Applied Panel Data Analysis – Lecture 9

Christopher F. Parmeter

AGRODEP
September 9-13th, 2013
Dakar, Senegal

UMIAMI

- Estimation and inference when dynamics are present
- Introduce the dynamic panel data model
- Discuss the issue of too many instruments in this setting

- Dynamic models abound in economics
- Baltagi and Levin (1986) analyze the dynamic demand for cigarettes
- Arellano and Bond (1991) study a dynamic model of employment
- Islam (1995) and Caselli, Esquivel and LeFort (1996) study dynamic growth models
- Dynamic model for pollution emissions or global warming are additional examples

- Our dynamic unobserved effects model is

$$y_{it} = \delta y_{i,t-1} + x'_{it}\beta + c_i + \varepsilon_{it} \tag{1}$$

- The presence of $y_{i,t-1}$ introduces additional statistical complications
- There are now two sources of time persistence, autocorrelation from the lagged dependent variable and individual specific heterogeneity

- First, note that since y_{it} is correlated with c_i, so to is $y_{i,t-1}$
- In the random effects framework where c_i is in the error term, this implies that $y_{i,t-1}$ is endogenous
- In this case pooled OLS is biased and consistent (recall it was only inefficient in random effects framework for the unobserved effects model)
- The fixed effect transformation will eliminate the c_i, however, the within transformed lagged dependent variable will be correlated with the within transformed error

- That is, $y_{i,t-1} - \bar{y}_{i,-1}$, where $\bar{y}_{i,-1} = (T-1)^{-1} \sum_{t=2}^{T} y_{i,t-1}$ is correlated with $\varepsilon_{i,t-1} - \bar{\varepsilon}_{i,-1}$
- The reason is that $y_{i,t-1}$ depends on $\varepsilon_{i,t-1}$, which is a component of $\bar{\varepsilon}_{i,-1}$
- While Nickell (1985) shows that the bias of the within estimator is proportional to T^{-1} in the classic microeconomic setting where N is large and T is small, this provides little assurance
- Even with $T = 30$, Judson and Owen (1999) find the bias of the within estimator for the dynamic unobserved effects model to be on the order of 20%
- The GLS estimator for the random effects framework will be biased for the same reason the within estimator is biased $\check{y}_{i,t-1}$ is correlated with $\check{u}_{i,t-1}$

- Given the endogeneity issues with the standard estimators of the dynamic unobserved effects model, a natural approach is to construct instruments to control for endogeneity
- Unlike a cross-sectional setting, there exist natural instruments in a dynamic panel setting, namely, further lags of the dependent variable
- Arellano and Bond (1991) proposed the dominant estimator for this model in applied economics
- There estimator is best explained omitting covariates

- The random effects framework for the simple dynamic unobserved effects model is

$$y_{it} = \delta y_{i,t-1} + c_i + \varepsilon_{it} \tag{2}$$

where $c \sim IID(0, \sigma_c^2)$ and $\varepsilon \sim IID(0, \sigma_\varepsilon^2)$

- The first step is to first-difference (**??**) to eliminate the unobserved effects

$$\begin{aligned} y_{it} - y_{i,t-1} =& \delta(y_{i,t-1} - y_{i,t-2}) + \varepsilon_{it} - \varepsilon_{i,t-1} \\ \triangle y_{it} =& \delta \triangle y_{i,t-1} + \triangle \varepsilon_{it} \end{aligned} \tag{3}$$

- If ε_{it} is IID, then $\triangle \varepsilon_{it}$ is $MA(1)$
- Note that for first differencing to be applicable we need $T > 2$, so more is required of the data to model a dynamic relationship

- Consider the first applicable period, $t = 3$
- Here we have

$$y_{i3} - y_{i2} = \delta(y_{i2} - y_{i1}) + \varepsilon_{i3} - \varepsilon_{i2} \tag{4}$$

- A valid instrument for $y_{i2} - y_{i1}$ is y_{i1}
- For $t = 4$ a valid instrument set for $y_{i3} - y_{i2}$ is (y_{i2}, y_{i1})
- Notice that as t increases we gain an additional instrument for each time period

- The instrument matrix is then $Z = [Z_1', \ldots, Z_N']'$ where

$$Z_i = \begin{bmatrix} [y_{i1}] & 0 & \cdots & 0 \\ 0 & [y_{i1}, y_{i2}] & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & [y_{i1}, \ldots, y_{i,T-2}] \end{bmatrix} \quad (5)$$

- Premultiplying the first differenced equation (**??**) by Z' results in our IV equation

$$Z' \triangle y = \delta Z' \triangle y_{-1} + Z' \triangle \varepsilon \tag{6}$$

- Before we can construct the IV estimator for δ we need to account for the $MA(1)$ structure of $\triangle \varepsilon$

- Note that $E(\triangle\varepsilon\triangle\varepsilon') = \sigma^2_\varepsilon(I_N \otimes G)$ where

$$G = \begin{bmatrix} 2 & -1 & 0 & \cdots & 0 & 0 & 0 \\ -1 & 2 & -1 & \cdots & 0 & 0 & 0 \\ 0 & -1 & 2 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 2 & -1 & 0 \\ 0 & 0 & 0 & \cdots & -1 & 2 & -1 \\ 0 & 0 & 0 & \cdots & 0 & -1 & 2 \end{bmatrix} \tag{7}$$

- G is known as a banded matrix, with 2 along the main diagonal and -1 along the main off-diagonals
- G is $(T-2) \times (T-2)$

- GLS estimation of (**??**) produces the preliminary Arellano and Bond (1991) dynamic panel data estimator

$$\hat{\delta} = \left[\triangle y_{-1}' P_{ZG} \triangle y_{-1}\right]^{-1} \triangle y_{-1}' P_{ZG} \triangle y \quad (8)$$

where $P_{ZG} = Z\left(Z'(I_N \otimes G)Z\right)^{-1} Z'$

- Notice that we do not need an estimator of σ_{ε}^2 to make this estimator feasible

- An optimal estimator for δ can be determined following Hansen (1982)
- In this case one replaces $Z'(I_N \otimes G)Z$ in P_{ZG} with

$$V_{Z\varepsilon} = Z'\triangle\varepsilon\triangle\varepsilon' Z \tag{9}$$

- The optimal estimator in this case is

$$\hat{\delta}^{opt} = \left[\triangle y'_{-1} Z \hat{V}_{Z\varepsilon}^{-1} Z' \triangle y_{-1}\right]^{-1} \triangle y'_{-1} Z \hat{V}_{Z\varepsilon}^{-1} Z' \triangle y \tag{10}$$

where $\triangle\varepsilon$ is replaced with the differenced residuals using the preliminary estimator

- A consistent estimator for the asymptotic variance of $\hat{\delta}^{opt}$ is

$$\widehat{var}(\hat{\delta}^{opt}) = \left[\triangle y'_{-1} Z \hat{V}_{Z\varepsilon}^{-1} Z' \triangle y_{-1}\right]^{-1} \tag{11}$$

- A concern in applied work is when $\delta \approx 1$
- In this case the lagged regressors will not be highly correlated with $\triangle y_{it}$, leading to a weak instrument problem
- Also, adding in all available instruments can exacerbate any perceived weakness of the instruments across time
- The Arellano and Bond (1991) estimator (and many other estimators) break down when there are weak instruments

- We can incorporate strictly exogenous regressors, x_{it} in (**??**) in the same fashion
- However, given the strict exogeneity condition, we can use the xs in each period as instruments for x_{it}
- That is, $z_{x_i} = [x'_{i1}, x'_{i2}, \ldots, x'_{iT}]$ should be added to each 'diagonal' element of Z_i

$$Z_i = \begin{bmatrix} [y_{i1}, z_{x_i}] & 0 & \cdots & 0 \\ 0 & [y_{i1}, y_{i2}, z_{x_i}] & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & [y_{i1}, \ldots, y_{i,T-2}, z_{x_i}] \end{bmatrix} \tag{12}$$

- The first differenced equation with strictly exogenous covariates, premultiplied by Z, becomes

$$Z'\triangle y = \delta Z'\triangle y_{-1} + Z'\triangle X\beta + Z'\triangle\varepsilon \tag{13}$$

- The preliminary estimator of (δ, β') is

$$\begin{pmatrix} \hat{\delta} \\ \hat{\beta} \end{pmatrix} = \left[(\triangle y_{-1}, \triangle X)' P_{ZG} (\triangle y_{-1}, \triangle X) \right]^{-1} (\triangle y_{-1}, \triangle X)' P_{ZG} \triangle y \tag{14}$$

which is identical to the simple dynamic estimator except for the inclusion of X

- Similarly, an optimal two-step estimation can be constructed by replacing $Z'(I_N \otimes G)Z$ in P_{ZG} with

$$V_{Z\varepsilon} = Z' \triangle\varepsilon \triangle\varepsilon' Z \tag{15}$$

- The unknown $\triangle\varepsilon$ can be replaced with the first differenced residuals obtain using the preliminary estimator

- If one faces a situation where strict exogeneity is too stringent of a condition an estimator can be constructed requiring only predeterminedness of the covariates
- Recall that we needed strict exogeneity in our earlier fixed/random effects frameworks
- Here we can relax this assumption and deploy and instrumental variables approach
- Need to be careful however, in practice many researchers use these types of instruments without putting much thought into what the economic relationship is (see Bazzi and Clemens, 2013)

- With predeterminedness we need a simple modification to our individual instrument matrix
- Instead of $z_{x_i} = [x'_{i1}, x'_{i2}, \ldots, x'_{iT}]$ being a valid matrix for each covariate in a given time period, now $[x'_{i1}, x'_{i2}, \ldots, x'_{i,t-1}]$ is a valid instrument vector for x_{it}
- In this case we have

$$Z_i = \begin{bmatrix} [y_{i1}, x'_{i1}, x'_{i2}] & 0 & \cdots & 0 \\ 0 & [y_{i1}, y_{i2}, x'_{i1}, x'_{i2}, x'_{i3}] & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & [y_{i1}, \ldots, y_{i,T-2}, x'_{i1}, x'_{i2}, \ldots, x'_{i,T-1}] \end{bmatrix}$$

- A preliminary estimator of (δ, β') can be found in exactly the same manner in this case
- Two other situations may arise in practice that could lead to more efficient estimation
  - There could be a combination of strictly exogenous and predetermined regressors
  - Not all of the x_{it} need to be correlated with c_i
- In both settings additional instrument matrices can be constructed

- There are two key inferential insights that should be investigate in the dynamic unobserved effects model: second order auto correlation of the error term and over-identifying restrictions
- Arellano and Bond (1991) provide tests of both; the AR(2) test is important because the consistency of the estimators described here hinge on the errors terms not having AR(2) serial correlation
- Primarily we are concerned with $E[\triangle\varepsilon_{it}\triangle\varepsilon_{i,t-2}] = 0$
- Can test AR(2) using a t-test from the regression of $\widehat{\triangle\varepsilon}_{it}$ on $\widehat{\triangle\varepsilon}_{i,t-2}$, where $\widehat{\triangle\varepsilon}_{it} = \triangle y_{it} - \hat{\delta}\triangle y_{i,t-1} - \triangle x_{it}'\hat{\beta}$
- Note that $T > 4$ for this test to be applicable; also should use heteroskedasticity and serial correlation robust standard errors when constructing the test statistic

- Sargan's over-identifying restrictions test proposed by Arellano and Bond (1991) is

$$\widehat{\triangle\varepsilon}'Z\left(Z'\widehat{\triangle\varepsilon}\widehat{\triangle\varepsilon}'Z\right)^{-1}Z'\widehat{\triangle\varepsilon} \sim \chi^2_{p-K-1} \qquad (16)$$

  where p is the column span of Z and K is the number of regressors
- Given the number of instruments increases with increases in T it is important to test for over-identifying restrictions
- Unfortunately, this test does not tell you which instruments are unnecessary

- A variety of alternative estimators have been proposed to estimate the dynamic unobserved effects model, all using different instrument sets
  - Keane and Runkle (1992)
  - Ahn and Schmidt (1995)
  - Arellano and Bover (1995)
  - Blundell and Bond (1998)

- Incorporated dynamics into the unobserved effects model
- Discussed estimation and inference issues
- Two new tests in this setup, over-identifying restrictions and AR(2) for the differenced error terms