Contents

Chapter 11

THE GENERALIZED METHOD OF MOMENTS ESTIMATION

Given that a random sample $x_1, x_2, \ldots, x_n$ are drawn from a population which is characterized by the parameter $\boldsymbol{\theta}$ whose true value is $\boldsymbol{\theta}_\circ$. If we can identify a vector of functions $\mathbf{g}(x;\boldsymbol{\theta})$ of the random variable x and the parameter $\boldsymbol{\theta}$ such that the true parameter value $\boldsymbol{\theta}_\circ$ uniquely solves the following *population moment condition*

$$\mathrm{E}\big[\mathbf{g}(x;\boldsymbol{\theta})\big] = \mathbf{0}, \tag{11.1}$$

while the estimator $\hat{\boldsymbol{\theta}}$ is the unique solution to the *sample moment condition*

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{g}(x_i;\boldsymbol{\theta}) = \mathbf{0}, \tag{11.2}$$

then, under some regularity conditions, we can show that $\hat{\boldsymbol{\theta}}$ is consistent and asymptotically normal

$$\hat{\boldsymbol{\theta}} \overset{\mathrm{A}}{\sim} \mathcal{N}(\boldsymbol{\theta}_\circ,\ \frac{1}{n}\,\mathbf{G}(\boldsymbol{\theta}_\circ)'^{-1}\boldsymbol{\Omega}(\boldsymbol{\theta}_\circ)\,\mathbf{G}(\boldsymbol{\theta}_\circ)^{-1}), \tag{11.3}$$

where $\boldsymbol{\theta}_\circ$ denotes the true value of the parameter $\boldsymbol{\theta}$,

$$\mathbf{G}(\boldsymbol{\theta}) = \mathrm{E}\left[\frac{\partial\mathbf{g}(x;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right], \tag{11.4}$$

and

$$\boldsymbol{\Omega}(\boldsymbol{\theta}) = \mathrm{E}\big[\mathbf{g}(x;\boldsymbol{\theta})\mathbf{g}(x;\boldsymbol{\theta})'\big]. \tag{11.5}$$

Any estimator defined in such a setup is referred to as a *Generalized Method of Moment* (GMM) estimator. The approach of first identifying some moment condition and then deriving the corresponding GMM estimator from its sample counterpart has become a very popular way of generating new estimators in econometrics.

A Simple Example Given the random sample $x_1, x_2, \ldots, x_n$ drawn from an unspecified population with a population mean μ and variance σ^2, we have derived the asymptotic properties of the sample mean $\bar{x}$ as an estimator of μ in such a case by directly applying law of large numbers and the central limit theorem. We now show that the asymptotic analysis of the sample mean can fit into the GMM framework.

Let's consider the function $g(x_i;\mu) \equiv x_i - \mu$ which gives the following population moment condition:

$$\mathrm{E}[g(x_i;\mu)] = \mathrm{E}(x_i) - \mu = 0.$$

It is obvious that the only solution of μ is the true value of the population mean μ to which $\mathrm{E}(x_i)$ is equal. The *sample counterpart* of the population moment condition is

$$\frac{1}{n}\sum_{i=1}^{n} g(x_i;\mu) = \frac{1}{n}\sum_{i=1}^{n} x_i - \mu = 0,$$

and the solution of μ is nothing but the sample mean $\bar{x}$. So the sample mean $\bar{x}$ is actually a GMM estimator of μ. Consequently, we can apply the general results for the GMM estimator to establish the consistency and asymptotic normality of the GMM estimator $\bar{x}$. It is also easy to prove that the asymptotic variance of the GMM estimator $\bar{x}$ is $n^{-1}\sigma^2$. Although the GMM argument here appears tedious, the idea is important and has wide applicability.

11.1 Consistency and Asymptotic Normality

The following argument for proving the consistency of the GMM estimator helps illustrate the key idea of the GMM approach.

We first note, if the second moment of $\mathbf{g}(x;\boldsymbol{\theta})$ exists, then law of large numbers implies

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{g}(x_i;\boldsymbol{\theta}) \quad \xrightarrow{\text{p}} \quad \mathrm{E}\big[\mathbf{g}(x;\boldsymbol{\theta})\big]. \tag{11.6}$$

It means that the population moment condition (11.1) can be approximated by the sample moment condition (11.2). If the estimator $\hat{\boldsymbol{\theta}}$ solves the sample moment condition (11.2) irrespective of the sample size, then its probability limit, say, $\boldsymbol{\theta}^*$ must also solve the probability limit of (11.2) which is the population moment condition (11.1). But by definition the true parameter value $\boldsymbol{\theta}_\circ$ uniquely solves the population moment condition (11.1), so the probability limit $\boldsymbol{\theta}^*$ must be equal to the true parameter value $\boldsymbol{\theta}_\circ$. That is, $\hat{\boldsymbol{\theta}}$ is a consistent estimator of $\boldsymbol{\theta}$. In other words, if we know the true parameter value is the solution to certain population moment condition, then the solution to its sample counterpart will be a consistent estimator.

The proof of asymptotic normality is based on Taylor expansion and the central limit theorem. Given that $\hat{\boldsymbol{\theta}}$ converges in probability to $\boldsymbol{\theta}_\circ$ and that $\mathbf{g}$ is differentiable with respect to $\boldsymbol{\theta}$, then for sufficiently large n the first-order Taylor expansion of (11.2) around the true value $\boldsymbol{\theta}_\circ$ gives the following approximation

$$\mathbf{0} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{g}(x_i;\hat{\boldsymbol{\theta}}) \approx \frac{1}{n}\sum_{i=1}^{n} \mathbf{g}(x_i;\boldsymbol{\theta}_\circ) + \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \mathbf{g}(x_i;\boldsymbol{\theta}_\circ)}{\partial \boldsymbol{\theta}}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_\circ) \tag{11.7}$$

or

$$\sqrt{n}\,(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_\circ) \approx -\left[\frac{1}{n}\sum_{i=1}^{n} \frac{\partial \mathbf{g}(x_i;\boldsymbol{\theta}_\circ)}{\partial \boldsymbol{\theta}}\right]^{-1} \sqrt{n}\,\frac{1}{n}\sum_{i=1}^{n} \mathbf{g}(x_i;\boldsymbol{\theta}_\circ). \tag{11.8}$$

Provided that the second moment of $\partial \mathbf{g}(x_i; \boldsymbol{\theta}_\circ)/\partial \boldsymbol{\theta}$ exists, law of large numbers again implies

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\partial \mathbf{g}(x_i; \boldsymbol{\theta}_\circ)}{\partial \boldsymbol{\theta}} \;\xrightarrow{\text{p}}\; \mathbf{G}(\boldsymbol{\theta}_\circ), \tag{11.9}$$

and the central limit theorem implies

$$\sqrt{n}\left\{\frac{1}{n}\sum_{i=1}^{n} \mathbf{g}(x_i; \boldsymbol{\theta}_\circ) - \mathrm{E}\big[\mathbf{g}(x; \boldsymbol{\theta}_\circ)\big]\right\} \;\xrightarrow{\text{d}}\; \mathbf{u} \;\sim\; \mathcal{N}(\mathbf{0},\ \boldsymbol{\Omega}(\boldsymbol{\theta}_\circ)), \tag{11.10}$$

where $\mathrm{E}\big[\mathbf{g}(x; \boldsymbol{\theta}_\circ)\big] = \mathbf{0}$. Consequently,

$$\sqrt{n}\,(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_\circ) \xrightarrow{\text{d}} \;-\mathbf{G}(\boldsymbol{\theta}_\circ)^{-1}\cdot\mathbf{u} \;\sim\; \mathcal{N}(\mathbf{0},\;\; \mathbf{G}(\boldsymbol{\theta}_\circ)^{-1}\boldsymbol{\Omega}(\boldsymbol{\theta}_\circ)\,\mathbf{G}(\boldsymbol{\theta}_\circ)'^{-1}), \tag{11.11}$$

which implies (11.3).

The Estimator of the Asymptotic Variance-Covariance Matrix Based on law of large numbers and (11.9), it is readily seen that the following statistic is a consistent estimator of the asymptotic variance-covariance matrix $\mathbf{G}(\boldsymbol{\theta}_\circ)^{-1}\boldsymbol{\Omega}(\boldsymbol{\theta}_\circ)\,\mathbf{G}(\boldsymbol{\theta}_\circ)'^{-1}$ of the GMM estimator $\hat{\boldsymbol{\theta}}$:

$$\frac{1}{n}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial \mathbf{g}(x_i; \hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}}\right]^{-1} \frac{1}{n}\sum_{i=1}^{n}\big[\mathbf{g}(x_i; \hat{\boldsymbol{\theta}})\mathbf{g}(x_i; \hat{\boldsymbol{\theta}})'\big]\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial \mathbf{g}(x_i; \hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}}\right]'^{-1}. \tag{11.12}$$

11.2 Regularity Conditions and Identification

In proving consistency and asymptotic normality of the GMM estimator, we have used law of large numbers and the central limit theorem. Obviously, certain assumptions are required before we can apply these theorems. The assumptions that ensure the validity of the GMM estimation are called regularity conditions and they can be divided into four categories:

1. Conditions that ensure the differentiability of $\mathbf{g}(x; \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. For example, $\mathbf{g}(x; \boldsymbol{\theta})$ is usually assumed to be twice continuously differentiable with respect to $\boldsymbol{\theta}$.

2. Conditions that restrict the moments of $\mathbf{g}(x; \boldsymbol{\theta})$ and its derivatives with respect to $\boldsymbol{\theta}$. For example, the second moments of $\mathbf{g}(x; \boldsymbol{\theta})$ and its first derivative are usually assumed to be finite.

3. Conditions that restrict the range of the possible values which the parameter $\boldsymbol{\theta}$ can take. For example, $\boldsymbol{\theta}$ is not allowed to have infinite value and the true value $\boldsymbol{\theta}_\circ$ may not be at the boundary of the permissible range of $\boldsymbol{\theta}$ (if $\boldsymbol{\theta}_\circ$ is on the boundary of the permissible range of $\boldsymbol{\theta}$, then convergence to $\boldsymbol{\theta}_\circ$ cannot take place freely from all directions).

4. The solution to the population moment condition $\mathrm{E}\big[\mathbf{g}(x; \boldsymbol{\theta})\big] = \mathbf{0}$ must be unique and the unique solution must be the true value $\boldsymbol{\theta}_\circ$ of the parameter.

The first three categories of regularity conditions are somewhat technical and are routinely assumed. However, we do need to make special efforts to check the validity of the last one in each application. This last condition is referred to as *the identification condition* because it allows us to identify the true parameter value $\boldsymbol{\theta}_\circ$ for estimation.

An obvious necessary condition for identification is that the row number, say, m of the vector $\mathbf{g}(x;\boldsymbol{\theta})$ is no less than the row number, say, k of the parameter vector $\boldsymbol{\theta}$. That is, the number of individual population moment conditions cannot be smaller than the number of parameters to be estimated. If $m < k$, then the population moment condition will have multiple solutions of which all but one can be the true value so that the resulting GMM estimator does not necessarily converge to the true parameter value. This is the so-called under-identification problem.

The identification condition is implicitly assumed in the previous analysis of the GMM estimation. In fact, we have made a stronger assumption that $m = k$ so that the derivative $\mathbf{G}(x;\boldsymbol{\theta})$ of $\mathbf{g}(x;\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ is a square matrix and invertible. This is the so-called just-identification case. In section 11.4 we will examine the over-identification case with $m > k$.

11.3 The GMM Interpretation of the OLS Estimation

For the linear regression model[1]

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i, \tag{11.13}$$

let's assume that the sample $\{y_i,\ \mathbf{x}_i'\}$, $i = 1, \ldots, n$, are i.i.d., and that $\mathrm{E}(\varepsilon_i) = 0$. In the present framework, the explanatory variables $\mathbf{x}_i$ are stochastic and, following the arguments in Chapter 10, we have to assume the following population moment condition:

$$\mathrm{E}(\mathbf{x}_i\varepsilon_i) = \mathrm{E}[\mathbf{x}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta})] = \mathbf{0}. \tag{11.14}$$

The dimensions of $\mathbf{x}_i$ and the zero vector on the right hand side are both k. So we have in fact k population moment conditions which are just enough for us to estimate the k parameters in $\boldsymbol{\beta}$, i.e., we have a just-identification case.[2] The corresponding sample moment condition is

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta}) = \mathbf{0}, \tag{11.15}$$

which can be written as

$$\frac{1}{n}\mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{0} \qquad \text{or} \qquad \mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{y}. \tag{11.16}$$

---

[1] We treat $\boldsymbol{\beta}$ not only as the notation for the regression coefficients but also as their true values. Such notational ambiguity has been existing throughout the earlier chapters. Better notations for the true values of the regression coefficients may be $\boldsymbol{\beta}_\circ$.

[2] If the $\mathbf{x}_i$ contains the constant term 1, then one of the moment conditions is $\mathrm{E}(y_i - \mathbf{x}_i'\boldsymbol{\beta}) = 0$ or $\mathrm{E}(y_i) = \mathrm{E}(\mathbf{x}_i)'\boldsymbol{\beta}$.

But this is equivalent to the first-order condition for the OLS estimation. Hence, the OLS estimator, which are solved from the above sample moment conditions, can be considered as a GMM estimator.

In order to apply the asymptotic theory for the GMM estimation, we need to first evaluate[3]

$$\boldsymbol{\Omega}(\boldsymbol{\beta}) \equiv \mathrm{E}[\mathbf{x}_i \varepsilon_i \varepsilon_i \mathbf{x}_i'] = \mathrm{E}\big[\mathrm{E}(\varepsilon_i^2 \,|\, \mathbf{x}_i)\mathbf{x}_i \mathbf{x}_i'\big] = \mathrm{E}(\sigma^2 \mathbf{x}_i \mathbf{x}_i') = \sigma^2 \mathrm{E}(\mathbf{x}_i \mathbf{x}_i') \tag{11.17}$$

and

$$\mathbf{G}(\boldsymbol{\beta}) \equiv \mathrm{E}\left[\frac{\partial \mathbf{x}_i (y_i - \mathbf{x}_i'\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right] = -\mathrm{E}(\mathbf{x}_i \mathbf{x}_i'). \tag{11.18}$$

It is important to note that one of the assumptions (Assumption 5) we made for a multiple linear regression model is

$$\lim_{n\to\infty} \frac{1}{n}\mathbf{X}'\mathbf{X} = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i' = \mathbf{Q},$$

with $\mathbf{Q}$ being a finite and p.d. matrix, we can equate $\mathrm{E}(\mathbf{x}_i \mathbf{x}_i')$ to $\mathbf{Q}$. That is, we have $\boldsymbol{\Omega}(\boldsymbol{\beta}) = \sigma^2\mathbf{Q}$ and $\mathbf{G}(\boldsymbol{\beta}) = -\mathbf{Q}$.

Now, following the general asymptotic theory for the GMM estimation, we then have that the OLS estimator $\mathbf{b}$ is consistent:

$$\mathbf{b} \xrightarrow{\ \mathrm{p}\ } \boldsymbol{\beta}. \tag{11.19}$$

and

$$\sqrt{n}(\mathbf{b} - \boldsymbol{\beta}) \xrightarrow{\ \mathrm{d}\ } \mathcal{N}(\mathbf{0},\ \sigma^2\mathbf{Q}^{-1}). \tag{11.20}$$

11.4 The GMM Interpretation of the MLE

Suppose the sample $\{x_i\}$, $i = 1, \ldots, n$ are i.i.d. with the density function $f(x|\boldsymbol{\theta}_\circ)$, where $\boldsymbol{\theta}_\circ$ is an unknown k-dimensional parameter to be estimated, then we have shown in Chapter 9 that

$$\mathrm{E}\left[\frac{\partial \ln f(x_i|\boldsymbol{\theta}_\circ)}{\partial \boldsymbol{\theta}}\right] = \mathbf{0}, \tag{11.21}$$

which can be viewed as k population moment conditions that are just enough for us to estimate the k-dimensional parameter $\boldsymbol{\theta}_\circ$. The corresponding sample counterpart is

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\partial \ln f(x_i|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{0}, \tag{11.22}$$

and the solution, denoted by $\hat{\boldsymbol{\theta}}$, is certainly the MLE of $\boldsymbol{\theta}_\circ$. In other words, the MLE can be viewed as a GMM estimator.

---

[3]Here, we have further assumed that $\mathrm{E}(\varepsilon_i^2 \,|\, \mathbf{x}_i) = \sigma^2$ (i.e., ε_i is *homoscedastic* with respect to $\mathbf{x}_i$), which will be true if $\mathbf{x}_i$ is assumed to be nonstochastic.

In order to apply the asymptotic theory for the GMM estimation, let's first define

$$\boldsymbol{\Omega}(\boldsymbol{\theta}) = \mathrm{E}\left[\frac{\partial \ln f(x_i|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]\left[\frac{\partial \ln f(x_i|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]' \tag{11.23}$$

and

$$\mathbf{G}(\boldsymbol{\theta}) = \mathrm{E}\left[\frac{\partial^2 \ln f(x_i|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\right]. \tag{11.24}$$

It has also been shown in Chapter 9 that $\boldsymbol{\Omega}(\boldsymbol{\theta}_\circ) = -\mathbf{G}(\boldsymbol{\theta}_\circ)$. Now following the general asymptotic theory for the GMM estimation, we then have the well-known results that the MLE $\hat{\boldsymbol{\theta}}$ is consistent and[4]

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_\circ) \quad \xrightarrow{\mathrm{d}} \quad \mathcal{N}(\mathbf{0},\ \boldsymbol{\Omega}(\boldsymbol{\theta}_\circ)^{-1}). \tag{11.25}$$

11.5 The GMM Estimation in the Over-Identification Case

If in the population moment condition

$$\mathrm{E}\big[\mathbf{g}(x;\boldsymbol{\theta}_\circ)\big] = \mathbf{0} \tag{11.26}$$

the row number of $\mathbf{g}$ is strictly greater than the row number of the parameter vector $\boldsymbol{\theta}$, then it is not possible to solve its sample counterpart

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{g}(x_i;\boldsymbol{\theta}) = \mathbf{0}, \tag{11.27}$$

because the number of equations is greater than the number of parameters to be solved. What we could do in such a case is to find a value of $\boldsymbol{\theta}$ that makes the sample moment condition as close to zero as possible based on the following quadratic form:

$$\min_{\boldsymbol{\theta}} \left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{g}(x_i;\boldsymbol{\theta})\right]' \mathbf{W} \left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{g}(x_i;\boldsymbol{\theta})\right] \tag{11.28}$$

where $\mathbf{W}$ is some positive definite weighting matrix of constants.

---

[4]In Chapter 9 we did not assume the sample to be identically distributed; i.e., the density functions $f_i(x_i|\boldsymbol{\theta}_\circ)$ have the subscript i, indicating they are all different. In such a case, the variance-covariance matrix of the asymptotical distribution is the inverse of

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[-\frac{\partial^2 \ln f_i(x_i|\boldsymbol{\theta}_\circ)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\right].$$

It is readily seen that such a matrix reduces to $\boldsymbol{\Omega}(\boldsymbol{\theta}_\circ) = -\mathbf{G}(\boldsymbol{\theta}_\circ)$ in the present i.i.d. case.

Given the assumption that $\mathbf{G}(\boldsymbol{\theta})$ has full column rank and some additional regularity conditions, $\hat{\boldsymbol{\theta}}$ is consistent. To see this we note that the first-order condition for the minimization problem (11.28) is

$$\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\mathbf{g}(x_i;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right]'\mathbf{W}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{g}(x_i;\boldsymbol{\theta})\right]=\mathbf{0}, \tag{11.29}$$

which can be viewed as the sample counterpart of the moment conditions

$$\mathbf{G}(\boldsymbol{\theta})'\mathbf{W}\,\mathrm{E}\big[\mathbf{g}(x;\boldsymbol{\theta})\big]=\mathbf{0}. \tag{11.30}$$

Given that $\mathbf{G}(\boldsymbol{\theta})$ has full column rank and that $\mathbf{W}$ is nonsingular, then only the ture parameter value $\boldsymbol{\theta}_\circ$ can satisfy these moment conditions, which in turns implies that the GMM estimator $\hat{\boldsymbol{\theta}}$ is consistent.

We can further show that $\hat{\boldsymbol{\theta}}$ is asymptotically normal:[5]

$$\hat{\boldsymbol{\theta}} \overset{\mathrm{A}}{\sim} \mathcal{N}\Big(\boldsymbol{\theta}_\circ,\ \frac{1}{n}\big[\mathbf{G}(\boldsymbol{\theta}_\circ)'\mathbf{W}\,\mathbf{G}(\boldsymbol{\theta}_\circ)\big]^{-1}\mathbf{G}(\boldsymbol{\theta}_\circ)'\mathbf{W}\,\boldsymbol{\Omega}(\boldsymbol{\theta}_\circ)\,\mathbf{W}\,\mathbf{G}(\boldsymbol{\theta}_\circ)\big[\mathbf{G}(\boldsymbol{\theta}_\circ)'\mathbf{W}\,\mathbf{G}(\boldsymbol{\theta}_\circ)\big]^{-1}\Big). \tag{11.31}$$

Obviously, different weighting matrix $\mathbf{W}$ will give different estimators with different asymptotic variance-covariance matrices. That is, the efficiency of the resulting GMM estimators

[5] Given that $\hat{\boldsymbol{\theta}}$ converges in probability to $\boldsymbol{\theta}_\circ$ and that $\mathbf{g}$ is twice differentiable with respect to $\boldsymbol{\theta}$, then for sufficiently large n the Taylor expansion of (11.29) around the true value $\boldsymbol{\theta}_\circ$ gives the following approximation

$$\begin{aligned}\mathbf{0}&=\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\mathbf{g}(x_i;\hat{\boldsymbol{\theta}})}{\partial\boldsymbol{\theta}}\right]'\mathbf{W}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{g}(x_i;\hat{\boldsymbol{\theta}})\right]\\ &\approx\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\mathbf{g}(x_i;\boldsymbol{\theta}_\circ)}{\partial\boldsymbol{\theta}}\right]'\mathbf{W}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{g}(x_i;\boldsymbol{\theta}_\circ)\right]+\left\{\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\mathbf{g}(x_i;\boldsymbol{\theta}_\circ)}{\partial\boldsymbol{\theta}}\right]'\mathbf{W}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\mathbf{g}(x_i;\boldsymbol{\theta}_\circ)}{\partial\boldsymbol{\theta}}\right]+\mathbf{S}\right\}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_\circ)\end{aligned}$$

or

$$\sqrt{n}\,(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_\circ)\approx\left\{-\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\mathbf{g}(x_i;\boldsymbol{\theta}_\circ)}{\partial\boldsymbol{\theta}}\right]'\mathbf{W}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\mathbf{g}(x_i;\boldsymbol{\theta}_\circ)}{\partial\boldsymbol{\theta}}\right]+\mathbf{S}\right\}^{-1}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\mathbf{g}(x_i;\boldsymbol{\theta}_\circ)}{\partial\boldsymbol{\theta}}\right]'\mathbf{W}\left[\sqrt{n}\,\frac{1}{n}\sum_{i=1}^{n}\mathbf{g}(x_i;\boldsymbol{\theta}_\circ)\right],$$

where $\mathbf{S}$ is a $k\times k$ matrix in which the jth column is

$$\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2\mathbf{g}(x_i;\boldsymbol{\theta}_\circ)}{\partial\boldsymbol{\theta}\partial\theta_j}\right]'\mathbf{W}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{g}(x_i;\boldsymbol{\theta}_\circ)\right].$$

We note that (11.6) implies that $\mathbf{S}$ converges in probability to zero. Thus, by (11.9) and (11.10), we have

$$\begin{aligned}\sqrt{n}\,(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_\circ)&\overset{\mathrm{d}}{\longrightarrow}-\big[\mathbf{G}(\boldsymbol{\theta}_\circ)'\mathbf{W}\,\mathbf{G}(\boldsymbol{\theta}_\circ)\big]^{-1}\mathbf{G}(\boldsymbol{\theta}_\circ)'\mathbf{W}\cdot\mathbf{u}\\ &\sim\ \mathcal{N}\Big(\mathbf{0},\ \big[\mathbf{G}(\boldsymbol{\theta}_\circ)'\mathbf{W}\,\mathbf{G}(\boldsymbol{\theta}_\circ)\big]^{-1}\mathbf{G}(\boldsymbol{\theta}_\circ)'\mathbf{W}\,\boldsymbol{\Omega}(\boldsymbol{\theta}_\circ)\,\mathbf{W}\,\mathbf{G}(\boldsymbol{\theta}_\circ)\big[\mathbf{G}(\boldsymbol{\theta}_\circ)'\mathbf{W}\,\mathbf{G}(\boldsymbol{\theta}_\circ)\big]^{-1}\Big).\end{aligned}$$

depends on the weighting matrix $\mathbf{W}$. It can be shown that[6]

$$\big[\mathbf{G}(\boldsymbol{\theta}_\circ)'\mathbf{W}\,\mathbf{G}(\boldsymbol{\theta}_\circ)\big]^{-1}\mathbf{G}(\boldsymbol{\theta}_\circ)'\mathbf{W}\,\boldsymbol{\Omega}(\boldsymbol{\theta}_\circ)\,\mathbf{W}\,\mathbf{G}(\boldsymbol{\theta}_\circ)\big[\mathbf{G}(\boldsymbol{\theta}_\circ)'\mathbf{W}\,\mathbf{G}(\boldsymbol{\theta}_\circ)\big]^{-1} \geq \big[\mathbf{G}(\boldsymbol{\theta}_\circ)'\boldsymbol{\Omega}(\boldsymbol{\theta}_\circ)^{-1}\mathbf{G}(\boldsymbol{\theta}_\circ)\big]^{-1} \tag{11.32}$$

for any positive definite $\mathbf{W}$. This finding implies that the most efficient GMM estimator $\hat{\boldsymbol{\theta}}$ can be obtained by setting $\mathbf{W} = \widetilde{\boldsymbol{\Omega}}^{-1}$ for any consistent estimator $\widetilde{\boldsymbol{\Omega}}$ of $\boldsymbol{\Omega}(\boldsymbol{\theta}_\circ) = \mathrm{E}\big[\mathbf{g}(x;\boldsymbol{\theta}_\circ)\mathbf{g}(x;\boldsymbol{\theta}_\circ)'\big]$ and then solving the following minimization problem:

$$\min_{\boldsymbol{\theta}} \left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{g}(x_i;\boldsymbol{\theta})\right]' \widetilde{\boldsymbol{\Omega}}^{-1} \left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{g}(x_i;\boldsymbol{\theta})\right]. \tag{11.33}$$

The resulitng GMM estimator is denoted again as $\hat{\boldsymbol{\theta}}$ which from now on will represent such an efficient GMM estimator. It is readily seen that $\hat{\boldsymbol{\theta}}$ is consistent and asymptotically normal

$$\hat{\boldsymbol{\theta}} \overset{\mathrm{A}}{\sim} \mathcal{N}\Big(\boldsymbol{\theta}_\circ,\ \frac{1}{n}\big[\mathbf{G}(\boldsymbol{\theta}_\circ)'\boldsymbol{\Omega}(\boldsymbol{\theta}_\circ)^{-1}\mathbf{G}(\boldsymbol{\theta}_\circ)\big]^{-1}\Big). \tag{11.34}$$

The derivation of the GMM estimator $\hat{\boldsymbol{\theta}}$ with over-identified moment condition essentially requires a two-stage procedure because a preliminary estimator is needed for calculating the weighting matrix $\widetilde{\boldsymbol{\Omega}}$. A particularly simple choice of $\widetilde{\boldsymbol{\Omega}}$ is

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{g}(x_i;\tilde{\boldsymbol{\theta}})\mathbf{g}(x_i;\tilde{\boldsymbol{\theta}})' \tag{11.35}$$

where $\tilde{\boldsymbol{\theta}}$ is a preliminary estimator of $\boldsymbol{\theta}$ which can be any *consistent* estimator of $\boldsymbol{\theta}$. A common one can be derived by solving the following simpler minimization problem

$$\min_{\boldsymbol{\theta}} \left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{g}(x_i;\boldsymbol{\theta})\right]' \left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{g}(x_i;\boldsymbol{\theta})\right]. \tag{11.36}$$

That is, the preliminary consistent estimator $\tilde{\boldsymbol{\theta}}$ itself is a GMM estimator based on an especially simple weighting matrix $\mathbf{W} = \mathbf{I}$.

The asymptotic variance-covariance matrix can be consistently estimated by

$$\frac{1}{n}\left\{\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\mathbf{g}(x_i;\hat{\boldsymbol{\theta}})}{\partial\boldsymbol{\theta}}\right]'\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{g}(x_i;\hat{\boldsymbol{\theta}})\mathbf{g}(x_i;\hat{\boldsymbol{\theta}})'\right]^{-1}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\mathbf{g}(x_i;\hat{\boldsymbol{\theta}})}{\partial\boldsymbol{\theta}}\right]\right\}^{-1}. \tag{11.37}$$

which can be compared to the one for the just-identified case in (11.12).

---

[6]Let $\mathbf{G} \equiv \mathbf{G}(\boldsymbol{\theta}_\circ)$ and $\boldsymbol{\Omega} \equiv \boldsymbol{\Omega}(\boldsymbol{\theta}_\circ)$, then $(\mathbf{G}'\mathbf{W}\mathbf{G})^{-1}\mathbf{G}'\mathbf{W}\boldsymbol{\Omega}\mathbf{W}\mathbf{G}(\mathbf{G}'\mathbf{W}\mathbf{G})^{-1} - (\mathbf{G}'\boldsymbol{\Omega}^{-1}\mathbf{G})^{-1} = (\mathbf{G}'\mathbf{W}\mathbf{G})^{-1}\big[\mathbf{G}'\mathbf{W}\boldsymbol{\Omega}\mathbf{W}\mathbf{G} - \mathbf{G}'\mathbf{W}\mathbf{G}(\mathbf{G}'\boldsymbol{\Omega}^{-1}\mathbf{G})^{-1}\mathbf{G}'\mathbf{W}\mathbf{G}\big](\mathbf{G}'\mathbf{W}\mathbf{G})^{-1} = (\mathbf{G}'\mathbf{W}\mathbf{G})^{-1}\mathbf{G}'\mathbf{W}\boldsymbol{\Omega}\big[\boldsymbol{\Omega}^{-1} - \boldsymbol{\Omega}^{-1}\mathbf{G}(\mathbf{G}'\boldsymbol{\Omega}^{-1}\mathbf{G})^{-1}\mathbf{G}'\boldsymbol{\Omega}^{-1}\big]\boldsymbol{\Omega}\mathbf{W}\mathbf{G}(\mathbf{G}'\mathbf{W}\mathbf{G})^{-1}$ which can be expressed in the form of $\mathbf{A}\boldsymbol{\Omega}\mathbf{A}'$ with $\mathbf{A} = (\mathbf{G}'\mathbf{W}\mathbf{G})^{-1}\mathbf{G}'\mathbf{W}\boldsymbol{\Omega}\big[\boldsymbol{\Omega}^{-1} - \boldsymbol{\Omega}^{-1}\mathbf{G}(\mathbf{G}'\boldsymbol{\Omega}^{-1}\mathbf{G})^{-1}\mathbf{G}'\boldsymbol{\Omega}^{-1}\big]$. Since $\mathbf{A}\boldsymbol{\Omega}\mathbf{A}'$ is necessarily a p.d. matrix, we therefore have $(\mathbf{G}'\mathbf{W}\mathbf{G})^{-1}\mathbf{G}\mathbf{W}'\boldsymbol{\Omega}\mathbf{W}\mathbf{G}(\mathbf{G}'\mathbf{W}\mathbf{G})^{-1} \geq (\mathbf{G}'\boldsymbol{\Omega}^{-1}\mathbf{G})^{-1}$.

11.6 The GMM Interpretation of the Instrumental Variable Estimation

For the linear regression model (11.13), suppose the stochastic explanatory variables $\mathbf{x}_i$ are *endogenous*; i.e.,

$$\mathrm{E}(\mathbf{x}_i \varepsilon_i) = \mathrm{E}[\mathbf{x}_i (y_i - \mathbf{x}_i' \boldsymbol{\beta})] \neq \mathbf{0}. \tag{11.38}$$

The analysis in Chapter 10 indicates that the OLS estimation will not be consistent. To estimate the regression coefficient $\boldsymbol{\beta}$, we need to employ certain *instrumental variables* $\mathbf{z}_i$ such that $\mathrm{Cov}(\mathbf{z}_i, \mathbf{x}_i) \neq \mathbf{O}$ and

$$\mathrm{Cov}(\mathbf{z}_i, \varepsilon_i) = \mathrm{E}(\mathbf{z}_i \varepsilon_i) = \mathrm{E}[\mathbf{z}_i (y_i - \mathbf{x}_i' \boldsymbol{\beta})] = \mathbf{0}. \tag{11.39}$$

Here, let's assume the dimensions m of $\mathbf{z}_i$ is greater than or equal to k, the number of explanatory variables in $\mathbf{x}_i$. The condition (11.39) can now be viewed as the (over-identified or just-identified) moment conditions we need for conducting the GMM estimation for the linear regression model (11.13).

In order to implement the GMM estimation, we need to first evaluate[7]

$$\boldsymbol{\Omega}(\boldsymbol{\theta}_\circ) \equiv \mathrm{E}(\mathbf{z}_i \varepsilon_i \varepsilon_i \mathbf{z}_i') = \mathrm{E}(\varepsilon_i^2 \mathbf{z}_i \mathbf{z}_i') = \mathrm{E}\big[\mathrm{E}(\varepsilon_i^2 \,|\, \mathbf{z}_i) \mathbf{z}_i \mathbf{z}_i'\big] = \mathrm{E}(\sigma^2 \mathbf{z}_i \mathbf{z}_i') = \sigma^2 \mathrm{E}(\mathbf{z}_i \mathbf{z}_i'). \tag{11.40}$$

The GMM estimation for $\boldsymbol{\beta}$ is then based on

$$\min_{\boldsymbol{\beta}} \left[\frac{1}{n} \sum_{i=1}^{n} \mathbf{z}_i (y_i - \mathbf{x}_i' \boldsymbol{\beta})\right]' \left(\sigma^2 \frac{1}{n} \sum_{i=1}^{n} \mathbf{z}_i \mathbf{z}_i'\right)^{-1} \left[\frac{1}{n} \sum_{i=1}^{n} \mathbf{z}_i (y_i - \mathbf{x}_i' \boldsymbol{\beta})\right]. \tag{11.41}$$

which can be written as[8]

$$\min_{\boldsymbol{\beta}} \frac{1}{n} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' \mathbf{Z} (\mathbf{Z}'\mathbf{Z})^{-1} \mathbf{Z}' (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \tag{11.42}$$

where $\mathbf{Z} = [\, \mathbf{z}_1 \, \mathbf{z}_2 \, \ldots \, \mathbf{z}_n \,]'$. It is readily seen that the solution to this minimization problem is

$$\hat{\boldsymbol{\beta}} = \big[\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}\big]^{-1} \mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}, \tag{11.43}$$

which is also referred to as the *instrumental variable* (IV) estimator of $\boldsymbol{\beta}$.[9]

---

[7] Here, we have further assumed that $\mathrm{E}(\varepsilon_i^2 \,|\, \mathbf{z}_i) = \sigma^2$; i.e., ε_i is *homoscedastic* with respect to $\mathbf{z}_i$.

[8] We drop the scalar σ^2 from the expression. Doing so will not affect the derivation of the GMM estimator of $\boldsymbol{\beta}$.

[9] In Chapter 10 we suggested a two-stage estimation for using the instrumental variables. It is readily seen that the resulting two-stage estimator is identical to (11.43).

In order to apply the asymptotic theory for the GMM estimation, we need

$$\mathbf{G}(\boldsymbol{\beta}) \equiv \mathrm{E}\left[\frac{\partial \mathbf{z}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right] = -\mathrm{E}(\mathbf{z}_i\mathbf{x}_i'). \tag{11.44}$$

The general asymptotic theory for the GMM estimation implies that the IV estimator $\hat{\boldsymbol{\beta}}$ is consistent and

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{\mathrm{d}} \mathcal{N}\left(\mathbf{0},\ \sigma^2\left\{\mathrm{E}(\mathbf{x}_i\mathbf{z}_i')\left[\mathrm{E}(\mathbf{z}_i\mathbf{z}_i')\right]^{-1}\mathrm{E}(\mathbf{z}_i\mathbf{x}_i')\right\}^{-1}\right). \tag{11.45}$$

Note that the asymptotic variance-covariance matrix for the IV estimator $\boldsymbol{\beta}$ can be estimated by

$$s^2\left[\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}\right]^{-1},$$

where s^2 is some consistent estimator of σ^2.

11.7 The Restricted GMM Estimation

Suppose other than the over-identified moment condition, we have another set of conditions that we believe the true parameter $\boldsymbol{\theta}_\circ$ should satisfy. Let's also assume such extraneous conditions can be expressed as an J-vector of functions of the parameter $\boldsymbol{\theta}$:

$$\mathbf{h}(\boldsymbol{\theta}_\circ) = \mathbf{0}. \tag{11.46}$$

We note these conditions do not involve the random variable x_i so that they are fundamentally different from the moment condition. If these conditions are true, then we certainly want the GMM estimator to satisfy them. The way to impose these conditions to the GMM estimator is to consider the restricted minimization with the condition $\mathbf{h}(\boldsymbol{\theta}_\circ) = \mathbf{0}$ imposed as a set of restrictions:

$$\min_{\boldsymbol{\theta}} \left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{g}(x_i;\boldsymbol{\theta})\right]' \widetilde{\boldsymbol{\Omega}}^{-1} \left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{g}(x_i;\boldsymbol{\theta})\right] \quad \text{subject to } \mathbf{h}(\boldsymbol{\theta}) = \mathbf{0}, \tag{11.47}$$

or

$$\min_{\boldsymbol{\theta}} \left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{g}(x_i;\boldsymbol{\theta})\right]' \widetilde{\boldsymbol{\Omega}}^{-1} \left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{g}(x_i;\boldsymbol{\theta})\right] + \mathbf{h}(\boldsymbol{\theta})'\boldsymbol{\lambda}, \tag{11.48}$$

where $\boldsymbol{\lambda}$ is an J-vector of Lagrange multipliers. The solution to such a problem, denoted as $\hat{\boldsymbol{\theta}}^*$, is called the restricted GMM estimator as opposed to the unrestricted GMM estimator $\hat{\boldsymbol{\theta}}$.

When we derive the GMM estimator based on the over-identified moment condition $\mathrm{E}\left[\mathbf{g}(x;\boldsymbol{\theta}_\circ)\right] = \mathbf{0}$, the moment condition is never exactly satisfied by either the restricted or the unrestricted GMM estimator. But it should be pointed out the restriction $\mathbf{h}(\boldsymbol{\theta}_\circ) = \mathbf{0}$, in contrast, is exactly satisfied by the restricted GMM estimator. So the moment condition and restriction are not treated symmetrically although both are conditions on the parameter value.

It can be proved that, just like the unrestricted GMM estimator, the restricted GMM estimator is consistent and has an asymptotic normal distribution. However, while both estimators are consistent, their asymptotic normal distributions are not the same. In particular, the asymptotic variance-covariance matrix of the restricted GMM estimator is always smaller than or equal to that of the unrestricted GMM estimator. This result simply reflects the fact that the restricted GMM estimator, by incorporating more information from the restriction $\mathbf{h}(\boldsymbol{\theta}_\circ) = \mathbf{0}$, is more (asymptotically) efficient. The present discussion is very similar to the one we had on the the relationship between the unrestricted MLE and the restricted MLE. As a matter of fact, the derivation of the asymptotic distribution of the restricted GMM estimator is parallel to that of the restricted MLE.

11.7.1 Comparing Restricted and Unrestricted GMM Estimators

To simplify our exposition here, let's denote (half of) the objective function for minimization in defining the GMM estimator with over-identified moment condition by

$$q(\boldsymbol{\theta}) \equiv \frac{1}{2}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{g}(x_i;\boldsymbol{\theta})\right]'\widetilde{\boldsymbol{\Omega}}^{-1}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{g}(x_i;\boldsymbol{\theta})\right]. \tag{11.49}$$

and the first order derivative of $q(\boldsymbol{\theta})$ by

$$\mathbf{s}(\boldsymbol{\theta}) \equiv \left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\mathbf{g}(x_i;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right]'\widetilde{\boldsymbol{\Omega}}^{-1}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{g}(x_i;\boldsymbol{\theta})\right]. \tag{11.50}$$

Because of the difference between the restricted GMM estimator $\hat{\boldsymbol{\theta}}^*$ and the unrestricted GMM estimator $\hat{\boldsymbol{\theta}}$, we observe the following inequalities

$$\mathbf{h}(\hat{\boldsymbol{\theta}}^*) = \mathbf{0} \neq \mathbf{h}(\hat{\boldsymbol{\theta}}), \qquad q(\hat{\boldsymbol{\theta}}^*) \geq q(\hat{\boldsymbol{\theta}}), \qquad \mathbf{s}(\hat{\boldsymbol{\theta}}^*) \neq \mathbf{0} = \mathbf{s}(\hat{\boldsymbol{\theta}}). \tag{11.51}$$

The second inequality is due to the fact that the restriction $\mathbf{h}(\boldsymbol{\theta}) = \mathbf{0}$ restricts the possible values of $\boldsymbol{\theta}$ for minimization. The third inequality results from the fact that the first order condition for the restricted minimization is

$$\mathbf{s}(\hat{\boldsymbol{\theta}}^*) + \mathbf{H}(\hat{\boldsymbol{\theta}}^*)'\hat{\boldsymbol{\lambda}} = \mathbf{0}, \qquad \text{where} \quad \mathbf{H}(\boldsymbol{\theta}) = \frac{\partial\mathbf{h}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}. \tag{11.52}$$

while the first order condition for the unrestricted minimization is

$$\mathbf{s}(\hat{\boldsymbol{\theta}}) = \mathbf{0}. \tag{11.53}$$

These three sets of inequalities in $\mathbf{h}$, q, and $\mathbf{s}$ hold for any random sample of a finite sample size.

Let's first denote the probability limit of the restricted GMM estimator $\hat{\boldsymbol{\theta}}^*$ by $\boldsymbol{\theta}^*$, then, like $\hat{\boldsymbol{\theta}}^*$ for every sample size, $\boldsymbol{\theta}^*$ must also satisfy the restriction: $\mathbf{h}(\boldsymbol{\theta}^*) = \mathbf{0}$. It is obvious that whether $\boldsymbol{\theta}^*$ is equal to $\boldsymbol{\theta}_\circ$, so that the restricted GMM estimator is consistent, depends on whether $\mathbf{h}(\boldsymbol{\theta}_\circ) = \mathbf{0}$ is correct or not. The theory for the restricted GMM estimation mentioned in the previous subsection is based on the implicit assumption that the restriction $\mathbf{h}(\boldsymbol{\theta}_\circ) = \mathbf{0}$ is correctly specified. We should also note that the unrestricted GMM estimator is always consistent irrespective of whether the restriction $\mathbf{h}(\boldsymbol{\theta}_\circ) = \mathbf{0}$ is correct or not.

We can now conclude that if the restriction $\mathbf{h}(\boldsymbol{\theta}_\circ) = \mathbf{0}$ is correct, then

$$\mathbf{h}(\boldsymbol{\theta}^*) = \mathbf{0} = \mathbf{h}(\boldsymbol{\theta}_\circ), \qquad q(\boldsymbol{\theta}^*) = q(\boldsymbol{\theta}_\circ), \qquad \mathbf{s}(\boldsymbol{\theta}^*) = \mathbf{0} = \mathbf{s}(\boldsymbol{\theta}_\circ). \tag{11.54}$$

But if the restriction $\mathbf{h}(\boldsymbol{\theta}_\circ) = \mathbf{0}$ is incorrect, then

$$\mathbf{h}(\boldsymbol{\theta}^*) = \mathbf{0} \neq \mathbf{h}(\boldsymbol{\theta}_\circ), \qquad q(\boldsymbol{\theta}^*) > q(\boldsymbol{\theta}_\circ), \qquad \mathbf{s}(\boldsymbol{\theta}^*) \neq \mathbf{0} = \mathbf{s}(\boldsymbol{\theta}_\circ). \tag{11.55}$$

The direct implication of the above inequalities is that, depending on whether $\mathbf{h}(\boldsymbol{\theta}_\circ)$ is equal to $\mathbf{0}$, or whether $q(\boldsymbol{\theta}^*)$ is greater than $q(\boldsymbol{\theta}_\circ)$, or whether $\mathbf{s}(\boldsymbol{\theta}^*)$ is equal to $\mathbf{0}$, we can judge whether the restriction $\mathbf{h}(\boldsymbol{\theta}_\circ) = \mathbf{0}$ is correct or not. Therefore, even though for a finite sample size we have $\mathbf{h}(\hat{\boldsymbol{\theta}}) \neq \mathbf{0}$, $q(\hat{\boldsymbol{\theta}}^*) \geq q(\hat{\boldsymbol{\theta}})$, and $\mathbf{s}(\hat{\boldsymbol{\theta}}^*) \neq \mathbf{0}$, the differences are expected to become small as the sample size becomes large if, and only if, the restriction $\mathbf{h}(\boldsymbol{\theta}_\circ) = \mathbf{0}$ is correct. This conclusion is important because it helps us formulate three formal tests for the hypothesis about the truthfulness of the restriction $\mathbf{h}(\boldsymbol{\theta}_\circ) = \mathbf{0}$, as will be explained next.

11.8 Hypothesis Testing

Given the GMM estimator $\hat{\boldsymbol{\theta}}$ that is based on over-identified moment condition, there are three asymptotically equivalent tests for testing

$$H_0\colon\ \mathbf{h}(\boldsymbol{\theta}_\circ) = \mathbf{0} \qquad \text{against} \qquad H_1\colon\ \mathbf{h}(\boldsymbol{\theta}_\circ) \neq \mathbf{0},$$

where $\mathbf{h}$ is a J-vector of functions of the parameter $\boldsymbol{\theta}$. To explain the motivation of the tests, we need to think the null hypothesis $\mathbf{h}(\boldsymbol{\theta}_\circ) = \mathbf{0}$ as a set of restrictions on the true parameter value $\boldsymbol{\theta}_\circ$.

11.8.1 Wald Test:

Wald test is based on the idea of using the difference between $\mathbf{h}(\hat{\boldsymbol{\theta}})$ and $\mathbf{0}$ to decide whether the null hypothesis is true. To determine whether $\mathbf{h}(\hat{\boldsymbol{\theta}})$ is significantly close to 0 or not, we need the following result which can be proved easily:

$$\mathbf{h}(\hat{\boldsymbol{\theta}}) \ \overset{\text{A}}{\sim}\ \mathcal{N}\Big(\mathbf{h}(\boldsymbol{\theta}_\circ),\ \frac{1}{n}\,\mathbf{H}(\boldsymbol{\theta}_\circ)\big[\mathbf{G}(\boldsymbol{\theta}_\circ)'\boldsymbol{\Omega}(\boldsymbol{\theta}_\circ)^{-1}\mathbf{G}(\boldsymbol{\theta}_\circ)\big]^{-1}\mathbf{H}(\boldsymbol{\theta}_\circ)'\Big), \qquad \text{where }\ \mathbf{H}(\boldsymbol{\theta}) = \frac{\partial \mathbf{h}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}. \tag{11.56}$$

When the null hypothesis is true so that $\mathbf{h}(\boldsymbol{\theta}_\circ) = \mathbf{0}$, then we have the following distribution result for the quadratic form W:

$$W \equiv n\,\mathbf{h}'(\hat{\boldsymbol{\theta}}) \left\{\mathbf{H}(\hat{\boldsymbol{\theta}})\big[\mathbf{G}(\hat{\boldsymbol{\theta}})'\,\boldsymbol{\Omega}(\hat{\boldsymbol{\theta}})^{-1}\mathbf{G}(\hat{\boldsymbol{\theta}})\big]^{-1}\mathbf{H}(\hat{\boldsymbol{\theta}})'\right\}^{-1} \mathbf{h}(\hat{\boldsymbol{\theta}}) \ \overset{\text{A}}{\sim}\ \chi^2(J), \tag{11.57}$$

where J is the number of restrictions or the number of rows in the vector h. This result forms the basis for the Wald test. Given the size of the test α and the corresponding critical value c_α from the $\chi^2(J)$ distribution, the null hypothesis is rejected if $\mathbf{h}(\hat{\boldsymbol{\theta}})$ is significantly different from 0 or, equivalently, the value of W is greater than the critical value c_α.

11.8.2 The Minimum χ^2 Test:

The minimum χ^2 test is based on the idea of using the difference between $q(\hat{\boldsymbol{\theta}}^*)$ and $q(\hat{\boldsymbol{\theta}})$ to decide whether the null hypothesis is true. Specifically, we have the following asymptotic result: if the null hypothesis $\mathbf{h}(\boldsymbol{\theta}_\circ) = \mathbf{0}$ is true, then

$$MC \equiv 2n\left[q(\hat{\boldsymbol{\theta}}^*) - q(\hat{\boldsymbol{\theta}})\right] \overset{A}{\sim} \chi^2(J). \tag{11.58}$$

Hence, the null hypothesis is rejected if the value of MC is greater than the critical value c_α.

11.8.3 The Lagrange Multiplier Test:

Lagrange multiplier test is based on the idea of using the difference between $\mathbf{s}(\hat{\boldsymbol{\theta}}^*)$ and $\mathbf{0}$ to decide whether the null hypothesis is true. It can be shown that the quadratic form

$$LM \equiv n\,\mathbf{s}(\hat{\boldsymbol{\theta}}^*)'\left[\mathbf{G}(\hat{\boldsymbol{\theta}}^*)'\boldsymbol{\Omega}(\hat{\boldsymbol{\theta}}^*)^{-1}\mathbf{G}(\hat{\boldsymbol{\theta}}^*)\right]^{-1}\mathbf{s}(\hat{\boldsymbol{\theta}}^*) \overset{A}{\sim} \chi^2(J), \tag{11.59}$$

if the null hypothesis is true. Hence, we reject the null hypothesis if the value of LM is greater than the critical value c_α.[10]

The three test statistics W, MC, and LM are asymptotically equivalent and have the same asymptotic distribution $\chi^2(J)$ when the null hypothesis is true. But in finite sample applications, these three tests may give conflicting results and there is no consensus about how to resolve such conflicts when they occur.

Finally, since the three tests are asymptotically equivalent, there is no need to compute all three test statistics all the time. We note the Wald test statistic W only requires the unrestricted GMM estimator $\hat{\boldsymbol{\theta}}$, the Lagrange multiplier test statistic LM only requires the restricted GMM estimator $\hat{\boldsymbol{\theta}}^*$, while the minimum χ^2 test statistic MC requires both restricted and unrestricted GMM estimators.

---

[10]The reason for the name *Lagrange-Multiplier test* is because the first-order condition for the restricted GMM estimator implies $\mathbf{s}(\hat{\boldsymbol{\theta}}^*) = -\mathbf{H}(\hat{\boldsymbol{\theta}}^*)'\hat{\boldsymbol{\lambda}}$ so that $LM \equiv n\,\hat{\boldsymbol{\lambda}}'\mathbf{H}(\hat{\boldsymbol{\theta}}^*)\left[\mathbf{G}(\hat{\boldsymbol{\theta}}^*)'\boldsymbol{\Omega}(\hat{\boldsymbol{\theta}}^*)^{-1}\mathbf{G}(\hat{\boldsymbol{\theta}}^*)\right]^{-1}\mathbf{H}(\hat{\boldsymbol{\theta}}^*)'\hat{\boldsymbol{\lambda}}$, which is a test statistic based on the Lagrange multiplier $\boldsymbol{\lambda}$.

11.9 The GMM Interpretation of the Restricted OLS Estimation

As mentioned earlier in Subsection 11.3, the OLS estimation of the multiple linear regression model is based on the just-identified population moment condition (11.14) and its sample counterpart (11.15). The immediate consequence of imposing the linear restriction

$$\mathbf{R}\boldsymbol{\beta} = \mathbf{q} \tag{11.60}$$

to the GMM estimation is that the restricted GMM estimator cannot be exactly solved from the just-identified moment condition alone since the linear restriction also needs to be satisfied. The restricted GMM estimation has changed from the just-identified case to an over-identified case.

To construct the objective function for deriving the GMM estimator in the over-identified case requires the sample counterpart of $\boldsymbol{\Omega}(\boldsymbol{\beta})$ in (11.17) which is

$$s^2 \frac{1}{n}\mathbf{X}'\mathbf{X}, \tag{11.61}$$

where s^2 is any consistent estimator of σ^2. The objective function for the over-identified GMM estimation is a quadratic function in the sample moments with the inverse of the above term as the weighting matrix:

$$q(\boldsymbol{\beta}) = \frac{1}{2ns^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \tag{11.62}$$

Recall that the objective function for the OLS estimation is $S(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$. It is easy to show

$$2ns^2 q(\boldsymbol{\beta}) = S(\boldsymbol{\beta}) - \mathbf{y}'\mathbf{M}\mathbf{y}, \tag{11.63}$$

where $\mathbf{M} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Because of the equivalence between $q(\boldsymbol{\beta})$ and $S(\boldsymbol{\beta})$ (both $2ns^2$ and $\mathbf{y}'\mathbf{M}\mathbf{y}$ do not involve $\boldsymbol{\beta}$), we conclude that the OLS estimator and the GMM estimator are identical. We should understand that in the present framework the OLS estimator is derived as the GMM estimator from the over-identified moment conditions. The approach is different from the one in Subsection 11.3 where the OLS estimator is derived as the GMM estimator from the just-identified moment conditions. It is interesting to note that if we plug the OLS estimator $\mathbf{b}$ into the quadratic objective function (11.62), we get $q(\mathbf{b}) = (\mathbf{y}'\mathbf{M}\mathbf{y} - \mathbf{y}'\mathbf{M}\mathbf{y})/2ns^2 = 0$, which is the smallest possible value of that quadratic function. This special result reflects that the objective function (11.62) actually is built from just-identified, instead of over-identified, moment conditions.

We now turn to the restricted GMM estimator subject to the linear restriction (11.60) which is to be solved from

$$\min_{\boldsymbol{\beta}} q(\boldsymbol{\beta}) \qquad \text{s.t. } \mathbf{R}\boldsymbol{\beta} = \mathbf{q}. \tag{11.64}$$

Because $q(\boldsymbol{\beta})$ and $S(\boldsymbol{\beta})$ are equivalent, the first-order condition for the restricted GMM estimation is also equivalent to the one for the restricted OLS estimation so that, similar to the case of the unrestricted estimation, the restricted GMM estimator is the same as the restricted OLS estimator $\mathbf{b}^*$.

Three Asymptotic Tests Given that the unrestricted and restricted OLS estimators both have the GMM interpretations, they can be used to construct the three asymptotically equivalent tests for testing

$$H_0\colon\ \mathbf{R}\boldsymbol{\beta} = \mathbf{q} \qquad \text{against} \qquad H_1\colon\ \mathbf{R}\boldsymbol{\beta} \neq \mathbf{q}.$$

1. Wald Test: based on the asymptotic result

$$\mathbf{Rb} \ \overset{\text{A}}{\sim}\ \mathcal{N}(\mathbf{q},\ \ \sigma^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'), \tag{11.65}$$

   we can immediately get the Wald test statistic which is

$$W = \frac{(\mathbf{Rb}-\mathbf{q})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb}-\mathbf{q})}{s^2} \ \overset{\text{A}}{\sim}\ \chi^2(m), \tag{11.66}$$

   under the null hypothesis H_0, where s^2 is any consistent estimator of σ^2.

2. The Minimum χ^2 Test: given the objective function (11.62) for the GMM estimation, the minimum χ^2 test statistic is

$$MC = 2n\big[q(\mathbf{b}^*) - q(\mathbf{b})\big] = 2n\,q(\mathbf{b}^*) = \frac{(\mathbf{Rb}-\mathbf{q})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb}-\mathbf{q})}{s^2} \ \overset{\text{A}}{\sim}\ \chi^2(m), \tag{11.67}$$

   under the null hypothesis H_0, where s^2 is any consistent estimator of σ^2. Note that $q(\mathbf{b})$ is identically equal to 0.

3. The Lagrange Multiplier Test: given the score function

$$\mathbf{s}(\boldsymbol{\beta}) = \frac{1}{ns^2}\mathbf{X}'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}), \tag{11.68}$$

   the Lagrange multiplier test statistic is[11]

$$\begin{aligned} LM &= \frac{(\mathbf{y}-\mathbf{Xb}^*)'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{y}-\mathbf{Xb}^*)}{s^2} \\ &= \frac{(\mathbf{Rb}-\mathbf{q})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb}-\mathbf{q})}{s^2} \ \overset{\text{A}}{\sim}\ \chi^2(m), \end{aligned} \tag{11.69}$$

   under the null hypothesis H_0, where s^2 is any consistent estimator of σ^2.

It is interesting to see that these three asymptotic tests are identically equal and

$$W = MC = LM = m\cdot F, \tag{11.70}$$

where F is the F test statistic discussed in Chapter 6 and m is the number of restrictions. It should also be pointed out that, in contrast to the F test, the three asymptotic tests do not hinge on the normality assumption and they are valid only when the sample size is sufficiently large.

---

[11]The Lagrange multiplier test statistic can also be derived from the fact that the asymptotic distribution of the Lagrange multiplier estimator $\mathbf{c}$ which is $\mathbf{c} = [\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb}-\mathbf{q}) \ \overset{\text{A}}{\sim}\ \mathcal{N}(\mathbf{0},\ \ \sigma^2[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1})$, under the null hypothesis H_0. See (6.120) in Chapter 6.