パネルデータ分析 II

今回から 4 回連続でパネルデータ分析の中級レベルの分析手法をご紹介します.

- Population-Averaged 推定量
- 操作変数法とハウスマン · テイラーモデルの推定
- ダイナミックパネルデータモデルの推定
- 非線形パネルデータモデルの推定

1 Population-Averaged 推定量

パネルデータ分析における固定効果モデルとランダム効果モデルの推定については, その考え方と操作方法をご存知の方も多いと思います. 今回紹介する Population-Averaged 推定量の問題意識の一つは, パネルデータ分析における残差の自己相関にあります. 主に Cameron and Trivedi (2010) による "Microeconometrics Using Stata, Revised Edition", Stata Press の第 8 章 Linear panel-data models:Basics のセクション 3 と 4 の内容を利用して解説を行います.

1.1 プーリングモデル

最初にパネルデータを用いた簡単な線形モデルの推定から話を始めます. 次のようなプーリングモデルを考えます.

$$y_{it} = \alpha + x_{it}'\beta + u_{it} \tag{1}$$

これはランダム効果または固定効果を含みません. しかし, u_{it} には自己相関の存在を許容するものとします.

利用するサンプルデータには賃金関数の推定に必要な変数が含まれています.

```
. use mus08psidextract, clear
```

データの構造を確認します.

```
. xtdes
```

```
        id:  1, 2, ..., 595                                    n =        595
         t:  1, 2, ..., 7                                      T =          7
             Delta(t) = 1 unit
             Span(t)  = 7 periods
             (id*t uniquely identifies each observation)
Distribution of T_i:   min      5%     25%       50%       75%     95%     max
                         7       7       7         7         7       7       7
     Freq.  Percent    Cum. |  Pattern
------------------------------+---------
       595   100.00  100.00 |  1111111
------------------------------+---------
       595   100.00         |  XXXXXXX
```

個人の ID が 1 から 595 まであり, 時点は 7 つであることが分かります. また, 欠損値はありません.

1.2 プーリング推定

ここでは個人毎のデータを連続して計測しています. 1 式を推定した場合, 誤差項に自己相関が存在すると考えられますので, 次のようなオプションを利用してモデル推定を実行します.

```
. regress lwage exp exp2 wks ed, vce(cluster id)
```

```
Linear regression                               Number of obs     =      4,165
                                                F(4, 594)         =      72.58
                                                Prob > F          =     0.0000
                                                R-squared         =     0.2836
                                                Root MSE          =     .39082

                                   (Std. Err. adjusted for 595 clusters in id)
------------------------------------------------------------------------------
             |               Robust
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         exp |    .044675   .0054385     8.21   0.000     .0339941     .055356
        exp2 |  -.0007156   .0001285    -5.57   0.000    -.0009679   -.0004633
         wks |    .005827   .0019284     3.02   0.003     .0020396    .0096144
          ed |   .0760407   .0052122    14.59   0.000     .0658042    .0862772
       _cons |   4.907961   .1399887    35.06   0.000     4.633028    5.182894
------------------------------------------------------------------------------
```

誤差項に相関がないと考える場合の推定結果を次に示します.

```
. regress lwage exp exp2 wks ed
```

```
      Source |       SS           df       MS      Number of obs   =     4,165
-------------+----------------------------------   F(4, 4160)      =    411.62
       Model |  251.491445         4  62.8728613   Prob > F        =    0.0000
    Residual |  635.413457     4,160  .152743619   R-squared       =    0.2836
-------------+----------------------------------   Adj R-squared   =    0.2829
       Total |  886.904902     4,164  .212993492   Root MSE        =    .39082

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         exp |    .044675   .0023929    18.67   0.000     .0399838    .0493663
        exp2 |  -.0007156   .0000528   -13.56   0.000    -.0008191   -.0006121
         wks |    .005827   .0011827     4.93   0.000     .0035084    .0081456
          ed |   .0760407   .0022266    34.15   0.000     .0716754     .080406
       _cons |   4.907961   .0673297    72.89   0.000     4.775959    5.039963
------------------------------------------------------------------------------
```

両者の標準誤差は明らかに異なります.

1.3 パネルデータの系列相関

賃金関数の残差に系列相関は存在するのでしょうか?ここではプーリングモデルを reg コマンドで推定しましたので, コマンドの対応関係で言えば, Breusch–Godfrey の系列相関の検定コマンドなどを利用することになりますが, それらは時系列データ用のコマンドであり, パネルデータでは利用できません. そこで単純に相関を調べる correlate コマンドを利用することにします. 最初に被説明変数の自己相関を調べます.

```
. sort id t
. correlate lwage l.lwage
```

```
(obs=3,570)
             |                   L.
             |    lwage     lwage
-------------+------------------
       lwage |
         --. |   1.0000
         L1. |   0.9189   1.0000
```

対数変換した賃金には強い自己相関があることが分かります. 次はプーリングモデルの残差について見ることにします. 次に示すコマンドのうち, 2.34. は自動的に画面表示されるので, ユーザが入力する必要はありません. 最後の } を入力すると, ループコマンドの完了を Stata が認識して繰り返し計算を実行します.

```
. regress lwage exp exp2 wks ed, vce(cluster id)
. predict uhat, residuals
. forvalues j = 1/6 {
2.    quietly corr uhat L‘j’.uhat
3.    display "Autocorrelation at lag ‘j’ = " %6.3f r(rho)
4.    }
```

```
Autocorrelation at lag 1 =  0.884
Autocorrelation at lag 2 =  0.838
Autocorrelation at lag 3 =  0.811
Autocorrelation at lag 4 =  0.786
Autocorrelation at lag 5 =  0.750
Autocorrelation at lag 6 =  0.729
```

ここでは賃金関数の残差 uhat の系列相関をラグ 1 から 6 までの範囲で繰り返し調べています. 計算結果からは時点間隔が離れるに従って相関が少しずつ弱まっていく様子が分かります.

次は現時点との相関ではなく, ラグは 1 に固定し, 基準となる時点を 1 期づつ動かして調べてみます.

```
. forvalues s = 2/7 {
2.    quietly corr uhat L1.uhat if t == ‘s’
3.    display "Autocorrelation at lag 1 in year ‘s’ = " %6.3f r(rho)
4.    }
```

```
 Autocorrelation at lag 1 in year 2 =  0.915

Autocorrelation at lag 1 in year 3 =  0.799
Autocorrelation at lag 1 in year 4 =  0.855
Autocorrelation at lag 1 in year 5 =  0.867
Autocorrelation at lag 1 in year 6 =  0.894
Autocorrelation at lag 1 in year 7 =  0.893
```

こちらは, 最初にやや減少しますが, 平均すると 0.87 程度の相関が保たれていることが分かります.

1.4 ランダム効果モデルの誤差相関

ランダム効果モデルを次に示します.

$$y_{it} = \alpha_i + x_{it}^{'}\beta + \epsilon_{it} \tag{2}$$

2 式の α_i は, 説明変数とは無相関であるという強い仮定が存在します. ここで $u_{it} = \alpha_i + \epsilon_{it}$ であり, α_i は平均 0 で分散を σ_α^2 とする i.i.d. であり, ϵ_{it} の分散は σ_ϵ^2 とします. したがって, $\mathrm{Var}\,(u_{it}) = \sigma_\alpha^2 + \sigma_\epsilon^2$, 共分散は $\mathrm{Cov}\,(u_{it}, u_{is}) = \sigma_\alpha^2, s \neq t$ となります. ランダム効果モデルでは次の相関のことを級内相関 (Intraclass correlation) と呼びます.

$$\rho_u = \mathrm{Cor}\,(u_{it}, u_{is}) = \sigma_\alpha^2 / \left(\sigma_\alpha^2 + \sigma_\epsilon^2\right), \text{ for all } s \neq t$$

つまり, ランダム効果モデルでは系列相関が許容され, 撹乱項の分散に比べ, ランダム効果の分散は大きいほど 1 に近くなります. ただし, この式からも明らかなようにランダム効果の級内相関は, 任意の時点間ですべて等しいことを前提としています.

Pooled OLS 推定量

2 式を次のように書き換えます.

$$y_{it} = \alpha + x_{it}^{'}\beta + (\alpha_i - \alpha + \epsilon_{it}) \tag{3}$$

3 式の推定値が一致推定量であるためには, 誤差項 $(\alpha_i - \alpha + \epsilon_{it})$ と x_{it} が無相関でなければなりません. つまり, 固定効果モデルとランダム効果モデルを推定し, ランダム効果モデルが支持される場合のみ, プーリング推定の推定値は一致性を持つことになります.

1.5 Pooled FGLS/population-averaged 推定量

しかし, OLS の代わりに Pooled FGLS という推定手法を利用すると, 効率性に優れていることが明らかになっています. つまり, 説明変数と誤差項に相関が無い状況では, Pooled FGLS を用いると, 一致性があり, 効率性に優れた推定量が得られます. また, **線形モデルの場合に限り**, ランダム効果モデルと PA モデルの推定量は漸近的に一致することが分かっています. ただし, PA モデルの場合は, 先のプーリング推定の項で紹介した誤差項の自己相関をモデル推定時に研究者が仮定できるという特徴があります.

1.6 xtreg,pa コマンド

PA モデルの誤差項において自己相関を考える場合, ユーザは誤差項について $T \times T$ の行列 (working matrix) を想定し, 制約を課す必要があります.

- corr(independent):$\rho_{ts} = 0$ for $s \neq t$. 異時点間での相関は無いものとします. PA モデルの推定量は pooling OLS 推定量に等しくなります.
- corr(exchangeable):$\rho_{ts} = \rho$ for $s \neq t$. 異時点間の相関はすべて等しいものと考えます. これはランダム効果モデルの考え方です. したがって, xtreg,pa と xtreg,re は漸近的に等しくなります.

- corr(unstructured):制約は課しません. T が短い場合には優れた推定量を提供しますが, 逆に T が大きくなると計算上の問題が生じることがあります.

xtreg,pa コマンドは xtgee コマンドで family(gaussian) オプションを利用した場合と同じ結果をもたらします. 非線形モデルで利用価値の高い xtgee コマンドは, また別の回に詳しく解説します.

1.7 線形モデルにおける xtreg,pa コマンドの利用例

例えば, AR(2) 過程を考慮してモデルを推定します.

```
. xtreg lwage exp exp2 wks ed, pa corr(ar 2) vce(robust) nolog
```

```
GEE population-averaged model                   Number of obs     =      4,165
Group and time vars:                 id t       Number of groups  =        595
Link:                              identity     Obs per group:
Family:                            Gaussian                   min =          7
Correlation:                          AR(2)                   avg =        7.0
                                                              max =          7
                                                Wald chi2(4)      =     873.28
Scale parameter:                   .1966639     Prob > chi2       =     0.0000
                                      (Std. Err. adjusted for clustering on id)
```

lwage	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
exp	.0718915	.003999	17.98	0.000	.0640535	.0797294
exp2	-.0008966	.0000933	-9.61	0.000	-.0010794	-.0007137
wks	.0002964	.0010553	0.28	0.779	-.001772	.0023647
ed	.0905069	.0060161	15.04	0.000	.0787156	.1022982
_cons	4.526381	.1056897	42.83	0.000	4.319233	4.733529

自己相関の AR(2) を仮定してモデル推定した結果, pooled OLS の推定値とはかなり異なる結果を得ました. ed を除く, その他の標準誤差は小さくなり, 効率的な推定量が得られました. この時の相関係数を調べてみると,

```
. matrix list e(R)
```

```
. matrix list e(R)
symmetric e(R)[7,7]
           c1         c2         c3         c4         c5         c6         c7
r1          1
r2  .89722058          1
r3  .84308581  .89722058          1
r4  .78392846  .84308581  .89722058          1
r5  .73064474  .78392846  .84308581  .89722058          1
r6   .6806209  .73064474  .78392846  .84308581  .89722058          1
r7  .63409777   .6806209  .73064474  .78392846  .84308581  .89722058          1
```

異時点間で調べた pooled OLS の相関は 0.88, 0.84, 0.81, 0.79, 0.75,0.73 でした.

ここまでのまとめ

ここまで解説したポイントを次にまとめておきます.

1. xtreg,pa コマンドはプーリング推定において一致推定量を提供する.
2. xtreg,pa コマンドは Pooled OLS 推定量よりも有効性の点で優れている.
3. 線形モデルの場合, xtreg,pa と xtreg,re の推定量は漸近的に等しくなる.
4. xtreg,pa は xtgee コマンドで family(gaussian) オプションを利用したものと等しい.

1.8 非線形モデルにおける population-averaged 推定量

ここからは非線形モデルにおける population-averaged 推定量の解説になります。次のタイトルでインターネット上に公開されている記事を用いて, xtreg コマンドの population-averaged 推定量の考え方を解説します. [1]

What is the difference between random-effects and population-averaged estimators?

Title	Comparing RE and PA models
Author	William Sribney, StataCorp
Date	January 1999; updated June 2013

ロジットモデルを用いた数値実験を通して, pa オプションを利用したときの考え方を理解することが目的です. 最初にランダム効果モデル (クラスタを考慮した推定量) を利用する場合の式を示します.

$$\Pr\left(Y_{ij}=1|X_{ij},u_i\right)=\mathrm{F}\left(X_{ij}b+u_i\right)$$

これに対し, PA(population-averaged) モデルは,

$$\Pr\left(Y_{ij}=1|X_{ij}\right)=\mathrm{G}\left(X_{ij}b^*\right)$$

b と b^* はそれぞれのモデルの母数です. データを用いた両者の推定では, 往々にして推定値は近しい値になります. ランダム効果モデルを真とすると, PA モデルは母集団の分布を完全に反映したものにはなりません. これは変量効果の無い, 周辺分布をモデル化するものです. 一方, クラスタを考慮したランダム効果モデルは母集団の分布の情報を完全に有しています. ここで変量効果 u_i は所与と考えます. そこで, 数値実験を使って, 両者の違いを直感的に説明します.

1.9 logit を用いた例題

変数を次のように設定します. X_{ij} も二値変数にするのは話が簡単という単純な理由からです.

被説明変数 Y_{ij}	:	就業中/失業中
説明変数 X_{ij}	:	既婚/独身

[1] http://www.stata.com/support/faqs/statistics/random-effects-versus-population-averaged/

クラスタを考慮した (cluster-specific) モデルは,

$$\text{logit}\ \ \Pr\left(Y_{ij}=1|X_{ij},u_i\right)=a+X_{ij}b+u_i$$

オッズ比は,

$$\text{OR}_{cs}=\frac{\Pr\left(Y_{ij}=1|X_{ij}=1,u_i\right)/\Pr\left(Y_{ij}=0|X_{ij}=1,u_i\right)}{\Pr\left(Y_{ij}=1|X_{ij}=0,u_i\right)/\Pr\left(Y_{ij}=0|X_{ij}=0,u_i\right)}=\exp\left(b\right)$$

これは既婚であることによって, 就業できるオッズがどのように変化するか表現するものですが, この時の基準は「ある人 i さん」です. i さんの婚姻状態 X_{ij} は時点 j によって変化すると考えます. ランダム効果モデルは個人のオッズ比の変化に注目することになります.

次に PA モデルについて考えます.

$$\text{logit}\ \ \Pr\left(Y_{ij}=1|X_{ij}\right)=a+X_{ij}b^*$$

オッズ比は,

$$\text{OR}_{pa}=\frac{\Pr\left(Y_{ij}=1|X_{ij}=1\right)/\Pr\left(Y_{ij}=0|X_{ij}=1\right)}{\Pr\left(Y_{ij}=1|X_{ij}=0\right)/\Pr\left(Y_{ij}=0|X_{ij}=0\right)}=\exp\left(b^*\right)$$

となりますが, これは変量効果を含まない平均的な既婚者が就業できるオッズと, 平均的な独身者が就業できるオッズの比をとったものになっています. ここでは平均的という言葉を使っていますが, これはランダムに選択した既婚者の就業オッズと, 同じくランダムに選択した独身者の就業オッズを比較すると言い換えることができます.

ここからは数値実験になります. 次のようなデータを考えます. これは標本でなく, 5 人の要素からなる母集団そのものとします.

i	j	X_{ij}	u_i	Z_{ij}	Pr_{cs}	Pr_{pa}
1	1	0	−0.2	−0.10	0.4750	0.5249
1	2	1	−0.2	0.50	0.6225	0.6674
2	1	0	−0.1	0.00	0.5000	0.5249
2	2	1	−0.1	0.60	0.6457	0.6674
3	1	0	0.0	0.10	0.5250	0.5249
3	2	1	0.0	0.70	0.6682	0.6674
4	1	0	0.1	0.20	0.5498	0.5249
4	2	1	0.1	0.80	0.6900	0.6674
5	1	0	0.2	0.30	0.5744	0.5249
5	2	1	0.2	0.90	0.7109	0.6674

ここで設定値は $a=0.1, b=0.6$ であり, $Z_{ij}=a+bX_{ij}+u_i$ とします. u_i は所与とします. 変量効果を考慮した Pr_{cs} の計算は次に示す 4 式を用いて行います.

$$\text{Pr}_{cs}=\frac{\exp\left(Z_{ij}\right)}{1+\exp\left(Z_{ij}\right)} \tag{4}$$

次に 5 式を用いて母集団 (population) に対して, PA 確率 $\Pr_{pa}$ を計算します. これは X_{ij} の値 (0,1) ごとに $\Pr_{cs}$ の平均 (average) を求めたものになります. つまり, $\Pr_{cs}$ あっての $\Pr_{pa}$ という訳です.

$$\begin{aligned}\Pr_{pa}(X_{ij}=1) &= \frac{1}{5}\times\sum_{i=1}^{5}\Pr_{cs}(x_{ij}=1) \\ &= \frac{1}{5}\times(0.6225+0.6457+0.6682+0.6900+0.7109) \\ &= 0.6674\end{aligned} \tag{5}$$

同じ要領で,

$$\Pr_{pa}(X_{ij}=0) = 0.5249$$

ここで利用した 5 つの $\Pr_{cs}$ に対してオッズ比は,

$$OR = \frac{\Pr_{cs}(x_{ij}=1)/(1-\Pr_{cs}(x_{ij}=1))}{\Pr_{cs}(x_{ij}=0)/(1-\Pr_{cs}(x_{ij}=0))}$$

として計算できますので, 各人で計算してみると,

$$\begin{aligned}&\text{Subject1:} \quad (0.6225/(1-0.6225))/(0.4750/(1-0.4750)) = 1.8221 \\ &\text{Subject2:} \quad (0.6457/(1-0.6457))/(0.5000/(1-0.5000)) = 1.8221 \\ &\text{Subject3:} \quad (0.6682/(1-0.6682))/(0.5250/(1-0.5250)) = 1.8221 \\ &\text{Subject4:} \quad (0.6900/(1-0.6900))/(0.5498/(1-0.5498)) = 1.8221 \\ &\text{Subject5:} \quad (0.7109/(1-0.7109))/(0.5744/(1-0.5744)) = 1.8221\end{aligned}$$

となり, すべて同じ値になります. これはランダム効果モデルや PA モデルかに関係なく, ロジスティック曲線を利用した時のオッズ比の特徴です. 実際, X_{ij} の係数に $b=0.6$ という設定値を利用してデータを作成したので,

$$OR = \exp(b) = \exp(0.6) = 1.8221$$

となることが分かります.

次に $\Pr_{pa}$ についてオッズ比を考えてみます. $\Pr_{pa}$ はデータに関係なく, 平均をとった値ですので,

$$\exp(b^*) = (0.6674/(1-0.6674))/(0.5249/(1-0.5249)) = 1.8169$$

つまり,

$$b^* = 0.5972$$

となります. 結果的に変量効果を考慮した $b=0.6$ と, 個体毎のアウトカムの平均を利用した PA モデルの $b^*=0.5972$ は非常に近しいことが分かります. xtreg,pa オプションについてもう少し理解を深めたいユーザは, 先ほどの数値実験を Excel を利用して是非, 行ってみてください.

次回は計量分析における操作変数法の一般的な考え方を解説し, パネルデータ分析における応用例であるハウスマン・テイラー法という推定手法を紹介します.

2016 年 9 月
株式会社 ライトストーン