Week 4: Nov. 24, 2022

Note Writer: Yu-Chieh Kuo[†]

[†]Department of Information Management, National Taiwan University

Recap

As usual, given the data

$$\begin{array}{lll} \underset{1\times 1}{y_1}, & y_2, \cdots, y_n & \text{dependent variables} \\ \underset{k\times 1}{x_1}, & x_2, \cdots, x_n & \text{independent variables,} \end{array}$$

we define

$$Y \equiv \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad \text{and} \quad X \equiv \begin{pmatrix} x_1' \\ x_2' \\ \vdots \\ x_n' \end{pmatrix},$$

where Y and X are $n \times 1$ and $n \times k$ matrices, respectively. Moreover, given the criterion / objective function $Q_n(\beta)$ and parameters of interest β ($k \times 1$ vector), for any β,

$$Q_n(\beta) = \frac{1}{n}\sum_{i=1}^{n}((y_i - \hat{y}_i))^2,$$

where $\hat{y}_i$ is the predicted y_i. Therefore for the linear and nonlinear case LS, we should obtain

$$\begin{array}{rl} \hat{y}_i = x_i'\hat{\beta} & \text{For the linear LS} \\ \hat{y}_i = f(x_i; \hat{\theta}) & \text{For the nonlinear LS.} \end{array}$$

where $\hat{\theta}$ is $p \times 1$. Here we want to specified the dimension ofr the estimation might not identical with the parameters. Now, $\hat{\theta}$ is defined by FOC

$$\frac{\partial Q_n(\hat{\theta})}{\partial \theta} = 0.$$

For the case of the linear LS, we have a closed-form solution. On the other hand, for the case of the nonlinear LS, we do not have the closed-form solution as

$$\frac{\partial Q_n(\theta)}{\partial \theta} = \frac{-2}{n}\sum_{i=1}^{n}\frac{\partial f_i}{\partial \theta}(y_i - f_i) \quad \text{and} \quad \frac{\partial^2 Q_n(\theta)}{\partial \theta \partial \theta'} = \frac{2}{n}\sum_{i=1}^{n}\frac{\partial f_i}{\partial \theta}\frac{\partial f_i}{\partial \theta'} - \frac{2}{n}\sum_{i=1}^{n}\frac{\partial^2 f_i}{\partial \theta \partial \theta'}\cancel{(y_i - f_i)}^{e_i}.$$

Note that $\frac{\partial Q_n(\theta)}{\partial \theta}$ and $\frac{\partial^2 Q_n(\theta)}{\partial \theta \partial \theta'}$ are $p \times 1$ and $p \times p$, respectively.

Remark. $\hat{\theta} = \arg\min Q_N(\theta)$ and $\theta_\infty = \arg\min Q_\infty(\theta)$. □

Now by applying the Mean Value Theorem, we obtain

$$\frac{\partial Q_n\big(\hat{\theta}\big)}{\partial \theta} = \frac{\partial Q_n(\theta_\infty)}{\partial \theta} + \frac{\partial^2 Q_n(\theta_m)}{\partial \theta \partial \theta'}\big(\hat{\theta} - \theta_\infty\big),$$

where $\theta_m \in [\hat{\theta}, \beta_\infty]$. Since $\hat{\theta} \xrightarrow{p} \theta_\infty$, therefore, it gives $\theta_m \xrightarrow{p} \theta_\infty$.

Thus, if we have

$$\sqrt{n}\frac{\partial Q_n(\theta_\infty)}{\partial \theta} \quad \xrightarrow{d} \quad \mathcal{N}(0, Cov),$$

where

$$Cov \quad \equiv \quad \operatorname*{plim}_{n\to\infty}\left(\sqrt{n}\frac{\partial Q_n(\theta_\infty)}{\partial \theta}\ \sqrt{n}\frac{\partial Q_n(\theta_\infty)}{\partial \theta'}\right).$$

If we next define $M \equiv \operatorname{plim}_{n\to\infty} \frac{\partial^2 Q_n(\theta_m)}{\partial\theta\partial\theta'}$, then

$$\sqrt{n}(\hat{\theta} - \theta_\infty) \quad \xrightarrow{d} \quad \mathcal{N}(0, MCovM').$$

Prediction Errors

Given the data $\mu_i = \text{Prob}\,(y_i \mid x_i)$, and the in-sample and out-sample models

$$y_i = \mu_i + e_i \quad \text{and} \quad y_i^{out} = \mu_i + e_i^{out},$$

where we have to notice that $x_i^{out} = x_i$. Additionally, we define the proejction matrix $P \equiv I_n - X(X'X)^{-1}X'$.

Linear LS

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(y_i\hat{y}_i)^2 \mid X\right] &= \sigma^2 - \frac{k}{n}\sigma^2 + \frac{1}{n}\mu'(I_n - P)\mu \\
\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(y_i^{out} - \hat{y}_i)^2 \mid X\right] &= \sigma^2 + \frac{k}{n}\sigma^2 + \frac{1}{n}\mu'(I_n - P)\mu,
\end{aligned}$$

where $\frac{1}{n}\mu'(I_n - P)\mu$ is called the approximation errors

Nonliear LS

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(y_i\hat{y}_i)^2 \mid X\right] &= \sigma^2 - \frac{p}{n}\sigma^2 + \text{nonlinear terms} \\
\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(y_i^{out} - \hat{y}_i)^2 \mid X\right] &= \sigma^2 + \frac{k}{n}\sigma^2 + \text{nonlinear terms.}
\end{aligned}$$

Remark. Note that θ_0 is the true parameter, and the objective function version is

$$\mathbb{E}[Q_n(\theta) \mid X] = \mathbb{E}[Q_n(\theta_0) \mid X] - \frac{1}{2}\,\mathbb{E}\left[\frac{\partial Q_n(\theta_0)}{\partial \theta'}\left(\frac{\partial^2 Q_n(\theta_0)}{\partial\theta\partial\theta'}\right)^{-1}\frac{\partial Q_n(\theta_0)}{\partial \theta} \mid X\right]$$

$$\mathbb{E}\left[Q_n^{out}(\theta) \mid X\right] = \mathbb{E}\left[Q_n^{out}(\theta_0) \mid X\right] - \frac{1}{2}\,\mathbb{E}\left[\frac{\partial Q_n^{out}(\theta_0)}{\partial \theta'}\left(\frac{\partial^2 Q_n^{out}(\theta_0)}{\partial\theta\partial\theta'}\right)^{-1}\frac{\partial Q_n^{out}(\theta_0)}{\partial \theta} \mid X\right].$$

□

Minimum Distance Estimations and Generalized Method of Moments (GMM) Estimations

Suppose we have an economic model which implies that the endogenous variables y_i''s are determined by a set of ℓ equations. That is,

$$\mathbb{E}[g(y_i, x_i, \beta)] = C,$$

where β is a $k \times 1$ vector, and $g(\cdot)$ and C are $\ell \times 1$ Assume $C = 0$ WLOG, we define

$$g_i \equiv g(y_i, x_i, \beta) \quad \text{and} \quad \hat{g}_i \equiv g(y_i, x_i, \hat{\beta})$$

For the case where $\ell = k$, we can just set

$$\frac{1}{n}\sum_{i=1}^{n} g(y_i, x_i, \hat{\beta}) = 0,$$

and we must notice that $\ell = k$ is a special case. Such a special case leads to

$$\mathbb{E}[x_i e_i = 0] \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^{n} x_i e_i = \frac{1}{n}\sum_{i=1}^{n} x_i(y_i - x_i'\beta) = 0.$$

For the case where $\ell > k$, there may not exist $\hat{\beta}$ such that

$$\frac{1}{n}\sum_{i=1}^{n} g(x_i, y_i, \hat{\beta}) = 0.$$

Hence, we find $\hat{\beta}$ such that the distance between $\frac{1}{n}\sum_{i=1}^{n} g_i$ and 0 ($\ell \times 1$ vector) is minimized. To elaborate, we want to minimize

$$\overbrace{Q_n(\beta)}^{\text{distance}} = \overbrace{\left(\frac{1}{n}\sum_{i=1}^{n} g_i - 0\right)'}^{1\times\ell} \overbrace{\hat{W}}^{\ell\times\ell} \overbrace{\left(\frac{1}{n}\sum_{i=1}^{n} g_i - 0\right)}^{\ell\times 1}.$$

Note that we want to select a weight matrix satisfying $\hat{W} \xrightarrow{p} W$ (W is some matrix, not the true model), symmetricity, and positive-definition. The weight matrix is our choice since the weight matrix sometimes is the function of our data, therefore, we want such a matrix holds some good properties.

Remark. Under some assumptions, the choice of the weight matrix will not affect the consistency. However, it affects the variance, and that's why we need to choose the proper weight matrix. □

Remark. For any $\ell \times \ell$ squared matrix M, if $V'MV > 0$ for any $\ell \times 1$ vector V, M is positive definite.

Now, consider $V'MV$ with $M = I_\ell$, we obtain

$$V'MV = V'V = \sum_{i=1}^{\ell} v_i^2.$$

Given $M \neq I_\ell$, it alters to

$$V'MV = \sum_{i=1}^{\ell}\sum_{j=1}^{\ell} v_i M_{ij} v_j.$$

□

Now define $\bar{g}_n \equiv \frac{1}{n}\sum_{i=1}^{n} g_i$, the objective function alters to $Q_n(\beta) = \bar{g}_n' \hat{W} \bar{g}_n$, where $\overline{g_n}'$ and $\hat{W}$ are $1 \times \ell$ and $\ell \times \ell$. The FOC gives

$$0 = \overbrace{\frac{\partial Q_n(\hat{\beta})}{\partial \beta}}^{k\times 1} = 2 \overbrace{\frac{\partial \hat{\overline{g_n}}'}{\partial \beta}}^{k\times\ell} \hat{W} \overbrace{\hat{\overline{g_n}}}^{\ell\times 1},$$

and the second derivative is

$$\overbrace{\frac{\partial^2 Q_n(\beta)}{\partial\beta\partial\beta'}}^{k\times k} = \frac{\partial}{\partial\beta}\left(\frac{\partial Q_n(\beta)}{\partial\beta}\right)' = 2 \overbrace{\frac{\partial \overline{g_n}'}{\partial \beta}}^{k\times\ell} \hat{W} \overbrace{\frac{\partial \overline{g_n}}{\partial \beta'}}^{\ell\times k} + 2 \overbrace{\frac{\partial}{\partial\beta}}^{k\times 1} \overbrace{\left(\frac{\partial \overline{g_n}'}{\partial \beta}\hat{W}\overline{g_n}\right)'}^{1\times k}.$$

Next whe apply the mean value theorem to $\overline{g_n}(\beta)$ with denoting β_0 and β_m by the true parameters and the mean value parameters. That is,

$$\hat{\overline{g_n}} \equiv \overline{g_n}(\hat{\beta}) = \overline{g_n}(\beta_0) + \frac{\partial \overline{g_n}(\beta_m)}{\partial \beta'}\left(\hat{\beta} - \beta_0\right).$$

Suppose $\hat{\beta} \xrightarrow{p} \beta_0$ and correspondingly $\beta_m \xrightarrow{p} \beta_0$ as well, substituting $\hat{\overline{g_n}}$ into the FOC gives

$$0 = \frac{\partial Q_n(\beta)}{\partial \beta} = 2\frac{\partial \overline{g_n}(\hat{\beta})'}{\partial \beta}\hat{W}\left(\overline{g_n}(\beta_0) + \frac{\partial \overline{g_n}(\beta_m)}{\partial \beta'}(\hat{\beta} - \beta_0)\right).$$

Rearranging the above equation gives

$$\sqrt{n}(\hat{\beta} - \beta_0) = \left(\overbrace{\frac{\partial \overline{g_n}(\hat{\beta})'}{\partial \beta}}^{k\times\ell}\hat{W}\frac{\partial \overline{g_n}(\beta_m)}{\partial \beta'}\right)^{-1}\left(-\frac{\partial \overline{g_n}(\hat{\beta})'}{\partial \beta}\hat{W}\overbrace{\sqrt{n}\overline{g_n}(\beta_0)}^{\ell\times 1}\right),$$

where

$$\begin{aligned} \sqrt{n}\overline{g_n}(\beta_0) &= \overbrace{\sqrt{n}\frac{1}{n}\sum_{i=1}^{n} g(y_i, x_i, \beta_0)}^{\xrightarrow{p}\mathbb{E}[g_i]=0} \\ &\xrightarrow{d} \mathcal{N}\left(0, \mathbb{E}[g_i(\beta_0)g_i(\beta_0)']\right) \\ &\equiv \mathcal{N}(0, Cov). \end{aligned}$$

In addition, we also denote the partial derivative by ∂G for convenience

$$\overbrace{\partial G}^{k\times\ell} \equiv \operatorname*{plim}_{n\to\infty} \frac{\partial \overline{g_n}(\hat{\beta})'}{\partial \beta} = \operatorname*{plim}_{n\to\infty} \frac{\partial \overline{g_n}(\beta_m)'}{\partial \beta}.$$

The distribution of $\sqrt{n}(\hat{\beta} - \beta_0)$ is, finally,

$$\begin{aligned} \sqrt{n}(\hat{\beta} - \beta_0) &\xrightarrow{d} (\partial G W \partial G')^{-1}(-\partial G W \mathcal{N}(0, Cov)) \\ &= \mathcal{N}\left(0, (\partial G W \partial G')^{-1}\partial G W Cov W' \partial G' (\partial G W \partial G')^{-1}\right). \end{aligned}$$

Remark. We did not prove the consistency here, but we know that $\beta_0 \equiv \arg\min Q_\infty(\beta)$ exists under some restrictions of Q_n and Q_∞, which might implies $\hat{\beta} \xrightarrow{p} \beta_0$. □

Remark. We can here easily observe that the weight matrix does not effect the consistency; however, it affects the efficiency. □

Most efficient weight matrix

Suppose $W = Cov^{-1}$, then the asymptotic covariance of $\sqrt{n}(\hat{\beta} - \beta_0)$ is

$$\left(\partial G Cov^{-1} \partial G'\right)^{-1} \partial G Cov^{-1} Cov Cov^{-1} \partial G' \left(\partial G Cov^{-1} \partial G'\right)^{-1} = \left(\partial G Cov^{-1} \partial G'\right)^{-1}.$$

Theorem. For all W,

$$(\partial G W \partial G')^{-1} \partial G W Cov W' \partial G' (\partial G W \partial G')^{-1} \geq \left(\partial G Cov^{-1} \partial G'\right)^{-1}.$$

The proof of the theorem did not be completed here. Go to wiki to see more, said by Hoho. □

Remark. FOr any two square matrices M_1, M_2, we say

$$M_1 \geq M_2 \iff M_1 - M_2 \geq 0,$$

Note that ≥ 0 for the matrix operation means the positive semi-definition. □

Procedures of efficient GMM

To clarify procedures of efficient GMM, we conclude

1. Choose $\hat{W} = W = I_\ell$, then we have the consistent estimator $\hat{\beta}$.
2. Obtain

$$\hat{Cov} \equiv \overbrace{\frac{1}{n} \sum_{i=1}^{n} g(y_i, x_i, \hat{\beta}) g(y_i, x_i, \hat{\beta})'}^{\hat{W}^{-1}} \xrightarrow{p} \overbrace{\mathbb{E}[g_i(\beta_0) g_i(\beta_0)']}^{W^{-1}} \equiv Cov.$$

3. Do the minimized distance estimation again with $\hat{W} = \hat{Cov}^{-1}$.

After the procedures above, we that obtain $\hat{\beta}_{efficient}$.

Notes on Consistency

We've already known that given some objective finctions, we have

$$Q_n(\theta) \xrightarrow{p} Q_\infty(\theta) \quad \text{and} \quad \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \xrightarrow{p} \mathbb{E}[y_i - \hat{y}_i]^2$$

for (point-wisely, i.e., for each θ_i) consistency. Extending such idea the the previous matrix gives

$$\left(\frac{1}{n} \sum_{i=1}^{n} g_i\right)' \hat{W} \left(\frac{1}{n} \sum_{i=1}^{n} g_i\right) \xrightarrow{p} (\mathbb{E}[g_i])' W (\mathbb{E}[g_i]) = 0.$$

Two commonly used theorems for consistency

Here we introduce two commonly used theorems for consistency.

1. If θ belongs to a compact (i.e., closed and bounded) set, it gives $Q_n(\theta) \xrightarrow{p} Q_\infty(\theta)$ uniformly. Such the convergence reveals $\hat{\theta}_n \xrightarrow{p} \theta_\infty$.

   Note that the point-wise convergence means $|Q_n(\theta) - Q_\infty(\theta)| \xrightarrow{p} 0$, and the uniform convergence means $\sup_\theta |Q_n(\theta) - Q_\infty(\theta)| \xrightarrow{p} 0$.

2. Assume $Q_n(\theta)$ is convex for all n, we can also claim $\hat{\theta}_n \xrightarrow{p} \theta_\infty$.

This subsection refers to Large Sample Theory in the Handbook of Econometrics, Hayashi Ch7-2.

General Notes

We here give some general notes on econometrics through the whole class.

1. The mathematical arguments are appliable to any extreme estimations. That is,

$$\begin{aligned}
\text{LS case:} \quad & Q_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \xrightarrow{p} \mathbb{E}[y_i - \hat{y}_i]^2 \\
\text{GMM case:} \quad & Q_n(\theta) = \overline{g_n}'\hat{W}\overline{g_n} \xrightarrow{p} 0 \\
\text{ML case:} \quad & Q_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\log f(y_i, x_i, \theta) \xrightarrow{p} \mathbb{E}[\log f(y_i, x_i, \theta)].
\end{aligned}$$

   - LS is a special case of GMM where $y_i = f(x_i, \beta) + e_i$, $g_i = x_i e_i$ (moment conditions), and $\mathbb{E}[g_i] = \mathbb{E}[x_i e_i] = 0$.
   - LS is a special case of ML where $y_i = f(x_i, \beta) + e_i$ and $e_i \sim$ Normal.
   - Instrumental variable estimation belongs to GMM, where $g_i = z_i e_i$ and $\mathbb{E}[g_i] = \mathbb{E}[z_i e_i] = 0$.
   - Mostly, ML is the most efficient in the sense of having the lowest asymptotic covariance. The reason is that ML use the information in the density.
   - Denote θ by a $k \times 1$ parameter vector, $\hat{\theta}$ by the $k \times 1$ estimator, θ_0 by the $k \times 1$ true parameters vector, and Cov by the $k \times k$ asymptotic covariance. We have

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, Cov) \quad \text{and} \quad \frac{\sqrt{n}(\hat{\theta}_j - \theta_{0,j})}{\sqrt{Cov_{jj}}} \xrightarrow{d} \mathcal{N}(0, 1).$$

   - Mostly, the objective function is with the quadratic form, which leads to there distribution asymptotes to chi-square distribution. **why??????**