The Generalised Method of Moments

1 Introduction

A model is represented by a set of DGPs. Each DGP in the model is characterised by a parameter vector, which we will normally denote by $\boldsymbol{\beta}$ in the case of regression functions and by $\boldsymbol{\theta}$ in the general case. The starting point for generalised method of moments (GMM) estimation is to specify functions, which, for any DGP in the model, depend both on the data generated by that DGP and on the model parameters. When these functions are evaluated at the parameters that correspond to the DGP that generated the data, their expectation must be zero.

As a simple example, consider the linear regression model $y_t = \mathbf{X}_t\boldsymbol{\beta} + u_t$. An important part of the model specification is that the error terms have mean zero. These error terms are unobservable, because the parameters $\boldsymbol{\beta}$ of the regression function are unknown. But we can define the residuals $u_t(\boldsymbol{\beta}) \equiv y_t - \mathbf{X}_t\boldsymbol{\beta}$ as functions of the observed data and the unknown model parameters, and these functions provide what we need for GMM estimation. If the residuals are evaluated at the parameter vector $\boldsymbol{\beta}_0$ associated with the true DGP, they have mean zero under that DGP, but if they are evaluated at some $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, they do not have mean zero. We used this fact to develop a method-of-moments (MM) estimator for the parameter vector $\boldsymbol{\beta}$ of the regression function. As we will see, the various GMM estimators of $\boldsymbol{\beta}$ include as a special case the MM (or OLS) estimator.[1]

2 GMM estimators for linear regression models

Consider the linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \quad \mathbb{E}[\mathbf{u}\mathbf{u}^\top] = \boldsymbol{\Omega}, \tag{1}$$

where there are n observations, and $\boldsymbol{\Omega}$ is an $n \times n$ covariance matrix. Some of the explanatory variables that form the $n \times k$ matrix $\mathbf{X}$ may not be predetermined with respect to the error terms $\mathbf{u}$. However, there is assumed to exist an $n \times l$ matrix of predetermined instrumental variables, $\mathbf{W}$ with $n > l$ and $n \geq k$ satisfying the condition $\mathbb{E}[u_t|\mathbf{W}_t] = 0$ for each row $\mathbf{W}_t$ of $\mathbf{W}$, $t = 1, ..., n$. Any column of $\mathbf{X}$ that is predetermined must also be a column of $\mathbf{W}$. In addition, we assume that, for all $t, s = 1, ..., n$, $\mathbb{E}[u_t u_s|\mathbf{W}_t, \mathbf{W}_s] = \omega_{ts}$, where ω_{ts} is the ts-th element of $\boldsymbol{\Omega}$. We will need this assumption later, because it allows us to see that:

$$\begin{aligned}
\text{Var}\left(\frac{1}{\sqrt{n}}\mathbf{W}^\top\mathbf{u}\right) &= \frac{1}{n}\mathbb{E}[\mathbf{W}^\top\mathbf{u}\mathbf{u}^\top\mathbf{W}] = \frac{1}{n}\sum_{t=1}^{\top} n \sum_{s=1}^{n}\mathbb{E}[u_t u s \mathbf{W}_t^\top\mathbf{W}_s] \\
&= \frac{1}{n}\sum_{t=1}^{n}\sum_{s=1}^{n}\mathbb{E}\left[\mathbb{E}[u_t u_s\mathbf{W}_t^\top\mathbf{W}_s|\mathbf{W}_t, \mathbf{W}_s]\right] \\
&= \frac{1}{n}\sum_{t=1}^{n}\sum_{s=1}^{n}\mathbb{E}[\omega_{ts}\mathbf{W}_t^\top\mathbf{W}_s] = \frac{1}{n}\mathbb{E}[\mathbf{W}^\top\boldsymbol{\Omega}\mathbf{W}].
\end{aligned} \tag{2}$$

[1] These notes are based on Chapter 9 from Davidson and MacKinnon (2004). But Hayashi (2000) provides very good treatment of GMM throughout the text too. I can also recommend Hansen's notes (which are still unpublished and online as of the time of writing). I, personally, am not a fan of Wooldridge's exposition when it comes to matrix notation, but his text works for some people too.

The assumption that $\mathbb{E}[u_t|\mathbf{W}_t] = 0$ implies that, for all $t = 1, ..., n$,

$$\mathbb{E}[\mathbf{W}_t^\top (y_t - \mathbf{X}_t\boldsymbol{\beta})] = \mathbf{0}. \tag{3}$$

These equations form a set of what we call theoretical moment conditions. These conditions were used when we looked at IV estimation as the starting point for MM estimation of the model (1). Each theoretical moment conditions corresponds to a sample moment, or empirical moment, of the form:

$$\frac{1}{n}\sum_{t=1}^{n} w_{ti}^\top (y_t - \mathbf{X}_t\boldsymbol{\beta}) = \frac{1}{n}\mathbf{w}_i^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \tag{4}$$

where $\mathbf{w}_i, i = 1, ..., l$, is the i-th column of $\mathbf{W}$, and w_{ti} is the ti-th element. When $l = k$, we can set these sample moments equal to zero and solve the resulting k equations to obtain the simple IV estimator:

$$\hat{\boldsymbol{\beta}}_{\text{IV}} = (\mathbf{W}^\top \mathbf{X})^{-1}\mathbf{W}^\top \mathbf{y}.$$

When $l > k$, we must select k independent linear combinations of the sample moments (4) in order to obtain an estimator.

Now let $\mathbf{J}$ be an $l \times k$ matrix with full column rank k, and consider the MM estimator obtained by using the k columns of $\mathbf{WJ}$ as instruments. This estimator solves the k equations:

$$\mathbf{J}^\top \mathbf{W}^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{0}, \tag{5}$$

which are referred to as sample moment conditions, or just moment conditions when there is no ambiguity. They are also sometimes called orthogonality conditions, since they require that the vector of residuals should be orthogonal to the columns $\mathbf{WJ}$. Assuming that we have data generated by a DGP which belongs to the model (1), with true coefficient vector $\boldsymbol{\beta}_0$ and variance-covariance matrix $\boldsymbol{\Omega}_0$. Under this assumption, we have the following expression, suitable for asymptotic analysis, for the estimator $\hat{\boldsymbol{\beta}}$ that solves (5):

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = \left(\frac{1}{n}\mathbf{J}^\top \mathbf{W}^\top \mathbf{X}\right)^{-1} \frac{1}{\sqrt{n}}\mathbf{J}^\top \mathbf{W}^\top \mathbf{u}. \tag{6}$$

From this, recalling (2), we find that the asymptotic variance-covariance matrix of $\hat{\boldsymbol{\beta}}$, that is, the variance-covariance matrix of the plim of $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$, is

$$\mathbb{E}\left[\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\right] = \left(\frac{1}{n}\mathbf{J}\mathbf{W}^\top \mathbf{X}\right)^{-1} \frac{1}{\sqrt{n}}\mathbf{J}^\top \mathbf{W}^\top \mathbf{u}\mathbf{u}^\top \mathbf{W}\mathbf{J}\frac{1}{\sqrt{n}}\left(\frac{1}{n}\mathbf{X}^\top \mathbf{W}\mathbf{J}\right)^{-1},$$

or after some simplification:

$$\text{AVar}\left(\hat{\boldsymbol{\beta}}\right) = \left(\operatorname*{plim}_{n\to\infty}\frac{1}{n}\mathbf{J}^\top \mathbf{W}^\top \mathbf{X}\right)^{-1}\left(\operatorname*{plim}_{n\to\infty}\frac{1}{n}\mathbf{J}^\top \mathbf{W}^\top \boldsymbol{\Omega}_0 \mathbf{W}\mathbf{J}\right)\left(\operatorname*{plim}_{n\to\infty}\frac{1}{n}\mathbf{X}^\top \mathbf{W}\mathbf{J}\right)^{-1}. \tag{7}$$

This matrix has the familiar sandwich form that we expect to see when an estimator is not asymptotically efficient.

The next step is to choose $\mathbf{J}$ so as to minimise the variance-covariance matrix (7). We want to get rid of the sandwich form, so the simplest choice for $\mathbf{J}$ which does is:

$$\mathbf{J} = \left(\mathbf{W}^\top \boldsymbol{\Omega}_0 \mathbf{W}\right)^{-1}\mathbf{W}^\top \mathbf{X}; \tag{8}$$

notice that, in the special case in which $\mathbf{\Omega}_0$ is proportional to $\mathbf{I}$, this expression reduces to:

$$\mathbf{J} = (\mathbf{W}^\top \mathbf{W})^{-1} \mathbf{W}^\top \bar{\mathbf{X}}, \tag{9}$$

in the case on generalised IV estimation, where $\bar{\mathbf{X}}$ is the expectation of $\mathbf{X}$ conditional on the information set $\mathbf{\Omega}_t$.[2] Therefore, (8) is the appropriate generalisation of (9) when $\mathbf{\Omega}$ is not proportional to an identity matrix. With $\mathbf{J}$ defined by (8), the variance-covariance matrix (7) becomes:

$$\operatorname*{plim}_{n\to\infty} \left(\frac{1}{n} \mathbf{X}^\top \mathbf{W} (\mathbf{W}^\top \mathbf{\Omega}_0 \mathbf{W})^{-1} \mathbf{W}^\top \mathbf{X} \right)^{-1}, \tag{10}$$

and the efficient GMM estimator is:

$$\hat{\boldsymbol{\beta}}_{\text{GMM}} = (\mathbf{X}^\top \mathbf{W} (\mathbf{W}^\top \mathbf{\Omega}_0 \mathbf{W})^{-1} \mathbf{W}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W} (\mathbf{W}^\top \mathbf{\Omega}_0 \mathbf{W})^{-1} \mathbf{W}^\top \mathbf{y}. \tag{11}$$

When $\mathbf{\Omega}_0 = \sigma^2 \mathbf{I}$, this estimator reduces to the generalised IV (GIV) estimator:

$$\hat{\boldsymbol{\beta}}_{\text{IV}} = (\mathbf{X}^\top \mathbf{P}_{\mathbf{W}} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{P}_{\mathbf{W}} \mathbf{y}. \tag{12}$$

We will see that a more efficient estimator is available if we know $\mathbf{\Omega}_0$ and are prepared to exploit that knowledge.

2.1 The GMM criterion function

With both GLS and IV estimation, we showed that the efficient estimators could also be derived by minimising an appropriate criterion function. The efficient GMM estimator (11) minimises the GMM criterion function:

$$Q(\boldsymbol{\beta}, \mathbf{y}) \equiv (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top \mathbf{W} (\mathbf{W}^\top \mathbf{\Omega}_0 \mathbf{W})^{-1} \mathbf{W}^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \tag{13}$$

as can be seen at once by noting that the first-order conditions for minimising (13) are:

$$\mathbf{X}^\top \mathbf{W} (\mathbf{W}^\top \mathbf{\Omega}_0 \mathbf{W})^{-1} \mathbf{W}^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{0}.$$

If we assume homoskedastic error variance and $\mathbf{\Omega}_0 = \sigma_0^2 \mathbf{I}$, then (13) reduces to the IV criterion function divided by σ_0^2:

$$Q(\boldsymbol{\beta}, \mathbf{y}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top \mathbf{P}_{\mathbf{W}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \tag{14}$$

We saw when we looked at IV estimation that the minimised value of the IV criterion function, divided by an estimate of σ^2, serves as the statistic for the Sargan test for overidentification. We will soon see that the GMM criterion function (13), with the usually unknown matrix $\mathbf{\Omega}_0$ replaced by a suitable estimate, can be used as a test statistic for overidentification.

The criterion function (13) is a quadratic form in the vector $\mathbf{W}^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$ of sample moments and the inverse of the matrix $\mathbf{W}^\top \mathbf{\Omega}_0 \mathbf{W}$. Equivalently, it is a quadratic form in $n^{-1/2} \mathbf{W}^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$ and the inverse of $n^{-1} \mathbf{W}^\top \mathbf{\Omega}_0 \mathbf{W}$, since the powers of n cancel. Under the sort of regularity conditions we have used earlier, $n^{-1/2} \mathbf{W} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_0)$ satisfies a central limit theorem (CLT), and so tends, as $n \to \infty$, to a normal random variable, with mean vector $\mathbf{0}$ and covariance matrix the limit of $n^{-1} \mathbf{W} \mathbf{\Omega}_0 \mathbf{W}^\top$. It follows that (13) evaluated using the true $\boldsymbol{\beta}_0$ and the true $\mathbf{\Omega}_0$ is asymptotically distributed as χ^2 with l degrees of freedom.[3]

[2] Of course, in practice, $\bar{\mathbf{X}}$ is never observed, and it should be replaced by something that estimates it consistently.

[3] Recall the following theorem (Theorem 4.1 from Davidson and MacKinnon (2004)):

This property of the GMM criterion function is simply a consequence of its structure as a quadratic form in the sample moments used for estimation and the inverse of the asymptotic covariance matrix of these moments evaluated at the true parameters. This makes the GMM criterion function useful for testing.

Provided that the instruments are predetermined, so that they satisfy the condition that $\mathbb{E}[u_t \mathbf{W}_t] = 0$, we still obtain a consistent estimator, even when the matrix $\mathbf{J}$ used to select linear combinations of the instruments is different from (8). Such a consistent, but in general inefficient, estimator can also be obtained by minimising a quadratic criterion function of the form:

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top \mathbf{W} \boldsymbol{\Lambda} \mathbf{W}^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \tag{15}$$

where the weighting matrix $\boldsymbol{\Lambda}$ is $l \times l$, positive definite, and must be at least asymptotically nonrandom. Without loss of generality, $\boldsymbol{\Lambda}$ can be taken to be symmetric. The inefficient GMM estimator is:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{W} \boldsymbol{\Lambda} \mathbf{W}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W} \boldsymbol{\Lambda} \mathbf{W}^\top \mathbf{y}, \tag{16}$$

from which it can be seen that the use of the weighting matrix $\boldsymbol{\Lambda}$ corresponds to the implicit choice $\mathbf{J} = \boldsymbol{\Lambda} \mathbf{W}^\top \mathbf{X}$. For a given choice of $\mathbf{J}$, there are various possible choices of $\boldsymbol{\Lambda}$ that give rise to the same estimator.

When $l = k$, the model is exactly identified, and $\mathbf{J}$ is a nonsingular square matrix which has no effect on the estimator. This is most easily seen by looking at the moment conditions (5), which are equivalent, when $l = k$, to those obtained by premultiplying them by $(\mathbf{J}^\top)^{-1}$. Similarly, if the estimator is defined by minising a quadratic form, it does not depend on the choice of $\boldsymbol{\Lambda}$ whenever $l = k$. To see this, consider the first-order conditions for minimising (15), which, up to a scalar factor, are:

$$\mathbf{X}^\top \mathbf{W} \boldsymbol{\Lambda} \mathbf{W}^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{0}.$$

If $l = k$, $\mathbf{X}^\top \mathbf{W}$ is a square matrix, and the first-order conditions can be premultiplied by $\boldsymbol{\Lambda}^{-1} (\mathbf{X}^\top \mathbf{W})^{-1}$. Therefore, the estimator is the solution to the equations $\mathbf{W}^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{0}$, independently of $\boldsymbol{\Lambda}$. This solution is just the simple IV estimator we've looked at before:

$$\hat{\boldsymbol{\beta}}_{\text{IV}} = (\mathbf{W}^\top \mathbf{X})^{-1} \mathbf{W}^\top \mathbf{y}.$$

When $l > k$, the model is overidentified, and the estimator (16) depends on the choice of $\mathbf{J}$ or $\boldsymbol{\Lambda}$. The efficient GMM estimator, for a given set of instruments, is defined in terms of the true variance-covariance matrix $\boldsymbol{\Omega}_0$, which is usually unknown. If $\boldsymbol{\Omega}_0$ is known up to a scalar multiplicative factor, so that $\boldsymbol{\Omega}_0 = \sigma^2 \boldsymbol{\Delta}_0$, with σ^2 unknown and $\boldsymbol{\Delta}_0$ known, then $\boldsymbol{\Delta}_0$ can be used in place of $\boldsymbol{\Omega}_0$ in either (11) or (13). This is true because multiplying $\boldsymbol{\Omega}_0$ by a scalar leaves (11) invariant, and it also leaves invariant the $\boldsymbol{\beta}$ that minimises (13).

2.2 GMM estimation with heteroskedasticity of unknown form

The assumption that $\boldsymbol{\Omega}_0$ is known, even up to a scalar factor, is often too strong. When we consider GMM estimation, $\boldsymbol{\Omega}_0$ appears only through the $l \times l$ matrix product $\mathbf{W}^\top \boldsymbol{\Lambda} \mathbf{W}$. In the context of heteroskedasticity consistent covariance matrix estimation (HCCME), n^{-1}

- If the m-vector $\mathbf{x}$ is distributed as $\mathrm{N}(\mathbf{0}, \boldsymbol{\Omega})$, then the quadratic form $\mathbf{x}^\top \boldsymbol{\Omega} \mathbf{x}$ is distributed as $\chi^2(m)$; and
- If $\mathbf{P}$ is a projection matrix with rank r and $\mathbf{z}$ is an n-vector that is distributed as $\mathrm{N}(\mathbf{0}, \mathbf{I})$, then the quadratic form $\mathbf{z}^\top \mathbf{P} \mathbf{z}$ is distributed as $\chi^2(r)$.

times such a matrix can be estimated consistently if $\boldsymbol{\Omega}_0$ is a diagonal matrix. What is needed is preliminary consistent estimate of the parameter vector $\boldsymbol{\beta}$, which furnishes residuals that are consistent estimates of the error terms.

The preliminary estimates of $\boldsymbol{\beta}$ must be consistent, but they need not be asymptotically efficient, and so we can obtain them by using any convenient choice of $\mathbf{J}$ or $\boldsymbol{\Lambda}$. One choice that is often convenient is $\boldsymbol{\Lambda} = (\mathbf{W}^\top \mathbf{W})^{-1}$, in which case the preliminary estimator is the GIV estimator (12). We can then use the preliminary estimate $\hat{\boldsymbol{\beta}}$ to calculate the residuals $\hat{u}_t \equiv y_t - \mathbf{X}\hat{\boldsymbol{\beta}}$. A typical element of the matrix $n^{-1}\mathbf{W}^{-1}\boldsymbol{\Omega}_0\mathbf{W}$ can then be estimated by:

$$\frac{1}{n}\sum_{t=1}^{n} \hat{u}_t^2 w_{ti} w_{tj}. \tag{17}$$

This estimator can be proved to be consistent by using arguments similar to what we did when looking at White HCCME.

The matrix with typical element (17) can be written as $n^{-1}\mathbf{W}^\top \hat{\boldsymbol{\Omega}}\mathbf{W}$, where $\hat{\boldsymbol{\Omega}}$ is an $n \times n$ matrix with typical element $\hat{u}_t^2$. Then the feasible efficient GMM (FGMM) estimator is:

$$\hat{\boldsymbol{\beta}}_{\text{FGMM}} = (\mathbf{X}^\top \mathbf{W}(\mathbf{W}^\top \hat{\boldsymbol{\Omega}}\mathbf{W})^{-1}\mathbf{W}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{W}(\mathbf{W}^\top \hat{\boldsymbol{\Omega}}\mathbf{W})^{-1}\mathbf{W}^\top \mathbf{y}, \tag{18}$$

which is just (11) with $\boldsymbol{\Omega}_0$ replaced with $\hat{\boldsymbol{\Omega}}$. Since $n^{-1}\mathbf{W}^\top \hat{\boldsymbol{\Omega}}\mathbf{W}$ consistently estimates $n^{-1}\mathbf{W}^\top \boldsymbol{\Omega}\mathbf{W}$, it follows that $\hat{\boldsymbol{\beta}}_{\text{FEGMM}}$ is asymptotically equivalent to (11).

Like other procedures that start from a preliminary estimate, this one can be iterated. The GMM residuals $y_t - \mathbf{X}\hat{\boldsymbol{\beta}}_{\text{FGMM}}$ can be used to calculate a new estimate of $\boldsymbol{\Omega}$, which can then be used to obtain second-round GMM estimates, which can then be used to calculate yet another estimate of $\boldsymbol{\Omega}$, and so on. This is known as 'continuously updated GMM'. Whether we stop after one round or continue until the procedure converges, the estimates have the same asymptotic distribution if the model is correctly specified. However, there is evidence that performing more iterations improves finite-sample performance. In practice, the covariance matrix is estimate by:

$$\widehat{\text{Var}}\left(\hat{\boldsymbol{\beta}}_{\text{FGMM}}\right) = \left(\mathbf{X}^\top \mathbf{W}(\mathbf{W}^\top \hat{\boldsymbol{\Omega}}\mathbf{W})^{-1}\mathbf{W}^\top \mathbf{X}\right)^{-1}. \tag{19}$$

It is not hard to see that n times the estimator (19) tends to the asymptotic covariance matrix (10) as $n \to \infty$.

2.3 Fully efficient GMM estimation

We choose a particular matrix of instrumental variables $\mathbf{W}$, which implies that we choose a representation of the information sets Ω_t appropriate for each observation in the sample. Since $\mathbf{W}_t \in \Omega_t$, it follows that any deterministic function – linear or nonlinear – of the elements of $\mathbf{W}_t$ also belongs to Ω_t. We know that we have to choose $\mathbf{W}$, and once this choice is made, our expression for $\mathbf{J}$ in (8) gives the optimal set of linear combinations of the columns of $\mathbf{W}$ to use for estimation. So how best to choose $\mathbf{W}$ given the information sets Ω_t?

When we looked at IV estimation, for the model (1) with $\boldsymbol{\Omega} = \sigma^2\mathbf{I}$, the best choice, by the criterion of the asymptotic covariance matrix, is the matrix $\bar{\mathbf{X}}$ by the defining condition that $\mathbb{E}[\mathbf{X}_t|\Omega_t] = \bar{\mathbf{X}}_t$, where $\mathbf{X}_t$ and $\bar{\mathbf{X}}_t$ are the t-th rows of $\mathbf{X}$ and $\bar{\mathbf{X}}$, respectively. However, this result does not hold when $\boldsymbol{\Omega}$ is not proportional to an identity matrix. Consider the GMM estimator (11) of which (18) is the feasible version. In the special case of exogenous

explanatory variables, for which the obvious choice of $\mathbf{W} = \mathbf{X}$, and for notational ease we write $\boldsymbol{\Omega}$ for the true covariance matrix $\boldsymbol{\Omega}_0$, (11) becomes:

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{\text{GMM}} &= (\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\boldsymbol{\Omega}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\boldsymbol{\Omega}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} \\ &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\Omega}\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\boldsymbol{\Omega}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} \\ &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\Omega}\mathbf{X}(\mathbf{X}^\top\boldsymbol{\Omega}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} \\ &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} = \hat{\boldsymbol{\beta}}_{\text{OLS}}.\end{aligned}$$

But we know that the efficient estimator is actually the GLS estimator:

$$\hat{\boldsymbol{\beta}}_{\text{GLS}} = (\mathbf{X}^\top\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\Omega}^{-1}\mathbf{y}, \tag{20}$$

which is usually different from $\hat{\boldsymbol{\beta}}_{\text{OLS}}$.

The GLS estimator (20) can be interpreted as an IV estimator, in which the instruments are the columns of $\boldsymbol{\Omega}^{-1}\mathbf{X}$. So when $\boldsymbol{\Omega}$ is not a multiple of the identity matrix, the optimal instruments are no longer the explanatory variables $\mathbf{X}$, but rather the columns of $\boldsymbol{\Omega}^{-1}\mathbf{X}$. When at least some of the explanatory variables in the matrix $\mathbf{X}$ are not predetermined, the optimal choice of instruments is given by $\boldsymbol{\Omega}^{-1}\bar{\mathbf{X}}$. This choice combines the optimality of the GLS estimator and the best instruments to use in place of explanatory variables that are not predetermined. It leads to the theoretical moment conditions:

$$\mathbb{E}[\bar{\mathbf{X}}^\top\boldsymbol{\Omega}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})] = \mathbf{0}. \tag{21}$$

Unfortunately, this solution does not always work, because the aforementioned moment conditions may not be correct. To see why not, suppose that the error terms are serially correlated, and that $\boldsymbol{\Omega}$ is consequently not a diagonal matrix. The i-th element of the matrix product in (21) can be expanded as:

$$\sum_{t=1}^{n}\sum_{s=1}^{n}\bar{\mathbf{X}}_{ti}\omega_{ts}(y_s - \mathbf{X}_s\boldsymbol{\beta}), \tag{22}$$

where ω_{ts} is the ts-th element of $\boldsymbol{\Omega}^{-1}$. If we evaluate at the true parameter vector $\boldsymbol{\beta}_0$, we find that $y_s - \mathbf{X}_s\boldsymbol{\beta}_0 = u_s$. But, unless the columns of the matrix $\bar{\mathbf{X}}$ are exogenous, it is not in general the case that $\mathbb{E}[u_s|\bar{\mathbf{X}}_t] = 0$ for $s \neq t$, and, if this condition is not satisfied, the expectations of (22) is not zero in general.

2.4 Choosing valid instruments

Like we did when we looked at GLS, we can construct an $n \times n$ matrix $\boldsymbol{\Psi}$, usually triangular, that satisfies the equation $\boldsymbol{\Omega}^{-1} = \boldsymbol{\Psi}\boldsymbol{\Psi}^\top$. We can premultiply our simple linear regression model by $\boldsymbol{\Psi}^\top$ to get:

$$\boldsymbol{\Psi}^\top\mathbf{y} = \boldsymbol{\Psi}^\top\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\Psi}^\top\mathbf{u}, \tag{23}$$

with the result that the variance covariance matrix of the transformed error vector, $\boldsymbol{\Psi}^\top\mathbf{u}$, is just the identity matrix. Suppose that we propose to use a matrix $\mathbf{Z}$ of instruments in order to estimate the transformed model, so that we are led to consider the theoretical moment conditions:

$$\mathbb{E}\left[\mathbf{Z}^\top\boldsymbol{\Psi}^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right] = \mathbf{0}. \tag{24}$$

If these conditions are correct, then what we need is that, for each t, $\mathbb{E}[(\boldsymbol{\Psi}^\top\mathbf{u})_t|\mathbf{Z}_t] = 0$, where the subscript t is used to select the t-th row of the corresponding vector or matrix.

If $\mathbf{X}$ is exogenous, optimal instruments are given by the matrix $\boldsymbol{\Omega}^{-1}\mathbf{X}$, and the moment conditions $\mathbb{E}[\mathbf{X}^\top\boldsymbol{\Omega}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})]=\mathbf{0}$, which can be written as:

$$\mathbb{E}\left[\mathbf{X}^\top\boldsymbol{\Psi}\boldsymbol{\Psi}^\top(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right]=\mathbf{0}. \tag{25}$$

Comparison with (24) shows that the optimal choice of $\mathbf{Z}$ is $\boldsymbol{\Psi}^\top\mathbf{X}$. Even if $\mathbf{X}$ is not exogenous, (25) is a correct set of moment conditions if:

$$\mathbb{E}\left[(\boldsymbol{\Psi}^\top\mathbf{u})_t|(\boldsymbol{\Psi}^\top\mathbf{X})_t\right]=0. \tag{26}$$

But this is not true in general when $\mathbf{X}$ is not exogenous. Consequently, we seek a new definition for $\bar{\mathbf{X}}$, such that (26) becomes true when $\mathbf{X}$ is replaced by $\bar{\mathbf{X}}$.

In most cases, we can choose $\boldsymbol{\Psi}$ so that $(\boldsymbol{\Psi}^\top\mathbf{u})_t$ is an innovation so that $\mathbb{E}[(\boldsymbol{\Psi}^\top\mathbf{u})_t|\Omega_t]=0$.[4] What is then required for condition (26) is that $(\boldsymbol{\Psi}^\top\bar{\mathbf{X}})_t$ should be predetermined in period t, and that it should belong to Ω_t. So we define $\bar{\mathbf{X}}$ implicitly by the equation:

$$\mathbb{E}[(\boldsymbol{\Psi}^\top\mathbf{X})_t|\Omega_t]=(\boldsymbol{\Psi}^\top\bar{\mathbf{X}})_t. \tag{27}$$

This implicit definition must be implemented on a case-by-case basis. By setting $\mathbf{Z}=\boldsymbol{\Psi}^\top\bar{\mathbf{X}}$, we find that the moment conditions (24) become:

$$\mathbb{E}[\bar{\mathbf{X}}^\top\boldsymbol{\Psi}\boldsymbol{\Psi}^\top(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})]=\mathbb{E}[\bar{\mathbf{X}}\boldsymbol{\Omega}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})]=\mathbf{0}. \tag{28}$$

These conditions do indeed use $\boldsymbol{\Omega}^{-1}\bar{\mathbf{X}}$ as instruments, albeit with a possibly redefined $\bar{\mathbf{X}}$. The estimator based on (28) is:

$$\hat{\boldsymbol{\beta}}_{\text{EGMM}}=(\bar{\mathbf{X}}^\top\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\bar{\mathbf{X}}^\top\boldsymbol{\Omega}^{-1}\mathbf{y}, \tag{29}$$

where EGMM denotes "fully efficient GMM". The asymptotic covariance matrix of (29) can be computed using (10), in which, on the basis of (28), we see that $\mathbf{W}$ is to be replaced by $\boldsymbol{\Psi}\bar{\mathbf{X}}$, $\mathbf{X}$ by $\boldsymbol{\Psi}^\top\mathbf{X}$, and $\boldsymbol{\Omega}$ by $\mathbf{I}$:

$$\text{AVar}\left(\hat{\boldsymbol{\beta}}_{\text{EGMM}}\right)=\operatorname*{plim}_{n\to\infty}\left(\frac{1}{n}\bar{\mathbf{X}}^\top\boldsymbol{\Omega}^{-1}\mathbf{X}\left(\frac{1}{n}\bar{\mathbf{X}}^\top\boldsymbol{\Omega}\bar{\mathbf{X}}\right)^{-1}\frac{1}{n}\mathbf{X}^\top\boldsymbol{\Omega}^{-1}\bar{\mathbf{X}}\right)^{-1}. \tag{30}$$

We find that for any matrix $\mathbf{Z}$ that satisfies $\mathbf{Z}_t\in\Omega_t$,

$$\operatorname*{plim}_{n\to\infty}\frac{1}{n}\mathbf{Z}^\top\boldsymbol{\Psi}^\top\mathbf{X}=\operatorname*{plim}_{n\to\infty}\frac{1}{n}\mathbf{Z}^\top\boldsymbol{\Omega}\bar{\mathbf{X}}. \tag{31}$$

Since $(\boldsymbol{\Psi}^\top\bar{\mathbf{X}})_t\in\Omega_t$, this implies that:

$$\begin{aligned}\operatorname*{plim}_{n\to\infty}\frac{1}{n}\bar{\mathbf{X}}^\top\boldsymbol{\Omega}^{-1}\mathbf{X}&=\operatorname*{plim}_{n\to\infty}\frac{1}{n}\bar{\mathbf{X}}^\top\boldsymbol{\Psi}\boldsymbol{\Psi}^\top\mathbf{X}\\&=\operatorname*{plim}_{n\to\infty}\frac{1}{n}\bar{\mathbf{X}}^\top\boldsymbol{\Psi}\boldsymbol{\Psi}^\top\bar{\mathbf{X}}\\&=\operatorname*{plim}_{n\to\infty}\frac{1}{n}\bar{\mathbf{X}}^\top\boldsymbol{\Omega}^{-1}\bar{\mathbf{X}}.\end{aligned}$$

[4] Consider models with AR(1) errors.

Therefore, the asymptotic covariance matrix (30) simplifies to:

$$\operatorname*{plim}_{n\to\infty}\left(\frac{1}{n}\bar{\mathbf{X}}^\top\boldsymbol{\Omega}^{-1}\bar{\mathbf{X}}\right)^{-1}. \tag{32}$$

Note that this matrix is even lesser of a sandwich than (10).

In most cases, $\bar{\mathbf{X}}$ is not observed, but it can be estimated consistently. We have an $n\times l$ matrix $\mathbf{W}$ of instruments, such that $\mathcal{S}(\bar{\mathbf{X}})\subseteq\mathcal{S}(\mathbf{W})$ and

$$(\boldsymbol{\Psi}^\top\mathbf{W})_t\in\Omega_t. \tag{33}$$

This condition is the form taken by the predeterminedness condition when $\boldsymbol{\Omega}$ is not proportional to the identity matrix. The theoretical moment conditions used for (overidentified) estimation are then

$$\mathbb{E}[\mathbf{W}^\top\boldsymbol{\Omega}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})]=\mathbb{E}[\mathbf{W}^\top\boldsymbol{\Psi}\boldsymbol{\Psi}^\top(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})]=\mathbf{0}, \tag{34}$$

where basically we are estimating the transformed model (23) using the transformed instruments $\boldsymbol{\Psi}^\top\mathbf{W}$. If indeed $\mathcal{S}(\bar{\mathbf{X}})\subseteq\mathcal{S}(\mathbf{W})$, the asymptotic covariance matrix of the resulting estimator is still (32).

The main obstacle to the use of the efficient estimation $\hat{\boldsymbol{\beta}}_{\text{EGMM}}$ is thus not the difficulty of estimating $\bar{\mathbf{X}}$, but rather the fact that $\boldsymbol{\Omega}$ is usually not known. As with GLS estimators, $\hat{\boldsymbol{\beta}}_{\text{EGMM}}$ cannot be calculated unless we either know $\boldsymbol{\Omega}$ or can estimate it consistently, usually by knowing the form of $\boldsymbol{\Omega}$ as a function of parameters that can be estimated consistently. But whenever there is heteroskedasticity or serial correlation of unknown form, this is impossible. The best we can then do, asymptotically, is to use the feasible efficient GMM estimator (18).

3 HAC covariance matrix estimation

Here we allow of the possibility of serial correlation of unknown form, which causes $\boldsymbol{\Omega}$ to have non-zero off-diagonal elements. When the pattern of the serial correlation is unknown, we can still, under fairly weak regularity conditions, estimate the covariance matrix of the sample moments by using a heteroskedasticity autocorrelation consistent (HAC) estimator of the matrix $n^{-1}\mathbf{W}^\top\boldsymbol{\Omega}\mathbf{W}$. This estimator, multiplied by n, can then be used in place of $\mathbf{W}^\top\hat{\boldsymbol{\Omega}}\mathbf{W}$ in the feasible efficient GMM estimator (18).

The asymptotic covariance matrix of the vector $n^{-1/2}\mathbf{W}^\top(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})$ of sample moments, evaluated at $\boldsymbol{\beta}=\boldsymbol{\beta}_0$, defined as follows:

$$\boldsymbol{\Sigma}\equiv\operatorname*{plim}_{n\to\infty}\frac{1}{n}\mathbf{W}^\top(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}_0)(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}_0)^\top\mathbf{W}=\operatorname*{plim}_{n\to\infty}\frac{1}{n}\mathbf{W}^\top\boldsymbol{\Omega}\mathbf{W}. \tag{35}$$

A HAC estimator of $\boldsymbol{\Sigma}$ is a matrix $\hat{\boldsymbol{\Sigma}}$ constructed so that $\hat{\boldsymbol{\Sigma}}$ consistently estimates $\boldsymbol{\Sigma}$ when the error terms u_t display any pattern of heteroskedasticity and/or autocorrelation that satisfies certain, generally quite weak, conditions. First, begin by rewriting the definition of $\boldsymbol{\Sigma}$:

$$\boldsymbol{\Sigma}=\lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\sum_{s=1}^{n}\mathbb{E}[u_tu_s\mathbf{W}_t^\top\mathbf{W}_s], \tag{36}$$

where we assume that a LLN can be used to justify replacing the probability limit in (35) by the expectations in (36).

For models with heteroskedasticity but no autocorrelation, only the terms with $t = s$ contribute to $\mathbf{\Sigma}$. Therefore, for such models, we can estimate $\mathbf{\Sigma}$ consistently by simply ignoring the expectations operator and replacing the error terms u_t by least-squares residuals $\hat{u}_t$, possibly with a modification designed to offset the tendency for such residuals to be too small. So we could drop the expectations operator and replace $u_t u_s$ by $\hat{u}_t \hat{u}_s$, where $\hat{u}_t$ denotes the t-th residual from some consistent but inefficient estimation procedure, such as generalised IV. Unfortunately, this approach does not work.

To see why this method does not work, let us rewrite (36) another way and define the autocovariance matrices of the $\mathbf{W}_t^\top u_t$ as follows:

$$\mathbf{\Gamma}(j) \equiv \begin{cases} \frac{1}{n} \sum_{t=j+1}^{n} \mathbb{E}[u_t u_{t-j} \mathbf{W}_t^\top \mathbf{W}_{t-j}] & \text{for } j \geq 0, \\ \frac{1}{n} \sum_{t=-j+1}^{n} \mathbb{E}[u_{t+j} u_t \mathbf{W}_{t+j}^\top \mathbf{W}_t] & \text{for } j < 0. \end{cases} \tag{37}$$

Because there are l moment conditions, these are $l \times l$ matrices. It is easy to check that $\mathbf{\Gamma}(j) = \mathbf{\Gamma}^\top(j)$. Then, in terms of the matrices $\mathbf{\Gamma}(j)$, expression (36) becomes:

$$\mathbf{\Sigma} = \lim_{n \to \infty} \sum_{j=-n+1}^{n-1} \mathbf{\Gamma}(j) = \lim_{n \to \infty} \left\{ \mathbf{\Gamma}(0) + \sum_{j=1}^{n-1} \left(\mathbf{\Gamma}(j) + \mathbf{\Gamma}^\top(j) \right) \right\}. \tag{38}$$

Therefore, in order to estimate $\mathbf{\Sigma}$, we apparently need to estimate all of the autocovariance matrices for $j = 0, ..., n-1$.

If $\hat{u}_t$ denotes a typical residual from some preliminary estimator, the sample autocovariance matrix of order j, $\hat{\mathbf{\Gamma}}(j)$, is just the appropriate expression in (37), without the expectation operator, and with the random variables u_t and u_{t-j} replaced by $\hat{u}_t$ and $\hat{u}_{t-j}$, respectively. For any $j \geq 0$, this is:

$$\hat{\mathbf{\Gamma}}(j) = \frac{1}{n} \sum_{t=j+1}^{n} \hat{u}_t \hat{u}_{t-j} \mathbf{W}_t^\top \mathbf{W}_{t-j}. \tag{39}$$

Unfortunately, the sample autocovariance matrix $\hat{\mathbf{\Gamma}}(j)$ of order j is not a consistent estimator of the true autocovariance matrix for arbitrary j. For example, consider $j = n-2$. Then we see that $\hat{\mathbf{\Gamma}}(n-2)$ has only two terms, and no conceivable law of large numbers can apply to only two terms. In fact, $\hat{\mathbf{\Gamma}}(n-2)$ must tend to zero as $n \to \infty$ because of the factor n^{-1} in its definition.

The solution to this problem is to restrict our attention to models for which the actual autocovariances mimic the behaviour of the sample autocovariances, and tend to zero as $j \to \infty$. We can drop most of the sample autocvariances matrices that appear in the sample analogue of (38) by eliminating ones for which $|j|$ is greater than some chosen threshold, say p. This yields the following estimator for $\mathbf{\Sigma}$:

$$\hat{\mathbf{\Sigma}}_{\text{HW}} = \hat{\mathbf{\Gamma}}(0) + \sum_{j=1}^{p} \left(\hat{\mathbf{\Gamma}}(j) + \hat{\mathbf{\Gamma}}^\top(j) \right). \tag{40}$$

This estimator is the Hansen-White estimator. For the purposes of asymptotic theory, it is necessary to let the parameter p, which is called the lag truncation parameter, go to infinity in (40) at some suitable rates as the sample size goes to infinity.

The Hansen-White estimator suffers from one very serious deficiency: In finite samples, it need not be positive definite or even positive semidefinite. If one happens to encounter a

data set that yields a nondefinite $\hat{\mathbf{\Gamma}}_{\text{HW}}$, then, since the weighting matrix for GMM must be positive definite, the Hansen-White estimator is unusable. In such a case, one can use the Newey-West estimator:

$$\hat{\mathbf{\Gamma}}_{\text{NW}} = \hat{\mathbf{\Gamma}}(0) + \sum_{j=1}^{p} \left(1 - \frac{j}{p+1}\right) \left(\hat{\mathbf{\Gamma}}(j) + \hat{\mathbf{\Gamma}}^{\top}(j)\right), \tag{41}$$

in which each sample autocovariance matrix $\hat{\mathbf{\Gamma}}(j)$ is multiplied by a weight $1 - j/(p+1)$ that decreases linearly as j increases. This estimator tends to underestimate the autocovariance matrices, especially for larger values of j. So p should almost certainly be larger for (41) than for (40). For both the Hansen-White and Newey-West estimators, p must increase as n does – typical rates rates are $n^{1/4}$ and $n^{1/3}$, respectively.

Both the Hansen-White and the Newey-West HAC estimators of $\mathbf{\Sigma}$ can be written in the form:

$$\hat{\mathbf{\Sigma}} = \frac{1}{n}\mathbf{W}^{\top}\hat{\mathbf{\Omega}}\mathbf{W}, \tag{42}$$

for an appropriate choice of $\hat{\mathbf{\Omega}}$. This face, which we will exploit in the next section, follows from the observation that there exist $n \times n$ matrices $\mathbf{U}(j)$ such that the $\hat{\mathbf{\Gamma}}(j)$ can be expressed in the form $n^{-1}\mathbf{W}^{\top}\mathbf{U}(j)\mathbf{W}$.

3.1 Feasible efficient GMM estimation

Efficient GMM estimation in the presence of heteroskedasticity and serial correlation of unknown form works as follows. As in the case with only heteroskedasticity in Section 2, we first obtain consistent but inefficient estimates, probably by using generalised IV. These estimates yield residuals $\hat{u}_t$, from which we next calculate a matrix $\hat{\mathbf{\Sigma}}$ that estimates $\mathbf{\Sigma}$ consistently, using either the Hansen-White or Newey-West estimator (or any other HAC estimator). The feasible efficient GMM estimator, which generalises (18) is then:

$$\hat{\boldsymbol{\beta}}_{\text{FGMM}} = (\mathbf{X}^{\top}\mathbf{W}\hat{\mathbf{\Sigma}}^{-1}\mathbf{W}^{\top}\mathbf{X})^{-1}\mathbf{X}^{\top}\mathbf{W}\hat{\mathbf{\Sigma}}^{-1}\mathbf{W}^{\top}\mathbf{y}. \tag{43}$$

As before, this procedure may be iterated. The first round GMM residuals may be used to obtain a new estimate for $\mathbf{\Sigma}$, which may be used to obtain a second round GMM estimate, and so on. For a correctly specified model, iteration should not affect the asymptotic properties of the estimates.

We can estimate the covariance matrix of $\hat{\boldsymbol{\beta}}_{\text{FGMM}}$ by:

$$\widehat{\text{Var}}\left(\hat{\boldsymbol{\beta}}_{\text{FGMM}}\right) = n(\mathbf{X}^{\top}\mathbf{W}\hat{\mathbf{\Sigma}}^{-1}\mathbf{W}^{\top}\mathbf{X})^{-1}. \tag{44}$$

The factor of n here is needed to offset the factor of n^{-1} in the definition of $\hat{\mathbf{\Sigma}}$. As usual, the covariance matrix estimator can be used to construct pseudo-t tests and other Wald tests, and asymptotic confidence intervals and confidence regions may also be based on it. The GMM criterion function that corresponds to (43) is:

$$\frac{1}{n}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\top}\mathbf{W}\hat{\mathbf{\Sigma}}^{-1}\mathbf{W}^{\top}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \tag{45}$$

The feasible efficient GMM (FGMM) estimator can be used even when all the columns of $\mathbf{X}$ are valid instruments and OLS would be the estimator of choice if the error terms are not heteroskedastic nor serially correlated. In this case, $\mathbf{W}$ typically consists of $\mathbf{X}$ augmented by a number of functions of the columns of $\mathbf{X}$, such as squares and cross-products, and $\hat{\mathbf{\Omega}}$ has squared OLS residuals on the diagonal. This estimator is asymptotically more efficient than OLS whenever $\mathbf{\Omega}$ is not proportional to an identity matrix.

4 Tests based on the GMM criterion function

For models estimated by IV, we saw that any set of r equality restrictions can be tested by taking the difference between the minimised values of the IV criterion function for the restricted and unrestricted models, and then dividing it by a consistent estimate of the error variance. The resulting test statistic was asymptotically distributed as $\chi^2(r)$. For models estimated by FGMM, a very similar testing procedure is available.

In this case, the difference between the constrained and unconstrained minima of the GMM criterion function is asymptotically distributed as $\chi^2(r)$. There is no need to divide by an estimate of σ^2, because the GMM criterion function already takes account of the covariance matrix of the error terms.

4.1 Tests of overidentifying restrictions

Whenever $l > k$, a model estimated by GMM involves $l - k$ overidentifying restrictions. As i nthe IV case, tests of these restrictions are even easier to perform than tests of other restrictions, because the minimised value of the optimal GMM criterion function (13), with $n^{-1}\mathbf{W}^\top\boldsymbol{\Omega}_0\mathbf{W}$ replaced by a HAC estimate, provides an asymptotically valid test statistic:

$$Q(\boldsymbol{\beta}, \mathbf{y}) \equiv (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top \mathbf{W}(\mathbf{W}^\top\hat{\boldsymbol{\Omega}}\mathbf{W})^{-1}\mathbf{W}^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \tag{46}$$

Since HAC estimators are consistent, the asymptotic distribution of this test statistic is the same whether we use the unknown true $\boldsymbol{\Omega}_0$ or a matrix $\hat{\boldsymbol{\Omega}}$ that provides a HAC estimate. For simplicity, we therefore use the true $\boldsymbol{\Omega}_0$ omitting the subscript 0 for ease of notation. The asymptotic equivalence of $\hat{\boldsymbol{\beta}}_{\text{FGMM}}$ of (18) or (43) and $\hat{\boldsymbol{\beta}}_{\text{GMM}}$ of (11) further implies that:

$$Q(\hat{\boldsymbol{\beta}}_{\text{GMM}}, \mathbf{y}; \boldsymbol{\Omega}) \Leftrightarrow Q(\hat{\boldsymbol{\beta}}_{\text{FGMM}}, \mathbf{y}; \hat{\boldsymbol{\Omega}}). \tag{47}$$

We remarked in Section 2 that $Q(\boldsymbol{\beta}_0, \mathbf{y})$ is asymptotically distributed as $\chi^2(l)$. In contrast, the minimised criterion function $Q(\hat{\boldsymbol{\beta}}_{\text{GMM}}, \mathbf{y})$ is distributed as $\chi^2(l-k)$, because we lose k degrees of freedom as a consequence of having estimated k parameters.

To demonstrate, first define a possibly triangular matrix $\boldsymbol{\Psi}$ that satisfies $\boldsymbol{\Omega}^{-1} = \boldsymbol{\Psi}\boldsymbol{\Psi}^\top$, or, equivalently:

$$\boldsymbol{\Omega} = (\boldsymbol{\Psi}^\top)^{-1}\boldsymbol{\Psi}^{-1}, \tag{48}$$

and then write an orthogonal projection matrix,

$$\mathbf{P}_\mathbf{A} \equiv \mathbf{A}(\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top,$$

where $\mathbf{A}$ is an $n \times l$ matrix defined as $\boldsymbol{\Psi}^{-1}\mathbf{W}$. Then we can write (46) in terms of the orthogonal projection matrix we just defined:

$$\begin{aligned} Q(\boldsymbol{\beta}, \mathbf{y}) &= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top\boldsymbol{\Psi}\boldsymbol{\Psi}^{-1}\mathbf{W}(\mathbf{W}(\boldsymbol{\Psi}^\top)^{-1}\boldsymbol{\Psi}^{-1}\mathbf{W})^{-1}\mathbf{W}^\top(\boldsymbol{\Psi}^\top)^{-1}\boldsymbol{\Psi}^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top\boldsymbol{\Psi}\mathbf{P}_\mathbf{A}\boldsymbol{\Psi}^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \end{aligned} \tag{49}$$

Since $\hat{\boldsymbol{\beta}}_{\text{GMM}}$ minimises $Q(\boldsymbol{\beta}, \mathbf{y})$, we see that one way to write it is:

$$\hat{\boldsymbol{\beta}}_{\text{GMM}} = (\mathbf{X}^\top\boldsymbol{\Psi}\mathbf{P}_\mathbf{A}\boldsymbol{\Psi}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\Psi}\mathbf{P}_\mathbf{A}\boldsymbol{\Psi}^\top\mathbf{y}. \tag{50}$$

Compared to (11), this expression makes it clear that $\hat{\boldsymbol{\beta}}_{\text{GMM}}$ can be thought of as a GIV estimator for the regression of $\boldsymbol{\Psi}^\top\mathbf{y}$ on $\boldsymbol{\Psi}^\top\mathbf{X}$ using instruments $\mathbf{A} \equiv \boldsymbol{\Psi}^{-1}\mathbf{W}$. One can show that:

$$\mathbf{P}_\mathbf{A}\boldsymbol{\Psi}^\top(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\text{GMM}}) = \mathbf{P}_\mathbf{A}(\mathbf{I} - \mathbf{P}_{\mathbf{P}_\mathbf{A}\boldsymbol{\Psi}^\top\mathbf{X}})\boldsymbol{\Psi}^\top\mathbf{y},$$

where $\mathbf{P_A\Psi^\top X}$ is the orthogonal projection on to the subspace $\mathcal{S}(\mathbf{P_A\Psi^\top X})$. It follows that

$$Q(\hat{\boldsymbol{\beta}}_{\text{GMM}}, \mathbf{y}) = \mathbf{y}^\top \boldsymbol{\Psi}(\mathbf{P_A} - \mathbf{P_{A\Psi^\top X}})\boldsymbol{\Psi}^\top \mathbf{y}, \tag{51}$$

which is the analogue for GMM estimation for the generalised IV case.[5]

Now, notice that:

$$\begin{aligned}
&(\mathbf{P_A} - \mathbf{P_{A\Psi^\top X}})\boldsymbol{\Psi}^\top \mathbf{X} \\
&= \mathbf{P_A\Psi^\top X} - \mathbf{P_A\Psi^\top X}(\mathbf{X^\top \Psi P_A \Psi^\top X})^{-1}\mathbf{X^\top \Psi P_A \Psi^\top X} \\
&= \mathbf{P_A\Psi^\top} - \mathbf{X} - \mathbf{P_A\Psi^\top X} = \mathbf{O}.
\end{aligned}$$

Since $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_0 + \mathbf{u}$ if the model we are estimating is correctly specified, this implies that (51) is equal to:

$$Q(\hat{\boldsymbol{\beta}}_{\text{GMM}}, \mathbf{y}) = \mathbf{u}^\top \boldsymbol{\Psi}(\mathbf{P_A} - \mathbf{P_{A\Psi^\top X}})\boldsymbol{\Psi}^\top \mathbf{u}. \tag{52}$$

This expression can be compared with the value of the criterion function evaluated at $\boldsymbol{\beta}_0$, which can be obtained directly from (49):

$$Q(\boldsymbol{\beta}_0, \mathbf{y}) = \mathbf{u}^\top \boldsymbol{\Psi}\mathbf{P_A}\boldsymbol{\Psi}^\top \mathbf{u}. \tag{53}$$

We can see why we lose k degrees of freedom when we look at (52) and (53): they are lost when we estimate $\boldsymbol{\beta}$. We know that $\mathbb{E}[\boldsymbol{\Psi}^\top \mathbf{u}] = \mathbf{0}$ and that $\mathbb{E}[\boldsymbol{\Psi}^\top \mathbf{u}\mathbf{u}^\top \boldsymbol{\Psi}] = \boldsymbol{\Psi}^\top \boldsymbol{\Omega}\boldsymbol{\Psi}^\top = \mathbf{I}$, by (48). The dimension of the space $\mathcal{S}(\mathbf{A})$ is equal to l. Therefore, we can conclude that (53) is asymptotically distributed as $\chi^2(l)$. Since $\mathcal{S}(\mathbf{P_A\Psi^\top X})$ is a k-dimensional subspace of $\mathcal{S}(A)$, it follows that $\mathbf{P_A} - \mathbf{P_{P_A\Psi^\top X}}$ is an orthogonal projection on to a space of dimension $l - k$, from whcih we can see that (52) is asymptotically distributed as $\chi^2(l-k)$. Replacing $\boldsymbol{\beta}_0$ by $\hat{\boldsymbol{\beta}}_{\text{GMM}}$ in (52) thus leads to the loss of the k dimensions of the space $\mathcal{S}(\mathbf{P_A\Psi^\top X})$, which are "used up" when we obtain $\hat{\boldsymbol{\beta}}_{\text{GMM}}$.

The statistic $Q(\hat{\boldsymbol{\beta}}_{\text{GMM}}, \mathbf{y})$ is the analogue, for efficient GMM estimation, of the Sargan test statistic that was discussed when we looked IV estimation. It is often called Hansen's overidentification statistic or Hansen's J statistic. Davidson and MacKinnon (2004) call it the Hansen-Sargan statistic.

As in the case of IV estimation, a Hansen-Sargan test may reject the null hypothesis for more than one reason. Perhaps the model is misspecified, either because one or more of the instruments should have been included among the regressors, or for some other reason. Perhaps one or more of the instruments is invalid because it is correlated with the error terms. Or perhaps the finite sample distribution of the test statistic just happens to differ substantially from its asymptotic distribution – which can happen in the case of feasible GMM estimation, especially involving HAC covariance matrices.

[5] Recall that that generalised IV criterion function (at the unrestricted estimates) is:

$$Q(\hat{\boldsymbol{\beta}}, \mathbf{y}) = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^\top \mathbf{P_W}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}),$$

which is equal to:

$$\begin{aligned}
&\mathbf{y}^\top(\mathbf{I} - \mathbf{P_W X}(\mathbf{X^\top P_W X})^{-1}\mathbf{X}^\top)\mathbf{P_W}(\mathbf{I} - \mathbf{X})(\mathbf{I} - \mathbf{X}(\mathbf{X^\top P_W X})^{-1}\mathbf{X^\top P_W})\mathbf{y} \\
&= \mathbf{y}^\top(\mathbf{P_W} - \mathbf{P_W X}(\mathbf{X^\top P_W X})^{-1}\mathbf{X^\top P_W})\mathbf{y} \\
&= \mathbf{y}^\top(\mathbf{P_W} - \mathbf{P_{P_W X}})\mathbf{y}.
\end{aligned}$$

4.2 Tests of linear regression

Consider the following model:

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{u}, \quad \mathbb{E}[\mathbf{u}\mathbf{u}^\top] = \boldsymbol{\Omega}, \tag{54}$$

where $\boldsymbol{\beta}_1$ is a k_1-vector and $\boldsymbol{\beta}_2$ is a k_2-vector, with $k = k_1 + k_2$. We wish to test the restrictions $\boldsymbol{\beta}_2 = \mathbf{0}$.

If we estimate (54) by feasible efficient GMM using $\mathbf{W}$ as the matrix of instruments, subject to the restriction that $\boldsymbol{\beta}_2 = \mathbf{0}$, we obtain the restricted estimates $\tilde{\boldsymbol{\beta}}_{\text{FGMM}} = \begin{bmatrix}\tilde{\boldsymbol{\beta}}_1 & \mathbf{0}\end{bmatrix}$. By the reasoning that leads to (52), we see that, if indeed $\boldsymbol{\beta}_2 = \mathbf{0}$, the constrained minimum of the criterion function is:

$$\begin{aligned} Q(\tilde{\boldsymbol{\beta}}_{\text{FGMM}}, \mathbf{y}) &= (\mathbf{y} - \mathbf{X}_1\tilde{\boldsymbol{\beta}}_1)^\top \mathbf{W}(\mathbf{W}^\top\hat{\boldsymbol{\Omega}}\mathbf{W})^{-1}\mathbf{W}^\top(\mathbf{y} - \mathbf{X}_1\tilde{\boldsymbol{\beta}}_1) \\ &= \mathbf{u}^\top\boldsymbol{\Psi}(\mathbf{P}_{\mathbf{A}} - \mathbf{P}_{\mathbf{P}_{\mathbf{A}}\boldsymbol{\Psi}^\top\mathbf{X}_1})\boldsymbol{\Psi}^\top\mathbf{u}. \end{aligned} \tag{55}$$

If we subtract (52) from (55), we find that the difference between the constrained and unconstrained minima of the criterion function is:

$$Q(\tilde{\boldsymbol{\beta}}_{\text{FGMM}}, \mathbf{y}) - Q(\hat{\boldsymbol{\beta}}_{\text{FGMM}}, \mathbf{y}) = \mathbf{u}^\top\boldsymbol{\Psi}(\mathbf{P}_{\mathbf{P}_{\mathbf{A}}\boldsymbol{\Psi}^\top\mathbf{X}} - \mathbf{P}_{\mathbf{P}_{\mathbf{A}}\boldsymbol{\Psi}^\top\mathbf{X}_1})\boldsymbol{\Psi}^\top\mathbf{u}. \tag{56}$$

Since $\mathcal{S}(\mathbf{P}_{\mathbf{A}}\boldsymbol{\Psi}^\top\mathbf{X}_1) \subseteq \mathcal{S}(\mathbf{P}_{\mathbf{A}}\boldsymbol{\Psi}^\top\mathbf{X})$, we see that $\mathbf{P}_{\mathbf{P}_{\mathbf{A}}\boldsymbol{\Psi}^\top\mathbf{X}} - \mathbf{P}_{\mathbf{P}_{\mathbf{A}}\boldsymbol{\Psi}^\top\mathbf{X}_1}$ is an orthogonal projection matrix of which the image is of dimension $k - k_1 = k_2$. Once again, we can say that the test statistic (56) is asymptotically distributed as $\chi^2(k_2)$ if the null hypothesis that $\boldsymbol{\beta}_2 = \mathbf{0}$ is true.

One interesting consequence of the form (56) is that we do not always need to bother estimating the unrestricted model. The test statistic (56) must always be less than the constrained minimum $Q(\tilde{\boldsymbol{\beta}}_{\text{FGMM}}, \mathbf{y})$. Therefore, if $Q(\tilde{\boldsymbol{\beta}}_{\text{FGMM}}, \mathbf{y})$ is less than the critical value for the $\chi^2(k_2)$ distribution at our chosen significance level, we can be sure that the actual test statistic is even smaller and would not lead us to reject the null.

5 The simulated method of moments

In this section we move onto the simulated method of moments, which will also involve estimation of nonlinear functions. The method of estimating functions employs the concept of an elementary zero function. Such a function plays the same role as a residual in an estimation of a regression model. It depends on observed variables, at least one fo which must be endogenous, and on a k-vector of parameters, $\boldsymbol{\theta}$. As with a residual, the expectation of an elementary zero function must vanish if it is evaluated at the true value of $\boldsymbol{\theta}$, but not in general otherwise. Suppose we want to use GMM to estimate elementary zero functions, but that they cannot be estimated analytically. Suppose they take the form

$$f_t(y_t, \boldsymbol{\theta}) = h_t(y_t) - m_t(\boldsymbol{\theta}), \quad t = 1, ..., n, \tag{57}$$

where the function $h_t(y_t)$ depends only on y_t and, possibly, on exogenous or predetermined variables. The function $m_t(\boldsymbol{\theta})$ depends only on exogenous or predetermined variables and on the parameters. Like a regression function, it is the expectation of $h_t(y_t)$, conditional on the information set Ω_t, under a DGP characterised by the parameter vector $\boldsymbol{\theta}$. Estimating such a model by GMM presents no special difficulty if the form of $m_t(\boldsymbol{\theta})$ is known analytically, but this need not be the case.

There are numerous situations in which $m_t(\boldsymbol{\theta})$ may not be known analytically. In particular, it may well occur in models which latent variables, that is, variables which are not observables by an econometrician. The variables that actually are observed are related to the latent variables in such a way that knowing the former does not permit the values of the latter to be fully recovered. One example is economic variables that censored, in the sense that they are observed only to a limited extent, for instance when only the sign of the variable is observed, or when all negative values are replaced by zeros. Even if the distributions of the latent variables are tractable, those of the observed variables may not be. In particular, it may not be possible to obtain analytic expressions for their expectations, or for the expectations of functions of them.

Even when analytic expressions are not available, it is often possible to obtain simulation-based estimates of the distributions of the observed variables. For example, suppose that an observed variable is equal to a latent variable plus a measurement error of some known distribution, possibly dependent on the parameter $\boldsymbol{\theta}$. Suppose further that, for a DGP characterised by $\boldsymbol{\theta}$, we can readily generate simulated values of the latent variable. Simulated values of the observed variable can then be generated by adding simulated measurement errors, drawn for their known distribution, to the simulated values of the latent variable. The mean of these drawings then provides an estimate of the expectation of the observed variable.

In general, an unbiased simulator for the unknown expectation $m_t(\boldsymbol{\theta})$ is any function is any function $m_t^*(u_t^*, \boldsymbol{\theta})$ of the model parameters, variables in Ω_t, and a random variable u_t^*, which either has a known distribution or can be simulated, such that, for all $\boldsymbol{\theta}$ in the parameter space, $\mathbb{E}[m_t^*(u_t^*, \boldsymbol{\theta})] = m_t(\boldsymbol{\theta})$. To simplify notation, we write u_t^* as a scalar random variable, but it may well be a vector of random variables in practical situations of interest.

The conceptually simplest unbiased simulator can be implemented as follows. For given $\boldsymbol{\theta}$, we obtain S simulated y_{ts}^* of the observed variable under the DGP characterised by $\boldsymbol{\theta}$, making use of S random numbers u_{ts}^*. Then we let $m_t^*(u_{ts}^*, \boldsymbol{\theta}) = h_t(y_{ts}^*)$. If (57) is indeed a zero function, then $h_t(y_{ts}^*)$ must have expectation $m_t(\boldsymbol{\theta})$, and it is obvious that the sample mean of the simulated values $h(y_{ts}^*)$ is a simulation-based estimate of that expectation.[6]

If an unbiased simulator is available, the elementary zero functions (57) can be replaced by the functions:

$$f_t^*(y_t, \boldsymbol{\theta}) = h_t(y_t) - \frac{1}{S}\sum_{s=1}^{S} m_t^*(u_{ts}^*, \boldsymbol{\theta}), \tag{58}$$

where the $u_{ts}^*, t = 1, ..., n, s = 1, ..., S$, are mutually independent draws. Since these draws are computer generated, they are evidently independent of the y_t. The functions (58) are legitimate elementary zero functions, even in the trivial case in which $S = 1$. If the true DGP is characterised by $\boldsymbol{\theta}_0$, then $\mathbb{E}[h_t(y_t)] = m_t(\boldsymbol{\theta}_0)$ by definition, and $\mathbb{E}[m_t^*(u_{ts}^*, \boldsymbol{\theta}_0)] = m_t(\boldsymbol{\theta}_0)$ for all s by construction. It follows that the expectation (58) is zero for $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, but not in general for other values of $\boldsymbol{\theta}$.

The application of GMM to the zero functions (58) is called the simulated method of moments (SMM) or method of simulated moments (MSM). We can use an $n \times l$ matrix $\mathbf{W}$ of appropriate instruments, with $l \geq k$, in order to form the empirical moments

$$\mathbf{W}^\top \mathbf{f}^*(\boldsymbol{\theta}), \tag{59}$$

[6] This simple estimator, is not the only possible simulator, and it may not be the most desirable one for some purposes. However, we will not more complicated simulators in this book.

in which the n-vector of functions $\mathbf{f}^*(\boldsymbol{\theta})$ has typical element $f_t^*(y_t, \boldsymbol{\theta})$. A GMM estimator that is efficient relative to this set of empirical moments may be obtained by minimising the quadratic form:

$$Q(\boldsymbol{\theta}, \mathbf{y}) \equiv \frac{1}{n}\mathbf{f}^*(\boldsymbol{\theta})^\top \mathbf{W}\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{W}^\top \mathbf{f}^*(\boldsymbol{\theta}), \tag{60}$$

with respect to $\boldsymbol{\theta}$, where $\hat{\boldsymbol{\Sigma}}$ consistently estimates the covariance matrix of $n^{-1/2}\mathbf{W}^\top \mathbf{f}^*(\boldsymbol{\theta})$.

Minimising (60) the criterion function with respect to $\boldsymbol{\theta}$ proceeds in the usual way, with one important proviso. Each evaluation of $\mathbf{f}^*(\boldsymbol{\theta})$ requires a large number of pseudo-random numbers (generally, at least nS of them). It is absolutely essential that the same set of random numbers be used every time $\mathbf{f}^*(\boldsymbol{\theta})$ is evaluated for a new value of the parameter vector $\boldsymbol{\theta}$. Otherwise, the criterion function would change not only as a result of changes in $\boldsymbol{\theta}$ but also as a result of changes in the random numbers used for the simulation. Therefore, if the algorithm happened to evaluate the criterion function twice at the same parameter vector, it would obtain two different values of $Q(\boldsymbol{\theta}, \mathbf{y})$, and it could not possibly tell where the minimum was located.

5.1 Asymptotic distribution of the SMM estimator

because the criterion function (60) is based on genuine zero functions, the estimator $\hat{\boldsymbol{\theta}}_{\text{SMM}}$ obtained by minimising it is consistent whenever the parameters are identified.

The first-order conditions for minimising (60), ignoring a factor of $2/n$, are

$$\mathbf{F}^*(\boldsymbol{\theta})^\top \mathbf{W}\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{W}^\top \mathbf{f}^*(\boldsymbol{\theta}) = \mathbf{0}, \tag{61}$$

where $\mathbf{F}^*(\boldsymbol{\theta})$ is the $n \times k$ matrix of w hich the ti-th element is $\partial f_t^*(y_t, \boldsymbol{\theta})/\partial\theta_i$. The solution to these equations is $\hat{\boldsymbol{\theta}}_{\text{SMM}}$.

From (61), it can be seen that the instruments effectively used by the SMM estimator are $\mathbf{W}\hat{\boldsymbol{\Sigma}}^{-1}(n^{-1}\mathbf{W}^\top \mathbf{F}_0^*)$, where $\mathbf{F}_0^* \equiv \mathbf{F}^*(\boldsymbol{\theta}_0)$, and a factor of n^{-1} has been used to keep the expression of order unity as $n \to \infty$. If we think of the effective instruments as $\mathbf{Z} = \mathbf{WJ}$, then $\mathbf{J} = \hat{\boldsymbol{\Sigma}}^{-1}(n^{-1}\mathbf{W}^\top \mathbf{F}_0^*)$.

The asymptotic covariance matrix of $\sqrt{n}(\hat{\boldsymbol{\theta}}_{\text{SMM}} - \boldsymbol{\theta}_0)$ can now be found by using the following general formula for the asymptotic covariance matrix of an efficient GMM estimator with unknown covariance matrix:

$$\left(\operatorname*{plim}_{n\to\infty} \frac{1}{n}\mathbf{J}^\top \mathbf{W}^\top \mathbf{F}_0\right)^{-1} \left(\operatorname*{plim}_{n\to\infty} \frac{1}{n}\mathbf{J}^\top \mathbf{W}^\top \boldsymbol{\Omega}\mathbf{WJ}\right) \left(\operatorname*{plim}_{n\to\infty} \frac{1}{n}\mathbf{F}_0^\top \mathbf{WJ}\right)^{-1}. \tag{62}$$

This is a sandwich estimator of the form $\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}$, and we find that:

$$\begin{aligned} \mathbf{A} &= \operatorname*{plim}_{n\to\infty} \left(\frac{1}{n}\mathbf{F}_0^{*\top}\mathbf{W}\right) \hat{\boldsymbol{\Sigma}}^{-1} \left(\frac{1}{n}\mathbf{W}^\top \mathbf{F}_0^*\right), \\ \mathbf{B} &= \operatorname*{plim}_{n\to\infty} \left(\frac{1}{n}\mathbf{F}_0^{*\top}\mathbf{W}\right) \hat{\boldsymbol{\Sigma}}^{-1} \left(\frac{1}{n}\mathbf{W}^\top \boldsymbol{\Omega}\mathbf{W}\right) \hat{\boldsymbol{\Sigma}}^{-1} \left(\frac{1}{n}\mathbf{W}^\top \mathbf{F}_0^{*\top}\right), \end{aligned} \tag{63}$$

where $\boldsymbol{\Omega}$ is the $n \times n$ covariance matrix of $\mathbf{f}^*(\boldsymbol{\theta}_0)$.

The ti-th element of $\mathbf{F}^*(\boldsymbol{\theta})$ is, from (58),

$$F_{ti}^*(\boldsymbol{\theta}) = -\frac{1}{S}\sum_{s=1}^{S} \frac{\partial m_t^*(u_{ts}^*, \boldsymbol{\theta})}{\partial\theta_i}.$$

If m_t^* is differentiable with respect to $\boldsymbol{\theta}$ in a neighbourhood of $\boldsymbol{\theta}$, then we can differentiate the relation $\mathbb{E}[m_t^*(u_t^*, \boldsymbol{\theta})] = m_t(\boldsymbol{\theta})$ to find that

$$\mathbb{E}\left[\frac{\partial m_t^*(u_t^*, \boldsymbol{\theta})}{\partial \theta_i}\right] = \frac{\partial m_t(\boldsymbol{\theta})}{\partial \theta_i}.$$

We denote by $\mathbf{M}(\boldsymbol{\theta})$ the $n \times k$ matrix with typical element $\partial m_t / \partial \theta_i(\boldsymbol{\theta})$. By a LLN, we then see that

$$\operatorname*{plim}_{n\to\infty} \frac{1}{n}\mathbf{W}^\top \mathbf{M}_0 = \operatorname*{plim}_{n\to\infty} \frac{1}{n}\mathbf{W}^\top \mathbf{M}_0,$$

where $\mathbf{M}_0 \equiv \mathbf{M}(\boldsymbol{\theta}_0)$.

Consider next the covariance matrix $\boldsymbol{\Omega}$ of $\mathbf{f}^*(\boldsymbol{\theta}_0)$. The original data y_t are of course completely independent of the simulated u_{ts}^*, and the simulated data are independent across simulations. Thus, from (58), we see that

$$\boldsymbol{\Omega} = \operatorname{Var}(\mathbf{h}(\mathbf{y})) + \frac{1}{S}\operatorname{Var}(\mathbf{m}^*(\boldsymbol{\theta}_0)), \tag{64}$$

where $\mathbf{h}(\mathbf{y})$ and $m^*(\boldsymbol{\theta})$ are the n-vectors with typical elements $h_t(y_t)$ and $m_t^*(u_t^*, \boldsymbol{\theta})$, respectively. We see that the covariance matrix $\boldsymbol{\Omega}$ has two components, one due to the randomness of the data and the other due to the randomness of the simulations. If the simulator $m_t^*(\cdot)$ is the simple one suggested above, then the simulated data $h_t(y_t^*)$ are generated from the DGP characterised by $\boldsymbol{\theta}$, which is also supposed to have generated the real data. Therefore, it is clear that $\operatorname{Var}(\mathbf{h}(\mathbf{y})) = \operatorname{Var}(\mathbf{m}^*(\boldsymbol{\theta}_0))$, and we conclude that $\boldsymbol{\Omega} = (1 + 1/S)\operatorname{Var}(\mathbf{h}(\mathbf{y}))$.

In general, the $n \times n$ matrix $\boldsymbol{\Omega}$ cannot be estimated consistently, but an HCCME or HAC estimator can be used to provide a consistent estimate of $\boldsymbol{\Sigma}$, the covariance matrix of $n^{-1/2}\mathbf{W}^\top \mathbf{f}^*(\boldsymbol{\theta}_0)$. For the simple simulator we have been discussing, $\hat{\boldsymbol{\Sigma}}$ is just $1 + 1/S$ times whatever HAC estimator of HCCME would be appropriate if there were no simulation involved. For other simulators, it may be a little harder to estimate (64). In any case, once $\hat{\boldsymbol{\Sigma}}$ is available, we use it to replace $n^{-1}\mathbf{W}^\top \boldsymbol{\Omega}\mathbf{W}$ in (63). We also replace $\operatorname{plim} n^{-1}\mathbf{W}^\top \mathbf{F}_0^*$ by $\operatorname{plim} n^{-1}\mathbf{W}^\top \mathbf{M}_0$. The sandwich estimator for the asymptotic covariance matrix then simplifies greatly, and we find that the asymptotic covariance matrix is just:

$$\operatorname*{plim}_{n\to\infty} \left(\frac{1}{n}\mathbf{M}_0^\top \mathbf{W}\hat{\boldsymbol{\Sigma}}^{-1}\frac{1}{n}\mathbf{W}^\top \mathbf{M}_0\right)^{-1}.$$

In practice, $\mathbf{M}_0$ can be estimated using either analytical or numerical derivatives of

$$\frac{1}{S}\sum_{s=1}^{S} m_t^*(u_{ts}^*, \hat{\boldsymbol{\theta}})\bigg|_{\hat{\boldsymbol{\theta}}_{\text{SMM}}}.$$

However, for this to be a reliable estimator, it is necessary for S to be reasonably large. If we let $\hat{\mathbf{M}}$ denote the estimate of $\mathbf{M}_0$, then in practice we use

$$\widehat{\operatorname{Var}}\left(\hat{\boldsymbol{\theta}}_{\text{SMM}}\right) = n\left(\hat{\mathbf{M}}^\top \mathbf{W}\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{W}^\top \hat{\mathbf{M}}\right)^{-1}. \tag{65}$$

This has the same form for the estimated covariance matrices for feasible efficient GMM estimators of linear regression and general nonlinear models, such as (44). The most important new feature of (65) is the factor of $1 + 1/S$, which is buried in $\hat{\boldsymbol{\Sigma}}$.

5.2 Example: The log-normal distribution

We now turn to a simple example (which actually does not require simulation at all, actually), where we demonstrate how GMM can be used to match moments of distributions. Moment matching can be done quite easily when the moments to be matched can be expressed analytically as functions of the parameters to be estimated, and no simulation is needed in such cases. If analytic expressions are not available, moment matching can still be done whenever we can simulate the random variables of which the expectations are the moments to be matched.

A random variable is said to follow the log-normal distribution if its logarithm is normally distributed. The log-normal distribution for a scalar random variable y thus depends on just two parameters, the expectation and the variance of $\log y$. Formally, if $z \sim N(\mu, \sigma^2)$, then the variable $y \equiv \exp(z)$ is log-normally distributed, with a distribution characterised by μ and σ^2.

Suppose we have an n-vector $\mathbf{y}$, of which the components y_t are IID, each log-normally distributed with unknown parameters μ and σ^2. The "right" way to estimate these unknown parameters is to take logs of each component of $\mathbf{y}$, thus obtaining an n-vector $\mathbf{z}$ with typical element z_t, and then to estimate μ and σ^2 by the sample mean and sample variance of the z_t. This can be done by regressing $\mathbf{z}$ on a constant.

The above estimation method implicitly matches the first and second moments of the log of y_t in order to estimate the parameters. It yields the parameter values that give theoretical moments equal to the corresponding moments in the sample. Since we have two parameters to estimate, we need at least two moments. But other sets of two moments could also be used in order to obtain MM estimators of μ and σ^2. So could sets of more than two moments, although the match could not be perfect, because there would implicitly be overidentifying restrictions.

We now consider precisely how we might estimate μ and σ^2 by matching the first moment of the y_t along with the first moment of the z_t. With this choice, it is once more possible to obtain an analytical answer, because, the expectation of y_t is $\exp(\mu + \frac{1}{2}\sigma^2)$. Thus, as before, we esimtate μ by using $\bar{z}$, the sample mean of the z_t, and then estimate σ^2 by solving the equation:

$$\log \bar{y} = \bar{z} + \frac{1}{2}\hat{\sigma}^2,$$

for $\hat{\sigma}^2$, where $\bar{y}$ is the sample mean of the y_t. The estimate is

$$\hat{\sigma}^2 = 2(\log \bar{y} - \bar{z}). \tag{66}$$

This estimate is not, generally, numerically equal to the estimate obtained by regressing $\mathbf{z}$ on a constant, and in fact it has a higher variance.

Let us formalise the estimation procedure described above in terms of zero functions and GMM. The moments used are the first moments of the y_t and the z_t, for $t = 1, ..., n$. For each observation, then, there are two elementary zero functions, which serve to express the expectations of the y_t and the z_t in terms of the parameters μ and σ^2. We write these elementary zero functions as follows:

$$\begin{aligned} f_{t1}(z_t, \mu, \sigma^2) &= z_t - \mu, \\ f_{t2}(y_t, \mu, \sigma^2) &= y_t - \exp\left(\mu + \frac{1}{2}\sigma^2\right). \end{aligned} \tag{67}$$

The derivatives of these functions with respect to the parameters are:

$$\begin{aligned} \frac{\partial f_{t1}}{\partial \mu} &= -1; \quad \frac{\partial f_{t1}}{\partial \sigma^2} = 0; \\ \frac{\partial f_{t2}}{\partial \mu} &= -\exp\left(\mu + \frac{1}{2}\sigma^2\right); \quad \frac{\partial f_{t2}}{\partial \sigma^2} = -\frac{1}{2}\exp\left(\mu + \frac{1}{2}\sigma^2\right). \end{aligned} \tag{68}$$

These derivatives, which are all deterministic, allow us to find the optimal instruments for the estimation of μ and σ^2 on the basis of the zero functions (67), provided that we can also obtain the covariance matrix $\mathbf{\Omega}$ of the zero functions.

Notice that here we have two elementary zero functions and no instruments. Nevertheless, we can set the problem up so that it looks like a standard one. Let $\mathbf{f}_1(\mu, \sigma^2)$ and $\mathbf{f}_2(\mu, \sigma^2)$ be two n-vectors with typical components $f_{t1}(z_t, \mu, \sigma^2)$ and $f_{t2}(y_t, \mu, \sigma^2)$, respectively. For notational simplicity, we suppress the explicit dependence of these vectors on the y_t and z_t. The $2n$-vector $\mathbf{f}(\mu, \sigma^2)$ of the full set of elementary zero functions, and the $2n \times 2n$ matrix $\mathbf{F}(\mu, \sigma^2)$ of the derivatives with respect to the parameters, can thus be written as

$$\mathbf{f}(\mu, \sigma^2) = \begin{bmatrix} \mathbf{f}_1(\mu, \sigma^2) \\ \mathbf{f}_2(\mu, \sigma^2) \end{bmatrix} \quad \text{and} \quad \mathbf{F}(\mu, \sigma^2) = -\begin{bmatrix} \boldsymbol{\iota} & \mathbf{0} \\ a\boldsymbol{\iota} & \frac{1}{2}a\boldsymbol{\iota} \end{bmatrix}, \tag{69}$$

where $a \equiv \exp(\mu + \frac{1}{2}\sigma^2)$. The constant vectors $\boldsymbol{\iota}$ in $\mathbf{F}(\mu, \sigma^2)$ arise because none of the derivatives in (68) depends on t, which is a consequence of the assumption that the data are IID.

Because $\mathbf{f}(\mu, \sigma^2)$ is a $2n$-vector, the covariance matrix $\mathbf{\Omega}$ is $2n \times 2n$. This matrix can be written as:

$$\mathbf{\Omega} = \mathbb{E}\left[\begin{bmatrix} \mathbf{f}_{10} \\ \mathbf{f}_{20} \end{bmatrix} \begin{bmatrix} \mathbf{f}_{10}^\top & \mathbf{f}_{20}^\top \end{bmatrix}\right],$$

where $\mathbf{f}_{i0}, i = 1, 2$ is $\mathbf{f}_i$ evaluated at the true values μ_0 and σ_0^2. Since the data are IID, $\mathbf{\Omega}$ can be partitioned as follows into four $n \times n$ blocks, each of which is proportional to an identity matrix. The result is:

$$\mathbf{\Omega} = \begin{bmatrix} \sigma_z^2\mathbf{I} & \sigma_{zy}\mathbf{I} \\ \sigma_{yz}\mathbf{I} & \sigma_y^2\mathbf{I} \end{bmatrix}, \tag{70}$$

where the coefficients of the identity matrices are the variances and covariances $\sigma_y^2 \equiv \text{Var}(y_t)$, $\sigma_z^2 \equiv \text{Var}(z_t)$, and $\sigma_{yz} = \sigma_{zy} \equiv \text{Cov}(y_t, z_t)$.

We now have everything we need to set up the efficient estimating equations as in (61):

$$\mathbf{F}^\top(\mu, \sigma^2)\mathbf{\Omega}^{-1}\mathbf{f}(\mu, \sigma^2) = \mathbf{0}, \tag{71}$$

where $\mathbf{f}(\cdot)$ and $\mathbf{F}(\cdot)$ are given by (69), and $\mathbf{\Omega}$ is given by (70). By explicitly performing the multiplications of partitioned matrices in (71), inverting $\mathbf{\Omega}$, and ignoring irrelevant scalar factors, we obtain:

$$\begin{bmatrix} \sigma_y^2 - a\sigma_{yz} & a\sigma_z^2 - \sigma_{zy} \\ -\frac{1}{2}a\sigma_{yz} & \frac{1}{2}a\sigma_z^2 \end{bmatrix} \begin{bmatrix} \boldsymbol{\iota}^\top & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\iota}^\top \end{bmatrix} \begin{bmatrix} \mathbf{f}_1(\mu, \sigma^2) \\ \mathbf{f}_2(\mu, \sigma^2) \end{bmatrix} = \mathbf{0}.$$

Since the leftmost factor above is a 2×2 nonsingular matrix, we see that these estimating equations are equivalent to:

$$\begin{aligned} \boldsymbol{\iota}^\top \mathbf{f}_1(\mu, \sigma^2) &= \mathbf{0} \\ \boldsymbol{\iota}^\top \mathbf{f}_2(\mu, \sigma^2) &= \mathbf{0}. \end{aligned} \tag{72}$$

The solution to these two equations is $\hat{\mu} = \bar{z}$ and $\hat{\sigma}^2$. given by (66).

Suppose we want to look at more than two moments. Consider the case where we want to match the first moments of the z_t and the y_t, but also the second moment of the y_t, or, equivalently, the first moment of the y_t^2. Since the log of y_t^2 is just $2z_t$, which is distributed as $N(2\mu, 4\sigma^2)$, the expectation of y_t^2 is $\exp(2(\mu + \sigma^2))$. We now have three elementary zero functions for each observation:

$$\begin{aligned} f_{t1}(z_t, \mu, \sigma^2) &= z_t - \mu, \\ f_{t2}(y_t, \mu, \sigma^2) &= y_t - \exp\left(\mu + \frac{1}{2}\sigma^2\right), \\ f_{t3}(y_t, \mu, \sigma^2) &= y_t^2 - \exp(2(\mu + \sigma^2)). \end{aligned}$$

The vector $\mathbf{f}(\cdot)$ and the matrix $\mathbf{F}(\cdot)$ now both have $3n$ rows. The latter still has two columns, both of which can be paritioned into three n-vectors, each proportional to $\boldsymbol{\iota}$. Further, the matrix $\boldsymbol{\Omega}$ grows to become $3n \times 3n$. It is then a matter of taste whether to set up a just identified estimation problem using as optimal instruments the two columns of $\boldsymbol{\Omega}^{-1}\mathbf{F}(\mu, \sigma^2)$, or to use three instruments, which are the columns of the matrix:

$$\mathbf{W} \equiv \begin{bmatrix} \boldsymbol{\iota} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\iota} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{\iota} \end{bmatrix}, \tag{73}$$

and to construct an optimal weighting matrix. Whichever choice is made, it is necessary to estimate $\boldsymbol{\Omega}$ in order to construct the optimal instruments for the first method, or the optimal weighting matrix for the second.

The procedures we have just described depend on the fact that we know the analytic forms of $\mathbb{E}[z_t]$, $\mathbb{E}[y_t]$, and $\mathbb{E}[y_t^2]$. In more complicated applications, comparable analytic expressions for the moments to be matched might not be available. In such cases, simulators can be used to replace such analytic expressions. We illustrate the method for the case of the log-normal distribution, matching the first moments of z_t and y_t pretending that we do not know the analytic expressions for their expectations.

For any given values of μ and σ^2, we can draw from the log-normal distribution characterised by these values by first using a random number generator to give a drawing u^* from $N(0, 1)$ and then computing $y^* = \exp(\mu + \sigma u^*)$. Thus, unbiased simulators for the expectations of $z \equiv \log y$ and of y it self are

$$\begin{aligned} m_1^*(u^*, \mu, \sigma^2) &\equiv \mu + \sigma u^*, \\ m_2^*(u^*, \mu, \sigma^2) &\equiv \exp(\mu + \sigma u^*). \end{aligned}$$

If we perform S simulations, the zero functions for SMM estimation can be written as:

$$\begin{aligned} f_{t1}^*(z_t, \mu, \sigma^2) &= z_t - \frac{1}{S}\sum_{s=1}^{S} m_1^*(u_{ts}^*, \mu, \sigma^2), \\ f_{t2}^*(y_t, \mu, \sigma^2) &= y_t - \frac{1}{S}\sum_{s=1}^{S} m_2^*(u_{ts}^*, \mu, \sigma^2), \end{aligned}$$

where the u_{ts}^* are IID standard normal. Comparison with (67) shows clearly how we replace analytic expressions for the moments, assumed to be unknown, by simulation-based estimates.

Since the data are IID, it might appear tempting to use just one set of random numbers, $u_s^*, s = 1, ..., S$, for all t. However, doing this would introduce dependence among the zero functions, greatly complicating the computation of their covariance matrix. As S becomes large, of course, the law of large numbers ensures that this effect becomes less and less important. Using just one set of random numbers would in any case not affect the consistency of the SMM estimator, merely that of the covariance matrix estimate.

By analogy with (72), we can see that the SMM estimating equations are:

$$\begin{aligned} \boldsymbol{\iota}^\top \mathbf{f}_1^*(\hat{\mu}, \hat{\sigma}^2) &= \mathbf{0}, \\ \boldsymbol{\iota}^\top \mathbf{f}_2^*(\hat{\mu}, \hat{\sigma}^2) &= \mathbf{0}. \end{aligned} \tag{74}$$

Here we have again grouped the elementary zero functions into two n-vectors $\mathbf{f}_1^*(\cdot)$ and $\mathbf{f}_2^*(\cdot)$. Recalling that the random numbers u_{ts}^* are drawn only once for the entire procedure, let us make the definitions:

$$\begin{aligned} m_{t1}(\mu, \sigma^2) &\equiv \frac{1}{S} \sum_{s=1}^{S} m_1^*(u_{ts}^*, \mu, \sigma^2) = \mu + \sigma \frac{1}{S} \sum_{s=1}^{S} u_{ts}^*, \\ m_{t2}(\mu, \sigma^2) &\equiv \frac{1}{S} \sum_{s=1}^{S} m_2^*(u_{ts}^*, \mu, \sigma^2) = \frac{1}{S} \exp(\mu + \sigma u_{ts}^*). \end{aligned} \tag{75}$$

It is clear that, as $S \to \infty$, these functions tend for all t to the limits of the expectations of z and y, respectively. It is also not hard to see that these limits are μ and $\exp(\mu + \frac{1}{2}\sigma^2)$.

On dividing by the sample size n and rearranging, the estimating equations (74) can be written as

$$\begin{aligned} \bar{m}_1(\mu, \sigma^2) &= \bar{z}, \\ \bar{m}_2(\mu, \sigma^2) &= \bar{y}, \end{aligned} \tag{76}$$

where $\bar{z}$ and $\bar{y}$ are the sample averages of the z_t and the y_t, respectively, and

$$\bar{m}_i(\mu, \sigma^2) \equiv \frac{1}{n} \sum_{t=1}^{n} m_{ti}(\mu, \sigma^2), \quad i = 1, 2.$$

Equations (76) can be solved in various ways. One approach is to turn the problem of solving them into a minimisation problem. Let

$$\mathbf{W} = \begin{bmatrix} \boldsymbol{\iota} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\iota} \end{bmatrix}. \tag{77}$$

Then it is not difficult to see that minimising the quadratic form

$$\begin{bmatrix} \mathbf{z} - \mathbf{m}_1(\mu, \sigma^2) \\ \mathbf{y} - \mathbf{m}_2(\mu, \sigma^2) \end{bmatrix}^\top \mathbf{W}\mathbf{W}^\top \begin{bmatrix} \mathbf{z} - \mathbf{m}_1(\mu, \sigma^2) \\ \mathbf{y} - \mathbf{m}_2(\mu, \sigma^2) \end{bmatrix}, \tag{78}$$

also solves equations (76). Here the n-vectors $\mathbf{m}_1(\cdot)$ and $\mathbf{m}_2(\cdot)$ have typical elements $m_{t1}(\cdot)$ and $m_{t2}(\cdot)$, respectively.

5.2.1 Newton's Method

Alternatively, we can use Newton's Method directly. Suppose that we wish to solve a set of k equations of the form $\mathbf{g}(\boldsymbol{\theta}) = \mathbf{0}$ for a k-vector of unknowns $\boldsymbol{\theta}$, where $\mathbf{g}(\cdot)$ is also a k-vector. The iterative step is:

$$\boldsymbol{\theta}_{j+1} = \boldsymbol{\theta}_j - \mathbf{G}^{-1}(\boldsymbol{\theta}_j)\mathbf{g}(\boldsymbol{\theta}_j), \tag{79}$$

where $\mathbf{G}(\boldsymbol{\theta})$ is the Jacobian matrix associated with $\mathbf{g}(\boldsymbol{\theta})$. This is $k \times k$ matrix contains the derivatives of the components of $\mathbf{g}(\boldsymbol{\theta})$ with respect to the elements of $\boldsymbol{\theta}$. For the estimating equations (76), the iterative step (79) becomes

$$\begin{bmatrix} \mu_{j+1} \\ \sigma^2_{j+1} \end{bmatrix} = \begin{bmatrix} \mu_j \\ \sigma^2_j \end{bmatrix} - \begin{bmatrix} \frac{\partial \bar{m}_1}{\partial \mu} & \frac{\partial \bar{m}_1}{\partial \sigma^2} \\ \frac{\partial \bar{m}_2}{\partial \mu} & \frac{\partial \bar{m}_2}{\partial \sigma^2} \end{bmatrix} \begin{bmatrix} \bar{m}_1(\mu_j, \sigma^2_j) - \bar{z} \\ \bar{m}_2(\mu_j, \sigma^2_j) - \bar{y} \end{bmatrix},$$

where all the partial derivatives are evaluated at (μ_j, σ^2_j).

To estimate the asymptotic covariance matrix of the SMM estimates, we can use any suitable estimator (such as (44)), so long as we multiply it by $1 + 1/S$ to account for the simulation randomness. The instrument matrix $mathbfW$ is just the matrix $\mathbf{W}$ of (77). We are pretending that we do not know the analytic form of the matrix $\mathbf{F}(\mu, \sigma^2)$ given in (69), and so instead we use the matrix of partial derivatives of $\mathbf{m}_1$ and $\mathbf{m}_2$, evaluated at $\hat{\mu}$ and $\hat{\sigma}^2$. This matrix is

$$\hat{\mathbf{F}} \equiv \begin{bmatrix} \frac{\partial \mathbf{m}_1}{\partial \mu}(\hat{\mu}, \hat{\sigma}^2) & \frac{\partial \mathbf{m}_1}{\partial \sigma^2}(\hat{\mu}, \hat{\sigma}^2) \\ \frac{\partial \mathbf{m}_2}{\partial \mu}(\hat{\mu}, \hat{\sigma}^2) & \frac{\partial \mathbf{m}_2}{\partial \sigma^2}(\hat{\mu}, \hat{\sigma}^2) \end{bmatrix}. \tag{80}$$

Note that each block of $\hat{\mathbf{F}}$ is an n-vector. If we use Newton's Method for the estimation, then all the partial derivatives in this matrix have already been computed. Finally, the covariance matrix $\boldsymbol{\Omega}$ of the elementary zero functions can be estimated using (70), by replacing the unknown quantities σ^2_z, σ^2_y, and σ_{zy} with their sample analogues. If we denote the result of this by $\hat{\boldsymbol{\Omega}}$, then our estimate of the covariance matrix of $\hat{\mu}$ and $\hat{\sigma}^2$ is:

$$\widehat{\text{Var}} \begin{bmatrix} \hat{\mu} \\ \hat{\sigma}^2 \end{bmatrix} = (\mathbf{W}^\top \hat{\mathbf{F}})^{-1} \mathbf{W}^\top \hat{\boldsymbol{\Omega}} \mathbf{W} (\hat{\mathbf{F}}^\top \mathbf{W})^{-1}, \tag{81}$$

with $\mathbf{W}$ given by (77) and $\hat{\mathbf{F}}$ given by (80).

References

Davidson, Russell and James G. MacKinnon (2004). *Econometric Theory and Methods*. Oxford University Press.

Hayashi, Fumio (2000). *Econometrics*. Princeton University Press.