Stata によるパネルデータ分析

一橋大学経済研究所

松浦寿幸

１．パネルデータとは

1.1. パネルデータ分析の考え方

表１－１のように、ある個体（たとえば、No.1）だけを時間軸で追跡したデータセットを時系列データ、ある１時点だけを取り出したデータセットをクロスセクション（横断面）データと呼びます。さらに、点線で囲まれた領域のように、時系列データとクロスセクションデータを組み合わせたデータセットをパネルデータと呼びます。パネルデータを利用することのメリットは、（１）個体の異質性をコントロールできる、（２）時間を通じた変化を評価することができるので因果関係の検証に適している、（３）サンプル数が増える、などの特徴があります。なお、個体が追跡できない複数時点のクロスセクションデータの場合（個体ごとに時系列でデータを追跡できない場合）、プーリングデータと呼びます。

表１－１

		1990	1991	1992	…	1995	…	1998	1999	2000
個体	1	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX
	2	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX
	3	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX
	4	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX
	5	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX
	6	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX
	⋮	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX
	n	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX

1.2. データセット利用のメリット

では、パネルデータの利用によりどんなメリットが期待できるのでしょうか。具体的な事例として、賃金（Y_i）と自己啓発費（X_i）の関係について考えてみましょう。労働者の自己啓発活動は、資格取得などを通じて仕事能力を高め、所得の上昇をもたらすと考えられます。この２変数の関係を計測するためには、下記のような回帰式を推定する方法が考えられます。

$$Y_i = a + b*X_i + u_i$$

しかし、そもそも、自己啓発活動に積極的な人もいれば消極的な人もいるので、仮に企業が強制的に自己啓発活動を義務付けても、全従業員の仕事能力の向上はあまり期待できないかもしれません。つまり、自己啓発費以外に賃金に影響を与える要因として、個人の

学習意欲という要因が重要な役割を担っていると考えられます。学習意欲が高い人は、仕事の習熟も速いため、賃金も高くなると考えられます。したがって、学習意欲を Z_i とすると、賃金関数は以下の数式で表されます。

$$Y_i = a + b*X_i + Z_i + u_i \qquad (1)$$

一般的に、自己啓発費と学習意欲の間には正の相関があると考えられます。というのは、学習意欲が高い人ほど自己啓発に費用を投入する、あるいは、自己啓発費が高い人ほど学習意欲が上昇するという因果関係が存在するからです。こういった相関関係を線形方程式で表したものが以下の式です。

$$Z_i = c*X_i + v_i \qquad (2)$$

学習意欲 Z_i は、通常数値として観察できない変数ですので、学習意欲 Z_i を無視して賃金関数を計測すると、推定されるパラメータは以下のような関係になります。

$$Y_i = a + (b+c)*X_i + (v_i + u_i) \qquad (3)$$

本当に知りたい効果は、自己啓発費が賃金に与える効果である b ですが、実際に計測されるのは、$b+c$ になります。

こういった状況で、パネルデータが利用可能であり、かつ、学習意欲 Z_i が時間に依存しないとすると、b を計測することができます。

$$Y_{it} = a + b*X_{it} + Z_i + u_{it} \qquad (4)$$

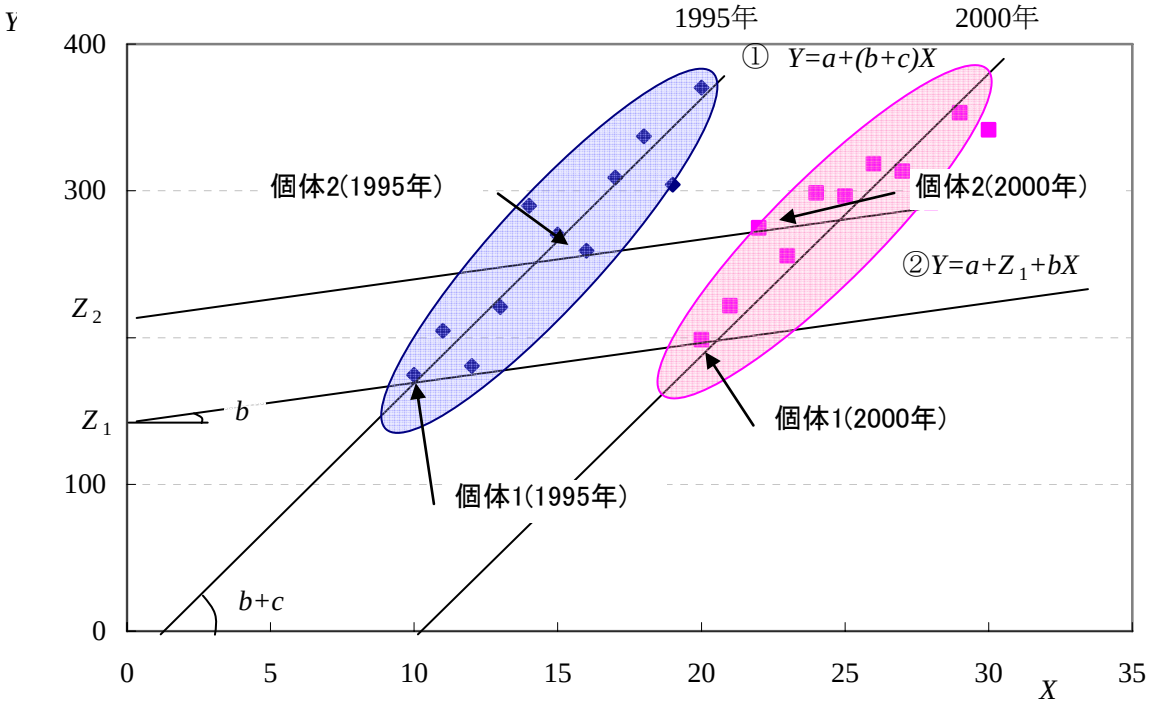
1995 年と 2000 年の自己啓発費（X）と賃金（Y）の散布図である図１－１で、（３）式と（４）式の関係をみてみましょう。図中の①式は、1995 年のクロスセクションデータにおける X と Y の関係を示すものですが、このときの傾きは $b+c$ になります。しかし、２時点以上のデータが利用できる場合、たとえば、図中の②式のように各個体で切片が異なると仮定し、個体ごとの自己啓発費（X）と賃金（Y）の値を結ぶことで、傾き b を計測することができます。実際の推定では、以下のように

このとき、Y と X の階差をとると、

$$\Delta Y_{it} = b*\Delta X_{it} + \Delta u_{it} \qquad (5)$$

となって、観察不可能な変数である学習意欲 Z_i は消えてしまいます。

図１－１



このほかにも、階差をとる代わりに、個体ごとのダミー変数を追加する方法でも同じ係数が得られることが知られています。

$$Y_{it} = a + b*X_{it} + \Sigma\, d_i*D_i + u_{it} \qquad (6)$$

この式の係数 d_i は、学習意欲を含む個人特有の仕事能力を示す係数であると考えられます。
この式の階差をとると、

$$\triangle Y_{it} = b*\triangle X_{it} + \triangle u_{it} \qquad (7)$$

という式が得られるので、ここで計測される b は、（５）式と等しいことが分かります。
この推定式を、固定効果モデル（Fixed Effect Model）と呼びます。
さらに、（４）式の両辺を個体ごとに合計すると

$$\Sigma_t Y_{it} = T*a + b*\Sigma_t X_{it} + T*Z_i + \Sigma_t u_{it} \qquad (8)$$

さらに両辺を T で割ると、

$$\overline{Y}_i = a + b * \overline{X}_i + Z_i + \overline{u}_i \qquad (9)$$

を得る。ただし、ここで、$\overline{Y}_i = \frac{1}{T}\sum_t Y_{it}$ 、 $\overline{X}_i = \frac{1}{T}\sum_t X_{it}$ 、 $\overline{u}_i = \frac{1}{T}\sum_t u_{it}$ とする。

（９）式を（４）式から引くと、

$$Y_{it} - \overline{Y}_i = a + b * \left(X_{it} - \overline{X}_i\right) + \overline{u}_i \qquad (10)$$

を得ます。この式も階差をとると、

$$\triangle Y_{it} = b * \triangle X_{it} + \triangle u_{it}$$

という式が得られるので、ここで計測される b は、（５）式と等しいことが分かります。

（１０）式の係数を特に、グループ内変動モデル(Within Estimator Model)と呼ぶことがあります。

なお、（２）式で、X と Z が相関しないと仮定し、Z を確率的な要因として誤差項 u に含めて推定する方法もあります。こちらの推定方法を変動効果モデル（Random Effect Model）と呼びます。変動効果モデルの詳細については、本書のレベルを超えますので割愛しますが、その特徴をまとめておきます。

- 個体固有効果 Z_i が確率的であると仮定されている
- 係数 b は、プーリング回帰モデルと固定効果モデルの中間になる
- データの時間軸が長くなればなるほど、係数 b は固定効果モデルに近くなる。

実際の推定に際して、変量効果モデルと固定効果モデルのいずれを用いるべきかは、ハウスマン検定（Hausman Specification Test）を用います。ハウスマン検定は、

帰無仮説：変量効果モデルが正しい
対立仮説：変量効果モデルは謝り（固定効果モデルが正しい）

という仮説の下、帰無仮説が正しいかどうかを調べることにより、変量効果モデルと固定効果モデルのいずれが望ましいかを判別します。

1.3. データセットの構造と作成方法

パネルデータを使って実際に分析する際のデータセットは、以下の表１－２のような構造になっています。ここで Y は、企業利益や賃金、家計所得のように個体により、時間により変化する変数です。X_1 は個々の個体により異なるが時間を通じて一定である変数、X_2

のように時間を通じてのみ変化する変数です。X_1 の例としては、たとえば各労働者の教育水準、X_2 は失業率などが考えられます。Z は、あるグループ（個体 1 と個体 2、個体 3 と個体 4）ごとに異なる変数で、かつ時間を通じて変化する変数です。Z の例としては、個体 1 と個体 2 が男性、個体 3 と個体 4 が女性とすれば、男性の平均賃金、女性の平均賃金などがこれにあてはまるでしょう。このデータセットを、破線で区切られたブロックごとにみていくと、クロスセクションデータを積み上げた形式になっていることが分かります。

表１－２

個体識別番号	年月	変数(1)	変数(2)	属性(1)	属性(2)	属性(3)
1	1990	$Y^1_{1,1990}$	$Y^1_{2,1990}$	X_1^1	$X_{2,1990}$	$Z_{(1)1990}$
1	1991	$Y^1_{1,1991}$	$Y^1_{2,1991}$	X_1^1	$X_{2,1991}$	$Z_{(1)1991}$
1	⋮	⋮	⋮	⋮	⋮	⋮
1	1995	$Y^1_{1,1995}$	$Y^1_{2,1995}$	X_1^1	$X_{2,1995}$	$Z_{(1)1995}$
2	1990	$Y^2_{1,1990}$	$Y^2_{2,1990}$	X_1^2	$X_{2,1990}$	$Z_{(1)1990}$
2	1991	$Y^2_{1,1991}$	$Y^2_{2,1991}$	X_1^2	$X_{2,1991}$	$Z_{(1)1991}$
2	⋮	⋮	⋮	⋮	⋮	⋮
2	1995	$Y^2_{1,1995}$	$Y^2_{2,1995}$	X_1^2	$X_{2,1995}$	$Z_{(1)1995}$
3	1990	$Y^3_{1,1990}$	$Y^3_{2,1990}$	X_1^3	$X_{2,1990}$	$Z_{(2)1990}$
3	1991	$Y^3_{1,1991}$	$Y^3_{2,1991}$	X_1^3	$X_{2,1991}$	$Z_{(2)1991}$
3	⋮	⋮	⋮	⋮	⋮	⋮
3	1995	$Y^3_{1,1995}$	$Y^3_{2,1995}$	X_1^3	$X_{2,1995}$	$Z_{(2)1995}$
4	1990	$Y^4_{1,1990}$	$Y^4_{2,1990}$	X_1^4	$X_{2,1990}$	$Z_{(2)1990}$
4	1991	$Y^4_{1,1991}$	$Y^4_{2,1991}$	X_1^4	$X_{2,1991}$	$Z_{(2)1991}$
4	⋮	⋮	⋮	⋮	⋮	⋮
4	1995	$Y^4_{1,1995}$	$Y^4_{2,1995}$	X_1^4	$X_{2,1995}$	$Z_{(2)1995}$

通常、パネルデータを用いた計量分析では、表１－２のようなデータセットが用いられます。（以下、LONG 形式のデータと呼びます。）このデータでは、個体と時間が縦方向に接続されたデータセットになっていますが、データ作成段階では、次の表１－３のように縦方向に並んだ個体データに、時系列データが横方向に接続されているデータセットが用いられることもあります。（以下、WIDE 形式のデータと呼びます。）

表１－３

個体識別番号	変数(1) 1990	変数(1) 1991	変数(2) 1990	変数(2) 1991
1	$Y^1_{1,1990}$	$Y^1_{1,1991}$	$Y^1_{2,1990}$	$Y^1_{2,1991}$
2	$Y^2_{1,1990}$	$Y^2_{1,1991}$	$Y^2_{2,1990}$	$Y^2_{2,1991}$
3	$Y^3_{1,1990}$	$Y^3_{1,1991}$	$Y^3_{2,1990}$	$Y^3_{2,1991}$
⋮	⋮	⋮	⋮	⋮
n	$Y^n_{1,1990}$	$Y^n_{1,1991}$	$Y^n_{2,1990}$	$Y^n_{2,1991}$

表１－２や表１－３のようなデータセットを用意するのは、若干煩雑な手間がかかります。というのは通常、データセットは、*Y* のような時間・個体によって異なる変数については、年次ごとのクロスセクションデータ（表１－４）、または、個体ごとの時系列データ（表１－５）として用意されていることが少なくありません。

表1－4

個体識別番号	年月	変数(1)
1	1990	$Y^1_{1,1990}$
2	1990	$Y^2_{1,1990}$
3	1990	$Y^3_{1,1990}$
4	1990	$Y^4_{1,1990}$

個体識別番号	年月	変数(1)
1	1991	$Y^1_{1,1991}$
2	1991	$Y^2_{1,1991}$
3	1991	$Y^3_{1,1991}$
4	1991	$Y^4_{1,1991}$

個体識別番号	年月	変数(1)
1	1995	$Y^1_{1,1995}$
2	1995	$Y^2_{1,1995}$
3	1995	$Y^3_{1,1995}$
4	1995	$Y^4_{1,1995}$

表1－5

個体識別番号	年月	変数(1)
1	1990	$Y^1_{1,1990}$
1	1991	$Y^1_{1,1991}$
1	⋮	⋮
1	1995	$Y^1_{1,1995}$

個体識別番号	年月	変数(1)
2	1990	$Y^2_{1,1990}$
2	1991	$Y^2_{1,1991}$
2	⋮	⋮
2	1995	$Y^2_{1,1995}$

個体識別番号	年月	変数(1)
3	1990	$Y^3_{1,1990}$
3	1991	$Y^3_{1,1991}$
3	⋮	⋮
3	1995	$Y^3_{1,1995}$

個体識別番号	年月	変数(1)
4	1990	$Y^4_{1,1990}$
4	1991	$Y^4_{1,1991}$
4	⋮	⋮
4	1995	$Y^4_{1,1995}$

しかも、Y_1とY_2が別のファイルで提供されることもあります（表１－６）。

表1－6

個体識別番号	年月	変数(1)
1	1990	$Y^1_{1,1990}$
1	1991	$Y^1_{1,1991}$
1	⋮	⋮
1	1995	$Y^1_{1,1995}$
2	1990	$Y^2_{1,1990}$
2	1991	$Y^2_{1,1991}$
2	⋮	⋮
2	1995	$Y^2_{1,1995}$
3	1990	$Y^3_{1,1990}$
3	1991	$Y^3_{1,1991}$
3	⋮	⋮
3	1995	$Y^3_{1,1995}$
4	1990	$Y^4_{1,1990}$
4	1991	$Y^4_{1,1991}$
4	⋮	⋮
4	1995	$Y^4_{1,1995}$

個体識別番号	年月	変数(2)
1	1990	$Y^1_{2,1990}$
1	1991	$Y^1_{2,1991}$
1	⋮	⋮
1	1995	$Y^1_{2,1995}$
2	1990	$Y^2_{2,1990}$
2	1991	$Y^2_{2,1991}$
2	⋮	⋮
2	1995	$Y^2_{2,1995}$
3	1990	$Y^3_{2,1990}$
3	1991	$Y^3_{2,1991}$
3	⋮	⋮
3	1995	$Y^3_{2,1995}$
4	1990	$Y^4_{2,1990}$
4	1991	$Y^4_{2,1991}$
4	⋮	⋮
4	1995	$Y^4_{2,1995}$

また、X_1のように個々の個体により異なるが時間を通じて一定である変数やX_2のように時間を通じてのみ変化する変数は、表１－７のようなクロスセクションデータ、もしくは表１－８のような時系列データとして、グループごとにことなる変数 Z は、表１－９のような時系列データの形式で提供されることがあります。

表1－7

個体識別番号	属性(1)
1	X_1^1
2	X_1^2
3	X_1^3
4	X_1^4

表1－8

年月	属性(2)
1990	$X_{2,1990}$
1991	$X_{2,1991}$
⋮	⋮
1995	$X_{2,1995}$

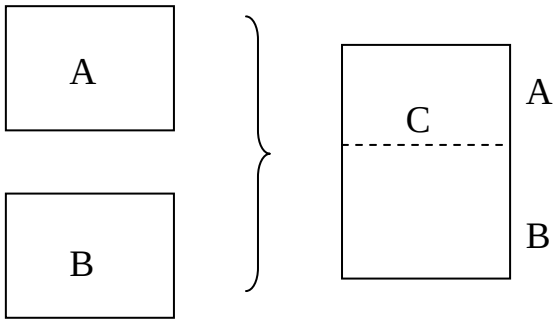
表１－９

グループ１（個体1、個体2）		グループ２（個体3、個体4）	
年月	変数	年月	変数
1990	$Z_{(1)1990}$	1990	$Z_{(2)1990}$
1991	$Z_{(1)1991}$	1991	$Z_{(2)1991}$
⋮	⋮	⋮	⋮
1995	$Z_{(1)1995}$	1995	$Z_{(2)1995}$

このように異なるデータファイルを統合し、分析用のデータセットを作成するためには、以下のような手順でデータを接続していきます。

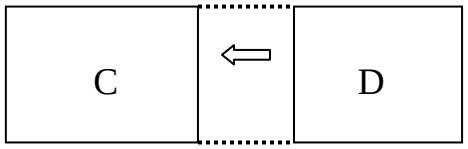
１）クロスセクションデータ・時系列データの縦方向の接続（表１－４、表１－５）

縦に接続する場合　⇒　append コマンド

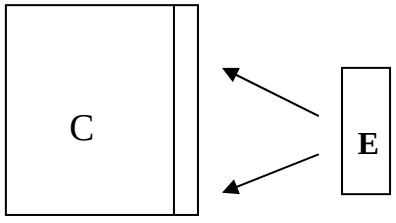


２）時間・個体によって異なる変数の接続（表１－６）

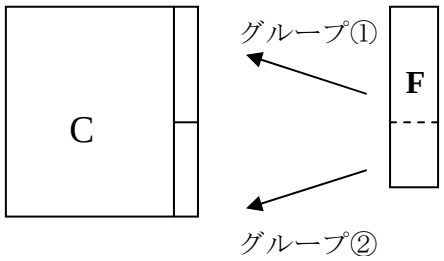
縦に接続する場合　⇒　merge コマンド



３）時間に依存しないクロスセクションデータ、個体に依存しない時系列データの接続（表１－７、表１－８）



４）グループ別の時系列データの接続（表１－９）



1.4. Stata によるデータセットの作成方法

ここでは、都道府県別 3 節の最後に１）～４）で示されたデータセットのパターンごとに Stata の操作方法を交えつつ、具体的にその方法を説明します。

１）データの縦方向の結合

まず、はじめに、複数の個体ごとのデータファイルを結合する方法を考えます。例として、都道府県ごとにファイルされたデータを一つにまとめる方法について考えましょう。

Hokkaido.dta

prefecture	year	production
1	1980	1200
1	1981	1310
1	1982	1450
⋮	⋮	⋮
1	2000	2560

Aomori.dta

prefecture	year	production
2	1980	800
2	1981	710
2	1982	1050
⋮	⋮	⋮
2	2000	1420

この２つのファイルを結合させる場合、append コマンドを用います。使用方法としては、一方のファイルを開いた状態で、もう一方のファイルを append で呼び出します。
具体的には以下のようになります。（２つのファイルは D:¥Data にあるとします。）

```
cd D:¥Data
use Hokkaido.dta
append Aomori.dta
save Production80-00.dta
```

完成したファイルは以下のようになります。

Prefecture	Year	Production
1	1980	1200
1	1981	1310
（省略）		
1	2000	2560
2	1980	800
2	1981	710
（省略）		
2	2000	1420

append を使う際の注意点として、必ず共通の変数には同じ変数名を付けておいてください。

２）時間・個体によって異なる変数の接続

次に、複数の個体のデータが変数ごとに各々のファイルに収録されている場合に、データを結合させる例を考えてみましょう。例として、都道府県別の生産額のデータに都道府県別の賃金のデータを接続する方法を考えます。

wage.dta

prefecture	wage
1	3.616
2	2.644
3	3.522
4	3.631
5	3.348
6	3.517
7	3.928
8	5.337
9	4.912
⋮	⋮
47	2.687

production.dta

prefecture	production
1	18954421
2	4634405
3	4678288
4	8429719
5	3901386
6	4095372
7	7692465
8	11374471
9	7739373
⋮	⋮
47	3268545

まず、接続する２つのファイルをキーとなる変数で sort しておく必要があります。wage.dta からみてみましょう。（二つのファイルは、D ドライブの Data フォルダーにあるものとします。）

```
. cd D:¥Data   (D ドライブ、Data フォルダーに移動)
. des

Contains data from D:¥Data¥Wage.dta
  obs:            47
 vars:             2                          21 Apr 2004 21:58
 size:           423 (99.9% of memory free)
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
prefecture      byte   %8.0g
wage            float  %9.0g
```

この場合、”Sorted by”の後ろに何も示されていないので、まだ sort されていないことがわかります。そこで、

```
sort prefecture
```

と Command ウインドウに入力し、データをソートしてから、もう一度、des で確認すると、以下のようになります。

```
. des

Contains data from D:¥Data¥Wage.dta
  obs:            47
 vars:             2                          21 Apr 2004 21:58
 size:           423 (99.9% of memory free)
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
prefecture            byte   %8.0g
wage                  float  %9.0g
```

この状態で、save しておきます。

save wage.dta,replace

上書きすることになるので、replace を忘れずに。
同様に、Production.dta も prefecture で sort し、save しておきます。これで準備完了です。
二つのファイルのうち、どちらを先に呼び出して構いませんが、Production.dta を先に呼び出すことにしましょう。

use Production.dta

データを接続するには、merge コマンドを使います。merge コマンドは、

merge [キー変数]　using [接続するファイル名]

となります。今の場合、接続のキーとなる変数は prefecture、接続するファイルは Wage.dta ですので、以下のようになります。

merge prefecture using Wage.dta

うまくいけば、データセットは以下のようになります。

prefecture	wage	production	_merge
1	3.616	18954421	3
2	2.644	4634405	3
3	3.522	4678288	3
4	3.631	8429719	3
5	3.348	3901386	3
6	3.517	4095372	3
7	3.928	7692465	3
8	5.337	11374471	3
9	4.912	7739373	3
⋮	⋮	⋮	⋮
47	2.687	3268545	3

ここで、_merge という新しい変数が生成されていますが、これについては後述します。
なお、続けて他のデータセットを merge する場合は、_merge を drop しておいてください。

３）時間に依存しないクロスセクションデータ、個体に依存しない時系列データの接続

さて、２）のケースでは、接続する２つのファイルの長さは等しくなっていました。しかし、現実のニーズとしては、１）で作成した都道府県×年次×項目のファイルに、表１－８のように年次別の全国一律のデータ、たとえば物価指数を接続するといった作業が必要になることもあります。このような場合はどうしたらいいのでしょうか？

例として、以下のような年次別の全国平均の物価指数を１）で作成したデータセットに接続する方法について考えましょう。

Price.dta

year	price
1980	100.0
1981	101.2
⋮	⋮
2000	132.2

接続方法は、基本的に２）と同じで、まず、接続する際のキーとなる変数で、接続する２つのファイルが sort されているかどうか確認します。この場合は、年次を示す Year がキー変数となります。問題がなければ、一方のデータを開いた状態で、merge を行います。

```
use Production80-00.dta
merge year using Price.dta
```

結果は、うまくいけば、以下の表のようになります。Year が同一のところには、必ず同じ Price の値が入っていることが確認できます。

prefecture	year	production	Price
1	1980	1200	100
2	1980	800	100
3	1980	1921	100
⋮	⋮	⋮	⋮
1	1981	1310	101.2
2	1981	1050	101.2
⋮	⋮	⋮	⋮
1	2000	2560	132.3
2	2000	1420	132.3

４）グループ別の時系列データの接続

１）～３）までのデータセットでは、２つのデータセットに含まれるキーとなる変数が完全な対応関係がありました。しかし、実際には、以下のようなキーとなる変数が部分的にしか対応していないケースがままあります。以下のような例を考えましょう。

even.dta

number	Even
5	10
6	12
7	14
8	16

odd.dta

number	Odd
1	1
2	3
3	5
4	7
5	9

この２つのファイルのキーとなる変数は number です。ですが、２つのファイルに重複する変数は、”5”だけです。このケースで、number をキーに merge すると以下のようになります。

use even.dta

merge number using odd.dta

number	Even	Odd	merge
5	10	9	3
6	12		1
7	14		1
8	16		1
1		1	2
2		3	2
3		5	2
4		7	2

この場合、even.dta と odd.dta の number 変数で共通なのは「５」のみなので、キーとして指定した変数が共通する場合のみ同じ行に odd.dta が接続され、異なる場合には異なる行に odd.dta を接続されます。

なお、merge コマンドを実行すると、_merge という変数が副産物として生成されます。_merge は、二つのデータの結合状態を表します。

_merge=3 ： キーに指定した変数が結合前の二つのファイル双方に存在していた場合。

_merge=1 ： キーにした変数が、merge 実行前に開いていたファイルのみに存在していた場合。

_merge=2 ： キーにした変数が、merge 実行時に呼び出しファイルのみに存在していた場合。

even.dta と odd.dta の接続を例にすると、

even.dta と odd.dta の両方のファイルに含まれていたデータ：_merge=3
even.dta のみに含まれていたデータ：_merge=1
odd.dta のみに含まれていたデータ：_merge=2

となります。

1.4. データ・フォーマットの変換

以上でみてきたように、merge コマンドと append コマンドを利用することで表１－２や表１－３のようなパネルデータを作成することができます。次に、Long 形式である表１－２のようなデータを Wide 形式に変換する方法、もしくは、Wide 形式のデータを Long けいしきに変換する方法を考えましょう。

ここでは、電気機械産業の上場企業の財務データを例に考えて見ましょう。

表１－１０．LONG 形式のデータ（例）

fid	year	labor	slsprofit	head-q
6501	1994	80493	0.0174	13
6501	1995	78368	0.0215	13
6501	1996	75590	0.0312	13
6501	1997	72193	0.0196	13
6501	1998	70375	0.0042	13
6501	1999	66046	-0.0304	13
6501	2000	58739	0.0084	13
6501	2001	54017	0.0140	13
6501	2002	48590	-0.0232	13
6502	1994	74558	0.0165	14
6502	1995	73463	0.0215	14

表１－１１．WIDE 形式のデータ（例）

fid	labor1994	labor1995	labor1996	labor1997	labor...
6501	80493	8368	5590	2193	…
6502	74558	3463	1170	8441	…
6503	49842	8421	7752	7372	…

なお、各変数の定義は、以下のとおり。

fid：株式コード

year：年次

labor：従業者数

slsprofit：売上利益率

head-q：本社所在地

（1）パネルデータ形式を変換する

パネルデータの『 LONG 形式　⇔ WIDE 形式 』 変換を **reshape** コマンドにより行うことができます。

【long 形式 ⇒ wide 形式】の変換

reshape wide labor slsprofit, i(fid) j(year)

変換対象の変数

wide に続いて変換したい変数名を記入します。個体ごとに時間を通じて一定の変数（たとえば、表の変数のうち、”head-q”のように個体ごとにみると、一定になっている変数）は記入する必要はありません。ただし、個体により、時間により異なる値をもつ変数がデータセットに含まれている（表の labor や slsprofit のような変数）にも関わらず、変換対象の変数として記述から漏れている時、データ形式変換は行われずエラー表示が返されます。コマンドライン中の wide 以下には time variant（時間について可変）な変数は全て記入するようにしましょう。

変換の軸となる個体を表わす変数 fid と、時間を表わす変数 year の全データが、一対一の関係であれば問題なく変換されます。誤植などにより重複してデータが存在する場合

```
year not unique within fid;
there are multiple observations at the same year within fid.
Type ″reshape error″ for a listing of the problem observations.
```

（例えば fid 番号 5948 の 1999 年のデータが１つ以上存在する場合など）は変換されず、上記のようなエラーが表示されます。このような場合、**reshape error** と入力することで、どのデータが重複しているか、詳細な情報を確認することができます。

なお、unbalanced panel である場合、データ変換に特に問題は生じません。欠損しているデータについては「.」の欠損を表わす記述が自動的に置き換わります。

【wide 形式⇒long 形式】の変換

reshape long labor slsprofit, i(fid) j(year)

変換するデータセットには､変換対称として指定する変数名(ここでは labor, slsprofit)と、その変数名に数値が続く変数（ここでは labor1994, labor11995,…）が存在する必要があります。指定変数名に続く数値が、 **j( )** で指定した時間軸変数の値として変換されます。全ての変数が正しく存在する時（共通した変数名があり、その各変数名に共通した数値系列が続いている場合)、細かい指定を省略し **reshape long** と記入するだけで、データ形式が変換されます。

後述のとおり、パネル計量分析を行うためには Long 形式のデータセットを用意する必要があります。しかし、個体ごとの変数の変化率などの変数を作成する際には、Wide 形式の

ほうが操作しやすく、Long 形式と Wide 形式をうまく使い分けることでデータセット作成の手間を省力化することができます。なお、Long 形式のデータセットで、個体ごとの変数の変化率を計算するには、後述するデータオペレーション関数を用います。

balance パネルと unbalance パネル

balanced panel とは、使用するデータセットの各個体の変数が全期間揃っている（欠損値を含まない）パネルデータセットであることを言います。反対に、ある個体のある時点のデータが欠損している場合は unbalanced panel と言います。

（２）パネルデータとしての認証

パネルデータによる分析を行う際、STATA にデータセットがパネルデータであるという情報を伝える必要があります。

tsset　var1　var2

「var1」には主体を表わす変数名を、「var2」には時間軸を表わす変数名を記述します。

```
. tsset fid year
       panel variable:  fid, 1909 to 359059
        time variable:  year, 1994 to 2002
```

パネルデータであることを伝えたら、パネルデータの形状を **xtdes** コマンドにより確認できます。

```
. iis fid
. tis year
. xtdes                                                          ①
                                                                 ↓
     fid:  1909, 1993, ..., 359059                       n =       334
    year:  1994, 1995, ..., 2002                        T =         9
           Delta(year) = 1; (2002-1994)+1 = 9
           (fid*year does not uniquely identify observations)

Distribution of T_i:   min      5%      25%       50%       75%      95%      max
                         1       7        9         9         9        9       27
                                                    ↑
     Freq.  Percent    Cum. |  Pattern              ②
 ---------------------------+-----------      ←③
       256     76.35   76.35 |  111111111    ←94-02 まで連続している標本が 256 社
        27      8.08   84.43 |  .11111111
        16      4.79   89.22 |  11111111.
        12      3.59   92.81 |  ..1111111
        10      2.99   95.81 |  ...111111
         3      0.90   96.71 |  1111111..
         1      0.30   97.60 |  ........1
```

① ここには、個体識別変数(fid)が 1909～359059 までの値の 334 社のデータが、1994～2002 年の 9 時点分あることを示しています。また、変数 fid と year が一対一の関係でないことも(**fid*year does not uniquely identify observations**)**で示しています**。そのため、データの重複を修正しなくては、Wide 形式に変換することも回帰分析することもできないことが分かります。

② ここには、データの欠損に関する情報が得られます。95%のデータは 9 時点のデータがあることを示していますが、5%のデータは 7 時点のデータであることが示されています。よって、このデータセットは unbalanced panel であることが分かります。

③ ②の情報を、より詳しく示しています。256 サンプルはデータは全期間連続しており、27 サンプルは 1 期目のデータが欠損していることを示しています。Pattern の列にしめされる「1」はデータ存在していることを示し、「.」はデータが存在していないことを示しています。

（3）データ・オペレータ・ファンクション

tsset の設定により STATA が時系列の概念を認識できるようになると、遅延演算子などのオペレーション・ファンクションを利用することが可能となります。

l. ファンクション 時系列方向のデータを含むデータを扱う際、**l.** を変数の前に付けることでラグ付変数として認識されます。

```
labor      ≡labor(t)
l.labor    ≡labor(t-1)
l2.labor   ≡labor(t-2)
  :          :
```

f. ファンクション **f.** を変数の前に付けることで一期前の値を参照します。

```
f.labor    ≡labor(t+1)
f2.labor   ≡labor(t+2)
  :          :
```

d. ファンクション **d.** を変数の前に付けると、前期値との差分変数として認識します。

```
d.labor    ≡labor(t)-labor(t-1)
```

これらのオペレーション・ファンクションは、個体ごとの時系列を参照して算出されます。その点が変数システムファンクション[_n-1]などと異なり、パネルデータを扱う際の極めて利便性の高いファンクションと言えます。以下に、l.ファンクションとシステムファンクション[_ｎ-1]との違いを例示しましょう。

```
. tsset fid year
. gen test1=l.labor
. gen test2=labor[_n-1]
. list
     +-------------------------------------+
     |   fid   year   labor   test1   test2 |
     |-------------------------------------|
  1. |  1909   1995     535       .       . |
  2. |  1909   1996     529     535     535 |
     |     (省略)                          |
  6. |  1909   2000     470     509     509 |
  7. |  1993   1994     922       .     470 |    ←test2 では個体変数別に
  8. |  1993   1995     929     922     922 |      データが作成がされない
     |     (省略)                          |      様子がわかります。
 15. |  1993   2002     773     799     799 |
```

データ・オペレータ・ファンクションにより Long 形式でもデータの扱いが容易になりますが、１－２.で紹介する回帰分析に、オペレータ・ファンクション付の変数を直接組み込むことはできません。回帰分析でラグ付変数などを使用したい場合は、まず一度 **gen** コマンドで新たな変数を作成し、その新変数を使って回帰分析を試みましょう。

2. パネルデータによる回帰分析

パネル計量分析を行う際、データの特性（**i**：個体を表わす変数、**t**:時間を表わす変数）に関する情報が必要です。１－１.（２）で指定した **tsset** から変更がなければ、回帰分析を行うコマンドラインごとに **i** や **t** を指定する必要はありません。ただし、データを加工したことで、新たな個体認識変数や時間変数が作成された場合などは、データ特性が変更された情報を STATA に伝えなければなりません。

iis varname

tis varname

iis コマンドは新たな個体認識の変数の指定、**tis** は新たな時間変数の指定を行います。この時、**tsset** で伝えていた情報は残されないため、元の特性を用いて分析し直したい時には、特性変数の再指定をする必要があります。なお、tsset を用いる場合、個体を認識するための変数は必ず単独の変数でを用いますが、iis の場合、複数の変数によって校正されていても構いません。たとえば、個体の識別番号が、都道府県番号（2 桁）、市区町村番号（3 桁）、事業所番号（4 桁）の３つの変数によって構成されているとき、tsset を用いる場合は、３

つを組み合わせた 9 桁の番号を作成する必要があります。しかし、iis の場合は、３つの変数を並べることで、個体の識別情報を Stata に認識させることができます。

（１）線形回帰分析（変量効果モデル、固定効果モデルなど）

```
xtreg depvar indepvar ,xx
```

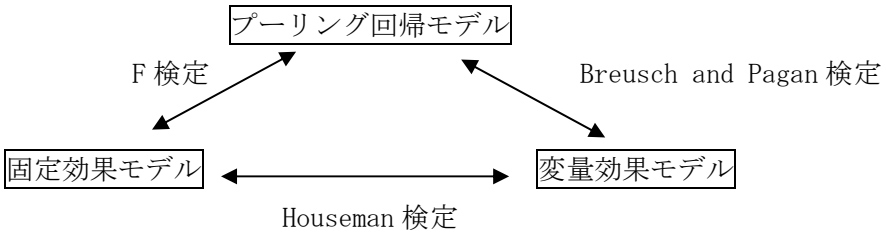
「**depvar**」部分に被説明変数を、「**indepvar**」部分に被説明変数（複数記入可）を記入します。「**,xx**」の **xx** 部分には、以下の得たい推定量を記入します。無記入の場合は変量効果モデル **re** が推定されます。

```
be      between-effects estimator
fe      fixed-effects estimator
re      GLS random-effects estimator
```

なお、areg コマンドを用いることで固定効果モデルと同じ結果を得ることができます。ただし、その際に出力される決定係数は、xtreg の場合は固定効果を含まない値が出力されるのに対して、areg の場合は固定効果を含む値が出力されます。そのため、後者のほうが１に近い値をとります。

（２）モデル選択の検定

パネルデータを用いる分析としては、プーリング回帰モデル、固定効果モデルと変量効果モデルの３つモデルがありますが、どの方法を用いるのが望ましいのかは統計的検定を用いて分析結果を比較する必要があります。具体的には以下の図のように、３つの統計的検定を用いて、分析結果を相互に比較します。



Houseman 検定

Housman 検定は、「固定効果モデルよりも変量効果モデルが正しい」という仮説を検定します。仮説が棄却されれば、固定効果モデルが正しいと考えます。

```
xtreg depvar indepvar, fe
est store fixed
xtreg depvar indepvar, re
hausman fixed .
```

なお、下線部分には適当な変数名を指定します。

Breusch and Pagan 検定

Breusch and Pagan 検定は、「変量効果モデルよりもプーリング回帰モデルが正しい」という仮説を検定します。仮説が棄却されれば、変量効果モデルが正しいと考えます。

```
xtreg depvar indepvar, re
xttest0
```

F 検定

F 検定は、「固定効果モデルよりもプーリング回帰モデルが正しい」という仮説を検定します。仮説が棄却されれば、固定効果モデルが正しいと考えます。なお、Stata では、F 検定の結果は、固定効果モデルの推定結果の一部として表示されます。

例題　パネルデータの作成とパネル回帰モデルの推定

1993-1994 年都道府県別データによる生産関数の推計
（出所：慶応大学・土居助教授の WEB ページ：http://www.econ.keio.ac.jp/staff/tdoi/）

y-prefec.csv: WIDE 形式
　　prefecture: 都道府県番号
　　Y1993: 県内総生産(1993)
　　Y1994: 県内総生産(1994)

l-prefec.csv: WIDE 形式
　　prefecture: 都道府県番号
　　L1993: 就業者数(1993)
　　L1994: 就業者数(1994)

k-prefec.csv: LONG 形式
　　prefecture: 都道府県番号
　　year: 年次
　　K: 資本ストック（沖縄県データなし)、

1) Stata で y-prefec.csv と l-prefec.csv を読み込み､merge コマンドで両者を統合し、prod-prefc.dta という stata-dta ファイルを作成せよ。
2) reshape コマンドを使って、long 形式に変更する。
3) k-prefec.csv を読み込み、merge コマンドで prod-prefc.dta ファイルと統合する。なお、k-prefec.csv には、沖縄県のデータが含まれていない。merge コマンド実行時に作成される変数"_merge"を tabulate コマンドでチェックすること。
4) 労働生産性(Y/L)と資本装備率（K/L）をそれぞれ対数変換する。そして、前者を被説明変数、後者を説明変数とするパネル回帰モデルを推定せよ。

プログラム例

```
/*
1993-1994 年都道府県別データによる生産関数の推計
y-prefec.csv:Y=県内総生産、WIDE 形式
l-prefec.csv:L=就業者数、WIDE 形式
k-prefec.csv1993:K=資本ストック（1993 年、沖縄県データなし）
k-prefec.csv1994:K=資本ストック（1994 年、沖縄県データなし）
   ↓
統合して、prod-prefec.dta を作成
*/
cd d:¥Data
clear
*  メモリーの割り当て：
set memory 10m
*  県内総生産の読み込み
insheet using y-prefec.csv,clear

sort prefecture
save prod-prefec.dta,replace


*  都道府県別就業者数の読み込み
insheet using l-prefec.csv,clear

**** merge コマンド：P.21～22 ****
* prod-prefec.dta と接続－あらかじめ prefecture で sort する－
sort prefecture
merge prefecture using prod-prefec.dta
tab _merge
drop _merge

reshape long l y,i(prefecture) j(year)

sort prefecture year
save prod-prefec.dta,replace


*  都道府県別資本ストックの読み込みと縦方向の接続
insheet using k-prefec1993.csv,clear
gen year=1993
save k-prefec.dta,replace
insheet using k-prefec1993.csv,clear
gen year=1994
append using k-prefec.dta
```

```
* prod-prefec.dta と接続
*  －あらかじめ prefecture と year で sort する－
sort prefecture year
merge prefecture year using prod-prefec.dta

tab _merge
drop _merge
sort prefecture
save prod-prefec.dta,replace



replace l=log(l)
replace k=log(k)
replace y=log(y)

gen yl=y-l
gen kl=k-l

*  パネルデータであることを宣言する
tsset prefecture year

****  パネルデータ回帰モデル  ****
*  固定効果モデルによる推定
xtreg yl kl,fe
areg yl kl,absorb(prefecture)
* Hauseman 検定の準備
* "fixed"のところは任意の変数名
est store fixed

*  変量効果モデルによる推定
xtreg yl kl
* Hauseman 検定
hausman fixed
* Breusch and Pagan 検定
xttest0
```

分析結果の見方

```
. * 固定効果モデルによる推定
. xtreg yl kl,fe

Fixed-effects (within) regression               Number of obs      =        92
Group variable (i): prefecture                  Number of groups   =        46

R-sq:  within  = 0.4097                         Obs per group: min =         2
       between = 0.4797                                        avg =       2.0
       overall = 0.4771                                        max =         2

                                                F(1,45)            =     31.24
corr(u_i, Xb)  = -0.9251                        Prob > F           =    0.0000

------------------------------------------------------------------------------
          yl |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          kl |   1.855042   .3319163     5.59   0.000     1.186528    2.523555
       _cons |  -2.611801   .8120736    -3.22   0.002    -4.247401   -.9762005
------------------------------------------------------------------------------
     sigma_u |  .24189795
     sigma_e |  .01326616
         rho |  .99700138   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(45, 45) =    95.93                 Prob > F = 0.0000

* Hausman 検定の準備
. est store growth_fixed
. areg yl kl,absorb(prefecture)

Linear regression, absorbing indicators         Number of obs =        92
                                                F(  1,    45) =     31.24
                                                Prob > F      =    0.0000
                                                R-squared     =    0.9946
                                                Adj R-squared =    0.9891
                                                Root MSE      =    .01327

------------------------------------------------------------------------------
          yl |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          kl |   1.855042   .3319163     5.59   0.000     1.186528    2.523555
       _cons |  -2.611801   .8120736    -3.22   0.002    -4.247401   -.9762005
-------------+----------------------------------------------------------------
  prefecture |         F(45, 45) =     95.931   0.000          (46 categories)
```

プーリング回帰 vs 固定効果の F 検定

Xtreg,fe と areg の係数は等しい

```
. * 変量効果モデルによる推定
. xtreg yl kl

Random-effects GLS regression                   Number of obs      =        92
Group variable (i): prefecture                  Number of groups   =        46

R-sq:  within  = 0.4097                         Obs per group: min =         2
       between = 0.4797                                        avg =       2.0
       overall = 0.4771                                        max =         2

Random effects u_i ~ Gaussian                   Wald chi2(1)       =     48.95
corr(u_i, X)       = 0 (assumed)                Prob > chi2        =    0.0000

------------------------------------------------------------------------------
          yl |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          kl |   .6007053   .0858632     7.00   0.000     .4324165    .7689942
       _cons |   .4570811    .210589     2.17   0.030     .0443343    .8698279
-------------+----------------------------------------------------------------
     sigma_u |    .092267
     sigma_e |  .01326616
         rho |  .97974597   (fraction of variance due to u_i)
------------------------------------------------------------------------------

. hausman fixed

                 ---- Coefficients ----
              |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
              |     fixed          .           Difference          S.E.
--------------+----------------------------------------------------------------
          kl |    1.855042     .6007053        1.254336        .3206181
------------------------------------------------------------------------------
                           b = consistent under Ho and Ha; obtained from areg
            B = inconsistent under Ha, efficient under Ho; obtained from xtreg

    Test:  Ho:  difference in coefficients not systematic

                  chi2(1) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                          =       15.31
                Prob>chi2 =      0.0001
```

帰無仮説「固定効果モデルよりも変量効果モデルが望ましい」が正しい確率は、0.01%しかないので、固定効果モデルが正しいと考えます。

```
* Breusch and Pagan 検定
. xttest0

Breusch and Pagan Lagrangian multiplier test for random effects:

        yl[prefecture,t] = Xb + u[prefecture] + e[prefecture,t]

        Estimated results:
                         |       Var     sd = sqrt(Var)
                ---------+-----------------------------
                      yl |   .0161337       .1270186
                       e |    .000176       .0132662
                       u |   .0085132        .092267

        Test:   Var(u) = 0
                                  chi2(1) =    43.46
                               Prob > chi2 =     0.0000
```

帰無仮説「変量効果モデルよりもプーリング回帰モデルが望ましい」が正しい確率は、0%なので、変量効果モデルが正しいと考えます。ただし、固定効果モデルの F 検定、前述の Hausman 検定の結果を踏まえると、プーリング回帰モデル＜変量効果モデル＜固定効果モデルという序列から、固定効果モデルの結果が最も望ましい結果といえます。