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。

- 統計分布や最尤法を前提としない多変量解析法も多い。

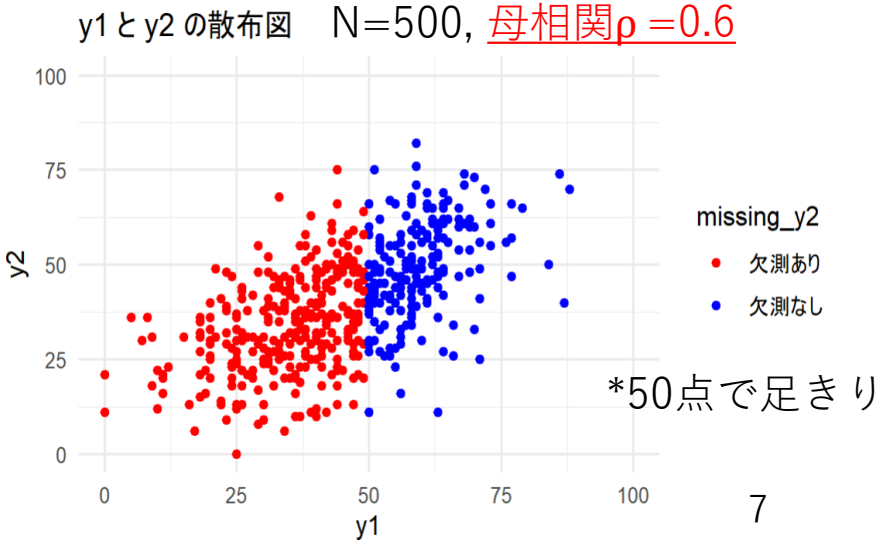
アウトライン

- 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

平均　[共]分散

データ $(10, 20)$ $\begin{pmatrix} 100 & 50 \\ 50 & 200 \end{pmatrix}$

モデル $\mu(\boldsymbol{\theta})$ $\Sigma(\boldsymbol{\theta})$

⇒

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

・ $\mu(\boldsymbol{\theta}) = (\mu_1, \alpha + \beta\, \mu_1)^{\mathrm{T}}$ ⇒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

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

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

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

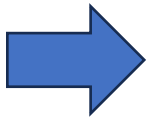
- FIMLの概要
- MIの概要
- 補助変数の活用
- まとめとMNARの場合

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	-

$$\hat{y}_2 = \hat{\alpha} + \hat{\beta} y_1$$



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。
- 残差分散や補完モデルの（切片や回帰係数、残差分散についての）推定誤差を考慮していない。

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。分析モデル内の変数は原則含める。

(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。

・ 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$y_1, y_3 \ldots, y_8$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)

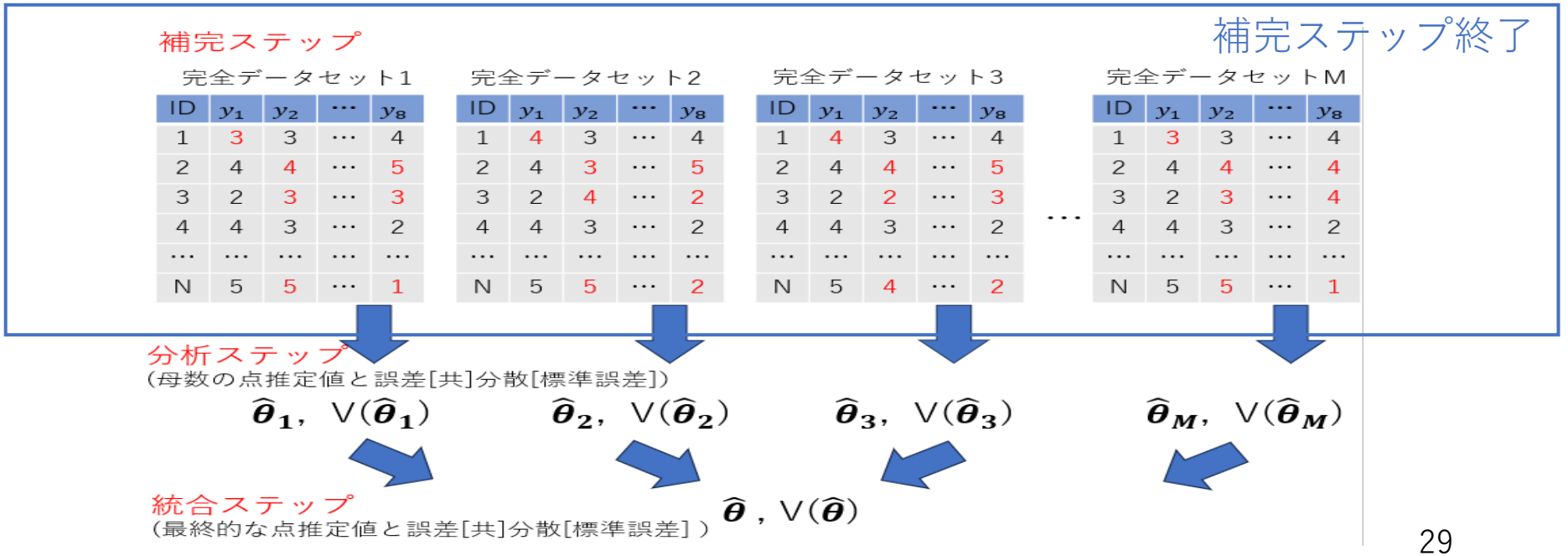
- y_8内にある欠測値を、疑似的な完全データセット内の$y_1, y_2 \ldots, y_7$を用いたy_8の補完モデルから生成された補完値により補完し更新する。

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

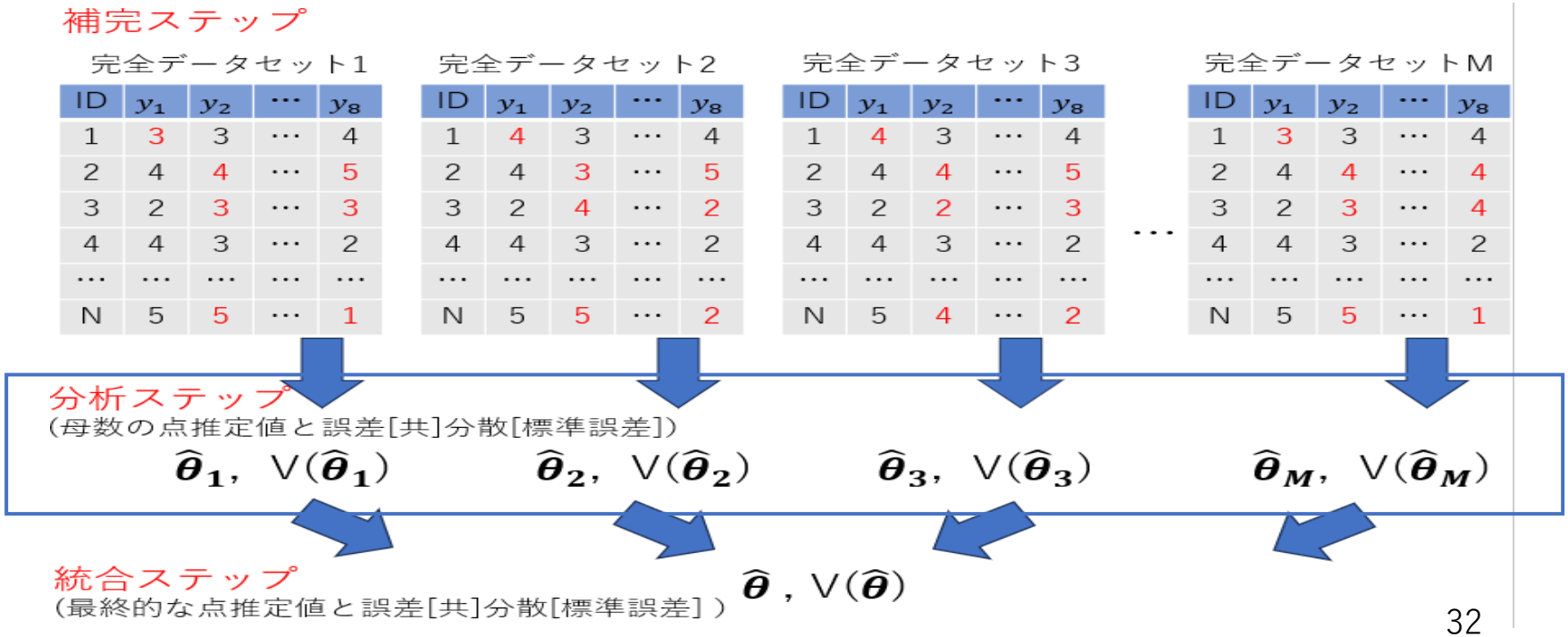
その中からランダムに選ばれた1名の観測値により補完（ここでは個人4）

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。



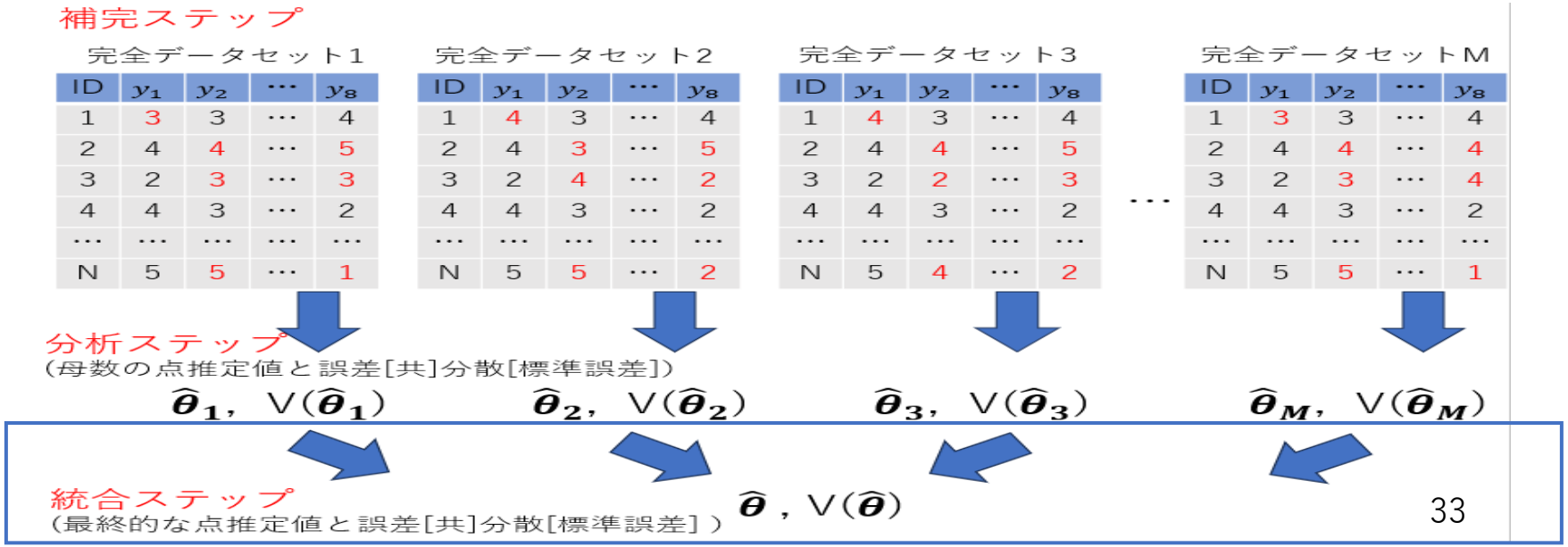
統合ステップ

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 − α)% 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100α% 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。

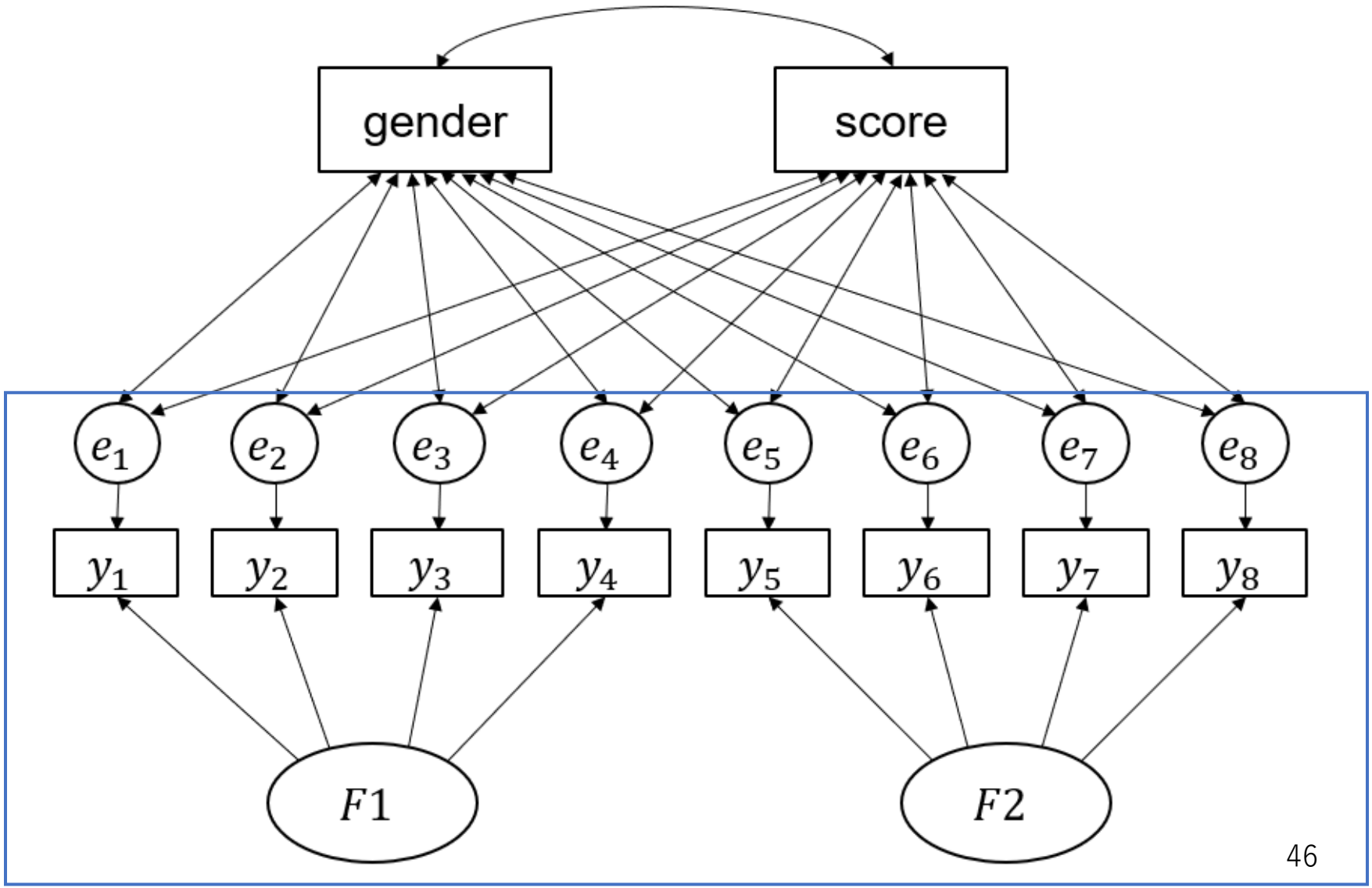
補助変数を考慮した分析

- 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, 4*8(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.