

# Rによるパネルデータモデルの推定<sup>†</sup>

Rを用いて、静学的パネルデータに対して Pooled OLS, LSDV (Least Squares Dummy Variable) 推定、F 検定(個別効果なしの F 検定)、GLS(Generalized Least Square :一般化最小二乗) 法による推定、およびハウスマン検定を行うやり方を、動学的パネルデータに対して 1 階階差 GMM とシステム GMM を行うやり方を、説明する。

## 1. パッケージ plm のインスツール

パネルデータモデルを分析するために R のパッケージ plm をインスツールする。パッケージとは通常の R には含まれていない、追加的な R のコマンドの集まりのようなものである。R には追加的に 600 以上のパッケージが用意されており、それぞれ分析の目的に応じて標準の R にパッケージを追加していくことになる。

インターネットに接続してあるパソコンで R を起動させ、

「パッケージ」→「パッケージのインスツール」→ (適当なミラーサイトを選ぶ。どれを選んでよい。

例えば Japan (Tokyo) ) →「OK」→「plm」

とクリックしていくと自動的にインスツールしてくれる。

インスツールが終わったら、次に実際に使用できるようにするために

```
> library(plm)
```

と入力する。

## 2. 静学的パネルデータ分析

データとしてパッケージ plm に含まれる Grunfeld データと呼ばれるアメリカの有名企業に関するパネルデータを用いる。これは 10 の企業 (N=10) に関する 20 期分 (T=20) のデータで、次のような変数を含んでいる、inv: 総投資 (単位:100 万ドル)、value: 企業価値 (単位:100 万ドル)、capital: 有形固定資産 (単位:100 万ドル)。今回はパッケージからデータを読み込むので Working ディレクトリの変更は必要ない。次のように入力する。

```
> data("Grunfeld", package="plm")
```

読み込んだデータの最初の 25 行を見るために

```
> head(Grunfeld, 25)
```

---

<sup>†</sup>この資料は私のゼミおよび講義で R の使用法を説明するために作成した資料です。ホームページ上で公開しており、自由に参照して頂いて構いません。ただし、内容について、一応検証してありますが、間違いがあるかもしれません。間違いがあった場合でもそれによって生じるいかなる損害、不利益について責任は負いかねますのでご了承ください。

と入力する。すると最初の 25 行が表示される。1 行目は変数の名前であり、2 列目は企業番号 (1, ..., 10), 3 列目は西暦 (1935, ..., 1954) を表している (他のパネルデータを分析する際もデータをこのように並べておかなくてはならない。つまりまず  $i=1$  を固定し、 $i=1$  の  $t$  に関するデータを並べ、並べ終わったら次は  $i=2$  を固定し、 $i=2$  の  $t$  に関するデータを並べるというような並べ方を  $i=1, \dots, N$  までやるという事。ここで  $N$  はクロスセクションの個体数)。ここでは投資がどのような要因によって決定されるかを分析するとして、以下のモデルを推定する。

$$\text{inv}_{it} = \alpha_i + \beta_1 \text{value}_{it} + \beta_2 \text{capital}_{it} + \varepsilon_{it}$$

$$i=1, \dots, 10, t=1, \dots, 20.$$

このモデルに対してまず pooled OLS で係数  $\beta_1$  と  $\beta_2$  を推定する (pooled OLS とは上記のモデルにおいて  $\alpha_1 = \alpha_2 = \dots = \alpha_N$  であると想定して推定したもの。これは普通の OLS と同じ)。

```
> result1=plm(inv~value+capital,data=Grunfeld,model="pooling")
```

ここで model="pooling" の部分が Pooled OLS で推定を行うということを R に指示している部分である。結果は

```
> summary(result1)
```

でみることができる。

```
Oneway (individual) effect Pooling Model
```

```
Call:
```

```
plm(formula = inv ~ value + capital, data = Grunfeld, model = "pooling")
```

```
Balanced Panel: n=10, T=20, N=200
```

```
Residuals :
```

```
  Min. 1st Qu.  Median 3rd Qu.  Max.
-292.0  -30.0     5.3   34.8   369.0
```

```
Coefficients :
```

```
              Estimate Std. Error t-value Pr(>|t|)
(Intercept) -42.7143694   9.5116760 -4.4907 1.207e-05 ***
value         0.1155622   0.0058357 19.8026 < 2.2e-16 ***
capital       0.2306785   0.0254758  9.0548 < 2.2e-16 ***
```

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Total Sum of Squares: 9359900
```

```
Residual Sum of Squares: 1755900
```

```
R-Squared      : 0.81241
```

```
Adj. R-Squared : 0.80022
```

```
F-statistic: 426.576 on 2 and 197 DF, p-value: < 2.22e-16
```

Coefficients のところにある Estimate の列が係数の推定値である。value の横の数字が  $\beta_1$  の推定値、capital の横の数字が  $\beta_2$  の推定値である。横の Std.Error, t-value

は推定量の標準誤差、t-値である、 $\Pr(>|t|)$ はP値である。ここでのF値はこの2つの係数がともに0であるという帰無仮説を検定するF値であり、p-valueはそのP値である。

ちなみにこれはただのOLS推定なので

```
> resultOLS=lm(inv~value+capital,data=Grunfeld)
```

でも同じ推定結果を得る(summary(resultOLS)で確認してみてください)。

次に固定効果モデルの推定をLSDV推定で行う(LSDV推定は別名としてWithin推定とも呼ばれる)。以下のように入力する。

```
> result2=plm(inv~value+capital,data=Grunfeld,model="within")
```

ここでmodel="within"の部分がLSDVで推定を行うという事をRに指示している部分である。推定結果を見るには先ほどと同様に

```
> summary(result2)
```

と入力する。結果は以下ようになる。

```
Oneway (individual) effect Within Model

Call:
plm(formula = inv ~ value + capital, data = Grunfeld, model = "within")

Balanced Panel: n=10, T=20, N=200

Residuals :
  Min. 1st Qu.  Median 3rd Qu.  Max.
-184.000 -17.600   0.563  19.200  251.000

Coefficients :
      Estimate Std. Error t-value Pr(>|t|)
value  0.110124  0.011857  9.2879 < 2.2e-16 ***
capital 0.310065  0.017355 17.8666 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    2244400
Residual Sum of Squares: 523480
R-Squared      : 0.76676
  Adj. R-Squared : 0.72075
F-statistic: 309.014 on 2 and 188 DF, p-value: < 2.22e-16
```

推定値の見方は先ほどと同様である。

個別効果 $\mu_i$ の推定値を見るには関数fixef()を用いる。

```
> mu=fixef(result2)
```

```

> summary(mu)
  Estimate Std. Error t-value Pr(>|t|)
1  -70.2967   49.7080 -1.4142  0.15730
2   101.9058   24.9383  4.0863 4.383e-05 ***
3  -235.5718   24.4316 -9.6421 < 2.2e-16 ***
4   -27.8093   14.0778 -1.9754  0.04822 *
5  -114.6168   14.1654 -8.0913 6.661e-16 ***
6   -23.1613   12.6687 -1.8282  0.06752 .
7   -66.5535   12.8430 -5.1821 2.194e-07 ***
8   -57.5457   13.9931 -4.1124 3.915e-05 ***
9   -87.2223   12.8919 -6.7657 1.327e-11 ***
10  -6.5678    11.8269 -0.5553  0.57867
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '

```

一番左の番号が  $i$  の番号である。例えば  $\alpha_1$  の推定値は  $-70.2967$ ,  $\alpha_3$  の推定値は  $-235.5718$  である。またそれぞれの  $t$  値は  $\alpha_i = 0$  かどうかの  $t$  検定用である。個別効果の平均は

```

> mean(mu)
[1] -58.74394

```

で見ることができる。

場合によっては個別効果のその平均からの乖離を見たい時がある(つまり  $\hat{\alpha}_i - N^{-1} \sum_{i=1}^N \hat{\alpha}_i$ 、ここで  $\hat{\alpha}_i$  は個別効果  $\alpha_i$  の推定値)。これは

```

> mu2=fixef(result2,type="dmean")

```

とすれば、mu2 がそれである。結果は

```

> summary(mu2)
  Estimate Std. Error t-value Pr(>|t|)
1  -11.5528   49.7080 -0.2324  0.816217
2   160.6498   24.9383  6.4419 1.180e-10 ***
3  -176.8279   24.4316 -7.2377 4.565e-13 ***
4   30.9346   14.0778  2.1974  0.027991 *
5  -55.8729   14.1654 -3.9443 8.003e-05 ***
6   35.5826   12.6687  2.8087  0.004974 **
7   -7.8095   12.8430 -0.6081  0.543136
8    1.1983   13.9931  0.0856  0.931758
9  -28.4783   12.8919 -2.2090  0.027174 *
10  52.1761   11.8269  4.4116 1.026e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'

```

となる。ここでの  $t$  値は、乖離が 0 かどうか、すなわち、それぞれの個別効果が個別効果全体の平均から異なるかどうかを検定している事に注意する必要がある。また、type として type="dfirst" とすると最初の個別効果との差が出力される。

次に個別効果があるかどうかの F 検定を行うやり方を見てみよう。帰無仮説は

$H_0: \alpha_1 = \alpha_2 = \dots = \alpha_N$  (全ての個別効果が同じ値である)

である。これは `pFtest()` 関数を用いて行う。今、`result1` に pooled OLS 推定の結果、`result2` に LSDV 推定の結果が保存されているとする。この時、この  $F$  検定は

```
> pFtest(result2,result1)
```

によって行う(**result1 が 2 番目になっている事に注意**)。結果は

```
F test for individual effects
```

```
data: inv ~ value + capital
F = 49.1766, df1 = 9, df2 = 188, p-value < 2.2e-16
alternative hypothesis: significant effects
```

と表示される。 $F$  値が 49.177 であり、これは第 1 自由度 9、第 2 自由度 188 の  $F$  分布に従う(誤差項が正規分布の場合、これは正確に成り立つ。ただし誤差項が正規分布に従わない場合もこの  $F$  検定は ( $T$  を固定し  $N$  を大きくしたとき) 漸近的には正しい事が示されている。また有限標本でも正規分布を仮定した  $F$  検定とほぼ同じ結果になるのでそのまま  $F$  検定でやっても特に問題ない)。 $P$  値が非常に小さいので、帰無仮説は棄却されることがわかる。つまり「個別効果はある」ということになる。

次に GLS 推定を行う。今回は  $\mu_i$  を確率変数とするので、モデルは

$$\text{inv}_{it} = \mu_i + \mu_\alpha + \beta_1 \text{value}_{it} + \beta_2 \text{capital}_{it} + \varepsilon_{it}, \quad i=1,\dots,10, t=1,\dots,20.$$
$$E(\mu_i) = 0, \quad \text{var}(\mu_i) = \sigma_\alpha^2, \quad E(\varepsilon_{it}) = 0, \quad \text{var}(\varepsilon_{it}) = \sigma_\varepsilon^2, \quad \text{cov}(\mu_i, \varepsilon_{it}) = 0$$

となる(より詳しくはスライド参照)。再び関数 `plm()` を用いる。

```
> result3=plm(inv~value+capital,data=Grunfeld,model="random")
> summary(result3)
Oneway (individual) effect Random Effect Model
(Swamy-Arora's transformation)

Call:
plm(formula = inv ~ value + capital, data = Grunfeld, model = "random")

Balanced Panel: n=10, T=20, N=200

Effects:
              var std.dev share
idiosyncratic 2784.46  52.77 0.282
individual    7089.80  84.20 0.718
theta: 0.8612

Residuals :
  Min. 1st Qu.  Median 3rd Qu.    Max.
-178.00 -19.70   4.69  19.50  253.00

Coefficients :
```

```

                Estimate Std. Error t-value Pr(>|t|)
(Intercept) -57.834415  28.898935 -2.0013  0.04674 *
value       0.109781   0.010493 10.4627 < 2e-16 ***
capital     0.308113   0.017180 17.9339 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:  2381400
Residual Sum of Squares: 548900
R-Squared      : 0.7695
Adj. R-Squared : 0.75796
F-statistic: 328.837 on 2 and 197 DF, p-value: < 2.22e-16

```

推定結果のうち Effects: の部分における var の列で idiosyncratic の横の数字が  $\sigma_e^2$  の推定値、individual の横の数字が  $\sigma_a^2$  の推定値となる。また Coefficients : の部分における Estimate の列の (Intercept) の横の数字が  $\mu_a$  の推定値、value の横の数字が  $\beta_1$  の推定値、capital の横の数字が  $\beta_2$  の推定値である。

GLS 推定では  $\mu_i$  と説明変数の間に相関があると推定量が一致性を失う(今までは相関なしを仮定していたが)よってこの相関があるかないかを確認するのは重要である。相関が0という帰無仮説を検定するための検定としてハウスマン検定と呼ばれる検定がある。この検定は以下のように phtest() 関数を用いる事によってできる。

```

> phtest(result2,result3)

Hausman Test

data:  inv ~ value + capital
chisq = 2.3304, df = 2, p-value = 0.3119
alternative hypothesis: one model is inconsistent

```

結果を見ると P 値が 0.3119 なので帰無仮説は棄却されないという事になる。つまり GLS による推定は問題ないという事になる。

### 3. 動学的パネルデータ分析

データとしてパッケージ plm に含まれる EmplUK データを用いる。これは 1976-1984 の 9 年間のイギリスの 140 の企業の就業者数に関するデータで、次の変数を含んでいる、emp: 就業者数 (i 番目の企業の t 年度末の値, 単位不明)、wage: 実質賃金 (計算方法は複雑なので省略、単位不明)、capital: 総資本 (計算方法は複雑なので省略、単位不明)、output: 生産量 (計算方法は複雑なので省略、単位不明)。データを読み込むには

```
> data("EmplUK", package="plm")
```

と入力する。上記のデータに対して ( $i=1, \dots, 140, t=1, \dots, 9$ )、以下のモデルを推定してみよう (log は自然対数)。

$$\begin{aligned} \log(\text{emp}_{it}) &= \alpha_i + \beta_1 \log(\text{emp}_{it-1}) + \beta_2 \log(\text{emp}_{it-2}) + \gamma_1 \log(\text{wage}_{it}) + \gamma_2 \log(\text{wage}_{it-1}) \\ &+ \delta_1 \log(\text{capital}_{it-1}) + \theta_1 \log(\text{output}_{it}) + \theta_2 \log(\text{output}_{it-1}) + \lambda_t + \varepsilon_{it}, \\ i &= 1, \dots, 140, \quad t = 3, \dots, 9, \end{aligned}$$

(ラグを 2 つとっているなので最初の 2 時点のデータが消えることに注意)ここで  $\lambda_t$  は  $t$  時点の時間効果を表す。このモデルでは説明変数に被説明変数の過去の値 (ラグ項) が入っていることに注意。このモデルにおいて個別効果を消去するために 1 階の階差を取ると

$$\begin{aligned} \Delta \log(\text{emp}_{it}) &= \beta_1 \Delta \log(\text{emp}_{it-1}) + \beta_2 \Delta \log(\text{emp}_{it-2}) + \gamma_1 \Delta \log(\text{wage}_{it}) + \gamma_2 \Delta \log(\text{wage}_{it-1}) \\ &+ \delta_1 \Delta \log(\text{capital}_{it}) + \theta_1 \Delta \log(\text{output}_{it}) + \theta_2 \Delta \log(\text{output}_{it-1}) + \Delta \lambda_t + \Delta \varepsilon_{it} \\ i &= 1, \dots, 140, \quad t = 4, \dots, 9, \end{aligned}$$

となる (1 階の階差を取ったのでさらに 1 つの時点のデータが使えなくなることに注意)。ここで  $\Delta x_{it} = x_{it} - x_{it-1}$  である。このモデルにおいて説明変数の  $\Delta \log(\text{emp}_{it-1})$  は誤差項  $\Delta \varepsilon_{it}$  と相関がある (ただしもう一つのラグ項  $\Delta \log(\text{emp}_{it-2})$  は  $\Delta \varepsilon_{it}$  と無相関。これは先決変数と考えられる) ので  $\Delta \log(\text{emp}_{it-1})$  に対する操作変数を用いて GMM で推定する。これを **1 階階差 GMM 推定** とよぶ。1 階階差 GMM 推定は `pgmm()` 関数を用いて推定することができる。上記のモデルの場合は

```
> result4=pgmm(log(emp)~lag(log(emp),1:2)+lag(log(wage),0:1)+
+log(capital)+lag(log(output),0:1)|lag(log(emp),2:99),data=EmplUK,
+effect="twoways", model="twosteps")
```

と入力することによって推定できる。ここで `lag(x, k:j)` は変数  $x$  の  $k$  から  $j$  までのラグ項  $x_{t-k}, \dots, x_{t-j}$  を含めるということ。ちなみに  $x_{t-k}$  だけを含めるのであれば `lag(x, k)` となる) 縦棒 "|" の右側の `lag(log(emp), 2:99)` は GMM 推定に追加的に用いる操作変数 (ここでは 2:99 は emp のラグを 2 時点前から用いるということでこのようになっている。99 の方は特に意味はなく十分な長さであれば何でもよいようである。8 (実際に使える最大の値) にしても推定結果は変わらなかった)、`effect="twoways"` は (もともとの) モデルに個別効果と時間効果両方を入れていることを意味し (個別効果だけであれば `effect="individual"` とする)、`model="twosteps"` は 2 段階 GMM で推定していることを意味する。結果は

```
> summary(result4)
```

によって見ることができる。また 2 段階 GMM 推定の通常の (計算方法で計算した) 標準誤差は実際の標準誤差を過小評価する傾向があるため、ここではロバストな標準誤差を用いてる。通常の標準誤差は

```
> summary(result4, robust=FALSE)
```

で見ることができる (デフォルトでは `robust=TRUE` になっている)。

推定した時間効果を見るには

```
> summary(result4, time.dummies=TRUE)
```

とする。ただし 1 階階差モデルに時間ダミーを入れて推定すると  $\Delta \lambda_t$  が推定されるが、ここでの時

間効果はモデルの最初の時点(今回の場合は  $t=4$ )の  $\Delta\lambda_t$  を  $\lambda_t$  と等しいと仮定して(言い換えると最初の時点の1つ前の時点の  $\lambda_t$  は0だと仮定して)以後、 $\lambda_{t+1}=\lambda_t+\Delta\lambda_{t+1}$  という関係式より  $\lambda_s, s>t$  を推定していることに注意。推定結果には  $\lambda_t$  の推定値が出力されている。

次に以下のモデルを**システム GMM 推定**してみよう。

$$\begin{aligned} \log(\text{emp}_{it}) = & \alpha_i + \beta_1 \log(\text{emp}_{it-1}) + \gamma_1 \log(\text{wage}_{it}) + \gamma_2 \log(\text{wage}_{it-1}) \\ & + \delta_1 \log(\text{capital}_t) + \delta_2 \log(\text{capital}_{t-1}) + \lambda_t + \varepsilon_{it} \\ & i=1, \dots, N, \quad t=3, \dots, T, \end{aligned}$$

以下のように入力する。

```
> result5=pgmm(log(emp)~lag(log(emp),1)+lag(log(wage),0:1)+
+lag(log(capital),0:1)|lag(log(emp),2:99)+lag(log(wage),2:99)+
+lag(log(capital),2:99),data=EmplUK,effect="twoways",
+model="onestep",transformation="ld")
```

ここで `transformation="ld"` がレベル式("l" は level の l)と階差式("d" が difference の d を表す)を合わせたもの、すなわちシステム GMM で推定することを意味する。また、先ほどは(追加的な)操作変数として `emp` のラグのみを用いたが、ここでは `wage` と `capital` のラグも用いている。

### 練習問題

資料ページの `chigin.csv` ファイル(これは 2003 年から 2007 年の地銀、第二地銀、信金 410 社のデータ<sup>1</sup>である)にある `badloan`(不良債権額)、`capgap`(自己資本比率)、`yokin`(預金額)、`public`(公的資本投資ダミー)、`kyujin`(本店所在地有効求人倍率)、`listed`(上場ダミー)、`keihi`(営業経費)というデータを用いて

$$\begin{aligned} \log(\text{badloan}_{it}) = & \alpha_i + \beta_1 \log(\text{yokin}_{i,t-1}) + \beta_2 \text{capgap}_{i,t-1} + \beta_3 \text{public}_{i,t-1} \\ & + \beta_4 \text{kyujin}_{it} + \beta_5 \text{listed}_{it} + \beta_6 \text{keihi}_{it} + \varepsilon_{it} \\ & i=1, \dots, 410, \quad t=1, \dots, 5. \end{aligned}$$

というパネルデータモデルに対して、Pooled OLS 推定、固定効果モデルの LSDV 推定、個別効果の有無についての F 検定、変量効果モデルの GLS 推定、ハウスマン検定、を行いなさい。ここで `yokini,t-1` は  $i$  番目の `yokin` データの  $t-1$  時点(年)の値を表している(`capgapi,t-1` および `publici,t-1` についても同様)。回帰式の説明変数に 1 時点前のデータが含まれていることに注意が必要である。動学的パネルデータ分析のところから出てきたようにラグ項を指定する。

また、同様に説明変数に不良債権の対数値の 1 期前の値をいれた

$$\begin{aligned} \log(\text{badloan}_{it}) = & \alpha_i + \rho \log(\text{badloan}_{it-1}) + \beta_1 \log(\text{yokin}_{i,t-1}) + \beta_2 \text{capgap}_{i,t-1} \\ & + \beta_3 \text{public}_{i,t-1} + \beta_4 \text{kyujin}_{it} + \beta_5 \text{listed}_{it} + \beta_6 \text{keihi}_{it} + \varepsilon_{it} \end{aligned}$$

というモデルを 1 階階差 GMM、システム GMM で推定しなさい。追加的な操作変数としては

<sup>1</sup> 松浦克己、コリン・マッケンジー(2012)『EViews による計量経済分析』、東洋経済新報社の提供データより

$\log(\text{badloan}_{it})$  の (使用可能な) 過去の値だけを (できるだけ多く) 用いなさい。