Chapter 8. Single-Equation GMM*

The LSE, the GLS estimator, the MLE, the IV estimator and the 2SLS estimator are all special cases of the generalized method of moments (GMM) estimator. This estimator is hinted in, e.g., Sargan (1958), Amemiya (1974) and White (1982b), but a formal development is usually credited to Hansen (1982). In statistics, a related estimator is the generalized estimating equations (GEE) estimator of Liang and Zeger (1986).

This chapter covers the single-equation generalized method of moments (GMM). Related materials can be found in Chapter 3 of Hayashi (2000), Chapters 21 and 22 of Ruud (2000), Chapter 6 of Cameron and Trivedi (2005), Chapter 8 of Wooldrige (2010), and Chapter 13 of Hansen (2022). For an intuitive introduction on GMM, see Alastair Hall (1993) and Wooldridge (2001); for discussion on empirical application issues, see Ogaki (1993); for a more comprehensive treatment of GMM, see Mátyás (1999) and Alastair Hall (2005).

1 GMM Estimator

We consider the linear model in this section. In a linear model,

$$E\left[g(\mathbf{w}_i, \boldsymbol{\beta})\right] = E\left[\mathbf{z}_i\left(y_i - \mathbf{x}_i'\boldsymbol{\beta}\right)\right] = \mathbf{0}, \tag{1}$$

where $g(\cdot,\cdot)$ is a set of moment conditions, and $\mathbf{w}_i = (y_i, \mathbf{x}_i', \mathbf{z}_i')'$. This is the instrument exogeneity condition $E\left[\mathbf{z}_i u_i\right] = \mathbf{0}$ in the endogenous linear regression model $y_i = \mathbf{x}_i'\boldsymbol{\beta} + u_i$ with $E[\mathbf{x}_i u_i] \neq \mathbf{0}$. Define the sample analog of (1)

$$\overline{g}_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} g_i(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i\left(y_i - \mathbf{x}_i'\boldsymbol{\beta}\right) = \frac{1}{n}\left(\mathbf{Z}'\mathbf{y} - \mathbf{Z}'\mathbf{X}\boldsymbol{\beta}\right).$$

When $l > k$, we cannot solve $\overline{g}_n(\boldsymbol{\beta}) = \mathbf{0}$ exactly as intuitively shown in Figure 1. The idea of the GMM is to define an estimator which sets $\overline{g}_n(\boldsymbol{\beta})$ "close" to zero.

For some $l \times l$ weight matrix $\mathbf{W}_n > 0$, let

$$J_n\left(\boldsymbol{\beta}\right) = n \cdot \overline{g}_n(\boldsymbol{\beta})'\mathbf{W}_n\overline{g}_n(\boldsymbol{\beta}). \tag{2}$$

This is a non-negative measure of the "length" of the vector $\overline{g}_n(\boldsymbol{\beta})$ under the inner product $\langle\cdot,\cdot\rangle_{\mathbf{W}_n}$ in Section 3 of Chapter 2. For example, if $\mathbf{W}_n = \mathbf{I}_l$, then, $J_n\left(\boldsymbol{\beta}\right) = n \cdot \overline{g}_n(\boldsymbol{\beta})'\overline{g}_n(\boldsymbol{\beta}) = n\left\|\overline{g}_n(\boldsymbol{\beta})\right\|^2$,

*Email: pingyu@hku.hk

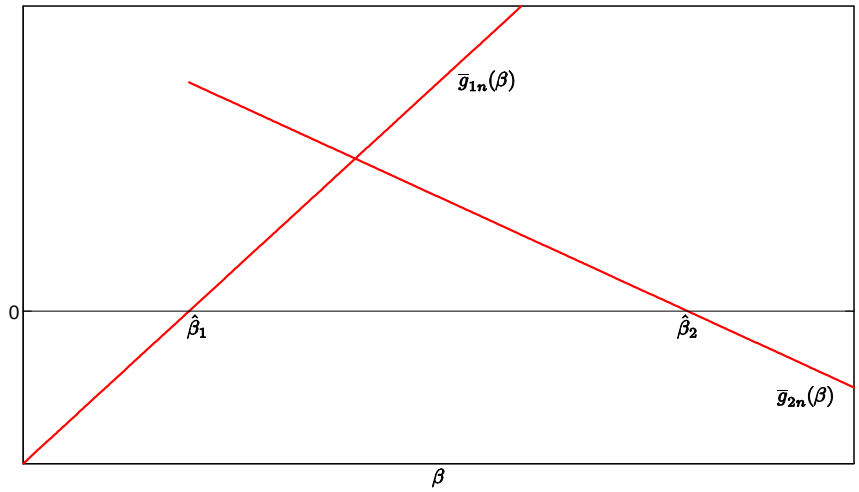

Figure 1: $\overline{g}_n(\beta) = 0$ Can Not Hold Exactly for Any β: $k = 1, l = 2$

the square of the Euclidean length. The GMM estimator minimizes $J_n(\boldsymbol{\beta})$. Note that if $l = k$, then $\overline{g}_n(\boldsymbol{\beta}) = \mathbf{0}$ can be solved exactly. The GMM estimator reduces to the MoM estimator (the IV estimator) and $\mathbf{W}_n$ is not required. The first order conditions for the GMM estimator are

$$\begin{aligned} \mathbf{0} &= \frac{\partial}{\partial \boldsymbol{\beta}} J_n\left(\widehat{\boldsymbol{\beta}}\right) = 2n \frac{\partial}{\partial \boldsymbol{\beta}} \overline{g}_n'(\widehat{\boldsymbol{\beta}}) \mathbf{W}_n \overline{g}_n(\widehat{\boldsymbol{\beta}}) \\ &= -2n \left(\frac{1}{n}\mathbf{X}'\mathbf{Z}\right) \mathbf{W}_n \left(\frac{1}{n}\left(\mathbf{Z}'\mathbf{y} - \mathbf{Z}'\mathbf{X}\widehat{\boldsymbol{\beta}}\right)\right), \end{aligned}$$

so

$$\widehat{\boldsymbol{\beta}}_{GMM} = \left[\left(\mathbf{X}'\mathbf{Z}\right)\mathbf{W}_n\left(\mathbf{Z}'\mathbf{X}\right)\right]^{-1}\left[\left(\mathbf{X}'\mathbf{Z}\right)\mathbf{W}_n\left(\mathbf{Z}'\mathbf{y}\right)\right]. \tag{3}$$

While the estimator depends on $\mathbf{W}_n$, the dependence is only up to scale, for if $\mathbf{W}_n$ is replaced by $c\mathbf{W}_n$ for some $c > 0$, $\widehat{\boldsymbol{\beta}}_{GMM}$ does not change. In Section 4 of Chapter 7, $\boldsymbol{\beta}$ is identified as $(\boldsymbol{\Gamma}'\mathbf{A}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}'\mathbf{A}\boldsymbol{\lambda} = \left(E\left[\mathbf{x}_i\mathbf{z}_i'\right]E[\mathbf{z}_i\mathbf{z}_i']^{-1}\mathbf{A}E[\mathbf{z}_i\mathbf{z}_i']^{-1}E\left[\mathbf{z}_i\mathbf{x}_i'\right]\right)^{-1}E\left[\mathbf{x}_i\mathbf{z}_i'\right]E[\mathbf{z}_i\mathbf{z}_i']^{-1}\mathbf{A}E[\mathbf{z}_i\mathbf{z}_i']^{-1}E[\mathbf{z}_i y_i]$, so there, $\mathbf{W}_n$ is the sample analog of $E[\mathbf{z}_i\mathbf{z}_i']^{-1}\mathbf{A}E[\mathbf{z}_i\mathbf{z}_i']^{-1}$. When $\mathbf{A} = E[\mathbf{z}_i\mathbf{z}_i']$, we obtain the 2SLS estimator, that is, $\mathbf{W}_n = (\mathbf{Z}'\mathbf{Z})^{-1}$.

From the FOCs of GMM estimation, we can see that although we cannot make $\overline{g}_n(\boldsymbol{\beta}) = \mathbf{0}$ exactly, we could let some of its linear combinations, say $\mathbf{B}_n\overline{g}_n(\boldsymbol{\beta})$, be zero, where $\mathbf{B}_n$ is a $k \times l$ matrix. For a weight matrix $\mathbf{W}_n$, $\mathbf{B}_n = \left(\frac{1}{n}\mathbf{X}'\mathbf{Z}\right)\mathbf{W}_n$. If $\mathbf{W}_n \xrightarrow{p} \mathbf{W} > 0$, and $\frac{1}{n}\mathbf{X}'\mathbf{Z} \xrightarrow{p} E\left[\mathbf{x}_i\mathbf{z}_i'\right] = \mathbf{G}'$, $\mathbf{B}_n$ converges to $\mathbf{B} = \mathbf{G}'\mathbf{W}$. So $\widehat{\boldsymbol{\beta}}_{GMM}$ is as if defined by a MoM estimator such that $\mathbf{B}\overline{g}_n(\widehat{\boldsymbol{\beta}}) = \mathbf{0}$. Equivalently, $\widehat{\boldsymbol{\beta}}_{GMM}$ is using k instruments $\mathbf{Bz}$, a linear combination of $\mathbf{z}$, to estimate $\boldsymbol{\beta}$.

2 Distribution of the GMM Estimator

Note that

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{GMM}-\boldsymbol{\beta} &= \left[\left(\frac{1}{n}\mathbf{X}'\mathbf{Z}\right)\mathbf{W}_n\left(\frac{1}{n}\mathbf{Z}'\mathbf{X}\right)\right]^{-1}\left[\left(\frac{1}{n}\mathbf{X}'\mathbf{Z}\right)\mathbf{W}_n\left(\frac{1}{n}\mathbf{Z}'\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)\right)\right]\\ &= \left[\left(\frac{1}{n}\mathbf{X}'\mathbf{Z}\right)\mathbf{W}_n\left(\frac{1}{n}\mathbf{Z}'\mathbf{X}\right)\right]^{-1}\left[\left(\frac{1}{n}\mathbf{X}'\mathbf{Z}\right)\mathbf{W}_n\left(\frac{1}{n}\mathbf{Z}'\mathbf{u}\right)\right].\end{aligned}$$

Now,

$$\left(\frac{1}{n}\mathbf{X}'\mathbf{Z}\right)\mathbf{W}_n\left(\frac{1}{n}\mathbf{Z}'\mathbf{X}\right)\xrightarrow{p}\mathbf{G}'\mathbf{W}\mathbf{G}$$

and

$$\left(\frac{1}{n}\mathbf{X}'\mathbf{Z}\right)\mathbf{W}_n\left(\frac{1}{\sqrt{n}}\mathbf{Z}'\mathbf{u}\right)\xrightarrow{d}\mathbf{G}'\mathbf{W}N\left(\mathbf{0},\boldsymbol{\Omega}\right),$$

where $\Omega = E\left[\mathbf{z}_i\mathbf{z}_i'u_i^2\right] = E\left[g_ig_i'\right]$ with $g_i = \mathbf{z}_iu_i$. So

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_{GMM}-\boldsymbol{\beta}\right)\xrightarrow{d}N\left(\mathbf{0},\mathbf{V}\right),$$

where

$$\mathbf{V}=\left(\mathbf{G}'\mathbf{W}\mathbf{G}\right)^{-1}\left(\mathbf{G}'\mathbf{W}\boldsymbol{\Omega}\mathbf{W}\mathbf{G}\right)\left(\mathbf{G}'\mathbf{W}\mathbf{G}\right)^{-1}. \tag{4}$$

In general, GMM estimators are asymptotically normal with "sandwich form" asymptotic variances. It is easy to check this asymptotic distribution is the same as the MoM estimator defined by $\mathbf{B}\overline{g}_n(\widehat{\boldsymbol{\beta}})=\mathbf{0}$.

Exercise 1 *Suppose the moment conditions are $E\left[g(\mathbf{w}_i,\boldsymbol{\beta})\right]=\mathbf{0}$ with $g(\mathbf{w},\boldsymbol{\beta})=g_1(\mathbf{w})-g_2(\mathbf{w})\boldsymbol{\beta}$. Set up $J_n(\boldsymbol{\beta})$ as in (2) and derive the asymptotic distribution of the corresponding GMM estimator of $\boldsymbol{\beta}$.*

A natural question is what is the optimal weight matrix $\mathbf{W}_0$ that minimizes $\mathbf{V}$. This turns out to be $\boldsymbol{\Omega}^{-1}$. The proof is left as an exercise. This yields the efficient GMM estimator:

$$\widehat{\boldsymbol{\beta}}=\left(\mathbf{X}'\mathbf{Z}\boldsymbol{\Omega}^{-1}\mathbf{Z}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Z}\boldsymbol{\Omega}^{-1}\mathbf{Z}'\mathbf{y},$$

which has the asymptotic variance $\mathbf{V}_0=\left(\mathbf{G}'\boldsymbol{\Omega}^{-1}\mathbf{G}\right)^{-1}$. This corresponds to the linear combination matrix $\mathbf{B}=\mathbf{G}'\boldsymbol{\Omega}^{-1}$.

$\mathbf{W}_0=\boldsymbol{\Omega}^{-1}$ is usually unknown in practice, but it can be estimated consistently. For any $\mathbf{W}_n\xrightarrow{p}\mathbf{W}_0$, we still call $\widehat{\boldsymbol{\beta}}$ the efficient GMM estimator, as it has the same asymptotic distribution.

Exercise 2 *In the linear model estimated by GMM with general weight matrix $\mathbf{W}$, the asymptotic variance of $\widehat{\boldsymbol{\beta}}_{GMM}$ is $\mathbf{V}$ in (4).*

(i) *Let $\mathbf{V}_0$ be this matrix when $\mathbf{W}=\boldsymbol{\Omega}^{-1}$. Show that $\mathbf{V}_0=\left(\mathbf{G}'\boldsymbol{\Omega}^{-1}\mathbf{G}\right)^{-1}$.*

(ii) *We want to show that for any* $\mathbf{W}$, $\mathbf{V} - \mathbf{V}_0 \geq 0$. *To do this, start by finding matrices* $\mathbf{A}$ *and* $\mathbf{B}$ *such that* $\mathbf{V} = \mathbf{A}'\mathbf{\Omega}\mathbf{A}$ *and* $\mathbf{V}_0 = \mathbf{B}'\mathbf{\Omega}\mathbf{B}$.

(iii) *Show that* $\mathbf{B}'\mathbf{\Omega}\mathbf{A} = \mathbf{B}'\mathbf{\Omega}\mathbf{B}$ *and therefore* $\mathbf{B}'\mathbf{\Omega}(\mathbf{A} - \mathbf{B}) = \mathbf{0}$.

(iv) *Use the expressions* $\mathbf{V} = \mathbf{A}'\mathbf{\Omega}\mathbf{A}$, $\mathbf{A} = \mathbf{B} + (\mathbf{A} - \mathbf{B})$, *and* $\mathbf{B}'\mathbf{\Omega}(\mathbf{A} - \mathbf{B}) = \mathbf{0}$ *to show that* $\mathbf{V} \geq \mathbf{V}_0$.

Exercise 3 *Show that when a new group of instrumental variables is added in, the optimal asymptotic variance matrix* $\mathbf{V}_0$ *will not increase. Discuss when the two asymptotic variance matrices will be equal. (Hint: use the result in Exercise 2.)*

In the homoskedastic case, $E\left[u_i^2|\mathbf{z}_i\right] = \sigma^2$, then $\mathbf{\Omega} = E\left[\mathbf{z}_i\mathbf{z}_i'\right]\sigma^2 \propto E\left[\mathbf{z}_i\mathbf{z}_i'\right]$ suggesting the weight matrix $\mathbf{W}_n = (\mathbf{Z}'\mathbf{Z})^{-1}$, which generates the 2SLS estimator. So the 2SLS estimator is the efficient GMM estimator under homoskedasticity. When the heteroskedasticity is present, the optimal weight matrix $\mathbf{\Omega}^{-1}$ explores also the potential information of correlation between the squared error, u_i^2, and the cross-products of the instrumental variables, $\mathbf{z}_i\mathbf{z}_i'$. Testing $E\left[u_i^2|\mathbf{z}_i\right] = \sigma^2$ can be similarly conducted as in testing homoskedasticity in linear regression. Nevertheless, we need to make an additional assumption to simplify the asymptotic arguments, i.e., $Cov(\mathbf{x}_i, u_i|\mathbf{z}_i)$ is a constant.[1] Without this assumption, the tests for heteroskedasticity are more complicated; see Wooldridge (1990) for the details.

Exercise 4 *Take the single equation*

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, E[\mathbf{u}|\mathbf{Z}] = \mathbf{0}.$$

Assume $E[u_i^2|\mathbf{z}_i] = \sigma^2$. *Show that if* $\widehat{\boldsymbol{\beta}}$ *is estimated by GMM with weight matrix* $\mathbf{W}_n = (\mathbf{Z}'\mathbf{Z})^{-1}$, *then*

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{d} N\left(0, \sigma^2\left(\mathbf{G}'\mathbf{M}^{-1}\mathbf{G}\right)^{-1}\right),$$

where $\mathbf{G} = E[\mathbf{z}_i\mathbf{x}_i']$ *and* $\mathbf{M} = E[\mathbf{z}_i\mathbf{z}_i']$.

The following example illustrates why $\mathbf{W}_0 = \mathbf{\Omega}^{-1}$.

Example 1 (Optimal Weight Matrix) *Suppose* $E[x_i] = E[y_i] = \mu$ *and* $Cov(x_i, y_i) = 0$. *We try to find an efficient GMM estimator for* μ. *First, sort out moment conditions* $E[g(\mathbf{w}_i, \mu)] = \mathbf{0}$, *where* $\mathbf{w}_i = (x_i, y_i)'$:

$$g(\mathbf{w}_i, \mu) = \begin{pmatrix} x_i - \mu \\ y_i - \mu \end{pmatrix}.$$

Since μ *appears in both moment conditions, we hope to find a better estimator than* $\overline{x}$ *or* $\overline{y}$ *which uses only one moment condition. Of course,* $E[g(\mathbf{w}_i, \mu)] = \mathbf{0}$ *uses extra information (*x_i *and* y_i *have a common mean) about* μ, *and not robust to such information.*

[1] When $\mathbf{z}_i = \mathbf{x}_i$ as in linear regression, $Cov(\mathbf{x}_i, u_i|\mathbf{z}_i) = \mathbf{0}$, so such an assumption is not required.

We can use $\overline{x}$ or $\overline{y}$ to estimate μ, but a weighted average may be better. Suppose $\widehat{\mu} = \omega\overline{x} + (1-\omega)\,\overline{y}$, and x_i and y_i are uncorrelated; then the asymptotic distribution of $\widehat{\mu}$ is

$$\sqrt{n}\,(\widehat{\mu} - \mu) \xrightarrow{d} N\left(0, \omega^2\sigma_x^2 + (1-\omega)^2\,\sigma_y^2\right),$$

where $\sigma_x^2 = Var(x)$ and $\sigma_y^2 = Var(y)$. Minimizing the asymptotic variance, we have

$$\omega = \frac{\sigma_y^2}{\sigma_x^2 + \sigma_y^2}.$$

That is, the sample (of x and y) with a larger variance is given a smaller weight, and the sample with a smaller variance is given a larger weight. (Check that $\widetilde{\mu} = \frac{\overline{x}+\overline{y}}{2}$, which corresponds to $\mathbf{W}_n = \mathbf{I}_2$ in (2), may have a larger asymptotic variance than $\overline{x}$ or $\overline{y}$). The asymptotic variance under this optimal weight is $\frac{\sigma_x^2\sigma_y^2}{\sigma_x^2+\sigma_y^2} \leq \min\left\{\sigma_x^2, \sigma_y^2\right\}$.

From Exercise 1 and 2, the optimal weight matrix

$$\mathbf{W}_0 = E[g(\mathbf{w}_i,\mu)g(\mathbf{w}_i,\mu)']^{-1} = \begin{pmatrix} E[(x_i-\mu)^2] & E[(x_i-\mu)(y_i-\mu)] \\ E[(x_i-\mu)(y_i-\mu)] & E[(y_i-\mu)^2] \end{pmatrix}^{-1} = \begin{pmatrix} \sigma_x^{-2} & 0 \\ 0 & \sigma_y^{-2} \end{pmatrix},$$

so

$$J_n(\mu) = n \cdot \overline{g}_n(\mu)'\mathbf{W}_0\overline{g}_n(\mu) = n\left(\frac{(\overline{x}-\mu)^2}{\sigma_x^2} + \frac{(\overline{y}-\mu)^2}{\sigma_y^2}\right),$$

and

$$\widehat{\mu} = \omega\overline{x} + (1-\omega)\,\overline{y}$$

is the same as the weighted average above.

In practice, σ_x^2 and σ_y^2 are unknown. In this simple example, they can be substituted by their sample analog. The next section deals with the general case. □

Exercise 5 *In Exercise 12 of Chapter 7, find the efficient GMM estimator of β based on the moment condition $E[\mathbf{z}_i(y_i - x_i\beta)] = \mathbf{0}$. Does it differ from 2SLS and/or OLS?*

3 Estimation of the Optimal Weight Matrix

Given any weight matrix $\mathbf{W}_n > 0$, the GMM estimator $\widehat{\boldsymbol{\beta}}_{GMM}$ is consistent yet inefficient. For example, we can set $\mathbf{W}_n = \mathbf{I}_l$. In the linear model, a better choice is $\mathbf{W}_n = (\mathbf{Z}'\mathbf{Z})^{-1}$ which corresponds to the 2SLS estimator. Given any such fist-step estimator, we can define the residuals $\widehat{u}_i = y_i - \mathbf{x}_i'\widehat{\boldsymbol{\beta}}_{GMM}$ and moment equations $\widehat{g}_i = \mathbf{z}_i\widehat{u}_i = g\left(\mathbf{w}_i, \widehat{\boldsymbol{\beta}}_{GMM}\right)$. Construct

$$\begin{aligned} \overline{g}_n &= \overline{g}_n(\widehat{\boldsymbol{\beta}}_{GMM}) = \frac{1}{n}\sum_{i=1}^n \widehat{g}_i, \\ \widehat{g}_i^* &= \widehat{g}_i - \overline{g}_n, \end{aligned}$$

and define

$$\mathbf{W}_n = \left(\frac{1}{n}\sum_{i=1}^{n} \widehat{g}_i^* \widehat{g}_i^{*\prime}\right)^{-1} = \left(\frac{1}{n}\sum_{i=1}^{n} \widehat{g}_i \widehat{g}_i' - \overline{g}_n \overline{g}_n'\right)^{-1}. \tag{5}$$

Then $\mathbf{W}_n \xrightarrow{p} \mathbf{\Omega}^{-1}$, and GMM using $\mathbf{W}_n$ as the weight matrix is asymptotically efficient.

Exercise 6 *Take the model $y_i = \mathbf{x}_i'\boldsymbol{\beta} + u_i$ with $E[\mathbf{z}_i u_i] = \mathbf{0}$. Let $\widehat{u}_i = y_i - \mathbf{x}_i'\widetilde{\boldsymbol{\beta}}$ where $\widetilde{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{\beta}$ (e.g., a GMM estimator with arbitrary weight matrix). Define the estimate of the optimal GMM weight matrix*

$$\mathbf{W}_n = \left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{z}_i \mathbf{z}_i' \widehat{u}_i^2\right)^{-1}.$$

Show that $\mathbf{W}_n \xrightarrow{p} \mathbf{\Omega}^{-1}$ where $\mathbf{\Omega} = E[\mathbf{z}_i \mathbf{z}_i u_i^2]$.

A common alternative choice is to set

$$\mathbf{W}_n = \left(\frac{1}{n}\sum_{i=1}^{n} \widehat{g}_i \widehat{g}_i'\right)^{-1}, \tag{6}$$

which uses the uncentered moment conditions. Since $E[g_i] = \mathbf{0}$, these two estimators are asymptotically equivalent under the hypothesis of correct specification. However, Alastair Hall (2000) has shown that the uncentered estimator is a poor choice. When constructing hypothesis tests, under the alternative hypothesis the moment conditions are violated, i.e. $E[g_i] \neq \mathbf{0}$, so the uncentered estimator will contain an undesirable bias term and the power of the test will be adversely affected. A simple solution is to use the centered moment conditions to construct the weight matrix.

Here is a simple way to compute the efficient GMM estimator for the linear model. First, set $\mathbf{W}_n = (\mathbf{Z}'\mathbf{Z})^{-1}$, estimate $\boldsymbol{\beta}$ using this weight matrix to get the first-step estimator $\widehat{\boldsymbol{\beta}}_{GMM}$, and construct the residual $\widehat{u}_i = y_i - \mathbf{x}_i'\widehat{\boldsymbol{\beta}}_{GMM}$. Then set $\widehat{g}_i = \mathbf{z}_i \widehat{u}_i$, and let $\widehat{g}$ be the associated $n \times l$ matrix. Then the efficient GMM estimator is

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{Z}\left(\widehat{g}'\widehat{g} - n\overline{g}_n\overline{g}_n'\right)^{-1}\mathbf{Z}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Z}\left(\widehat{g}'\widehat{g} - n\overline{g}_n\overline{g}_n'\right)^{-1}\mathbf{Z}'\mathbf{y}.$$

In most cases, when we say "GMM" we actually mean "efficient GMM". There is little point in using an inefficient GMM estimator when the efficient estimator is easy to compute. An estimator of the asymptotic variance of $\widehat{\boldsymbol{\beta}}$ can be seen from the above formula. Set

$$\widehat{\mathbf{V}} = n\left(\mathbf{X}'\mathbf{Z}\left(\widehat{g}'\widehat{g} - n\overline{g}_n\overline{g}_n'\right)^{-1}\mathbf{Z}'\mathbf{X}\right)^{-1}.$$

Asymptotic standard errors are given by the square roots of the diagonal elements of $\widehat{\mathbf{V}}/n$.

Exercise 7 *Suppose we want to estimate σ^2 in Exercise 4. (i) Show that $n^{-1}\sum_{i=1}^{n}\widehat{u}_i^2$ is consistent,*

where $\widehat{u}_i = y_i - \mathbf{x}_i\widehat{\boldsymbol{\beta}}$. *(ii) In the structural model*

$$\begin{aligned} \mathbf{y} &= \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{u}, \\ \mathbf{X}_2 &= \mathbf{X}_1\boldsymbol{\Gamma}_{12} + \mathbf{Z}_2\boldsymbol{\Gamma}_{22} + \mathbf{V}, \end{aligned}$$

show that $\frac{1}{n-l_2}\left(1, -\widehat{\boldsymbol{\beta}}'_{2,2SLS}\right)\mathbf{Y}'\mathbf{M}_\mathbf{Z}\mathbf{Y}\left(1, -\widehat{\boldsymbol{\beta}}'_{2,2SLS}\right)'$ *is also a consistent estimator of* σ^2, *where* $\mathbf{Y} = (\mathbf{y}, \mathbf{X}_2)$.

Given the efficient estimator $\widehat{\boldsymbol{\beta}}$, we can continue to reestimate $\mathbf{W}_n$ by replacing $\widehat{g}_i$ by $g\left(\mathbf{w}_i, \widehat{\boldsymbol{\beta}}\right)$ and construct a new estimator of $\boldsymbol{\beta}$. This is repeated until the $\boldsymbol{\beta}$ estimator converges or enough iterations are conducted. The estimator generated from this procedure is called the *iterative estimator*.

4 Nonlinear GMM

Suppose the moment conditions are

$$E\left[g(\mathbf{w}_i, \boldsymbol{\theta}_0)\right] = \mathbf{0},$$

where $g(\cdot,\cdot) \in \mathbb{R}^l$ is a general nonlinear function of $\theta \in \mathbb{R}^k$, $l \geq k$. The GMM estimator $\widehat{\boldsymbol{\theta}}$ minimizes

$$J_n(\boldsymbol{\theta}) = n \cdot \overline{g}_n(\boldsymbol{\theta})'\mathbf{W}_n\overline{g}_n(\boldsymbol{\theta}),$$

where $\overline{g}_n(\boldsymbol{\theta}) = n^{-1}\sum_{i=1}^n g(\mathbf{w}_i, \boldsymbol{\theta}) \equiv n^{-1}\sum_{i=1}^n g_i(\boldsymbol{\theta})$, and $\mathbf{W}_n$ is a consistent estimator of $\boldsymbol{\Omega}^{-1} \equiv E\left[g_i(\boldsymbol{\theta}_0)g_i(\boldsymbol{\theta}_0)'\right]^{-1}$ which is the optimal weight matrix. For example,

$$\mathbf{W}_n = \left(\frac{1}{n}\sum_{i=1}^n \widehat{g}_i\widehat{g}'_i - \overline{g}_n\overline{g}'_n\right)^{-1}$$

with $\widehat{g}_i = g_i\left(\widetilde{\boldsymbol{\theta}}\right)$ constructed using a preliminary consistent estimator $\widetilde{\boldsymbol{\theta}}$, perhaps obtained by first setting $\mathbf{W}_n = \mathbf{I}_n$. Since the GMM estimator depends upon the first-stage estimator, often the weight matrix $\mathbf{W}_n$ is updated, and then $\widehat{\boldsymbol{\theta}}$ recomputed. This estimator can be iterated if needed.

In Appendix A, we show $\widehat{\boldsymbol{\theta}}$ is CAN under some regularity conditions based on Newey and McFadden (1994). More specifically,

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \xrightarrow{d} N\left(\mathbf{0}, \left(\mathbf{G}'\boldsymbol{\Omega}^{-1}\mathbf{G}\right)^{-1}\right) \equiv N\left(\mathbf{0}, \mathbf{V}\right), \tag{7}$$

where $\mathbf{G} = E\left[\partial g_i(\boldsymbol{\theta}_0)/\partial\boldsymbol{\theta}'\right]$. The asymptotic covariance matrix of $\widehat{\boldsymbol{\theta}}$ can be consistently estimated by $\widehat{\mathbf{V}} \equiv \left(\widehat{\mathbf{G}}'\widehat{\boldsymbol{\Omega}}^{-1}\widehat{\mathbf{G}}\right)^{-1}$, where $\widehat{\boldsymbol{\Omega}} = n^{-1}\sum_{i=1}^n g_i^*(\widehat{\boldsymbol{\theta}})g_i^*(\widehat{\boldsymbol{\theta}})'$ with $g_i^*(\theta) = g_i(\theta) - \overline{g}_n(\boldsymbol{\theta})$, and $\widehat{\mathbf{G}} = n^{-1}\sum_{i=1}^n \partial g_i(\widehat{\boldsymbol{\theta}})/\partial\boldsymbol{\theta}'$.

(*) For the MLE, the one-step estimator is efficient. Similarly, the one-step GMM estimator is

efficient. Such an estimator is defined as

$$\widehat{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}} - \left(\mathbf{G}_n \left(\widetilde{\boldsymbol{\theta}} \right)' \boldsymbol{\Omega}_n^{-1} \left(\widetilde{\boldsymbol{\theta}} \right) \mathbf{G}_n \left(\widetilde{\boldsymbol{\theta}} \right) \right)^{-1} \mathbf{G}_n \left(\widetilde{\boldsymbol{\theta}} \right)' \boldsymbol{\Omega}_n^{-1} \left(\widetilde{\boldsymbol{\theta}} \right) \overline{g}_n(\widetilde{\boldsymbol{\theta}}),$$

where $\widetilde{\boldsymbol{\theta}} = \boldsymbol{\theta}_0 + O_p(1/\sqrt{n})$, $\mathbf{G}_n(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}'} \overline{g}_n(\boldsymbol{\theta})$, and $\boldsymbol{\Omega}_n(\boldsymbol{\theta}) = 1/n \sum_{i=1}^n g(\mathbf{w}_i, \boldsymbol{\theta}) g(\mathbf{w}_i, \boldsymbol{\theta})'$.

Exercise 8 *The equation of interest is*

$$y_i = m(\mathbf{x}_i, \boldsymbol{\beta}) + u_i, E[\mathbf{z}_i u_i] = \mathbf{0}.$$

The observed data is $(y_i, \mathbf{x}_i, \mathbf{z}_i)$, $\mathbf{z}_i$ *is* $l \times 1$ *and* $\boldsymbol{\beta}$ *is* $k \times 1$, $l \geq k$. *Show how to construct an efficient GMM estimator for* $\boldsymbol{\beta}$.

5 Hypothesis Testing

This section summarizes the tests in GMM. We first discuss two specification tests - the Hausman test for the presence of endogeneity and the J test for the validity of overidentifying restrictions. The J test is also called the *Sargan-Hansen test* due to a special case established by Sargan (1958) and the general case by Hansen (1982). We then consider the three asymptotically equivalent tests in the GMM framework - the Wald, Lagrange Multiplier (or Rao's Score), and Likelihood Ratio test. The LR test is also called the *distance test* or the *Newey-West test* due to Newey and West (1987a). These tests are counterparts of those in the likelihood framework (see Section 4 of Chapter 4).

It should be emphasized that a specification test is a test for the *whole* model, not only for the restrictions of interest. Only if it is guaranteed that the rest of the model (e.g., the rank condition) is specified correctly, a rejection of the null is a sign of violation of the interested restrictions. Nevertheless, a rejection of the null is typically cause for concern.

5.1 Testing for Exogeneity: The Hausman Test (*)

The null is $E[\mathbf{x}u] = \mathbf{0}$, i.e., $\mathbf{x}$ is exogenous. If the null is true, then no instruments are needed. Suppose the model is homoskedastic under the null, that is, $E[u^2|\mathbf{x}] = \sigma^2$. Under the null, the LSE is efficient, while under the alternative it is inconsistent. On the other hand, the 2SLS estimator is consistent under both the null and alternative. The Hausman test examines the null by checking for a statistically significant difference between the OLS and 2SLS estimate of $\boldsymbol{\beta}$. There are various versions of this test, three of which are

$$T_j = \left(\widehat{\boldsymbol{\beta}}_{2,2SLS} - \widehat{\boldsymbol{\beta}}_{2,OLS} \right)' \widehat{\mathbf{V}}_j^{-1} \left(\widehat{\boldsymbol{\beta}}_{2,2SLS} - \widehat{\boldsymbol{\beta}}_{2,OLS} \right),$$

where

$$\begin{aligned}
\widehat{\mathbf{V}}_1 &= \left(\widetilde{\mathbf{X}}_2'\mathbf{P}_{\widetilde{\mathbf{Z}}_2}\widetilde{\mathbf{X}}_2\right)^{-1}\widehat{\sigma}^2_{2SLS} - \left(\widetilde{\mathbf{X}}_2'\widetilde{\mathbf{X}}_2\right)^{-1}\widehat{\sigma}^2_{OLS}, \\
\widehat{\mathbf{V}}_2 &= \left[\left(\widetilde{\mathbf{X}}_2'\mathbf{P}_{\widetilde{\mathbf{Z}}_2}\widetilde{\mathbf{X}}_2\right)^{-1} - \left(\widetilde{\mathbf{X}}_2'\widetilde{\mathbf{X}}_2\right)^{-1}\right]\widehat{\sigma}^2_{2SLS}, \\
\widehat{\mathbf{V}}_3 &= \left[\left(\widetilde{\mathbf{X}}_2'\mathbf{P}_{\widetilde{\mathbf{Z}}_2}\widetilde{\mathbf{X}}_2\right)^{-1} - \left(\widetilde{\mathbf{X}}_2'\widetilde{\mathbf{X}}_2\right)^{-1}\right]\widehat{\sigma}^2_{OLS},
\end{aligned}$$

$\widehat{\sigma}^2_{2SLS}$ and $\widehat{\sigma}^2_{OLS}$ are estimates of σ^2 based on the 2SLS and OLS residuals, respectively, $\widetilde{\mathbf{X}}_2 = \mathbf{M}_{\mathbf{X}_1}\mathbf{X}_2$, and $\widetilde{\mathbf{Z}}_2 = \mathbf{M}_{\mathbf{X}_1}\mathbf{Z}_2$. T_2 was proposed by Wu (1973; his T_3 statistic) and by Hausman (1978); T_3 was proposed by Durbin (1954). Under the null, $T_j \xrightarrow{d} \chi^2_\kappa$, where κ =rank$(\mathbf{V}_j)$ and $\mathbf{V}_j$ =plim$\left(\widehat{\mathbf{V}}_j\right)$. See also Smith (1994) for several asymptotically equivalent limited information tests.

5.2 Testing Overidentifying Restrictions: The J Test

The hypotheses are

$$H_0 : \exists\ \boldsymbol{\beta}_0 \text{ s.t. } E[g(\mathbf{w}_i, \boldsymbol{\beta}_0)] = \mathbf{0} \tag{8}$$

versus

$$H_1 :\ \forall\ \boldsymbol{\beta} \in \mathcal{B},\ E[g(\mathbf{w}_i, \boldsymbol{\beta})] \neq \mathbf{0},$$

where $\mathcal{B}$ is the parameter space. When $l = k$, there always exists a $\boldsymbol{\beta}_0 \in \mathcal{B}$ such that $E[g(\mathbf{w}_i, \boldsymbol{\beta}_0)] = \mathbf{0}$. So only if $l > k$, we need this test - to test whether the overidentifying restrictions are valid.

For example, take the linear model $y_i = \mathbf{x}_{1i}'\boldsymbol{\beta}_1 + \mathbf{x}_{2i}'\boldsymbol{\beta}_2 + u_i$ with $E[\mathbf{x}_{1i}u_i] = \mathbf{0}$ and $E[\mathbf{x}_{2i}u_i] = \mathbf{0}$. It is possible that $\boldsymbol{\beta}_2 = \mathbf{0}$, so that the linear equation may be written as $y_i = \mathbf{x}_{1i}'\boldsymbol{\beta}_1 + u_i$. However, it is possible that $\boldsymbol{\beta}_2 \neq \mathbf{0}$, and in this case it would be impossible to find a value of $\boldsymbol{\beta}_1$ so that $E[\mathbf{x}_{1i}\,(y_i - \mathbf{x}_{1i}'\boldsymbol{\beta}_1)] = \mathbf{0}$ and $E[\mathbf{x}_{2i}\,(y_i - \mathbf{x}_{1i}'\boldsymbol{\beta}_1)] = \mathbf{0}$ hold simultaneously. In this sense an exclusion restriction ($\boldsymbol{\beta}_2 = \mathbf{0}$) can be seen as an overidentifying restriction.

Note that $\overline{g}_n(\widehat{\boldsymbol{\beta}}) \xrightarrow{p} E[g_i(\boldsymbol{\beta}_0)]$, and thus $\overline{g}_n(\widehat{\boldsymbol{\beta}})$ can be used to assess whether or not the hypothesis that $E[g_i(\boldsymbol{\beta}_0)] = \mathbf{0}$ is true or not. The test statistic is the criterion function at the parameter estimates

$$J_n = J_n\left(\widehat{\boldsymbol{\beta}}\right) = n\overline{g}_n(\widehat{\boldsymbol{\beta}})'\mathbf{W}_n\overline{g}_n(\widehat{\boldsymbol{\beta}}) = n^2\overline{g}_n(\widehat{\boldsymbol{\beta}})'\left(\widehat{g}'\widehat{g} - n\overline{g}_n\overline{g}_n'\right)^{-1}\overline{g}_n(\widehat{\boldsymbol{\beta}}),$$

where $\mathbf{W}_n$ is defined in (5). Under the hypothesis of correct specification,

$$J_n \xrightarrow{d} \chi^2_{l-k}.$$

The degrees of freedom of the asymptotic distribution are the number of over-identifying restrictions. If the statistic J_n exceeds the chi-square critical value, we can reject the model.

Exercise 9 *Take the linear model*

$$\begin{aligned} y_i &= \mathbf{x}_i'\boldsymbol{\beta} + u_i, \\ E[\mathbf{z}_i u_i] &= \mathbf{0}, \end{aligned}$$

and consider the GMM estimator $\widehat{\boldsymbol{\beta}}$ *of* $\boldsymbol{\beta}$*. Let*

$$J_n = n\overline{g}_n\left(\widehat{\boldsymbol{\beta}}\right)' \widehat{\boldsymbol{\Omega}}^{-1}\overline{g}_n\left(\widehat{\boldsymbol{\beta}}\right)$$

denote the test of overidentifying restrictions. Show that $J_n \xrightarrow{d} \chi^2_{l-k}$ *by demonstrating each of the following:*

(i) *Since* $\boldsymbol{\Omega} > 0$*, we can write* $\boldsymbol{\Omega}^{-1} = \mathbf{C}\mathbf{C}'$ *and* $\boldsymbol{\Omega} = \mathbf{C}'^{-1}\mathbf{C}^{-1}$*.*

(ii) $J_n = n\left(\mathbf{C}'\overline{g}_n\left(\widehat{\boldsymbol{\beta}}\right)\right)'\left(\mathbf{C}'\widehat{\boldsymbol{\Omega}}\mathbf{C}\right)^{-1}\mathbf{C}'\overline{g}_n\left(\widehat{\boldsymbol{\beta}}\right)$*.*

(iii) $\mathbf{C}'\overline{g}_n\left(\widehat{\boldsymbol{\beta}}\right) = \mathbf{D}_n\mathbf{C}'\overline{g}_n\left(\boldsymbol{\beta}_0\right)$ *where*

$$\begin{aligned} \mathbf{D}_n &= \mathbf{I}_l - \mathbf{C}'\left(\frac{1}{n}\mathbf{Z}'\mathbf{X}'\right)\left[\left(\frac{1}{n}\mathbf{X}'\mathbf{Z}\right)\widehat{\boldsymbol{\Omega}}^{-1}\left(\frac{1}{n}\mathbf{Z}'\mathbf{X}\right)\right]^{-1}\left(\frac{1}{n}\mathbf{X}'\mathbf{Z}\right)\widehat{\boldsymbol{\Omega}}^{-1}\mathbf{C}'^{-1}, \\ \overline{g}_n\left(\boldsymbol{\beta}_0\right) &= \frac{1}{n}\mathbf{Z}'\mathbf{u}. \end{aligned}$$

(iv) $\mathbf{D}_n \xrightarrow{p} I_l - \mathbf{R}(\mathbf{R}'\mathbf{R})^{-1}\mathbf{R}'$ *where* $\mathbf{R} = \mathbf{C}'E[\mathbf{z}_i\mathbf{x}_i'] = \mathbf{C}'\mathbf{G}$*.*

(v) $n^{1/2}\mathbf{C}'\overline{g}_n\left(\boldsymbol{\beta}_0\right) \xrightarrow{d} Z \sim N(\mathbf{0}, \mathbf{I}_l)$*.*

(vi) $J_n \xrightarrow{d} Z'\left(\mathbf{I}_l - \mathbf{R}(\mathbf{R}'\mathbf{R})^{-1}\mathbf{R}'\right)Z$*.*

(vii) $Z'\left(\mathbf{I}_l - \mathbf{R}(\mathbf{R}'\mathbf{R})^{-1}\mathbf{R}'\right)Z \sim \chi^2_{l-k}$*.*

(Hint: $\mathbf{I}_l - \mathbf{R}(\mathbf{R}'\mathbf{R})^{-1}\mathbf{R}'$ *is a projection matrix. Note also the difference in the notation* Z *and* $\mathbf{Z}$*.)*

An alternative way to understand the J test by Sargan (1958) is to show that it is actually an F test in the homoskedastic linear model

$$\begin{aligned} y_i &= \mathbf{x}_{1i}'\boldsymbol{\beta}_1 + \mathbf{x}_{2i}'\boldsymbol{\beta}_2 + u_i, \\ E\left[\mathbf{z}_i u_i\right] &= 0,\ E[u_i^2|\mathbf{z}_i] = \sigma^2, \end{aligned} \tag{9}$$

where $\mathbf{z}_i = (\mathbf{x}_{1i}', \mathbf{z}_{2i}')'$. Exogeneity of the instruments means that they are uncorrelated with u_i, which suggests that the instruments should be approximately uncorrelated with $\widehat{u}_i$, where $\widehat{u}_i = y_i - \mathbf{x}_{1i}'\widehat{\boldsymbol{\beta}}_1 - \mathbf{x}_{2i}'\widehat{\boldsymbol{\beta}}_2$ with $\widehat{\boldsymbol{\beta}} = \left(\widehat{\boldsymbol{\beta}}_1', \widehat{\boldsymbol{\beta}}_2'\right)'$ being the 2SLS estimator. So we expect in the regression

$$\widehat{u}_i = \mathbf{x}_{1i}'\boldsymbol{\delta}_1 + \mathbf{z}_{2i}'\boldsymbol{\delta}_2 + v_i, \tag{10}$$

the estimate of $\boldsymbol{\delta} \equiv \left(\boldsymbol{\delta}_1', \boldsymbol{\delta}_2'\right)'$ is close to zero. Let F denote the homoskedasticity-only F statistic testing $\boldsymbol{\delta}_2 = \mathbf{0}$; then $l_2 F$ converges to $\chi^2_{l_2-k_2} = \chi^2_{l-k}$.

Exercise 10 *In the homoskedastic linear model (9), show that (i)* $J_n = nR_u^2 = \frac{\widehat{\mathbf{u}}'\mathbf{P}_{\widetilde{\mathbf{Z}}_2}\widehat{\mathbf{u}}}{\widehat{\mathbf{u}}'\widehat{\mathbf{u}}/n}$, *where* R_u^2 *is the uncentered* R^2 *in the regression (10), and* $\widetilde{\mathbf{Z}}_2 = \mathbf{M}_{\mathbf{X}_1} Z_2$*; (ii)* $l_2 F$ *has the same asymptotic distribution as* J_n*.*

To further appreciate the idea of the J test, consider the linear model (9) again. Suppose we have one endogenous variable x_{2i} and two instruments $\mathbf{z}_{2i}$, and then we can use either instrument to estimate $\boldsymbol{\beta} \equiv \left(\boldsymbol{\beta}_1', \boldsymbol{\beta}_2'\right)'$. If H_0 holds, we expect that these two instruments will generate similar estimates. If the two estimates are very different, then we suspect H_0 fails. The J test implicitly makes this comparison; see Newey (1985b) for such a Hausman test interpretation of the J test.[2] In other words, the J test is testing whether the estimates from different sets of instruments are consistent with each other; it is not testing whether $\mathbf{z}$ is exogenous. Acceptance is consistent with all of the instruments being endogenous, while failure is consistent with a subset being exogenous. Passing an overidentification test does not validate instrumentation. In other words, the null hypothesis of the GMM over-identification test is refutable but nonverifiable as a test of exogeneity of $\mathbf{z}$ (see Breusch (1986)) - if the null is rejected, then we are sure that some (although need not all) instruments are not exogenous; while even if the null is not rejected, all or some of the instruments can be endogenous.

The following exercise rigorously shows that the J test does not have power against some direction of violation of the null.

Exercise 11 *We use the same setup and notations as in Exercise 9 except that* $E[g(\mathbf{w}, \boldsymbol{\beta}_n) = E[\mathbf{z}(y - \mathbf{x}'\boldsymbol{\beta}_n)] = E[\mathbf{z}u] = \boldsymbol{\delta}/\sqrt{n}$*. Show the following results.*

(i) $\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_n\right) \xrightarrow{d} N\left(\left(\mathbf{G}'\boldsymbol{\Omega}^{-1}\mathbf{G}\right)^{-1}\mathbf{G}'\boldsymbol{\Omega}^{-1}\boldsymbol{\delta}, \mathbf{V}\right)$*, where* $\mathbf{V} = \left(\mathbf{G}'\boldsymbol{\Omega}^{-1}\mathbf{G}\right)^{-1}$*.*

(ii) $J_n \xrightarrow{d} \chi^2_{l-k}(\lambda)$*, where* $\lambda = \boldsymbol{\delta}'\boldsymbol{\Omega}^{-1}\left(\boldsymbol{\Omega} - \mathbf{G}\mathbf{V}\mathbf{G}'\right)\boldsymbol{\Omega}^{-1}\boldsymbol{\delta}$*.*

(iii) $\lambda = 0$ *if and only if* $\boldsymbol{\delta} \in span(\mathbf{G})$*.*

From this exercise, when $\boldsymbol{\delta}$ falls in $span(\mathbf{G})$, a k-dimensional space, the J test does not have power. Intuitively, when the null (8) holds, $\boldsymbol{\delta} = \sqrt{n}E[g(\mathbf{w}, \boldsymbol{\beta}_n)] = \sqrt{n}\left(E[g(\mathbf{w}, \boldsymbol{\beta}_n)] - E[g(\mathbf{w}, \boldsymbol{\beta}_0)]\right) = \mathbf{G}\sqrt{n}\left(\boldsymbol{\beta}_n - \boldsymbol{\beta}_0\right)$ stays in $span\left(\mathbf{G}\right)$, a k-dimensional space, where we denote the $\boldsymbol{\beta}$ under the local alternative as $\boldsymbol{\beta}_n$ and that under the null as $\boldsymbol{\beta}_0$. Since for whatever value $E\left[g(\mathbf{w}, \boldsymbol{\beta})\right]$ takes, we can always find some $\boldsymbol{\beta}$ such that $\mathbf{G}'\boldsymbol{\Omega}^{-1}E\left[g(\mathbf{w}, \boldsymbol{\beta})\right] = \mathbf{0}$ and we actually estimate $\boldsymbol{\beta}$ using these moment conditions, the J test obtains power only from the remaining $l - k$ moment conditions not covered by $\mathbf{G}'\boldsymbol{\Omega}^{-1}E\left[g(\mathbf{w}, \boldsymbol{\beta})\right] = \mathbf{0}$. This is why the degree of freedom of J_n is $l - k$ and also why only when $\boldsymbol{\delta}$ falls in a $(l-k)$-dimensional space, the J test has power. On the other hand, if we use the inner product $\langle \mathbf{x}, \mathbf{z} \rangle = \mathbf{x}'\boldsymbol{\Omega}^{-1}\mathbf{z}$, then when $\boldsymbol{\delta}$ falls in $span(\mathbf{G})^{\perp}$, the asymptotic bias of $\widehat{\boldsymbol{\beta}}$,

[2] See also Angrist (1991) for the dummy instrument case.

$\left(\mathbf{G}'\mathbf{\Omega}^{-1}\mathbf{G}\right)^{-1}\mathbf{G}'\mathbf{\Omega}^{-1}\boldsymbol{\delta}$, is zero. Intuitively, the estimation of $\boldsymbol{\beta}$ is based on $\mathbf{G}'\mathbf{\Omega}^{-1}E\left[g(\mathbf{w},\boldsymbol{\beta})\right]=\mathbf{0}$ as mentioned in Section 1, so only the deviation $\boldsymbol{\delta}$ such that $\mathbf{G}'\mathbf{\Omega}^{-1}\boldsymbol{\delta}\neq\mathbf{0}$ would introduce bias to $\widehat{\boldsymbol{\beta}}$. In this sense, the bias of $\boldsymbol{\beta}$ and the power of the J test satisfy some orthogonality property.

The J test is a very useful by-product of the GMM methodology, and it is advisable to report the statistic J_n whenever GMM is used. When over-identified models are estimated by GMM, it is customary to report the J_n statistic as a general test of model adequacy.

(**) As reported in the July 1996 issue of the *Journal of Business and Economic Statistics*, the J test tends to over-reject in finite samples. On the other hand, Tauchen (1986) reported cases in which $J_n(\boldsymbol{\beta})$ evaluated at the iterative estimator led to underrejection of the overidentifying restrictions.

When the J test rejects the null, there are two responses, either screening correct moments based on some moment selection procedure as in Section 6.1 or estimating the misspecified model directly. We briefly discuss the second response here based on Hall and Inoue (2003). It can be shown that (i) the probability limit of the GMM estimator depends on the limit of the weighting matrix; (ii) the limiting distribution of the GMM estimator depends on the limiting distribution of the elements of the weighting matrix (especially the rate of converence to its limit); (iii) the iterated estimators are not asymptotically equivalent; (iv) the three asymptotically equivalent tests discussed in the next subsection are not asymptotically equivalent or asymptotically χ^2-distributed under the null. They propose statistics for testing hypotheses about the pseudo-parameters which have limiting χ^2 distributions under the null. (**)

5.3 Three Asymptotically Equivalent Tests: The Wald, LM and Distance Test

Suppose we want to test

$$H_0:\underset{(q\times 1)}{\mathbf{r}(\boldsymbol{\beta})=\mathbf{0}}\quad\text{vs}\quad H_1:\underset{(q\times 1)}{\mathbf{r}(\boldsymbol{\beta})\neq\mathbf{0}}.$$

We impose the same regularity conditions, Assumption RLS.1′, as in Section 4 of Chapter 5 on $r(\cdot)$. Specifically, we assume that $\mathbf{r}(\cdot)$ is continuously differentiable at the true value β and $\mathbf{R}=\frac{\partial}{\partial\boldsymbol{\beta}}\mathbf{r}(\beta)'$ has rank q.

We described before how to construct estimates of the asymptotic covariance matrix of the GMM estimates. These may be used to construct Wald tests of statistical hypotheses. Specifically,

$$W_n=n\cdot\mathbf{r}\left(\widehat{\boldsymbol{\beta}}\right)'\left[\widehat{\mathbf{R}}'\widehat{\mathbf{V}}\widehat{\mathbf{R}}\right]^{-1}\mathbf{r}\left(\widehat{\boldsymbol{\beta}}\right),$$

where $\widehat{\boldsymbol{\beta}}$ is the unrestricted estimator

$$\widehat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}}J_n(\boldsymbol{\beta}),$$

and for a given weight matrix $\mathbf{W}_n$ in Section 3, the GMM criterion function

$$J_n(\boldsymbol{\beta}) = n \cdot \overline{g}_n(\boldsymbol{\beta})' \mathbf{W}_n \overline{g}_n(\boldsymbol{\beta}),$$

and $\widehat{\mathbf{R}} = \partial \mathbf{r}\left(\widehat{\boldsymbol{\beta}}\right)' / \partial \beta$.

The principal advantage of the Wald test is that it only requires the unconstrained estimator to compute it. Its principal disadvantage is that it is not invariant to reparametrization as discussed in Section 10 of Chapter 5. When the hypothesis is non-linear, a better approach is to directly use the GMM criterion function. This is sometimes called the GMM Distance statistic, and sometimes called a LR-like statistic. The idea was first put forward by Newey and West (1987a). Define the restricted estimator $\widetilde{\boldsymbol{\beta}}$ as

$$\widetilde{\boldsymbol{\beta}} = \arg \min_{\mathbf{r}(\boldsymbol{\beta})=\mathbf{0}} J_n(\boldsymbol{\beta}).$$

The two minimizing criterion functions for $\widehat{\boldsymbol{\beta}}$ and $\widetilde{\boldsymbol{\beta}}$ are $J_n(\widehat{\boldsymbol{\beta}})$ and $J_n(\widetilde{\boldsymbol{\beta}})$. The GMM distance statistic is the difference

$$D_n = J_n(\widetilde{\boldsymbol{\beta}}) - J_n(\widehat{\boldsymbol{\beta}}).$$

As discussed before, if $\mathbf{r}$ is non-linear, the Wald statistic can work quite poorly. In contrast, current evidence suggests that the D statistic appears to have quite good sampling properties, and is the preferred test statistic; see Hansen (2006) for a comparison of its higher-order properties with the Wald statistic. Newey and West (1987a) suggested to use the same weight matrix $\mathbf{W}_n$ for both null and alternative, as this ensures that $D_n \geq 0$. This reasoning is not compelling, however, and some current research suggests that this restriction is not necessary for good performance of the test. This test shares the useful feature of LR tests in that it is a natural by-product of the computation of alternative models.

Another test is the Lagrange multiplier (LM) or the score test. Its test statistic is constructed as

$$LM_n = n\left[\overline{g}_n\left(\widetilde{\boldsymbol{\beta}}\right)' \mathbf{W}_n \mathbf{G}_n\left(\widetilde{\boldsymbol{\beta}}\right)\right] \widetilde{\mathbf{V}} \left[\mathbf{G}_n\left(\widetilde{\boldsymbol{\beta}}\right)' \mathbf{W}_n \overline{g}_n\left(\widetilde{\boldsymbol{\beta}}\right)\right],$$

where

$$\widetilde{\mathbf{V}} = \left[\mathbf{G}_n\left(\widetilde{\boldsymbol{\beta}}\right)' \mathbf{W}_n \mathbf{G}_n\left(\widetilde{\boldsymbol{\beta}}\right)\right]^{-1},$$

and $\mathbf{G}_n\left(\widetilde{\boldsymbol{\beta}}\right)' \mathbf{W}_n \overline{g}_n\left(\widetilde{\boldsymbol{\beta}}\right)$ is the first-order derivative of $J_n(\cdot)$ at $\widetilde{\boldsymbol{\beta}}$ and plays the role of the score function in the likelihood framework. As the LM test statistic in the likelihood framework, we need only calculate the restricted estimator $\widetilde{\boldsymbol{\beta}}$, while we need to calculate both $\widehat{\boldsymbol{\beta}}$ and $\widetilde{\boldsymbol{\beta}}$ in the distance statistic.

As in Section 11 of Chapter 5, we can also consider the minimum distance (or minimum chi-square) test. Define

$$\widehat{\boldsymbol{\beta}}_{EMD} = \arg \min_{\mathbf{r}(\boldsymbol{\beta})=\mathbf{0}} \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \widehat{\mathbf{V}}^{-1} \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right),$$

where $\widehat{\boldsymbol{\beta}}$ is the optimal GMM estimator, and $\widehat{\mathbf{V}}$ is an estimator of its asymptotic covariance matrix.

It can be shown that $\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_{EMD}-\widetilde{\boldsymbol{\beta}}\right)=o_p(1)$; see, e.g., Proposition 2 of Newey and West (1987a). The minimum chi-square statistic is

$$MC_n = n\left(\widehat{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}_{EMD}\right)\widehat{\mathbf{V}}^{-1}\left(\widehat{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}_{EMD}\right).$$

Like the distance statistic, it requires two minimizations to compute.

5.3.1 The Trinity in GMM

As in the likelihood case, we can show that the Wald, LM and distance tests are asymptotically equivalent. Actually, we can show they are asymptotically equivalent even under the local alternatives and when the moment conditions are nonlinear in $\boldsymbol{\beta}$. Moreover, they are even asymptotically equivalent to the minimum chi-square test. This result is rigorously stated in Theorem 2 of Newey and West (1987a).

Proposition 1 *Under some regularity conditions, and the local alternatives* $\boldsymbol{\beta}_n=\boldsymbol{\beta}+n^{-1/2}\mathbf{b}$,

$$W_n \xrightarrow{d} \chi^2_q(\lambda),$$

where $\lambda=\mathbf{b}'\mathbf{R}\left(\mathbf{R}'\mathbf{V}\mathbf{R}\right)^{-1}\mathbf{R}\mathbf{b}$. *In addition,* $W_n-D_n=o_p(1)$, $W_n-LM_n=o_p(1)$, *and* $W_n-MC_n=o_p(1)$.

It should be emphasized that the optimal weight matrix is used in the construction of D_n; otherwise, D_n is not asymptotically chi-squared and is not asymptotically equivalent to W_n. This is parallel to the result in the misspecified likelihood case. Also, the form of the LM statistic would be more complicated, and would in general involve the Jacobian matrix $\mathbf{R}$ of the constraints. So it is strongly suggested to use the optimal weight matrix in the hypothesis testing of GMM.

We now consider some special cases. The following proposition follows from Proposition 1, 3 and 4 of Newey and West (1987a).

Proposition 2 *(i) When the model is just-identified,* $LM_n=D_n$. *(ii) When* $g(\mathbf{w},\boldsymbol{\beta})=g_1(\mathbf{w})-g_2(\mathbf{w})\boldsymbol{\beta}$, $D_n=LM_n=MC_n$. *(iii) When* $g(\mathbf{w},\boldsymbol{\beta})=g_1(\mathbf{w})-g_2(\mathbf{w})\boldsymbol{\beta}$ *and* $\mathbf{r}(\boldsymbol{\beta})=\mathbf{R}'\boldsymbol{\beta}-\mathbf{c}$, $W_n=D_n=LM_n=MC_n$.

Result (i) can be easily proved. In the just-identified case, $\overline{g}_n\left(\widehat{\boldsymbol{\beta}}\right)=\mathbf{0}$, so $D_n=J_n(\widetilde{\boldsymbol{\beta}})=n\cdot\overline{g}_n\left(\widetilde{\boldsymbol{\beta}}\right)'\mathbf{W}_n\overline{g}_n\left(\widetilde{\boldsymbol{\beta}}\right)$. On the other hand, given $\mathbf{G}_n\left(\widetilde{\boldsymbol{\beta}}\right)$ is invertible,

$$\begin{aligned}
LM_n &= n\left[\overline{g}_n\left(\widetilde{\boldsymbol{\beta}}\right)'\mathbf{W}_n\mathbf{G}_n\left(\widetilde{\boldsymbol{\beta}}\right)\right]\widetilde{\mathbf{V}}\left[\mathbf{G}_n\left(\widetilde{\boldsymbol{\beta}}\right)'\mathbf{W}_n\overline{g}_n\left(\widetilde{\boldsymbol{\beta}}\right)\right] \\
&= n\cdot\overline{g}_n\left(\widetilde{\boldsymbol{\beta}}\right)'\mathbf{W}_n\mathbf{G}_n\left(\widetilde{\boldsymbol{\beta}}\right)\mathbf{G}_n\left(\widetilde{\boldsymbol{\beta}}\right)^{-1}\mathbf{W}_n^{-1}\mathbf{G}_n\left(\widetilde{\boldsymbol{\beta}}\right)'^{-1}\mathbf{G}_n\left(\widetilde{\boldsymbol{\beta}}\right)'\mathbf{W}_n\overline{g}_n\left(\widetilde{\boldsymbol{\beta}}\right) \\
&= n\cdot\overline{g}_n\left(\widetilde{\boldsymbol{\beta}}\right)'\mathbf{W}_n\overline{g}_n\left(\widetilde{\boldsymbol{\beta}}\right).
\end{aligned}$$

The equivalence in (ii) does not include W_n because it involves the Jacobian of the constraints when $\mathbf{r}(\cdot)$ is nonlinear. We will provide some intuition on $D_n = LM_n$ at the end of the next subsection. The following exercise shows Result (iii) in the linear instrumental variables case. The minimum distance statistic J_n^* in Section 11 of Chapter 5 is numerically equivalent to D_n and LM_n even if the restrictions are nonlinear since it can be put in case (ii) of Proposition 2.

Exercise 12 *Take the linear model*

$$\begin{aligned} y_i &= \mathbf{x}_i'\boldsymbol{\beta} + u_i, \\ E[\mathbf{z}_i u_i] &= \mathbf{0}, \end{aligned}$$

and consider the unrestricted GMM estimator $\widehat{\boldsymbol{\beta}}$ *and restricted GMM estimator* $\widetilde{\boldsymbol{\beta}}$ *of* $\boldsymbol{\beta}$ *under the linear constraints* $\mathbf{R}'\boldsymbol{\beta} = \mathbf{c}$*. Define*

$$J_n(\boldsymbol{\beta}) = n \cdot \overline{g}_n(\boldsymbol{\beta})\, \widehat{\boldsymbol{\Omega}}^{-1} \overline{g}_n(\boldsymbol{\beta}),$$

and then $\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} J_n(\boldsymbol{\beta})$ *and* $\widetilde{\boldsymbol{\beta}} = \arg\min_{\mathbf{R}'\boldsymbol{\beta}=\mathbf{c}} J_n(\boldsymbol{\beta})$*. Define the Lagrangian*

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \frac{1}{2} J_n(\boldsymbol{\beta}) + \boldsymbol{\lambda}' \left(\mathbf{R}'\boldsymbol{\beta} - \mathbf{c}\right).$$

(i) *Show that*

$$\begin{aligned} \widetilde{\boldsymbol{\beta}} &= \widehat{\boldsymbol{\beta}} - \left(\mathbf{X}'\mathbf{Z}\widehat{\boldsymbol{\Omega}}^{-1}\mathbf{Z}'\mathbf{X}\right)^{-1} \mathbf{R} \left[\mathbf{R}' \left(\mathbf{X}'\mathbf{Z}\widehat{\boldsymbol{\Omega}}^{-1}\mathbf{Z}'\mathbf{X}\right)^{-1} \mathbf{R}\right]^{-1} \left(\mathbf{R}'\widehat{\boldsymbol{\beta}} - \mathbf{c}\right), \\ \widehat{\boldsymbol{\lambda}} &= \frac{1}{n} \left[\mathbf{R}' \left(\mathbf{X}'\mathbf{Z}\widehat{\boldsymbol{\Omega}}^{-1}\mathbf{Z}'\mathbf{X}\right)^{-1} \mathbf{R}\right]^{-1} \left(\mathbf{R}'\widehat{\boldsymbol{\beta}} - \mathbf{c}\right). \end{aligned}$$

(ii) *Derive the asymptotic distribution of* $\widetilde{\boldsymbol{\beta}}$ *under the null.*

(iii) *Show that*

$$J_n\left(\widetilde{\boldsymbol{\beta}}\right) = J_n\left(\widehat{\boldsymbol{\beta}}\right) + \frac{1}{n}\left(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}\right)' \mathbf{X}'\mathbf{Z}\widehat{\boldsymbol{\Omega}}^{-1}\mathbf{Z}'\mathbf{X}\left(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}\right).$$

(iv) *Show that the distance statistic is equal to the Wald statistic.*

(**)The numerical equivalence result suggests a convenient way to calculate D_n in the linear model

$$\begin{aligned} y_i &= \mathbf{x}_i'\boldsymbol{\beta} + u_i, \\ E[\mathbf{z}_i u_i] &= \mathbf{0}, \end{aligned}$$

Define $\mathbf{z}_i^* = \mathbf{z}_i \widehat{u}_i$, and $\mathbf{Z}^* = (\mathbf{z}_1^*, \cdots, \mathbf{z}_n^*)'$, $y_i^* = y_i/\widehat{u}_i$, $\mathbf{y}^* = (y_1^*, \cdots, y_n^*)'$, $\mathbf{x}_i^* = \mathbf{x}_i/\widehat{u}_i$ and $\mathbf{X}^* = (\mathbf{x}_1^*, \cdots, \mathbf{x}_n^*)'$. The 2SLS estimator $\widehat{\boldsymbol{\beta}}$ for a regression of $\mathbf{y}^*$ on $\mathbf{X}^*$ with instruments $\mathbf{Z}^*$ is an efficient

estimator of $\boldsymbol{\beta}$, where $\mathbf{W}_n$ in (6) is used as the efficient weight matrix (why?). Define $\widehat{u}_i^* = y_i^* - \mathbf{x}_i^{*\prime}\widehat{\boldsymbol{\beta}}$, and $\widehat{\mathbf{u}}^* = (\widehat{u}_1^*, \cdots, \widehat{u}_n^*)'$. The sum of squares of the predicted values of a regression of $\widehat{u}_i^*$ on $\mathbf{z}_i^*$ is

$$\widehat{S}_n = \widehat{\mathbf{u}}^{*\prime}\mathbf{Z}^*(\mathbf{Z}^{*\prime}\mathbf{Z}^*)^{-1}\mathbf{Z}^{*\prime}\widehat{\mathbf{u}}^* = (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})'\mathbf{Z}\mathbf{W}_n\mathbf{Z}'(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})/n = J_n(\widehat{\boldsymbol{\beta}}).$$

Substitute out the constraints $\mathbf{R}'\boldsymbol{\beta} - \mathbf{c} = \mathbf{0}$, and repeat the procedure above for the restricted estimator.[3] Let $\widetilde{S}_n$ be the counterpart of $\widehat{S}_n$; then $D_n = \widehat{S}_n - \widetilde{S}_n$.

The numerical equivalence result seems puzzling given the ranking $W \geq LR \geq LM$ in the normal regression model (see Section 5 of Chapter 4). Part of the explanation is that we assume that all statistics use the same estimate $\widehat{\mathbf{V}}$ of $\mathbf{V}$. Also, our objective function uses the FOCs rather than the log-likelihood itself. (**)

5.3.2 Summary of the Trinity in GMM and the M-estimation (*)

We generally consider the *extremum estimator* which is defined as

$$\begin{aligned} \widehat{\boldsymbol{\theta}} &= \arg\max\ Q_n(\boldsymbol{\theta}) \\ \text{s.t. } \boldsymbol{\theta} &\in \Theta \subset \mathbb{R}^k \end{aligned}$$

where $Q_n(\cdot)$ is a general criterion function. To test $H_0 : \mathbf{r}(\boldsymbol{\theta}) = \mathbf{0}$, we sometimes need the *constrained estimator*, denoted as $\widetilde{\boldsymbol{\theta}}$, which solves

$$\begin{aligned} &\max\ Q_n(\boldsymbol{\theta}) \\ &\text{s.t. } \mathbf{r}(\boldsymbol{\theta}) = \mathbf{0} \end{aligned}$$

Among extremum estimators, the most popular cases are M-estimators (e.g., the MLE) and GMM estimators. Their criterion functions are as follows:

1. M-estimators: $Q_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} m(\mathbf{w}_i; \boldsymbol{\theta})$, that is, the objective function is a sample average.
2. GMM: $Q_n(\boldsymbol{\theta}) = -\frac{1}{2}\underset{(1\times l)}{\overline{g}_n(\boldsymbol{\theta})'}\underset{(l\times l)}{\mathbf{W}_n}\underset{(l\times 1)}{\overline{g}_n(\boldsymbol{\theta})}$

Here, we define $Q_n(\boldsymbol{\theta}) = -\frac{1}{2n}J_n(\boldsymbol{\theta})$ to give an analog of the average log-likelihood in the ML case.

[3] When the constraints are nonlinear, we cannot substitute out for the constraints and use the 2SLS calculation to obtain the residuals for the restricted estimates.

The following notation will be used throughout this section:

$$\begin{aligned}
\underset{(k\times 1)}{s(\mathbf{w}_i;\boldsymbol{\theta})} &= \frac{\partial m(\mathbf{w}_i;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}, \quad \underset{(k\times k)}{\mathbf{H}(\mathbf{w}_i;\boldsymbol{\theta})} = \frac{\partial s(\mathbf{w}_i;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} = \frac{\partial^2 m(\mathbf{w}_i;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}, \\
\underset{(l\times 1)}{\overline{g}_n(\boldsymbol{\theta})} &= \frac{1}{n}\sum_{i=1}^{n} g(\mathbf{w}_i;\boldsymbol{\theta}), \underset{(k\times q)}{\mathbf{R}(\boldsymbol{\theta})} = \partial \mathbf{r}\left(\boldsymbol{\theta}\right)'/\partial \boldsymbol{\theta} \\
\underset{(l\times k)}{\mathbf{G}_n(\boldsymbol{\theta})} &= \frac{\partial \overline{g}_n(\mathbf{w}_i;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}, \quad \underset{(l\times k)}{\mathbf{G}} = E\left[\frac{\partial g(\mathbf{w}_i;\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}'}\right], \\
\underset{(l\times l)}{\boldsymbol{\Omega}_n\left(\boldsymbol{\theta}\right)} &= \frac{1}{n}\sum_{i=1}^{n} g(\mathbf{w}_i;\boldsymbol{\theta})g(\mathbf{w}_i;\boldsymbol{\theta})', \underset{(l\times l)}{\boldsymbol{\Omega}} = E[g(\mathbf{w}_i;\boldsymbol{\theta}_0)g(\mathbf{w}_i;\boldsymbol{\theta}_0)'].
\end{aligned}$$

We summarize the asymptotic approximation of these statistics and tests in the following two tables, which are Tables 7.1 and 7.2 of Hayashi (2000). Table 1 provides the correspondence of Taylor expansion for the sampling error between M-estimators and GMM estimators. Table 2 gives the components of the three asymptotically equivalent test statistics in the ML and GMM estimation. Also, Figure 2 provides an intuitive explanation for these three tests. W_n can be interpreted as twice the difference in the criterion function at the two estimates, using a quadratic approximation to the criterion function at $\widehat{\boldsymbol{\theta}}$; LM_n can be interpreted as twice the difference in the criterion function at the two estimates, using a quadratic approximation to the criterion function at $\widetilde{\boldsymbol{\theta}}$; and D_n is precisely twice the difference in the criterion function at the unconstrained and constrained estimates. Note that the same covariance matrix $\widehat{\boldsymbol{\Omega}}$ is used in $Q_n(\widehat{\boldsymbol{\theta}})$ and $Q_n(\widetilde{\boldsymbol{\theta}})$.

Exercise 13 *Suppose $\theta \in \mathbb{R}$, H_0 is $\theta - c = 0$, and the curvature of $Q_n(\theta)$ at $\widetilde{\theta}$ is $-\widetilde{\Sigma}$ and at $\widehat{\theta}$ is $-\widehat{\Sigma}$. Show that (i) W_n is twice the difference in $Q_n(\theta)$ at $\widetilde{\theta}$ and $\widehat{\theta}$, using a quadratic approximation to $Q_n(\theta)$ at $\widehat{\theta}$; (ii) LM_n is twice the difference in $Q_n(\theta)$ at $\widetilde{\theta}$ and $\widehat{\theta}$, using a quadratic approximation to $Q_n(\theta)$ at $\widetilde{\theta}$.*

$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right) = -\boldsymbol{\Psi}^{-1}\sqrt{n}\frac{\partial Q_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} + o_p(1)$, $\sqrt{n}\frac{\partial Q_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \xrightarrow{d} N(\mathbf{0},\boldsymbol{\Sigma})$, $\text{Avar}\left(\widehat{\boldsymbol{\theta}}\right) = \boldsymbol{\Psi}^{-1}\boldsymbol{\Sigma}\boldsymbol{\Psi}^{-1}$		
Terms for substitution	M-estimators	GMM
$Q_n\left(\boldsymbol{\theta}\right)$	$\frac{1}{n}\sum_{i=1}^{n} m\left(\mathbf{w}_i;\boldsymbol{\theta}\right)$	$-\frac{1}{2}\overline{g}_n(\boldsymbol{\theta})'\mathbf{W}_n\overline{g}_n(\boldsymbol{\theta})$
$\sqrt{n}\frac{\partial Q_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}$	$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} s\left(\mathbf{w}_i;\boldsymbol{\theta}_0\right)$	$-\left[\mathbf{G}_n(\boldsymbol{\theta}_0)\right]'\mathbf{W}_n\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g\left(\mathbf{w}_i;\boldsymbol{\theta}_0\right)$
$\boldsymbol{\Psi}$	$E\left[\mathbf{H}(\mathbf{w}_i;\boldsymbol{\theta}_0)\right]$	$-\mathbf{G}'\mathbf{W}\mathbf{G}$
$\boldsymbol{\Sigma}$	$E\left[s(\mathbf{w}_i;\boldsymbol{\theta}_0)s(\mathbf{w}_i;\boldsymbol{\theta}_0)'\right]$	$\mathbf{G}'\mathbf{W}\boldsymbol{\Omega}\mathbf{W}\mathbf{G}$

Table 1: Taylor Expansion for the Sampling Error

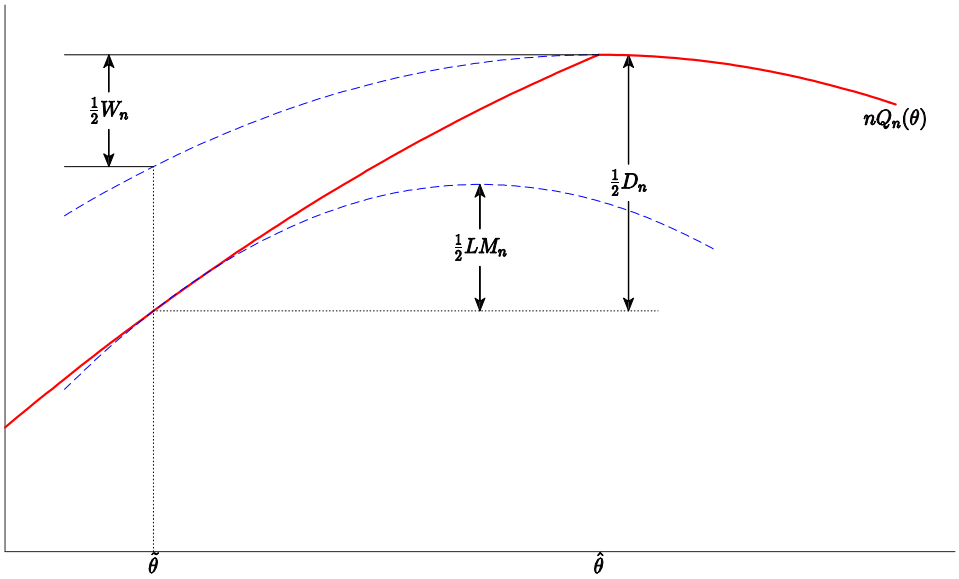

Figure 2: Trinity

Wald: $n \cdot \mathbf{r}\left(\widehat{\boldsymbol{\theta}}\right)' \left[\mathbf{R}(\widehat{\boldsymbol{\theta}})'\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{R}(\widehat{\boldsymbol{\theta}})\right]^{-1} \mathbf{r}\left(\widehat{\boldsymbol{\theta}}\right)$		
LM: $n\left(\frac{\partial Q_n(\widetilde{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}}\right)' \widetilde{\boldsymbol{\Sigma}}^{-1} \left(\frac{\partial Q_n(\widetilde{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}}\right)$		
LR: $2n \cdot \left[Q_n(\widehat{\boldsymbol{\theta}}) - Q_n(\widetilde{\boldsymbol{\theta}})\right]$ [4]		
Terms for substitution	Conditional ML	Efficient GMM
$Q_n(\boldsymbol{\theta})$	$\frac{1}{n}\sum_{i=1}^{n} \log f(y_i \mid \mathbf{x}_i; \boldsymbol{\theta})$	$-\frac{1}{2}\overline{g}_n(\boldsymbol{\theta})'\widehat{\boldsymbol{\Omega}}^{-1}\overline{g}_n(\boldsymbol{\theta})$
$\widehat{\boldsymbol{\Sigma}}$	$-\frac{\partial^2 Q_n(\widehat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}$ or $\frac{1}{n}\sum_{i=1}^{n} s(\mathbf{w}_i; \widehat{\boldsymbol{\theta}})s(\mathbf{w}_i; \widehat{\boldsymbol{\theta}})'$ [5]	$\widehat{\mathbf{G}}'\widehat{\boldsymbol{\Omega}}^{-1}\widehat{\mathbf{G}}$, $\underset{(l\times k)}{\widehat{\mathbf{G}}} \equiv \mathbf{G}_n(\widehat{\boldsymbol{\theta}})$, $\underset{(l\times l)}{\widehat{\boldsymbol{\Omega}}} = \boldsymbol{\Omega}_n\left(\widehat{\boldsymbol{\theta}}\right)$
$\widetilde{\boldsymbol{\Sigma}}$	replace $\widehat{\boldsymbol{\theta}}$ by $\widetilde{\boldsymbol{\theta}}$ in above	$\widetilde{\mathbf{G}}'\widetilde{\boldsymbol{\Omega}}^{-1}\widetilde{\mathbf{G}}$, $\underset{(l\times k)}{\widetilde{\mathbf{G}}} \equiv \mathbf{G}_n(\widetilde{\boldsymbol{\theta}})$, $\underset{(l\times l)}{\widetilde{\boldsymbol{\Omega}}} = \boldsymbol{\Omega}_n\left(\widetilde{\boldsymbol{\theta}}\right)$

Table 2: Trinity

From Figure 2, we can understand why $D_n = LM_n$ when $g(\mathbf{w}, \boldsymbol{\beta})$ is linear in $\boldsymbol{\beta}$ even if $\mathbf{r}(\cdot)$ is nonlinear in Proposition 2(ii). This is because in this case, $Q_n(\boldsymbol{\beta})$ is exactly quadratic if $g(\mathbf{w}, \boldsymbol{\beta})$ is linear in $\boldsymbol{\beta}$, so the quadratic approximation of the LM statistic (the lower blue line) coincides with $Q_n(\boldsymbol{\beta})$ itself (the red line). As a result, the differences at the two estimates in $Q_n(\boldsymbol{\beta})$ and the quadratic approximation are the same.

[4] The same weighting matrix is used in $Q_n(\widehat{\boldsymbol{\theta}})$ and $Q_n(\widetilde{\boldsymbol{\theta}})$ for GMM to guarantee LR is greater than 0 in finite samples. Also, we can see $D_n = J_n(\widetilde{\boldsymbol{\theta}}) - J_n(\widehat{\boldsymbol{\theta}})$.

[5] or let $\mathbf{I}(\boldsymbol{\theta}) = E\left[s(\mathbf{w}_i; \boldsymbol{\theta})s(\mathbf{w}_i; \boldsymbol{\theta})'\right]$, and $\widehat{\boldsymbol{\Sigma}} = \mathbf{I}(\widehat{\boldsymbol{\theta}})$.

5.4 Confidence Regions

By inverting the test statistics, we can construct confidence regions for $\boldsymbol{\theta}$. A straightforward choice of the test statistic is the Wald statistic. However, as mentioned above, the distance statistic may perform better in some cases of hypothesis testing. We expect the confidence region by reverting the distance statistic would inherit its good properties in testing. Suppose we want to construct confidence region for $\boldsymbol{\theta}_2$, where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2')' \in \mathbb{R}^k$ and $\boldsymbol{\theta}_2 \in \mathbb{R}^{k_2}$ is a subvector of $\boldsymbol{\theta}$. We need to find $\boldsymbol{\theta}_2$ such that

$$J_n\left(\widetilde{\boldsymbol{\theta}}_1\left(\boldsymbol{\theta}_2\right), \boldsymbol{\theta}_2\right) - J_n\left(\widehat{\boldsymbol{\theta}}\right) \leq \chi^2_{k_2,\alpha},$$

where $\widetilde{\boldsymbol{\theta}}_1\left(\boldsymbol{\theta}_2\right) = \arg\min_{\boldsymbol{\theta}_1} J_n\left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2\right)$ for a given $\boldsymbol{\theta}_2$, the df of the χ^2 limiting distribution is k_2 because the df of $J_n\left(\widetilde{\boldsymbol{\theta}}_1\left(\boldsymbol{\theta}_2\right), \boldsymbol{\theta}_2\right)$ is $l - k_1$ and the df of $J_n\left(\widehat{\boldsymbol{\theta}}\right)$ is $l - k$ so the difference is $(l - k_1) - (l - k) = k - k_1 = k_2$. Of course, we can construct confidence region for $\boldsymbol{\theta}_2$ by collecting $\boldsymbol{\theta}_2$'s such that $J_n\left(\widetilde{\boldsymbol{\theta}}_1\left(\boldsymbol{\theta}_2\right), \boldsymbol{\theta}_2\right) \leq \chi^2_{l-k_1,\alpha}$ directly. However, by observing that $J_n\left(\widetilde{\boldsymbol{\theta}}_1\left(\boldsymbol{\theta}_2\right), \boldsymbol{\theta}_2\right) = \left[J_n\left(\widetilde{\boldsymbol{\theta}}_1\left(\boldsymbol{\theta}_2\right), \boldsymbol{\theta}_2\right) - J_n\left(\widehat{\boldsymbol{\theta}}\right)\right] + J_n\left(\widehat{\boldsymbol{\theta}}\right)$, we can conclude that this confidence region is based on the joint test of overidentification and $\boldsymbol{\theta}_2 = \boldsymbol{\theta}_{20}$. If the model is misspecified so that the overidentifying conditions are invalid, this confidence region can be null.

6 Moment Selection (*)

If the J test rejects the null, we suspect there are some moment conditions which are invalid. Thus, it may be useful to employ a moment selection procedure to estimate which moments are correct and which are incorrect. On the other hand, as mentioned in Section 9, we may need to select moments among many valid ones to improve finite-sample inference. We use Andrews (1999) and Donald and Newey (2001) to exemplify these two scenarios.

6.1 Andrews (1999)

Andrews (1999) develops parallel information criteria (IC) as in the least squares environment; he labels these criteria by adding the prefix GMM. He shows that the GMM-version IC is indeed analogue of the usual IC. To see why, note that for a moment selection vector $\mathbf{c}$ (which is a vector of 0 and 1 with 1 indicating that the corresponding moment equation is included),

$$J_n^*(\mathbf{c}) = n \inf_{\boldsymbol{\theta}\in\Theta, \boldsymbol{\mu}\in\mathbb{R}^{l-k}} \left(\overline{g}_n(\boldsymbol{\theta}) - \mathbf{D}_{\mathbf{c}}\boldsymbol{\mu}\right)' \mathbf{W}_n \left(\overline{g}_n(\boldsymbol{\theta}) - \mathbf{D}_{\mathbf{c}}\boldsymbol{\mu}\right) \tag{11}$$

is the same as (with a $o_p(1)$ difference)[6]

$$J_n(\mathbf{c}) = n \inf_{\boldsymbol{\theta}\in\Theta} \left(\overline{g}_{n\mathbf{c}}(\boldsymbol{\theta})\right)' \mathbf{W}_n(\mathbf{c}) \left(\overline{g}_{n\mathbf{c}}(\boldsymbol{\theta})\right),$$

[6] Actually, by the results of Back and Brown (1993), the minimizers in these two minimization problems are exactly the same.

where $\mathbf{D_c}\boldsymbol{\mu}$ sets the moment conditions with zeros in $\mathbf{c}$ equal to $\boldsymbol{\mu}$ ($\boldsymbol{\mu} \in \mathbb{R}^{l-k}$ corresponds to the just-identified model, or the smallest model allowed), $\overline{g}_{n\mathbf{c}}$ includes the elements of $\overline{g}_n$ with ones in $\mathbf{c}$, and $\mathbf{W}_n$ and $\mathbf{W}_n(\mathbf{c})$ are constructed in the same fashion as in Section 3. The number of parameters with selection vector $\mathbf{c}$ is $k + l - |\mathbf{c}|$, where $|\mathbf{c}|$ is the number of moment conditions selected by $\mathbf{c}$, and $l - |\mathbf{c}|$ is the number of excluded moment conditions. Rewrite $k + l - |\mathbf{c}|$ as $l - q_{\mathbf{c}}$, where $q_{\mathbf{c}} = |\mathbf{c}| - k$ is the number of "over-identifying restrictions". All parameters in J_n^* are $(\boldsymbol{\theta}', \boldsymbol{\mu}')' \in \mathbb{R}^l$. With the selection parameter $\mathbf{c}$, the last $q_{\mathbf{c}}$ parameters are set as zero. Thus, different selection vectors correspond to the setting of different parameters equal to zero in (11), just as different models correspond to the setting of different parameters equal to zero in the likelihood environment of Section 4 in Chapter 6. Actually, $J_n^*(\mathbf{c})$ ($= J_n(\mathbf{c}) + o_p(1) \xrightarrow{d} \chi^2_{q_{\mathbf{c}}}$ under "correct" selection) plays the role of $2n\left[\ell_n(\mathcal{M}_2) - \ell_n(\mathcal{M}_1)\right]$, and $q_{\mathbf{c}}$ plays the role of k_2 there.

The general moment selection criteria (MSC) is specified as

$$MSC_n(\mathbf{c}) = J_n(\mathbf{c}) - h(|\mathbf{c}|)\kappa_n,$$

where $\mathbf{c} \in \mathcal{C}$ is a moment selection vector, $\mathcal{C}$ is the selection set which may not include all possible selection vectors, $h(\cdot)$ is strictly increasing, $\kappa_n \to \infty$ and $\kappa_n = o(n)$. Take $h(x) = x - k$, where k is the number of parameters; then

GMM-BIC: $\kappa_n = \ln n$ and $MSC_{\mathrm{BIC},n}(\mathbf{c}) = J_n(\mathbf{c}) - (|\mathbf{c}| - k)\ln n$;

GMM-AIC: $\kappa_n = 2$ and $MSC_{\mathrm{AIC},n}(\mathbf{c}) = J_n(\mathbf{c}) - 2(|\mathbf{c}| - k)$;

GMM-HQIC: $\kappa_n = Q \ln\ln n$ for some $Q > 2$ and $MSC_{\mathrm{HQIC},n}(\mathbf{c}) = J_n(\mathbf{c}) - Q(|\mathbf{c}| - k)\ln\ln n$.

The MSC estimator, $\widehat{\mathbf{c}}_{MSC}$, is the minimizer of $MSC_n(\mathbf{c})$. It is shown that $\widehat{\mathbf{c}}_{MSC}$ is consistent in the sense that $\widehat{\mathbf{c}}_{MSC} \xrightarrow{p} \mathbf{c}_0$, where $\mathbf{c}_0$ is the set of moments whose expectation is zero for some $\boldsymbol{\theta} \in \Theta$ and whose number is largest among all such sets of moments. We assume $\mathbf{c}_0 \in \mathcal{C}$ and is unique; otherwise, $\widehat{\mathbf{c}}_{MSC}$ will converge to a set that includes all possible selection vectors which maximize the number of valid moments. $\widehat{\mathbf{c}}_{MSC}$ determines when there are no over-identification restrictions. In GMM-AIC, $\kappa_n = 2 \nrightarrow \infty$, so the GMM-AIC procedure is not consistent. It has positive probability even asymptotically of selecting too few moments.

A simple method can be used to detect whether an MSC is reliable. In cases where an MSC performs poorly, there are typically two or more selection vectors that yield MSC values close to the minimum and that yield parameter estimates differing noticeably from each other. In cases where a moment selection procedure performs well, the latter typically does not occur.

Andrews also considers two testing procedures to select correct moment conditions: downward testing (DT) and upward testing (UT) procedures. These procedures are similar to informal methods based on the J test often employed by empirical researchers to determine which moments to use. In the DT procedure, we start with $\mathbf{c} \in \mathcal{C}$ for which $|\mathbf{c}|$ is the largest, and then test with progressively smaller $|\mathbf{c}|$ until the null cannot be rejected. Let $\widehat{k}_{DT}$ be this value of $|\mathbf{c}|$; $\widehat{\mathbf{c}}_{DT}$ is the selection vector that minimizes $J_n(\mathbf{c})$ over $\mathbf{c} \in \mathcal{C}$ with $|\mathbf{c}| = \widehat{k}_{DT}$. If the critical values used in the

J test diverge to infinity at a slower rate than n, then $\widehat{\mathbf{c}}_{DT}$ is consistent. The UT procedure is a converse procedure of the DT procedure. Under additional restrictions which avoids the procedure stopping early, $\widehat{\mathbf{c}}_{UT}$ is also consistent, where $\widehat{\mathbf{c}}_{UT}$ is similarly defined as $\widehat{\mathbf{c}}_{DT}$ with $\widehat{k}_{UT}$, the largest k such that for all $|\mathbf{c}| \leq k$ there is at least one $\mathbf{c}$ for which the null is not rejected, replacing $\widehat{k}_{DT}$.

Monte Carlo results show that GMM-BIC, DT, and UT procedures perform best and about equally well, the GMM-HQIC procedure is next best and the GMM-AIC procedure is worst overall.

Andrews and Lu (2001) extend the results to cover simultaneous moment and model selection; they apply their procedures to dynamic panel models. Hong et al. (2003) use GEL-statistics to provide an alternative interpretation of Andrews' MSCs. Chen et al. (2007) propose a nonparametric likelihood ratio testing procedure for choosing between a parametric (likelihood) model and a moment condition model when both models could be misspecified; their procedure is based on comparing the Kullback-Leibler Information Criterion (KLIC) between the parametric model and moment condition model. Hall et al. (2007) propose an entropy-based moment selection procedure to select relevant moments among valid moment conditions (with possibly weak identification). DiTraglia (2013) extends the focused information criterion (FIC) of Claeskens and Hjort (2003) to FMSC and shows that the use of an invalid but highly relevant instrument can substantially improve inference in finite samples. Caner (2009) studies the LASSO-type GMM estimator and Liao (2013) extends to the general shrinkage estimator.

6.2 Donald and Newey (2001)

Different from Andrews (1999) who is searching for the largest set of valid instruments, Donald and Newey (2001) propose a simple method to choose among *valid* instruments, by minimizing approximate MSE. Their emphasis is on MSE approximation when the instruments may be weak and the number of instruments may be large.

The model is

$$\begin{aligned} y_i &= \mathbf{z}_{1i}'\boldsymbol{\beta}_1 + \mathbf{x}_{2i}'\boldsymbol{\beta}_2 + u_i = \mathbf{x}_i'\boldsymbol{\beta} + u_i, E[u_i|\mathbf{z}_i] = 0, \\ \mathbf{x}_i &= \begin{pmatrix} \mathbf{z}_{1i} \\ \mathbf{x}_{2i} \end{pmatrix} = f(\mathbf{z}_i) + \mathbf{v}_i = \begin{pmatrix} \mathbf{z}_{1i} \\ E\left[\mathbf{x}_{2i}|\mathbf{z}_i\right] \end{pmatrix} + \begin{pmatrix} \mathbf{0} \\ \mathbf{v}_{2i} \end{pmatrix}, \end{aligned}$$

where $\mathbf{z}_i$ can be a few continuous variables, many dummy variables or even be infinite dimensional, and $(u_i, \mathbf{v}_{2i}')$ is *homoskedastic* (see Donald, Imbens and Newey (2009) for extensions). Let

$$\boldsymbol{\psi}_i^K = \psi^K(\mathbf{z}_i) = (\psi_{1K}(\mathbf{z}_i), \cdots, \psi_{KK}(\mathbf{z}_i))$$

be a vector of instruments, where $\boldsymbol{\psi}_i^K$ includes $\mathbf{z}_{1i}$. Because $E[u_i|\mathbf{z}_i] = 0$, $E\left[\boldsymbol{\psi}_i^K u_i\right] = \mathbf{0}$. Different values of K correspond to different instrument sets. Usually, we specify $\boldsymbol{\psi}_i^K$ such that the earliest terms have the biggest impact on the reduced form. Given that $\boldsymbol{\psi}_i^K$ depends on K, the instrument sets need not form a nested sequence. K cannot be too large to avoid too variable estimator due to the selection process.

Define $\mathbf{\Psi}^K = \left(\boldsymbol{\psi}_1^K, \cdots, \boldsymbol{\psi}_n^K\right)'$, and $\mathbf{P}^K = \mathbf{\Psi}^K \left(\mathbf{\Psi}^{K\prime}\mathbf{\Psi}^K\right)^- \mathbf{\Psi}^{K\prime}$, where $\mathbf{A}^-$ denotes a generalized inverse.[7] $\mathbf{y}$,X_2,Z_1,X are defined by stacking the corresponding vectors. $\widehat{\Lambda}$ is the minimum of $(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\mathbf{P}^K(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})/(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})$, and $\overline{\Lambda} = (K-l_1-2)/n$. The estimators considered are

$$\begin{aligned}
\text{2SLS:} \qquad & \widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{P}^K\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}^K\mathbf{y}, \\
\text{LIML:} \qquad & \widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{P}^K\mathbf{X} - \widehat{\Lambda}\mathbf{X}'\mathbf{X})^{-1}\left(\mathbf{X}'\mathbf{P}^K\mathbf{y} - \widehat{\Lambda}\mathbf{X}'\mathbf{y}\right), \\
\text{B2SLS:} \qquad & \widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{P}^K\mathbf{X} - \overline{\Lambda}\mathbf{X}'\mathbf{X})^{-1}\left(\mathbf{X}'\mathbf{P}^K\mathbf{y} - \overline{\Lambda}\mathbf{X}'\mathbf{y}\right),
\end{aligned}$$

where B2SLS is a bias adjusted version of 2SLS (see Nagar (1959) and Rothenberg (1984)), it is a K-estimator with $\widehat{\kappa} = \frac{n}{n-(K-l_1)+2}$ and is shown to be unbiased to second order in the fixed-instrument, normal-error model in Rothenberg (1984). The instrument selection is based on minimizing the approximate MSE of a linear combination $\widehat{\boldsymbol{\lambda}}'\widehat{\boldsymbol{\beta}}$ of the IV estimator, where $\widehat{\boldsymbol{\lambda}}$ is some vector of estimated linear combination coefficients.

Estimating the MSE requires preliminary estimates of some of the parameters of the model and a goodness of fit criterion for estimation of the (first stage) reduced form using the instrument $\boldsymbol{\psi}_i^K$. The preliminary estimator can be either an IV estimator with only as many instruments as right-hand side variables or an IV estimator where instruments are chosen to minimize one of the first stage goodness of fit criteria below. Note that the preliminary estimator does not depend on K. Given some preliminary estimator of $\boldsymbol{\beta}$, say $\widetilde{\boldsymbol{\beta}}$, define

$$\widehat{\sigma}_u^2 = \widetilde{\mathbf{u}}'\widetilde{\mathbf{u}}/n,\ \widehat{\sigma}_{\boldsymbol{\lambda}}^2 = \widetilde{\mathbf{v}}_{\boldsymbol{\lambda}}'\widetilde{\mathbf{v}}_{\boldsymbol{\lambda}}/n,\ \widehat{\sigma}_{\boldsymbol{\lambda}u} = \widetilde{\mathbf{v}}_{\boldsymbol{\lambda}}'\widetilde{\mathbf{u}}/n,$$

where $\widetilde{\mathbf{u}} = \mathbf{y} - \mathbf{X}\widetilde{\boldsymbol{\beta}}$, $\widetilde{\mathbf{v}}_{\boldsymbol{\lambda}} = \widetilde{\mathbf{V}}\widetilde{\mathbf{H}}^{-1}\widetilde{\boldsymbol{\lambda}}$, $\widetilde{\mathbf{V}} = (\mathbf{I}-\mathbf{P}^K)\mathbf{X}$, and $\widetilde{\mathbf{H}} = \mathbf{X}'\mathbf{P}^K\mathbf{X}/n$ for some fixed K is an estimator of $\mathbf{f}'\mathbf{f}/n$ with $\mathbf{f} = [f(\mathbf{z}_1), \cdots, f(\mathbf{z}_n)]'$. As to goodness of fit criterion, cross-validation and Mallows (1973) reduced form goodness of fit criteria are considered. The cross-validation criterion is

$$\widehat{R}_{\boldsymbol{\lambda}}^{cv}(K) = \frac{1}{n}\sum_{i=1}^{n}\frac{\left(\widehat{\mathbf{v}}_{\boldsymbol{\lambda}i}^K\right)^2}{\left(1-\mathbf{P}_{ii}^K\right)^2}.$$

The Mallows criterion is

$$\widehat{R}_{\boldsymbol{\lambda}}^{m}(K) = \frac{\widehat{\mathbf{v}}_{\boldsymbol{\lambda}}^{K\prime}\widehat{\mathbf{v}}_{\boldsymbol{\lambda}}^K}{n} + \widehat{\sigma}_{\boldsymbol{\lambda}}^2\frac{2K}{n},$$

where $\widehat{\mathbf{v}}_{\boldsymbol{\lambda}}^K = \widehat{\mathbf{V}}^K\widetilde{\mathbf{H}}^{-1}\widehat{\boldsymbol{\lambda}}$ with $\widehat{\mathbf{V}}^K = (\mathbf{I}-\mathbf{P}^K)\mathbf{X}$, and $\mathbf{A}_{ij}$ denotes the i,jth element of a matrix $\mathbf{A}$.

[7] The generalized inverse of $\mathbf{A}$ is defined in Appendix C of Chapter 2.

Given these preparations, the approximate MSE of the estimators are

$$\begin{aligned}
\text{2SLS:} \qquad & \widehat{S}_{\boldsymbol{\lambda}}(K) = \widehat{\sigma}^2_{\boldsymbol{\lambda}u}\frac{K^2}{n} + \widehat{\sigma}^2_u\left(\widehat{R}_{\boldsymbol{\lambda}}(K) - \widehat{\sigma}^2_{\boldsymbol{\lambda}}\frac{K}{n}\right), \\
\text{LIML:} \qquad & \widehat{S}_{\boldsymbol{\lambda}}(K) = \widehat{\sigma}^2_u\left(\widehat{R}_{\boldsymbol{\lambda}}(K) - \frac{\widehat{\sigma}^2_{\boldsymbol{\lambda}u}}{\widehat{\sigma}^2_u}\frac{K}{n}\right), \\
\text{B2SLS:} \qquad & \widehat{S}_{\boldsymbol{\lambda}}(K) = \widehat{\sigma}^2_u\left(\widehat{R}_{\boldsymbol{\lambda}}(K) + \frac{\widehat{\sigma}^2_{\boldsymbol{\lambda}u}}{\widehat{\sigma}^2_u}\frac{K}{n}\right),
\end{aligned}$$

where the approach to calculating the approximate MSE is similar to Nagar's (1959). 2SLS includes a bias term $\widehat{\sigma}^2_{\boldsymbol{\lambda}u}\frac{K^2}{n}$, while LIML and B2SLS include only variance terms. The first variance term $\widehat{\sigma}^2_u\widehat{R}_{\boldsymbol{\lambda}}(K)$ is common and comes from approximating the reduced form $f(\mathbf{z})$ by linear combinations of $\boldsymbol{\psi}^K(\mathbf{z})$. The second variance term comes from the FOCs of the three estimators:

$$\begin{aligned}
\text{2SLS:} \qquad & \mathbf{X}'\mathbf{P}^K(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) = \mathbf{0}, \\
\text{LIML:} \qquad & \left(\mathbf{X} - \widehat{\mathbf{u}}\widehat{\boldsymbol{\alpha}}'\right)'\mathbf{P}^K(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) = \mathbf{0}, \widehat{\boldsymbol{\alpha}} = \mathbf{X}'\widehat{\mathbf{u}}/\widehat{\mathbf{u}}'\widehat{\mathbf{u}}, \\
\text{B2SLS:} \qquad & \mathbf{X}'\mathbf{P}^K(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) - (K - l_1 - 2)\mathbf{X}'\widehat{\mathbf{u}} = \mathbf{0},
\end{aligned}$$

where $\widehat{\mathbf{u}} = \mathbf{y} - \mathbf{X}'\widehat{\boldsymbol{\beta}}$ for each estimator $\widehat{\boldsymbol{\beta}}$, $\widehat{\mathbf{u}}\widehat{\boldsymbol{\alpha}}'$ in LIML eliminates an important source of 2SLS bias arising from the correlation between $\mathbf{X}$ and $\mathbf{u}$, and the bias correction for B2SLS subtracts an estimate of the bias from 2SLS FOCs. The first variance term decreases with K while the second increases with K. For each estimator, the $\widehat{K}$ that minimizes the corresponding $\widehat{S}_{\boldsymbol{\lambda}}(K)$ will result in $\widehat{\boldsymbol{\lambda}}'\widehat{\boldsymbol{\beta}}$ that has relatively small MSE asymptotically. For 2SLS, $\widehat{K}$ accounts for a trade-off between bias and variance, while for LIML and B2SLS, $\widehat{K}$ accounts for a trade-off only between variance terms. When $\dim(x_2) = 1$, the value of K that minimizes these criteria will not depend on $\widehat{\boldsymbol{\lambda}}$.

Donald and Newey show that their method can improve the finite sample properties of the three IV estimators in the sense that

$$\frac{S_{\boldsymbol{\lambda}}(\widehat{K})}{\min_K S_{\boldsymbol{\lambda}}(K)} \xrightarrow{p} 1,$$

where $S_{\boldsymbol{\lambda}}(\cdot)$ is the true dominant term of the exact MSE. They also compare the approximate MSE of these estimators, and find the LIML is best.

Their results also apply to the choice of nonlinear functions to use in the efficient semiparametric instrumental variables estimator of Newey (1990b). In this case instrument choice is analogous to choosing the smoothing parameter in semiparametric estimation. In Donald and Newey's framework, Kuersteiner and Okui (2010) extend the model averaging of Hansen (2007b) to the 2SLS estimation.

7 Extensions to GMM (*)

Note that for the asymptotic arguments in the previous sections to go through, we need some critical assumptions on the data generating process. Relaxing these assumptions is the task of

current econometric practices. We will overview some extensions in the literature. These extensions are interacted with each other and also with the alternative inference methods mentioned above.

From Appendix A, the asymptotic approximation in (7) requires at least the following six assumptions. We list these assumptions and the relevant literature to relax them.

(i) $\mathbf{w}_i$, $i = 1, \cdots, n$, is a random sample. If $\mathbf{w}_i$, $i = 1, \cdots, n$, are time series $\mathbf{w}_t$, $t = 1, \cdots, T$, such that $g(\mathbf{w}_t, \boldsymbol{\theta})$ are correlated, then the optimal

$$\begin{aligned} \boldsymbol{\Omega} &= TE\left[\overline{g}_T(\boldsymbol{\theta}_0)\overline{g}_T(\boldsymbol{\theta}_0)'\right] \\ &= \sum_{v=-\infty}^{\infty} E\left[g(\mathbf{w}_t, \boldsymbol{\theta}_0)g(\mathbf{w}_{t-v}, \boldsymbol{\theta}_0)'\right] \equiv \sum_{v=-\infty}^{\infty} \boldsymbol{\Omega}_v. \end{aligned}$$

A consistent estimator of $\boldsymbol{\Omega}$ is often called the *heteroskedasticity and autocorrelation consistent* (HAC) estimator. Reading starts from Newey and West (1987b, 1994), Andrews (1991), Andrews and Monahan (1992) and Phillips (2005).

(ii) $g(\mathbf{w}, \boldsymbol{\theta})$ is smooth in θ. When g is nondifferentiable and/or discontinuous in $\boldsymbol{\theta}$ (e.g., the moment conditions in quantile regression), the asymptotic arguments in Section 4 and the usual calculation algorithm for the GMM estimator may be problematic; see Pakes and Pollard (1989), Andrews (1997a) and Chernozhukov and Hong (2003) for classical references.[8]

(iii) $\mathbf{G}$ is full column rank. When this assumption fails, there is the weak or partial identification problem as mentioned in Chapter 1. Especially, when $\mathbf{G} \approx n^{-1/2}\mathbf{C}$, the instruments are weak, and $\boldsymbol{\theta}$ cannot be consistently estimated.[9] The 2SLS estimator is close to the LSE so suffers a serious bias problem. This strand of literature starts from Nelson and Startz (1990a,b) and Bound et al. (1995). Classical references include Staiger and Stock (1997) and Stock and Wright (2000) on estimation and Wang and Zivot (1998) [LR and LM tests], Moreira (2003) [two conditional tests] and Kleibergen (2002, 2005) [K-test] on inference. See Stock et al. (2002), Dufour (2003), Hahn and Hausman (2003), Andrews and Stocks (2007), Mikusheva (2013), Andrews et al. (2019) and Keane and Neal (2024) for summaries.

(iv) l is fixed. When l can go to infinity, there are many moment conditions which will increase the bias of the GMM estimator and deteriorates the estimation of $\boldsymbol{\Omega}$. Reading starts from Bekker (1994), Chao and Swanson (2005), Han and Phillips (2006), Newey and Windmeijer (2009) and Mikusheva and Sun (2022). See Anatolyev (2019) for a summary.

(v) k is fixed. When k can go to infinity, there are nonparametric parameters in the moment conditions. For identification, we need infinite moment conditions. Reading starts from Newey and Powell (2003), Ai and Chen (2003), Blundell and Powell (2003), Hall and Horowitz (2005) and Darolles et al. (2011). See Chen (2007) for a summary.

[8] Especially, Andrews (1997a) discusses the asymptotic properties of the J statistic while Pakes and Pollard (1989) do not.

[9] If $\mathbf{G} \approx \mathbf{C}n^{-1/2}$, then $E[g(\mathbf{w}_i, \boldsymbol{\beta})] \approx \mathbf{G}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) \approx n^{-1/2}\mathbf{C}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)$. As a result, $J_n(\boldsymbol{\beta})/n \approx 0$ even for $\boldsymbol{\beta} \neq \boldsymbol{\beta}_0$, and the identification fails.

(vi) There are only moment equalities. If there are moment inequalities, $\boldsymbol{\theta}$ can only be partially identified. Reading starts from Chernozhukov et al. (2007), Beresteanu and Molinari (2008), Pakes et al. (2015). See Molchanov and Molinari (2014, 2018), Bontemps and Magnac (2017), Canay and Shaikh (2017), Ho and Rosen (2017) and Molinari (2020) for summaries.

In the following, we illustrate the difficulties introduced by (iii) and (iv), i.e., weak instruments and many instruments, and we also show how to test whether the instruments are weak and how to conduct inference with weak instruments. Our discussions follow Sections 12.35-12.39 of Hansen (2022).

7.1 Identification Failure

Recall the reduced form equation

$$\mathbf{x}_2 = \boldsymbol{\Gamma}_{12}'\mathbf{z}_1 + \boldsymbol{\Gamma}_{22}'\mathbf{z}_2 + \mathbf{v}_2.$$

The parameter $\boldsymbol{\beta}$ fails to be identified if $\boldsymbol{\Gamma}_{22}$ has deficient rank. The consequences of identification failure for inference are quite severe.

Take the simplest case where $k_1 = 0$ and $k_2 = l_2 = 1$. Then the model may be written as

$$\begin{aligned} y &= x\beta + u. \\ x &= z\gamma + v, \end{aligned} \tag{12}$$

and $\boldsymbol{\Gamma}_{22} = \gamma = E[zx]/E\left[z^2\right]$. We see that β is identified if and only if $\gamma \neq 0$, which occurs when $E[zx] \neq 0$. Thus identification hinges on the existence of correlation between the excluded exogenous variable and the included endogenous variable.

Suppose this condition fails. In this case $\gamma = 0$ and $E[zx] = 0$. We now analyze the distribution of the least squares and IV estimators of β. For simplicity we assume conditional homoskedasticity and normalize the variances of u, v and z to unity. Thus

$$Var\left(\left.\begin{pmatrix} u \\ v \end{pmatrix}\right| z\right) = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}. \tag{13}$$

The errors have non-zero correlation $\rho \neq 0$ when the variables are endogenous.

By the CLT we have the joint convergence

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\begin{pmatrix} z_i u_i \\ z_i v_i \end{pmatrix} \xrightarrow{d} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} \sim N\left(0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right).$$

It is convenient to define $\xi_0 = \xi_1 - \rho\xi_2$ which is normal and independent of ξ_2.

As a benchmark it is useful to observe that the least squares estimator of β satisfies

$$\widehat{\beta}_{\mathrm{OLS}} - \beta = \frac{n^{-1}\sum_{i=1}^{n} v_i u_i}{n^{-1}\sum_{i=1}^{n} u_i^2} \xrightarrow{p} \rho \neq 0,$$

so endogeneity causes $\widehat{\beta}_{\mathrm{OLS}}$ to be inconsistent for β. Under identification failure $\gamma = 0$ the asymptotic distribution of the IV estimator is

$$\widehat{\beta}_{\mathrm{IV}} - \beta = \frac{n^{-1/2}\sum_{i=1}^{n} z_i u_i}{n^{-1/2}\sum_{i=1}^{n} z_i x_i} \xrightarrow{d} \frac{\xi_1}{\xi_2} = \rho + \frac{\xi_0}{\xi_2}.$$

This asymptotic convergence result uses the continuous mapping theorem which applies since the function ξ_1/ξ_2 is continuous everywhere except at $\xi_2 = 0$, which occurs with probability equal to zero.

This limiting distribution has several notable features. First, $\widehat{\beta}_{\mathrm{IV}}$ does not converge in probability to a limit, rather it converges in distribution to a random variable. Thus the IV estimator is inconsistent. Indeed, it is not possible to consistently estimate an unidentified parameter and β is not identified when $\gamma = 0$. Second, the ratio ξ_0/ξ_2 is symmetrically distributed about zero so the median of the limiting distribution of $\widehat{\beta}_{\mathrm{IV}}$ is $\beta + \rho$. This means that the IV estimator is median biased under endogeneity. Thus under identification failure the IV estimator does not correct the centering (median bias) of least squares. Third, the ratio ξ_0/ξ_2 of two independent normal random variables is Cauchy distributed. This is particularly nasty as the Cauchy distribution does not have a finite mean. The distribution has thick tails, meaning that extreme values occur with higher frequency than the normal. Inferences based on the normal distribution can be quite incorrect. Together, these results show that $\gamma = 0$ renders the IV estimator particularly poorly behaved – it is inconsistent, median biased, and non-normally distributed.

We can also examine the behavior of the t-statistic. For simplicity consider the classical (homoskedastic) t-statistic. The error variance estimate has the asymptotic distribution

$$\begin{aligned}
\widehat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}\left(y_i - x_i\widehat{\beta}_{\mathrm{IV}}\right)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} u_i^2 - \frac{2}{n}\sum_{i=1}^{n} u_i x_i\left(\widehat{\beta}_{\mathrm{IV}} - \beta\right) + \frac{1}{n}\sum_{i=1}^{n} x_i^2\left(\widehat{\beta}_{\mathrm{IV}} - \beta\right)^2 \\
&\xrightarrow{d} 1 - 2\rho\frac{\xi_1}{\xi_2} + \left(\frac{\xi_1}{\xi_2}\right)^2.
\end{aligned}$$

Thus the t-statistic has the asymptotic distribution

$$t_n = \frac{\widehat{\beta}_{\mathrm{IV}} - \beta}{\sqrt{\widehat{\sigma}^2 \frac{1}{n}\sum_{i=1}^{n} z_i^2} / \left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n} z_i x_i\right|} \xrightarrow{d} \frac{\xi_1/\xi_2}{\sqrt{1 - 2\rho\frac{\xi_1}{\xi_2} + \left(\frac{\xi_1}{\xi_2}\right)^2} / |\xi_2|} = \operatorname{sign}(\xi_2)\frac{\xi_1}{\sqrt{1 - 2\rho\frac{\xi_1}{\xi_2} + \left(\frac{\xi_1}{\xi_2}\right)^2}}.$$

The limiting distribution is non-normal, meaning that inference using the normal distribution will

be (considerably) incorrect. This distribution depends on the correlation ρ. The distortion is increasing in ρ. Indeed as $\rho \to 1$ we have $\xi_1/\xi_2 \xrightarrow{p} 1$ and the unexpected finding $\widehat{\sigma}^2 \xrightarrow{p} 0$. The latter means that the conventional standard error $s\left(\widehat{\beta}_{\text{IV}}\right)$ for $\widehat{\beta}_{\text{IV}}$ also converges in probability to zero. This implies that the t-statistic diverges in the sense $|t_n| \xrightarrow{p} \infty$. In this situations users may incorrectly interpret estimates as precise despite the fact that they are highly imprecise.

7.2 Weak Instruments

In the previous subsection we examined the extreme consequences of full identification failure. Similar problems occur when identification is weak in the sense that the reduced form coefficients are of small magnitude. In this section we derive the asymptotic distribution of the OLS, 2SLS, and LIML estimators when the reduced form coefficients are treated as weak. We show that the estimators are inconsistent and the 2SLS and LIML estimators remain random in large samples.

To simplify the exposition we assume that there are no included exogenous variables (no $\mathbf{x}_1$) so we write $\mathbf{x}_2, \mathbf{z}_2$, and $\boldsymbol{\beta}_2$ simply as $\mathbf{x}, \mathbf{z}$, and $\boldsymbol{\beta}$. The model is

$$\begin{aligned} y &= \mathbf{x}'\boldsymbol{\beta} + u, \\ \mathbf{x} &= \boldsymbol{\Gamma}'\mathbf{z} + \mathbf{v}_2. \end{aligned} \tag{14}$$

Recall the reduced form error vector $\mathbf{f} = (e, \mathbf{v}_2')'$ and asume its covariance matrix

$$E\left[\mathbf{f}\mathbf{f}'|\mathbf{z}\right] = \boldsymbol{\Sigma} = \begin{pmatrix} \Sigma_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}. \tag{15}$$

Recall that the structural error is $u = e - \boldsymbol{\beta}'\mathbf{v}_2 = \boldsymbol{\gamma}'\mathbf{f}$, which has variance $E\left[u^2|\mathbf{z}\right] = \boldsymbol{\gamma}'\boldsymbol{\Sigma}\boldsymbol{\gamma}$, where $\boldsymbol{\gamma} = \left(1, -\boldsymbol{\beta}'\right)'$. Also define the covariance $\boldsymbol{\Sigma}_{2u} = E\left[\mathbf{v}_2 u|\mathbf{z}\right] = \boldsymbol{\Sigma}_{21} - \boldsymbol{\Sigma}_{22}\boldsymbol{\beta}$.

In the last subsection we assumed complete identification failure in the sense that $\boldsymbol{\Gamma} = \mathbf{0}$. We now want to assume that identification does not completely fail but is weak in the sense that $\boldsymbol{\Gamma}$ is small. A rich asymptotic distribution theory has been developed to understand this setting by modeling $\boldsymbol{\Gamma}$ as "local-to-zero". The seminal contribution is Staiger and Stock (1997). The theory was extended to nonlinear GMM estimation by Stock and Wright (2000).

The technical device introduced by Staiger and Stock (1997) is to assume that the reduced form parameter is **local-to-zero**, specifically

$$\boldsymbol{\Gamma} = n^{-1/2}\mathbf{C}. \tag{16}$$

where $\mathbf{C}$ is a free matrix. The $n^{-1/2}$ scaling is picked because it provides just the right balance to allow a useful distribution theory. The local-to-zero assumption (16) is not meant to be taken literally but rather is meant to be a useful distributional approximation. The parameter $\mathbf{C}$ indexes the degree of identification. Larger $\|\mathbf{C}\|$ implies stronger identification; smaller $\|\mathbf{C}\|$ implies weaker identification.

We now derive the asymptotic distribution of the least squares, 2SLS, and LIML estimators

under the local-to-unity assumption (16). The least squares estimator satisfies

$$\widehat{\boldsymbol{\beta}}_{\text{OLS}} - \boldsymbol{\beta} = \left(n^{-1}\mathbf{X}'\mathbf{X}\right)^{-1}\left(n^{-1}\mathbf{X}'\mathbf{u}\right) = \left(n^{-1}\mathbf{V}_2'\mathbf{V}_2\right)^{-1}\left(n^{-1}\mathbf{V}_2'\mathbf{u}\right) + o_p(1) \xrightarrow{p} \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{2u}.$$

Thus the least squares estimator is inconsistent for $\boldsymbol{\beta}$.

To examine the 2SLS estimator, by the central limit theorem

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{z}_i\mathbf{f}_i' \xrightarrow{d} \boldsymbol{\xi} = (\boldsymbol{\xi}_1, \boldsymbol{\xi}_2),$$

where $\text{vec}(\boldsymbol{\xi}) \sim N\left(0, E\left[\mathbf{f}\mathbf{f}' \otimes \mathbf{z}\mathbf{z}'\right]\right)$. This implies

$$\frac{1}{\sqrt{n}}\mathbf{Z}'\mathbf{u} \xrightarrow{d} \boldsymbol{\xi}_u = \boldsymbol{\xi}'\boldsymbol{\gamma}.$$

We also find that

$$\frac{1}{\sqrt{n}}\mathbf{Z}'\mathbf{X} = \frac{1}{n}\mathbf{Z}'\mathbf{Z}\mathbf{C} + \frac{1}{\sqrt{n}}\mathbf{Z}'\mathbf{V}_2 \xrightarrow{d} \mathbf{Q}_{\mathbf{z}}\mathbf{C} + \boldsymbol{\xi}_2.$$

Thus

$$\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\mathbf{X} = \left(\frac{1}{\sqrt{n}}\mathbf{X}'\mathbf{Z}\right)\left(\frac{1}{n}\mathbf{Z}'\mathbf{Z}\right)^{-1}\left(\frac{1}{\sqrt{n}}\mathbf{Z}'\mathbf{X}\right) \xrightarrow{d} (\mathbf{Q}_{\mathbf{z}}\mathbf{C} + \boldsymbol{\xi}_2)'\,\mathbf{Q}_{\mathbf{z}}^{-1}\,(\mathbf{Q}_{\mathbf{z}}\mathbf{C} + \boldsymbol{\xi}_2)$$

and

$$\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\mathbf{u} = \left(\frac{1}{\sqrt{n}}\mathbf{X}'\mathbf{Z}\right)\left(\frac{1}{n}\mathbf{Z}'\mathbf{Z}\right)^{-1}\left(\frac{1}{\sqrt{n}}\mathbf{Z}'\mathbf{u}\right) \xrightarrow{d} (\mathbf{Q}_{\mathbf{z}}\mathbf{C} + \boldsymbol{\xi}_2)'\,\mathbf{Q}_{\mathbf{z}}^{-1}\boldsymbol{\xi}_u.$$

We find that the 2SLS estimator has the asymptotic distribution

$$\widehat{\boldsymbol{\beta}}_{\text{2SLS}} - \boldsymbol{\beta} = \left(\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\mathbf{X}\right)^{-1}\left(\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\mathbf{u}\right) \xrightarrow{d} \left[(\mathbf{Q}_{\mathbf{z}}\mathbf{C} + \boldsymbol{\xi}_2)'\,\mathbf{Q}_{\mathbf{z}}^{-1}\,(\mathbf{Q}_{\mathbf{z}}\mathbf{C} + \boldsymbol{\xi}_2)\right]^{-1}(\mathbf{Q}_{\mathbf{z}}\mathbf{C} + \boldsymbol{\xi}_2)'\,\mathbf{Q}_{\mathbf{z}}^{-1}\boldsymbol{\xi}_u. \tag{17}$$

As in the case of complete identification failure we find that $\widehat{\boldsymbol{\beta}}_{\text{2SLS}}$ is inconsistent for $\boldsymbol{\beta}$, it is asymptotically random, and its asymptotic distribution is non-normal. The distortion is affected by the coefficient $\mathbf{C}$. As $\|\mathbf{C}\| \to \infty$, the distribution in (17) converges in probability to zero suggesting that $\widehat{\boldsymbol{\beta}}_{\text{2SLS}}$ is consistent for $\boldsymbol{\beta}$. This corresponds to the classic "strong identification" context.

Now consider the LIML estimator. The reduced form is $\mathbf{Y} = \mathbf{Z}\boldsymbol{\Pi} + \mathbf{V}$. This implies $\mathbf{M}_{\mathbf{Z}}\mathbf{Y} = \mathbf{M}_{\mathbf{Z}}\mathbf{V}$ and by standard asymptotic theory

$$\frac{1}{n}\mathbf{Y}'\mathbf{M}_{\mathbf{Z}}\mathbf{Y} = \frac{1}{n}\mathbf{V}'\mathbf{M}_{\mathbf{Z}}\mathbf{V} \xrightarrow{p} \boldsymbol{\Sigma} = E\left[\mathbf{v}\mathbf{v}'\right].$$

Define $\overline{\boldsymbol{\beta}} = (\boldsymbol{\beta}, \mathbf{I}_k)$ so that the reduced form coefficients equal $\boldsymbol{\Pi} = (\boldsymbol{\Gamma}\boldsymbol{\beta}, \boldsymbol{\Gamma}) = n^{-1/2}\mathbf{C}\overline{\boldsymbol{\beta}}$. Then

$$\frac{1}{\sqrt{n}}\mathbf{Z}'\mathbf{Y} = \frac{1}{n}\mathbf{Z}'\mathbf{Z}\mathbf{C}\overline{\boldsymbol{\beta}} + \frac{1}{\sqrt{n}}\mathbf{Z}'\mathbf{V} \xrightarrow{d} \mathbf{Q}_{\mathbf{z}}\mathbf{C}\overline{\boldsymbol{\beta}} + \boldsymbol{\xi}$$

and

$$\mathbf{Y}'\mathbf{Z}\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{Y} \xrightarrow{d} \left(\mathbf{Q}_{\mathbf{z}}\mathbf{C}\overline{\boldsymbol{\beta}}+\boldsymbol{\xi}\right)'\mathbf{Q}_{\mathbf{z}}^{-1}\left(\mathbf{Q}_{\mathbf{z}}\mathbf{C}\overline{\boldsymbol{\beta}}+\boldsymbol{\xi}\right).$$

This allows us to calculate that by the continuous mapping theorem

$$n\widehat{\mu} = \min_{\boldsymbol{\gamma}} \frac{\boldsymbol{\gamma}'\mathbf{Y}'\mathbf{Z}\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{Y}\boldsymbol{\gamma}}{\boldsymbol{\gamma}'\mathbf{Y}'\mathbf{M}_{\mathbf{Z}}\mathbf{Y}\boldsymbol{\gamma}} \xrightarrow{d} \min_{\boldsymbol{\gamma}} \frac{\boldsymbol{\gamma}'\left(\mathbf{Q}_{\mathbf{z}}\mathbf{C}\overline{\boldsymbol{\beta}}+\boldsymbol{\xi}\right)'\mathbf{Q}_{\mathbf{z}}^{-1}\left(\mathbf{Q}_{\mathbf{z}}\mathbf{C}\overline{\boldsymbol{\beta}}+\boldsymbol{\xi}\right)\boldsymbol{\gamma}}{\boldsymbol{\gamma}'\boldsymbol{\Sigma}\boldsymbol{\gamma}} \equiv \mu^*,$$

which is function of $\boldsymbol{\xi}$ and thus random. We deduce that the asymptotic distribution of the LIML estimator is

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{\text{LIML}} - \boldsymbol{\beta} &= \left(\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\mathbf{X} - n\widehat{\mu}\frac{1}{n}\mathbf{X}'\mathbf{M}_{\mathbf{Z}}\mathbf{X}\right)^{-1}\left(\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\mathbf{u} - n\widehat{\mu}\frac{1}{n}\mathbf{X}'\mathbf{M}_{\mathbf{Z}}\mathbf{u}\right) \\ &\xrightarrow{d} \left[\left(\mathbf{Q}_{\mathbf{z}}\mathbf{C}+\boldsymbol{\xi}_2\right)'\mathbf{Q}_{\mathbf{z}}^{-1}\left(\mathbf{Q}_{\mathbf{z}}\mathbf{C}+\boldsymbol{\xi}_2\right) - \mu^*\boldsymbol{\Sigma}_{22}\right]^{-1}\left[\left(\mathbf{Q}_{\mathbf{z}}\mathbf{C}+\boldsymbol{\xi}_2\right)'\mathbf{Q}_{\mathbf{z}}^{-1}\boldsymbol{\xi}_u - \mu^*\boldsymbol{\Sigma}_{22}\boldsymbol{\Sigma}_{2u}\right].\end{aligned}$$

Similarly to 2SLS, the LIML estimator is inconsistent for $\boldsymbol{\beta}$, is asymptotically random, and non-normally distributed.

We summarize

Theorem 1 *Under (16),*

$$\begin{aligned}&\widehat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta} \xrightarrow{p} \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{2u}, \\ &\widehat{\boldsymbol{\beta}}_{2SLS} - \boldsymbol{\beta} \xrightarrow{d} \left[\left(\mathbf{Q}_{\mathbf{z}}\mathbf{C}+\boldsymbol{\xi}_2\right)'\mathbf{Q}_{\mathbf{z}}^{-1}\left(\mathbf{Q}_{\mathbf{z}}\mathbf{C}+\boldsymbol{\xi}_2\right)\right]^{-1}\left(\mathbf{Q}_{\mathbf{z}}\mathbf{C}+\boldsymbol{\xi}_2\right)'\mathbf{Q}_{\mathbf{z}}^{-1}\boldsymbol{\xi}_u,\end{aligned}$$

and

$$\widehat{\boldsymbol{\beta}}_{LIML} - \boldsymbol{\beta} \xrightarrow{d} \left[\left(\mathbf{Q}_{\mathbf{z}}\mathbf{C}+\boldsymbol{\xi}_2\right)'\mathbf{Q}_{\mathbf{z}}^{-1}\left(\mathbf{Q}_{\mathbf{z}}\mathbf{C}+\boldsymbol{\xi}_2\right) - \mu^*\boldsymbol{\Sigma}_{22}\right]^{-1}\left[\left(\mathbf{Q}_{\mathbf{z}}\mathbf{C}+\boldsymbol{\xi}_2\right)'\mathbf{Q}_{\mathbf{z}}^{-1}\boldsymbol{\xi}_u - \mu^*\boldsymbol{\Sigma}_{22}\boldsymbol{\Sigma}_{2u}\right],$$

where

$$\mu^* = \min_{\boldsymbol{\gamma}} \frac{\boldsymbol{\gamma}'\left(\mathbf{Q}_{\mathbf{z}}\mathbf{C}\overline{\boldsymbol{\beta}}+\boldsymbol{\xi}\right)'\mathbf{Q}_{\mathbf{z}}^{-1}\left(\mathbf{Q}_{\mathbf{z}}\mathbf{C}\overline{\boldsymbol{\beta}}+\boldsymbol{\xi}\right)\boldsymbol{\gamma}}{\boldsymbol{\gamma}'\boldsymbol{\Sigma}\boldsymbol{\gamma}}$$

and $\overline{\boldsymbol{\beta}} = (\boldsymbol{\beta}, \mathbf{I}_k)$.

All three estimators are inconsistent. The 2SLS and LIML estimators are asymptotically random with non-standard distributions, similar to the asymptotic distribution of the IV estimator under complete identification failure explored in the previous subsection. The difference under weak identification is the presence of the coefficient matrix $\mathbf{C}$.

7.3 Many Instruments

Some applications have available a large number l of instruments. If they are all valid, using a large number should reduce the asymptotic variance relative to estimation with a smaller number of instruments. Is it then good practice to use many instruments? Or is there a cost to this practice? Bekker (1994) initiated a large literature investigating this question by formalizing the

idea of "many instruments". Bekker proposed an asymptotic approximation which treats the number of instruments l as proportional to the sample size, that is, $l = \alpha n$, or equivalently that $l/n \to \alpha \in [0, 1)$. The distributional theory obtained is similar in many respects to the weak instrument theory outlined in the previous subsection. Consequently the impact of "weak" and "many" instruments is similar.

Again for simplicity we assume that there are no included exogenous regressors so that the model is (14). We also make the simplifying assumption that the reduced form errors are conditionally homoskedastic, i.e., (15). In addition we assume that the conditional fourth moments are bounded

$$E\left[\|\mathbf{f}\|^4 |\mathbf{z}\right] \leq B < \infty. \tag{18}$$

The idea that there are "many instruments" is formalized by the assumption that the number of instruments is increasing proportionately with the sample size

$$\frac{l}{n} \to \alpha. \tag{19}$$

The best way to think about this is to view α as the ratio of l to n in a given sample. Thus if an application has $n = 100$ observations and $l = 10$ instruments, then we should treat $\alpha = 0.10$.

Suppose that there is a single endogenous regressor x. Calculate its variance using the reduced form: $Var(x) = Var(\mathbf{z}'\boldsymbol{\Gamma}) + Var(v)$. Suppose as well that $Var(x)$ and $Var(u)$ are unchanging as l increases. This implies that $Var(\mathbf{z}'\boldsymbol{\Gamma})$ is unchanging even though the dimension l is increasing. This is a useful assumption as it implies that the population R^2 of the reduced form is not changing with l. We don't need this exact condition, rather we simply assume that the sample version converges in probability to a fixed constant. Specifically, we assume that

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\Gamma}'\mathbf{z}_i\mathbf{z}_i'\boldsymbol{\Gamma} \xrightarrow{p} \mathbf{H} > 0. \tag{20}$$

Again, this essentially implies that the R^2 of the reduced form regressions for each component of $\mathbf{x}$ converge to constants.

As a baseline it is useful to examine the behavior of the least squares estimator of $\boldsymbol{\beta}$. First, observe that the variance of $\text{vec}\left(n^{-1}\sum_{i=1}^{n} \boldsymbol{\Gamma}'\mathbf{z}_i\mathbf{f}_i'\right)$ conditional on $\mathbf{Z}$, is

$$\boldsymbol{\Sigma} \otimes n^{-2}\sum_{i=1}^{n} \boldsymbol{\Gamma}'\mathbf{z}_i\mathbf{z}_i'\boldsymbol{\Gamma} \xrightarrow{p} \mathbf{0}$$

by (20). Thus it converges in probability to zero:

$$n^{-1}\sum_{i=1}^{n} \boldsymbol{\Gamma}'\mathbf{z}_i\mathbf{f}_i' \xrightarrow{p} \mathbf{0}. \tag{21}$$

Combined with (20) and the WLLN we find

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i u_i = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\Gamma}'\mathbf{z}_i u_i + \frac{1}{n}\sum_{i=1}^{n}\mathbf{v}_{2i}u_i \xrightarrow{p} \boldsymbol{\Sigma}_{2u}$$

and

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i' = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\Gamma}'\mathbf{z}_i\mathbf{z}_i'\boldsymbol{\Gamma} + \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\Gamma}'\mathbf{z}_i\mathbf{v}_{2i}' + \frac{1}{n}\sum_{i=1}^{n}\mathbf{v}_{2i}\mathbf{z}_i'\boldsymbol{\Gamma} + \frac{1}{n}\sum_{i=1}^{n}\mathbf{v}_{2i}\mathbf{v}_{2i}' \xrightarrow{p} \mathbf{H} + \boldsymbol{\Sigma}_{22}.$$

Hence

$$\widehat{\boldsymbol{\beta}}_{\text{ols}} - \boldsymbol{\beta} = \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i u_i\right) \xrightarrow{p} (\mathbf{H} + \boldsymbol{\Sigma}_{22})^{-1}\boldsymbol{\Sigma}_{2u},$$

i.e., the LSE is inconsistent for $\boldsymbol{\beta}$.

Now consider the 2SLS estimator:

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{\text{2SLS}} - \boldsymbol{\beta} &= \left(\frac{1}{n}\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\mathbf{X}\right)^{-1}\left(\frac{1}{n}\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\mathbf{u}\right) \\ &= \left(\frac{1}{n}\boldsymbol{\Gamma}'\mathbf{Z}'\mathbf{Z}\boldsymbol{\Gamma} + \frac{1}{n}\boldsymbol{\Gamma}'\mathbf{Z}'\mathbf{V}_2 + \frac{1}{n}\mathbf{V}_2'\mathbf{Z}\boldsymbol{\Gamma} + \frac{1}{n}\mathbf{V}_2'\mathbf{P}_{\mathbf{Z}}\mathbf{V}_2\right)^{-1}\left(\frac{1}{n}\boldsymbol{\Gamma}'\mathbf{Z}'\mathbf{u} + \frac{1}{n}\mathbf{V}_2'\mathbf{P}_{\mathbf{Z}}\mathbf{u}\right), \quad (22)\end{aligned}$$

In the expression on the right-side of (22) several of the components have been examined in (20) and (21). We now examine the remaining components $\frac{1}{n}\mathbf{V}_2'\mathbf{P}_{\mathbf{Z}}\mathbf{V}_2$ and $\frac{1}{n}\mathbf{V}_2'\mathbf{P}_{\mathbf{Z}}\mathbf{u}$ which are sub-components of the matrix $\frac{1}{n}\mathbf{V}'\mathbf{P}_{\mathbf{Z}}\mathbf{V}$. Take the jk^{th} element $\frac{1}{n}\mathbf{V}_j'\mathbf{P}_{\mathbf{Z}}\mathbf{V}_k$.

First, take its expectation. We have (given under the conditional homoskedasticity assumption (15))

$$E\left[\frac{1}{n}\mathbf{V}_j'\mathbf{P}_{\mathbf{Z}}\mathbf{V}_k\middle|\,\mathbf{Z}\right] = \frac{1}{n}\text{tr}\left(E\left[\mathbf{P}_{\mathbf{Z}}\mathbf{V}_k\mathbf{V}_j'\middle|\,\mathbf{Z}\right]\right) = \frac{1}{n}\text{tr}\left(\mathbf{P}_{\mathbf{Z}}\right)\boldsymbol{\Sigma}_{jk} = \frac{l}{n}\boldsymbol{\Sigma}_{jk} \to \alpha\boldsymbol{\Sigma}_{jk} \quad (23)$$

using $\text{tr}(\mathbf{P}_{\mathbf{Z}}) = l$.

Second, we calculate its variance which is a more cumbersome exercise. Let $P_{im} = \mathbf{Z}_i'\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}_m$ be the im^{th} element of $\mathbf{P}_{\mathbf{Z}}$. Then $\mathbf{V}_j'\mathbf{P}_{\mathbf{Z}}\mathbf{V}_k = \sum_{i=1}^{n}\sum_{m=1}^{n}u_{ji}u_{km}P_{im}$. The matrix $\mathbf{P}_{\mathbf{Z}}$ is idempotent. It therefore has the properties $\sum_{i=1}^{n}P_{ii} = \text{tr}(\mathbf{P}_{\mathbf{Z}}) = l$ and $0 \leq P_{ii} \leq 1$. The property

$\mathbf{P_Z P_Z} = \mathbf{P_Z}$ also implies $\sum_{m=1}^{n} P_{im}^2 = P_{ii}$. Then

$$\begin{aligned}
Var\left(\frac{1}{n}\mathbf{V}_j'\mathbf{P_Z}\mathbf{V}_k\middle|\mathbf{Z}\right) &= \frac{1}{n^2}E\left[\left(\sum_{i=1}^{n}\sum_{m=1}^{n}\left(u_{ji}u_{km} - E\left[u_{ji}u_{km}\right]1\left\{i=m\right\}\right)P_{im}\right)^2\middle|\mathbf{Z}\right] \\
&= \frac{1}{n^2}E\left[\sum_{i=1}^{n}\sum_{m=1}^{n}\sum_{q=1}^{n}\sum_{r=1}^{n}\left(u_{ji}u_{km} - \mathbf{\Sigma}_{jk}1\left\{i=m\right\}\right)P_{im}\left(u_{jq}u_{kr} - \mathbf{\Sigma}_{jk}1\left\{q=r\right\}\right)P_{qr}\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\left(u_{ji}u_{ki} - \mathbf{\Sigma}_{jk}\right)^2 P_{ii}^2 \\
&\quad + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{m\neq i}E\left[u_{ji}^2u_{km}^2\right]P_{im}^2 + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{m\neq i}E\left[u_{ji}u_{km}u_{jm}u_{ki}\right]P_{im}^2 \\
&\leq \frac{B}{n^2}\left(\sum_{i=1}^{n}P_{ii}^2 + 2\sum_{i=1}^{n}\sum_{m=1}^{n}P_{im}^2\right) \\
&\leq \frac{3B}{n^2}\sum_{i=1}^{n}P_{ii} = 3B\frac{l}{n^2} \to 0.
\end{aligned}$$

The third equality holds because the remaining cross-products have zero expectation as the observations are independent and the errors have zero mean. The first inequality is (18). The second uses $\sum_{i=1}^{n} P_{ii}^2 \leq \sum_{i=1}^{n} P_{ii}$ and $\sum_{m=1}^{n} P_{im}^2 = P_{ii}$. The final equality is $\sum_{i=1}^{n} P_{ii} = l$.

Using (19), (23), Markov's inequality, and combining across all j and k we deduce that

$$\frac{1}{n}\mathbf{V}'\mathbf{P_Z}\mathbf{V} \to \alpha\mathbf{\Sigma}. \tag{24}$$

Returning to the 2SLS estimator (22) and combining (20), (21), and (24), we find

$$\widehat{\boldsymbol{\beta}}_{2\text{SLS}} - \boldsymbol{\beta} \xrightarrow{p} \left(\mathbf{H} + \alpha\mathbf{\Sigma}_{22}\right)^{-1}\alpha\mathbf{\Sigma}_{2u}.$$

Thus 2SLS is also inconsistent for $\boldsymbol{\beta}$. The limit, however, depends on the magnitude of α.

We finally examine the LIML estimator. (24) implies

$$\frac{1}{n}\mathbf{Y}'\mathbf{M_Z}\mathbf{Y} = \frac{1}{n}\mathbf{V}'\mathbf{V} - \frac{1}{n}\mathbf{V}'\mathbf{P_Z}\mathbf{V} \xrightarrow{p} \left(1-\alpha\right)\mathbf{\Sigma}.$$

Similarly,

$$\begin{aligned}
\frac{1}{n}\mathbf{Y}'\mathbf{Z}\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{Y} &= \overline{\boldsymbol{\beta}}'\mathbf{\Gamma}'\left(\frac{1}{n}\mathbf{Z}'\mathbf{Z}\right)\mathbf{\Gamma}\overline{\boldsymbol{\beta}} + \overline{\boldsymbol{\beta}}'\mathbf{\Gamma}'\left(\frac{1}{n}\mathbf{Z}'\mathbf{V}\right) + \left(\frac{1}{n}\mathbf{V}'\mathbf{Z}\right)\mathbf{\Gamma}\overline{\boldsymbol{\beta}} + \frac{1}{n}\mathbf{V}'\mathbf{P_Z}\mathbf{V} \\
&\xrightarrow{p} \overline{\boldsymbol{\beta}}'\mathbf{H}\overline{\boldsymbol{\beta}}' + \alpha\mathbf{\Sigma}.
\end{aligned}$$

Hence

$$\widehat{\mu} = \min_{\boldsymbol{\gamma}}\frac{\boldsymbol{\gamma}'\mathbf{Y}'\mathbf{Z}\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{Y}\boldsymbol{\gamma}}{\boldsymbol{\gamma}'\mathbf{Y}'\mathbf{M_Z}\mathbf{Y}\boldsymbol{\gamma}} \xrightarrow{p} \min_{\boldsymbol{\gamma}}\frac{\boldsymbol{\gamma}'\left(\overline{\boldsymbol{\beta}}'\mathbf{H}\overline{\boldsymbol{\beta}}' + \alpha\mathbf{\Sigma}\right)\boldsymbol{\gamma}}{\boldsymbol{\gamma}'\left(1-\alpha\right)\mathbf{\Sigma}\boldsymbol{\gamma}} = \frac{\alpha}{1-\alpha}$$

and

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{\text{LIML}}-\boldsymbol{\beta} &= \left(\frac{1}{n}\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\mathbf{X}-\widehat{\mu}\frac{1}{n}\mathbf{X}'\mathbf{M}_{\mathbf{Z}}\mathbf{X}\right)^{-1}\left(\frac{1}{n}\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\mathbf{u}-\widehat{\mu}\frac{1}{n}\mathbf{X}'\mathbf{M}_{\mathbf{Z}}\mathbf{u}\right)\\ &\xrightarrow{p}\left(\mathbf{H}+\alpha\boldsymbol{\Sigma}_{22}-\frac{\alpha}{1-\alpha}(1-\alpha)\boldsymbol{\Sigma}_{22}\right)^{-1}\left(\alpha\boldsymbol{\Sigma}_{2u}-\frac{\alpha}{1-\alpha}(1-\alpha)\boldsymbol{\Sigma}_{2u}\right)\\ &= \mathbf{H}^{-1}\mathbf{0}=\mathbf{0}.\end{aligned}$$

Thus LIML is consistent for $\boldsymbol{\beta}$ unlike 2SLS.

We state these results formally.

Theorem 2 *In model (14), under assumptions (15), (18) and (19), then as $n\to\infty$,*

$$\begin{aligned}&\widehat{\boldsymbol{\beta}}_{OLS}-\boldsymbol{\beta}\xrightarrow{p}\left(\mathbf{H}+\boldsymbol{\Sigma}_{22}\right)^{-1}\boldsymbol{\Sigma}_{2u},\\ &\widehat{\boldsymbol{\beta}}_{2SLS}-\boldsymbol{\beta}\xrightarrow{p}\left(\mathbf{H}+\alpha\boldsymbol{\Sigma}_{22}\right)^{-1}\alpha\boldsymbol{\Sigma}_{2u},\\ &\widehat{\boldsymbol{\beta}}_{LIML}-\boldsymbol{\beta}\xrightarrow{p}\mathbf{0}.\end{aligned}$$

This result is quite insightful. It shows that while endogeneity ($\boldsymbol{\Sigma}_{2u}\neq\mathbf{0}$) renders the least squares estimator inconsistent, the 2SLS estimator is also inconsistent if the number of instruments diverges proportionately with n. The limit in Theorem 2 shows a continuity between least squares and 2SLS. The probability limit of the 2SLS estimator is continuous in α with the extreme case ($\alpha=1$) implying that 2SLS and least squares have the same probability limit. The general implication is that the inconsistency of 2SLS is increasing in α. The theorem also shows that unlike 2SLS the LIML estimator is consistent under the many instruments assumption. Effectively, LIML makes a bias-correction.

Theorem 1 (weak instruments) and 2 (many instruments) tell a cautionary tale. They show that when instruments are weak and/or many, the 2SLS estimator is inconsistent. The degree of inconsistency depends on the weakness of the instruments (the magnitude of the matrix $\mathbf{C}$ in Theorem 1) and the degree of overidentification (the ratio α in Theorem 2). The theorems also show that the LIML estimator is inconsistent under the weak instrument assumption but with a bias-correction, and is consistent under the many instrument assumption. This suggests that LIML is more robust than 2SLS to weak and many instruments.

An important limitation of the results in Theorem 2 is the assumption of conditional homoskedasticity. It appears likely that the consistency of LIML fails in the many instrument setting if the errors are heteroskedastic.

In applications users should be aware of the potential consequences of the many instrument framework. It is useful to calculate the "many instrument ratio" $\alpha=l/n$. While there is no specific rule-of thumb for α which leads to acceptable inference a minimum criterion is that if $\alpha\geq 0.05$ you should be seriously concerned about the many-instrument problem. In general, when α is large it seems preferable to use LIML instead of 2SLS.

7.4 Testing for Weak Instruments

In the previous subsections we found that weak instruments results in non-standard asymptotic distributions for the 2SLS and LIML estimators. In practice how do we know if this is a problem? Is there a way to check if the instruments are weak?

This question was addressed in an influential paper by Stock and Yogo (2005) as an extension of Staiger and Stock (1997). Stock-Yogo focus on two implications of weak instruments: (1) estimation bias and (2) inference distortion. They show how to test the hypothesis that these distortions are not "too big". They propose F-tests for the excluded instruments in the reduced form regressions with non-standard critical values. In particular, when there is one endogenous regressor and a single instrument the Stock-Yogo test rejects the null of weak instruments when this F-statistic exceeds 10. While Stock and Yogo explore two types of distortions, we focus exclusively on inference as that is the more challenging problem. In this subsection we describe the Stock-Yogo theory and tests for the case of a single endogenous regressor ($k_2 = 1$). In the following subsection we describe their method for the case of multiple endogeneous regressors.

While the theory in Stock and Yogo allows for an arbitrary number of exogenous regressors and instruments, for the sake of clear exposition we will focus on the very simple case of no included exogenous variables ($k_1 = 0$) and just one exogenous instrument ($l_2 = 1$) which is model (12). Furthermore, as in Section 7.1 we assume conditional homoskedasticity and normalize the variances as in (13). Since the model is just-identified the 2SLS, LIML, and IV estimators are all equivalent.

The question of primary interest is to determine conditions on the reduced form under which the IV estimator of the structural equation is well behaved, and secondly, what statistical tests can be used to learn if these conditions are satisfied. As in Section 7.2 we assume that the reduced form coefficient Γ is local-to-zero, specifically $\Gamma = n^{-1/2}\mu$. The asymptotic distribution of the IV estimator is presented in Theorem 1. Given the simplifying assumptions the result is

$$\widehat{\beta}_{\mathrm{IV}} - \beta \xrightarrow{d} \frac{\xi_u}{\mu + \xi_2},$$

where (ξ_u, ξ_2) are bivariate normal, and $\xi_u = \xi_1$ in Section 7.1. For inference we also examine the behavior of the classical (homoskedastic) t-statistic for the IV estimator. Note

$$\begin{aligned}
\widehat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}\left(y_i - x_i\widehat{\beta}_{\mathrm{IV}}\right)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} u_i^2 - \frac{2}{n}\sum_{i=1}^{n} u_i x_i\left(\widehat{\beta}_{\mathrm{IV}} - \beta\right) + \frac{1}{n}\sum_{i=1}^{n} x_i^2\left(\widehat{\beta}_{\mathrm{IV}} - \beta\right)^2 \\
&\xrightarrow{d} 1 - 2\rho\frac{\xi_u}{\mu + \xi_2} + \left(\frac{\xi_u}{\mu + \xi_2}\right)^2.
\end{aligned}$$

Thus

$$t_n = \frac{\widehat{\beta}_{\mathrm{IV}} - \beta}{\sqrt{\widehat{\sigma}^2 \frac{1}{n}\sum_{i=1}^n z_i^2 / \left|\frac{1}{\sqrt{n}}\sum_{i=1}^n z_i x_i\right|}} \xrightarrow{d} \frac{\xi_u / (\mu + \xi_2)}{\sqrt{1 - 2\rho\frac{\xi_u}{\mu+\xi_2} + \left(\frac{\xi_u}{\mu+\xi_2}\right)^2} / |\mu + \xi_2|}$$

$$= \operatorname{sign}(\mu + \xi_2) \frac{\xi_u}{\sqrt{1 - 2\rho\frac{\xi_u}{\mu+\xi_2} + \left(\frac{\xi_u}{\mu+\xi_2}\right)^2}} \equiv S. \tag{25}$$

In general, S is non-normal and its distribution depends on the parameters ρ and μ.

Can we use the distribution S for inference on β? The distribution depends on two unknown parameters and neither is consistently estimable. This means we cannot use the distribution in (25) with ρ and μ replaced with estimates. To eliminate the dependence on ρ one possibility is to use the "worst case" value which turns out to be $\rho = 1$. By worst-case we mean the value which causes the greatest distortion away from normal critical values. Setting $\rho = 1$ we have the considerable simplification

$$S = S_1 = \xi\left(1 + \frac{\xi}{\mu}\right)\operatorname{sign}(\mu), \tag{26}$$

where $\xi \sim N(0,1)$. When the model is strongly identified (so $|\mu|$ is very large) then $S_1 \overset{d}{\approx} \xi$ is standard normal, consistent with classical theory. However when μ is very small (but non-zero) $|S_1| \approx \xi^2/|\mu|$ (in the sense that this term dominates), which is a scaled χ^2 and quite far from normal. As $|\mu| \to 0$ we find the extreme case $|S_1| \xrightarrow{p} \infty$.

While (26) is a convenient simplification it does not yield a useful approximation for inference as the distribution in (26) is highly dependent on the unknown μ. If we take the worst-case value of μ, which is $\mu = 0$, we find that $|S_1|$ diverges and all distributional approximations fail.

To break this impasse Stock and Yogo (2005) recommended a constructive alternative. Rather than using the worst-case μ they suggested finding a threshold such that if μ exceeds this threshold then the distribution (26) is not "too badly" distorted from the normal distribution.

Specifically, the Stock-Yogo recommendation can be summarized by two steps. First, the distribution result (26) can be used to find a threshold value τ^2 such that if $\mu^2 \geq \tau^2$ then the size of the nominal 5% test "Reject if $|t_n| \geq 1.96$" has asymptotic size $P(|S_1| \geq 1.96) \leq 0.15$. This means that while the goal is to obtain a test with size 5%, we recognize that there may be size distortion due to weak instruments and are willing to tolerate a specific distortion. For example, a 10% distortion means we allow the actual size to be up to 15%. Second, they use the asymptotic distribution of the reduced-form (first stage) F-statistic to test if the actual unknown value of μ^2 exceeds the threshold τ^2. These two steps together give rise to the rule-of-thumb that the first-stage F-statistic should exceed 10 in order to achieve reliable IV inference. (This is for the case of one instrumental variable. If there is more than one instrument then the rule-of-thumb changes.) We now describe the steps behind this reasoning in more detail.

The first step is to use the distribution (25) to determine the threshold τ^2. Formally, the goal is to find the value of $\tau^2 = \mu^2$ at which the asymptotic size of a nominal 5% test is actually a given

r (e.g., $r = 0.15$), thus $P(|S_1| \geq 1.96) \leq r$. By some algebra and the quadratic formula the event $\left|\xi\left(1+\frac{\xi}{\mu}\right)\right| < x$ is the same as

$$\frac{\mu^2}{4} - x\mu < \left(\xi + \frac{\mu}{2}\right)^2 < \frac{\mu^2}{4} + x\mu$$

as $\mu > 0$. The random variable between the inequalities is distributed $\chi^2_1(\mu^2/4)$, a noncentral chi-square with one degree of freedom and noncentrality parameter $\mu^2/4$. Thus

$$\begin{aligned} P(|S_1| \geq x) &= P\left(\chi^2_1(\mu^2/4) \geq \frac{\mu^2}{4} + x\mu\right) + P\left(\chi^2_1(\mu^2/4) \leq \frac{\mu^2}{4} - x\mu\right) \\ &= 1 - G\left(\frac{\mu^2}{4} + x\mu, \frac{\mu^2}{4}\right) + G\left(\frac{\mu^2}{4} - x\mu, \frac{\mu^2}{4}\right), \end{aligned} \tag{27}$$

where $G(u, \lambda)$ is the distribution function of $\chi^2_1(\lambda)$. Hence the desired threshold τ^2 solves

$$1 - G\left(\frac{\tau^2}{4} + 1.96\tau, \frac{\tau^2}{4}\right) + G\left(\frac{\tau^2}{4} - 1.96\tau, \frac{\tau^2}{4}\right) = r$$

or effectively

$$G\left(\frac{\tau^2}{4} + 1.96\tau, \frac{\tau^2}{4}\right) = 1 - r$$

because $\frac{\tau^2}{4} - 1.96\tau < 0$ for relevant values of τ. The numerical solution (computed with the non-central chi-square distribution function, e.g. ncx2cdf in MATLAB) is $\tau^2 = 1.70$ when $r = 0.15$. (That is, the command

$$\mathsf{ncx2cdf}(1.7/4 + 1.96 * \mathsf{sqrt}(1.7), 1, 1.7/4)$$

yields the answer 0.85. Stock and Yogo (2005) approximate the same calculation using simulation methods and report $\tau^2 = 1.82$.) This calculation means that if the reduced form satisfies $\mu^2 > 1.7$, or equivalently if $\Gamma^2 \geq 1.7/n$, then the asymptotic size of a nominal 5% test on the structural parameter is no larger than 15%.

To summarize the Stock-Yogo first step, we calculate the minimum value τ^2 for μ^2 sufficient to ensure that the asymptotic size of a nominal 5% t-test does not exceed r, and find that $\tau^2 = 1.70$ for $r = 0.15$.

The Stock-Yogo second step is to find a critical value for the first-stage F-statistic sufficient to reject the hypothesis that $H_0 : \mu^2 = \tau^2$ against $H_1 : \mu^2 > \tau^2$. We now describe this procedure.

They suggest testing $H_0 : \mu^2 = \tau^2$ at the 5% size using the first stage F-statistic. If the F-statistic is small so that the test does not reject then we should be worried that the true value of μ^2 is small and there is a weak instrument problem. On the other hand if the F-statistic is large so that the test rejects then we can have some confidence that the true value of μ^2 is sufficiently large that the weak instrument problem is not too severe.

To implement the test we need to calculate an appropriate critical value. It should be calculated under the null hypothesis $H_0 : \mu^2 = \tau^2$. This is different from a conventional F-test which is

calculated under $H_0 : \mu^2 = 0$.

We start by calculating the asymptotic distribution of F. Since there is one regressor and one instrument in our simplified setting the first-stage F-statistic is the squared t-statistic from the reduced form. Given our previous calculations it has the asymptotic distribution

$$F = \frac{\widehat{\gamma}^2}{s\left(\widehat{\gamma}\right)^2} = \frac{\left(\sum_{i=1}^n z_i x_i\right)^2}{\left(\sum_{i=1}^n z_i^2\right)\widehat{\sigma}_v^2} \xrightarrow{d} \left(\mu + \xi_2\right)^2 \sim \chi_1^2\left(\mu^2\right). \tag{28}$$

This is a non-central chi-square distribution $G\left(u, \mu^2\right)$ with one degree of freedom and non-centrality parameter μ^2.

To test $H_0 : \mu^2 = \tau^2$ against $H_1 : \mu^2 > \tau^2$ we reject for $F \geq c$ where c is selected so that the asymptotic rejection probability satisfies

$$P\left(F \geq c | \mu^2 = \tau^2\right) \to P\left(\chi_1^2\left(\tau^2\right) \geq c\right) = 1 - G\left(c, \tau^2\right) = 0.05$$

for $\tau^2 = 1.70$, equivalently $G\left(c, 1.7\right) = 0.95$. This is found by inverting the non-central chi-square quantile function, e.g. the function $Q\left(p, d\right)$ which solves $G(Q(p, d), d) = p$. We find that $c = Q(0.95, 1.7) = 8.7$. In MATLAB, this can be computed by ncx2inv$(.95, 1.7)$. Stock and Yogo (2005) report $c = 9.0$ because they used $\tau^2 = 1.82$.

This means that if $F > 8.7$ we can reject $H_0 : \mu^2 = 1.7$ against $H_1 : \mu^2 > 1.7$ with an asymptotic 5% test. In this context we should expect the IV estimator and tests to be reasonably well behaved. However, if $F < 8.7$ then we should be cautious about the IV estimator, confidence intervals, and tests. This finding led Staiger and Stock (1997) to propose the informal "rule of thumb" that the first stage F statistic should exceed 10. Notice that F exceeding 8.7 (or 10) is equivalent to the reduced form t-statistic exceeding 2.94 (or 3.16), which is considerably larger than a conventional check if the t-statistic is "significant". Equivalently, the recommended rule-of-thumb for the case of a single instrument is to estimate the reduced form and verify that the t-statistic for exclusion of the instrumental variable exceeds 3 in absolute value.

Does the proposed procedure control the asymptotic size of a 2SLS test? The first step has asymptotic size bounded below r (e.g., 15%). The second step has asymptotic size 5%. By the Bonferroni bound the two steps together have asymptotic size bounded below $r + 0.05$ (e.g., 20%). We can thus call the Stock-Yogo procedure a rigorous test with asymptotic size $r + 0.05$ (or 20%).

Our analysis has been confined to the case $k_2 = l_2 = 1$. Stock and Yogo (2005) also examine the case $l_2 > 1$ (which requires numerical simulation to solve) and both the 2SLS and LIML estimators. They show that the F-statistic critical values depend on the number of instruments l_2 as well as the estimator. Their critical values (calculated by simulation) are in their paper and posted on Motohiro Yogo's webpage. We report a subset in Table 1.

	Maximal Size r							
	2SLS				LIML			
l_2	0.10	0.15	0.20	0.25	0.10	0.15	0.20	0.25
1	16.4	9.0	6.7	5.5	16.4	9.0	6.7	5.5
2	19.9	11.6	8.7	7.2	8.7	5.3	4.4	3.9
3	22.3	12.8	9.5	7.8	6.5	4.4	3.7	3.3
4	24.6	14.0	10.3	8.3	5.4	3.9	3.3	3.0
5	26.9	15.1	11.0	8.8	4.8	3.6	3.0	2.8
6	29.2	16.2	11.7	9.4	4.4	3.3	2.9	2.6
7	31.5	17.4	12.5	9.9	4.2	3.2	2.7	2.5
8	33.8	18.5	13.2	10.5	4.0	3.0	2.6	2.4
9	36.2	19.7	14.0	11.1	3.8	2.9	2.5	2.3
10	38.5	20.9	14.8	11.6	3.7	2.8	2.5	2.2
15	50.4	26.8	18.7	12.2	3.3	2.5	2.2	2.0
20	62.3	32.8	22.7	17.6	3.2	2.3	2.1	1.9
25	74.2	38.8	26.7	20.6	3.8	2.2	2.0	1.8
30	86.2	44.8	30.7	23.6	3.9	2.2	1.9	1.7

Table 1: 5% Critical Value for Weak Instruments, $k_2 = 1$

One striking feature about these critical values is that those for the 2SLS estimator are strongly increasing in l_2 while those for the LIML estimator are decreasing in l_2 (except for $r = 0.10$ where the critical values are increasing for large l_2). This means that when the number of instruments l_2 is large, 2SLS requires a much stronger reduced form (larger μ^2) in order for inference of β_2 to be reliable, but this is not the case for LIML. This is direct evidence that LIML inference is less sensitive to weak instruments than 2SLS. This makes a strong case for LIML over 2SLS, especially when l_2 is large or the instruments are potentially weak.

We now summarize the recommended Staiger-Stock/Stock-Yogo procedure for $k_1 \geq 1$, $k_2 = 1$, and $l_2 \geq 1$. The structural equation and reduced form equations are

$$\begin{aligned} y_1 &= \mathbf{z}_1'\boldsymbol{\beta}_1 + y_2\beta_2 + u, \\ y_2 &= \mathbf{z}_1'\boldsymbol{\gamma}_1 + \mathbf{z}_2'\gamma_2 + v, \end{aligned}$$

where following the literature, the endogenous variable x_2 is denoted as y_2. The structural equation is estimated by either 2SLS or LIML. Let F be the F-statistic for $H_0 : \gamma_2 = 0$ in the reduced form equation. Let $s\left(\widehat{\beta}_2\right)$ be a standard error for β_2 in the structural equation. The procedure is:

1. Compare F with the critical values c in Table 1 with the row selected to match the number of excluded instruments l_2 and the columns to match the estimation method (2SLS or LIML) and the desired size r.

2. If $F > c$ then report the 2SLS or LIML estimates with conventional inference.

There are possible extensions to the Stock-Yogo procedure.

One modest extension is to use the information to convey the degree of confidence in the accuracy of a confidence interval. Suppose in an application you have $l_2 = 5$ excluded instruments and have estimated your equation by 2SLS. Now suppose that your reduced form F-statistic equals 12. You check Table 1 and find that $F = 12$ is significant with $r = 0.20$. Thus we can interpret the conventional 2SLS confidence interval as having coverage of 80% (or 75% if we make the Bonferroni correction). On the other hand if $F = 27$ we would conclude that the test for weak instruments is significant with $r = 0.10$, meaning that the conventional 2SLS confidence interval can be interpreted as having coverage of 90% (or 85% after Bonferroni correction). Thus the value of the F-statistic can be used to calibrate the coverage accuracy.

A more substantive extension, which we now discuss, reverses the steps. Unfortunately this discussion will be limited to the case $l_2 = 1$. First, use the reduced form F-statistic to find a one-sided confidence interval for μ^2 of the form $[\mu_L^2, \infty)$. Second, use the lower bound μ_L^2 to calculate a critical value c for S_1 such that the 2SLS test has asymptotic size bounded below 0.05. This produces better size control than the Stock-Yogo procedure and produces more informative confidence intervals for β_2. We now describe the steps in detail.

The first goal is to find a one-sided confidence interval for μ^2. This is found by test inversion. As we described earlier, for any τ^2 we reject $H_0 : \mu^2 = \tau^2$ in favor of $H_0 : \mu^2 > \tau^2$ if $F > c$ where $G(c, \tau^2) = 0.95$. Equivalently, we reject if $G(F, \tau^2) > 0.95$. By the test inversion principle an asymptotic 95% confidence interval $[\mu_L^2, \infty)$ is the set of all values of τ^2 which are not rejected. Since $G(F, \tau^2) \leq 0.95$ for all τ^2 in this set, the lower bound μ_L^2 satisfies $G(F, \tau^2) = 0.95$, and is found numerically. In MATLAB, the solution is mu2 when ncx2cdf(F,1,mu2) returns 0.95.

The second goal is to find the critical value c such that $P(|S_1| \geq c) = 0.05$ when $\mu^2 = \mu_L^2$. From (27) this is achieved when

$$1 - G\left(\frac{\mu_L^2}{4} + c\mu_L, \frac{\mu_L^2}{4}\right) + G\left(\frac{\mu_L^2}{4} - c\mu_L, \frac{\mu_L^2}{4}\right) = 0.05. \tag{29}$$

This an be solved as

$$G\left(\frac{\mu_L^2}{4} + c\mu_L, \frac{\mu_L^2}{4}\right) = 0.95.$$

(The third term on the left-hand-side of (29) is zero for all solutions so can be ignored.) Using the non-central chi-square quantile function $Q(p, d)$, this c equals

$$c = \frac{Q(0.95, \frac{\mu_L^2}{4}) - \frac{\mu_L^2}{4}}{\mu_L}.$$

For example, in MATLAB this is found as c=(ncx2inv(.95,1,mu2/4)-mu2/4)/sqrt(mu2). 95% confidence intervals for β_2 are then calculated as $\widehat{\beta}_2 \pm c \cdot s\left(\widehat{\beta}_2\right)$.

We can also calculate a p-value for the t-statistic t_n for β_2. This is

$$p = 1 - G\left(\frac{\mu_L^2}{4} + |t_n|\,\mu_L, \frac{\mu_L^2}{4}\right) + G\left(\frac{\mu_L^2}{4} - |t_n|\,\mu_L, \frac{\mu_L^2}{4}\right),$$

where the third term equals zero if $|t_n| > \mu_L/4$. In MATLAB, for example, this can be calculated by the commands

```
T1= mu2/4+abs(T) *sqrt(mu2);
T2= mu2/4-abs(T) *sqrt(mu2);
p= -ncx2cdf(T1,1,mu2/4) +ncx2cdf(T2,1,mu2/4);
```

These confidence intervals and p-values will be larger than the conventional intervals and p-values, reflecting the incorporation of information about the strength of the instruments through the first-stage F-statistic. Also, by the Bonferroni bound these tests have asymptotic size bounded below 10% and the confidence intervals have asymptotic converage exceeding 90%, unlike the Stock-Yogo method which has size of 20% and coverage of 80%.

We have described an extension to the Stock-Yogo procedure for the case of one instrumental variable $l_2 = 1$. This restriction was due to the use of the analytic formula (29) for the asymptotic distribution which is only available when $l_2 = 1$. In principle the procedure could be extended using simulation or bootstrap methods but this has not been done to my knowledge.

The weak instrument methods described here are important for applied econometrics as they discipline researchers to assess the quality of their reduced form relationships before reporting structural estimates. The theory, however, has limitations and shortcomings, in particular the strong assumption of conditional homoskedasticity. Despite this limitation, in practice researchers apply the Stock-Yogo recommendations to estimates computed with heteroskedasticity-robust standard errors. This is an active area of research so the recommended methods may change in the years ahead.

7.5 Weak Instruments with $k_2 > 1$

When there is more than one endogenous regressor ($k_2 > 1$) it is better to examine the reduced form as a system. Staiger and Stock (1997) and Stock and Yogo (2005) provided an analysis of this case and constructed a test for weak instruments. The theory is considerably more involved than the $k_2 = 1$ case so we briefly summarize it here excluding many details, emphasizing their suggested methods.

The structural equation and reduced form equations are

$$\begin{aligned} y_1 &= \mathbf{z}_1'\boldsymbol{\beta}_1 + \mathbf{y}_2'\boldsymbol{\beta}_2 + u, \\ \mathbf{y}_2 &= \boldsymbol{\Gamma}_{12}'\mathbf{z}_1 + \boldsymbol{\Gamma}_{22}'\mathbf{z}_2 + \mathbf{v}_2, \end{aligned}$$

where following the literature, the endogenous variables $\mathbf{x}_2$ is denoted as $\mathbf{y}_2$. As in the previous subsection we assume that the errors are conditionally homoskedastic.

Identification of $\boldsymbol{\beta}_2$ requires the matrix $\boldsymbol{\Gamma}_{22}$ to be full rank. A necessary condition is that each row of $\boldsymbol{\Gamma}_{22}$ is non-zero but this is not sufficient.

We focus on the size performance of the homoskedastic Wald statistic for the 2SLS estimator of $\boldsymbol{\beta}_2$. For simplicity assume that the variance of u is known and normalized to one. Using representation in Exercise 15 of Chapter 7, the Wald statistic can be written as

$$W_n = \mathbf{u}'\widetilde{\mathbf{Z}}_2\left(\widetilde{\mathbf{Z}}_2'\widetilde{\mathbf{Z}}_2\right)^{-1}\widetilde{\mathbf{Z}}_2'\mathbf{Y}_2\left(\mathbf{Y}_2'\widetilde{\mathbf{Z}}_2\left(\widetilde{\mathbf{Z}}_2'\widetilde{\mathbf{Z}}_2\right)^{-1}\widetilde{\mathbf{Z}}_2'\mathbf{Y}_2\right)^{-1}\left(\mathbf{Y}_2'\widetilde{\mathbf{Z}}_2\left(\widetilde{\mathbf{Z}}_2'\widetilde{\mathbf{Z}}_2\right)^{-1}\widetilde{\mathbf{Z}}_2'\mathbf{u}\right),$$

where $\widetilde{\mathbf{Z}}_2 = \mathbf{M}_{\mathbf{Z}_1}\mathbf{Z}_2$.

Recall from Section 7.2 that Stock and Staiger model the excluded instruments $\mathbf{z}_2$ as weak by setting $\boldsymbol{\Gamma}_{22} = n^{-1/2}\mathbf{C}$ for some matrix $\mathbf{C}$. In this framework we have the asymptotic distribution results

$$\frac{1}{n}\widetilde{\mathbf{Z}}_2'\widetilde{\mathbf{Z}}_2 \xrightarrow{p} \mathbf{Q} = E\left[\mathbf{z}_2\mathbf{z}_2'\right] - E\left[\mathbf{z}_2\mathbf{z}_1'\right]\left(E\left[\mathbf{z}_1\mathbf{z}_1'\right]\right)^{-1}E\left[\mathbf{z}_1\mathbf{z}_2'\right]$$

and

$$\frac{1}{\sqrt{n}}\widetilde{\mathbf{Z}}_2'\mathbf{u} \xrightarrow{d} \mathbf{Q}^{1/2}\boldsymbol{\xi}_0,$$

where $\boldsymbol{\xi}_0 \sim N(\mathbf{0}, \mathbf{I})$. Furthermore, setting $\boldsymbol{\Sigma} = E\left[\mathbf{v}_2\mathbf{v}_2'\right]$ and $\overline{\mathbf{C}} = \mathbf{Q}^{1/2}\mathbf{C}\boldsymbol{\Sigma}^{-1/2}$,

$$\frac{1}{\sqrt{n}}\widetilde{\mathbf{Z}}_2'\mathbf{Y}_2 = \frac{1}{n}\widetilde{\mathbf{Z}}_2'\widetilde{\mathbf{Z}}_2\mathbf{C} + \frac{1}{\sqrt{n}}\widetilde{\mathbf{Z}}_2'\mathbf{V}_2 \xrightarrow{d} \mathbf{Q}^{1/2}\overline{\mathbf{C}}\boldsymbol{\Sigma}^{1/2} + \mathbf{Q}^{1/2}\boldsymbol{\xi}_2\boldsymbol{\Sigma}^{1/2},$$

where $\boldsymbol{\xi}_2$ is a matrix normal variate whose columns are independent $N(\mathbf{0}, \mathbf{I})$. The variables $\boldsymbol{\xi}_0$ and $\boldsymbol{\xi}_2$ are correlated. Together we obtain the asymptotic distribution of the Wald statistic

$$W_n \xrightarrow{d} S = \boldsymbol{\xi}_0'\left(\overline{\mathbf{C}} + \boldsymbol{\xi}_2\right)\left[\left(\overline{\mathbf{C}} + \boldsymbol{\xi}_2\right)'\left(\overline{\mathbf{C}} + \boldsymbol{\xi}_2\right)\right]^{-1}\left(\overline{\mathbf{C}} + \boldsymbol{\xi}_2\right)'\boldsymbol{\xi}_0.$$

Note that $\overline{\mathbf{C}}'\boldsymbol{\xi}_0 \sim N\left(\mathbf{0}, \overline{\mathbf{C}}'\overline{\mathbf{C}}\right)$, $\mathrm{vec}\left(\overline{\mathbf{C}}'\boldsymbol{\xi}_2\right) \sim N\left(0, \mathbf{I}_{k_2} \otimes \left(\overline{\mathbf{C}}'\overline{\mathbf{C}}\right)\right)$, and by the spectral decomposition, $\overline{\mathbf{C}}'\overline{\mathbf{C}} = \mathbf{H}'\boldsymbol{\Lambda}\mathbf{H}$ with $\mathbf{H}'\mathbf{H} = \mathbf{I}$ and $\boldsymbol{\Lambda}$ diagonal. So the asymptotic distribution of the Wald statistic is non-standard and a function of the model parameters only through the eigenvalues of $\overline{\mathbf{C}}'\overline{\mathbf{C}}$ and the correlations between the normal variates $\boldsymbol{\xi}_0$ and $\boldsymbol{\xi}_2$.[10] The worst-case can be summarized by the maximal correlation between $\boldsymbol{\xi}_0$ and $\boldsymbol{\xi}_2$ and the smallest eigenvalue of $\overline{\mathbf{C}}'\overline{\mathbf{C}}$. To mimic the F-statistic, we rescale the latter by dividing by the number of excluded instruments l_2. Define

$$\mathbf{G} = \overline{\mathbf{C}}'\overline{\mathbf{C}}/l_2 = \boldsymbol{\Sigma}^{-1/2}\mathbf{C}'\mathbf{Q}\mathbf{C}\boldsymbol{\Sigma}^{-1/2}/l_2$$

and

$$g = \lambda_{\min}\left(\mathbf{G}\right) = \lambda_{\min}\left(\boldsymbol{\Sigma}^{-1/2}\mathbf{C}'\mathbf{Q}\mathbf{C}\boldsymbol{\Sigma}^{-1/2}\right)/l_2,$$

where $\overline{\mathbf{C}}'\overline{\mathbf{C}}$ is the matrix counterpart of the concentration parameter $\boldsymbol{\Gamma}'\mathbf{Z}\mathbf{Z}\boldsymbol{\Gamma}/\boldsymbol{\sigma}_v^2$ as $l_1 = k_1 = 0$ and $k_2 = 1$.

[10] Rigorously, this result is shown in the many instruments framework with $l_2^4/n \to 0$.

This can be estimated from the reduced-form regression

$$\mathbf{y}_2 = \widehat{\mathbf{\Gamma}}_{12}'\mathbf{z}_1 + \widehat{\mathbf{\Gamma}}_{22}'\mathbf{z}_2 + \widehat{\mathbf{v}}_2.$$

The estimator is

$$\begin{aligned}
\widehat{\mathbf{G}} &= \widehat{\mathbf{\Sigma}}^{-1/2}\widehat{\mathbf{\Gamma}}_{22}'\left(\widetilde{\mathbf{Z}}_2'\widetilde{\mathbf{Z}}_2\right)\widehat{\mathbf{\Gamma}}_{22}\widehat{\mathbf{\Sigma}}^{-1/2}/l_2 = \widehat{\mathbf{\Sigma}}^{-1/2}\left(\mathbf{Y}_2'\widetilde{\mathbf{Z}}_2\right)\left(\widetilde{\mathbf{Z}}_2'\widetilde{\mathbf{Z}}_2\right)^{-1}\left(\widetilde{\mathbf{Z}}_2\mathbf{Y}_2\right)\widehat{\mathbf{\Sigma}}^{-1/2}/l_2, \\
\widehat{\mathbf{\Sigma}} &= \frac{1}{n-l}\sum_{i=1}^{n}\widehat{\mathbf{v}}_{2i}\widehat{\mathbf{v}}_{2i}' = \frac{1}{n-l}\mathbf{Y}_2'\mathbf{M}_{\mathbf{Z}}\mathbf{Y}_2 = \frac{1}{n-l}\widetilde{\mathbf{Y}}_2'\mathbf{M}_{\widetilde{\mathbf{Z}}_2}\widetilde{\mathbf{Y}}_2, \\
\widehat{g} &= \lambda_{\min}\left(\widehat{\mathbf{G}}\right).
\end{aligned}$$

$\widehat{\mathbf{G}}$ is a matrix F-type statistic for the coefficient matrix $\widehat{\mathbf{\Gamma}}_{22}$, and has the limiting distribution $\left(\overline{\mathbf{C}}+\boldsymbol{\xi}_2\right)'\left(\overline{\mathbf{C}}+\boldsymbol{\xi}_2\right)/l_2$. In summary, the asymptotic distribution of $l_2\widehat{\mathbf{G}}$ is a noncentral Wishart distribution of dimension k_2 with degree of freedom l_2 and noncentrality matrix $\overline{\mathbf{C}}'\overline{\mathbf{C}}$. Here, the noncentral Wishart distribution is a multiple dimensional generalization of the noncentral chi-square distribution; comparing to $\chi_1^2\left(\mu^2\right)$ in (28), we can see $\overline{\mathbf{C}}'\overline{\mathbf{C}}$ plays the role of μ^2 in the $k_2 = 1$ case. By the CMT, $\widehat{g}$ converges to the minimum eigenvalue of a noncentral Wishart distribution, divided by l_2.

The statistic $\widehat{g}$ was proposed by Cragg and Donald (1993) as a test for underidentification. Stock and Yogo (2005) use it as a test for weak instruments. Using simulation methods they determined critical values for $\widehat{g}$ similar to those for $k_2 = 1$. For given size $r > 0.05$ there is a critical value c (reported in the table below) such that if $\widehat{g} > c$ then the 2SLS (or LIML) Wald statistic W_n for $\widehat{\boldsymbol{\beta}}_2$ has asymptotic size bounded below r. On the other hand, if $\widehat{g} \leq c$ then we cannot bound the asymptotic size below r and we cannot reject the hypothesis of weak instruments.

Critical values (calculated by simulation) are reported in their paper and posted on Motohiro Yogo's webpage. We report a subset for the case $k_2 = 2$ in Table 2. The methods and theory applies to the cases $k_2 > 2$ as well but those critical values have not been calculated. As for the $k_2 = 1$ case the critical values for 2SLS are dramatically increasing in l_2. Thus when the model is over-identified, we need a large value of $\widehat{g}$ to reject the hypothesis of weak instruments. This is a strong cautionary message to check the $\widehat{g}$ statistic in applications. Furthermore, the critical values for LIML are generally decreasing in l_2 (except for $r = 0.10$ where the critical values are increasing for large l_2). This means that for over-identified models LIML inference is less sensitive to weak instruments than 2SLS and may be the preferred estimation method.

	Maximal Size r							
	2SLS				LIML			
l_2	0.10	0.15	0.20	0.25	0.10	0.15	0.20	0.25
2	7.0	4.6	3.9	3.6	7.0	4.6	3.9	3.6
3	13.4	8.2	6.4	5.4	5.4	3.8	3.3	3.1
4	16.9	9.9	7.5	6.3	4.7	3.4	3.0	2.8
5	19.4	11.2	8.4	6.9	4.3	3.1	2.8	2.6
6	21.7	12.3	9.1	7.4	4.1	2.9	2.6	2.5
7	23.7	13.3	9.8	7.9	3.9	2.8	2.5	2.4
8	25.6	14.3	10.4	8.4	3.8	2.7	2.4	2.3
9	27.5	15.2	11.0	8.8	3.7	2.7	2.4	2.2
10	29.3	16.2	11.6	9.3	3.6	2.6	2.3	2.1
15	38.0	20.6	14.6	11.6	3.5	2.4	2.1	2.0
20	46.6	25.0	17.6	13.8	3.6	2.4	2.0	1.9
25	55.1	29.3	20.6	16.1	3.6	2.4	1.97	1.8
30	63.5	33.6	23.5	18.3	4.1	2.4	1.95	1.7

Table 2: 5% Critical Value for Weak Instruments, $k_2 = 2$

Exercise 14 *When $\sigma_u^2 \equiv E\left[u^2\right] \neq 1$, what is the form of the Wald statistic, and what is its asymptotic distribution?*

Exercise 15 *In the LIML estimator, what is the asymptotic distribution of $n\left(\widehat{\kappa} - 1\right)$ when $\mathbf{\Gamma}_{22} = n^{-1/2}\mathbf{C}$?*

7.6 Tests with Nonhomoskedatic Errors

The results of Stock and Yogo (2005) rely heavily on the assumption of homoskedasticity, i.e., the data are independent, and (15) holds. When $k_2 = 1$ (for simplicity, assume $k_1 = 0$), the asymptotic variance of $\left(\widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\gamma}}\right)$ can be written as the Kronecker product of a 2×2 matrix with a $l \times l$ matrix. In the nonhomoskedastic case (e.g., heteroskedastic $\mathbf{f}$, and/or data dependency across i due to clustering or time-series correlation), this is not the case such that Stock and Yogo's procedure cannot be applied. As a robust alternative, Monteil Olea and Pflueger (2013) propose the effective F-statistic, which adds a multiplicative correction to the conventional first-stage F-statistic for testing $\boldsymbol{\gamma} = \mathbf{0}$ in models with homoskedastic errors.

The nonhomoskedasticity-robust F-statistics

$$F_R = \frac{1}{l}\widehat{\boldsymbol{\gamma}}'\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}^{-1}\widehat{\boldsymbol{\gamma}}$$

for $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}$ a robust estimator for the variance of $\widehat{\boldsymbol{\gamma}}$, and the traditional nonrobust F-statistics

$$F_N = \frac{1}{l}\widehat{\boldsymbol{\gamma}}'\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\gamma}\boldsymbol{\gamma},N}^{-1}\widehat{\boldsymbol{\gamma}} = \frac{n}{k\widehat{\sigma}_v^2}\widehat{\boldsymbol{\gamma}}'\widehat{\mathbf{Q}}_{\mathbf{z}}\widehat{\boldsymbol{\gamma}}$$

for $\widehat{\mathbf{\Sigma}}_{\boldsymbol{\gamma\gamma},N} = \frac{\widehat{\sigma}_v^2}{n}\widehat{\mathbf{Q}}_{\mathbf{z}}^{-1}$ with $\widehat{\mathbf{Q}}_{\mathbf{z}} = \frac{1}{n}\mathbf{Z}'\mathbf{Z}$ and $\widehat{\sigma}_v^2 = \frac{1}{n}\sum_{i=1}^n \widehat{v}_i^2$. There is no theoretical justification for the use of either F_N or F_R to gauge instrument strength in nonhomoskedastic settings. As an alternative, the effective F-statistic of Monteil Olea and Pflueger (2013) is

$$F_{\text{Eff}} = \frac{\widehat{\boldsymbol{\gamma}}'\widehat{\mathbf{Q}}_{\mathbf{z}}\widehat{\boldsymbol{\gamma}}}{\text{tr}\left(\widehat{\mathbf{\Sigma}}_{\boldsymbol{\gamma\gamma}}\widehat{\mathbf{Q}}_{\mathbf{z}}\right)} = \frac{k\widehat{\sigma}_v^2/n}{\text{tr}\left(\widehat{\mathbf{\Sigma}}_{\boldsymbol{\gamma\gamma}}\widehat{\mathbf{Q}}_{\mathbf{z}}\right)}F_N = \frac{\text{tr}\left(\widehat{\mathbf{\Sigma}}_{\boldsymbol{\gamma\gamma},N}\widehat{\mathbf{Q}}_{\mathbf{z}}\right)}{\text{tr}\left(\widehat{\mathbf{\Sigma}}_{\boldsymbol{\gamma\gamma}}\widehat{\mathbf{Q}}_{\mathbf{z}}\right)}F_N.$$

In cases with homoskedastic errors, F_{Eff} reduces to F_N, while in cases with nonhomoskedastic errors it incorporates a multiplicative correction that depends on the robust variance estimate. Likewise, in the just-identified case, F_{Eff} reduces the F_R, while in the nonhomoskedastic case, it weights $\widehat{\boldsymbol{\gamma}}$ by $\widehat{\mathbf{Q}}_{\mathbf{z}}$ rather than $\widehat{\mathbf{\Sigma}}_{\boldsymbol{\gamma\gamma}}^{-1}$.

It is easy to see that

$$\widehat{\beta}_{\text{2SLS}} = \left(\widehat{\boldsymbol{\gamma}}'\widehat{\mathbf{Q}}_{\mathbf{z}}\widehat{\boldsymbol{\gamma}}\right)^{-1}\widehat{\boldsymbol{\gamma}}'\widehat{\mathbf{Q}}_{\mathbf{z}}\widehat{\boldsymbol{\lambda}},$$

which behaves badly when its denominator, $\widehat{\boldsymbol{\gamma}}'\widehat{\mathbf{Q}}_{\mathbf{z}}\widehat{\boldsymbol{\gamma}}$, is close to zero. The statistic F_N measures the same object, but it gets the standard error wrong and so does not have a noncentral χ^2 distribution; in the nonhomoskedastic case, F_N can be extremely large with high probability even when $\widehat{\boldsymbol{\gamma}}'\widehat{\mathbf{Q}}_{\mathbf{z}}\widehat{\boldsymbol{\gamma}}$ is small. By contrast, the statistic F_R measures the wrong population object, $\boldsymbol{\gamma}'\mathbf{\Sigma}_{\boldsymbol{\gamma\gamma}}^{-1}\boldsymbol{\gamma}$ rather than $\boldsymbol{\gamma}'\mathbf{Q}_{\mathbf{z}}\boldsymbol{\gamma}$, so while it has a noncentral χ^2 distribution, its noncentrality parameter does not correspond to the distribution of $\widehat{\beta}_{\text{2SLS}}$. Finally, F_{Eff} measures the right object and gets the standard errors right on average. More precisely, F_{Eff} is distributed as a weighted average of noncentral χ^2 variables where the weights, given by the eigenvalues of $\widehat{\mathbf{\Sigma}}_{\boldsymbol{\gamma\gamma}}^{1/2}\widehat{\mathbf{Q}}_{\mathbf{z}}\widehat{\mathbf{\Sigma}}_{\boldsymbol{\gamma\gamma}}^{1/2}/\text{tr}\left(\widehat{\mathbf{\Sigma}}_{\boldsymbol{\gamma\gamma}}\widehat{\mathbf{Q}}_{\mathbf{z}}\right)$, are positive and sum to one. Monteil Olea and Pflueger (2013) show that the distribution of F_{Eff} can be approximated by a noncentral χ^2 distribution and formulate tests for weak instruments as defined based on the Nagar (1959) approximation to the bias of two-stage least squares and LIML. Their test rejects when the effective F_{Eff} exceeds a critical value as listed in Table 1 of Monteil Olea and Pflueger (2013). It seems that conventional asymptotic approximations appear reasonable in specifications where F_{Eff} exceeds 10, so 10 is a good rule-of-thumb critical value when $l > 1$.

In conclusion, F_{Eff}, not F_R or F_N, is the preferred statistic for detecting weak instruments in overidentified, nonhomoskedastic settings with one endogenous variable where one uses 2SLS or LIML.[11] When $l = 1$, F_{Eff} ($= F_R$) can be compared to Stock and Yogo (2005) critical values based on t-test size (the mean of the IV estimator does not exist when $l = 1$ so the critical values cannot be based on its bias).

7.7 Valid Inference with Weak IVs

In this subsection, we review four valid inference methods for β with weak IVs. The first three are the AR test of Anderson and Rubin (1949), the LM test of Kleibergen (2002) and Moreira (2009), and the conditional LR (CLR) test of Moreira (2003). The fourth is the tF method of Lee et al.

[11] There is no analog of F_{Eff} when $k_2 > 1$.

(2022).

We follow the notations commonly used in this literature. Suppose

$$\begin{aligned} y_1 &= y_2\beta + X\gamma_1 + u, \\ y_2 &= Z\pi + X\xi + v_2, \end{aligned}$$

where $\beta \in \mathbb{R}$ is the parameter of interest, $y_2 \in \mathbb{R}^n$ is the observations for the only endogenous regressor, $X \in \mathbb{R}^{n\times k_1}$ and $Z \in \mathbb{R}^{n\times l_2}$ include observations for k_2 included IVs and l_2 excluded IVs, respectively, and $u, v_2 \in \mathbb{R}^n$ are unobserved errors. We assume $Z'X = 0$, so that Z is orthogonal to X; otherwise, Z denotes the residual matrix after projecting Z onto X. The reduced-form equations are

$$\begin{aligned} y_1 &= Z\pi\beta + X\gamma + v_1, \\ y_2 &= Z\pi + X\xi + v_2, \end{aligned}$$

where $\gamma = \gamma_1 + \xi\beta$, and $v_1 = u + v_2\beta$, or in matrix form

$$Y = Z\pi a' + X\eta + V,$$

where $Y = (y_1, y_2) \in \mathbb{R}^{n\times 2}$, $V = (v_1, v_2) \in \mathbb{R}^{n\times 2}$, $a = (\beta, 1)'$, and $\eta = (\gamma, \xi)$. Denote the ith row of a matrix by a subscript i. We assume

$$\begin{pmatrix} v_{1i} \\ v_{2i} \end{pmatrix} \sim N(0, \Omega),$$

and we first assume $\Omega = \begin{pmatrix} \omega_{11} & \omega_{12} \\ \omega_{12} & \omega_{22} \end{pmatrix}$ to be known for simplicity. This assumption implies a simple asymptotic variance structure of $n^{-1/2}Z'v_1$ and $n^{-1/2}Z'v_2$; specifically,

$$\lim_{n\to\infty} Var \begin{pmatrix} n^{-1/2}Z'v_1 \\ n^{-1/2}Z'v_2 \end{pmatrix} = \Omega \otimes E\left[Z_i Z_i'\right].$$

Our hypotheses are

$$H_0 : \beta = \beta_0 \text{ vs. } H_1 : \beta \neq \beta_0.$$

7.7.1 AR, LM and CLR

Notice that $Z'Y \in \mathbb{R}^{l_2 \times 2}$ is the sufficient statistic for $(\beta, \pi')'$, so the test can be based on $Z'Y$. The nuisance parameter π remains. To eliminate the dependence on π, decompose $Z'Y$ into

$$\begin{aligned} S &= \left(Z'Z\right)^{-1/2} Z'Y b_0 \cdot \left(b_0'\Omega b_0\right)^{-1/2}, \\ T &= \left(Z'Z\right)^{-1/2} Z'Y\Omega^{-1} a_0 \cdot \left(a_0'\Omega^{-1} a_0\right)^{-1/2}, \\ b_0 &= (1, -\beta_0)', a_0 = (\beta_0, 1)', \end{aligned}$$

where $Z'Yb_0 = Z'(y_1 - y_2\beta_0)$ in S appears also in the AR test statistic, and $a_0'b_0 = 0$. Then

$$\begin{aligned} S &\sim N(c_\beta \mu_\pi, I_{l_2}), \\ T &\sim N(d_\beta \mu_\pi, I_{l_2}), \end{aligned}$$

and S and T are independent, where

$$\begin{aligned} \mu_\pi &= \left(Z'Z\right)^{-1/2} \pi \in \mathbb{R}^{l_2}, \\ c_\beta &= (\beta - \beta_0) \cdot \left(b_0'\Omega b_0\right)^{-1/2} \in \mathbb{R}, \\ d_\beta &= a'\Omega^{-1} a_0 \cdot \left(a_0'\Omega^{-1} a_0\right)^{-1/2} \in \mathbb{R}. \end{aligned}$$

S has a null distribution not dependent on π, and T has a null distribution dependent on π. That is, we decompose the information in $Z'Y$ into two parts, where under the null, the second part T is a sufficient statistic for π, and the first part S is independent of π.

The AR, LM and CLR test statistics are functions of

$$Q = (S, T)'(S, T) = \begin{pmatrix} S'S & S'T \\ T'S & T'T \end{pmatrix} = \begin{pmatrix} Q_S & Q_{ST} \\ Q_{ST} & Q_T \end{pmatrix}.$$

The distribution of Q depends on π only through $\lambda = \pi' Z'Z\pi \geq 0$, where $\lambda^{1/2}$ plays the role of $n^{1/2}$ in the strong IV case. Specifically,

$$\begin{aligned} AR &= \frac{Q_S}{l_2} = \frac{S'S}{l_2}, \\ LM &= \frac{Q_{ST}^2}{Q_T} = S'T\left(T'T\right)^{-1} T'S = S'P_T S, \\ LR &= \frac{1}{2}\left(Q_S - Q_T + \sqrt{\left(Q_S - Q_T\right)^2 + 4Q_{ST}^2}\right), \end{aligned}$$

where P_T is the projection matrix on T. Under the null, $AR \sim \chi_{l_2}^2/l_2$ and $LM \sim \chi_1^2$ are pivotal. On the other hand, LR depends on Q_T under the null, so the corresponding test cannot be similar (i.e., the type-I error rate is invariant to π) if a fixed critical value is used. The CLR test uses critical values that depend on Q_T (which is sufficient for π under the null), i.e., it rejects the null

when

$$LR > c_\alpha\left(Q_T\right),$$

where $c_\alpha\left(Q_T\right)$ satisfies

$$P_{\beta_0}\left(LR > c_\alpha\left(Q_T\right)|Q_T = q_T\right) = \alpha$$

for any q_T, and note that the conditional distribution of LR given Q_T depends only on the conditional distribution of $Q_1 = \left(Q_S, Q_{ST}\right)'$ given Q_T which does not depend on π under the null. It turns out that $c_\alpha\left(q_T\right) \to \chi^2_{1,\alpha}$ (the critical value of LM) as $q_T \to \infty$ (strong IV) and $c_\alpha\left(q_T\right) \to \chi^2_{l_2,\alpha}$ (the critical value of AR times l_2) as $q_T \to 0$ (weak IV).

When Ω is unkown, we can estimate it by

$$\widehat{\Omega} = \frac{1}{n-l}\widehat{V}'\widehat{V},$$

where $\widehat{V} = Y - P_Z Y - P_X Y$.

The AR test is inefficient under strong identification in over-identified models, so Kleibergen (2002) and Moreira (2009) propose the LM test or the K test which is efficient under strong identification. Kleibergen (2005) generalizes this statistic (and also the CLR test) to GMM. As shown in Kleibergen (2005), his K test is a score test based on the CUE objective function.[12] However, the power function of the LM test is not monotone. The LM test fails to reject some nonlocal alternatives. Kleibergen (2002) explains that this is because the LM statistic equals zero at two points, both satisfying the quadratic (in β_0) expression:

$$a_0'\widehat{\Omega}^{-1}Y'P_Z Y b_0 = 0.$$

AMS mention that this is due to the switch of the sign of d_β ad β moves through the value β_{AR}, where $\beta_{\text{AR}} = \left(\omega_{11} - \omega_{12}\beta_0\right)/\left(\omega_{12} - \omega_{22}\beta_0\right)$, provided $\omega_{12} - \omega_{22}\beta_0 \neq 0$, satisfies $d_{\beta_{\text{AR}}} = 0$. For a deeper explanation, see Example I on page 2166 of Andrews (2016).

AMS determine a two-sided power envelope for invariant similar tests, where an invariant test $\phi\left(S,T\right)$ satisfies $\phi\left(S,T\right) = \phi\left(FS,FT\right)$ for all $l_2 \times l_2$ orthogonal matrices F, and F constitutes the group of transformation G on $\left(S,T\right)$. Actually, Q is a maximal invarint for G, which is why all the three test statistics are only functions of Q. They find that the power curve of the CLR test is quite close to this power envelope. This power envelope is also the power envelope under weak IV asymptotics. Under the strong IV asymptoics, the CLR test is asymptotically equivalent (same asymptotic distribution and same asymptotic critical value) to the LM test, so is also asymptotically efficient.

When $l_2 = 1$, i.e., the model is just-identified, $LR = LM = AR$. Actually, the AR test is a uniformly most powerful (UMP) two-sided invariant similar test, so are the LM and CLR tests. Moreira (2009) also shows that these tests are UMP unbiased.

[12] The estimation counterpart of the CLR test is the LIML in the homoskedastic case and the CUE in the heterskedastic case.

7.7.2 tF

The subsection reviews the valid t-ratio inference in Lee et al. (2022). In their setup, $k_2 = l_2 = 1$ (a common case for applied work), so we consider the simple model (12) in this subsection (with included IVs being allowed). The motivation is that practitioners often use the rule-of-thumb F-statistic threshold of 10 in the first stage and if $F > 10$ then claim the interval $\widehat{\beta} \pm 1.96 \cdot s\left(\widehat{\beta}\right)$ as the 95% CI (which is actually only a 80% CI). Rather than relying on a fixed pretesting threshold value, they show how to smoothly adjust 2SLS t-ratio inference based on the first-stage F-statistic to achieve the target coverage (not the reduced coverage like 80%), where the 2SLS instead of the LIML is employed because practitioners seem prefer the former. In its simplest form, this amounts to applying an adjustment factor to 2SLS standard errors based on the first-stage F with the adjustment factors provided in Table 3. They refer to this procedure as the tF procedure, and this procedure is robust to nonhomoskedastic errors.

Stock and Yogo's procedure implies a "step function" critical value function in

$$P\left(t^2 > c^*, F > F^*\right) \leq \alpha,$$

where we understand t and F as the limit random variables of t_n and the F statistic above; specifically, if $F < F^*$, set $c^* = \infty$ (accept the null), and otherwise, use the value c^* as the critical value for t_n^2. For example, setting $F^* = 16.4$ and $c^* = 1.96^2$ achieves $\alpha = 15\%$. Equivalently, this implies a CI procedure that sets the CI to the entire real line if $F < F^*$, and otherwise uses $\pm\sqrt{c^*} \cdot s\left(\widehat{\beta}\right)$ for the CI. As a "smooth" alternative, the tF critical value function $c_\alpha(F)$ satisfies

$$P\left(t^2 > c_\alpha(F)\right) \leq \alpha$$

for a prespecified significance level α, say, 5%, where $c_\alpha(F)$ is a smooth function of F.

Table 3: Selected values of tF critical values, $\sqrt{c_{0.05}(F)}$, and tF s.e. adjustments, $\sqrt{c_{0.05}(F)}/1.96$

Table 3 reports $c_{0.05}(F)$ as a function of F. $c_{0.05}(F)$ tends to infinity as F tends to 1.96^2 from above, and it is strictly decreasing in F until reaching a minimum, 1.96^2, when F reaches around 104.7. The unreported values of $c_{0.05}(F)$ can be calculated by linear interpolation, which is conservative as $c_{0.05}(F)$ is a convex function of F. In practice, the 95% CI can be constructed as $\widehat{\beta} \pm 1.96 \cdot$"0.05 tF s.e.", where the 0.05 tF s.e. is $s\left(\widehat{\beta}\right) \cdot \sqrt{c_{0.05}(F)}/1.96$, and the adjustment factor $\sqrt{c_{0.05}(F)}/1.96$ is reported in the third row of Table 3. For a different α value, $c_\alpha(F)$ is a different function of F. For example, $c_{0.01}(F)$ ends at 2.726^2 rather than the chi-square critical value 2.576^2.

The tF inference has significant power advantages over inference using constant thresholds c^* and F^*. Also, the tF and Anderson-Rubin (AR) tests have similar power, but neither uniformly dominates the other.[13] To compare them, we consider the expected length of these two CIs conditional on $F > 1.96^2$, where we condition on the event $F > 1.96^2$ because when $F \leq 1.96^2$ the

[13] Although AR has known power optimality among unbiased tests, tF is not unbiased.

F	4.000	4.008	4.015	4.023	4.031	4.040	4.049	4.059	4.068	4.079
$\sqrt{c_{0.05}(F)}$	18.656	18.236	17.826	17.425	17.033	16.649	16.275	15.909	15.551	15.201
$\sqrt{c_{0.05}(F)}/1.96$	9.519	9.305	9.095	8.891	8.691	8.495	8.304	8.117	7.934	7.756
	4.090	4.101	4.113	4.125	4.138	4.151	4.166	4.180	4.196	4.212
	14.859	14.524	14.197	13.878	13.566	13.260	12.962	12.670	12.385	12.107
	7.581	7.411	7.244	7.081	6.922	6.766	6.614	6.465	6.319	6.177
	4.229	4.247	4.265	4.285	4.305	4.326	4.349	4.372	4.396	4.422
	11.834	11.568	11.308	11.053	10.804	10.561	10.324	10.091	9.864	9.642
	6.038	5.902	5.770	5.640	5.513	5.389	5.268	5.149	5.033	4.920
	4.449	4.477	4.507	4.538	4.570	4.604	4.640	4.678	4.717	4.759
	9.425	9.213	9.006	8.803	8.605	8.412	8.222	8.037	7.856	7.680
	4.809	4.701	4.595	4.492	4.391	4.292	4.195	4.101	4.009	3.919
	4.803	4.849	4.897	4.948	5.002	5.059	5.119	5.182	5.248	5.319
	7.507	7.338	7.173	7.011	6.854	6.699	6.549	6.401	6.257	6.117
	3.830	3.744	3.660	3.578	3.497	3.418	3.341	3.266	3.193	3.121
	5.393	5.472	5.556	5.644	5.738	5.838	5.944	6.056	6.176	6.304
	5.979	5.844	5.713	5.584	5.459	5.336	5.216	5.098	4.984	4.872
	3.051	2.982	2.915	2.849	2.785	2.723	2.661	2.602	2.543	2.486
	6.440	6.585	6.741	6.907	7.085	7.276	7.482	7.702	7.940	8.196
	4.762	4.655	4.550	4.448	4.348	4.250	4.154	4.061	3.969	3.880
	2.430	2.375	2.322	2.270	2.218	2.169	2.120	2.072	2.025	1.980
	8.473	8.773	9.098	9.451	9.835	10.253	10.711	11.214	11.766	12.374
	3.793	3.707	3.624	3.542	3.463	3.385	3.309	3.234	3.161	3.090
	1.935	1.892	1.849	1.808	1.767	1.727	1.688	1.650	1.613	1.577
	13.048	13.796	14.631	15.566	16.618	17.810	19.167	20.721	22.516	24.605
	3.021	2.953	2.886	2.821	2.758	2.696	2.635	2.576	2.518	2.461
	1.542	1.507	1.473	1.440	1.407	1.376	1.345	1.315	1.285	1.256
	27.058	29.967	33.457	37.699	42.930	49.495	57.902	68.930	83.823	104.67
	2.406	2.352	2.299	2.247	2.197	2.147	2.099	2.052	2.006	1.96
	1.228	1.200	1.173	1.147	1.121	1.096	1.071	1.047	1.024	1.00

Notes: The top number in each of the ten rows is the first-stage F-statistic, the middle number is the corresponding critical value, $\sqrt{c_{0.05}(F)}$, and the bottom number in each row is the corresponding value of $\sqrt{c_{0.05}(F)}/1.96$, where we write 1.96 as a shorthand for $\Phi^{-1}(0.975)$. Numerical values in each pair are rounded up (e.g., 4.0051 rounds up to 4.006).

expected length of both CIs is infinite (see Dufour (1997)). It turns out that the conditional expected length of the AR CI is infinite, while that of the tF CI is finite. In other words, conditional on the event that they produce bounded intervals, the expected length of the tF CI will always be shorter than that of the AR CI.

(**) We provide more details on the results above. We first connect t_n in (25) with the AR test statistic. For a hypothesized value β_0,

$$t_n = \frac{\widehat{\beta} - \beta_0}{s\left(\widehat{\beta}\right)}$$

with $\widehat{\beta}$ being the IV estimator, and the first-stage t test statistic is

$$f_n = \frac{\widehat{\pi}}{s\left(\widehat{\pi}\right)},$$

and $F = f_n^2$. Recall that the AR test is to test whether $\pi\left(\beta - \beta_0\right) = 0$, so the corresponding t test statistic is

$$t_n^{AR}\left(\beta_0\right) = \frac{\widehat{\pi}\left(\widehat{\beta} - \beta_0\right)}{s\left(\widehat{\pi}\left(\widehat{\beta} - \beta_0\right)\right)}.$$

It turns out that in this special case, the AR F-statistic $F\left(\beta_0\right) = t_n^{AR}\left(\beta_0\right)^2$. It can be shown that

$$t_n^2 = \frac{t_n^{AR}\left(\beta_0\right)^2}{1 - 2\widehat{\rho}\left(\beta_0\right)\frac{t_n^{AR}\left(\beta_0\right)}{f_n} + \frac{t_n^{AR}\left(\beta_0\right)^2}{f_n^2}},$$

where $\widehat{\rho}\left(\beta_0\right)$ is the sample analog of $\rho\left(\beta_0\right) = Corr\left(ze^*, zv\right)$ with e^* being the AR error (i.e., the error in regressing $y - x\beta_0$ on z, which is equal to u under the null). Under the assumption that $\pi = n^{-1/2}\mu$,

$$t_n^2 \xrightarrow{d} t^2 = t^2\left(t_{AR}\left(\beta_0\right), f, \rho\left(\beta_0\right)\right) := \frac{t_{AR}\left(\beta_0\right)^2}{1 - 2\rho\left(\beta_0\right)\frac{t_{AR}\left(\beta_0\right)}{f} + \frac{t_{AR}\left(\beta_0\right)^2}{f^2}},$$

where

$$\begin{pmatrix} t_{AR}\left(\beta_0\right) \\ f \end{pmatrix} \sim N\left(\begin{pmatrix} f_0\frac{\Delta(\beta_0)}{\sqrt{1+2\rho\cdot\Delta(\beta_0)+\Delta(\beta_0)^2}} \\ f_0 \end{pmatrix}, \begin{pmatrix} 1 & \rho\left(\beta_0\right) \\ \rho\left(\beta_0\right) & 1 \end{pmatrix}\right)$$

with $f_0 = \pi/\sqrt{AVar\left(\widehat{\pi}\right)}$, $\Delta\left(\beta_0\right) = \frac{\sqrt{Var(zv)}}{\sqrt{Var(zu)}}\left(\beta - \beta_0\right)$, $\rho = Corr\left(zu, zv\right)$ (which is equal to $Corr\left(u, v\right)$ in the homoskedastic case), and $\rho\left(\beta_0\right) = \frac{\rho+\Delta(\beta_0)}{\sqrt{1+2\rho\cdot\Delta(\beta_0)+\Delta(\beta_0)^2}}$. As a result, for the critical value function $\kappa\left(F\right)$, the rejection probability

$$P_{\Delta(\beta_0),\rho,f_0}\left(t^2 > k\left(F\right)\right) = \int\int 1\left(t^2\left(x, y, \rho\left(\beta_0\right)\right) > \kappa\left(y^2\right)\right)\varphi\left(x - f_0\frac{\Delta\left(\beta_0\right)}{\sqrt{1 + 2\rho\cdot\Delta\left(\beta_0\right) + \Delta\left(\beta_0\right)^2}}, y - f_0; \rho\left(\beta_0\right)\right)dxdy, \tag{30}$$

where $\varphi(\cdot,\cdot;r)$ is the bivariate normal density with means zero, unit variances, and correlation r.

To obtain the tF critical value function $c_\alpha(F)$, we need to find a decreasing function $c_\alpha(\cdot)$ for $F \in (q_{1-\alpha}, \overline{F}]$, followed by a flat function beyond $\overline{F}$ for some $\overline{F}$ (e.g., 104.67 in Table 3), such that

$$P_{\Delta(\beta_0)=0,\rho,f_0}\left(t^2 > c_\alpha(F)\right) \leq \alpha$$

for all ρ and $f_0 \neq 0$, where $q_{1-\alpha}$ is the $(1-\alpha)$th quantile of χ^2_1. Because $\rho = \pm 1$ is the least favorable case, we need only to obtain a function $c_\alpha(F)$ such that

$$P_{\Delta(\beta_0)=0,|\rho|=1,f_0}\left(t^2 > c_\alpha(F)\right) = \alpha$$

for some set of small values of f_0. For simplicity, take $\rho = 1$ and $f_0 > 0$ as an example, and $\rho = -1$ and/or $f_0 < 0$ can be symmetrically addressed. When $\rho = 1$, $t_{AR}(\beta_0)$ is a linear function f which follows $N(f_0, 1)$, so

$$t^2 = \frac{f^2(f-f_0)^2}{f_0^2},$$

which is a quartic function, uniquely indexed by f_0. This quartic function has the shape of a "W", with one trough located at $f = 0$, the other trough at $f = f_0$, and an interior peak at $f = f_0/2$. Furthermore, the magnitude of the location and height of the interior peak of the "W" function is monotonically increasing in $|f_0|$. This greatly simplifies the expression of the null rejection probability for any critical value function; the null rejection probability is the probability that f takes on a value for which the quartic t^2 curve is above the critical value function. For any continuous and decreasing (in f^2) critical value function (that eventually plateaus), there exists an interval of values of f_0 for which the "W" curve and the critical value function intersect only twice. The acceptance probability is then simply $\Phi\left(\overline{f}(f_0) - f_0\right) - \Phi\left(\underline{f}(f_0) - f_0\right)$, where the intersections between the two curves are denoted by $\overline{f}(f_0)$ and $\underline{f}(f_0)$; see Figure 7 of Lee et al. for an illustration. As a result, the function $c_\alpha(F)$ satisfies the following system of equations:

$$\begin{aligned}
\frac{\overline{f}(f_0)^2\left(\overline{f}(f_0)-f_0\right)^2}{f_0^2} &= c_\alpha\left(\overline{f}(f_0)\right), \\
\Phi\left(\overline{f}(f_0)-f_0\right) - \Phi\left(\underline{f}(f_0)-f_0\right) &= 1-\alpha, \\
\frac{\underline{f}(f_0)^2\left(\underline{f}(f_0)-f_0\right)^2}{f_0^2} &= c_\alpha\left(\underline{f}(f_0)\right)
\end{aligned}$$

for a set of small values of f_0. They prove the existence of the $c_\alpha(\cdot)$ function, and develop an interative algorithm to solve the system of equations. The remaining question is the determination of $\overline{F}$; they suggest the lowest possible plateau because a lower plateau will lead to a more powerful test. To determine $\overline{F}$ and verify size control for all ρ and f_0 values (the construction of $c_\alpha(F)$ above considers only small values of f_0), they employ numerical integration of (30) to compute these rejection probabilities.

To understand why

$$E\left[L_{AR}|F > q_{1-\alpha}\right] = \infty \text{ and } E\left[L_{tF}|F > q_{1-\alpha}\right] < \infty,$$

they show that conditional on $F > q_{1-\alpha}$,

$$L_{AR} = \frac{\sqrt{F}\sqrt{F - q_{1-\alpha}\left(1 - \widetilde{\rho}^2\right)}}{F - q_{1-\alpha}} L_{IV}, \text{ and } L_{tF} = \sqrt{\frac{c_\alpha\left(F\right)}{q_{1-\alpha}}} L_{IV}$$

are both inflated versions of L_{IV}, where L_{AR} is the limit random variable of CI length based on the AR test, L_{IV} and L_{tF} are similarly defined, and

$$\widetilde{\rho}^2 = \frac{\left(-t_{AR}\left(\beta\right) + \rho f\right)^2}{\left(f^2 - 2\rho t_{AR}\left(\beta\right) f + t_{AR}\left(\beta\right)^2\right)}.$$

It turns out that the L_{AR} inflation factor explodes as F approaches $q_{1-\alpha}$ from above, and even accounting for the other parts of the inflation factor, the denominator $F - q_{1-\alpha}$ leads to an infinite conditional expected length. As for L_{tF}, the inflation factor does not grow as quickly as F approaches $q_{1-\alpha}$ from above, and in particular grows slowly enough that conditional expected length is finite. The key to this result is

$$\lim_{F \downarrow q_{1-\alpha}} c_\alpha\left(F\right)\left(F - q_{1-\alpha}\right) = q_{1-\alpha}^3$$

such that

$$\sqrt{c_\alpha\left(F\right)} = \sqrt{\frac{c_\alpha\left(F\right)\left(F - q_{1-\alpha}\right)}{F - q_{1-\alpha}}} \leq \frac{M}{\sqrt{F - q_{1-\alpha}}}$$

and in a neighborhood of $q_{1-\alpha}$, $1/\sqrt{F - q_{1-\alpha}}$ is integrable (although $1/\left(F - q_{1-\alpha}\right)$ in L_{AR} is not). (**)

7.8 Many Weak Instruments

We still focus on the case with only one endogenous regressor. In the homoskedastic case, Chao and Swanson (2005) show that LIML and B2SLS are consistent when $r_n/\sqrt{K_n} \to \infty$, while 2SLS is consistent only $r_n/K_n \to \infty$, where r_n is the rate of growth of concentration parameter $\pi' Z' Z \pi / \sigma_v^2$, and K_n is the number of instruments. Hansen, Hausman, and Newey (2008) argue that the 2SLS estimator should not be used in applications with many instruments as it becomes very biased. They also argue that a low first-stage F statistic is not always indicative of a weak identification issue, and t-statistics inferences based on more appropriate estimators other than 2SLS, along with corrected standard errors, such as LIML and Fuller's estimator under Bekker asymptotics with nonnormal errors, may still be reliable.

In a heteroskedastic model, Chao et al. (2012) show that the consistency of 2SLS, B2SLS and LIML require $r_n/K_n \to \infty$, while JIVE remains consistent when $r_n/\sqrt{K_n} \to \infty$. Mikusheva and

Sun (2022) show that the condition of $r_n/\sqrt{K_n} \to \infty$ is also necessary for consistency. They develop a jackknife AR test statistic which is valid under heteroskedasticity and weak identification:

$$AR(\beta_0) = \frac{1}{\sqrt{K_n}\sqrt{\widehat{\Phi}}} \sum_{i=1}^{n} \sum_{j \neq i} P_{ij} u_i(\beta_0) u_j(\beta_0),$$

where P_{ij} is the (i,j) element of P_Z, $u_i(\beta_0) = y_i - x_i\beta_0$, and $\widehat{\Phi}$ is a consistent estimator of the asymptotic variance of $\frac{1}{\sqrt{K_n}} \sum_{i=1}^{n} \sum_{j \neq i} P_{ij} u_i(\beta_0) u_j(\beta_0)$. The new test uses an asymptotic approximation based on a Central Limit Theorem (CLT) for quadratic forms in Chao et al. (2012). It has uniformly correct size and good power properties due to the novel cross-fit variance etimator $\widehat{\Phi}$:

$$\widehat{\Phi} = \frac{2}{K_n} \sum_{i=1}^{n} \sum_{j \neq i} \frac{P_{ij}^2}{M_{ii}M_{jj} + M_{ij}^2} \left[u_i(\beta_0) M_i \mathbf{u}(\beta_0)\right] \left[u_j(\beta_0) M_j \mathbf{u}(\beta_0)\right],$$

where M_{ij} is the (i,j) element of $M_Z = I - P_Z$, M_i is the ith row of M_Z, and $\mathbf{u}(\beta_0)$ is the column vector collecting $u_i(\beta_0)$. Different from the naive estimator of Φ, say $\widehat{\Phi}_1 = \frac{2}{K_n} \sum_{i=1}^{n} \sum_{j \neq i} P_{ij}^2 u_i^2(\beta_0) u_j^2(\beta_0)$, $\widehat{\Phi}$ is consistent under both the null and the alterantive. In the spirit of Stock and Yogo (2005), they develop a pre-test for weak identification, that can help to assess the reliability of the JIVE-Wald test. Specifically, this is a two-step procedure: accept the null $\beta = \beta_0$ if the pre-test F statistic,

$$\widetilde{F} = \frac{1}{\sqrt{K_n}\sqrt{\widehat{\Upsilon}}} \sum_{i=1}^{n} \sum_{j \neq i} P_{ij} x_i x_j,$$

is greater than 4.14 and the JIVE-Wald test statistic

$$W(\beta_0) = \frac{\left(\widehat{\beta}_{\text{JIVE}} - \beta_0\right)^2}{\widehat{V}}$$

is greater than 3.84 or if $\widetilde{F} < 4.14$ and $AR(\beta_0) < 1.96$, where

$$\widehat{\beta}_{\text{JIVE}} = \frac{\sum_{i=1}^{n} \sum_{j \neq i} P_{ij} y_i x_j}{\sum_{i=1}^{n} \sum_{j \neq i} P_{ij} x_i x_j}$$

is $\widehat{\beta}_{\text{JIVE2}}$ in Chapter 7, $\widehat{\Upsilon}$ is a consistent variance etimator of $F = \frac{1}{\sqrt{K_n}} \sum_{i=1}^{n} \sum_{j \neq i} P_{ij} x_i x_j$, and $\widehat{V}$ is a consistent variance estimator of $\widehat{\beta}_{\text{JIVE}}$. Note that $F \xrightarrow{d} N\left(\frac{\mu^2}{\sqrt{K_n}}, \Upsilon\right)$, where $\mu^2 = \pi' Z' Z \pi - \sum_{i=1}^{n} P_{ii} (\pi' z_i)^2 \sim \pi' Z' Z \pi$, i.e., F indeed includes the information of identification in the first stage. This procedure has an asymptotic size smaller than 15%.

Besides JIVE, Hausman et al. (2012) show that heteroscedasticity-robust Fuller is also a good choice under strong identification. For overidentification tests under homoskedasticity, see Anatolyev and Gospodinov (2011), and under heteroskedasticity, see Chao et al. (2014).

8 Alternative Inference Procedures (*)

Monte Carlo studies have shown that estimated asymptotic standard errors of the efficient two-step GMM estimator can be severely downward biased in small samples. A key observation for the source of this bias is that the weight matrix used in the calculation of the efficient two-step GMM estimator is based on initial consistent parameter estimates whose variation is not embodied in the asymptotic covariance matrix estimation. Windemeijer (2005) shows that when the moment conditions used are linear in the parameters, the extra variation due to the presence of these estimated parameters in the weight matrix accounts for much of the difference between the finite sample and the usual asymptotic variance of the two-step GMM estimator.

To this problem, there are a few reactions in the literature. First, nonlinear procedures, especially the generalized empirical likelihood (GEL) estimation, are proposed. Inferences based on these nonlinear procedures are more accurate because they circumvent the estimation of the optimal weight matrix.[14] Second, linear procedures are proposed to incorporate the variation in the first-stage estimator explicitly. Third, bootstrap procedures are put forward to refine the inferences based on the two-step GMM estimator.

In this section, we will concentrate on two GEL estimators, the continuously-updated estimator and the empirical likelihood estimator; we will also briefly discuss the inferences based on the "general" GEL approach.

8.1 Continuously-Updated Estimator

There is an important alternative to the two-step GMM estimator. Specifically, we can let the weight matrix be considered as a function of $\boldsymbol{\theta}$. The criterion function is then

$$J_n(\boldsymbol{\theta}) = n \cdot \overline{g}_n(\boldsymbol{\theta})' \left(\frac{1}{n} \sum_{i=1}^{n} g_i^*(\boldsymbol{\theta}) g_i^*(\boldsymbol{\theta})' \right)^{-1} \overline{g}_n(\boldsymbol{\theta}),$$

where

$$g_i^*(\boldsymbol{\theta}) = g_i(\boldsymbol{\theta}) - \overline{g}_n(\boldsymbol{\theta}).$$

The $\widehat{\boldsymbol{\theta}}$ which minimizes this function is called the **continuously-updated estimator** (CUE) of GMM, and was introduced by Hansen et al. (1996). An advantage of this estimator relative to the two-step estimator is that it is *invariant* to how the moment conditions are scaled even when parameter-dependent scale factors are introduced. This estimator appears to have some better properties (e.g., smaller bias) than traditional GMM, but can be numerically tricky to obtain in some cases, e.g., its objetive function may possess multiple local minima and may produce extreme estimates (i.e., its distribution has a fat tail); see Section 4 of Hansen et al. (1996) for more details. Donald and Newey (2000) interpret the CUE as a jackknife estimator to explain why the CUE is less biased. Essentially, the CUE makes the estimator of Jacobian $\mathbf{G}$ asymptotically uncorrelated

[14] Actually, the bias is also smaller.

with $\overline{g}_n\left(\widehat{\boldsymbol{\theta}}\right)$, which eliminates an important source of nonzero expectations for the FOCs, and hence of bias. Another way to look at the CUE is that it is the heteroskedasticity-robust version of the LIML estimator.

Exercise 16 *(i) Write out the objective function of the CUE in the linear homoskedastic endogenous model. (ii) Show that this CUE is equivalent to the LIML estimator. (iii) Show that if $g_i^*(\boldsymbol{\theta})$ is replaced by $g_i(\boldsymbol{\theta})$ in $J_n(\boldsymbol{\theta})$, then the new objective function $\widetilde{J}_n(\boldsymbol{\theta}) = J_n(\boldsymbol{\theta})/(1 + J_n(\boldsymbol{\theta}))$.*

8.2 EL Estimator

The idea of **empirical likelihood** (EL) is due to Owen (1988, 1990); see also Qin and Lawless (1994) and Imbens (1997) for its GMM extension. It is a non-parametric analog of likelihood estimation.

The idea is to construct a multinomial distribution $F(p_1, \cdots, p_n)$ which places probability p_i at each observation. To be a valid multinomial distribution, these probabilities must satisfy the requirements that $p_i \geq 0$ and

$$\sum_{i=1}^{n} p_i = 1. \tag{31}$$

Since each observation is observed once in the sample, the log-likelihood function for this multinomial distribution is

$$\log L(p_1, \cdots, p_n) = \sum_{i=1}^{n} \log(p_i). \tag{32}$$

First let us consider a just-identified model. In this case the moment condition places no additional restrictions on the multinomial distribution. The maximum likelihood estimators of the probabilities $(p_1, \cdots, p_n)$ are those which maximize the log-likelihood subject to the constraint (31). This is equivalent to maximizing

$$\sum_{i=1}^{n} \log(p_i) - \mu\left(\sum_{i=1}^{n} p_i - 1\right),$$

where μ is a Lagrange multiplier. The n FOCs are $0 = p_i^{-1} - \mu$. Combined with the constraint (31) we find that the MLE is $p_i = n^{-1}$ yielding the log-likelihood $-n\log(n)$.

Now consider the case of an over-identified model with moment condition $E[g(\mathbf{w}_i, \boldsymbol{\theta}_0)] \equiv E[g_i(\boldsymbol{\theta}_0)] = \mathbf{0}$. The multinomial distribution which places probability p_i at each observation $\mathbf{w}_i$ will satisfy this condition if and only if

$$\sum_{i=1}^{n} p_i g_i(\boldsymbol{\theta}) = \mathbf{0}. \tag{33}$$

The EL estimator is the value of $\boldsymbol{\theta}$ which maximizes the multinomial log-likelihood (32) subject to the restrictions (31) and (33).

The Lagrangian for this maximization problem is

$$\mathcal{L}(\boldsymbol{\theta}, p_1, \cdots, p_n, \boldsymbol{\lambda}, \mu) = \sum_{i=1}^{n} \log(p_i) - \mu\left(\sum_{i=1}^{n} p_i - 1\right) - n\boldsymbol{\lambda}' \sum_{i=1}^{n} p_i g_i(\boldsymbol{\theta}),$$

where $\boldsymbol{\lambda}$ and μ are Lagrange multipliers. The FOCs of $\mathcal{L}$ with respect to p_i, μ and $\boldsymbol{\lambda}$ are

$$\frac{1}{p_i} = \mu + n\boldsymbol{\lambda}' g_i(\boldsymbol{\theta}),\ \sum_{i=1}^{n} p_i = 1 \text{ and } \sum_{i=1}^{n} p_i g_i(\boldsymbol{\theta}) = 0.$$

Multiplying the first equation by p_i, summing over i and using the second and third equations, we find $\mu = n$ and

$$p_i = \frac{1}{n\left(1 + \boldsymbol{\lambda}' g_i(\boldsymbol{\theta})\right)}.$$

Substituting into $\mathcal{L}$ we find

$$R(\boldsymbol{\theta}, \boldsymbol{\lambda}) = -n\log(n) - \sum_{i=1}^{n} \log\left(1 + \boldsymbol{\lambda}' g_i(\boldsymbol{\theta})\right).$$

For given $\boldsymbol{\theta}$ the Lagrange multiplier $\boldsymbol{\lambda}(\boldsymbol{\theta})$ minimizes $R(\boldsymbol{\theta}, \boldsymbol{\lambda})$:

$$\boldsymbol{\lambda}(\boldsymbol{\theta}) = \arg\min_{\boldsymbol{\lambda}} R(\boldsymbol{\theta}, \boldsymbol{\lambda}).$$

This minimization problem is the dual of the constrained maximization problem. The solution (when it exists) is well defined since $R(\boldsymbol{\theta}, \boldsymbol{\lambda})$ is a convex function of $\boldsymbol{\lambda}$. The solution cannot be obtained explicitly, but must be obtained numerically. This yields the (profile) empirical log-likelihood function for $\boldsymbol{\theta}$:

$$R(\boldsymbol{\theta}) = R(\boldsymbol{\theta}, \boldsymbol{\lambda}(\boldsymbol{\theta})) = -n\log(n) - \sum_{i=1}^{n} \log\left(1 + \boldsymbol{\lambda}(\boldsymbol{\theta})' g_i(\boldsymbol{\theta})\right).$$

The EL estimate $\widehat{\boldsymbol{\theta}}$ is the value which maximizes $R(\boldsymbol{\theta})$, or equivalently minimizes its negative

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \left[-R(\boldsymbol{\theta})\right].$$

Numerical methods are required for calculation of $\widehat{\boldsymbol{\theta}}$.

As a by-product of estimation, we also obtain the Lagrange multiplier $\widehat{\boldsymbol{\lambda}} = \boldsymbol{\lambda}\left(\widehat{\boldsymbol{\theta}}\right)$, probabilities

$$\widehat{p}_i = \frac{1}{n\left(1 + \widehat{\boldsymbol{\lambda}}' g_i(\widehat{\boldsymbol{\theta}})\right)},$$

and maximized empirical likelihood

$$R\left(\widehat{\boldsymbol{\theta}}\right)=\sum_{i=1}^{n}\widehat{p}_i. \tag{34}$$

In Appendix B, we show that $\widehat{\boldsymbol{\theta}}$ has the same asymptotic distribution as the usual GMM estimator in Section 4, so the EL estimator is asymptotically efficient (without estimating the optimal weight matrix). Also,

$$\sqrt{n}\widehat{\boldsymbol{\lambda}}\xrightarrow{d}\boldsymbol{\Omega}^{-1}N\left(\mathbf{0},\mathbf{V}_{\boldsymbol{\lambda}}\right),$$

where $\mathbf{V}_{\boldsymbol{\lambda}}=\boldsymbol{\Omega}-\mathbf{G}\left(\mathbf{G}'\boldsymbol{\Omega}^{-1}\mathbf{G}\right)^{-1}\mathbf{G}'$. Furthermore, $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\lambda}}$ are asymptotically independent.

In a parametric likelihood context, tests are based on the difference in the log likelihood functions. The same statistic can be constructed for empirical likelihood. In this case, the unrestricted and restricted empirical log-likelihoods are

$$\begin{aligned}\mathcal{L}_n^{\mathrm{r}} &= \sum_{i=1}^{n}\ln(\widehat{\pi}_i)=-n\log n-\sum_{i=1}^{n}\log\left(1+\widehat{\boldsymbol{\lambda}}'g_i(\widehat{\boldsymbol{\theta}})\right),\\ \mathcal{L}_n^{\mathrm{ur}} &= -n\log n,\end{aligned}$$

so twice the difference between these two log-likelihoods is

$$LR_n=2\left(\mathcal{L}_n^{\mathrm{ur}}-\mathcal{L}_n^{\mathrm{r}}\right)=2\sum_{i=1}^{n}\log\left(1+\widehat{\boldsymbol{\lambda}}'g_i(\widehat{\boldsymbol{\theta}})\right).$$

Under the null (8), $LR_n\xrightarrow{d}\chi^2_{l-k}$, which is shown in Appendix B. Kitamura (2001) shows that this empirical likelihood ratio test of the overidentifying restrictions satisfies the optimal criterion of Hoeffding (1965). The EL overidentification test is similar to the GMM overidentification test. They are asymptotically first-order equivalent, and have the same interpretation. The overidentification test is a very useful by-product of EL estimation, and it is advisable to report the statistic LR_n whenever EL is the estimation method.

8.3 GEL Estimators

Both the CUE and the EL estimator are special cases of the GEL estimator of Smith (1997); see Newey and Smith (2004) and Smith (2011) for further discussions on the GEL estimator. To describe GEL let the **carrier function** $\rho(v)$ be a function of a scalar v that is concave on its domain, an open interval $\mathcal{V}$ containing 0. Let $\widehat{\Lambda}_n(\boldsymbol{\theta})=\left\{\boldsymbol{\lambda}|\boldsymbol{\lambda}'g_i(\boldsymbol{\theta})\in\mathcal{V},\ i=1,\cdots,n\right\}$. The estimator is the solution to a saddle point problem

$$\widehat{\boldsymbol{\theta}}_{GEL}=\arg\min_{\boldsymbol{\theta}\in\Theta}\sup_{\boldsymbol{\lambda}\in\widehat{\Lambda}_n(\boldsymbol{\theta})}\sum_{i=1}^{n}\rho\left(\boldsymbol{\lambda}'g_i(\boldsymbol{\theta})\right)=\arg\min_{\boldsymbol{\theta}\in\Theta}R(\boldsymbol{\theta}), \tag{35}$$

where Θ denotes the parameter space. The EL estimator is a special case with $\rho(v)=\log(1-v)$ and $\mathcal{V}=(-\infty,1)$. The **exponential tilting** (ET) estimator of Kitamura and Stutzer (1997) and

Imbens, Spady and Johnson (1998) is a special case with $\rho(v) = -e^v$ which has the computational advantage, relative to EL, of having an unrestricted domain, although the restricted domain of EL is usually not a problem in practice.[15] The CUE is a special case with $\rho(v) = -(1+v)^2/2$.[16] Associated with each GEL estimator are empirical probabilities for observations. Specifically,

$$\widehat{\pi}_i = \rho_1(\widehat{\boldsymbol{\lambda}}'\widehat{g}_i)/\sum_{j=1}^n \rho_1(\widehat{\boldsymbol{\lambda}}'\widehat{g}_j),$$

where ρ_1 is the first derivative of ρ, $\widehat{g}_i = g_i(\widehat{\boldsymbol{\theta}})$, and

$$\widehat{\boldsymbol{\lambda}} = \arg \max_{\boldsymbol{\lambda} \in \widehat{\Lambda}_n(\widehat{\boldsymbol{\theta}})} \sum_{i=1}^n \rho\left(\boldsymbol{\lambda}'\widehat{g}_i\right)/n. \tag{36}$$

For EL and ET,

$$\widehat{\pi}_i = \frac{1}{n(1-\widehat{\boldsymbol{\lambda}}'\widehat{g}_i)} \text{ and } \frac{e^{\widehat{\boldsymbol{\lambda}}'\widehat{g}_i}}{\sum_{j=1}^n e^{\widehat{\boldsymbol{\lambda}}'\widehat{g}_j}},$$

respectively. These empirical probabilities $\widehat{\pi}_i$ sum to one by construction, satisfy the sample moment condition $\sum_{i=1}^n \widehat{\pi}_i \widehat{g}_i = \mathbf{0}$ when the FOCs for $\widehat{\boldsymbol{\lambda}}$ hold, and are positive when $\widehat{\boldsymbol{\lambda}}'\widehat{g}_i$ is small uniformly in i.

Another formulation of these estimators is through the minimum discrepancy (MD) estimator of Corcoran (1998). The MD estimator is defined as

$$\begin{aligned} \overline{\boldsymbol{\theta}} &= \arg \min_{\boldsymbol{\theta} \in \Theta,\, \pi} \sum_{i=1}^n h(\pi_i) \\ \text{s.t. } \sum_{i=1}^n \pi_i g_i(\boldsymbol{\theta}) &= \mathbf{0},\ \sum_{i=1}^n \pi_i = 1, \end{aligned} \tag{37}$$

where $\pi = (\pi_1, \cdots, \pi_n)$. When $h(\pi) = -\ln(\pi)$, $\pi \ln(\pi)$ (the Kullback-Leibler information criterion), and π^2 we get the EL, ET and CUE, respectively. For each MD estimator there is a dual GEL estimator when $h(\pi)$ is a member of the Cressie and Read (1984) family of discrepancies in which $h(\pi) = [\gamma(\gamma+1)]^{-1}\left[(n\pi)^{\gamma+1} - 1\right]/n$.[17] The corresponding $\rho(v) = -(1+\gamma v)^{(\gamma+1)/\gamma}/(\gamma+1)$. When $\gamma = -1, 0$ and 1, we get the h functions for the EL, ET and CUE, respectively. $\boldsymbol{\lambda}$ in (35) is proportional to the Lagrange multipliers for the first (moment) constraint in (37). The maximization problem for $\boldsymbol{\lambda}$ in (36) is considerably easier than the MD problem, having much smaller dimension (l versus n) and being a simple concave programming problem. For $\gamma = -1$ and 0, there

[15] Another GEL estimator is the minimum Hellinger distance estimator (MHDE) of Kitamura et al. (2013).

[16] Note that for EL, ET and CUE, their ρ satisfies $\rho_1(0) = \rho_2(0) = -1$, which can be a general normalization for any ρ, where ρ_j is the jth derivative of ρ. Also, to normalize ρ such that $\rho(0) = 0$, sometimes let $\rho(v) = 1 - e^v$ in ET and $\rho(v) = 1/2 - (1+v)^2/2$ in CUE.

[17] For two discrete distributions with common support $p = (p_1, \cdots, p_n)$ and $q = (q_1, \cdots, q_n)$, the Cressie-Read power-divergence statistic is defined as $I_\gamma(p,q) = \frac{1}{\gamma(\gamma+1)} \sum_{i=1}^n p_i \left[\left(\frac{p_i}{q_i}\right)^{\gamma+1} - 1\right]$. $\sum_{i=1}^n h(\pi_i)$ measures the distance between π and the empirical distribution π_{unif}. $\gamma = 0$ defines the Kullback-Leibler distance from π_{unif} to π; $\gamma = -1$ defines the Kullback-Leibler distance from π to π_{unif}; $\gamma = -1/2$ defines the Hellinger distance between π and π_{unif}.

are no explicit solutions for $\widehat{\boldsymbol{\lambda}}$ and $\widehat{\pi}$. When $\gamma = 1$, explicit formulae for $\widehat{\boldsymbol{\lambda}}$ and $\widehat{\pi}$ are possible, but they usually involve one or more of the $\widehat{\pi}_i$'s being negative and so give rise to problems of interpretation.

Although the first-order properties of all GEL estimators are the same as the efficient GMM estimator, their higher-order properties are quite different. For example, Newey and Smith (2004) show that EL generally exhibits a smaller $O\left(n^{-1}\right)$ bias than any other member of GEL unless the centered third moments of the distribution of $g_i \equiv g_i(\boldsymbol{\theta}_0)$ happen to all vanish, in which case all GEL estimators have the same $O\left(n^{-1}\right)$ bias. The intuition for this result is that the GEL FOCs are

$$\left[\sum_{i=1}^{n} \widehat{\pi}_i \mathbf{G}_i\left(\widehat{\boldsymbol{\theta}}\right)\right]\left[\sum_{i=1}^{n} \widehat{k}_i g_i\left(\widehat{\boldsymbol{\theta}}\right) g_i\left(\widehat{\boldsymbol{\theta}}\right)'\right]^{-1} \overline{g}_n\left(\widehat{\boldsymbol{\theta}}\right) = \mathbf{0}$$

where $\mathbf{G}_i(\boldsymbol{\theta}) = \partial g_i(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}'$, $\widehat{k}_i = k(\widehat{v}_i) / \sum_{j=1}^{n} k(\widehat{v}_j)$ with $k(v) = [\rho_1(v) + 1] / v$ when $v \neq 0$ and $k(0) = -1$, and $\widehat{v}_i = \widehat{\boldsymbol{\lambda}}' \widehat{g}_i$. For EL, $\widehat{k}_i = \widehat{\pi}_i$ while for CUE, $\widehat{k}_i = 1/n$. The bias of $\widehat{\boldsymbol{\theta}}$ includes three parts: $B_I = \mathbf{H}\left(-\mathbf{a} + E\left[\mathbf{G}_i \mathbf{H} g_i\right]\right)/n$ is the asymptotic bias from the estimator based on the optimal linear combination $\mathbf{G}'\boldsymbol{\Omega}^{-1} g(\mathbf{w}, \boldsymbol{\theta}) = \mathbf{0}$, $B_{\mathbf{G}} = -\boldsymbol{\Sigma} E\left[\mathbf{G}_i \mathbf{P} g_i\right]/n$ comes from the correlation between the Jacobian $\mathbf{G}$ estimator with $\overline{g}_n\left(\widehat{\boldsymbol{\theta}}\right)$, and $B_{\boldsymbol{\Omega}} = \mathbf{H} E\left[g_i g_i' \mathbf{P} g_i\right]$ comes from the correlation between the outer product matrix $\boldsymbol{\Omega}$ estimator with $\overline{g}_n\left(\widehat{\boldsymbol{\theta}}\right)$, where $\boldsymbol{\Sigma} = \left(\mathbf{G}'\boldsymbol{\Omega}^{-1}\mathbf{G}\right)^{-1}$, $\mathbf{H} = \boldsymbol{\Sigma}\mathbf{G}'\boldsymbol{\Omega}^{-1}$, $\mathbf{P} = \boldsymbol{\Omega}^{-1} - \boldsymbol{\Omega}^{-1}\mathbf{G}\boldsymbol{\Sigma}\mathbf{G}'\boldsymbol{\Omega}^{-1}$, $a_j = \text{tr}\left(\boldsymbol{\Sigma} E\left[\partial^2 g_{ij}(\boldsymbol{\theta}_0) / \partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'\right]\right)/2$ represents the bias from the nonlinearity of $g_{ij}(\boldsymbol{\theta}) = g_j(\mathbf{w}_i, \boldsymbol{\theta})$, and $\mathbf{G}_i = \mathbf{G}_i(\boldsymbol{\theta}_0)$. The CUE effectively eliminates $B_{\mathbf{G}}$ by using the weights $\widehat{\pi}_i$ in the Jacobian estimation but not $B_{\boldsymbol{\Omega}}$ because $\widehat{k}_i = 1/n$, while the EL eliminates both by using "effective" weights $\widehat{\pi}_i$ and $\widehat{k}_i$ (that are different from $1/n$). For a general GEL estimator, the bias is $B_I + (1 + \rho_3/2) B_{\boldsymbol{\Omega}}$. EL can eliminate $B_{\boldsymbol{\Omega}}$ because its $\rho_3 = -2$. If $\rho_3 \neq -2$, but $B_{\boldsymbol{\Omega}} = \mathbf{0}$ (or equivalently, $E\left[g_i g_i' g_{ij}\right] = \mathbf{0}$), then this source of bias can still be eliminated. For example, in the linear IV setting of Section 1, when disturbances are symmetrically distributed, this can happen. They also show that the bias-corrected EL is higher-order efficient, possessing an $O\left(n^{-2}\right)$ variance that is no greater than that of any other bias-corrected method of moments estimators. The proof of this result combines the arguments of Chamberlain (1987), as explained in Section 9, and the third-order efficiency of the parametric MLE (see Rao (1963) and Pfanzagl and Wefelmeyer (1978)).

If we want to test H_0: $\mathbf{r}(\boldsymbol{\theta}) = \mathbf{0}$, we can use the LR statistic

$$LR_n = 2\left[R\left(\widetilde{\boldsymbol{\theta}}\right) - R\left(\widehat{\boldsymbol{\theta}}\right)\right],$$

which follows χ^2_q under the null, where $R(\cdot)$ is defined in (35), and

$$\widetilde{\boldsymbol{\theta}} = \arg \min_{\mathbf{r}(\boldsymbol{\theta}) = \mathbf{0}} R(\boldsymbol{\theta}).$$

As a result, the $(1-\alpha)$ confidence region for $\boldsymbol{\theta}_2$, a k_2-subvector of $\boldsymbol{\theta}$, is

$$2\left[R\left(\widetilde{\boldsymbol{\theta}}_1\left(\boldsymbol{\theta}_2\right),\boldsymbol{\theta}_2\right)-R\left(\widehat{\boldsymbol{\theta}}\right)\right]\leq\chi^2_{k_2,\alpha},$$

where $\widetilde{\boldsymbol{\theta}}_1\left(\boldsymbol{\theta}_2\right)=\arg\min_{\boldsymbol{\theta}_1}R\left(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2\right)$ for a given $\boldsymbol{\theta}_2$. This Wilks's phenomenon is invariant to the choice of $\rho\left(\cdot\right)$ (or equivalently, $h(\cdot)$), but second-order results depend intimately on the choice of ρ. For example, the GEL is Bartlett correctable if and only if $\rho(v)=\log(1-v)$. Also, such a confidence region has the advantages of having natural shape and respects the range of $\boldsymbol{\theta}_2$.

We can also use the LR statistic to test the overidentifying restrictions as in the EL case; all such GEL tests are asymptotically first-order equivalent, but may have different higher-order properties depending on h. Smith (1997) puts forward an alternative overidentifiction test based on ρ rather than h. Specifically, he shows

$$2n\left[\sum_{i=1}^{n}\rho\left(\widehat{\boldsymbol{\lambda}}'g_i(\widehat{\boldsymbol{\theta}})\right)/n-\rho\left(0\right)\right]\xrightarrow{d}\chi^2_{l-k},$$

under the null that the model is correctly specified, where note that the true value of $\widehat{\boldsymbol{\lambda}}$ is zero, which explains the presence of $\rho\left(0\right)$.

8.4 Other Inference Procedures

Windmeijer (2005) proposes a finite-sample correction for the variance of linear efficient two-step GMM estimators. His correction explicitly incorporates the variation in the first-stage estimator. Details are included in Appendix C. Chao and Swanson (2001) derive the bias and MSE of 2SLS under weak-instrument asymptotics, modified to allow the number of instruments to increase with the sample size. They report improvements in Monte Carlo simulations by incorporating bias adjustments. As to the bootstrap inference, there are basically two methods attributed to Hall and Horowitz (1996) and Brown and Newey (2002) respectively; see also Lee (2014). Bond and Windmeijer (2002) report problems with the bootstrap procedures when the weight matrix is a poor estimate of the covariance matrix of the moment conditions, which occurs for example when there are a large number of overidentifying restrictions.

9 Conditional Moment Restrictions

In many cases, the model may imply conditional moment restrictions

$$E[u\left(\mathbf{w},\boldsymbol{\beta}_0\right)|\mathbf{x}]=\mathbf{0},$$

where $u\left(\mathbf{w},\boldsymbol{\beta}\right)$ is some $s\times1$ function of the observation and the parameters. For example, in linear regression, $u\left(\mathbf{w},\boldsymbol{\beta}\right)=y-\mathbf{x}'\boldsymbol{\beta}$, $\mathbf{w}=(y,\mathbf{x}')'$, and $s=1$; in a joint model of conditional mean and

variance,

$$u(\mathbf{w}, \boldsymbol{\beta}) = \begin{pmatrix} y - \mathbf{x}'\boldsymbol{\beta}_1 \\ (y - \mathbf{x}'\boldsymbol{\beta}_1)^2 - f(\mathbf{x})'\boldsymbol{\beta}_2 \end{pmatrix}$$

for a specification $Var(y|\mathbf{x}) = f(\mathbf{x})'\boldsymbol{\beta}_2$, where $\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1', \boldsymbol{\beta}_2'\right)'$, so $s = 2$.

Conditional moment restrictions imply infinite unconditional moment conditions, since for any function of $\mathbf{x}$, say $\phi(\mathbf{x})$, $E[\phi(\mathbf{x})u(\mathbf{w}, \boldsymbol{\beta}_0)] = \mathbf{0}$. So a natural question is which instruments are optimal, or what is the semiparametric efficiency (or variance) bound for $\boldsymbol{\beta}_0$. Chamberlain (1987) derived this bound by approximating the CDF $F(\mathbf{x})$ and the conditional CDF $F(\mathbf{w}|\mathbf{x})$ with multinomial distributions; see Appendix D. It turns out that the optimal instruments are

$$\mathbf{A}(\mathbf{x}) = \mathbf{G}(\mathbf{x})'\boldsymbol{\Omega}(\mathbf{x})^{-1},$$

where $\mathbf{G}(\mathbf{x}) = E\left[\partial u(\mathbf{w}, \boldsymbol{\beta}_0)/\partial \boldsymbol{\beta}' \middle| \mathbf{x}\right]$, and $\boldsymbol{\Omega}(\mathbf{x}) = E\left[u(\mathbf{w}, \boldsymbol{\beta}_0)\, u(\mathbf{w}, \boldsymbol{\beta}_0)' \middle| \mathbf{x}\right]$. $\mathbf{A}(\mathbf{x})$ is similar to the optimal linear combination $\mathbf{B}$ in the unconditional moment case, but now we condition every random variable on $\mathbf{x}$. Using the optimal instruments, the unconditional moment conditions are

$$E[m(\mathbf{w}, \boldsymbol{\beta}_0)] = E[\mathbf{A}(\mathbf{x})u(\mathbf{w}, \boldsymbol{\beta}_0)] = \mathbf{0}.$$

Applying the formula of the asymptotic variance for the MoM estimator, we have the semiparametric efficiency bound for $\boldsymbol{\beta}_0$

$$\begin{aligned} & E\left[\mathbf{A}(\mathbf{x})\partial u(\mathbf{w}, \boldsymbol{\beta}_0)/\partial \boldsymbol{\beta}'\right]^{-1} E\left[\mathbf{A}(\mathbf{x})u(\mathbf{w}, \boldsymbol{\beta}_0)\, u(\mathbf{w}, \boldsymbol{\beta}_0)'\, \mathbf{A}(\mathbf{x})'\right] E\left[\mathbf{A}(\mathbf{x})\partial u(\mathbf{w}, \boldsymbol{\beta}_0)/\partial \boldsymbol{\beta}'\right]'^{-1} \\ = \; & E\left[\mathbf{G}(\mathbf{x})'\boldsymbol{\Omega}(\mathbf{x})^{-1}\mathbf{G}(\mathbf{x})\right]^{-1}. \end{aligned}$$

In the linear regression case, $\mathbf{G}(\mathbf{x}) = \mathbf{x}'$, and $\boldsymbol{\Omega}(\mathbf{x}) = \sigma^2(\mathbf{x})$, so the optimal instrument is $\mathbf{x}/\sigma^2(\mathbf{x})$, which corresponds to the generalized least squares estimator, and the semiparametric efficiency bound for $\boldsymbol{\beta}_0$ is $E\left[\mathbf{x}\mathbf{x}'/\sigma^2(\mathbf{x})\right]$.

The optimal instruments involve the conditional mean estimation. This will use nonparametric estimation techniques which are not covered by this course; see Newey (1990b) for such estimations. In practice, we may only want to select a group of instruments that need not be (asymptotically) optimal. But given an infinite list of potential instruments, which should be used? This is essentially a model selection problem, and will be briefly discussed in Section 6.2.

(*) Kitamura et al. (2004) study the empirical likelihood-based inference in conditional moment restrictions models. Andrews and Shi (2013) and Chernozhukov et al. (2013) study inference based on conditional moment inequalities.

Exercise 17 (Empirical) *Continue the empirical exercise in the last chapter.*

(d) *Re-estimate the model by efficient GMM. I suggest that you use the 2SLS estimates as the first-step to get the weight matrix, and then calculate the GMM estimator from this weight matrix without further iteration. Report the estimates and standard errors.*

(e) *Calculate and report the J statistic for overidentification.*

(f) *Discuss your findings.*

Appendix A: Asymptotics for the Nonlinear GMM

The following two theorems show that the GMM estimator is CAN.

Theorem 3 *Suppose that* $\mathbf{w}_i$, $i = 1, 2, \cdots$, *are i.i.d.,* $\mathbf{W}_n \xrightarrow{p} \mathbf{W}$, *and (i)* $\mathbf{W} \geq 0$ *and* $\mathbf{W}E[g(\mathbf{w}, \boldsymbol{\theta})] = \mathbf{0}$ *only if* $\boldsymbol{\theta} = \boldsymbol{\theta}_0$; *(ii)* $\boldsymbol{\theta}_0 \in \Theta$, *which is compact; (iii)* $g(\mathbf{w}, \boldsymbol{\theta})$ *is continuous at each* $\boldsymbol{\theta} \in \Theta$ *with probability one; (iv)* $E[\sup_{\boldsymbol{\theta} \in \Theta} \|g(\mathbf{w}, \boldsymbol{\theta})\|] < \infty$. *Then* $\widehat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$.

Proof. We prove this theorem similarly as in the proof of the consistency of the MLE. Let $Q_n(\boldsymbol{\theta})$ be $-\overline{g}_n(\boldsymbol{\theta})'\mathbf{W}_n\overline{g}_n(\boldsymbol{\theta})$; we need to verify the four conditions at the end of Section 3.1 of Chapter 4. Condition (I): Let $\mathbf{R}$ be such that $\mathbf{R}'\mathbf{R} = \mathbf{W}$. If $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$, then $\mathbf{0} \neq \mathbf{W}g(\boldsymbol{\theta}) = \mathbf{R}'\mathbf{R}g(\boldsymbol{\theta})$ implies $\mathbf{R}g(\boldsymbol{\theta}) \neq 0$ and hence $Q(\boldsymbol{\theta}) = -(\mathbf{R}g(\boldsymbol{\theta}))'(\mathbf{R}g(\boldsymbol{\theta})) < Q(\boldsymbol{\theta}_0) = 0$ for $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$, where $g(\boldsymbol{\theta}) = E[g(\mathbf{w}, \boldsymbol{\theta})]$. Condition (II) is (ii) of the theorem. Given (ii), (iii) and (iv), Mickey's Theorem implies $g(\boldsymbol{\theta})$ is continuous and $\sup_{\boldsymbol{\theta} \in \Theta} \|\overline{g}_n(\boldsymbol{\theta}) - g(\boldsymbol{\theta})\| \xrightarrow{p} 0$. Thus condition (III) holds by $Q(\boldsymbol{\theta}) = -g(\boldsymbol{\theta})'\mathbf{W}g(\boldsymbol{\theta})$ continuous. By Θ compact, $g(\boldsymbol{\theta})$ is bounded on Θ, and by the triangle and Cauchy-Schwartz inequalities,

$$\begin{aligned} & |Q_n(\boldsymbol{\theta}) - Q(\boldsymbol{\theta})| \\ \leq\ & \left|[\overline{g}_n(\boldsymbol{\theta}) - g(\boldsymbol{\theta})]'\mathbf{W}_n[\overline{g}_n(\boldsymbol{\theta}) - g(\boldsymbol{\theta})]\right| + \left|g(\boldsymbol{\theta})'\left(\mathbf{W}_n + \mathbf{W}_n'\right)[\overline{g}_n(\boldsymbol{\theta}) - g(\boldsymbol{\theta})]\right| + \left|g(\boldsymbol{\theta})'(\mathbf{W}_n - \mathbf{W})g(\boldsymbol{\theta})\right| \\ \leq\ & \|\overline{g}_n(\boldsymbol{\theta}) - g(\boldsymbol{\theta})\|^2 \|\mathbf{W}_n\| + 2\|g(\boldsymbol{\theta})\| \|\overline{g}_n(\boldsymbol{\theta}) - g(\boldsymbol{\theta})\| \|\mathbf{W}_n\| + \|g(\boldsymbol{\theta})\|^2 \|\mathbf{W}_n - \mathbf{W}\|, \end{aligned}$$

so that $\sup_{\boldsymbol{\theta} \in \Theta} |Q_n(\boldsymbol{\theta}) - Q(\boldsymbol{\theta})| \xrightarrow{p} 0$, and condition (IV) holds. ■

Theorem 4 *Suppose that the conditions in the above theorem hold,* $\mathbf{W}_n \xrightarrow{p} \mathbf{W}$, *and (i)* $\boldsymbol{\theta}_0 \in$*interior of* Θ; *(ii)* $g(\mathbf{w}, \boldsymbol{\theta})$ *is continuously differentiable in a neighborhood of* $\mathcal{N}$ *of* $\boldsymbol{\theta}_0$, *with probability approaching one; (iii)* $E[g(\mathbf{w}, \boldsymbol{\theta}_0)] = \mathbf{0}$ *and* $E\left[\|g(\mathbf{w}, \boldsymbol{\theta}_0)\|^2\right] < \infty$; *(iv)* $E[\sup_{\boldsymbol{\theta} \in \mathcal{N}} \|\nabla_{\boldsymbol{\theta}} g(\mathbf{w}, \boldsymbol{\theta})\|] < \infty$, *where* $\nabla_{\boldsymbol{\theta}} g(\mathbf{w}, \boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}'} g(\mathbf{w}, \boldsymbol{\theta})$ *; (v)* $\mathbf{G}'\mathbf{W}\mathbf{G}$ *is nonsingular for* $\mathbf{G} = E[\nabla_{\boldsymbol{\theta}} g(\mathbf{w}, \boldsymbol{\theta}_0)]$. *Then for* $\boldsymbol{\Omega} = E[g(\mathbf{w}, \boldsymbol{\theta}_0) g(\mathbf{w}, \boldsymbol{\theta}_0)']$, $\sqrt{n}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \xrightarrow{d} N(0, \mathbf{V})$, *where* $\mathbf{V} = (\mathbf{G}'\mathbf{W}\mathbf{G})^{-1}\mathbf{G}'\mathbf{W}\boldsymbol{\Omega}\mathbf{W}\mathbf{G}(\mathbf{G}'\mathbf{W}\mathbf{G})^{-1}$.

Proof. By (i), (ii) and (iii), the FOC $2\mathbf{G}_n\left(\widehat{\boldsymbol{\theta}}\right)'\mathbf{W}_n\overline{g}_n\left(\widehat{\boldsymbol{\theta}}\right) = \mathbf{0}$ is satisfied with probability approaching one, where $\mathbf{G}_n(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}\overline{g}_n(\boldsymbol{\theta})$. Expanding $\overline{g}_n(\widehat{\boldsymbol{\theta}})$ around $\boldsymbol{\theta}_0$, multiplying through by $\sqrt{n}$, and solving gives

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) = -\left[\mathbf{G}_n\left(\widehat{\boldsymbol{\theta}}\right)'\mathbf{W}_n\mathbf{G}_n\left(\overline{\boldsymbol{\theta}}\right)\right]\mathbf{G}_n\left(\widehat{\boldsymbol{\theta}}\right)'\mathbf{W}_n\sqrt{n}\overline{g}_n(\boldsymbol{\theta}_0),$$

where $\overline{\boldsymbol{\theta}}$ is the mean value. By (iv), $\mathbf{G}_n\left(\widehat{\boldsymbol{\theta}}\right) \xrightarrow{p} \mathbf{G}$ and $\mathbf{G}_n\left(\overline{\boldsymbol{\theta}}\right) \xrightarrow{p} \mathbf{G}$, so that by (v),

$$\left[\mathbf{G}_n\left(\widehat{\boldsymbol{\theta}}\right)'\mathbf{W}_n\mathbf{G}_n\left(\overline{\boldsymbol{\theta}}\right)\right]\mathbf{G}_n\left(\widehat{\boldsymbol{\theta}}\right)'\mathbf{W}_n \xrightarrow{p} \left(\mathbf{G}'\mathbf{W}\mathbf{G}\right)^{-1}\mathbf{G}'\mathbf{W}.$$

The conclusion then follows by Slutsky's theorem. ■

The complicated asymptotic variance formula simplifies to $\left(\mathbf{G}'\boldsymbol{\Omega}^{-1}\mathbf{G}\right)^{-1}$ when $\mathbf{W} = \boldsymbol{\Omega}^{-1}$. As shown in Hansen (1982), this value for $\mathbf{W}$ is optimal in the sense that it minimizes the asymptotic variance matrix of the GMM estimator. $\mathbf{V}$ can be consistently estimated by its sample analog,

$$\widehat{\mathbf{V}} = \left(\widehat{\mathbf{G}}'\mathbf{W}_n\widehat{\mathbf{G}}\right)^{-1}\widehat{\mathbf{G}}'\mathbf{W}_n\widehat{\boldsymbol{\Omega}}\mathbf{W}_n\widehat{\mathbf{G}}\left(\widehat{\mathbf{G}}'\mathbf{W}_n\widehat{\mathbf{G}}\right)^{-1},$$

where $\widehat{\mathbf{G}} = \mathbf{G}_n\left(\widehat{\boldsymbol{\theta}}\right)$, and $\widehat{\boldsymbol{\Omega}} = n^{-1}\sum_{i=1}^n g(\mathbf{w}_i, \widehat{\boldsymbol{\theta}})g(\mathbf{w}_i, \widehat{\boldsymbol{\theta}})'$. To prove the consistency of $\widehat{\mathbf{V}}$, we first prove the following lemma, which is Lemma 4.3 of Newey and McFadden (1994).

Lemma 1 *If $\mathbf{w}_i$ is i.i.d., $a(\mathbf{w}, \boldsymbol{\theta})$ is continuous at $\boldsymbol{\theta}_0$ with probability one, and there is a neighborhood $\mathcal{N}$ of $\boldsymbol{\theta}_0$ such that $E\left[\sup_{\boldsymbol{\theta}\in\mathcal{N}} \|a(\mathbf{w}, \boldsymbol{\theta})\|\right] < \infty$, then for any $\widehat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$, $n^{-1}\sum_{i=1}^n a\left(\mathbf{w}_i, \widehat{\boldsymbol{\theta}}\right) \xrightarrow{p} E\left[a\left(\mathbf{w}, \boldsymbol{\theta}_0\right)\right]$.*

Proof. By consistency of $\widehat{\boldsymbol{\theta}}$ there is $\delta_n \to 0$ such that $\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right\| \leq \delta_n$ with probability approaching one. Let $\Delta_n(\mathbf{w}) = \sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\leq\delta_n} \|a\left(\mathbf{w}, \boldsymbol{\theta}\right) - a\left(\mathbf{w}, \boldsymbol{\theta}_0\right)\|$. By continuity of $a\left(\mathbf{w}, \boldsymbol{\theta}\right)$ at $\boldsymbol{\theta}_0$, $\Delta_n(\mathbf{w}) \to 0$ with probability one, while by the dominance condition, for n large enough $\Delta_n(\mathbf{w}) \leq 2\sup_{\boldsymbol{\theta}\in\mathcal{N}} \|a(\mathbf{w}, \boldsymbol{\theta})\|$. Then by the dominated convergence theorem,[18] $E\left[\Delta_n(\mathbf{w})\right] \to 0$, so by the Markov inequality,[19] $P\left(\left|n^{-1}\sum_{i=1}^n \Delta_n(\mathbf{w}_i)\right| > \varepsilon\right) \leq E\left[\Delta_n(\mathbf{w})\right]/\varepsilon \to 0$ for all $\varepsilon > 0$, giving $n^{-1}\sum_{i=1}^n \Delta_n(\mathbf{w}_i) \xrightarrow{p} 0$. By the LLN, $n^{-1}\sum_{i=1}^n a\left(\mathbf{w}_i, \boldsymbol{\theta}_0\right) \xrightarrow{p} E\left[a\left(\mathbf{w}, \boldsymbol{\theta}_0\right)\right]$. Also, with probability approaching one,

$$\left\|n^{-1}\sum_{i=1}^n a\left(\mathbf{w}_i, \widehat{\boldsymbol{\theta}}\right) - n^{-1}\sum_{i=1}^n a\left(\mathbf{w}_i, \boldsymbol{\theta}_0\right)\right\| \leq n^{-1}\sum_{i=1}^n \left\|a\left(\mathbf{w}_i, \widehat{\boldsymbol{\theta}}\right) - a\left(\mathbf{w}_i, \boldsymbol{\theta}_0\right)\right\| \leq n^{-1}\sum_{i=1}^n \Delta_n(\mathbf{w}_i) \xrightarrow{p} 0,$$

so the conclusion follows by the triangle inequality. ■

The conditions in this lemma are weaker than those of Mickey's theorem, because the conclusion is simply uniform convergence at the true parameter. In particular, the function is only required to be continuous at the true parameter.

Theorem 5 *If the assumptions in the last theorem are satisfied, and for neighborhood $\mathcal{N}$ of $\boldsymbol{\theta}_0$, $E[\sup_{\boldsymbol{\theta}\in\mathcal{N}} \|g(\mathbf{w}, \boldsymbol{\theta})\|^2] < \infty$, then $\widehat{\mathbf{V}} \xrightarrow{p} \mathbf{V}$.*

Proof. Applying the above lemma to $a(\mathbf{w}, \boldsymbol{\theta}) = g(\mathbf{w}, \boldsymbol{\theta})g(\mathbf{w}, \boldsymbol{\theta})'$, we get $\widehat{\boldsymbol{\Omega}} \xrightarrow{p} \boldsymbol{\Omega}$, and applying to $a(\mathbf{w}, \boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} g(\mathbf{w}, \boldsymbol{\theta})$, we get $\widehat{\mathbf{G}} \xrightarrow{p} \mathbf{G}$. The conclusion follows from the CMT and continuity of matrix inversion and multiplication. ■

[18] Dominated convergence theorem: If $X_n \xrightarrow{p} X$ and for any n, $|X_n| \leq Y$ with $E\left[Y\right] < \infty$, then $E[X_n] \to E[X]$.
[19] Markov's inequality: For any nonnegative random variable X and $a > 0$, $P\left(X > a\right) \leq E\left[X\right]/a$.

Appendix B: Asymptotics for EL

We first show the asymptotic distribution of $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\lambda}}$. Note that $\left(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\lambda}}\right)$ jointly solve

$$\mathbf{0} = \frac{\partial}{\partial \boldsymbol{\lambda}} R\left(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\lambda}}\right) = -\sum_{i=1}^{n} \frac{g_i(\widehat{\boldsymbol{\theta}})}{1 + \widehat{\boldsymbol{\lambda}}' g_i(\widehat{\boldsymbol{\theta}})} \tag{38}$$

and

$$\mathbf{0} = \frac{\partial}{\partial \boldsymbol{\theta}} R\left(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\lambda}}\right) = -\sum_{i=1}^{n} \frac{\mathbf{G}_i(\widehat{\boldsymbol{\theta}})' \widehat{\boldsymbol{\lambda}}}{1 + \widehat{\boldsymbol{\lambda}}' g_i(\widehat{\boldsymbol{\theta}})}, \tag{39}$$

where $\mathbf{G}_i(\boldsymbol{\theta}) = \partial g_i(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}'$. Let $\mathbf{G}_n = n^{-1} \sum_{i=1}^{n} \mathbf{G}_i(\boldsymbol{\theta}_0) = \mathbf{G}_n(\boldsymbol{\theta}_0)$, $\overline{g}_n = n^{-1} \sum_{i=1}^{n} g_i(\boldsymbol{\theta}_0) = \overline{g}_n(\boldsymbol{\theta}_0)$ and $\boldsymbol{\Omega}_n = n^{-1} \sum_{i=1}^{n} g_i(\boldsymbol{\theta}_0) g_i(\boldsymbol{\theta}_0)' = \boldsymbol{\Omega}_n(\boldsymbol{\theta}_0)$.

Expanding (39) around $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ and $\boldsymbol{\lambda} = \boldsymbol{\lambda}_0 = \mathbf{0}$ yields

$$\mathbf{0} \approx \mathbf{G}_n' \left(\widehat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_0\right) = \mathbf{G}_n' \widehat{\boldsymbol{\lambda}}. \tag{40}$$

Expanding (38) around $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ and $\boldsymbol{\lambda} = \boldsymbol{\lambda}_0 = \mathbf{0}$ yields

$$\mathbf{0} \approx -\overline{g}_n - \mathbf{G}_n \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) + \boldsymbol{\Omega}_n \widehat{\boldsymbol{\lambda}}. \tag{41}$$

Premultiplying by $\mathbf{G}_n' \boldsymbol{\Omega}_n^{-1}$ and using (40) yields

$$\begin{aligned} \mathbf{0} &\approx -\mathbf{G}_n' \boldsymbol{\Omega}_n^{-1} \overline{g}_n - \mathbf{G}_n' \boldsymbol{\Omega}_n^{-1} \mathbf{G}_n \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) + \mathbf{G}_n' \boldsymbol{\Omega}_n^{-1} \boldsymbol{\Omega}_n \widehat{\boldsymbol{\lambda}} \\ &= -\mathbf{G}_n' \boldsymbol{\Omega}_n^{-1} \overline{g}_n - \mathbf{G}_n' \boldsymbol{\Omega}_n^{-1} \mathbf{G}_n \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right). \end{aligned}$$

Solving for $\widehat{\boldsymbol{\theta}}$ and using the WLLN and CLT yields

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \approx -\left(\mathbf{G}_n' \boldsymbol{\Omega}_n^{-1} \mathbf{G}_n\right)^{-1} \mathbf{G}_n' \boldsymbol{\Omega}_n^{-1} \overline{g}_n \xrightarrow{d} \left(\mathbf{G}' \boldsymbol{\Omega}^{-1} \mathbf{G}\right)^{-1} \mathbf{G}' \boldsymbol{\Omega}^{-1} N(\mathbf{0}, \boldsymbol{\Omega}) = N(\mathbf{0}, \mathbf{V}). \tag{42}$$

Solving (41) for $\widehat{\boldsymbol{\lambda}}$ and using (42) yields

$$\begin{aligned} \sqrt{n}\widehat{\boldsymbol{\lambda}} &\approx \boldsymbol{\Omega}_n^{-1} \left(\mathbf{I} - \mathbf{G}_n \left(\mathbf{G}_n' \boldsymbol{\Omega}_n^{-1} \mathbf{G}_n\right)^{-1} \mathbf{G}_n' \boldsymbol{\Omega}_n^{-1}\right) \sqrt{n} \overline{g}_n \\ &\xrightarrow{d} \boldsymbol{\Omega}^{-1} \left(\mathbf{I} - \mathbf{G} \left(\mathbf{G}' \boldsymbol{\Omega}^{-1} \mathbf{G}\right)^{-1} \mathbf{G}' \boldsymbol{\Omega}^{-1}\right) N(\mathbf{0}, \boldsymbol{\Omega}) \\ &= \boldsymbol{\Omega}^{-1} N(\mathbf{0}, \mathbf{V}_{\boldsymbol{\lambda}}) \end{aligned} \tag{43}$$

Furthermore, since

$$\boldsymbol{\Omega}^{-1} \left(\mathbf{I} - \mathbf{G} \left(\mathbf{G}' \boldsymbol{\Omega}^{-1} \mathbf{G}\right)^{-1} \mathbf{G}' \boldsymbol{\Omega}^{-1}\right) \boldsymbol{\Omega} \boldsymbol{\Omega}^{-1} \mathbf{G} \left(\mathbf{G}' \boldsymbol{\Omega}^{-1} \mathbf{G}\right)^{-1} = \mathbf{0},$$

$\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\lambda}}$ are asymptotically uncorrelated and hence independent.

We now show the asymptotic null distribution of LR_n. First, by a Taylor expansion, (42), and

(43),

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\overline{g}_n\left(\widehat{\boldsymbol{\theta}}\right) &\approx \sqrt{n}\left(\overline{g}_n+\mathbf{G}_n\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right)\right)\\ &\approx \left(\mathbf{I}-\mathbf{G}_n\left(\mathbf{G}_n'\boldsymbol{\Omega}_n^{-1}\mathbf{G}_n\right)^{-1}\mathbf{G}_n'\boldsymbol{\Omega}_n^{-1}\right)\sqrt{n}\overline{g}_n\\ &\approx \boldsymbol{\Omega}_n\sqrt{n}\widehat{\boldsymbol{\lambda}}.\end{aligned}$$

Second, since $\log(1+u)\approx u-u^2/2$ for u small,

$$\begin{aligned}LR_n &= \sum_{i=1}^{n}2\log\left(1+\widehat{\boldsymbol{\lambda}}'g_i(\widehat{\boldsymbol{\theta}})\right)\approx 2\widehat{\boldsymbol{\lambda}}'\sum_{i=1}^{n}\widehat{g}_i-\widehat{\boldsymbol{\lambda}}'\sum_{i=1}^{n}g_i(\widehat{\boldsymbol{\theta}})g_i(\widehat{\boldsymbol{\theta}})'\widehat{\boldsymbol{\lambda}}\\ &\approx n\widehat{\boldsymbol{\lambda}}'\boldsymbol{\Omega}_n\widehat{\boldsymbol{\lambda}}\xrightarrow{d}N\left(\mathbf{0},\mathbf{V}_{\boldsymbol{\lambda}}\right)'\boldsymbol{\Omega}^{-1}N\left(\mathbf{0},\mathbf{V}_{\boldsymbol{\lambda}}\right)=\chi^2_{l-k}.\end{aligned}$$

Exercise 18 *Show that* $N\left(\mathbf{0},\mathbf{V}_{\boldsymbol{\lambda}}\right)'\boldsymbol{\Omega}^{-1}N\left(\mathbf{0},\mathbf{V}_{\boldsymbol{\lambda}}\right)=\chi^2_{l-k}$.

Appendix C: Linear Procedure of Windmeijer (2005)

Consider only the linear-in-parameter model. Suppose the step-one estimator is $\widehat{\boldsymbol{\theta}}_1$, and $\mathbf{W}_n$ depends on $\widehat{\boldsymbol{\theta}}_1$ through

$$\mathbf{W}_n\left(\widehat{\boldsymbol{\theta}}_1\right)=\frac{1}{n}\sum_{i=1}^{n}g_i\left(\widehat{\boldsymbol{\theta}}_1\right)g_i\left(\widehat{\boldsymbol{\theta}}_1\right)'.$$

The step-two estimator satisfies

$$\begin{aligned}\widehat{\boldsymbol{\theta}}_2-\boldsymbol{\theta}_0 &= -\left(\mathbf{G}_n'\mathbf{W}_n^{-1}\left(\widehat{\boldsymbol{\theta}}_1\right)\mathbf{G}_n\right)^{-1}\mathbf{G}_n'\mathbf{W}_n^{-1}\left(\widehat{\boldsymbol{\theta}}_1\right)\overline{g}_n\left(\boldsymbol{\theta}_0\right)\\ &= -\left(\mathbf{G}_n'\mathbf{W}_n^{-1}\left(\boldsymbol{\theta}_0\right)\mathbf{G}_n\right)^{-1}\mathbf{G}_n'\mathbf{W}_n^{-1}\left(\boldsymbol{\theta}_0\right)\overline{g}_n\left(\boldsymbol{\theta}_0\right)+\mathbf{D}_{\boldsymbol{\theta}_0,\mathbf{W}_n(\boldsymbol{\theta}_0)}\left(\widehat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_0\right)+o_p\left(n^{-1}\right),\end{aligned}$$

where $\mathbf{G}_n=\partial\overline{g}_n\left(\boldsymbol{\theta}\right)/\partial\boldsymbol{\theta}'$ does not depend on $\boldsymbol{\theta}$, and the jth column of $\mathbf{D}_{\boldsymbol{\theta}_0,\mathbf{W}_n(\boldsymbol{\theta}_0)}$ is given by

$$\begin{aligned}\mathbf{D}_{\boldsymbol{\theta}_0,\mathbf{W}_n(\boldsymbol{\theta}_0)}[\cdot,j] &= -\left(\mathbf{G}_n'\mathbf{W}_n^{-1}\left(\boldsymbol{\theta}_0\right)\mathbf{G}_n\right)^{-1}\left[\mathbf{G}_n'\mathbf{W}_n^{-1}\left(\boldsymbol{\theta}_0\right)\left.\frac{\partial\mathbf{W}_n\left(\boldsymbol{\theta}\right)}{\partial\theta_j}\right|_{\boldsymbol{\theta}_0}\mathbf{W}_n^{-1}\left(\boldsymbol{\theta}_0\right)\mathbf{G}_n\right]\\ &\quad\cdot\left(\mathbf{G}_n'\mathbf{W}_n^{-1}\left(\boldsymbol{\theta}_0\right)\mathbf{G}_n\right)^{-1}\mathbf{G}_n'\mathbf{W}_n^{-1}\left(\boldsymbol{\theta}_0\right)\overline{g}_n\left(\boldsymbol{\theta}_0\right)\\ &\quad+\left(\mathbf{G}_n'\mathbf{W}_n^{-1}\left(\boldsymbol{\theta}_0\right)\mathbf{G}_n\right)^{-1}\mathbf{G}_n'\mathbf{W}_n^{-1}\left(\boldsymbol{\theta}_0\right)\left.\frac{\partial\mathbf{W}_n\left(\boldsymbol{\theta}\right)}{\partial\theta_j}\right|_{\boldsymbol{\theta}_0}\mathbf{W}_n^{-1}\left(\boldsymbol{\theta}_0\right)\overline{g}_n\left(\boldsymbol{\theta}_0\right).\end{aligned}$$

$\mathbf{D}_{\boldsymbol{\theta}_0,\mathbf{W}_n(\boldsymbol{\theta}_0)}\left(\widehat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_0\right)=O_p(n^{-1})$, so taking account of this term will result in a more accurate approximation of the variance of $\widehat{\boldsymbol{\theta}}_2$ in finite samples. Note that when the model is just identified, the correction disappears.

A step-one linear estimator satisfies

$$\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_0 = -\left(\mathbf{G}_n'\mathbf{W}_n^{-1}\mathbf{G}_n\right)^{-1}\mathbf{G}_n'\mathbf{W}_n^{-1}\overline{g}_n\left(\boldsymbol{\theta}_0\right)$$

and the finite sample corrected estimate of the variance of $\widehat{\boldsymbol{\theta}}_2$ can be obtained as

$$\begin{aligned}\widehat{Var}\left(\widehat{\boldsymbol{\theta}}_2\right) &= \frac{1}{n}\left(\mathbf{G}_n'\mathbf{W}_n^{-1}\left(\widehat{\boldsymbol{\theta}}_1\right)\mathbf{G}_n\right)^{-1} + \mathbf{D}_{\widehat{\boldsymbol{\theta}}_2,\mathbf{W}_n(\widehat{\boldsymbol{\theta}}_1)}\widehat{Var}\left(\widehat{\boldsymbol{\theta}}_1\right)\mathbf{D}_{\widehat{\boldsymbol{\theta}}_2,\mathbf{W}_n(\widehat{\boldsymbol{\theta}}_1)}' \\ &\quad + \frac{1}{n}\left[\mathbf{D}_{\widehat{\boldsymbol{\theta}}_2,\mathbf{W}_n(\widehat{\boldsymbol{\theta}}_1)}\left(\mathbf{G}_n'\mathbf{W}_n^{-1}\left(\widehat{\boldsymbol{\theta}}_1\right)\mathbf{G}_n\right)^{-1} + \left(\mathbf{G}_n'\mathbf{W}_n^{-1}\left(\widehat{\boldsymbol{\theta}}_1\right)\mathbf{G}_n\right)^{-1}\mathbf{D}_{\widehat{\boldsymbol{\theta}}_2,\mathbf{W}_n(\widehat{\boldsymbol{\theta}}_1)}'\right],\end{aligned}$$

where the first term is the conventional estimate of the asymptotic variance; the first term of $\mathbf{D}_{\widehat{\boldsymbol{\theta}}_2,\mathbf{W}_n(\widehat{\boldsymbol{\theta}}_1)}$ is zero since $\left(\mathbf{G}_n'\mathbf{W}_n^{-1}\left(\widehat{\boldsymbol{\theta}}_1\right)\mathbf{G}_n\right)^{-1}\mathbf{G}_n'\mathbf{W}_n^{-1}\left(\widehat{\boldsymbol{\theta}}_1\right)\overline{g}_n\left(\widehat{\boldsymbol{\theta}}_2\right) = \mathbf{0}$ from the FOCs, so

$$\mathbf{D}_{\widehat{\boldsymbol{\theta}}_2,\mathbf{W}_n(\widehat{\boldsymbol{\theta}}_1)} = \left(\mathbf{G}_n'\mathbf{W}_n^{-1}\left(\widehat{\boldsymbol{\theta}}_1\right)\mathbf{G}_n\right)^{-1}\mathbf{G}_n'\mathbf{W}_n^{-1}\left(\widehat{\boldsymbol{\theta}}_1\right)\left.\frac{\partial\mathbf{W}_n\left(\boldsymbol{\theta}\right)}{\partial\theta_j}\right|_{\widehat{\boldsymbol{\theta}}_1}\mathbf{W}_n^{-1}\left(\widehat{\boldsymbol{\theta}}_1\right)\overline{g}_n\left(\widehat{\boldsymbol{\theta}}_2\right)$$

and

$$\widehat{Var}\left(\widehat{\boldsymbol{\theta}}_1\right) = \frac{1}{n}\left(\mathbf{G}_n'\mathbf{W}_n^{-1}\mathbf{G}_n\right)^{-1}\mathbf{G}_n'\mathbf{W}_n^{-1}\mathbf{W}_n\left(\widehat{\boldsymbol{\theta}}_1\right)\mathbf{W}_n^{-1}\mathbf{G}_n\left(\mathbf{G}_n'\mathbf{W}_n^{-1}\mathbf{G}_n\right)^{-1}.$$

$\widehat{Var}\left(\widehat{\boldsymbol{\theta}}_2\right)$ will provide a better finite sample estimate of $Var\left(\widehat{\boldsymbol{\theta}}_2\right)$ by taking into account the finite sample variation of $\widehat{\boldsymbol{\theta}}_1$.

Appendix D: Semiparametric Efficiency Bound

Recall from Section 3.3 of Chapter 4, there are two criteria of efficiency. Chamberlain (1986) uses the first criterion, i.e., best among regular estimators, while Chamberlain (1987, 1992) use the second criterion, i.e., local asymptotic minimaxity (LAM) among all estimators. We will concentrate on Chamberlain (1987) in this appendix. A key advantage of Chamberlain (1987)'s method is that it provides a conjecture for the form of the semiparametric efficiency bound (and thus the form of the efficient influence function) so sidesteps the need to directly calculate a complicated projection problem in using the first criterion. More discussions on the first criterion can be found in Newey (1990a) and Bickel et al. (1998).

Following the notations in the main text, suppose we have the following conditional moment conditions in hand,

$$E[u\left(\mathbf{w},\boldsymbol{\beta}_0\right)|\mathbf{x}] = \mathbf{0}.$$

First (without loss of generality) transform the moments so they have unit conditional variances and are uncorrelated conditional on $\mathbf{x}$,

$$u^*\left(\mathbf{w},\boldsymbol{\beta}\right) = \boldsymbol{\Omega}^{-1/2}(\mathbf{x})u\left(\mathbf{w},\boldsymbol{\beta}\right),$$

where $\boldsymbol{\Omega}(\mathbf{x}) = E\left[u\left(\mathbf{w},\boldsymbol{\beta}_0\right)u\left(\mathbf{w},\boldsymbol{\beta}_0\right)'\middle|\mathbf{x}\right]$. Then we ask how to choose $\mathbf{A}^*(\mathbf{x}) = \mathbf{A}(\mathbf{x})\boldsymbol{\Omega}^{1/2}(\mathbf{x})$ such

that the moment conditions $E[m(\mathbf{w},\boldsymbol{\beta})] = E[\mathbf{A}^*(\mathbf{x})u^*(\mathbf{w},\boldsymbol{\beta})]$ is optimal, where note that each moment in u^* is on an "equal footing".

From the intuition in Section 2 of Chapter 4, the larger $\frac{\partial}{\partial \boldsymbol{\beta}} u_j^*(\mathbf{w},\boldsymbol{\beta}_0)$, $j = 1,\cdots,s$, is, the more informative is the jth mometn for estimating $\boldsymbol{\beta}_0$. So one might choose weights for each $u_j^*(\mathbf{w},\boldsymbol{\beta}_0)$ to be proportional to $\frac{\partial}{\partial \boldsymbol{\beta}} u_j^*(\mathbf{w},\boldsymbol{\beta}_0)$. Thus one might consider

$$m(\mathbf{w},\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}'} u^*(\mathbf{w},\boldsymbol{\beta}_0)' u^*(\mathbf{w},\boldsymbol{\beta}).$$

But then we might have $E[m(\mathbf{w},\boldsymbol{\beta}_0)] \neq \mathbf{0}$ and so GMM would not necessarily be consistent. To ensure consistency, use a function of $\mathbf{x}$ "close" to $\frac{\partial}{\partial \boldsymbol{\beta}'} u^*(\mathbf{w},\boldsymbol{\beta}_0)$, i.e.,

$$\mathbf{A}^*(\mathbf{x}) = E\left[\frac{\partial}{\partial \boldsymbol{\beta}'} u^*(\mathbf{w},\boldsymbol{\beta}_0)\middle|\mathbf{x}\right]' = \mathbf{G}(\mathbf{x})',$$

so that $m(\mathbf{w},\boldsymbol{\beta}) = \mathbf{A}^*(\mathbf{x})u^*(\mathbf{w},\boldsymbol{\beta})$ and $E[m(\mathbf{w},\boldsymbol{\beta}_0)] = \mathbf{0}$. This implies

$$\begin{aligned}
\mathbf{A}(\mathbf{x}) &= \mathbf{A}^*(\mathbf{x})\boldsymbol{\Omega}^{-1/2}(\mathbf{x}) = E\left[\frac{\partial}{\partial \boldsymbol{\beta}'} u^*(\mathbf{w},\boldsymbol{\beta}_0)\middle|\mathbf{x}\right]' \boldsymbol{\Omega}^{-1/2}(\mathbf{x}) \\
&= E\left[\frac{\partial}{\partial \boldsymbol{\beta}'} u(\mathbf{w},\boldsymbol{\beta}_0)\middle|\mathbf{x}\right]' \boldsymbol{\Omega}^{-1/2}(\mathbf{x})\boldsymbol{\Omega}^{-1/2}(\mathbf{x}) = \mathbf{G}(\mathbf{x})'\boldsymbol{\Omega}^{-1}(\mathbf{x}),
\end{aligned}$$

and

$$m(\mathbf{w},\boldsymbol{\beta}) = \mathbf{G}(\mathbf{x})'\boldsymbol{\Omega}^{-1}(\mathbf{x})u(\mathbf{w},\boldsymbol{\beta}).$$

Chamberlain (1987) assumes that $\mathbf{x}$ has a multinomial (finite discrete) distribution. Then the conditional moment restrictions are equivalent to a set of unconditional moment conditions. Using the optimal weight matrix choice, we can then derive the asymptotic lower bound for the GMM estimator using these unconditional moment conditions. We then argue that this bound applies to the conditional moment restriction case since an arbitrary distribution for $\mathbf{x}$ can be approximated as well as desired by a multinomial distribution.

Suppose $\{\tau_1,\cdots,\tau_J\}$ is the support for $\mathbf{x}$, where $\pi_j = P(\mathbf{x} = \tau_j)$ for $j = 1,\cdots,J$. Note that

$$\begin{aligned}
E[1(\mathbf{x} = \tau_j)u(\mathbf{w},\boldsymbol{\beta})] &= E[1(\mathbf{x} = \tau_j)E[u(\mathbf{w},\boldsymbol{\beta})|\mathbf{x}]] \\
&= P(\mathbf{x} = \tau_j)E[u(\mathbf{w},\boldsymbol{\beta})|\mathbf{x} = \tau_j] \\
&= \pi_j E[u(\mathbf{w},\boldsymbol{\beta})|\mathbf{x} = \tau_j].
\end{aligned}$$

So $E[u(\mathbf{w},\boldsymbol{\beta}_0)|\mathbf{x}] = \mathbf{0}$ iff $E[u(\mathbf{w},\boldsymbol{\beta})|\mathbf{x} = \tau_j] = \mathbf{0}$ for $j = 1,\cdots,J$ iff $E[1(\mathbf{x} = \tau_j)u(\mathbf{w},\boldsymbol{\beta})] = \mathbf{0}$ for $j = 1,\cdots,J$, where iff means "if and only if".

If $h(\mathbf{x}) = (1(\mathbf{x} = \tau_1),\cdots,1(\mathbf{x} = \tau_J))'$, then let $m(\mathbf{w},\boldsymbol{\beta}) = h(\mathbf{x}) \otimes u(\mathbf{w},\boldsymbol{\beta})$ (so $E[m(\mathbf{w},\boldsymbol{\beta})] = \mathbf{0}$), where $\otimes$ is the Kronecker product which is defined for two matrices $\mathbf{A}_{M\times N} = (a_{ij})$ and $\mathbf{B}_{K\times L} =$

(b_{ij}) as

$$\mathbf{A} \otimes \mathbf{B} = \begin{pmatrix} a_{11}\mathbf{B} & \cdots & a_{1N}\mathbf{B} \\ \vdots & \ddots & \vdots \\ a_{M1}\mathbf{B} & \cdots & a_{MN}\mathbf{B} \end{pmatrix}.$$

Using the optimal weight matrix, the best asymptotic variance for the GMM estimator using these (unconditional) moments is $\left(\mathbf{G}'\mathbf{\Omega}^{-1}\mathbf{G}\right)^{-1}$, where

$$\mathbf{G} = \frac{\partial}{\partial \boldsymbol{\beta}'} E\left[m\left(\mathbf{w}, \boldsymbol{\beta}_0\right)\right] = E\left[h\left(\mathbf{x}\right) \otimes \frac{\partial}{\partial \boldsymbol{\beta}'} u\left(\mathbf{w}, \boldsymbol{\beta}_0\right)\right] = \begin{pmatrix} \pi_1 E\left[\frac{\partial}{\partial \boldsymbol{\beta}'} u\left(\mathbf{w}, \boldsymbol{\beta}_0\right) \middle| \mathbf{x} = \tau_1\right] \\ \vdots \\ \pi_J E\left[\frac{\partial}{\partial \boldsymbol{\beta}'} u\left(\mathbf{w}, \boldsymbol{\beta}_0\right) \middle| \mathbf{x} = \tau_J\right] \end{pmatrix}_{Js \times k}$$

and

$$\begin{aligned} \mathbf{\Omega} &= E\left[m\left(\mathbf{w}, \boldsymbol{\beta}_0\right) m\left(\mathbf{w}, \boldsymbol{\beta}_0\right)'\right] = E\left[h\left(\mathbf{x}\right) h\left(\mathbf{x}\right)' \otimes u\left(\mathbf{w}, \boldsymbol{\beta}\right) u\left(\mathbf{w}, \boldsymbol{\beta}_0\right)'\right] \\ &= E\left[\operatorname{diag}\left(h\left(\mathbf{x}\right)\right) \otimes u\left(\mathbf{w}, \boldsymbol{\beta}_0\right) u\left(\mathbf{w}, \boldsymbol{\beta}_0\right)'\right] \\ &= \operatorname{diag}\left(\pi_1 E\left[u\left(\mathbf{w}, \boldsymbol{\beta}_0\right) u\left(\mathbf{w}, \boldsymbol{\beta}_0\right)' \middle| \mathbf{x} = \tau_1\right], \cdots, \pi_J E\left[u\left(\mathbf{w}, \boldsymbol{\beta}_0\right) u\left(\mathbf{w}, \boldsymbol{\beta}_0\right)' \middle| \mathbf{x} = \tau_J\right]\right). \end{aligned}$$

So

$$\begin{aligned} & \left(\mathbf{G}'\mathbf{\Omega}^{-1}\mathbf{G}\right)^{-1} \\ = & \left\{\sum_{j=1}^{J} \pi_j E\left[\frac{\partial}{\partial \boldsymbol{\beta}'} u\left(\mathbf{w}, \boldsymbol{\beta}_0\right) \middle| \mathbf{x} = \tau_j\right]' E\left[u\left(\mathbf{w}, \boldsymbol{\beta}_0\right) u\left(\mathbf{w}, \boldsymbol{\beta}_0\right)' \middle| \mathbf{x} = \tau_j\right]^{-1} E\left[\frac{\partial}{\partial \boldsymbol{\beta}'} u\left(\mathbf{w}, \boldsymbol{\beta}_0\right) \middle| \mathbf{x} = \tau_j\right]\right\}^{-1} \\ = & E\left\{E\left[\frac{\partial}{\partial \boldsymbol{\beta}'} u\left(\mathbf{w}, \boldsymbol{\beta}_0\right) \middle| \mathbf{x}\right]' E\left[u\left(\mathbf{w}, \boldsymbol{\beta}_0\right) u\left(\mathbf{w}, \boldsymbol{\beta}_0\right)' \middle| \mathbf{x}\right]^{-1} E\left[\frac{\partial}{\partial \boldsymbol{\beta}'} u\left(\mathbf{w}, \boldsymbol{\beta}_0\right) \middle| \mathbf{x}\right]\right\}^{-1} \\ = & E\left[\mathbf{G}(\mathbf{x})'\mathbf{\Omega}(\mathbf{x})^{-1}\mathbf{G}(\mathbf{x})'\right]^{-1}. \end{aligned}$$

Since the multinomial distribution can closely approximate any distribution for $\mathbf{x}$, this is the bound for the conditional moment restriction model. Hence $\mathbf{A}(\mathbf{x})$ is in fact optimal as given above.