14.382 L3. STRUCTURAL EQUATIONS MODELS AND GMM

VICTOR CHERNOZHUKOV AND IVÁN FERNÁNDEZ-VAL

ABSTRACT. Here we analyze a system of simultaneous equations arising in the supply-demand analysis. We derive a system of moment conditions that potentially identify the structural parameters and naturally arrive at a generalized method of moments (GMM) estimator. We end up outlining the general properties of the GMM estimator, and formally verifying these properties for linear moment condition models.

1. WRIGHT'S SUPPLY-DEMAND SYSTEM OF SIMULTANEOUS EQUATIONS

Following the original work of Wright in 1928 on demand and supply for tobacco, we consider the following system of equations where quantities and prices are in logs:

$$\begin{aligned} Y_p^d &= \alpha_1 p + \alpha_2' Z^d + \alpha_3' W + \epsilon^d, \quad \epsilon^d \perp Z^d, Z^s, W, \\ Y_p^s &= \beta_1 p + \beta_2' Z^s + \beta_3' W + \epsilon^s, \quad \epsilon^s \perp Z^d, Z^s, W. \end{aligned} \tag{Wright's M}$$

Potential price is denoted by p, and potential quantities demanded and supplied at price p are denoted by Y_p^d and Y_p^s respectively. Thus, the curve $p \mapsto Y_p^d$ is a random aggregate demand curve, and $p \mapsto Y_p^s$ is a random aggregate supply curve. Each curve is shifted by observable and unobservable variables: elements of W are common shifters (that include a constant), variables in Z^d shift demand only, and variables in Z^s shift only the supply. The shocks ϵ^d and ϵ^s capture all the unobservable shifters of demand and supply curves, and are assumed to be orthogonal to all observable shifters.

We can also think of (Wright's M) as of a collection of pairs of regression equations indexed by p:

$$\begin{aligned} Y_p^d - \alpha_1 p &= \alpha_2' Z^d + \alpha_3' W + \epsilon^d, & \epsilon^d \perp Z^d, Z^s, W \\ Y_p^s - \beta_1 p &= \beta_2' Z^s + \beta_3' W + \epsilon^s, & \epsilon^s \perp Z^d, Z^s, W, \end{aligned}$$

which clarifies the role of the observables in explaining parts of the fluctuations of the supply and demand curves. Here the decision of which W's and Z^d and Z^s to include in the regression specification is similar to specification analysis we employ the regression. It is critical to find useful demand and supply shifters, which often involves creativity and good data collection (for example, the supply of fish is affected by weather conditions at sea, and so you'd need to collect that data if you want to estimate the demand for fish).

The maps $p \mapsto \alpha_1 p$ and $p \mapsto \beta_1 p$ describe the deterministic parts of the random demand and supply functions. This part is structural and the specification here corresponds to the Cobb-Douglas form (since we are working in logs); we could entertain other structural specifications as well.

The structural equations in (Wright's M) determine the equilibrium (log) quantity and (log) price (Y, P) determined via the market clearing condition:

$$Y_P^s = Y_P^d =: Y.$$

Note that in the treatment effects framework, commonly used in data analysis, Y_p^s and Y_p^d are called the *potential outcomes* indexed by p, and the observed outcome Y is obtained after plugging the observed index $p = P$, namely $Y = Y_P^s = Y_P^d$. Inserting the equilibrium quantities (Y, P) in the original equations we obtain the following SEM:

$$\begin{aligned} Y - \alpha_1 P - \alpha_2' Z^d - \alpha_3' W = \epsilon^d, \quad \epsilon^d \perp Z^d, Z^s, W, \\ Y - \beta_1 P - \beta_2' Z^s - \beta_3' W = \epsilon^s, \quad \epsilon^s \perp Z^d, Z^s, W. \end{aligned} \tag{Wright}$$

It should be clear from Figure 1 that we can not hope to identify the structural elasticity parameters α_1 and β_1 from the projection coefficients of Y on P or of P on Y, as equilibrium quantities won't necessarily trace either demand or supply curve. On the other hand, as illustrated in Figure 2, if we have supply shifters that move the supply curve, without affecting the demand curve, then we conclude that the data contain some *quasi-experimental* fluctuations that can be used to trace out the demand curve and thus identify the demand elasticity α_1. We can make a similar observation regarding the demand shifters helping us identify the supply elasticity β_1. We proceed to identify and estimate these and other parameters systematically.

In what follows, we could take the indirect least squares approach we used previously: namely, we could solve for (Y, P) in terms of exogenous variables and shocks, creating a reduced form and could take the indirect least squares approach to identification and estimation of structural parameters, where we first estimate the reduced form parameters by least squares method, then back out the structural parameters from the reduced form parameters. This is an excellent approach to take, but since we had already done it for a closely related model, we take the opportunity to pursue another approach which will take us directly to the GMM.

The crux of the method lies in the realization that the following moment conditions hold as the result of the assumed orthogonality conditions in (Wright) :

$$\begin{aligned} \mathrm{E}(Y - \alpha_1 P - \alpha_2' Z^d - \alpha_3' W)(Z^{s\prime}, Z^{d\prime}, W')' = 0, \\ \mathrm{E}(Y - \beta_1 P - \beta_2' Z^s - \beta_3' W)(Z^{s\prime}, Z^{d\prime}, W')' = 0. \end{aligned} \tag{1.1}$$

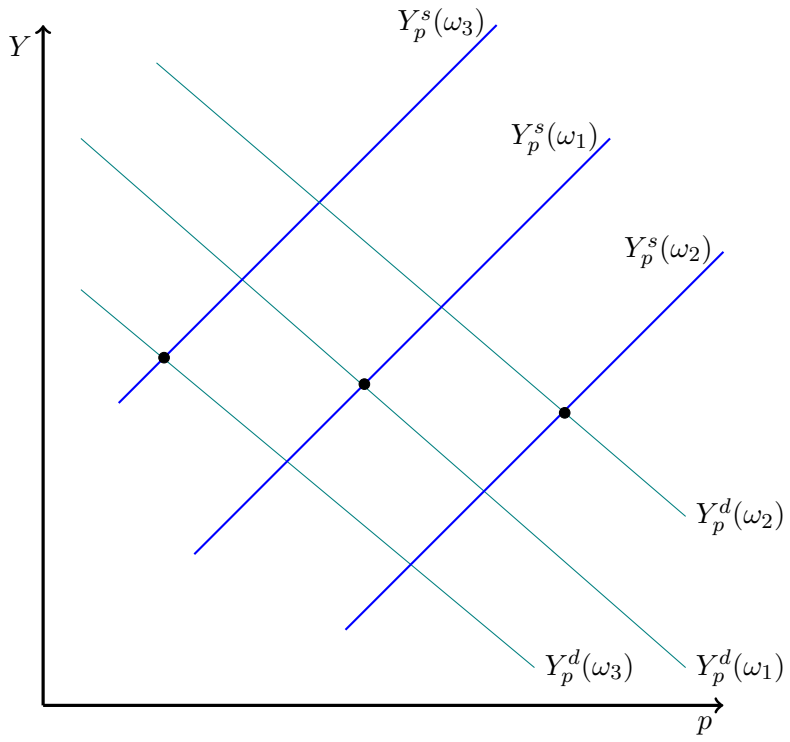

FIGURE 1. Here we see three different realizations of random supply and demand curves. The equilibrium prices (Y, P) determined by the points of intersection of pairs curves $p \mapsto Y_p^s(\omega)$ and $p \mapsto Y_p^d(\omega)$ at a given point in time or in a given market.

This is a system of deterministic equations, where the number of unknown parameter values is smaller than or equal to the number of equations, so there is some hope that the system identifies these parameter values. Therefore we can set-up an empirical analog of these equations and solve the resulting empirical equations, at least approximately, to get a good estimator of the parameter values.

In order to develop the approach systematically, we need to set-up some notation. Let

$$\theta_0 := (\alpha', \beta')' := (\alpha_1, \alpha_2', \alpha_3', \beta_1, \beta_2', \beta_3')',$$

denote the true value of our parameter vector that satisfies (1.1), and let

$$\theta := (a', b')' = (a_1, a_2', a_3', b_1, b_2', b_3')',$$

denote the potential other values it could take in the parameter space Θ. Define random vectors

$$D_1 := (P, Z^{d\prime}, W')', \quad D_2 := (P, Z^{s\prime}, W')',$$

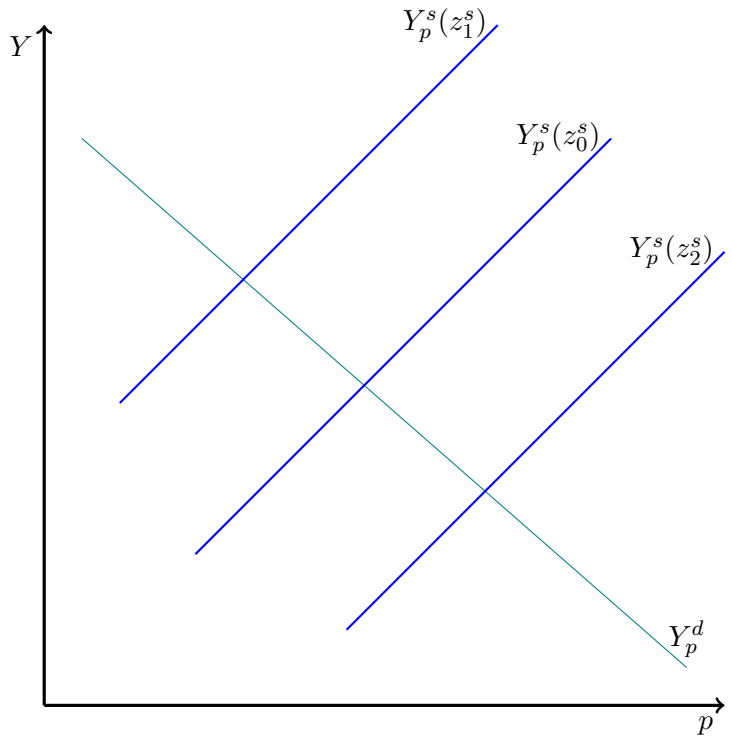

FIGURE 2. Here we observe realizations driven by changes in the observed supply shifters, which shift the supply curve but not the demand curve. The equilibrium prices (Y, P) are determined by points of the intersection of pairs curves $p \mapsto Y_p^s$ and $p \mapsto Y_p^d$. Here changes in the supply shifters allow us to trace out the demand curve.

and let

$$X := (Y, P, Z')', \quad Z := (Z^{d\prime}, Z^{s\prime}, W')',$$

be a vector of observables and a vector of technical instruments.

Define the "score" function:

$$g(X, \theta) := \left[\begin{array}{c} g_1(X, a) \\ g_2(X, b) \end{array} \right] := \left[\begin{array}{c} (Y - D_1' a) Z \\ (Y - D_2' b) Z \end{array} \right].$$

We thus have the moment condition:

$$\mathrm{E} g(X, \theta_0) = \left[\begin{array}{c} \mathrm{E} g_1(X, \alpha) \\ \mathrm{E} g_2(X, \beta) \end{array} \right] = 0.$$

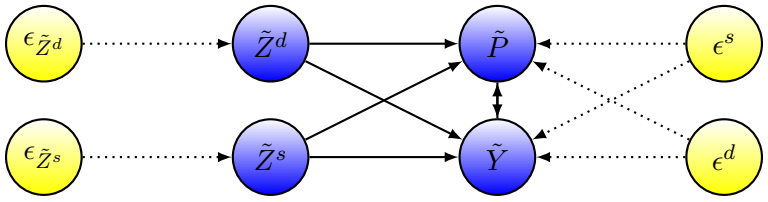

FIGURE 3. Graph-theoretic representation of Wright's model after partialling out the effect of W. Random vectors are given as nodes or vertices of the graph. Observed nodes are shaded and latent nodes are not. Directed edges represent causal channels. The absence of links between latent nodes signifies the lack of correlation among nodes: the instrument shocks $\epsilon_{\tilde{Z}^s}$ and $\epsilon_{\tilde{Z}^d}$ are uncorrelated with structural errors ϵ^s and ϵ^d.

This neat system of moment equations is equivalent to (1.1). The true parameter value $\theta_0 = (\alpha', \beta')'$ is identified by these linear equations if the following matrix

$$G := \mathrm{E}\frac{\partial}{\partial \theta'} g(X, \theta) =: \begin{bmatrix} G_1 & 0 \\ 0 & G_2 \end{bmatrix} = \begin{bmatrix} \mathrm{E}ZD_1' & 0 \\ 0 & \mathrm{E}ZD_2' \end{bmatrix}$$

is full rank. The technical full rank condition on $\mathrm{E}ZD_1'$ and $\mathrm{E}ZD_2'$ in turn entails the requirement that the excluded variables Z^d and Z^s have predictive power for the endogenous variable P (and hence also Y). This generalizes the previous identification condition in L2 of there being a non-trivial "first stage".

We use the empirical analog of this moment condition to set-up estimation. What follows below is no longer specific to our working example and so we begin a new section.

2. GMM

We describe a general framework here. The first building block is the following assumption.

We have a random vector X and a score function $g(X, \theta)$ that is a vector-valued function of X and parameter vector θ. For the moment function

$$g(\theta) := \mathrm{E}g(X, \theta),$$

we assume that the true parameter value $\theta_0 \in \Theta \subset \mathbb{R}^d$ satisfies:

$$g(\theta_0) = 0.$$

This was the case in Wright's model for an *appropriately chosen* score function $g(X, \theta)$, based on instrumental variables. As in Wright's model, we should always make sure that the true parameter value that we want to identify and estimate satisfies the condition above. For this it is important to choose the score functions appropriately.

Note that g maps $\Theta \subseteq \mathbb{R}^d$ to $\mathbb{R}^m$, where $d = \dim(\theta) \leq m$. Econometricians say that the system of equations is

- "exactly identified" if $d = m$,
- "over identified" if $d < m$,
- "under identified" if $d > m$.

This is a conventional terminology, so we need to know it in order to understand our colleagues, although this terminology seems to be misleading to suggest that just having enough equations is sufficient to identify the true parameter value. Perhaps, it would be better to call the first case "exactly determined", the second "overdetermined", and the third "underdetermined". In the last case we can not hope to identify θ_0 but we might still be able to "set-identify" θ_0 to lie in the identified set Θ_0 which is a strict subset of Θ. We don't study this case in this course.

We have data $\{X_i\}_{i=1}^n$, which are identical copies of X, and, as a leading case we assume that they are also independent (i.i.d.). We form the empirical moment function:

$$\hat{g}(\theta) = \mathbb{E}_n g(X, \theta).$$

Then our estimator $\hat{\theta}$ of θ_0 is

$$\hat{\theta} \in \arg\min_{\theta \in \Theta} \ \hat{g}(\theta)' \hat{A} \hat{g}(\theta), \tag{GMM}$$

where $\hat{A}$ is a positive-definite matrix, possibly data-dependent, that converges to a non-stochastic positive-definite matrix A, that is $\hat{A} \to_P A$.

Thus, the estimator $\hat{\theta}$ sets $\hat{g}(\theta)$ close to zero with respect to the quadratic discrepancy function. In "over-identified" systems it is generally not possible to set $\hat{g}(\theta)$ exactly to zero. The choice of quadratic discrepancy to measure deviations from zero, is both convenient and good, since it does deliver an optimal way to combine moments, with an appropriately chosen A. We note that there are other, asymptotically equivalent ways of combining moments: for instance, the methods called generalized empirical likelihood and continuous updating estimators are first-order equivalent to the GMM estimator, and they also have some refined properties in some circumstances.

An important practical and theoretical matter is the choice of the weighting matrix A. It makes sense to choose A in a way that gives more weight to precisely estimated moments and less weight to imprecisely estimated moments, while also taking into account the correlation of the moments.

The optimal weighting matrix A for GMM takes the form

$$A = \Omega^{-1}, \quad \Omega := \operatorname{Var}(\sqrt{n}\hat{g}(\theta_0)),$$

where in the case of i.i.d. data,

$$\Omega = \operatorname{Var} g(X, \theta_0) = \mathrm{E} g(X, \theta_0) g(X, \theta_0)'.$$

We will establish below the precise sense in which $A = \Omega^{-1}$ is optimal. Note that Ω is unknown. For this reason, we often use the following algorithm to compute the GMM.

1. Set $\hat{A} = \hat{\Omega}^{-1}$, for $\hat{\Omega} = I$ (or some other reasonable initialization) and obtain $\hat{\theta}$.
2. Set $\hat{A} = \hat{\Omega}^{-1}$, for $\hat{\Omega} = \widehat{\operatorname{Var}}(\sqrt{n}\hat{g}(\theta))|_{\theta=\hat{\theta}}$, and obtain $\hat{\theta}$.
3. Repeat the previous step several times.

In step 1, we could use other reasonable initializations, for example, given a parameter guess $\bar{\theta}$ we could use $\hat{\Omega} = \widehat{\operatorname{Var}}(\sqrt{n}\hat{g}(\theta))|_{\theta=\bar{\theta}}$ instead, and this would arguably be a more clever choice in some cases. In step 2, we need to specify a variance estimator: under i.i.d. sampling we can use the variance estimator

$$\widehat{\operatorname{Var}}(\sqrt{n}\hat{g}(\theta)) = \mathbb{E}_n g(X_i, \theta) g(X_i, \theta)'.$$

For dependent data, we can use the Newey-West variance estimator. Two steps are sufficient to reach the full efficiency, although an additional step might be desirable to use a more efficient estimator of the variance in an effort to improve the finite-sample properties of the estimator.

3. A Helicopter Tour of GMM Properties

In order for GMM estimator to be consistent for the true parameter values, these values need to be identifiable in the population (that is when infinite amount of data is available). The assumption of identifiability is the following.

We assume that the identification condition holds: $g(\theta) = 0$ if and only if $\theta = \theta_0$.

This is a non-trivial assumption to impose and it will take us some time to understand what makes this assumption hold. We begin to build our understanding with the discussion given below.

We consider the following Jacobian matrix that plays an important role:

$$G = \frac{\partial}{\partial \theta'} g(\theta_0).$$

We assume that the Jacobian matrix G has full column rank.

It turns out that in the linear models, where $\theta \mapsto g(\theta)$ is linear, the full rank assumption on G is equivalent to the identification condition. In non-linear models, this assumption is not equivalent to the identification condition. The latter point is quite delicate, so we postpone a detailed discussion of this point to L4. In linear IV models, like Wright's model, this assumption translates into instruments having a predictive power over endogenous variables. If instruments have no predictive power, then the full rank assumption fails, and θ_0 is no longer identified. This assumption makes intuitive sense since instruments must create quasi-experimental fluctuations in the endogenous variables in order for us to identify the structural parameters.

By way of preview, we would like to note that the GMM estimator is root-n consistent and asymptotically normal under a set of plausible conditions that we shall develop in what follows. Just like in L2 we need to distinguish strongly and weakly identified cases.

We say that we have a strongly identified case if the smallest eigenvalue of $G'AG$ is bounded away from zero. Otherwise, we say that we have a weakly identified case.

In the context of linear models, like Wright's, we can use the F-statistics as diagnostics for weak identification, as discussed in Lecture 2. In general GMM formulation a simple diagnostic of this type has been developed by [2]. We briefly discuss inference under weak identification in GMM in the next lecture.

Assertion 1 (General Properties of GMM under Strong Identification)**.** *Suppose that $g(\theta) = 0$ for $\theta \in \Theta$ if and only if $\theta = \theta_0$. Under strong identification and other plausible*

regularity conditions,

$$\sqrt{n}(\hat{\theta} - \theta_0) \overset{\text{a}}{\sim} (G'AG)^{-1}G'A\epsilon, \quad \epsilon \sim N(0, \Omega)$$

where

$$(G'AG)^{-1}G'A\epsilon \sim N(0, V_A),$$

where

$$V_A := (G'AG)^{-1}G'A\Omega AG(G'AG)^{-1}.$$

If $A = \Omega^{-1}$*, then*

$$V_A = V_{\Omega^{-1}} = (G'\Omega^{-1}G)^{-1}.$$

In order to make the best use of this result and understand it better, let us make several remarks.

First, we note that if $m = d$, that is, in the "exactly identified" case, the weighting matrix vanishes from the variance formula:

$$V_A = G^{-1}\Omega(G')^{-1}.$$

Thus the choice of A is irrelevant in exactly identified cases, which makes a lot of sense, since in this case we solve each equation by equation and it is unimportant how we weigh them together.

If $m > d$, that is, in the "over-identified" case, the weighting matrix does matter and does not vanish from the formula. In particular, as noted above, using weighting matrix Ω^{-1} simplifies the variance matrix dramatically.

Moreover, Ω^{-1} is the *optimal weighting matrix for GMM* in the sense that

$$V_A \geq V_{\Omega^{-1}}$$

for all $A \geq 0$, where the inequality means that $V_A - V_{\Omega^{-1}} \geq 0$, i.e. positive definite. In words, the optimally weighted GMM has smaller variance matrix asymptotically than a suboptimally weighted GMM.

The claim is immediate from the following observation, which can be verified by a simple calculation:

$$0 \leq \operatorname{Var}((G'AG)^{-1}G'A\epsilon - (G'\Omega^{-1}G)^{-1}G'\Omega^{-1}\epsilon) = V_A - V_{\Omega^{-1}}, \tag{3.1}$$

Second, above we assert the following asymptotic representation of the GMM:

$$\sqrt{n}(\hat{\theta} - \theta_0) \overset{\text{a}}{\sim} (G'AG)^{-1}G'A\epsilon,$$

where $\epsilon \sim N(0, \Omega)$ is the asymptotic sampling error, corresponding to $\sqrt{n}\hat{g}(\theta_0)$. This is a very convenient observation because it connects us to the least squares analysis: We can view GMM in large samples as a weighted least squares estimator

$$\tilde{\theta} = \arg\min_{\theta \in \mathbb{R}^d} \ (U - G\theta)' A (U - G\theta),$$

in a "limit experiment", where we observe

$$U = G\theta_0 + \epsilon/\sqrt{n}, \quad \epsilon \sim N(0, \Omega),$$

and we try to learn θ_0. For this experiment, the weighted least squares estimator can be represented as

$$\sqrt{n}(\tilde{\theta} - \theta_0) = (G'AG)^{-1} G'A\epsilon,$$

and the optimal weighted least squares (generalized least squares) estimator is the one that uses $A = \Omega^{-1}$. We know that from the Gauss-Markov theorem that the generalized least squares is optimal. In fact, (3.1) is a quick re-proof of this theorem. Note that the calculation (3.1) also shows that the difference of the variances of the suboptimal and optimal estimators is equal to the variance of the difference of the two estimators.

We can also think of the creating optimal linear combination of moments $\bar{g}(X, \theta) = L'g(X, \theta)$, where L is such that the GMM estimator based on $\bar{g}(X, \theta)$ has the smallest asymptotic variance. Using previous calculations, we can deduce that the optimal linear combination of moments is given by

$$L = G\Omega^{-1}.$$

4. Asymptotic Properties of Linear GMM

The linear case refers to the case where

$$\theta \mapsto g(X, \theta)$$

is an affine map with respect to θ. This means, in particular, that we can write

$$g(X, \theta) = g(X, 0) + G(X)(\theta - 0).$$

For example, the Wright's model has a score function that is affine in this sense.

In the affine case we have that

$$G = \frac{\partial}{\partial \theta'} \mathrm{E} g(X, \theta) = \mathrm{E} G(X).$$

The GMM estimator has the closed form solution in the linear case. For

$$\hat{G} = \mathbb{E}_n G(X), \quad \hat{g}(0) = \mathbb{E}_n g(X, 0),$$

we have that

$$\hat{\theta} = -(\hat{G}'\hat{A}\hat{G})^{-1}\hat{G}'\hat{A}\hat{g}(0), \tag{4.1}$$

provided the pre-factor is invertible.

This claim follows by solving the first-order conditions for the (GMM) problem:

$$0 = \hat{G}'\hat{A}\hat{g}(\hat{\theta}) = \hat{G}'\hat{A}(\hat{g}(0) + \hat{G}\hat{\theta}).$$

The linear case is great because it is analytically tractable and allows us to derive the following formal result.

Theorem 1. *Consider the linear GMM problem where dimensions d and m are fixed, and assume that we have $\hat{A} \to_P A > 0$. Suppose that the law of large numbers holds, namely $\hat{G} \to_P G$, where G is of full column rank. Suppose also that the central limit theorem holds,*

$$\sqrt{n}\hat{g}(\theta_0) \overset{\text{a}}{\sim} N(0, \Omega).$$

Here G, A, and Ω are assumed not to depend on n. Then all the conclusions of Assertion 1 hold. When $(X_i)_{i=1}^{\infty}$ are i.i.d. copies of X, where X does not depend on n, it suffices to have $\mathrm{E}\|G(X)\| < \infty$ for the law of large numbers and $\mathrm{E}\|g(X, \theta_0)\|^2 < \infty$ for the central limit theorem.

Proof. We set-up the first order condition for the GMM problem

$$0 = \hat{G}'\hat{A}\hat{g}(\hat{\theta}).$$

Since $\hat{g}(\hat{\theta}) = \hat{g}(\theta_0) + \hat{G}(\hat{\theta} - \theta_0)$ by linearity

$$0 = \hat{G}'\hat{A}[\hat{g}(\theta_0) + \hat{G}(\hat{\theta} - \theta_0)].$$

As explained below, we can solve these equations for the deviation of the estimator from the estimand:

$$\sqrt{n}(\hat{\theta} - \theta_0) = -(\hat{G}'\hat{A}\hat{G})^{-1}\hat{G}'\hat{A}\sqrt{n}\hat{g}(\theta_0). \tag{4.2}$$

Then we notice that by the assumed law of large numbers and the central limit theorem and the continuous mapping theorem that

$$-(\hat{G}'\hat{A}\hat{G})^{-1}\hat{G}'\hat{A}\sqrt{n}\hat{g}(\theta_0) \overset{\text{a}}{\sim} -(G'AG)^{-1}G'AN(0, \Omega). \tag{4.3}$$

In what follows we explain some details. We have by the full rank assumption and by $A > 0$ that $G'AG > 0$, so that inverse of $G'AG$ exists. We have by the continuous mapping

theorem and the law of large numbers

$$\hat{G} \to_P G, \quad \hat{A} \to_P A, \quad \hat{G}'\hat{A} \to_P G'A, \quad (\hat{G}'\hat{A}\hat{G})^{-1} \to_P (G'AG)^{-1}.$$

Thus $(\hat{G}'\hat{A}\hat{G})^{-1}$ exists with probability converging to 1, which is what enabled us to solve for the deviation in (4.2). The convergence results were combined with $\sqrt{n}\hat{g}(\theta_0) \overset{a}{\sim} N(0, \Omega)$ to arrive at the conclusion (4.3) by the continuous mapping theorem. ■

5. Back to Wright's System of Equations, 2SLS and 3SLS as GMM

We now can see that we can estimate Wright's model by using the score function we have specified in Section 1. In principle we could stop here, but it is very interesting to explore the details of the linear GMM for this model.

In fact Wright's model provides a very interesting framework in which we can take several routes to estimation. Recall that in L2 we had exact identification and all estimation routes discussed led to the same estimator, which we called the IV estimator. In Wright's model, we have over-identification in general, and the set of IV estimators we could consider is quite rich. We can consider all of them as special cases of GMM.

We consider two broad approaches to estimation:

- *limited-information* or *equation-by-equation* approach, where we treat estimation of α and β separately, based on each block of equations; we can view this approach artificially as a joint estimation with a block-diagonal weighting matrix;

- *full-information* or *systems* approach, where we treat estimation of α and β jointly, employing (jointly) optimal weighting matrices.

It turns out that both approaches could be treated as part of the general GMM approach, although the second approach is generally more efficient (if the regularity conditions hold for the joint estimation problem). In what follows we explain the features of the two approaches in details. This material can and should be skipped on the first reading.

- We assume i.i.d. sampling in what follows.

5.1. Limited Information or Equation-by-Equation Approach.

Here we estimate α and β separately, similarly to what we have done in L2. The "separate" estimation of α and β can be carried out by formulating a GMM estimator for each of the two sets of equations:

$$\mathrm{E} g_1(X, \alpha) = 0, \quad \mathrm{E} g_2(X, \beta) = 0.$$

Consider the first set of equation first. Here

$$g_1(X,a) = (Y - D_1'a)Z, \quad G_1(X) = \frac{\partial}{\partial a'} g_1(X,a) = -ZD_1',$$

so that this gives us the following quantities:

$$\hat{g}_1(a) = \mathbb{E}_n(Y_i - D_{1i}'a)Z_i, \quad \hat{g}_1(0) = \mathbb{E}_n Y_i Z_i, \quad \hat{G}_1 = \mathbb{E}_n G_1(X_i) = -\mathbb{E}_n Z_i D_{1i}'.$$

Given a weighting matrix $\hat{A}_1 \to_P A_1$ the explicit solution for the *limited-information GMM estimator* of α is given by

$$\hat{\alpha}_{LI} = -(\hat{G}_1'\hat{A}_1\hat{G}_1)^{-1}\hat{G}_1'\hat{A}_1\hat{g}_1(0).$$

The optimal weighting matrix (in the limited-information sense) and its estimator are

$$A_1 = (\mathrm{E} g_1(X,\alpha)g_1(X,\alpha)')^{-1}, \quad \hat{A}_1 = (\mathbb{E}_n g_1(X,\tilde{\alpha})g_1(X,\tilde{\alpha})')^{-1},$$

where $\tilde{\alpha}$ is some preliminary estimator. It is also important to mention the following (canonical) choice of the weighting matrix and its estimator:

$$A_1^c = (\mathrm{E}ZZ')^{-1}, \quad \hat{A}_1^c = (\mathbb{E}_n ZZ')^{-1}.$$

It turns out that under this choice the estimator $\hat{\alpha}$ becomes the (canonical) *two stage least squares estimator* (2SLS):

$$\hat{\alpha}_{2SLS} = (\mathbb{E}_n D_{1i}Z'(\mathbb{E}_n ZZ')^{-1}\mathbb{E}_n ZD_{1i}')^{-1}\mathbb{E}_n D_{1i}Z_i'(\mathbb{E}_n ZZ')^{-1}\mathbb{E}_n YZ.$$

Under overidentification 2SLS is not optimal in general compared to the estimator $\hat{\alpha}$ that uses optimal weighting matrices.[1] Note however that 2SLS could be used as a preliminary estimator $\tilde{\alpha}$ in the computation of the optimal estimator. In linear IV models, like Wright's model, we can call the optimal estimator $\hat{\alpha}_{LI}$ the limited information *three stage least squares* (3SLS) estimator. The properties of $\hat{\alpha}_{LI}$ and $\hat{\alpha}_{2SLS}$ follow from the general properties of GMM summarized in Assertion 1.

All of the above also works to define the limited information estimator for β. Here

$$g_2(X,b) = (Y - D_2'b)Z, \quad G_2(X) = \frac{\partial}{\partial b'} g_2(X,b) = -ZD_2',$$

so that this gives us the following quantities:

$$\hat{g}_2(b) = \mathbb{E}_n(Y - D_{2i}'b)Z, \quad \hat{g}_2(0) = \mathbb{E}_n YZ, \quad \hat{G}_2 = \mathbb{E}_n G_2(X) = -\mathbb{E}_n ZD_{2i}'.$$

[1] It has some limited-information optimality under homoscedasticity of the structural errors, which we don't use in this course, because such assumptions seem practically irrelevant.

Given the weighting matrix $\hat{A}_2 \to_P A_2$ the explicit solution for GMM is given by

$$\hat{\beta}_{LI} = -(\hat{G}_2' \hat{A}_2 \hat{G}_2)^{-1} \hat{G}_2' \hat{A}_2 \hat{g}_2(0).$$

The optimal weighting matrix (in limited-information sense) is

$$A_2 = (\mathrm{E} g_2(X,\beta) g_2(X,\beta)')^{-1}, \quad \hat{A}_2 = (\mathbb{E}_n g_2(X,\tilde{\beta}) g_2(X,\tilde{\beta})')^{-1},$$

where $\tilde{\beta}$ is a preliminary estimator of β. The canonical choice of the weighting matrix and its estimators are $A_2^c = (\mathrm{E} ZZ')^{-1}$ and $\hat{A}_2^c = (\mathbb{E}_n ZZ')^{-1}$, using which the estimator becomes the (canonical) 2SLS. The properties of $\hat{\beta}_{LI}$ follow from the general properties of GMM summarized in Assertion 1.

It turns out that the joint properties of $\hat{\alpha}_{LI}$ and $\hat{\beta}_{LI}$ could be obtained by stacking the two estimation problems into common GMM estimation problem with

$$g(X,\theta) = [g_1(X,a)', g_2(X,b)']', \quad \hat{g}(\theta) = \mathbb{E}_n g(X,\theta),$$

and

$$\hat{g}(0) = \begin{bmatrix} \hat{g}_1(0) \\ \hat{g}_2(0) \end{bmatrix}, \quad \hat{G} = \begin{bmatrix} \hat{G}_1 & 0 \\ 0 & \hat{G}_2 \end{bmatrix}, \qquad \hat{A} = \begin{bmatrix} \hat{A}_1 & 0 \\ 0 & \hat{A}_2 \end{bmatrix},$$

where the block-diagonal weighting matrix mimics the "separation" of the two problems. Then

$$\hat{\theta}_{LI} = -(\hat{G}' \hat{A} \hat{G})^{-1} \hat{G}' \hat{A} \hat{g}(0) = [\hat{\alpha}_{LI}', \hat{\beta}_{LI}']'.$$

The properties of this estimator, in particular the joint variance matrix, then follow from the general properties of GMM given in Assertion 1.

5.2. **Full-Information or Systems Approach.** Here we start out in the same way as in the last paragraph: We have

$$g(X,\theta) = [g_1(X,a)', g_2(X,b)']', \quad \hat{g}(\theta) = \mathbb{E}_n g(X,\theta),$$

so that, as before,

$$\hat{g}(0) = \begin{bmatrix} \hat{g}_1(0) \\ \hat{g}_2(0) \end{bmatrix}, \quad \hat{G} = \begin{bmatrix} \hat{G}_1 & 0 \\ 0 & \hat{G}_2 \end{bmatrix}$$

but instead of using block-diagonal weighting matrices, we employ the optimal weighting matrix and its estimator:

$$A = (\mathrm{E} g(X,\theta_0) g(X,\theta_0)')^{-1}, \quad \hat{A} = (\mathbb{E}_n g(X,\tilde{\theta}) g(X,\tilde{\theta})'])^{-1},$$

where $\tilde{\theta}$ is a preliminary estimator of θ_0, for example, we could use the limited-information estimators.

Then the full-information GMM estimator is given by

$$\hat{\theta}_{FI} = [\hat{\alpha}_{FI}', \hat{\beta}_{FI}']' = -(\hat{G}' \hat{A} \hat{G})^{-1} \hat{G}' \hat{A} \hat{g}(0).$$

The properties of this estimator then follow from the general properties of GMM summarized in Assertion 1.

In linear IV systems like Wright's model, we can call the optimal GMM estimator the full-information or system *three-stage least squares* (3SLS), reflecting the fact that we are aggregating information from the entire system of equations. In general, in over-identified problems, A does *not* reduce to the block-diagonal structure in the previous subsection, and so by the GMM efficiency theory, the full information approach is generally more efficient than the limited information approach.

At the same time, the full-information approach is less robust than the limited information approach. For example, if we mess up estimation of first block of equations, for example, due to weak identification or misspecification, we generally end up messing up estimation in second block of equations. Some of these problems can be avoided using the limited information approach on the second block of equations. This is one of the main reason why limited information approach is more frequently used in empirical work.

To summarize, the full information or systems approach is more efficient than the limited information or equation approach; at the same time the latter is more robust to departure from strong identification and misspecification.

6. Demand and Supply of Whiting Fish in Fulton Market

We illustrate the estimation and inference of systems of simultaneous equations through a demand and supply empirical application. We construct an artificial data set calibrated to the data set of [1] on demand and supply of whiting fish in the Fulton Fish Market.[2] This is simulated data which we can term bit-data and refer to the commodity in this data as bit-fish. The data include 1,554 daily observations on the log of quantity sold of whiting fish in pounds as Y; the log of the average price in dollars per pound as p; two indicators of the weather conditions on shore (cold and rainy) and four day-of-the-week indicators (day1–day4) as the demand shifters Z^d; and two indicators of the conditions of the sea (stormy and mixed) as the supply shifters Z^s. Note that the system is overidentified with a total of $m = 18$ moment conditions for $d = 12$ parameters. We deem the shifters as strong instruments because the first stage F-statistics are 148.44 for the demand and 19.62 for the supply.

Table 1 reports limited and full information estimates of the demand and supply elasticities α_1 and β_1, together with 95% confidence intervals. In this data set where the error

[2] We artificially increase the sample size by a factor of 14 because the demand shifters are very weak in the original data set with only 111 observations.

TABLE 1. Demand and Supply of Whiting Bit-Fish

		Demand Elasticity		Supply Elasticity	
Estimate	Altern. Name	Estimate	Std. Error	Estimate	Std. Error
OLS	–	-0.50	0.05	-0.32	0.05
LI-GMM (1 step)	"LI-2SLS"	-0.92	0.12	0.94	0.22
LI-GMM (2 step)	"LI-3SLS"	-0.93	0.12	0.86	0.22
FI-GMM(2 step)	"FI-3SLS"	-0.93	0.11	0.91	0.21

terms are homoskedastic by construction, there is no efficiency gain of doing 2 step over 1 step in limited information estimation. The full information estimator, however, is more precise than the limited information estimators because of the over identification and the correlation between the supply and demand shocks.

NOTES

The foundational work of Wright, who was an economist, is quintessential to econometrics – please see a nice article by J. Stock on Wright's work.

APPENDIX A. PROBLEMS

(1) Consider the regression model $Y = X'\theta_0 + \epsilon, \epsilon \perp X$. Analyze estimation of the regression parameter θ_0 from the GMM point of view: write the score function, write the GMM estimator, write its asymptotic distribution. Compare your results to the standard results on OLS estimator of θ_0.

(2) Work through the proof of Theorem 1 several times. Write down the proof by memory, without looking at the notes.

(3) Prove the equality in equation (3.1). State in words what you have proved.

(4) Write down the details of optimal weighting matrix estimation in implementation of GMM for the case of i.i.d. data. How would you adjust it if the data were a time series?

(5) Replicate estimation results for the simulated Fulton fish market data, which is provided. Explain limited information GMM estimators as well as full information GMM.

(6) Provide estimation results for the real Fulton fish market data, which is provided. Explain limited information GMM estimators as well as full information GMM. Notice that this application has weak instruments, so the inferential results drawn using strong-instrument asymptotics is not reliable. You can try Anderson-Rubin approach for inference and see what you get.

References

[1] Joshua D Angrist, Kathryn Graddy, and Guido W Imbens. The interpretation of instrumental variables estimators in simultaneous equations models with an application to the demand for fish. *The Review of Economic Studies*, 67(3):499–527, 2000.

[2] Jonathan H Wright et al. Detecting lack of identification in gmm. *Econometric theory*, 19(2):322–330, 2003.

14.382 Econometrics
Spring 2017