欠損データ分析 (missing data analysis)
-完全情報最尤推定法と多重代入法-

村山 航*

1/16/2011

調査研究，特に縦断調査などを行うときには，欠損値 (missing data, missing value) は頭の痛い問題である。伝統的に欠損値はリストワイズ法 (list-wise deletion) やペアワイズ法 (pair-wise deletion) などを用いて対処されてきた。また，平均値や回帰による代入法などを用いる場合もある。しかし，こうした伝統的な方法は，後述するように，欠損値が完全にランダムに生じるような状況でない限り，推定値にバイアスが生じることが分かっている。

一方，近年では伝統的な方法に代わる方法として，完全情報最尤推定法 (full information maximum likelihood method, FIML) や多重代入法 (multiple imputation method) が脚光を浴びてきた[1]。これらの方法は，欠損値に関して，伝統的な方法より弱い仮定のもとでバイアスのない推定値を与えてくれるものであり，近年のデータ解析のスタンダードになりつつある方法である。本稿では，まず欠損値の種類（欠損メカニズム）として，missing completely at random (MCAR)，missing at random (MAR) そして missing not at random (MNAR) の区別を解説する。そのなかで FIML 法や多重代入法のメリットを述べ，それぞれの方法についてできるだけ具体的な解説を行うことを目的とする。その際，補助変数 (auxiliary variable) の役割についても説明したい。なお，本稿は Enders (2010) の解説をベースにしているので，興味のある人はそちらも読むとよい。

1 MCAR, MAR そして MNAR

欠損値はその性質（欠損値発生のメカニズム）によって次の 3 つのタイプに分けることができる (Rubin, 1976)。図 1 はその概念図である。

1.1 Missing Completely At Random (MCAR)

MCAR とは，いわゆる欠損値が完全にランダムに生じているようなケースである。すなわち，欠損値の有無が他の分析に含まれる変数や，その変数自体の値とは無関係であるケースを示している。図 1 の左側のパネルにその概念図を記した。Y というのが欠損値のある変数，X がその他の分析に含まれる変数である。R は，欠損したときには 0，観測されたときには 1 の値をとる確率変数である。欠損が生じるかどうかの確率変数 R が，X とも Y とも無関係であることが分かるだろう。これが MCAR である。

*e-mail: murakou@orion.ocn.ne.jp 覚書なので，不正確な部分があるかもしれません。気づいた方はご連絡ください。

[1]ただし，これらの方法自体は古くから知られており，近年の SEM や HLM のソフトウェアに実装されるようになったことで，応用研究での適用が格段に増えたと言ったほうが正確だろう。

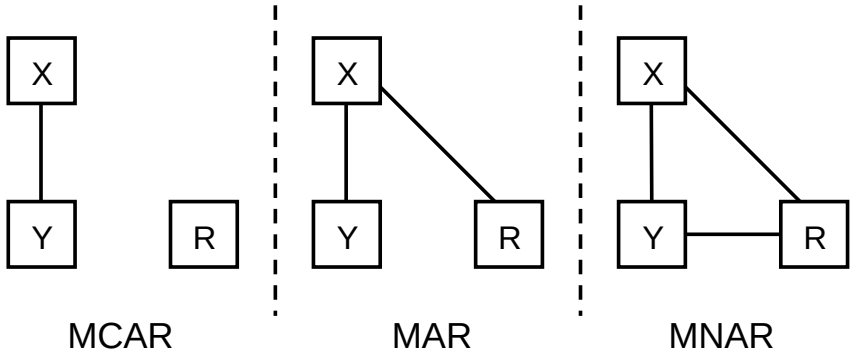

Figure 1: MCAR, MAR, MNAR の概念図。Y が欠損値のある変数。R が欠損値の有無をコーディングした確率変数。X は分析に含まれるその他の変数

1.2 Missing At Random (MAR)

MAR は MCAR と聞いた感じが似ているかもしれないが，MCAR よりもずっと制約が弱まっている。具体的には，Y における欠損値の有無は，他の変数 X と関係しているが，(X を統制したとき）その変数 Y 自体の値とは無関係であるケースを示している。MCAR と違うのは，他の変数と欠損値の有無とが関係することを許容している点である [2]。図 1 の真ん中のパネルがその概念図である。R は X と関係しているが，欠損値を持つ Y 自体とは関係していない。

たとえば，Table 1 の例を考えてみる[3]。この例では，極端ではあるが，IQ テストを全員に実施し，その得点が低かった人にはその後の適性検査を実施しなかった場合の測定値である（またその他の変数として動機づけも測定した)。また，もし測定した場合に得られたはずのデータも記してある。このデータは MCAR ではない。なぜなら，適性検査（図でいう Y ）の欠損の有無は IQ (図でいう X) と関連しているからである。実際，欠損値のある人とない人を比較すると，IQ の値には大きな違いがある。しかし，MAR だということができる。なぜなら，欠損は IQ によって生じただけであり，この要因を統制すると，欠損の有無はテストの成績自体とはまったく関係がなくなるからである。

ここで 1 つ重要な点がある。**Y と R が無関係であることが MAR の大切な要件だが，それはあくまでも分析に含まれる変数 X を統制した後のことである**。たとえばこの例では，IQ と適性検査との間に正の相関関係がある。その場合，IQ が低い人が欠損値になっているのだから，IQ を欠損している人は適性検査の得点も低いはずである（表の「欠損なしの場合」をみると分かるだろう)。すなわち，欠損値の有無 R と，欠損値のある変数 Y の値は相関しており，一見 MAR の仮定が成り立っていないようにみえる。しかし，このことは MAR にとって問題にはならない。大切なのは，分析に含まれる X を統制したときに，Y と欠損値の有無に関係が残るかどうかである。この例だと，IQ が欠損値の唯一の規定因であるので，IQ を統制すると，適性検査の得点と欠損の有無に関係はみられなくなる。したがって，MAR が満たされていると考えることができるのである。図 1 における Y と R の間を結ぶ線は X を統制したあとの，一種の偏相関のようなものだと考えると理解しやすい。

このポイントは，さらに重要なことを示唆している。それは，**MAR が満たされるかどうかは，**

[2]むしろ FIML や多重代入法では，この関係が強いほど，他の変数から欠損値の情報を借りることができるため，望ましくなる。
[3]R で適当な相関を持つ乱数を発生させている。

分析に含める変数に依存するということである。上の例で言えば，IQ が分析に含まれればテストの欠損値は MAR になるが，IQ が分析に含まれなければ（たとえば動機づけと適性検査だけが分析に含まれた場合），欠損値の有無が Y の値自身と相関することになり，MAR とはみなしえなくなる。MAR とは欠損値を持つ変数自体に関するものではなく，分析に含まれる他の変数との関係に依存するものであるという点には注意しなくてはならない。

Table 1: MAR データの例

id	動機づけ	IQ	適性検査	適性検査（欠損なしの場合）
1	3	83	n/a	93
2	4	85	n/a	99
3	5	95	n/a	98
4	2	96	n/a	103
5	5	103	128	128
6	3	104	102	102
7	2	109	111	111
8	6	112	113	113
9	3	115	117	117
10	3	116	133	133
平均値	3.6	101.8	117.3	111.7

1.3 Missing Not At Random (MNAR)

MNAR とは，分析に含まれる他の変数を統制した後でも，欠損値の有無が欠損値を持つ変数自身と関係を持つケースを示している。図 1 の右側のパネルがその状況を示している。X を完全に統制したあとでも，Y と欠損値の有無である R の間に関係が残っている。

1.4 補助変数 (auxiliary variable)

MAR の仮定が満たされるかどうかは分析に含まれる変数に依存することを述べた。言い換えれば，MNAR であっても，適切な変数を分析に含むことで MAR の仮定を満たすこともありうる。そこで，分析には直接関係のない変数であっても，MAR の仮定を満たすために，分析に意図的に組み込むというアイディアが出てくる（図 2 参照）。こうした変数（分析には直接関係ないが，MAR を満たすために組み込む変数のこと）のことを補助変数 (auxiliary variable) と呼び，こうした変数を組み込んだ分析アプローチを inclusive analysis strategy と呼ぶ (Enders, 2010; Rubin, 1996; Schafer & Graham, 2002)。補助変数はあくまで MAR の仮定を満たすためのものであり，研究者が想定しているモデル自体にはインパクトを持たないような形で組み込まれる。当然のことながら，補助変数は，欠損値の有無と相関が高い変数ほどよい。具体的に補助変数を（FIML や多重代入法で）どのように組み込めばよいのかは後述する。

補助変数は，原理的にいくつあってもよい。極端な話，適切な補助変数が分からない場合，質問紙で測定したすべての変数を（分析に用いなくても）補助変数として投入し，データをできるだけ MAR に近づけるというアプローチも可能である。実際，これまでのシミュレーション研究では，

たとえゴミのような補助変数が多量に含まれていたとしても，メインの分析の推定値の精度は低下しないことが示されており (Enders, 2008)，このようなストラテジーの妥当性が示されている。ただし，補助変数を多くすると，計算が複雑になり計算的負荷が高まる。また，FIML では多数の補助変数を分析に組み込むにはやや労力が必要となる（後述）。したがって，ある程度適切な補助変数をあらかじめ考えた上で選択して，補助変数として投入する方が現実的であろう。

なお，いくら補助変数を投入しても，欠損値の有無 R とその変数 Y との間に直接的な因果関係があった場合には，MAR の仮定を満たすことができない（偏相関を用いても因果関係が残ることと同じである）。たとえば，上の例でいえば，適性検査の得点が低そうな受験者は最初から諦めてテストを受けに来ないかもしれない。このような因果関係があれば，補助変数を入れても MAR は満たされないことがある。その点は留意しておく必要がある。

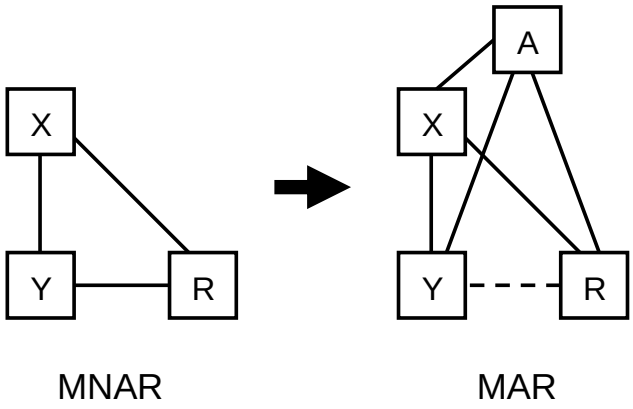

Figure 2: 補助変数を含めた inclusive analysis strategy の概念図。A が新たに加えた補助変数 (auxiliary variables)。A を加えることによって，欠損パタンが MNAR から MAR になる。

1.5 欠損値のタイプと分析方法

伝統的な欠損値処理法であるリストワイズ法やペアワイズ法は，MCAR の前提のもとでのみバイアスのない推定値を与えてくれる。逆にいえば，欠損値発生のメカニズムが MAR であるならば，この伝統的な方法は不適切だということになる。実際，Table 1 をみると，リストワイズ法やペアワイズ法を用いて適性検査の平均値を算出すると 117.3 点であり，実際の平均値 111.7 を過大評価していることが分かる。一方，Rubin は，完全情報最尤推定法 (full information maximum likelihood method; FIML) や多重代入法を用いると，欠損値の発生メカニズムが MCAR だけでなく MAR であってもバイアスのない推定値を得ることができることを示した。したがって，MAR のときには FIML や多重代入法を使う必要がある。

MCAR であれば，FIML や多重代入法を使わなくていいかというとそういうわけでもない。リストワイズ法は，基本的に欠損値のあるケースを削除する方法なので，MCAR であっても統計的検定力が低下するという問題点がある。一方，FIML や多重代入法は，もちろん欠損値が多いと検定力が落ちるが，欠損値以外の情報をすべて用いて統計的分析に生かすため，統計的検定力は相対的に高くなる。したがって，基本的に MCAR の場合であっても，FIML や多重代入法よりもリストワイズ法などを選択する必然性はない。

MNAR の場合は，FIML や多重代入法を用いても，推定値にバイアスが生じる。この場合，欠損値のメカニズムを明示的に組み込んだ分析が必要になる。こうした分析には，Heckman (1979) の selection model や，GHlynn, Laird & Rubin (1986) の pattern mixture model などがある。しかし，MNAR に基づいた分析方法というのは検証不能な前提に多く依存しており，まだ十分に発展していない。また，MAR の仮定を著しく逸脱するような欠損データは稀であることも指摘されている (e.g., Schafer & Graham, 2002)。したがって，MAR にもとづいた FIML や多重代入法を用いるのが多少バイアスがあっても現実的にはベターである場合が多い。また，大切なのは，欠損値が MNAR であっても，上に書いたように適切な補助変数を分析に含むことで，MAR の仮定をみたすことがありえるということである。したがって，欠損値のメカニズムが分からない場合には，とりあえず MAR-based の FIML や多重代入法を用い，補助変数を併用することで，これらの方法を用いる正当性を高めるというストラテジーも有効だろう (Schafer, 2003, p. 30)。

なお，欠損値が MAR の仮定を満たしているかを直接検討する方法は存在しない。あくまで欠損値発生のメカニズムをデータ収集の状況から推測するしかない。一方，MCAR は，たとえば欠損値のある人たちとない人たちで，他の変数の平均値に差があるかどうかを調べるといった方法などで，直接検討することが可能である [4]。MCAR であることが示唆されれば，MAR の仮定も満たしているので，FIML や多重代入法を正当化する根拠になる。

2 完全情報最尤推定法

完全情報最尤推定法 (full maximum likelihood method; FIML) とは，仰々しい名前がついているが，実は普通の最尤推定法である。より正確にいうと，ケースごとに欠損パタンに応じた個別の尤度関数を仮定した最尤推定法であるが，通常の最尤推定と何ら変わりないことはこれから明らかになるだろう。そうであるにも関わらず，応用研究では FIML という名称が用いられることが多いので，本稿でもこの呼称を用いることにする。

2.1 多変量正規分布のもとでの最尤推定

FIML を解説する前に，通常の多変量正規分布を用いた平均と分散共分散の最尤推定法を考える。多変量正規分布とは，p 変数の確率変数ベクトル $\boldsymbol{x}$ （ $p \times 1$ の縦ベクトル）の密度関数が，

$$f(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \times \exp\left(-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) \tag{1}$$

という形をしている分布のことである。$\boldsymbol{\mu}$ は確率変数の平均（ベクトル）を，$\boldsymbol{\Sigma}$ はこの確率変数の分散共分散行列である。最尤推定では，データがこの分布にしたがっていると仮定した上で，データから $\boldsymbol{\mu}$ と $\boldsymbol{\Sigma}$ を求める作業にほかならない。

ここである 1 人からデータ $\boldsymbol{x}_1$ が得られたとしよう。すると，平均 $\boldsymbol{\mu}$，分散共分散 $\boldsymbol{\Sigma}$ である多変量正規分布から，このデータが得られる確率 [5]は

$$f(\boldsymbol{x}_1|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \times \exp\left(-\tfrac{1}{2}(\boldsymbol{x}_1-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_1-\boldsymbol{\mu})\right) \tag{2}$$

[4]他の方法として Little (1988) の MCAR test などもある。
[5]厳密には確率密度だが，煩雑になるので以下では基本的に確率と表現する。

のようになる。同じように，2 番目，3 番目，... i 番目の人からデータ $\boldsymbol{x}_i$ が得られたとき，このデータが同じ多変量正規分布のもとで得られる確率はそれぞれ

$$f(\boldsymbol{x}_i|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \times \exp\left(-\tfrac{1}{2}(\boldsymbol{x}_i-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu})\right) \tag{3}$$

のように書ける。$\boldsymbol{x}_1$ が $\boldsymbol{x}_i$ に変わっただけである。したがって，データが独立だと仮定すると，この多変量正規分布のもとで，$i = 1, 2, \ldots, N$ のデータが得られる同時確率は，これらを掛け合わせて

$$f(\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_N|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \prod_{i=1}^{N} \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \times \exp\left(-\tfrac{1}{2}(\boldsymbol{x}_i-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu})\right) \tag{4}$$

となる。この確率は，残念ながら平均 $\boldsymbol{\mu}$，分散共分散 $\boldsymbol{\Sigma}$ が未知であるために実際には計算することができない。しかし，$\boldsymbol{\mu}$ や $\boldsymbol{\Sigma}$ の値を仮に代入してやると，その値の元でのデータを得る確率をその値ごとに計算してやることができる。ここで発想を転換して，得られたデータを最も確率的に得やすい平均 $\boldsymbol{\mu}$ と分散共分散 $\boldsymbol{\Sigma}$ の値は何かということを考える。これが最尤法の基本的発想である。この値を求めるためには，$\boldsymbol{\mu}$ と $\boldsymbol{\Sigma}$ をシステマティックに変化させて，そのたびごとに式 (4) を用いて確率を計算し，その値が最大になる $\boldsymbol{\mu}$ と $\boldsymbol{\Sigma}$ を採用すればいいだろう。つまり，式 (4) を，$\boldsymbol{\mu}$ と $\boldsymbol{\Sigma}$ の関数と見なして，その最大値を求めるという発想である。このとき，式 (4) を $\boldsymbol{\mu}$ と $\boldsymbol{\Sigma}$ に関する尤度関数 (likelihood function) と呼び，$L(\boldsymbol{\mu},\boldsymbol{\Sigma})$ のように表現する。また，同じように個人ごとの確率を示す関数である式 (3) も，個人のデータ $\boldsymbol{x}_i$ が与えられたときの $\boldsymbol{\mu}$ と $\boldsymbol{\Sigma}$ の尤度関数だと考えられるので，これを $L_i(\boldsymbol{\mu},\boldsymbol{\Sigma})$ と表現する。

尤度関数はこのままの形だと，積の連なりであるため値が非常に小さくなってしまい扱いにくい。そこで，尤度関数の対数を取って計算をしやすくする。

$$\log L(\boldsymbol{\mu},\boldsymbol{\Sigma}) = \log \prod_{i=1}^{N} L_i(\boldsymbol{\mu},\boldsymbol{\Sigma}) = \sum_{i=1}^{N} \log L_i(\boldsymbol{\mu},\boldsymbol{\Sigma}) \tag{5}$$

式 (5) を最大化することは，式 (4) を最大化することと同じである。この式を最大化するとき，実際には，$\boldsymbol{\mu}$ と $\boldsymbol{\Sigma}$ のすべての値を組織的に調べるのではなく，最適化計算という方法を用いて，この値を効率的に探ることになる。今の例では，$\boldsymbol{\mu}$ と $\boldsymbol{\Sigma}$ を求めるのが目的であったが，たとえば共分散構造分析 (SEM) であれば，$\boldsymbol{\Sigma}$ の中身がモデルから imply されたパラメタの関数になるだけであり，本質的なロジックはまったく同じである。

2.2 欠損値がある場合の最尤推定

この多変量正規分布の最尤推定で，データに欠損があった場合はどうすればいいのだろうか。i 番目の人の尤度関数をもう 1 度考えてみよう。

$$L_i(\boldsymbol{\mu},\boldsymbol{\Sigma}) = f(\boldsymbol{x}_i|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \times \exp\left(-\tfrac{1}{2}(\boldsymbol{x}_i-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu})\right) \tag{6}$$

この人から得られたデータは $\boldsymbol{x}_i$ である。もし 3 つの変数があるモデルなら，$\boldsymbol{x}_i$ は 3×1 のベクトルであり，推定すべき $\boldsymbol{\mu}$ は 3×1 のベクトル，$\boldsymbol{\Sigma}$ は 3×3 の行列となる。たとえば，Table 1 の例で id $= 5$ の人は

$$\boldsymbol{x}_5 = \begin{pmatrix} 5 \\ 103 \\ 148 \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 \end{pmatrix} \tag{7}$$

尤度関数ではデータはもう与えられているが，パラメタベクトル $\boldsymbol{\mu}$ と $\boldsymbol{\Sigma}$ は未知だということを確認しておこう。ここで，適性検査の得点が欠損しており，実際は動機づけと IQ の 2 変数しか測定されなかった人を考えよう（たとえば id $= 1$ の人）。その場合，この人から 3 変数のデータが得られる尤度は当然算出することができない。しかし，欠損している変数を除いた 2 変数の値が得られる尤度はまだ計算できる。式 (6) において，下のように $\boldsymbol{x}_i$ を 2×1 のベクトル，推定すべき $\boldsymbol{\mu}$ を 2×1 のベクトル，$\boldsymbol{\Sigma}$ を 2×2 の行列にしてやればいいだけである。

$$\boldsymbol{x}_1 = \begin{pmatrix} 3 \\ 83 \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix} \tag{8}$$

この式を用いることで，この人のデータ 動機づけ $= 3$ と IQ $= 83$ だけが与えられたときの，$\boldsymbol{\mu}$ と $\boldsymbol{\Sigma}$ の尤度を求めることができる。重要なのは，このように個人ごとに（欠損によって）データのサイズが違っていても，個人ごとの尤度を求めることが可能であり，さらに式 (4) や (5) に基づいて全体の尤度を求めることもできるという点である。そして，欠損値があったとしても，この全体の尤度を最大化するものが最尤推定値であるということには変わりがない。つまり，最尤法は欠損値があったとしてもなかったとしても，同じ原理（個々人の対数尤度の和を最大化するように推定値を求める原理）で最尤推定値を求めることができるのである。最尤推定法は，個々人の尤度関数を個別に定義することで，欠損値の問題を非常に自然に解決している。欠損のある人のデータも，上に示したように欠損のない部分の情報を完全に利用しているので，完全情報最尤推定法（FIML）と呼ばれることがあるが，基本的には通常の最尤法と原理は何ら変わりがない[6]。

先述したように，最尤法（もしくは FIML）を用いると，MAR のもとでもバイアスのない推定値が得られることが明らかになっている。実際，Table 1 の例で平均値を最尤推定すると，動機づけ $= 3.6$，IQ $= 101.8$，そして適性検査 $= 110.9$ となり[7]，適性検査の平均点がより真の値 (111.7) に近づいているのが分かるだろう[8]。欠損値のない変数（たとえば IQ）に関しては，バイアスがなぜ生じないかは自明だろう。もしリストワイズ法を用いたならば，IQ が低い人はすべて削除されてしまう（IQ が低い人は適性検査が欠損しているため）。したがって，IQ の平均値の推定にはバイアスが生じる。しかし，FIML を用いれば，IQ しかない人のデータも，IQ に関する部分だけは尤度関数に組み込まれるため，IQ の平均値に関してはバイアスのない推定ができる。一方，欠損値のある適性検査に関する平均値推定で，どうしてバイアスがなくなるのかに関しては，少し直感的理解が難しいかもしれない。少なくとも，単純に適性検査が測定された人の標本平均を算出しただけではバイアスが残ってしまう（標本平均は 117.3）。IQ が低い人が欠損値になっているの

[6] すべてのデータに欠損値がなければ，尤度関数は十分統計量を用いて簡略化して表現できるため，推定は楽になる。この簡略化した尤度関数しか扱えないアルゴリズムを用いるのなら，欠損値には対処できない。そういった意味で，FIML と（かなり狭義の）最尤推定は少し違うと考えることもできるかもしれない。

[7] Mplus で推定

[8] ちなみに IQ を除いて動機づけと適性検査の得点だけを含めたモデルで適性検査の平均を推定すると平均の推定値は 117.2 となりバイアスが除去されていない。適切な変数をモデルに含めることが重要だということがわかるだろう。

だから，IQ と正の相関がある適性検査得点も，その得点が低い人に偏って欠損値が生じているからである。したがって，測定された適性検査の情報だけを用いても，MAR であれば推定値にはバイアスが生じてしまう。しかし，FIML は適性検査の得点だけでなく IQ や動機づけの情報も用いて，その同時関数を最大化する推定値を求めようとしている点が重要である。つまり，IQ や動機づけの平均の推定値によって，尤度を最大化する適性検査の平均の推定値も変わってくるのであり，単純に適性検査の標本平均値が平均の推定値にはならないのである。別の表現をすれば，最尤法では他の変数の情報を借りて (borrow the information)，欠損値のある変数のパラメタを推定することができる。そして，その結果として，MAR が満たされていれば欠損のある変数のパラメタも偏りなく推定することができる。少し分かりにくいかもしれないが，この点に関しては，Enders (2010) により具体的な数値例に基づいた詳しい議論がなされているので，より詳しく知りたい人はそちらを参照して欲しい。

ここでは多変量正規分布の $\boldsymbol{\mu}$ と $\boldsymbol{\Sigma}$ を推定することを考えたが，たとえば（多変量正規分布を仮定した） SEM であれば $\boldsymbol{\Sigma}$ の中身がパス図のパラメタで構造化された式になるだけであり，本質的な考え方は変わらない[9]。多変量正規分布でないモデルの場合であっても，尤度関数の形が変わるだけであり，やはり考え方は同じである。

2.3 ソフトウェア

FIML は，共分散構造分析のソフトウェアに実装されている (e.g., AMOS, LISREL, EQS, Mplus, and Mx)。したがって，たとえば回帰分析などで FIML を行いたい場合でも，これらのソフトウェアを使うとよいだろう。汎用ソフトウェアの SAS などでも，たとえば mixed model であるならばデフォルトで最尤推定が用いられている。

2.4 補助変数を用いたアプローチ (inclusive analysis strategy)

繰り返しになるが，MAR を仮定した分析では，欠損パタンと相関のあるような適切な変数を分析に入れることで，推定値の精度が高まる（バイアスや標準誤差が低下する）。しかし，そうした変数が分析の上では必要ない場合もある。たとえば，これまでの例だと IQ が MAR の仮定を満たすために重要な変数であったが，分析者は動機づけと適性検査の関係にのみ関心があり，IQ は分析モデルには含めたくないかもしれない。こうしたとき，IQ をモデル自体には影響を与えない形で補助変数 (auxiliary variables) として分析に含める必要がある。以下に，SEM のソフトウェアでモデルの推定を行うことを前提として，その方法を示そう (Enders, 2008)。

Figure 3 に，観測変数だけを扱った場合に，補助変数をどう組み入れるのかを図示した。基本的には，補助変数はモデルに含まれているすべての変数と相関を持つように組み入れる。内生変数

[9] FIML では個人ごとに持っている情報量の定義が違うため，N が定義できなくなる（100 変数のうち 1 つしか回答しなかった人とすべてを回答した人を同じように N=1 として扱うのは直感的にも変である）。そのため，適合度の基礎となる χ^2 統計量を算出するための公式

$$\chi^2 = (N-1)\log L(\hat{\theta}) \tag{9}$$

を使うことができない。そのため，FIML を用いる場合は，まず飽和モデル（適合度が完全であるはずのモデル）における尤度 L_0 をもとめ，そこからどれくらい調べようとしているモデルの尤度 L_1 が離れているかを考える。具体的には，χ^2 統計量は次のように算出でき，統計的検定や適合度の算出が可能になる。

$$\chi^2 = -2(\log L_1 - \log L_0) \tag{10}$$

これは，いわゆる尤度比検定（この場合飽和モデルをベースラインにした尤度比検定）と同じ考え方である。

には相関を指定できないので誤差と相関を持たせる。補助変数はすべての変数との相関が仮定されているので，補助変数に関するモデルはいわゆる飽和であり，モデルの χ^2 統計量や自由度には影響を与えない。逆にいうと，このモデルとオリジナルのモデルを比較して，χ^2 統計量や自由度に違いがあれば，組み込み方を間違えているということになる。補助変数を加えることで推定の精度は変化するので，パス係数や検定結果などには変化があるかもしれないが，当然これは問題ではない。また，これも当然であるが，補助変数はモデルの解釈には影響を与えない。

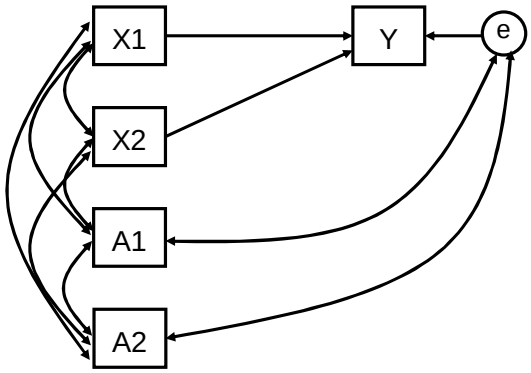

Figure 3: 補助変数 (auxiliary variables) の組み込み方（観測変数だけのモデルの場合)。$A1$ と $A2$ が補助変数である。Enders (2008) をもとに作成。

Figure 4 には，潜在変数が含まれているモデルを推定した場合の補助変数の組み込み方を示した。基本的なアイディアは同じであり，他の変数との相関（内生変数の場合は誤差相関）を仮定すればよい。ただし，潜在変数に対してではなく，あくまで観測変数に仮定する。

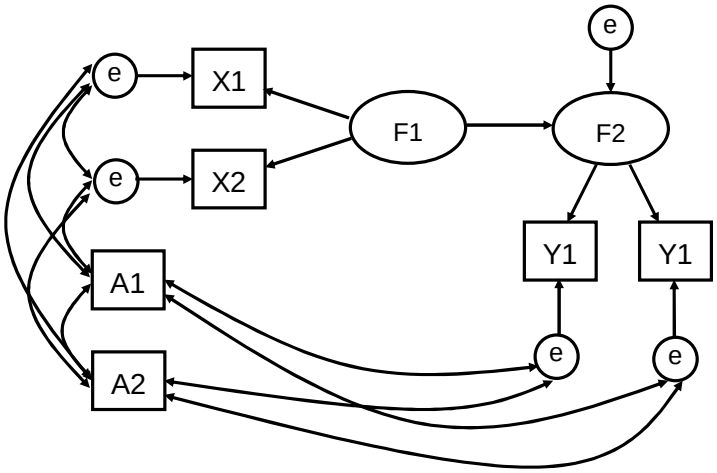

Figure 4: 補助変数の組み込み方（潜在変数を含むモデルの場合)。$A1$ と $A2$ が補助変数である。Enders (2008) をもとに作成。

こうした inclusive analysis strategy は推定の精度を上げるために有用なストラテジーであるが，FIML (SEM) の場合，大量の補助変数をモデルに投入するのはシンタックスを書く上で骨が折れるという問題点がある。Mplus を用いている場合，VARIABLE コマンドで auxiliary = と書き，

補助変数を入れたい変数を指定していけば自動的に分析に含めてくれる（詳しくはマニュアル参照のこと）。しかし，他のソフトウェアの場合，図に示したようなパス図をそのたびごとに指定しなくてはいけない。さらに，変数の数が増えると推定がうまくいかなくなる可能性もあり，注意しなくてはいけない。

なお，補助変数を入れたモデルでは CFI や TLI といった incremental fit index の値が不正確になる。CFI や TLI は，すべての変数間に相関が存在しないという独立モデル (independent model) と現在のモデルの適合度を比較して算出するが，補助変数を組み入れることで，独立モデルの適合度が変化してしまうからである [10]。補助変数を入れた分析を行う場合には，分析者が自分の手で独立モデルを構築し，その適合度をメモした上で，手計算で正しい CFI と TLI を求める必要がある。詳しくは Enders (2010) を参照して欲しい。

3　多重代入法

多重代入法 (multiple imputation method) とは，FIML と並んで，近年の欠損値データ解析のスタンダードになりつつある手法である。伝統的に，欠損値のあるデータ解析では種々の代入法 (imputation method) がリストワイズ・ペアワイズ法と並んで使用されてきた。たとえば平均値代入法や回帰代入法などである。平均値代入法とは，欠損値のある人に対して，その変数の欠損値のない人の平均値を欠損値の変わりに代入する方法である。回帰代入法とは，まず欠損値のある変数を残りの変数から予測する回帰モデルを推定し，欠損値のある人に対しては，この回帰モデルの予測値（推定された回帰モデルの独立変数にその人の測定値を代入した値）を欠損値に代入する方法である。これらの方法は，測定に伴う不確定性を考慮していないため，たとえば分散などを過小推定してしまうという問題点がある。実際，回帰代入をした場合を考えてみると，代入された値はすべて単一の回帰直線の上に乗ることになるため，測定値の不確定性が考慮されていないことは明らかだろう。

この問題に対処する 1 つの方法は，欠損値を回帰モデルで予測した後，その予測値にランダム誤差（回帰モデルの誤差分散に相当）を加えてその誤差が加わったものを欠損値の代わりに代入する方法である。これを stochastic regression imputation と呼ぶ。この stochastic regression imputation は欠損値への対処法としては比較的よい方法であり，MAR のもとでもバイアスのない推定値を与えてくれる。しかし，この方法の欠陥点は，代入したデータセットを唯一絶対のものとして分析の対象にすることである。その結果，欠損値があることによる推定の不確定性が考慮されていない。直感的に考えると，欠損値が多い方が推定が不安定になることが考えられるが，stochastic regression imputation では，欠損値がどの程度あろうと，代入された単一のデータをそのまま分析するので，もともと欠損値が多かったかどうかが統計的推測の過程に反映されない（そもそも代入されたデータをみてもどの程度欠損値があったかが分からない）。すなわち，stochastic regression imputation では，MAR の仮定のもとで推定値にはバイアスがないが，欠損値が多い場合に標準誤差を過小評価してしまう。

Rubin (1987) の多重代入法は，その名の通り，欠損値を代入したデータセットを複数作成し，その結果を統合することで欠損値データの統計的推測を行う方法である。stochastic regression imputation と似たステップを踏むが，データセットを複数作成することによって，欠損値による推定の不安定性を結果に反映させている点が大きく違う。

[10] CFI や TLI はこのように，統計ソフトウェアが自動的に仮定する独立モデルが間違っている場合には，補助変数のモデルに限らず，誤った値を出力する。1 つの代表的な例が潜在曲線モデルである (see Wu, West, & Taylor, 2009)。

Figure 5 にその概念図を示した。多重代入法は大きく 3 つのステップがある。最初は代入ステップ (imputation step) であり，欠損値に何らかの値が代入された擬似完全データセットを複数作成する。次は擬似完全データをもとに目的の分析 (regression, anova, sem, etc...) を行うステップである。同じ分析がすべての擬似完全データに対して実施される。擬似完全データは欠損値がないので，分析は straightforward なはずである。また，擬似完全データは欠損値の部分の値がデータセットごとに異なるので，このステップでは少し異なった推定値と標準誤差が擬似完全データセットごとに得られるはずである。最後が統合ステップ (posterior step) であり，これらの推定値と標準誤差を統合し，単一の推定値と標準誤差を算出する。第二ステップは通常の統計分析と同じなのでここでは省略し，以下では代入ステップと統合ステップについて解説する。

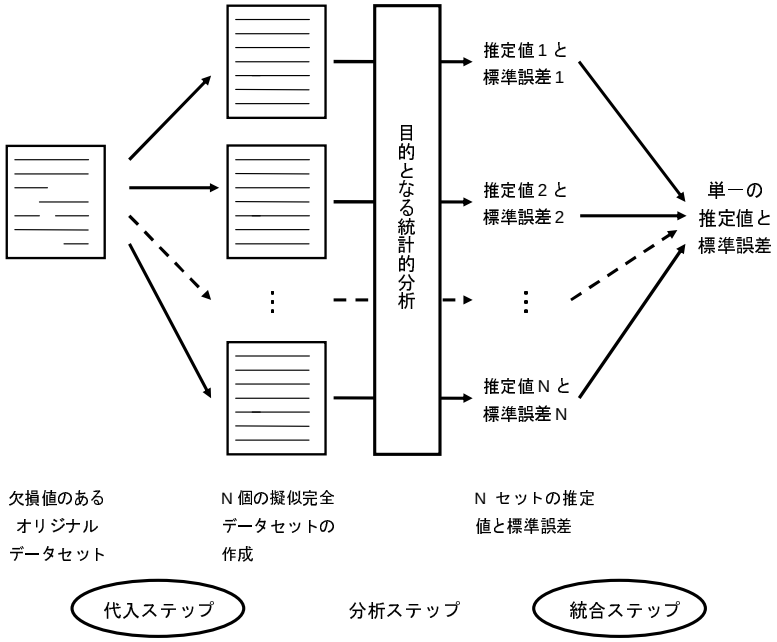

Figure 5: 多重代入法 (multiple imputation method) の概念図

3.1 代入ステップ (imputation step)

どのように欠損値を代入し，複数の擬似完全データを作成するかについては，いくつもの方法が提案されている。ここではそのなかでも，特にポピュラーであるデータ拡大法 (data augmentation method) に基づいた方法を説明する。SAS の proc MI や，多重代入法のソフトウェアである NORM [11] はこのアルゴリズムを用いている。一方，SPSS の multiple imputation module は，sequential regression approach (or chained equations approach) というものを用いているが，こちらを知りたい人は van Buuren (2007) を参照して欲しい。

データ拡大法による擬似完全データの作成は，基本的にベイズ統計学の考えて大きく依拠している。一言でいうと，欠損値の事後予測分布 (posterior predictive distribution) から乱数を発生させ，それを欠損値に代入したデータセットを複数作るというアイディアである。事後予測分布から

[11] Schafer (1997) を参照。http://www.stat.psu.edu/ jls/misoftwa.html

の乱数の発生にはマルコフ連鎖モンテカルロ法 (Markov chain monte carlo; MCMC) を用いる。具体的には，次のようなステップを踏む。

1. データの平均・分散共分散行列の初期値を定める
2. 得られた平均・分散共分散行列をもとに，欠損値のあるデータを他のすべての変数で予測する回帰式を推定する
3. 回帰式をもとに stochastic regression model と同じように予測値にシミュレートした誤差を加えて欠損値を代入する
4. 得られた完全データから平均・分散共分散行列を求める
5. 平均・分散共分散行列に乱数を加えた値をシミュレートし，それを新たな平均・分散共分散行列とする
6. 上の 2 - 5 を繰り返す

3. はベイズ統計学でいうと，事後予測分布からの乱数の発生であり，5 は事後分布からの乱数の発生である。ベイズ統計学の流儀で，よりフォーマルに表現すると，3. は

$$p(Y^t|\boldsymbol{\mu}_*^{t-1}, \boldsymbol{\sigma}_*^{t-1}, Y_{obs}) \tag{11}$$

から１セットの乱数を発生させることである。Y_{obs} はデータのなかで欠損していない部分，Y^t は欠損値がどういう値になるかに関する確率変数である。得られた乱数は，固定された値であることを強調して以下では Y_*^t と表現しよう。t はサイクル数であり，$\boldsymbol{\mu}_*^{t-1}, \boldsymbol{\sigma}_*^{t-1}$ はその前のサイクルで得られた平均と分散共分散行列パラメタの値である（これも固定された値なので * がついている)。最初のサイクルの場合，$\boldsymbol{\mu}_*^0$ と $\boldsymbol{\sigma}_*^0$ は 1. で設定した初期値である。2 回目以降のサイクルでは，5. で得られた平均と分散共分散行列の値のことである。

もう少し具体的にいうと，3. では，欠損値を他の変数から予測する回帰モデルを構成してやり，その回帰係数を $\boldsymbol{\mu}_*^{t-1}$,$\boldsymbol{\sigma}_*^{t-1}$ を用いて計算してやる（回帰係数は平均値ベクトルや分散共分散行列があったら求められることを思い返して欲しい)。回帰係数が定まった回帰分析モデルが得られたら欠損値を予測する事後予測分布を計算することが可能であるので，その事後予測分布を計算し（それが $p(Y^t|\boldsymbol{\mu}_*^{t-1}, \boldsymbol{\sigma}_*^{t-1}, Y_{obs})$)，そこから乱数を発生させて欠損値に代入する（この代入された値が Y_*^t である)。この段階で，擬似的な完全データセットができることをイメージして欲しい。回帰分析で事後予測分布をどのように導出するかはベイズ統計学の本をみれば書いてある。

いったん擬似的な完全データセットができたなら，その完全データセットから平均値ベクトルと分散共分散行列ベクトルの事後分布を推定することが可能になる。その事後分布から新たな平均値ベクトルと分散共分散行列の値をシミュレートするのが上の 4. と 5. のステップである。

よりフォーマルに表現すると，まず平均値の事後分布

$$p(\boldsymbol{\mu}^t|Y_*^t, \boldsymbol{\sigma}_*^{t-1}, Y_{obs}) \tag{12}$$

を求めて，そこから $\boldsymbol{\mu}$ の値を１セットシミュレートする（これが $\boldsymbol{\mu}_*^t$ である)。次のこのシミュレートした値を利用して，分散共分散行列の事後分布

$$p(\boldsymbol{\sigma}^t|Y_*^t, \boldsymbol{\mu}_*^t, Y_{obs}) \tag{13}$$

を求めて，$\boldsymbol{\sigma}$ の値を 1 セットシミュレートする（得られた値が $\boldsymbol{\sigma}_*^t$ である）。平均値と分散共分散行列の事後分布をどのように求めるかに関してはベイズ統計学のテキストに必ず書いてあるのでそちらを参照して欲しい [12]。

このようにして新しい $\boldsymbol{\mu}_*^t$ と $\boldsymbol{\sigma}_*^t$ が得られたら，式 (11)（つまりステップ 2-3）に戻って新たな Y をシミュレートすることになる（つまり別の値で埋めた新たな擬似データセットを作成することになる）。

この過程を繰り返すことで

$$\boldsymbol{\mu}_*^0, \boldsymbol{\sigma}_*^0, Y_*^1, \boldsymbol{\mu}_*^1, \boldsymbol{\sigma}_*^0, Y_*^2, \boldsymbol{\mu}_*^2, \boldsymbol{\sigma}_*^2, Y_*^3, \ldots \tag{14}$$

という，Y_* や $\boldsymbol{\mu}_*$,$\boldsymbol{\sigma}_*$ の連鎖が得られる。この連鎖を繰り返していくことで，たくさんの欠損値の代入値 Y_* が何セットも得られる。このうち，最初の方の Y_* は，初期値依存性が高いのでそのまま捨てる。これをバーンイン (burn-in) と呼ぶ。また，連続して得られた欠損値の代入値セット（たとえば Y_*^{200} と Y_*^{201}）は相互の相関が高いので，できるだけ独立の代入値セットを得るために，ある程度の間隔を空けて Y_* を拾っていく [13]。以上のような手続きで，擬似完全データセットを複数得ることができるのである。

このように sequential な乱数を次々と発生させることで，特定の分布にしたがう乱数を得る方法がマルコフ連鎖モンテカルロ法である。ベイズ統計学やマルコフ連鎖モンテカルロ法に馴染みがある人は，気づいたと思うが，この手続きは，マルコフ連鎖モンテカルロ法でよく用いられるギブスサンプラー (Gibbs sampler) に非常に似ている。実際，データ拡大法はギブスサンプラーの variant の 1 つだと捉えることができる [14]。データ拡大法は，ギブスサンプラーを用いて，欠損値の事後予測分布から乱数を発生させている手続きだと捉えることができる。

マルコフ連鎖モンテカルロ法（もしくはその一種であるギブスサンプラー）では，得られた乱数が本当に目的としている分布からの乱数であるか（目的の分布に収束しているか）をチェックする必要がある。そのために，得られた値の時系列プロットを描いてみたり，autocorrelation function plot というものを描いてみて収束を診断することが多い。また，初期値を変えてみて，同じような分布に収束するかを調べることもある。したがって，データ拡大法で多重代入法を行う場合には，こうした時系列プロットや収束を判定する指標などがいろいろと出力される。こういったことに関しては，MCMC のテキストなどに詳しく載っているのでそちらを参照して欲しい（たとえば豊田，2008 など）。

3.2 統合ステップ (posterior/integration step)

複数の擬似完全データセットが得られたら，それぞれのデータセットに関して，目的の分析（回帰分析，ANOVA, SEM, 階層線形モデル，etc...）を実施する。すると，データセットの分だけ出力が得られるだろう。欠損値の部分はデータセットごとに少しずつ違うので，その結果も違うはずである。そこで，この統合ステップでは，これらの出力を統合することになる。具体的には，パラメタの推定値と標準誤差をこれらの結果を統合して求めるのが目的である。

[12] 多変量正規分布の場合，多変量正規分布と逆ウィシャート分布にしたがう。

[13] この間隔が長いほど基本的にデータ間の独立性が強くなるが，長すぎるとシミュレートするのに時間がかかってしまうという問題もある。

[14] 歴史的には先にデータ拡大法があり，あとでそれがギブスサンプラーの一種だということが分かったらしい。

3.2.1 パラメタ推定値の統合

パラメタ推定値の統合は簡単である。得られた値を平均すればよい。$\hat{\theta}_t$ が t 番目のデータセットから得られたパラメタ推定値であり，擬似完全データセットが m 個あったとすると，

$$\bar{\theta} = \frac{1}{m}\sum_{t=1}^{m}\hat{\theta}_t \tag{15}$$

が統合されたパラメタ推定値である。なお，この統合の式は，パラメタ推定値が正規分布にしたがうことを仮定している。したがって，たとえば相関係数や分散の推定値など，標本分布が正規分布ではないパラメタの場合には，N が少ない場合にバイアスが生じる可能性もある。

3.2.2 標準誤差の統合

標準誤差を統合するときには，まず下の値を計算する。

$$V_W = \frac{1}{m}\sum_{t=1}^{m}SE_t^2 \tag{16}$$

SE_t は t 番目の擬似完全データセットにおける標準誤差の値である。したがってこれは，標準誤差の二乗（標本分布の分散）の平均値を算出したに過ぎない。これを within-imputation variance と呼ぶ。直感的にはこの値の平方根を算出すれば，標準誤差の統合された値が得られそうな気がするが，実際にはもう１つ次の値を計算する必要がある。

$$V_B = \frac{1}{m-1}\sum_{t=1}^{m}(\hat{\theta}_t - \bar{\theta}) \tag{17}$$

これは，標準誤差が擬似データセット間でどれくらいばらついているかを示している。したがって，between-imputation variance と呼ばれる。V_W と V_B が ANOVA における級内分散と級間分散に類似している点に注目して欲しい。

多重代入法における標準誤差は，この２つの値を組み合わせた形で算出される。具体的には，

$$V_T = V_W + V_B + \frac{V_B}{m} \tag{18}$$

を算出し，その平方根が標準誤差になる。

$$SE = \sqrt{V_T} \tag{19}$$

この式で大切なのは，統合された標準誤差は標準誤差の平均である V_W だけでなく，標準誤差の擬似データセット間のばらつき V_B も加わっていることである。これは何を意味しているのだろうか。あるデータに欠損値が大きいと，擬似完全データセット間にもばらつきが大きくなり，V_B も大きくなる。つまり，V_B は，欠損値に多さに応じたパラメタ推定の不確定性を意味していると考えられる。単一の擬似データセットしか使わない stochastic regression 法だと，欠損値の多寡を考慮できず，標準誤差が過小評価されると書いたが，多重代入法ではこの点がこの V_B によってクリアできているのである。なお，式 (18) に V_B/m の項があることから，擬似データセットの数が多いほど標準誤差が小さくなるということも分かる。

3.3 多重代入法におけるいくつかの留意事項

3.3.1 擬似完全データセットの数

いくつの擬似完全データセットが必要か，ということに関しては歴史的には 3-5 と言われていた (e.g., Rubin, 1987)。しかし，Graham, Olchowski, & Gilreath (2007) が示しているように，擬似完全データセットの数が少ないと検出力が低くなることがある。そうしたことなどを考えると，擬似完全データセットはより多いほうがいいと考えられる。Enders (2010) は 20 くらいを目安として挙げている。

3.3.2 交互作用に興味があるときの注意点

補助変数のように，擬似完全データセットを作成する際には，分析とは直接関係ないが MAR の仮定のために重要だと思われる変数は積極的に入れたほうがよい。多重代入法の便利なところは，代入ステップでこの点を抑えておけば，その後の分析では補助変数などをあえて分析に含める必要がない点である。完全情報最尤推定法における補助変数のアプローチでは，モデルを構成するたびに補助変数を手で組み入れなければならず，かなり煩雑であったが，多重代入法ではそういった煩雑さはない。

ここで 1 つ重要なことは，分析者が交互作用などに興味があるときには，交互作用を代入ステップでしっかりとモデルに含めておく点である。交互作用がないモデルで擬似完全データセットを作り，そのデータセットで交互作用を検定して結果を統合すると，検出力が低くなっている（有意な結果が出にくい）可能性がある。なぜなら，擬似完全データセットをつくるときに，基本的には回帰モデルを用いているため，変数間に交互作用があることが仮定されていないからである。

同じような問題として，階層的なデータ (nested data, hierarchical data) の問題がある。階層線形モデル (hierarchical linear model, HLM) などの分析を行う場合には，多重代入法もデータの階層性を考慮した方法を用いた方がよい。しかし，データの階層性を考慮したモデルで擬似完全データセットを作成するソフトウェアは非常に乏しい [15]。現実的には，階層性を無視した形で多重代入法を用いることになるだろうが，そのときにこのような問題点があることは意識しておく必要がある。

3.3.3 Rounding をすべきか

たとえばリッカート尺度で欠損値があった場合，imputed value が 2.33 といった（実際にはありえない）小数の値が得られたり，尺度の上限値や下限値を超えた値が得られることもある。こうした場合，rounding をしたり，尺度の上限値や下限値にコードし直すことが考えられる。しかし，一般的には，こうしたことはしない方が，よりバイアスの少ない推定値・標準誤差を得られることが分かっている。ただし，ダミー変数を扱っていたりする場合には，その測定値を何らかの形でどこかのカテゴリーに割り振らなければいけないこともあるだろう。こうした場合には，Allison (2002) が，便利なルールを提唱しているのでそちらを参照するとよい。

[15] Norm というソフトウェアの variant にあるみたいだが (http://www.stat.psu.edu/ jls/misoftwa.html)，使ったことはない。なお，あとで発見したことだが，Mplus は version 6 からベイズ推定のオプションを使って階層データの擬似データセット作成も可能になった。

3.3.4 尺度レベルで代入すべきか

通常の質問紙研究だと，複数の項目を平均してある尺度得点を用いることが多い。このような場合に多重代入法を行うとき，多重代入法は項目レベルで行うべきだろうか（つまりすべての項目をデータセットに含めて代入ステップを実施する），それとも尺度レベル（つまり尺度得点を算出したデータセットで代入ステップを実施する）で行うべきだろうか。尺度レベルで多重代入法を行った場合，たとえばある尺度を構成する項目が 10 個あったとすると，そのうち 1 個欠損しただけでその人は欠損値として扱われてしまう。したがって，尺度レベルの多重代入法だと，残りの 9 項目の情報がまったく生かされない。その結果，分析の検出力が低くなってしまう。よって，理想的には項目レベルで多重代入法は行ったほうがよい（尺度得点は擬似データセットを得てから算出する）。しかし，項目数が非常に多い場合には結果が収束しなかったり，そもそもデータ代入ステップの実施が不可能な場合もでてくる（基本的にデータ代入ステップは回帰分析で代入をするので，変数の数が回答者の数を超えてはならない）。

こうしたとき，Enders (2010) は duplicate-scale imputation と呼ぶ方法が有用であるとした。この方法では，尺度レベルで多重代入法を行う。しかし，たとえばある尺度 X に欠損値があった場合，2 通りの変数を作成して，その両方をデータ代入ステップに含める。1 つは（X_1 とする），通常の尺度得点の変数であり，欠損がある人はそのまま欠損値として扱う。この変数では，たとえば X を構成する 8 項目のうち 1 項目でも欠損があれば，そこは欠損になっているわけである。もう 1 つの変数（X_2 とする）は，やはり尺度得点の変数であるが，下位項目のいくつかに欠損があった場合，欠損のない項目で尺度得点を算出した変数である。たとえばある人が 8 項目のうち 2 項目ほど欠損していたとすると，その人の X_1 は欠損値であるが，X_2 は残り 6 項目の平均値が観測値として入っていることになる。X_1 と X_2 は非常に似ているが（欠損値のない人の場合は同一），欠損値がある人に関しては，少し違ったものになる。ここで X_1 と X_2 を代入ステップで同時に用いると，X_1 の欠損部分が代入された擬似完全データセットを得ることができるが，X_1 に代入された値は X_2 の情報を利用しているので（X_2 が一種の補助変数として働く），項目レベルの情報も含まれることになる。つまり，X_1 のみで多重代入法をするよりも，より精度の高い擬似データセットができるというわけである。擬似完全データセットが得られたなら，X_2 は補助変数として利用していただけなのでデータセットから削除してやり，X_1 を用いて分析をすればよい。

duplication-scale imputation は非常に有用だが，X_1 と X_2 の相関が高くなりがちなので，データ代入ステップがうまく収束しないことも多い。その場合には，事前分布を変えるなど，何らかの工夫をする必要が出てくる。また，Little et al. (2008) [16] は，項目レベルで代入を行うが，すべてを項目レベルで行うのではなく，データをいくつかの項目ブロックに分割し，ブロック内では項目レベルで，ブロック間では尺度レベルでのデータを用いて代入を行うという three-step approach というのを提唱している。この方法はやや煩雑ではあるが，duplicate-scale method にあるような収束の問題を回避できるというメリットがある。

3.4 ソフトウェア

多重代入法は一見するとかなり煩雑そうにみえるが，SAS や SPSS のパッケージでは，代入ステップと統合ステップを自動的に行ってくれるコマンドがついているので，実際には簡単に実行できる。上に書いたように SPSS ではデータ拡大法ではなく，sequential regression model を代入ステップに用いている。また，擬似完全データセットを手軽に出力してくれるソフトとして Schafer

[16] これは http://www.crmda.ku.edu/pdf/11._Imputation_with_Large_Data_Sets.pdf よりダウンロード可能。

(1997) の Norm がある [17]。SEM や HLM のソフトウェアも，ほぼすべてにおいて多重代入法のオプションがついている。ただし，統合ステップだけであり，擬似完全データは基本的に作成してくれないので，代入ステップは SAS や Norm などを使って実行しておく必要がある[18]。

4 終わりに

以上で，完全情報最尤推定法と多重代入法を紹介してきた。近年ではこれらの方法が欠損値データ解析のスタンダードになりつつあり，それは今後もしばらくは変わらないであろう。この両者は MAR の仮定のもとでは非常に似た結果になるはずなので，どちらを選ぶかは分析ソフトウェアの都合などで決めればよいと思われる。たとえば，SEM であるならば FIML がデフォルトでついていることが多いので，補助変数の扱いが面倒でなければ完全情報最尤推定法の方が便利だろう。マルチレベル分析ならば，FIML が実装されているソフトが Mplus しかないので，多重代入法が便利かもしれない（ただし前に書いたように，階層モデルのときには通常の多重代入法だと推定にバイアスが生じる恐れがある）。

5 参考文献

Allison, P. D. (2002). *Missing data.*, Newbury Park, CA: Sage.

Enders, C. K. (2008). A note on the use of missing auxiliary variables in FIML-based structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *15*, 434-448.

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

Glynn, R. J., Laird, N. M., & Rubin, D. B. (1986). Selection modeling versus mixture modeling with nonignorable nonresponse. In H. Weiner (Ed.), *Drawing inferences from self-selected samples* (pp. 116-142). Berlin Springer-Verlag.

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.

Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement*, *5*, 475-492.

Little, R. J. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Associatin*, *83*, 1198-1202.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data.* New York: Wiley.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* New York: Wiley.

Rubin, D.B. (1996). Multiple imputation after 18+ years (with discussion). *Journal of the American Statistical Association*, *91*, 473-489.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data.* Chapman & Hall, London.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*, 147-177.

[17] http://www.stat.psu.edu/ jls/misoftwa.html

[18] ただし Mplus は version 6 から，擬似データセットの作成もできるようになった。

豊田秀樹（編） (2008). マルコフ連鎖モンテカルロ法 朝倉書店

van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, *16*, 219-242.

Wu, W., West, S. G., & Taylor, A. B. (2009). Evaluating model fit for growth curve models: Integration of fit indices from SEM and MLM frameworks. *Psychological Methods*, *14*, 183-201.