# パネルデータ Part 1: 基礎と実践

講師：遠山祐太

最終更新：2024-11-17

# はじめに

# パネルデータ（panel data）とは？

- パネルデータ：クロスセクション(cross-sectional) と時系列（time-series）の組み合わせ
- 具体例: 年 $t$ における個人 $i$ の所得
- データ構造例：

| 個人 $i$ \ 時間 $t$ | 1995 | 1996 | 1997 |
|---|---|---|---|
| 1 | $x_{1,1995}$ | $x_{1,1996}$ | $x_{1,1997}$ |
| 2 | $x_{2,1995}$ | $x_{2,1996}$ | $x_{2,1997}$ |

- パネルデータがなぜ有用か？
  1. 変動が多い（クロスセクションかつ時系列の変動）
  2. **時間を通じて変化しない未観測の要因**に対応できる

# 講義の流れ（全２回）

- 今回：フレームワーク
  - モデル・固定効果の導入
  - 推定：標準誤差、ハウスマン検定
  - Rでの実装
- 次回：差の差分法（difference in differences; DID）

# フレームワーク

# パネルデータにおけるフレームワーク

- 個人(クロスセクションユニット) $i$ の 時間 $t$ に関するモデル

$$y_{it} = \beta' x_{it} + \epsilon_{it}, E[\epsilon_{it}|x_{it}] = 0$$

  ただし $\beta = (\beta_0, \beta_1, \ldots, \beta_K)'$, $x_{it} = (1, x_{1it}, \ldots, x_{Kit})'$ は $(K+1)$ 次元ベクトルである。

- もし $x_{it}$ と $\epsilon_{it}$ の間に相関がなければ、このモデルを最小二乗法で推定することができる。
- 懸念：欠落変数バイアス

# 固定効果の導入

- $\epsilon_{it}$ が

$$\epsilon_{it} = \alpha_i + u_{it}$$

のように分割できるとする。

- $\alpha_i$ を**ユニットの固定効果（unit fixed effect）**と呼び、非時変的な未観測の異質性を表す。
- 各ユニット $i$ のダミー変数を組み込むことで固定効果を統制することができる！

$$y_{it} = \beta' x_{it} + \gamma_2 D2_i + \cdots + \gamma_n Dn_i + u_{it}$$

ここで $Dl_i$ は $l = i$ のとき、またその場合に限り 1 をとる変数である。

# 固定効果モデル

- モデル

$$y_{it} = \beta' x_{it} + \alpha_i + u_{it}$$

- 仮定
  1. $u_{it}$ は $(x_{i1}, \cdots, x_{iT})$ と 相関しない、すなわち $E[u_{it}|x_{i1}, \cdots, x_{iT}] = 0$ である。
  2. $(Y_{it}, x_{it})$ は個人 $i$ の間で独立である。
  3. 外れ値がない。
  4. 説明変数 $x_{it}$ と固定効果 $\alpha_i$ に完全な多重共線性がない。

# 仮定１：平均独立

- 仮定１は最小二乗法の仮定より弱い。
- ここで、ユニットに関する時間を通じて一定の要因は $\alpha_i$ が捉えている。
- 上記の $\alpha_i$ をコントロールしたうえで、 $x_{it}$ と $u_{it}$ が無相関であればOK.

# 仮定４：完全な多重共線性がない

- 次のモデルを考えよう。

$$wage_{it} = \beta_0 + \beta_1 experience_{it} + \beta_2 male_i + \beta_3 white_i + \alpha_i + u_{it}$$

  - $experience_{it}$ は、労働者 $i$ が $t$ 期より前に就労していた年数を表す。
- $male_i$ と $white_i$ が多重共線性の問題をもたらす。
- 直観：**非時変的な要因を $\alpha_i$ が捉えているがゆえに、** $\beta_2$ と $\beta_3$ を推定することができない。

# モデルの拡張：様々な固定効果

- **時間固定効果**を加えることもできる。

$$y_{it} = \beta' x_{it} + \alpha_i + \gamma_t + u_{it}$$

- この回帰は、**各時間特有の、全員に共通なショック**を統制している。
- パネルデータは、固定効果を含めることでさまざまな観測されないショックを捉えるのに便利である。

# 推定方法

# 固定効果を伴う推定

- 各個人のダミー変数を追加することにより、モデルを推定できる。
  - **最小二乗ダミー変数（least squares dummy variables: LSDV）推定量**
  - クロスセクションのユニット数が多いと計算負荷が大きい
- 次の**ユニット内変換（within transformation）**がよく用いられる。

# ユニット内変換による推定

- ユニット内変換を行うため、新しい変数 $\tilde{Y}_{it}$ を、

$$\tilde{Y}_{it} = Y_{it} - \bar{Y}_i$$

のように定義する。ここで $\bar{Y}_i = \frac{1}{T}\sum_{t=1}^{T} Y_{it}$ である。

- ユニット内変換を回帰式に適用すると、固定効果 $\alpha_i$ を取り除ける。

$$\tilde{Y}_{it} = \beta' \tilde{X}_{it} + \tilde{u}_{it}$$

この式を最小二乗法で推定する。

- FWL定理と同じ。変数を個人ダミーに回帰して、その残差で更に回帰する。

# 推定におけるユニット内変動の重要性

- 説明変数の変動は精緻な推定に欠かせない。
- ユニット内変換は時間を通じて一定の未観測要因を取り除く。
- しかし、ユニット内変換は $X_{it}$ の変動 (variation) を吸収してしまう。
- ユニット内変換

$$\tilde{X}_{it} = X_{it} - \bar{X}_i$$

  を思い出そう。
  - 変換後の変数 $\tilde{X}_{it}$ は時点 $t$ 間の、ユニット $i$ 内の変動をもつ。
  - もし $X_{it}$ がユニット $i$　内で時間的に固定ならば $\tilde{X}_{it} = 0$ となり、変動がなくなる。

# クラスターに頑健な標準誤差

- 最小二乗法では、2 種類の誤差構造について考えた。
  1. 分散均一性： $Var(u_i) = \sigma^2$
  2. 分散不均一性： $Var(u_i|x_i) = \sigma(x_i)$
- これまで：観測の間の独立性、すなわち $Cov(u_i, u_{i'}) = 0$ を仮定していた。
- パネルデータの設定では、**自己相関（autocorrelation）**に注意しなければならない。
  - 各個人 $i$ の、時点間での $u_{it}$ と $u_{it'}$ の相関
- **クラスターに頑健な標準誤差（cluster-robust standard error）**を利用する
  - クラスター内の誤差同士が相関することを加味して計算した標準誤差
  - パネルデータではユニット $i$ に基づくクラスターを設定することが多い。

# 固定効果モデルとランダム効果モデル

- ランダム効果モデル：ユニット $i$ の固定効果がないモデル。
  - 誤差項の構造を考慮した一般化最小二乗法で推定する。
  - 単純なOLSよりも統計的な意味で効率的な推定量

- しばしば、「固定効果」か「ランダム効果」どちらを選ぶか？が論点となる。

- 伝統的によく用いられている方法：**ハウスマン検定**によって選択する。
  - ハウスマン検定を棄却できなかったらランダム効果
  - 棄却したら固定効果モデル

- ポイント
  1. ハウスマン検定で選ぶべきではない！
  2. 固定効果モデルか否かは分析の設定に基づくべき。

# ポイント１：ハウスマン検定の概略

- ハウスマン検定

$$J = (\hat{\beta}_{FE} - \hat{\beta}_{RE})' Var(\hat{\beta}_{FE} - \hat{\beta}_{RE})(\hat{\beta}_{FE} - \hat{\beta}_{RE}) \sim \chi^2_k$$

  ここで $k$ はパラメタの個数。FE: 固定効果、RE: ランダム効果

- アイデア：FEとREの推定値が大きく異なっていれば棄却、すなわちFEを支持する。
  - 似ているならば、REの仮定が妥当（＝固定効果は不要）
  - ただし、その「違い」の計算に際して、標準誤差を考慮。

# ハウスマン検定をモデル選択に用いる問題点

- FEとREの推定値が異なっていても、FEの標準誤差が大きいと $J$ は小さくなり棄却できない。
  - そしてFEでは、固定効果によってVariationがなくなって標準誤差が大きくなる。
- 本当は交絡因子の除去のために固定効果が必要にも関わらず「検定を棄却した」ということに基づいてREを用いる。
- 一般にREでは推定量の標準誤差が小さく推定されやすい。よって結果も有意に出やすい。
  - 結果、「本来は効果がないのに効果があった」という偽陽性
  - 計量理論におけるプレ・テストという問題として知られる。
- 以上の詳細については奥井「固定効果と変量効果」を参照

# ポイント２：固定効果モデルか否かは分析の設定次第

- 個人の異質性という交絡因子を取りに除くために固定効果モデルを用いる。
- もし固定効果を入れて、推定値が非精緻になる（標準誤差が大きくなる）ならば、それはリサーチデザインやデータの限界。
- 分析やリサーチデザイン上、**固定効果が交絡因子としては重要でない**と説得的に言える場合にのみランダム効果モデルを使う。

# Rでの実践

# 準備

- パネルデータモデルの推定に`fixest`パッケージを使う。それ以外はこれまでと同様

```
rm(list = ls())
library(AER)
library(tidyverse)
library(fixest)
```

# パネルデータを用いた回帰

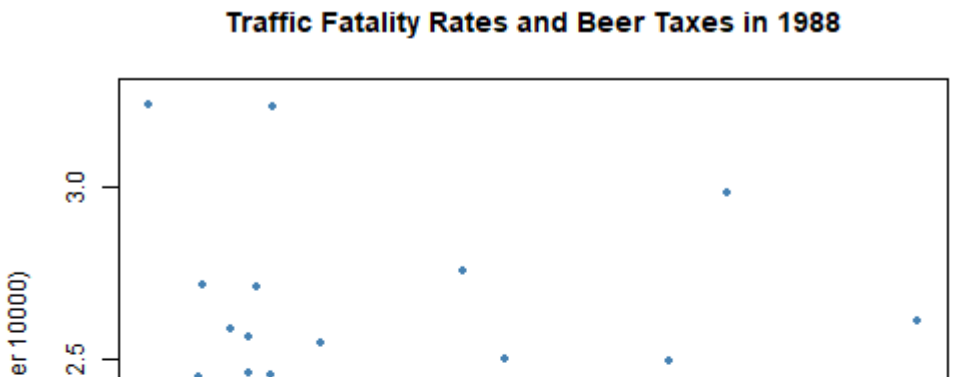
- AER パッケージにある Fatalities データを用いる。詳細はここ

```
## 'data.frame':    336 obs. of  10 variables:
##  $ state   : Factor w/ 48 levels "al","az","ar",..: 1 1 1 1 1 1 1 2 2 2 ...
##  $ year    : Factor w/ 7 levels "1982","1983",..: 1 2 3 4 5 6 7 1 2 3 ...
##  $ spirits : num  1.37 1.36 1.32 1.28 1.23 ...
##  $ unemp   : num  14.4 13.7 11.1 8.9 9.8 ...
##  $ income  : num  10544 10733 11109 11333 11662 ...
##  $ emppop  : num  50.7 52.1 54.2 55.3 56.5 ...
##  $ beertax : num  1.54 1.79 1.71 1.65 1.61 ...
##  $ baptist : num  30.4 30.3 30.3 30.3 30.3 ...
##  $ mormon  : num  0.328 0.343 0.359 0.376 0.393 ...
##  $ drinkage: num  19 19 19 19.7 21 ...
```

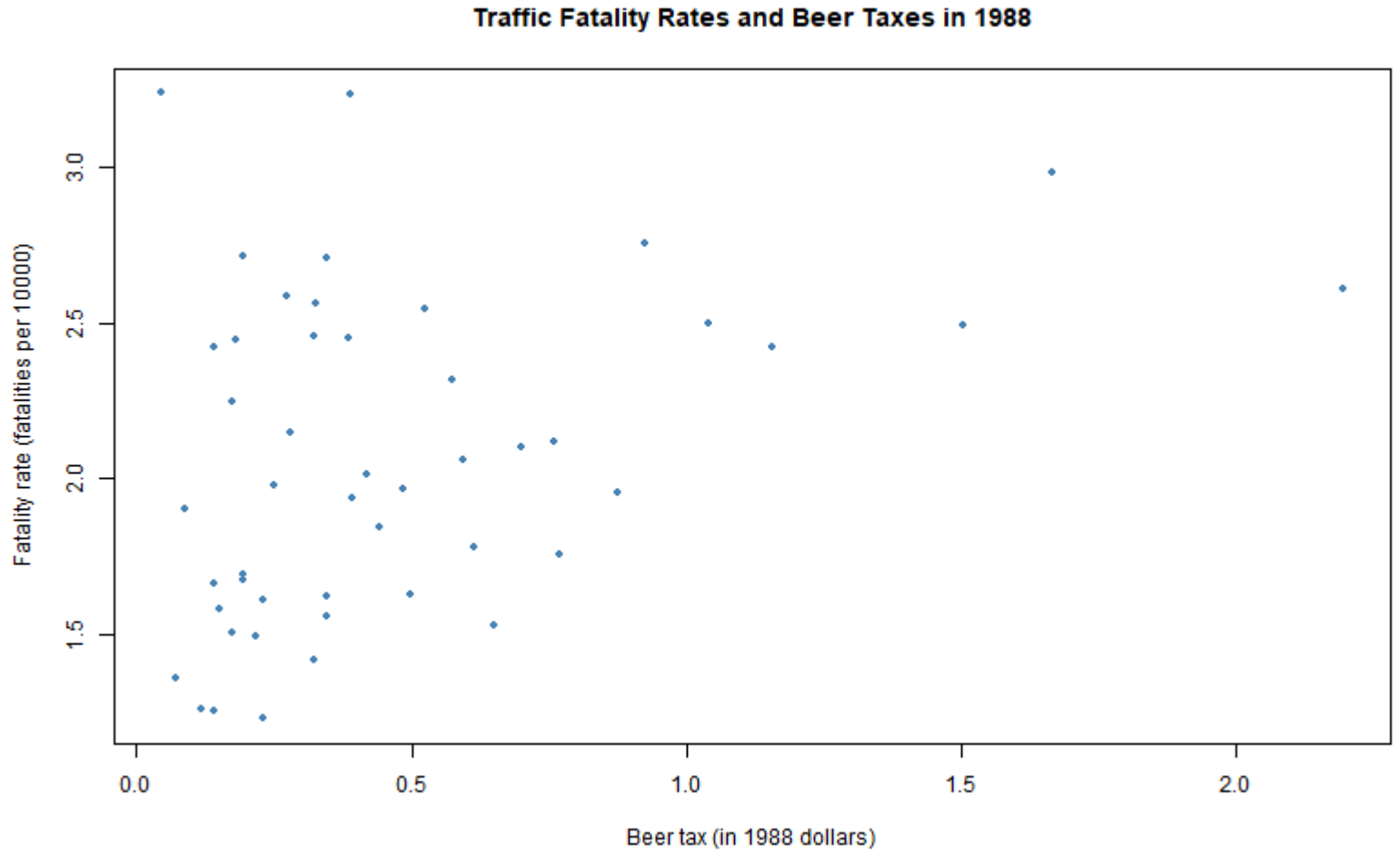
# 自動車事故の死亡率と1988年のビール税の関係

```
data <- Fatalities %>%
    mutate(fatal_rate = fatal / pop * 10000) %>%
    filter(year == "1988")
```

```
plot(x = data$beertax,
     y = data$fatal_rate,
     xlab = "Beer tax (in 1988 dollars)",
     ylab = "Fatality rate (fatalities per 10000)",
     main = "Traffic Fatality Rates and Beer Taxes in 1988",
     pch = 20,
     col = "steelblue")
```



# 酒税と交通事故の相関


Traffic Fatality Rates and Beer Taxes in 1988
Fatality rate (fatalities per 10000)
1.5
2.0
2.5
3.0
0.0
0.5
1.0
1.5
2.0
Beer tax (in 1988 dollars)

# 固定効果モデルの推定

- 固定効果モデルを、`fixest` パッケージの `feols()` 関数で推定する。詳しくはこちら

```
data <- Fatalities %>%
  mutate(fatal_rate = fatal / pop * 10000)

# 通常の最小二乗法
result_ols <- feols(fatal_rate ~ beertax | 0,
                    se = "hetero", # 分散不均一性に頑健な誤差
                    data = data)
```

# 最小二乗法の結果 (固定効果なし)

```
##                          result_ols
## Dependent Var.:          fatal_rate
##
## Constant          1.853*** (0.0471)
## beertax         0.3646*** (0.0529)
## _______________ __________________
## S.E. type       Heteroskedas.-rob.
## Observations                   336
## R2                         0.09336
## Adj. R2                    0.09065
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

# 州の固定効果

```
# 州の固定効果
result_stateFE <- feols(fatal_rate ~ beertax | state,
                        cluster = "state", data = data)

etable(result_stateFE)
```

```
##                    result_stateFE
## Dependent Var.:         fatal_rate
##
## beertax          -0.6559* (0.2919)
## Fixed-Effects:   -----------------
## state                         Yes
## _______________ _________________
## S.E.: Clustered          by: state
## Observations                   336
## R2                         0.90501
## Within R2                  0.04075
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

# 州と年の固定効果

```
result_bothFE <- feols(fatal_rate ~ beertax | state + year,
                       cluster = "state",  data = data)

etable(result_bothFE)
```

```
##                     result_bothFE
## Dependent Var.:         fatal_rate
##
## beertax          -0.6400. (0.3571)
## Fixed-Effects:   -----------------
## state                         Yes
## year                          Yes
## ________________ _________________
## S.E.: Clustered           by: state
## Observations                  336
## R2                        0.90893
## Within R2                 0.03606
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

# 結果の比較

```
etable(list(result_ols, result_stateFE, result_bothFE))
```

```
##                            model 1           model 2           model 3
## Dependent Var.:         fatal_rate        fatal_rate        fatal_rate
##
## Constant          1.853*** (0.0471)
## beertax         0.3646*** (0.0529) -0.6559* (0.2919) -0.6400. (0.3571)
## Fixed-Effects:  ------------------ ----------------- -----------------
## state                          No               Yes               Yes
## year                           No                No               Yes
## _______________ __________________ _________________ _________________
## S.E. type       Heteroskedas.-rob.         by: state         by: state
## Observations                   336               336               336
## R2                         0.09336           0.90501           0.90893
## Within R2                       --           0.04075           0.03606
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

# 標準誤差の比較計算

```
# 誤差がクラスター頑健でない
result_wo_CRS <- feols(fatal_rate ~ beertax | state,
                       vcov = "hetero", # heteroskedasticity-robustのみ
                       data = data)

# 誤差がクラスター頑健
result_w_CRS <- feols(fatal_rate ~ beertax | state,
                      cluster = "state", # cluster-robust
                      data = data )
```

# 標準誤差の比較の結果

```
etable(list(result_ols, result_stateFE))
```

```
##                            model 1           model 2
## Dependent Var.:         fatal_rate        fatal_rate
##
## Constant          1.853*** (0.0471)
## beertax         0.3646*** (0.0529) -0.6559* (0.2919)
## Fixed-Effects:  ------------------ -----------------
## state                           No               Yes
## _______________ __________________ _________________
## S.E. type       Heteroskedas.-rob.          by: state
## Observations                   336                336
## R2                         0.09336            0.90501
## Within R2                       --            0.04075
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```