Generalized Method of Moments

Topics in Advanced Econometrics (ResEcon 703)

Matt Woerman
Resource Economics, UMass Amherst

1 Overview of method of moments

Generalized method of moments (GMM) and its special case, method of moments (MM), are two common estimation methods in modern empirical economics. These methods are often used to estimate structural econometric models and also have frequent applications in the fields of macroeconomics and finance. A primary reason that these methods are commonly used is that they are semi-parametric, making them more flexible than fully parametric methods, such as maximum likelihood estimation. In particular, MM and GMM require assumptions about only moments of the data, rather than the full distribution of the data. Additionally, many other estimation methods—ordinary least squares, two-stage least squares, generalized least squares, maximum likelihood estimation, etc.—generate moment conditions that can be used with MM and GMM, meaning that GMM is a highly general estimation method that encompasses most other estimators typically found in applied microeconomics.

1.1 Moment conditions

The central assumption of (generalized) method of moments estimation is that we know moment conditions about the population from which our data are drawn. These moment conditions are functions of parameters and data that equal zero in expectation when evaluated at the true parameter values. Simple examples of moments conditions are generated by the mean and variance of an i.i.d. random variable, as well as the covariance of jointly distributed random variables. These moment conditions are shown in the following equations and described in more detail below.

$$\begin{aligned}
\text{Mean:} \quad & \mu_y = E[y] && \Rightarrow \; E[y - \mu_y] = 0 \\
\text{Variance:} \quad & \sigma_y^2 = E[(y - \mu_y)^2] && \Rightarrow \; E[(y - \mu_y)^2 - \sigma_y^2] = 0 \\
\text{Covariance:} \quad & \sigma_{xy} = E[(y - \mu_y)(x - \mu_x)] && \Rightarrow \; E[(y - \mu_y)(x - \mu_x) - \sigma_{xy}] = 0
\end{aligned}$$

Mean Let μ_y be the mean of i.i.d. random variable y. Then, by definition, μ_y is the expectation of y, or $\mu_y = E[y]$. Rearranging this expression gives us a moment condition, $E[y - \mu_y] = 0$. This expression is a moment condition because it is a function of parameters and data, $y - \mu_y$, that equals zero in expectation.

Variance Let σ_y^2 be the variance of i.i.d. random variable y. Then, by definition, σ_y^2 is the expectation of $(y - \mu_y)^2$, or $\sigma_y^2 = E[(y - \mu_y)^2]$. Rearranging this expression gives us a moment condition, $E[(y - \mu_y)^2 - \sigma_y^2] = 0$. This expression is a moment condition because it is a function of parameters and data, $(y - \mu_y)^2 - \sigma_y^2$, that equals zero in expectation.

Covariance Let σ_{xy} be the covariance of jointly distributed random variables x and y. Then, by definition, σ_{xy} is the expectation of $(y-\mu_y)(x-\mu_x)$, or $\sigma_{xy} = E[(y-\mu_y)(x-\mu_x)]$, where μ_x is the mean of x. Rearranging this expression gives us a moment condition, $E[(y-\mu_y)(x-\mu_x) - \sigma_{xy}] = 0$. This expression is a moment condition because it is a function of parameters and data, $(y-\mu_y)(x-\mu_x)-\sigma_{xy}$, that equals zero in expectation.

These example moment conditions are all generated by definitions of statistical objects, but we can generate moment conditions in many other ways. A common example is the first-order conditions of an economic model, which are functions of parameters and data that equal zero, meaning they can be used as moment conditions. Moment conditions can also come from econometric assumptions, such as the assumption that instruments must be orthogonal to model errors. A final example is a model fit criterion, such as ensuring that the predicted market shares of the model equal realized market shares in the data, which can be expressed as a moment condition. These are some common examples of moment conditions in structural econometric models, but it is far from an exhaustive list.

1.2 Intuition of method of moments

The basic intuition of method of moments estimation is best illustrated through an example. Suppose we have five random draws, $\boldsymbol{y} = \{5, 10, 9, 14, 7\}$, but we do not know the distribution from which these data were generated. Because we do not know the distribution, we cannot make the kind of distributional assumption required by maximum likelihood estimation.

Suppose we are interested in finding the mean of this unknown distribution, which we denote μ. By definition, this mean gives us the expectation of random draw y_i, $E[y_i] = \mu$. We can rearrange this expression to yield a moment condition:

$$E[y_i - \mu] = 0$$

If this moment condition holds in expectation for the population, then we should expect an analogous condition to hold on average in a sample drawn from that population

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \mu) = 0$$

We can estimate the population parameter by finding the parameter value that solves this empirical expression for our sample of data.

This simple example illustrates the basic intuition of method of moments estimation and the method of moments estimator—we know that certain moment conditions hold in expectation for the population, so we find the set of parameters that makes analogous moment conditions hold on average for our sample of data.

2 Method of moments estimator

Method of moments estimation requires moment conditions about the population from which our data are drawn. We assume that our data, $\{y, \boldsymbol{x}, \boldsymbol{z}\}$, are drawn from a population with K moment conditions that are functions of K parameters, $\boldsymbol{\theta}$, and these moment conditions are given by

$$E[\boldsymbol{m}(y, \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\theta})] = \mathbf{0}$$

where $\boldsymbol{m}(\cdot)$ is a vector of K functions, $\boldsymbol{\theta}$ is a vector of K parameters, y is a dependent variable, $\boldsymbol{x}$ is a vector of independent variables, and $\boldsymbol{z}$ is a vector of exogenous instruments.[1] Note that we are making assumptions about the moments of this population but not its full density.

We seek to find the set of parameters, $\boldsymbol{\theta}$, that define these moment conditions and describe this population. We do not observe the full population, however, only a sample of data, so we cannot directly calculate the parameters. Instead, we replace the population expectation with its empirical analog—the sample mean—to generate the corresponding K sample moment conditions as a function of the K parameters:

$$\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{m}(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i, \boldsymbol{\theta}) = \mathbf{0}$$

The method of moments estimator, which we denote $\widehat{\boldsymbol{\theta}}$, is the set of parameters that solves the system of equations given by these sample moment conditions:

$$\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{m}(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i, \widehat{\boldsymbol{\theta}}) = \mathbf{0}$$

In some cases, this system of equations will yield closed form expressions for the MM estimator. For more complex estimation, we can reformulate the MM estimator as the solution to a minimization problem and use numerical optimization. In this case, we define the MM estimator as

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\, Q(\boldsymbol{\theta})$$

where the objective function, $Q(\boldsymbol{\theta})$, is given by

$$Q(\boldsymbol{\theta}) = \left[\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{m}(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i, \boldsymbol{\theta})\right]' \left[\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{m}(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i, \boldsymbol{\theta})\right]$$

3 Examples of method of moments estimation

3.1 Population mean

Suppose we have five random draws from a population:

$$\boldsymbol{y} = \{5, 10, 9, 14, 7\}$$

We do not know the distribution of this population but want to find its mean, which we denote μ. By definition, if y is i.i.d., then μ is equal to the expectation of y, or $\mu = E[y]$. Rearranging this expression gives a population moment condition:

$$E[y - \mu] = 0$$

We replace this population expectation with its empirical analog—the sample mean—yielding a sample moment condition. The MM estimator, $\widehat{\mu}$, is the value that solves this sample moment condition for our data:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{\mu}) = 0$$

1. Independent variables that are exogenous may appear in both $\boldsymbol{x}$ and $\boldsymbol{z}$.

We rearrange this sample moment condition to get an expression for the MM estimator, $\widehat{\mu}$:

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

This expression defines the MM estimator, $\widehat{\mu}$, as a function of our data. In fact, the MM estimator for a population mean is simply the sample mean of the data. Plugging in our data yields the MM estimator:

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i = 9$$

In other words, we assume that the moment condition holds in expectation for the population, and a population mean of $\widehat{\mu} = 9$ makes the analogous moment condition hold on average for our sample of the data.

3.2 Ordinary least squares regression

In the previous example, we estimated the parameter that defined the moment condition for a single random variable. In most econometric applications, however, we model some outcome data, y, as a function of both parameters, $\boldsymbol{\theta}$, and other data, $\boldsymbol{x}$. An example is the general ordinary least squares (OLS) regression model:

$$y_i = \boldsymbol{\beta}'\boldsymbol{x}_i + \varepsilon_i$$

where $\boldsymbol{\beta}$ is a vector of K parameters and $\boldsymbol{x}_i$ is a vector of K variables.[2] Note that OLS is a special case of MM estimation, so the MM estimator will be equivalent to the traditional OLS estimator, but we will find those parameter values in a different way.

In the OLS regression model, we assume the error term, ε_i, is orthogonal to the data, $\boldsymbol{x}_i$:

$$E[\boldsymbol{x}_i\varepsilon_i] = 0$$

Note that the error term, ε_i, can be expressed as the difference between the dependent variable and the fitted model value, $y_i - \boldsymbol{\beta}'\boldsymbol{x}_i$. Replacing ε_i with $y_i - \boldsymbol{\beta}'\boldsymbol{x}_i$ gives K population moment conditions:

$$E[\boldsymbol{x}_i(y_i - \boldsymbol{\beta}'\boldsymbol{x}_i)] = \mathbf{0}$$

We replace the population expectation with its empirical analog—the sample mean—yielding K sample moment conditions. The MM estimator, $\widehat{\beta}$, is the set of parameter values that solves these sample moment conditions for our data:

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

We rearrange these sample moment conditions to get an expression for the MM estimator, $\widehat{\beta}$, as a function of our data:

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\left(\sum_{i=1}^{n} \boldsymbol{x}_i y_i\right)$$

This MM estimator for an OLS regression is equivalent to the traditional OLS estimator. Thus, we can interpret OLS parameters as not only minimizing the sum of squared error but also achieving orthogonality between the error terms and data.

2. One of these K variables could be a constant or intercept term.

3.3 Maximum likelihood estimation

One advantage of (generalized) method of moments estimation over maximum likelihood estimation is that we are not required to make an assumption about the full distribution of the population from which our data are drawn, only the moments of that population. If we do make a distributional assumption, however, then we can use MM estimation to find a MM estimator that is equivalent to the maximum likelihood estimator.

We assume that a random variable has conditional density $f(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta})$. Then the log-likelihood function of $\boldsymbol{\theta}$ conditional on $\boldsymbol{y}$ and $\boldsymbol{X}$ is

$$\ln L(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{X}) = \sum_{i=1}^{n} \ln f(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta})$$

When we find the maximum likleihood estimator, we maximize this log-likelihood function, which yields K first-order conditions, one for each of the K parameters:

$$\mathbf{0} = \frac{\partial \ln L(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{X})}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{n} \frac{\partial \ln f(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

If we divide this expression by n, we get what look like sample moment conditions:

$$\frac{1}{n} \sum_{i=1}^{n} \frac{\partial \ln f(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{0}$$

These K sample moment conditions can be viewed as the empirical analogs of K population moment conditions:

$$E\left[\frac{\partial \ln f(y \mid \boldsymbol{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right] = \mathbf{0}$$

If we used MM estimation with these population moment conditions, the MM estimator would satisfy the first-order conditions of the maximum likelihood estimator, so these two estimators would be equivalent. Thus, maximum likelihood estimation can be motivated as a special case of MM estimation with the population moment conditions given above.

4 Generalized method of moments estimator

The method of moments estimator requires that we have the same number of moment conditions and parameters to be estimated. In many empirical settings, however, we have more moment conditions than parameters to be estimated.[3] When this occurs, we cannot use MM estimation and must instead use its more general form, generalized method of moments.

Generalized method of moments estimation requires moment conditions about the population from which our data are drawn. We assume that our data, $\{y, \boldsymbol{x}, \boldsymbol{z}\}$, are drawn from a population with L

3. Additional moment conditions can come from many sources. A common example is estimating a model with more exogenous instruments than independent variables. Another example is that one economic model parameter might appear in multiple first-order conditions, which could generate more moments than parameters. These are only two of many possible sources of additional moment conditions.

moment conditions that are functions of K parameters, $\boldsymbol{\theta}$, with $L \geq K$.[4] These moment conditions are given by

$$E[\boldsymbol{m}(y, \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\theta})] = \mathbf{0}$$

where $\boldsymbol{m}(\cdot)$ is a vector of L functions, $\boldsymbol{\theta}$ is a vector of K parameters, y is a dependent variable, $\boldsymbol{x}$ is a vector of independent variables, and $\boldsymbol{z}$ is a vector of exogenous instruments.[5] Note that we are making assumptions about the moments of this population but not its full density.

We seek to find the set of parameters, $\boldsymbol{\theta}$, that define these moment conditions and describe this population. We do not observe the full population, however, only a sample of data, so we cannot directly calculate the parameters. Instead, we replace the population expectation with its empirical analog—the sample mean—to generate the corresponding L sample moment conditions as a function of the K parameters:

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{m}(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i, \boldsymbol{\theta}) = \mathbf{0}$$

We cannot solve for L unique sample moment conditions with only K parameters when $L > K$; we describe this model as being overidentified. Instead, we will minimize the weighted sum of the sample moments, effectively getting as close as possible to solving all L sample moment conditions.

The generalized method of moments estimator, which we denote $\widehat{\boldsymbol{\theta}}$, is the set of parameters that minimizes the weighted sum of the sample moments:

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\, Q(\boldsymbol{\theta})$$

where the objective function, $Q(\boldsymbol{\theta})$, is given by

$$Q(\boldsymbol{\theta}) = \left[\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{m}(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i, \boldsymbol{\theta})\right]' \boldsymbol{W} \left[\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{m}(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i, \boldsymbol{\theta})\right]$$

and $\boldsymbol{W}$ is a $L \times L$ positive definite weighting matrix. We will discuss the choice of weighting matrix after looking at an example and describing the properties of the GMM estimator.

5 Example of generalized method of moments estimation

A common example of an econometric model that could be overidentified is an instrumental variables linear regression model:

$$y_i = \boldsymbol{\beta}'\boldsymbol{x}_i + \varepsilon_i$$

where $\boldsymbol{\beta}$ is a vector of K parameters and $\boldsymbol{x}_i$ is a vector of K variables,[6] some of which are endogenous. Because of the endogeneity of $\boldsymbol{x}_i$, estimating this model by OLS would yield biased and inconsistent estimates of the true causal parameters, $\boldsymbol{\beta}$. Instead, we instrument for the endogenous variables with a set of L exogenous instruments, which we denote $\boldsymbol{z}_i$.[7] We must have at least as many exogenous

4. If we have fewer moments than parameters—that is, $L < K$—then the parameters of the model are not identified. If we have the same number of moments and parameters—that is, $L = K$—then we could use either MM estimation or GMM estimation to obtain the same estimator.

5. Independent variables that are exogenous may appear in both $\boldsymbol{x}$ and $\boldsymbol{z}$.

6. One of these K variables could be a constant or intercept term.

7. Independent variables that are exogenous may appear in both $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$.

instruments as independent variables, so $L \geq K$.

Valid instruments must be orthogonal to the error term from the above regression model, ε_i:

$$E[\boldsymbol{z}_i \varepsilon_i] = 0$$

Note that the error term, ε_i, can be expressed as the difference between the dependent variable and the fitted model value, $y_i - \boldsymbol{\beta}' \boldsymbol{x}_i$. Replacing ε_i with $y_i - \boldsymbol{\beta}' \boldsymbol{x}_i$ gives the population moment conditions:

$$E[\boldsymbol{z}_i (y_i - \boldsymbol{\beta}' \boldsymbol{x}_i)] = \mathbf{0}$$

We have L instruments and, hence, L population moment conditions that are functions of K parameters, with $L \geq K$. Thus, this model could be overidentified, so we will use GMM estimation. If we replace the population expectation with its empirical analog—the sample mean—we obtain L sample moment conditions that are functions of K parameters.

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

We cannot solve a system of L unique equations with K parameters if $L > K$. Instead, we find the GMM estimator, $\widehat{\boldsymbol{\beta}}$, by minimizing the weighted sum of the sample moments:

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \, Q(\boldsymbol{\beta})$$

where the objective function, $Q(\boldsymbol{\beta})$, is given by

$$Q(\boldsymbol{\beta}) = \left[\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{z}_i (y_i - \boldsymbol{\beta}' \boldsymbol{x}_i) \right]' \boldsymbol{W} \left[\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{z}_i (y_i - \boldsymbol{\beta}' \boldsymbol{x}_i) \right]$$

Note that this objective function depends on the choice of weighting matrix, $\boldsymbol{W}$, so the GMM estimator also depends on the weighting matrix. If we use a weighting matrix of

$$\boldsymbol{W} = \sum_{i=1}^{n} \boldsymbol{z}_i \boldsymbol{z}_i'$$

then the GMM estimator, $\widehat{\boldsymbol{\beta}}$, is equivalent to the 2SLS estimator, indicating that the 2SLS parameters achieve orthogonality between the error terms and exogenous instruments. If we use a different weighting matrix, however, then we will obtain a different GMM estimator.

The dependence of the GMM estimator on the weighting matrix, $\boldsymbol{W}$, highlights the importance of the choice of weighting matrix. In order to describe the optimal weighting matrix, we first need to know the properties of the GMM estimator.

6 Properties of the generalized method of moments estimator

The GMM estimator is consistent and asymptotically normal when the assumed population moment conditions are valid and and the empirical moments and weighting matrix meet certain conditions. Note

that, unlike the maximum likelihood estimator, the GMM estimator is not efficient.[8]

The required conditions on the empirical moments and weighting matrix are shown below. In this and future sections, $\widehat{\boldsymbol{\theta}}$ denotes the GMM estimator, $\boldsymbol{\theta}_0$ denotes the true values of the parameters, $\boldsymbol{\theta}$ denotes any arbitrary set of parameters, and the 0 subscript denotes an object based on the true parameters values.

1. The empirical moments obey the law of large numbers:

$$\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{m}(y_i,\boldsymbol{x}_i,\boldsymbol{z}_i,\boldsymbol{\theta}_0)\xrightarrow{p}\mathbf{0}$$

2. The derivatives of the empirical moments converge:

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\boldsymbol{m}(y_i,\boldsymbol{x}_i,\boldsymbol{z}_i,\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\xrightarrow{p}\boldsymbol{G}_0$$

3. The empirical moments obey the central limit theorem:

$$\frac{\sqrt{n}}{n}\sum_{i=1}^{n}\boldsymbol{m}(y_i,\boldsymbol{x}_i,\boldsymbol{z}_i,\boldsymbol{\theta}_0)\xrightarrow{d}\mathcal{N}(0,\boldsymbol{S}_0)$$

   where

$$\boldsymbol{S}_0=\operatorname{plim}\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\boldsymbol{m}(y_i,\boldsymbol{x}_i,\boldsymbol{z}_i,\boldsymbol{\theta}_0)\boldsymbol{m}(y_j,\boldsymbol{x}_j,\boldsymbol{z}_j,\boldsymbol{\theta}_0)'$$

4. The parameters are identified:

$$\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{m}(y_i,\boldsymbol{x}_i,\boldsymbol{z}_i,\boldsymbol{\theta}_1)=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{m}(y_i,\boldsymbol{x}_i,\boldsymbol{z}_i,\boldsymbol{\theta}_2)\quad\Leftrightarrow\quad\boldsymbol{\theta}_1=\boldsymbol{\theta}_2$$

5. The weighting matrix converges to a finite symmetric positive definite matrix:

$$\boldsymbol{W}\xrightarrow{p}\boldsymbol{W}_0$$

These conditions ensure that the empirical moments and weighting matrix meet some basic requirements for being "well-behaved." When these conditions are met and the population moments are valid, the GMM estimator has the properties listed above, which are described in more detail below.[9]

Consistency The GMM estimator converges in probability to the true parameters values:

$$\widehat{\boldsymbol{\theta}}\xrightarrow{p}\boldsymbol{\theta}_0$$

In words, as the sample size increases to infinity, the GMM estimator becomes vanishingly close to the true parameter values.

8. The distributional assumption of maximum likelihood estimation provides additional information that leads to its efficiency.

9. For the proofs of these properties, see William H. Greene. 2018. *Econometric Analysis*. Eighth Edition. Pearson.

Asymptotic normality The GMM estimator is asymptotically normal with a mean at the true parameter values and known variance:

$$\widehat{\boldsymbol{\theta}} \overset{a}{\sim} \mathcal{N}\left(\boldsymbol{\theta}_0, \frac{1}{n}(\boldsymbol{G}_0'\boldsymbol{W}_0\boldsymbol{G}_0)^{-1}(\boldsymbol{G}_0'\boldsymbol{W}_0\boldsymbol{S}_0\boldsymbol{W}_0\boldsymbol{G}_0)(\boldsymbol{G}_0'\boldsymbol{W}_0\boldsymbol{G}_0)^{-1}\right)$$

where

$$\begin{aligned}
\boldsymbol{G}_0 &= \operatorname{plim} \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \boldsymbol{m}(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \\
\boldsymbol{W}_0 &= \operatorname{plim} \boldsymbol{W} \\
\boldsymbol{S}_0 &= \operatorname{plim} \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} \boldsymbol{m}(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i, \boldsymbol{\theta}_0)\boldsymbol{m}(y_j, \boldsymbol{x}_j, \boldsymbol{z}_j, \boldsymbol{\theta}_0)'
\end{aligned}$$

In words, the asymptotic variance of the GMM estimator depends on the probability limit of the gradients of the empirical moments evaluated at the true parameter values, the probability limit of the weighting matrix, and the probability limit of the variance-covariance matrix of the empirical moments evaluated at the true parameter values.

7 Optimal generalized method of moments estimator

From these properties of the GMM estimator, we see that the consistency of the GMM estimator does not depend on the choice of weighting matrix, $\boldsymbol{W}$, but the asymptotic variance of the GMM estimator does. Thus, every weighting matrix yields a GMM estimator that is consistent, but the choice of weighting matrix will affect the efficiency of the estimator. We would ideally like to use the weighting matrix that minimizes the variance of the GMM estimator.

The asymptotic variance of the GMM estimator is minimized when the probability limit of the weighting matrix, $\boldsymbol{W}_0$, equals the inverse of $\boldsymbol{S}_0$, which was defined above:

$$\boldsymbol{W}_0 = \boldsymbol{S}_0^{-1} = \left[\operatorname{plim} \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} \boldsymbol{m}(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i, \boldsymbol{\theta}_0)\boldsymbol{m}(y_j, \boldsymbol{x}_j, \boldsymbol{z}_j, \boldsymbol{\theta}_0)'\right]^{-1}$$

This weighting matrix causes many terms to cancel in the asymptotic variance, resulting in a simpler expression for the asymptotic variance:

$$Var(\widehat{\boldsymbol{\theta}}) = \frac{1}{n}(\boldsymbol{G}_0'\boldsymbol{S}_0^{-1}\boldsymbol{G}_0)^{-1}$$

Note that $\boldsymbol{S}_0$ is the probability limit of the variance-covariance matrix of the empirical moments evaluated at the true parameter values, so we minimize the the asymptotic variance of the GMM estimator by effectively weighting each moment inversely to its variance.

This GMM estimator, which has the smallest variance of all GMM estimators, is called the optimal GMM estimator. To estimate it, we want to use a weighting matrix, $\boldsymbol{W}$, that converges in probability to $\boldsymbol{S}_0$, the probability limit of the variance-covariance matrix of the empirical moments evaluated at the true parameter values. We do not know the true parameter values, however, so we cannot directly evaluate this matrix. Instead, we can use a two-step procedure to first estimate $\boldsymbol{S}_0$ and then use this

estimate to find the optimal GMM estimator, which is sometimes called the two-step GMM estimator.[10]

In the first step, we find a GMM estimator using any arbitrary weighting matrix—a common choice is the identity matrix. We denote this first-step GMM estimator as $\widetilde{\boldsymbol{\theta}}$. Because the GMM estimator is consistent for any weighting matrix, $\widetilde{\boldsymbol{\theta}}$ is a consistent estimate of the true parameter values, $\boldsymbol{\theta}_0$. We then use this first-step GMM estimator, $\widetilde{\boldsymbol{\theta}}$, to construct an estimator of $\boldsymbol{S}_0$. If we assume observations are independent, this estimator is

$$\widetilde{\boldsymbol{S}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{m}(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i, \widetilde{\boldsymbol{\theta}})\boldsymbol{m}(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i, \widetilde{\boldsymbol{\theta}})'$$

In the second step, we find the optimal GMM estimator using a weighting matrix based on this estimate of $\boldsymbol{S}_0$, $\boldsymbol{W} = \widetilde{\boldsymbol{S}}^{-1}$. That is, we define the optimal GMM estimator as

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\, Q(\boldsymbol{\theta})$$

where the objective function, $Q(\boldsymbol{\theta})$, is given by

$$Q(\boldsymbol{\theta}) = \left[\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{m}(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i, \boldsymbol{\theta})\right]' \widetilde{\boldsymbol{S}}^{-1} \left[\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{m}(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i, \boldsymbol{\theta})\right]$$

The GMM estimator we obtain in this second step, $\widehat{\boldsymbol{\theta}}$, also gives a consistent estimate of the true parameter values, $\boldsymbol{\theta}_0$, and it has the minimum variance among all GMM estimators. Note that the optimal GMM estimator is not efficient among all estimators—for example, when compared to the maximum likelihood estimator—but it is efficient among GMM estimators.

The optimal GMM estimator has an asymptotic variance of

$$Var(\widehat{\boldsymbol{\theta}}) = \frac{1}{n}(\boldsymbol{G}_0'\boldsymbol{S}_0^{-1}\boldsymbol{G}_0)^{-1}$$

To calculate this variance-covariance matrix, we must evaluate the probability limit of the gradients of the empirical moments and the probability limit of the variance-covariance matrix of the empirical moments evaluated at the true parameter values. We do not know the true parameter values, however, so we cannot use this expression directly. Instead, an estimator for the variance-covariance matrix is

$$\widehat{Var}(\widehat{\boldsymbol{\theta}}) = \frac{1}{n}\left(\widehat{\boldsymbol{G}}'\widehat{\boldsymbol{S}}^{-1}\widehat{\boldsymbol{G}}\right)^{-1}$$

where the components of this matrix are estimated by

$$\widehat{\boldsymbol{G}} = \frac{1}{n}\sum_{i=1}^{n} \left.\frac{\partial \boldsymbol{m}(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}$$

$$\widehat{\boldsymbol{S}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{m}(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i, \widehat{\boldsymbol{\theta}})\boldsymbol{m}(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i, \widehat{\boldsymbol{\theta}})'$$

That is, we estimate the variance-covariance matrix of the optimal GMM estimator by evaluating the gradients and variance-covariance matrix of the empirical moments at the optimal GMM estimator.[11]

10. This procedure can be viewed as a generalization of a feasible generalized least squares regression.

11. There are alternate variance estimators for the GMM estimator that are more robust, but this most basic estimator will be sufficient for this course.

8 Specification and hypothesis tests

8.1 Overidentifying restrictions test

When we have more moment conditions than parameters, the model is overidentified and we cannot ensure all sample moment conditions hold. We can, however, perform an overidentifying restrictions test to determine if the model is misspecified and, hence, not all population moment conditions hold. The null hypothesis of the overidentifying restrictions test is

$$H_0\text{: } E[\boldsymbol{m}(y, \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\theta}_0)] = \mathbf{0}$$

When $\widehat{\boldsymbol{\theta}}$ is the optimal GMM estimator, the overidentifying restrictions test statistic is equal to the GMM objective function evaluated at the optimal GMM estimator, $Q(\widehat{\boldsymbol{\theta}})$:

$$OIR = \left[\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{m}(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i, \widehat{\boldsymbol{\theta}})\right]' \widetilde{\boldsymbol{S}}^{-1} \left[\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{m}(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i, \widehat{\boldsymbol{\theta}})\right]$$

This test statistic is asymptotically distributed χ^2 with $L - K$ degrees of freedom

$$OIR \overset{a}{\sim} \chi^2(L - K)$$

If the test statistic is sufficiently large, then we reject that the population moment conditions are valid, concluding that the model is misspecified and the GMM estimator is inconsistent for the true parameter values, $\boldsymbol{\theta}_0$.

8.2 Hypothesis tests

There are three common test procedures to test hypotheses about parameters that use the results of GMM estimation: difference test, Wald test, and Lagrange multiplier test.[12] Consider the test of hypotheses:

$$H_0\text{: } \boldsymbol{h}(\boldsymbol{\theta}_0) = \mathbf{0}$$

where $\boldsymbol{h}(\boldsymbol{\theta}_0)$ is any set of J parameter restrictions. This specification of hypotheses is fully general. For example, it can represent a test that parameters equal zero:

$$\boldsymbol{h}(\boldsymbol{\theta}_0) = \begin{pmatrix}\theta_1 \\ \theta_2\end{pmatrix} = \begin{pmatrix}0 \\ 0\end{pmatrix}$$

or a test that parameters equal one another:

$$\boldsymbol{h}(\boldsymbol{\theta}_0) = \begin{pmatrix}\theta_1 - \theta_3 \\ \theta_2 - \theta_3\end{pmatrix} = \begin{pmatrix}0 \\ 0\end{pmatrix}$$

or any other hypothesis test about the parameters.

The following test statistics apply to the optimal GMM estimator of the unrestricted model and the GMM estimator of the restricted model obtained using the unrestricted optimal weighting matrix. More complex test statistics apply to other GMM estimators.

12. The difference test for GMM is analogous to the likelihood ratio test for maximum likelihood estimation. The Wald and Lagrange multiplier tests for GMM are analogous to the same-named tests for maximum likelihood estimation.

Difference test If the hypotheses are true, then the objective function value, $Q(\boldsymbol{\theta})$, should be approximately the same whether the hypothesized restrictions are imposed or not. That is, $Q(\widehat{\boldsymbol{\theta}}_R) \approx Q(\widehat{\boldsymbol{\theta}}_U)$ where $\widehat{\boldsymbol{\theta}}_R$ is the GMM estimator of the restricted model that is obtained with the hypothesized restrictions imposed and $\widehat{\boldsymbol{\theta}}_U$ is the GMM estimator of the unrestricted model that is obtained without those restrictions. The difference test draws on this intuition and tests if $Q(\widehat{\boldsymbol{\theta}}_R)$ is sufficiently close to $Q(\widehat{\boldsymbol{\theta}}_U)$. The difference test statistic is

$$D = n\left(Q(\widehat{\boldsymbol{\theta}}_R) - Q(\widehat{\boldsymbol{\theta}}_U)\right)$$

The difference test statistic has an asymptotic chi-squared distribution with degrees of freedom equal to the number of restrictions, J:

$$D \overset{a}{\sim} \chi^2(J)$$

Note that the difference test requires estimating two models—the restricted model that is obtained with the hypothesized restrictions imposed and the unrestricted model that is obtained without these restrictions—and calculating the objective function value of each model. Once these objective function values are obtained, the test statistic is simple to calculate.

Wald test If the hypotheses are true, then the same functional transformations applied to the GMM estimator should be close to zero. That is, $\boldsymbol{h}(\widehat{\boldsymbol{\theta}}) \approx \mathbf{0}$. The Wald test draws on this intuition and tests if $\boldsymbol{h}(\widehat{\boldsymbol{\theta}})$ is sufficiently close to $\mathbf{0}$. The Wald test statistic is

$$W = n\boldsymbol{h}(\widehat{\boldsymbol{\theta}})'\left[\widehat{\boldsymbol{H}}\left(\widehat{\boldsymbol{G}}'\widehat{\boldsymbol{S}}^{-1}\widehat{\boldsymbol{G}}\right)^{-1}\widehat{\boldsymbol{H}}'\right]^{-1}\boldsymbol{h}(\widehat{\boldsymbol{\theta}})$$

where $\widehat{\boldsymbol{G}}$ and $\widehat{\boldsymbol{S}}$ are as defined above and $\widehat{\boldsymbol{H}}$ is the $J \times K$ matrix of derivatives of $\boldsymbol{h}(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ evaluated at the GMM estimator:

$$\widehat{\boldsymbol{H}} = \left.\frac{\partial \boldsymbol{h}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}$$

The Wald test statistic has an asymptotic chi-squared distribution with degrees of freedom equal to the number of restrictions, J:

$$W \overset{a}{\sim} \chi^2(J)$$

Note that the Wald test requires estimating only the unrestricted model, but the additional calculations of the test statistic are more involved than those of the difference test.

Lagrange multiplier test If the hypotheses are true, then the GMM estimator of the restricted model should be close to the GMM estimator of the unrestricted model. The slope of the objective function at the unrestricted GMM estimator is zero, so the slope of the objective function at the restricted GMM estimator should be close to zero. That is, $\partial Q(\boldsymbol{\theta})/\partial\boldsymbol{\theta} \approx \mathbf{0}$ when evaluated at the restricted GMM estimator, $\widehat{\boldsymbol{\theta}}_R$. The Lagrange multiplier test draws on this intuition and tests if $\partial Q(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ evaluated at $\widehat{\boldsymbol{\theta}}_R$ is sufficiently close to $\mathbf{0}$. The Lagrange multiplier test statistic is

$$LM = n\left[\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{m}(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i, \widehat{\boldsymbol{\theta}}_R)\right]'\widehat{\boldsymbol{S}}^{-1}\widehat{\boldsymbol{G}}_R\left(\widehat{\boldsymbol{G}}_R'\widehat{\boldsymbol{S}}^{-1}\widehat{\boldsymbol{G}}_R\right)^{-1}\widehat{\boldsymbol{G}}_R'\widehat{\boldsymbol{S}}^{-1}\left[\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{m}(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i, \widehat{\boldsymbol{\theta}}_R)\right]$$

where $\widehat{\boldsymbol{S}}$ is as defined above and $\widehat{\boldsymbol{G}}_R$ is similar to above but evaluated at the restricted GMM estimator, $\widehat{\boldsymbol{\theta}}_R$. The Lagrange multiplier test statistic has an asymptotic chi-squared distribution with degrees of freedom equal to the number of restrictions, J:

$$LM \overset{a}{\sim} \chi^2(J)$$

Note that the Lagrange multiplier test requires estimating only the restricted model, but the additional calculations of the test statistic are more involved than those of the difference test.