Stataによる

パネル調査データ分析の実践

麦山 亮太 Ryota MUGIYAMA

学習院大学法学部政治学科

ryota.mugiyama@gakushuin.ac.jp

2024/03/04 2023年度CSRDA計量分析セミナー・春

自己紹介

現所属

2021/04– 学習院大学法学部政治学科

経歴

2019/03 東京大学大学院人文社会系研究科修了、博士（社会学）

2019/04–2021/03 日本学術振興会特別研究員PD・一橋大学経済研究所

専門

社会階層・社会移動、労働市場、家族形成

*より詳しい業績などはこちら：http://ryotamugiyama.com

日本でもパネル調査データは飛躍的に増えている

東大社研パネル調査（2007–）

親子の生活と学びに関する調査（2015–）

老研・ミシガン全国高齢者パネル調査（1987–）

全国就業実態パネル調査（2016–）

日本家計パネル調査（旧慶應義塾家計パネル調査、2004–）

消費生活に関するパネル調査（1993–2020）

くらしと健康の調査（JSTAR）（2007–）

大阪大学くらしの好みと満足度についてのアンケート（JHPS-CPS）

21世紀出生児／成年者／中高年者縦断調査（2001–）　……etc

繰り返しクロスセクション調査とパネル調査：特徴

繰り返しクロスセクション調査 Repeated cross-section survey：

異なる時点で、同じ調査項目（を含む調査）を、**母集団から都度新しくサンプリングして**調査を実施してデータを収集する

パネル調査 Panel survey：

異なる時点で、同じ調査項目（を含む調査）を、**すでに抽出したサンプルに対して**再度実施してデータを収集する

* 繰り返しクロスセクション調査であると同時にパネル調査にもなるような場合もある（e.g. 個人識別番号と国勢調査が紐付いている社会）

繰り返しクロスセクション調査とパネル調査：強みと弱み

繰り返しクロスセクション調査 Repeated cross-section survey：

- サンプルと母集団のずれはサンプリング時点でのみ生じる。そのため、**異なる時点間で集団の特徴を比較**して記述するのに適している
- 同じ個人を複数回調査しているわけではないので、個人の変化を分析することはできない

パネル調査 Panel survey：

- 同一個人から繰り返しデータを収集するため、**個人の変化**を分析できる
- サンプルと母集団のずれはサンプリング時点とその後の継続調査の**両方**で生じる。そのため、異なる時点間で集団の特徴を比較して記述する目的には劣る

個人の変化を分析する：問いの例

同一個人内の従属変数の変化を記述・説明する

- 貧困状態の人が翌年に貧困でなくなる確率はどれくらいか？

→ 生存分析・イベントヒストリー分析（移行を従属変数とした分析）

同一個人内の時系列変化を記述・類型化する

- 学校を出てから10年間の持ち家の履歴はどのように変化／類型化できるか？

→ ランダム効果（成長曲線）モデル、シークエンス分析・集団軌跡モデル

同一個人内の独立変数の変化が従属変数に与える効果を知る

→今回扱う内容

今回扱う内容

パネルデータを使って、**独立変数の変化が従属変数に与える効果**を明らかにするための方法を学ぶ

- 非正規雇用になると、正規雇用と比べて主観的Well-beingは低くなるのか？
- 出産を経験すると、出産以前と比べて所得は低くなるのか？

→ 固定効果モデル、差分の差法／イベントスタディデザイン

さらに、分析のためにどのようにデータを準備すればよいのか、その手順を学ぶ

下準備：パッケージのインストール

0_install_2024-03-04.doを開き、コードを順に実行しよう

ユーザーが作ったパッケージについての補足

- 一度インストールすると、Stata自体を新しくインストールし直したりしない限りは、再びダウンロードする必要はない
- すでにインストール済のパッケージをインストールしようとすると警告やエラーが出ることがある

目次

パネルデータの構造と作成手順

演習用データ：JLPS2007–2019

データ：東大社研・若年壮年パネル調査、2007–2019年

対象者：2007年時点で20–40歳の男女（2019年には32–52歳になる）。今回は2011年からの追加サンプル、2019年からの新規サンプルは使用しない

働き方とライフスタイルの変化に関する全国調査

パネルデータとは

パネルデータ：同一個人（個体）の複数時点にわたる観察（observation）からなるデータ。1つの行が1つの個体を表すのではなく、1つの観察を表す。

id	wave	income
1	1	112.5
1	2	200
1	3	200
1	4	300
2	1	50
3	1	700
3	2	525
3	3	525

wide形式とlong形式

パネルデータ分析を行う場合には、（ほぼ常に）1つの行が1つの観察を表すようになっている必要がある。

wide形式（1つの行が1つの個体）

id	income1	income2	income3	income4
1	112.5	200	200	300
2	50	.	.	.
3	700	525	525	.

long形式（1つの行が1つの観察）

id	wave	income
1	1	112.5
1	2	200
1	3	200
1	4	300
2	1	50
3	1	700
3	2	525
3	3	525

パネルデータを構築する手順

個人所得の回答区間の中点をとって作成した連続変数を作りたいとする：

問47. 過去一年間の収入についてうかがいます。あなた個人、配偶者、世帯全体の収入はそれぞれどれくらいでしょうか。臨時収入、副収入も含めてお答えください。

	あなた個人	配偶者	世帯全体
1. なし	1	1	1
2. 25 万円未満	2	2	2
3. 50 万円くらい（25～75 万円未満）	3	3	3
4. 100 万円くらい（75～150 万円未満）	4	4	4
5. 200 万円くらい（150～250 万円未満）	5	5	5
6. 300 万円くらい（250～350 万円未満）	6	6	6
7. 400 万円くらい（350～450 万円未満）	7	7	7
8. 500 万円くらい（450～600 万円未満）	8	8	8
9. 700 万円くらい（600～850 万円未満）	9	9	9
10. 1, 000 万円くらい（850～1, 250 万円未満）	10	10	10
11. 1, 500 万円くらい（1, 250～1, 750 万円未満）	11	11	11
12. 2, 000 万円くらい（1, 750～2, 250 万円未満）	12	12	12
13. 2, 250 万円以上	13	13	13
14. わからない	14	14	14
15. 配偶者はいない		15	

パネルデータにたどり着くまでの3つの工程

（パネル調査データ特有の工程）

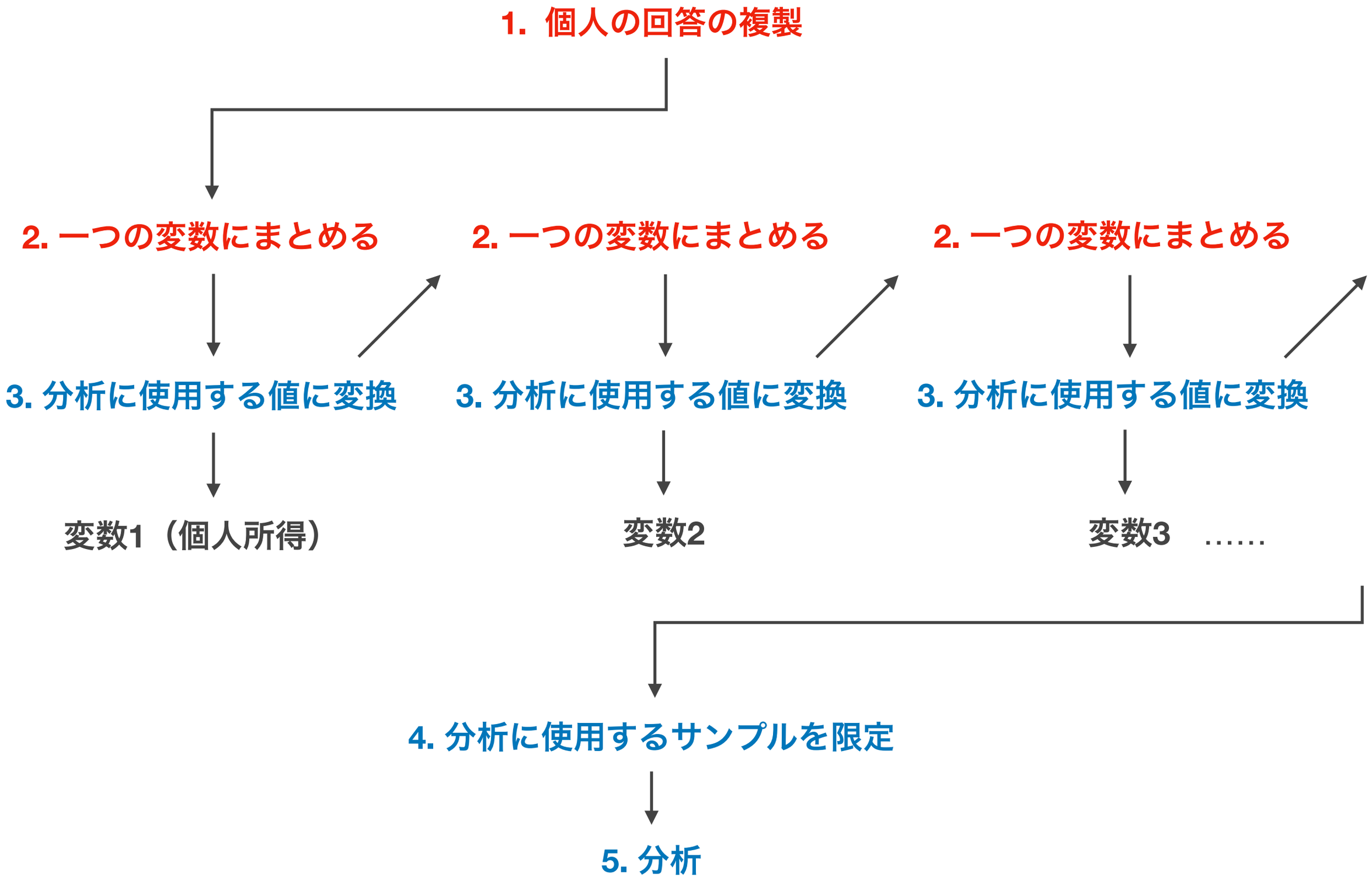
1. 個人の回答の複製

2. 一つの変数にまとめる

id	wave	w1q1	w2q1	w3q1	w4q1	q1	income
1	1	4	5	5	6	4	112.5
1	2	4	5	5	6	5	200
1	3	4	5	5	6	5	200
1	4	4	5	5	6	6	300
2	1	3	.	.	.	3	50
3	1	8	7	7	.	8	700
3	2	8	7	7	.	7	525
3	3	8	7	7	.	7	525

3. 値を変換した変数の作成

データ構築手順のフローチャート



自分が1〜3のどの段階にいるのかを確かめる

もともとのデータの整理のされ方によって、1〜3のどの段階からデータ構築作業を始めればよいかが異なる

1 個人の回答の複製 から始める例：

東大社研・若年壮年パネル調査、子どもの生活と学びに関する親子調査

* 回顧調査をパネルデータに変換する場合もここから（例：SSM調査）

2 一つの変数にまとめる から始める例：

全国就業実態パネル調査

3 値を変換した変数の作成 から始める例：

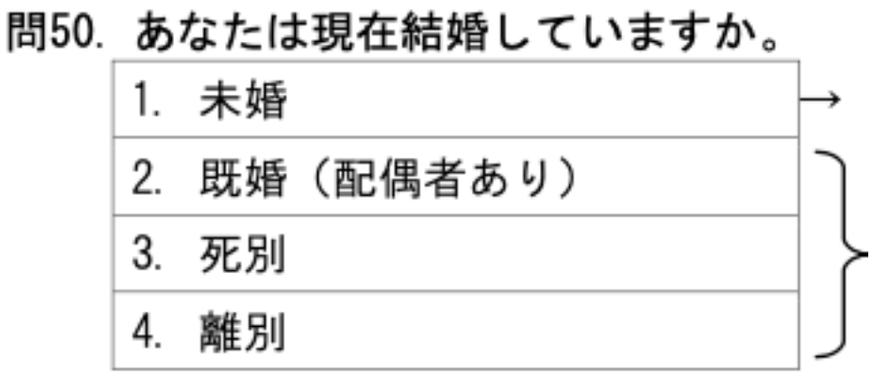
日本家計パネル調査、消費生活に関するパネル調査

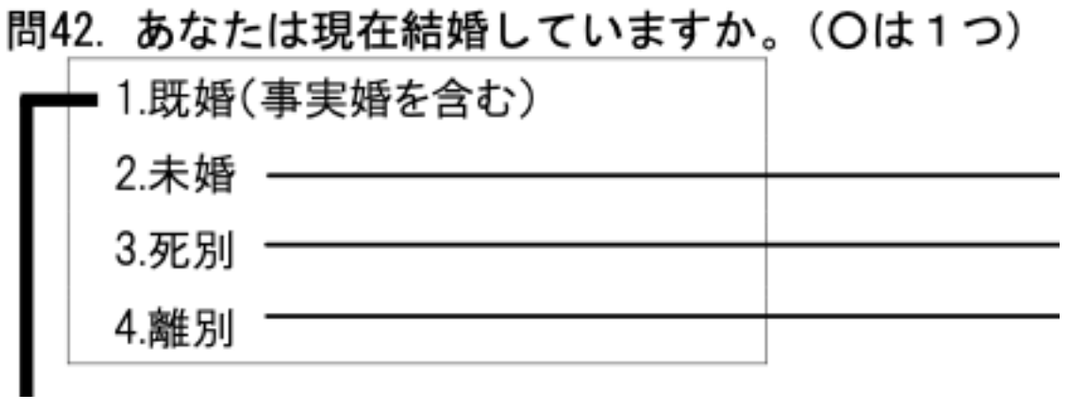
一つの変数にまとめるときの注意点

正しく同じ質問をまとめているかを繰り返し確認する

同じ質問項目でも、選択肢の番号や順序が変わっていないかを確認する

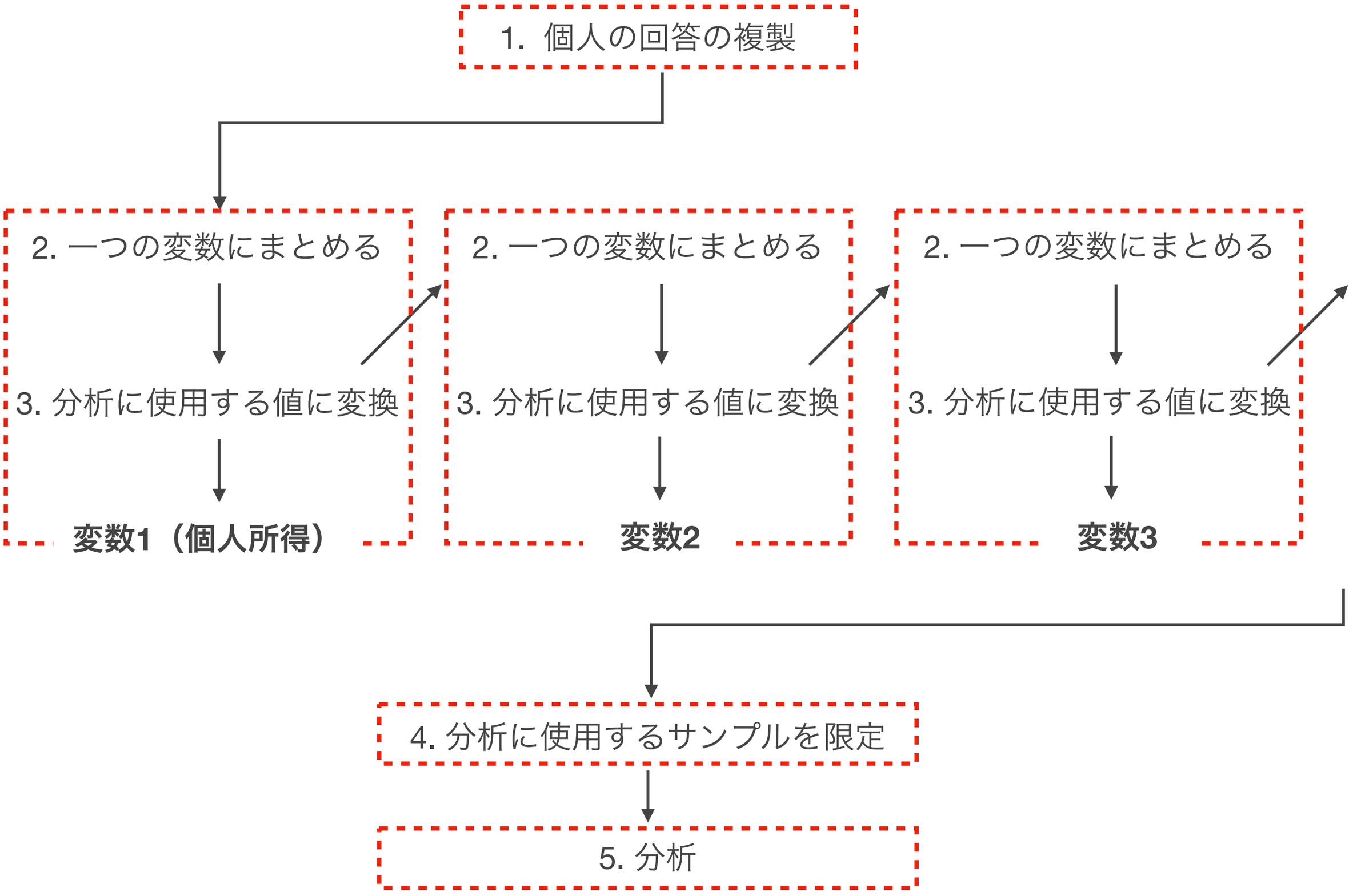
例）東大社研パネル調査の婚姻状態の質問項目はWave 1とWave 2以降で選択肢の順序が異なる





*理想的には、個々の分析者が1や2の段階をしなくていいように提供時からデータが整理されていることが望ましい

困難とdoファイルは分割せよ



doファイルの整理

- master_2024-03-04.do　　全体を統合するdo-file
- 1_expand_2024-03-04.do　　個人の回答を複製するdo-file
- 2_variable　　変数の作成に関わるdo-fileをまとめたフォルダ
  - age_2024-03-04.do　　年齢の変数を作成するdo-file
  - child_2024-03-04.do　　子どもの有無の変数を作成するdo-file……
- 3_sample_2024-03-04.do　　サンプルを限定するdo-file
- 4_analysis_fe　　固定効果モデルの節に関わるdo-fileをまとめたフォルダ
  - 4_1_descriptive.do　　記述的分析に使用するdo-file
  - 4_2_fixedeffect.do　　固定効果モデルを推定するdo-file……

演習：作業ディレクトリの設定

分析前に、分析のコードを走らせる場所（=作業ディレクトリ）をPCに教える。

- File → Change working directory
- または、作業ディレクトリに設定したいフォルダ内にあるdoファイル

  （**master_2024-03-04.do**）を開く

今回は「code」フォルダを作業ディレクトリとして指定。Stataの画面の下部が次のように（末尾が「/code」に）なるはず

doファイルの整理方法についてのtips

- doファイルは何に関する、いつ作成した（編集した）ものなのかがわかるような名前をつけるのがおすすめ
- 類似する作業に関わるコードはまとめてフォルダに入れて管理するとわかりやすい。相対パスを書けばmaster do-fileから同じように走らせることができる
  - ../というふうにすると、作業ディレクトリから1つ上の階層に戻ることを表す。例：`use "../data/JLPS.dta", clear`
- 上から順番にコードを走らせればすべての結果を常に同じように出力できる状態が望ましい（100行目から120行目は飛ばして〜みたいなのはダメ）
- 大きな変更があったときには日付を更新した新しいファイルを作るとよい。過去の日付のコードを走らせれば過去の分析結果を再現できるというのが理想

doファイルの開き方（Mac/Windows）

Mac

- Stataのウインドウは1つしか開かない
- do-fileをクリックして開くと当該フォルダが作業ディレクトリとして設定される（ver 18以降？）が、新しいウィンドウは開かない

Windows

- Stataのウインドウをたくさん開くことができる
- PC上のdo-fileをクリックして開くと、当該のdo-fileが入っているフォルダが自動的に作業ディレクトリとして設定され、新しいウィンドウが開いてしまう

- 作業ディレクトリを変えずに複数のdo-fileを開きたいときには、do-file上のメニューからファイル > 開く > 開きたいdo-fileを選んで開く

演習：回答を複製する

1_expand_2024-03-04.doを開き、コードを順に実行しよう

次のような場合、回答していないwaveについてもデータを複製しておくと良い

- 脱落確率のウェイトを作る場合：各waveで回答したかどうかを従属変数とするモデルを推定するため
- 一度脱落した個人が再度調査に回答するようになる場合：隣接時点の変化に関する変数を作る際に、誤って隣接していない調査時点の値を参照してしまうことがあるため

演習：変数を作成する

2_variableフォルダに入っているdo-fileをそれぞれ確認し、どのように変数を作成しているのかをみてみよう

パネルデータの加工・設定でよく使うコマンド

sort　データの並び替えを行う。個人→時点、の順にソートしてあるとみやすい

by id:　idごとに何らかの変数を作ったり計算をしたりする際に用いる

browse　データを見る（目で見てきちんとできているかを確認するのが大事）

forvalues　指定した値に対して繰り返し処理を実行

variable[_n+1]　ある変数の1行後ろの値を参照する際に用いる。

例）**by id: change = 1 if marriage == 0 & marriage[_n+1] == 1**

xtset id wave　idを個体、waveが時点を表すパネルデータであることを宣言

L.　変数の前につけることで1時点前の値を参照することができる。**L2.**とすると、2時点前の値を参照できる。1時点前の値が存在しない場合は、. を返す。

F.　変数の前につけることで1時点後の値を参照することができる。上に同じ。

おすすめしない方法：reshape long

1. 変数名の変換

wide状態のデータの変数の名前を変更して、末尾に対応するwaveの数値を記載するようにする

id	q1w1	q1w2	q1w3	q1w4
1	4	5	5	6
2	3	.	.	.
3	8	7	7	.

2. long形式への変換

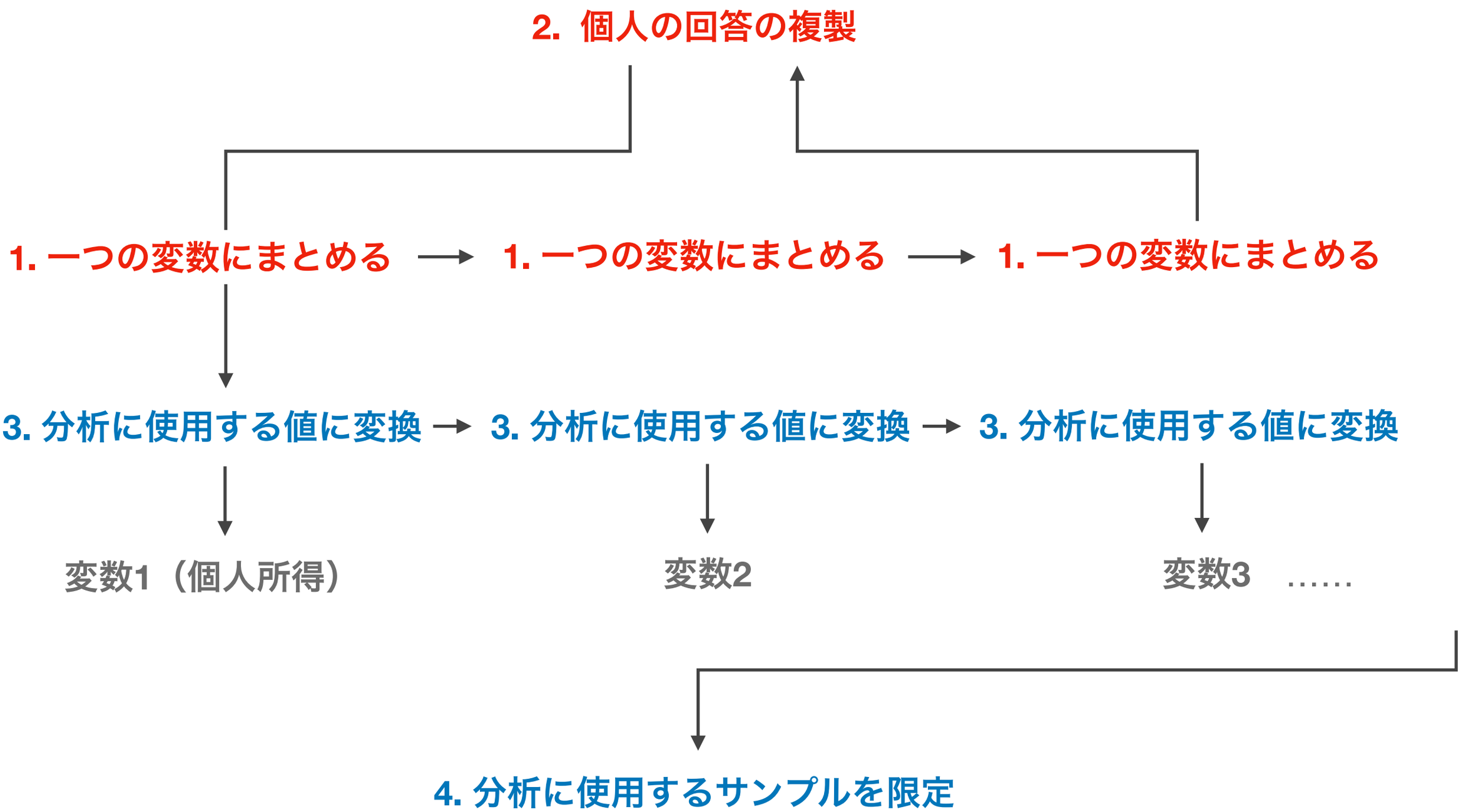
```
reshape long q1w, i(id) j(wave)
```

3. 値を書き換えて変数を作成

```
recode q1w …, gen(income)
```

id	wave	q1w	income
1	1	4	112.5
1	2	5	200
1	3	5	200
1	4	6	300
2	1	3	50
3	1	8	700
3	2	7	525
3	3	7	525

reshape longを使ったデータ構築手順のフローチャート



reshape longを使った方法は変数の作成のコードと回答の複製のコードが分割されるため、一覧性に難があり、管理しにくい

演習：サンプルを限定する

3_sample_variable_2024-03-04.doを開き、コードを順に実行しよう

3_sample_select_2024-03-04.doを開き、コードを順に実行しよう

（後に説明）固定効果モデルを使用する場合、分析に使用するサンプルを作成したのち、2時点以上回答していない個人をサンプルから除外するほうがよい

→1時点しか回答のない個人は推定から除外されてまったく分析に使われないため

固定効果モデル

非正規雇用は主観的well-beingを低くするのか

非正規雇用者は正規雇用者とくらべて主観的well-being（日本語では生活満足度とも、以下SWB）が低いといわれている。

では、非正規雇用はSWBを**低くする**のだろうか？

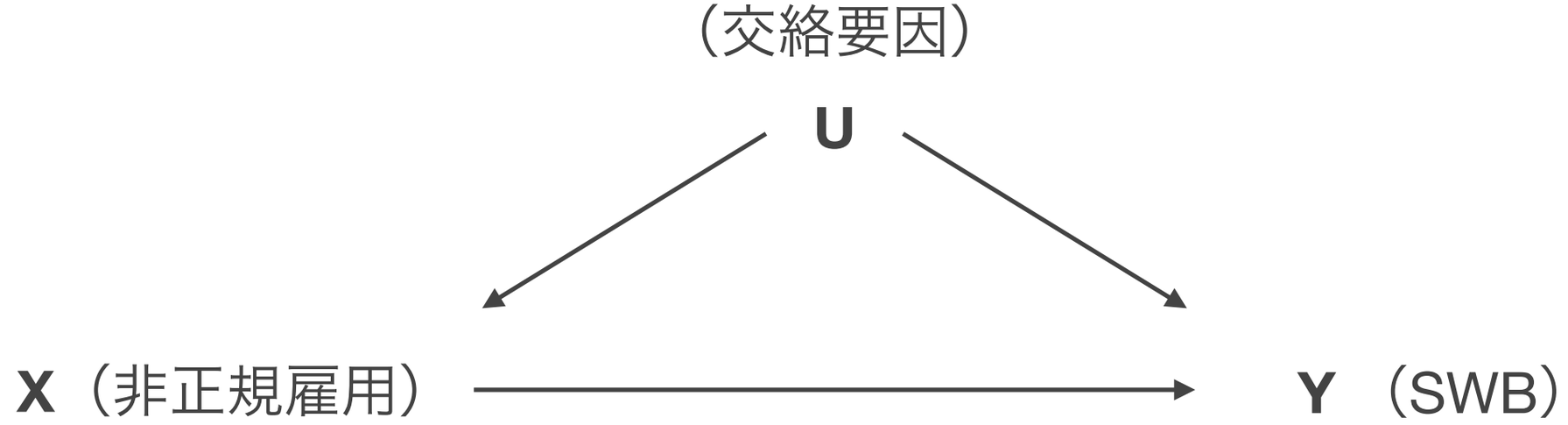
問い. 日本の男性は、非正規雇用になると、正規雇用であるときと比べてSWBが低くなるのだろうか？

Wang, Jia, and James M. Raymo. 2021. “Nonstandard Employment, Gender, and Subjective Well-Being in Japan.” *Journal of Marriage and Family* 83(3):845–64.

交絡の問題

非正規雇用に「なる」ことの効果：現実に非正規雇用の人が正規雇用になったとしたら、現実に正規雇用の人と（平均的に）同じSWBになるし、逆もまた然り

けれど実際には、そうではないかも。現実の両集団は、年齢、世代、学歴、所得、パーソナリティ、子ども時代の経験……などさまざまな点で異なる



もし交絡要因があるなら、両集団の平均の差は非正規雇用に「なる」ことの効果

= 平均処置効果と一致しない

補足：平均処置効果 Average treatment effect

統計的意味で因果関係という場合、そこで知りたいのは、ある個人 i をある時点t において非正規雇用（1）にしたときには、正規雇用（0）と比べてどれくらいSWBが変わるのか、という**処置効果**である：

$$\text{Treatment Effect}_{it} = Y_{it}(1) - Y_{it}(0)$$

しかし実際には2つの状態を同時に観察することはできない。また、より知りたいのは、ある個人ではなく集団における処置効果、つまり**平均処置効果**である：

$$\text{Average Treatment Effect}_{it} = E(Y_{it}(1)) - E(Y_{it}(0))$$

補足：平均処置効果 Average treatment effect

現実の雇用形態			
非正規	$X_{it}=1$	$E(Y_{it}(1)\mid X_{it}=1)$	$E(Y_{it}(0)\mid X_{it}=1)$
正規	$X_{it}=0$	$E(Y_{it}(1)\mid X_{it}=0)$	$E(Y_{it}(0)\mid X_{it}=0)$

平均処置効果を推定する場合には、以下がわかればよい：

$$E(Y_{it}(1))-E(Y_{it}(0))=\left[X_{it}E(Y_{it}(1)\mid X_{it}=1)+(1-X_{it})E(Y_{it}(1)\mid X_{it}=0)\right]$$
$$-\left[X_{it}E(Y_{it}(0)\mid X_{it}=1)+(1-X_{it})E(Y_{it}(0)\mid X_{it}=0)\right]$$

しかし現実には灰色の部分はわからないため、次の値を平均処置効果とみなす：

$$\delta=X_{it}E(Y_{it}(1)\mid X_{it}=1)-(1-X_{it})E(Y_{it}(0)\mid X_{it}=0)$$

***ただし $(1-X_{it})E(Y_{it}(1)\mid X_{it}=0)=X_{it}E(Y_{it}(0)\mid X_{it}=1)$ という仮定の下で！**

線形回帰モデル Pooled OLS

個人iの時点tにおける従属変数をY_{it}、独立変数を$X_{it1}\cdots, X_{itk}$と表記する。このとき、線形回帰モデルは次のように表すことができる：

$$Y_{it} = \beta_0 + \beta_1 X_{it1} + \cdots + \beta_k X_{itk} + r_{it}, \quad r_{it} \sim N(0, \sigma_r^2)$$

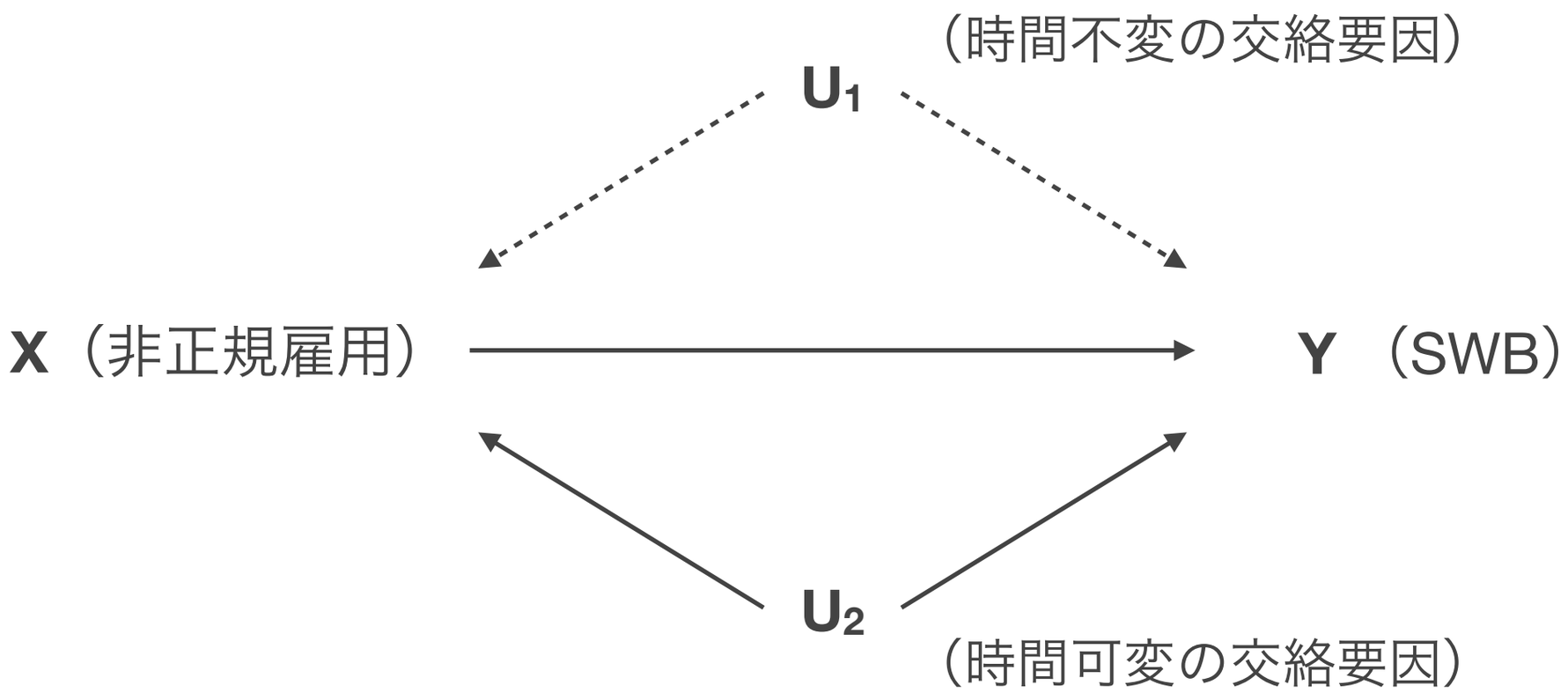
係数β_jは、他の独立変数を一定としたうえで、独立変数X_{itj}が1単位高いときに従属変数Y_{it}がどれだけ高いかを表す

パネルデータを使って通常の線形回帰モデルを推定することを指して Pooled OLS（POLS）と呼ぶ

もし**関心のある独立変数と従属変数の間の交絡要因をすべて統制することができたなら（極めて強い仮定）**、係数は平均処置効果に一致する

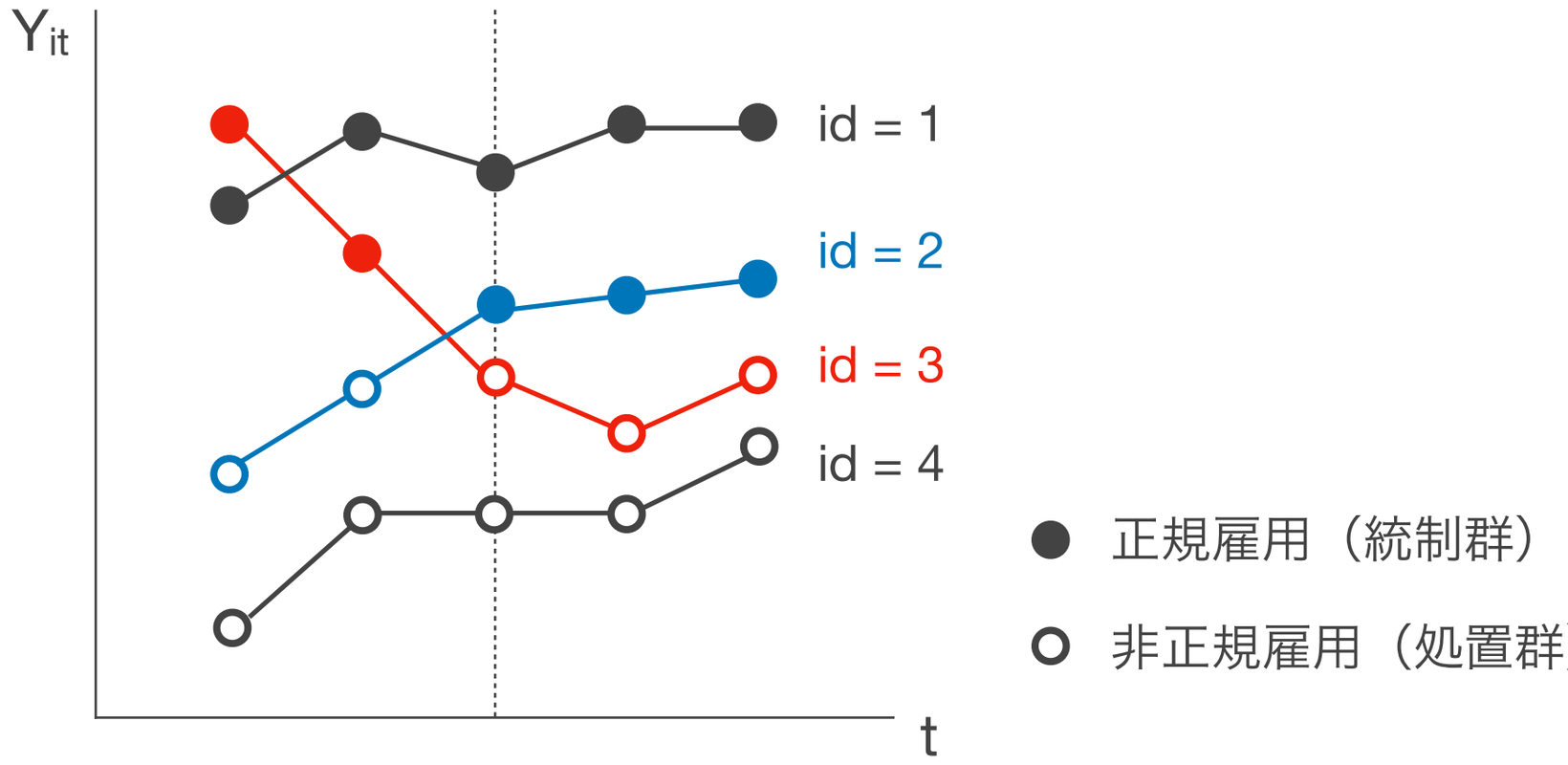
交絡の問題を減らす：パネルデータの強み

同一個人を複数回観察したデータを用いれば、観察期間中に変動しない個人要因（U_1）を統制でき、交絡要因をすべて統制するという条件に**近づく**



もちろん、観察期間中に変動し、XとYの両者に影響する要因（U_2）が十分統制されていない場合、係数は平均処置効果に一致しない（e.g. 事故や病気）

パネルデータの考え方とクロスセクションの考え方



クロスセクション：ある時点における異なる個人を比較する。処置群と統制群の差は個人間の差から生じているかもしれないし、個人内の差から生じているかもしれない。

パネルデータ：同じ個人内の異なる時点を比較できる。この場合、処置群と統制群の差は個人内の差のみによって生じる。

時間可変・不変、個人内・個人間分散

変数の種類はその性質によって時間可変のものと時間不変のものに分けられる

- **時間可変 time-varying**：同一個人内で、観察期間中に値が変化しうる（例：所得、年齢、幸福度）
- **時間不変 time-invariant**：同一個人内で、観察期間中に値が変化しない（例：出生年、15歳の時の暮らし向き）

ある変数の分散 variance は、個人内のものと個人間のものに分けられる

- **個人内分散 within-variance**：同一個人内で生じる値のばらつき
- **個人間分散 between-variance**：異なる個人間で生じる値のばらつき

個人内分散を使う推定の発想

変数の総分散は個人間分散と個人内分散にわけられる：

$$s^2_{overall} \simeq s^2_{between} + s^2_{within}$$

$$\frac{1}{NT-1}\sum_i\sum_t (Y_{it}-\overline{Y})^2 \simeq \frac{1}{NT-1}\sum_i (\overline{Y}_i-\overline{Y})^2 + \frac{1}{NT-1}\sum_i\sum_t (Y_{it}-\overline{Y}_i)^2$$

個人内分散**だけ**を係数の推定に使えば、平均処置効果（＝ある個人に処置を与えたときの従属変数の変化量の平均）に**近い**係数の推定値を得ることができる、というのが個人内分散を使う推定（within-estimation）の発想

演習：変数の個人内・個人間分散

4_1_descriptive.doを開き、書かれているコードを順に実行しよう

Variable		Mean	Std. dev.	Min	Max	Observations
swb	overall	3.661552	.918132	1	5	N = 12643
	between		.7599055	1	5	n = 1575
	within		.5901941	.4115518	6.361552	T-bar = 8.0273
age	overall	37.34058	6.45454	20	52	N = 12643
	between		5.858402	20.5	48.5	n = 1575
	within		3.388444	29.34058	44.89614	T-bar = 8.0273
hincome	overall	7.220003	4.089233	0	31.5	N = 12643
	between		3.49992	.5	31.5	n = 1575
	within		2.242449	-13.03	30.83111	T-bar = 8.0273

	Overall		Between		Within
status	Freq.	Percent	Freq.	Percent	Percent
Regular	11479	90.79	1476	93.71	94.10
Non-regu	1164	9.21	331	21.02	56.23
Total	12643	100.00	1807	114.73	87.16
			(n = 1575)		

Between percent：観察期間中に一度でも当該の雇用形態を経験した人が何%いるか

Within percent：一度でも当該の雇用形態を経験した人に関して、平均して何%の観察が当該の雇用形態であるか

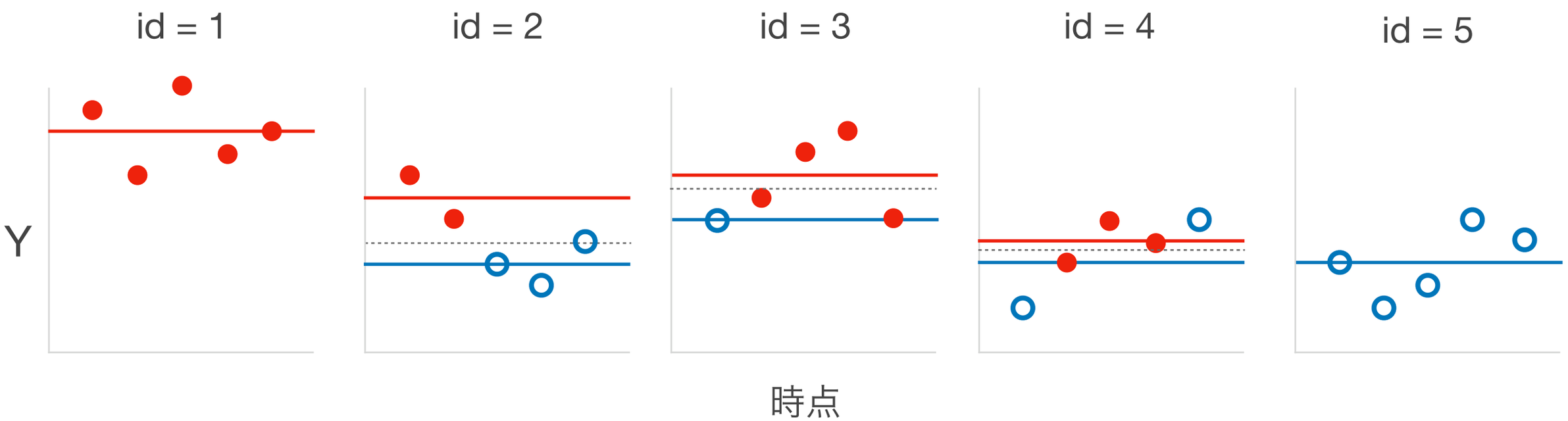
固定効果モデル Fixed-effects model

次のように、個人を表す項 u_i をモデルに含める：

$$Y_{it} = \beta_1 X_{it1} \cdots + \beta_k X_{itk} + u_i + e_{it}$$

- u_i は独立変数を統制したうえでの残差r_{it}（→Pooled OLS）の個人内平均を表す
- 時間不変の独立変数は u_i と区別できないため、推定から除外される。個人がわかれば必ずその値もわかる = 完全な共線性がある
- 係数 β_j は、他の時間可変の独立変数と時間不変の個人要因を統制したうえで、独立変数X_{itj}が1単位高いときに従属変数 Y_{it} がどれだけ高いかを表す。**個人内効果 within-effect/within-estimator**とも言われる

個人内効果（within estimator）のおおまかなイメージ



従属変数の変化と、独立変数の変化（この場合正規雇用と非正規雇用）を経験した人（id 2–4）それぞれについて、以下の値を計算する：

非正規雇用のときのYの平均値 – 正規雇用のときのYの平均値

個人内効果は、これを全個人について重み付け平均をとった値だとイメージすれば良い

推定の2つの方法

Demeaning：各変数および残差の個人内平均を引いた値を用いる

$Y_{it} = \beta_1 X_{it1} + \cdots + \beta_k X_{itk} + u_i + e_{it}$ の個人内平均を取った式は

$$\overline{Y}_i = \beta_1 \overline{X}_{i1} + \cdots + \beta_k \overline{X}_{ik} + u_i + \overline{e}_i$$

上式から下式を引いて得られた以下の式の係数を推定する：

$$Y_{it} - \overline{Y}_i = \beta_1 (X_{it1} - \overline{X}_{i1}) + \cdots + \beta_k (X_{itk} - \overline{X}_{ik}) + (e_{it} - \overline{e}_i)$$

LSDV（Least-Squares Dummy Variable）：個人を表すダミー変数をN（人数）個含める

$$Y_{it} = \beta_1 X_{it} + \cdots + \beta_k X_{kit} + \delta_1 I_{i1} + \cdots + \delta_N I_{iN} + e_{it}$$

通常 $\delta_1, \cdots, \delta_N$ には興味がないので、結果には掲載しない

Demeaned OLS（xtreg, fe）の出力結果例

```
. xtreg swb i.status, fe

Fixed-effects (within) regression               Number of obs      =     12,643
Group variable: id                              Number of groups   =      1,575

R-squared:                                      Obs per group:
     Within  = 0.0021                                         min =          2
     Between = 0.0632                                         avg =        8.0
     Overall = 0.0309                                         max =         13

                                                F(1, 11067)        =      23.02
corr(u_i, Xb) = 0.1607                          Prob > F           =     0.0000
```

swb	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
status						
Non-regular employment	-.172531	.0359618	-4.80	0.000	-.2430226	-.1020394
_cons	3.677436	.0065091	564.97	0.000	3.664677	3.690195
sigma_u	.7492212					
sigma_e	.63013943					
rho	.58569246	(fraction of variance due to u_i)				

F test that all u_i=0: F(1574, 11067) = 9.49 Prob > F = 0.0000

Demeaned OLS（xtreg, fe）の出力結果例

```
. xtreg swb i.status, fe

Fixed-effects (within) regression
Group variable: id

R-squared:
     Within  = 0.0021
     Between = 0.0632
     Overall = 0.0309

corr(u_i, Xb) = 0.1607                  Prob > F          =     0.0000

Number of obs     =     12,643
Number of groups  =      1,575

Obs per group:
                  min =          2
                  avg =        8.0
                  max =         13

F(1, 11067)       =      23.02
```

within R^2

観察数（人×時点）および人数

swb	Coefficient	Std. err.	t	P>|t|	[95% conf. interval]
status					
Non-regular employment	-.172531	.0359618	-4.80	0.000	-.2430226 -.1020394
_cons	3.677436	.0065091	564.97	0.000	3.664677 3.690195
sigma_u	.7492212				
sigma_e	.63013943				
rho	.58569246	(fraction of variance due to u_i)			

残差の個人間分散と個人内分散の大きさ

```
F test that all u_i=0: F(1574, 11067) = 9.49                     Prob > F = 0.0000
```

個人効果u_iが従属変数の分散を有意に説明するかどうかのF検定の結果

LSDV（reghdfe）の出力結果例

```
. reghdfe swb i.status, absorb(id)
(MWFE estimator converged in 1 iterations)

HDFE Linear regression                            Number of obs   =     12,643
Absorbing 1 HDFE group                            F(   1,  11067) =      23.02
                                                  Prob > F        =     0.0000
                                                  R-squared       =     0.5876
                                                  Adj R-squared   =     0.5290
                                                  Within R-sq.    =     0.0021
                                                  Root MSE        =     0.6301
```

swb	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
status						
Non-regular employment	-.172531	.0359618	-4.80	0.000	-.2430226	-.1020394
_cons	3.677436	.0065091	564.97	0.000	3.664677	3.690195

Absorbed degrees of freedom:

Absorbed FE	Categories	- Redundant	= Num. Coefs
id	1575	0	1575

LSDV（reghdfe）の出力結果例

```
. reghdfe swb i.status, absorb(id)
(MWFE estimator converged in 1 iterations)

HDFE Linear regression                            Number of obs   =     12,643
Absorbing 1 HDFE group                            F(   1,  11067) =      23.02
                                                  Prob > F        =     0.0000
                                                  R-squared       =     0.5876
                                                  Adj R-squared   =     0.5290
                                                  Within R-sq.    =     0.0021
                                                  Root MSE        =     0.6301
```

観察数

R^2

within R^2

swb	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
status						
Non-regular employment	-.172531	.0359618	-4.80	0.000	-.2430226	-.1020394
_cons	3.677436	.0065091	564.97	0.000	3.664677	3.690195

Absorbed degrees of freedom:

Absorbed FE	Categories	- Redundant	= Num. Coefs
id	1575	0	1575

個人ダミーとその係数の個数

2つの方法の決定係数R^2の違いについて

Demeaned OLS：個人内分散 s^2_{within} の説明量

$Y_{it} - \overline{Y}_i$ の分散がどの程度 $X_{1it} - \overline{X}_{1i}, \cdots, X_{kit} - \overline{X}_{ki}$ によって説明できるかを表す。StataではWithin R-squaredと表記される

LSDV：総分散 $s^2_{overall}$ の説明量

Y_{it} の分散がどの程度$X_{1it}, \cdots, X_{kit}, I_{i1}, \cdots, I_{iN}$ によって説明できるかを表す（いわゆる通常のR^2）

それぞれの決定係数は違った情報を与えるものであり、どちらが良いというものではない

演習：固定効果モデルを推定する

4_2_fixedeffect.doを開き、コードを順に実行しよう

	(1) POLS	(2) Demeaned	(3) LSDV
Regular employment	0.000	0.000	0.000
	(.)	(.)	(.)
Non-regular employ~t	-0.247***	-0.119***	-0.119***
	(0.027)	(0.036)	(0.036)
age	-0.009***	-0.000	-0.000
	(0.001)	(0.002)	(0.002)
Never married	0.000	0.000	0.000
	(.)	(.)	(.)
Married	0.605***	0.490***	0.490***
	(0.019)	(0.033)	(0.033)
Separated/divorced	-0.034	0.021	0.021
	(0.049)	(0.065)	(0.065)
Household income (~)	0.035***	0.008**	0.008**
	(0.002)	(0.002)	(0.002)
Constant	3.341***	3.288***	3.288***
	(0.046)	(0.062)	(0.062)
Observations	12643	12643	12643
r2	0.141	0.029	0.599

Standard errors in parentheses
* p<0.05, ** p<0.01, *** p<0.001

残差の自己相関とクラスター・ロバスト標準誤差

回帰分析では残差が独立に同一の分布に従う（i.i.d.）という仮定のもとで標準誤差を計算するが、パネルデータの場合にはこの仮定が成り立たない可能性が非常に高い = 残差の自己相関 autocorrelation, あるいは系列相関 serial correlation

一般にパネルデータ分析の際には、個人（クラスター）内の相関を許容するクラスター・ロバスト標準誤差の使用が強く推奨される*

ついでにクラスター・ロバスト標準誤差は残差の不均一分散に対しても頑健

*clusterの数が小さい（50未満？）とバイアスが生じることが知られているが（Cameron and Miller 2015）、パネル調査データで個人をクラスタとする場合には、ほとんど心配する必要はない

Cameron, A. Colin, and Douglas L. Miller. 2015. “A Practitioner’s Guide to Cluster-Robust Inference.” *Journal of Human Resources* 50(2):317–72.

演習：クラスター・ロバスト標準誤差

4_2_fixedeffect.doを開き、コードを順に実行しよう

	(1) POLS	(2) Demeaned	(3) LSDV
Regular employment	0.000	0.000	0.000
	(.)	(.)	(.)
Non-regular employ~t	-0.247***	-0.119**	-0.119**
	(0.057)	(0.046)	(0.046)
age	-0.009***	-0.000	-0.000
	(0.002)	(0.002)	(0.002)
Never married	0.000	0.000	0.000
	(.)	(.)	(.)
Married	0.605***	0.490***	0.490***
	(0.041)	(0.047)	(0.047)
Separated/divorced	-0.034	0.021	0.021
	(0.087)	(0.088)	(0.088)
Household income (~)	0.035***	0.008**	0.008**
	(0.003)	(0.003)	(0.003)
Constant	3.341***	3.288***	3.288***
	(0.087)	(0.086)	(0.086)
Observations	12643	12643	12643
N_clust	1575.000	1575.000	1575.000
r2	0.141	0.029	0.599

Standard errors in parentheses
* p<0.05, ** p<0.01, *** p<0.001

二方向固定効果モデル Two-way fixed-effects model

調査時点固有の効果（e.g. 景気変動）が結果に影響する可能性がある。このような場合、次のように時点の効果（時点を表すダミー変数）を統制するモデルを推定する：

$$Y_{it} = \beta_1 X_{1it} + \cdots + \beta_k X_{kit} + u_i + \tau_t + e_{it}$$

u_i をとくに個人固定効果、τ_t を時点固定効果などという。

強い理論的仮定がない限り、基本的には時点を表す変数は統制したほうがよい

時点と区別できない変数はモデルから除かれる（e.g. 同一コーホートを追跡したデータにおける年齢の係数）

個別トレンド固定効果モデル Fixed-effects individual-slope model

個人によって（Xを条件づけたうえでの）平均値が異なるだけではなく、時点間の傾きも異なると想定する：

$$Y_{it} = \beta_1 X_{1it} + \cdots + \beta_k X_{kit} + u_i + \gamma_i t + e_{it}$$

さらに時点固定効果を加えた以下のモデルを推定することもできる：

$$Y_{it} = \beta_1 X_{1it} + \cdots + \beta_k X_{kit} + u_i + \gamma_i t + \tau_t + e_{it}$$

3時点以上の観察をもつ個体だけが独立変数の係数の推定値に貢献する

時間可変の個人効果を統制したモデルと考えればよいが、それが線形であるという強い仮定を置いている点に注意が必要

Rüttenauer, Tobias, and Volker Ludwig. 2023. “Fixed Effects Individual Slopes: Accounting and Testing for Heterogeneous Effects in Panel Data or Other Multilevel Models.” *Sociological Methods & Research* 52(1):43–84.

演習：二方向固定効果、個別トレンド固定効果

4_3_twfe_feis.doを開き、コードを順に実行しよう

	(1) TWFE - Dem~d		(2) TWFE - LSDV	
Regular employment	0.000	(.)	0.000	(.)
Non-regular employ~t	-0.122**	(0.045)	-0.122**	(0.045)
Never married	0.000	(.)	0.000	(.)
Married	0.489***	(0.047)	0.489***	(0.047)
Separated/divorced	0.019	(0.089)	0.019	(0.089)
Household income (~)	0.007**	(0.003)	0.007**	(0.003)
wave=1	0.000	(.)		
wave=2	-0.091**	(0.029)		
wave=3	0.084**	(0.029)		
wave=4	-0.020	(0.032)		
wave=5	0.022	(0.032)		
wave=6	0.008	(0.032)		
wave=7	-0.075*	(0.032)		
wave=8	0.022	(0.033)		
wave=9	-0.021	(0.033)		
wave=10	0.013	(0.034)		
wave=11	-0.015	(0.035)		
wave=12	-0.021	(0.035)		
wave=13	0.005	(0.034)		
Constant	3.292***	(0.041)	3.286***	(0.041)
Observations	12643		12643	
N_clust	1575.000		1575.000	
r2	0.035		0.601	

Standard errors in parentheses
* p<0.05, ** p<0.01, *** p<0.001

固定効果モデルの注意点

Xが変化していない個人はXの係数の推定に（間接的にしか）貢献しない

- 観察回数が少ないほど、また変化しにくい変数ほど、効果を検出するためにより多くのサンプルサイズが必要。有意でないから効果がないとは即断できない（statistical significance ≠ scientific significance）
- 個人内効果はATTであって、平均処置効果（ATE）とは一致しないかも（e.g. 雇用形態が変わる人は、非正規の負の効果が小さい人かも？）

時間可変の要因は統制しきれない

- 交絡となりうる観察可能な時間可変の変数を適切にモデルに含める必要

変化の方向を区別しない

- 現実にどのような移動が起こっているのかをチェックしておく必要

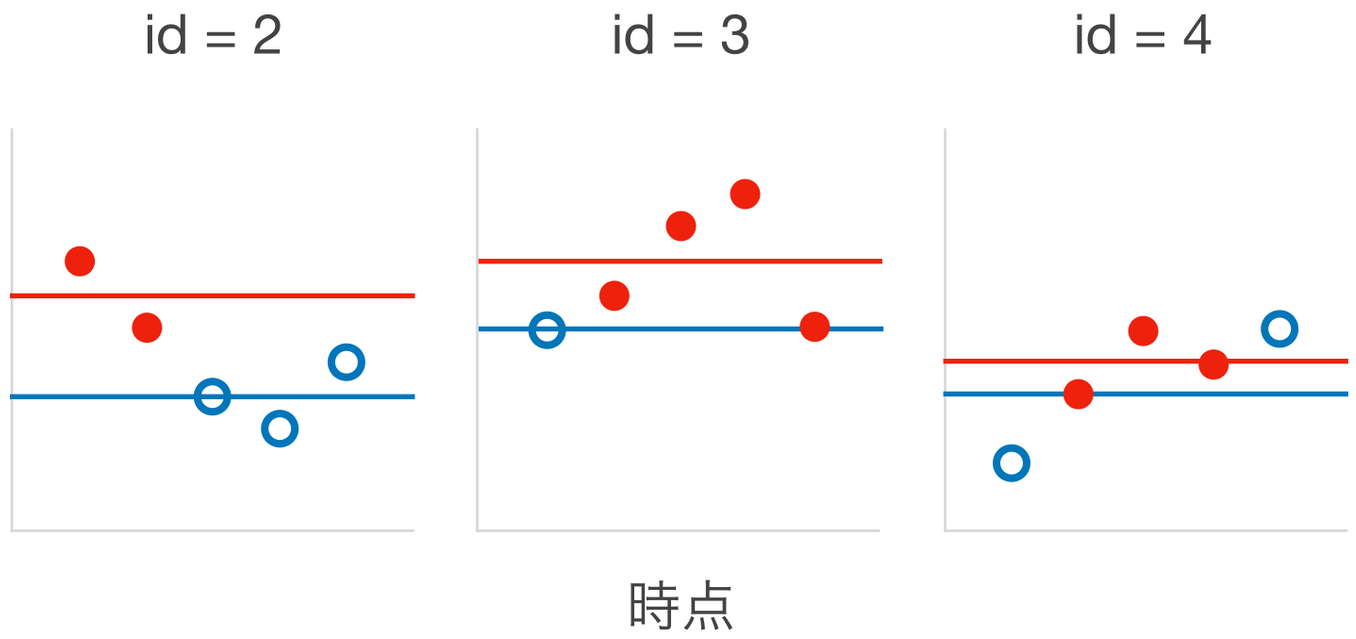
大久保将貴，2021，「パネルデータ分析における固定効果モデルの取扱説明書」『社会科学研究』72(2): 55–68. など

演習：移行行列

4_4_transitionmatrix.doを開き、コードを順に実行しよう

Employment status	Employment status 1	2	Total
1	9,913	156	10,069
	98.45	1.55	100.00
2	207	792	999
	20.72	79.28	100.00
Total	10,120	948	11,068
	91.43	8.57	100.00

補足：変化の方向の問題



固定効果モデルの推定値 = （他の時間可変の独立変数を統制したうえでの）同一個人内の正規雇用のときの平均値と非正規雇用のときの平均値の差を集めたもの

変化を経験した個人のなかには、「正規雇用から非正規雇用になった人」と「非正規雇用から正規雇用になった人」がいる

補足：変化の方向の問題

固定効果モデルで得られた非正規雇用の係数は負であった場合、非正規雇用であると、正規雇用であるときと比べてSWBが低いといえる

しかし、それが「正規から非正規になると下がる」からなのか、「非正規から正規になると上がる」からなのかは厳密には区別できない

ただし、事実上一方向の変化が主である場合には、実質的意味は変わる

変化の向きを区別したモデルなども提唱されているが（Allison 2019; 有田・仲 2021）、安易に使用しないほうがいい。それが本当に必要な問いなのか、だとしても研究デザインで解決すべき問題ではないのかをよくよく考える

Allison, Paul D. 2019. “Asymmetric Fixed-Effects Models for Panel Data.” *Socius* 5:2378023119826441.
有田伸・仲修平，2021，「変化の向き等を区別したパネルデータ分析の実践：それでも使いたいあなたに」『東京大学社会科学研究所パネル調査プロジェクト ディスカッションペーパーシリーズ』134.

イベントの効果推定

子どもをもつ女性は労働市場で不利を被るのか

出産を経ることで女性は就業面でさまざまな不利（就業中断、所得や賃金の減少など）を被るといわれている。では実際、その大きさはどの程度なのだろうか？

X（子ども） ⟶ **Y**
（所得／就業／賃金）

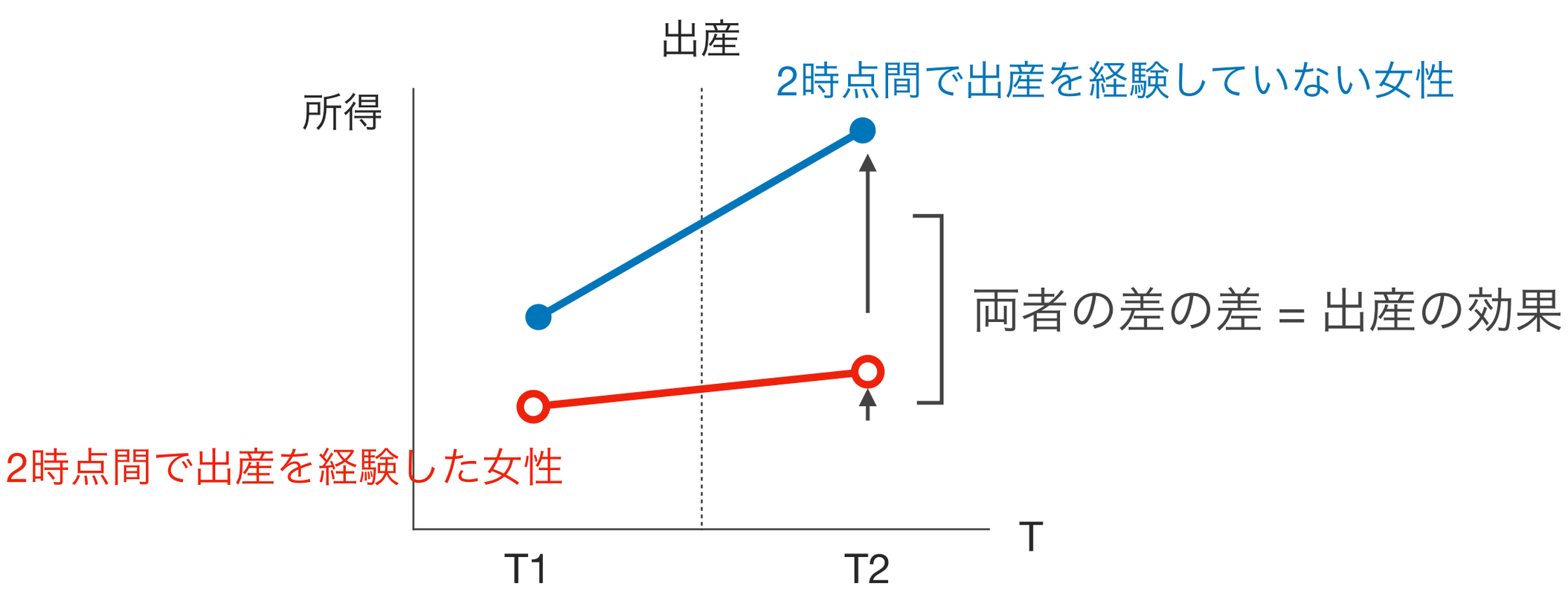
問い. 日本において、出産を経ることは女性の個人所得に対してどのような影響を持つのか？その効果はどれほど続くのか？

Hsu, Chen-Hao. 2021. “Parity-Specific Motherhood Penalties: Long-Term Impacts of Childbirth on Women’s Earnings in Japan.” *Advances in Life Course Research* 50:100435.
成年者縦断調査を用いた例：https://www.mof.go.jp/pri/research/conference/fy2021/shigoto_report03.pdf

イベントの効果推定の考え方

ある状態（D＝0とする）から別の状態（D＝1とする）へと変化し、それ以前の状態に戻ることはないような変化（イベント）の効果を推定する

イベントを経験した人の前後の所得変化を、イベントを経験していない人の平均的な所得変化と比較し、その差をもってイベントの効果とみなす



差分の差法 Difference-in-differences (DD)

2時点のパネルデータがあり、D_{it}はイベントを経験していないときは0、したあとは1をとるダミー変数とする。次のモデルで、イベントの効果を推定できる：

$$Y_{it} = \alpha D_{it} + \beta_1 T2_t + u_i + e_{it}$$

イベントの発生が外生的である（イベントの発生が処置／統制群の別と相関しない）と仮定できるなら、個人固定効果 u_i を気にする必要はなく*、以下のように単純化できる：

$$Y_{it} = \alpha D_{it} + \beta_1 T2_t + \beta_2 \text{Treated}_i + e_{it}$$

このモデルは繰り返しクロスセクションでも推定できる。

*ただし、係数の大小に影響しないとしても、統制することで標準誤差を小さくすることはできるかもしれない。

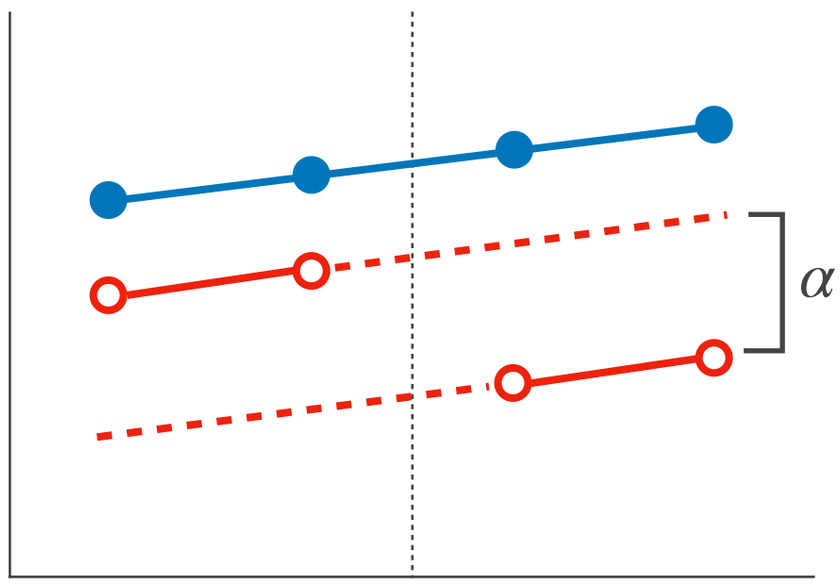
多時点のDD

二方向固定効果モデル（TWFE）は3時点以上のときのDDとして理解できる：

$$Y_{it} = \alpha D_{it} + \beta_1 X_{it1} + \cdots + \beta_k X_{itk} + \tau_t + u_i + e_{it}$$

D_{it}はイベントの発生前はずっと0、発生後はずっと1を取るダミー変数。

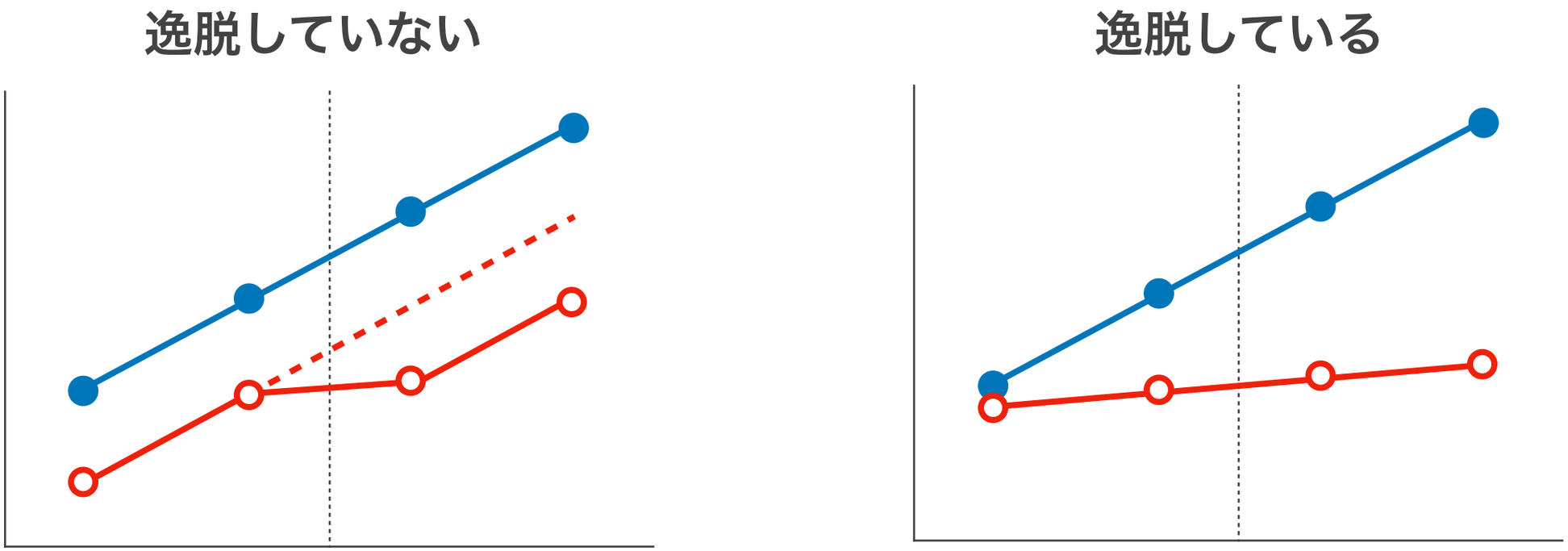
係数αは（イベントを経験しなかったときに得られるだろう）値の平均と、イベント発生後の値の平均の差を捉えたものと解釈できる



Goodman-Bacon, Andrew. 2021. “Difference-in-Differences with Variation in Treatment Timing.” *Journal of Econometrics* 225(2):254–77.

平行トレンドの仮定 parallel trend assumption

平行トレンドの仮定：イベントを経験した人としていない人は、仮にイベント経験がなかったならば平均的に同じトレンドをたどるだろう（イベント経験前後の差分の差はイベント経験のみによって生じる）という仮定



仮定を直接検証するのは不可能だが、それ以前のトレンドを確認したり（→後述）、統制したり（→synthetic control method）、偽の処置に関する変数を作ったり（→placebo test）することで仮定が正しいか推測することはできる

Motherhood penaltyの推定にSynthetic control methodを用いた例として、Vagni, G., & Breen, R. (2021). Earnings and income penalties for motherhood: Estimates for British women using the individual synthetic control method. *European Sociological Review*, *37*(5), 834–848.

演習：イベント変数の作成

5_1_eventvariable.doを開き、イベント経験の変数の作りかたを確認しよう

	id	year	child1birt~r	birthexp	birthgap	birth_lead10	birth_lead9
1	1	2007	.	0	.	0	0
2	1	2008	.	0	.	0	0
3	1	2009	.	0	.	0	0
4	1	2010	.	0	.	0	0
5	1	2011	.	0	.	0	0
6	1	2012	.	0	.	0	0
7	1	2013	.	0	.	0	0
8	1	2014	.	0	.	0	0
9	1	2015	.	0	.	0	0
10	1	2016	.	0	.	0	0
11	1	2017	.	0	.	0	0
12	2	2007	2000	.	.	0	0
13	2	2008	2000	.	.	0	0
14	2	2009	2000	.	.	0	0
15	2	2010	2000	.	.	0	0
16	2	2011	2000	.	.	0	0
17	2	2012	2000	.	.	0	0
18	2	2013	2000	.	.	0	0
19	2	2014	2000	.	.	0	0
20	2	2015	2000	.	.	0	0
21	2	2016	2000	.	.	0	0
22	2	2017	2000	.	.	0	0
23	3	2007	.	0	.	0	0
34	4	2007	2012	1	-5	0	0
35	4	2008	2012	1	-4	0	0
36	4	2009	2012	1	-3	0	0
37	4	2010	2012	1	-2	0	0
38	4	2011	2012	1	-1	0	0
39	4	2012	2012	1	0	0	0
40	4	2013	2012	0	1	0	0
41	4	2014	2012	0	2	0	0
42	4	2015	2012	0	3	0	0

演習：差分の差法

5_2_diff_in_diffs.doを開き、コードを順に実行しよう

```
HDFE Linear regression                            Number of obs   =     10,750
Absorbing 2 HDFE groups                           F(   2,   1135) =     171.90
Statistics robust to heteroskedasticity           Prob > F        =     0.0000
                                                  R-squared       =     0.7121
                                                  Adj R-squared   =     0.6776
                                                  Within R-sq.    =     0.0841
Number of clusters (id)      =      1,136         Root MSE        =     1.0427

                                      (Std. err. adjusted for 1,136 clusters in id)
```

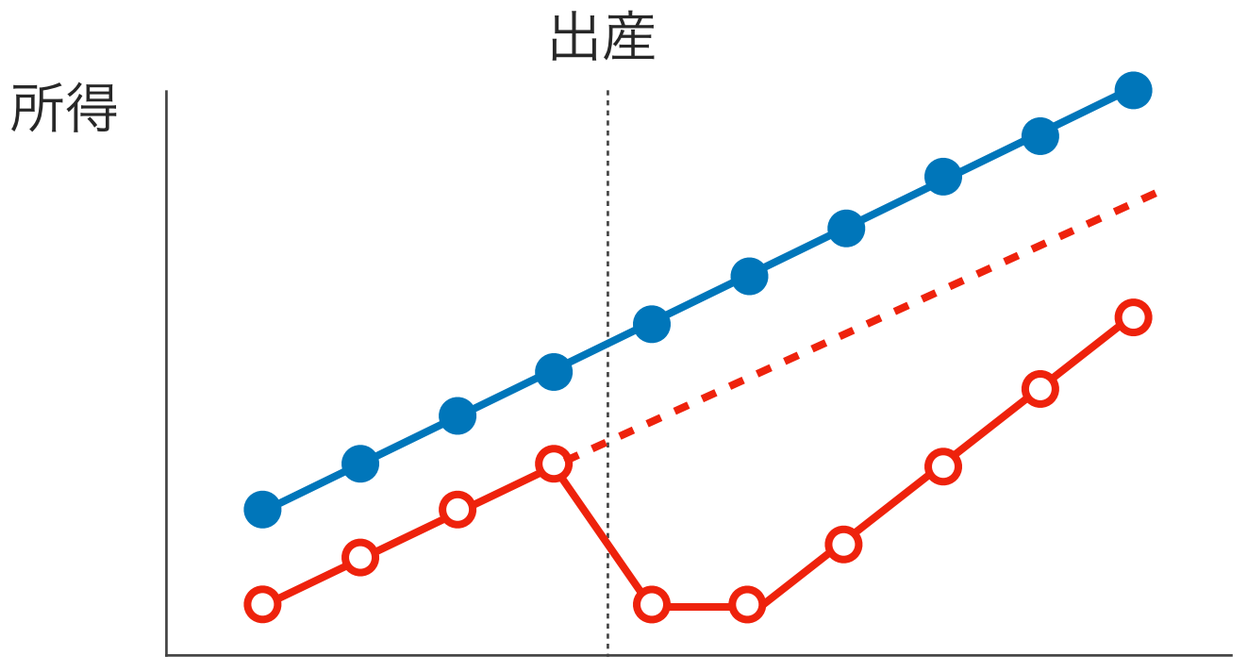
income	Coefficient	Robust std. err.	t	P>|t|	[95% conf.	interval]
birthexp	-1.325537	.0715132	-18.54	0.000	-1.46585	-1.185224
age	0	(omitted)				
c.age#c.age	-.0012942	.0004735	-2.73	0.006	-.0022233	-.0003651
_cons	4.24809	.5793805	7.33	0.000	3.111313	5.384868

Absorbed degrees of freedom:

Absorbed FE	Categories	- Redundant	= Num. Coefs	
id	1136	1136	0	*
year	13	0	13	

* = FE nested within cluster; treated as redundant for DoF computation

前後比較の拡張：長期効果 long-term effect の検証



子どもを持つことが所得に与える効果は出産経験後ずっと一定と考えるよりも、出産直後に最も大きく、その後は小さくなっていくと考えるほうが妥当かもしれない（e.g. 就業中断後の再就職）

効果が経過時間によって変化する様子をみたいならば、イベントからの時間を区別できるモデリングが必要

Motherhood penaltyに関連して時間経過による効果の変化を社会学の理論に引きつけて考えている論文として、Cheng, S. (2016). The Accumulation of (Dis)advantage: The Intersection of Gender and Race in the Long-Term Wage Effect of Marriage. *American Sociological Review*, *81*(1), 29–56.

イベントスタディ Event-study design

$$Y_{it} = \sum_{p=2}^{P} \alpha_p D_{itp}^{lead} + \sum_{q=0}^{Q} \alpha_q D_{itq}^{lag} + \beta_1 X_{iti} + \cdots + \beta_k X_{itk} + u_i + \tau_t + e_{it}$$

D_{itp}^{lead} ：個人iがイベントを経験した年から数えてp年前であることを示すダミー変数。イベントを経験していない人は常に0をとる

D_{itp}^{lag} ：個人iがイベントを経験した年から数えてq年後であることを示すダミー変数。イベントを経験していない人は常に0をとる

α_p, α_q：基準年（この場合イベントの1年前）と比較して、どれくらいY_{it}の値が高いかを示す係数

イベントスタディの強み

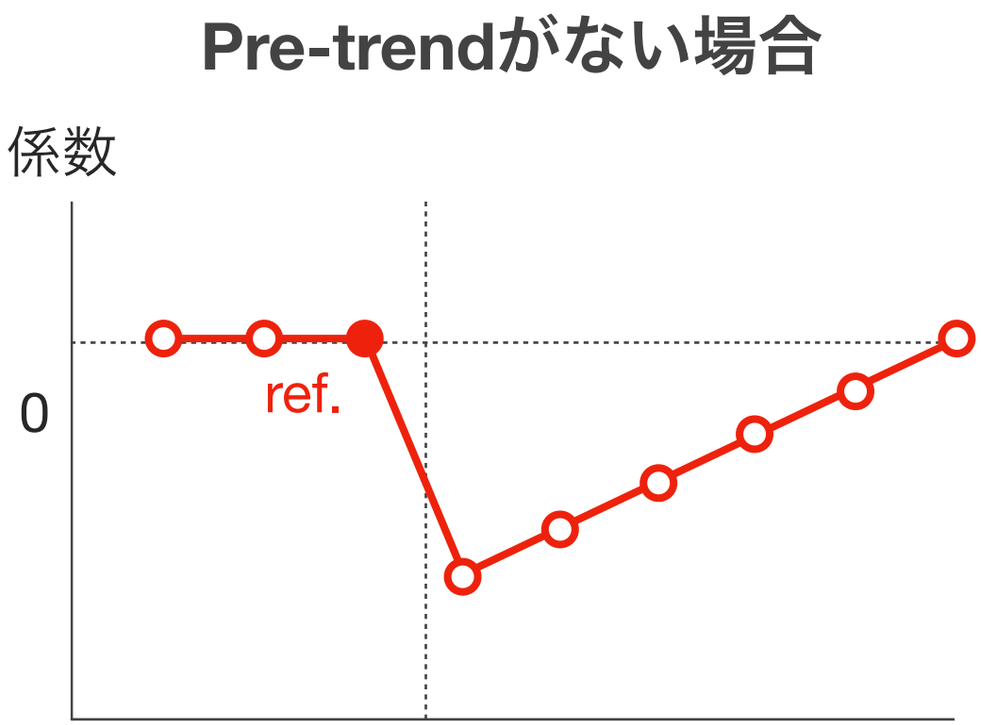
DD: $$Y_{it} = \alpha D_{it} + \beta_1 X_{it1} + \cdots + \beta_k X_{itk} + \tau_t + u_i + e_{it}$$

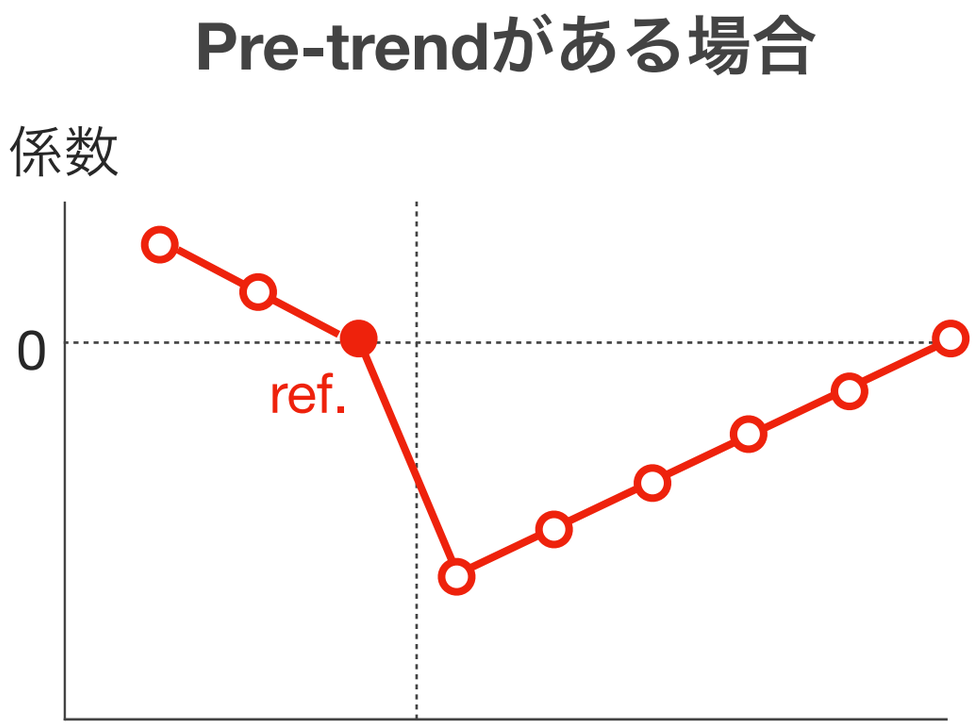
Event: $$Y_{it} = \sum_{p=2}^{P} \alpha_p D_{itp}^{lead} + \sum_{q=0}^{Q} \alpha_q D_{itq}^{lag} + \beta_1 X_{iti} + \cdots + \beta_k X_{itk} + u_i + \tau_t + e_{it}$$

1. イベントの効果が時間によって変化する場合には、イベントスタディのほうがより正確な効果を記述できる。イベントの効果が時間によって変化するにもかかわらずDDを使った場合、効果の大きさは観察期間の長さに依存する（イベントからの経過時間の分布が変わるため）

2. イベント以前のトレンド（pre-trend）を見ることで、間接的に平行トレンドの仮定を検証できる

Pre-trendは平行トレンドの逸脱を見破る鍵





1. 除外変数バイアス：処置群を統制群を分け、かつ従属変数に影響する要因が存在する
2. 逆因果あるいはセレクション：従属変数が低下（上昇）傾向にある人は処置群に入りやすい
3. 予期効果：将来の処置を予期して現在の行動を変える

Brüderl, Josef, and Volker Ludwig. 2019. “Applied Panel Data Analysis.” Retrieved February 21, 2023 (https://www.ls3.soziologie.uni-muenchen.de/studium-lehre/archiv/teaching-marterials/panel-analysis_april-2019.pdf).

イベントスタディによる出産の効果推定の事例

第一子出産直前と比較して女性の勤労所得（非就業者も含む）は大きく低下し、その後多少上昇するものの、10年後も低下の影響が残る

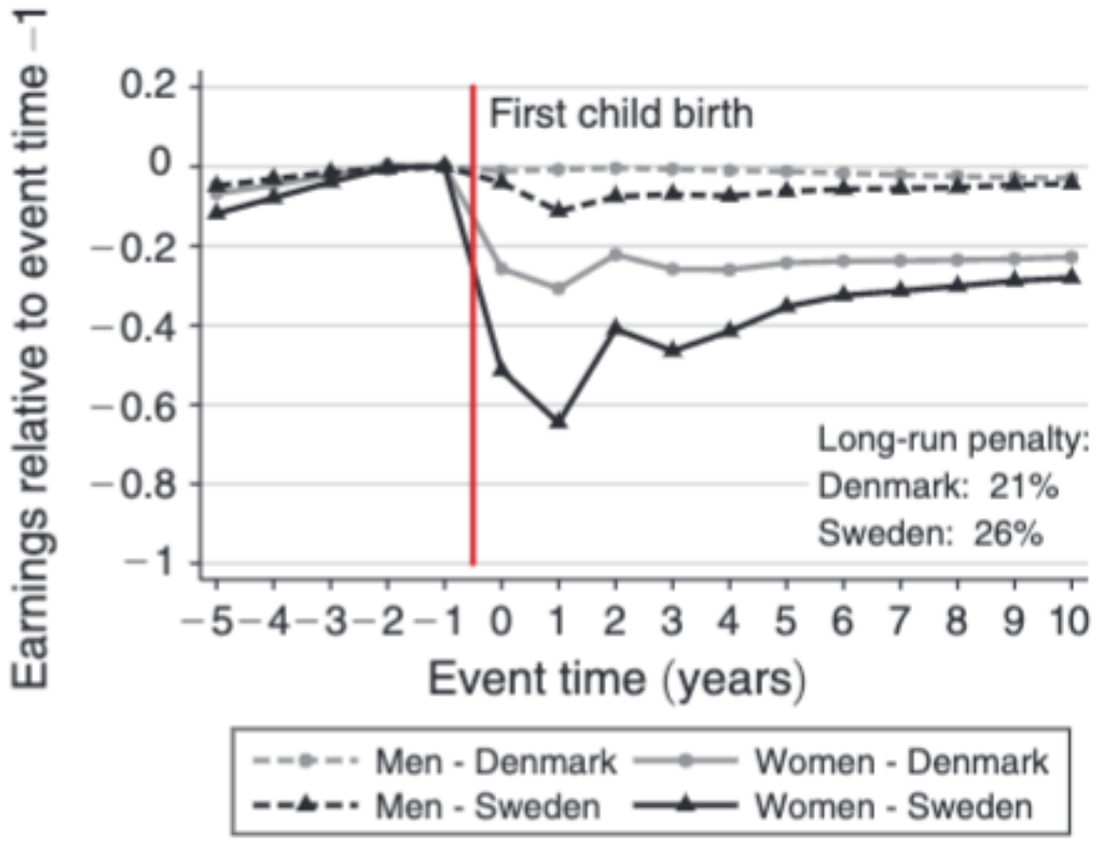


FIGURE 1. CHILD PENALTIES IN EARNINGS IN SCANDINAVIAN COUNTRIES

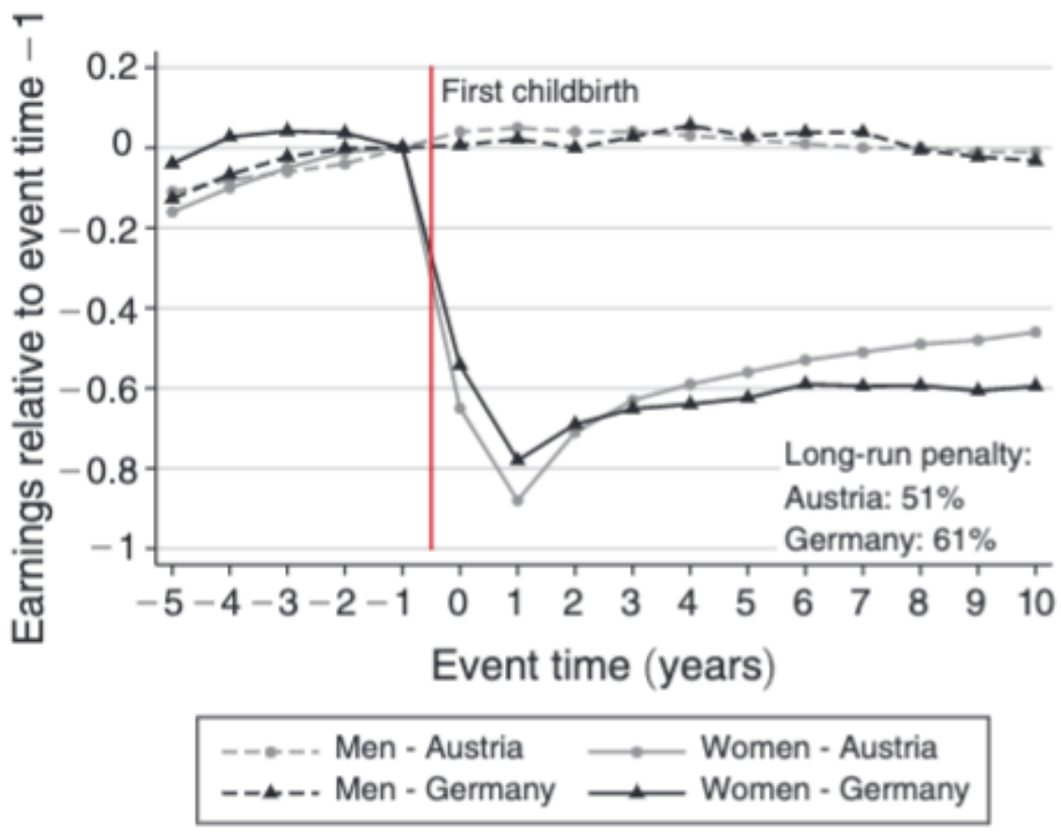
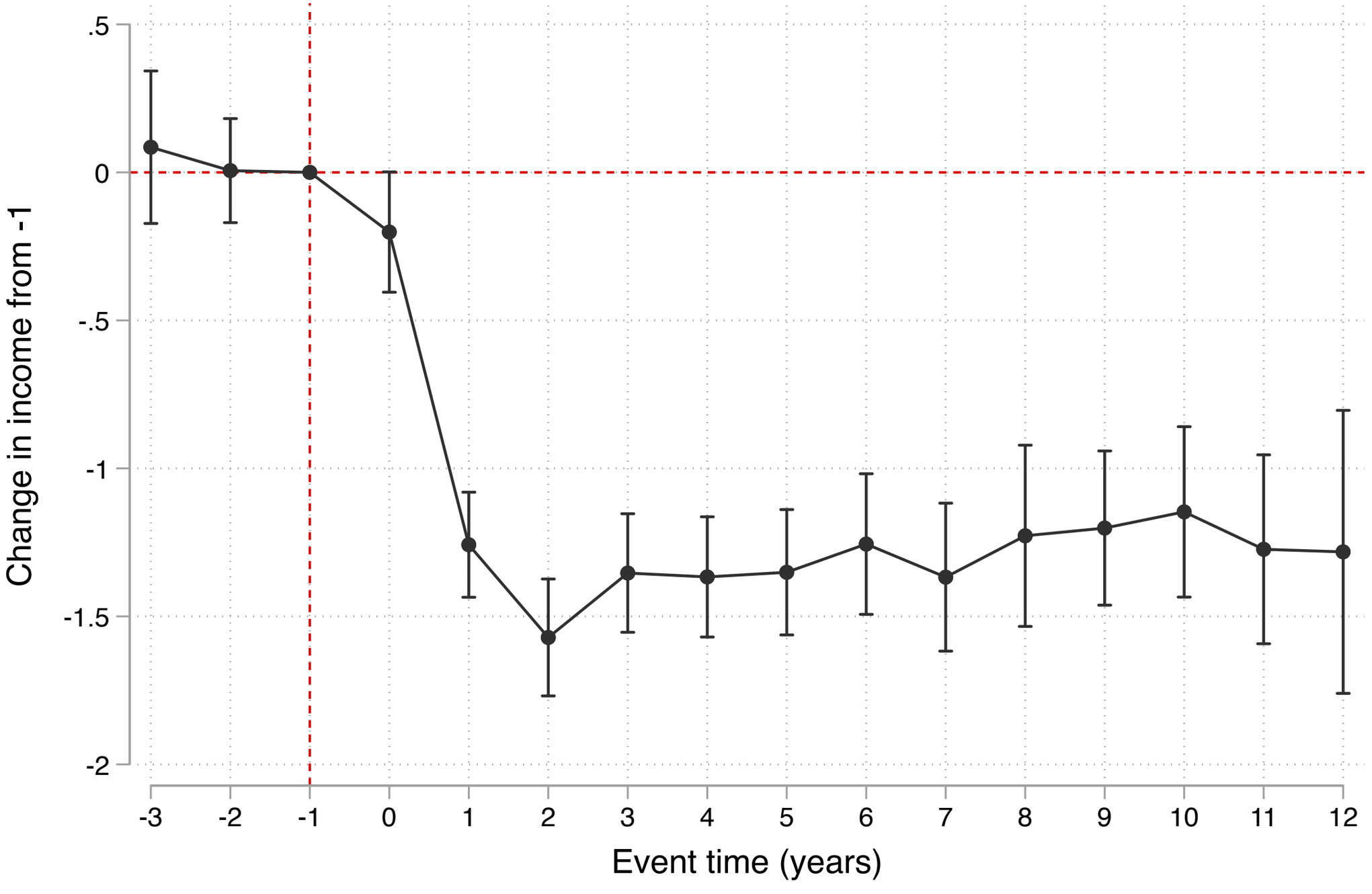


FIGURE 3. CHILD PENALTIES IN EARNINGS IN GERMAN-SPEAKING COUNTRIES

Kleven, Henrik, Camille Landais, Johanna Posch, Andreas Steinhauer, and Josef Zweimüller. 2019. “Child Penalties across Countries: Evidence and Explanations.” *AEA Papers and Proceedings* 109:122–26.

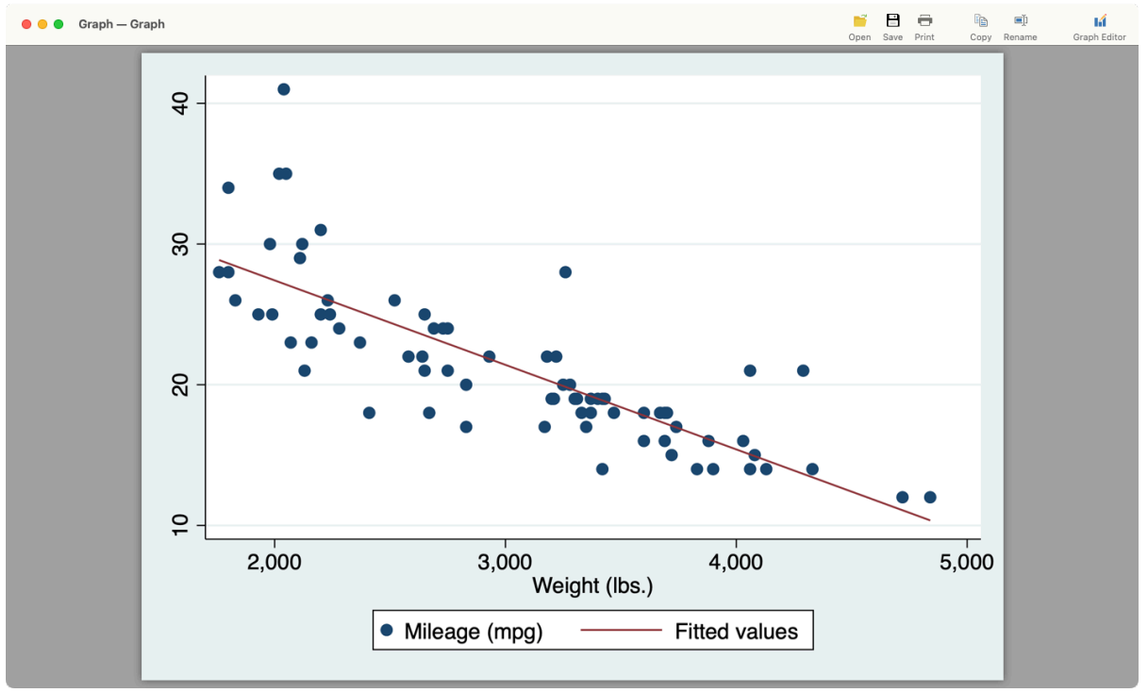
演習：イベントスタディ

5_3_eventstudy.doを開き、コードを順に実行しよう

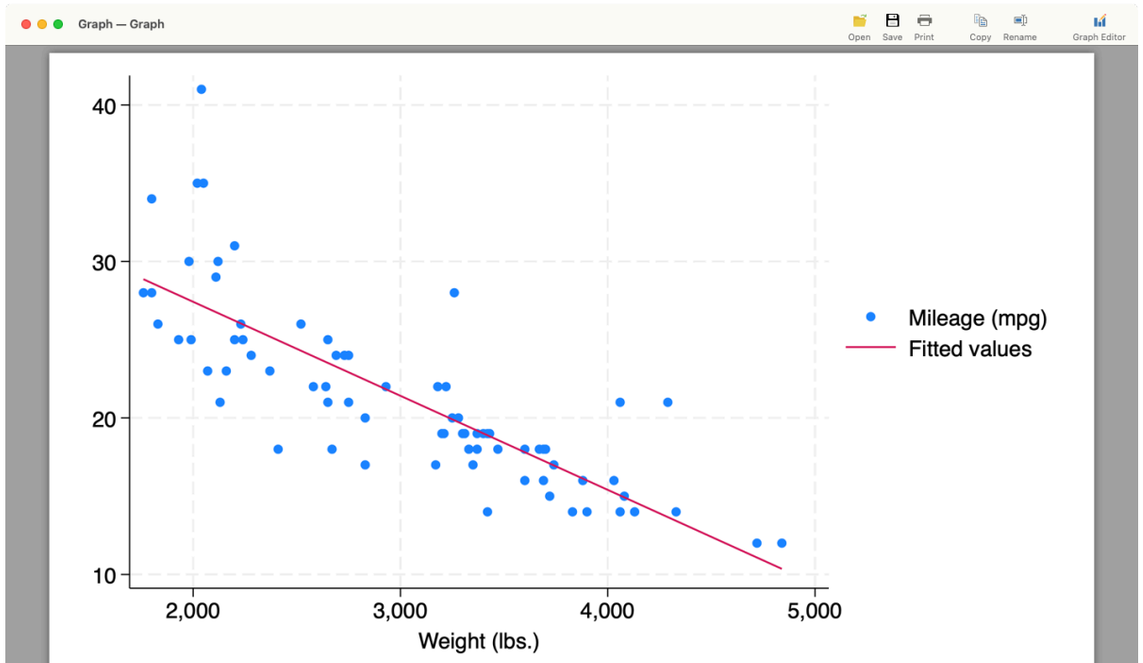


Stata 18のグラフデザイン変更

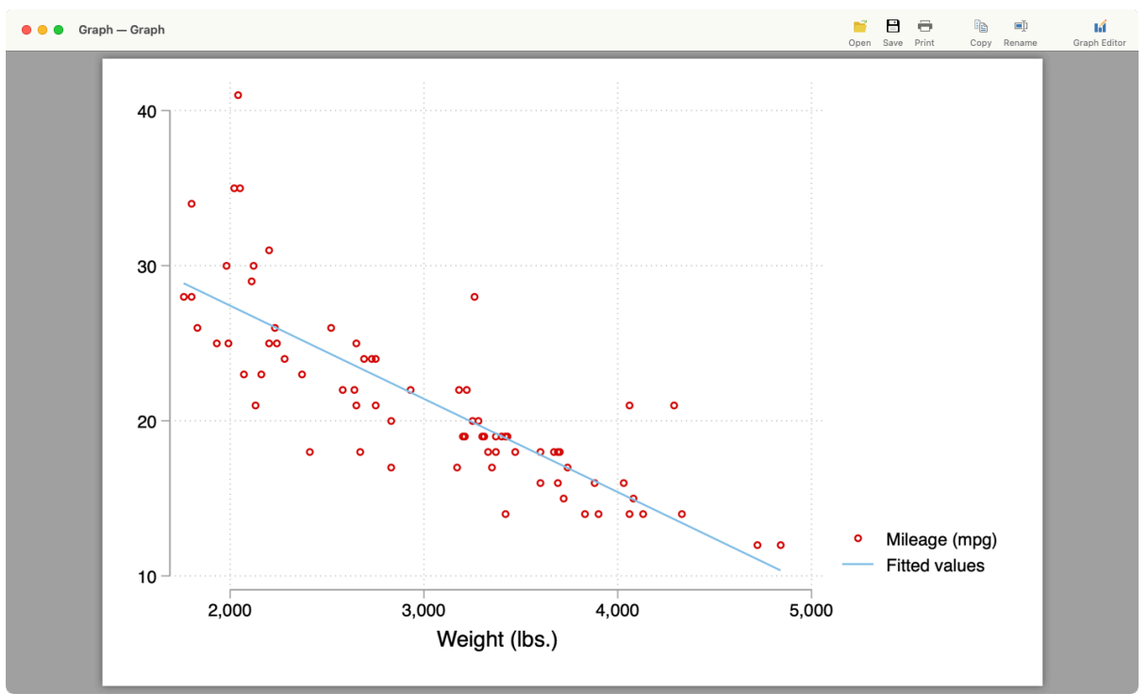
Stata 17まで



Stata 18以降

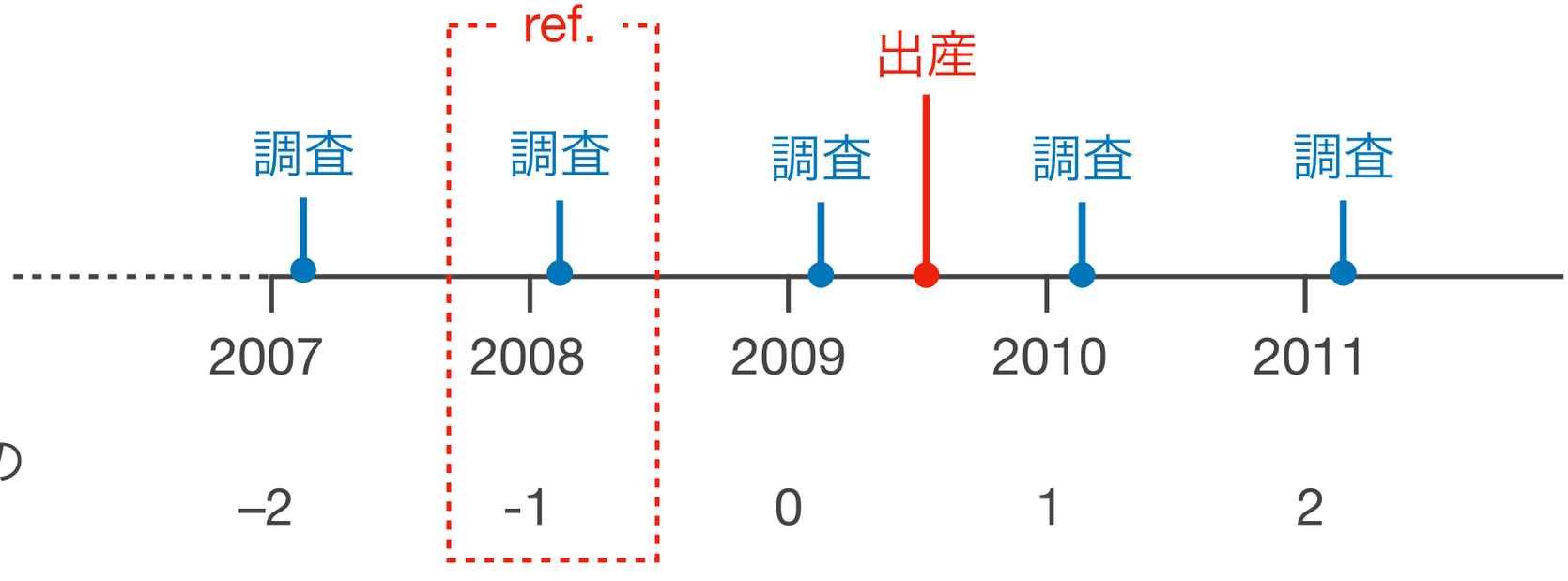


scheme(cleanplots)



イベントの発生時点の定義と参照時点の選択

今回の定義で2009年6月に第一子を産んだ対象者を考えると：



どの時点を比較対象時点として、どこと比較するのかは自明でない

「0年」時点には、間もなく出産するような人もいれば、まだ妊娠もしていないような人も含まれている。

パネル調査データでイベントの効果推定をする際の注意点

イベントからの経過年数はどのように分布しているか？

- 今回のデータで出産後12年目を経験しているのは、2007年に出産した人だけ
- 出産後数年以上を経過した観察は分析結果を表示しないなど、一般化の範囲を限定するとよいかもしれない

適切な統制群（比較対象）を選んでいるか？

- 今回は「2007–2019年に第一子を出産した人」と「2007–2019年に第一子を出産しなかった人（調査途中で脱落したので出産したかわからない人含む）」を比較した。「2007年以前に第一子を出産した人」はサンプルに含まない
- イベントが起こりうる対象を意識する（c.f. 転職の効果分析の妥当な対象は？）

Miller, Douglas L. 2023. “An Introductory Guide to Event Study Models.” *Journal of Economic Perspectives* 37(2):203–30.

ランダム効果モデル

固定効果モデル Fixed-effects model（再掲）

次のように、個人を表す項 u_i をモデルに含める：

$$Y_{it} = \beta_1 X_{it1} \cdots + \beta_k X_{itk} + u_i + e_{it}$$

- u_i は独立変数を統制したうえでの残差r_{it}（→Pooled OLS）の個人内平均を表す
- 時間不変の独立変数は u_i と区別できないため、推定から除外される。個人がわかれば必ずその値もわかる = 完全な共線性がある
- 係数 β_j は、他の時間可変の独立変数と時間不変の個人要因を統制したうえで、独立変数X_{itj}が1単位高いときに従属変数 Y_{it} がどれだけ高いかを表す。**個人内効果 within-effect/within-estimator**とも言われる

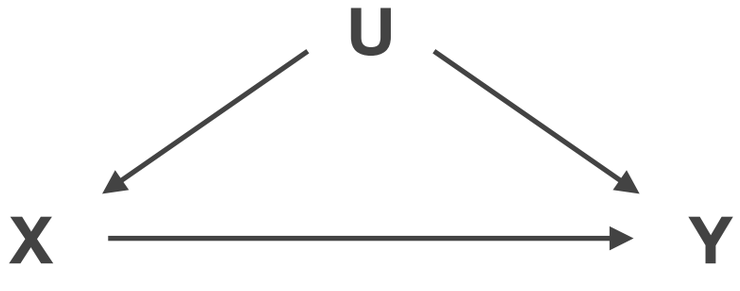
ランダム効果（変量効果）モデル Random-effects model

以下のようにPOLSの残差を個人間残差と個人内残差に分けた式を考える：

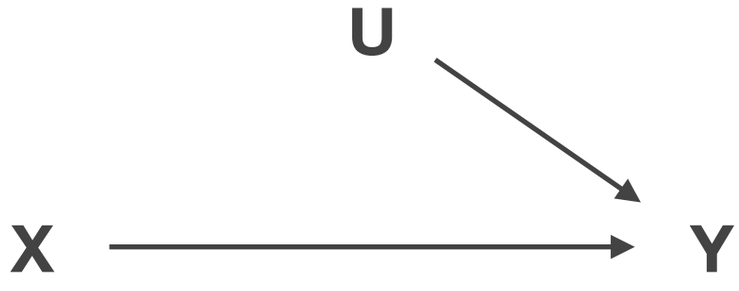
$$Y_{it} = \beta_0 + \beta_1 X_{it1} \cdots + \beta_k X_{itk} + u_i + e_{it}, \quad u_i \sim N(0, \sigma_u^2),\ e_{it} \sim N(0, \sigma_e^2)\,.$$

- u_i は残差であり、$\mathrm{Cov}(X_{itj}, u_i) = 0$ と仮定される（下図）。当然、独立変数と残差に相関があれば、係数にはバイアスが生じる

- 時間不変の要因をすべて統制する固定効果モデルとは異なり、時間不変の変数をモデルに含める余地が残っている



固定効果モデル



ランダム効果モデル

ランダム効果モデルに似たモデルのさまざまな呼び名

階層線形モデル Hierarchical linear model

マルチレベルモデル Multilevel model

ランダム切片／傾きモデル Random-intercept/random-slope model

成長曲線モデル Growth curve model

いずれも切片（あるいは傾き）に**独立変数とは相関しないと仮定されたばらつきを認める**という点でランダム効果モデルと共通している

「ランダム効果モデルと成長曲線モデルは同じなのか（違うのか）」という質問への回答は「場合による」。式を見てはじめて判断できる

ランダム効果モデルの位置づけ

REはquasi-demeaningともよばれる。詳細な証明（Wooldridge 2010）は省略するが……

$$Y_{it} - \theta\overline{Y}_i = \beta_0(1-\theta) + \beta_1(X_{it1} - \theta\overline{X}_{i1}) + \cdots + \beta_k(X_{itk} - \theta\overline{X}_{ik}) + (1-\theta)u_i + (e_{it} - \theta\overline{e}_i),$$

$$\text{ただし } \theta = 1 - \sqrt{\frac{\sigma_e^2}{\sigma_e^2 + T\sigma_u^2}}$$

POLS **RE** **FE**

$\theta = 0$ ⟷ $\theta = 1$

残差の個人内平均 u_i のばらつきσ_u^2が小さい
（人によって平均水準が異ならない）

残差の個人内平均 u_i のばらつきσ_u^2が大きい
（人によって平均水準が異なる）

REにおける時間可変の変数の係数はPOLSとFEの間にあり、σ_u^2が小さければ

POLSに近い推定値となり、大きければFEに近い推定値となる。

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data, 2nd Edition*. The MIT Press.

ランダム効果モデルと固定効果モデル

パネルデータを分析する最大の動機は、観察期間中に変化しない個人の要因を除いたうえで、独立変数の変化の効果（個人内効果）を推定すること

→この目的であれば、常にREよりもFEのほうが望ましい

にもかかわらずREを選択しなければいけないのはどんな場面か？

- 時間可変の独立変数の効果に関心があるが、（観察期間が短い、分散が少ないなどの理由から）個人内分散が少なく満足な推定ができない。それでもなお、ちょっとだけでも「効果」に近い推定値を得たい
- 時間不変の独立変数の効果に関心がある（普通のクロスセクションのデータと同じ分析がしたいだけ）

おすすめしない古のモデル選択法：Hausman検定

Hausman (1978) 検定：REとFEの係数が等しい（H_0: $\beta^{FE} = \beta^{RE}$）という帰無仮説を検定し、帰無仮説が棄却されればFE、棄却されなければREを用いる

この方法でモデル選択をすべきではない

- H_0 が棄却されなかったからといってREが一致推定量（よりバイアスが少ないという意味）を与えるということにはならない
- データの限界（観察期間が少ないなど）によってFEの係数が安定して推定できていないだけの可能性がある

Wooldridge, Jefrey. 2019. *Introductory Econometrics, 7th Edition*. Cengage Learning.
西山慶彦・新谷元嗣・川口大司・奥井亮，2019，『計量経済学』有斐閣.

ランダム効果モデルとPooled OLS

FEではなくREを使うという条件のもとで、REとPOLSではどちらを使えばよいのだろうか？

→常にPOLSよりもREを使うべき

REは残差のうちu_iに由来する部分を（不完全ではあれ）多少統制している。そのため、POLSよりも一致推定量に近く、かつ有効性（係数のばらつき = 標準誤差が小さいという意味）の高い推定値が得られる

*ただし、非線形モデルを使っていて、かつ集団レベルの予測確率を求めることを関心とする場合にはREでなくPooledモデルを選択することがある

Wooldridge, Jefrey. 2019. *Introductory Econometrics, 7th Edition*. Cengage Learning.

おすすめしない古のモデル選択法：BP検定

Breusch-Pagan (1980) 検定：REにおいて $H_0: \sigma_u^2 = 0$ という帰無仮説を検定し、棄却されればRE、棄却できなければPOLSを用いる

この方法でモデル選択をするべきではない

- H_0 が棄却されなかったからといってPOLSが一致推定量を与えるということにはならない
- モデルの特定化の誤りによって帰無仮説が棄却されないだけの可能性がある

Wooldridge, Jefrey. 2019. *Introductory Econometrics, 7th Edition*. Cengage Learning.

演習：ランダム効果モデルの推定

6_1_randomeffect.doを開き、コードを順に実行しよう

```
Random-effects GLS regression                   Number of obs     =     12,643
Group variable: id                              Number of groups  =      1,575

R-squared:                                      Obs per group:
     Within  = 0.0288                                         min =          2
     Between = 0.2246                                         avg =        8.0
     Overall = 0.1376                                         max =         13

                                                Wald chi2(8)      =     433.64
corr(u_i, X) = 0 (assumed)                      Prob > chi2       =     0.0000

                                    (Std. err. adjusted for 1,575 clusters in id)
```

swb	Coefficient	Robust std. err.	z	P>|z|	[95% conf.	interval]
status						
Non-regular employment	-.1805531	.0398703	-4.53	0.000	-.2586975	-.1024087
age	-.0013678	.0022194	-0.62	0.538	-.0057177	.002982
marstat						
Married	.562682	.0344864	16.32	0.000	.4950899	.6302742
Separated/divorced	.0296589	.0759129	0.39	0.696	-.1191278	.1784455
hincome	.0160392	.0025092	6.39	0.000	.0111213	.0209571
cohort						
1971-75	.0290965	.046894	0.62	0.535	-.062814	.121007
1976-80	.0198208	.0586287	0.34	0.735	-.0950894	.134731
1981-86	.1385114	.0598788	2.31	0.021	.0211512	.2558716
_cons	3.15558	.1014709	31.10	0.000	2.956701	3.354459
sigma_u	.60696986					
sigma_e	.62155446					
rho	.48813003	(fraction of variance due to u_i)				

Within-between random-effects model* (REWB)

ランダム効果モデルに時間可変の独立変数とその個人内平均を含めることで、時間可変の独立変数については個人内効果を求めつつ、時間不変の部分の効果（個人間効果）を求めることができる。

$$Y_{it} = \beta_0 + \beta_1^w X_{it1} \cdots + \beta_k^w X_{itk} + \beta_1^b \overline{X}_{1i} \cdots + \beta_k^b \overline{X}_{ik} + \gamma_1 Z_{i1} + \cdots + \gamma_k Z_{ik} + u_i + e_{it},$$
$$u_t \sim N(0, \sigma_u^2)\,.$$

または

$$Y_{it} = \beta_0 + \beta_1^w (X_{it1} - \overline{X}_{i1}) \cdots + \beta_k^w (X_{itk} - \overline{X}_{ik}) + \beta_1^b \overline{X}_{1i} \cdots + \beta_k^b \overline{X}_{ik}$$
$$+ \gamma_1 Z_{i1} + \cdots + \gamma_k Z_{ik} + u_i + e_{it}, \quad u_i \sim N(0, \sigma_u^2)\,.$$

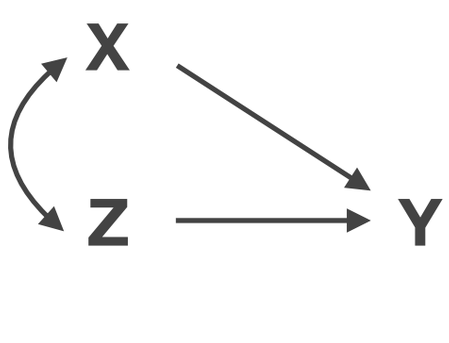
β_j^wは個人内効果を表す係数、β_j^bは個人間効果を表す係数

*1つ目の式はMundlak model や Correlated random effects model、2つめの式は Hybrid model（Allison 2009）などともよばれることがある。Allison, P. D. (2009). Fixed Effects Regression Models. Sage.

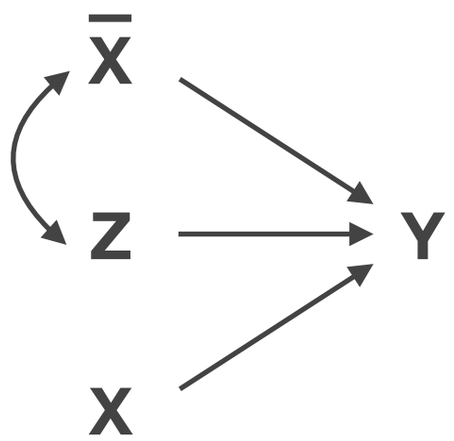
REWBをいつ使うか

パネルデータを使っていて、かつ、時間不変の変数の係数に関心があるならば、REWBはよい選択肢かもしれない

REの想定



REWBの想定



時間不変の変数Zは、時間不変という定義上、時間可変の変数のうち時間不変の部分とのみ相関するはずなので、REWBのほうが理に叶っているかもしれない

Andreß, H.-J., Golsch, K., & Schmidt, A. W. (2013). *Applied Panel Data Analysis for Economic and Social Surveys.* Springer.

REWBをいつ使うか

REWBはFEの特徴もREの特徴を併せ持つ（より一般的な）モデルで、これを使うことをデフォルトにすることを推奨する一派もあるが（Bell and Jones 2015; Bell et al. 2019）、「併せ持つ」ことによってはじめて答えられる問いはあるのかと考えると、よくわからない

知りたい問いが個人内の変化の効果に関するものならわざわざ複雑なモデルを使う必要はないので、FEを使うとよい

REWBを使いたい衝動に駆られたら、自分の問いが何かを改めて考えよう

Bell, Andrew, and Kelvyn Jones. 2015. “Explaining Fixed Effects: Random Effects Modeling of Time-Series Cross-Sectional and Panel Data.” *Political Science Research and Methods* 3(1):133–53.
Bell, Andrew, Malcolm Fairbrother, and Kelvyn Jones. 2019. “Fixed and Random Effects Models: Making an Informed Choice.” *Quality and Quantity* 53(2):1051–74.

演習：within-between random effects modelの推定

6_2_wbre.doを開き、コードを順に実行しよう

Variable	model
swb	
R__cohort_d2	0.1109
	0.0537
	0.0389
R__cohort_d3	0.1861
	0.0836
	0.0260
R__cohort_d4	0.3671
	0.1100
	0.0008
W__status_d2	-0.1188
	0.0456
	0.0092
W__age	-0.0001
	0.0024
	0.9600
W__marstat_d2	0.4902
	0.0472
	0.0000
W__marstat_d3	0.0211
	0.0883
	0.8114
W__hincome	0.0076
	0.0028
	0.0076
D__status_d2	-0.1478
	0.0877
	0.0919

カテゴリ変数を従属変数にする

線形確率モデル Linear probability model を使う

最もシンプルな方法は、これまでの連続変数を従属変数とする固定効果モデルをそのまま適用すること

$$Y_{it} = \beta_1 X_{it1} + \cdots + \beta_k X_{itk} + u_i + e_{it}$$

係数β_jは、独立変数X_{itj}が1ポイント高いと、従属変数が1をとる確率（割合）が何ポイント高いのかを表す

固定効果モデルで2値変数を扱う場合にはわりとデファクトスタンダードとして使われている印象

よく言われている線形確率モデルの注意点（Mood, 2010）

1. 予測値が確率の定義上あり得ない数値（0未満、あるいは1より大きい）になることがある

   →普通の回帰分析でもこういうことはある

2. 残差が正規分布しない（不均一分散）ため標準誤差にバイアスが生じる

   →ロバスト標準誤差（頑健標準誤差）を使うことで対処可能

3. **関数型の誤り**：もし真の関係が非線形――従属変数が1をとる確率が低い個人と中程度の個人で、ある独立変数が1単位増えることによる確率の増加量が異なるなど――のであれば、変数の効果を正しく推定できない

Mood, Carina. 2010. “Logistic Regression: Why We Cannot Do What We Think We Can Do, and What We Can Do about It.” *European Sociological Review* 26(1):67–82.

ロジスティック回帰分析 / ロジットモデル

二値変数を従属変数とするときには、ロジットモデル（あるいはプロビットモデル）がしばしば用いられる。ロジットモデルは次のように表記される：

$$\log \frac{\Pr(Y_{it}=1)}{1-\Pr(Y_{it}=1)} = \beta_0 + \beta_1 X_{it1} + \cdots + \beta_k X_{itk}$$

または

$$\Pr(Y_{it}=1) = \frac{\exp(\beta_0 + \beta_1 X_{it1} + \cdots + \beta_k X_{itk})}{1+\exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_{itk})}$$

係数β_jは、X_{itj}が1単位増加したときの従属変数の対数オッズの増加量を示す。

潜在変数による定式化

$$y_{it}^* = \beta_0 + \beta_1 X_{it1} + \cdots + \beta_k X_{itk} + r_{it}\,, \quad y_{it} = \begin{cases} 1 & \text{if } Y_{it}^* > 0 \\ 0 & \text{otherwise} \end{cases}$$

ただしr_{it}は平均0、分散$\pi^2/3$のロジスティック分布にしたがう。残差が固定されているということから、以下の解釈上の注意を要する：

- 異なるサンプルからなるモデル間で係数の大きさを比較できない
- 異なる独立変数を含むモデル間で係数の大きさを比較できない

ロジットモデルの結果を解釈するときには、平均限界効果 Average marginal effectなどを併せて使うことが強く推奨される

Mood, Carina. 2010. “Logistic Regression: Why We Cannot Do What We Think We Can Do, and What We Can Do about It.” *European Sociological Review* 26(1):67–82.

Long, J. Scott, and Jeremy Freese. 2014. Regression Models for Categorical Dependent Variables Using Stata, Third Edition. Stata Press.

Mize, Trenton D., Long Doan, and J. Scott Long. 2019. “A General Framework for Comparing Predictions and Marginal Effects across Models.” *Sociological Methodology* 49(1):152–89.

固定効果／ランダム効果二項ロジットモデル

固定効果ロジットモデル Fixed-effects logit

$$\log \frac{\Pr(Y_{it}=1)}{1-\Pr(Y_{it}=1)} = \beta_1 X_{it1} + \cdots + \beta_k X_{itk} + u_i$$

係数は観察期間中は変化しない個人要因を統制したうえでの独立変数の効果（個人内効果）を表すものと解釈できる

観察期間中に従属変数にまったく変化がない個人は分析から除外される

ランダム効果ロジットモデル Random-effects logit

$$\log \frac{\Pr(Y_{it}=1)}{1-\Pr(Y_{it}=1)} = \beta_1 X_{1it} + \cdots + \beta_k X_{kit} + u_i, \quad u_i \sim N(0,\sigma_u^2)$$

u_i は独立変数と相関しない残差と仮定される。

固定効果／ランダム効果ロジットモデルの問題

対数オッズの解釈のしにくさ

固定効果／ランダム効果ロジットモデルでは平均限界効果（Average marginal effects）を計算できないため、係数を対数オッズとして解釈するしかない。そのため、実質的な効果の大きさをつかみにくい

潜在変数の分散変化

固定効果ロジットモデルやランダム効果ロジットモデルはPooled logitよりも従属変数をよく予測する

しかし潜在変数の残差e_{it}は常に$\pi^2/3$で固定されているため、潜在変数の分散が大きくなり、見かけ上係数の絶対値が大きくなる

演習：線形確率モデルとロジットモデル

6_3_binary.doを開き、コードを順に実行しよう

	LPM-pooled	LPM-FE	LPM-RE	Logit-pooled	Logit-FE	Logit-RE
main						
-3	0.035*	0.010	0.016	0.496	0.312	0.524
	(0.017)	(0.019)	(0.017)	(0.276)	(0.404)	(0.420)
-2	0.025	0.004	0.009	0.279	0.044	0.256
	(0.019)	(0.018)	(0.017)	(0.227)	(0.326)	(0.301)
0	-0.156***	-0.138***	-0.143***	-0.870***	-1.614***	-1.568***
	(0.029)	(0.028)	(0.026)	(0.165)	(0.258)	(0.251)
1	-0.344***	-0.333***	-0.338***	-1.621***	-3.486***	-3.300***
	(0.031)	(0.030)	(0.028)	(0.166)	(0.286)	(0.270)
2	-0.314***	-0.288***	-0.296***	-1.499***	-2.987***	-2.878***
	(0.032)	(0.030)	(0.028)	(0.167)	(0.284)	(0.262)
3	-0.291***	-0.268***	-0.272***	-1.400***	-2.767***	-2.656***
	(0.032)	(0.030)	(0.028)	(0.172)	(0.288)	(0.268)
4	-0.246***	-0.226***	-0.230***	-1.212***	-2.367***	-2.261***
	(0.034)	(0.032)	(0.030)	(0.180)	(0.298)	(0.292)
5	-0.214***	-0.189***	-0.193***	-1.080***	-1.963***	-1.873***
	(0.035)	(0.032)	(0.030)	(0.188)	(0.308)	(0.295)
6	-0.165***	-0.148***	-0.151***	-0.865***	-1.643***	-1.521***
	(0.036)	(0.033)	(0.031)	(0.197)	(0.326)	(0.309)
7	-0.153***	-0.122***	-0.127***	-0.807***	-1.425***	-1.295***
	(0.039)	(0.034)	(0.033)	(0.210)	(0.341)	(0.334)
8	-0.132**	-0.083*	-0.090**	-0.703**	-1.064**	-0.920**
	(0.041)	(0.036)	(0.034)	(0.225)	(0.358)	(0.353)
9	-0.107*	-0.057	-0.063	-0.585*	-0.860*	-0.703
	(0.045)	(0.039)	(0.036)	(0.248)	(0.393)	(0.388)

その他のStataのコマンド

さまざまな場面に応じたコマンドが準備されている

- 固定効果／ランダム効果**多項ロジット**モデル：`xtmlogit`（Stata 17から）
- 固定効果**順序ロジット**モデル：`feologit`（Baetschmann et al., 2020）
- ランダム効果**順序ロジット**モデル：`xtologit`
- 固定効果／ランダム効果**ポワソン回帰**：`xtpoisson`
- ランダム効果**トービットモデル**：`xttobit`

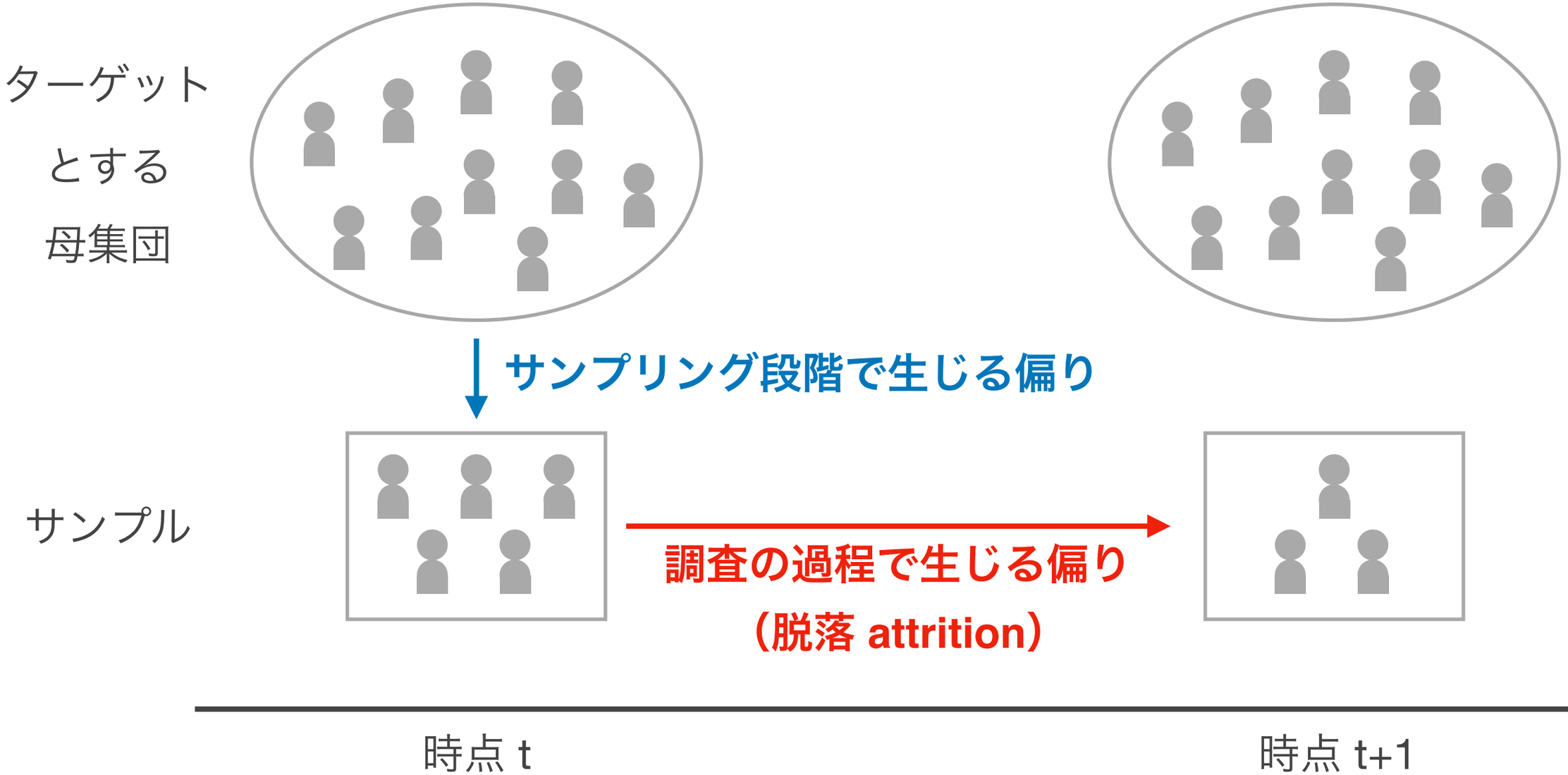
これらはいずれも線形回帰のように単純ではなく、推定も収束しないことが多く、あまり使われない（自分は論文でこれまでに一度も見たことがない）。

Baetschmann, G., Ballantyne, A., Staub, K. E., & Winkelmann, R. (2020). feologit: A new command for fitting fixed-effects ordered logit models. *The Stata Journal*, *20*(2), 253–275.

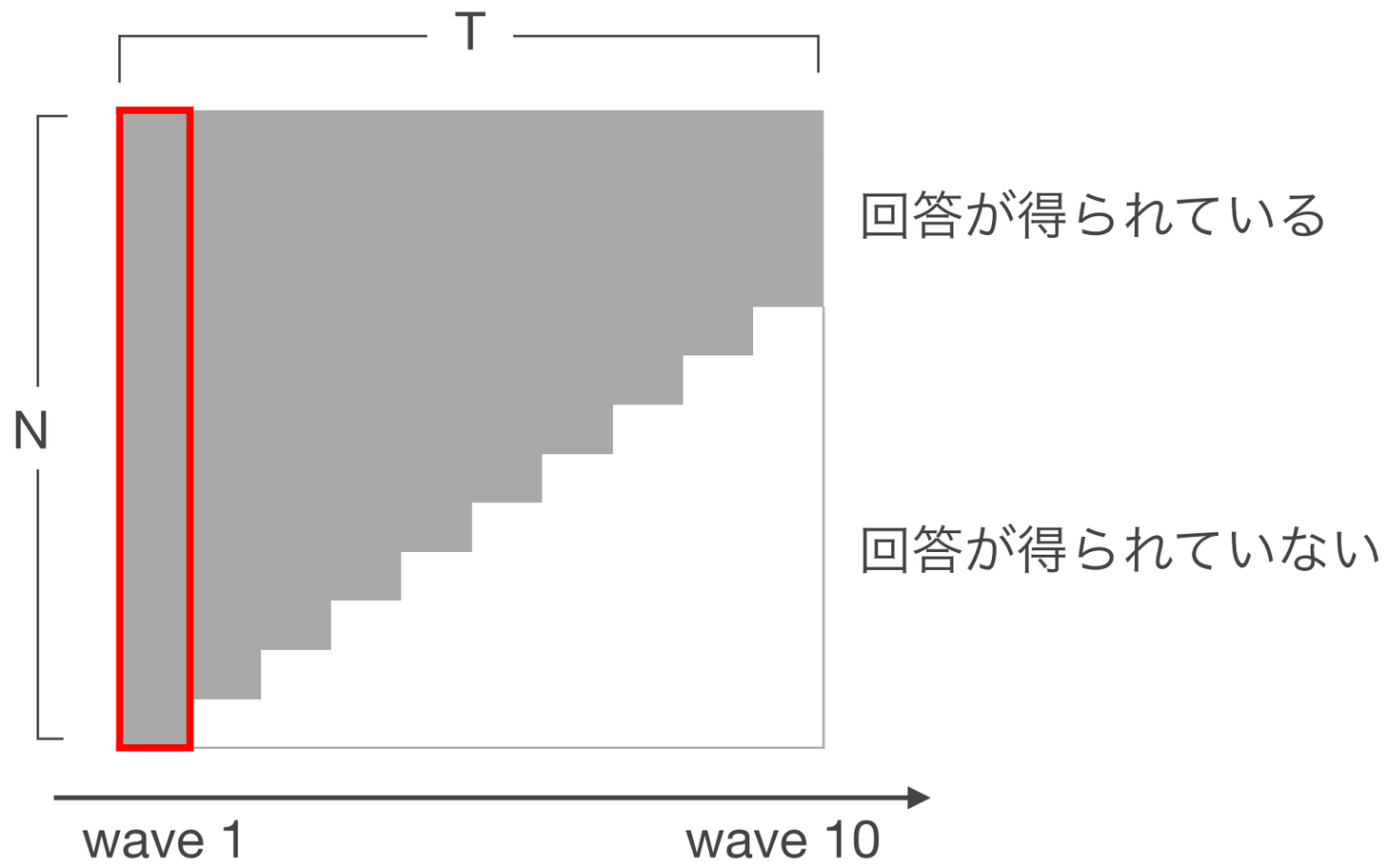
脱落の問題と対処

パネル調査データにおける偏りの出処

パネル調査データにおける母集団からの偏りは以下の2つの過程で生じる：



脱落がなぜ問題となるのか



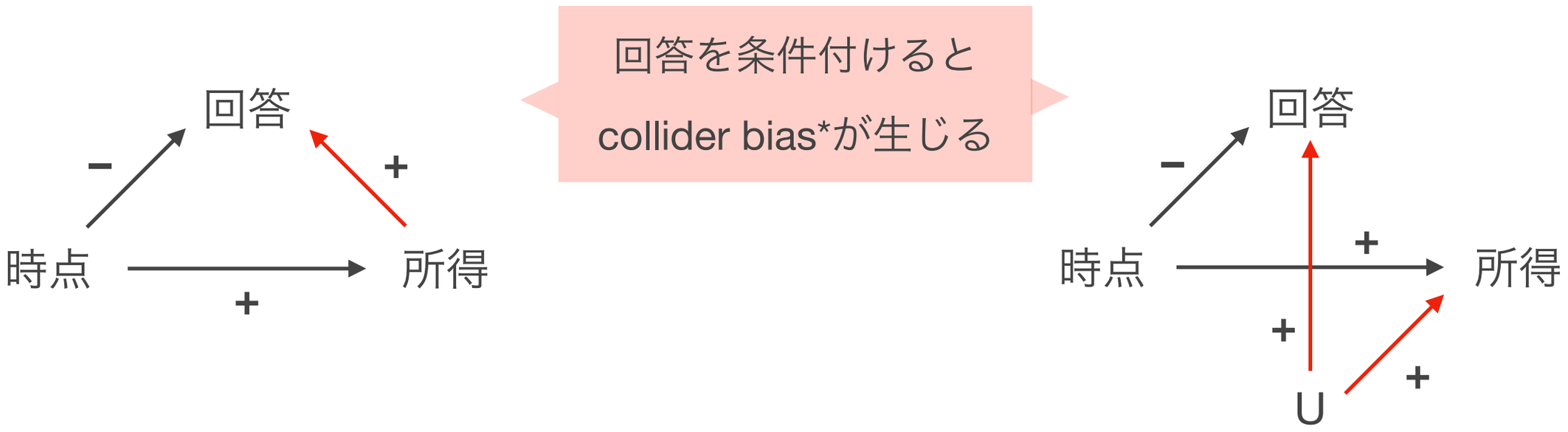
調査を重ねるにつれて、回答が得られているサンプルと回答が得られていないサンプルが生まれる

回答が得られるかどうかが系統的に異なると、サンプルはもともとのターゲット母集団から乖離する。そうすると、サンプルから得た結果が必ずしも母集団に一般化できなくなるかもしれない

脱落が問題になるとき：記述

「1966–86年生まれコーホートの所得は2007年から2019年にかけてどのように推移したのか？」という**記述的問い**であれば、脱落はほぼ常に問題

- 所得が回答確率に影響する。たとえば所得が低い人ほど継続回答しにくいならば、時間が経つにつれて所得（の平均値など）は過大推計となる
- 回答確率と所得の両者と相関する何らかの要因がある。この場合、所得は過大または過小推計となる

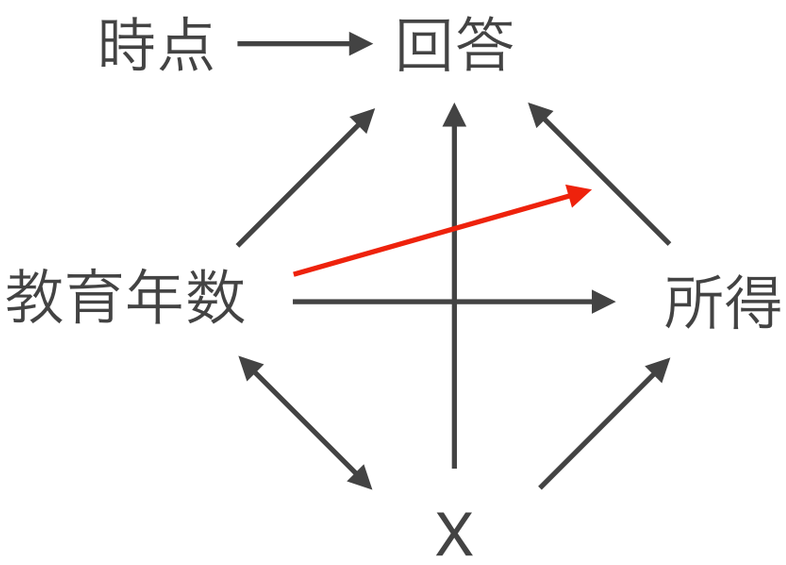


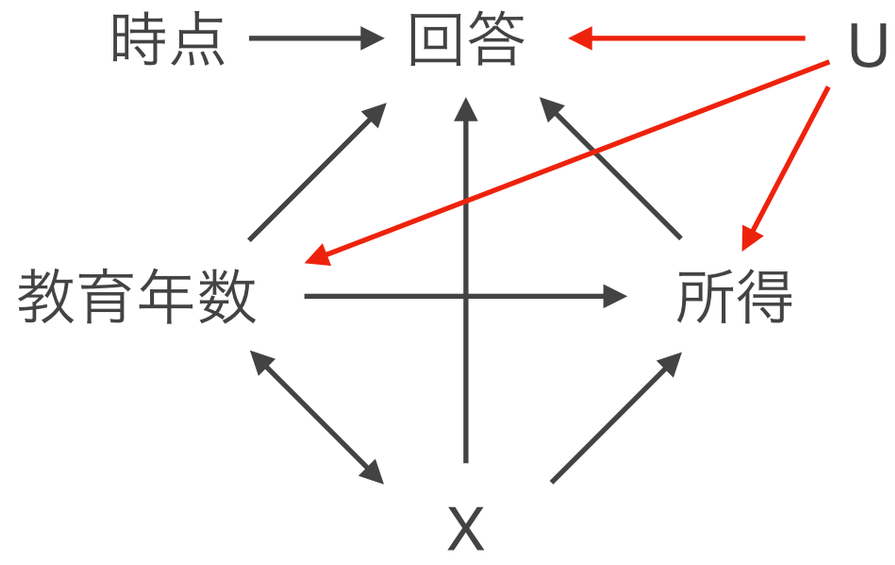
Elwert, F., & Winship, C. (2014). Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable. *Annual Review of Sociology*, *40*, 31–53.

脱落が問題になるとき：固定効果モデルでない回帰分析

「他の変数を統制した上で、教育年数が所得に与える影響はどの程度であるのか？」というような**回帰分析による問い**ならば、次のようなとき脱落は問題

- 他の変数を統制したうえで、学歴と所得の組み合わせが脱落確率に影響する。たとえば学歴が低くかつ所得が低い人ほど継続回答しにくい場合、他の変数を統制したうえでの教育年数の係数は過小推計となる
- 他の変数を統制したうえで、回答確率、学歴、所得のいずれとも相関する観察されない要因がある。





固定効果モデルは脱落の影響を小さくできる

時間不変の個人要因を統制する固定効果モデルは、回答確率に影響しうる時間不変の個人要因をすべて統制できる。そのため、一般に固定効果モデルは脱落の問題に対して比較的頑健

ただし、時間可変の要因が脱落に影響することについては、統制できない（e.g. 観察期間中の転居）

脱落の問題のまとめ

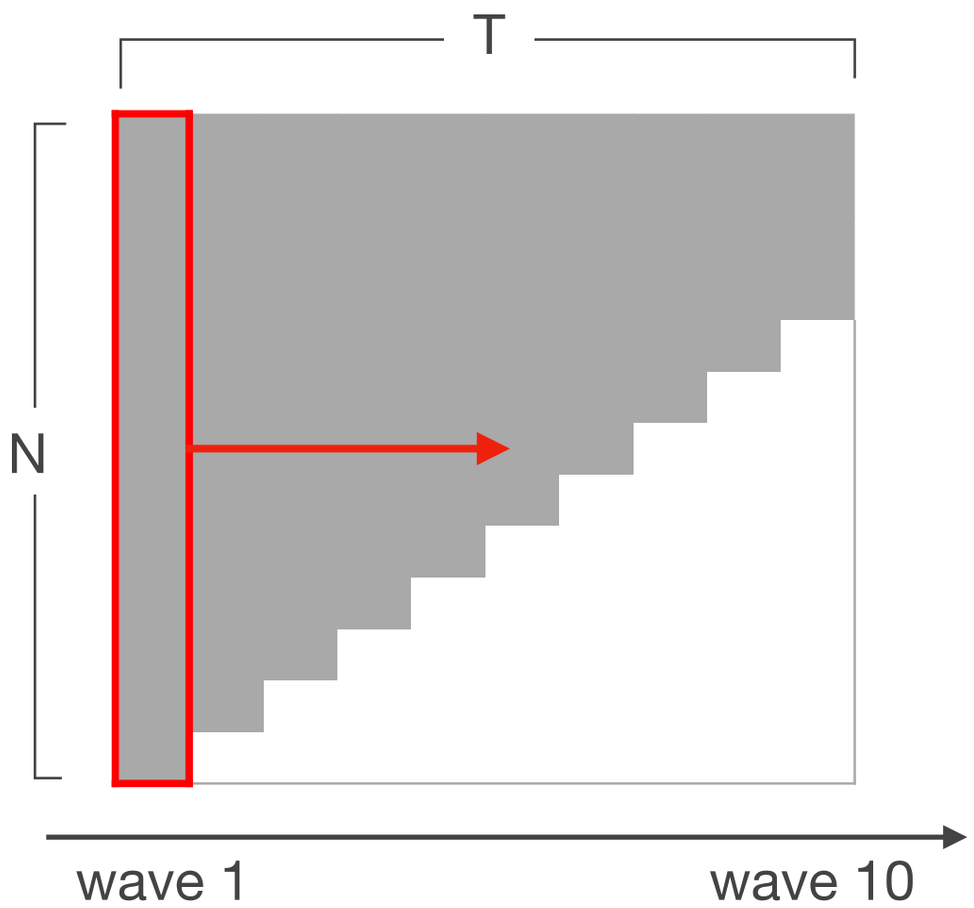
脱落が問題になるかどうかは目的に依存する

- 少ない変数での記述を目的とする場合：脱落は常に考慮すべき問題。個人内変化を伴わない集団の変化の記述が目的なら、繰り返しクロスセクション調査のほうが好ましいことも多い
- 複数の変数を用いて回帰分析／固定効果モデルを使う場合：脱落が大きく結果を左右する可能性はさほど高くない

パネル調査データにおける偏りが結果に与える影響は脱落よりもむしろ最初のサンプリング時点での歪みのほうが大きいとの指摘もある（小川 2019）

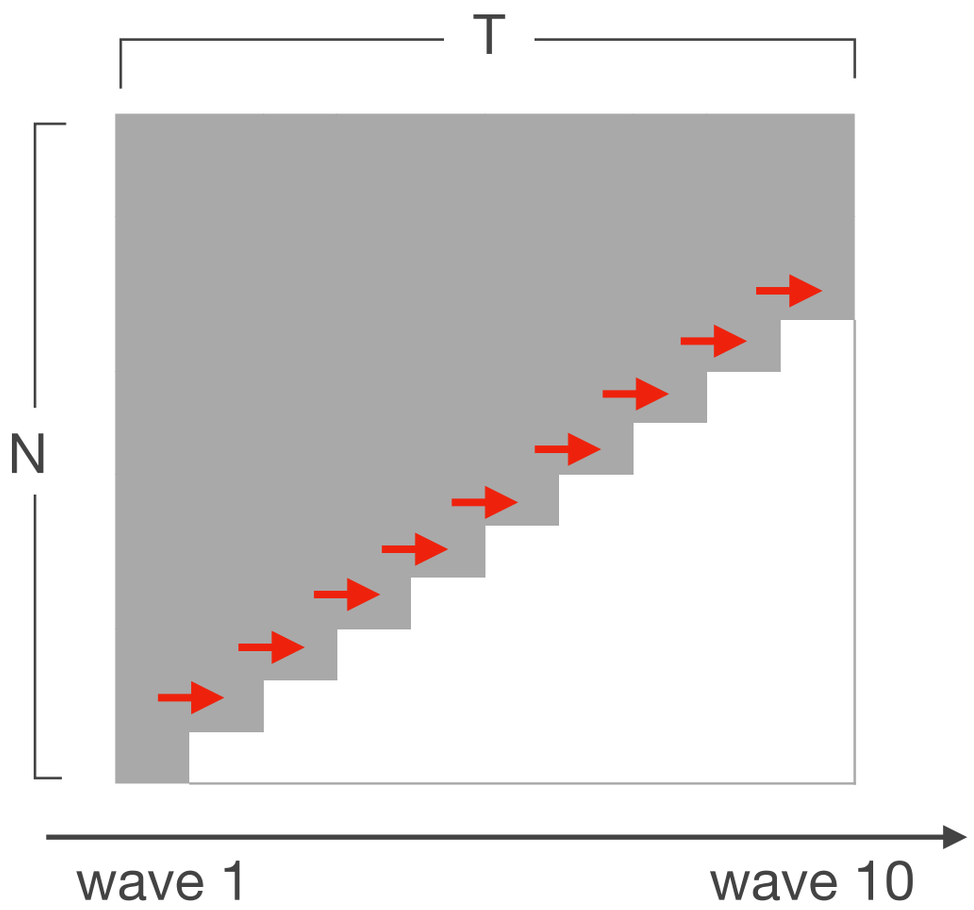
小川和孝，2019，「パネルデータにおける脱落と収入の分析：クロスセクションデータとの比較から」『年報社会学論集』32: 73–83.

脱落補正のバリエーション：wave 1への補正



1. N×Tのサンプルを作成し（→回答の複製）、各時点で回答しているか否かを従属変数、wave 1時点の独立変数で回帰するロジット／プロビットモデルを推定し、回答の予測確率を求める
2. 予測確率の逆確率を求め、回答しているサンプルにウェイトをかける

脱落補正のバリエーション：隣接waveへの補正



1. wave tに回答したという条件のもとで、wave t+1に回答するか否かを示す変数を作成し、wave t時点の独立変数で回帰するロジット／プロビットモデルを推定し、回答の予測確率を求める
2. 予測確率の逆確率を求め、回答しているサンプルにウェイトをかける

演習：脱落の予測確率計算とウェイト（wave 1への補正）

6_4_attrition.doを開き、コードを順に実行しよう

id	wave	response	sex	cohort	prob	inv_prob
8	8	1	Women	1971-75	.7311258	1.367754
8	9	1	Women	1971-75	.7046357	1.419173
8	10	0	Women	1971-75	.6927152	1.443595
8	11	0	Women	1971-75	.6821192	1.466019
9	1	1	Men	1971-75	.8017136	1
9	2	1	Men	1971-75	.8017136	1.247328
9	3	0	Men	1971-75	.7405141	1.350413
9	4	0	Men	1971-75	.6548347	1.527103
9	5	0	Men	1971-75	.6744186	1.482759
9	6	0	Men	1971-75	.6401469	1.562142
9	7	0	Men	1971-75	.619339	1.614625
9	8	0	Men	1971-75	.6070991	1.647177
9	9	0	Men	1971-75	.5997552	1.667347
9	10	0	Men	1971-75	.5728274	1.745727
9	11	0	Men	1971-75	.5618115	1.779956
10	1	1	Women	1976-80	.8138528	1
10	2	1	Women	1976-80	.8138528	1.228723
10	3	0	Women	1976-80	.7770563	1.286908
10	4	0	Women	1976-80	.7294372	1.37092
10	5	0	Women	1976-80	.7294372	1.37092
10	6	0	Women	1976-80	.7186147	1.391566
10	7	0	Women	1976-80	.7056277	1.417178

まとめ：パネル調査データの可能性

パネル調査データの最大の強みは、**個人の変化**をみられること

パネル調査データは因果関係を証明する万能のツールではないが、適切な問いと分析があれば、これまでわからなかった実態や因果関係に近づける

パネル調査データを加工するためには通常のクロスセクションのデータよりも手間が必要であるものの、基本的な手順は同じであり、決して難しくない

まとめ：明確な問いを立てることの重要性

近年の因果推論等の発展によって、データ分析はよりシンプルで分かりやすくなっている（脱魔術化している）

必要なのは高度なモデルやテクニックではなく、よいリサーチデザイン（問い）

パネル調査データは一見情報量が多く複雑なため、何をやっているのか、何をやればいいのか、わからなくなりやすい

パネルデータではじめて答えられる**明確な問いを立てられているか？どのような問いに答えたいのか？今行っている分析は、答えたい問いにとって必要なのか？**を考えることが重要

今後の学習のための参考文献

計量経済学関連

Wooldridge, Jefrey. 2019. *Introductory Econometrics, 7th Edition.* Cengage Learning.

西山慶彦・新谷元嗣・川口大司・奥井亮, 2019, 『計量経済学』有斐閣.

Huntington-Klein, Nick. 2021. *The Effect: An Introduction to Research Design and Causality*. https://theeffectbook.net/

Cunningham, Scott. 2021. *Causal Inference: The Mixtape.* Yale University Press. https://mixtape.scunning.com/（加藤真大ほか訳, 2023, 『因果推論入門――ミックステップ：基礎から現代的アプローチまで』技術評論社. ）

因果推論関連

伊藤公一朗，2017，『データ分析の力：因果関係に迫る思考法』光文社新書.

松林哲也，2021，『政治学と因果推論』岩波書店.

Morgan, Stephan and Christopher Winchip. 2015. Counterfactuals and Causal Inference: Methods and Principles for Social Research, 2nd Edition. Cambridge University Press.（落海浩訳，2024，『反事実と因果推論』朝倉書店．）

わかりやすい論文や資料など

パネルデータ分析の講義資料

Brüderl, Josef, and Volker Ludwig. 2019. “Applied Panel Data Analysis.” Retrieved February 21, 2023 (https://www.ls3.soziologie.uni-muenchen.de/studium-lehre/archiv/teaching-marterials/panel-analysis_april-2019.pdf).

イベントスタディの実践的ガイド

Miller, Douglas L. 2023. “An Introductory Guide to Event Study Models.” *Journal of Economic Perspectives* 37(2):203–30.

モデルを考えるときの指針

Lundberg, Ian, Rebecca Johnson, and Brandon M. Stewart. 2021. “What Is Your Estimand? Defining the Target Quantity Connects Statistical Evidence to Theory.” *American Sociological Review* 86(3):532–65.

今回扱わず、先の書籍で扱われていないような論点の一例

時変の変数どうしの交互作用

Giesselmann, Marco, and Alexander W. Schmidt-Catran. 2022. “Interactions in Fixed Effects Regression Models.” *Sociological Methods & Research* 51(3):1100–1127.

交差ラグ効果を考慮した因果関係の方向性の検証

Leszczensky, Lars, and Tobias Wolbring. 2022. “How to Deal With Reverse Causality Using Panel Data? Recommendations for Researchers Based on a Simulation Study.” *Sociological Methods & Research* 51(2):837–65.

今回扱わなかった他の方法

Singer, Judith D. and John B. Willett. 2003. *Applied longitudinal data analysis: Modeling change and event occurrence.* Oxford University Press.（菅原ますみ監訳，2014，『縦断データの分析I / II』朝倉書店．）

Cornwell, Benjamin. 2015. *Social sequence analysis: Methods and application, second edition.* Cambridge University Press.

Nagin, Daniel S. 2005. *Group-based modeling of development.* Harvard University Press.

Blossfeld, Hans-Peter, Gotz Rohwer, and Thorsten Schneider. 2019. *Event history analysis with Stata, second edition.* Routledge.

Rabe-Hesketh, Sophhia and Anders Skrondal. 2022. *Multilevel and longitudinal modeling using Stata, Volume I/II, fourth edition.* Stata Press.