平成 26 年度 中級計量経済学・応用計量経済学
講義ノート 4 パネルデータ分析

このノートでは、パネルデータの分析で使われる統計手法を解説する。パネルデータを使用することにより、時間を通じて一定だが個人間では異なる欠落変数の影響を取り除くことができる。この利点により、横断面データを使用した場合にはうまく推定のできない効果を推定することが可能になる。一方で、横断面データの分析の場合とは違った統計理論上の注意点も出てくる。

4.1 パネルデータ

パネルデータ とは、n 人の個人を T 期間観察することから得られるデータである。(人ではなく企業や国のこともある。)

表記 : y_{it} のように二つの添え字を用いる。最初の添え字の i は個人を表し、次の添え字の t は時点を表す。

- 調和パネル: データに含まれるすべての個人と時点について、すべての変数が観測できている。
- 非調和パネル: 少なくとも一人の個人の一つの時点について、ある変数が観測できていない。

この授業では、調和パネルのみを取り扱う。ただ以下に取り扱う固定効果推定量を非調和パネルに拡張するのは容易である。しかし、動学パネルモデルの操作変数推定量を非調和パネルに拡張する際には注意が必要である。

4.2 固定効果回帰

次のような線形回帰モデルを考える。しかし、時間を通じて一定な回帰変数 $\mathbf{Z}_i$ が観測できておらず、しかも時間とともに変化する回帰変数 $\mathbf{X}_{it}$ と相関している場合を考える。

$$y_{it} = \beta_0 + \beta_1' \mathbf{X}_{it} + \beta_2' \mathbf{Z}_i + u_{it}. \tag{1}$$

もし $\mathbf{Z}_i$ を無視して OLS でこのモデルを推定すると、欠落変数のバイアスが起こり、β_1 の推定量は一致性をもたない。しかし、パネルデータを用いると β_1 の一致推定が可能になる。部分的であるにせよこのような欠落変数の問題を解決することができることがパネルデータを使用する一つの利点であり、近年の実証分析では、パネルデータを用いることが多くなってきている。

- 例として、広告費と売り上げの関係を調べたいとする。上の表記に従えば、y_{it} が売上高であり、$\mathbf{X}_{it}$ が広告費などの変数になる。しかし、単に売上高を広告費に回帰しても、広告費の効果を正しく推定できるとは思えない。規模の大きい企業ほど、売上高も広告費も大きいであろうから、もし、規模の大きさを示す変数が回帰に含まれていない場合は、欠落変数の問題が起こる。ここで、パネルデータがあり、たとえば資本や従業員数等の企業規模を決める変数が時間を通じて一定とする (上の表記のように $\mathbf{Z}_i$ と書ける場合である) と、広告費の効果を一致推定することができる。

このモデルは、n 個の切片のあるモデルと考えることもできる。$\alpha_i = \beta_0 + \beta_2' \mathbf{Z}_i$ とすると、

$$y_{it} = \beta_1' \mathbf{X}_{it} + \alpha_i + u_{it} \tag{2}$$

と書くことができる。これが、固定効果回帰モデルである。ただし、$\mathbf{X}_{it}$ の中に時間を通じて変化しない変数を含めることができないことに注意が必要である。もし含めてしまうと、α_i と多重共線性を起こしてしまうからである。

さらにこのモデルは n 個の 2 項変数からなるモデルとして書くこともできる。$D1_{it}$ を $i=1$ なら 1 をとり、$i \neq 1$ なら 0 をとる変数とする。$D2_{it}$ から Dn_{it} も同様に定義する。すると、モデルは

$$y_{it} = \beta_1' \mathbf{X}_{it} + \alpha_1 D1_{it} + \alpha_2 D2_{it} + \alpha_3 D3_{it} + \cdots + \alpha_n Dn_{it} + u_{it} \tag{3}$$

として書くことができる。D にも添え字 t がつけてあるが、実際には t に依存しない。定数項が含まれたままだと多重共線性の問題を起こすため、抜かれていることに注意。

パネルデータを扱う際のモデルには、固定効果モデル以外に変量効果モデルと呼ばれるものがある。これは、個別効果 $\mathbf{Z_i}$ が確率変数である場合で、特に u_{it} と相関をもたない場合は個別効果 $\mathbf{Z_i}$ を推定モデルに含めなくても OLS によって β_1 を一致推定できる。変量効果モデルも固定効果回帰によって一致推定が可能であり、この授業ではこれ以上は立ち入らない。

推定　上で書いた 2 項変数による固定効果回帰モデルの表現に基づいて、OLS での推定を行うことができる。

また、次のように推定量を求めることもできる。まず、$\bar{y}_i = \sum_{t=1}^{T} Y_{it}/T$ とする。$\bar{\mathbf{X}}_i$ と $\bar{u}_i$ も同じように定義する。すると、

$$\bar{y}_i = \beta_1' \bar{\mathbf{X}}_i + \alpha_i + \bar{u}_i \tag{4}$$

である。この各個人の平均の式を、それぞれの観測値の式から引くことにより、

$$y_{it} - \bar{y}_i = \beta_1' (\mathbf{X}_{it} - \bar{\mathbf{X}}_i) + (u_{it} - \bar{u}_i) \tag{5}$$

となる。固定効果が消えていることに注意すること。この変換を固定効果変換と呼ぶ。この式を、OLS で推定する。

二つの推定量が同じになることは以下のようにして証明できる。まず、2 項変数を入れたモデルによる最小二乗推定量は

$$\sum_{i=1}^{n} \sum_{t=1}^{T} (y_{it} - \beta' \mathbf{X}_{it} - \alpha_1 D1_{it} - \cdots - \alpha_n Dn_{it})^2 \tag{6}$$

を最小化する。ここで、

$$\sum_{i=1}^{n} \sum_{t=1}^{T} (y_{it} - \beta' \mathbf{X}_{it} - \alpha_1 D1_{it} - \cdots - \alpha_n Dn_{it})^2 = \sum_{i=1}^{n} \left(\sum_{t=1}^{T} (y_{it} - \beta' \mathbf{X}_{it} - \alpha_i)^2 \right) \tag{7}$$

である。β を固定したときに上の関数を最小化する γ_i の値は、標本平均が最小二乗推定量であるという結果を使うと、

$$\hat{\alpha}_i = \frac{1}{T} \sum_{t=1}^{T} (y_{it} - \beta' \mathbf{X}_{it}) = \bar{y}_i - \beta' \bar{\mathbf{X}}_i \tag{8}$$

ということがわかる。したがって、β の最小二乗推定量は、

$$\sum_{i=1}^{n} \left(\sum_{t=1}^{T} (y_{it} - \beta' \mathbf{X}_{it} - \bar{y}_i + \beta' \bar{\mathbf{X}}_i)^2 \right) = \sum_{i=1}^{n} \sum_{t=1}^{T} (y_{it} - \bar{y}_i - \beta' (\mathbf{X}_{it} - \bar{\mathbf{X}}_i))^2 \tag{9}$$

を最小化するものであるが、これは、$y_{it} - \bar{y}_i$ を $\mathbf{X}_{it} - \beta' \bar{\mathbf{X}}_i$ に回帰した時の最小二乗推定量である。

この推定量はいろいろな名称で呼ばれる。

- 固定効果推定量 (FE);
- 群間推定量 (WG);
- ダミー変数最小二乗推定量 (LSDV)。

4.3 時間固定効果

また、個人間で相違がないが、時間を通じて変わっていく変数から引き起こされる欠落変数のバイアスを避けたい場合もある。

例: 酒税と交通事故死の数の関係を調べる。酒税が上がれば飲酒が減って、交通事故死亡率が下がると予想される。他方、技術進歩によって自動車の安全性が向上しており、その効果を含めないと推定にバイアスが生ずるかもしれない。

時間固定効果を推定式に含めることでこの問題を取り扱うことができる。

$$y_{it} = \beta_0 + \beta_1' \mathbf{X}_{it} + \beta_2' \mathbf{Z}_i + \beta_4' \mathbf{S}_t + u_{it}. \tag{10}$$

- 時間固定効果のみのモデル:

$$y_{it} = \beta_1' \mathbf{X}_{it} + \lambda_t + u_{it} \tag{11}$$

とすると、λ_t は時間固定効果である。このモデルは、T 個の 2 項変数を使い表現できる。$B1_{it}$ を $t = 1$ なら 1 とし、$t \neq 1$ なら 0 である 2 項変数とする。$B2_{it} = 1$ は $t = 2$ なら 1 で $t \neq 2$ なら 0 な変数とする。他の変数も同じように定義する。

$$y_{it} = \beta_1' \mathbf{X}_{it} + \lambda_1 B1_{it} + \lambda_2 B2_{it} + \cdots + \lambda_T BT_{it} + u_{it}. \tag{12}$$

このモデルは最小二乗推定することができる。また各時点での平均を引くことで、時間固定効果を取り除くことができる。

$$y_{it} - \bar{y}_t = \beta_1' (\mathbf{X}_{it} - \bar{\mathbf{X}}_t) + u_{it} - \bar{u}_t. \tag{13}$$

ここで、$\bar{y}_t = \sum_{i=1}^n y_{it}/n$ である。$\bar{\mathbf{X}}_t$ と $\bar{u}_t$ も同様に定義する。

- 時間固定効果と個人固定効果の両方が入ったモデル:

$$y_{it} = \beta' \mathbf{X}_{it} + \alpha_i + \lambda_t + u_{it}. \tag{14}$$

やはり 2 項変数を使って

$$y_{it} = \beta_1' \mathbf{X}_{it} + \alpha_1 D1_{it} + \alpha_2 D2_{it} + \cdots + \alpha_n Dn_{it} + \lambda_2 B2_{it} + \cdots + \lambda_T BT_{it} + u_{it} \tag{15}$$

と表現できる。多重共線性を避けるために、$B1_{it}$ を式に含めていないことに注意すること。もちろん、$B1_{it}$ に限らず、どれか一つのダミー変数を抜けばよい。α_i と λ_t をモデルから除くためには、

$$y_{it} - \bar{y}_i - \bar{y}_t + \bar{y} = \beta' (\mathbf{X}_{it} - \bar{\mathbf{X}}_i - \bar{\mathbf{X}}_t + \bar{\mathbf{X}}) + u_{it} - \bar{u}_i - \bar{u}_t + \bar{u} \tag{16}$$

という変換を行うとよい。ここで、$\bar{y} = \sum_{i=1}^n \sum_{t=1}^T y_{it}/nT$ であり、$\bar{X}$ も同様に定義する。やはり、OLS で推定する。

4.4 固定効果推定量の漸近的性質

固定効果推定量の漸近的性質を導出する。ここでは、横断面 (クロスセクション) の標本数 n が比較的大きく、時系列の長さ T が比較的短い場合を考える。その場合、$n \to \infty$ だが T は固定しているような漸近理論を考える。

固定効果推定量のための仮定

1. $E(u_{it}|\mathbf{X}_{i1}, \mathbf{X}_{i2}, \ldots, \mathbf{X}_{iT}, \alpha_i) = 0$ (強外生の仮定と呼ばれる)。
2. $(\mathbf{X}_{i1}, \mathbf{X}_{i2}, \ldots, \mathbf{X}_{iT}, u_{i1}, u_{i2}, \ldots, u_{iT})$, $i = 1, \ldots, n$ は、個人間で i.i.d. である。
3. $(\mathbf{X}_{it}, u_{it})$ は、4 次までの有界なモーメントをもつ。
4. $\mathbf{X}_{it} - \bar{\mathbf{X}}_i$ に多重共線性がない。

注意点:

- 誤差項の条件付き平均をとる際には、“すべての t” の X の値について条件をとる。単に t での X の値だけ考えればよいわけではない。T 個の時点すべての X の値について条件付き平均を考えることは、平均を引いた回帰変数と平均を引いた誤差項が無相関であることを示すために重要である。
- 多重共線性に関する仮定により、$\mathbf{X}_{it}$ には時間を通じて変化する変数しか含めることができないことがわかる。また Bt_{it} は $\mathbf{X}_{it}$ に含めることができる。

漸近分布　固定効果推定量は次のように表現できる。

$$\hat{\beta}^{FE} = \left(\sum_{i=1}^{n} \sum_{t=1}^{T} (\mathbf{X}_{it} - \bar{\mathbf{X}}_i)(\mathbf{X}_{it} - \bar{\mathbf{X}}_i)' \right)^{-1} \sum_{i=1}^{n} \sum_{t=1}^{T} (\mathbf{X}_{it} - \bar{\mathbf{X}}_i)(y_{it} - \bar{y}_i). \tag{17}$$

$\hat{\beta}^{FE}$ の漸近分布は、簡単に導出でき、それは、

$$\sqrt{n}(\hat{\beta}^{FE} - \beta_1) \to_d N(\mathbf{0}, V_{FE}) \tag{18}$$

である。ただし、

$$V_{FE} = \left(E\left\{ \sum_{t=1}^{T} (\mathbf{X}_{it} - \bar{\mathbf{X}}_i)(\mathbf{X}_{it} - \bar{\mathbf{X}}_i)' \right\} \right)^{-1} E\left\{ \sum_{t=1}^{T} \sum_{s=1}^{T} u_{it} u_{is} (\mathbf{X}_{it} - \bar{\mathbf{X}}_i)(\mathbf{X}_{is} - \bar{\mathbf{X}}_i)' \right\} \tag{19}$$

$$\times \left(E\left\{ \sum_{t=1}^{T} (\mathbf{X}_{it} - \bar{\mathbf{X}}_i)(\mathbf{X}_{it} - \bar{\mathbf{X}}_i)' \right\} \right)^{-1} \tag{20}$$

である。また、$\sum_{t=1}^{T} (\mathbf{X}_{it} - \bar{\mathbf{X}}_i)(u_{it} - \bar{u}_i) = \sum_{t=1}^{T} (\mathbf{X}_{it} - \bar{\mathbf{X}}_i) u_{it}$ であることに注意すること。

標準誤差　$\hat{u}_{it} = y_{it} - \bar{y}_i - \hat{\beta}^{FE\prime}(\mathbf{X}_{it} - \bar{\mathbf{X}}_i)$ とする。漸近分散推定量は、

$$\hat{V}_{FE} = \left(\frac{1}{n} \sum_{i=1}^{n} \sum_{t=1}^{T} (\mathbf{X}_{it} - \bar{\mathbf{X}}_i)(\mathbf{X}_{it} - \bar{\mathbf{X}}_i)' \right)^{-1} \frac{1}{n} \sum_{i=1}^{n} \sum_{t=1}^{T} \sum_{s=1}^{T} \hat{u}_{it}\hat{u}_{is}(\mathbf{X}_{it} - \bar{\mathbf{X}}_i)(\mathbf{X}_{is} - \bar{\mathbf{X}}_i)' \quad (21)$$

$$\times \left(\frac{1}{n} \sum_{i=1}^{n} \sum_{t=1}^{T} (\mathbf{X}_{it} - \bar{\mathbf{X}}_i)(\mathbf{X}_{it} - \bar{\mathbf{X}}_i)' \right)^{-1} \quad (22)$$

を使うことができる。この漸近分散推定量を用いて、$\hat{\beta}^{FE}$ の標準誤差を計算することができる。

- この漸近分散推定量は、自己相関と分散不均一に頑健な漸近分散推定量と呼ばれる。クラスタリングに頑健な推定量とも呼ばれる。
- Eviews では、上で紹介した漸近分散推定量や標準誤差は、“White period” というオプションを使用することによって計算できる。
- もし、誤差項に系列相関がない (時間の異なる誤差項は無相関になっている) という仮定があると、漸近分散はもっと簡単に表現されて、

$$V_{FE} = \left(E\left\{ \sum_{t=1}^{T} (\mathbf{X}_{it} - \bar{\mathbf{X}}_i)(\mathbf{X}_{it} - \bar{\mathbf{X}}_i)' \right\} \right)^{-1} E\left\{ \sum_{t=1}^{T} u_{it}^2 (\mathbf{X}_{it} - \bar{\mathbf{X}}_i)(\mathbf{X}_{it} - \bar{\mathbf{X}}_i)' \right\} \quad (23)$$

$$\times \left(E\left\{ \sum_{t=1}^{T} (\mathbf{X}_{it} - \bar{\mathbf{X}}_i)(\mathbf{X}_{it} - \bar{\mathbf{X}}_i)' \right\} \right)^{-1} \quad (24)$$

  となる。しがし、Eviews では “White diagonal” というオプションで計算できる次の漸近分散推定量は、不一致になる。

$$\hat{V}_{FE}^{HR} = \left(\frac{1}{n} \sum_{i=1}^{n} \sum_{t=1}^{T} (\mathbf{X}_{it} - \bar{\mathbf{X}}_i)(\mathbf{X}_{it} - \bar{\mathbf{X}}_i)' \right)^{-1} \frac{1}{n} \sum_{i=1}^{n} \sum_{t=1}^{T} \hat{u}_{it}^2 (\mathbf{X}_{it} - \bar{\mathbf{X}}_i)(\mathbf{X}_{it} - \bar{\mathbf{X}}_i)' \quad (25)$$

$$\times \left(\frac{1}{n} \sum_{i=1}^{n} \sum_{t=1}^{T} (\mathbf{X}_{it} - \bar{\mathbf{X}}_i)(\mathbf{X}_{it} - \bar{\mathbf{X}}_i)' \right)^{-1}. \quad (26)$$

  不一致になる原因は、T が固定されている漸近理論では、α_i を一致性を持って推定することができないからである。この問題についてのより詳しい説明と、一致性があり、系列無相関の情報を生かした漸近分散推定量については、Stock and Watson (2008) を参照のこと。
- なお、自己相関と分散不均一に頑健な漸近分散推定量は、系列相関があってもなくても使用可能である。

4.5　動学的パネルデータモデル

パネルデータを用いることで、ある経済変数の系列相関の原因を理解することができる。たとえば、所得は時間を通じて相関している。つまり、今期高い所得を得る人は、来期も高い所得を得ることが多く、今期所得の低い人は来期も低い所得であることが多い。この現象には主に次の 2 つの解釈がなりたつ。

1. 異質性: 所得を得る能力は人ごとに異なり、系列相関はそうした人々の能力の違いの現れである。

2. 状態依存: ある人が偶然高い所得を得ることができた場合、その人の能力に関係なく、労働市場の硬直性などの原因によって、次の期も高い所得を得る可能性が高くなる。

これらの二つの解釈から得られる政策への含意は大きく異なる。

この問題をデータを使って調べるためには、動学的パネルデータモデルが有用である。代表的なモデルは、パネル自己回帰モデルであり、その一次のものは、

$$y_{it} = \beta y_{i,t-1} + \alpha_i + u_{it} \tag{27}$$

とかける。これをパネル AR(1) モデルと呼ぶ。ここで、y_{it} は観察できる確率変数で、その動学的性質に興味がある。また、α_i は観察できない個人効果である。誤差項 u_{it} は、時間を通じては相関していないと仮定する。

推定　固定効果推定量は、不一致になる。固定効果推定量が一致性を持つための条件には、強外生の条件というものがある。しかし、y_{it} と u_{it} は相関しているため、その条件はこのモデルでは満たすことができない。実際、固定効果変換をすると、回帰変数は $y_{i,t-1} - \sum_{s=0}^{T-1} \bar{y}_{i,s}/T$ となり、誤差項は $u_{it} - \sum_{s=1}^{T} u_{is}/T$ となるが、これらは相関している。

ここでは、Anderson and Hsiao (1981)　による推定量を考える。まず、一次の階差をとり、個人効果を取り除くと次のようになる。

$$y_{it} - y_{i,t-1} = \beta(y_{i,t-1} - y_{i,t-2}) + u_{it} - u_{i,t-1}. \tag{28}$$

$y_{it-1} - y_{i,t-2}$ と $u_{it} - u_{i,t-1}$ は相関しているが、$y_{i,t-2}$ と $u_{it} - u_{i,t-1}$ は無相関であるとわかる。そこで、$y_{i,t-2}$ を操作変数として使うことで、β の推定を行うことができる。推定量は

$$\hat{\beta}^{AH} = \frac{\sum_{i=1}^{n}\sum_{t=3}^{T} y_{i,t-2}(y_{it} - y_{i,t-1})}{\sum_{i=1}^{n}\sum_{t=3}^{T} y_{i,t-2}(y_{i,t-1} - y_{i,t-2})} \tag{29}$$

とかける。

また、2 期以上のラグをとった被説明変数はすべて操作変数として使用可能である。そうしたすべての操作変数を使用した推定量については、Arellano and Bond (1991) を参照のこと。これらの推定量は、EViews や STATA 等のソフトで計算することができる。

先決変数　同じような推定法は、一般に先決変数を含むモデルにおいても使用可能である。次のモデルを考える。

$$y_{it} = \beta x_{it} + \alpha_i + u_{it}. \tag{30}$$

- もし、$s \geq 0$ について $E(x_{it-s}u_{it}) = 0$ ならば、x_{it} は先決変数と呼ばれる。このとき、仮に $x_{i,t}$ と $u_{i,t}$ に相関があっても、$x_{i,t-s}$ を ＩＶに使うことによって β の一致推定が可能である。

ラグ付き被説明変数は $y_{i,t-1}$ は先決変数の代表的な例である。