Sequential GMM

Bryan S. Graham, UC - Berkeley & NBER

January 21, 2021

Consider an oligopolistic market consisting of T firms competing on quantity. That is each firm supplies a certain amount of some homogenous product to a common market. For example, Genesove & Mullin (1998) study the East Coast market for refined sugar at the turn of the 20th century. A more contemporary example might be the market for NAND flash memory, where a handful for firms – Hynix, Samsung, KIOXIA (Toshiba), Sandisk, Micron and Intel – dominate global production. Other commodity markets are well described by quantity-based competition. Associated with such markets are a variety of important questions. Are firms engaging in Cournot competition or are they colluding such that they behave like a cartel (cf., Porter, 1983, 2005)? How far do prices depart from marginal costs? What is the nature of the production technology in the industry? How do firms' marginal costs vary with their scale of production, nameplate capacity, R&D expenditures, past cumulative output (as might be important if there is learning by doing), regulatory environments and so on?

Some of these questions require information on firms' cost functions. Unfortunately a firm's marginal costs of production are difficult to directly observe. The type of accounting cost data provided in, say, a firm's balance sheet may not be a reliable indicator of marginal production costs as needed for economic analysis. Conveniently, empirical industrial organization economists have shown that, under certain assumptions, it is possible to recover firms' marginal costs from price and quantity data (see Bresnahan (1989) for a review of early studies of this type). A key insight of this approach is that if firms engage in static profit maximization, then they will choose production levels to equate marginal revenue with marginal cost. Marginal revenue depends on the price elasticity of demand – estimable from market-level price and quantity data – as well as the nature of firm competition (e.g., Cournot vs. cartel). Hence, under assumptions about the nature of competition, marginal revenues, and hence, marginal costs are recoverable.

You have probably already encountered static models of competition in prices and quantity in the context of learning game theory and/or industrial organization (see Tirole (1988) for a classic textbook treatment). While by the standards of contemporary empirical IO, these models are simple, they remain useful in many circumstances. For our purposes they also provide an interesting example around which to organize a review of the generalized method of moments (GMM) approach to estimation. This particular application of GMM will be *sequential* in nature. It is applications of this type which are the focus of this note. Multistep estimation procedures arise frequently in empirical work.

In the first step of a sequential GMM procedure one set of parameters is estimated. Here these parameters consist of those describing the market demand schedule; specifically, the price elasticity of demand. In a second step the demand elasticity is combined with an assumption about firm conduct – Cournot competition or collusion – to infer firms' marginal costs, which are then related to observed firm characteristics (e.g., R&D expenditure). In most applications the parameters estimated in the second step are of primary interest, while those estimated in the first step are of secondary interest; sometimes first step parameters are even called *nuisance* parameters.

Several interesting questions arise in this simple set-up. First, how does sampling uncertainty, about the true values of the first step parameters, affect the precision with which we can learn the second step parameters? Put differentially how much reduction in uncertainty about the second step parameter would be available – quantified in terms of a smaller asymptotic variance – if the first step parameter was known? An answer to this question is essential for constructing accurate standard errors (e.g., Newey, 1984). Moreover it provides useful insights on how to efficiently combine all available information to construct a precise estimate of the target, second step, parameter (e.g., Graham, 2011).

Market setting

Let $i = 1, \ldots, N$ index independent oligopolistic markets. In practice these markets may consist of a cross section of spatially distinct markets or a time series of observations on a single market. In each market information on all $t = 1, \ldots, T_i$ competing firms is observed. Specifically, suppressing the market-level subscript for now, we observe firm t's output, Q_t, as well as a vector of exogenous firm attributes, W_t, that the econometrician believes enter the cost function. Also observed is the market-level price, $\mathbf{P}$, and a vector of market-level demand shifters, $\mathbf{X}$. Market-level quantity, $\mathbf{Q}$, is simply the aggregate of the firm-level quantities: $\mathbf{Q} = \sum_t Q_t$. Below we will also have occasion to use the leave-own-out quantity: $\mathbf{Q}_{-t} = \sum_{s \neq t} Q_s = \mathbf{Q} - Q_t$.

The inverse demand schedule for a (random) market is

$$\mathbf{P}=p\left(\mathbf{Q};\mathbf{X},\mathbf{U}\right), \tag{1}$$

with $\mathbf{U}$ an unobserved demand-shifter.

The cost of firm t supplying quantity Q_t to the market is given by

$$C_t=c\left(Q_t;W_t,V_t\right), \tag{2}$$

where, again, W_t is a set of firm-specific variables entering the cost function (e.g., nameplate capacity, R&D expenditures) and V_t is an unobserved firm-specific cost/productivity shifter. Firms choose quantity Q_t to maximize the profit function (assumed concave)

$$\Pi\left(q_t,\mathbf{Q}_{-t}\right)=p\left(q_t+\mathbf{Q}_{-t};\mathbf{X},\mathbf{U}\right)q_t-c\left(q_t;W_t,V_t\right). \tag{3}$$

The first order condition (FOC) for (static) profit maximization is

$$0=\frac{\partial\Pi\left(Q_t,\mathbf{Q}_{-t}\right)}{\partial q_t}=\frac{\partial p\left(\mathbf{Q};\mathbf{X},\mathbf{U}\right)}{\partial\mathbf{Q}}\left(1+\frac{\partial\mathbf{Q}_{-t}}{\partial Q_t}\right)Q_t+\mathbf{P}-\frac{\partial c\left(Q_t;W_t,V_t\right)}{\partial Q_t}.$$

The term $\frac{\partial\mathbf{Q}_{-t}}{\partial Q_t}$ is sometimes called the firm's conjectural variation (CV); it represents a firm's beliefs about how its competitors will revise their chosen output level given a change in own output. Under Cournot/Nash competition this term is zero since firms' choose output levels to maximize profits taking other firms chosen output levels as fixed. Assuming Cournot conduct yields, after some manipulation, the FOCs

$$\frac{\partial c\left(Q_t;W_t,V_t\right)}{\partial Q_t}=\left(\frac{S_t}{\epsilon}+1\right)\mathbf{P},\ t=1,\ldots,T \tag{4}$$

with $S_t=\frac{Q_t}{\mathbf{Q}}$ firm t's market share and $\epsilon=\frac{\partial\mathbf{Q}}{\partial\mathbf{P}}\frac{\mathbf{P}}{\mathbf{Q}}$ the elasticity of demand. We will proceed under the assumption of Cournot competition in what follows.

Parameterizing the cost function and demand schedule

Consider the following isoelastic firm-specific cost function

$$c\left(Q_t;W_t,V_t\right)=\frac{1}{1+\eta}Q_t^{1+\eta}\exp\left(\lambda+W_t'\delta+V_t\right), \tag{5}$$

such that $\eta>1$ indicates increasing returns-to-scale and $\eta<1$ decreasing returns.

Differentiating (5) to get the marginal cost, taking logs, and substituting into (4) yields

$$\ln\left(\frac{S_t}{\epsilon}+1\right)+\ln\mathbf{P}=\lambda+W_t'\delta+\eta\ln Q_t+V_t. \tag{6}$$

Assume, for now, that the elasticity of demand in each market is known. In such a scenario we might attempt to estimate $\beta=(\lambda,\delta',\eta)'$ by the least squares fit of $\ln\left(\frac{S_{it}}{\epsilon_i}+1\right)+\ln\mathbf{P}_i$ onto a constant, W_{it}, and Q_{it}. This would not yield a consistent estimate of β, however, because more productive firms, those with low values of V_{it}, will produce more output, Q_{it}, such that $\mathbb{C}\left(Q_{it},V_{it}\right)<0$. The OLS fit might lead the econometrician to, for example, erroneously conclude that there exist decreasing returns to scale when, in fact, there are increasing returns.

If we are willing to assume that the observed firm cost-shifters $\{W_{it}\}_{t=1}^{T_i}$ vary independently of unobserved productivity $\{V_{it}\}_{t=1}^{T_i}$, then we can use the leave-own-out aggregate $\mathbf{W}_{i,-t}=\sum_{s\neq t}W_{is}$ as an excluded instrumental variable (IV) for $\ln Q_{it}$ in the linear IV fit of of $\ln\left(\frac{S_{it}}{\epsilon_i}+1\right)+\ln\mathbf{P}_i$ onto a constant, W_{it}, and Q_{it}.

Since the elasticity of demand is unknown, we must replace it with a consistent estimate to form a feasible version of the estimation procedure described above. To this end assume that the market demand schedule takes the form.

$$\ln\mathbf{Q}^D\left(\mathbf{p}\right)=\kappa+\epsilon\ln\mathbf{p}+\mathbf{X}'\gamma+\mathbf{U}. \tag{7}$$

Since $\mathbf{P}$, the market clearing price, will not vary independently of the unobserved demand shock, $\mathbf{U}$, in equilibrium we again require an instrumental variable. Let $\mathbf{Z}$ be a market-level supply-shifter which varies independently of $\mathbf{U}$. The linear instrumental variables fit of $\ln\mathbf{Q}_i$ onto a constant, $\ln\mathbf{P}_i$ and $\mathbf{X}_i$ with $\mathbf{Z}_i$ serving as an excluded instrument for $\ln\mathbf{P}_i$ provides a consistent estimate of $\alpha=(\kappa,\epsilon,\gamma')'$. Here α is our first-step parameter vector, while β is our second step one. Note that the log-log demand schedule specification implies that the price-elasticity of demand is constant. This may not be realistic in practice. It also has the implication of a rather strong relationship between observed firm market shares and unobserved firm marginal costs. In practice we might want to fit a more flexible demand schedule model.

Sequential GMM formulation

Let $\mathbf{U}_i\left(\alpha\right)=\ln\mathbf{Q}_i-\kappa-\epsilon\ln\mathbf{P}_i-\mathbf{X}_i'\gamma$, such that the actual unobserved market-level demand shock equals $\mathbf{U}_i\left(\alpha\right)$ at $\alpha=\alpha_0$ (the "0" subscript denotes the true/population parameter value). Similarly, let $V_{it}(\alpha,\beta)=\ln\left(\frac{S_{it}}{\epsilon}+1\right)+\ln\mathbf{P}_i-\lambda-W_{it}'\delta-\eta\ln Q_{it}$, such that the unobserved firm-specific cost-shifter equals $V_{it}(\alpha,\beta)$ when $\alpha=\alpha_0$ and $\beta=\beta_0$. Next define the following pair of moment functions

$$\underset{J\times 1}{\psi_{1i}}(\alpha)=\mathbf{U}_i(\alpha)\begin{pmatrix}1\\ \mathbf{Z}_i\\ \mathbf{X}_i\end{pmatrix} \tag{8}$$

$$\underset{K\times 1}{\psi_{2i}}(\alpha,\beta)=\sum_{t=1}^{T_i}V_{it}(\alpha,\beta)\begin{pmatrix}1\\ W_{it}\\ \mathbf{W}_{i,-t}\end{pmatrix}. \tag{9}$$

Under the assumptions outlined above we have the *unconditional* moment restrictions

$$\mathbb{E}\left[\begin{array}{c}\psi_{1i}(\alpha_0)\\ \psi_{2i}(\alpha_0,\beta_0)\end{array}\right]=0. \tag{10}$$

The expectation in (10) is across *independent* markets. Our asymptotic sampling experiment will involve observing an increasing number of (independent) markets $i=1,\ldots,N$ each consisting of a finite number of firms. Our set-up makes no assumptions about the structure of dependence across unobservables within the same market. We allow, for example, V_{it} and V_{is} to covary, as might occur in the presence of market-specific productivity shocks. By defining the second moment function at the market-level – by summing over firms with in it – we implicitly allow for within-market dependence across $V_{i1},\ldots,V_{iT_i}$. The standard errors that will be suggested by our textbook GMM analysis will be, for example, “clustered” by default/design.

Actually our assumptions imply the stronger pair of *conditional* moment restrictions

$$\begin{aligned}\mathbb{E}\left[\mathbf{U}(\alpha_0)\middle|\,\mathbf{X}_i,\mathbf{Z}_i\right]=&0 \qquad (11)\\ \mathbb{E}\left[\sum_{t=1}^{T_i}V_{it}(\alpha_0,\beta_0)\middle|\,W_{i1},\ldots,W_{iT_t}\right]=&0.\end{aligned}$$

The study of how to efficiently use all the information contained in (11) in order to estimate $\theta_0=(\alpha_0',\beta_0')'$ is a classic question in econometrics (e.g., Chamberlain, 1987); with open questions remaining (e.g., Hristache & Patilea, 2016). We will study efficient estimation of conditional moment problems later in this course.

To keep things simple assume that $\dim(\psi_{1i}(\alpha_0))=\dim(\alpha_0)=J$ and $\dim(\psi_{2i}(\alpha_0,\beta_0))=\dim(\beta_0)=K$. The model is just-identified. We estimate θ_0 in two steps. In step one we find $\hat{\alpha}$ as the solution to

$$\frac{1}{N}\sum_{i=1}^{N}\psi_{1i}\left(\hat{\alpha}\right)=0.$$

We then plug $\hat{\alpha}$ into $\psi_{2i}\left(\hat{\alpha},\beta\right)$ and in a second step find $\hat{\beta}$ as the solution to

$$\frac{1}{N}\sum_{i=1}^{N}\psi_{2i}\left(\hat{\alpha},\hat{\beta}\right)=0.$$

Under just-identification this two step procedure is numerically identical to the joint GMM estimator that solves both sample moments simultaneously. Standard GMM theory applies (e.g., Wooldridge, 2010; Newey & McFadden, 1994). The Jacobian matrix of this joint GMM estimator is

$$\frac{1}{N}\sum_{i=1}^{N}\left[\begin{array}{cc}\frac{\partial\psi_{1i}(\hat{\alpha})}{\partial\alpha'} & 0\\ \frac{\partial\psi_{2i}\left(\hat{\alpha},\hat{\beta}\right)}{\partial\alpha'} & \frac{\partial\psi_{2i}\left(\hat{\alpha},\hat{\beta}\right)}{\partial\beta'}\end{array}\right]\xrightarrow{p}\left[\begin{array}{cc}\Gamma_{1\alpha} & 0\\ \Gamma_{2\alpha} & \Gamma_{2\beta}\end{array}\right]=\Gamma_{0},$$

while the variance-covariance of the combined moment vector $\psi_i\overset{def}{\equiv}\psi_i\left(\alpha_0,\beta_0\right)=\left(\psi_i\left(\alpha_0\right)',\psi_{2i}\left(\alpha_0,\beta_0\right)'\right)'$ is

$$\mathbb{E}\left[\psi_i\psi_i'\right]=\Omega_0=\left[\begin{array}{cc}\mathbb{E}\left[\psi_{1i}\psi_{1i}'\right] & \mathbb{E}\left[\psi_{1i}\psi_{2i}'\right]\\ \mathbb{E}\left[\psi_{2i}\psi_{1i}'\right] & \mathbb{E}\left[\psi_{2i}\psi_{2i}'\right]\end{array}\right]=\left[\begin{array}{cc}\Omega_{11} & \Omega_{12}\\ \Omega_{12}' & \Omega_{22}\end{array}\right].$$

Where we defined $\psi_i\overset{def}{\equiv}\psi_i\left(\alpha_0,\beta_0\right)$, $\psi_{1i}\overset{def}{\equiv}\psi_{1i}\left(\alpha_0\right)$ and $\psi_{2i}\overset{def}{\equiv}\psi_{2i}\left(\alpha_0,\beta_0\right)$ (we will often assume that, when the parameters of a function are omitted in the notation, that the function is evaluated at the true population parameter values).

A standard argument then gives, under regularity conditions,

$$\sqrt{N}\left(\begin{array}{c}\hat{\alpha}-\alpha_0\\ \hat{\beta}-\beta_0\end{array}\right)\xrightarrow{p}N\left(0,\Lambda_0\right),\ \Lambda_0=\Gamma_0^{-1}\Omega_0\Gamma_0^{-1\prime}.$$

See, for example, Newey & McFadden (1994).

The inferential consequences of sequential estimation

We are interested in the form of the asymptotic sampling variance for $\hat{\beta}$, our estimate of the target parameter of interest. The partitioned inverse formula yields an inverse Jacobian matrix of

$$\left[\begin{array}{cc}\Gamma_{1\alpha} & 0\\ \Gamma_{2\alpha} & \Gamma_{2\beta}\end{array}\right]^{-1}=\left[\begin{array}{cc}\Gamma_{1\alpha}^{-1} & 0\\ -\Gamma_{2\beta}^{-1}\Gamma_{2\alpha}\Gamma_{1\alpha}^{-1} & \Gamma_{2\beta}^{-1}\end{array}\right],$$

and hence, after tedious multiplication, a partition of the variance-covariance matrix Λ_0 equal to

$$\Lambda_0 = \left[\begin{array}{cc} \Gamma_{1\alpha}^{-1}\Omega_{11}\Gamma_{1\alpha}^{-1\prime} & \Gamma_{1\alpha}^{-1}\Omega_{12}\Gamma_{2\beta}^{-1\prime} - \Gamma_{1\alpha}^{-1}\Omega_{11}\Gamma_{1\alpha}^{-1\prime}\Gamma_{2\alpha}^{\prime}\Gamma_{2\beta}^{-1\prime} \\ \Gamma_{2\beta}^{-1}\Omega_{12}^{\prime}\Gamma_{1\alpha}^{-1\prime} - \Gamma_{2\beta}^{-1}\Gamma_{2\alpha}\Gamma_{1\alpha}^{-1}\Omega_{11}\Gamma_{1\alpha}^{-1\prime} & \Gamma_{2\beta}^{-1}\Gamma_{2\alpha}\Gamma_{1\alpha}^{-1}\Omega_{11}\Gamma_{1\alpha}^{-1\prime}\Gamma_{2\alpha}^{\prime}\Gamma_{2\beta}^{-1\prime} - \Gamma_{2\beta}^{-1}\Omega_{12}^{\prime}\Gamma_{1\alpha}^{-1\prime}\Gamma_{2\alpha}^{\prime}\Gamma_{2\beta}^{-1\prime} - \Gamma_{2\beta}^{-1}\Gamma_{2\alpha}\Gamma_{1\alpha}^{-1}\Gamma_{2\beta}^{-1}\Omega_{12}\Gamma_{2\beta}^{-1\prime} + \Gamma_{2\beta}^{-1}\Omega_{22}\Gamma_{2\beta}^{-1\prime} \end{array} \right]. \tag{12}$$

Let $\phi_{\beta i}$ denote the influence function for $\hat{\beta}$. Inspection of the lower-right-hand block of Λ_0 indicates that

$$\phi_{\beta i} = \Gamma_{2\beta}^{-1}\left\{\psi_{2i} - \Gamma_{2\alpha}\Gamma_{1\alpha}^{-1}\psi_{1i}\right\}. \tag{13}$$

Imagine instead, that α_0 was known and we estimated β_0 by $\tilde{\beta}$, the solution to the second step moment with α set equal to its population value:

$$\frac{1}{N}\sum_{i=1}^{N}\psi_{2i}\left(\alpha_0, \tilde{\beta}\right) = 0.$$

The influence function for this estimate is $\tilde{\phi}_{\beta i} = \Gamma_{2\beta}^{-1}\psi_{2i}$. Comparing this "oracle" influence function with (13) indicates that the second term in (13) reflects the influence of the first step estimation of α_0 on the asymptotic sampling properties of $\hat{\beta}$. A default intuition is that the effect of first step estimation is to increase sampling variance. That is we expect that $\mathbb{V}\left(\sqrt{N}\left(\hat{\beta} - \beta_0\right)\right) \geq \mathbb{V}\left(\sqrt{N}\left(\tilde{\beta} - \beta_0\right)\right)$, where a "$\geq$ " relationship between two matrices denotes that their difference is positive semi-definite. In fact the situation is a bit more complicated. This arises because ψ_{1i} and ψ_{2i} may covary.

Consider the multivariate regression of ψ_{2i} onto ψ_{1i} :

$$\begin{aligned} \mathbb{E}^*\left[\psi_{2i}\middle|\,\psi_{1i}\right] &= \mathbb{E}\left[\psi_{2i}\psi_{1i}^{\prime}\right] \times \mathbb{E}\left[\psi_{1i}\psi_{1i}^{\prime}\right]^{-1}\psi_{1i} \\ &= \Omega_{12}^{\prime}\Omega_{11}^{-1}\psi_{1i}. \end{aligned}$$

By the properties of multivariate linear predictors we have that $\mathbb{C}\left(\psi_{2i} - \Omega_{12}^{\prime}\Omega_{11}^{-1}\psi_{1i}, \psi_{1i}^{\prime}\right) = 0$. This implies the following orthogonal decomposition of the influence function (13):

$$\phi_{\beta i} = \Gamma_{2\beta}^{-1}\left\{\left[\psi_{2i} - \Omega_{12}^{\prime}\Omega_{11}^{-1}\psi_{1i}\right] - \left[\Gamma_{2\alpha}\Gamma_{1\alpha}^{-1} - \Omega_{12}^{\prime}\Omega_{11}^{-1}\right]\psi_{1i}\right\}. \tag{14}$$

Next observe that the variance of the first term in (14) is

$$\Gamma_{2\beta}^{-1}\left[\Omega_{22}-\Omega_{12}\Omega_{11}^{-1}\Omega_{12}\right]\Gamma_{2\beta}^{-1\prime}\leq\Gamma_{2\beta}^{-1}\Omega_{22}\Gamma_{2\beta}^{-1\prime}$$

implying a variance reduction relative to the oracle estimator $\tilde{\beta}$. This variance reduction arises because, even in the case where α_0 is known, it would be efficient to use ψ_{1i} to reduce the variance of ψ_{2i}. This is related to GMM estimation with auxiliary moments, which we will study next (e.g., Imbens & Lancaster, 1994). It is also closely related to the theory of efficient estimation of expectations (e.g., Brown & Newey, 1998). The second term in (14) captures the penalty associated with imperfect knowledge of the nuisance parameter α_0. This term is a consequence of sampling error in $\hat{\alpha}$ and leads to an increase in the variance of $\hat{\beta}$.

There are two cases of special interest. In the first case

$$\Gamma_{2\alpha}=0, \tag{15}$$

such that Λ_0 is block diagonal

$$\Lambda_0=\left(\begin{array}{cc}\Gamma_{1\alpha}^{-1}\Omega_{11}\Gamma_{1\alpha}^{-1\prime} & 0\\ 0 & \Gamma_{2\beta}^{-1}\Omega_{22}\Gamma_{2\beta}^{-1\prime}\end{array}\right)$$

and $\phi_{\beta i}=\tilde{\phi}_{\beta i}=\Gamma_{2\beta}^{-1}\psi_{2i}$. This condition implies that we can, in a certain sense, completely ignore the first step of our estimation procedure when conducting inference. This condition represents a strong lack of sensitivity of our identifying moment to the value of α. Specifically we have the first order approximation:

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}\psi_{2i}\left(\alpha,\beta_0\right) &\approx \frac{1}{N}\sum_{i=1}^{N}\psi_{2i}\left(\alpha_0,\beta_0\right)+\Gamma_{2\alpha}\left(\alpha-\alpha_0\right)\\ &\approx \frac{1}{N}\sum_{i=1}^{N}\psi_{2i}\left(\alpha_0,\beta_0\right),\end{aligned}$$

for α not "too far" from α_0.

Condition (15) does arise is some important settings; most famously in the context of estimating so-called optimal instrumental variables (we'll study this case later in the context of our discussion of efficiency in conditional moment models). A virtue of (15) is that it is very easy to check. If it holds, you are free to ignore the effects of first step estimation when conducting inference. If it fails, then the safe thing to do is to base inference on a consistent estimate of the full variance-covariance matrix (12).

A second interesting case arises when the following equality holds

$$\Omega_{12} = \Omega_{11}\Gamma_{1\alpha}^{-1\prime}\Gamma_{2\alpha}^{\prime}. \tag{16}$$

Under this condition we have

$$\Lambda_0 = \begin{pmatrix} \Gamma_{1\alpha}^{-1}\Omega_{11}\Gamma_{1\alpha}^{-1\prime} & 0 \\ 0 & \Gamma_{2\beta}^{-1}\left[\Omega_{22} - \Omega_{12}\Omega_{11}^{-1}\Omega_{12}\right]\Gamma_{2\beta}^{-1\prime} \end{pmatrix}.$$

As with the first case, there is asymptotic independence between the first and second step parameters – so called block diagonality – but now the two-step estimator doesn't have the same asymptotic variance for $\hat{\beta}$ as the oracle estimate $\tilde{\beta}$. Condition (16) is harder to check, but it does arise is some important examples. Perhaps most famously in the context of program evaluation estimators using the propensity score. Early on this caused some confusion – or more accurately was a puzzle – as it seemed to suggest that using the true propensity score was worse – in terms of asymptotic precision – than using an estimated one (e.g., Wooldridge, 2007). The resolution to this puzzle is that $\mathbb{E}\left[\psi_{1i}\left(\alpha_0\right)\right] = 0$ does more than just identify the nuisance parameter, α_0; this condition can be used to construct a more efficient estimate of $\mathbb{E}\left[\psi_{2i}\left(\alpha_0, \beta_0\right)\right] = 0$ (and hence of β_0). This is because ψ_{1i} and ψ_{2i} covary. This idea is developed formally in Graham (2011). We'll develop this point more full when we discuss GMM estimation with auxiliary moments. Prokhorov & Schmidt (2009) have an extensive discussion of block diagonality in sequential GMM estimation problems.

Returning to our Cournot oligopoly example, a simple derivative calculation yields

$$\underset{K\times J}{\Gamma_{2\alpha}} = \left[\begin{array}{ccc} 0 & -\mathbb{E}\left[\sum_{t=1}^{T_i}\frac{1/\epsilon}{1+\epsilon/S_t}\begin{pmatrix} 1 \\ W_{it} \\ \mathbf{W}_{i,-t}\end{pmatrix}\right] & 0 \end{array}\right],$$

indicating a lack of orthogonality between the two steps of our estimation procedure. Correct standard errors could be based on a consistent estimate of (12).

Further reading

Characterizing the asymptotic sampling variance of multi-step estimation procedures is a common problem faced by empirical researchers. You may have already encountered the problem of "generated regressors" in your own, or a classmate's, work (for example). Pierce

(1982) and Newey (1984) are key early references. Condition (15) has come into new prominence in the recent literature on orthogonal machine learning, where it has been re-branded "Neyman Orthogonality". One insight from this literature, itself drawn from earlier work on semiparametrically efficient estimation (e.g., Newey, 1990), is that it can be advantageous to base estimation of the target parametric of interest on an estimate of the "first-step corrected influence" function, (13) above. In a large class of problems, this influence function is, in fact, the efficient influence function for β_0.

Consider the following two step estimator. In step one we estimate α_0 as before:

$$\frac{1}{N}\sum_{i=1}^{N}\psi_{1i}\left(\hat{\alpha}\right)=0.$$

Let $\hat{\Gamma}_{2\alpha}\left(\beta\right)=\frac{1}{N}\sum_{i=1}^{N}\frac{\partial\psi_{2i}(\hat{\alpha},\beta)}{\partial\alpha'}$ to be the natural analog estimate of the Jacobian matrix of the second step moment with respect to the first step parameter, α, *written as a function of the target parameter, β.* Similarly define $\hat{\Gamma}_{1\alpha}=\frac{1}{N}\sum_{i=1}^{N}\frac{\partial\psi_{1i}(\hat{\alpha})}{\partial\alpha'}$ and $\hat{\Gamma}_{2\beta}\left(\beta\right)=\frac{1}{N}\sum_{i=1}^{N}\frac{\partial\psi_{2i}(\hat{\alpha},\beta)}{\partial\beta'}$ Next construct the following analog estimate of the influence function for $\hat{\beta}$ given in equation (13) above.

$$\hat{\phi}_{\beta i}\left(\beta\right)=\hat{\Gamma}_{2\beta}^{-1}\left(\beta\right)\left\{\psi_{2i}\left(\hat{\alpha},\beta\right)-\hat{\Gamma}_{2\alpha}\left(\beta\right)\hat{\Gamma}_{1\alpha}^{-1}\psi_{1i}\left(\hat{\alpha}\right)\right\}.$$

In step two we choose $\check{\beta}$ to set the sample mean of this estimated influence function equal to zero

$$\frac{1}{N}\sum_{i=1}^{N}\hat{\phi}_{\beta i}\left(\check{\beta}\right)=0.$$

Next observe that

$$\begin{aligned}\mathbb{E}\left[\frac{\partial\phi_{\beta i}\left(\alpha_0,\beta_0\right)}{\partial\alpha}\right] &= \Gamma_{2\beta}^{-1}\Gamma_{2\alpha}-\Gamma_{2\beta}^{-1}\Gamma_{2\alpha}\Gamma_{1\alpha}^{-1}\Gamma_{1\alpha}\\ &= \Gamma_{2\beta}^{-1}\Gamma_{2\alpha}-\Gamma_{2\beta}^{-1}\Gamma_{2\alpha}=0.\end{aligned}$$

This suggests that if we base our second step on (an estimate of) the "adjusted" moment function $\phi_{\beta i}\left(\alpha,\beta\right)$, then we don't need to make further corrections for first step estimation error.[1] It turns out that this is very advantageous in settings where the first step parameter α_0 is high dimensional and estimated by machine learning methods (which may have complicated asymptotic properties and require tuning parameter choices). The main idea is that we can reduce sensitivity of our estimate of β_0 to first step estimation error in $\hat{\alpha}$ by working with an estimating equation which is orthogonal to this error. See Chernozhukov

[1] It turns out we can also ignore the estimation error in the Jacobian matrices appearing in $\hat{\phi}_{\beta i}\left(\beta\right)$.

et al. (2017) for discussion and several important examples.

References

Bresnahan, T. F. (1989). *Handbook of Industrial Organization*, volume 2, chapter Empirical studies of industries with market power, (pp. 1011 – 1057). North-Holland: Amsterdam.

Brown, B. W. & Newey, W. K. (1998). Efficient semiparametric estimation of expectations. *Econometrica*, 66(2), 453 – 464.

Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics*, 34(3), 305 – 334.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2017). Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal*, 21(C1 - C68).

Genesove, D. & Mullin, W. P. (1998). Testing static oligopoly models: conduct and cost in the sugar industry, 1890-1914. *Rand Journal of Economics,*, 29(2), 355 – 377.

Graham, B. S. (2011). Efficiency bounds for missing data models with semiparametric restrictions. *Econometrica*, 79(2), 437 – 452.

Hristache, M. & Patilea, V. (2016). Semiparametric efficiency bounds for conditional moment restriction models with different conditioning variables. *Econometric Theory*, 32(4), 917 – 946.

Imbens, G. W. & Lancaster, T. (1994). Combining micro and macro data in microeconometric models. *Review of Economic Studies*, 61(4), 665 – 680.

Newey, W. K. (1984). A method of moments interpretation of sequential estimators. *Economics Letters*, 14(2-3), 201 – 206.

Newey, W. K. (1990). Semiparametric efficiency bounds. *Journal of Applied Econometrics*, 5(2), 99 – 135.

Newey, W. K. & McFadden, D. (1994). *Handbook of Econometrics*, volume 4, chapter Large sample estimation and hypothesis testing, (pp. 2111 – 2245). North-Holland: Amsterdam.

Pierce, D. A. (1982). The asymptotic effect of substituting estimators for parameters in certain types of statistics. *Annals of Statistics*, 10(2), 475 – 478.

Porter, R. H. (1983). A study of cartel stability: the joint executive committee, 1880-1886. *Bell Journal of Economics*, 14(2), 301 – 314.

Porter, R. H. (2005). Detecting collusion. *Review of Industrial Organization*, 26(2), 147 – 167.

Prokhorov, A. & Schmidt, P. (2009). Gmm redundancy results for general missing data problems. *Journal of Econometrics*, 151(1), 47 – 55.

Tirole, J. (1988). *The Theory of Industrial Organization*. Cambridge, MA: The MIT Press.

Wooldridge, J. M. (2007). Inverse probability weighted estimation for general missing data problems. *Journal of Econometrics*, 141(2), 1281 – 1301.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press, 2nd edition.