Unbalanced Panel Data Models

Chapter 9 from Baltagi: Econometric Analysis of Panel Data (2005)

by András Malasics (0404248)

Introduction

- ‘balanced’ or ‘complete’ panels:
  - a panel data set where data/observations are available for all cross-sectional units in the entire sample period
- ‘unbalanced’ or ‘incomplete’ panels:
  - a panel data set where some data/observations are missing for some cross-sectional units in the sample period
- Randomly missing observations

Complete panel

person	year	income	age	sex
1	2003	1500	27	1
1	2004	1700	28	1
1	2005	2000	29	1
2	2003	2100	35	2
2	2004	2200	36	2
2	2005	2000	37	2

Incomplete panel

person	year	income	age	sex
1	2003	1500	27	1
1	2004	1700	28	1
2	2003	2000	31	2
2	2004	2100	32	2
2	2005	2200	33	2
3	2004	2000	30	1

The unbalanced one-way error component model

- The model: $y_{it} = \alpha + X_{it}^{'}\beta + u_{it}$ where $\mu_i \sim IIN(0,\sigma_\mu^2)$

  $u_{it} = \mu_i + \nu_{it}$ $\quad \upsilon_{it} \sim IIN(0,\sigma_\upsilon^2)$

  $i = 1,\ldots,N$

  $t = 1,\ldots,T_i$

- Vector form: $\underbrace{y}_{nx1} = \alpha\iota_n + X\beta + u = \underbrace{Z}_{nxK}\delta + u$ where $Z = (\iota_n, X)$

  $u = Z_\mu\mu + \upsilon$ $\quad \delta' = (\alpha', \beta')$

  $n = \sum T_i$

  $Z_\mu = diag(\iota_{T_i})$

- The OLS on the unbalanced data is given by: $\hat{\delta}_{OLS} = (Z'Z)^{-1}Z'y$
  - This is BLUE when $\sigma_\mu^2 = 0$.
  - If σ_μ^2 >0 then OLS is still unbiased and consistent, but its standard errors are biased.

Complete panel case

$$Z_\mu = I_N \otimes \iota_T = \begin{pmatrix} \left.\begin{matrix}1\\ \vdots\\ 1\end{matrix}\right\}T & \left.\begin{matrix}0\\ \vdots\\ 0\end{matrix}\right\}T & \cdots & 0 \\ \left.\begin{matrix}0\\ \vdots\\ 0\end{matrix}\right\}T & \left.\begin{matrix}1\\ \vdots\\ 1\end{matrix}\right\}T & & \vdots \\ & & \ddots & \\ \vdots & & & \\ 0 & \cdots & \left.\begin{matrix}0\\ \vdots\\ 0\end{matrix}\right\}T & \left.\begin{matrix}1\\ \vdots\\ 1\end{matrix}\right\}T \end{pmatrix}$$

$$Z = (\underbrace{\iota_{NT}}_{NTx1}, \underbrace{X}_{NTx(K-1)}) = \underbrace{\begin{bmatrix} \left.\begin{matrix}1\\ \vdots\\ 1\end{matrix}\right\}T & \underbrace{X_1}_{Tx(K-1)} \\ \left.\begin{matrix}1\\ \vdots\\ 1\end{matrix}\right\}T & \underbrace{X_2}_{Tx(K-1)} \\ \vdots & \vdots \\ \left.\begin{matrix}1\\ \vdots\\ 1\end{matrix}\right\}T & \underbrace{X_N}_{Tx(K-1)} \end{bmatrix}}_{NTxK}$$

Incomplete panel case

$$Z_\mu = I_N \otimes \iota_{T_i} = \begin{pmatrix} \left.\begin{matrix}1\\ \vdots\\ 1\end{matrix}\right\}T_1 & \left.\begin{matrix}0\\ \vdots\\ 0\end{matrix}\right\}T_1 & \cdots & 0 \\ \left.\begin{matrix}0\\ \vdots\\ 0\end{matrix}\right\}T_2 & \left.\begin{matrix}1\\ \vdots\\ 1\end{matrix}\right\}T_2 & & \vdots \\ & & \ddots & \\ \vdots & & & \\ 0 & \cdots & \left.\begin{matrix}0\\ \vdots\\ 0\end{matrix}\right\}T_N & \left.\begin{matrix}1\\ \vdots\\ 1\end{matrix}\right\}T_N \end{pmatrix}$$

$$Z = (\underbrace{\iota_n}_{nx1}, \underbrace{X}_{nx(K-1)}) = \underbrace{\begin{bmatrix} \left.\begin{matrix}1\\ \vdots\\ 1\end{matrix}\right\}T_1 & \underbrace{X_1}_{T_1x(K-1)} \\ \left.\begin{matrix}1\\ \vdots\\ 1\end{matrix}\right\}T_2 & \underbrace{X_2}_{T_2x(K-1)} \\ \vdots & \vdots \\ \left.\begin{matrix}1\\ \vdots\\ 1\end{matrix}\right\}T_N & \underbrace{X_N}_{T_Nx(K-1)} \end{bmatrix}}_{nxK}$$

- Within estimator: $\tilde{\beta} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{y} = (X'QX)^{-1}X'Qy$ where $E_{T_i} = I_{T_i} - \bar{J}_{T_i}$

$$\tilde{X} = QX$$

$$\tilde{y} = Qy$$

$$Q = diag(E_{T_i})$$

$$\bar{J}_{T_i} = \frac{J_{T_i}}{T_i}$$

$$J_{T_i} = \begin{pmatrix} \overbrace{\begin{matrix} 1 & \cdots & 1 \\ \vdots & & \vdots \\ 1 & \cdots & 1 \end{matrix}}^{T_1 x T_1} & & & 1 \\ & \left.\begin{matrix} 1 & \cdots & 1 \\ \vdots & & \vdots \\ 1 & \cdots & 1 \end{matrix}\right\} T_2 x T_2 & & \\ & & \ddots & \\ 1 & & & \underbrace{\begin{matrix} 1 & \cdots & 1 \\ \vdots & & \vdots \\ 1 & \cdots & 1 \end{matrix}}_{T_N x T_N} \end{pmatrix}$$

$$I_{T_i} = \underbrace{\begin{pmatrix} \left.\begin{matrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{matrix}\right\} I_{T_1} & 0 & 0 & 0 \\ 0 & \left.\begin{matrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{matrix}\right\} I_{T_2} & & 0 \\ 0 & & \ddots & \\ 0 & 0 & & \left.\begin{matrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{matrix}\right\} I_{T_N} \end{pmatrix}}_{nxn}$$

Incomplete panel

$$Q = diag(E_{T_i}) = \begin{pmatrix} \left.\begin{matrix} 1-\frac{1}{T_1} & -\frac{1}{T_1} & \cdots & -\frac{1}{T_1} \\ -\frac{1}{T_1} & 1-\frac{1}{T_1} & & \vdots \\ \vdots & & \ddots & \\ -\frac{1}{T_1} & \cdots & & 1-\frac{1}{T_1} \end{matrix}\right\} T_1 x T_1 & & & 0 \\ & \left.\begin{matrix} 1-\frac{1}{T_2} & & & -\frac{1}{T_2} \\ & 1-\frac{1}{T_2} & & \\ & & \ddots & \\ -\frac{1}{T_2} & & & 1-\frac{1}{T_2} \end{matrix}\right\} T_2 x T_2 & & \\ & & \ddots & \\ 0 & & & \left.\begin{matrix} 1-\frac{1}{T_N} & & & -\frac{1}{T_N} \\ & 1-\frac{1}{T_N} & & \\ & & \ddots & \\ -\frac{1}{T_N} & & & 1-\frac{1}{T_N} \end{matrix}\right\} T_N x T_N \end{pmatrix}$$

$$Q = I_{NT} - T^{-1} I_N \otimes J_T = \begin{pmatrix} \left.\begin{matrix} 1-\frac{1}{T} & -\frac{1}{T} & \cdots & -\frac{1}{T} \\ -\frac{1}{T} & 1-\frac{1}{T} & & \vdots \\ \vdots & \ddots & & \\ -\frac{1}{T} & \cdots & & 1-\frac{1}{T} \end{matrix}\right\} TxT & & & 0 \\ & \left.\begin{matrix} 1-\frac{1}{T} & & & -\frac{1}{T} \\ & 1-\frac{1}{T} & & \\ & & \ddots & \\ -\frac{1}{T} & & & 1-\frac{1}{T} \end{matrix}\right\} TxT & & \\ & & \ddots & \\ 0 & & & \left.\begin{matrix} 1-\frac{1}{T} & & & -\frac{1}{T} \\ & 1-\frac{1}{T} & & \\ & & \ddots & \\ -\frac{1}{T} & & & 1-\frac{1}{T} \end{matrix}\right\} TxT \end{pmatrix}$$

Complete panel

- Within residuals $\tilde{u}$ for the unbalanced panels are given by:

$$\tilde{u} = y - \tilde{\alpha}\iota_N - X\tilde{\beta}$$
$$\tilde{\alpha} = (\bar{y}_{..} - \bar{X}_{..}\tilde{\beta})$$
$$\bar{y}_{..} = \frac{\sum\sum y_{it}}{n}$$

- The Between estimator and the Between residuals are obtained as follows:

$$\hat{\delta}_{Between} = (Z'PZ)^{-1}Z'Py$$
$$P = diag[\bar{J}_{T_i}]$$

$$\hat{u}^b = y - Z\hat{\delta}_{Between}$$

The unbalanced two-way error component model

- The fixed effects model:
  - Wansbeek and Kapteyn (1989) consider the following model:

$$y_{it} = \alpha + X_{it}^{'}\beta + u_{it}$$

$$u_{it} = \mu_i + \lambda_t + \nu_{it}$$

$$i = 1,\ldots,N_t$$

$$t = 1,\ldots,T$$

  - Furthermore define the matrix Δ that gives the dummy-variable structure for the incomplete data model:

$$\Delta = (\Delta_1, \Delta_2) \equiv \begin{bmatrix} D_1 & D_1 \iota_N & & \\ \vdots & & \ddots & \\ D_T & & & D_T \iota_N \end{bmatrix}$$

$$\Delta_1 = \underbrace{(D_1^{'},\ldots,D_T^{'})'}_{nxN}$$

$$\Delta_2 = diag\left[D_t \iota_N\right] = \underbrace{diag\left[\iota_{N_t}\right]}_{nxT}$$

- D_t is N_txN matrix obtained from I_N by omitting the rows corresponding to individuals not observed in year t
- For complete panels:

$$\Delta_1 = (\iota_T \otimes I_N)$$

$$\Delta_2 = I_T \otimes \iota_N$$

- If μ_i and λ_t are fixed, one has to run the regression with the matrix of dummies on the previous slide
- However, it is infeasible for large panels → Within transformation needed

- Incomplete case: ⟷ • Complete case:

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

$$Q_{[\Delta]} = Q_{[\Delta_1]} - Q_{[\Delta_1]}\Delta_2(\Delta'_2 Q_{[\Delta_1]}\Delta_2)^{-1}\Delta'_2 Q_{[\Delta_1]}$$

$Q_{[\Delta]} = I_n - P_{[\Delta]}$ where $P_{[\Delta]} = \Delta(\Delta'\Delta)^{-1}\Delta'$

and

$$\Delta_N \equiv \Delta'_1\Delta_1 = diag[T_i]$$
$$\Delta_T \equiv \Delta'_2\Delta_2 = diag[N_t]$$
$$\Delta_{TN} \equiv \Delta'_2\Delta_1$$

⟷

$$\Delta_N = TI_N$$
$$\Delta_T = NI_T$$
$$\Delta_{TN} = \iota_T\iota'_N = J_{TN}$$

- Extensions to higher-order error component model (e.g. 3-way): (Davis (2001))

$$\Delta = [\Delta_1, \Delta_2, \Delta_3] \qquad Q_{[\Delta]} = Q_{[A]} - P_{[B]} - P_{[C]} \qquad \begin{aligned} A &= \Delta_1 \\ B &= Q_{[A]}\Delta_2 \\ C &= Q_{[B]}Q_{[A]}\Delta_3 \end{aligned}$$

- The random effects model:
  - Vector form of the model: $u = \Delta_1 \underbrace{\mu}_{Nx1} + \Delta_2 \underbrace{\lambda}_{Tx1} + \underbrace{\upsilon}_{nx1}$ where

$$\mu \sim (0, \sigma_\mu^2)$$
$$\lambda \sim (0, \sigma_\lambda^2)$$
$$\upsilon \sim (0, \sigma_\upsilon^2)$$

  - Variance matrix: $\Omega = E(uu') = \sigma_\upsilon^2 I_n + \sigma_\mu^2 \Delta_1 \Delta'_1 + \sigma_\lambda^2 \Delta_2 \Delta'_2 =$

$$= \sigma_\upsilon^2 (I_n + \phi_1 \Delta_1 \Delta'_1 + \phi_2 \Delta_2 \Delta'_2) = \sigma_\upsilon^2 \Sigma$$

where $\phi_1 = \sigma_\mu^2 / \sigma_\upsilon^2$

$\phi_2 = \sigma_\lambda^2 / \sigma_\upsilon^2$

  - Using the general expression for the inverse of (I+X'X), one obtains

$$\Sigma^{-1} = V - V\Delta_2 \tilde{P}^{-1} \Delta'_2 V$$

where

$$V = I_n - \Delta_1 \tilde{\Delta}_N^{-1} \Delta'_1$$
$$\tilde{P} = \tilde{\Delta}_T - \Delta_{TN} \tilde{\Delta}_N^{-1} \Delta'_{TN}$$
$$\tilde{\Delta}_N = \Delta_N + (\sigma_\upsilon^2 / \sigma_\mu^2) I_N$$
$$\tilde{\Delta}_T = \Delta_T + (\sigma_\upsilon^2 / \sigma_\lambda^2) I_T$$

  - Davis (2001) shows that this result can be generalized to an arbitrary number of random error components, e.g. 3-way model:

$$\Omega = E(uu') = \sigma_\upsilon^2 I_n + \sigma_\mu^2 \Delta_1 \Delta'_1 + \sigma_\lambda^2 \Delta_2 \Delta'_2 + \sigma_\eta^2 \Delta_3 \Delta'_3$$

- Wansbeek and Kapteyn suggest an ANOVA-type quadratic unbiased estimator of the variance components based on the Within residuals
- Let $e = y - X\tilde{\beta}$ and define

$$q_W = e'Q_{[\Delta]}e$$

$$q_N = e'\Delta_2\Delta_T^{-1}\Delta'_2\,e = e'P_{[\Delta_2]}e$$

$$q_T = e'\Delta_1\Delta_N^{-1}\Delta'_1\,e = e'P_{[\Delta_1]}e$$

- By equating q_W, q_N, q_T to their expected values and solving these three equations one gets QUE of $\sigma_\upsilon^2, \sigma_\mu^2, \sigma_\lambda^2$

Testing for individual and time effects using unbalanced panel data

- Baltagi and Li (1990) derived a corresponding LM test for the unbalanced two-way error component model
- Under normality of disturbances the LM statistic is given by

$$LM = (\frac{1}{2})n^2[A_1^2/(M_{11}-n) + A_2^2/(M_{22}-n)]$$

$$A_r = [(\tilde{u}'\Delta_r\Delta'_r\tilde{u}/\tilde{u}'\tilde{u})-1]$$

$$M_{11} = \sum_{i=1}^{N} T_i^2$$

$$M_{22} = \sum_{t=1}^{T} N_t^2$$

which is asymptotically distributed as χ_2^2 under the null hypothesis ($H_0 : \sigma_\mu^2 = \sigma_\lambda^2 = 0$)

- If $H_0 : \sigma_\mu^2 = 0$ then $LM = (\frac{1}{2})n^2[A_1^2/(M_{11}-n)]$ and it is asymptotically distributed as χ_1^2
- If $H_0 : \sigma_\lambda^2 = 0$ then $LM = (\frac{1}{2})n^2[A_2^2/(M_{22}-n)]$ and it is asymptotically distributed as χ_1^2

- These variance components cannot be negative and therefore the alternative hypotheses are (Moulton and Randolph (1989)):
$H_A : \sigma_\mu^2 > 0$ and $H_A : \sigma_\lambda^2 > 0$ and the one-sided LM statistics are given by:

$$SLM = \frac{LM_1 - E(LM_1)}{\sqrt{\mathrm{var}(LM_1)}} = \frac{d - E(d)}{\sqrt{\mathrm{var}(d)}} \qquad \text{and} \qquad LM_2 = n[2((M_{22} - n)]^{-1/2} A_2$$

$$LM_1 = n[2(M_{11} - n)]^{-1/2} A_1$$

$$d = (\tilde{u}' D \tilde{u}) / \tilde{u}'\tilde{u}$$

$$D = \Delta_1 \Delta'_1$$

- Under the null hypothesis they have asymptotic N(0,1) distribution
- Honda's (1985) one-sided for the two-way model with unbalanced data is simply: $HO = (LM_1 + LM_2)/\sqrt{2}$
- Baltagi, Chang and Li (1998): the locally mean most powerful one-sided test for unbalanced two-way error component model is given by (King and Wu (1997)):

$$KM = \frac{\sqrt{M_{11} - n}}{\sqrt{M_{11} + M_{22} - 2n}} LM_1 + \frac{\sqrt{M_{22} - n}}{\sqrt{M_{11} + M_{22} - 2n}} LM_2$$

- Both tests are asymptotically distributed as N(0,1) under H_0

- These tests can be standardized and the resulting SLM given by:
  - For Honda's version: $D = \frac{1}{2}\frac{n}{\sqrt{M_{11}-n}}(\Delta_1\Delta'_1) + \frac{1}{2}\frac{n}{\sqrt{M_{22}-n}}(\Delta_2\Delta'_2)$
  - For the King and Wu version: $D = \frac{n}{\sqrt{2}\sqrt{M_{11}+M_{22}-2n}}[(\Delta_1\Delta'_1) + (\Delta_2\Delta'_2)]$
- Since LM1 and LM2 can be negative for a specific application, especially when one or both variance components are small or close to zero → GHM test (Gourieroux, Holly and Monfort (1982)):

$$\chi_m^2 = \begin{cases} LM_1^2 + LM_2^2 & \text{if LM1>0, LM2>0} \\ LM_1^2 & \text{if LM1>0, LM2<=0} \\ LM_2^2 & \text{if LM1<=0, LM2>0} \\ 0 & \text{if LM1<=0, LM2<=0} \end{cases}$$

and $\chi_m^2 \sim \left(\frac{1}{4}\right)\chi^2(0) + \left(\frac{1}{2}\right)\chi^2(1) + \left(\frac{1}{4}\right)\chi^2(2)$

- Recommendation:
  - The use of standardized version of these tests

Thank you for your attention!