Mistape Chapter8: Panel Data

牧野泰季
山岸翔一朗
伊藤璃杏

Overview

1- Panel Data

- パネルデータとは何か？
- パネルデータでできること

2- Panel Data Analysis

- Pooled OLS model プールしたOLS推計モデル
- Fixed Effects model　固定効果モデル
- Random Effects model 変量効果モデル
- Which one to use? 分析手法の使い分け

3- Panel Data Analysis- in practice-

- Panel Data Analysis_ Example

4- Appendix

Panel Data

パネルデータとは何か？

データ構造の比較

	一時点のデータ At one point in time	複数時点のデータ At many points in time
変数は1つ Data from one unit	-	時系列データ Time series data
変数は複数 Data from many units	クロスセクションデータ Cross sectional data	パネルデータ Panel data (/ Time series cross section)

パネルデータとは何か？

データ構造の比較

横断面データ/クロスセクションデータ (Cross-section data)

- ある時点における場所・グループ別などに記録した、複数の項目（個体）を集めたデータ
  - e.g.,) ID, 性別, 年齢, 雇用体系, 配偶者の有無 etc.

・横断調査：観察時点が一点（固定）で、複数の項目を観察する

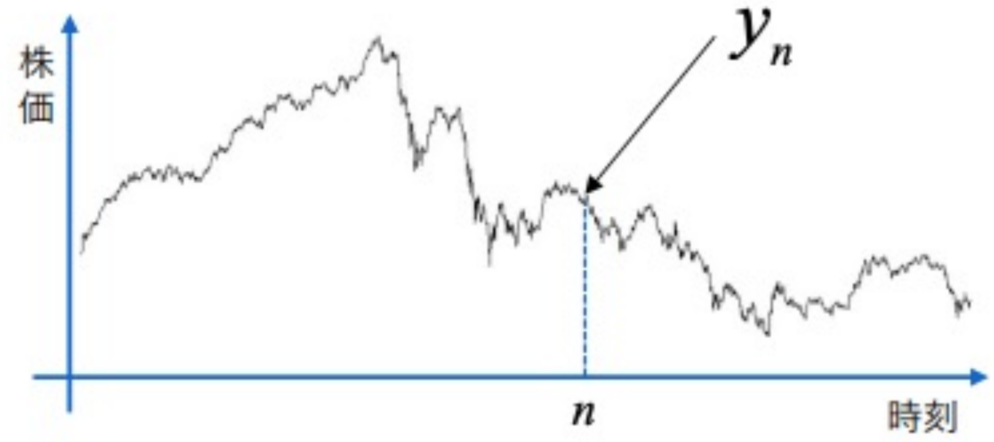
ID	性別	年齢	従業上の地位	配偶者の有無
1	男性	30	正規雇用	有配偶
2	女性	28	正規雇用	無配偶
3	男性	46	自営業	有配偶
…	…	…	…	…

パネルデータとは何か？

データ構造の比較

時系列データ（Time-series data）

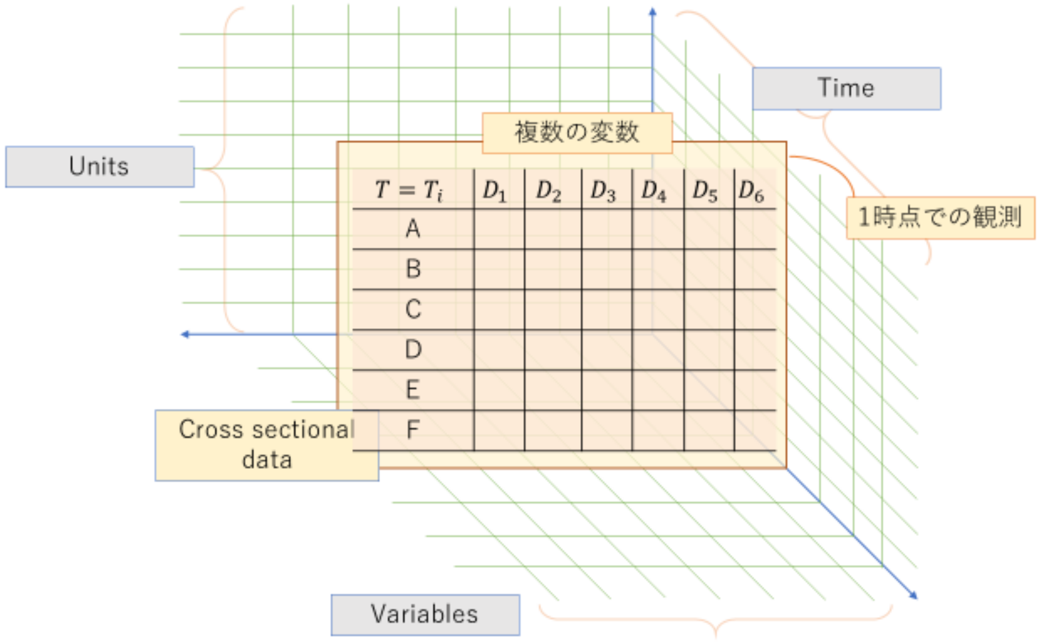
- 時間的に変化した情報を持つデータ
- 4つの変動：トレンド, 季節変動, 循環変動, 不規則変動



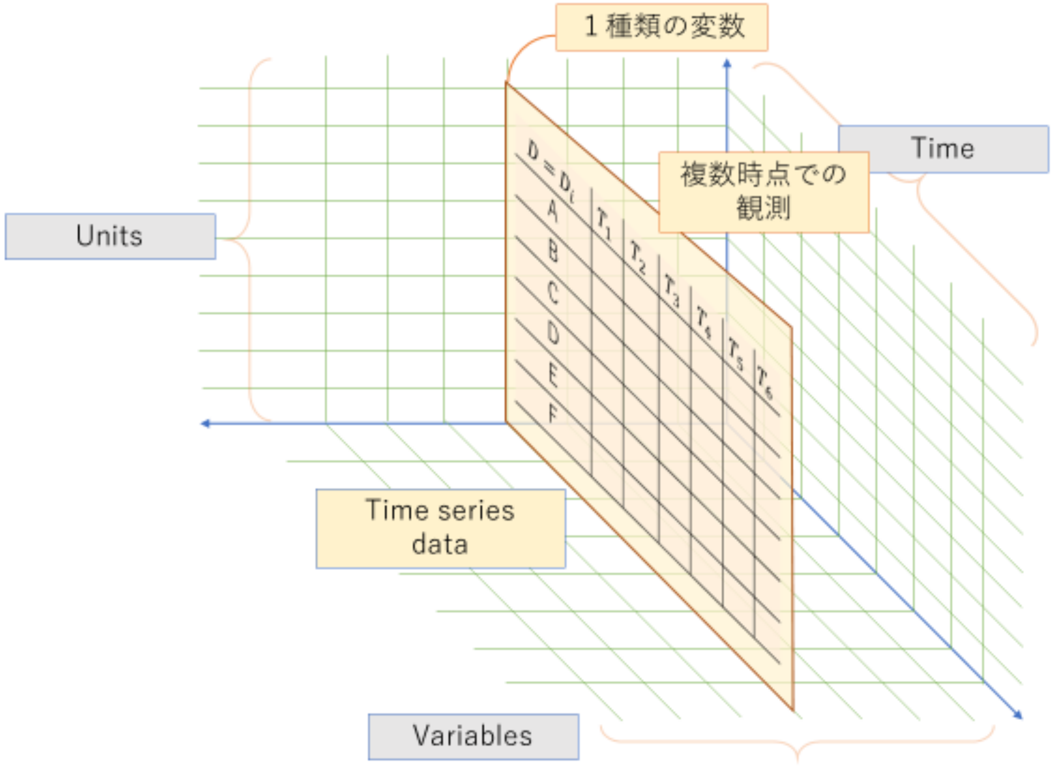
・時系列調査：一つの項目について継続的に観察する

パネルデータとは何か？

データ構造の比較



横断面データ/クロスセクションデータ (Cross-section data)



時系列データ(Time-series data)

パネルデータとは何か？

データ構造の比較

パネルデータ（Panel data）

- 同一のサンプルに対して複数の項目を複数時点にわたって調査したデータ

  → クロスセクションデータ・時系列データの特徴を併せ持つデータ、と言える！

- ・パネル調査：「同一の時点で複数個体に対して行う調査」を複数時点にわたって繰り返し行う

複数項目

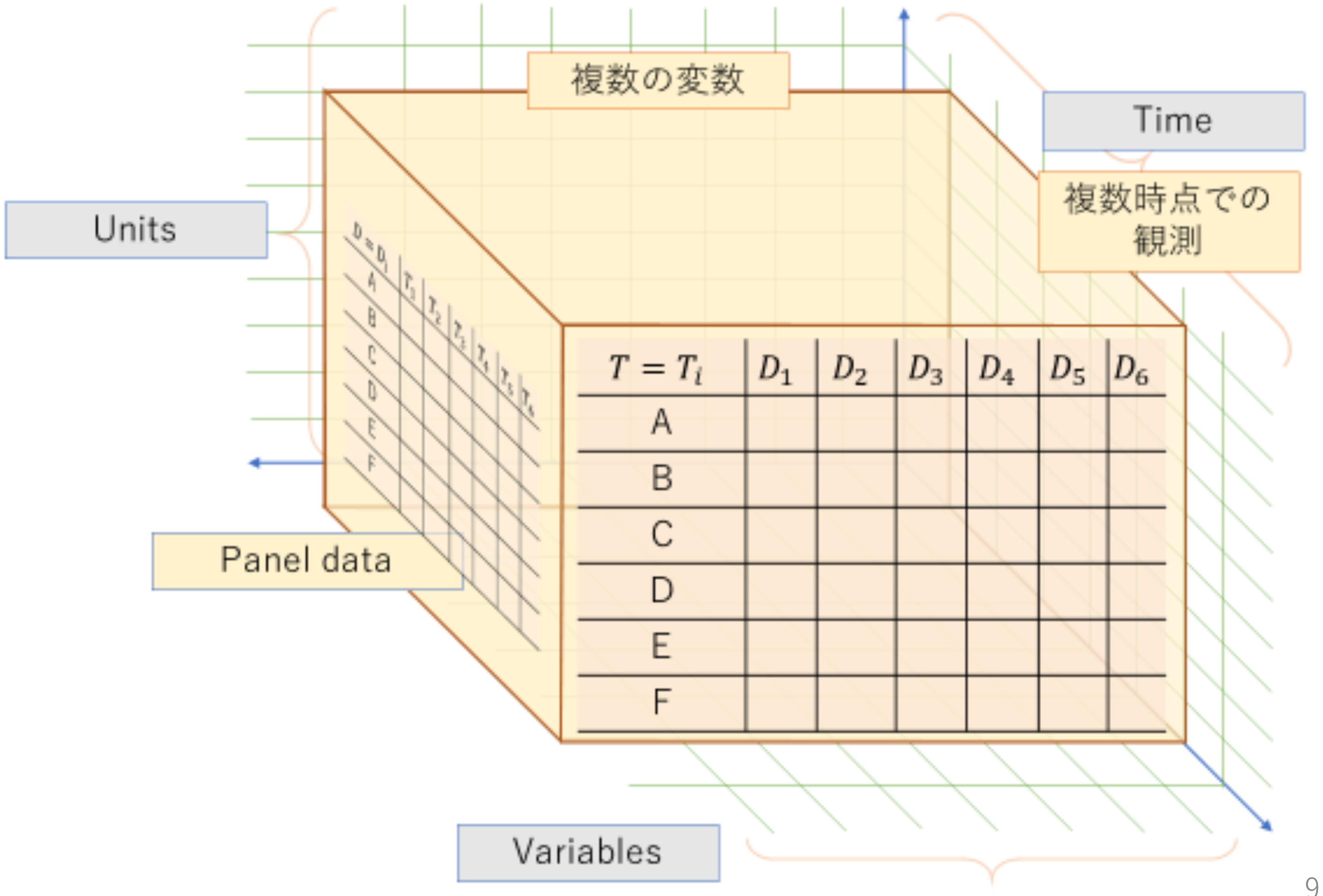
		D_1	D_2	D_3
A	T_1			
B	T_1			
C	T_1			
A	T_2			
B	T_2			
C	T_2			
A	T_3			
B	T_3			
C	T_3			

パネルデータのイメージ↑

パネルデータとは何か？

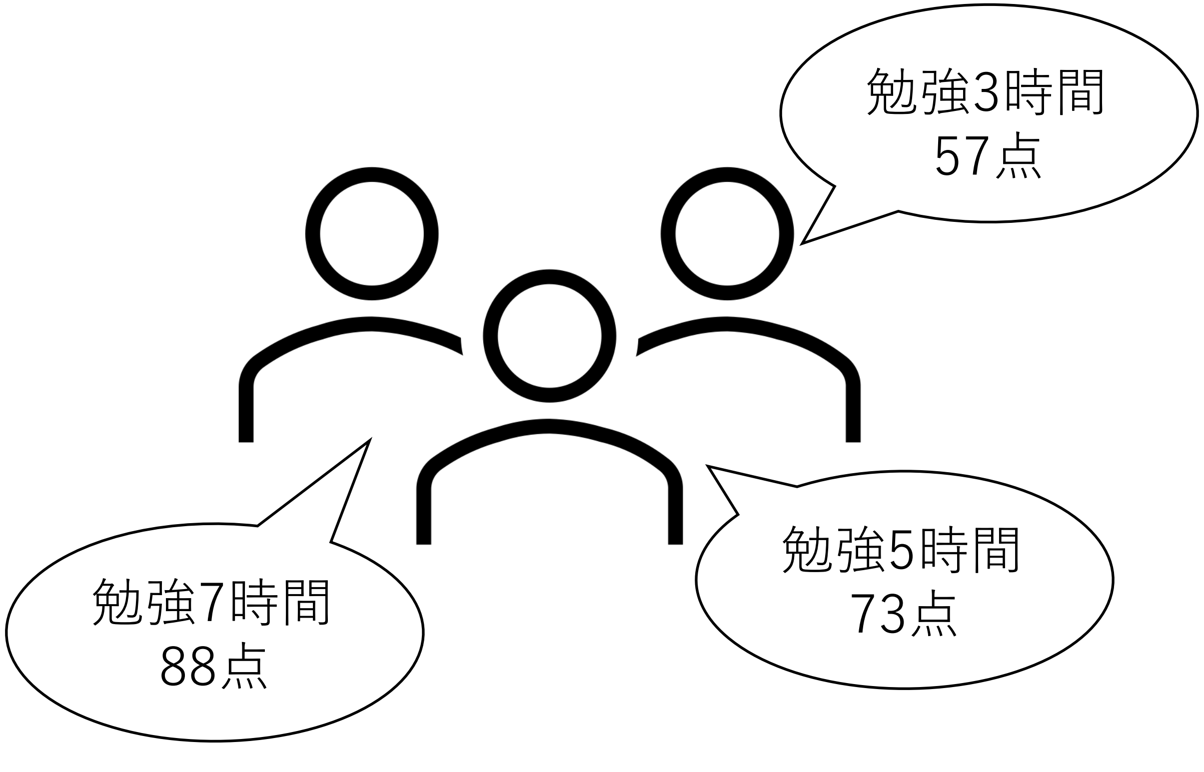
データ構造の比較

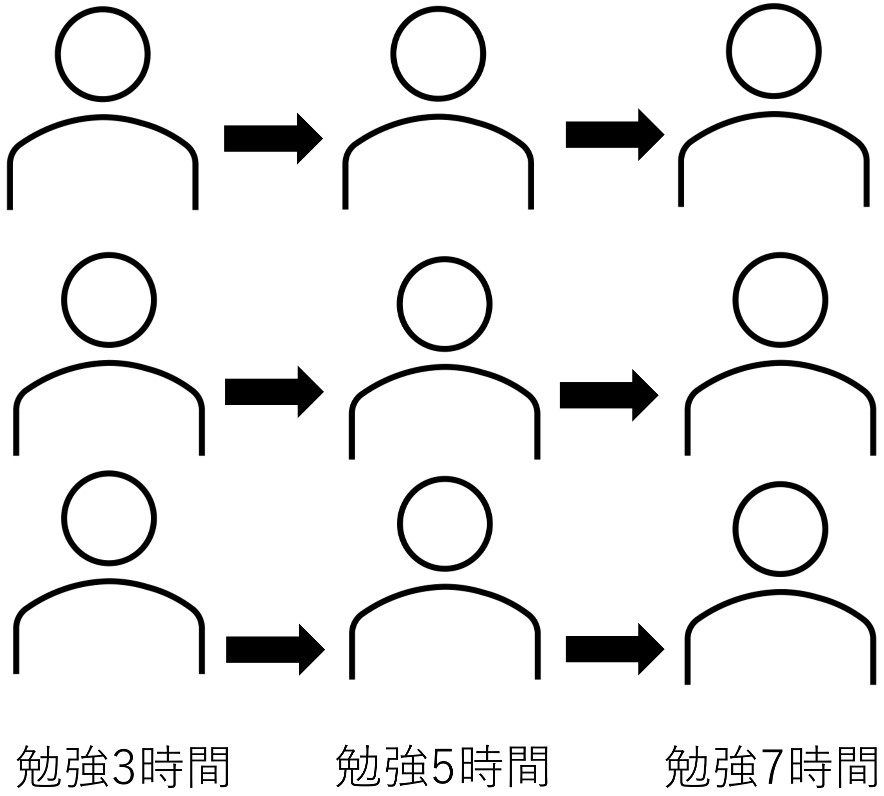
パネルデータ(Panel data)



パネルデータでできること

横断調査とパネル調査





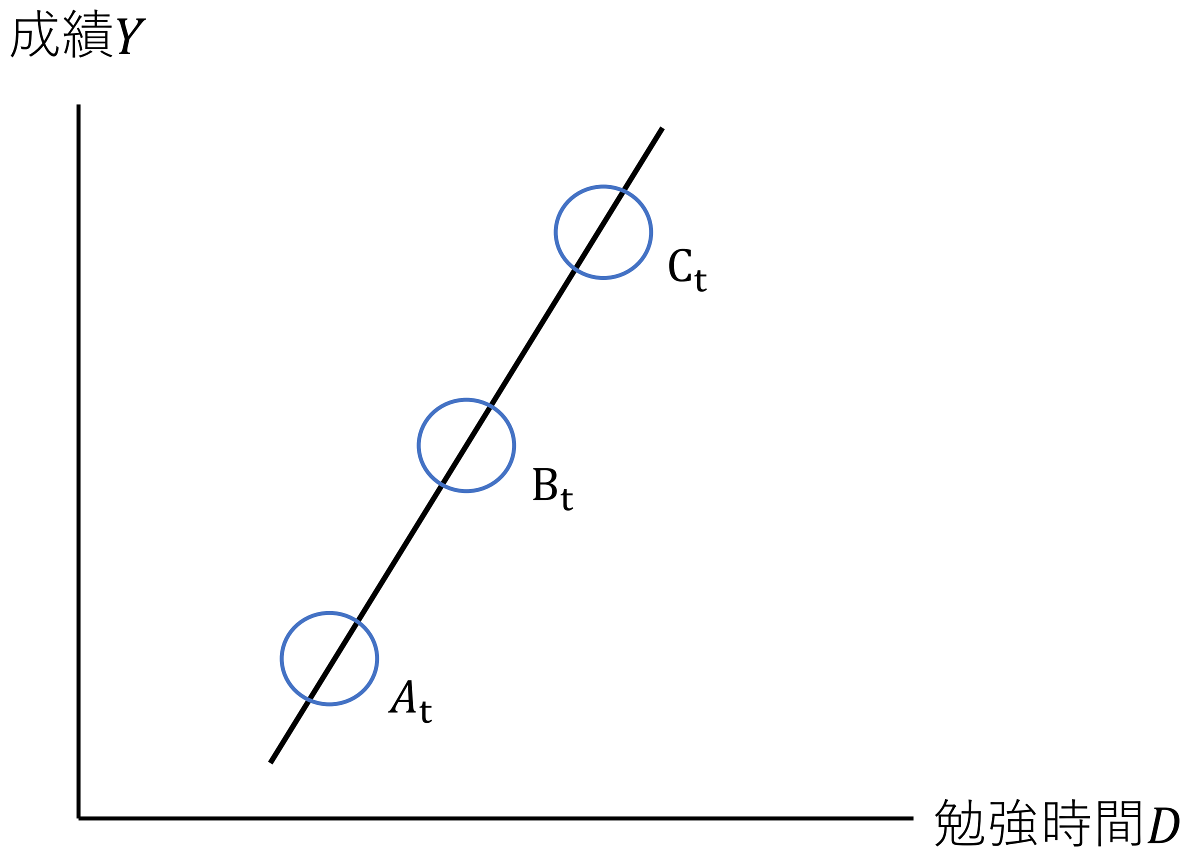
パネルデータでできること

横断調査

例：勉強時間が成績に及ぼす影響

t時点のA、B、Cのデータから
勉強時間Dと成績Yの関係を明らかにしよう
としている

勉強時間の短い個体と長い個体、という別
個体の成績を比べることになる



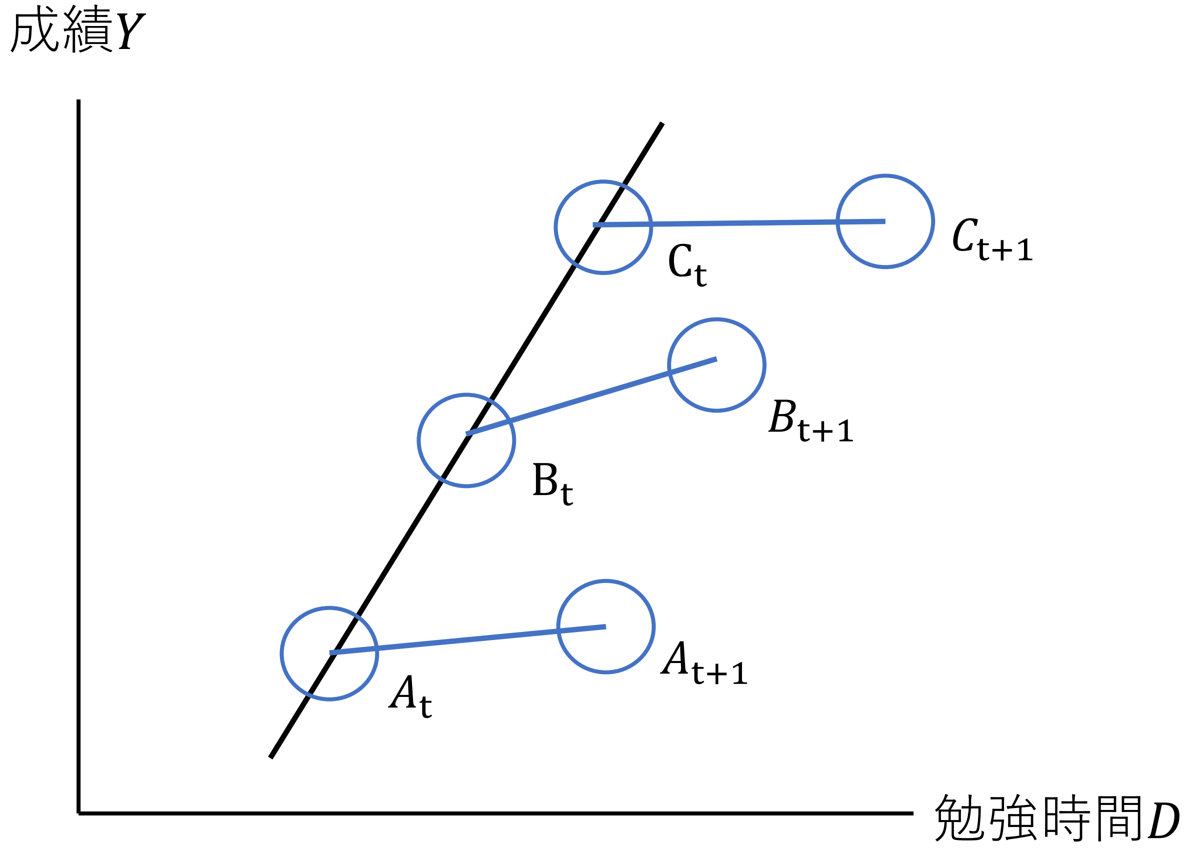
パネルデータでできること

パネル調査

例：勉強時間が成績に及ぼす影響

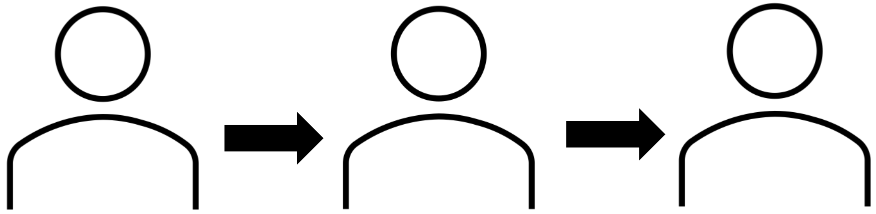
A、B、Cそれぞれについて
t時点のデータとt+1時点のデータを
比較する

A、B、C個々の変化をそれぞれ観察する
ことができる



パネルデータでできること

パネル調査の特徴



- パネルデータの目的
  ①変化の記述
  ②因果関係の立証

- 各個体を各項目（変数）ごとに継続的に追跡したデータが得られる

- 長期にわたって関心が持続される問題や分析の精度（より正確な因果の測定などが求められる問題に向いている
  ↔時事的な即応性が求められるときは不向き

パネルデータでできること

パネル調査の強み

- ✓ これまで観察不可能であった潜在変数を推定したり, 経済主体のダイナミックな変動などを理解することもできるようになる

- ✓ 個体内部の変化が同じ個体内部の変化に与える影響を推定できる

- ✓ **観察されない個々の異質性**をコントロールした上で、注目する要因の影響を明らかにできる

  →観察されない個々の異質性によって引き起こされる**欠落変数バイアス**の一部を取り除くことができる

パネルデータでできること

●**観察されない個々の異質性（Unobserved heterogeneity）：**

直接観察できない個人特性のこと。

性格、能力etc.

- これによって引き起こされるのが**欠落変数バイアス (Omitted variable bias)**

●**欠落変数バイアス (Omitted variable bias)：**

介入や処置と相関の考えられる、本来コントロールすべき(回帰式に加えるべき)

共変量をモデルから取り除いてしまうことで生じるバイアス

パネルデータでできること

パネルデータの強みを活かせる分析：

- 「変化する変数で変化する変数を説明する」調査
  aからa'になった時に、bからb'に変化する

- 向いてる調査、向いていない調査（例）
  ◎ 夫婦の労働時間の変化が家事分担に及ぼす影響
  △親の学歴が子供の行動に与える影響

パネルデータでできること

パネル調査の注意点

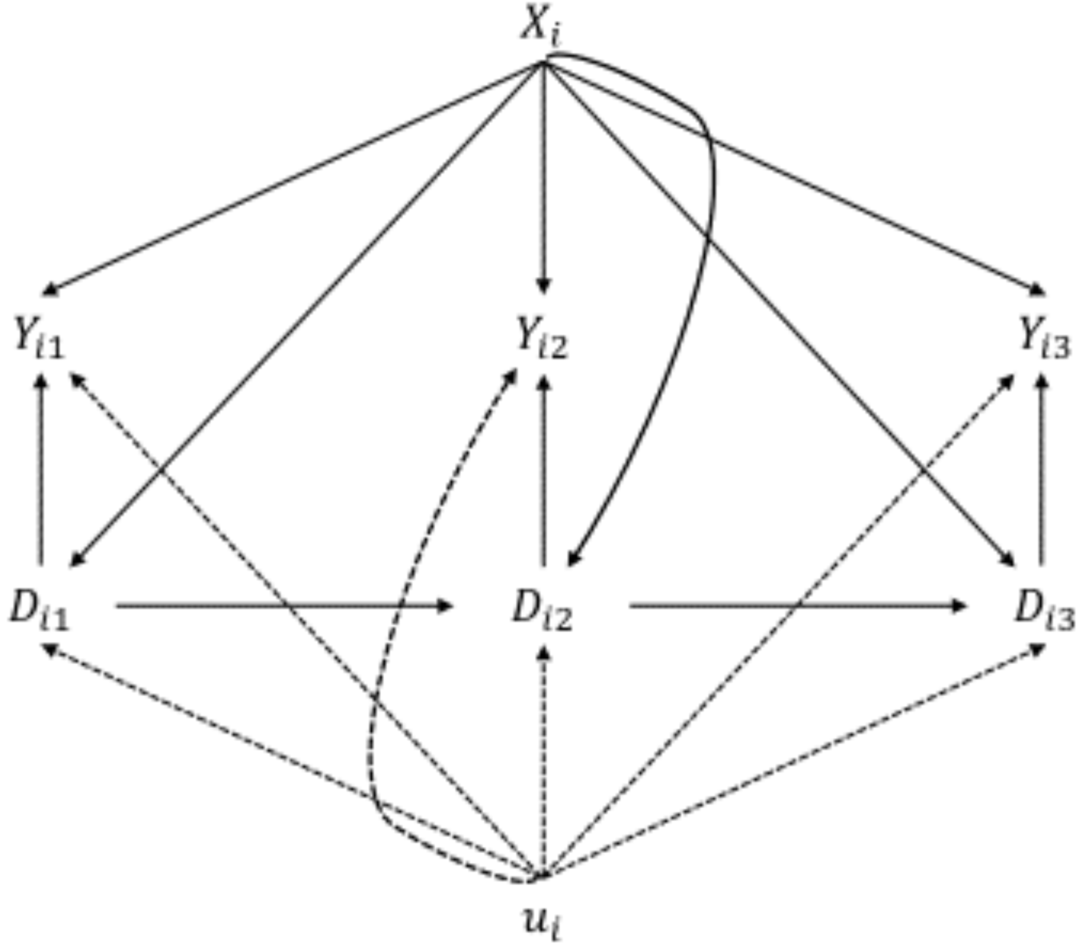
- パネル調査の観察はイベントの発生や個人が経験する変化に応じてなされているわけではない。
  - それほど変化が生じていない複数時点の観察が同一個人から得られることもある
  - サンプルサイズの水増しにより誤差が過小に推定される
    → 回帰分析における**誤差の独立性**が侵害される

- 解決策：類似度の度合いに応じて誤差を数理的に調整する手続きを取る

Panel Data Analysis

DAGによる整理

Y_{it} ：t期における、標本iの被説明変数
D_{it}：t期における、標本iの説明変数
X_i ：標本iに固有の効果（観測可能）
u_i ：標本iに固有の効果（観測不可能）

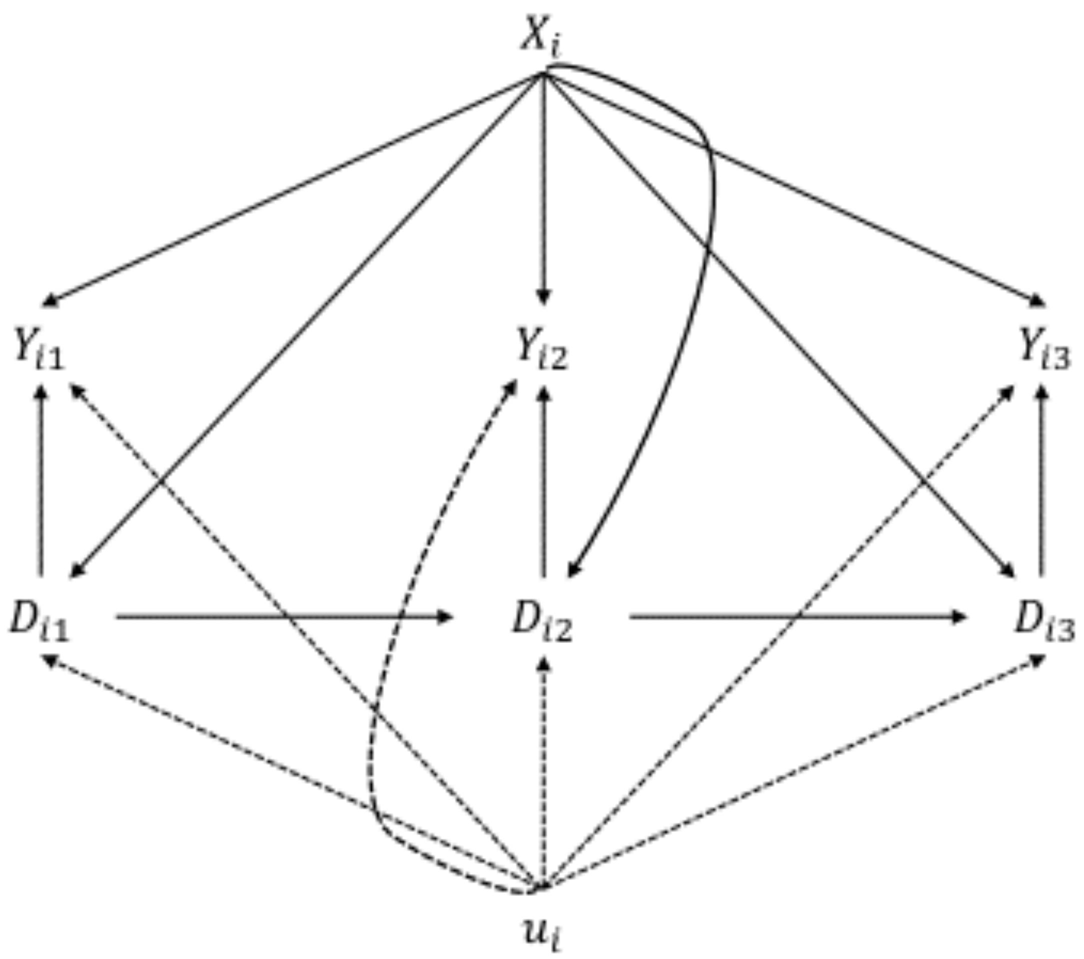


DAGによる整理

Y_{it} ：t期における、標本iの被説明変数
D_{it}：t期における、標本iの説明変数
X_i ：標本iに固有の効果（観測可能）
u_i ：標本iに固有の効果（観測不可能）

① D_{i1}はY_{i1}と次の期間の推定値D_{i2}の両方に影響を与える

$Y_{i1} \leftarrow \mathrm{D}_{i1} \rightarrow \mathrm{D}_{i2}$

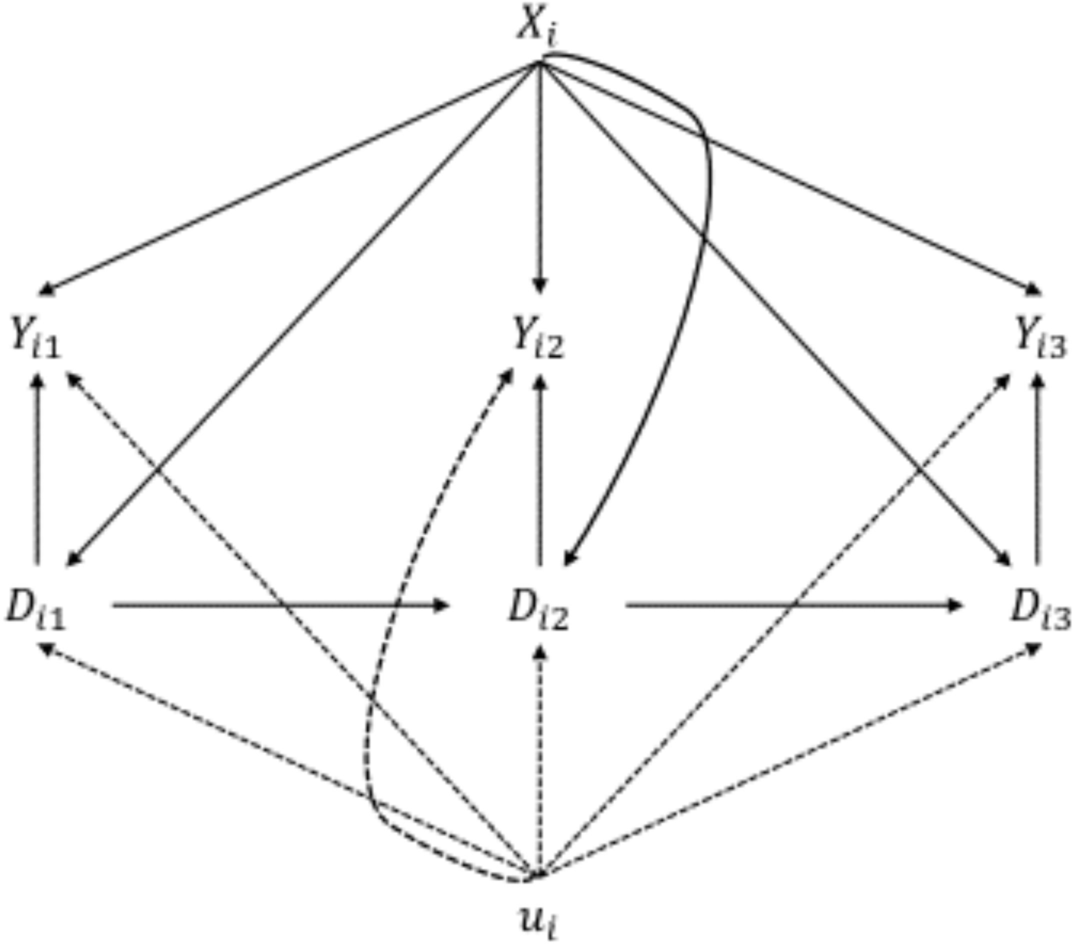


DAGによる整理

Y_{it}　：t期における、標本iの被説明変数
D_{it}：t期における、標本iの説明変数
X_i　：標本iに固有の効果（観測可能）
u_i　：標本iに固有の効果（観測不可能）

②観察されない交絡変数u_iは
すべてのYとDの値を決める

- u_i "は観測されずに回帰分析の誤差項 (structural error term)に含まれる
＝ D_{it} "は内生的

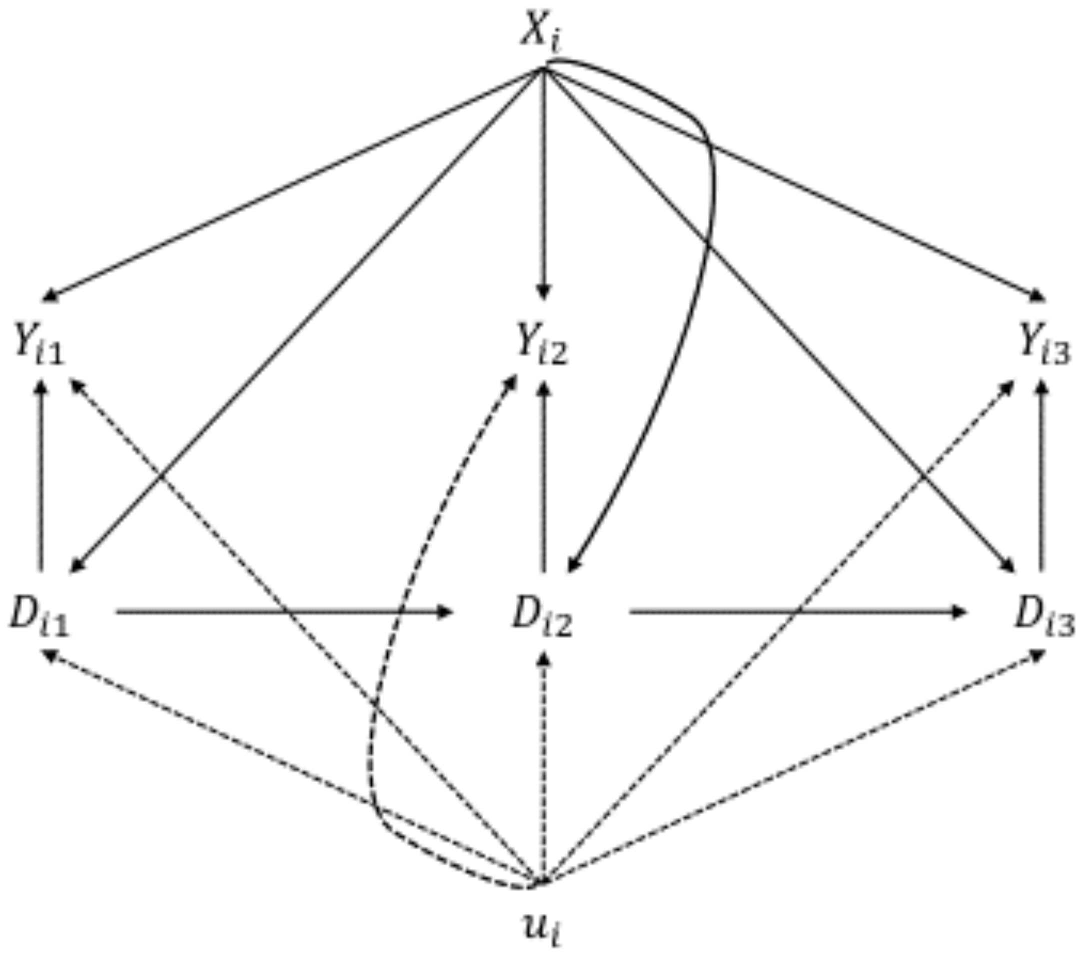


DAGによる整理

Y_{it} ：t期における、標本iの被説明変数
D_{it}：t期における、標本iの説明変数
X_i ：標本iに固有の効果（観測可能）
u_i ：標本iに固有の効果（観測不可能）

③時間によって変化する、観察不可能な交絡因子（time-varying unobserved confounder）はない

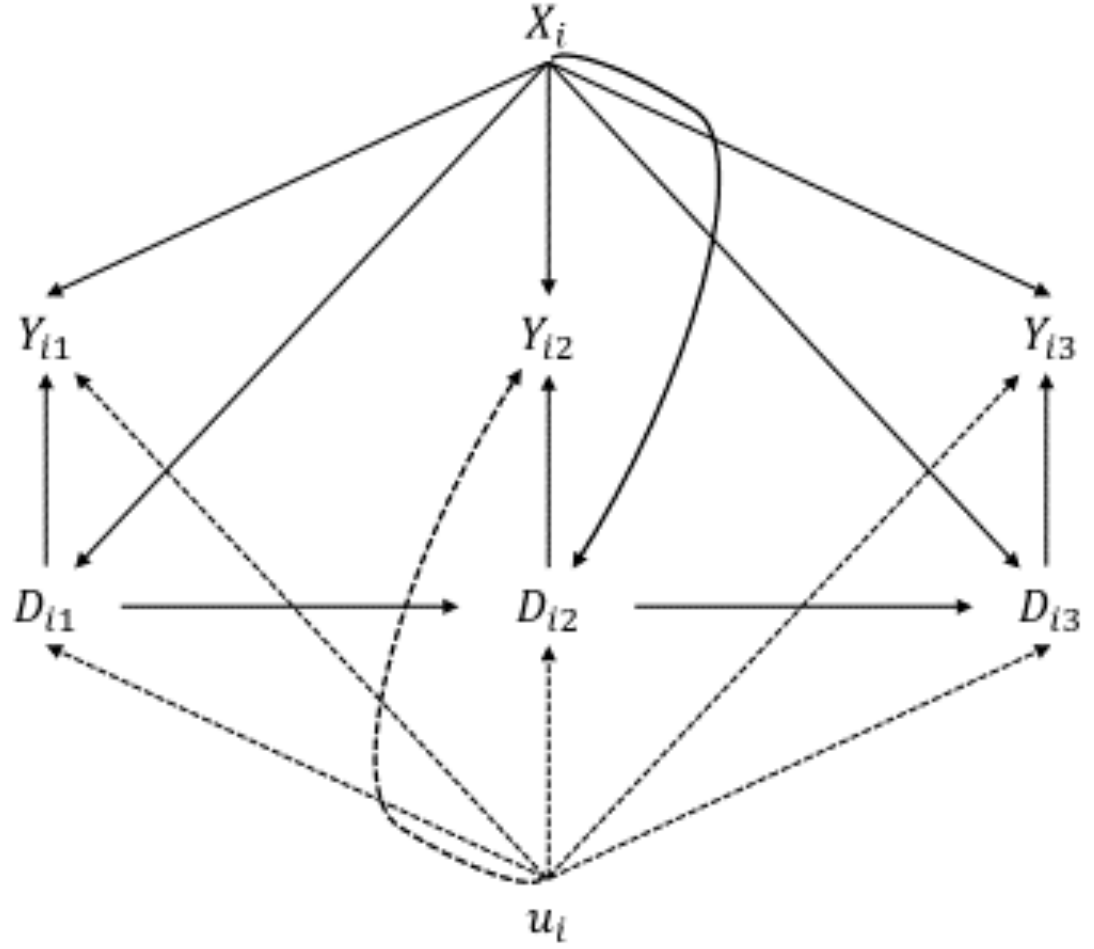
•唯一の交絡因子はu_i
→これを観察されない異質性（unobserved heterogeneity）と呼ぶ



DAGによる整理

Y_{it} ：t期における、標本iの被説明変数
D_{it}：t期における、標本iの説明変数
X_i ：標本iに固有の効果（観測可能）
u_i ：標本iに固有の効果（観測不可能）

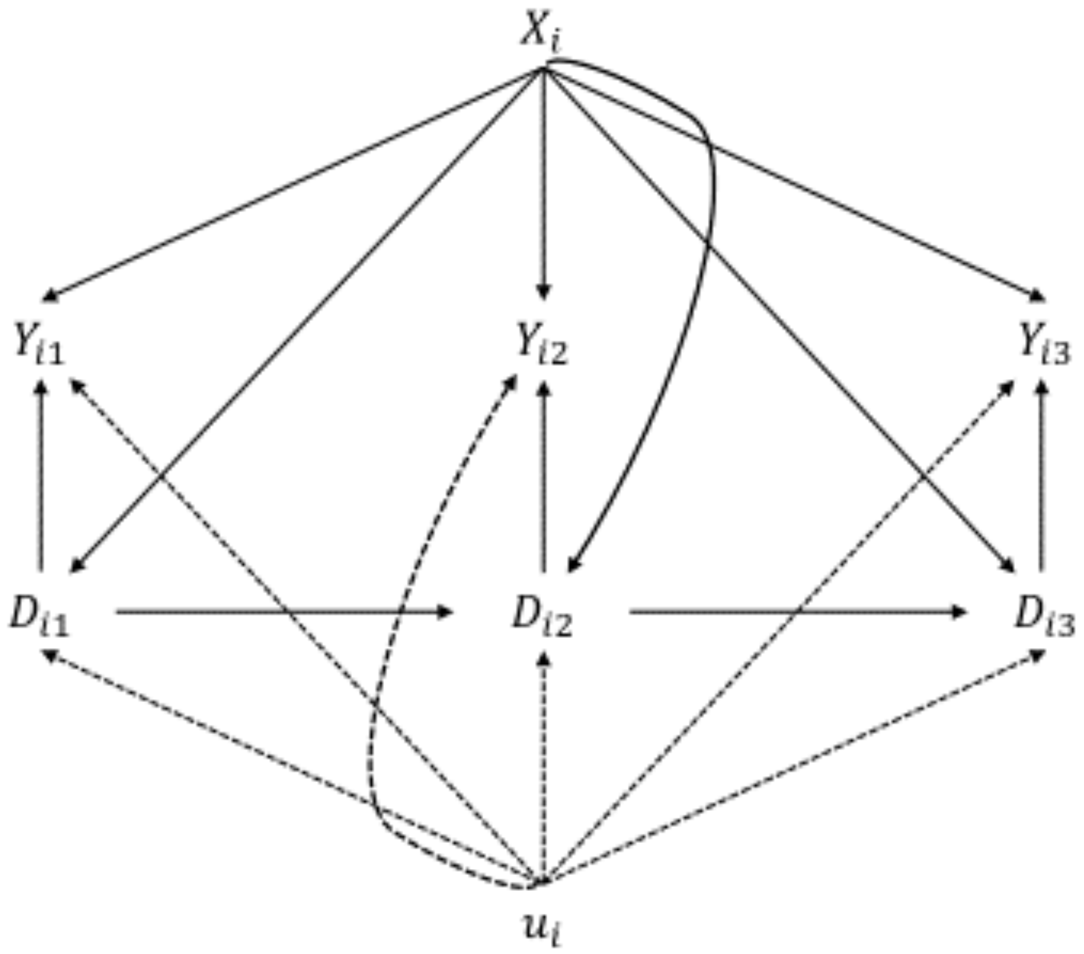
④過去の結果は現在の結果に直接的な影響を及ばさない（Y_{t-1}はY_tに直接的な影響はない）



DAGによる整理

Y_{it} ：t期における、標本iの被説明変数
D_{it} ：t期における、標本iの説明変数
X_i ：標本iに固有の効果（観測可能）
u_i ：標本iに固有の効果（観測不可能）

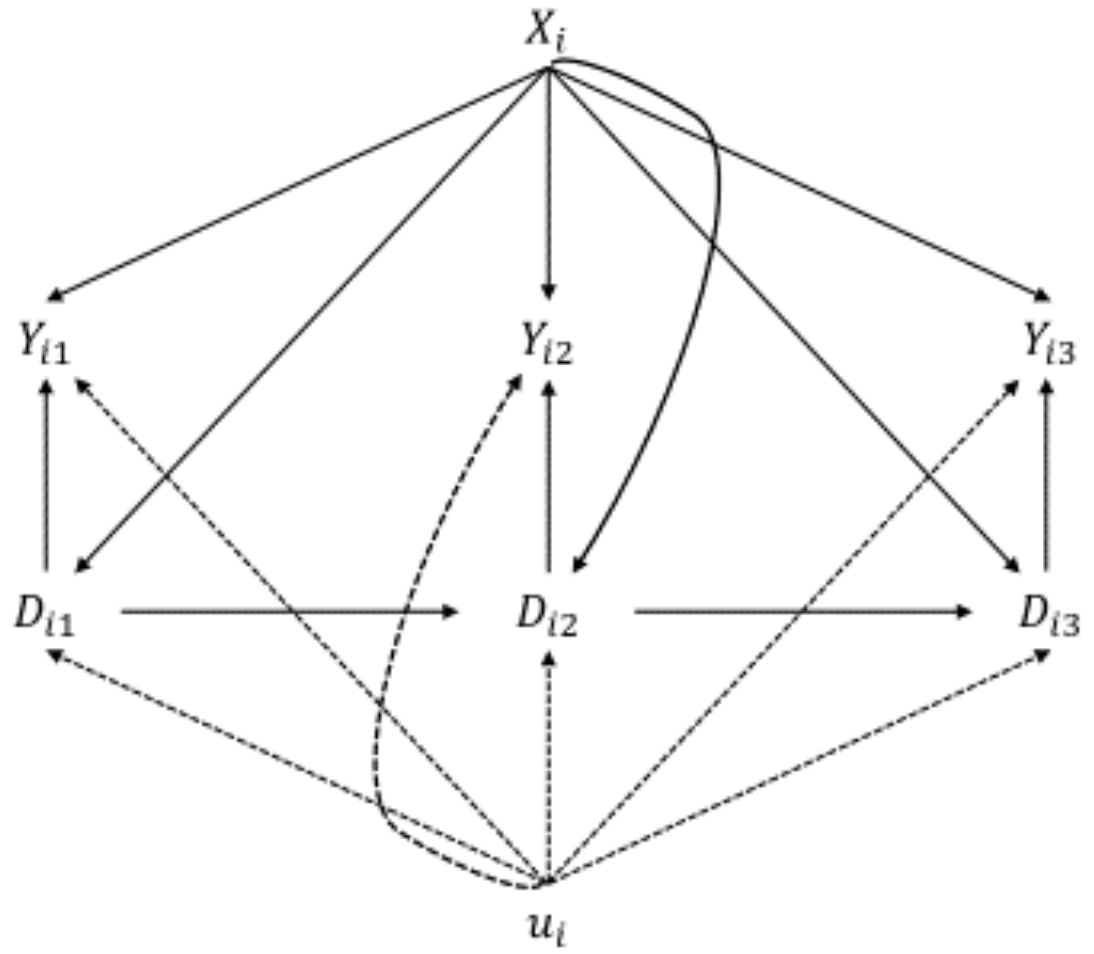
⑤過去の結果は現在の処置に直接的な影響を及ぼさない（Y_{t-1}はD_tに直接的な影響はない）



DAGによる整理

Y_{it} ：t期における、標本iの被説明変数
D_{it}：t期における、標本iの説明変数
X_i ：標本iに固有の効果（観測可能）
u_i ：標本iに固有の効果（観測不可能）

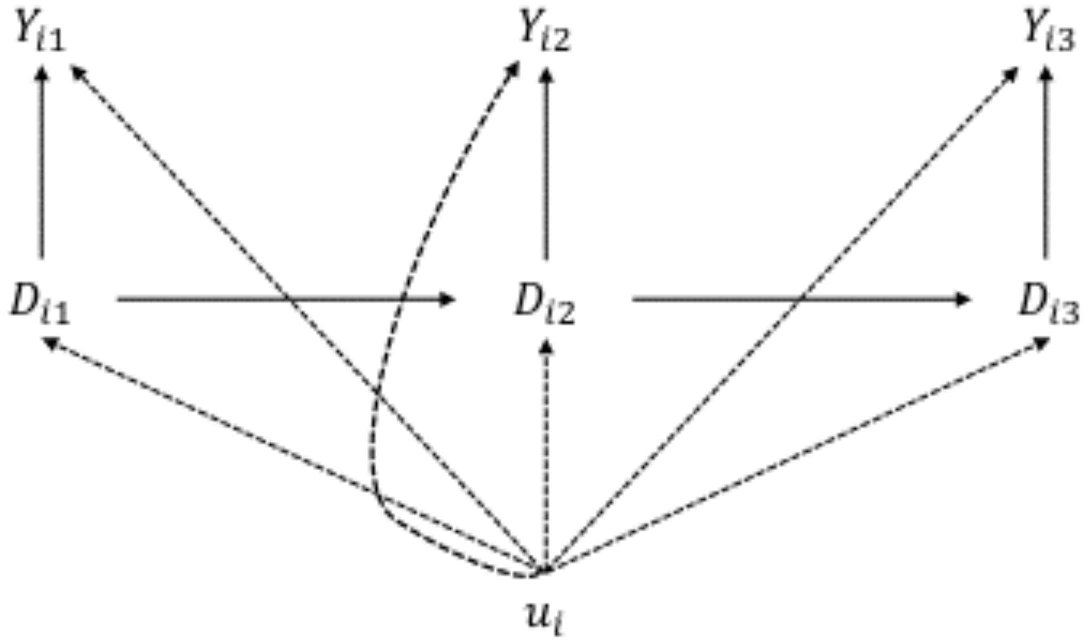
⑥過去の処置は現在の結果に直接的な影響を及ぼさない（ D_{t-1}はY_tに直接的な影響はない）



DAGによる整理

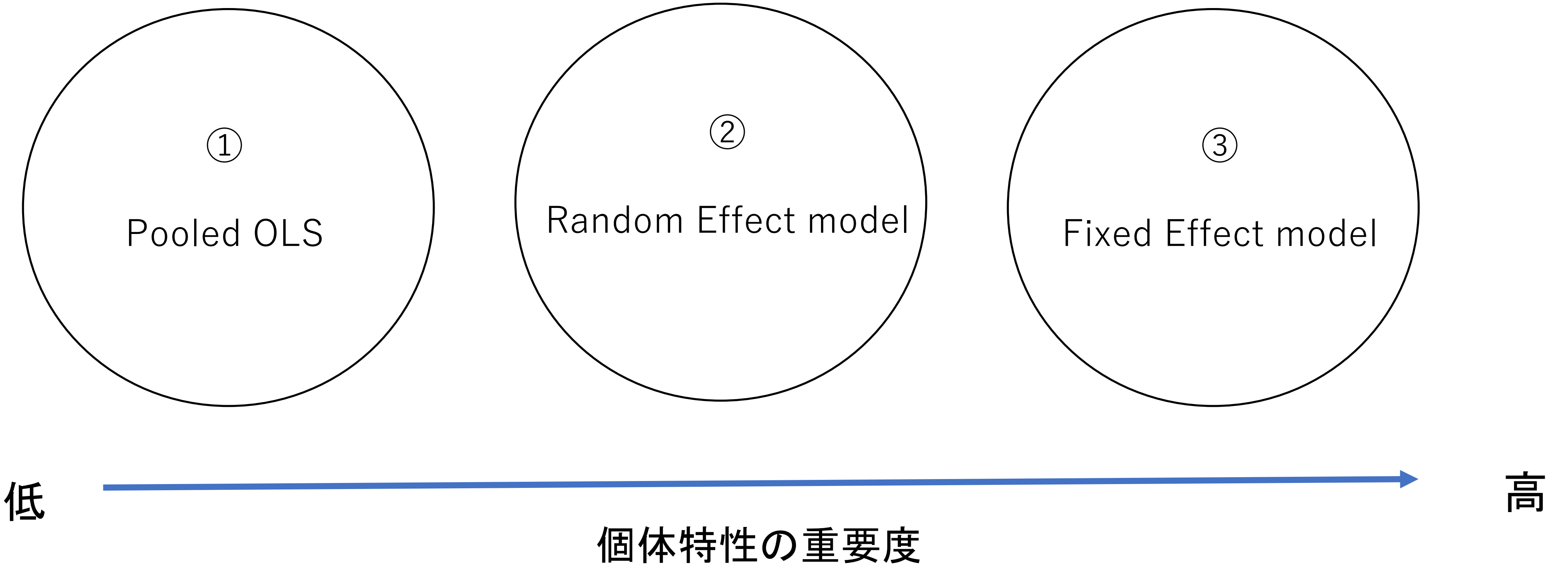
※Mixtape では、理論部分の説明においてX_iを除いて回帰分析を行っている為、ここでもそれに倣います。（X_iを含んでいてもやることと結果はほぼ同じですが、納得いかなければAppendixへ。）

※これ以降の計算はすべて行列計算ですが、
普通の一次式と見ていただいても特に問題ありません。数学苦手な方もご安心下さい。
（行列の場合の各項の姿を見たい方もAppendixへ。）



パネルデータを分析しよう

- パネルデータの分析方法は主に3つある



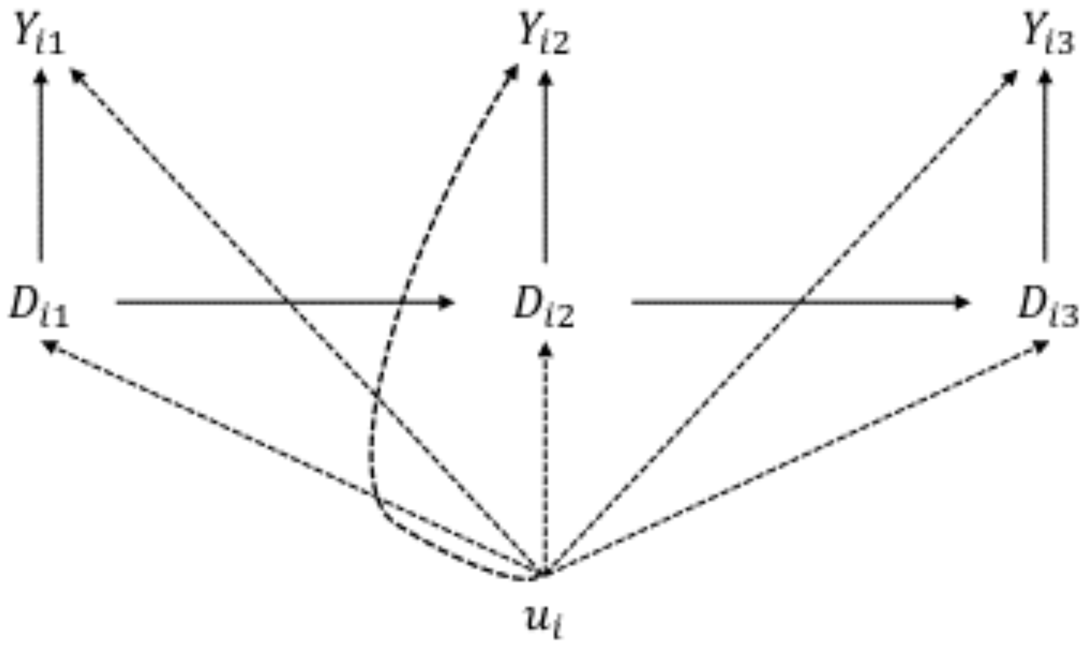
パネルデータを分析しよう

例：学校教育が収入に与える影響

$$Y_{it} = D_{it}\delta_i + u_i + \varepsilon_{it}$$

Y_{it} ：t期における、iの収入(log)
D_{it}：t期における、iの受ける学校教育
u_i ：iの知性(など、遺伝的要因)

イメージ：
$$\log(earning_{it}) = schooling_{it}\delta_i + intelligence_i + \varepsilon_{it}$$



用語：u_i …unobserved heterogeneity
ε_{it}…idiosyncratic error

個体固有効果 … individual specific effects

Panel Data Analysis 1: Pooled OLS model

Pooled OLS model

(Note)
Y_{it} ：$\log(earning_{it})$
D_{it} ：$schooling_{it}$
u_i ：$intelligence_{it}$

まずは素直にOLSで。

$$Y_{it} = D_{it}\delta_i + u_i + \varepsilon_{it}$$

u_iは非観測誤差（時間を通じて個人に一定の効果）

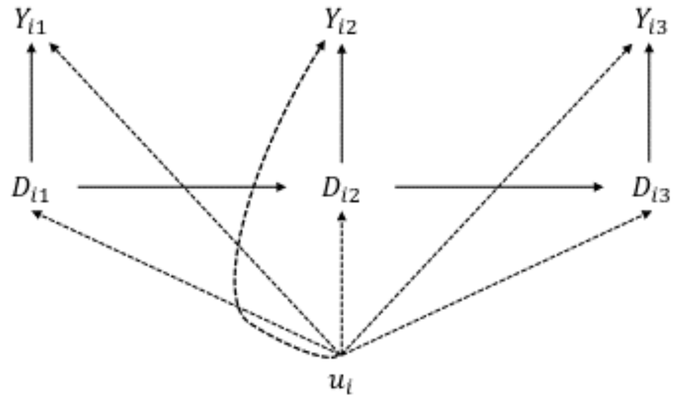
●集計最小二乗法(Pooled OLS)

・データをCross-sectional なものと仮定して分析

要は、**D_{it}とu_iが独立であることを仮定**して分析

（u_iがジャマなので、もともと無かったことに）

$$Y_{it} = D_{it}\delta_i + \eta_{it}$$



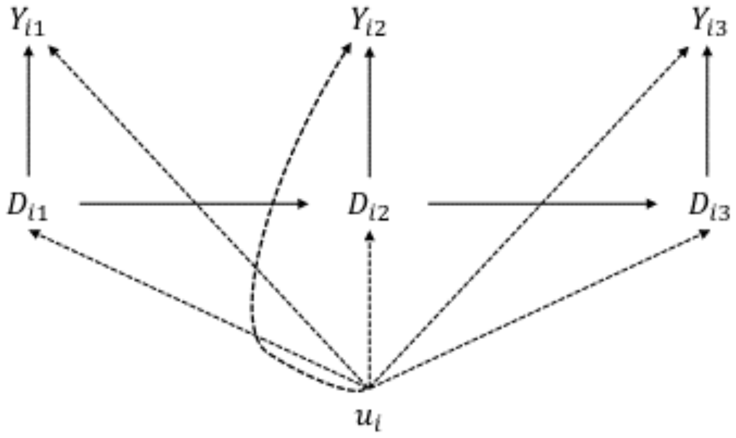
=> 係数の推定はできる！…けど、合ってる？

Pooled OLS model

欠落変数バイアスの問題

就学することD_{it}は時間経過で変化しない個体特性u_iの影響を受けている
→統制しないと欠落変数バイアス (omitted variable bias) が発生

しかし、
D_{it}とη_{it}に相関がないことを仮定
＝全ての時間軸において、u_iと就学することD_{it}に相関がないと仮定することになる。



Pooled OLS model

(Note)
Y_{it} ：$\log(earning_{it})$
D_{it} ：$schooling_{it}$
u_i ：$intelligence_{it}$

欠落変数バイアスの問題

$$Y_{it} = D_{it}\delta_i + \eta_{it} \quad \dots \quad (\eta_{it} = u_i + \varepsilon_{it})$$

・実際には、D_{it}とu_iは独立ではなかった

$$\widehat{\delta_i} = \delta_i + \frac{Covariance(D_{it}, u_i)}{Variance(D_{it})}, \qquad Covariance(D_{it}, u_i) \neq 0$$

↑ ココの分、欠落変数バイアスが生じる => ズレる…

Pooled OLS model

不均一分散に関して頑健な誤差の問題

●不均一分散に関して頑健な誤差（Heteroscedastic Robust Standard Errors）
残差の均一性の仮定に対して、頑健な標準誤差の推定を行う

頑健性のある標準誤差（Robust Standard Error）
＊頑健性のある・ロバストな・（robust）：必要としている条件または仮定を少々満たしていないようなデータにおいてもほぼ妥当な結果を与える

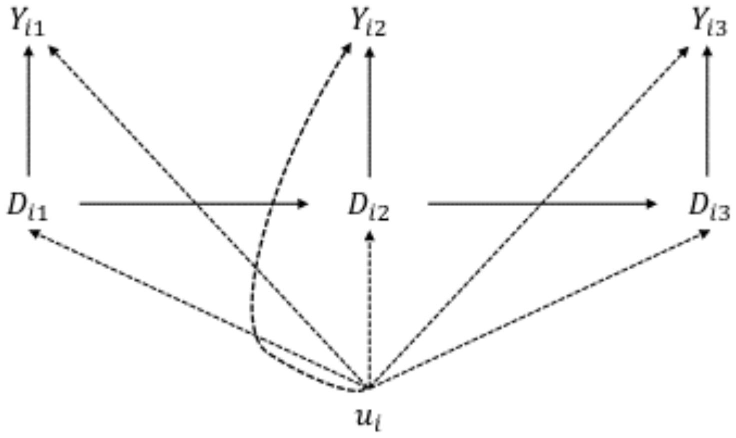
Pooled OLS model

不均一分散に関して頑健な誤差の問題

u_iは各t期それぞれに存在するため、 $\eta_{it} \equiv (c_i + \epsilon_{it})$と iとが系列相関を持っているといえる。

●系列相関（serial correlation):
回帰分析において、誤差項（残差、実測値と理論値の差）に自己相関があること

●自己相関(autocorrelation)：
ある変数の系列中において、個々の変数値が一定間隔を置いて相関を持つこと。元データと時間をずらしたデータとの相関。



Pooled OLS model

不均一分散に関して頑健な誤差の問題

$$Var(\widehat{\delta}_i) = \frac{\sum_{t=1}^{T}(D_{it} - \overline{D_i})^2 Var(\eta_{it}|D_{it})}{(\sum_{t=1}^{T}(D_{it} - \overline{D_i})^2)^2}$$

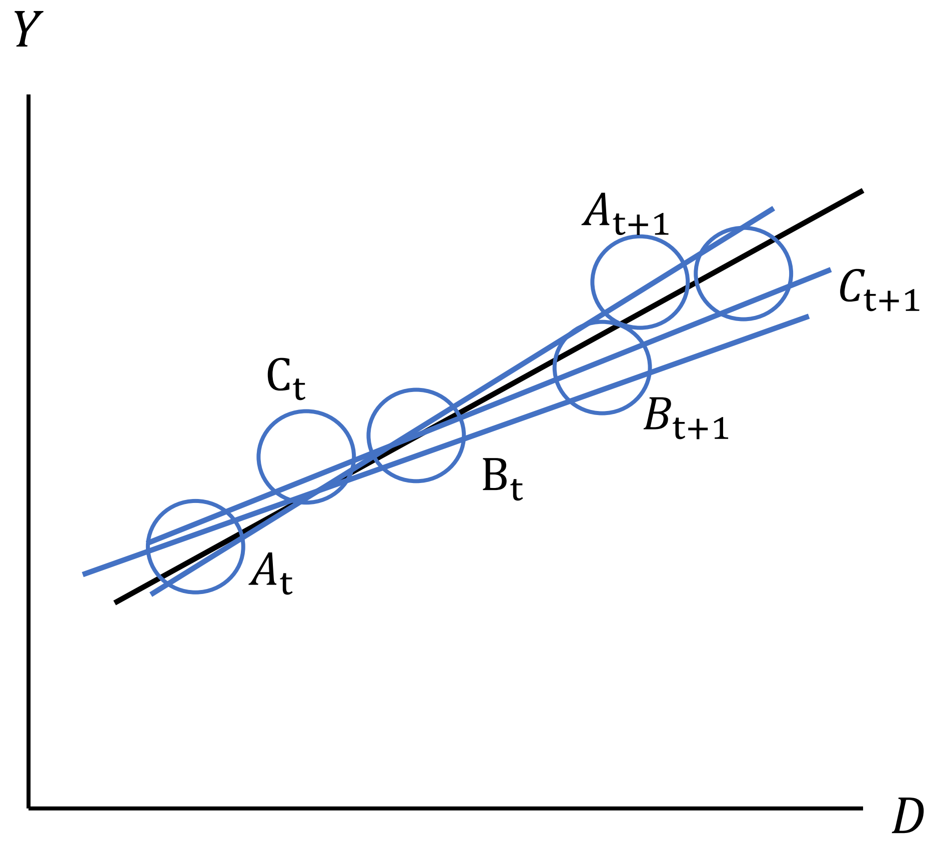
u_iの存在による、η_{it}の自己相関により、不均一分散に関して頑健な誤差が過少に推定されやすい

Pooled OLS model

ちなみに…

個体特性が無視できるほど小さければ
問題はない。

しかし、現実そのような場面はほとんどない。



Pooled OLS model

ほとんどの場合、そのまま素直にPooled OLSで分析すると、うまくいかなさそう。

・原因：u_iを無視したこと

だが、u_iが含まれたままでも、それはそれで問題

$\because$ u_iは観測できない

=>なら、「合法的に」u_iに退場してもらえば良い！　＝　Fixed Effects / Random Effects model の導入

Panel Data Analysis 2: Fixed Effects model

1. Fixed Effects model でできること

Fixed Effects model（FE model）

●固定効果モデル（Fixed Effects Model）

- unobserved individual effects(観測不可能な個体固有効果)を、
「時間平均をとって引く」ことで取り除き、推定を行う手法

→まずは、現在の状況を整理

<image>
Year: 1989 - 2003

状態

時間に関して一定の非観測誤差u_iあり（POLS耐性）

$$Y_{it} = D_{it}\delta_i + u_i + \varepsilon_{it}$$

$$(\log(earning_{it}) = schooling_{it}\delta_i + intelligence_i + \varepsilon_{it})$$

あ！やせいの
かいきしきが とびだしてきた！ ▼

わざ

▻ Pooled OLS ↘　▻ FE ↗　▻ RE

~~なげる~~
匙
~~タオル~~

Fixed Effects model（FE model）

(Note)
Y_{it} ：$\log(earning_{it})$
D_{it} ：$schooling_{it}$
u_i ：$intelligence_{it}$

$$Y_{it} = D_{it}\delta_i + u_i + \varepsilon_{it} \quad \ldots (1)$$

FE （固定効果法）で分析したい … 具体的にどうするの？

⇒ “time-demeaning” する！

(1)式においてu_iは（時間に関して一定であることから）「固定された効果(fixed effect)」と見なすことが出来る。

→ 両辺の時間平均をとると、$\sum_{t=1}^{T} u_i = Tu_i$より、$u_i$だけが同じ値のまま残る！

⇒ 〃 をそれぞれ辺々引けば、u_iだけが消えるハズ！

Fixed Effects model（FE model）

(Note)
Y_{it} ：$\log(earning_{it})$
D_{it} ：$schooling_{it}$
u_i ：$intelligence_{it}$

$$Y_{it} = D_{it}\delta_i + u_i + \varepsilon_{it} \quad \ldots(1)$$
$$\overline{Y_i} = \overline{D_i}\delta_i + \overline{u_i} + \overline{\varepsilon_i} \quad \ldots(2)$$
$$(\overline{X_i} = \frac{1}{T}\sum_{t=1}^{T} X_{it}, X = Y, D, u, \varepsilon)$$

(1) − (2)より、

$$Y_{it} - \overline{Y_i} = (D_{it} - \overline{D_i})\delta_i + \underbrace{(u_i - \overline{u_i})}_{=0} + (\varepsilon_{it} - \overline{\varepsilon_i})$$

$= 0 \quad \because u_i$：定数

$$\therefore \boxed{\ddot{Y}_{it} = \ddot{D}_{it}\delta_i + \ddot{\varepsilon}_{it}}$$
$$(\ddot{X}_{it} = X_{it} - \overline{X_i}, \qquad X = Y, D, \varepsilon)$$

Fixed Effects model（FE model）

(Note)

$$\bar{X}_i = \frac{1}{T}\sum_{t=1}^{T} X_{it}$$

＜一般形＞(Name : Fixed effects estimator / within estimator)

$$\ddot{Y}_{it} = \ddot{D}_{it}\delta_i + \ddot{\varepsilon}_{it}$$

$$(\ddot{X}_{it} = X_{it} - \bar{X}_i, \qquad X = Y, D, \varepsilon)$$

具体例：

$$\log(\ddot{earning}_{it}) = \ddot{Schooling}_{it} + \ddot{\varepsilon}_{it}$$

➡　u_iが消去されたので、欠落変数バイアスを回避して分析可能！
　$Corr(D_i, u_i) \neq 0$、つまり、D_iとu_iの間に相関があってもOK！

用語：「2FE model」
回帰式が次のような形をしている場合もあります。(γ_tという、時間固有効果が含まれています。)

$$Y_{it} = D_{it}\delta_i + u_i + \gamma_t + \varepsilon_{it}$$

このようなモデルのことを two way fixed effects model (2FE model) と呼びます。やることは同じです。

Fixed Effects model（FE model）

FE model は（パネルデータ分析）いつも使える？
→　以下の条件が必要！($\widehat{\delta_{FE}}$が「一致不偏推定量」になってくれる）

1）「強外生性」の仮定が成立

$$E[\varepsilon_{it}|D_i, u_i] = 0 \quad for^{\forall} t \in \{1,2,\dots,T\} \qquad (D_i \equiv (D_{i1}, D_{i2}, \dots, D_{iT}))$$

2）　D_{it}は時間に関して一定の項ではない　<=　当たり前なので、気にする必要は特になし。

$$rank\left(\sum_{t=1}^{T} E\left[\ddot{D_{it}}' \ddot{D_{it}}\right]\right) = K \quad \text{(full rank)}$$

（一応、数式的にはD_{it}が正則行列であることと同値。）

用語：「一致推定量」、「不偏推定量」（復習）

一致推定量：　$\operatorname*{plim}_{N\to\infty} \hat{\delta} = \delta$

不偏推定量：　$E\left[\widehat{\sigma^2}\right] = \sigma^2,\ \ \sigma^2 = Var(\varepsilon|D)$

FE model – weak point

FE model を使わない方が良い（もしくは使えない）のはいつか？

1. Simultaneity (or reverse causality) や 「時間経過で変化する individual specific effects」には対処がきかない

2. 非観測誤差(heterogeneity)u_iが「時間に関して一定ではない」時はバイアスが生じる

3. どうしても係数を推定できない場合がある

FE model – weak point 1

弱点1)Simultaneity (or reverse causality) や 「時間経過で変化するindividual specific effects」には対処がきかない

e.g.,) Cornwell & Trumbull (1994) “Estimating the Economic Model of Crime with Panel Data.”

・North Carolina州の郡ごとのパネルデータを用いて、人口当たりの警察組織の規模や犯罪の検挙率といった要因が、犯罪の発生率に与える影響を分析した研究。

$$R_{it} = X_{it}\beta + P_{it}\gamma + \alpha_i + \varepsilon_{it}$$

R_{it}：i郡におけるt期のcrime rate
X_{it}：i郡におけるt期のrelative return to legal activities & P_{it}以外でR_{it}に関係しそうな要因(observable)
P_{it}：i郡におけるt期の検挙や有罪、投獄確率や、刑罰の重さなど(observable)
α_i：i郡における郡固有効果(unobservable)、時間に関して一定

FE model – weak point 1

弱点1)

$$R_{it} = X_{it}\beta + P_{it}\gamma + \alpha_i + \varepsilon_{it} \quad \dots (3)$$

(3)を、「個体効果の影響が無い(cross-sectional / Pooled OLS)」場合、「個体効果がある(FE model)」場合,の指標にした上で、回帰分析を行った。

cross-sectional な場合の指標：

$$\overline{R_i} = \overline{X_i}\beta + \overline{P_i}\gamma + \alpha_i + \overline{\varepsilon_i} \quad \dots \quad (between\ estimator)$$

Fixed Effects estimator

$$\ddot{R_{it}} = \ddot{X_{it}}\beta + \ddot{P_{it}}\gamma + \ddot{\varepsilon_{it}} \quad \dots \quad (within\ estimator)$$

$$\left(\because \overline{R_i} = \frac{1}{T}\sum\nolimits_t R_{it}, \qquad \ddot{R_{it}} = R_{it} - \overline{R_i}\right)$$

(Note)
R_{it}：i郡におけるt期のcrime rate
X_{it}：i郡におけるt期のrelative return to legal activities & P_{it}以外でR_{it}に関係しそうな要因(observable)
P_{it}：i郡におけるt期の検挙や有罪、投獄確率や、刑罰の重さなど (observable)
α_i：i郡における郡固有効果(unobservable)、時間に関して一定

FE model – weak point 1

弱点1)

$$R_{it} = X_{it}\beta + P_{it}\gamma + \alpha_i + \varepsilon_{it} \quad \ldots (3)$$

以下、説明の簡単のために、(3)式を次のように限定する

$$R_{it} = Police_{it}\beta + P_{A_{it}}\gamma + \alpha_i + \varepsilon_{it}$$

$Police_{it}$ ： i郡におけるt期の警察の規模
$P_{A_{it}}$ ： i郡におけるt期の犯罪の検挙率
(possibility of arrest)

(Note)
R_{it}：i郡におけるt期のcrime rate
X_{it}：i郡におけるt期のrelative return to legal activities & P_{it}以外でR_{it}に関係しそうな要因 (observable)
P_{it}：i郡におけるt期の検挙や有罪、投獄確率や、刑罰の重さなど (observable)
α_i：i郡における郡固有効果(unobservable)、時間に関して一定

FE model – weak point 1

弱点1）
各指標は次のように変化します。

Pooled OLS:

$$\overline{R_i} = \overline{Police_i}\beta + \overline{P_{A_i}}\gamma + \alpha_i + \overline{\varepsilon_i} \quad \ldots \quad (between\ estimator)$$

Fixed Effects:

$$\ddot{R_{it}} = Pol\ddot{i}ce_{it}\beta + \ddot{P_{A_{it}}}\gamma + \ddot{\varepsilon_{it}} \quad \ldots \quad (within\ estimator)$$

今回は、$H_0: E[P|X] = 0$が検定により棄却されるため、本分析においては「固定効果の影響あり」とするのが妥当とされている。（検定については後程説明。）

(Note)
R_{it}：i郡におけるt期のcrime rate
$Police_{it}$： i郡におけるt期の警察の規模
$P_{A_{it}}$　： i郡におけるt期の犯罪の検挙率
α_i：i郡における郡固有効果(unobservable)、時間に関して一定
$\overline{R_i} = \frac{1}{T}\sum_t R_{it}$, $\ddot{R_{it}} = R_{it} - \overline{R_i}$

FE model – weak point 1

弱点1）

$$\overline{R_i} = \overline{Police_i}\beta + \overline{P_{A_i}}\gamma + \alpha_i + \overline{\varepsilon_i} \quad \ldots \quad (between\ estimator)$$

$$\ddot{R}_{it} = Pol\ddot{i}ce_{it}\beta + \ddot{P}_{A_{it}}\gamma + \ddot{\varepsilon}_{it} \quad \ldots \quad (within\ estimator)$$

ここで、$R, Police, P_A$の関係に注目。（次頁）

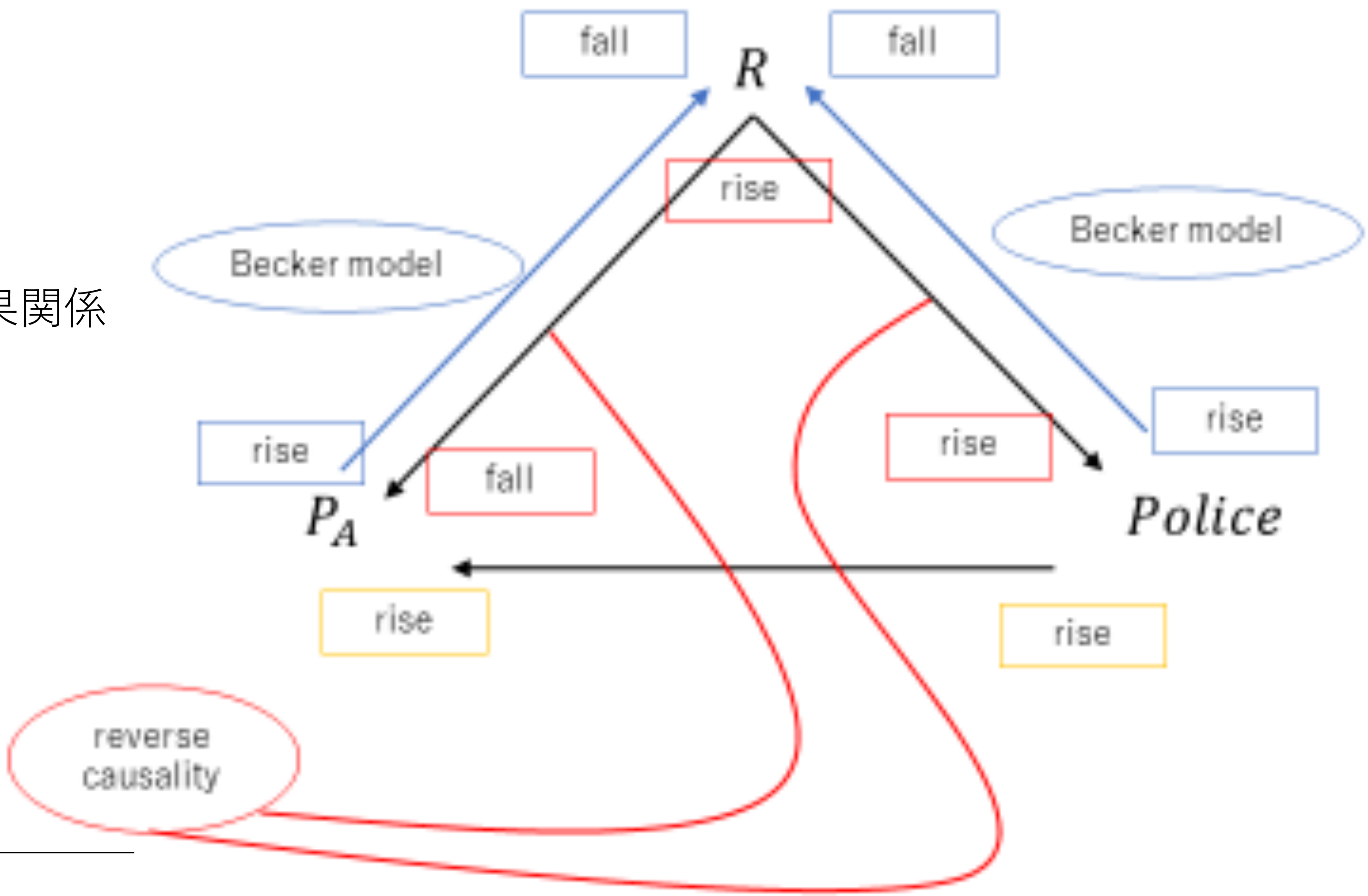
(Note)
R_{it}：i郡におけるt期のcrime rate
$Police_{it}$： i郡におけるt期の警察の規模
$P_{A_{it}}$ ： i郡におけるt期の犯罪の検挙率
α_i：i郡における郡固有効果(unobservable)、時間に関して一定
$\overline{R_i} = \frac{1}{T}\sum_t R_{it}$, $\ddot{R}_{it} = R_{it} - \overline{R_i}$

Simultaneity

(Note)
R ：crime rate
$Police$： 警察の規模
P_A ： 犯罪の検挙率

青矢印：既存の理論で説明されている因果関係
黒　〃　：実際に観測されている因果関係

・Simultaneity の問題が発生



用語：「Simultaneity」と「Reverse causality」
- Simultaneity(同時性)　$X \leftrightarrow Y$
- Reverse Causality(逆因果)　$Y \rightarrow X$　$(but, expected\ as\ X \rightarrow Y)$

FE model – weak point 1

弱点1)

$$\overline{R_i} = \overline{Police_i}\beta + \overline{P_{A_i}}\gamma + \alpha_i + \overline{\varepsilon_i} \quad \ldots \quad (between\ estimator)$$
$$\ddot{R_{it}} = \ddot{Police_{it}}\beta + \ddot{P_{A_{it}}}\gamma + \ddot{\varepsilon_{it}} \quad \ldots \quad (within\ estimator)$$

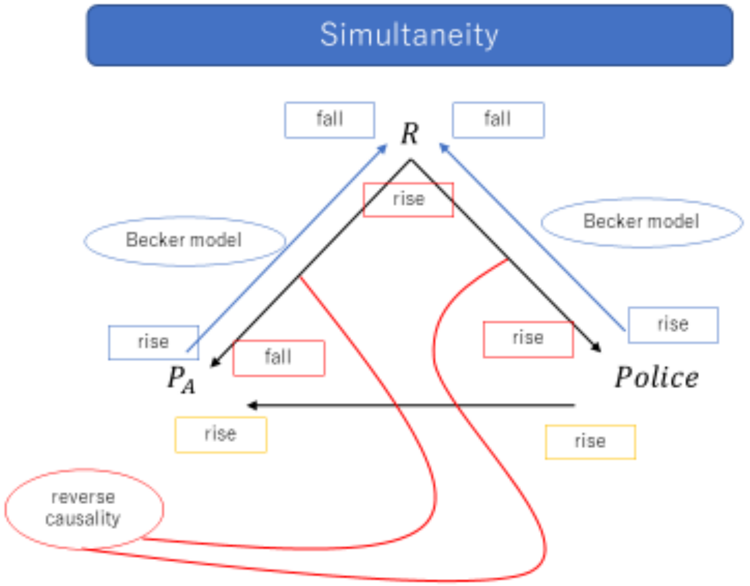
ここで、$R, Police, P_A$の関係に注目。

=>Simultaneity の問題が発生

=>内生性の問題($E[\varepsilon|Police, P_A] \neq 0$)

(↑Wu - Hausman 検定でも、$H_0: E \neq 0$が棄却できなかった)

=>2SLS(二段階最小二乗法)でも分析



用語：「Simultaneity」と「Reverse causality」

Simultaneity(同時性)　$X \leftrightarrow Y$

Reverse Causality(逆因果)　$Y \rightarrow X \quad (but, expected\ as\ X \rightarrow Y)$

(Note)

R_{it}：i郡におけるt期のcrime rate

$Police_{it}$：i郡におけるt期の警察の規模

P_{Ait}：i郡におけるt期の犯罪の検挙率

α_i：i郡における郡固有効果(unobservable)、時間に関して一定

$\overline{R_i} = \frac{1}{T}\sum_t R_{it}$, $\ddot{R_{it}} = R_{it} - \overline{R_i}$

FE model – weak point 1

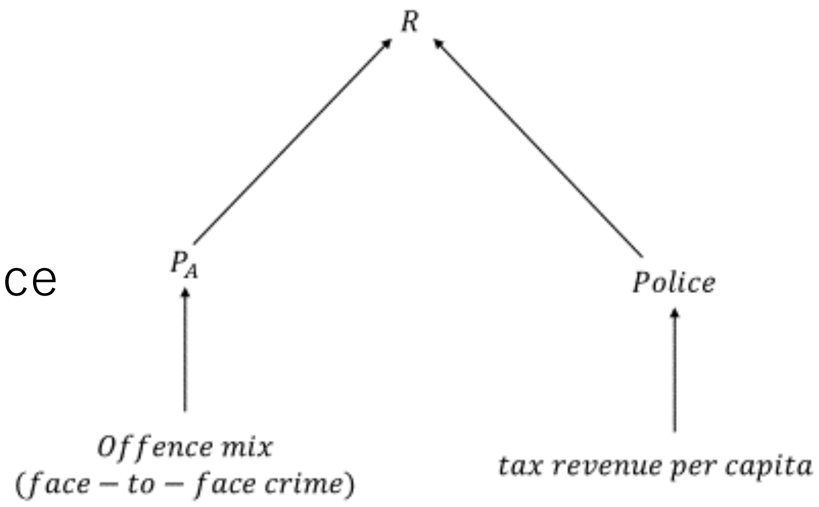
弱点1)

$$\overline{R_i} = \overline{Police_i}\beta + \overline{P_{A_i}}\gamma + \alpha_i + \overline{\varepsilon_i} \quad \ldots \quad (between\ estimator)$$

$$\ddot{R}_{it} = \ddot{Police}_{it}\beta + \ddot{P}_{A_{it}}\gamma + \ddot{\varepsilon}_{it} \quad \ldots \quad (within\ estimator)$$

（2SLS）

操作変数IVとして、それぞれ「Offence mix(強盗など、face-to-face で起こった犯罪)」、
「tax revenue per capita」を利用。



=>普通の固定効果モデルなどでは解決出来ていなかった、同時性の問題の解決を図る

(Note)
R_{it}：i郡におけるt期のcrime rate
$Police_{it}$： i郡におけるt期の警察の規模
$P_{A_{it}}$　： i郡におけるt期の犯罪の検挙率
α_i：i郡における郡固有効果(unobservable)、時間に関して一定
$\overline{R_i} = \frac{1}{T}\sum_t R_{it}$， $\ddot{R}_{it} = R_{it} - \overline{R_i}$

FE model – weak point 1

“no longer statistically significant”

TABLE 3.—RESULTS FROM ESTIMATION
(standard errors in parentheses)

	Between	Within	2SLS (fixed effects)	2SLS (no fixed eff[illegible])
CONSTANT	−2.097			[illegible].719
	(2.822)			(8.189)
P_A	−0.648	−0.355	−0.455	−0.507
	(0.088)	(0.032)	(0.618)	(0.[illegible])
P_C	−0.528	−0.282	−0.336	−[illegible]30
	(0.067)	(0.021)	(0.371)	([illegible].110)
P_P	0.297	−0.173	−0.196	0.200
	(0.231)	(0.032)	(0.200)	(0.343)
S	−0.236	−0.00245	−0.0298	−0.218
	(0.174)	(0.02612)	(0.0300)	(0.185)
POLICE	0.364	0.413	0.504	0.419
	(0.060)	(0.027)	(0.617)	(0.218)
DENSITY	0.168	0.414	0.291	0.226
	(0.077)	(0.283)	(0.785)	(0.103)

The scale of police is positively correlated to crime rate …!?
(because of simultaneity)

Table :
Cornwell, Christopher, and William N. Trumbull. 1994. “Estimating the Economic Model of Crime with Panel Data.” *Review of Economics and Statistics* 76 (2): 360–66, Retrieved from: https://www.amherst.edu/media/view/121570/original/CornwellTrumbullCrime%2BElasticities.pdf

⇒　逆因果関係や同時性がある時には、Fixed Effects modelは上手く適用し難い

FE model – weak point 2

弱点2)非観測誤差(heterogeneity)u_iが「時間に関して一定ではない」時はバイアスが生じる

$$Y_{it} = D_{it}\delta_i + \boxed{u_{it}} + \varepsilon_{it} \quad \ldots(1)$$
$$\overline{Y_i} = \overline{D_i}\delta_i + \overline{u_i} + \overline{\varepsilon_i} \quad \ldots(2)$$

$(1)-(2)$より、

$$Y_{it} - \overline{Y_i} = (D_{it} - \overline{D_i})\delta_i + \underline{(u_{it} - \overline{u_i})} + (\varepsilon_{it} - \overline{\varepsilon_i})$$
$$\therefore \ddot{Y_{it}} = \ddot{D_{it}}\delta_i + \boxed{\ddot{u_{it}}} + \ddot{\varepsilon_{it}}$$

u_{it}が観測できない以上、$\ddot{u_{it}}$も分からない。→　仮に$\ddot{\eta_{it}} = \ddot{u_{it}} + \ddot{\varepsilon_{it}}$とまとめて回帰をしようものなら、欠落変数バイアス

$$\frac{Covariance(\ddot{D_{it}}, \ddot{u_{it}})}{Var(\ddot{D_{it}})}$$

が生じてしまう

FE model – weak point 3

弱点3)どうしても係数を推定できない場合がある（参考）

・以下のようなタイプの項に関しては、FE modelでは係数推定が出来ない

・時間変化をしない項

e.g.,) u_i, 定数項, 属性ダミー(i.e., 人種 $race$, etc.)

属性ダミーについては、時点時点ダミー(i.e., 1994年ダミー)との交差項をとることで係数の推定は可能になる。ただし、解釈の仕方には注意。

・時間変化の幅が一定な項

e.g.,) 年齢, 経験年数（1年経過毎に、「すべての標本に対して」+1ずつ増加）

すべての年度の時間ダミーを含むようなモデルを考えるときには、このタイプの項は推定が出来なくなる。

Panel Data Analysis 3: Random Effects model

Random Effects Model（RE model）(参考)

●変量効果モデル（**Random Effect Model**）

- 観測されない個体固有効果を独立変数とは独立としてランダムに発生しているものとみなして推計を行う。

→ 個体特性を誤差項$\boldsymbol{u_i}$として扱う

Fixed Effects model では扱えなかった部分（定数項、時間に関して一定な変数の係数）を推定することが出来る

Random Effects model（RE model）（参考）

$$Y_{it} = D_{it}\delta_i + u_i + \varepsilon_{it}$$

今度はu_iを「確率変数」と見なして、定式化を図る。

→　すべてのu_iに共通の平均値μ_iを中心に分散$\sigma_{u_i}^2$で分布する確率変数、と見る。

＜定式化＞（詳細は割愛。）

$$Y_{it} = D_{it}\delta_i + (\mu_i + \alpha) + \varepsilon_{it}$$

$$\therefore Y_{it} = D_{it}\delta_i + \alpha + (\mu_i + \varepsilon_{it})$$

ここから変数変換を施して、最終的に以下を得る。

$$Y_{it}^* = D_{it}^*\delta_i + (1-\theta) + \varepsilon_{it}$$

$$\left(Y_{it}^* = Y_{it} - \theta\bar{Y}_i, X_{it}^* = X_{it} - \theta\bar{X}_i, \qquad \theta = 1 - \sqrt{\frac{\sigma_\varepsilon^2}{T\sigma_u^2 + \sigma_\varepsilon^2}} \right)$$

用語：

合成誤差(a composite error term）：$\mu_i + \varepsilon_{it}$のこと

Generalized Least Square (GLS, 一般化最小二乗法）：ガウス・マルコフの仮定を満たすように変数変換を施してからOLSを適用する手法のこと。今回はθの推定が必要なのだが、このような推定の過程を含むGLSをFeasible GLS (FGLS)と呼ぶ。

Random Effects model (RE model) (参考)

RE model は パネルデータ分析ならいつでも使えるのか？
→　こちらも以下の条件が必要

(i)

$$Corr(D_{it}, u_i) = 0$$

（要は、D_i, u_iの間には相関が無い、ということ。）

(ii)強外生性

$$E[\varepsilon_{it} | D_i, u_i] = 0$$

（RE推定量が一致不偏推定量であるために必要。）

Panel Data Analysis: which one to use?

分析手法の使い分け　（参考）

$$Y_{it} = D_{it}\delta_i + u_i + \varepsilon_{it}$$

これまでに見てきた手法：Pooled OLS / Fixed effects model / Random effects model

・もし個体特有の効果（ u_i ）が無視できる程度（ $u_i \approx 0$）のモノだとしたら、Pooled OLSも使える。

… では、どれを使えばいいのか？

→検定の利用

※あくまで参考指標です。検定でこうなったから絶対これしか使わない、ではなく、検定でこうなったからコレが有効そうかな、くらいの捉え方が推奨されます。

<image>
Year: 1994 - 2022

状態
？？？

$$Y_{it} = D_{it}\delta_i + u_i + \varepsilon_{it}$$

あ！やせいの
かいきしきが とびだしてきた！ ▼

わざ
▻ Pooled OLS ▻ FE ▻ RE

どうぐ
▻ F検定(Breusch-Pegan 検定)
▻ Wu-Hausman 検定

1) そもそもPooled OLSをつかって良さそうか？

状態

$u_i = 0 \ or \neq 0$?

$$Y_{it} = D_{it}\delta_i + u_i + \varepsilon_{it}$$

せつめい：
個体個別効果u_iの有無を確認できる。
帰無仮説H_0：「個別効果なし$u_i \approx 0$」が棄却されればFE, REが、反対にH_0が棄却されなければPooled OLSも有効になる。

H_0棄却不可のとき

H_0棄却時

わざ

▻ Pooled OLS

▻ FE　▻ RE

どうぐ

▻ F検定(Breusch-Pegan 検定)

▻ Wu-Hausman 検定

2) FE or RE?

状態

$$u_i \neq 0 \quad , \quad Corr(D_i, u_i) = 0 \;\; or \;\; \neq 0$$

$$Y_{it} = D_{it}\delta_i + u_i + \varepsilon_{it}$$

H_0棄却時

H_0棄却不可のとき

せつめい：
D_iとu_iの相関の有無を確認できる。
帰無仮説H_0：「$Corr(D_i, u_i) = 0$」
が棄却されればFEが、反対にH_0が棄却されなければREも有効になる。

わざ

▻ ~~Pooled OLS~~ ▻ FE ▻ RE

どうぐ

▻ F検定(Breusch-Pegan 検定)
▻ Wu-Hausman 検定

Panel Data Analysis -in practice-

Panel data analysis _ example

- Cornwell & Rupert (1997) “Unobservable Individual Effects, Marriage and the Earnings of Young Men”

・男性に関して、婚姻が賃金に与える影響を分析。「結婚が生産性（賃金）を向上させる」は本当か？

$$y_{it} = \delta + X_{it}\beta + M_{it}\gamma + \alpha_i + \varepsilon_{it}$$
$$(\log(wage_{it}) = \delta + experience_{it}\beta_1 + \cdots + education_{it}\beta_k + Married_{it}\gamma_1 + \cdots + \alpha_i + \varepsilon_{it})$$

y_{it}：t期におけるiの賃金（log）
X_{it}： t期におけるiの諸々の説明変数
M_{it}： t期におけるiの婚姻状況（ダミー変数）
α_i： t期におけるiに固有な効果（unobserved）

(Note)
y_{it}：t期におけるiの賃金（log）
X_{it}： t期におけるiの諸々の説明変数
M_{it}： t期におけるiの婚姻状況（ダミー変数）
α_i： t期におけるiに固有な効果（unobserved）

・分析手法の選択

$$y_{it} = \delta + X_{it}\beta + M_{it}\gamma + \alpha_i + \varepsilon_{it}$$

RE model (FGLS) vs. FE model (Within)

→ Wu-Hausman検定
$Corr(X_i, \alpha_i) = 0 \;\; or \;\; Corr(X_i, \alpha_i) \neq 0$?

今回は$\chi^2 = 111.9$となっており、帰無仮説
H_0: $Corr(X_i, \alpha_i) = 0$ が棄却される。

→ FE model の方が適切と判断

Table :
Cornwell & Rupert.(1997). “Unobserved individual effects, marriage and the earnings of young men”. Economic inquiry. 1997, vol. 35, no. 2, p. 285–294.

TABLE II
Estimated Wage Regressions
(Standard Errors in Parentheses)

Variable	(1) FGLS	(2) Within	(3) Within	(4) Within
Married	0.083	0.056	0.051	0.033
	(0.022)	(0.026)	(0.026)	(0.028)
Divorced	0.064	0.062	0.057	0.040
	(0.033)	(0.036)	(0.036)	(0.038)
Years Married				–0.005
				(0.006)
Years Married^2				–0.0003
				(0.0003)
Years Divorced				–0.014
				(0.008)
Experience	0.027	0.027	0.024	0.021
	(0.004)	(0.004)	(0.004)	(0.005)
Experience^2	–0.001	–0.001	–0.001	–0.001
	(0.0001)	(0.0002)	(0.0002)	(0.0002)
Tenure			0.013	0.011
			(0.004)	(0.004)
Tenure^2			–0.0006	–0.0005
			(0.0002)	(0.0002)
South	–0.091	–0.121	–0.117	–0.118
	(0.019)	(0.034)	(0.034)	(0.034)
Urban	0.137	0.057	0.059	0.059
	(0.017)	(0.024)	(0.024)	(0.024)
Union	0.109	0.106	0.102	0.103
	(0.015)	(0.018)	(0.018)	(0.018)
Dependents	0.052	0.052	0.048	0.047
	(0.017)	(0.019)	(0.019)	(0.020)
No High School	–0.325			
	(0.057)			
Some High School	–0.148			
	(0.032)			
Some College	0.091			
	(0.028)			
College Grad	0.278			
	(0.034)			
Post-College	0.322			
	(0.041)			
Standard error	0.215	0.212	0.212	0.211
χ^2_{27}	111.9			

Panel data analysis _ example

y_{it}：t期におけるiの賃金（log）
X_{it}： t期におけるiの諸々の説明変数
M_{it}： t期におけるiの婚姻状況（ダミー変数）
α_i： t期におけるiに固有な効果（unobserved）

・選択した手法で分析

$$y_{it} = \delta + X_{it}\beta + M_{it}\gamma + \alpha_i + \varepsilon_{it}$$

$$(\log(wage_{it}) = \boxed{\delta} + experience_{it}\beta_1 + \cdots + \boxed{education_{it}}\beta_k + Married_{it}\gamma_1 + \cdots + \boxed{\alpha_i} + \varepsilon_{it})$$

➡ Fixed Effects model：

$$\ddot{y}_{it} = \ddot{X}_{it}\beta^* + \ddot{M}_{it}\gamma + \ddot{\varepsilon}_{it} \quad (within\ estimator)$$

$$\left(\log(\ddot{wage}_{it}) = \ddot{experience}_{it}\beta_1 + \cdots + \ddot{Married}_{it}\gamma_1 + \cdots + \ddot{\varepsilon}_{it}\right)$$

※時間に関して一定であった項は within estimator ではすべて消えることに注意！
e.g.,) δ, $education$, α

著者はFixed Effects model での分析を
1) そのまま　2) 変数$Tenure, Tenure^2$を追加　3) 変数$YearsMarried, YearsMarried^2, YearsDivorced$を追加
の3通り行っている。

<分析方法>
(1)RE model での分析
(2)FE model での分析その１
(3)FE model での分析（Tenureを追加）
(4)FE model での分析（Year Marriedなど追加）

＜RQ＞
結婚は、男性の生産性（賃金）を本当に向上させるのか。

Cornwell & Rupert.(1997). “Unobserved individual effects, marriage and the earnings of young men”. Economic inquiry. 1997, vol. 35, no. 2, p. 285–294.

TABLE II
Estimated Wage Regressions
(Standard Errors in Parentheses)

Variable	(1) FGLS	(2) Within	(3) Within	(4) Within
Married	0.083 (0.022)	0.056 (0.026)	0.051 (0.026)	0.033 (0.028)
Divorced	0.064 (0.033)	0.062 (0.036)	0.057 (0.036)	0.040 (0.038)
Years Married				−0.005 (0.006)
Years Married^2				−0.0003 (0.0003)
Years Divorced				−0.014 (0.008)
Experience	0.027 (0.004)	0.027 (0.004)	0.024 (0.004)	0.021 (0.005)
Experience^2	−0.001 (0.0001)	−0.001 (0.0002)	−0.001 (0.0002)	−0.001 (0.0002)
Tenure			0.013 (0.004)	0.011 (0.004)
Tenure^2			−0.0006 (0.0002)	−0.0005 (0.0002)
South	−0.091 (0.019)	−0.121 (0.034)	−0.117 (0.034)	−0.118 (0.034)
Urban	0.137 (0.017)	0.057 (0.024)	0.059 (0.024)	0.059 (0.024)
Union	0.109 (0.015)	0.106 (0.018)	0.102 (0.018)	0.103 (0.018)
Dependents	0.052 (0.017)	0.052 (0.019)	0.048 (0.019)	0.047 (0.020)
No High School	−0.325 (0.057)			
Some High School	−0.148 (0.032)			
Some College	0.091 (0.028)			
College Grad	0.278 (0.034)			
Post-College	0.322 (0.041)			
Standard error	0.215	0.212	0.212	0.211
χ^2_{27}	111.9			

Panel data analysis _ example

結果：

$Married$の係数　…　FGLS(1) >> FE(2) > FE(3) > FE(4)

結論：

△　結婚が生産性（賃金）を向上させる

○　賃金上昇は、婚姻関係と賃金、双方に相関のある、観測されない個体固有効果の影響

この時点で、5%水準で、統計的に有意ではなくなっている

Variable	(1) FGLS		(2) Within		(3) Within		(4) Within
Married	0.083 (0.022)	>>	0.056 (0.026)	>	0.051 (0.026)	>	0.033 (0.028)

Cornwell & Rupert.(1997). "Unobserved individual effects, marriage and the earnings of young men". Economic inquiry. 1997, vol. 35, no. 2, p. 285–294.

Appendix

Appendix 1: 回帰式の行列形と、$\boldsymbol{X_i}$の影響

$$Y_{it} = D_{it}\delta_i + X_i\gamma_i + u_i + \varepsilon_{it}$$

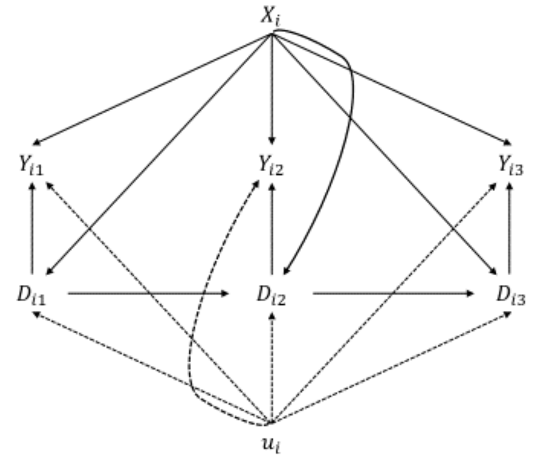
$$Y_i = \begin{pmatrix} Y_{i1} \\ \vdots \\ Y_{it} \\ \vdots \\ Y_{iT} \end{pmatrix}_{T\times 1},\ D_{it} = (D_{it,1} \quad \cdots \quad D_{it,j} \quad \cdots \quad D_{it,K})_{1\times K},\ D_i = \begin{pmatrix} D_{i1} \\ \vdots \\ D_{it} \\ \vdots \\ D_{iT} \end{pmatrix}_{T\times K},\ \delta_i = \begin{pmatrix} \delta_{i1} \\ \vdots \\ \delta_{ij} \\ \vdots \\ \delta_{iK} \end{pmatrix}_{K\times 1},\ \varepsilon_i = \begin{pmatrix} \varepsilon_{i1} \\ \vdots \\ \varepsilon_{it} \\ \vdots \\ \varepsilon_{iT} \end{pmatrix}_{T\times 1},$$

X_i, γ_i, u_iは行列ではない普通の項

cross-sectional な見方でこれを分析すると、

$$Y_{it} = D_{it}\delta_{it} + \eta_{it}, \qquad \eta_{it} = X_i\gamma_i + u_i + \varepsilon_{it}$$

$$\widehat{\delta_{it}} = \delta_{it} + \frac{Covariance(D_{it}, X_i\gamma_i + u_i)}{Var(D_{it})}$$



DAGより、 $Covariance(D_{it}, X_i\gamma_i + u_i) \neq 0$ なので、欠落変数バイアスはかかります。

Appendix 1: 回帰式の行列形と$\boldsymbol{X_i}$の影響(続)

Fixed Effects model を適用すると、t に依らず一定の項 $X_i\gamma_i + u_i$ は相変わらず消えるので、

$$\ddot{Y}_{it} = \ddot{D}_{it}\delta_i + \ddot{\varepsilon}_{it}$$

後の議論は同じものが適用できます。

Appendix 2: FE推定量=一致不偏推定量

Condition

1）個体固有誤差(idiosyncratic error) ε_{it} と独立変数 D_i, 非観測誤差 u_i は独立
「強外生性(strict exogeneity)」

$$E[\varepsilon_{it}|D_i, u_i] = 0 \quad for\ t = 1,2,\dots,T \quad D_i \equiv (D_{i1}, D_{i2}, \dots, D_{iT})$$

2) 説明変数 $D_i \equiv (D_{i1}, D_{i2}, \dots, D_{iT})$ は、すべて時間変化する D_{it} で構成されている

$$rank \sum_{t=1}^{T} E[\,{}^t\ddot{D}_{it}\ddot{D}_{it}] = K \quad (\dots\text{full rank})$$

<Prf.>
もし、時間変化しない物が入っていると、$K\times K$行列 ${}^t\ddot{D}_{it}\ddot{D}_{it}$のどこかの列、ないし行に0のみで構成される物が、全てのtに関して現れる。（しかも、全てのtについて同じ行、列に出てくる。）なので、tに関して足し合わせても、どうしても全ての要素が0の行、または列が出てくる。フルランクにならない。（実際、D_{it}が正則であることと同値。））

Appendix 3: Fixed Effects model - $\hat{\delta}$の計算

$$Y_{it} = D_{it}\delta_i + +u_i + \varepsilon_{it}$$

$$Y_i = \begin{pmatrix} Y_{i1} \\ \vdots \\ Y_{it} \\ \vdots \\ Y_{iT} \end{pmatrix}_{T\times 1}, D_{it} = \begin{pmatrix} D_{it,1} & \cdots & D_{it,j} & \cdots & D_{it,K} \end{pmatrix}_{1\times K}, \quad D_i = \begin{pmatrix} D_{i1} \\ \vdots \\ D_{it} \\ \vdots \\ D_{iT} \end{pmatrix}_{T\times K},$$

$$\delta_i = \begin{pmatrix} \delta_{i1} \\ \vdots \\ \delta_{ij} \\ \vdots \\ \delta_{iK} \end{pmatrix}_{K\times 1}, \quad \varepsilon_i = \begin{pmatrix} \varepsilon_{i1} \\ \vdots \\ \varepsilon_{it} \\ \vdots \\ \varepsilon_{iT} \end{pmatrix}_{T\times 1}$$

$$\left(\hat{\delta}, \widehat{u_1}, \dots \widehat{u_N}\right) = \arg\min_{\beta, m_1, \dots m_N} \sum_{i=1}^{N} \sum_{t=1}^{T} (Y_{it} - D_{it}\beta - m_i)^2$$

これを満たす$\hat{\delta}$を求める。（別途、参考資料としてPDFを掲載しています。）