

エコノメトリックス II '07 第16回講義補足資料

パネルデータの分析

2008年2月8日

§ 7.5 パネルデータの分析

いままで扱ってきた観測データは、クロスセクション（横断面）データのように一時点において観測される様々な個体からなるもの、又は時系列データのように一つの系列が一定時間毎に観測されるもの、のどちらかであった。

しかし、観測される軸として、時間軸と横断（面）軸の二つを共に有するデータも存在する。「上場企業の当期純利益」を例にとって考えてみると、各企業の利益データは、決算期毎に存在する。このような二つの観測軸を持つデータを

- panel data
- time series & cross section data
- longitudinal data (biostatistics ではこちらを主に用いる)
- repeated measurement (biostatistics)

のように呼んでいる。

パネルデータのモデル

観測軸が二つあることから、回帰モデルは次のように表現できる。

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + u_{it} \quad (i = 1, \dots, m \text{ and } t = 1, \dots, T)$$

この回帰モデルの一番簡単な推定法は、パネルデータがもつ潜在的な構造を無視し、 u_{it} が i.i.d. である、すなわち

$$\begin{aligned} E(u_{it}^2) &= \sigma^2 \quad \forall i \text{ and } t \\ E(u_{it}u_{js}) &= 0 \quad \forall (i \neq j \text{ or } t \neq s) \end{aligned}$$

であるものと仮定して、そのまま OLS を適用するものであり、pooling regression と呼ばれている。

しかし、 u_{it} についてこのような仮定をおくことは、

- (1) 分散不均一性 (heteroscedasticity)
- (2) 時点効果 (ないし個体効果)

の存在を無視している点で問題がある。

もし、この二点が存在するときに OLS 推定を行うと、

- (a) OLSE は inefficient
- (b) 撓乱項の分散共分散行列の定式化を誤る (misspecified)

という問題が起こりうる。

そこで、パネルデータの分析では、撓乱項 u_{it} に関して次のようにモデル化することが行われる。

$$u_{it} = e_t + v_i + \varepsilon_{it}$$

time effect
時点効果

individual effect
個体効果

iid random dist
ランダムエラー

これは、分散分析 (ANOVA) の二元配置モデルと同じ構造であることが分かる。(ただし、繰り返し数が 1 であり、交互作用がない点が異なる点である。)

諸効果のモデル化の方法

諸効果のモデルとして次の三つが用いられる。

(1) fixed effect (固定効果)

e_t や v_i をパラメータとして扱う。

説明変数に 1 (全ての要素が 1 からなるベクトル) を含む場合、パラメータの識別のために $\sum_{t=1}^T e_t = 0$ かつ $\sum_{i=1}^m v_i = 0$ という制約を課す必要がある。

(2) random effect (ランダム効果)

e_t や v_i を確率変数として扱う。

$e_t \perp v_i$ であり、 $e_t \sim (0, \sigma_e^2)$, $v_i \sim (0, \sigma_v^2)$ 。

このモデル化の利点としてパラメータが少数で済むことが挙げられる。

(3) mixed effect (混合効果)

fixed effect と random effect の mixture。

biostat でよく用いられるモデル。

固定効果 (fixed effect) モデルの推定

fixed effect モデルの推定は、ダミー変数を用いた OLS 推定により行う。(以下では、説明の簡略化のために $e_t = 0 \ \forall t$ と仮定する。)

fixed effect モデルは次式で表現できる。

$$\underset{n \times 1}{\mathbf{y}} = \underset{n \times k}{\mathbf{X}} \underset{k \times 1}{\boldsymbol{\beta}} + \underset{n \times m}{\mathbf{D}} \underset{m \times 1}{\boldsymbol{\eta}} + \underset{n \times 1}{\boldsymbol{\varepsilon}} \quad \boldsymbol{\varepsilon} \sim \text{iid}(\mathbf{0}, \sigma_{\varepsilon}^2 \mathbf{I}_n)$$

ここで、 $n = mT$ であり、 $\mathbf{y}$ 、 $\mathbf{X}$ 、 $\boldsymbol{\varepsilon}$ および $\mathbf{D}$ 、 $\boldsymbol{\eta}$ は以下の形をとるベクトルおよび行列である。

$$\mathbf{y} = \begin{pmatrix} y_{11} \\ \vdots \\ y_{1T} \\ y_{21} \\ \vdots \\ y_{mT} \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} \mathbf{x}'_{11} \\ \vdots \\ \mathbf{x}'_{1T} \\ \mathbf{x}'_{21} \\ \vdots \\ \mathbf{x}'_{mT} \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{1T} \\ \varepsilon_{21} \\ \vdots \\ \varepsilon_{mT} \end{pmatrix},$$

$$\mathbf{D} = \begin{pmatrix} \mathbf{1} & \mathbf{0} & \cdots & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{1} & \ddots & & \vdots \\ \vdots & \mathbf{0} & \mathbf{1} & \ddots & \vdots \\ \vdots & & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \cdots & \mathbf{0} & \mathbf{1} \end{pmatrix}, \quad \boldsymbol{\eta} = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_m \end{pmatrix}$$

FWL 定理より、fixed effect モデルにおける $\boldsymbol{\beta}$ の OLS 推定量は

$$\mathbf{M}_D \mathbf{y} = \mathbf{M}_D \mathbf{X} \boldsymbol{\beta} + \text{error} \quad (\text{i})$$

における $\boldsymbol{\beta}$ の OLS 推定量に一致する。($\mathbf{M}_D = \mathbf{I}_n - \mathbf{D}(\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}'$ より $\mathbf{M}_D \mathbf{D} = \mathbf{0}$ となるため。) よって、 $\boldsymbol{\beta}$ の OLS 推定量とその分散は

$$\hat{\boldsymbol{\beta}}_{LSDV} = (\mathbf{X}'\mathbf{M}_D\mathbf{X})^{-1}\mathbf{X}'\mathbf{M}_D\mathbf{y},$$

$$\text{Var}(\hat{\boldsymbol{\beta}}_{LSDV}) = \sigma_{\varepsilon}^2(\mathbf{X}'\mathbf{M}_D\mathbf{X})^{-1}$$

として得られる。

ここで注意したいのは、 $\mathbf{D}'\mathbf{D} = T\mathbf{I}_m$ であるので、

$$\mathbf{M}_D = \mathbf{I} - \frac{1}{T}\mathbf{D}\mathbf{D}' = \begin{pmatrix} \mathbf{I}_T & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_T & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{I}_T \end{pmatrix} - \frac{1}{T} \begin{pmatrix} \mathbf{1}\mathbf{1}' & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{1}\mathbf{1}' & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{1}\mathbf{1}' \end{pmatrix}.$$

になっている。ゆえに、 $n \times 1$ ベクトルである $\mathbf{M}_D \mathbf{y}$ の $(it, 1)$ 要素は、 $\bar{y}_i \equiv (1/T) \sum_{t=1}^T y_{it}$ で表すと

$$(\mathbf{M}_D \mathbf{y})_{it,1} = (y_{it} - \bar{y}_i)$$

であることが分かる。この性質より、 $\hat{\boldsymbol{\beta}}_{LSDV}$ を 'within estimator' (グループ内変動推定量) と呼ぶ。

以下板書の説明を大幅に変更

ランダム効果 (random effect) モデルの推定

次に、 v_i をパラメータではなく $v_i \sim (0, \sigma_v^2)$ である確率変数として考えることにしよう。
このとき、グループ i 毎に T 期間の平均を求めると、回帰モデルは

$$\begin{aligned} \frac{1}{T} \sum_{t=1}^T y_{it} &= \left[\frac{1}{T} \sum_{t=1}^T x_{it} \right]' \beta + \frac{1}{T} \sum_{t=1}^T v_i + \frac{1}{T} \sum_{t=1}^T \varepsilon_{it} \\ &= \left[\frac{1}{T} \sum_{t=1}^T x_{it} \right]' \beta + v_i + \frac{1}{T} \sum_{t=1}^T \varepsilon_{it} \end{aligned}$$

と書くことができる。この式を行列を用いて表すと、

$$P_D y = P_D X \beta + \text{error} \quad (\text{ii})$$

となり、その OLS 推定量は

$$\hat{\beta}_{BG} = (X' P_D X)^{-1} X' P_D y$$

で得ることができる。 $P_D y$ の $(it, 1)$ 要素は、

$$(P_D y)_{it,1} = (\bar{y}_i)$$

であることから、 $\hat{\beta}_{BG}$ を 'between estimator' (グループ間変動推定量) と呼ぶ。

さて、分散分析においては

$$\text{全変動} = \text{グループ内変動} + \text{グループ間変動}$$

であった。同様に pooling regression における β の OLS 推定量を変形すると、

$$\begin{aligned} \hat{\beta} &= (X' X)^{-1} X' y \\ &= (X' X)^{-1} X' (M_D + P_D) y \\ &= (X' X)^{-1} X' M_D X \hat{\beta}_{LDV} + (X' X)^{-1} X' P_D X \hat{\beta}_{BG} \end{aligned}$$

のように、'within estimator' と 'between estimator' の線形結合として得られる。

$v_i \perp \varepsilon_{it}$ であるから、(ii) 式の誤差項の分散共分散行列は、 $\Xi \equiv \sigma_v^2 \mathbf{1}\mathbf{1}' + \frac{1}{T} \sigma_\varepsilon^2 I_T$ とおくと、

$$\begin{pmatrix} \Xi & O & \cdots & O \\ O & \Xi & \ddots & \vdots \\ \vdots & \ddots & \ddots & O \\ O & \cdots & O & \Xi \end{pmatrix}$$

で表される。これより $T \rightarrow \infty$ に近づくと、(ii) 式の誤差項の分散共分散における σ_ε^2 の寄与は小さくなるが、 σ_v^2 の部分はそのまま残ることがわかる。したがって、漸近効率性を持つ推定量を得るためにも、GLS 推定を行う必要がある。

ランダム効果モデルの GLS 推定

GLS 推定を行うために、まず回帰モデル

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$

における誤差項 $\mathbf{u}$ の分散共分散行列 ($n \times n$)

$$\boldsymbol{\Omega} \equiv \text{E}(\mathbf{u}\mathbf{u}')$$

の構造を知っておく必要がある。

$$\mathbf{u} = \begin{pmatrix} \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_m \end{pmatrix}, \quad \text{ここで } \mathbf{u}_i = \begin{pmatrix} u_{i1} \\ \vdots \\ u_{iT} \end{pmatrix} = \begin{pmatrix} v_i + \varepsilon_{i1} \\ \vdots \\ v_i + \varepsilon_{iT} \end{pmatrix}$$

であるから、 $v_i \perp \varepsilon_{it}$ かつ $v_i \sim (0, \sigma_v^2)$, $\varepsilon_{it} \sim (0, \sigma_\varepsilon^2)$ という仮定より

$$\boldsymbol{\Sigma} \equiv \text{E}(\mathbf{u}_i \mathbf{u}_i') = \sigma_v^2 \mathbf{1}\mathbf{1}' + \sigma_\varepsilon^2 \mathbf{I}_T$$

という $T \times T$ の行列を用いると、

$$\boldsymbol{\Omega} = \begin{pmatrix} \boldsymbol{\Sigma} & \mathbf{O} & \cdots & \mathbf{O} \\ \mathbf{O} & \boldsymbol{\Sigma} & \ddots & \vdots \\ \vdots & \ddots & & \mathbf{O} \\ \mathbf{O} & \cdots & \mathbf{O} & \boldsymbol{\Sigma} \end{pmatrix} \quad (\text{iii})$$

という構造になっていることがわかる。

次に、 $\boldsymbol{\Omega}^{-1/2}$ を求めることにしよう。(iii) 式より、 $\boldsymbol{\Sigma}^{-1/2}$ が求まれば、自動的に求まることが分かる。 $\mathbf{P}_1 \equiv \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1} = 1/T \mathbf{1}\mathbf{1}'$ であることを用いると、

$$\boldsymbol{\Sigma} = T\sigma_v^2 \mathbf{P}_1 + \sigma_\varepsilon^2 \mathbf{I}_T$$

となる。さらに、適当なスカラー λ を用いて、

$$\boldsymbol{\Sigma}^{-1/2} = (1/\sigma_\varepsilon) (\mathbf{I}_T - \lambda \mathbf{P}_1)$$

と表現できるものとする、

$$\boldsymbol{\Sigma}^{-1} = \frac{1}{\sigma_\varepsilon^2} (\mathbf{I}_T + (\lambda^2 - 2\lambda) \mathbf{P}_1)$$

となるので、

$$\begin{aligned} \mathbf{I}_T &= \boldsymbol{\Sigma} \boldsymbol{\Sigma}^{-1} \\ &= \frac{T\sigma_v^2}{\sigma_\varepsilon^2} \mathbf{P}_1 + \mathbf{I}_T + \frac{T\sigma_v^2}{\sigma_\varepsilon^2} (\lambda^2 - 2\lambda) \mathbf{P}_1 + (\lambda^2 - 2\lambda) \mathbf{P}_1 \end{aligned}$$

これを整理すると、

$$\begin{aligned} \left(1 + \frac{T\sigma_v^2}{\sigma_\varepsilon^2}\right) (\lambda^2 - 2\lambda + 1) &= 1 \\ (\lambda - 1)^2 &= \left(1 + \frac{T\sigma_v^2}{\sigma_\varepsilon^2}\right)^{-1} \\ \lambda &= 1 \pm \left(1 + \frac{T\sigma_v^2}{\sigma_\varepsilon^2}\right)^{-1/2} \end{aligned}$$

となり、

$$\Omega^{-1/2} = \frac{1}{\sigma_\varepsilon} (\mathbf{I}_n - \lambda \mathbf{P}_D)$$

が求められた。よって GLS 推定量は

$$\hat{\beta}_{GLS} = (\mathbf{X}'(\mathbf{I}_n - \lambda \mathbf{P}_D)^2 \mathbf{X})^{-1} \mathbf{X}'(\mathbf{I}_n - \lambda \mathbf{P}_D)^2 \mathbf{y} \quad (\text{iv})$$

であり、その分散共分散行列は

$$\text{Var}(\hat{\beta}_{GLS}) = \sigma_\varepsilon^2 (\mathbf{X}'(\mathbf{I}_n - \lambda \mathbf{P}_D)^2 \mathbf{X})^{-1}$$

となる。

このように、GLS が実行可能 (feasible) であるためには、 σ_v^2 と σ_ε^2 の値が分かっている必要がある。(λ だけではダメな事に注意。)

(iv) 式より、 $\lambda = 0$ の時には、GLS 推定量と OLS 推定量は同じになり、また $\lambda = 1$ の時には、GLS 推定量は 'within estimator' と同じになることが分かる。

Ω の推定

σ_v^2 と σ_ε^2 が既知でない場合、GLS が実行可能となるためには、これらを推定する必要がある。

まず、(i) 式の OLS 残差二乗和を $n - k - m$ で割って $\hat{\sigma}_\varepsilon^2$ を得る。次に、(ii) 式の OLS 残差二乗和を $(m - k)T$ で割ると、 $\sigma_v^2 + \widehat{\sigma_\varepsilon^2}/T$ が求まるので、

$$\hat{\sigma}_v^2 = \sigma_v^2 + \widehat{\sigma_\varepsilon^2}/T - \frac{1}{T} \hat{\sigma}_\varepsilon^2$$

として求める。

ランダム効果モデルの推定上の注意

ランダム効果モデルを適用する場合、確率変数である 'individual effect' v_i と説明変数 x_{it} とが無相関である必要がある。もし、相関がある場合には説明変数と攪乱項 (誤差項) の間に相関があることになり、GLS 推定量は不偏性、一致性を持たなくなる。

相関が存在する場合でも、 $\hat{\beta}_{LSDV}$ は不偏推定量 (但し効率性はもたない) であるから、この性質を用いた検定が提案されている。

Hausman 検定は、specification test と呼ばれるものの一つである。 v_i と x_{it} とが無相関であるという帰無仮説 H_0 の下では、 $\hat{\beta}_{LSDV}$ も $\hat{\beta}_{GLS}$ もどちらも不偏である。

$$\Psi \equiv \text{Var}(\hat{\beta}_{LSDV} - \hat{\beta}_{GLS}) = \text{Var}(\hat{\beta}_{LSDV}) - \text{Var}(\hat{\beta}_{GLS})$$

として、次のように検定統計量を構成する。

$$W = (\hat{\beta}_{LSDV} - \hat{\beta}_{GLS})' \Psi^{-1} (\hat{\beta}_{LSDV} - \hat{\beta}_{GLS})$$

W は帰無仮説 H_0 の下で自由度 $(k - 1)$ の χ^2 分布に従うので、この性質を用いて検定を行う。

勿論、Hausman 検定によって帰無仮説 H_0 が棄却された場合には、GLS 推定量は一致性がないと判断されるので、用いてはならない。