right)^\top \widehat{\Omega}^{-1} \overline{g}_n\left(b\right).$$

  and if it is large, we reject H_0.
- The test statistic is just $J\left(\widehat{\beta}_{\mathrm{gmm}}\right)$. This is known as the J-statistic. The overidentification test is referred to as the Sargan test.
- Under H_0, $J\left(\widehat{\beta}_{\mathrm{gmm}}\right) \to_d \chi^2_{l-k}$. We reject H_0 if $J\left(\widehat{\beta}_{\mathrm{gmm}}\right) > c_{1-\alpha}$ with $c_{1-\alpha}$ being the $1-\alpha$ quantile of the χ^2_{l-k} distribution: $F_{\chi^2_{l-k}}\left(c_{1-\alpha}\right) = 1 - \alpha$.

Maximum likelihood

- Let $(X_1, ..., X_n)$ be a random (i.i.d.) sample on a continuous with a density function $f(\cdot;\theta)$, $\theta \in \Theta \subseteq \mathbb{R}^k$. Let x_i be the observed value of X_i. Then we call

$$L(\theta; x_1, ..., x_n) = \prod_{i=1}^{n} f(x_i; \theta)$$

the likelihood function of θ given $(x_1, x_2, ..., x_n)$, and we call the value of θ that maximizes $L(\theta; X_1, ..., X_n)$ the maximum likelihood (ML) estimator.

- The log-likelihood function:

$$\ell(\theta; x_1, ..., x_n) = \sum_{i=1}^{n} \log f(x_i; \theta).$$

- The ML estimator: $\widehat{\theta}_{\mathrm{ml}} = \mathrm{argmax}_{\theta \in \Theta} \ell(\theta; X_1, ..., X_n)$.
- The model $\{f(\cdot;\theta) : \theta \in \Theta\}$ is correctly specified if there exists $\theta_* \in \Theta$ so that $f(\cdot;\theta_*) = f_X$, where f_X denotes the true density of X_i.

Kullback–Leibler divergence

- The Kullback–Leibler (KL) divergence from a density f to another density g:

$$\mathbb{D}_{\mathrm{kl}}\left(f \mid g\right) = \int \log\left(\frac{f\left(x\right)}{g\left(x\right)}\right) f\left(x\right) \mathrm{d}x.$$

- $\mathbb{D}_{\mathrm{kl}}\left(f \mid g\right) \geq 0$ and $\mathbb{D}_{\mathrm{kl}}\left(f \mid g\right) = 0$ if and only if $f = g$.
- Jensen's inequality: Let X be a random variable and h be a strictly concave function. That is,

$$h\left(\lambda a + \left(1 - \lambda\right) b\right) > \lambda h\left(a\right) + \left(1 - \lambda\right) h\left(b\right)$$

  for any $a < b$ and $0 < \lambda < 1$. Then $\mathrm{E}\left[h\left(X\right)\right] < h\left(\mathrm{E}\left[X\right]\right)$.
- If $f \neq g$,

$$\begin{aligned}\int \log\left(\frac{f\left(x\right)}{g\left(x\right)}\right) f\left(x\right) \mathrm{d}x &= -\int \log\left(\frac{g\left(x\right)}{f\left(x\right)}\right) f\left(x\right) \mathrm{d}x \\ &> -\log\left(\int g\left(x\right) \mathrm{d}x\right) = 0.\end{aligned}$$

- $\mathbb{D}_{\mathrm{kl}}\left(f_X \mid f\left(\cdot;\theta\right)\right) \geq 0$ and $\mathbb{D}_{\mathrm{kl}}\left(f_X \mid f\left(\cdot;\theta_*\right)\right) = 0$. This is equivalent to

$$\begin{aligned}
\theta_* &= \underset{\theta\in\Theta}{\operatorname{argmin}} \, \mathbb{D}_{\mathrm{kl}}\left(f_X \mid f\left(\cdot;\theta\right)\right) \\
&= \underset{\theta\in\Theta}{\operatorname{argmax}} \, -\int \log\left(\frac{f_X\left(x\right)}{f\left(x;\theta\right)}\right) f_X\left(x\right) \mathrm{d}x \\
&= \underset{\theta\in\Theta}{\operatorname{argmax}} \, \mathrm{E}\left[\log f\left(X;\theta\right)\right].
\end{aligned}$$

- A natural estimator from the perspective of KL divergence is given by $\operatorname{argmax}_{\theta\in\Theta} Q_n\left(\theta\right)$ with $Q_n\left(\theta\right) = n^{-1}\sum_{i=1}^{n} \log f\left(X_i;\theta\right)$, which is just the ML estimator.
- By LLN, we know that for each θ, $Q_n\left(\theta\right) \to_p Q\left(\theta\right)$ where $Q\left(\theta\right) = \mathrm{E}\left[\log f\left(X;\theta\right)\right]$. The ML estimator is defined to be the maximizer of $Q_n\left(\theta\right)$. We expect the maximizer should converge to the maximizer of its limit $Q\left(\theta\right)$ in probability.

Nonparametric likelihood

- The moment equation model is nonparametric in the sense that we do not fully specify the distribution of the observed variables.
- Rather than specifying a parametric model for X_i, we assume the variables follow a discrete distribution supported on the observations $X_1, ..., X_n$.
- The parameters corresponding to this "model" is $p_1, ..., p_n$ with $(p_1, ..., p_n) \in \Delta$, where

$$\Delta = \left\{ (p_1, ..., p_n) : \sum_{i=1}^{n} p_i = 1,\, p_i \geq 0,\, i = 1, ..., n \right\}.$$

- The nonparametric log-likelihood is

$$\ell\left(p_1, ..., p_n; X_1, ..., X_n\right) = \sum_{i=1}^{n} \log\left(n \cdot p_i\right),\, (p_1, ..., p_n) \in \Delta.$$

- The maximum of the above log-likelihood function is attained at $p_i = 1/n$, $\forall i$, which is the empirical distribution.

- Maximizing the nonparametric log-likelihood is equivalent to minimizing the KL divergence from the empirical distribution $(1/n, ..., 1/n)$ to $(p_1, ..., p_n)$: $\sum_{i=1}^{n} n^{-1} \log\left(n^{-1}/p_i\right)$.
- Consider the moment equation model:

$$\mathrm{E}\left[g_i\left(\beta\right)\right] = \int g\left(w, \beta\right) f_W\left(w\right) \mathrm{d}w = 0,$$

  where f_W denotes the true density of W_i.
- The model imposes a restriction on f_W.
- The empirical likelihood method is a constrained nonparametric likelihood with the constraint $\sum_{i=1}^{n} p_i g_i\left(b\right) = 0$ imposed.

Empirical likelihood (EL)

- The EL criterion function:

$$\ell_{\text{el}}(b) := \max_{p_1,\dots,p_n} 2\sum_{i=1}^{n} \log(n \cdot p_i)$$
$$\text{subject to } \sum_{i=1}^{n} p_i g_i(b) = 0,\ (p_1, ..., p_n) \in \Delta.$$

- The EL estimator $\widehat{\beta}_{\text{el}}$ is the maximizer of $\ell_{\text{el}}(b)$.
- $\sqrt{n}\left(\widehat{\beta}_{\text{el}} - \beta\right) \to_d \mathrm{N}(0, V_\beta)$, where V_β is the asymptotic variance of the efficient GMM (Qin and Lawless, 1994).
- The EL estimator is efficient and avoids estimating the optimal weighting matrix in the first step.

EL ratio inference

- The EL ratio statistic:

$$LR(\theta) = \max_b \ell_{\mathrm{el}}(b) - \max_{r(b)=\theta} \ell_{\mathrm{el}}(b).$$

  Then, $LR(\theta) \to_d \chi^2_q$.
- Estimation of the asymptotic variance is not needed.
- The EL confidence set:

$$\widehat{C} = \{\theta \in \mathbb{R}^q : LR(\theta) \le c_{1-\alpha}\}.$$

- The shape of the EL confidence set is data-driven.
- The EL method has many other favorable properties relative to efficient GMM. See Kitamura (2006) for a review.

Duality

- It seems that the high dimensionality of the parameter space makes the maximization problem infeasible in practice.
- Instead of directly solving it, we fix b first and use the Lagrange multiplier method to solve

$$\ell_{\mathrm{el}}(b) := \max_{p_1,...,p_n} 2 \sum_{i=1}^{n} \log(n \cdot p_i)$$

$$\text{subject to } \sum_{i=1}^{n} p_i g_i(b) = 0,\ (p_1, ..., p_n) \in \Delta.$$

- The Lagrangian associated with the constrained optimization problem is

$$\mathcal{L}(p_1, ..., p_n, \lambda) = \sum_{i=1}^{n} \log(p_i) + \gamma \left(1 - \sum_{i=1}^{n} p_i\right) - n \cdot \lambda^\top \sum_{i=1}^{n} p_i g_i(b),$$

where $\gamma \in \mathbb{R}$ and $\lambda \in \mathbb{R}^l$ are Lagrange multipliers.

- The first-order conditions:

$$
\begin{aligned}
0 &= \frac{1}{p_i} - \gamma - n\left(\lambda^\top g_i(b)\right) \\
0 &= 1 - \sum_{i=1}^{n} p_i \\
0 &= n\sum_{i=1}^{n} p_i g_i(b).
\end{aligned}
$$

- The first-order conditions are solved by $\gamma = n$ and $(p_1, ..., p_n, \lambda)$ are given by the solution to

$$
\begin{aligned}
p_i &= \frac{1}{n\left(1+\lambda^\top g_i(b)\right)} \\
0 &= \sum_{i=1}^{n} \frac{g_i(b)}{1+\lambda^\top g_i(b)}.
\end{aligned}
$$

- The l equations $0 = \sum_{i=1}^{n} g_i(b) / \left(1+\lambda^\top g_i(b)\right)$ are the first-order conditions of the convex minimization problem $\min_\lambda -\sum_{i=1}^{n} \log\left(1+\lambda^\top g_i(b)\right)$.

- The EL estimator is therefore

$$\widehat{\beta}_{\mathrm{el}} = \underset{b}{\operatorname{argmax}} \min_{\lambda} - \sum_{i=1}^{n} \log\left(1 + \lambda^{\top} g_i\left(b\right)\right).$$

- For fixed b, $\min_{\lambda} - \sum_{i=1}^{n} \log\left(1 + \lambda^{\top} g_i\left(b\right)\right)$ is a convex minimization problem, for which a simple Newton algorithm works.
- The maximization of $\min_{\lambda} - \sum_{i=1}^{n} \log\left(1 + \lambda^{\top} g_i\left(b\right)\right)$ with respect to b is harder to solve. It is solved by a nonlinear optimization algorithm.

Implied probabilities

- Once $\widehat{\beta}_{\mathrm{el}}$ is calculated, we get the implied probabilities

$$\widehat{p}_i = \frac{1}{n\left(1 + \widehat{\lambda}^\top g_i\left(\widehat{\beta}_{\mathrm{el}}\right)\right)},$$

where $\widehat{\lambda}$ is the solution of the equations

$$0 = \sum_{i=1}^{n} \frac{g_i\left(\widehat{\beta}_{\mathrm{el}}\right)}{1 + \widehat{\lambda}^\top g_i\left(\widehat{\beta}_{\mathrm{el}}\right)}.$$

- Suppose that we are interested in estimating $\mathrm{E}\left[h\left(W_i\right)\right]$, where $h\left(\cdot\right)$ is a known function.
- $\sum_i \widehat{p}_i h\left(W_i\right)$ is an efficient estimator of $\mathrm{E}\left[h\left(W_i\right)\right]$ relative to the sample mean $n^{-1}\sum_i h\left(W_i\right)$ (Brown and Newey, 1998).
- $\left(\widehat{p}_1, ..., \widehat{p}_n\right)$ is also a more efficient estimator than the empirical distribution, from which we do bootstrap resampling.

- We have a small dataset on $W_i = (Y_i, X_i)$, $i = 1, ..., n$, but X_i includes a rich set of variables so that the regression model is not suffering from the omitted variable bias.
- Suppose that M_i is the vector collecting a small subset of variables in W_i. We have another auxiliary dataset on M_i. Such a dataset has a very large sample size N.
- We can calculate the implied probabilities

$$\widehat{p}_i = \frac{1}{n\left(1 + \widehat{\lambda}^\top \left(M_i - \overline{M}\right)\right)}$$

where $\widehat{\lambda}$ is the solution of the equations

$$0 = \sum_{i=1}^{n} \frac{M_i - \overline{M}}{1 + \widehat{\lambda}^\top \left(M_i - \overline{M}\right)},$$

where $\overline{M}$ is the sample mean of M_i computed by using the auxiliary dataset.
- The reweighted estimator $\left(\sum_{i=1}^n \widehat{p}_i X_i X_i^\top\right)^{-1} \left(\sum_{i=1}^n \widehat{p}_i X_i Y_i\right)$ is more efficient than the OLS (Hellerstein and Imbens, 1999).

Cressie-Read divergence

- EL can be thought of as minimizing the KL divergence (distance) of the empirical distribution and the discrete distribution supported on the sample with a constraint.
- We can consider other distance. E.g., $\sum_{i=1}^{n} p_i \log\left(n^{-1}/p_i\right)$ (reverse KL divergence, exponential tilting, Kitamura and Stutzer, 1998) and $\sum_{i=1}^{n}\left(n \cdot p_i - 1\right)^2$ (Euclidean distance, continuously-updated GMM/Euclidean likelihood).
- Cressie-Read divergence:

$$\frac{1}{\gamma\left(\gamma+1\right)}\sum_{i=1}^{n}\left[\left(n \cdot p_i\right)^{-\gamma} - 1\right], \gamma \in \mathbb{R}.$$

  Special cases: $\gamma = -2$, continuously-updated GMM; $\gamma = -1$, exponential tilting; $\gamma = 0$, EL among many others.
- In the literature, various papers show that some method has certain advantages over other methods, from different perspectives.