LECTURE 1

REVIEW OF GMM FOR LINEAR MODELS

Definition and asymptotic properties

Suppose that an econometrician observes the data $\{(Y_i, X_i', Z_i') : i = 1, \ldots, n\}$, and the model is given by

$$\begin{aligned} Y_i &= X_i'\beta + U_i, \text{ and} \\ E(Z_i U_i) &= 0, \end{aligned} \tag{1}$$

where $\beta \in R^k$ is the vector of unknown parameters, X_i is a random k-vector of regressors, and Z_i is a random l-vector of instrumental variables (IVs). In the usual regression model, we have $X_i = Z_i$. In the IV regression model, X_i or some of its elements might be endogenous, i.e. $E(X_i U_i) \neq 0$, which requires using (weakly) exogenous instruments Z_i (exogenous elements of X_i may be a part of Z_i as well). A simple rule is that for every endogenous regressor in X_i there should be at least one instrument in Z_i that is excluded from X_i. For the formal requirement, see the rank condition for $E(Z_i X_i')$ below.

When $k = l$, we can estimate β by the value that solves the sample analogue of (1), $\hat{\beta}_n$:

$$0 = n^{-1} \sum_{i=1}^{n} Z_i \left(Y_i - X_i' \hat{\beta}_n \right).$$

This provides us with k equations in k unknowns, which can be solved to obtain $\hat{\beta}_n$:

$$\hat{\beta}_n = \left(\sum_{i=1}^{n} Z_i X_i' \right)^{-1} \sum_{i=1}^{n} Z_i Y_i,$$

provided that $\sum_{i=1}^{n} Z_i X_i'$ is invertible. However, when $l > k$, in general, there is no $\hat{\beta}_n \in R^k$ that solves all l equations exactly. In this case, we can choose the value that makes the sample moments as close to zero as possible.

Let A_n be a (possibly random) $l \times l$ weight matrix such that $A_n \to_p A$, where A is non-random and has full rank (l). For a given choice of the weight matrix A_n, the *Generalized Method of Moments (GMM) estimator* of β is the value of b that minimizes the weighted distance of $n^{-1} \sum_{i=1}^{n} Z_i (Y_i - X_i' b)$ from zero:

$$\begin{aligned} \widehat{\beta}_n(A_n) &= \arg\min_b \left\| A_n n^{-1} \sum_{i=1}^{n} Z_i (Y_i - X_i' b) \right\|^2 \\ &= \arg\min_b \left(n^{-1} \sum_{i=1}^{n} Z_i (Y_i - X_i' b) \right)' A_n' A_n \left(n^{-1} \sum_{i=1}^{n} Z_i (Y_i - X_i' b) \right). \end{aligned} \tag{2}$$

The GMM actually produces a class of estimators indexed by A_n, i.e. different choices of A_n produce different estimators. Note that $A'A$ is positive definite.

By obtaining the first order conditions and solving for $\widehat{\beta}_n(A_n)$, one obtains

$$\widehat{\beta}_n(A_n) = \left(\sum_{i=1}^{n} X_i Z_i' (A_n' A_n) \sum_{i=1}^{n} Z_i X_i' \right)^{-1} \sum_{i=1}^{n} X_i Z_i' (A_n' A_n) \sum_{i=1}^{n} Z_i Y_i.$$

We will show next that the GMM estimator is consistent. We will use the following assumptions.

- $\{(Y_i, X_i, Z_i) : i \geq 1\}$ are iid.
- $Y_i = X_i'\beta + U_i$, $\beta \in R^k$.

- $EZ_iU_i = 0$.
- $E(Z_iX_i')$ has rank k.
- $A_n \to_p A$, where A has rank $l \geq k$.
- $EX_{i,j}^2 < \infty$ for all $j = 1, \ldots, k$.
- $EZ_{i,j}^2 < \infty$ for all $j = 1, \ldots, l$.

Write

$$\widehat{\beta}_n(A_n) = \beta + \left(n^{-1}\sum_{i=1}^n X_iZ_i'(A_n'A_n)\,n^{-1}\sum_{i=1}^n Z_iX_i'\right)^{-1} n^{-1}\sum_{i=1}^n X_iZ_i'(A_n'A_n)\,n^{-1}\sum_{i=1}^n Z_iU_i.$$

The last two of the above assumptions and Cauchy-Schwartz inequality imply that

$$\begin{aligned} E|X_{i,r}Z_{i,s}| &\leq \sqrt{EX_{i,r}^2 EZ_{i,s}^2} \\ &< \infty \end{aligned}$$

for all $r = 1, \ldots, k$ and $s = 1, \ldots, l$. Therefore, by the Weak Law of Large Numbers (WLLN),

$$n^{-1}\sum_{i=1}^n X_iZ_i' \to_p EX_iZ_i'.$$

Since $A_n \to_p A$, we also have that

$$n^{-1}\sum_{i=1}^n X_iZ_i'(A_n'A_n)\,n^{-1}\sum_{i=1}^n Z_iX_i' \to_p EX_iZ_i'(A'A)\,EZ_iX_i'.$$

Further, since $E(Z_iX_i')$ has rank k, A has rank $l \geq k$, it follows that the $k \times k$ matrix $EX_iZ_i'(A'A)\,EZ_iX_i'$ has a full rank k and, therefore, invertible. Consequently, by the Slutsky's Theorem,

$$\left(n^{-1}\sum_{i=1}^n X_iZ_i'(A_n'A_n)\,n^{-1}\sum_{i=1}^n Z_iX_i'\right)^{-1} \to_p \left(EX_iZ_i'(A'A)\,EZ_iX_i'\right)^{-1}.$$

Next, by the WLLN,

$$\begin{aligned} n^{-1}\sum_{i=1}^n Z_iU_i &\to_p EZ_iU_i \\ &= 0, \end{aligned}$$

and thus $\widehat{\beta}_n(A_n) \to_p \beta$.

In order to show asymptotic normality, we will use the following three assumptions in addition to the above.

- $EZ_{i,j}^4 < \infty$ for all $j = 1, \ldots, l$.
- $EU_i^4 < \infty$,
- $EU_i^2Z_iZ_i'$ is positive definite.

Write

$$n^{1/2}\left(\widehat{\beta}_n(A_n)-\beta\right)=\left(n^{-1}\sum_{i=1}^n X_iZ_i'(A_n'A_n)\,n^{-1}\sum_{i=1}^n Z_iX_i'\right)^{-1} n^{-1}\sum_{i=1}^n X_iZ_i'(A_n'A_n)\,n^{-1/2}\sum_{i=1}^n Z_iU_i.$$

The two additional assumptions imply that the variance of Z_iU_i, $EU_i^2Z_iZ_i'$ is finite:

$$\begin{aligned} E\left|U_i^2Z_{i,r}Z_{i,s}\right| &\leq \sqrt{EU_i^4E\left(Z_{i,r}Z_{i,s}\right)^2} \\ &\leq \sqrt{EU_i^4\sqrt{EZ_{i,r}^4EZ_{i,s}^4}} \\ &< \infty, \end{aligned} \tag{3}$$

for all $r,s=1,\ldots,l$. Hence, we can apply the Central Limit Theorem (CLT) to obtain that

$$n^{-1/2}\sum_{i=1}^n Z_iU_i \to_d N\left(0,EU_i^2Z_iZ_i'\right).$$

Let's define

$$\begin{aligned} Q &= EZ_iX_i', \\ \Omega &= EU_i^2Z_iZ_i'. \end{aligned}$$

Combining the above results, we have

$$n^{1/2}\left(\widehat{\beta}_n(A_n)-\beta\right) \to_d N\left(0,V(A)\right), \tag{4}$$

where the asymptotic variance $V(A)$ takes a sandwich form:

$$V(A)=\left(Q'A'AQ\right)^{-1}Q'A'A\Omega A'AQ\left(Q'A'AQ\right)^{-1}.$$

Estimation of the asymptotic variance

The variance-covariance matrix $V(A)$ can be estimated consistently by replacing A, Q and Ω with their consistent estimators A_n, $\widehat{Q}_n$ and $\widehat{\Omega}_n$ respectively:

$$\begin{aligned} \widehat{V}_n(A_n) &= \left(\widehat{Q}_n'A_n'A_n\widehat{Q}_n\right)^{-1}\widehat{Q}_n'A_n'A_n\widehat{\Omega}_nA_n'A_n\widehat{Q}_n\left(\widehat{Q}_n'A_n'A_n\widehat{Q}_n\right)^{-1} \\ \widehat{Q}_n &= n^{-1}\sum_{i=1}^n Z_iX_i', \\ \widehat{\Omega}_n &= n^{-1}\sum_{i=1}^n \widehat{U}_i^2Z_iZ_i', \text{ where} \\ \widehat{U}_i &= Y_i-X_i'\widehat{\beta}_n(A_n) \\ &= U_i-X_i'\left(\widehat{\beta}_n(A_n)-\beta\right). \end{aligned}$$

In order to show consistency of $\widehat{V}_n(A_n)$, we need to show that $\widehat{\Omega}_n \to_p \Omega$. To make the notation shorter, I will suppress the dependence of the GMM estimator of β on A_n. Write

$$\begin{aligned}
\widehat{\Omega}_n &= n^{-1}\sum_{i=1}^{n}\widehat{U}_i^2 Z_i Z_i' \\
&= n^{-1}\sum_{i=1}^{n}U_i^2 Z_i Z_i' - 2R_{1,n} + R_{2,n}, \text{ where} \\
R_{1,n} &= n^{-1}\sum_{i=1}^{n}\left(\left(\widehat{\beta}_n-\beta\right)' X_i U_i\right) Z_i Z_i', \\
R_{2,n} &= n^{-1}\sum_{i=1}^{n}\left(\left(\widehat{\beta}_n-\beta\right)' X_i\right)^2 Z_i Z_i'.
\end{aligned}$$

From (3) we know that $EU_i^2 Z_i Z_i'$ is finite, and, therefore,

$$\begin{aligned}
n^{-1}\sum_{i=1}^{n}U_i^2 Z_i Z_i' &\to_p EU_i^2 Z_i Z_i' \\
&= \Omega.
\end{aligned}$$

Thus, we need to show that both $R_{1,n}$ and $R_{2,n}$ converge in probability to zero. In addition to the above assumptions we will assume that

- $EX_{i,j}^4 < \infty$ for all $j = 1, \ldots, k$.

Define the matrix norm of a $r \times c$ matrix B as

$$\begin{aligned}
\|B\| &= \left(tr\left(B'B\right)\right)^{1/2} \\
&= \left(\sum_{i=1}^{r}\sum_{j=1}^{c}B_{ij}^2\right)^{1/2},
\end{aligned}$$

where B_{ij} denotes the (i,j)-th element of the matrix B. We have

$$\begin{aligned}
\|R_{1,n}\| &\le n^{-1}\sum_{i=1}^{n}\left\|\left(\left(\widehat{\beta}_n-\beta\right)' X_i U_i\right) Z_i Z_i'\right\| \\
&= n^{-1}\sum_{i=1}^{n}\left|\left(\widehat{\beta}_n-\beta\right)' X_i\right| |U_i| \, tr\left(Z_i Z_i' Z_i Z_i'\right)^{1/2} \\
&= n^{-1}\sum_{i=1}^{n}\left|\left(\widehat{\beta}_n-\beta\right)' X_i\right| |U_i| \, \|Z_i\| \, tr\left(Z_i Z_i'\right)^{1/2} \\
&= n^{-1}\sum_{i=1}^{n}\left|\left(\widehat{\beta}_n-\beta\right)' X_i\right| |U_i| \, \|Z_i\|^2 \\
&\le \left\|\widehat{\beta}_n-\beta\right\| n^{-1}\sum_{i=1}^{n}|U_i| \, \|X_i\| \, \|Z_i\|^2 .
\end{aligned}$$

Now, $\left\|\widehat{\beta}_n-\beta\right\| \to_p 0$. Hence, $\|R_{1,n}\| \to_p 0$ provided that $n^{-1}\sum_{i=1}^{n}|U_i| \, \|X_i\| \, \|Z_i\|^2$ does not diverge as $n \to \infty$. Next, we will use the Holder's inequality. According to Holder's inequality, for $p > 1, q > 1$ such that $1/p + 1/q = 1$,

$$E|XY| \le \left(E|X|^p\right)^{1/p}\left(E|Y|^q\right)^{1/q}.$$

Set $p = 4$ and $q = 4/3$ to obtain

$$E|U_i| \|X_i\| \|Z_i\|^2 \leq \left(E|U_i|^4\right)^{1/4} \left(E\|X_i\|^{4/3} \|Z_i\|^{8/3}\right)^{3/4}.$$

Apply the Holder's inequality once again with $p = 3$ and $q = 3/2$,

$$E\|X_i\|^{4/3} \|Z_i\|^{8/3} \leq \left(E\|X_i\|^4\right)^{1/3} \left(E\|Z_i\|^4\right)^{2/3}.$$

Thus,

$$\begin{aligned} E|U_i| \|X_i\| \|Z_i\|^2 &\leq \left(E|U_i|^4\right)^{1/4} \left(E\|X_i\|^4\right)^{1/4} \left(E\|Z_i\|^4\right)^{1/2} \\ &< \infty, \end{aligned}$$

and, therefore, by the WLLN

$$n^{-1} \sum_{i=1}^{n} |U_i| \|X_i\| \|Z_i\|^2 \to_p E|U_i| \|X_i\| \|Z_i\|^2,$$

a finite constant. In a similar way, one can show that $\|R_{2,n}\| \to_p 0$.

Efficient GMM

An efficient GMM estimator is such that it has the smallest asymptotic variance $V(A)$ among the class of GMM estimators $\widehat{\beta}_n(A_n)$. The lower bound on $V(A)$ is given by $\left(Q'\Omega^{-1}Q\right)^{-1}$. This can be shown by proving that $\left(Q'\Omega^{-1}Q\right)^{-1} - V(A)$ is negative semi-definite or, alternatively, that $Q'\Omega^{-1}Q - (V(A))^{-1}$ is positive semi-definite for all A's of full rank l. Consider

$$Q'\Omega^{-1}Q - (V(A))^{-1} = Q'\Omega^{-1}Q - Q'A'AQ\left(Q'A'A\Omega A'AQ\right)^{-1} Q'A'AQ. \tag{5}$$

Assume that Ω is positive definite. We can write

$$\Omega^{-1} = C'C,$$

where C is invertible as well. Write (5) as

$$\begin{aligned} & Q'C'CQ - Q'A'AQ\left(Q'A'AC^{-1}(C')^{-1}A'AQ\right)^{-1} Q'A'AQ \\ = \; & Q'C'\left(I - (C')^{-1}A'AQ\left(Q'A'AC^{-1}(C')^{-1}A'AQ\right)^{-1} Q'A'AC^{-1}\right)CQ. \end{aligned} \tag{6}$$

Define

$$H = (C')^{-1}A'AQ,$$

and note that, using this definition, (6) becomes

$$Q'C'\left(I - H(H'H)^{-1}H'\right)CQ.$$

The above matrix is positive semi-definite if $I - H(H'H)^{-1}H'$ is positive semi-definite. Next,

$$\begin{aligned} & \left(I - H(H'H)^{-1}H'\right)\left(I - H(H'H)^{-1}H'\right) \\ = \; & I - 2H(H'H)^{-1}H' + H(H'H)^{-1}H'H(H'H)^{-1}H' \\ = \; & I - H(H'H)^{-1}H'. \end{aligned}$$

Therefore, $I - H\left(H'H\right)^{-1}H'$ is idempotent and, consequently, positive semi-definite.

The lower bound is achieved if $A_n'A_n \to_p A'A = \Omega^{-1}$. A natural choice for such $A_n'A_n$ is $\widehat{\Omega}_n^{-1}$. Thus, the efficient GMM estimator is given by

$$\widehat{\beta}_n = \left(\sum_{i=1}^n X_i Z_i' \widehat{\Omega}_n^{-1} \sum_{i=1}^n Z_i X_i'\right)^{-1} \sum_{i=1}^n X_i Z_i' \widehat{\Omega}_n^{-1} \sum_{i=1}^n Z_i Y_i. \tag{7}$$

It can be constructed using a two-step procedure.

1. Choose some A_n. For example, $A_n'A_n = I_l$. Obtain the corresponding (inefficient) estimates of β, say $\widetilde{\beta}_n$.

2. Using $\widetilde{\beta}_n$, construct $\widehat{\Omega}_n$:

$$\begin{aligned} \widehat{\Omega}_n &= n^{-1}\sum_{i=1}^n \widehat{U}_i^2 Z_i Z_i', \text{ where} \\ \widehat{U}_i &= Y_i - X_i'\widetilde{\beta}_n, \end{aligned}$$

   Obtain the efficient GMM estimates of β using (7).

2SLS and control function approach

Consider the homoskedastic case:

$$E\left(U_i^2|Z_i\right) = \sigma^2,$$

where σ^2 is some constant. In this case, $\Omega = \sigma^2 E Z_i Z_i'$ and $\widehat{\Omega}_n = \widehat{\sigma}_n^2 n^{-1}\sum_{i=1}^n Z_i Z_i'$, where $\widehat{\sigma}_n^2$ is some consistent estimator of σ^2. In this case, the efficient GMM estimator is the 2SLS estimator:

$$\widehat{\beta}_n^{2SLS} = \left(\sum_{i=1}^n X_i Z_i' \left(\sum_{i=1}^n Z_i Z_i'\right)^{-1} \sum_{i=1}^n Z_i X_i'\right)^{-1} \sum_{i=1}^n X_i Z_i' \left(\sum_{i=1}^n Z_i Z_i'\right)^{-1} \sum_{i=1}^n Z_i Y_i.$$

In fact, in the homoskedastic case, one does not need to perform the first step involving an inefficient GMM estimator, since the optimal weight matrix can be obtained directly.

Define:

$$X = \begin{pmatrix} X_1' \\ \vdots \\ X_n' \end{pmatrix}, Z = \begin{pmatrix} Z_1' \\ \vdots \\ Z_n' \end{pmatrix}, Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \text{ and } U = \begin{pmatrix} U_1 \\ \vdots \\ U_n \end{pmatrix}.$$

In matrix notation, the 2SLS estimator can be written as

$$\begin{aligned} \widehat{\beta}_n^{2SLS} &= \left(X'Z(Z'Z)^{-1}Z'X\right)^{-1} X'Z(Z'Z)^{-1}Z'Y \\ &= (X'P_Z X)^{-1} X'P_Z Y \\ &= \left(\hat{X}'\hat{X}\right)^{-1}\hat{X}'Y, \end{aligned}$$

where

$$\begin{aligned} P_Z &= Z(Z'Z)^{-1}Z', \\ \hat{X} &= P_Z X. \end{aligned}$$

Here, P_Z is the orthogonal projection matrix for the sub-space of $\mathbb{R}^l$ spanned by the columns of Z. Note that $\hat{X} = P_Z X$ has an interpretation of fitted values from a regression of the columns of X against Z:

$$X = Z\Pi + V,$$

where Π is a $l \times k$ matrix of parameters, and V is an $n \times k$ matrix of errors uncorrelated with Z's. The above equation is often referred to as the first-stage. The matrix of parameters is consistently estimated by $\hat{\Pi}_n = (Z'Z)^{-1}Z'X$, and we have $\hat{X} = P_Z X = Z\hat{\Pi}_n$.

When $k = l$, one can easily show that the 2SLS estimator reduces to

$$\widehat{\beta}_n^{IV} = \left(\sum_{i=1}^n Z_i X_i'\right)^{-1} \sum_{i=1}^n Z_i Y_i = (Z'X)^{-1} Z'Y,$$

which is often called the IV estimator.

The 2SLS estimator can be viewed as a basic example of the so-called "fitted values method" (see Blundell and Powell "Endogeneity in Nonparametric and Semiparametric Regression Models". According to this method, the matrix of endogenous regressors X in $Y = X\beta + U$ should be replaced with fitted values from the first-stage. To justify this approach, note that substitution of the first-stage in $Y = X\beta + U$ produces

$$Y = Z\Pi\beta + U + V\beta.$$

Since $Z\Pi$ is uncorrelated with U and V, one could estimate β by OLS using $Z\Pi$ as regressors if Π were known. Since Π is unknown, one has to use $\hat{\Pi}_n$ instead, which leads to 2SLS. Note that the fitted values method can only be used with linear models. If the true model for Y is nonlinear, e.g. $Y = m(X, \beta) + U$, substitution of the first-stage will produce $Y = m(Z\Pi + V, \beta) + U$, where $m(Z\Pi + V, \beta) \neq m(Z\Pi, \beta) + V$ unless $m(\cdot, \beta)$ is a linear function. Thus with nonlinear models, one cannot expect that the fitted values approach will lead to consistent estimation of β.

The 2SLS estimator can also be viewed as an example of the so-called "control function" approach to estimation. Consider again the first-stage equation

$$X = Z\Pi + V.$$

Since Z is exogenous and X is endogenous, X can be correlated with U only through V. Define:

$$\lambda = (EV_iV_i')^{-1}EV_iU_i,$$

and write

$$\begin{aligned} Y_i &= X_i'\beta + U_i \\ &= X_i'\beta + V_i'\lambda + (U_i - V_i'\lambda) \\ &= X_i'\beta + V_i'\lambda + \epsilon_i, \end{aligned}$$

where $V_i'\lambda$ is the so-called *control function*, and $\epsilon_i = U_i - V_i'\lambda$ is the new error term. Adding the control function to the regression eliminates the endogeneity problem. First of all, by construction V_i and ϵ_i are uncorrelated:

$$\begin{aligned} EV_i\epsilon_i &= EV_i(U_i - V_i'\lambda) \\ &= EV_iU_i - EV_iV_i'\lambda \\ &= EV_iU_i - EV_iV_i'(EV_iV_i')^{-1}EV_iU_i, \\ &= 0. \end{aligned}$$

As a result, X_i and ϵ_i are uncorrelated as well:[1]

$$\begin{aligned} EX_i\epsilon_i &= E\left(\Pi'Z_i + V_i\right)\epsilon_i \\ &= \Pi'E(Z_i\epsilon_i) + E(V_i\epsilon_i) \\ &= 0. \end{aligned}$$

[1] $X = Z\Pi + V$ can be re-written in observation to observation notation as $X_i' = Z_i'\Pi + V_i'$, or $X_i = \Pi'Z_i + V_i$.

Hence, β can be consistently estimated by regressing Y against X and V. Since V is unobservable, one should replace it with first-stage fitted residuals:

$$\hat{V} = M_Z X = (I_n - P_Z) X.$$

Define:

$$M_{\hat{V}} = I_n - \hat{V}\left(\hat{V}'\hat{V}\right)^{-1}\hat{V}'.$$

Using the partitioned regression result, the control function estimator of β can be written as

$$\hat{\beta}_n^{CF} = \left(X' M_{\hat{V}} X\right)^{-1} X' M_{\hat{V}} Y.$$

Next, consider $X' M_{\hat{V}} X$:

$$\begin{aligned}
X' M_{\hat{V}} X &= X'\left(I_n - \hat{V}\left(\hat{V}'\hat{V}\right)^{-1}\hat{V}'\right)X \\
&= X'X - X'\hat{V}\left(\hat{V}'\hat{V}\right)^{-1}\hat{V}'X \\
&= X'X - X'M_Z X\left(X'M_Z X\right)^{-1} X'M_Z X \\
&= X'X - X'M_Z X \\
&= X'P_Z X.
\end{aligned}$$

Similarly, one can show

$$X' M_{\hat{V}} Y = X' P_Z Y,$$

and, therefore, in the case of the linear IV model, the control function and fitted values approaches produce the same 2SLS estimator:

$$\hat{\beta}_n^{CF} = \hat{\beta}_n^{2SLS}.$$

Unlike the fitted values approach, the control function approach can be used in nonlinear regression models for construction of consistent estimators.

Confidence intervals and hypothesis testing

It is a common practice to construct symmetric confidence intervals using the normal approximation (4). Confidence intervals for the j-th element of β, β_j, with asymptotic coverage probability $1-\alpha$ can be constructed as follows:

$$CI_{n,j,1-\alpha} = \left[\widehat{\beta}_{jn}(A_n) - z_{1-\alpha/2}\sqrt{\left[\widehat{V}_n(A_n)\right]_{jj}/n}, \widehat{\beta}_{jn}(A_n) + z_{1-\alpha/2}\sqrt{\left[\widehat{V}_n(A_n)\right]_{jj}/n}\right],$$

where $\left[\widehat{V}_n(A_n)\right]_{ij}$ denotes the element (i,j) of the matrix $\widehat{V}_n(A_n)$, and z_α is the α quantile of the standard normal distribution, i.e. for $Z \sim N(0,1)$, $P(Z \leq z_\alpha) = \alpha$. One can show that

$$P\left(\beta_j \in CI_{n,j,1-\alpha}\right) \to 1 - \alpha. \tag{8}$$

Note that the random element in the above expression is $CI_{n,j,1-\alpha}$.

Suppose the econometrician is interested in testing $H_0 : \beta = \beta_0$ against the alternative $H_1 : \beta \neq \beta_0$. A test that has asymptotic size α can be based on the Wald statistic:

$$W_n = n\left(\widehat{\beta}_n(A_n) - \beta_0\right)'\left(\widehat{V}_n(A_n)\right)^{-1}\left(\widehat{\beta}_n(A_n) - \beta_0\right).$$

One can show that under the null,

$$W_n \to_d \chi_k^2, \tag{9}$$

and, therefore, the null should be rejected when $W_n > \chi^2_{k,1-\alpha}$, where $\chi^2_{k,1-\alpha}$ is the $(1-\alpha)$ quantile of the χ^2_k distribution.

The power properties of the test (rejection of the null when it is wrong) can be considered in fixed or local alternatives frameworks. In the case of fixed alternatives, we assume that

$$H_1 : \beta = \beta_0 + \delta,$$

where $\delta \in R^k$, $\|\delta\| > 0$ gives the deviation from the null. In this case, $\widehat{\beta}_n(A_n) - \beta_0 \to_p \delta$, and, therefore,

$$\begin{aligned} W_n/n &\to_p \delta' (V(A))^{-1} \delta \\ &> 0. \end{aligned}$$

As a result,

$$P\left(W_n > \chi^2_{k,1-\alpha}\right) \to 1. \quad (10)$$

We say that a test based on W_n is consistent, since the probability to reject the null when it is wrong approaches 1 with the sample size.

In the case of local alternatives, it is assumed that

$$H_1 : \beta = \beta_0 + \delta/\sqrt{n}.$$

Thus, we consider only local to H_0 or very small deviations from the null hypothesis, where "small" is defined relatively to the sample size n. The $1/\sqrt{n}$-rate is chosen so that the statistic would have a non-degenerate asymptotic distribution when $n \to \infty$. Write

$$n^{1/2}\left(\widehat{\beta}_n(A_n) - \beta_0\right) = \delta + \left(n^{-1}\sum_{i=1}^n X_i Z_i' (A_n' A_n) n^{-1} \sum_{i=1}^n Z_i X_i'\right)^{-1} n^{-1} \sum_{i=1}^n X_i Z_i' (A_n' A_n) n^{-1/2} \sum_{i=1}^n Z_i U_i.$$

One can show that under the local alternative,

$$n^{1/2}\left(\widehat{\beta}_n(A_n) - \beta_0\right) \to_d N(\delta, V(A)), \quad (11)$$

where $V(A)$ has the same expression as before. Thus, under the local alternative, the asymptotic distribution of $n^{1/2}\left(\widehat{\beta}_n(A_n) - \beta_0\right)$ is no longer centered around zero. Naturally, noncentrality in the distribution of $n^{1/2}\left(\widehat{\beta}_n(A_n) - \beta_0\right)$ translates to that of the Wald statistic. Note that we still have that $\widehat{\beta}_n(A_n) \to_p \beta_0$ and $\widehat{V}_n(A_n) \to_p V(A)$. In order to see that the last result is true, write $\widehat{U}_i = U_i - X_i'\left(\widehat{\beta}_n(A_n) - \beta_0\right) + X_i'\delta/\sqrt{n}$, and use it together with $\widehat{\beta}_n(A_n) - \beta_0 \to_p 0$ to show that $\widehat{\Omega}_n \to_p \Omega$, where $\Omega = EU_i^2 Z_i Z_i'$ as before.

Let V_c be a random k-vector such that $V_c \sim N(c, I_k)$. The distribution of $V_c'V_c$ is called the noncentral χ^2_k distribution with the noncentrality parameter $\|c\|^2$ (written as $V_c'V_c \sim \chi^2_k\left(\|c\|^2\right)$). If $c = 0$ we have the usual (central) χ^2_k distribution. The probability $P\left(V_c'V_c > \chi^2_{k,1-\alpha}\right) > \alpha$, where $\chi^2_{k,1-\alpha}$ is the $1-\alpha$ quantile of the central χ^2_k distribution and $c \neq 0$. Further, $P\left(V_c'V_c > \chi^2_{k,1-\alpha}\right)$ is an increasing function of $\|c\|^2$.

To show that the distribution of $V_c'V_c$ depends only on $\|c\|^2$ (a scalar parameter), first, write $V_c = c + \mathcal{Z}$, where $\mathcal{Z} \sim N(0, I_k)$. Next, we will show that there is a standard normal random vector $\mathcal{X} \sim N(0, I_k)$ such that, for any $u \in \mathbb{R}$,

$$P\left(\sum_{j=1}^k (c_j + \mathcal{Z}_j)^2 > u\right) = P\left((\|c\| + \mathcal{X}_1)^2 + \sum_{j=2}^k {\mathcal{X}_j}^2 > u\right).$$

Let $b_1, \ldots, b_k$ be orthonormal k-vectors ($b_i' b_i$=1 and $b_i' b_j = 0$ for $i \neq j$) such that $b_1 = c / \|c\|$. Define a $k \times k$ matrix

$$B = \begin{bmatrix} \frac{c'}{\|c\|} \\ b_2' \\ \vdots \\ b_k' \end{bmatrix},$$

and note that $BB' = I_k$, $(B^{-1})' B^{-1} = I_k$, and

$$Bc = \begin{bmatrix} \|c\| \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Next, define

$$\mathcal{X} = B \mathcal{Z}.$$

It follows that

$$\mathcal{X} \sim N(0, BB') =^d N(0, I_k).$$

Furthermore,

$$\begin{aligned} (c + \mathcal{Z})'(c + \mathcal{Z}) &= (B(c + \mathcal{Z}))' \left(B^{-1}\right)' B^{-1} (B(c + \mathcal{Z})) \\ &= (Bc + \mathcal{X})'(Bc + \mathcal{X}) \\ &= (\|c\| + \mathcal{X}_1)^2 + \sum_{j=2}^{k} \mathcal{X}_j{}^2, \end{aligned}$$

and the result follows. Note also that $P\left((\|c\| + \mathcal{X}_1)^2 + \sum_{j=2}^{k} \mathcal{X}_j{}^2 > u\right)$ is increasing in $\|c\|$.

From (11), one can show that under the local alternative $H_1 : \beta = \beta_0 + \delta/\sqrt{n}$,

$$W_n \to_d \chi_k^2 \left(\delta' \left(V(A)\right)^{-1} \delta\right). \tag{12}$$

Hence,

$$\begin{aligned} P\left(W_n > \chi^2_{k,1-\alpha}\right) &\to \rho\left(\delta' \left(V(A)\right)^{-1} \delta\right) \\ &\equiv P\left(\left(\sqrt{\delta' \left(V(A)\right)^{-1} \delta} + \mathcal{X}_1\right)^2 + \sum_{j=2}^{k} \mathcal{X}_j{}^2 > \chi^2_{k,1-\alpha}\right) \end{aligned}$$

where $\mathcal{X}_1, \ldots, \mathcal{X}_k$ are iid $N(0,1)$, and therefore $\rho(\cdot)$ is a non-decreasing function such that

$$\alpha \leq \rho\left(\delta' \left(V(A)\right)^{-1} \delta\right) \leq 1,$$

with $\rho(\delta' \left(V(A)\right)^{-1} \delta) = \alpha$ if and only if $\delta = 0$. Consequently, when the null hypothesis is wrong, a test based on the Wald statistic will reject the null with asymptotic probability higher than α (the asymptotic size of the test). In this case, we say that the test has nontrivial power against local alternatives. The probability to reject the null depends on the magnitude of the noncentrality parameter $\delta' \left(V(A)\right)^{-1} \delta$. Thus, it increases with the distance from the null δ and decreases with the variance V.

The local alternatives framework can be used to compare various tests. For example, since the power of the Wald test increases with the noncentrality parameter, as we mentioned above, the asymptotically efficient test (in the GMM sense) corresponds to $V(A) = \left(Q' \Omega^{-1} Q\right)^{-1}$.

Testing overidentified restrictions

In this section, we discuss a *specification test* that allows one to test whether the moment condition $EZ_iU_i = 0$ holds. Contrary to the tests discussed before, this is not a test of whether β takes on some specific value, but rather whether the model, as defined by the moment conditions, is correctly specified. The null hypothesis is that there exists some β such that $EZ_i\left(Y_i - X_i'\beta\right) = 0$. The alternative hypothesis is that $EZ_i\left(Y_i - X_i'\beta\right) \neq 0$ for all $\beta \in R^k$. Note that, when the model is exactly identified, the system of k equations in k unknowns $EZ_i\left(Y_i - X_i'\beta\right) = 0$ can be solved exactly. Thus, in this sense an exactly identified model is never misspecified (see, for example, Hall and Inoue, *Journal of Econometrics*, 2003). Thus, we can test validity of moment restrictions only if the model is overidentified. In the linear framework discussed here, the overidentified restrictions test is often interpreted as a test of whether the instruments are exogenous.

When the model is overidentified, in general, it is impossible to choose b such that $n^{-1}\sum_{i=1}^n Z_i\left(Y_i - X_i'b\right)$ is exactly zero. However, if the moment condition holds, we should expect that $n^{-1}\sum_{i=1}^n Z_i\left(Y_i - X_i'\beta\right)$ is close to zero, and further,

$$n^{-1/2}\sum_{i=1}^n Z_i\left(Y_i - X_i'\beta\right) \to_d \ N\left(0, \Omega\right).$$

If we use the efficient matrix A_n, then

$$A_n'A_n \to_p \Omega^{-1}. \tag{13}$$

In this case, the weighted distance

$$\left(n^{-1/2}\sum_{i=1}^n Z_i\left(Y_i - X_i'\beta\right)\right)' A_n'A_n \left(n^{-1/2}\sum_{i=1}^n Z_i\left(Y_i - X_i'\beta\right)\right)$$

asymptotically has the χ^2_l distribution (the degrees of freedom are determined by the l moment restrictions). It turns out that, when β is replaced by its *efficient* GMM estimator $\widehat{\beta}_n^{GMM}$, the degrees of freedom change from l to $l-k$. We have the following result. Under the null hypothesis $H_0 : EZ_i\left(Y_i - X_i'\beta\right) = 0$ for some $\beta \in R^k$, and provided that A_n satisfies (13) and $\widehat{\beta}_n^{GMM}$ is efficient,

$$\left(n^{-1/2}\sum_{i=1}^n Z_i\left(Y_i - X_i'\widehat{\beta}_n^{GMM}\right)\right)' A_n'A_n \left(n^{-1/2}\sum_{i=1}^n Z_i\left(Y_i - X_i'\widehat{\beta}_n^{GMM}\right)\right) \to_d \chi^2_{l-k}.$$

The reason for change in degrees of freedom is that we have to estimate k parameters β before construction the test statistic. Another explanation is that we need k restrictions to estimate β. Thus, we can test only additional (overidentified) $l-k$ restrictions.

Consider the linear and homoskedastic case. The efficient GMM estimator is the 2SLS estimator, and the efficient weight matrix is given by $\left(\sum_{i=1}^n Z_iZ_i'\right)^{-1}$. One should reject the null of correctly specified model if

$$\begin{aligned} & n^{-1/2}\sum_{i=1}^n \widehat{U}_iZ_i'\left(n^{-1}\sum_{i=1}^n Z_iZ_i'\right)^{-1} n^{-1/2}\sum_{i=1}^n \widehat{U}_iZ_i/\widehat{\sigma}_n^2 \\ = & \left(\sum_{i=1}^n\left(Y_i - X_i'\widehat{\beta}_n^{GMM}\right)Z_i\right)'\left(\sum_{i=1}^n Z_iZ_i'\right)^{-1}\left(\sum_{i=1}^n\left(Y_i - X_i'\widehat{\beta}_n^{GMM}\right)Z_i\right)/\widehat{\sigma}_n^2 \\ > & \chi^2_{l-k,1-\alpha}, \end{aligned}$$

where $\widehat{\sigma}_n^2$ is any consistent estimator of $\sigma^2 = EU_i^2$, such as $n^{-1}\sum_{i=1}^n\left(Y_i - X_i'\widehat{\beta}_n^{GMM}\right)^2$.

Efficient instruments

The presentation in this section follows Shinichi Sakata's lecture notes.

In section, we consider a linear IV regression model, however, with a stronger exogeneity condition:

$$\begin{aligned} Y_i &= X_i'\beta + U_i, \\ E\left(U_i|Z_i\right) &= 0. \end{aligned}$$

The above restriction on the conditional mean of U_i is also known as "strong exogeneity" as opposed to "weak exogeneity" in (1). While mean independence in the conditional moment restriction is substantially stronger than just uncorrelatedness of weak exogeneity, in practice, exogeneity of instruments is often motivated by arguing statistical independence between instruments and errors, which also implies mean independence. It is obvious from the previous discussion that the conditional moment restriction is stronger than what is need to obtain a consistent estimator of β. However, we will see that the conditional mean restriction has important efficiency implications.

Strong exogeneity implies that Z_i can be potentially used as in an instrument, as mean independence implies uncorrelatedness. However, it also implies that functions of Z_i are uncorrelated with U_i as well:

$$Eg(Z_i)U_i = EE\left(g(Z_i)U_i|Z_i\right) = E\left(g(Z_i)E\left(U_i|Z_i\right)\right) = 0,$$

Hence, in this framework we have potentially infinitely many instruments.

In practice, of course, one can construct a GMM estimator only with finitely many instruments. An interesting question is, therefore, if there exists a vector-valued function $g^*(Z_i)$ that can be used as an instrument to produce the most efficient GMM estimator of β among infinitely many possible GMM estimators corresponding to different functions $g(\cdot)$.

Suppose we use $g(Z_i)$ as an instrument, i.e. we construct an asymptotically efficient GMM estimator for β based on the following *unconditional* moment restriction:

$$Eg(Z_i)\left(Y_i - X_i'\beta\right) = 0.$$

Let $\hat{\beta}_{g,n}$ denote an asymptotically efficient GMM estimator corresponding to $g(\cdot)$. Its asymptotic variance is

$$V_g = \left(Q_g'\Omega_g^{-1}Q_g\right)^{-1},$$

where

$$\begin{aligned} Q_g &= Eg(Z_i)X_i', \\ \Omega_g &= EU_i^2 g(Z_i)g(Z_i)', \end{aligned}$$

and we assume that Q_g has rank k and Ω_g is positive definite.

Define

$$g^*(Z_i) = \frac{E(X_i|Z_i)}{E(U_i^2|Z_i)},$$

and note that $g^*(\cdot)$ is a k-vector valued function. Next, let

$$\begin{aligned} Q^* &= Eg^*(Z_i)X_i', \\ \Omega^* &= EU_i^2 g^*(Z_i)g^*(Z_i)'. \end{aligned}$$

We have:

$$\begin{aligned} Q^* &= E\left(\frac{E(X_i|Z_i)}{E(U_i^2|Z_i)}X_i'\right) \\ &= E\left(\frac{E(X_i|Z_i)E(X_i|Z_i)'}{E(U_i^2|Z_i)}\right), \end{aligned}$$

where the last equality holds by the law of iterated expectation. Further,

$$\begin{aligned}\Omega^* &= E\left(U_i^2 \frac{E(X_i|Z_i)E(X_i|Z_i)'}{\left(E(U_i^2|Z_i)\right)^2}\right) \\ &= E\left(\frac{E(X_i|Z_i)E(X_i|Z_i)'}{E(U_i^2|Z_i)}\right) \\ &= Q^*.\end{aligned}$$

Hence, since with $g^*(Z_i)$ the model is exactly identified, its corresponding efficient GMM estimator $\hat{\beta}^*$ has the asymptotic variance

$$V^* = \left(Q^{*'}\left(\Omega^*\right)^{-1} Q*\right)^{-1} = \left(Q^*\right)^{-1} = \left(\Omega^*\right)^{-1}.$$

We will show next that

$$V_g - V^* \geq 0. \tag{14}$$

To prove the claim, we will use the following lemma.

Lemma 1 *Let e and ε be two random l-vectors such that $Ee\varepsilon' = Eee'$. Then $E\varepsilon\varepsilon' - Eee' \geq 0$.*

Proof. Consider $E(\varepsilon - e)(\varepsilon - e)'$. It is a matrix of second moments and, therefore, is positive semidefinite. Hence,

$$\begin{aligned}0 &\leq E(\varepsilon - e)(\varepsilon - e)' \\ &= E\varepsilon\varepsilon' + Eee' - Ee\varepsilon' - E\varepsilon e' \\ &= E\varepsilon\varepsilon' - Eee',\end{aligned}$$

where the equality in the last line follows by $Ee\varepsilon' = Eee' = E\varepsilon e'$. ■

To show (14) using Lemma 1, let

$$\begin{aligned}\varepsilon &\equiv \left(Q_g'\Omega_g^{-1}Q_g\right)^{-1} Q_g'\Omega_g^{-1}U_i g(Z_i), \\ e &\equiv \left(\Omega^*\right)^{-1} U_i g^*(Z_i).\end{aligned}$$

We have

$$\begin{aligned}E\varepsilon\varepsilon' &= \left(Q_g'\Omega_g^{-1}Q_g\right)^{-1} = V_g, \\ Eee' &= \left(\Omega^*\right)^{-1} = V^*.\end{aligned}$$

Next,

$$E\varepsilon e' = \left(Q_g'\Omega_g^{-1}Q_g\right)^{-1} Q_g'\Omega_g^{-1} E\left(U_i^2 g(Z_i) g^{*'}(Z_i)\right)\left(\Omega^*\right)^{-1}.$$

Furthermore,

$$\begin{aligned}E\left(u_i^2 g(Z_i) g^{*'}(Z_i)\right) &= E\left(\frac{U_i^2 g(Z_i) E(X_i'|Z_i)}{E(U_i^2|Z_i)}\right) \\ &= E\left(\frac{E(U_i^2|Z_i) g(Z_i) E(X_i'|Z_i)}{E(U_i^2|Z_i)}\right) \\ &= E\left(g(Z_i)E(X_i'|Z_i)\right) \\ &= E\left(g(Z_i)X_i'\right) \\ &= Q_g,\end{aligned}$$

where the equalities in the second and fourth lines follow by the law of iterated expectation. Therefore, $E\varepsilon e' = \left(\Omega^*\right)^{-1} = V^*$, and (14) follows by Lemma 1.

In general, efficient instrument $E(X_i|Z_i)/E(U_i^2|Z_i)$ is unknown as it depends on the unknown joint distributions of X_i and Z_i, and U_i and Z_i. It is however can be estimated nonparametric kernel methods. To illustrate the approach, suppose for simplicity that $l = k = 1$. An estimator for $E(X_i|Z_i = c)$ is

$$\hat{m}(c) = \frac{\sum_{j=1}^n X_j K((Z_j - c)/h)}{\sum_{j=1}^n K((Z_j - c)/h)},$$

where $K(\cdot)$ is a *kernel function*, and h is the *bandwidth*. It is usually assumed that $K(\cdot)$ is non-negative, symmetric around zero, integrates to one, and $K(u) = 0$ outside $[-1, 1]$ interval. For example, $K(u) = 1\{-1 \leq u \leq 1\}/2$ (uniform kernel). Bandwidth h is a tuning parameter assigned some small positive value. Since $K(\cdot)$ is compactly supported on $[-1, 1]$, only observations j with $c - h \leq Z_j \leq c + h$ will have a non-zero weight in the formula for $\hat{m}(c)$. Hence, $\hat{m}(c)$ is a *local* average of X's that have Z's close to c.

Estimator $\hat{m}(c)$ is an example of *nonparametric* regression. Consistency of $\hat{m}$ is obtained by requiring that $h \to 0$ and $nh \to \infty$ as $n \to \infty$. One can think about nh as the effective sample size. Because of that, the convergence rate of $\hat{m}(c)$ is therefore $1/\sqrt{nh}$, which is slower than the usual rate of $1/\sqrt{n}$.

Note that to construct an efficient instrument, one has to compute $\hat{m}(Z_i)$ for each $i = 1, \ldots, n$. To estimate $E(U_i^2|Z_i)$, one can use, first, a consistent (but inefficient) estimator of β to construct $\hat{U}_i$. Then, $E(U_i^2|Z_i)$ can be estimated as a nonparametric regression of $\hat{U}_i^2$ against Z_i.

When Z_i is a vector, one can use products of univariate kernels to construct multivariate kernels.