R 使い方入門：パネルデータ分析とクラスターロバスト標準偏差つき固定効果モデルの実践

データの読み込みと整形

このドキュメントでは、西山慶彦・新谷元嗣・川口大司・奥井亮 (2019)『計量経済学』有斐閣の**第 6 章「パネルデータ分析」**で紹介されている実証例を再現します。

分析テーマは、**既婚女性の労働供給**に関するパネルデータを用いて、**保育所の整備（資本率: cap.rate）が女性の雇用率（emp.rate）に与える影響**を検証することです。これは労働経済学や公共政策の分野で重要なテーマであり、「保育所が増えると女性の就業率は上がるのか？」という因果関係を検証する実証分析の一例です。データは山口（2009）に基づいています。

ここでは `yamaguchi.csv` を読み込み、2000 年以降かつ世帯タイプが `all` の観測だけを残したデータを使って、雇用率と資本率の関係を検討します。

```
# データ読み込みと確認
yamaguchi <- read.csv("yamaguchi.csv")
head(yamaguchi)
#>     pref year  emp.rate  cap.rate      age  age.hus emp.rate.hus      urate
#> 1 北海道 1990 0.2701991 0.1875791 31.64818 34.17721    0.9895076 0.03622666
#> 2 青森県 1990 0.4507844 0.3631905 31.40027 34.24343    0.9820385 0.04486930
#> 3 岩手県 1990 0.5267243 0.2618665 31.54298 34.25208    0.9900675 0.02627886
#> 4 宮城県 1990 0.4206343 0.1327991 31.50385 34.19711    0.9902523 0.02730289
#> 5 秋田県 1990 0.5473229 0.2682563 31.46496 34.20592    0.9918102 0.02717173
#> 6 山形県 1990 0.6777410 0.2228247 31.45862 34.23097    0.9949089 0.01745351
#>    nuc.rate  numhh hh.type
#> 1 0.8240965 250086     all
#> 2 0.5649628  68647     all
#> 3 0.4966100  63126     all
#> 4 0.5790675 107307     all
#> 5 0.4144711  51772     all
#> 6 0.3202995  57159     all
```

```r
yamaguchi <- subset(yamaguchi, year > 1999 & hh.type == "all")
head(yamaguchi)
#>       pref year  emp.rate  cap.rate      age  age.hus emp.rate.hus      urate
#> 95  北海道 2000 0.2862369 0.2125679 32.04964 34.29423    0.9772923 0.04775749
#> 96  青森県 2000 0.4631657 0.4071130 32.06907 34.44721    0.9748427 0.05423297
#> 97  岩手県 2000 0.4862405 0.2966232 32.11034 34.45501    0.9833432 0.04025419
#> 98  宮城県 2000 0.3743681 0.1612467 31.96046 34.25884    0.9757594 0.04895517
#> 99  秋田県 2000 0.5195577 0.3449625 32.30129 34.67442    0.9823769 0.04312861
#> 100 山形県 2000 0.6099338 0.2706212 32.24717 34.69922    0.9860293 0.03341050
#>      nuc.rate  numhh hh.type
#> 95  0.8908764 202266     all
#> 96  0.6382392  55451     all
#> 97  0.5975216  52291     all
#> 98  0.7008059  92572     all
#> 99  0.5464666  40061     all
#> 100 0.4617841  46669     all
```

モデル定義と概要

パネルデータ分析では、**固定効果（fixed effects）**を導入することで、観察できない個体差や時点差をコントロールできます。以下では単純な OLS から、クラスタリングや固定効果を導入したモデル、加えて制御変数を含むモデルまで段階的に推定します。それぞれのモデルは表で比較します。

estimatr パッケージのインストールと読み込み

`estimatr` パッケージは、頑健な標準誤差（ロバスト標準誤差）やクラスター標準誤差を簡単に計算できるパッケージです。`lm_robust()` 関数を使うことで、クロスセクション分析だけでなく、今回のようなパネルデータ分析においても、固定効果モデルやクラスタリングを含む回帰分析を効率的に実行できます。

主な機能：

- **lm_robust()**：頑健標準誤差付きの線形回帰
- **se_type**：標準誤差の種類を指定（“stata” は Stata と同じ HC1 タイプ）
- **clusters**：クラスター標準誤差の計算
- **fixed_effects**：固定効果の指定（ダミー変数を明示的に作らなくてよい）

注意：`estimatr` パッケージはクロスセクション分析で使う `lm()` と同じ感覚で固定効果モデルを推定できる便利なパッケージですが、パネルデータ専用の `plm` パッケージとは異なり、1 階差分法（First Difference）や変量効果モデル（Random Effects）には対応していません。これらの手法が必要な場合は `plm` パッケージを使用してください。

```
library(estimatr)
```

ポイント：

- **クラスター標準誤差**：都道府県内で誤差が相関している可能性があるため、都道府県でクラスタリングして標準誤差を計算します。
- **固定効果**：都道府県固定効果は「各県固有の特性」を、年固定効果は「全国共通のマクロショック」をコントロールします。

クラスターロバスト標準誤差とは？
通常の回帰分析では、各観測値の誤差項が**互いに独立**であると仮定します。しかし、パネルデータでは同じ都道府県のデータが複数年にわたって含まれるため、この仮定が成り立たない可能性があります。

なぜ誤差が相関するのか？

例えば、東京都の 2000 年、2005 年、2010 年のデータを考えてみましょう。東京都には観察できない固有の特性（例：都市化の程度、産業構造、住民の価値観など）があり、これらは 3 時点すべてに影響します。このような観察できない要因が誤差項に含まれると、同じ都道府県内の誤差は互いに相関してしまいます。

問題点：誤差の相関を無視すると、標準誤差が過小評価され、t 値が過大になり、「本当は有意でない変数が有意に見えてしまう」という問題が生じます。

解決策：クラスターロバスト標準誤差

クラスターロバスト標準誤差は、「同じクラスター（ここでは都道府県）内では誤差が相関していても構わない」という緩い仮定のもとで、正しい標準誤差を計算する方法です。

- **クラスター内**：誤差の相関を許容
- **クラスター間**：誤差は独立と仮定

`lm_robust()` では `clusters = pref` と指定することで、都道府県ごとにクラスタリングした標準誤差を計算できます。

注意点：クラスターロバスト標準誤差が適切に機能するためには、クラスターの数がある程度多い（一般的には 30〜50 以上）ことが望ましいとされています。今回のデータでは 47 都道府県がクラスターとなっており、この条件を満たしています。

モデル 1: 単純 OLS（固定効果なし）
まずは固定効果を入れない単純な回帰から始めます。

```
# モデル 1: 単純 OLS
fm1 <- lm_robust(emp.rate ~ cap.rate,
    data = yamaguchi, se_type = "stata")
```

```
summary(fm1)
#>
#> Call:
#> lm_robust(formula = emp.rate ~ cap.rate, data = yamaguchi, se_type = "stata")
#>
#> Standard error type:  HC1
#>
#> Coefficients:
#>             Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper  DF
#> (Intercept)   0.2081    0.01814   11.47 7.493e-22   0.1723   0.2440 139
#> cap.rate      0.6155    0.04391   14.02 2.244e-28   0.5287   0.7023 139
#>
#> Multiple R-squared:  0.5336 ,    Adjusted R-squared:  0.5302
#> F-statistic: 196.5 on 1 and 139 DF,  p-value: < 2.2e-16
```

結果の解釈：cap.rate の係数は約 **0.615** で、1% 水準で有意です。これは「資本率（保育所整備率）が 1 ポイント上がると、雇用率が約 0.615 ポイント上がる」という正の相関を示しています。ただし、この段階では因果関係とは言えません。

モデル 2: 都道府県固定効果

```
# モデル 2: クラスタリングと都道府県固定効果
fm2 <- lm_robust(
  emp.rate ~ cap.rate,
  data = yamaguchi,
  se_type = "stata",
  clusters = pref,
  fixed_effects = pref
)
summary(fm2)
#>
#> Call:
#> lm_robust(formula = emp.rate ~ cap.rate, data = yamaguchi, clusters = pref,
#>     fixed_effects = pref, se_type = "stata")
#>
#> Standard error type:  stata
#>
#> Coefficients:
#>          Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper DF
```

```
#> cap.rate   0.8051    0.05531   14.56 7.736e-19   0.6938   0.9164 46
#>
#> Multiple R-squared:  0.9836 ,    Adjusted R-squared:  0.9754
#> Multiple R-squared (proj. model):  0.859 ,   Adjusted R-squared (proj. model):  0.7878
#> F-statistic (proj. model): 211.9 on 1 and 46 DF,  p-value: < 2.2e-16
```

結果の解釈：都道府県固定効果を入れると、cap.rate の係数は約 **0.805** に上昇し、依然として 1% 水準で有意です。これは各県固有の特性（産業構造、文化など）をコントロールした上でも、保育所整備と雇用率の間に強い正の関係があることを示しています。

モデル 3: 年固定効果

```
# モデル 3: 年の固定効果
fm3 <- lm_robust(
  emp.rate ~ cap.rate,
  data = yamaguchi,
  se_type = "stata",
  clusters = pref,
  fixed_effects = year
)
summary(fm3)
#>
#> Call:
#> lm_robust(formula = emp.rate ~ cap.rate, data = yamaguchi, clusters = pref,
#>     fixed_effects = year, se_type = "stata")
#>
#> Standard error type:  stata
#>
#> Coefficients:
#>          Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper DF
#> cap.rate   0.5848    0.07302    8.01 2.864e-10   0.4378   0.7318 46
#>
#> Multiple R-squared:  0.5461 ,    Adjusted R-squared:  0.5361
#> Multiple R-squared (proj. model):  0.4897 ,  Adjusted R-squared (proj. model):  0.4785
#> F-statistic (proj. model): 64.15 on 1 and 46 DF,  p-value: 2.864e-10
```

結果の解釈：年固定効果のみを入れた場合、cap.rate の係数は約 **0.585** で有意です。全国共通の時間トレンド（景気変動など）を除去しても正の関係が残っています。

モデル 4: 都道府県 + 年 の二元固定効果（Two-way FE）

```
# モデル 4: 都道府県 + 年 の固定効果
fm4 <- lm_robust(
  emp.rate ~ cap.rate,
  data = yamaguchi,
  se_type = "stata",
  clusters = pref,
  fixed_effects = pref + year
)
summary(fm4)
#>
#> Call:
#> lm_robust(formula = emp.rate ~ cap.rate, data = yamaguchi, clusters = pref,
#>     fixed_effects = pref + year, se_type = "stata")
#>
#> Standard error type:  stata
#>
#> Coefficients:
#>          Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper DF
#> cap.rate  0.09032    0.07357   1.228   0.2258 -0.05776   0.2384 46
#>
#> Multiple R-squared:  0.9946 ,    Adjusted R-squared:  0.9917
#> Multiple R-squared (proj. model):  0.02776 , Adjusted R-squared (proj. model):  -0.4958
#> F-statistic (proj. model): 1.507 on 1 and 46 DF,  p-value: 0.2258
```

結果の解釈（重要！）：都道府県と年の両方の固定効果を入れると、cap.rate の係数は約 **0.090** に大幅に低下し、**統計的に有意ではなくなります**（p 値 $= 0.226$）。これは非常に重要な発見です。

なぜ係数が小さくなったのか？

都道府県固定効果は「県ごとの平均的な違い」を、年固定効果は「全国共通の時間変化」を吸収します。両方を入れると、「ある県で、ある年に、他の県や年と比べて保育所が増えたとき、雇用率がどう変わるか」という、より厳密な因果関係に近い推定になります。その結果、単純な OLS で見られた強い正の相関の多くは、実は県固有の特性や全国的なトレンドによるものだった可能性があります。

モデル 5: 制御変数を追加（年固定効果）

```
# モデル 5: 制御変数を追加（年固定効果）
fm5 <- lm_robust(
```

```
  emp.rate ~ cap.rate + age + age.hus + emp.rate.hus + urate,
  data = yamaguchi,
  se_type = "stata",
  clusters = pref,
  fixed_effects = year
)
summary(fm5)
#>
#> Call:
#> lm_robust(formula = emp.rate ~ cap.rate + age + age.hus + emp.rate.hus +
#>     urate, data = yamaguchi, clusters = pref, fixed_effects = year,
#>     se_type = "stata")
#>
#> Standard error type:  stata
#>
#> Coefficients:
#>               Estimate Std. Error  t value  Pr(>|t|) CI Lower CI Upper DF
#> cap.rate      0.535270    0.06821  7.84698 4.976e-10   0.3980   0.6726 46
#> age          -0.003658    0.09312 -0.03928 9.688e-01  -0.1911   0.1838 46
#> age.hus      -0.050556    0.08120 -0.62260 5.366e-01  -0.2140   0.1129 46
#> emp.rate.hus  0.065030    0.44035  0.14768 8.832e-01  -0.8214   0.9514 46
#> urate        -0.204517    1.05466 -0.19392 8.471e-01  -2.3274   1.9184 46
#>
#> Multiple R-squared:  0.5887 ,    Adjusted R-squared:  0.567
#> Multiple R-squared (proj. model):  0.5376 ,  Adjusted R-squared (proj. model):  0.5132
#> F-statistic (proj. model): 26.79 on 5 and 46 DF,  p-value: 1.395e-12
```

結果の解釈：年固定効果に加えて、妻の年齢（age）、夫の年齢（age.hus）、夫の雇用率（emp.rate.hus）、失業率（urate）を制御変数として追加しました。cap.rate の係数は約 **0.535** で依然として有意です。制御変数を加えてもなお正の効果が確認されます。

モデル 6: 制御変数を追加（二元固定効果）

```
# モデル 6: 制御変数を追加（都道府県 + 年 固定効果）
fm6 <- lm_robust(
  emp.rate ~ cap.rate + age + age.hus + emp.rate.hus + urate,
  data = yamaguchi,
  se_type = "stata",
  clusters = pref,
```

```
  fixed_effects = pref + year
)
summary(fm6)
#>
#> Call:
#> lm_robust(formula = emp.rate ~ cap.rate + age + age.hus + emp.rate.hus +
#>     urate, data = yamaguchi, clusters = pref, fixed_effects = pref +
#>     year, se_type = "stata")
#>
#> Standard error type:  stata
#>
#> Coefficients:
#>              Estimate Std. Error t value Pr(>|t|)  CI Lower CI Upper DF
#> cap.rate      0.11402    0.07314  1.5590  0.12584 -0.033195  0.26124 46
#> age          -0.02661    0.03393 -0.7844  0.43681 -0.094905  0.04168 46
#> age.hus       0.04088    0.03341  1.2236  0.22732 -0.026366  0.10812 46
#> emp.rate.hus  0.31107    0.15765  1.9731  0.05451 -0.006268  0.62841 46
#> urate        -0.66126    0.30333 -2.1800  0.03441 -1.271829 -0.05070 46
#>
#> Multiple R-squared:  0.996 , Adjusted R-squared:  0.9936
#> Multiple R-squared (proj. model):  0.2848 ,  Adjusted R-squared (proj. model):  -0.1508
#> F-statistic (proj. model): 2.468 on 5 and 46 DF,  p-value: 0.04618
```

結果の解釈：二元固定効果に制御変数を加えたモデルでは、cap.rate の係数は約 **0.114** で、やはり**有意ではありません**（p 値 $= 0.126$）。一方、失業率（urate）は係数 **-0.661** で 5% 水準で有意です。これは「失業率が高い地域・時期ほど女性の雇用率が低い」という直感に合った結果です。

modelsummary パッケージのインストールと読み込み

`modelsummary` パッケージは、回帰分析の結果を見やすい表形式で出力するためのツールです。複数のモデルを横に並べて比較したり、記述統計を作成したりできます。論文やレポートで使える高品質な表を簡単に作成できるため、実証分析では非常に便利です。

主な機能：

- **modelsummary()**：複数の回帰モデルを 1 つの表にまとめて比較
- **datasummary()**：記述統計表の作成
- **coef_omit / gof_omit**：不要な係数や適合度指標を非表示にするオプション

```
library(modelsummary)
```

記述統計（主要変数）

分析に使用する主要変数の記述統計を示します。141 の観測値（47 都道府県 × 3 年＝ 2000 年、2005 年、2010 年）があります。

- **emp.rate（母親就業率）**：平均 0.431、標準偏差 0.097、範囲 0.229〜0.640。既婚女性の約 43% が就業しています。
- **cap.rate（保育所定員率）**：平均 0.363、標準偏差 0.115、範囲 0.138〜0.655。保育所の整備状況を表します。
- **age（母親平均年齢）**：平均 32.730 歳、標準偏差 0.670。
- **age.hus（父親平均年齢）**：平均 34.834 歳、標準偏差 0.531。
- **emp.rate.hus（父親就業率）**：平均 0.965、標準偏差 0.018。父親のほとんど（約 97%）が就業しています。
- **urate（失業率）**：平均 0.057（5.7%）、標準偏差 0.015、範囲 0.030〜0.119。地域の労働市場の状況を反映します。

```
library(tidyverse)
library(kableExtra)
vars <- yamaguchi %>%
  select(emp.rate, cap.rate, age, age.hus, emp.rate.hus, urate) %>%
  rename(
    "母親就業率" = emp.rate,
    "保育所定員率" = cap.rate,
    "母親平均年齢" = age,
    "父親平均年齢" = age.hus,
    "父親就業率" = emp.rate.hus,
    "失業率" = urate
  )

table63 <- datasummary(
  All(vars) ~ N + Mean + SD + Min + Max,
  data = yamaguchi,
  fmt = 3)
## 表出力（フォーマット別に処理）
table63
```

日本語を反映させると以下になる。

```
# library(gt)
# colnames(table63) <- c("変数", "サンプルサイズ", "平均", "標準偏差", "最小値", "最大値")
# # 変数名
```

	N	Mean	SD	Min	Max
母親就業率	141	0.431	0.097	0.229	0.640
保育所定員率	141	0.363	0.115	0.138	0.655
母親平均年齢	141	32.730	0.670	31.490	34.764
父親平均年齢	141	34.834	0.531	33.877	36.859
父親就業率	141	0.965	0.018	0.878	0.989
失業率	141	0.057	0.015	0.030	0.119

```
# table63[,1] <- c("母親就業率", "保育所定員率", "母親平均年齢", "父親平均年齢", "父親就業率", "失業率")
# # 表を出力
# gt(table63)
```

年別記述統計（主要変数）

母親就業率と保育所定員率の年別推移を示します。2000 年から 2010 年にかけて、両変数とも増加傾向にあることがわかります。

- **母親就業率**：2000 年の平均 0.392 → 2005 年 0.430 → 2010 年 0.471 と、10 年間で約 8 ポイント上昇。
- **保育所定員率**：2000 年の平均 0.320 → 2005 年 0.361 → 2010 年 0.407 と、同様に約 9 ポイント上昇。

この時系列的な共変動が、単純 OLS で強い正の相関が見られる一因と考えられます。

```
table64 <- datasummary(emp.rate * (Mean + SD) + cap.rate * (Mean + SD) ~ factor(year),data = yamaguc
table64
#>                  2000  2005  2010
#> 1 emp.rate Mean 0.392 0.430 0.471
#> 2            SD 0.096 0.091 0.090
#> 3 cap.rate Mean 0.320 0.361 0.407
#> 4            SD 0.102 0.110 0.119
```

日本語を反映させると以下になる。

```
library(gt)
# 列名
colnames(table64) <- c(" ", "変数", "2000", "2005", "2010") # tibble の都合上 1 列目は空白 1 文字とする
# 変数名
table64[,1] <- c("母親就業率", "", "保育所定員率", "")
# 統計量を日本語に直す
table64[,2] <- c("平均", "標準偏差", "平均", "標準偏差")
# 表を出力
```

	変数	2000	2005	2010
母親就業率	平均	0.392	0.430	0.471
	標準偏差	0.096	0.091	0.090
保育所定員率	平均	0.320	0.361	0.407
	標準偏差	0.102	0.110	0.119

```
gt(table64)
```

回帰結果の表（定数項を表示しない）

下の表は 6 つのモデルを横に並べて比較したものです。`coef_omit` により定数項（Intercept）は表示していません。

```
models <- list("(1)" <- fm1,"(2)" <- fm2,"(3)" <- fm3,
               "(4)" <- fm4,"(5)" <- fm5,"(6)" <- fm6)

table65 <- modelsummary(
  models,
  stars = c(`***` = 0.01, `**` = 0.05, `*` = 0.1),
  gof_omit = "IC|Log|AIC|BIC|RMSE|Within|Pseudo|Std.Errors",
  coef_omit = "\\(Intercept\\)",
  output= "table65.png",
  coef_map = c(
    "cap.rate" = "保育所定員率",
    "age" = "母親平均年齢",
    "age.hus" = "父親平均年齢",
    "emp.rate.hus" = "父親就業率",
    "urate" = "失業率"
  ),
  fmt = 3
)
```

```
knitr::include_graphics("table65.png")
```

	(1)	(2)	(3)	(4)	(5)	(6)
保育所定員率	0.615***	0.805***	0.585***	0.090	0.535***	0.114
	(0.044)	(0.055)	(0.073)	(0.074)	(0.068)	(0.073)
母親平均年齢					-0.004	-0.027
					(0.093)	(0.034)
父親平均年齢					-0.051	0.041
					(0.081)	(0.033)
父親就業率					0.065	0.311*
					(0.440)	(0.158)
失業率					-0.205	-0.661**
					(1.055)	(0.303)
Num.Obs.	141	141	141	141	141	141
R2	0.534	0.984	0.546	0.995	0.589	0.996
R2 Adj.	0.530	0.975	0.536	0.992	0.567	0.994

* p < 0.1, ** p < 0.05, *** p < 0.01

簡単な解釈

分析結果のまとめ

モデル	cap.rate 係数	有意性	固定効果
(1) 単純 OLS	0.615	***	なし
(2) 県 FE	0.805	***	都道府県
(3) 年 FE	0.585	***	年
(4) 二元 FE	0.090	n.s.	都道府県 + 年
(5) 年 FE+ 制御変数	0.535	***	年
(6) 二元 FE+ 制御変数	0.114	n.s.	都道府県 + 年

教育的ポイント

1. **固定効果の重要性**：単純な OLS では強い正の相関（0.615）が見られますが、二元固定効果を入れると効果は消えます（0.090、有意でない）。これは「見かけの相関」と「因果関係」の違いを示す好例です。

2. **欠落変数バイアス（Omitted Variable Bias）**：県の特性や時間トレンドをコントロールしないと、推定にバイアスが生じます。例えば、「女性の就業意欲が高い県は保育所も充実している」という場合、保育所の効果を過大評価してしまいます。

3. **クラスター標準誤差**：パネルデータでは同一個体（ここでは都道府県）内で誤差が相関している可能性があるため、クラスター頑健標準誤差を使うことが重要です。

4. **政策的含意**：この分析結果だけでは「保育所を増やしても女性の就業率は上がらない」とは結論づけられません。二元固定効果モデルは非常に厳しい条件での推定であり、効果がゼロというよりも「このデータ・手法では検出できなかった」という解釈が適切です。

(注) 出力の形式や不要な係数をさらに除外したい場合は、`modelsummary()` の `coef_map` や `coef_omit` に正規表現を渡して調整してください。