The Annual Report of Educational Psychology in Japan
2025, Vol. 64, 331-341

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

Missing Data Analysis in Educational Psychology Research

MASAYUKI SUZUKI, SATOSHI USAMI AND TAKETOSHI SUGISAWA

企画・司会・話題提供　鈴 木 雅 之（横 浜 国 立 大 学）
企画・話題提供　宇佐美　慧（東 京 大 学）
話 題 提 供　杉 澤 武 俊（早 稲 田 大 学）

企画趣旨

教育心理学研究において，データの欠測は悩ましい問題である。欠測データの処理としては，削除法（リストワイズ削除・ペアワイズ削除）や，各変数の代表値などを代入する単一代入法がしばしば利用されてきた。しかし，これらの方法を利用した場合には，推定のバイアスや検定力の低下など，推定上のさまざまな問題が生じうる。とりわけ削除法について，Wilkinson & Task Force on Statistical Inference APA Board of Scientific Affairs（1999）は，基本的な統計ソフトウェアに実装されている方法の中で「最悪の方法」と評している。そのため，削除法や単一代入法のような古典的な方法ではなく，完全情報最尤推定法や多重代入法を利用することが望ましいとされている。これら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方法とされている（Baraldi & Enders, 2010）。しかしながら，『教育心理学研究』では，欠測値の存在やその割合等について説明していない論文や，削除法を用いている論文が多いという問題がある（杉澤, 2011; 鈴木, 2018）。そこで本チュートリアル・セミナーでは，欠測が生じるメカニズムと，古典的な方法の問題点について理解を深め，完全情報最尤推定法と多重代入法の考え方，および統計ソフトウェアRでの実行方法について理解することを目的とする。

話題提供

欠測が生じるメカニズムと古典的な方法

杉澤武俊

不完全データと欠測データ処理の古典的方法

無回答などの何らかの原因により，変数の一部のデータが値として得られなかったものを欠測値といい，欠測値を含まないデータセットは完全データ，欠測値を含むデータセットは不完全データと呼ばれる。従来適用されてきた古典的な欠測データ処理の方法として，削除法と単一代入法がある。削除法はさらに，欠測値が1つでも含まれる観測対象（実験参加者など）を全ての分析から除外する「リストワイズ削除」（listwise deletion）と，分析ごとに，使用する変数に欠測が生じている観測対象のみ削除する「ペアワイズ削除」（pairwise deletion）の2種類に分けることができる。また，単一代入法とは，欠測となったデータに対して，何らかの方法により求めた1つの値を代入する方法である。単一代入法にはさまざまな手法が提案されており，欠測値の全てに当該変数の平均値を代入する「平均値代入法」（mean imputation），他の観測された変数を説明変数とする回帰式によって求めた予測値を，誤差を加味せずそのまま代入する「確定的回帰代入法」（deterministic regression imputation）などが知られている。これらの古典的な方法による欠測データ処理は，多くの場合で推定結果にバイアスが生じ，さらに，削除法ではサンプルサイズが減少することで標準誤差が増大して推定精度や検定力の低下が生じ，単一代入法でも標準誤差の評価が適切になされないなどの望ましくない性質を持つことが知られている。

欠測が生じるメカニズム

欠測データ処理の方法が妥当であるか否かには，当該データにおける欠測メカニズムが関係する。欠測メカニズムとは，欠測値にも本来得られるはずであった真の値が存在するものとして，欠測が生じなければ得られるはずであった「完全データ」を所与としたときに，どのデータが欠測するかという欠測の出現パターンに関する条件付き確率として表される。すなわち，本来得られるはずであった完全データ行列をYとし，Yの各要素について実際に観測されるものを1，欠測となるものを0と置き換えた「回答指標行列」をR,

実際の観測データを Y_{obs} と表すと，欠測メカニズムは $P(R|Y)$ と表すことができる。欠測データ処理に関する文献では，Little & Rubin（2002）が提案した枠組みに準じた欠測メカニズムの分類が一般的に用いられる。その分類とは，（1）欠測となる確率がデータのあらゆる要素と無関係に定まる，つまり，変数ごとに一定の確率で欠測となるデータをランダムサンプリングしたと見なすことができ，$P(R|Y)=P(R)$ となる「完全にランダムな欠測」（missing completely at random: MCAR），（2）欠測となる確率が実際に観測されたデータの情報だけで定まり，他の変数の実際に観測された値を一定に統制すれば欠測した変数自体の値とは関係がない，$P(R|Y)=P(R|Y_{obs})$ となる「（条件付きで）ランダムな欠測」（missing at random: MAR），（3）欠測となる確率が欠測したデータの値に依存し，実際に観測されたデータの情報だけでは決まらず $P(R|Y)\neq P(R|Y_{obs})$ となる「ランダムでない欠測」（missing not at random: MNAR，または not missing at random: NMAR と呼ばれることもある）の3種類である。これらの欠測メカニズムは，（1），（2），（3）の順に，後にいくほど妥当な分析結果を得るためにより高度な手法を要求する。

なお，ここで注意すべきは，（2）の MAR について，欠測が生じるか否かは必ずしもその変数の値自体と独立である必要はないということである。たとえば，入試成績と入学後の成績の関係を調べる状況において，入試成績で合格ライン未満である人については入学後の成績が全て欠測となるので，両変数間に強い正の相関関係があれば，欠測となるのは入学後の成績が相対的に低い人ということになる。この場合であっても，入試成績が同点である下位集団ごとに見たときに，合格ライン未満の人は欠測確率が1，合格ライン以上の人は欠測確率が0となる（あるいは，入学後の成績が悪い人ほどドロップアウトしやすいというような傾向がなく，下位集団内では一定の確率で欠測が生じる）のであれば，MAR ということになる。

実際の研究で得られたデータの欠測メカニズムがどれにあたるかについて，MCAR であるか否かは Little（1988）の検定などの観測データに基づく判定法が提案されている。しかし，MCAR とはいえない場合に，MAR と MNAR の区別を観測データのみから行うことは極めて困難であるといえる。

古典的な欠測データ処理に関するシミュレーション

古典的な欠測処理の方法を適用した場合に実際にどのようなことが起こるかについて，人工データを用いたシミュレーションによって確認する。2変数 y_1 と y_2 があり，各変数は平均がいずれも50，分散がいずれも100，相関係数が.60である2変量正規分布に従う母集団からサンプルサイズ200の標本を抽出するとき，y_2 にのみ欠測が生じる状況を想定する。欠測メカニズムとして，y_1 の値にかかわらず確率.5で y_2 が欠測する MCAR 条件と $y_1<50$ のとき確率.1，$y_1\geqq 50$ のとき確率.9で y_2 が欠測する MAR 条件の2条件を設定する。欠測値処理の方法として，リストワイズ削除（以下では欠測のない y_1 のみを用いた推定を行わないのでペアワイズ削除と見なすこともできる），平均値代入法，確定的回帰代入法の3手法を適用し，それぞれの処理を行った疑似的な完全データに対して，y_2 の平均と分散，および，y_1 と y_2 の相関係数に関する推定を行う。平均については R 言語の mean 関数と t.test 関数による点推定値および95%信頼区間（CI），分散と相関係数についてはそれぞれ R 言語の var 関数と cor 関数による点推定値を採用する。完全データの生成から，不完全データの生成，欠測値処理，各種統計量の算出に至る一連の処理を10,000回反復し，その結果をまとめたものが Table 1 である。Table 1 では，平均，分散，相関係数の点推定については10,000個の推定値の平均を，平均の区間推定では10,000個の95%信頼区間について，信頼区間の幅の平均と，真の母平均である50を含んでいるものの割合（被覆率）を示している。さらに，欠測を発生させる前の完全データで同様の推定を行った結果につい

Table 1
古典的な方法による推定結果のシミュレーション

	完全データ	MCAR			MAR		
		リストワイズ削除	平均値代入法	確定的回帰代入法	リストワイズ削除	平均値代入法	確定的回帰代入法
平均	50.0	50.0	50.0	50.0	53.7	53.7	50.0
分散	100.1	99.9	49.7	68.1	85.2	43.7	69.6
相関係数	.598	.597	.421	.725	.494	.271	.716
95% CI 幅	2.79	3.96	1.96	2.29	3.60	1.84	2.32
CI 被覆率	.952	.951	.670	.789	.019	.001	.703

ても参考として掲載している。

Table 1から，まず，欠測メカニズムがMCARであるとき，リストワイズ削除で欠測値処理を行った場合には，平均，分散，相関係数の点推定値の平均と母数の真値がほぼ一致しており（すなわち，バイアスが生じていない），95％信頼区間の被覆率も実質的に理論値通りのほぼ95％となっていることがわかる。ただし，リストワイズ削除によりサンプルサイズの期待値が100に半減することに伴って標準誤差が理論的に $\sqrt{2}$ 倍となることを反映して，信頼区間の幅が約 $\sqrt{2}$ 倍になっている。標準誤差の増大は，検定の文脈では検定力の低下につながるものである。また，MCARのもとで平均値代入法や確定的回帰代入法を適用した場合は，平均の点推定のみバイアスのない結果が得られているが，分散や相関係数の推定ではバイアスが生じ，また，平均の信頼区間に関する結果からは標準誤差を過小評価して誤った推測が行われていることがわかる。標準誤差の過小評価は検定の文脈では第1種の誤りを設定した有意水準以上の確率で犯すことにつながる。

次に，欠測メカニズムがMARであるとき，確定的回帰代入法により処理したときの平均の点推定値のみ，バイアスのない結果が得られているが，標準誤差を正しく評価できずに誤った信頼区間が得られている。これ以外の欠測値処理方法と推定量の組合せでは，いずれも点推定値にはバイアスが生じ，標準誤差の大きさも正しく評価されていない。

以上より，削除法や単一代入法による古典的な欠測値処理では，欠測メカニズムや対象となる母数の組合せによっては妥当な結果が得られることもあるものの，その条件は限定的であり，多くの場合に無視できない不適切な分析結果が得られるということができる。

実際のデータ分析における欠測値処理

欠測メカニズムがMCARの場合，一般的な研究の文脈では，標準誤差や信頼区間を求めたり検定を行ったりすることが多い現状をふまえれば，妥当な分析結果が得られるリストワイズ削除はひとつの選択肢となり得るが，常に標準誤差の増大を伴うことから望ましい方法とはいえない。ペアワイズ削除はリストワイズ削除よりも標準誤差を小さく抑えられる可能性はあるものの，3変数以上の場合に完全データからは得られることのない（具体的には，半正定値でない）相関係数行列が得られて多変量解析の適用に支障をきたすことがある。したがって，MCARの場合であっても古典的方法による欠測値処理は推奨されず，MCARおよびMARの場合においてバイアスのない推定が可能であり，標準誤差を小さく抑えることのできる完全情報最尤推定法や多重代入法が現在推奨されている。MNARの場合は，各データの欠測確率を説明する適切な補助変数を加えることでMARと見なせるならMARと同様の処理となる。そうでなければ回答指標行列 R（あるいは R の0と1を入れ替えた欠測指標行列としても同じ）の具体的な確率モデルを指定する必要がある。これらの手法についてはこの後続く話題提供を参照されたい。

完全情報最尤推定法と多重代入法

宇佐美　慧

MARに基づく欠測下では，リストワイズ削除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である。

FIMLは，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であり，補完（代入）は伴わない。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では，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。このように，補完モデルと分析モデルが明確に区別されるのが特徴である。

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（つまり，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（e.g., 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）を分析モデルとする場合にFIMLの実装は容易である（Newsom, 2015)。一方，MIの方が有用，あるいはその使用が現実的と考えられる場合も多くある。例えば，自治体によるデータの二次利用を目的とした補完を行う場合のような，補完の実行者と分析者が異なる状況が関係する。このとき，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）に欠測が仮に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である（高井他, 2016)。他にも，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は直接的で有用であり，またそもそも統計分布や最尤法を前提としない多変量解析法も多い。

チュートリアルでは以上のように，FIMLとMIのそれぞれの基本的な特徴と使い分けのポイントについて概観してから，各方法の詳細を説明した。その後，補助変数の活用についての補足とともにまとめを行った。当日の発表スライドは https://usami-lab.com/教心チュートリアル2024_宇佐美.pdf のリンクから参照可能である。

FIML

上述の通り，FIMLでは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する。その際，個人ごとの尤度を考える。またy_1とy_2の間の母相関係数を推定したい状況を考える。直接尤度は，個人1の尤度×個人2の尤度×個人3の尤度…×個人Nの尤度，と表現できる。仮想データに基づく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）に欠測が生じる場合を考え，リストワイズ削除による分析とFIMLを行った場合の結果を示した。すなわち，リストワイズ削除では母相関係数が大幅に過少推定され，一方で，FIMLの場合では母相関係数に近い値が推定される。

（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。例えば，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はSEMの範疇で表現可能である。また，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である。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では，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θを推定する。後者は平均構造$\mu(\theta)$，共分散構造$\Sigma(\theta)$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θの推定を行う。欠測がある場合の直接尤度は，各個人で観測された変数に対応する$\mu(\theta)$，$\Sigma(\theta)$の一部要素を利用することで表現できる。

チュートリアルでは具体例として，SEMに基づく回帰分析モデルの推定と直接尤度の表現を説明した。もちろん，回帰分析モデルに限らず，SEMで表現できる下位モデルであれば個々のモデルに応じた$\mu(\theta)$，$\Sigma(\theta)$の表現が可能なので，欠測があっても先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。

チュートリアルでは主として，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θ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）。

MI

例えば，確定的回帰代入を行う（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）方法は，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。結果として，例えば先の例で挙げた，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）に欠測が生じる場合では，y_2の分散が過少推定される。

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であり，3つのステップから構成される。すなわち，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補完ステップ，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θを推定する分析ステップ，そして得られた複数のθの推定結果を統合する統合ステップである。冒頭で述べたように，MIでは，補完モデルと分析モデルが明確に区別される。またMIは，SEMの文脈でも実装は容易であるが（lavaanやMplus），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。

MIには，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。対して，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方法であり，汎用性が高い。特に，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。チュートリアルではMICEに基づく多重代入法について説明した。

MIの最初のステップである補完ステップは，補完モデルの設定，初期値の設定，連鎖方程式による補完値の更新と反復，から成る。補完モデルの設定に際して，扱う変数が連続変数の場合，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。また補完モデルには，分析モデルに含まれない変数を含めてよいが，一方で分析モデル内の変数は原則含めることが求められる。初期値に関しては，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。

補完ステップで最も肝要なのが，連鎖方程式による補完値の更新である。例えば，$y_1, y_2, \dots y_8$ の8種類の変数それぞれに欠測を伴うデータを扱う場合を考える。最初に，y_1 内にある欠測値を，疑似的な完全データセット内の $y_2, y_3, \dots y_8$ 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。次に，y_2 内にある欠測値を，疑似的な完全データセット内の $y_1, y_3, \dots y_8$ 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。y_3 以降の変数についても同様の作業を繰り返していく。より具体的な更新方法については上述の URL および Rubin（1987），野間（2017）も参照されたい。

初期値は通常ラフなものであり，T 回の反復（R 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。

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。これは，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。

分析ステップでは，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。結果，分析モデル内の母数 θ 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。

最後が統合ステップであり，Rubin's rule に基づいて説明した。すなわち，点推定値（$\hat{\theta}$）として，各完全データセットから得られた推定値（$\hat{\theta}_m$; $m = 1, 2, \dots M$）の平均を利用する。$\hat{\theta}$ の誤差共分散行列 $V(\hat{\theta})$ については，各完全データセットから得られた $\hat{\theta}_m$ の誤差共分散行列の推定値 $V(\hat{\theta}_m)$ に基づいて評価できる，補完値内・補完値間の共分散行列を利用することで得られる（上述の URL および，高井他, 2016, pp. 117-118 も参照されたい）。特定の母数 θ に関する標準誤差の推定値 $se(\hat{\theta})$ は，$V(\hat{\theta})$ の対応する対角要素の正の平方根に等しい。チュートリアルでは特定の母数 θ に関する帰無仮説検定の方法や信頼区間の推定，およびそれらに付随する自由度の推定量についても説明した。

疑似的な完全データセット数 M については，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 が必要である（野間, 2017, p. 69）。比較的最近の研究の例としては，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。一般に，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，本チュートリアルでは，少なくとも $M = 20$，可能であれば $M = 50, 100$ 程度は確保する必要があることを述べた。

補助変数の活用

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）。

特に欠測の割合が大きいとき，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。

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., Enders, 2025, pp. 325-326; Newsom, 2015, pp. 18-25），飽和した相関アプローチ（saturated correlated approach）は簡便であり，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。ただし，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ことには留意する必要がある。

まとめと MNAR の場合

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）も有用である。一方，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である。

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（本チュートリアルの分析例，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。

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。

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である。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, 2022; Newsom, 2015; 高井他, 2016）。

ただし現状において，絶対的に優れた方法があるとは言えず，そのため MNAR に基づく欠測が想定される場合には，感度分析の実行が重要とされている。すなわち，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。一方，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。

統計ソフトウェア R での分析例

鈴木雅之

本稿では，心理尺度を用いて調査を行った研究を想定し，項目得点に欠測が生じている状況において，確認的因子分析，尺度得点の算出，回帰分析を統計ソフトウェア R で行う方法について紹介する[1]。FIML 法による分析を行うためのパッケージとしては lavaan や sem，MI 法による分析を行うためのパッケージとしては mice や Amelia など様々なものがあるが，本稿では FIML 法と MI 法を同一の枠組みで行うことを目的に，lavaan（Rosseel, 2012）と mice（van Buuren & Groothuis-Oudshoorn, 2011），semTools パッケージ（Jorgensen et al., 2022）を用いて分析する方法を紹介する[2]。semTools パッケージには，lavaan パッケージの関数 cfa() や sem() と同様の方法で，mice パッケージや Amelia パッケージで補完を行ったデータに対して分析・統合をするための関数がある[3]。

Figure 1

「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

1　データファイルおよび R のスクリプトは電子付録で公開している。

2　mice パッケージの使い方の詳細は高橋・渡辺（2017），lavaan と semTools パッケージについて豊田（2014）などを参照されたい。

3　厳密には，補完・分析・統合を一括で行うための関数もあるが，本稿では紹介しない。

Table 2
各変数の欠測状況

	school	gender	y1	y2	y3	y4	y5	y6	y7	y8	score
欠測数（人数）	0	0	4	5	4	9	4	17	17	24	14
欠測割合（%）	0	0	1.3	1.7	1.3	3.0	1.3	5.7	5.7	8.0	4.7

中学生300名を対象に，内発的動機づけと外発的動機づけに関する質問紙調査（5件法），および学力テスト（100点満点）を実施したとする。調査とテストの結果が「example.csv」として保存されている（欠測を生じさせる前の完全データは「complete.csv」として保存されている）。このファイルをExcelで開いたものをFigure 1，データの欠測状況をTable 2に示す。schoolは生徒が所属する学校の種別（1＝公立，2＝国立，3＝私立），genderは性別（0＝男性，1＝女性），y1―y4は内発的動機づけを測定するための項目の得点，y5―y8は外発的動機づけを測定するための項目の得点，scoreはテスト得点であり，「.」は欠測をあらわす。

このCSVファイルを，関数read.csv()を用いて，dat_misというオブジェクトに保存する。その際，引数na.strings＝″.″で，ピリオド（.）が欠測値であることを指定する。

```
# データの読み込み
dat_mis <- read.csv("example.csv", na.strings = ".")
```

次に，本稿で紹介する分析を行うために必要なパッケージを，関数library()を用いて読み込む。

```
# パッケージの読み込み
library(mice)
library(lavaan)
library(semTools)
library(mitml)
```

多重代入法による補完

miceパッケージの関数mice()を用いて補完を行った結果を，dat_MIというオブジェクトに保存する。

```
# 関数miceによる補完
dat_MI <- mice(data = dat_mis, method = "pmm",
m = 100, maxit = 30, seed = 23109, printFlag =
FALSE)
```

引数dataでデータの指定，引数methodで代入法の指定を行う。ここでは，予測平均マッチング（pmm）を指定している。引数methodでは，変数ごとに代入法の指定をすることも可能であり，他にも線形回帰（norm）や，ロジスティック回帰（logreg）などを指定することができる。引数mでは作成する疑似完全データの数，引数maxitでは反復回数を指定する。ここでは，疑似完全データの数を100（デフォルトでは5），反復回数を30回としている（デフォルトでは5）。引数seedでは，再現性を確保するための任意のシード値を設定する。最後に，引数printFlag＝FALSEでは，コンソール画面に反復情報を出力しないように設定している。

疑似完全データの確認をしたい場合には，関数complete()を利用する。たとえば，1つ目の疑似完全データを確認するためには，complete(dat_MI, 1)，3つめの疑似完全データの最初の8行を確認するためには，head(complete(dat_MI, 3), 8)とする。

確認的因子分析

FIML法による確認的因子分析をlavaanパッケージの関数cfa()，MI法による確認的因子分析をsemToolsパッケージの関数cfa.mi()を用いて行う。いずれの場合も，まずはモデルの記述を行う。ここでは，内発的動機づけに関する4つの変数（y1―y4）が因子「F1」，外発的動機づけに関する4つの変数（y5―y8）が因子「F2」から影響を受けることを仮定したモデルを記述し，CFA_modelというオブジェクトに保存する。

```
# モデルの記述（確認的因子分析）
CFA_model <- '
F1 =~ y1 + y2 + y3 + y4
F2 =~ y5 + y6 + y7 + y8
'
```

モデルを記述する部分は，クォーテーション（'）で囲む。lavaanパッケージでは，各変数がどの因子から影響を受けているかを「=~」を用いて表現する。具体的には，「=~」の左側に因子名，右側にその因子から影響を受ける観測変数の名前を記述する。

モデルを記述したら母数の推定を行う。まず，FIML法による確認的因子分析について，関数cfa()による

Table 3
確認的因子分析の推定結果のまとめ（因子負荷と標準誤差）

	完全データ		FIML		FIML（補助変数あり）		MI		リストワイズ削除	
	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
F1 → y1	0.612	0.082	0.634	0.094	0.627	0.094	0.629	0.084	0.462	0.102
F1 → y2	0.755	0.073	0.744	0.079	0.746	0.079	0.746	0.075	0.736	0.100
F1 → y3	0.532	0.077	0.560	0.087	0.557	0.087	0.553	0.079	0.471	0.098
F1 → y4	0.581	0.074	0.562	0.078	0.565	0.078	0.564	0.076	0.543	0.097
F2 → y5	0.958	0.071	0.958	0.072	0.959	0.072	0.959	0.073	0.917	0.089
F2 → y6	1.222	0.067	1.214	0.069	1.213	0.068	1.210	0.069	1.155	0.083
F2 → y7	0.825	0.061	0.846	0.063	0.848	0.064	0.839	0.063	0.863	0.075
F2 → y8	0.703	0.077	0.694	0.082	0.699	0.082	0.694	0.079	0.669	0.094
F1 ⇔ F2	0.403	0.067	0.393	0.070	0.395	0.070	0.399	0.069	0.398	0.087

注）　推定値は非標準化推定値である。

推定結果を CFA_FIML というオブジェクトに保存し，関数 summary()　で結果を出力する。

```
# FIML 法による母数の推定（確認的因子分析）
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)
```

最初の引数で，記述したモデルを代入したオブジェクト（CFA_model）を指定し，引数 data でデータを指定する。また，missing = "fiml" では，FIML 法による分析をすることを指定している。最後に，引数 std.lv = TRUE では，因子の分散を 1 に固定して分析をすることを指定している。デフォルトでは，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が 1 に固定されて推定が行われる[4]。

次に関数 summary() では，最初の引数で，分析結果を代入したオブジェクト（CFA_FIML）を指定し，引数 fit.measures = TRUE で適合度指標，引数 standardized = TRUE で標準化解，引数 ci = TRUE で 95％信頼区間の出力を指定している。

補助変数を含める場合には，semTools パッケージの関数 cfa.auxiliary() を用いる。関数の使い方は関数 cfa() と同様であり，引数 aux で補助変数を指定する。たとえば，性別とテスト得点を補助変数とする場合には，引数に aux = c("gender", "score") を加える。

次に，MI 法による確認的因子分析について，semTools パッケージの関数 cfa.mi() は，関数 cfa() とほとんど同一の方法で使うことができる。関数 cfa() との違いは，補完後のデータを指定すること，引数の missing = "fiml" が不要であること，関数 summary で結果を統合・出力する際に，引数 output = "data.frame" を加えること，である。

```
# MI 法による母数の推定（確認的因子分析）
CFA_MI <- cfa.mi(CFA_model, data = dat_MI,
std.lv = TRUE)
# 結果の統合・出力
summary(CFA_MI, fit.measures = TRUE,
standardized = TRUE, ci = TRUE, output = "data.
frame")
```

以上の方法で分析した結果と，リストワイズ削除により分析を行った結果，欠測のない完全データで分析を行った結果を Table 3 に示す。リストワイズ削除では，他と比べて推定値に乖離（過小推定）が生じているとともに，標準誤差も完全データの場合と比べて大きくなっていることが確認できる。

重回帰分析

FIML 法による重回帰分析を lavaan パッケージの関数 sem()，MI 法による重回帰分析を semTools パッケージの関数 sem.mi() を用いて行う。分析に先立ち，関数 rowMeans() を用いて尺度得点（1 項目あたりの平均値）を算出する。y1—y4 の平均値を内発的動機づけ得点として Int という変数名，y5—y8 の平均値を外発的動機づけ得点として Ext という変数名で保存する。

[4]　因子負荷の非標準化推定値を比較するため，ここでは因子の分散を 1 に固定して推定を行っている。

```
# 尺度得点の作成
dat_mis$Int <- rowMeans(dat_mis[c(3:6)])
dat_mis$Ext <- rowMeans(dat_mis[c(7:10)])
```

関数 mice() で作成した疑似完全データについて尺度得点を算出するためには，まず，mitml パッケージ（Grund et al., 2023）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する。次に，疑似完全データの数だけ計算を繰り返すため，for 文による繰り返し処理を利用する。

```
# 尺度得点の作成（疑似完全データ）
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)])
}
```

重回帰分析も確認的因子分析と同様，まずモデルを記述する。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 というオブジェクトに保存する。

```
# モデルの記述（重回帰分析）
reg_model <- '
score ~ Int + Ext
Int ~~ Int
Ext ~~ Ext
Int ~~ Ext
'
```

「~」の左側に従属変数，右側に独立変数の名前を記述する。また，「~~」は共分散をあらわし，「Int ~~ Int」は内発的動機づけの分散，「Int ~~ Ext」は内発的動機づけと外発的動機づけの共分散をあらわす。lavaan パッケージでは，独立変数の分散・共分散の推定を指定しないと，独立変数に欠測のあるケースは除外される。

FIML 法による重回帰分析について，関数 sem() による推定結果を reg_FIML というオブジェクトに保存し，関数 summary() で結果を出力する。関数の使い方は確認的因子分析の時と同様であり，ここでは標準化解と決定係数，95％信頼区間を出力する。

```
# FIML 法による母数の推定（重回帰分析）
reg_FIML <- sem(reg_model, data = dat_mis,
missing = "fiml")
# 結果の出力
summary(reg_FIML, standardized = TRUE,
rsquare = TRUE, ci = TRUE)
```

MI 法による重回帰分析は，semTools パッケージの関数 sem.mi() を用いる。

```
# MI 法による母数の推定（重回帰分析）
reg_MI <- sem.mi(reg_model, data = dat_MI_list)
# 結果の統合・出力
summary(reg_MI, standardized = TRUE, rsquare =
TRUE, ci = TRUE, output = "data.frame")
```

以上の方法で分析した結果と，リストワイズ削除により分析を行った結果，欠測のない完全データで分析を行った結果を Table 4 に示す。リストワイズ削除では，完全データの場合と比較して，推定値に乖離が生じているとともに，標準誤差も大きくなっていることが確認できる。また，FIML 法についても，推定値の乖離と標準誤差の増大がみられる。これは，項目得点に 1 つでも欠測があると尺度得点が欠測していることになり，内発的動機づけ得点と外発的動機づけ得点の

Table 4
重回帰分析の推定結果のまとめ（偏回帰係数と標準誤差）

	完全データ		FIML		MI		リストワイズ削除	
	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
内発的動機づけ	3.521	0.622	3.378	0.665	3.384	0.633	3.379	0.762
外発的動機づけ	0.131	0.474	−0.260	0.567	0.009	0.483	−0.495	0.562
R^2	.105		.091		.095		.086	

注） 推定値は非標準化推定値である。

欠測割合が大きくなるためである（内発的動機づけ得点の欠測の割合は7.3%，外発的動機づけ得点の欠測の割合は18.3%）。換言すると，１つの下位尺度を構成する４項目のうち１項目だけ欠測している場合，残りの３項目に関する情報が活用されないために，推定の精度は低下してしまう。同様に，尺度得点を算出し，尺度得点だけを用いてMI法を利用した場合も，推定の精度は低下する。そのため，項目レベルで欠測が生じている場合には，尺度得点を算出してから補完を行うのではなく，本稿で紹介したように，項目レベルで補完を行い，その後に尺度得点を算出する方法が有用といえる。ただし，項目数が非常に多い場合には結果が収束しなかったり，そもそも補完ができなかったりすることがある。こうした問題への対処については，Enders（2010）で詳しく議論されている。

引 用 文 献

Baraldi, A. N., & Enders, C. K. (2010). An introduction to modern missing data analyses. *Journal of School Psychology*, *48*, 5–37. https://doi.org/10.1016/j.jsp.2009.10.001

Enders, C. K. (2010). *Applied missing data analysis*. Guilford.

Enders, C. K. (2022). *Applied missing data analysis* (2nd ed.). Guilford.

Enders, C. K. (2025). Missing data: An update on the state of the art. *Psychological Methods*, *30*(2), 322–339. 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. https://doi.org/10.1146/annurev.psych.58.110405.085530

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. https://doi.org/10.1007/s11121-007-0070-9

Grund, S., Robitzsch, A., & Luedtke, O. (2023). mitml: Tools for Multiple Imputation in Multilevel Modeling. R package version 0.4-5. https://CRAN.R-project.org/package=mitml

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*, Article 168. https://doi.org/10.1186/s12874-018-0615-6

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., & Rosseel, Y. (2022). semTools: Useful tools for structural equation modeling. R package version 0.5-6. https://CRAN.R-project.org/package=semTools

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. https://doi.org/10.1037/met0000381

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*, 1198–1202. https://doi.org/10.1080/01621459.1988.10478722

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). John Wiley & Sons, Inc.

Newsom, J. T. (2015). *Longitudinal structural equation modeling: A comprehensive introduction*. 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. https://doi.org/10.5023/jappstat.46.67

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. John Wiley & Sons, Inc.

杉澤武俊（2011）. 測定・評価に関する動向と方法論研究のススメ　教育心理学年報, *50*, 126–135. https://doi.org/10.5926/arepj.50.126

鈴木雅之（2018）. 測定・評価・研究法に関する研究動向と展望―統計的分析手法の利用状況と評価リテラシーの育成に向けて　教育心理学年報, *57*, 136–154. https://doi.org/10.5926/arepj.57.136

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

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

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

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3), 1–67. https://doi.org/10.18637/jss.v045.i03

Wilkinson, L., & Task Force on Statistical Inference APA Board of Scientific Affairs. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*(8), 594-604. https://doi.org/10.1037/0003-066X.54.8.594

Yuan, K.-H., & Bentler, P. M. (2010). Consistency of normal-distribution-based pseudo maximum likelihood estimates when data are missing at random. *The American Statistician*, *64*(3), 263-267. https://doi.org/10.1198/tast.2010.09203

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. https://doi.org/10.1146/annurev-orgpsych-041621-031401