第５講　不完備パネルデータ分析

これまで、パネルデータはすべて揃っていて欠損がない完備パネルデータを想定していた。しかし、実際のパネルデータは個人や企業が回答拒否して観察値が欠落していることがある（これを attrition 問題と呼ぶ)。また、さらには、企業であれば倒産したり、新規参入してくることもあるし、個人であれば、死亡したり、移転して追跡不可能になることもある。むしろ、パネルデータは不完備な状態の方が当たり前とさえ言える。では、不完備パネルデータを利用するために注意すべき問題点は何だろうか。

データの問題として、無作為（ランダム）にデータが欠測する場合と、有為に欠測する場合（例えば、企業倒産や個人のサンプルからの脱落）とでは意味が違ってくる。無作為（ランダム）欠測の場合、一般に不完備パネルデータであっても、その平均、分散の計算をデータサイズを適切に考慮して計算し、データサイズに応じたウェイト付けした加重最小二乗法 (weighted least square=WLS) を用いて推定すれば問題はない。問題はデータの欠測に何らかの法則性 (self-selection reasons) があり、残ったサンプルが元のサンプルの性格と違ってくる場合である。この場合にはいわゆるサンプル・セレクション・バイアス問題に直面する[1]。

誤差項の分散に関する推定は ANOVA(分散分析) 法[2]や最尤 (ML) 法[3]が用いられている。ANOVA 法は、完備データに対しては最良不偏推定が得られることが知られている、不完備データに関しては、推定は誤差項の分散の関数として表されているが（Townsend and Searle (1971))、不偏推定を得ることは可能である。しかし、等分散性、無相関性は保障されていないので、最良不偏推定とはならない。最尤法は十分統計量の関数となり、一致推定であり、漸近的に有効推定となることが示されているが、誤差項の分散を推定するために多くの自由度が失われている。

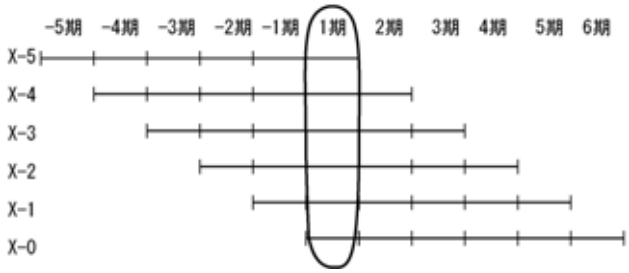
Baltagi and Chang (1994) は不完備パネルデータを用いて一元配置誤差項モデルのモンテカルロ実験を行った。その結果、次のようなことが明らかにされた。（１）推定されたパラメータに関しては ANOVA 法による一般化最小二乗法の推定も、最尤法の推定もほとんどかわらないこと。（２）誤差項の個別分散推定においては ANOVA 法による推定は最尤法に比べて精度が低い。

[1] よく知られている事例は、ニュージャージー (New Jersey) およびインディアナ州ギャリー（Gary）における所得維持政策実験である。ここでは、家計簿をつけることに便益を感じない参加者が脱落し、軍隊に召集された人も脱落し、さらにこの実験から何の恩恵も受けない高額所得者も脱落した。このように、一定の傾向を持った人々が脱落することで実験計画の無作為化と局所管理の原則が破られていった。Hausman and Wise (1979) はこの脱落問題の引き起こすバイアスを推計している。彼らによれば脱落のバイアスは小さいが有意であることが明らかにされている。

[2] ANOVA 法については (Searle (1971)、Townsend and Searle (1971)、Wallace and Hussain (1969)、Swamy and Arora (1972)、Fuller and Battese (1974)、Henderson (1953) などを参照。

[3] 最尤法については Jennrich and Sampson (1976)、Harville (1977)、Das (1979)、Corbeil and Searle (1976a,b)、Hocking (1985) などを参照。

図１　ローテーション・パネル・データの構造



とりわけ、データの不完備度が高かったり、分散構成比 (variance component ratio) が１より大きい場合には、それが顕著となる。（３）不完備データから完備データ部分だけを摘出して推定することは、有効性を大幅に失う。

これらの結果より、不完備データだからといって、一概にそのサブセットである完備データにまで情報量を落とすことは薦められないし、現在では一般に用いられているパネル・データ推定プログラムでも不完備データに応じて自動的に推定を調整してくれるようになり、推定量が完備データと比べれば最良ではないとしても、不完備データの問題は大幅に縮小されるようになっている[4]。

パネルデータの不完備性はいくつかの理由で発生するが、調査対象サンプルの一定比率を順次入れ替えるローテーション方式 (rotating panel）が採用されている場合にも発生する。

例えば、6 期間パネル調査を行っていて、毎期 6 分の 1 サンプルずつ入れ替えるという方式である。このような調査方法をとるのは、サンプルを一度に全部取り替えると前期からの連続性が全く失われてしまうが、6 分の 5 サンプルが残っていれば連続性はある程度確保される。このローテーションの 6 分の 1 に入るサンプルがランダムに選ばれているとすると (実際の家計調査では全国から選ばれているサンプルが無作為に入れ替えられる)。

もう一つの理由はサンプルから脱落してしまう (attrition）場合である。これが特定の質問に対して拒否したい人が回答拒否ということで脱落したとすれば、これはシステマティックなものであり、ランダムな脱落とは違う取扱が必要になる。

さらに、サンプルから脱落するのではなく、一部の質問に答えないあるいは答えられないという問題（incidental truncation problem）が生じる。Heckman の賃金関数のサンプルセレクション・バイアスの問題などがこれに入る (調査のある時点で失業して賃金所得がなくなるケースなど)。

[4] もちろん、不完備データにも程度があり、あまりにデータの欠落が多いようだと利用上問題がでてくることもあることには注意を要する。

5.1 不完備パネルデータの固定効果推計

次のようなモデルを考える。[5]

$$y_{it} = \mathbf{x}_{it}\beta + \alpha_i + u_{it} \quad t = 1, \ldots, T \tag{1}$$

$\mathbf{x}_{it}$ は $1 \times k$ 行列、β は $k \times 1$ 行列。α_i は $\mathbf{x}_{it}$ と相関している。

ここで、ある時点においていくつかの主体 i が観察されないケースを考えよう。具体的には $t = 1$ 期にはすべての主体 i がサンプルに入っているとする。主体 i がランダムに欠落する場合、$\mathbf{s}_i \equiv (s_{i1}, \ldots, s_{iT})'$ を $T \times 1$ 行列の選択指標（selection indicators) を考える。ここで $(\mathbf{x}_{it}, y_{it})$ が観察されれば $s_{it} = 1$、観察されなければ $s_{it} = 0$ とする。

$\{(\mathbf{x}_i, \mathbf{y}_i, \mathbf{s}_i) : i = 1, 2, \ldots, N\}$ を母集団からランダムに選ばれたサンプルであるとする。選択指標 s_i はどの期にどの i が欠落しているかを示す。

(1) 式の β の推定量は次のように表せる。

$$\begin{aligned}\hat{\beta} &= \left(N^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{\mathbf{x}}_{it}'\ddot{\mathbf{x}}_{it}\right)^{-1}\left(N^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{\mathbf{x}}_{it}'\ddot{y}_{it}\right) \\ &= \beta + \left(N^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{\mathbf{x}}_{it}'\ddot{\mathbf{x}}_{it}\right)^{-1}\left(N^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{\mathbf{x}}_{it}'u_{it}\right)\end{aligned} \tag{2}$$

ここで

$$\ddot{\mathbf{x}}_{it} \equiv x_{it} - T_i^{-1}\sum_{r=1}^{T} s_{ir}\mathbf{x}_{ir}, \quad \ddot{y}_{it} \equiv y_{it} - T_i^{-1}\sum_{r=1}^{T} s_{ir}y_{ir}, \quad T_i \equiv \sum_{t=1}^{T} s_{it}$$

不完備パネルデータの固定効果が一致推定になるためには、$E(S_{it}\ddot{\mathbf{x}}_{it}'u_{it}) = 0 \quad \text{for all } t$ が成り立つ必要がある。$\ddot{\mathbf{x}}_{it}$ はすべての $\mathbf{x}_i$ と $\mathbf{s}_i$ に依存しているので次のような強い外生性条件が満たされる必要がある。

条件 (a) $E(u_{it} \mid \mathbf{x}_i, \mathbf{s}_i, \alpha_i) = 0 \quad t = 1, 2, \ldots, T$

条件 (b) $\sum_{t=1}^{T} E(s_{it}\ddot{\mathbf{x}}_{it}'\ddot{\mathbf{x}}_{it})$ が非ゼロ行列である

条件 (c) $E(\mathbf{u}_i\mathbf{u}_i' \mid \mathbf{x}_i, \mathbf{s}_i, \alpha_i) = \sigma_u^2\mathbf{I}_T$

条件 (a) と条件 (b) が満たされていれば、固定効果推定は不完備データのもとでも一致推定となる。

ランダムローテーションパネルやその他のランダムな欠落パネルの場合には、$\mathbf{s}_i$ は $(\mathbf{u}_i, \mathbf{x}_i, \alpha_i)$ から独立しており、一般的な固定効果推定の下では $E(\mathbf{u}_{it} \mid \mathbf{x}_i, \alpha_i)$ が成り立つ。

条件 (c) が加われば、固定効果の推定は有効となる。条件 (a),(c) より

[5] 本節は Wooldridge(2002, chap17), pp.578-579 を参照している。

$$Var\left(\sum_{t=1}^{T} s_{it}\ddot{\mathbf{x}}'_{it}u_{it}\right) = \sigma_u^2\left[\sum_{t=1}^{T} E(s_{it}\ddot{\mathbf{x}}'_{it}\ddot{\mathbf{x}}_{it})\right] \tag{3}$$

固定効果推定の漸近分散は次のように表せる。

$$\hat{\sigma}_u^2\left(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\ddot{\mathbf{x}}'_{it}\ddot{\mathbf{x}}_{it}\right)^{-1} \tag{4}$$

$\hat{\sigma}_u^2$ は次のように求められる。

$$E\left(\sum_{t=1}^{T} s_{it}\ddot{\mathbf{x}}'_{it}\ddot{u}_{it}^2\right) = E\left(\sum_{t=1}^{T} s_{it}E(\ddot{u}_{it}^2 \mid \mathbf{s}_i)\right) = E\left\{T_i\left[\sigma_u^2(1-1/T_i)\right]\right\} = \sigma_u^2 E\left[(T_i-1)\right] \tag{5}$$

$s_{it}=1$ の時固定効果推定の誤差は $\hat{u}_{it} = \ddot{y}_{it} - \ddot{\mathbf{x}}_{it}\hat{\beta}$ とする。
$N^{-1}\sum_{i=1}^{N}(T_i-1) \xrightarrow{p} E(T_i-1)$ であるので

$$\hat{\sigma}_u^2 = \left[N^{-1}\sum_{i=1}^{N}(T_i-1)\right]^{-1} N^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\hat{u}_{it}^2 = \left[\sum_{i=1}^{N}(T_i-1)\right]^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T} s_{it}\hat{u}_{it}^2 \tag{6}$$

したがって $plim_{N\to\infty}\hat{\sigma}_u^2 = \sigma_u^2$ が成り立つ。

ランダム効果推定の場合、$\mathbf{s}_i$ と α_i が独立でなければならない。$\mathbf{s}_i$ と α_i が何らかのセレクションバイアスを持っていればランダム効果推定は不一致となる。

5.2 脱落サンプル問題（attrition problem）

脱落サンプルはランダムではなく何らかの理由を持って脱落していると考える。[6]一度脱落したサンプルが復帰することも考えられるがこれは問題を難しくするのでここでは一度脱落したサンプルは戻らないと仮定する。

s_{it} を選択指標とし $(\mathbf{x}_{it}, y_{it})$ が観察されるとき、$s_{it}=1$ とし、一度サンプルから脱落したら復帰しないのだから $s_{it}=1$ は $s_{ir}=1 \quad for \quad r<t$ を意味する。

観察できない効果を除去するためには一階の階差を取る。

$$\triangle y_{it} = \triangle\mathbf{x}_{it}\beta + \triangle u_{it} \quad t=2,\ldots,T \tag{7}$$

$s_{it-1}=1$ の下で選択指標は次のように表せる。

[6]本節は Wooldridge (2002, chap17), pp.585-587 を参照している。

$$s_{it} = 1\left[\mathbf{w}_{it}\delta_i + v_{it} > 0\right]$$
$$v_{it} \mid \{\mathbf{w}_{it}s_{it-1}\} \sim N(0.1) \tag{8}$$

$\mathbf{w}_{it}$ は t 期に観察される変数を含んでおりその中には $s_{it-1} = 1$ で $\mathbf{x}_{it-1}$ や $\mathbf{x}_{it}$ も含まれる。

$\mathbf{x}_{it}$ は強い外生変数であり $\mathbf{w}_{it}$ でコントロールすれば選択指標 s_{it} は $\triangle\mathbf{x}_{it}$ に依存しないという条件の下では次のような関係が成り立つ。

$$E(\triangle u_{it} \mid \triangle\mathbf{x}_{it}, \mathbf{w}_{it}, v_{it}, s_{it-1} = 1) = E(\triangle u_{it} \mid \triangle v_{it}) = \rho_t v_{it} \tag{9}$$

すなわち、

$$E(\triangle y_{it} \mid \triangle\mathbf{x}_{it}, \mathbf{w}_{it}, s_{it} = 1) = \triangle\mathbf{x}_{it}\beta + \rho_t\lambda(\mathbf{w}_{it}\delta_t) \qquad t = 2 \ldots T \tag{10}$$

ここで $\hat{\lambda}_{it}$ は $T-1$ のクロスセクションのプロビット推定 ((8) 式) より得られた the inverse Mill's ratio であり、それを用いて (10) 式をプーリング OLS 推定すれば β_1 と ρ_t は一致推定となる。

仮説検定 $H_0 : \rho_t = 0 \quad t = 2, \ldots T$ は脱落（attrition）バイアステストになり得る。

ここで問題となるのは $\mathbf{w}_{it}$ でコントロールすれば s_{it} は $\mathbf{x}_{it}$ の影響を受けないという仮定と $\mathbf{x}_{it}$ の強外生性の仮定である。これらを操作変数法を用いることで緩和してやる。

$\mathbf{z}_{it}$ を選択指標式に含まれている外生変数で $\mathbf{z}_{it}$ は $\mathbf{x}_{it} \quad for \quad r < t$ を含むものとする。

$$\triangle y_{it} = \triangle\mathbf{x}_{it}\beta + \rho_z dz_t\lambda_{it} + \ldots + \rho_t dz_t\lambda_{it} + \varepsilon_{it}) \tag{11}$$

$(z_{it}, dz_t\hat{\lambda}_{it}, \ldots dT_t\hat{\lambda}_{it})$ を操作変数としてプーリング 2LSL で (11) 式を推計してやるとパラメータは一致推定となり誤差は正規分布に従う。

仮説検定 $H_0 : \rho_t = 0 \quad t = 2, \ldots T$ が代替的な脱落バイアステストとなる。

参考文献

Baltagi, B.H. and Y.J. Chang (1994) "Incomplete Panels: A Comparative Study of Alternative Estimators for the Unbalanced One-Way Error Components Regression Model," Journal of Econometrics, 62, pp.67-89.

Corbeil, R.R. and S.R. Searle (1976a) "A Comparison of Variance Component Estimators," Biometrics, 32, pp.779-791.

Corbeil, R.R. and S.R. Searle (1976b) "Restricted Maximum Likelihood (REML), Estimation of Variance Components in the Mixed Model," Technometrics 18, pp.31-38.

Das, K. (1979) Asymptotic Optimality of Restricted Maximum Likelihood Estimates for the Mixed Model," Calcutta Statistical Association Bulletin 28, pp.125-142.

Fuller, W.A. and G.E. Battese (1974) "Estimation of Linear Models with Cross-Error Structure," Journal of Econometrics, 2, pp.67-78.

Hausman, J.A. and D. Wise (1979) "Attrition Bias in Experimental and Panel Data: the Gary Income Maintenance Experiment," Econometrica, 47, pp.455-473.

Harville, D.A. (1977) "Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems," Journal of the American Statistical Association 72, pp.320-340.

Henderson, C.R., Jr. (1953) "Estiation of Variance Components," Biometrics, 9, pp.226-252.

Hocking, R.R. (1985) The Analysis of Linear Models, Monterey: Brooks/Cole Company.

Jenrich, R.I. and P.F. Sampson (1976) "Newton-Raphson and Related Algorithms for Maximum Likelihood Variance Component Estimation," Technometrics 18, pp.11-17.

Searle, S.R. (1971) Linear Models, John Wiley, New York.

Swamy, P.A.V.B. and S.S. Arora (1972) "The Exact Finite Sample Properties of the Estimators of Coefficients in the Error Components Regression Models," Econometrica, 40, pp.261-275.

Townsend, E.C. and S.R. Searle (1971) "Best Quadratic Unbiased Estimation of Variance Components from Unbalanced Data in the One-Way Classification," Biometrics 27, pp.643-657.

Wallace, T.D. and A. Hussain (1969) "The Use of Error Components Models in Combining Cross-Section and time-Series Data, Econometrica, 37, pp.55-72.

Wooldridge, J.M. (2002) Econometirc Analysis of Cross Section and Panel Data, The MIT Press: Cambridge.