日本教育心理学会 第66回総会
学会企画チュートリアル・セミナー

教育心理学研究のための欠測データ処理

資料目次（PDFのページ）

日本教育心理学会第66回総会
学会企画チュートリアル・セミナー
2024年9月15日（日）15:30～17:30

教育心理学研究のための欠測データ処理

話題提供　杉澤 武俊（早稲田大学）
話題提供　宇佐美 慧（東京大学）
話題提供　鈴木 雅之（横浜国立大学）
司　　会　鈴木 雅之（横浜国立大学）

はじめに：お願い

- 本チュートリアルの後半では，統計ソフトウェアRを用いた分析例を紹介します
  - R の操作に関する基礎的な説明も行いますが，初歩的な操作経験はお持ちであることを前提といたします
  - 操作に関する補助等を行うこともできかねますので，予めご理解くださいますようお願いいたします

- 実際にお手元で操作をされる場合には，事前にデータのダウンロードと，R のパッケージのインストールをお願いいたします
  - 使用するパッケージは以下の通りです
    - lavaan, mice, mitml, multcomp, semTools

企画の背景

- 多くの調査・実験研究において，何らかの欠測が生じる
  - 欠測の割合や処理方法が明記されていない論文が存在
  - 処理方法として削除法（リストワイズ削除・ペアワイズ削除）や単一代入法（代表値の代入など）が広く利用されてきた
    - これらの方法では，推定のバイアスや検定力低下など，推測上のさまざまな問題が生じうる

> *“The two popular methods for dealing with missing data that are found in basic statistics packages — listwise and pairwise deletion of missing values — are among the worst methods available for practical applications.”*
> (Wilkinson & Task Force on Statistical Inference APA Board of Scientific Affairs, 1999, p.600)

チュートリアル・セミナーの目的

- より適切な研究の実践ができるよう，欠測データ処理について理解を深める
  - 欠測が生じるメカニズム
  - 削除法や単一代入法などの古典的な方法の問題点
  - 完全情報最尤法と多重代入法の考え方
  - Rを用いた，完全情報最尤法と多重代入法による分析例
    - これら2つの方法が十分に機能しない場合もあるが，古典的な方法と比較すれば基本的にはより望ましい

> *“Unfortunately, virtually every mainstream missing data technique performs poorly with MNAR data, although maximum likelihood and multiple imputation tend to fare better than most traditional approaches*.”
> （Baraldi & Enders, 2010, p.8）

チュートリアル・セミナーの流れ

担当	内容	時間
杉澤	欠測が生じるメカニズムと古典的な方法	30分
宇佐美	完全情報最尤推定法と多重代入法	50分
鈴木	統計ソフトウェアRでの分析例	30分

参考文献

Baraldi, A. N., & Enders, C. K. (2010). An introduction to modern missing data analyses. *Journal of School Psychology*, *48*, 5-37.

Wilkinson, L., & Task Force on Statistical Inference APA Board of Scientific Affairs. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594–604

文献ガイド

- 欠測全般を扱った書籍
  - Enders, C. K. (2022). *Applied missing data analysis* (2nd Ed.). Guilford Publications.
  - Little, R. J. A., & Rubin, D. B. (2020). *Statistical analysis with missing data* (3rd ed.). Wiley.
  - 高井啓二・星野崇宏・野間久史 (2016). 欠測データの統計科学―医学と社会科学への応用　岩波書店
- 欠測に関する最新のレビュー
  - Enders, C.K. (2023). Missing data: An update on the state of the art. *Psychological Methods*. Advance online publication.
- 多重代入に関する書籍・論文
  - van Buuren, S. (2018). *Flexible imputation of missing data* (2nd ed.). Chapman and Hall.
  - 野間久史 (2017). 連鎖方程式による多重代入法　応用統計学 *46*, 67-86.
  - 高橋将宜・渡辺美智子 (2017). 欠測データ処理―Rによる単一代入法と多重代入法　共立出版

欠測が生じるメカニズムと古典的な方法

早稲田大学人間科学学術院　杉澤武俊

日本教育心理学会第66回（2024年）総会
学会企画チュートリアルセミナー「教育心理学研究のための欠測データ処理」話題提供

本話題提供の目的

- 欠測値処理の古典的方法とその問題点を理解する
- 欠測が生じるメカニズム，特に「ランダムな欠測」（MAR）の考え方を理解する

欠測値と完全データ・不完全データ

- 欠測値（欠損値；missing value）：
  何らかの原因（無回答など）により，変数の一部のデータが値として得られなかったもの
- 完全データ（complete data）：
  欠測値を含まないデータセット
- 不完全データ（incomplete data）：
  欠測値を含んだデータセット

完全データ

ID	y_1	y_2	y_3
1	5	9	4
2	4	6	2
3	7	1	3
4	2	5	1
5	3	6	7
6	8	2	5

不完全データ

ID	y_1	y_2	y_3
1	-	-	-
2	4	6	-
3	7	-	3
4	-	5	1
5	3	6	7
6	8	2	5

欠測値処理の古典的方法

- 削除法：欠測値を含む観測対象のデータを分析から除外することで，疑似的に完全データとする方法
- 単一代入法：欠測値を何らかの値で置き換えて，疑似的に完全データとする方法

削除法の例

- リストワイズ削除（listwise deletion / complete case analysis ）：欠測値を1つでも含む観測対象のデータを全て削除（ID 1～4を全ての分析から除外）
- ペアワイズ削除（pairwise deletion / available case analysis）：計算に使用する変数のデータがない対象をその都度削除（y_2とy_3の相関係数を求める際はID 1～3を除外しID 4は含める）

ID	y_1	y_2	y_3
1	-	-	-
2	4	6	-
3	7	-	3
4	-	5	1
5	3	6	7
6	8	2	5

単一代入法（single imputation）の例

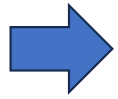
- 確定的単一代入法
  1. 平均値代入法(mean imputation)
  2. 確定的回帰代入法(deterministic regression imputation)
  3. 比率代入法(ratio imputation)
  4. ホットデック法(hot deck imputation)
- 確率的単一代入法
  5. 確率的回帰代入法(stochastic regression imputation)

参考：高橋・渡辺(2017)

平均値代入法

- 欠測を含む変数の観測値を使って求めた平均値を，その変数の全ての欠測値に代入する

ID	y_1
1	-
2	4
3	7
4	-
5	3
…	…
N	-
平均	5.5



ID	y_1
1	5.5
2	4
3	7
4	5.5
5	3
…	…
N	5.5
平均	5.5

確定的回帰代入法

- 回帰モデルから算出した予測値を欠測値の代わりとする
- 変数 y_2 に欠測があり，変数 y_1 は完全に観測されているなら，$\hat{y}_2 = \hat{\alpha} + \hat{\beta} y_1$ を欠測値に代入する（y_1の値が同じ人には同じ値が代入される。重回帰も可）
- さらにランダムな誤差項を加えたものが確率的回帰代入法

ID	y_1	y_2
1	5	4.44
2	4	6
3	7	3.60
4	6	5
5	3	6
6	8	2
…	…	…
N	7	3.60

$$\hat{y}_2 = 6.54 - 0.42 y_1$$

古典的方法の問題点

- 結果にバイアスが生じる場合がある
- 削除法ではサンプルサイズの減少に伴い，推定精度・検定力の低下（標準誤差の増大）が生じる
- 単一代入法では標準誤差を正しく評価できない

ペアワイズ削除の問題

- y_1, y_2, y_3の3変数間の相関係数を求める
  - y_1とy_2の相関：ID 1, 2のデータ
  - y_2とy_3の相関：ID 3, 4のデータ
  - y_3とy_1の相関：ID 5, 6のデータ

…各相関係数間に共通する対象がいない

→現実にはあり得ない相関行列が得られる可能性

ID	y_1	y_2	y_3
1	1	5	-
2	5	1	-
3	-	1	5
4	-	5	1
5	5	-	1
6	1	-	5

欠測メカニズムの分類

- 欠測値処理が適切に行えるか否かに欠測メカニズムが大きく関係する
- 各データが欠測となる確率がどのようになっているのか
- 欠測値にも本来得られるはずであった「真値」が存在するものとして，
  完全データとして得られるはずだったデータセットとの関係性を考える

1. 完全にランダムな欠測（Missing Completely At Random; MCAR）
2. （条件付きで）ランダムな欠測（Missing At Random; MAR）
3. ランダムでない欠測（Missing Not At Random; MNAR）

(Little & Rubin, 2020)

条件付き確率による欠測メカニズムの表現

Y: （観測対象）×（変数）からなる完全データ行列

R: 回答指標行列（Yの各要素について，観測=1，欠測=0としたもの）

Y_{obs}: 観測データ（Observed Data）

欠測メカニズム＝ $P(R|Y)$

Y

ID	y_1	y_2	y_3
1	5	2	8
2	4	6	7
3	7	3	3
4	6	5	1

R

ID	y_1	y_2	y_3
1	1	0	0
2	1	1	0
3	1	0	1
4	1	1	1

Y_{obs}

ID	y_1	y_2	y_3
1	5	-	-
2	4	6	-
3	7	-	3
4	6	5	1

完全にランダムな欠測（MCAR）

- ある変数における欠測の有無が，その変数自身や，データセットに含まれる他の変数の値とは無関係

→完全データから，各変数において欠測するデータが単純無作為抽出によって選ばれる

$$P(R|Y) = P(R)$$

（条件付きで）ランダムな欠測（MAR）

- ある変数における欠測の有無が，同じデータセットに含まれる他の変数の**実際に観測された値**とは関係するが，他の変数の値を統制したときにはその変数自体の値とは無関係

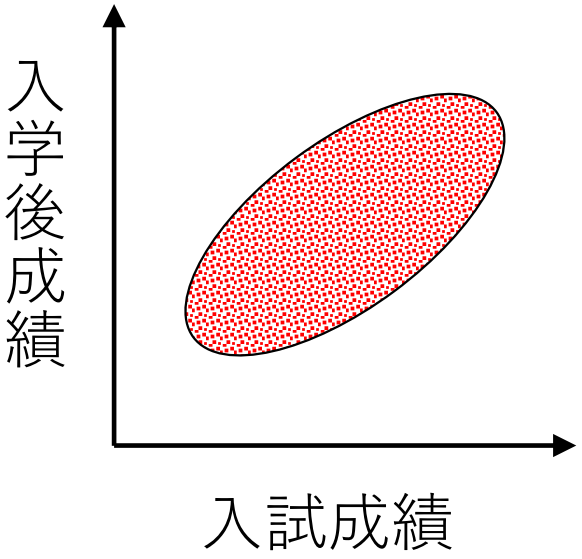
（Conditionally missing at random；CMAR; Graham, 2009）

→欠測するかどうかはどの欠測値にも依存しない

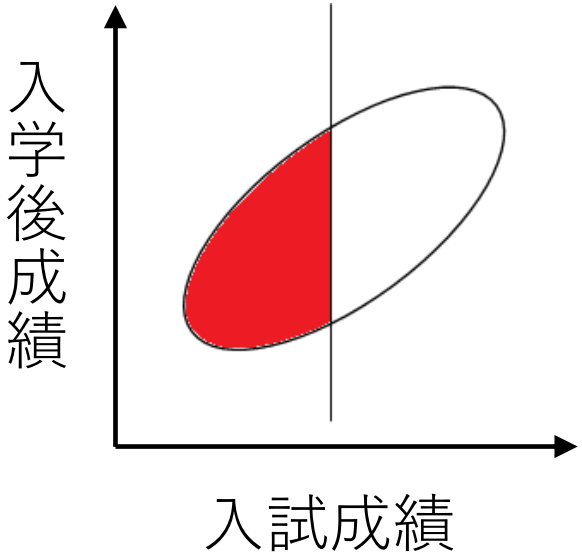
$$P(R|Y) = P(R|Y_{obs})$$

※数学的な詳細は，例えば，狩野(2019)

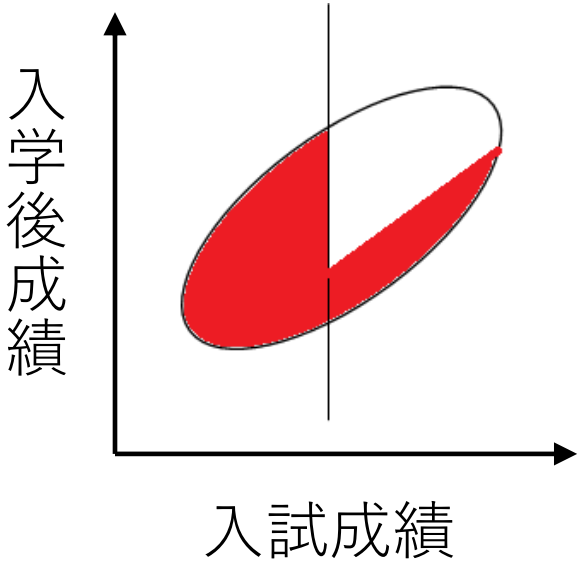
完全に
ランダムな欠測
（MCAR）



ランダムな欠測
（MAR）



ランダムな欠測
（MAR）でない



ランダムでない欠測（MNAR）

- 同じデータセットに含まれる他の変数を統制しても，ある変数における欠測の有無がその変数の値自体と関係

→欠測値自体が本来どのような値であったかに依存

$$P(R|Y) \neq P(R|Y_{obs})$$

- NMAR（Not Missing At Random）とも呼ばれる(Little & Rubin, 2002)

【参考】欠測メカニズムの判定

- MCARか否か

  Little(1988)の検定など

- MARかMNARか

  観測データのみから区別することは極めて困難

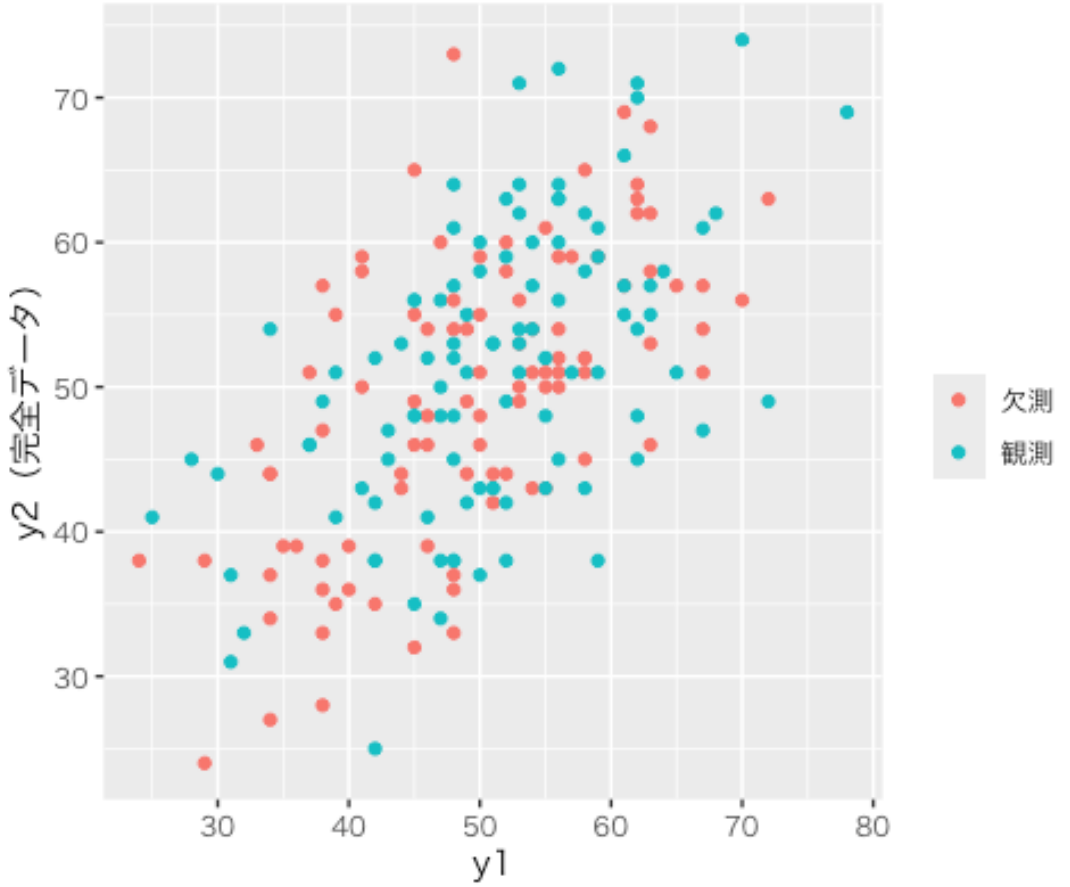
欠測メカニズムと欠測値処理

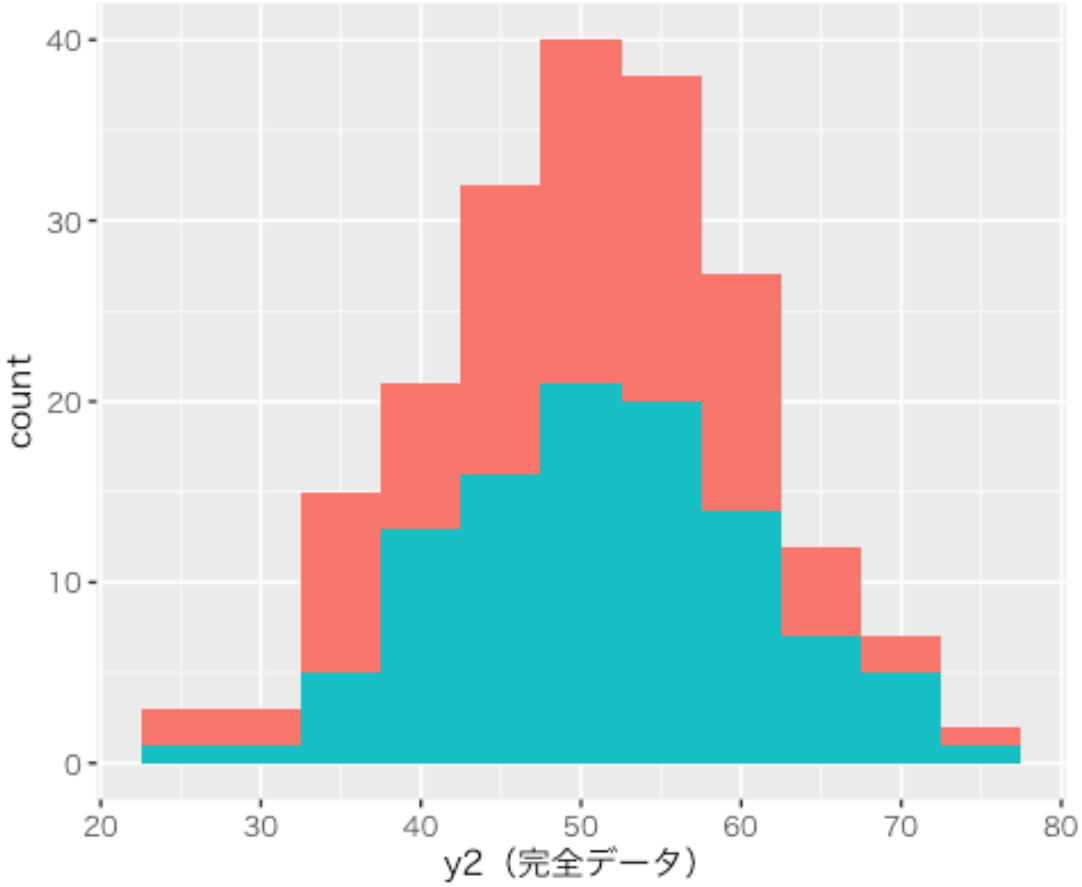
- MCARの場合，リストワイズ削除を行っても偏りのない推定が可能
- MARの場合，尤度に基づく方法（ベイズ推定含む）では回答指標を無視しても，偏りのない推定が可能
- MNARの場合，尤度の計算のために欠測メカニズムとしての具体的な確率モデルを指定する必要あり。適切な補助変数を分析に含めることでMARとみなせる可能性

シミュレーションによる古典的方法の評価

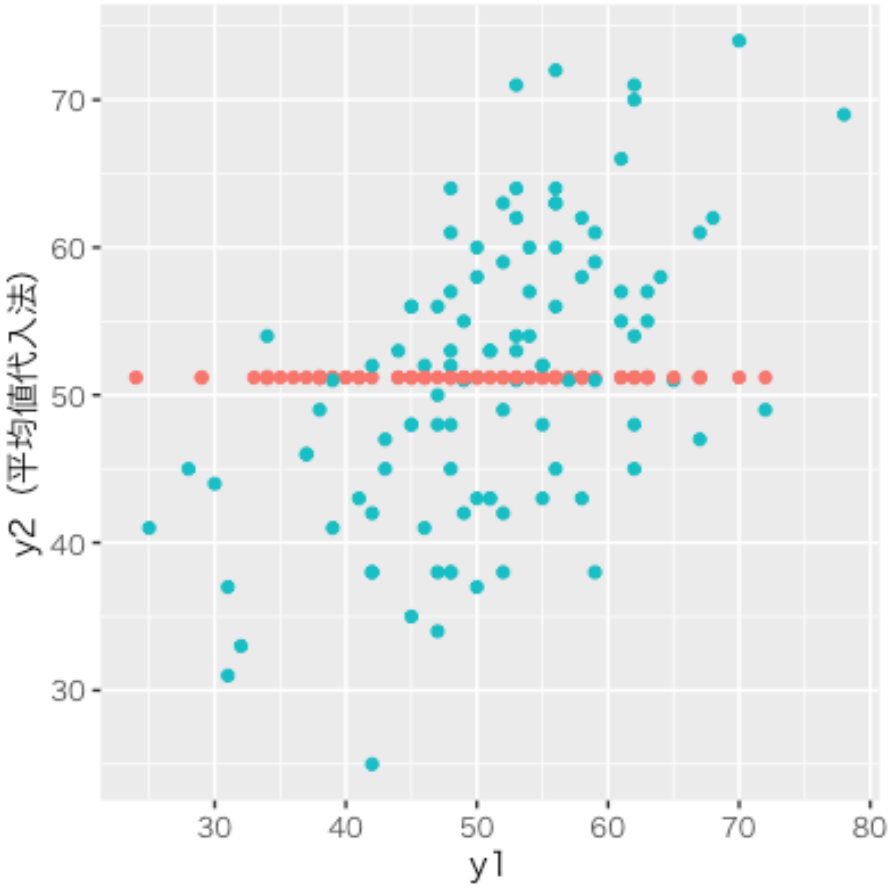
- y_1とy_2の 2 変数について，y_2にのみ欠測が生じている状況を想定
- N=200の完全データを生成（固定）
- 以下の条件で不完全データを生成
  - MCAR条件：y_1の値に関わらず確率0.5で欠測
  - MAR条件：y_1が中央値以上→欠測確率0.1，中央値未満→欠測確率0.9
- 欠測値処理：リストワイズ削除，平均値代入法，確定的回帰代入法
- 統計処理：平均値(y_2)，分散(y_2)，相関係数(y_1, y_2)
- 不完全データを生成し直して10000回反復

MCAR：完全データ

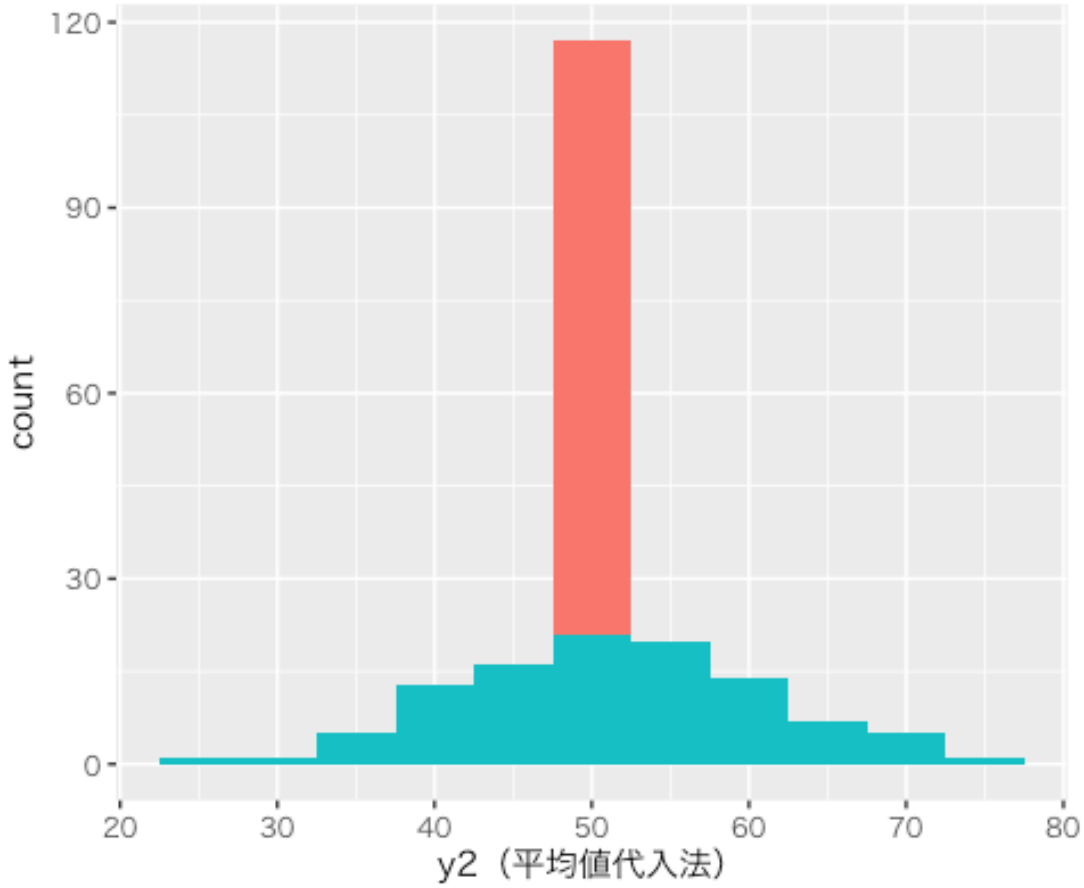




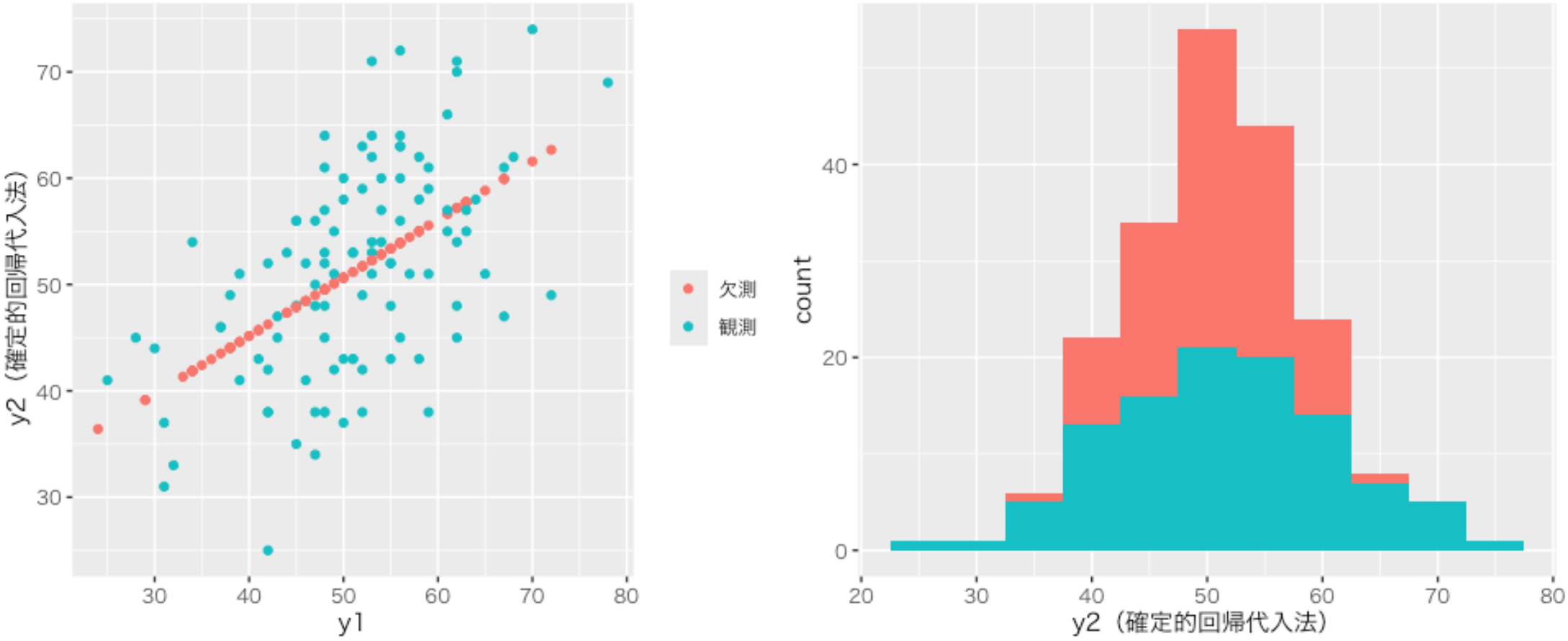
MCAR：平均値代入法


70
60
50
40
30
y2（平均値代入法）
30
40
50
60
70
80
y1

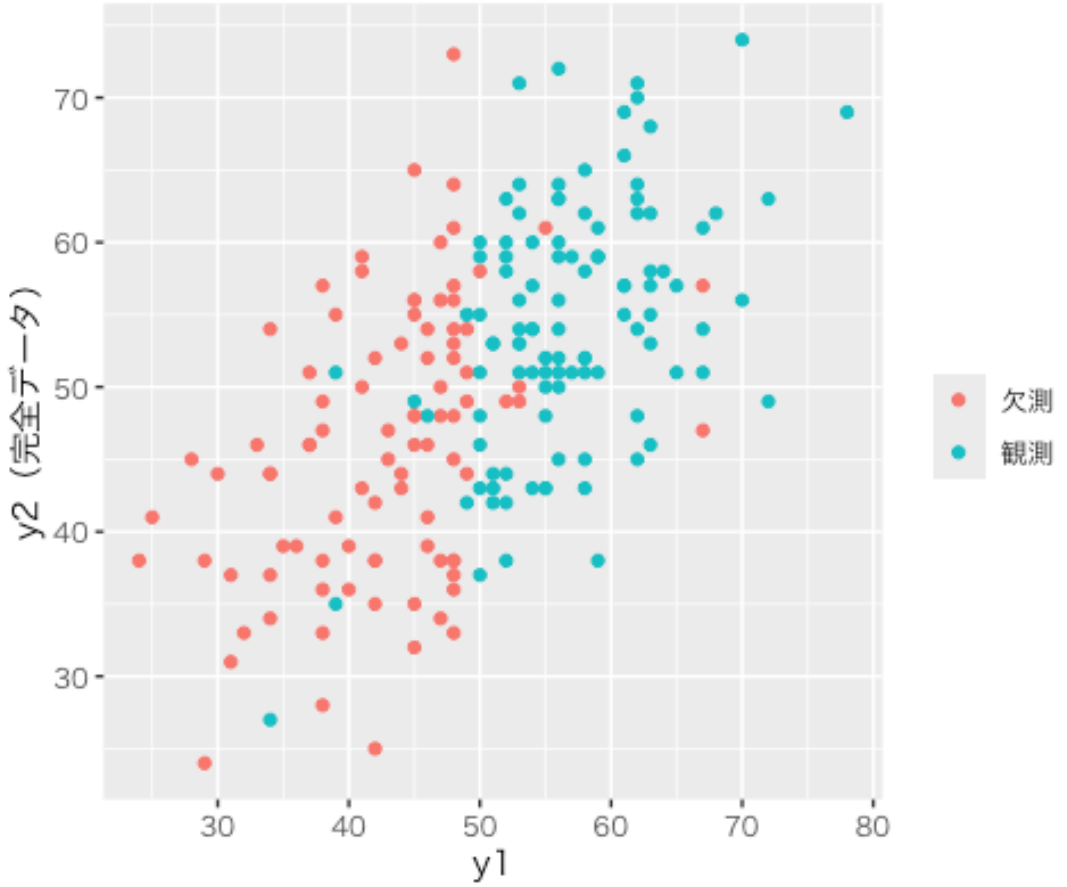

欠測
観測

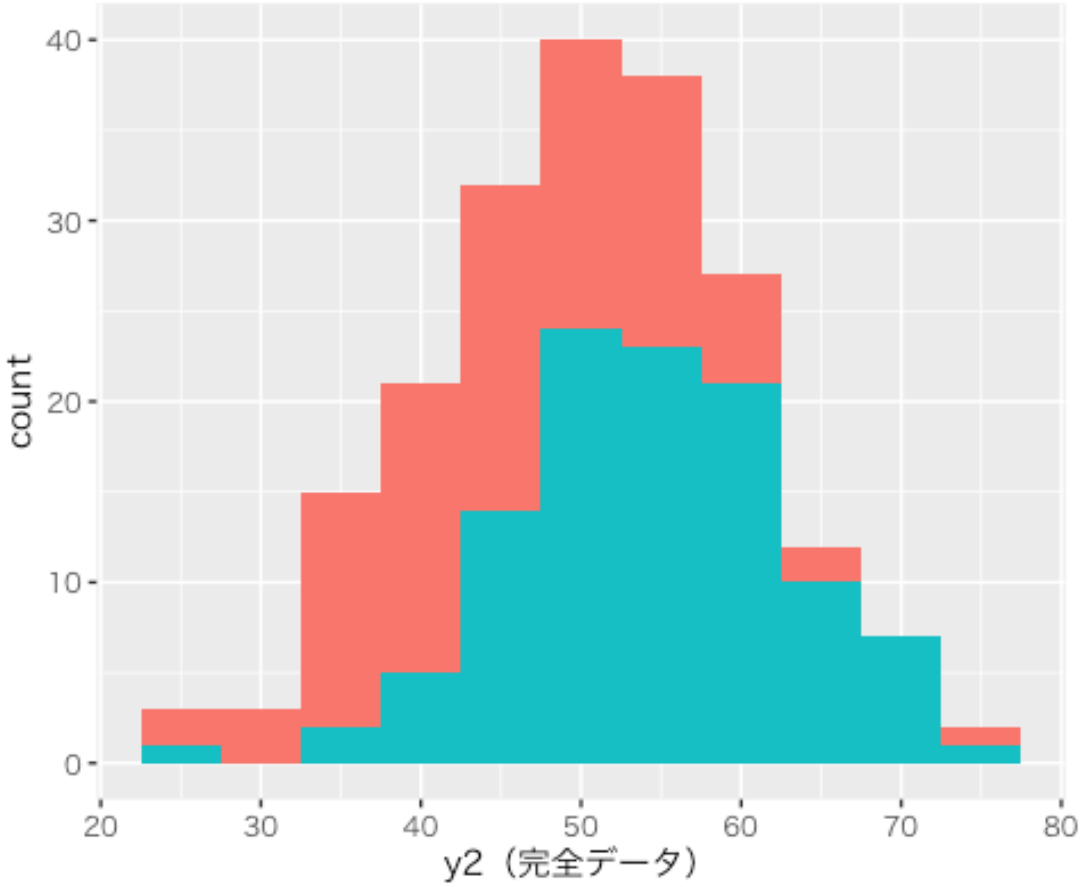

120
90
60
30
0
count
20
30
40
50
60
70
80
y2（平均値代入法）

MCAR：確定的回帰代入法

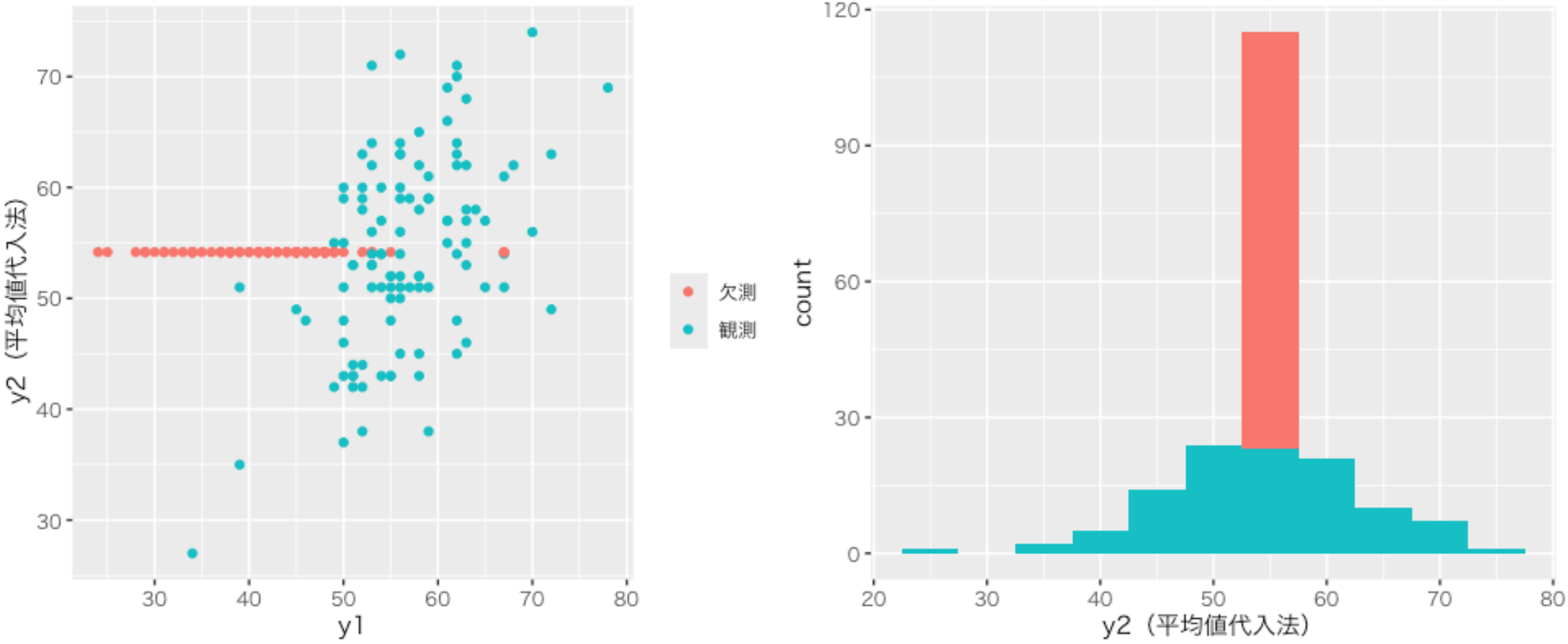


MAR：完全データ

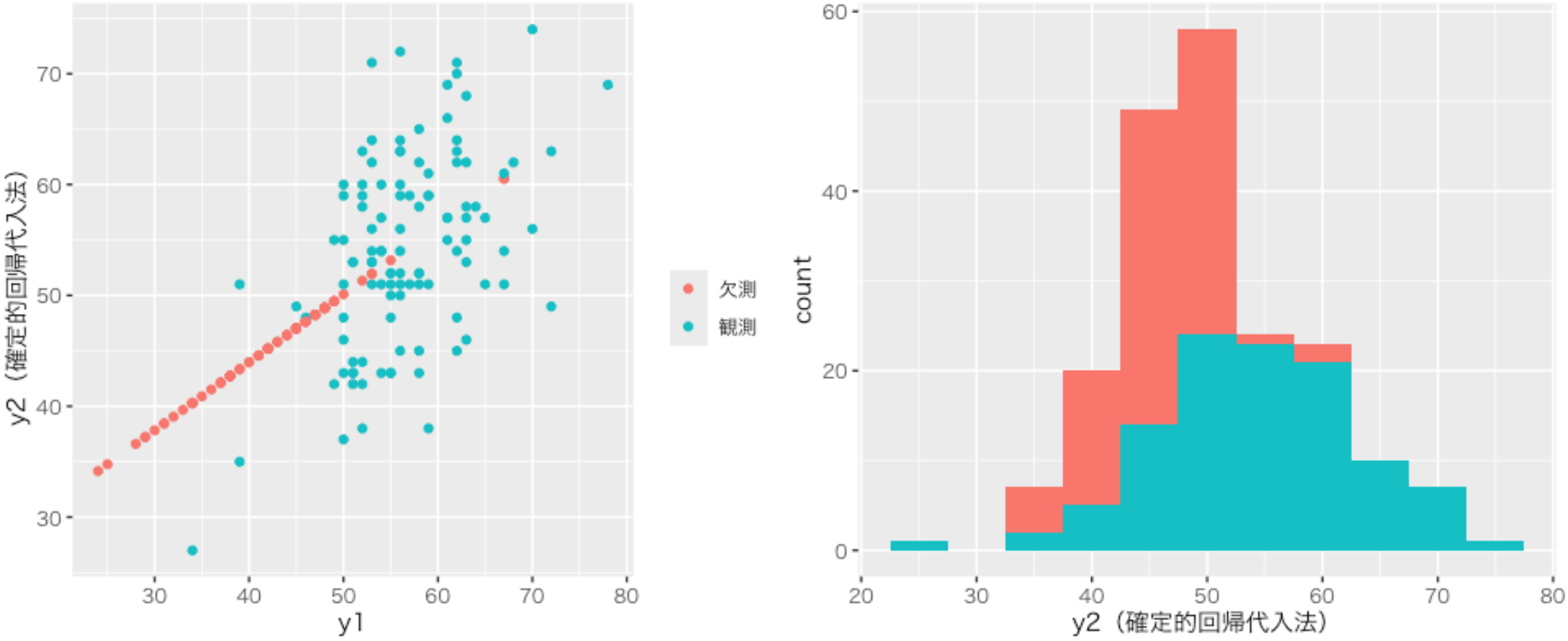

70
60
50
40
30
y2（完全データ）
30
40
50
60
70
80
y1
欠測
観測


40
30
20
10
0
count
20
30
40
50
60
70
80
y2（完全データ）

MAR：平均値代入法



MAR：確定的回帰代入法



シミュレーションの結果（10000回の平均）

		MCAR			MAR		
	完全データ	リストワイズ削除	平均値代入法	回帰代入法	リストワイズ削除	平均値代入法	回帰代入法
平均	50.3	50.3	50.3	50.3	53.9	53.9	50.8
分散	99.0	99.0	49.3	66.3	74.0	39.2	59.1
相関係数	.582	.581	.410	.711	.451	.244	.683

引用文献

- Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, *60*, 549-576.
- 狩野裕 (2019). 欠測データ解析のmissとmyth　日本統計学会誌, *48*(2), 199-214.
- Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, *83*(404), 1198–1202.
- Little, R. J. A. & Rubin, D. B. (2002). *Statistical Analysis with Missing Data* (2nd ed.). Wiley.
- Little, R. J. A. & Rubin, D. B. (2020). *Statistical Analysis with Missing Data* (3rd ed.). Wiley.
- 高橋将宜・渡辺美智子 (2017). 欠測データ処理—Rによる単一代入法と多重代入法 共立出版

日本教育心理学会第66回（2024年）総会
学会企画チュートリアルセミナー
「教育心理学研究のための欠測データ処理」

完全情報最尤推定法と多重代入法

宇佐美　慧
（東京大学）

Email: usami_s@p.u-tokyo.ac.jp
HP: http://usami-lab.com/

復習

- MAR：ある変数(y_1)が欠測するかどうかは、別の観測変数($y_2, y_3, \ldots$)にのみ依存し、欠測値(y_1)そのものには依存しない。

- MAR に基づく欠測下では、リストワイズ法によってデータの削除を行うと多くの場合に推測上のバイアスが生じ、標準誤差も不当に大きくなる。

- 完全情報最尤推定法（full information maximum likelihood: FIML）や多重代入法（multiple imputation: MI）はMARに基づく欠測下において有用な処理法。

FIMLとMI

- FIML：各個人（対象）の観測データのみを用いて母数を最尤推定する方法。補完（代入）を伴わない。

- 観測データのみに基づく尤度は直接尤度（direct likelihood; または観測尤度や完全情報尤度）と呼ばれる。 *仮に欠測がない場合、通常の尤度関数は直接尤度に対応する。

- MI:補完モデルと乱数を用いて欠測値を補完し、疑似的な完全データセットを複数作成する。そして、関心のある分析モデルをそれぞれあてはめ、推定結果を統合する方法。

  *「欠測値を復元して、1つの尤もらしい完全データセットを作成・統合し分析する方法」ではない。

- 補完モデルと分析モデルが明確に区別される。

はじめに：FIMLとMIの使いわけ

- MARが仮定でき、また分布仮定を含めモデルを正しく設定できれば、一般に最尤推定量（FIML）は良い特徴（e.g., 標準誤差の小ささ）をもつ。

- 特に、SEM（構造方程式モデリング・共分散構造分析）で表現可能な下位モデルを分析モデルとする場合にFIMLの実装は容易（Newsom, 2015）。

- 回帰分析モデル、因子分析モデル、パス分析、潜在成長モデルなどの種々の縦断モデル。

はじめに：FIMLとMIの使いわけ

- 補完の実行者と分析者が異なるケースがある（自治体によるデータの二次利用を目的とした補完）。例えばMIでは、個人情報が特定される恐れのある共変量（補助変数）に欠測が依存している場合でも、このような情報を含めない（複数の）完全データセットを提供可能（高井他, 2016）。

- テストや心理尺度等を通して、その項目和得点を用いた分析や実践を行う場合、補完を行うMIは直接的で有用。

- ソフトウェアの観点からは、特にSEMで直接表現できないモデルを扱う際に、MI の方が容易に実装できる状況も多い（e.g., Asparouhov & Muthen, 2022）。例えば階層線形モデル（マルチレベルモデル）や種々の非線形モデル。

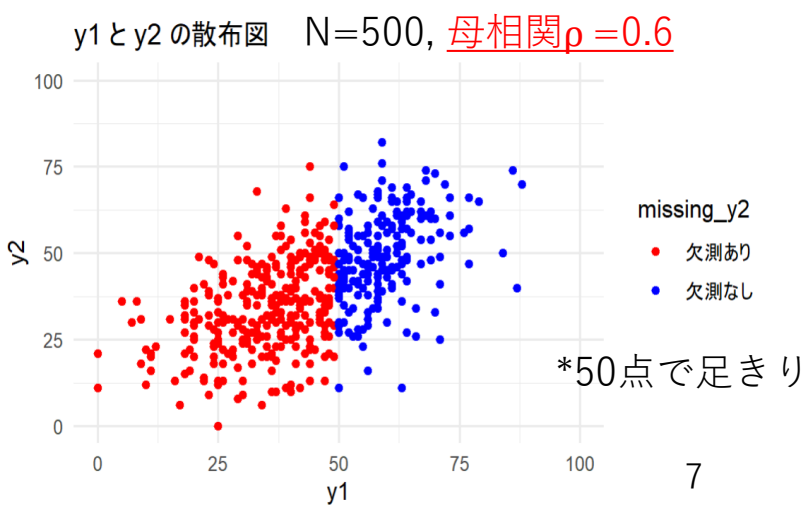
アウトライン

- FIMLの概要
- MIの概要
- 補助変数の活用
- まとめとMNARの場合

FIML

- 観測データのみを用いた直接尤度を構成し母数を最尤推定。
- 個人（対象）ごとの尤度を考える。
- MAR の欠測例として、1 次試験の得点(y_1)が低い受験者が足きりにより2 次試験の得点(y_2)が欠測している場合を考える。
- y_1とy_2のあいだの母相関係数ρを推定したい。

ID	y_1	y_2
1	55	28
2	36	-
3	20	-
4	69	62
…	…	…
500	26	-



直接尤度の例

- 個人iの変数y_1, y_2に関するデータをy_{i1}, y_{i2}とする。

- 直接尤度＝個人1の尤度×個人2の尤度×個人3の尤度…×個人Nの尤度

- 数式で書けば、

$$f(y_{11}, y_{12}) \times f(y_{21}) \times f(y_{31}) \times f(y_{41}, y_{42}) \times \cdots \times f(y_{N1})$$

*母数を表す記号は省略

ID	y_1	y_2
1	55	28
2	36	-
3	20	-
4	69	62
…	…	…
N=500	26	-

相関係数の推定

・リストワイズ削除の場合　$f(y_{11}, y_{12}) \times f(y_{41}, y_{42}) \times \cdots$

（足きりを受けていない受験者集団に限定した分析）

$\hat{\rho} = 0.347$

ID	y_1	y_2
1	55	28
4	69	62
…	…	…

・FIMLの場合　$f(y_{11}, y_{12}) \times f(y_{21}) \times f(y_{31}) \times f(y_{41}, y_{42}) \times \cdots \times f(y_{N1})$

$\hat{\rho} = 0.593$　（母相関ρ =0.6に近い）

SEMとFIML

- （教育）心理学研究では、SEMを用いたモデルの推定や評価は広くなされている。

- 回帰モデル、因子分析モデル、パス分析、媒介モデル、多母集団モデル、潜在成長モデル・交差遅延モデル等の縦断モデル。

- SEMでは一般に、複数の潜在変数と観測変数を伴う線形モデルの表現が可能で、現在でも様々な拡張が行われている。

・最尤法はSEM （共分散構造分析）で最もよく利用される推定法であり、直接尤度（FIML)の構成も直接的かつ容易。Rのlavaanパッケージ（Rosseel, 2012）、Mplus 等のSEMの標準的なソフトウェアではFIMLに基づく推測が容易に実行できる。

SEMの推定の考え方と直接尤度

- データの標本平均・（共）分散と、分析モデルの平均・（共）分散が「近く」なるように、分析モデル内の母数$\boldsymbol{\theta}$を推定する。後者は平均構造$\boldsymbol{\mu(\theta)}$、共分散構造$\boldsymbol{\Sigma(\theta)}$と呼ばれる。

- 最尤法では通常、多変量正規分布に基づく尤度の最大化によって、これらが「近く」なるような母数$\boldsymbol{\theta}$の推定を行う。

＜尤度関数＞*各個人のデータをまとめて$\boldsymbol{y_i}$と表記。

$$f(\boldsymbol{y_1}|\mu(\boldsymbol{\theta}),\Sigma(\boldsymbol{\theta})) \times f(\boldsymbol{y_2}|\mu(\boldsymbol{\theta}),\Sigma(\boldsymbol{\theta})) \times \cdots \times f(\boldsymbol{y_N}|\mu(\boldsymbol{\theta}),\Sigma(\boldsymbol{\theta}))$$

$$f(\boldsymbol{y_i}|\mu(\boldsymbol{\theta}),\Sigma(\boldsymbol{\theta})) = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma(\boldsymbol{\theta})|^{\frac{1}{2}}} \exp[-\frac{1}{2}(\boldsymbol{y_i}-\mu(\boldsymbol{\theta}))^{\mathrm{T}}\Sigma(\boldsymbol{\theta})^{-1}(\boldsymbol{y_i}-\mu(\boldsymbol{\theta}))]$$

*p次の多変量正規分布の密度関数。Tは転置。

＜欠測がある場合の直接尤度＞

各個人で観測された変数に対応する $\mu(\boldsymbol{\theta}),\Sigma(\boldsymbol{\theta})$ の一部要素を利用。

具体例：SEMに基づく回帰分析モデルの推定

$$y_2 = \alpha + \beta y_1 + \varepsilon$$

α :切片、 β :回帰係数、 ε :残差(平均0, 分散σ^2)

・確率変数y_1とy_2の多変量正規性を仮定

・ y_1の母平均はμ_1 、母分散はσ_1^2

⇒

・ $\boldsymbol{\theta} = (\mu_1, \sigma_1^2, \alpha, \beta, \sigma^2)^{\mathrm{T}}$ ⇒母数は5種類。

・ $\mu(\boldsymbol{\theta}) = (\mu_1, \alpha + \beta y_1)^{\mathrm{T}}$ ⇒順に、モデルに基づくy_1 とy_2の平均。

・ $\Sigma(\boldsymbol{\theta}) = \begin{pmatrix} \sigma_1^2 & \beta\,\sigma_1^2 \\ \beta\,\sigma_1^2 & \beta^2\sigma_1^2 + \sigma^2 \end{pmatrix}$ ⇒ $\begin{pmatrix} y_1\text{の分散} & y_1, y_2\text{の共分散} \\ y_1, y_2\text{の共分散} & y_2\text{の分散} \end{pmatrix}$

＊y_1とy_2の関係を記述する回帰モデルの設定を通して、(暗に)各変数の平均や[共]分散を母数$\boldsymbol{\theta}$の関数で記述している。

・尤度関数 $f(\boldsymbol{y_1}|\mu(\boldsymbol{\theta}), \Sigma(\boldsymbol{\theta})) \times f(\boldsymbol{y_2}|\mu(\boldsymbol{\theta}), \Sigma(\boldsymbol{\theta})) \times \cdots \times f(\boldsymbol{y_N}|\mu(\boldsymbol{\theta}), \Sigma(\boldsymbol{\theta}))$

具体例：SEMに基づく回帰分析モデルの推定

- 欠測がある場合の直接尤度

$$f(y_{11}, y_{12}|\mu(\boldsymbol{\theta}), \sum(\boldsymbol{\theta})) \times f(y_{21}, |\mu_1, \sigma_1^2) \times f(y_{31}, |\mu_1, \sigma_1^2) \times$$
$$f(y_{41}, y_{42}|\mu(\boldsymbol{\theta}), \sum(\boldsymbol{\theta})) \times \cdots \times f(y_{N1}, |\mu_1, \sigma_1^2)$$

- $\mu(\boldsymbol{\theta}) = (\mu_1, \alpha + \beta y_1)^{\mathrm{T}}$
- $\Sigma(\boldsymbol{\theta}) = \begin{pmatrix} \sigma_1^2 & \beta\,\sigma_1^2 \\ \beta\,\sigma_1^2 & \beta^2\sigma_1^2 + \sigma^2 \end{pmatrix}$

y_2 に欠測のある個人については、$\mu(\boldsymbol{\theta}), \Sigma(\boldsymbol{\theta})$の$y_1$に対応する要素（$\mu_1, \sigma_1^2$）のみを利用。

たとえば、

$$f(y_{21}|\mu_1, \sigma_1^2) = \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp[-\frac{(y_{21} - \mu_1)^2}{2\sigma_1^2}]$$

ID	y_1	y_2
1	55	28
2	36	-
3	20	-
4	69	62
…	…	…
N=500	26	-

補足

- 回帰分析モデルに限らず、SEMで表現できる下位モデル（e.g., 因子分析モデル、パス分析）であれば個々のモデルに応じた$\boldsymbol{\mu(\theta)}$, $\boldsymbol{\Sigma(\theta)}$の表現が可能なので、欠測があってもさきと同様の方式の下で直接尤度を構成し母数を推定できる。

- 分析モデルが適切に設定できれば、 MARに基づく欠測データ処理法として、 SEMの文脈では基本的にFIMLの利用で問題ない。そのため、この文脈ではMIについての説明を割愛している文献もある（Newsom, 2015）。

補足

- 一般に、より多くの変数について観測データが得られた個人の方が全体の推定に与える影響は大きくなる。

- 回帰分析の例では、どの変数が観測されているかに関するパタンの総数は2通りなので，集団全体が2つの群に分かれていると見做せる。FIMLはこのような複数の群のデータを扱う多群モデル（多母集団モデル）としても位置付けられる。

- 観測データの多変量正規性を仮定したSEMのFIMLを説明したが、MAR に基づく欠測下では通常、変数間が線形的な関係であれば、分布が非正規である場合にも$\boldsymbol{\theta}$の推定値は一致性をもつ（Nが大きくなれば真の値に確率収束する）ことが知られている（e.g., Yuan & Bentler, 2010）。

補足

- Zhang & Savalei (2023)…欠測がある場合のFIML における適合度指標（RMSEA やCFI）の算出に関して。

- Savalei & Rosseel (2022)…分布が正規および非正規であるデータに欠測がある際の標準誤差やモデルの検定統計量の算出に関する包括的なサマリー。

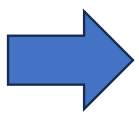
アウトライン

- FIMLの概要
- MIの概要
- 補助変数の活用
- まとめとMNARの場合

単一代入の問題

- 例えば、確定的回帰代入（補完モデルとしての回帰分析モデルから得られる条件付平均による予測値を補完に用いる方法）を行う場合。

ID	y_1	y_2
1	55	28
2	36	-
3	20	-
4	69	62
…	…	…
500	26	-



ID	y_1	y_2
1	55	28
2	36	35.3
3	20	25.7
4	69	62
…	…	…
500	26	29.3
平均	44.3 (45)	40.2 (40)
分散	236 (225)	142 (225)
共分散		140 (135)

*カッコ内の数字は真の値。

- y_2の分散の過小推定。
- 予測誤差、すなわち残差分散や補完モデルの（切片や回帰係数等についての）推定誤差を考慮していないことに起因。

MI

- MI（Rubin, 1987）はベイズ統計学の枠組の下で構築された、汎用性の高い欠測データ処理法。

- 3つのステップ：

補完ステップ：補完モデルと乱数を用いて欠測値を補完し、疑似的な完全データセットを複数作成。

分析ステップ：各完全データセットに対し関心がある（確認的因子分析モデルなどの）分析モデルをそれぞれあてはめ母数$\boldsymbol{\theta}$を推定。

統合ステップ： 得られた複数の$\boldsymbol{\theta}$の推定結果を統合。

・補完モデルと分析モデルが明確に区別される。

・MIは、例えばSEMによる表現が困難なモデルに対しても汎用的に利用できる。SEMの文脈でもMIは実装可能（lavaanやMplus）。

仮想的なデータセット（”-”部分が欠測）

ID	y_1	y_2	…	y_8
1	-	3	…	4
2	4	-	…	-
3	2	-	…	-
4	4	3	…	2
…	…	…	…	…
N	5	-	…	-

MIの流れ

補完ステップ

完全データセット1

ID	y_1	y_2	…	y_8
1	3	3	…	4
2	4	4	…	5
3	2	3	…	3
4	4	3	…	2
…	…	…	…	…
N	5	5	…	1

完全データセット2

ID	y_1	y_2	…	y_8
1	4	3	…	4
2	4	3	…	5
3	2	4	…	2
4	4	3	…	2
…	…	…	…	…
N	5	5	…	2

完全データセット3

ID	y_1	y_2	…	y_8
1	4	3	…	4
2	4	4	…	5
3	2	2	…	3
4	4	3	…	2
…	…	…	…	…
N	5	4	…	2

…

完全データセットM

ID	y_1	y_2	…	y_8
1	3	3	…	4
2	4	4	…	4
3	2	3	…	4
4	4	3	…	2
…	…	…	…	…
N	5	5	…	1

分析ステップ

(母数の点推定値と誤差[共]分散[標準誤差])

$\widehat{\boldsymbol{\theta}}_{\boldsymbol{1}}$, $\mathrm{V}(\widehat{\boldsymbol{\theta}}_{\boldsymbol{1}})$　　$\widehat{\boldsymbol{\theta}}_{\boldsymbol{2}}$, $\mathrm{V}(\widehat{\boldsymbol{\theta}}_{\boldsymbol{2}})$　　$\widehat{\boldsymbol{\theta}}_{\boldsymbol{3}}$, $\mathrm{V}(\widehat{\boldsymbol{\theta}}_{\boldsymbol{3}})$　　$\widehat{\boldsymbol{\theta}}_{\boldsymbol{M}}$, $\mathrm{V}(\widehat{\boldsymbol{\theta}}_{\boldsymbol{M}})$

統合ステップ

(最終的な点推定値と誤差[共]分散[標準誤差])

$\widehat{\boldsymbol{\theta}}$, $\mathrm{V}(\widehat{\boldsymbol{\theta}})$

*「1つの尤もらしい完全データセットを作成・統合し分析する方法」ではない

補完ステップ：連鎖方程式MICE

- 大別して、欠測のある変数についての同時事後分布を用いる方法（joint modeling: JM）と、完全条件付分布を用いる方法（fully conditional specification: FCS）の2つがある。

- JMでは通常、欠測のある変数が多変量正規分布に従うことを仮定する。

- FCS では、欠測のある変数について、他の全ての変数が所与の下での完全条件付分布を用いて補完し、その作業を各変数に対して行う。汎用性が高い。

- FCS のアルゴリズムとして、連鎖方程式によるMI（multiple imputation by chained equation: MICE, van Buuren & Groothuis-Oudshoorn, 2011）は近年特に広く利用されている。

MICEによる補完ステップ

(i) 補完モデルの設定

(ii) 初期値の設定

(iii) 連鎖方程式による補完値の更新

(iv)(iii)の反復

ID	y_1	y_2	…	y_8
1	-	3	…	4
2	4	-	…	-
3	2	-	…	-
4	4	3	…	2
…	…	…	…	…
N	5	-	…	-

MICEによる補完ステップ

(i) 補完モデルの設定

・yが連続変数の場合、補完モデルとして線形回帰モデルが用いられることが多い。

・補完モデルには、分析モデルにない変数（たとえば、z_1, $z_2, \ldots$）も含めてよい。<u>分析モデル内の変数は原則含める。</u>

(ii) 初期値の設定

・単一代入など、適当な方法で得た初期値により欠測値を補完して疑似的な完全データセットを作成する。

ID	y_1	y_2	…	y_8
1	2.7	3	…	4
2	4	2.6	…	1.9
3	2	2.1	…	3.2
4	4	3	…	2
…	…	…	…	…
N	5	4.3	…	2.3

MICEによる補完ステップ

(iii) 連鎖方程式による補完値の更新(y_1)

- y_1内にある欠測値を、疑似的な完全データセット内のy_2, $y_3 . . . , y_8$を用いたy_1の補完モデルから生成された補完値により補完し更新する。

ID	y_1	y_2	…	y_8
1	2.7 ⇒ **3.1**	3	…	4
2	4	2.6	…	1.9
3	2	2.1	…	3.2
4	4	3	…	2
…	…	…	…	…
N	5	4.3	…	2.3

*補完モデルとして線形回帰モデルを用いた場合の例。
*補完モデルには、分析モデルにない変数（たとえば、$z_1, z_2, \ldots$）も含めてよい。

補足：補完モデルが線形回帰モデルの場合
(Rubin, 1987; 野間, 2017)

・モデル内の母数（偏回帰係数$\boldsymbol{\beta}$、残差分散σ^2）のサンプルを得て、それを用いて補完値を乱数により生成する。

ポイント：予測誤差、すなわち残差分散や補完モデルの（回帰係数等についての）推定誤差を考慮して補完。

・ q (=8-1=7)個の独立変数を含む線形回帰モデル($y_1 = \alpha + \beta_2 y_2 + \beta_3 y_3 + \cdots + \beta_8 y_8 + \varepsilon$)において、$\hat{\boldsymbol{\beta}}, \hat{\sigma}^2$を完全ケース(サイズ$n_{obs}$)からの推定値、$\hat{\boldsymbol{V}}$を$\hat{\boldsymbol{\beta}}$の共分散行列の推定値とする。

・ $\hat{\boldsymbol{\beta}}, \hat{\sigma}^2$の標本分布からのサンプルは、

$$\sigma^* = \hat{\sigma}\sqrt{\frac{n_{obs}-q}{g}}, \qquad \boldsymbol{\beta}^* = \hat{\boldsymbol{\beta}} + \frac{\sigma^*}{\hat{\sigma}} u_1 \hat{\boldsymbol{V}}^{\frac{1}{2}}$$

g: $n_{obs} - q$を自由度とするカイ二乗分布からの乱数。
u_1: q次の多変量正規分布からの乱数。

で得られ、y_1で欠測が生じている個人iのデータy_{i1}^*は、標準正規分布からの乱数u_{i2}を用いて、以下から生成される。

$$y_{i1}^* = \alpha^* + \beta_2^* y_{i2} + \beta_3^* y_{i3} + \cdots + \beta_8^* y_{i8} + u_{i2}\sigma^*$$

MICEによる補完ステップ

(iii) 連鎖方程式による補完値の更新(y_2)

- y_2内にある欠測値を、疑似的な完全データセット内のy_1, $y_3 . . . , y_8$を用いたy_2の補完モデルから生成された補完値により補完し更新する。

ID	y_1	y_2	…	y_8
1	**3.1**	3	…	4
2	4	2.6⇒**2.9**	…	1.9
3	2	2.1⇒**1.8**	…	3.2
4	4	3	…	2
…	…	…	…	…
N	5	4.3⇒**4.6**	…	2.3

MICEによる補完ステップ

(iii) 連鎖方程式による補完値の更新(y_8)

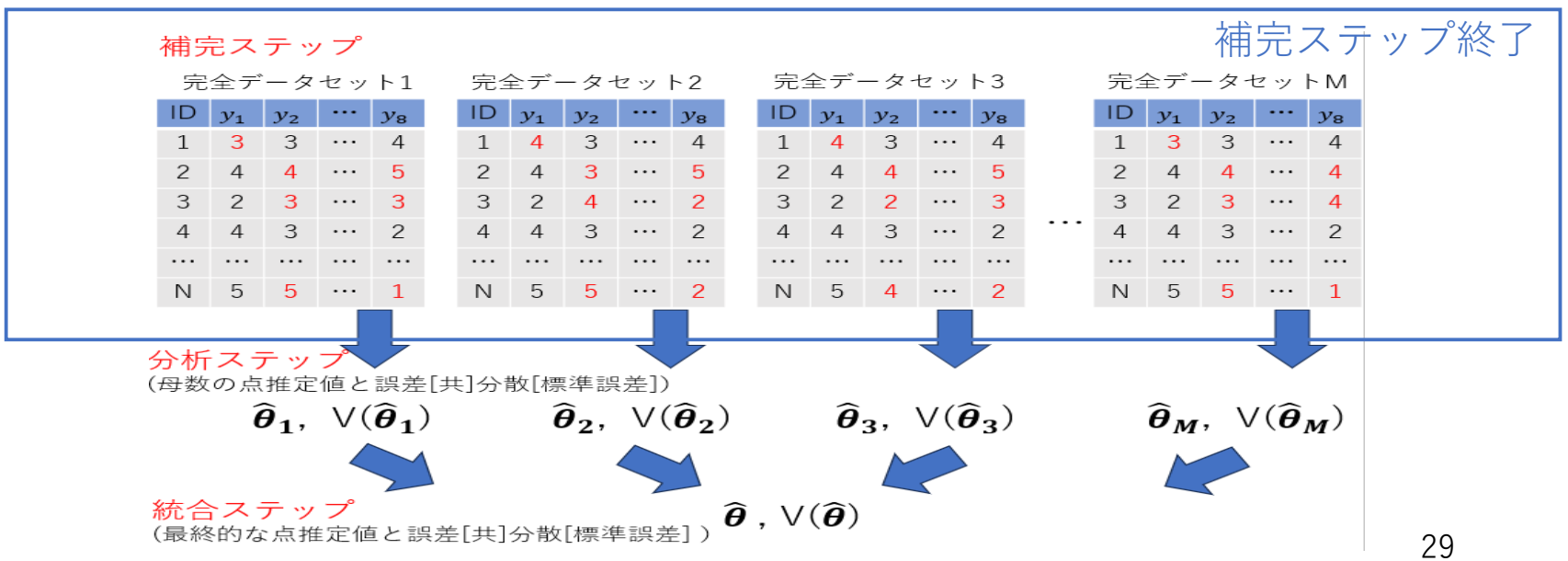
- y_8内にある欠測値を、疑似的な完全データセット内のy_1, $y_2 . . . , y_7$を用いたy_8の補完モデルから生成された補完値により補完し更新する。

ID	y_1	y_2	…	y_8
1	**3.1**	3	…	4
2	4	**2.9**	…	1.9⇒**1.7**
3	2	**1.8**	…	3.2⇒**3.5**
4	4	3	…	2
…	…	…	…	…
N	5	**4.6**	…	2.3⇒**2.0**

MICEによる補完ステップ

(iv)(iii)の更新の反復

(ii)の初期値は通常ラフなものであり、T回の反復（後述のmice関数では、”maxit”に対応）を経て単一の疑似的な完全データセットを得る。そして、ここまでの一連の作業をM回行いM個の完全データセットを得る。



補足：予測平均マッチング

- 線形回帰モデルを用いた補完で、特に残差の非正規性や変数間の非線形的な関係が疑われる場合には、予測平均マッチング（predictive mean matching: PMM）が利用されることも多い。

- 変数 yに欠測のある個人i について、補完モデルを基に生成された予測値y_i^*と、yが観測されている個人について計算された予測値$\hat{y}$との距離が近い個人を複数人選択し、そこからランダムに選ばれた1 名の個人i'の観測値$y_{i'}$を用いて個人iの欠測値を補完する。

- このように補完値として観測値を利用することで、元々のデータの分布を反映した補完が実現できる。

補足：予測平均マッチング（y_1を補完する場合）

個人1のデータが欠測

ID	y_1	y_2	…	y_8
1	-	3	…	4
2	4	2.6	…	1.9
3	2	2.1	…	3.2
4	4	3	…	2
…	…	…	…	…
N	5	4.3	…	2.3

各個人の予測値を算出（赤字部分）

ID	y_1	y_2	…	y_8
1	**3.2**	3	…	4
2	4 (3.8)	2.6	…	1.9
3	2 (2.4)	2.1	…	3.2
4	4 (3.3)	3	…	2
…	…	…	…	…
N	5 (3.1)	4.3	…	2.3

予測値が近い個人を複数選択

ID	y_1	y_2	…	y_8
1	**3.2**	3	…	4
2	4 (3.8)	2.6	…	1.9
3	2 (2.4)	2.1	…	3.2
4	4 (3.3)	3	…	2
…	…	…	…	…
N	5 (3.1)	4.3	…	2.3

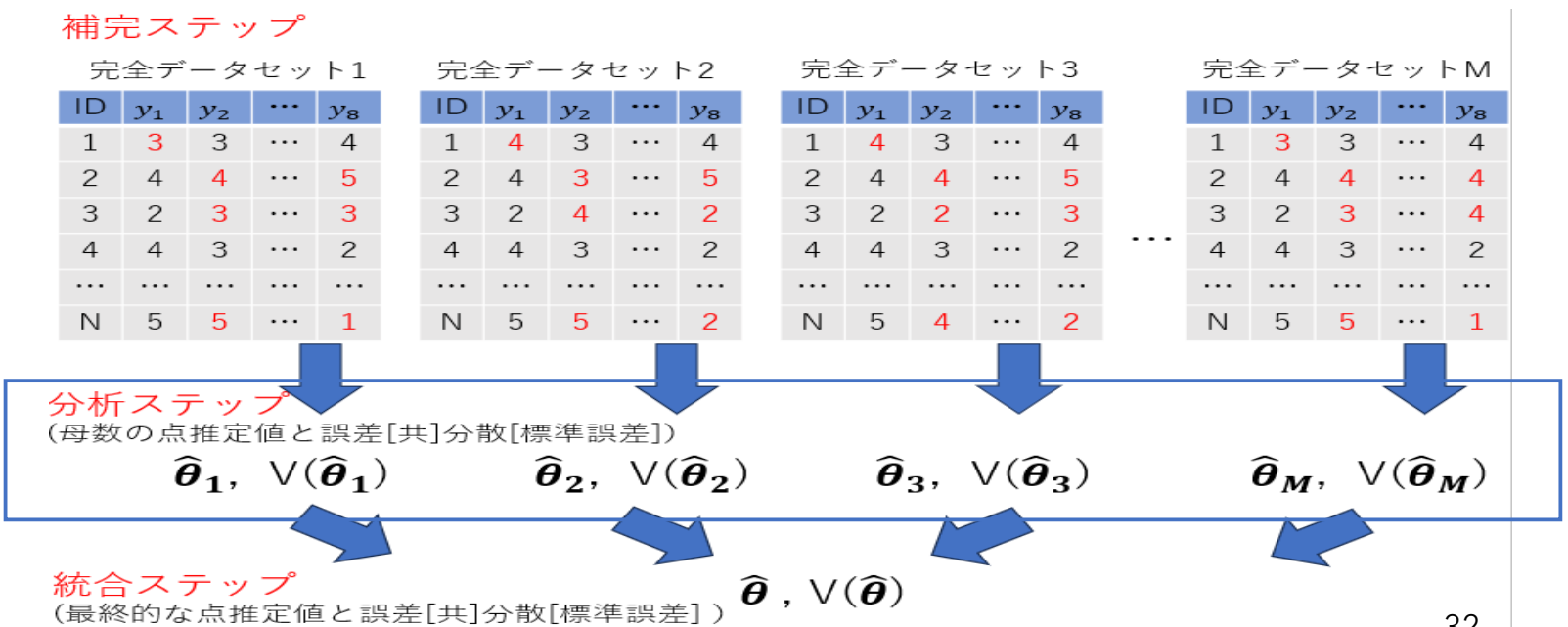
その中からランダムに選ばれた1名の観測値により補完（ここでは個人4）

ID	y_1	y_2	…	y_8
1	4	3	…	4
2	4	2.6	…	1.9
3	2	2.1	…	3.2
4	4	3	…	2
…	…	…	…	…
N	5	4.3	…	2.3

分析ステップ

- 分析の段階：得られたM個の完全データセットに対して、分析モデル（e.g.,確認的因子分析モデル）をそれぞれあてはめる。

- 分析モデル内の母数$\boldsymbol{\theta}$について、M種類の点推定値と標準誤差（または誤差共分散行列）が得られる。



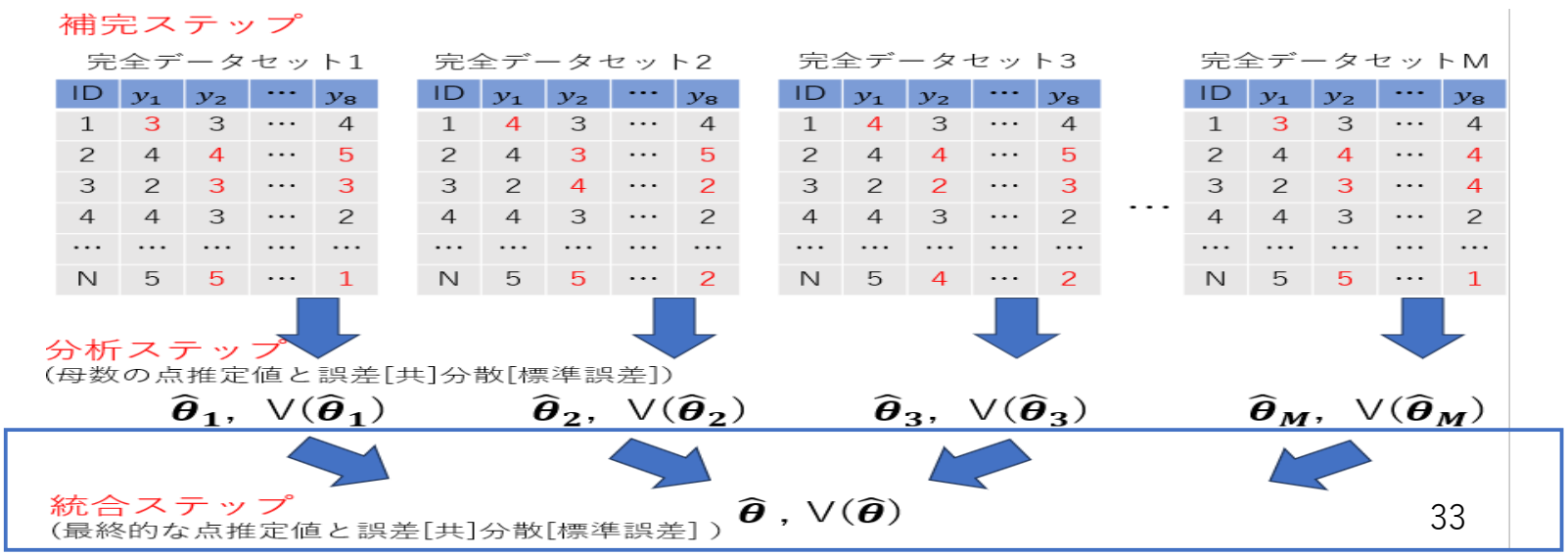
統合ステップ

統合の方法（Rubin’ s rule）：

- 点推定値（$\hat{\boldsymbol{\theta}}$）として、各完全データセットから得られた推定値（$\hat{\boldsymbol{\theta}}_{\boldsymbol{m}}; m = 1,2 \ldots M$)の平均を利用する。すなわち、

$$\hat{\boldsymbol{\theta}} = \frac{1}{M}\sum_{m=1}^{M}\hat{\boldsymbol{\theta}}_{\boldsymbol{m}}$$

である。



統合ステップ

- $\hat{\boldsymbol{\theta}}$の誤差共分散行列$\mathrm{V}(\hat{\boldsymbol{\theta}})$は、各完全データセットから得られた$\hat{\boldsymbol{\theta}}_{\boldsymbol{m}}$の誤差共分散行列の推定値$\mathrm{V}(\hat{\boldsymbol{\theta}}_{\boldsymbol{m}})$を利用して、

$$\mathrm{V}(\hat{\boldsymbol{\theta}}) = \boldsymbol{W_M} + (1 + \frac{1}{M})\,\boldsymbol{B_M}$$

となる(e.g., 高井他, 2016, pp.117-118)。ここで、

$$\boldsymbol{W_M} = \frac{1}{M}\sum_{m=1}^{M} \mathrm{V}(\hat{\boldsymbol{\theta}}_{\boldsymbol{m}}) \quad , \qquad \boldsymbol{B_M} = \frac{1}{M-1}\sum_{m=1}^{M} (\hat{\boldsymbol{\theta}}_{\boldsymbol{m}} - \hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}_{\boldsymbol{m}} - \hat{\boldsymbol{\theta}})^{\mathrm{T}}$$

であり、$\boldsymbol{W_M}$および$\boldsymbol{B_M}$はそれぞれ、補完値内・補完値間の共分散行列と呼ばれる。 *M個の誤差分散の推定値を単に平均するだけではなく、M個の推定値間の変動も考慮。

・特定の母数θに関する標準誤差の推定値 $\mathrm{se}(\hat{\theta})$は、$\mathrm{V}(\hat{\boldsymbol{\theta}})$の対応する対角要素の正の平方根に等しい。

統合ステップ

- 特定の母数θに関する帰無仮説（H_0: $\theta = 0$）の検定：

$$t = \frac{\hat{\theta}}{\mathrm{se}(\hat{\theta})}$$

- θ に関する$100(1 - \alpha)\%$ 信頼区間：

$$\hat{\theta} \pm t_{v,\alpha/2}\ \mathrm{se}(\hat{\theta})$$

＊$t_{v,\alpha}$は自由度v のt分布の上側$100\alpha\%$ 点

自由度vの1つの推定量として、

$$v = (M-1)(1+\frac{1}{r})^2, \qquad r = (1+\frac{1}{M})\frac{B_M}{W_M}$$

＊B_M, W_Mは、対応する$\boldsymbol{B_M}$および$\boldsymbol{W_M}$の（対角）要素

補足

- SEMにより表現可能な分析モデルを扱う場合、例えばR のsemTools パッケージやmitml パッケージ（Grund, Robitzsch, & Ludtke, 2021）を用いて、推定結果の統合や検定を行うことができる。

- Enders(2023, p.9) のレビューでは、尤度比検定を行う場合も含め、統合ステップでの推測法に関する最新の知見がまとめられている。

- Lee & Cai (2012) およびEnders & Mansolf (2018) …MI を適用した際のSEM の検定統計量および適合度指標の算出について。⇒R のsemTools パッケージが利用できる。

- Liu et al (2021)…順序データの欠測に対してMI を適用した際の適合度の評価について。

補足：疑似的な完全データセット数M について

- 従来M =5,10程度で十分とされてきたが、近似推測法であるMIにおいては、十分な数のMが必要（野間、2017, p.69）。

- Graham et al. (2007) は$M = 20$ を推奨し、またHuque et al. (2018) のシミュレーションでは$M = 40$ である。

- 野間（2017, p.69）では、$M = 100 - 1000$ 程度であっても現在の計算機環境であれば必ずしも大きな負荷とならず、そのため十分な数のMを設定することが望ましいと述べている。

- 特に欠測の割合が高いときには、より大きなM が求められる。大まかな目安として、少なくとも$M = 20$、可能であれば$M = 50, 100$程度は確保したい。

FIMLとMIの分析例（確認的因子分析）

・動機づけに関する計8変数を含む人工データ(*N*=300)。 MARに基づく欠測を仮定し、 y_1 ～y_8全体で欠測の割合は3.5%。

・ y_1～y_4が内発的動機づけ因子（Int）を、 y_5 ～y_8が外発的動機づけ因子（Ext）を反映する2因子の確認的因子分析モデル(CFA)の推定に関心がある状況を考える。

・FIML とMI（MICE とPMMによる補完、 $M = 100$）を使って、CFA内の母数を推定。

・MI において、各完全データセットにモデルをあてはめる際には最尤推定を用いた。また、例証のため、リストワイズ法(N=209)、および（通常は得られない）欠測のない完全データ(N=300）に基づく分析（いずれも最尤推定）も実施。

	A	B	C	D	E	F	G	H	I	J	K
1	school	gender	y1	y2	y3	y4	y5	y6	y7	y8	score
2	1	0	3	3	3	5	5	5	5	5	49
3	1	0	4	4	4	3	5	2	4	3	51
4	1	0	2	3	1	4	4	4	4	NA	51
5	1	0	4	3	3	4	4	4	4	3	59
6	1	0	4	3	4	3	5	2	4	5	49
7	1	0	3	3	3	4	4	3	4	3	65
8	1	0	4	3	3	4	3	4	4	4	53
9	1	0	5	5	5	5	5	NA	5	3	66
10	1	0	4	4	4	5	5	5	5	4	59
11	1	0	3	3	3	4	3	NA	3	5	53
12	1	0	3	3	3	3	3	3	3	3	50
13	1	0	3	3	3	3	3	3	3	3	55
14	1	0	4	4	5	3	4	4	NA	NA	50
15	1	0	5	5	5	5	5	5	5	5	67

推定結果（因子負荷と因子間相関）

	MI (N=300)		FIML (N=300)		Listwise (N=209)		Complete (N=300)	
	推定値	標準誤差	推定値	標準誤差	推定値	標準誤差	推定値	標準誤差
Int ⇒x1	0.632	0.084	0.634	0.094	0.462	0.102	0.612	0.082
Int ⇒x2	0.745	0.075	0.744	0.079	0.736	0.100	0.755	0.073
Int ⇒x3	0.554	0.079	0.560	0.087	0.471	0.098	0.532	0.077
Int ⇒x4	0.562	0.076	0.562	0.078	0.543	0.097	0.581	0.074
Ext ⇒x5	0.958	0.073	0.958	0.072	0.917	0.089	0.958	0.071
Ext ⇒x6	1.209	0.069	1.214	0.069	1.155	0.083	1.222	0.067
Ext ⇒x7	0.838	0.062	0.846	0.063	0.863	0.075	0.825	0.061
Ext ⇒x8	0.698	0.079	0.694	0.082	0.669	0.094	0.703	0.077
Int⇔Ext	0.398	0.069	0.393	0.070	0.398	0.087	0.403	0.067
CFI	0.906		0.909		0.896		0.913	
RMSEA	0.099		0.098		0.099		0.099	
SRMR	0.065		0.064		0.068		0.064	

*Completeは（通常は得られない）欠測値のない完全データセットを分析した場合
*CFI, RMSEA, SRMRはモデルの適合度指標。

推定結果

- 各方法においてCFAのあてはまりは良好であり、またMI とFIML の推定値には大きな違いは見られない。

- いまMAR に基づく欠測であることを反映して、完全データ（Complete）とMI およびFIML の点推定値は類似している。欠測があることを反映して、これらにおける標準誤差は完全データの場合と比べて若干ではあるが大きくなる。

- リストワイズ法では他と比べて推定値に乖離（過小推定）が生じている。標準誤差も、完全データの場合と比べて概ね10%-20%程大きくなっている。⇒検定力、更には研究の結論にも影響し得る。

アウトライン

- FIMLの概要
- MIの概要
- 補助変数の活用
- まとめとMNARの場合

補助変数ーMARかMNARかー

- FIML やMI ではMAR に基づく欠測を仮定している。これらの分析が正当化されるためには、欠測の生起（r）を説明できる観測変数（y_{obs}）が適切に分析モデル内に投入される必要がある。

- 一方で、欠測の生起（r）および欠測値（y_{mis}）を説明できる変数が実際にどの程度観測でき、また分析モデルに反映されているのかに関する度合いには幅がある。

- その意味で、MAR の仮定が実際にどれだけ充たされているのかという問いは、程度問題と言える（Graham, 2009; Newsom, 2015）。

補助変数の意義

- 特に欠測の割合が多いとき、分析モデルには元々含まれていないが、rやy_{mis}と相関があると考えられる観測変数を収集しモデルに投入することで、MAR の蓋然性を高められる可能性がある。

- このような観測変数は補助変数（auxiliary variable）と呼ばれ、仮にそれが欠測の直接的な原因となっていなくとも、投入により推定値のバイアスが低減し標準誤差も小さくなることが期待される。

- 縦断デザインにおいて（分析モデルに投入されていない）過去のラグ付き変数やベースラインの情報はしばしば有用。また、特に大規模調査データ・ビッグデータを扱う場合は補助変数の候補は多数あり得る。

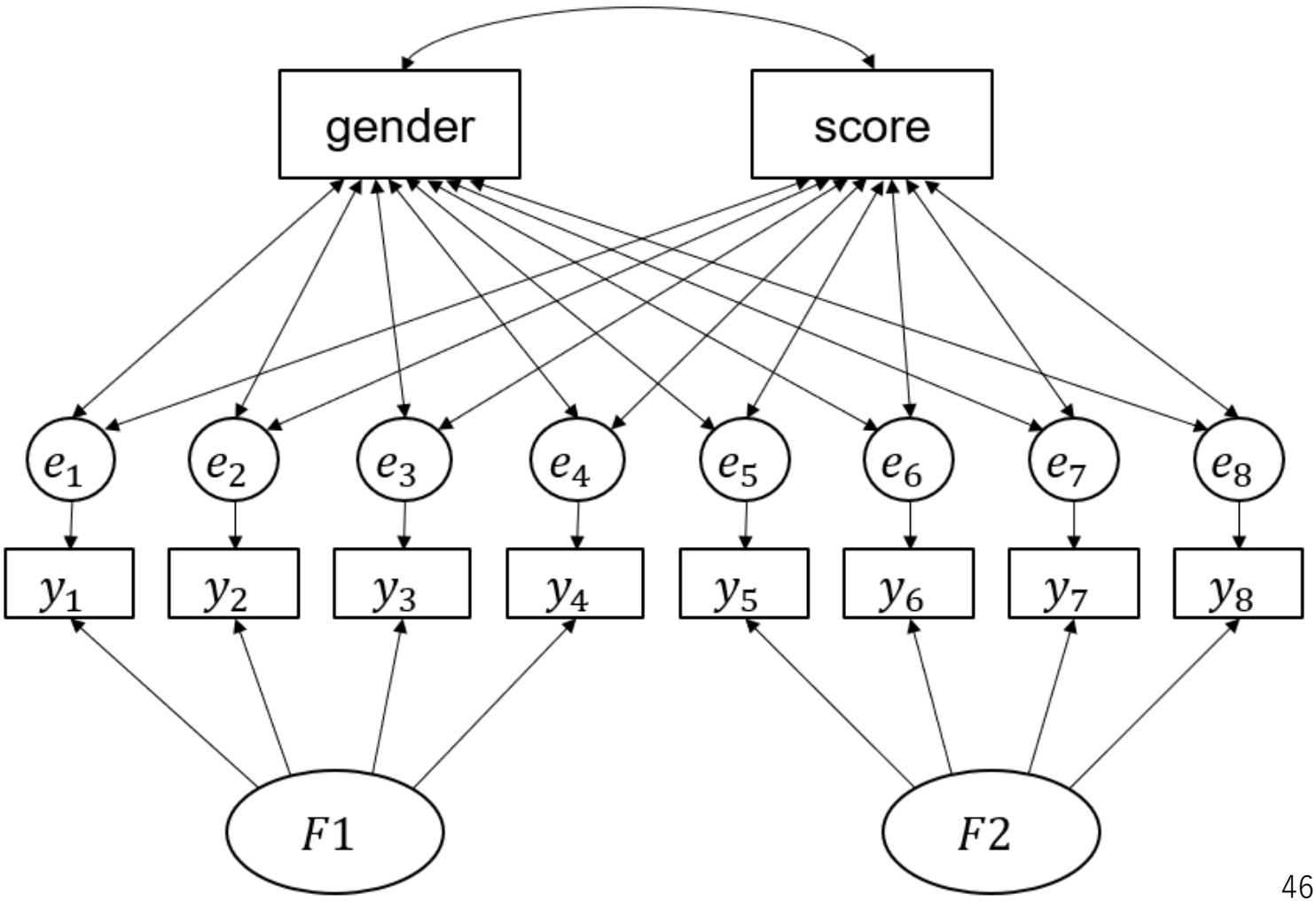
補助変数を考慮した分析

- MI では分析モデルと補完モデルが明確に区別されているため、収集した補助変数を補完モデルに含めて分析を実行すればよい。

- SEM のFIML において補助変数を考慮した分析アプローチは幾つか知られているが（e.g., Newsom, 2015, pp.18-25; Enders, 2023; pp.5-6）、飽和した相関アプローチ（saturated correlated approach）は簡便。Mplus やR のSemTools パッケージで実装できる。

- この方法では、モデル内に元々投入されている変数（の残差）と補助変数間の相関を仮定したモデルを新たに設定することで、当初の分析モデルの構造に影響を与えずに補助変数を考慮する。

飽和した相関アプローチ（CFA）



補足

・経験的に、rやy_{mis}との相関がかなり高い補助変数が投入されない限り、分析結果に与える変化は小さいことが多い。

・SEMで表現可能な分析モデルを扱う際、モデル適合に関する検定統計量（χ^2）や自由度は補助変数の投入前後で変わらないため、RMSEA の指標値も変化しない。

・CFI の算出に際しては、あくまで（投入前の）元々の分析モデルの適合を吟味することが目的のため、補助変数により導入された相関を含めたモデルを独立モデルとして設定する。

・補助変数に関わる共分散行列内の要素は（飽和しているため）適合が完全になり、SRMR を算出すると適合が過大評価される。そのため、補助変数を除外した上での算出が推奨される。

・RのSemTools パッケージでは以上の点を考慮した指標値が返される。

アウトライン

- FIMLの概要
- MIの概要
- 補助変数の活用
- まとめとMNARの場合

まとめ

・特にSEMの下位モデルを扱う場合のように、直接尤度の設定・評価が容易に実行できる状況では、FIMLを利用すればよい。モデルが正しく設定されていれば有効性（標準誤差）の観点からも優れている。

・特に欠測の割合が大きいとき、分析モデルに含まれないが、欠測を説明するのに有用な補助変数があれば、それを含めた分析（e.g., 飽和した相関アプローチ）も有用。

・補完モデルと分析モデルを明確に区別したMI、特にMICEは汎用性の高い方法であり、様々な分析モデルに対して柔軟に適用可能であり、ソフトウェア上の実装も容易。

補足ーFIMLとMIの比較と選択ー

・モデルが正しく設定されていれば，FIMLとMIが互いにかなり類似した結果を示すことは経験的にもよく知られている（本資料の分析例、およびGraham, 2009; Lee & Shi, 2021）。

・一方で、実際にはモデルの誤設定を避けることは非常に困難であり、このときFIMLとMIの間で推定結果に大きな乖離が生じる可能性もある（e.g., Lee & Shi, 2021）。このような点を含めたFIMLとMIの比較と選択については、現在でも研究・議論の余地がある。

よりよい分析実践のために

- 欠測データの分析に際しては、欠測データメカニズムや各分析法に内在する仮定を吟味しながら適切な分析方法を選択していくことが求められる。

- 分析結果の報告に際して、欠測の割合やその処理方法が明記されていないケースは多い。例えば、経営学や心理学領域での文献調査を行ったZyphur et al (2023) では、処理方法について説明があった論文は全体の34% であったことを報告している。

- また、分析上の工夫だけではなく、様々なデータ収集上の工夫も重要である。

- たとえば質問紙調査の場合に、内容が不必要に複雑で理解や回答のしにくい質問項目を修正・削除することで欠測が生じるリスクを下げることや、欠測の有無を説明できる補助変数を予め吟味し収集することなどが挙げられる。

MNARの可能性と感度分析

- MAR（およびMCAR）に基づく欠測とは考えられず、また有力な補助変数の情報が十分得られない（または、提供されている多重代入データを使う場合に補助変数の情報が十分反映されていない）場合、すなわちMNAR に基づく欠測である場合は、FIML やMI による推測結果には大きなバイアスを伴う可能性がある。

- ＭＮARにおいては、欠測指標rについてのモデリングが必要。

MNARの可能性と感度分析

- MNARの場合の分析法として、選択モデル、混合モデルなどがある（Enders, 2011; Newsom, 2015; 高井他, 2016）。

- ただし現状、絶対的に優れた方法があるとは言えない。

- MNAR に基づく欠測が想定される場合には、感度分析の実行は有用。異なる方法に基づく推定結果の間に大きな乖離が見られないのであれば、方法の選択如何が最終的な結論に与える影響は小さいものと結論づけられる。

- もし推定結果に大きな乖離が見られるのであれば、実質科学的な見地や先行研究等の外的な情報も踏まえながら、判断され得る結論の範囲を示すことが求められる。

引用文献

- Asparouhov, T. & Muthen, B. (2022). Multiple imputation with Mplus. https://www.statmodel.com/download/Imputations7.pdf
- Enders, C. K. (2023). Missing data: An update on the state of the art. *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000563
- Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology, 60*, 549–576.
- Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science, 8*, 206–213.
- Grund, S., Robitzsch, A., & Ludtke, O. (2021). Package "mitml" [Computer software]. https://cran.r-project.org/web/packages/mitml/mitml.pdf
- Huque, M. H., Carlin, J. B., Simpson, J. A. & Lee, K. J. (2018). A comparison of multiple imputation methods for missing data in longitudinal studies. *BMC Medical Research Methodology, 18*, 1-16.
- Lee, T., & Cai, L. (2012). Alternative multiple imputation inference for mean and covariance structure modeling. *Journal of Educational and Behavioral Statistics, 37*, 675–702.
- Lee, T., & Shi, D. (2021). A comparison of full information maximum likelihood and multiple imputation in structural equation modeling with missing data. *Psychological Methods, 26* (4), 466–485.
- Liu, Y., Sriutaisuk, S., & Chung, S. (2021). Evaluation of model fit in structural equation models with ordinal missing data: A comparison of the D2 and MI2S methods. *Structural Equation Modeling: A Multidisciplinary Journal, 28* (5), 740–762.

引用文献

- Newsom, J. T. (2015). *Longitudinal structural equation modeling: A comprehensive introduction*. New York: Routledge.
- 野間久史 (2017). 連鎖方程式による多重代入法　応用統計学, *46*(2), 67-86.
- Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*(2), 1–36.
- Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons Inc., New York.
- Savalei, V., & Rosseel, Y. (2022). Computational options for standard errors and test statistics with incomplete normal and nonnormal data. *Structural Equation Modeling: A Multidisciplinary Journal*, *29* (2), 163–181.
- 高井啓二・星野崇宏・野間久史 (2016). 欠測データの統計科学―医学と社会科学への応用 岩波書店
- Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). Mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45* (3), 1–67.
- Yuan, K.-H., & Bentler, P. M. (2010). Consistency of normal distribution based pseudo maximum likelihood estimates when data are missing at random. *American Statistician*, *64* (3), 263–267.
- Zhang, X., & Savalei, V. (2023). New computations for RMSEA and CFI following FIML and TS estimation with missing data. *Psychological Methods*, *28*(2), 263–283.
- Zyphur, M. J., Bonner, C. V., & Tay, L. (2023). Structural equation modeling in organizational research: The state of our science and some proposals for its future. *Annual Review of Organizational Psychology and Organizational Behavior*, *10*, 495–517.

日本教育心理学会第66回総会
学会企画チュートリアル・セミナー
「教育心理学研究のための欠測データ処理」

統計ソフトウェアRでの分析例

鈴木 雅之（横浜国立大学）

欠測データ処理のための代表的なパッケージ

- 完全情報最尤推定法（FIML法）
  - lavaan, sem
    - いずれも，SEMによる分析を行うためのパッケージ
- 多重代入法（MI法）
  - Amelia
    - EMB（Expectation-Maximization with Bootstrapping）によるMI
  - mice
    - FCS（Fully Conditional Specification）によるMI
  - semTools (+ lavaan)
    - AmeliaとmiceパッケージでMI作成された疑似完全データに対する分析・統合が可能

発表の概要

- 因子分析と回帰分析，*t* 検定，分散分析を行う方法を紹介（ただし，探索的因子分析はFIML，分散分析はMIのみ）
  - FIML法とMI法による分析を同じ枠組みで行えるよう，主にlavaanとsemTools, mice パッケージを用いる
  - *t* 検定と分散分析は，回帰分析の枠組みで行う
    - semTools では，mice で作成された疑似完全データに対して，SEM の枠組みで分析・統合が可能
    - semTools の関数runMI() では，補完・分析・統合が同時に実行できるが，本発表では取り上げない
    - mice やAmelia の詳細な使い方は高橋・渡辺（2017），lavaan やsemTools は豊田（2014）などを参照のこと

データの内容

- 300名の人工データ（欠測箇所には「.」（ピリオド）が挿入）
  - school: 学校種（1＝公立，2＝国立，3＝私立）
  - gender: 性別（0＝男性，1＝女性）
  - y1-y4: 内発的動機づけを測定する項目（5件法）
  - y5-y8: 外発的動機づけを測定する項目（5件法）
  - score: テスト得点

「example.csv」（一部抜粋）

	A	B	C	D	E	F	G	H	I	J	K
1	school	gender	y1	y2	y3	y4	y5	y6	y7	y8	score
2	1	0	3	3	3	5	5	5	5	5	49
3	1	0	4	4	4	3	5	2	4	3	51
4	1	0	2	3	1	4	4	4	4	.	51
5	1	0	4	3	3	4	4	4	4	3	59
6	1	0	4	3	4	3	5	2	4	5	49
7	1	0	3	3	3	4	4	3	4	3	65
8	1	0	4	3	3	4	3	4	4	4	53
9	1	0	5	5	5	5	5	.	5	3	66
10	1	0	4	4	4	5	5	5	5	4	59
11	1	0	3	3	3	4	3	.	3	5	53

データの読み込み

- 関数read.csv() の引数na.strings で欠測値を指定

例）データをdat_mis というオブジェクトに保存

```
dat_mis <- read.csv(file.choose(), na.strings = ".")
```

「example.csv」（一部抜粋）

	A	B	C	D	E	F	G	H	I	J	K
1	school	gender	y1	y2	y3	y4	y5	y6	y7	y8	score
2	1	0	3	3	3	5	5	5	5	5	49
3	1	0	4	4	4	3	5	2	4	3	51
4	1	0	2	3	1	4	4	4	4	.	51
5	1	0	4	3	3	4	4	4	4	3	59
6	1	0	4	3	4	3	5	2	4	5	49
7	1	0	3	3	3	4	4	3	4	3	65
8	1	0	4	3	3	4	3	4	4	4	53
9	1	0	5	5	5	5	5	.	5	3	66
10	1	0	4	4	4	5	5	5	5	4	59
11	1	0	3	3	3	4	3	.	3	5	53

データフレームの確認

- 関数head() を利用して，最初の数行を表示
  - head(データフレーム, 表示する行数)

例）最初の8行を表示

```
head(dat_mis, 8)
```

R Console

```
> # 最初の8行を表示
> head(dat_mis, 8)
  school gender y1 y2 y3 y4 y5 y6 y7 y8 score
1      1      0  3  3  3  5  5  5  5  5    49
2      1      0  4  4  4  3  5  2  4  3    51
3      1      0  2  3  1  4  4  4  4 NA    51
4      1      0  4  3  3  4  4  4  4  3    59
5      1      0  4  3  4  3  5  2  4  5    49
6      1      0  3  3  3  4  4  3  4  3    65
7      1      0  4  3  3  4  3  4  4  4    53
8      1      0  5  5  5  5  5 NA  5  3    66
>
```

欠測値はNAと表示される

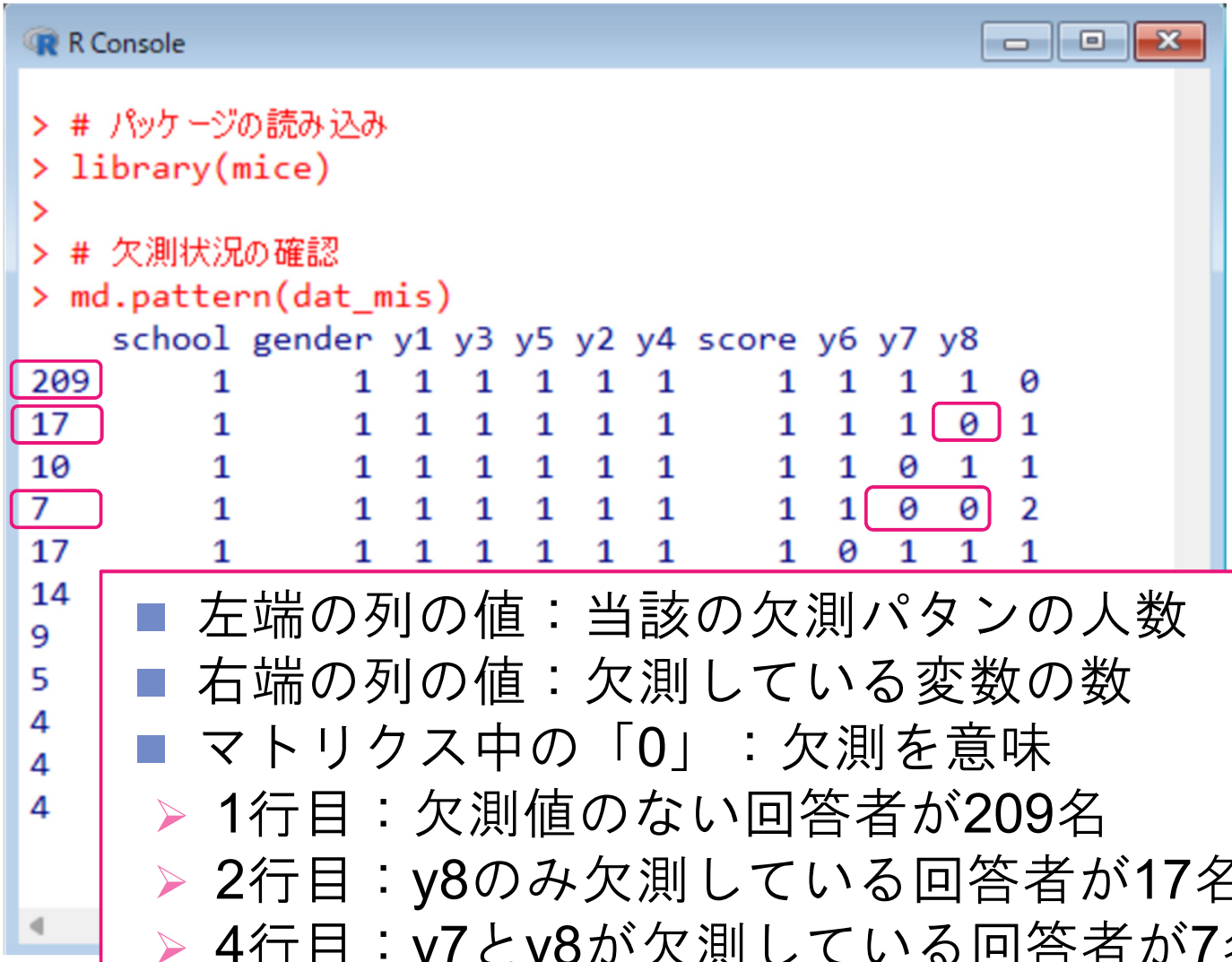
miceパッケージを用いた欠測の補完

欠測状況の確認①

- miceパッケージの関数md.pattern() を利用して，データの欠測状況を確認
  - md.pattern(データフレーム)

例）library(mice)
　　md.pattern(dat_mis)

欠測状況の確認②



- 左端の列の値：当該の欠測パタンの人数
- 右端の列の値：欠測している変数の数
- マトリクス中の「0」：欠測を意味
  - 1行目：欠測値のない回答者が209名
  - 2行目：y8のみ欠測している回答者が17名
  - 4行目：y7とy8が欠測している回答者が7名

欠測状況の確認③

R Console

```
> # パッケージの読み込み
> library(mice)
>
> # 欠測状況の確認
> md.pattern(dat_mis)
     school gender y1 y3 y5 y2 y4 score y6 y7 y8
209       1      1  1  1  1  1  1     1  1  1  0
17        1      1  1  1  1  1  1     1  1  1  1
10        1      1  1  1  1  1  1     1  1  0  1
7         1      1  1  1  1  1  1     1  1  0  0
17        1      1  1  1  1  1  1     1  0  1  1
14        1      1  1  1  1
9         1      1  1  1  1
5         1      1  1  1  1
4         1      1  1  1  0
4         1      1  1  0  1
4         1      1  0  1  1  1  1     1  1  1  1
          0      0  4  4  4  5  9    14 17 17 24 98
```

- school と gender は欠測なし
- y1とy3, y5は4名が欠測
- y2は5名が欠測

欠測状況の報告例

Motiv Emot (2018) 42:178–189
DOI 10.1007/s11031-017-9646-2

ORIGINAL PAPER

Basic psychological needs and work motivation: A longitudinal test of directionality

Anja H. Olafsen[1] · Edward L. Deci[1,2,3] · Hallgeir Halvari[1]

Table 2 Item non-response and wave non-response across the 15-month study period

Variable	Item non-response (of the composite variable)				Wave non-response			
	Time 1 (%)	Time 2 (%)	Time 3 (%)	Time 4 (%)	Time 1 (%)	Time 2 (%)	Time 3 (%)	Time 4 (%)
Need support	1.5	3.2	1.3	0.0	0.0	30.7	43.1	56.9
Need satisfaction	4.1	1.6	2.6	0.0	0.0	30.7	43.1	56.9
Autonomous motivation	4.5	4.3	3.9	0.0	0.0	30.7	43.1	56.9

AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Educational Psychology

ISSN: 0022-0663

2024, Vol. 116, No. 6, 1019–1033
https://doi.org/10.1037/edu0000898

What Are the Long-Term Prospects for Children With Comprehension Weaknesses? A Registered Report Investigating Education and Employment Outcomes

Emma James[1, 2], Paul A. Thompson[3], Lucy Bowes[1], and Kate Nation[1]

Table S3

Data availability (n; %) in covariate and outcome variables for each reading group

	Comprehension weakness (n = 947)	Word reading weakness (n = 1383)	No reading weakness (n = 4516)
Covariates			
Sex	947 (100%)	1382 (99.93%)	4505 (99.76%)
Maternal education	876 (92.50%)	1271 (91.90%)	4104 (90.88%)
Free school meal status	800 (84.48%)	1127 (81.49%)	3780 (83.70%)
Outcome variables			
SATs (Year 6)[a]	807 (85.22%)	1171 (84.67%)	3882 (85.96%)
SATs (Year 9)[a]	722 (76.24%)	1007 (72.81%)	3372 (74.67%)
GCSEs[a]	794 (83.84%)	1113 (80.48%)	3738 (82.77%)
NEET status	396 (41.82%)	625 (45.19%)	2001 (44.31%)

欠測の補完①

- miceパッケージの関数mice() を利用
  - mice(データフレーム, method = 代入方法,
    m = 作成するデータセットの数, maxit = 反復回数,
    seed = シード値)
    - 代入法には予測平均マッチング（pmm）や
      線形回帰（norm），ロジスティック回帰（logreg）など

    → 変数のタイプに応じてデフォルトの方法は変わる
    （連続変数はpmm，2値変数はlogregなど）
    - 作成するデータセット数のデフォルトは5
    - 反復回数のデフォルトは5
    - 再現性を確保するための任意のシード値を設定

欠測の補完②

例）以下の条件で補完を行い，dat_MI に保存

- 代入法：予測平均マッチング（pmm）
- 作成するデータセット数：100
- 反復回数：30
- シード値：23109（任意の値）
- printFlag = FALSE：コンソール画面に反復情報を出力しないように設定

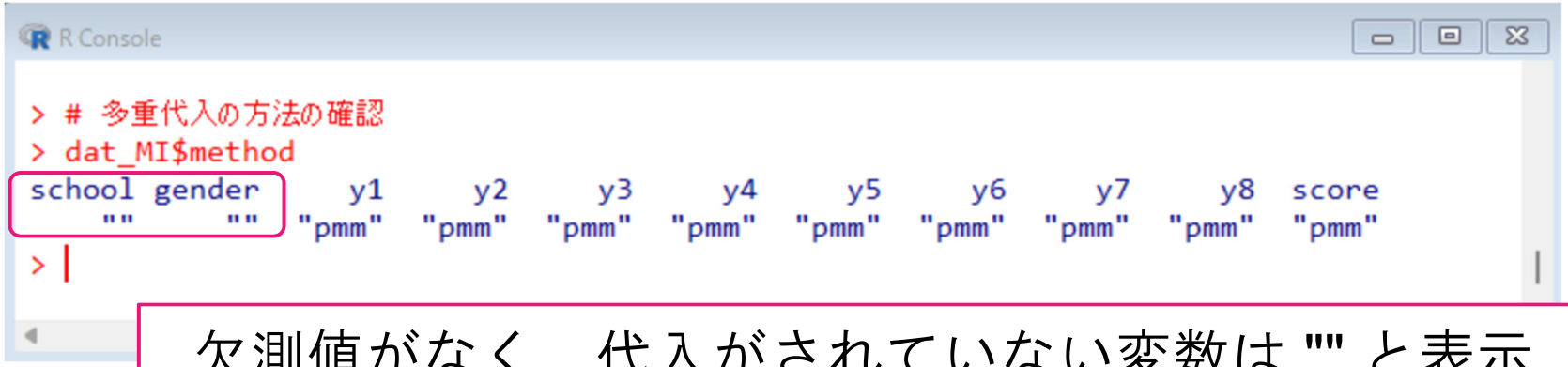
チュートリアル中，試しに実行する場合は20程度を推奨

```
dat_MI <- mice(data = dat_mis, method = "pmm",
m = 100, maxit = 30, seed = 23109, printFlag = FALSE)
```

代入法の確認

- 代入法の確認
  - 疑似完全データ$method

例）dat_MI$method



代入された値の確認①

- 代入された値の確認
  - 特定の1つの変数について確認
    - 疑似完全データimp変数名
  - 複数の変数について確認
    - 疑似完全データ$imp[c(列番号, 列番号…)]

例）変数y1 に代入された値の確認

dat_MIimpy1

例）3～11列目の変数に代入された値の確認

dat_MI$imp[c(3:11)]

代入された値の確認②

R Console

```
> ## 変数y1 に代入された値の確認
> dat_MI$imp$y1
    1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29
59  3  4  4  2  4  4  4  4  4  4  5  3  4  3  3  3  3  3  3  1  3  3  3  3  2  2  3  4  1
64  3  3  5  4  4  5  5  3  5  3  4  3  3  5  5  3  4  4  4  3  5  4  5  3  2  4  4  5  4
190 3  5  5  3  5  3  4  3  5  5  4  5  4  5  5  5  4  5  5  4  3  5  5  5  5  5  5  5  5
265 2  3  3  4  3  4  3  2  5  3  4  3  4  3  4  4  4  2  3  3  2  3  4  2  3  3  3  4  3
    30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55
59   4  4  3  4  3  4  3  3  3  3  4  1  4  4  4  3  3  4  4  2  3  3  4  4  3  3
64   3  4  5  4  2  5  3  4  4  4  2  1  4  3  2  2  1  5  4  3  3  4  3  1  3  4
190  5
265  3
    56
59   1
64   4
190  3
265  5  3  3  3  3  3  3  4  3  3  3  3  3  3  3  3  4  4  3  4  3  3  4  3  1  3
    82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100
59   4  4  4  3  4  3  3  3  4  2  4  4  3  3  4  3  3  3   3
64   4  3  3  5  4  1  5  4  2  4  5  5  4  5  3  3  3  3   3
190  5  5  4  3  5  5  1  3  5  4  5  5  5  5  5  5  3  4   3
265  3  3  2  4  2  4  1  4  3  2  3  1  3  2  3  1  3  2   3
> 
```

- 変数y1 について，59, 64, 190, 265行目の4名が欠測
  - 1つ目のデータセットでは，59, 64, 190行目の回答者は「3」，265行目は「2」が代入された

疑似完全データの確認

- 疑似完全データの確認
  - complete(疑似完全データ, データの番号)

例）1つ目の疑似完全データの確認

```
complete(dat_MI, 1)
```

例）3つ目の疑似完全データの最初の8行の確認

```
head(complete(dat_MI, 3), 8)
```

【参考】変数ごとに代入法を指定

- 関数mice() の引数method では，変数ごとに代入法の指定が可能
  - 補完の不要な変数については，ダブルクォーテーション（" "）の中をブランクにする

例）y1～y4は線形回帰（norm），
y5～y8とscore は予測平均マッチング（pmm）で補完
（school とgender は補完しない）

```
dat_MI2 <- mice(data = dat_mis, m = 100, maxit = 30, seed = 23109,
method = c(" ", " ", "norm", "norm", "norm", "norm",
"pmm", "pmm", "pmm", "pmm", "pmm"),  printFlag = FALSE)
```

確認的因子分析

lavaanパッケージの基本的な使い方①

- モデルの記述
  - シングルクオーテーション（'）で囲んだ部分で
    モデルを記述し，オブジェクトに保存
    - 3種類の記号を用いて，変数間の関係を記述
    - 変数が複数ある場合には「+」でつなぐ

記号	意味	例	例の意味
=~	測定方程式	f =~ x1 + x2 + x3	因子f は，x1とx2, x3の 3つの観測変数に影響する
~	構造方程式	y ~ x1 + x2	yは，x1とx2からパスを受ける
~~	共分散	x1 ~~ x2	x1 とx2 の間の共分散
		x1 ~~ x1	x1 の分散

lavaanパッケージの基本的な使い方②

- 母数の推定
  - lavaanパッケージの関数sem() や関数cfa() を用いて，モデルを記述したオブジェクトを指定して推定
    - sem(モデルを記述したオブジェクト，
      data = データフレーム)

      → 引数に「missing = "fiml"」を加えることでFIML法による推定

lavaanパッケージの基本的な使い方③

- 結果の出力
  - 関数summary()を用いて結果の出力
    - summary(推定結果のオブジェクト)
    - 引数

      fit.measures = TRUE：適合度指標

      standardized = TRUE：標準化解

      ci = TRUE：95%信頼区間

      rsquare = TRUE：決定係数

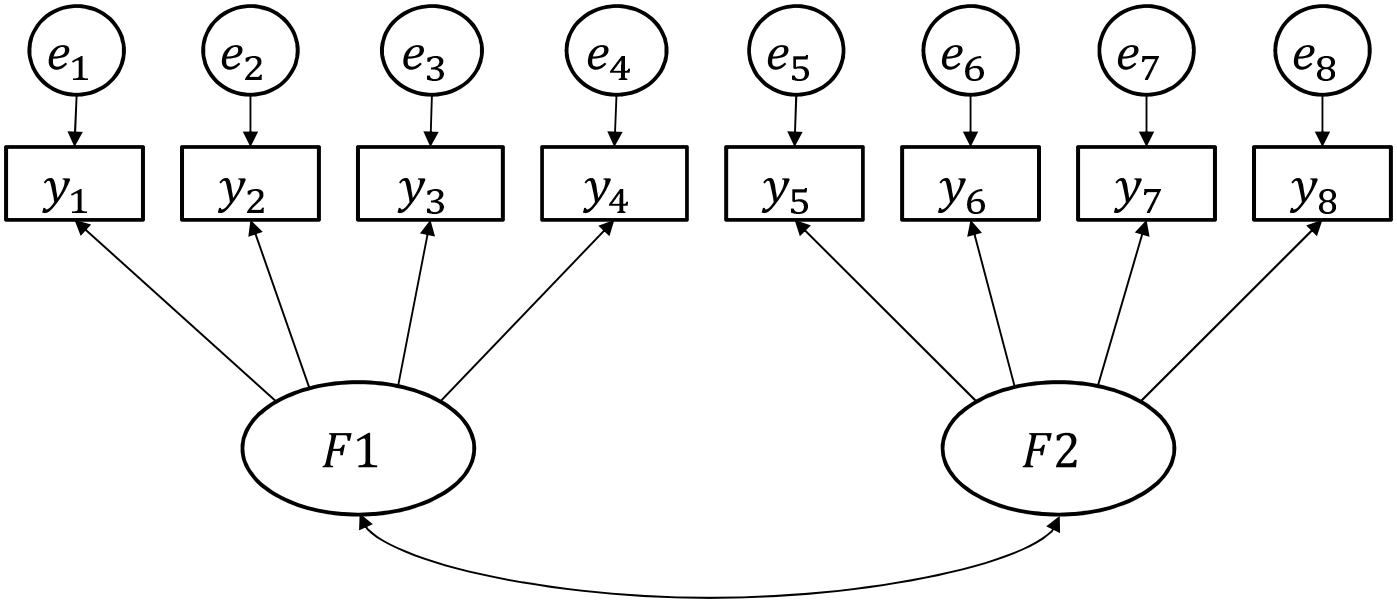
確認的因子分析：モデルの記述

- モデルの記述
  - シングルクオーテーション（'）で囲んだ部分でモデルを記述
    - 「因子 =~ 観測変数」
  - 記述したモデルはオブジェクトに保存

例）1つの因子（F1）が y1～y4，別の因子（F2）が y5～y8 で構成されることを記述し，CFA_model に保存

```
CFA_model <- '
F1 =~ y1 + y2 + y3 + y4
F2 =~ y5 + y6 + y7 + y8
'
```

確認的因子分析のモデル



FIML法による確認的因子分析①

- lavaanパッケージの関数cfa() を用いて母数の推定
  - cfa(モデルを記述したオブジェクト,
    data = データフレーム,
    missing = "fiml", std.lv = TRUE）
    - std.lv = TRUE：因子の分散を1に固定して分析

    → FALSEの場合，各因子を構成する観測変数のうち，最初に記述された観測変数の因子負荷を1に固定

- 関数summary() を用いて結果の出力
  - ここでは，適合度指標と標準化解，信頼区間を出力
    - summary(推定結果のオブジェクト,
      fit.measures = TRUE,
      standardized = TRUE, ci = TRUE)

FIML法による確認的因子分析②

例）母数を推定した結果を CFA_FIML に保存し，
　　結果を出力（適合度指標と標準化解，信頼区間を出力）

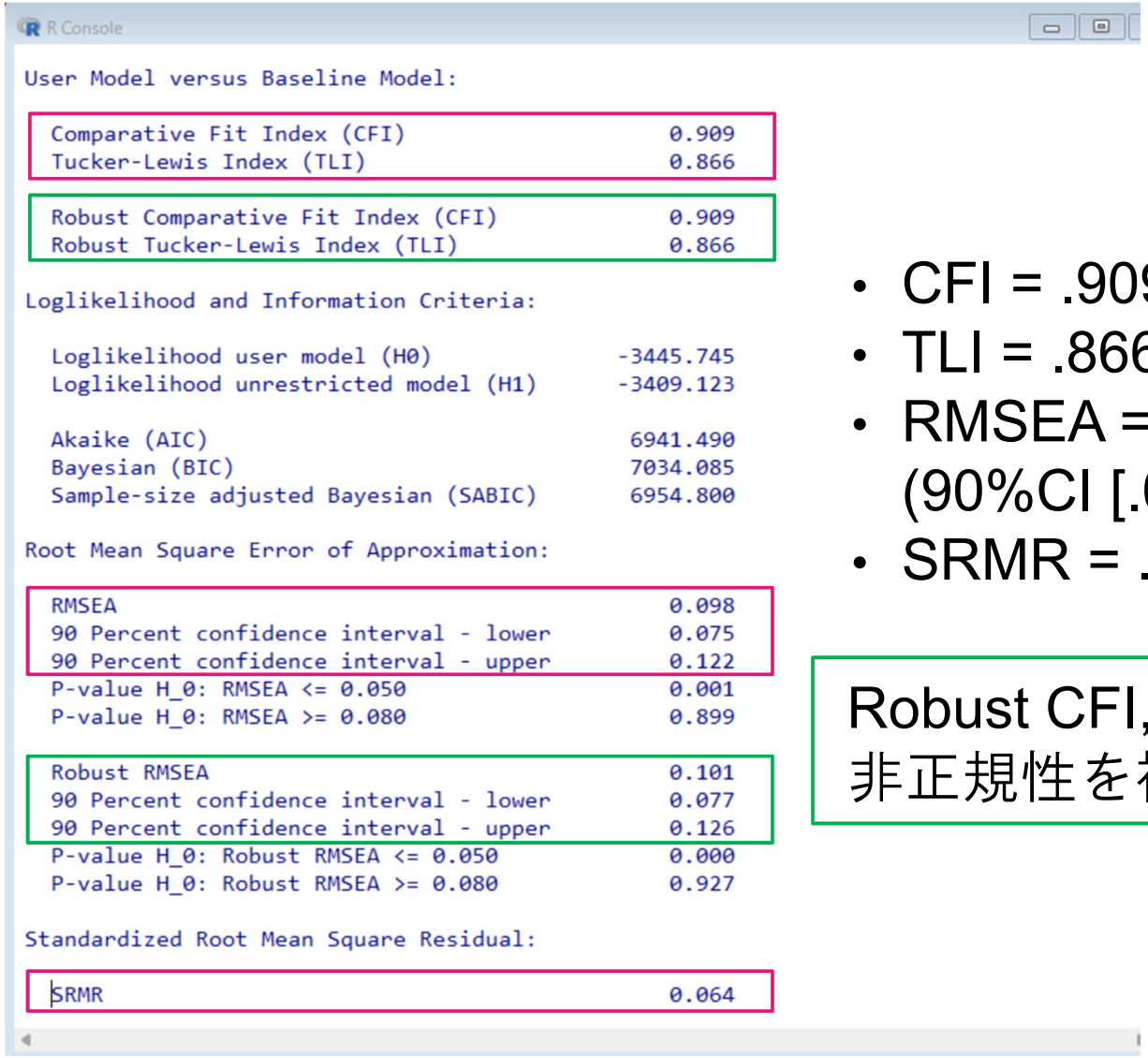
```
library(lavaan)
CFA_FIML <- cfa(CFA_model, data = dat_mis,
missing = "fiml", std.lv = TRUE)
summary(CFA_FIML, fit.measures = TRUE,
standardized = TRUE, ci = TRUE)
```

分析結果（FIML法）

R Console

```
> # モデルの記述
> CFA_model <-'
+ F1 =~ y1 + y2 + y3 + y4
+ F2 =~ y5 + y6 + y7 + y8
+ '
>
> # パッケージの読み込み
> library(lavaan)
>
> # FIML法による確認的因子分析
> ## 母数推定
> CFA_FIML <- cfa(CFA_model, data = dat_mis, missing = "fiml", std.lv = TRUE)
> ## 結果の出力
> summary(CFA_FIML, fit.measures = TRUE, standardized = TRUE, ci = TRUE)
```

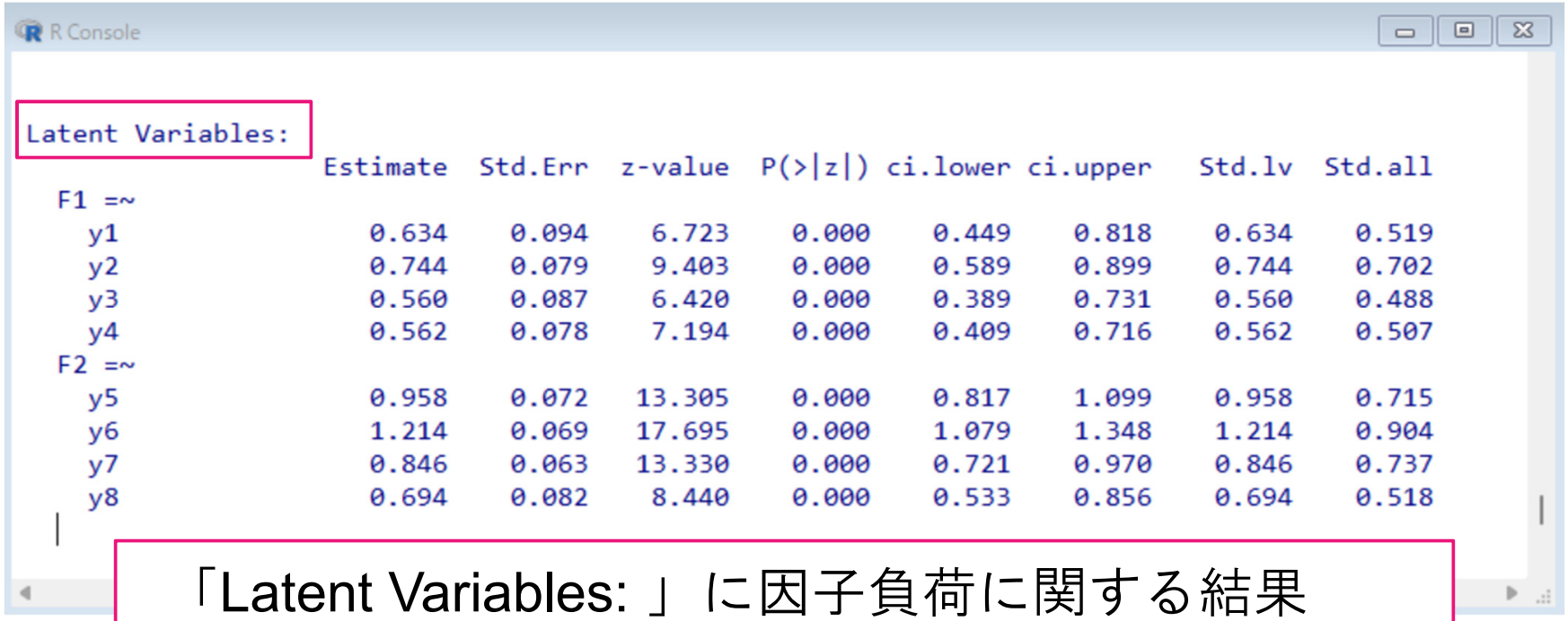
分析結果：適合度指標（FIML法）



- CFI = .909
- TLI = .866
- RMSEA = .098 (90%CI [.075, .122])
- SRMR = .064

Robust CFI, TLI, RMSEA は，非正規性を補正した指標

分析結果：因子負荷（FIML法）



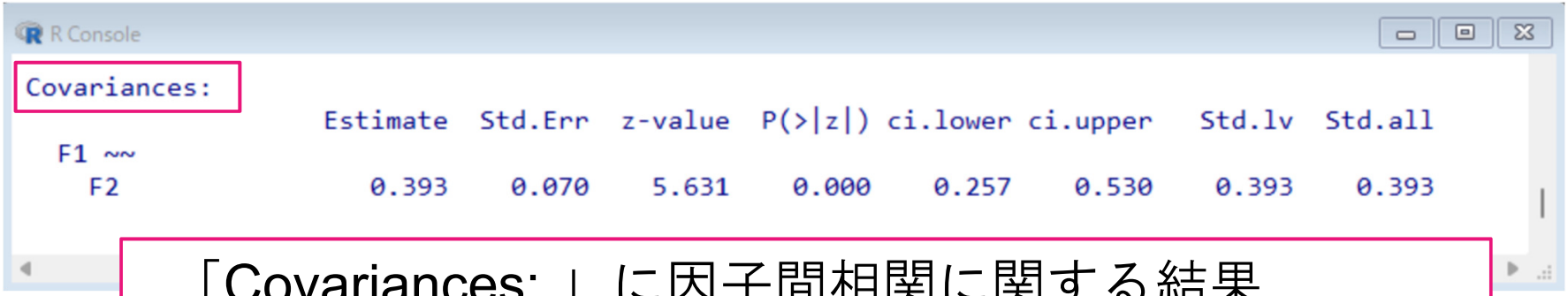
R Console

```
Latent Variables:
                   Estimate  Std.Err  z-value  P(>|z|) ci.lower ci.upper   Std.lv  Std.all
  F1 =~
    y1                0.634    0.094    6.723    0.000    0.449    0.818    0.634    0.519
    y2                0.744    0.079    9.403    0.000    0.589    0.899    0.744    0.702
    y3                0.560    0.087    6.420    0.000    0.389    0.731    0.560    0.488
    y4                0.562    0.078    7.194    0.000    0.409    0.716    0.562    0.507
  F2 =~
    y5                0.958    0.072   13.305    0.000    0.817    1.099    0.958    0.715
    y6                1.214    0.069   17.695    0.000    1.079    1.348    1.214    0.904
    y7                0.846    0.063   13.330    0.000    0.721    0.970    0.846    0.737
    y8                0.694    0.082    8.440    0.000    0.533    0.856    0.694    0.518
```

「Latent Variables: 」に因子負荷に関する結果

- Estimate：非標準化推定値
- Std.Err：非標準化推定値の標準誤差
- ci.lower, ci.upper：95%信頼区間の下限と上限
- Std.all：標準化推定値

分析結果：因子間相関（FIML法）



「Covariances: 」に因子間相関に関する結果

- Estimate：共分散
- Std.all：相関係数

→ ここでは，因子の分散を1に固定しているため，「共分散＝相関係数」

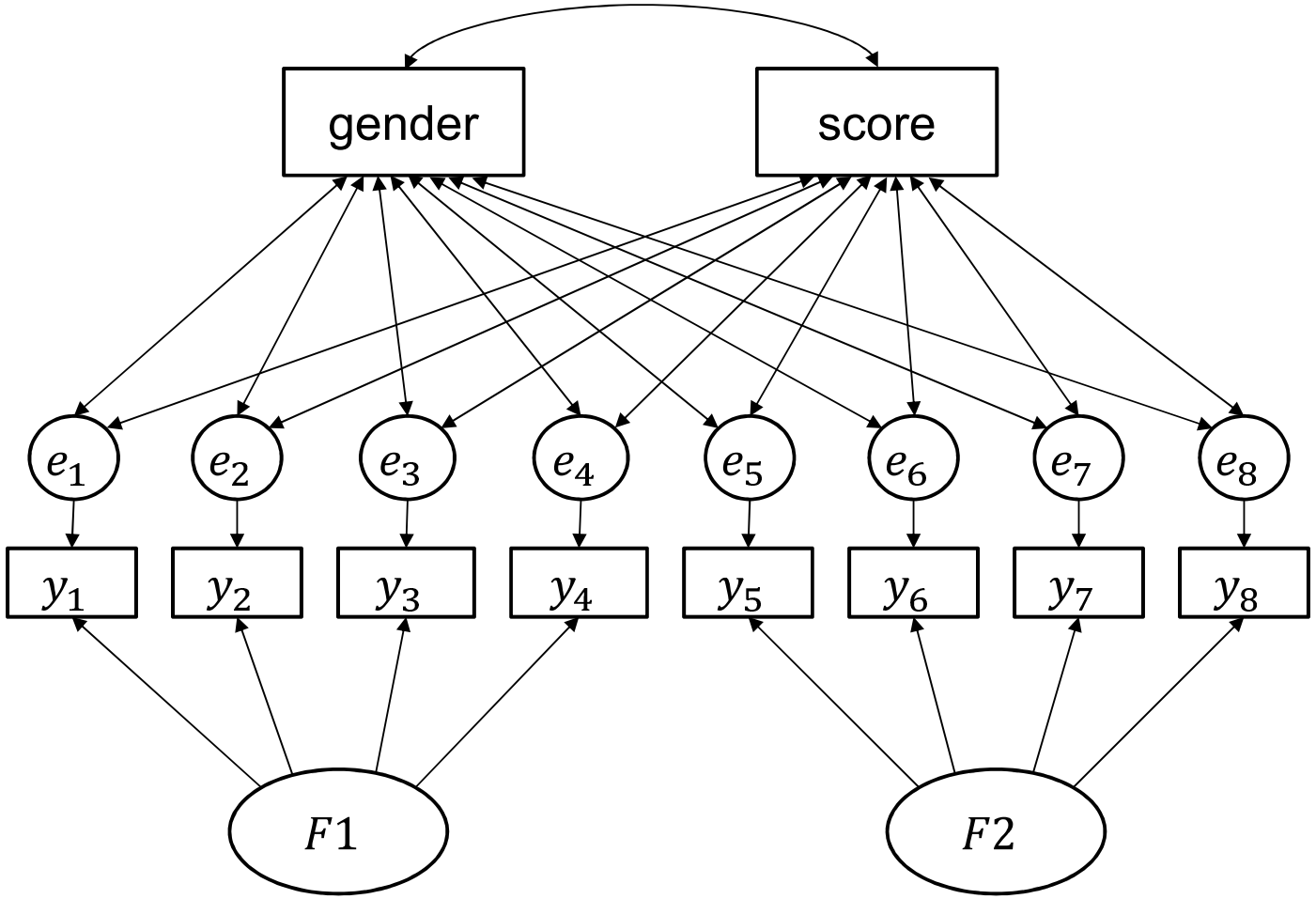
補助変数を用いた分析

- semToolsパッケージの関数cfa.auxiliary() を用いて推定
  - cfa.auxiliary (モデルを記述したオブジェクト,
    data = データフレーム,
    missing = "fiml", std.lv = TRUE,
    aux = "補助変数名"）
- 関数summary () を用いて結果の出力

例）性別とテスト得点を補助変数とした確認的因子分析

```
library(semTools)
CFA_FIML_aux <- cfa.auxiliary(CFA_model, data = dat_mis,
missing = "fiml", std.lv = TRUE, aux = c("gender", "score"))
summary(CFA_FIML_aux, fit.measures = TRUE,
standardized = TRUE, ci = TRUE)
```

補助変数を含めた確認的因子分析のモデル



分析結果：補助変数との相関

R Console

```
Covariances:
                   Estimate  Std.Err  z-value  P(>|z|) ci.lower ci.upper   Std.lv  Std.all
  F1 ~~
    F2                0.395    0.070    5.675    0.000    0.259    0.532    0.395    0.395
  gender ~~
    score             1.202    0.263    4.565    0.000    0.686    1.718    1.202    0.278
   .y1               -0.071    0.036   -1.948    0.051   -0.142    0.000   -0.071   -0.135
  score ~~
   .y1                2.407    0.650    3.703    0.000    1.133    3.680    2.407    0.267
  gender ~~
   .y2
  score ~~
   .y2
  gender ~~
   .y3
  score ~~
   .y3
  gender ~~
   .y4
  score ~~
   .y4
  gender ~~
   .y5
  score ~~
   .y5
  gender ~~
   .y6               -0.038    0.039   -0.985    0.325   -0.114    0.038   -0.038   -0.133
  score ~~
   .y6                0.565    0.687    0.822    0.411   -0.782    1.911    0.565    0.114
  gender ~~
   .y7               -0.016    0.033   -0.477    0.633   -0.082    0.050   -0.016   -0.041
  score ~~
   .y7                1.165    0.597    1.953    0.051   -0.004    2.335    1.165    0.174
  gender ~~
   .y8               -0.114    0.040   -2.830    0.005   -0.193   -0.035   -0.114   -0.199
  score ~~
   .y8               -0.205    0.700   -0.292    0.770   -1.577    1.168   -0.205   -0.021
```

変数名の前に「.」がついているものは内生変数
→ その変数の残差の結果であることを意味

例）score ~~ .y1
→ scoreとy1の残差（e_1）との共分散・相関係数
（Estimate は共分散，Std.all は相関係数）

MI法による確認的因子分析

- semToolsパッケージの関数cfa.mi() を用いて母数の推定
  - cfa.mi(モデルを記述したオブジェクト,
    data = 疑似完全データ, std.lv = TRUE)
- 関数summary() を用いて結果の出力
  - 引数にoutput = "data.frame"を追加

例）母数を推定した結果を CFA_MI に保存し，結果を出力
（適合度指標と標準化解，信頼区間を出力）

```
CFA_MI <- cfa.mi(CFA_model, data = dat_MI, std.lv = TRUE)
summary(CFA_MI, fit.measures = TRUE,
standardized = TRUE, ci = TRUE, output = "data.frame")
```

分析結果：因子負荷（MI法）

R Console

	lhs	op	rhs	est	se	t	df	pvalue	ci.lower	ci.upper	std.lv	std.all
1	F1	=~	y1	0.629	0.084	7.487	Inf	0	0.465	0.794	0.629	0.517
2	F1	=~	y2	0.746	0.075	9.981	Inf	0	0.600	0.893	0.746	0.705
3	F1	=~	y3	0.553	0.079	6.963	Inf	0	0.398	0.709	0.553	0.481
4	F1	=~	y4	0.564	0.076	7.412	Inf	0	0.415	0.713	0.564	0.512
5	F2	=~	y5	0.959	0.073	13.060	Inf	0	0.815	1.103	0.959	0.718
6	F2	=~	y6	1.210	0.069	17.557	Inf	0	1.075	1.345	1.210	0.901
7	F2	=~	y7	0.839	0.063	13.425	Inf	0	0.717	0.962	0.839	0.734
8	F2	=~	y8	0.694	0.079	8.770	Inf	0	0.539	0.849	0.694	0.517
9	y1	~~	y1	1.087	0.111	9.783	Inf	0	0.870	1.305	1.087	0.733
10	y2	~~	y2	0.562	0.090	6.249	Inf	0	0.386	0.738	0.562	0.502
11	y3	~~	y3	1.016	0.100	10.180	Inf	0	0.820	1.211	1.016	0.768
12	y4	~~	y4	0.898	0.091	9.849	Inf	0	0.719	1.076	0.898	0.738
13	y5	~~	y5	0.866	0.089	9.694	Inf	0	0.691	1.041	0.866	0.485
14	y6	~~	y6	0.338	0.081	4.148	Inf	0	0.178	0.497	0.338	0.187

15
16
17
18
19
> |

「=~」に因子負荷に関する結果

- est：非標準化推定値
- se：非標準化推定値の標準誤差
- ci.lower, ci.upper：95%信頼区間の下限と上限
- std.all：標準化推定値

分析結果：因子間相関（MI法）

R Console

	lhs	op	rhs	est	se	t	df	pvalue	ci.lower	ci.upper	std.lv	std.all
1	F1	=~	y1	0.629	0.084	7.487	Inf	0	0.465	0.794	0.629	0.517
2	F1	=~	y2	0.746	0.075	9.981	Inf	0	0.600	0.893	0.746	0.705
3	F1	=~	y3	0.553	0.079	6.963	Inf	0	0.398	0.709	0.553	0.481
4	F1	=~	y4	0.564	0.076	7.412	Inf	0	0.415	0.713	0.564	0.512
5	F2	=~	y5	0.959	0.073	13.060	Inf	0	0.815	1.103	0.959	0.718
6	F2	=~	y6	1.210	0.069	17.557	Inf	0	1.075	1.345	1.210	0.901
7	F2	=~	y7	0.839	0.063	13.425	Inf	0	0.717	0.962	0.839	0.734
8	F2	=~	y8	0.694	0.079	8.770	Inf	0	0.539	0.849	0.694	0.517
9	y1	~~	y1	1.087	0.111	9.783	Inf	0	0.870	1.305	1.087	0.733
10	y2	~~	y2	0.562	0.090	6.249	Inf	0	0.386	0.738	0.562	0.502
11	y3	~~	y3	1.016	0.100	10.180	Inf	0	0.820	1.211	1.016	0.768
12	y4	~~	y4	0.898	0.091	9.849	Inf	0	0.719	1.076	0.898	0.738
13	y5	~~	y5	0.866	0.089	9.694	Inf	0	0.691	1.041	0.866	0.485
14	y6	~~	y6	0.338	0.081	4.148	Inf	0	0.178	0.497	0.338	0.187
15	y7	~~	y7	0.604	0.064	9.421	5261.583	0	0.479	0.730	0.604	0.462
16	y8	~~	y8	1.321	0.117	11.266	Inf	0	1.091	1.550	1.321	0.733
17	F1	~~	F1	1.000	0.000	NA	NA	NA	1.000	1.000	1.000	1.000
18	F2	~~	F2	1.000	0.000	NA	NA	NA	1.000	1.000	1.000	1.000
19	F1	~~	F2	0.399	0.069	5.770	Inf	0	0.263	0.534	0.399	0.399

>

「~~」に因子間相関に関する結果

- est：因子間の共分散
- std.all：因子間の相関係数

推定結果のまとめ（因子負荷と因子間相関）

因子の分散を1に固定したときの非標準化推定値と標準誤差

	完全データ		FIML		FIML+補助変数		MI		リストワイズ	
	推定値	標準誤差	推定値	標準誤差	推定値	標準誤差	推定値	標準誤差	推定値	標準誤差
Int → y1	0.612	0.082	0.634	0.094	0.627	0.094	0.629	0.084	0.462	0.102
Int → y2	0.755	0.073	0.744	0.079	0.746	0.079	0.746	0.075	0.736	0.100
Int → y3	0.532	0.077	0.560	0.087	0.557	0.087	0.553	0.079	0.471	0.098
Int → y4	0.581	0.074	0.562	0.078	0.565	0.078	0.564	0.076	0.543	0.097
Ext → y5	0.958	0.071	0.958	0.072	0.959	0.072	0.959	0.073	0.917	0.089
Ext → y6	1.222	0.067	1.214	0.069	1.213	0.068	1.210	0.069	1.155	0.083
Ext → y7	0.825	0.061	0.846	0.063	0.848	0.064	0.839	0.063	0.863	0.075
Ext → y8	0.703	0.077	0.694	0.082	0.699	0.082	0.694	0.079	0.669	0.094
Int ↔ Ext	0.403	0.067	0.393	0.070	0.395	0.070	0.399	0.069	0.398	0.087

注）「FIML+補助変数」では，性別とテスト得点を補助変数とした

探索的因子分析

FIML法による探索的因子分析①

- lavaanパッケージの関数efa()を利用
  - efa(データフレーム, nfactors = 因子数,
    rotation = "回転法", missing = "fiml")
    - 回転法には，バリマックス回転（varimax），
      プロマックス回転（promax）など

      → デフォルトはジェオミン回転（geomin）

- 結果の出力には関数summary()を利用
  - summary(推定結果のオブジェクト, cutoff = 0)
    - cutoff = 0：すべての因子負荷の値を表示

- 適合度の出力には関数fitMeasures()を利用
  - fitMeasures(推定結果のオブジェクト)

FIML法による探索的因子分析②

例）動機づけに関する変数（3～10列目）について
探索的因子分析（2因子解，オブリミン回転）

```
EFA_FIML <- efa(dat_mis[c(3:10)], nfactors = 2,
rotation = "oblimin", missing = "fiml")
summary(EFA_FIML, cutoff = 0)
```

分析結果：適合度と固有値（FIML法）

R Console

```
> ## 探索的因子分析(2因子解･オブリミン回転)
> EFA_FIML <- efa(dat_mis[c(3:10)], nfactors = 2, rotation = "oblimin", missing = "fiml")
> ## 結果の出力
> summary(EFA_FIML, cutoff = 0)
This is lavaan 0.6-18 -- running exploratory factor analysis

  Estimator                                         ML
  Rotation method                      OBLIMIN OBLIQUE
  Oblimin gamma                                     0
  Rotation algorithm (rstarts)                GPA (30)
  Standardized metric                             TRUE
  Row weights                                     None

  Number of observations                           300
  Number of missing patterns                        10

Fit measures:
                   aic      bic    sabic  chisq df pvalue   cfi rmsea
  nfactors = 2 6926.488 7041.306 6942.992 46.242 13      0 0.948 0.104

Eigenvalues correlation matrix:

    ev1     ev2     ev3     ev4     ev5     ev6     ev7     ev8
  2.951   1.651   0.898   0.675   0.577   0.506   0.483   0.259
```

Fit measures：適合度指標（の一部）

Eigenvalues：固有値

分析結果：因子負荷・因子間相関（FIML法）

R Console

```
Standardized loadings: (* = significant at 1% level)

       f1      f2       unique.var   communalities
y1  0.658* -0.042            0.579           0.421
y2  0.501*  0.203*           0.660           0.340
y3  0.645* -0.048            0.597           0.403
y4  0.321*  0.215*           0.818           0.182
y5 -0.094   0.755*           0.454           0.546
y6  0.020   0.877*           0.222           0.778
y7  0.131*  0.707*           0.439           0.561
y8 -0.103   0.560*           0.703           0.297
```

- f1：第1因子の因子負荷
- f2：第2因子の因子負荷
- unique.var：独自性
- communalities：共通性

```
                              f2    f1 total
Sum of sq (obliq) loadings 2.267 1.261 3.528
Proportion of total        0.643 0.357 1.000
Proportion var             0.283 0.158 0.441
Cumulative var             0.283 0.441 0.441
```

- Sum of sq loadings：因子寄与
- Proportion Var：因子寄与率
- Cumulative Var：累積寄与率

```
Factor correlations: (* = significant at 1% level)

       f1      f2
f1  1.000
f2  0.236*  1.000
```

Factor correlations：因子間の相関係数

推定結果のまとめ（因子負荷と因子間相関）

	完全データ		FIML		リストワイズ	
	Int	Ext	Int	Ext	Int	Ext
y1	.659	-.050	.658	-.042	.578	-.019
y2	.511	.200	.501	.203	.409	.235
y3	.623	-.058	.645	-.048	.659	-.023
y4	.344	.221	.321	.215	.224	.255
y5	-.089	.757	-.094	.755	-.176	.735
y6	.012	.883	.020	.877	.041	.837
y7	.131	.697	.131	.707	.155	.726
y8	-.079	.557	-.103	.560	-.100	.532
因子間相関	.247		.236		.153	

回帰分析

合成変数の作成①

- 関数rowMeans() を利用して平均値を算出
  - データフレーム$新しい変数の名前
    <- rowMeans(データフレーム[c(列番号, 列番号…)])
    - 項目に欠測のある回答者については，平均値も欠測することになる

例）y1～y4（3～6列目）の平均値をInt，y5～y8（7～10列目）の平均値をExt という変数名でデータフレームに保存

```
dat_mis$Int <- rowMeans(dat_mis[c(3:6)])
dat_mis$Ext <- rowMeans(dat_mis[c(7:10)])
```

合成変数の作成②

R Console

```
> # 合成変数の作成
> dat_mis$Int <- rowMeans(dat_mis[c(3:6)])
> dat_mis$Ext <- rowMeans(dat_mis[c(7:10)])
> head(dat_mis, 8)
  school gender y1 y2 y3 y4 y5 y6 y7 y8 score  Int  Ext
1      1      0  3  3  3  5  5  5  5  5    49 3.50 5.00
2      1      0  4  4  4  3  5  2  4  3    51 3.75 3.50
3      1      0  2  3  1  4  4  4  4 NA    51 2.50   NA
4      1      0  4  3  3  4  4  4  4  3    59 3.50 3.75
5      1      0  4  3  4  3  5  2  4  5    49 3.50 4.00
6      1      0  3  3  3  4  4  3  4  3    65 3.25 3.50
7      1      0  4  3  3  4  3  4  4  4    53 3.50 3.75
8      1      0  5  5  5  5  5 NA  5  3    66 5.00   NA
> |
```

合成変数の作成（疑似完全データ）

- mitmlパッケージの関数mids2mitml.list() により，miceパッケージで作成されたオブジェクトをコンバートし，for 文を利用して合成変数を作成
  - オブジェクト <- mids2mitml.list(疑似完全データ)

例）y1～y4（3～6列目）の平均値をInt，y5～y8（7～10列目）の平均値をExt という変数名でデータフレームに保存

```
library(mitml)
dat_MI_list <- mids2mitml.list(dat_MI)
for(i in 1:100){
  dat_MI_list[[i]]$Int <- rowMeans(dat_MI_list[[i]][c(3:6)])
  dat_MI_list[[i]]$Ext <- rowMeans(dat_MI_list[[i]][c(7:10)])
}
```

作成した疑似データセット数に応じて設定

【参考】for 文による繰り返し処理

- 「for(変数 in 値) { } 」とすると，「変数」が in の後に書かれた全ての「値」を取りながら，{ } で囲まれた処理を繰り返す

例）for(i in 1:5) {
    print(i)
}

  - 変数i が1から5の値を順に取りながらprint(i) を反復（i が1のときの { } 内の処理を実行し，それが終わるとi が2に変わって{ } 内の処理を実行し，i が5になるまで{ } 内の処理を実行する）

回帰分析：モデルの記述①

- モデルの記述
  - シングルクオーテーション（'）で囲んだ部分で
    モデルを記述
    - 「従属変数 ~ 独立変数」
    - 「独立変数 ~~ 独立変数」

    → 独立変数の分散・共分散の推定をすることで，
    FIML法による分析をする際に，独立変数が欠測して
    いるケースも分析に含まれる

    （推定しないと，独立変数に欠測のあるケースは除外される）
  - 記述したモデルはオブジェクトに保存

回帰分析：モデルの記述②

例）内発的動機づけ（Int）と外発的動機づけ（Ext）を
独立変数，テスト得点（score）を従属変数とする
重回帰分析のモデルを記述し，reg_model に保存

```
reg_model <- '
score ~ Int + Ext
Int ~~ Int
Ext ~~ Ext
Int ~~ Ext
'
```

FIML法による回帰分析

- lavaanパッケージの関数sem() を用いて母数の推定
  - sem(モデルを記述したオブジェクト,
    data = データフレーム, missing = "fiml")
- 関数summary() を用いて結果の出力
  - summary(推定結果のオブジェクト)
    - ここでは，標準化解と決定係数，信頼区間を出力

例）母数を推定した結果をreg_FIML に保存し，結果を出力
（標準化解と決定係数，信頼区間を出力）

```
reg_FIML <- sem(reg_model, data = dat_mis, missing = "fiml")
summary(reg_FIML, standardized = TRUE, rsquare = TRUE,
ci = TRUE)
```

分析結果：偏回帰係数（FIML法）

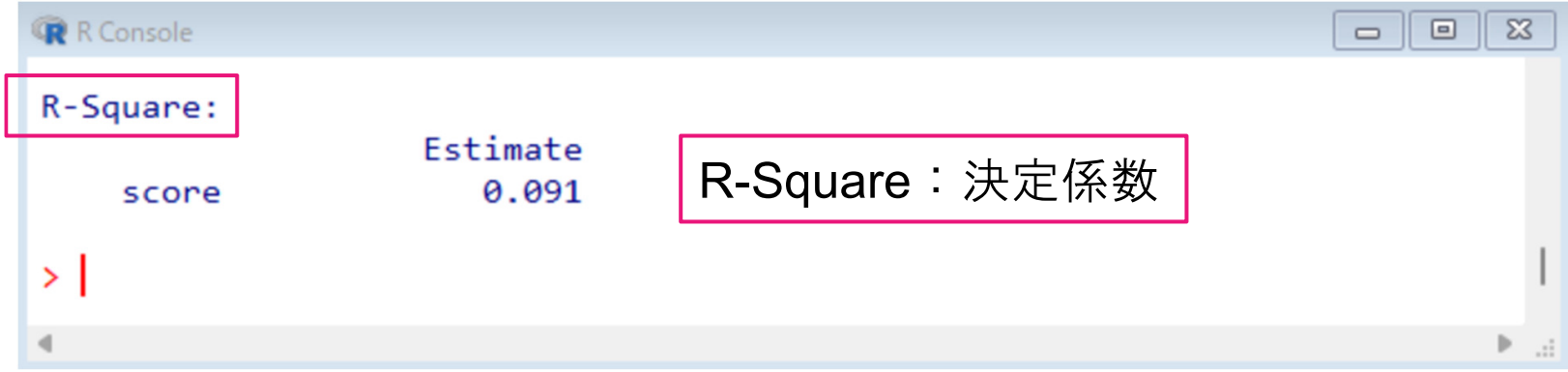
R Console

```
> # モデルの記述
> reg_model <-'
+ score ~ Int + Ext
+ Int ~~ Int
+ Ext ~~ Ext
+ Int ~~ Ext
+ '
> # FIML法による回帰分析
> ## 母数推定
> reg_FIML <- sem(reg_model, data = dat_mis, missing = "fiml")
> ## 結果の出力
> summary(reg_FIML, standardized = TRUE, rsquare = TRUE, ci = TRUE)
lavaan 0.6-18 ended normally after 28 iterations

  Estimator                                         ML
  Optimization method                           NLMINB
  Number of model parameters                         9

  Number of observations
  Number of missing patterns

Model Test User Model:

  Test statistic
  Degrees of freedom

Parameter Estimates:

  Standard errors
  Information
  Observed information based on                Hessian

Regressions:
                   Estimate  Std.Err  z-value  P(>|z|) ci.lower ci.upper   Std.lv  Std.all
  score ~
    Int               3.378    0.665    5.083    0.000    2.076    4.681    3.378    0.308
    Ext              -0.260    0.567   -0.458    0.647   -1.370    0.851   -0.260   -0.030
```

「Regressions:」に偏回帰係数に関する結果

- Estimate：非標準化偏回帰係数
- Std.Err：非標準化偏回帰係数の標準誤差
- ci.lower, ci.upper：95%信頼区間の下限と上限
- Std.all：標準化偏回帰係数

分析結果：決定係数（FIML法）



MI法による回帰分析

- semToolsパッケージの関数sem.mi() を用いて母数の推定
  - sem.mi(モデルを記述したオブジェクト,
    data = 疑似完全データ)
- 関数summary() を用いて結果の出力
  - summary(推定結果のオブジェクト)
    - 引数にoutput = "data.frame"を追加

例）母数を推定した結果をreg_MI に保存し，結果を出力
（標準化解と決定係数，信頼区間を出力）

```
reg_MI <- sem.mi(reg_model, data = dat_MI_list)
summary(reg_MI, standardized = TRUE,
rsquare = TRUE, ci = TRUE, output = "data.frame")
```

分析結果：偏回帰係数（MI法）

R Console

```
> # MI法による回帰分析
> ## 母数推定
> reg_MI <- sem.mi(reg_model, data = dat_MI_list)
> ## 結果の出力
> summary(reg_MI, standardized = TRUE, rsquare = TRUE, ci = TRUE, output = "data.frame")
    lhs op    rhs    est    se      t  df pvalue ci.lower ci.upper std.lv std.all
1 score  ~    Int  3.384 0.633  5.346 Inf  0.000    2.143    4.624  3.384   0.308
2 score  ~    Ext  0.009 0.483  0.018 Inf  0.985   -0.938    0.956  0.009   0.001
3   Int ~~    Int  0.617 0.051 12.062 Inf  0.000    0.517    0.718  0.617   1.000
4   Ext ~~    Ext  1.059 0.088 12.062 Inf  0.000    0.887    1.231  1.059   1.000
5   Int ~~    Ext
6 score ~~ score
7   Int r2    Int
8   Ext r2    Ext
9 score r2 score
>
```

「~」に偏回帰係数に関する結果

- est：非標準化偏回帰係数
- se：非標準化偏回帰係数の標準誤差
- ci.lower, ci.upper：95%信頼区間の下限と上限
- std.all：標準化偏回帰係数

分析結果：決定係数（MI法）

R Console

```
> # MI法による回帰分析
> ## 母数推定
> reg_MI <- sem.mi(reg_model, data = dat_MI_list)
> ## 結果の出力
> summary(reg_MI, standardized = TRUE, rsquare = TRUE, ci = TRUE, output = "data.frame")
    lhs op   rhs    est    se      t  df pvalue ci.lower ci.upper std.lv std.all
1 score  ~   Int  3.384 0.633  5.346 Inf  0.000    2.143    4.624  3.384   0.308
2 score  ~   Ext  0.009 0.483  0.018 Inf  0.985   -0.938    0.956  0.009   0.001
3   Int ~~   Int  0.617 0.051 12.062 Inf  0.000    0.517    0.718  0.617   1.000
4   Ext ~~   Ext  1.059 0.088 12.062 Inf  0.000    0.887    1.231  1.059   1.000
5   Int ~~   Ext  0.204 0.049  4.169 Inf  0.000    0.108    0.300  0.204   0.252
6 score ~~ score 67.360 5.585 12.061 Inf  0.000   56.414   78.307 67.360   0.905
7   Int r2   Int  0.000    NA     NA  NA     NA       NA       NA     NA      NA
8   Ext r2   Ext  0.000    NA     NA  NA     NA       NA       NA     NA      NA
9 score r2 score  0.095    NA     NA  NA     NA       NA       NA     NA      NA
>
```

「score r2 score」のest：決定係数

推定結果のまとめ（非標準化偏回帰係数と決定係数）

	完全データ		FIML		MI		リストワイズ	
	推定値	標準誤差	推定値	標準誤差	推定値	標準誤差	推定値	標準誤差
Int	3.521	0.622	3.378	0.665	3.384	0.633	3.379	0.762
Ext	0.131	0.474	-0.260	0.567	0.009	0.483	-0.495	0.562
R^2	.105		.091		.095		.086	

項目和得点を用いた分析では，完全データと比較してFIMLでは標準誤差が大きくなっていることから，（項目レベルで代入してから尺度得点を算出した）MI法が有用といえる

【参考】関数with() を用いたMI法による回帰分析①

- 関数with() と関数lm() を用いたMI法による回帰分析
  - with(疑似完全データ, lm(従属変数 ~ 独立変数))
- 関数summary() と pool() を用いた結果の統合
  - summary(pool(推定結果のオブジェクト))

例）内発的動機づけ（Int）と外発的動機づけ（Ext）を独立変数，テスト得点（score）を従属変数とする重回帰分析

```
reg_MI_lm <- with(dat_MI_list, lm(score ~ Int + Ext))
summary(pool(reg_MI_lm))
```

【参考】関数with() を用いたMI法による回帰分析②

- miceパッケージの関数pool.r.squared() を用いて決定係数を算出
  - pool.r.squared(統合後の推定値のオブジェクト)
    - 引数に adjusted = TRUE を加えることで,自由度調整済み決定係数を出力

例）統合後の推定値をest に代入し,
決定係数と自由度調整済み決定係数を出力

```
est <- pool(reg_MI_lm)
pool.r.squared(est)
pool.r.squared(est, adjusted = TRUE)
```

t 検定

t 検定：モデルの記述

- モデルの記述
  - シングルクオーテーション（'）で囲んだ部分でモデルを記述
    - 「従属変数 ~ 独立変数」（※ 独立変数は2値変数）
    - 「独立変数 ~~ 独立変数」

    → 独立変数の分散を推定することで，独立変数が欠測しているケースも分析に含まれる

例）性別（gender）を独立変数，テスト得点（score）を従属変数とするモデルを記述し，t_model に保存

```
t_model <- '
score ~ gender
gender ~~ gender
'
```

FIML法による*t* 検定

- lavaanパッケージの関数sem() を用いて母数の推定
  - sem(モデルを記述したオブジェクト,
    data = データフレーム, missing = "fiml")
- 関数summary() を用いて結果の出力
  - summary(推定結果のオブジェクト)
    - ここでは，信頼区間を出力

例）母数を推定した結果を t_FIML に保存し，結果を出力
（信頼区間を出力）

```
t_FIML <- sem(t_model, data = dat_mis, missing = "fiml")
summary(t_FIML, ci = TRUE)
```

分析結果（FIML法）

```
> # モデルの記述
> t_model <-'
+ score ~ gender
+ gender ~~ gender
+ '
> # FIML法によるt検定
> ## 母数推定
> t_FIML <- sem(t_model, data = dat_mis, missing = "fiml")
> ## 結果の出力
> summary(t_FIML, ci = TRUE)
lavaan 0.6-18 ended normally after 21 iterations

  Estimator                                         ML
  Optimization method                           NLMINB
  Number of model parameters                         5

  Number of observations                           300
  Number of missing patterns                         2

Model Test User Model:

  Test statistic                                 0.000
  Degrees of freedom                                 0

Parameter Estimates:

  Standard errors
  Information
  Observed information based on

Regressions:
                   Estimate  Std.Err  z-value  P(>|z|) ci.lower ci.upper
  score ~
    gender            4.863    0.979    4.968    0.000    2.945    6.782
```

Estimateの値は差得点を意味
→ 女性（1）は男性（0）
よりも得点が4.863点高い

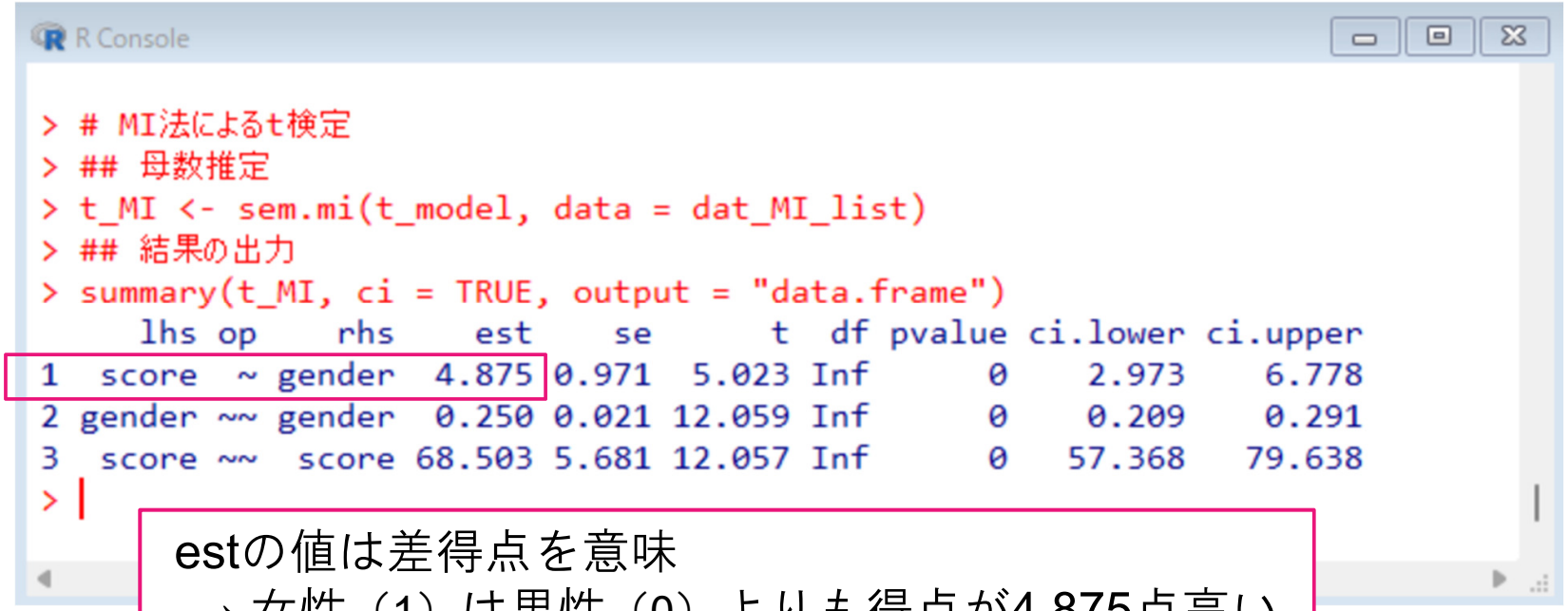
MI法による*t* 検定

- semToolsパッケージの関数sem.mi() を用いて母数の推定
  - sem.mi(モデルを記述したオブジェクト,
    data = 疑似完全データ)
- 関数summary() を用いて結果の出力
  - summary(推定結果のオブジェクト)
    - 引数にoutput = "data.frame"を追加

例）母数を推定した結果を t_MI に保存し，結果を出力
（信頼区間を出力）

```
t_MI <- sem.mi(t_model, data = dat_MI_list)
summary(t_MI, ci = TRUE, output = "data.frame")
```

分析結果（MI法）


R Console
> # MI法によるt検定
> ## 母数推定
> t_MI <- sem.mi(t_model, data = dat_MI_list)
> ## 結果の出力
> summary(t_MI, ci = TRUE, output = "data.frame")
lhs op rhs est se t df pvalue ci.lower ci.upper
1 score ~ gender 4.875 0.971 5.023 Inf 0 2.973 6.778
2 gender ~~ gender 0.250 0.021 12.059 Inf 0 0.209 0.291
3 score ~~ score 68.503 5.681 12.057 Inf 0 57.368 79.638
>
estの値は差得点を意味
→ 女性（1）は男性（0）よりも得点が4.875点高い

推定結果のまとめ

	完全データ		FIML		MI		リストワイズ	
	推定値	標準誤差	推定値	標準誤差	推定値	標準誤差	推定値	標準誤差
女子ダミー	4.853	0.957	4.863	0.979	4.875	0.971	4.439	1.126

【参考】関数with() を用いたMI法による*t* 検定

- 関数with() と関数lm() を用いたMI法による*t* 検定
  - with(疑似完全データ, lm(従属変数 ~ 独立変数))
- 関数summary() と pool() を用いた結果の統合
  - summary(pool(推定結果のオブジェクト))

例）性別（gender）を独立変数，テスト得点（score）を従属変数とする*t* 検定

```
t_MI_lm <- with(dat_MI_list, lm(score ~ gender))
summary(pool(t_MI_lm))
```

分散分析

MI法による分散分析①

- 独立変数を要因型に変換した後，関数with() とlm() を利用して分散分析
  - 例として，学校種を独立変数とする分散分析
    - 変数「school」は数値型のため，関数as.factor() を利用して要因型に変換してから分析

例）学校種（school）を要因型に変換

```
for(i in 1:100){
  dat_MI_list[[i]]$school <- as.factor(dat_MI_list[[i]]$school)
}
```

作成した疑似データセット数に応じて設定

MI法による分散分析②

- 関数with() とlm() を利用した分散分析
  - with(疑似完全データ, lm(従属変数 ~ 独立変数))
- mitmlパッケージの関数testEstimates() を利用して結果を統合
  - testEstimates(分散分析の結果を代入したオブジェクト)

例）学校種（school）を独立変数，テスト得点（score）を従属変数とする分散分析

```
anova_MI <- with(dat_MI_list, lm(score ~ school))
testEstimates(anova_MI)
```

分散分析の結果（MI法）

R Console

```
> # 学校種を要因型に変換
> for(i in 1:100){
+  dat_MI_list[[i]]$school <- as.factor(dat_MI_list[[i]]$school)
+ }
> # 分散分析
> anova_MI <- with(dat_MI_list, lm(score ~ school))
> ## 結果の統合
> testEstimates(anova_MI)

Call:

testEstimates(model = anova_MI)

Final parameter estimates and inferences obtained from 100 imputed data sets.

             Estimate Std.Error   t.value        df   P(>|t|)       RIV       FMI
(Intercept)    54.909     0.818    67.131 3.133e+05     0.000     0.018     0.018
school2         7.376     1.159     6.366 2.228e+05     0.000     0.022     0.021
school3         2.494     1.173     2.126 5.008e+04     0.033     0.047     0.045

Unadjus

> |
```

- school 2：school = 1（公立）とschool = 2（国立）との比較
  → 公立と比較して国立は7.376点高い
- school 3：school = 1（公立）とschool = 3（私立）との比較
  → 公立と比較して私立は2.494点高い

MI法による多重比較

- multcompパッケージの関数glht() を利用（テューキー法）
  - lapply(分散分析の結果を代入したオブジェクト,
    glht, linfct = mcp(独立変数 = "Tukey"))
- mitmlパッケージの関数testEstimates() を利用して結果を統合
  - testEstimates(多重比較の結果を代入したオブジェクト)

例）多重比較の結果をanova_MI_pairwise に保存し，
関数testEstimates() により結果の統合

```
library(multcomp)
anova_MI_pairwise <- lapply(anova_MI,
glht, linfct = mcp(school = "Tukey"))
testEstimates(anova_MI_pairwise)
```

多重比較の結果（MI法）

R Console

```
> # 多重比較
> ## パッケージの読み込み
> library(multcomp)
> ## 多重比較
> anova_MI_pairwise <- lapply(anova_MI, glht, linfct = mcp(school = "Tukey"))
> ## 結果の統合
> testEstimates(anova_MI_pairwise)

Call:

testEstimates(model = anova_MI_pairwise)

Final parameter estimates and inferences obtained from 100 imputed data sets.

       Estimate Std.Error   t.value        df   P(>|t|)       RIV       FMI
2 - 1     7.376     1.159     6.366 2.228e+05     0.000     0.022     0.021
3 - 1     2.494     1.173     2.126 5.008e+04     0.033     0.047     0.045
3 - 2    -4.882     1.177    -4.148 3.765e+04     0.000     0.054     0.051

Unadjusted

> |
```

- 2－1：school = 2（国立）とschool = 1（公立）の差
  → 公立と比較して国立は7.376点高い
- 3－1：school = 3（私立）とschool = 1（公立）の差
  → 公立と比較して私立は2.494点高い
- 3－2：school = 3（私立）とschool = 2（国立）の差
  → 国立と比較して私立は4.882点低い

推定結果のまとめ

	完全データ		MI		リストワイズ	
	推定値	標準誤差	推定値	標準誤差	推定値	標準誤差
国立ダミー	7.570	1.145	7.376	1.159	7.210	1.298
私立ダミー	2.850	1.145	2.494	1.173	2.268	1.371

【参考】MI法による反復測定デザインの分散分析

- 関数lm() ではなく，lme4パッケージの関数lmer() を利用
  - with(疑似完全データ,
    lmer(従属変数 ~ 独立変数 + (1|個人を識別する変数)))
  - 階層線形モデルによる分析も同じ要領で可能

例）従属変数がscore，独立変数がcondition，個人を識別する変数がID，疑似完全データを代入したオブジェクトがimp

```
# 分散分析
fit <- with(imp, lmer(score ~ condition + (1|ID)))
testEstimates(fit)
# 多重比較
fit.pairwise <- lapply(fit, glht, linfct = mcp(condition = "Tukey"))
testEstimates(fit.pairwise)
```

引用文献

高橋将宜・渡辺美智子 (2017). 欠測データ処理—Rによる単一代入法と多重代入法　共立出版

豊田秀樹編 (2014). 共分散構造分析［R編］—構造方程式モデリング　東京図書