R 使い方入門：パネルデータ分析と差の差法（DID）の実践

データの読み込みと概要

分析テーマは、**健康ショック（shock）が生活満足度（life）に与える影響**を、パネルデータを用いて検証することです。ここでは**差の差法（Difference-in-Differences: DID）**と呼ばれる手法を用いて、処置（健康ショック）の因果効果を識別します。

データの説明

- **id**: 個人識別番号
- **t**: 時点（1 = 処置前、2 = 処置後）
- **life**: 生活満足度（0〜4 の 5 段階）
- **shock**: 健康ショックを受けたグループかどうか（1 = 処置群、0 = 対照群）
- **y2**: 時点ダミー（1 = 2 期目、0 = 1 期目）
- **shock_y2**: 処置群 ×2 期目の交差項（DID の識別変数）
- **income**: 所得（制御変数）

```
# データ読み込みと確認
life9 <- read.csv("9_2_life_xt.csv")
head(life9)
#>   id t income life shock y2 shock_y2
#> 1  1 1    300    3     0  0        0
#> 2  1 2    300    1     0  1        0
#> 3  2 1    300    3     1  0        0
#> 4  2 2    300    4     1  1        1
#> 5  3 1     50    3     1  0        0
#> 6  3 2    300    2     1  1        1
```

データ構造を確認します。

```
str(life9)
#> 'data.frame':    6040 obs. of  7 variables:
#>  $ id      : int  1 1 2 2 3 3 4 4 5 5 ...
#>  $ t       : int  1 2 1 2 1 2 1 2 1 2 ...
```

```
#>  $ income  : num  300 300 300 300 50 300 400 400 200 200 ...
#>  $ life    : int  3 1 3 4 3 2 3 3 3 3 ...
#>  $ shock   : int  0 0 1 1 1 1 1 1 0 0 ...
#>  $ y2      : int  0 1 0 1 0 1 0 1 0 1 ...
#>  $ shock_y2: int  0 0 0 1 0 1 0 1 0 0 ...
```

サンプルサイズを確認します。

```
# 個人数と観測数
n_individuals <- length(unique(life9$id))
n_obs <- nrow(life9)
cat("個人数:", n_individuals, "\n")
#> 個人数: 3020
cat("総観測数:", n_obs, "\n")
#> 総観測数: 6040
cat("1 人あたり観測数:", n_obs / n_individuals, "\n")
#> 1 人あたり観測数: 2
```

差の差法（DID）の考え方

DID とは何か？

差の差法（Difference-in-Differences）は、**処置群と対照群の変化の差**を比較することで、処置の因果効果を識別する手法です。

基本的なアイデア：

1. 処置群の変化（2 期目 − 1 期目）を計算
2. 対照群の変化（2 期目 − 1 期目）を計算
3. 両者の差を取る → これが DID 推定量

なぜ「差の差」なのか？

単純に処置群と対照群を比較するだけでは、両グループのもともとの違い（セレクションバイアス）を除去できません。また、単純に前後比較するだけでは、時間とともに変化する共通のトレンド（時間効果）を除去できません。

DID は「差を取る」ことを 2 回行うことで、個人固有の効果と時間効果の両方を除去し、処置の純粋な効果を識別します。

DID 推定のための回帰モデル

DID は以下の回帰式で表現できます：

$$life_{it} = \alpha + \beta \cdot shock_i + \gamma \cdot y2_t + \delta \cdot (shock_i \times y2_t) + \varepsilon_{it}$$

- β：処置群と対照群のベースラインの差
- γ：時間効果（全員に共通する 2 期目の変化）
- δ：**DID 推定量**（処置の因果効果）

δ の解釈：処置群が 2 期目に受けた追加的な効果、つまり健康ショックの因果効果です。

plm パッケージによるパネルデータ分析

plm パッケージのインストールと読み込み

`plm` パッケージは、パネルデータ分析のための標準的な R パッケージです。固定効果モデル、変量効果モデル、1 階差分法など、さまざまなパネルデータ推定法を提供します。

主な機能：

- **plm()**：パネルデータ回帰の推定
- **model**：推定方法の指定（“pooling”, “within”, “random”, “fd” など）
- **effect**：固定効果の種類（“individual”, “time”, “twoways”）
- **pFtest(), phtest()**：モデル選択のための検定

```
library(plm)
```

モデル 1: 1 階差分法（First Difference）

1 階差分法は、各個人の 2 時点間の差を取ることで、時間不変の個人固有効果を除去する方法です。

$$\Delta life_i = life_{i2} - life_{i1} = \gamma + \delta \cdot shock_i + \Delta income_i + \Delta \varepsilon_i$$

`plm()` で `model = "fd"` を指定します。

```
# 1 階差分法（fd）による回帰
preg_fd <- plm(life ~ shock_y2 + income,
    data = life9,
    effect = "individual",
    model = "fd",
    index = c("id", "t")
)
summary(preg_fd)
#> Oneway (individual) effect First-Difference Model
#>
#> Call:
```

```
#> plm(formula = life ~ shock_y2 + income, data = life9, effect = "individual",
#>     model = "fd", index = c("id", "t"))
#>
#> Balanced Panel: n = 3020, T = 2, N = 6040
#> Observations used in estimation: 3020
#>
#> Residuals:
#>      Min.   1st Qu.    Median   3rd Qu.      Max.
#> -4.260022 -1.075247 -0.097576  0.806964  3.969410
#>
#> Coefficients:
#>                Estimate  Std. Error t-value  Pr(>|t|)
#> (Intercept)  0.21536463  0.03135500  6.8686 7.842e-12 ***
#> shock_y2    -0.14011741  0.04844452 -2.8923  0.003851 **
#> income       0.00022329  0.00016141  1.3834  0.166659
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Total Sum of Squares:    5186.5
#> Residual Sum of Squares: 5168.7
#> R-Squared:      0.0034335
#> Adj. R-Squared: 0.0027729
#> F-statistic: 5.19735 on 2 and 3017 DF, p-value: 0.0055809
```

結果の解釈：

- **shock_y2** の係数（約**-0.140**）が DID 推定量です。これは健康ショックを受けた人が 2 期目に経験した生活満足度への追加的な影響を表します。係数が負であることは、健康ショックが生活満足度を約 0.14 ポイント低下させることを意味します（1% 水準で有意）。
- 1 階差分法では、時間不変の個人特性（性格、遺伝的要因など）が自動的に除去されます。

モデル 2: 固定効果モデル（Within 推定量）

固定効果モデルは、各個人の平均からの偏差を用いて推定を行います。1 階差分法と同様に、時間不変の個人固有効果を除去できます。

```
# 固定効果モデル（within）を用いた回帰
preg_fe <- plm(life ~ shock + y2 + shock_y2 + income,
    data = life9,
    effect = "individual",
    model = "within",
```

```
    index = c("id", "t")
)
summary(preg_fe)
#> Oneway (individual) effect Within Model
#>
#> Call:
#> plm(formula = life ~ shock + y2 + shock_y2 + income, data = life9,
#>     effect = "individual", model = "within", index = c("id",
#>         "t"))
#>
#> Balanced Panel: n = 3020, T = 2, N = 6040
#>
#> Residuals:
#>     Min.  1st Qu.   Median  3rd Qu.     Max.
#> -2.13001 -0.45121  0.00000  0.45121  2.13001
#>
#> Coefficients:
#>             Estimate  Std. Error t-value  Pr(>|t|)
#> y2        0.21536463  0.03135500  6.8686 7.842e-12 ***
#> shock_y2 -0.14011741  0.04844452 -2.8923  0.003851 **
#> income    0.00022329  0.00016141  1.3834  0.166659
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Total Sum of Squares:    2634
#> Residual Sum of Squares: 2584.4
#> R-Squared:      0.018844
#> Adj. R-Squared: -0.96394
#> F-statistic: 19.3148 on 3 and 3017 DF, p-value: 2.1065e-12
```

結果の解釈：

- 固定効果モデルでは、`shock` は時間不変なので推定されません（個人固定効果に吸収されます）。
- `y2` の係数（約 **0.215**）は時間効果を表し、`shock_y2`（約**-0.140**）が DID 推定量です。
- 2 時点パネルの場合、1 階差分法と固定効果モデルは**同じ推定量**を与えます。実際に両モデルで shock_y2 と income の係数が一致していることを確認できます。

モデル 3: プーリング OLS

比較のため、個人固有効果を無視したプーリング OLS も推定します。

```r
# プーリング OLS を用いた回帰
preg_p <- plm(life ~ shock + y2 + shock_y2 + income,
    data = life9,
    effect = "individual",
    model = "pooling",
    index = c("id", "t")
)
summary(preg_p)
#> Pooling Model
#>
#> Call:
#> plm(formula = life ~ shock + y2 + shock_y2 + income, data = life9,
#>     effect = "individual", model = "pooling", index = c("id",
#>         "t"))
#>
#> Balanced Panel: n = 3020, T = 2, N = 6040
#>
#> Residuals:
#>     Min.  1st Qu.   Median  3rd Qu.     Max.
#> -2.97119 -0.59113  0.30241  0.48499  1.54578
#>
#> Coefficients:
#>               Estimate  Std. Error t-value  Pr(>|t|)
#> (Intercept)  2.4542e+00  2.6219e-02 93.6050 < 2.2e-16 ***
#> shock        1.5318e-02  3.4994e-02  0.4377  0.661606
#> y2           2.1298e-01  3.1689e-02  6.7207 1.974e-11 ***
#> shock_y2    -1.3977e-01  4.9478e-02 -2.8250  0.004744 **
#> income       3.0399e-04  5.1187e-05  5.9389 3.028e-09 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Total Sum of Squares:    5478.3
#> Residual Sum of Squares: 5393.6
#> R-Squared:      0.015466
#> Adj. R-Squared: 0.014814
#> F-statistic: 23.7011 on 4 and 6035 DF, p-value: < 2.22e-16
```

プーリング OLS vs 固定効果モデル：

プーリング OLS は個人固有効果を無視するため、もし個人固有効果が説明変数と相関していれば、推定量に

バイアスが生じます。

F 検定：固定効果の有意性

固定効果モデルとプーリング OLS のどちらが適切かを検定します。

```
# F 検定：固定効果の有意性
pFtest(preg_fe, preg_p)
#>
#>  F test for individual effects
#>
#> data:  life ~ shock + y2 + shock_y2 + income
#> F = 1.0866, df1 = 3018, df2 = 3017, p-value = 0.01125
#> alternative hypothesis: significant effects
```

結果の解釈：

- 帰無仮説：「すべての個人固定効果がゼロ」（プーリング OLS が適切）
- p 値 $= 0.011$ で 5% 水準で有意なので、帰無仮説を棄却し、固定効果モデルが適切と判断します。
- 固定効果が有意であれば、個人差を考慮した分析が必要ということになります。

モデル 4: 変量効果モデル

変量効果モデルは、個人固有効果を確率変数として扱います。固定効果モデルより効率的ですが、個人固有効果と説明変数が無相関という強い仮定が必要です。

```
# 変量効果モデル（random）を用いた回帰
preg_re <- plm(life ~ shock + y2 + shock_y2 + income,
    data = life9,
    effect = "individual",
    model = "random",
    index = c("id", "t")
)
summary(preg_re)
#> Oneway (individual) effect Random Effect Model
#>    (Swamy-Arora's transformation)
#>
#> Call:
#> plm(formula = life ~ shock + y2 + shock_y2 + income, data = life9,
#>     effect = "individual", model = "random", index = c("id",
#>         "t"))
#>
```

```
#> Balanced Panel: n = 3020, T = 2, N = 6040
#>
#> Effects:
#>                   var std.dev share
#> idiosyncratic 0.85660 0.92553 0.958
#> individual    0.03723 0.19294 0.042
#> theta: 0.04082
#>
#> Residuals:
#>     Min.  1st Qu.   Median  3rd Qu.     Max.
#> -2.89772 -0.57890  0.29064  0.47237  1.55096
#>
#> Coefficients:
#>                Estimate  Std. Error z-value  Pr(>|z|)
#> (Intercept)  2.4544e+00  2.6336e-02 93.1956 < 2.2e-16 ***
#> shock        1.5308e-02  3.4995e-02  0.4374  0.661800
#> y2           2.1300e-01  3.1025e-02  6.8652 6.639e-12 ***
#> shock_y2    -1.3978e-01  4.8437e-02 -2.8857  0.003905 **
#> income       3.0332e-04  5.2024e-05  5.8304 5.528e-09 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Total Sum of Squares:    5250.9
#> Residual Sum of Squares: 5169
#> R-Squared:      0.015598
#> Adj. R-Squared: 0.014946
#> Chisq: 95.628 on 4 DF, p-value: < 2.22e-16
```

固定効果 vs 変量効果の違い：

両モデルとも個人固有効果 α_i を確率変数として扱いますが、説明変数との相関についての仮定が異なります。

特徴	固定効果モデル	変量効果モデル
α_i と誤差項の相関	相関があってもよい	無相関を仮定
推定方法	Within 変換（個人平均からの偏差）	GLS（一般化最小二乗法）
効率性	低い	高い（仮定が正しければ）
一致性	常に一致推定量	無相関の仮定が満たされれば一致
時間不変変数	推定不可（Within 変換で消える）	推定可能

ポイント：固定効果モデルは「個人効果と説明変数が相関している」という現実的な状況でも一致推定量を与えるため、因果推論ではより頑健です。一方、変量効果モデルは無相関の仮定が正しければより効率的（標準誤差が小さい）な推定が可能です。過去には固定効果を未知の定数、変量効果を確率変数とみなしたが、現在はどちらも確率変数とかんげて、違いは誤差項と相関があるかどうかの仮定が異なるだけです。

Hausman 検定：固定効果 vs 変量効果

どちらのモデルが適切かを検定します。

```
# Hausman 検定
phtest(preg_re, preg_fe)
#>
#>  Hausman Test
#>
#> data:  life ~ shock + y2 + shock_y2 + income
#> chisq = 0.27439, df = 3, p-value = 0.9648
#> alternative hypothesis: one model is inconsistent
```

結果の解釈：

- 帰無仮説：「個人固有効果と説明変数は無相関」（変量効果モデルが適切）
- p 値 $= 0.965$ と非常に大きいため、帰無仮説を棄却できません。この場合、変量効果モデルが適切と判断できます。
- 変量効果モデルは固定効果モデルより効率的なので、Hausman 検定で棄却されなければ変量効果モデルを使うことが推奨されます。

教育的ポイント

1. DID の識別仮定

DID が因果効果を正しく識別するためには、**平行トレンド仮定**が必要です：

> 処置がなかった場合、処置群と対照群は同じトレンドで変化していたはず

この仮定が満たされない場合（例：健康ショックを受けやすい人はもともと生活満足度が低下傾向にあった）、DID 推定量にはバイアスが生じます。

2. 固定効果と 1 階差分の関係

2 時点パネルでは：

- **1 階差分法 $=$ 固定効果モデル**（同じ推定量）

3 時点以上のパネルでは：

- 誤差項に系列相関がなければ、固定効果モデルの方が効率的
- 誤差項がランダムウォークなら、1 階差分法の方が効率的

3. **モデル選択の指針**
   1. **pFtest**：固定効果が有意か？ → 有意なら固定効果モデルを使用
   2. **phtest**（Hausman 検定）：個人効果と説明変数は無相関か？
      - 棄却 → 固定効果モデル
      - 棄却されない → 変量効果モデル（効率性のため）

4. **政策評価への応用**

DID は政策評価で広く使われます：

- **例**：ある県で保育所が増設された → 女性の就業率への効果
- **処置群**：保育所が増えた県
- **対照群**：保育所が増えなかった県
- **DID 推定量**：政策の因果効果

ただし、政策の実施が無作為でない場合（例：就業意欲が高い県ほど保育所を増やした）、平行トレンド仮定が満たされない可能性があります。