University of Illinois
Fall 2016

Department of Economics
Roger Koenker

Economics 536

Lecture 18

Incidental Parameters and Dynamic Bias in Panel Data

In 1948 Jerzy Neyman and Elizabeth Scott published an *Econometrica* paper called "Consistent Estimates Based on Partially Consistent Observations."[1] It introduced some puzzling examples of situations in which the MLE delivers inconsistent estimates of what they called a structural parameter. As Tony Lancaster remarks in his valuable survey of the subsequent literature, their examples continue to puzzle. In its simplest form the Neyman and Scott incidental parameter problem can be formulated as,

$$y_{it} = \alpha_i + u_{it} \quad t = 1, 2, \cdots T, \quad i = 1, 2, \cdots n,$$

with u_{it} iid $\mathcal{N}(0, \sigma^2)$. The MLE of the structural parameter, σ^2, is therefore,

$$\hat{\sigma}^2 = (nT)^{-1} \sum \sum (y_{it} - \bar{y}_i)^2 \sim \sigma^2 \chi^2_{n(T-1)}/Tn$$

but $\mathbb{E}\hat{\sigma}^2 = \sigma^2(T-1)/T$ and is therefore inconsistent, unless $T \to \infty$ with n. Elaborating on this example, Neyman and Scott also consider the case that $\mathbb{V}(u_{it}) = \sigma_i^2$ differ with i, so we have incidental variance parameters as well as means. This model brings us into close proximity to the empirical Bayes methods introduced by Robbins a few years later. Of course in its simplest form it is easy to fix the inconsistency described above, however we will see that it is more of a challenge to deal with similar problems introduced by consideration of dynamics in closely related panel data models to which we now turn.

In the previous lecture we found that (apparently) $\hat{\beta}_w$ was *safe* in the sense that it provided a consistent estimate of the parameters as $T \to \infty$ and $n \to \infty$ regardless of whether there was correlation between individual effects and the included explanatory variables. The situation is less comforting when T is fixed and $n \to \infty$ as we might view as typical in many econometric panel data problems. (expanding n is relatively easy, expanding T is usually not.) Chamberlain (1980) and Nickell (1981) consider the following model:

$$y_{it} = \gamma y_{it-1} + \alpha_i + u_{it}$$

[1] The paper was the only paper in the first issue of volume 16.

then the within estimator is

$$\begin{aligned}\hat{\gamma}_w &= \frac{\sum\sum(y_{it}-\bar{y}_i)(y_{it-1}-\bar{y}_{i,-1})}{\sum\sum(y_{i,t-1}-\bar{y}_{i,-1})^2} = \sum\sum w_{it}(y_{it}-\bar{y}_i) \\ &= \gamma + \sum\sum(u_{it}-\bar{u}_i)w_{it}\end{aligned}$$

repeatedly substituting we have,

$$y_{it} = u_{it} + \gamma u_{it-1} + \cdots + \gamma^{t-1}u_{it} + \gamma^t y_{0i} + \frac{1-\gamma^t}{1-\gamma}\alpha_i$$

so

$$\begin{aligned}\sum_{t=1}^{T} y_{it-1} &= (1+\gamma+\gamma^2+\ldots+\gamma^{T-1})y_{0i} \quad + \left(\frac{T-1-T\gamma+\gamma^T}{(1-\gamma)^2}\right)\alpha_i \\ &\quad + \frac{1-\gamma^{T-1}}{1-\gamma}u_{i1} + \frac{1-\gamma^{T-2}}{1-\gamma}u_{i2} + \ldots + u_{i,T-1}.\end{aligned}$$

Similar computations for the denominator of $\hat{\gamma}_w$ eventually yield

$$\hat{\gamma}_w \to \gamma - \underbrace{\frac{1-\gamma}{T-1}\left(1 - T^{-1}\frac{1-\gamma^T}{1-\gamma}\right)\left\{1 - \frac{2\gamma}{(1-\gamma)(T-1)}\left[1 - \frac{1-\gamma^T}{T(1-\gamma)}\right]^{-1}\right\}}_{\text{bias}}$$

Thus for example we have in the simple version of the model with $\gamma = .5$,

$T = 2$	ABias $= -(1+\gamma)/2$	$-3/4$
$T = 3$	ABias $= -(2+\gamma)(1+\gamma)/(2(3+\gamma))$	$-.53$
$T = 10$		$-.16$

These asymptotic biases are obviously very large relative to the true $\gamma = .5$.

Yet Another GMM Interlude

There is an emerging consensus that the best approach to dealing with the problems we have just seen in dynamic panel data models is based on generalized method of moments (GMM) methods. We are already familiar with many important examples of GMM, although we may not have explicitly recognized this. For example, my usual simplified derivation of OLS and IV estimators proceeds by first assuming we have an orthogonality condition between observed x's and unobserved u's, and then impose this orthogonality on the sample to get OLS:

$$EX^\top u = 0 \Rightarrow X^\top \hat{u} = 0 \Rightarrow \hat{\beta} = (X^\top X)^{-1}X^\top y$$

In the instrumental variables version of this, X isn't orthogonal to u, but we have exactly the right number of IV's, say Z, and we obtain

$$EZ^\top u = 0 \Rightarrow Z^\top \hat{u} = 0 \Rightarrow \hat{\beta} = (Z^\top X)^{-1}Z^\top y$$

and finally, if we have too many IV's we would like to impose the orthogonality condition $EZ^\top u = 0$ on the sample, but $Z^\top \hat{u} = 0$ in this case is expecting us to solve $q > p$ equations in only p unknowns, which is not generally feasible, so we need a new idea.

One approach which suggests itself is to minimize the length of the vector $Z^\top \hat{u}$. This sounds reasonable and is also suggested by least squares ideas, so we would solve

$$\min_b \hat{u}(b)^\top Z Z^\top \hat{u}(b)$$

which yields

$$\hat{\beta} = (X^\top Z Z^\top X)^{-1} X^\top Z Z^\top y$$

What is wrong with this? What is missing if we want to get 2SLS? How do we rationalize the 2SLS choice

$$\hat{\beta} = (X^\top P_Z X)^{-1} X^\top P_Z y$$

Well, let's work backward. We see immediately that if we had minimized instead,

$$\min_b \hat{u}(b)^\top Z (Z^\top Z)^{-1} Z^\top \hat{u}(b),$$

we would get 2SLS, does this make any particular sense? Maybe.

Suppose we had something like the following

$$M(\theta) \rightsquigarrow \mathcal{N}(0, V)$$

for example $M(\theta) = M\theta$, and we wanted to estimate θ with V known. What would we do? What might be the argument for solving

$$\min_\theta M(\theta)^\top V^{-1} M(\theta)?$$

Suppose, first that V were diagonal, then this would weight the coordinates so that they all had χ^2_1 behavior. A better, more general, idea would be to say "let's think of this as nonlinear regression." The model is then,

$$y_i = M_i(\theta) + v_i \qquad\qquad i = 1, \ldots, P$$

where $Evv^\top = V$, so the GLS estimator minimizes the weighted sum of squares.

Now in the 2SLS context we need to compute $V = V(Z^\top u)$. This is easy if we assume that $E(uu^\top | Z) = \sigma^2 I$ as usual, then we get

$$V = V(Z^\top u) = E(Z^\top u u^\top Z) = \sigma^2 Z^\top Z,$$

so we do indeed get back to 2SLS, by taking this route. Note that if $E(uu^\top | Z) = \Omega$, then we get the the GIVE estimate, as discussed in an earlier lecture.

This justifies GMM as GLS for a nonlinear regression model. Note that the *assumption* of exact normality is rather implausible, but *approximate* normality is easy to justify since one would hope/expect that

$$n^{-1/2}Z^{\top}\hat{u}$$

would satisfy conditions for a CLT. So in practice, we have approximate normality and we solve

$$\min_{\theta} M(\theta)^1 V_n^{-1} M(\theta)$$

where $V_n \to V$ in probability.

Now, in considerably more general situations than 2SLS we may think of orthogonality conditions generating a set of $\perp$ conditions

$$M(\theta) = 0$$

with $V = EMM^{\top}$ and we can, on the same principle as we have just developed suggest using

$$\hat{\theta} = \arg\min_{\theta} M(\theta)^{\top} V^{-1} M(\theta)$$

Suppose we had some consistent estimator of θ, say $\hat{\theta}_0$, then by Taylor expansion

$$M(\theta) = M(\theta_0) + (\theta - \hat{\theta}_0)^{\top} \nabla M(\tilde{\theta}_0)$$

and a one-step estimation of θ would minimize

$$\min_{\theta} (M(\hat{\theta}_0) - \nabla M(\tilde{\theta}_0)^{\top}(\theta - \hat{\theta}_0))^{\top} V^{-1} (M(\hat{\theta}_0) - \nabla M(\tilde{\theta}_0)(\theta - \hat{\theta}_0))$$
$$\Rightarrow \hat{\theta}_1 = \hat{\theta}_0 + (\nabla M^{\top} V^{-1} \nabla M)^{-1} \nabla M^{\top} V^{-1} M(\hat{\theta}_0).$$

one could continue to iterate this solution, which would yield (eventually) a solution to the original problem.

Returning to Dynamic Panel Models

Now, we are ready to consider the use of GMM methods in panel data. To fix some ideas, consider our very simple dynamic panel model

$$y_{it} = \alpha y_{it-1} + \eta_i + \nu_{it}$$

where $E\nu_{it}\nu_{is} = 0$ for $t \neq s$. We are interested in estimating the vector α and to do so we would like to find an exhaustive list of available, valid moment conditions.

The first problem is that the η_i's generate dependence over time, the second problem is that if we pursue the HT strategy of applying the Q transformation to get rid of η_i, we lose the time-series structure of the data. What to do? Consider first differencing the data. We get

$$\Delta y_{it} = \alpha \Delta y_{it-1} + \Delta \nu_{it}$$

But if the ν_{it} are iid, then y_{t-2} is independent of $\Delta\nu_{it}$, and so is y_{t-3}, etc. So we may collect these conditions to write

$$E[(\Delta y_{it} - \alpha \Delta y_{it-1}) y_{i(t-j)}] = 0$$

$$t = 3, \ldots, T; \qquad j = 2, \ldots, t-1$$

From these conditions we may design an estimator *à la* GMM. Anderson and Hsiao (1981) suggest estimating $(*)$ by IV using either y_{it-2} or Δy_{it-2} as an IV. Since we obviously have more serviceable instruments it (may) make sense to use more instruments. AB(1991) suggests using all the $\perp$ conditions and GMM.

They write

$$Z_i = \begin{pmatrix} y_{i1} & & & \\ & (y_{i1}, y_{i2}) & & \\ & & \ddots & \\ & & & \ldots (y_{i1}, y_{i2}, \ldots, y_{i(T-2)}) \end{pmatrix}$$

$$(T-2) \times (m = (T-2)(T-1)/2)$$

Note that $1 + 2 + \ldots + (T-2) = m$. And $\tilde{v}_i = (\Delta v_{i3}\ \Delta v_{i4}\ \ldots\ \Delta v_{iT})'$ The $\perp$ conditions say that

$$E Z_i' \tilde{v}_i = 0$$

so GMM suggests that we minimize

$$(\sum_{i=1}^{n} \tilde{v}_i(\alpha)^\top Z_i) A_n (\sum_{i=1}^{n} Z_i^\top \tilde{v}_i(\alpha))$$

for some appropriate choice of A_n. Which one? This would yield the estimator (4) in AB (1991)

$$\hat{\alpha} = \frac{\Delta y_{-1} Z A Z^\top \Delta y}{\Delta \bar{y}_{-1} Z A Z^\top \Delta y_{-1}}$$

Consider $V(n^{-1} \sum Z_i' \tilde{v}_i) = n^{-1} \sum Z_i (E \tilde{v}_i \tilde{v}_i') Z_i$ where

$$E \tilde{v}_i \tilde{v}_i^\top = \sigma_u^2 \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ 0 & -1 & 2 & -1 & \\ & & & \ddots & \\ & \ldots & \ldots & -1 & 2 \end{pmatrix}$$

T	3	6	10	20
q	1	10	36	171
N	20	20,000	933,000	100,000,000
n	7	3,333	93,000	5,000,000

Since

$$\begin{aligned} E\tilde{v}_{it}\tilde{v}_{it} &= E(v_{it} - v_{it-1})(v_{it} - v_{it-1}) \\ &= Ev_i^2 - 2v_{it}v_{it-1} + v_{it-1}^2 \\ &= 2\sigma_v^2 \\ E\tilde{v}_{it}\tilde{v}_{it-1} &= E(v_{it} - v_{it-1})(v_{it-1} - v_{it-2}) \\ &= -\sigma_v^2 \end{aligned}$$

If there is heteroscedasticity, then things are more complicated. Of course

$$E\tilde{v}_{it}\tilde{v}_{it-s} = 0 \qquad \text{for } s \geq 2.$$

This gives a one-step estimator. A two step estimator may be constructed using the White type estimator,

$$\hat{V}_n = n^{-1} \sum Z_i^\top \hat{\tilde{v}}_i \hat{\tilde{v}}_i^\top Z_i$$

In effect these estimators are like the Anderson-Hsiao estimator, but (i) They use more IV's (ii) They use a better $\hat{V}_n$. It is interesting to consider how the number of IV's grows with T in the AB model. A simple computation yields the following table which describes the situation. Here we let N denote a rule of thumb for choosing the full sample necessary to justify the use of q moment conditions. The rule, which is developed in Koenker and Machado (1999) requires that $N = 20q^3$, where 20 is viewed as a reasonable initial sample size for estimating a scalar parameter. The last row of the table is simply, $n = [N/T]$.

It is reasonably straightforward to consider adding exogenous variables. Consider

$$y_{it} = \alpha y_{it-1} + x_{it}^\top \beta + \eta_i + u_{it}$$

if we don't wish to assume $x_{it} \perp \eta_i$, then we get Z like previous case except that in addition to $y_{it}, \ldots, y_{is}$ we have $x_{i1}, \ldots, x_{is+1}$. for predetermined x's and $x_{i1} \ldots x_{iT}$ for strictly exogenous x's. More generally, as in HT, we could partition x's into x_1, x_2 with $x_1 \perp \eta$, then we get even more $\perp$ conditions which could be exploited. These methods are available in R, in the package `plm` by Yves Croissant. See the function `pgmm` in particular.

Postscript

All of the foregoing discussion is closely related to recent developments in semiparametric econometrics and especially to the empirical Bayes approach to models with many incidental parameters. From this viewpoint it is reasonable to think of the α_i parameters as drawn from a distribution, and if we focus on MLE's that attempt to estimate this *distribution* as well as the structural parameters of interest, there is an opportunity to restore some sanity to the ML method. This is close to the viewpoint espoused in two influential papers by Lancaster(2000, 2002), who adopts the view that integrating out the α_i's via some reasonable prior can provide a similar resolution. In effect the empirical Bayes approach simply offers a way to "estimate this prior from the data."

References

Hsiao, C. (1986) it Analysis of Panel Data, Cambridge

Nickell, S. (1981) Biases in Dynamic Models with Fixed Effects, *Econometrica*, 1417-26..

Chamberlain (1984). *Handbook of Econometrics*, Elsevier.

Neyman, J. and E. Scott, (1948) Consistent Estimates Based on Partially Consistent Observations, *Econometrica*, 16, 1-32.

Lancaster, T. (2000) The incidental parameter problem since 1948, *J. of Econometrics*, 95, 391-413.

Lancaster, T. (2002) Orthogonal parameters and Panel Data, *REStud*, 69, 647-666.

Arellano, M. and P. Bond (1991). Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations *REStud*, 277-97.

Keane-Runkle (1992). On the Estimation of Panel-data Models With Serial Correlation When Instruments Are Not Strictly Exogenous, *JBES*, 1-9.

Koenker R. Machado, J.A.F. (1999) GMM Inference when the Number of Moment Conditions is Large, *J. of Econometrics*,(1999), 93, 327-344.