Generalised Method of Moments

jessekelighine.com

December 12, 2023

Consider the normal linear regression form:

$$y_i = \begin{bmatrix} \mathbf{z}_{i1} & \cdots & \mathbf{z}_{iL} \end{bmatrix} \begin{bmatrix} \delta_1 \\ \vdots \\ \delta_L \end{bmatrix} + \varepsilon_i \quad \rightsquigarrow \quad y_i = \mathbf{z}_i'\delta + \varepsilon_i \tag{1}$$

where y_i is the dependent variable, $\mathbf{z}_i$ is the independent variable vector, δ is the parameter vector. We are interested in the parameters δ.

If $\mathbf{z}_i$ is exogenous, then we can simply use the OLS estimator. However, if $\mathbf{z}_i$ is not exogenous, then *instrument variables* can be used. Let $\mathbf{x}_i$ be a vector of valid instruments (relevant and exogenous) with dimension $K \times 1$. Consider the following:

$$y_i = \mathbf{z}_i'\delta + \varepsilon_i \implies \mathbf{x}_i y_i = \mathbf{x}_i\mathbf{z}_i'\delta + \mathbf{x}_i\varepsilon_i \tag{2}$$

$$\implies \mathbb{E}[\mathbf{x}_i y_i] = \mathbb{E}[\mathbf{x}_i\mathbf{z}_i'\delta + \mathbf{x}_i\varepsilon_i] \tag{3}$$

$$(\mathbf{x}_i \text{ is exogenous}) \implies \mathbb{E}[\mathbf{x}_i y_i] = \mathbb{E}[\mathbf{x}_i\mathbf{z}_i']\,\delta \tag{4}$$

$$(\mathbf{x}_i \text{ is relevant, } \mathbb{E}[\mathbf{x}_i\mathbf{z}_i'] \text{ invertible}) \implies \mathbb{E}[\mathbf{x}_i\mathbf{z}_i']^{-1}\,\mathbb{E}[\mathbf{x}_i y_i] = \delta \tag{5}$$

Notice that were $\mathbf{z}_i$ to be exogenous, then the result would be identical to OLS by replacing $\mathbf{x}_i$ with $\mathbf{z}_i$. If we replace the *moments* in eq.(5) with the corresponding estimators, then we obtain the familiar "instrument variable estimator" estimator $\hat{\delta}_{\text{IV}} = \mathbf{S}_{\mathbf{xz}}^{-1}\mathbf{s}_{\mathbf{x}y} \xrightarrow{p} \mathbb{E}[\mathbf{x}_i\mathbf{z}_i']^{-1}\,\mathbb{E}[\mathbf{x}_i y_i]$ where

$$\mathbf{S}_{\mathbf{xz}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{z}_i' \quad \text{and} \quad \mathbf{s}_{\mathbf{x}y} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i y_i$$

in which n is the number of samples.

However, the glaring problems of the IV method is that $\mathbb{E}[\mathbf{x}_i\mathbf{z}_i']$ is not invertible. It could be that $\mathbb{E}[\mathbf{x}_i\mathbf{z}_i']$ is not full rank, or that $\mathbb{E}[\mathbf{x}_i\mathbf{z}_i']$ is not a square matrix if $K \neq L$. Before considering how to solve for a solution, we first consider under which circumstances the solution exists and is unique.

Definition 1 (Identification)**.** *The $K \times L$ matrix $\mathbb{E}[\mathbf{x}_i\mathbf{z}_i']$ is said to satisfy identification condition if it is of full* column *rank (rank $= L$). Denote this matrix by $\mathbf{\Sigma}_{\mathbf{xz}}$.*

From eq.(4) it is clear that the identification condition is crucial for obtaining a unique solution. Let's consider three cases of K, L and identification conditions:

($K > L$) **under-identified**: If there are more regressor than instruments, then it is clear from eq.(4) that there are infinitely many solutions to δ.

($K = L$) **just-identified**: If there are exactly the same number of regressors and instruments (and the instruments are valid), then we can simply use eq.(5).

($K < L$) **over-identified**: If there are more instruments than regressors, the solution to δ cannot be obtained by eq.(5), but we know the solution exists if the identification conditions is met.

For the first case, we can do nothing; for the second case, we can use eq.(5); for the third case, we have nothing yet. Therefore, the GMM (Generalised Method of Moments) is introduced to find a solution in the third case.

1 The Method

Let's rewrite eq.(4) by moving everything to one side and also the version in which the population moments are replaced with their estimators:

$$\mathbb{E}[\mathbf{x}_i y_i] - \mathbb{E}[\mathbf{x}_i \mathbf{z}_i'] \, \delta = \mathbf{0}. \tag{6}$$

Notice that we no not actually need to invert $\mathbf{\Sigma_{xz}}$ to get a solution, we just need to find a δ in the solution space that satisfies eq.(6). Similar to the IV case, we replace the moments with their estimators and let it be denoted by $\mathbf{g}_n(\tilde{\delta})$:

$$\mathbf{g}_n(\tilde{\delta}) \coloneqq \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i (y_i - \mathbf{z}_i' \tilde{\delta}) = \mathbf{s_{xy}} - \mathbf{S_{xz}} \tilde{\delta} \tag{7}$$

where $\tilde{\delta}$ is something we want to find to make $\mathbf{g}_n(\tilde{\delta})$ to be as close to $\mathbf{0}$ as possible. To achieves this, we do not actually need to find solutions of δ in L-dimensional space, we can achieve the equivalent by minimising the *norm* of $\mathbf{g}_n(\tilde{\delta})$. We can pick any norm that is well-defined. GMM uses the well-known quadratic form:

$$\mathbf{g}_n(\tilde{\delta})' \hat{\mathbf{W}} \mathbf{g}_n(\tilde{\delta})$$

where $\hat{\mathbf{W}}$ can be any symmetric positive definite matrix. (It is often the case that the choice of $\hat{\mathbf{W}}$ depends on the data) Notice that if we choose $\hat{\mathbf{W}} = I_K$ where I_K is an $K \times K$ identity matrix, then the quadratic form reduces to Euclidean norm squared.

Now we can state the GMM estimator explicitly:

Definition 2 (GMM Estimator)**.** *Let $\hat{\mathbf{W}}$ be a $K \times K$ symmetric positive definite matrix (possible dependent on the sample) s.t. $\hat{\mathbf{W}} \xrightarrow{p} \mathbf{W}$ also symmetric*

positive definite as the sample size $n \to \infty$. The GMM estimator of δ, denoted $\hat{\delta}(\hat{\mathbf{W}})$, is

$$\hat{\delta}(\hat{\mathbf{W}}) = \arg\min_{\tilde{\delta}} \mathcal{J}(\tilde{\delta}, \hat{\mathbf{W}}) \quad \textit{where} \quad \mathcal{J}(\tilde{\delta}, \hat{\mathbf{W}}) \coloneqq n \cdot \mathbf{g}_n(\tilde{\delta})' \hat{\mathbf{W}} \mathbf{g}_n(\tilde{\delta})$$

To obtain the explicit form of $\hat{\delta}(\hat{\mathbf{W}})$, we consider the first order condition (check for gradient $= \mathbf{0}$):

$$\frac{\partial \mathcal{J}(\tilde{\delta}, \hat{\mathbf{W}})}{\partial \tilde{\delta}} \equiv \begin{bmatrix} \frac{\partial \mathcal{J}(\tilde{\delta}, \hat{\mathbf{W}})}{\partial \tilde{\delta}_1} \\ \frac{\partial \mathcal{J}(\tilde{\delta}, \hat{\mathbf{W}})}{\partial \tilde{\delta}_2} \\ \vdots \\ \frac{\partial \mathcal{J}(\tilde{\delta}, \hat{\mathbf{W}})}{\partial \tilde{\delta}_K} \end{bmatrix} = 2n \cdot \mathbf{S}'_{\mathbf{xz}} \hat{\mathbf{W}} (\mathbf{s}_{\mathbf{x}y} - \mathbf{S}_{\mathbf{xz}} \tilde{\delta}) \overset{\text{let}}{=} \mathbf{0}.$$

And by rearranging, we have

$$\hat{\delta}(\hat{\mathbf{W}}) = (\mathbf{S}'_{\mathbf{xz}} \hat{\mathbf{W}} \mathbf{S}_{\mathbf{xz}})^{-1} \mathbf{S}'_{\mathbf{xz}} \hat{\mathbf{W}} \mathbf{s}_{\mathbf{x}y}. \tag{8}$$

Notice that here we rely on $\mathbf{S}'_{\mathbf{xz}} \hat{\mathbf{W}} \mathbf{S}_{\mathbf{xz}}$ being invertible. This is justified *asymptotically*. Given some sufficiently large n, then we can guarantee $\mathbf{S}_{\mathbf{xz}}$ is full column rank. Since it is required that $\hat{\mathbf{W}}$ is positive definite, it is invertible. (Since its kernel has dimension 0.) Therefore, $\mathbf{S}'_{\mathbf{xz}} \hat{\mathbf{W}} \mathbf{S}_{\mathbf{xz}}$ is invertible However, sufficiently large n is nearly impossible to obtain. Thus, in practice, we do not necessarily rely on this explicit form.

2 Properties

2.1 Sampling Error

The sampling error can be obtained by multiplying the true model eq.(1) by $\mathbf{x}_i$ on both sides and taking the average:

$$y_i = \mathbf{z}_i' \delta + \varepsilon_i \Longrightarrow \mathbf{s}_{\mathbf{x}y} = \mathbf{S}_{\mathbf{xz}} \delta + \bar{\mathbf{g}} \quad \text{where} \quad \bar{\mathbf{g}} \coloneqq \mathbf{g}_n(\delta) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \varepsilon_i$$

and then substitute $\mathbf{s}_{\mathbf{x}y}$ it into eq.(8):

$$\begin{aligned} \hat{\delta}(\hat{\mathbf{W}}) &= (\mathbf{S}'_{\mathbf{xz}} \hat{\mathbf{W}} \mathbf{S}_{\mathbf{xz}})^{-1} \mathbf{S}'_{\mathbf{xz}} \hat{\mathbf{W}} (\mathbf{S}_{\mathbf{xz}} \delta + \bar{\mathbf{g}}) \\ \Longrightarrow \hat{\delta}(\hat{\mathbf{W}}) - \delta &= (\mathbf{S}'_{\mathbf{xz}} \hat{\mathbf{W}} \mathbf{S}_{\mathbf{xz}})^{-1} \mathbf{S}'_{\mathbf{xz}} \hat{\mathbf{W}} \bar{\mathbf{g}} \xrightarrow{p} 0 \end{aligned} \tag{9}$$

The consistency of GMM immediately follows from the above.

2.2 Asymptotic Distribution and its Estimation

Since this is a classical theory, the asymptotic distribution is normal. Consider the explicit form of GMM eq.(9) and multiply both sides by $\sqrt{n}$:

$$\sqrt{n}(\hat{\delta}(\hat{\mathbf{W}}) - \delta) = (\mathbf{S}'_{\mathbf{xz}} \hat{\mathbf{W}} \mathbf{S}_{\mathbf{xz}})^{-1} \mathbf{S}'_{\mathbf{xz}} \hat{\mathbf{W}} \left(\sqrt{n} \bar{\mathbf{g}}\right) \xrightarrow{d} \mathcal{N}\left(0, \operatorname{Avar}(\hat{\delta}(\hat{\mathbf{W}}))\right)$$

where $\text{Avar}(\hat{\delta}(\hat{\mathbf{W}}))$ is the asymptotic covariance matrix. [1] It remains to find the explicit expression of $\text{Avar}(\hat{\delta}(\hat{\mathbf{W}}))$. To obtain the expression, the only way is to expand out the square of $(\hat{\delta}(\hat{\mathbf{W}}) - \delta)$:

$$(\hat{\delta}(\hat{\mathbf{W}}) - \delta)(\hat{\delta}(\hat{\mathbf{W}}) - \delta)' = (\mathbf{S}_{\mathbf{xz}}'\hat{\mathbf{W}}\mathbf{S}_{\mathbf{xz}})^{-1}\mathbf{S}_{\mathbf{xz}}'\hat{\mathbf{W}}\bar{\mathbf{g}}\bar{\mathbf{g}}'\hat{\mathbf{W}}\mathbf{S}_{\mathbf{xz}}(\mathbf{S}_{\mathbf{xz}}'\hat{\mathbf{W}}\mathbf{S}_{\mathbf{xz}})^{-1}$$

Notice that everything in the above expression, apart from $\bar{\mathbf{g}}\bar{\mathbf{g}}'$, converges to something we know (or something we defined). Thus, we consider $\bar{\mathbf{g}}\bar{\mathbf{g}}'$, the covariance matrix of $\bar{\mathbf{g}}$:

$$\bar{\mathbf{g}}\bar{\mathbf{g}}' = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{x}_i\varepsilon_i\varepsilon_j\mathbf{x}_j' = \underbrace{\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2\mathbf{x}_i\mathbf{x}_i'}_{\xrightarrow{p}\mathbb{E}\left[\varepsilon_i^2\mathbf{x}_i\mathbf{x}_i'\right]} + \underbrace{\frac{1}{n}\sum_{i}\sum_{j\neq i}\mathbf{x}_i\varepsilon_i\varepsilon_j\mathbf{x}_j'}_{\xrightarrow{p}0}$$

where we let $\mathbf{\Omega} \coloneqq \mathbb{E}\left[\varepsilon_i^2\mathbf{x}_i\mathbf{x}_i'\right]$. Therefore, the explicit form of $\text{Avar}(\hat{\delta}(\hat{\mathbf{W}}))$ is

$$\text{Avar}(\hat{\delta}(\hat{\mathbf{W}})) = (\mathbf{\Sigma}_{\mathbf{xz}}'\mathbf{W}\mathbf{\Sigma}_{\mathbf{xz}})^{-1}\mathbf{\Sigma}_{\mathbf{xz}}'\mathbf{W}\mathbf{\Omega}\mathbf{W}\mathbf{\Sigma}_{\mathbf{xz}}(\mathbf{\Sigma}_{\mathbf{xz}}'\mathbf{W}\mathbf{\Sigma}_{\mathbf{xz}})^{-1}.$$

Therefore, the estimator for $\text{Avar}(\hat{\delta}(\hat{\mathbf{W}}))$ is simply

$$\widehat{\text{Avar}(\hat{\delta}(\hat{\mathbf{W}}))} = (\mathbf{S}_{\mathbf{xz}}'\hat{\mathbf{W}}\mathbf{S}_{\mathbf{xz}})^{-1}\mathbf{S}_{\mathbf{xz}}'\hat{\mathbf{W}}\hat{\mathbf{\Omega}}\hat{\mathbf{W}}\mathbf{S}_{\mathbf{xz}}(\mathbf{S}_{\mathbf{xz}}'\hat{\mathbf{W}}\mathbf{S}_{\mathbf{xz}})^{-1}$$

provided we have an unbiased estimator $\hat{\mathbf{\Omega}}$ of $\mathbf{\Omega}$. The obvious choice for $\hat{\mathbf{\Omega}}$ is the same as OLS:

$$\hat{\mathbf{\Omega}} = \frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2\mathbf{x}_i\mathbf{x}_i' \quad \text{where} \quad \hat{\varepsilon}_i^2 \coloneqq y_i - \mathbf{z}_i'\hat{\delta}$$

Note that $\hat{\delta}$ can be any consistent estimator of δ.

So far the procedure is very similar to OLS. Once we obtained the asymptotic distribution, then we can perform hypothesis testing in the regular manner. Two questions remain to be answered:

(i) How to choose $\hat{\mathbf{W}}$?

(ii) Can we somehow check the two conditions, relevancy and exogeneity, of instruments?

We will answer the two questions in the following two sections.

[1] Here we rely on $\sqrt{n}\bar{\mathbf{g}} \xrightarrow{d} \mathcal{N}(0, \mathbf{\Omega})$. This in turn relies on some assumptions on the property of $\bar{\mathbf{g}}$. The assumption is that $\mathbf{g}_i$ is a martingale difference sequence, *a fortiori*, $\mathbb{E}[\mathbf{g}_i] = 0$.

3 How to Choose $\hat{\mathbf{W}}$?

Although different choices of $\hat{\mathbf{W}}$ will not alter the consistency of the estimator, but the efficiency of it will differ. Thus, a natural question is whether we can find some optimal $\hat{\mathbf{W}}$ that minimises the asymptotic variance $\text{Avar}(\hat{\delta}(\hat{\mathbf{W}}))$. By "minimise" we mean trying to find a choice for $\hat{\mathbf{W}}$, say $\mathbf{X}$, such that every element in $\text{Avar}(\hat{\delta}(\mathbf{X}))$ is no larger than $\text{Avar}(\hat{\delta}(\hat{\mathbf{W}}))$ $\forall \hat{\mathbf{W}}$.

Before a serious attempt of minimising $\text{Avar}(\hat{\delta}(\hat{\mathbf{W}}))$, it tempting to choose $\mathbf{W} = \boldsymbol{\Omega}^{-1}$ since

$$\begin{aligned}
\text{Avar}(\hat{\delta}(\boldsymbol{\Omega}^{-1})) &= (\boldsymbol{\Sigma}'_{\mathbf{xz}}\boldsymbol{\Omega}^{-1}\boldsymbol{\Sigma}_{\mathbf{xz}})^{-1}\boldsymbol{\Sigma}'_{\mathbf{xz}}\cancel{\boldsymbol{\Omega}^{-1}\boldsymbol{\Omega}\boldsymbol{\Omega}^{-1}}\boldsymbol{\Sigma}_{\mathbf{xz}}(\boldsymbol{\Sigma}'_{\mathbf{xz}}\boldsymbol{\Omega}^{-1}\boldsymbol{\Sigma}_{\mathbf{xz}})^{-1} \\
&= (\boldsymbol{\Sigma}'_{\mathbf{xz}}\boldsymbol{\Omega}^{-1}\boldsymbol{\Sigma}_{\mathbf{xz}})^{-1}\cancel{\boldsymbol{\Sigma}'_{\mathbf{xz}}\boldsymbol{\Omega}^{-1}\boldsymbol{\Sigma}_{\mathbf{xz}}}\cancel{(\boldsymbol{\Sigma}'_{\mathbf{xz}}\boldsymbol{\Omega}^{-1}\boldsymbol{\Sigma}_{\mathbf{xz}})^{-1}} \\
&= (\boldsymbol{\Sigma}'_{\mathbf{xz}}\boldsymbol{\Omega}^{-1}\boldsymbol{\Sigma}_{\mathbf{xz}})^{-1}
\end{aligned}$$

is a big simplification. And this can be achieved by some $\hat{\mathbf{W}} \xrightarrow{p} \boldsymbol{\Omega}^{-1}$. In fact, it can be shown that this choice of $\hat{\mathbf{W}}$ is an optimal choice:

Proposition 1 (optimal choice of weighting matrix)**.** *A lower bound for the asymptotic variance of GMM estimators indexed by $\hat{\mathbf{W}}$ is given by $(\boldsymbol{\Sigma}'_{\mathbf{xz}}\boldsymbol{\Omega}^{-1}\boldsymbol{\Sigma}_{\mathbf{xz}})^{-1}$. This lower bound can be achieved by choosing $\hat{\mathbf{W}} \xrightarrow{p} \boldsymbol{\Omega}^{-1}$.*

Proof. (sketch) We want to show that all entries in matrix

$$\text{Avar}(\hat{\delta}(\mathbf{W})) - \text{Avar}(\hat{\delta}(\boldsymbol{\Omega}^{-1}))$$

are non-negative $\forall$ symmetric positive definite $\mathbf{W}$. We achieve this goal by stripping $\text{Avar}(\hat{\delta}(\hat{\mathbf{W}}))$ on both sides until only I remains:

$$\begin{aligned}
&(\boldsymbol{\Sigma}'_{\mathbf{xz}}\mathbf{W}\boldsymbol{\Sigma}_{\mathbf{xz}})^{-1}\boldsymbol{\Sigma}'_{\mathbf{xz}}\mathbf{W}\boldsymbol{\Omega}\mathbf{W}\boldsymbol{\Sigma}_{\mathbf{xz}}(\boldsymbol{\Sigma}'_{\mathbf{xz}}\mathbf{W}\boldsymbol{\Sigma}_{\mathbf{xz}})^{-1} - (\boldsymbol{\Sigma}'_{\mathbf{xz}}\boldsymbol{\Omega}^{-1}\boldsymbol{\Sigma}_{\mathbf{xz}})^{-1} \\
&= \bigcirc\Big[\boldsymbol{\Sigma}'_{\mathbf{xz}}\mathbf{W}\boldsymbol{\Omega}\mathbf{W}\boldsymbol{\Sigma}_{\mathbf{xz}} - (\boldsymbol{\Sigma}'_{\mathbf{xz}}\mathbf{W}\boldsymbol{\Sigma}_{\mathbf{xz}})(\boldsymbol{\Sigma}_{\mathbf{xz}}\boldsymbol{\Omega}^{-1}\boldsymbol{\Sigma}_{\mathbf{xz}})^{-1}(\boldsymbol{\Sigma}'_{\mathbf{xz}}\mathbf{W}\boldsymbol{\Sigma}_{\mathbf{xz}})\Big]\bigcirc \\
&= \bigcirc\Big[\boldsymbol{\Omega} - \boldsymbol{\Sigma}_{\mathbf{xz}}(\boldsymbol{\Sigma}'_{\mathbf{xz}}\boldsymbol{\Omega}^{-1}\boldsymbol{\Sigma}_{\mathbf{xz}})^{-1}\boldsymbol{\Sigma}'_{\mathbf{xz}}\Big]\bigcirc \\
&= \bigcirc\Big[I - \boldsymbol{\Omega}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{\mathbf{xz}}(\boldsymbol{\Sigma}'_{\mathbf{xz}}\boldsymbol{\Omega}^{-1}\boldsymbol{\Sigma}_{\mathbf{xz}})^{-1}\boldsymbol{\Sigma}'_{\mathbf{xz}}\boldsymbol{\Omega}^{-\frac{1}{2}}\Big]\bigcirc \\
&= A[I - B]A'
\end{aligned}$$

It is clear that B is symmetric and idempotent. Therefore, all entries in

$$A[I - B]A' = A(I - B)(I - B)'A'$$

are positive. □

However, there are is glaring problems of choosing $\mathbf{W} = \boldsymbol{\Omega}^{-1}$: If we set $\hat{\mathbf{W}} = \hat{\boldsymbol{\Omega}}^{-1}$, we have to obtain $\hat{\boldsymbol{\Omega}}$ first, and to obtain $\boldsymbol{\Omega}$, we have to obtain an estimation for δ first. In practice, there are a number of methods that overcomes this problem, either by doing a initial estimation of δ first or update $\hat{\mathbf{W}}$ iteratively.

4 How to Check Instruments?

Checking for instrument validity is very hard. Consistent with the GMM idea, the J-test is proposed:

Proposition 2 (Hansen's test of overidentifying restrictions)**.** *If all assumptions for GMM hold, and there is a consistent estimator for* $\mathbf{\Omega}^{-1}$*, then*

$$\mathcal{J}(\hat{\delta}(\hat{\mathbf{\Omega}}^{-1}), \hat{\mathbf{\Omega}}^{-1}) \xrightarrow{d} \chi^2_{K-L}$$

Remark 1. *Notice that this test is for* all *assumptions. That is, if the test fails, any assumption could be false. Furthermore, it is an necessary condition for the validity of the assumptions, i.e., even if the test holds, it is not necessary that the assumptions hold.*

This J-test is what we've learnt in year two, the difference is that what we've learned is the homoskedastic version. The one stated above is the general version.

5 Implications of Homoskedasticity

If we impose homoskedasticity, then GMM becomes TSLS. This is why we say the TSLS error is not robust.

6 Conclusion

References

[1] Econometrics (2000) Fumio Hayashi. Chapter 3. ■