Chapter 1 Introduction
Chapter 2 Managing Panel Data

Andreß, Hans-Jürgen, Katrin Golsch, and Alexander W. Schmidt. 2013. *Applied Panel Data Analysis for Economic and Social Surveys*. Springer.

麦山 亮太 (mugiyama@l.u-tokyo.ac.jp)

東京大学大学院人文社会系研究科
社会学専門分野博士課程

目次

0 はじめに

1 Introduction

2 Managing Panel Data

- 2.1 The Nature of Panel Data
- 2.2 The Basics of Panel Data Management
- 2.3 Three Case Studies on Poverty in Germany
- 2.4 How to Represent a Population with Panel Data?

目次

0 はじめに

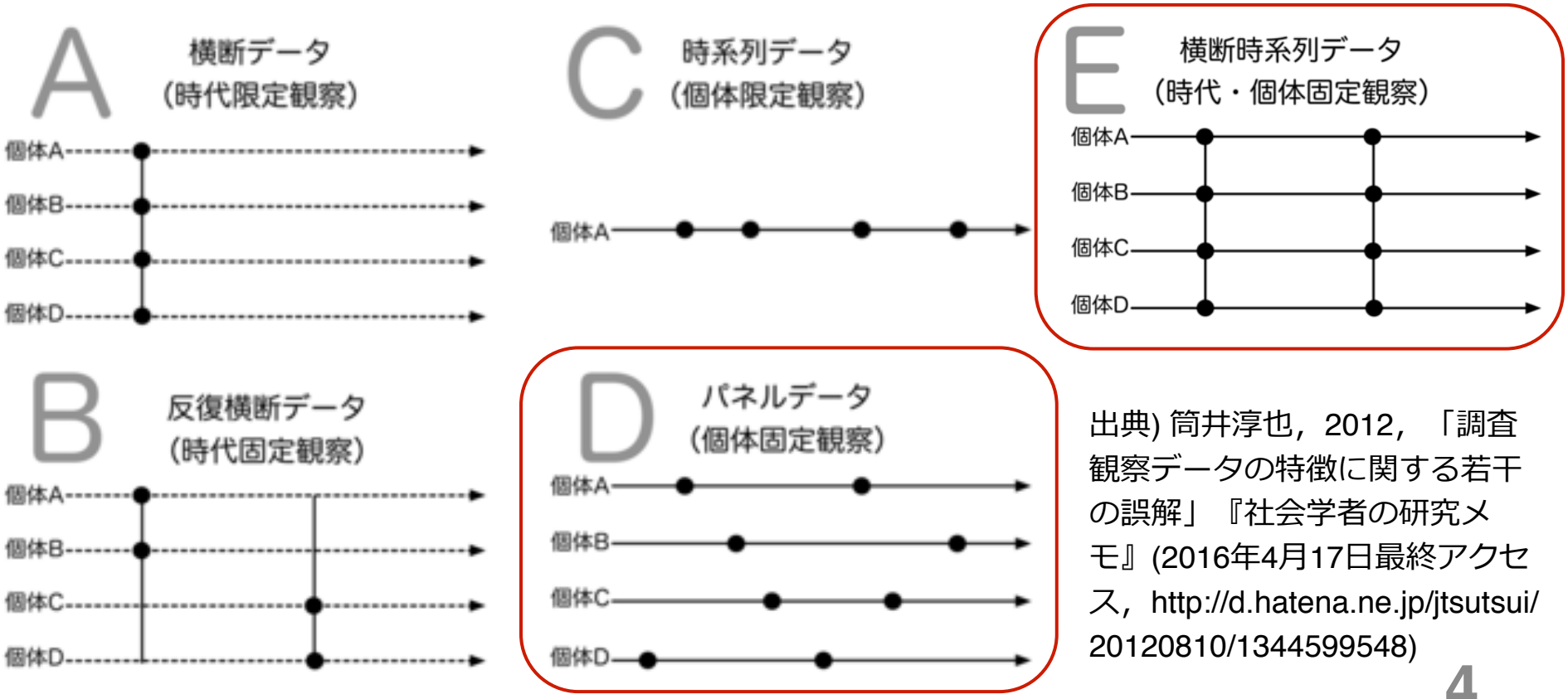
1 Introduction

2 Managing Panel Data

- 2.1 The Nature of Panel Data
- 2.2 The Basics of Panel Data Management
- 2.3 Three Case Studies on Poverty in Germany
- 2.4 How to Represent a Population with Panel Data?

パネルデータとは何か?

A research design that collects information of the same units repeatedly over time is called a ***panel***. (p.1)



出典) 筒井淳也，2012，「調査観察データの特徴に関する若干の誤解」『社会学者の研究メモ』(2016年4月17日最終アクセス，http://d.hatena.ne.jp/jtsutsui/20120810/1344599548)

国内の利用可能なパネル調査データの例

調査名	実施主体	調査概要
東京大学社会科学研究所パネル調査（JLPS）	東京大学社会科学研究所	高卒パネル（調査時高校3年生を対象、2004年~）、若年・壮年パネル（20~40歳の男女を対象、2007年~）の2種類。毎年実施され、就業、結婚、交際、意識、行動などの幅広い項目を扱う。
日本家計パネル調査（JHPS / KHPS）	慶応義塾大学パネルデータ設計・解析センター	慶應義塾パネル調査（20~69歳の男女 + 配偶者票、2004年~）と日本家計パネル調査（20~69歳の男女 + 配偶者票、2009年~）が統合された。毎年実施され、収入・支出・資産などの経済関係の項目が中心。
消費生活に関するパネル調査	家計経済研究所	調査時点で24~34歳の女性を対象として、1993年以降毎年実施。ライフイベント、家族構造、意識、経済活動などを扱う。
老研・ミシガン大学 全国高齢者パネル調査	東京都健康長寿医療センター研究所・ミシガン大学	調査時点で60歳以上の男女を対象として、1987年以降3年に1度ずつ実施。健康、ライフイベント、社会関係資本、生活状態などを扱う。

パネルデータを使う目的(1)

Halaby (2003, 2004)によれば、パネルデータを使った問いは以下の2つのタイプに整理できる

1. **因果のプロセスを明らかにする**
   …**ある状態の違いが他の状態に与える影響を知りたい**
   例) 出産は女性の賃金をどの程度低下させるか?（Budig and England 2001）失業は所得をどの程度低下させるか?（DiPrete and McManus 2000）

この流れで発展してきたモデルは、**固定効果モデルfixed-effects model, 一階差分モデルfirst difference model**と呼ばれることが多い。

パネルデータを使う目的(2)

2. **変数の軌道trajectoryを記述する**

   …時間の経過にともなう変数の値の軌道と、その軌道が集団によってどのように異なるかが知りたい

   例) 人種・性別によって、賃金の伸び方はどのように異なるか?（Rosenfeld 1980）両親の離婚タイミングによって、子どものメンタルヘルスの軌道は異なるか?（Cherlin et al. 1998）

この流れで発展してきたモデルは、**成長曲線モデルgrowth-curve model (random-effects model, multilevel model, hierarchical linear model)**と呼ばれることが多い。

* 実際には、これらの**2つの系統に属するモデルは同一の枠組みから導出できる**（Andreß et al. 2013）

Andreß et al. (2013) の特長

1. 徹底してエンドユーザー（実際に分析を行う研究者）向けに書かれている
2. 社会調査のパネルデータの分析を念頭に置いている
3. 「水準 level」に関するモデルと「変化 change」に関するモデルを分けている
4. 従属変数がカテゴリカルな場合のモデルも丁寧に説明している
5. 分析の実例が豊富

パネルデータを使った論文を読む

アメリカ・ヨーロッパの雑誌

- American Sociological Review
- American Journal of Sociology

2000年頃のパネルデータ分析の論文は比較的読みやすい印象

- Social Forces
- European Journal of Sociology

領域別

- Research in Social Stratification and Mobility（階層）
- Work and Occupations / Work, Employment & Society（労働）
- Journal of Marriage and Family（家族）
- Sociology of Education（教育）
- Demography（人口）

目次

0 はじめに

1 Introduction

2 Managing Panel Data

2.1 The Nature of Panel Data

2.2 The Basics of Panel Data Management

2.3 Three Case Studies on Poverty in Germany

2.4 How to Represent a Population with Panel Data?

パネルデータの利点

1. **個人レベルの変化の測定**
2. **年齢効果とコホート効果の区別**
3. **除外変数バイアスの統制**
4. **因果の方向性の評価**
5. 大きなサンプルサイズの確保 → 省略
6. 測定誤差の軽減 → 省略

利点1｜個人レベルの(非)変化の測定

Table 1.1 相対的貧困線付近の、子どもを持つ家族の等価可処分所得の前年から次年にかけての変化（所得中央値の40%, 50%, 60%を基準）

	t + 1時点				
t時点	< 40	40 - 50	50 - 60	60 ≦	All
< 40	**9.7**	1.8	0.8	1.3	13.6
40 - 50	2.1	2.0	1.1	1.5	6.7
50 - 60	1.1	1.4	2.1	3.0	7.5
60 ≦	1.6	1.5	3.4	**65.6**	72.2
All	14.5	6.6	7.4	71.4	100.0

- t時点の最貧困層のうち、71.3%(= 9.7 / 13.6 × 100)はt + 1時点においても最貧困の状態に留まる
- t時点の非貧困層のうち、90.9%(= 65.6 / 72.2 × 100)はt + 1時点においても非貧困の状態に留まる

利点2｜加齢効果とコホート効果の区別

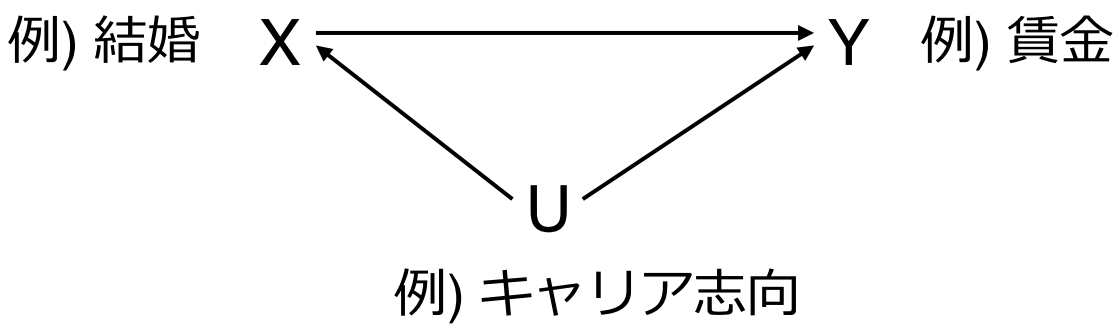
APCの**完全線形従属性**(perfect linear dependency)

$$\text{Cohort} = \text{Period} - \text{Age}$$

- パネルデータの場合、複数のPeriodが得られるため、完全線形従属性は成り立たず、Ageの効果とCohortの効果を区別できる
- Pooled cross sectionデータでも同じことができるが、パネルデータの場合は個人レベルの純粋なAgeの効果を取り出せる点で優れている

利点3｜除外変数バイアスの統制

除外変数バイアス (Omitted variable bias)
関心のある独立変数と従属変数の両者と相関する変数を統制しないことで、独立変数の係数にバイアスが生じる



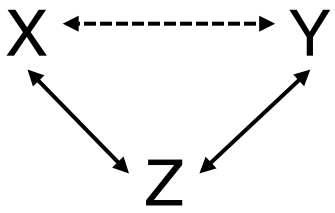
パネルデータの場合、処置前(X = 0)の状態と処置後(X = 1)の状態を比較することで、観察されない変数の一部を統制することができる

利点4｜因果の方向性の評価

共変関係

X ⟷ Y

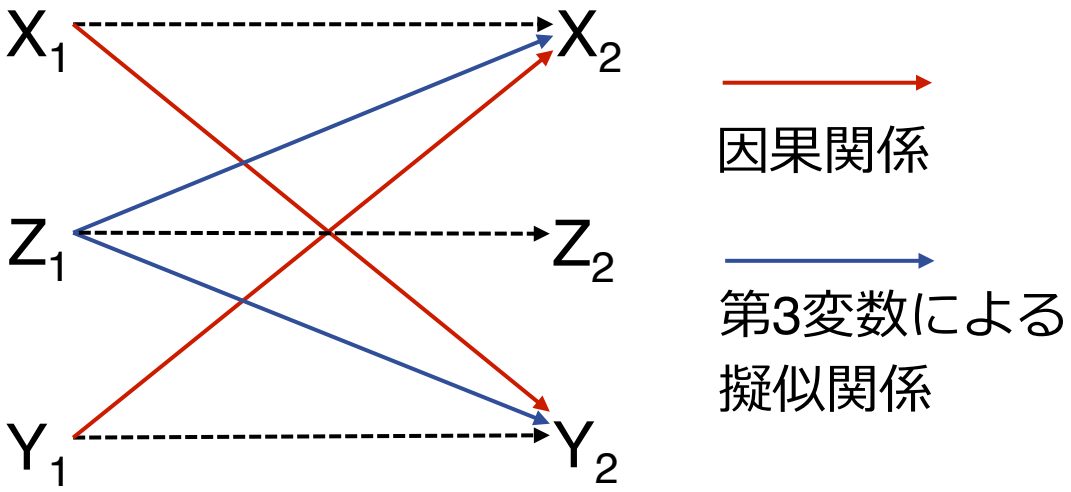
擬似関係



因果関係

X → Y

複数時点の観察を用いることで、変数間の因果の方向性や強さを評価することができる



パネルデータの困難・課題

1. **いかにサンプルの代表性を保ち続けるか?**
   パネルの脱落 (attrition)・新規サンプルの追加に伴い、母集団とサンプルのずれが大きくなる
2. **いかに有効な測定値を得るか?**
   同じ質問を繰り返し尋ねることによって回答者が「学習」してしまう（panel conditioning）
   時代が変わっても同じ意味の質問として通用するか
3. 資金の問題 → 省略

目次

0 はじめに

1 Introduction

2 Managing Panel Data

2.1 The Nature of Panel Data

2.2 The Basics of Panel Data Management

2.3 Three Case Studies on Poverty in Germany

2.4 How to Represent a Population with Panel Data?

wide形式とlong形式

パネルデータは、個人i (i = 1, …, N)、時点t (t = 1, …, T)、変数v (v = 1, …, V) の3次元からなる

wideデータ（N行のレコード）

id	income 2004	income 2005	income 2006
21	1245	1245	814
41	480	502	524

longデータ（N×T行のレコード）

id	year	income
21	2004	1245
21	2005	1245
21	2006	814
41	2004	480
41	2005	502
41	2006	524

パネルデータにおいては、**個体**を示す変数（id）と、**時点**を示す変数（year）という2つのkey変数がある

longデータ作成過程の2つのパターン

パネルデータ分析のためには、longデータを作る必要がある。ではどのように作ればよいのか？

- **Merge, Append**

各調査時点ごとにデータファイルが分かれており、同一個人を結びつけるlink idが記録されている場合（GSOEP）

→ **各年の同一個人を結びつけた統合データ（wide or long）を作成**

- **Reshape long**

すでにすべての調査時点がそろったwide形式のデータがある場合（JLPS）

→ **変数を指定し、wide形式のデータをlong形式に変換**

appendの論理

現在のdatasetにrowを加える操作

“same variables, different observations”

adult08

ID	HHNR	Year	Sex
1	10	2008	0
2	11	2008	1

append

child08

ID	HHNR	Year	Sex
3	10	2008	1
4	11	2008	0

ID	HHNR	Year	Sex
1	10	2008	0
2	11	2008	1
3	10	2008	1
4	11	2008	0

mergeの論理

現在のdatasetにcolumnを加える操作

“different variables, same observations”

adult08

ID	HHNR	Year	Sex
1	10	2008	0
2	11	2008	1

merge

house08

HHNR	Year	Income
10	2008	3000
11	2008	4000

ID	HHNR	Year	Sex	Income
1	10	2008	0	3000
2	11	2008	1	4000

append / merge のシンタクス

```
** how to use append ************************
append using filename [,options]
/*example*/
use “adult08.dta”   /*read data*/
append using child08   /*append*/
** how to use merge ***************************
merge 1:1 varlist using filename [,options]
/*example*/
use “house08.dta”   /*read data*/
save house08, replace   /*save data*/
use adult08   /*read new data*/
merge 1:1 HHNR using house08   /*merge*/
```

append / mergeを使う際の注意点

以下のようになっていないかをあらかじめ確認

append

1. 実質的に同じ変数なのに同じ名前になっていない
2. 同じ名前なのに実質的には異なる変数を指している

merge

1. 2つの観察が同じ個人に属しているはずなのに、同じ番号が振られていない
2. 2つの観察は異なる個人に属しているはずなのに、同じ番号が振られている

reshape long

id	income2004	income2005	income2006
21	1245	1245	814
41	480	502	524

`reshape long`

id	year	income
21	2004	1245
21	2005	1245
21	2006	814
41	2004	480
41	2005	502
41	2006	524

reshape longのシンタクス

```
** how to use reshape long ***************
reshape long varlist, i(i) j(j)
/*i is ID variable, j is new variable */
/*example*/
reshape long income, i(id) j(year)
```

事例紹介｜Cross-section, pooled cross-section, and panel analysis

1. **Cross-sectionのRQ**
   2004年において、何%の人が貧困**である**か?

2. **Pooled cross-sectionのRQ**
   2004年以降、貧困は**増加傾向にある**か?

3. **Panel analysisのRQ**
   2004-6年において、貧困**になる**リスクはどの程度大きいか?

逆に言えば、**3.**のようなRQでない限り、パネルデータ分析をする必要はないと言える

重みづけの問題

サンプルから母集団のパラメータを推定するときの仮定

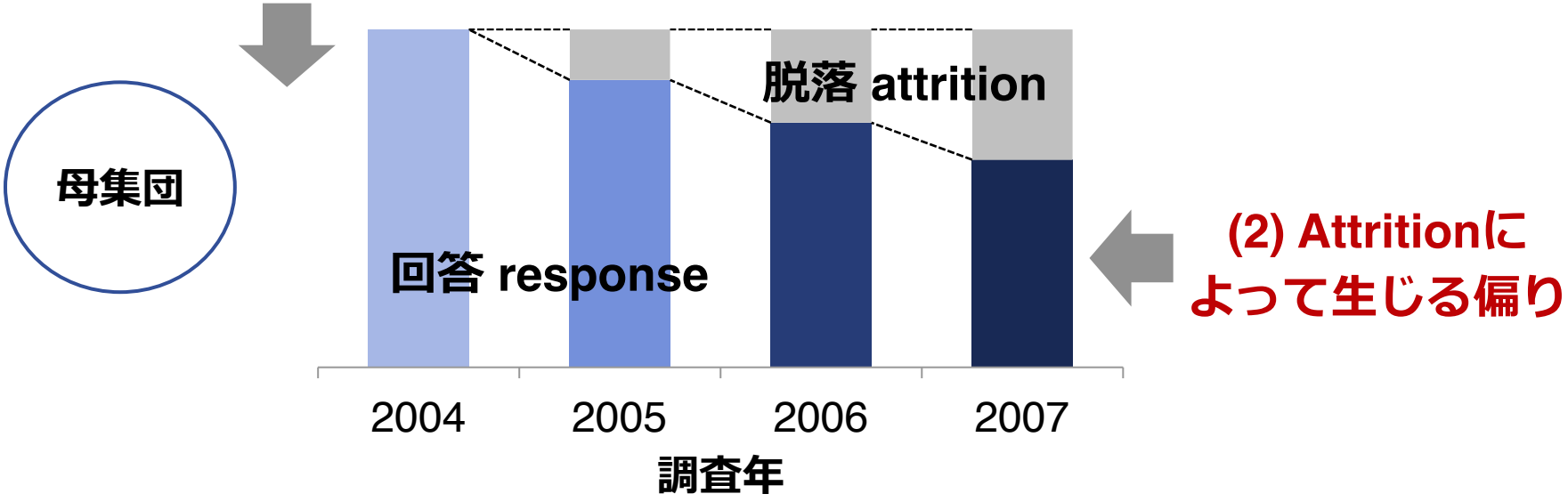
1. 各個体が無作為抽出されている
2. すべての個体の抽出確率が等しい
3. ある個体の抽出が他の個体の抽出確率に影響しない

社会調査データの場合、これらの仮定が満たされることはほとんどない（例: 多段抽出、オーバーサンプリング）

サンプルに入る確率が低い（高い）observationがより大きく（小さく）結果に反映されるようにする = **重みづけ(weighting)**

2つのバイアスと重みづけ

(1) Sampling design
の段階で生じる偏り



(2) Attritionに
よって生じる偏り

2004年時点でサンプリングが適切になされかつ回答者に偏りがなく、先の仮定を満たしていたとしても、サンプルの脱落によってサンプルの性質が母集団から乖離する可能性がある

パネルデータにおける重みづけの考えかた

(1) Sampling designの段階で生じる偏り

$$weight_t = designweight_t (\text{or } populationweight_t) \ (\text{if } t = 1)$$

(2) Attritionによって生じる偏り

$$weight_t = weight_{t-1} \times responserate_t^{-1} \ (\text{if } t \geq 2)$$

関心のある変数に関して(1)の偏りがある場合には、必ず重みづけしなければならない。

(2)の偏りがある場合、それがどのようなタイプの偏りであるかによって重みづけ等の対処をすべきかが異なる。

いつ重みづけをすべきか?

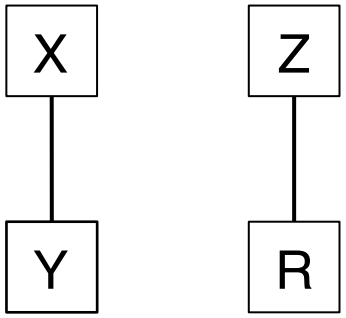
- **記述的推論**｜興味のある変数の分布や統計量に関して、母集団への一般化を志向するとき ＝ **重みづけ**が必要

  例) どの程度の人びとが貧困にさらされているか？

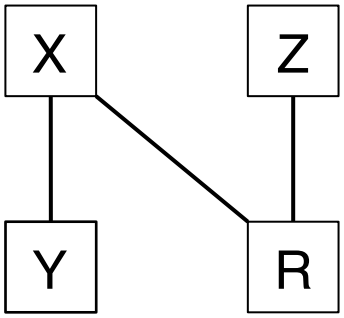
- **分析的推論**｜興味のある変数と、他の変数との関連（パラメータ）を測定・検定するとき ＝ セレクションと欠損確率を統制できるような**モデルの特定化**が必要

  例) 貧困であることと家族構成の間にはいかなる関連があるか？

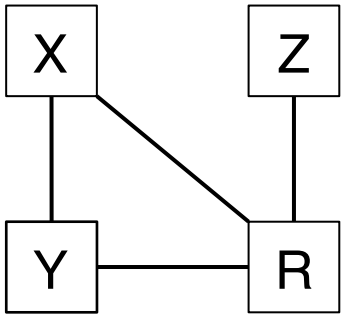
欠損データの3つのタイプ



MCAR
(Missing Completely At Random)



MAR
(Missing At Random)



MNAR
(Missing Not At Random)

Y: 従属変数　X: 独立変数　Z: その他の変数　R: 欠損確率

欠損確率Rが、他のどの変数と関連を持っているかがポイント

よくある状況のうち対処可能な、**欠損がMARの場合**について考える

欠損確率がMARの場合

欠損確率に影響を与える変数を統制変数としてモデルに組み込むことで対処できる

Heckmanの二段階推定を用いた例

1. プロビットモデルを用いて**欠損確率を推定**

$$\hat{\Pr}(R_{it} = 0 | R_{i,t-1} = 1) = \boldsymbol{Z}_{i,t-1}\hat{\gamma} \text{ for } t \geq 2$$

2. 欠損確率のモデルから求めた**ミルズ比をメインのモデルに投入**

$$Y_{it} = \boldsymbol{X}_{it}\boldsymbol{\beta} + \sigma\lambda(\boldsymbol{Z}_{i,t-1}\hat{\gamma}) + \varepsilon_{it}, \quad \lambda(\boldsymbol{Z}_{i,t-1}\hat{\gamma}) = \frac{\phi(\boldsymbol{Z}_{i,t-1}\hat{\gamma})}{\Phi(\boldsymbol{Z}_{i,t-1}\hat{\gamma})}$$

この手続きの場合、欠損確率についての重みづけを用いる必要はなくなる。

重みづけを用いる場合には、ロバスト標準誤差の利用が推奨

Balanced panelsと復活サンプル

- **Balanced panels**｜すべての個体について同一の回数の観察値を得ている

→現実のデータは確実にUnbalanced panelsであり、これに対応したデータ加工を行うのが現実的かつ望ましい

- **復活サンプル**｜ある時点で回答を得られなくなったが、それ以降で再び回答を得られるようになったケース。この場合、復活以降については重みづけが難しくなる。考えうる対処法は以下の3つ。
  1. すべてのサンプルについて重みづけをしない
  2. あらゆる無回答のパターンについてweightをそれぞれ計算
  3. サンプルから一度抜けた場合はそれ以降をすべて分析から除外

参考文献

Budig, Michelle J. and Paula England. 2001. “The Wage Penalty for Motherhood.” *American Sociological Review* 66(2):204–25.

Cherlin, Andrew J., P. Lindsay Chase-Lansdale, and Christine McRae. 1998. “Effects of Parental Divorce on Mental Health throughout the Life Course.” *American Sociological Review* 63(2):239–49.

DiPrete, Thomas A. and Patricia A. McManus. 2000. “Family Change, Employment Transitions, and the Welfare State: Household Income Dynamics in the United States and Germany.” *American Sociological Review* 65(3):343–70.

Halaby, Charles N. 2003. “Panel Models for the Analysis of Change and Growth in Life Course Studies.” Pp. 503–27 in *Handbook of the Life Course*, edited by Jeylan T. Mortimer and Michael J. Shanahan. New York: Kluwer Academic/Plenum Publishers.

Halaby, Charles N. 2004. “Panel Models in Sociological Research: Theory into Practice.” *Annual Review of Sociology* 30(1):507–44.

Rosenfeld, Rachel A. 1980. “Race and Sex Differences in Career Dynamics.” *American Sociological Review* 45(4):583–609.