Applied Panel Data Analysis – Lecture 10

Christopher F. Parmeter

AGRODEP
March 10-14th, 2015
Addis Ababa, Ethiopia

UMIAMI

- We will cover how to deal with unbalanced panel data
- Discuss the implications of having unbalanced panel data

- Our discussion the entire class so far has dealt with balanced panels, there are T observations for each of the N individuals
- It is more likely that one will have access to an unbalanced panel, where there are individuals with $T_i < T$ observations
- Examples include individuals dying or moving out of the sampling area or with cross-country studies, many countries have incomplete data prior to some date

- Conceptually an unbalanced panel introduces some notational complications, but the estimators we have discussed so far still operate in the same fashion
- We will assume that our panel is unbalanced completely at random
- When observations are missing in a systematic fashion this introduces econometric issues that need to be explicitly handled

- Consider a model with two cross sections with an unequal number of time series
- Assume that individual 2 has $T = T_1 + T_2$ observations while individual 1 has just T_1 observations
- The stacked model is

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \beta + \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}$$

- X_1 is of dimension $T \times K$ and X_2 is of dimension $T \times K$

- The variance-covariance matrix of the error vector is

$$\Omega = \begin{bmatrix} \sigma_\varepsilon^2 I_{T_1} + \sigma_c^2 J_{T_1} & 0 & 0 \\ 0 & \sigma_\varepsilon^2 I_{T_1} + \sigma_c^2 J_{T_1} & \sigma_c^2 J_{T_1 T_2} \\ 0 & \sigma_c^2 J_{T_1 T_2} & \sigma_\varepsilon^2 I_{T_2} + \sigma_c^2 J_{T_2} \end{bmatrix}$$

  where J_T is a $T \times T$ matrix of ones while $J_{T_1 T_2}$ is a $T_1 \times T_2$ matrix of ones
- Notice that all off-diagonal, non zero elements of Ω are σ_c^2
- This extends to more than two individuals

- Ω in the n individual setting has a block diagonal structure with jth block

$$\Omega_j = (T_j\sigma_c^2 + \sigma_\varepsilon^2)\bar{J}_{T_j} + \sigma_\varepsilon^2 E_{T_j} \tag{1}$$

where $\bar{J}_{T_j} = J_{T_j}/T_j$ and $E_{T_j} = I_{T_j} - \bar{J}_{T_j}$

- To apply GLS we again use the spectral decomposition, but at the block level, which gives us

$$\Omega_j^r = (T_j\sigma_c^2 + \sigma_\varepsilon^2)^r \bar{J}_{T_j} + (\sigma_\varepsilon^2)^r E_{T_j} \tag{2}$$

- Let $\sigma_{1j}^2 = T_j\sigma_c^2 + \sigma_\varepsilon^2$, then our unbalanced random effects framework transformation is

$$\sigma_\varepsilon \Omega_j^{-1/2} = (\sigma_\varepsilon/\sigma_{1j})\bar{J}_{T_j} + E_{T_j} = I_{T_j} - \theta_j \bar{J}_{T_j} \tag{3}$$

  where $\theta_j = 1 - \sigma_\varepsilon/\sigma_{1j}$
- Our transformation works as $\check{z}_{it} = z_{it} - \theta_j \bar{z}_{i\cdot}$ where $\bar{z}_{i\cdot} = T_j^{-1}\sum_{t=1}^{T_j} z_{it}$
- Unlike the balanced panel case for the random effects framework, here our weighting is individual specific
- Individuals with larger T_j will have a θ_j that is smaller
- This different weighting has important implications for the random versus fixed effects framework setup

- Both the within and between estimators work in similar fashion
- Our Q matrix for the within transformation is now $Q = diag(E_{T_j})$ instead of $I_N \otimes E_T$
- Our P matrix for the between transformation is now $P = diag(\bar{J}_{T_j})$ instead of $I_N \otimes \bar{J}_T$
- The only issue that remains is how to estimate the variance components in the unbalanced case

- As in the balanced case we will use $u'Qu$ and $u'Pu$ to estimate our variance components; here Q and P are in unbalanced form
- This leads to complications in the solutions for $\hat{\sigma}_1^2$ and $\hat{\sigma}_\varepsilon^2$
- Amemiya's (1971) approach is to replace u in each of the quadratic forms with the unbalanced within transformation residuals

- Amemiya's estimators are

$$\widehat{\sigma}_\varepsilon^2 = \frac{\tilde{\varepsilon}' Q \tilde{\varepsilon}}{\sum_{j=1}^{N} T_j - N - K + 1} \tag{4}$$

$$\widehat{\sigma}_1^2 = n \frac{\tilde{\varepsilon}' P \tilde{\varepsilon} - \left(N - 1 + tr[A] - tr[B]\right) \widehat{\sigma}_\varepsilon^2}{n^2 - \sum_{j=1}^{N} T_j^2} \tag{5}$$

where $n = \sum_{j=1}^{N} T_j$, $A = \left(\tilde{X}' \tilde{X}\right)^{-1} X' P X$ and $B = \left(\tilde{X}' \tilde{X}\right)^{-1} X' \bar{J} X$

- Baltagi and Chang (1994) conducted a Monte Carlo study using an unbalanced panel
- They found that balancing the panel leads to losses in inefficiency that are not recommended in practice
- There are two main ways to balance the sample, either the largest total number of observations or the largest number of individuals
- It is recommended to use an unbalanced panel rather than balance the panel as the observations that are lost are not dropped at random

- Consider the unbalanced two-way unobserved effects model

$$y_{it} = x_{it}'\beta + c_i + d_t + \varepsilon_{it} \tag{6}$$

  for $i = 1, \ldots, N_t$ and $t = 1, \ldots, T$
- Here N_t denotes the number of individuals that are observed in period t
- This is different than how we described the one-way unobserved effects model
- N will still denote the total number of individuals in the sample

- Let D_t be the $N_t \times N$ matrix obtained from I_N by omitting the rows of the individuals that are not observed in year t
- Next define

$$\Delta = \begin{bmatrix} D_1 & D_1 \imath_N & 0 & \cdots & 0 \\ D_2 & 0 & D_2 \imath_N & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ D_T & 0 & 0 & \cdots & D_T \imath_N \end{bmatrix} = [\Delta_1, \Delta_2] \quad (7)$$

- Letting $n = \sum_{t=1}^{T} N_t$ signify the total number of observations in the unbalanced panel, Δ_1 is $n \times N$ while Δ_2 is $n \times T$
- One key difference with this setup is that the fast index here is individuals and the slow index is time; this is the opposite from both the balanced case and the one-way unbalanced case

- Δ is simply the matrix of time and individual dummies, just for a different arrangement of the data
- For n large it will be infeasible to incorporate these dummies directly into the model (in the fixed effects framework)
- A few interesting properties of the Δ matrices:
  - $\Delta_1'\Delta_1 = diag[T_i]$, the matrix that describes the number of years each individual appears in the sample
  - $\Delta_2'\Delta_2 = diag[N_t]$, the matrix that describes the number of observations in each year of the sample
  - $\Delta_2'\Delta_1$ is the $T \times N$ matrix of zeros and ones that indicates the absence/presence of an individual in a given year
- For the balanced panel case $\Delta_1'\Delta_1 = TI_N$, $\Delta_2'\Delta_2 = NI_T$ and $\Delta_2'\Delta_1 = \imath_T \imath_N' = J_{TN}$

- To construct the two-way transformation we define

$$P_{[\Delta]} = \Delta \left(\Delta' \Delta \right)^{-} \Delta'$$

- The within transformation is then $Q_{[\Delta]} = I_n - P_{[\Delta]}$
- Using matrix algebra one can show that

$$P_{[\Delta]} = P_{[\Delta_1]} + P_{[Q_{[\Delta_1]} \Delta_2]} \tag{8}$$

- Why is this important?
- Davis (2001) showed that this formulation for $P_{[\Delta]}$ is recursive; therefore if you have higher order panel data that is unbalanced, this technique is useful
- As an example consider matched employee-employer data, there you have a time effect, a firm effect and a worker effect
- Or consider cross-country trade databases, where you have an importer, an exporter and year effects
- Consider Δ_1, Δ_2 and Δ_3, then the decomposition would be

$$P_{[\Delta]} = P_{[\Delta_1]} + P_{[Q_{[\Delta_1]}\Delta_2]} + P_{[Q_{[Q_{[\Delta_1]}\Delta_2]}Q_{[\Delta_1]}\Delta_3]} \tag{9}$$

- In the random effects framework we write our error component as

$$u = \Delta_1 c + \Delta_2 d + \varepsilon \tag{10}$$

with variance-covariance matrix

$$\begin{aligned}\Omega =& \sigma_\varepsilon^2 I_n + \sigma_c^2 \Delta_1 \Delta_1' + \sigma_d^2 \Delta_2 \Delta_2' \\ =& \sigma_\varepsilon^2 \left(I_n + \phi_1 \Delta_1 \Delta_1' + \phi_2 \Delta_2 \Delta_2' \right) = \sigma_\varepsilon^2 \Sigma \end{aligned} \tag{11}$$

- Σ is an $n \times n$ matrix so direct inversion will typically not be computationally easy
- Wansbeek and Kapteyn (1989) use results for $(I + WW')^{-1}$ to show
$$\Sigma^{-1} = V - V\Delta_2 \tilde{P}^{-1} \Delta_2' V \tag{12}$$
where $V = I_n - \Delta_1 \Delta_N^{-1} \Delta_1'$, $P = \Delta_T - \imath_T \imath_N' \Delta_N^{-1} \imath_N \imath_T'$, $\Delta_N = TI_N + (\sigma_\varepsilon^2/\sigma_c^2)I_N$ and $\Delta_T = NI_T + (\sigma_\varepsilon^2/\sigma_d^2)I_T$
- Unfortunately, matrix analytic solutions for σ_ε^2, σ_c^2 and σ_d^2 do not exist in the unbalanced two-way case

- Tests for significance of the unobserved effects can be formulated as well as a Hausman test
- However, these tests have complicated structures given the unbalanced nature of the panel data
- Sound testing may reveal that a two-way effects model is statistically indifferent from a one-way error component model, in which case the notation is easier to handle

- Unbalanced panel leads to notational complications not present in the balanced panel case
- Closed form transformations exist in the one-way effect case but not in the two-way effects setup
- Should avoid balancing the panel as this can dramatically distort estimates