応用統計学　Vol. 34, No. 1 (2005), 1–20

総 合 報 告

連鎖方程式による多重代入法

統計数理研究所　野　間　久　史

要　旨　一般的な調査・実験研究において，欠測はほとんど避けられない問題であり，統計解析において，適切な処理を行わなくては，バイアス・推定精度の低下が起こり得る．ほとんどの研究において，欠測は複数の変数にまたがって，個人ごとに異なるパターンで起こることが一般的であるが，このような条件下で，汎用的な統計ソフトウェアで実行することができる不完全データの解析手法は，現状ではわずかしかない．連鎖方程式による多重代入法（multiple imputation by chained equation; MICE）は，このような条件下で有効な解析を行うために開発された方法であり，その実践的な有用性から，近年，多くの統計ソフトウェアに実装され，さまざまな研究領域において普及しつつある．本稿では，非統計家を含めた，データ解析に携わる実務家・研究者を対象として，邦文による MICE についての実践的な解説を行う．また，Clark and Altman (2003, *J. Clin. Epidemiol.* 56, 28-37) による卵巣がんの予後因子研究を事例として，R のパッケージ `mice` を用いた解析方法について紹介する．

1. はじめに

不完全データの統計解析において，欠測データに適当な補完値を代入（impute）して解析を行うことは，最も単純かつ直感的なアプローチであり，古くから多くの理論・応用に関する研究が行われてきた．中でも，多重代入法（multiple imputation; Little and Rubin, 2001; Rubin, 1987）は，近年の計算機性能の著しい向上と統計ソフトウェアの普及によって，実践的な方法として大きく発展した方法であり，丸尾，五所 (2017) によって紹介されている米国学術評議会による学術報告書（National Research Council, 2010）においても推奨されている方法のひとつとなっている．

多重代入法とは，その名の通り，単一の補完値ではなく，複数の補完値を利用した解析方法である．欠測値の補完に基づく解析方法は，原則として，「実際には観測することができなかった，欠測してしまったデータ」を，適当な予測方法によって予測し，代替的な補完値としてこれを埋めた擬似的な完全データの解析を行うという原理に基づく方法であるが，当然ながら，その妥当性は，欠測データの予測方法の正確性に依存する．しかしながら，完全に正確な予測方法を構築することは原理的に不可能であり，生成された補完値には，必ず予測誤差に基づく不確実性が伴う．結果として，最終的な推測の結果にも，その不確実性が加わることになる．多重代入法は，この予測誤差を反映した複数の補完値を生成し，この不確実性を適切に考慮した推測を行うための方法である．

一方，多重代入法の補完値の生成アルゴリズムの多くは，後述の通り，データセットの中で，1つの変数のみに欠測が起こったものという仮定のもと，観測データからそれを予測するというも

のとなっている．しかしながら，人を対象とした多くの調査・実験研究において，特定の 1 つの変数に限定して欠測が生じるという都合のよい欠測パターンが起こることはほとんどなく，実際には，複数の変数において，個人ごとに異なるパターンで欠測が生じる場合がほとんどである．連鎖方程式に基づく多重代入法（multiple imputation by chained equation; MICE）は，このように，対象となる解析データセットにおいて，複数の変数にまたがって非単調な欠測が生じたときに，すべての利用可能なデータを用いて多重代入法を実行できる方法として，近年，その計算上の有用性からも大きな関心を集めている．この数年ほどで，SAS, Stata, R などの標準的な統計ソフトウェアにおいても，計算パッケージが実装・整備されており（Berglund and Heeringa, 2014; Royston and White, 2011; van Buuren and Groothuis-Oudshoorn, 2011），実践においても多くの研究で採用されている（例えば，van Buuren and Groothuis-Oudshoorn (2011) のレビューなどをご参照いただきたい）．

本稿の目的は，非統計家を含めた，データ解析に携わる実務家・研究者を対象として，邦文による MICE についての実践的な解説を行うことである．2 節で，多重代入法の原理について簡潔な導入を行い，3 節で，さまざまな型の変数（連続変数，カテゴリカル変数，非対称な分布を持つ変数など）ごとの補完値の生成方法について，そして，4 節で，MICE の理論とアルゴリズムについて述べる．5 節で，補完値の生成モデルの構築方法，6 節で，補完値を代入した後のデータセットの解析におけるモデルの構築・評価方法について解説する．7 節において，具体的な応用事例として，R のライブラリ `mice` を用いた，Clark et al. (2001) による卵巣がんの予後因子研究（Clark et al., 2001）の解析例を紹介する．

2. 多重代入法の原理

多重代入法（Little and Rubin, 2001; Rubin, 1987）は，MAR の欠測メカニズムのもとで妥当性が担保される，最も代表的な不完全データの解析方法である．この方法では，不完全データの欠測値に，適当な方法によって作成した複数の補完値を代入した擬似的な完全データの組を生成し，そのそれぞれを個別に解析した結果を統合することによって推測を行う．古典的な単一の補完値を用いた解析でも，一定の条件下で，偏りのない推定量を得ることは可能であるが，一般的に，推定量の分散の評価が困難であることが問題となる．多重代入法では，複数の補完値に基づく解析結果を統合することにより，この補完値の不確実性を含めた統計的な誤差の評価を行う明快な枠組みを与えることができる．以降では，補完値の組の数を m 回とし，関心のあるパラメータを $\boldsymbol{\theta}$ と表記する．

多重代入法は，以下の 3 段階のステップによって，推定量を構成するアルゴリズムである．

Step 1.（補完値の生成）
欠測値に代入する補完値のデータセットを，m 組，適当な方法によって独立に生成する（具体的な補完値の生成方法の詳細は，3 節に示す）．

Step 2.（m 組の擬似的な完全データの解析）
Step 1 によって生成した m 組の補完値のデータセットを欠測データに代入し，m 組の擬似的な完全データを作る．この m 組のデータセットをそれぞれ独立に解析することによって，パラメータ $\boldsymbol{\theta}$ の推定値 $\hat{\boldsymbol{\theta}}_1, ..., \hat{\boldsymbol{\theta}}_m$ とその共分散行列の推定値 $\hat{\mathbf{V}}_1, ..., \hat{\mathbf{V}}_m$ を得る．

Step 3.（m 組の解析結果の統合）

Step 1, 2 によって得られた，m 組の推定値を，以下の公式によって統合する．

$$\hat{\boldsymbol{\theta}}_{\mathrm{IM}} = \frac{1}{m}\sum_{j=1}^{m}\hat{\boldsymbol{\theta}}_j \tag{1}$$

同様に，共分散行列の推定量も，$\hat{\boldsymbol{\theta}}_j$ の共分散行列の推定値 $\hat{\mathbf{V}}_j$ $(j = 1, 2, ..., m)$ をもとにして，

$$\hat{\mathbf{V}}(\hat{\boldsymbol{\theta}}_{\mathrm{IM}}) = \mathbf{W}_{\mathrm{IM}} + (1 + m^{-1})\mathbf{B}_{\mathrm{IM}} \tag{2}$$

$$\mathbf{W}_{\mathrm{IM}} = \frac{1}{m}\sum_{j=1}^{m}\hat{\mathbf{V}}_j, \mathbf{B}_{\mathrm{IM}} = \frac{1}{m-1}\sum_{j=1}^{m}(\hat{\boldsymbol{\theta}}_j - \hat{\boldsymbol{\theta}}_{\mathrm{IM}})(\hat{\boldsymbol{\theta}}_j - \hat{\boldsymbol{\theta}}_{\mathrm{IM}})^T \tag{3}$$

として得られる．$\mathbf{W}_{\mathrm{IM}}, \mathbf{B}_{\mathrm{IM}}$ は，それぞれ補完値内・補完値間での共分散行列（within-/between-imputation covariance matrices）である．擬似 Wald 流の検定・信頼区間の構成においては，適当な近似自由度を持つ t 分布を参照分布に用いるのが一般的である．自由度の設定については，Barnard and Rubin (1999)，Rubin (1987) の方法が一般的である．上付きの "T" は，行列の転置記号を表す．

上記の枠組みは，当初，ベイズ流の枠組みでの事後推測を近似するものとして提案された (Rubin, 1987)．実際，Step 1 において，補完値のデータセットが，$\boldsymbol{\theta}$ について適当な無情報事前分布を仮定したもとで，観測データを所与としたもとでの事後予測分布からのサンプルとして生成されたものとすると，Step 2 で得られる $\boldsymbol{\theta}$ の推定値 $\hat{\boldsymbol{\theta}}_1, ..., \hat{\boldsymbol{\theta}}_m$ を求めるプロセスは，近似的な Gibbs サンプリング（Gelfand and Smith, 1990）と見なすことができ，これらは $\boldsymbol{\theta}$ の事後分布からのサンプルと見なすことができる．これにより，$\hat{\boldsymbol{\theta}}_{\mathrm{IM}}$ は事後平均，$\hat{\mathbf{V}}(\hat{\boldsymbol{\theta}}_{\mathrm{IM}})$ は事後共分散行列のモンテカルロ推定量と解釈することができる．したがって，直感的には，$\hat{\boldsymbol{\theta}}_{\mathrm{IM}}$ は，$\boldsymbol{\theta}$ に無情報事前分布を仮定したもとで，観測データ尤度に基づく最尤推定量を近似した漸近有効な推定量と解釈することができる．より厳密には，Robins and Wang (2000)，Wang and Robins (1998) によって，詳細な漸近的評価が行われており，より一般的な正則条件のもとで，一致性を持つ推定量となることが示されている．またモデル誤特定のもとでのロバスト分散の推定量も与えられている．

繰り返し回数 m は，従来は，欠測の占める割合がそれほど大きくない場合には 5〜10 程度で十分であるとされてきたが，多重代入法は，上記の通り，原理的にはモンテカルロ法による近似推測法であり，十分な数の繰り返しを行わなくては，モンテカルロ誤差を制御することができない．Carpenter and Kenward (2013)， Royston and White (2011) によると，正確な結果を得るためには，概ね 100〜1000 回の繰り返しを必要とすると述べられている．現在の計算機環境では，繰り返し回数 m を大きくしても，それほど負担にならないことも多いため，十分な回数での繰り返しを行うことが望ましい．

3. 補完値の生成方法

ここでは，多重代入法による代表的な補完値の生成方法について解説する．1 節で述べた通り，多くの代表的な補完値の生成方法は，欠測を起こした変数が 1 つに限定されている状況をあらかじめ想定している．ここでは，データセットの中で，部分的に欠測を含む変数を z と表記するこ

ととし，これをすべての対象者において完全なデータが観測されている共変量 $\mathbf{x} = (x_1, \cdots, x_q)^T$ の情報から予測して，補完値を生成するものとする．n_{obs} を z において欠測が起こらなかった対象者の数とし，一般性を失うことなく，x_1 は，回帰モデルによる予測を行う際の切片項に対応する，値が 1 の変数であるとする．

3.1. 連続変数

3.1.1. 線形回帰モデル

前節において，多重代入法は，ベイズ流の理論的枠組みのもとで正当化されると述べたが，その理論的妥当性を保持するためには，対象となる事後予測分布からのサンプルを近似した補完値を用いるのが望ましい．そのための最も率直なアプローチとしては，以下の「回帰モデルによる補完値の生成方法」がある．いま，z が連続変数であり，$\mathbf{x}$ を説明変数とした，以下の線形回帰モデルによって予測モデルを構築することを考える．

$$z|\mathbf{x}, \boldsymbol{\beta} \sim N(\boldsymbol{\beta}^T\mathbf{x}, \sigma^2) \tag{4}$$

いま，$\hat{\boldsymbol{\beta}}, \hat{\sigma}^2$ を，最小二乗法による $(z, \mathbf{x})$ の欠測データを除いた集団（完全ケース）におけるパラメータの推定値とし，$\hat{\mathbf{V}}$ を $\hat{\boldsymbol{\beta}}$ の共分散行列の推定値とする．このとき，g を $n_{obs} - q$ を自由度とするカイ二乗分布からの乱数，u_1 を q 次の多変量標準正規分布からの乱数とすると，$\hat{\boldsymbol{\beta}}, \hat{\sigma}^2$ の標本分布からのサンプルは次式によって得ることができる．

$$\sigma^* = \hat{\sigma}\sqrt{\frac{n_{obs} - q}{g}}, \boldsymbol{\beta}^* = \hat{\boldsymbol{\beta}} + \frac{\sigma^*}{\hat{\sigma}}u_1\hat{\mathbf{V}}^{\frac{1}{2}} \tag{5}$$

これは，無情報事前分布を仮定したもとでの $(\boldsymbol{\beta}, \sigma^2)$ の同時事後分布からのサンプルと見なすことができる (Rubin, 1987)．したがって，個々の欠測データ z_i の補完値 z_i^* は，標準正規分布からの乱数 $u_{2i} \sim N(0, 1)$ を用いて，

$$z_i^* = \boldsymbol{\beta}^{*T}\mathbf{x}_i + u_{2i}\sigma^* \tag{6}$$

として生成することができる．

z が，正規性を仮定しにくいような非対称な分布をとる場合には，あらかじめ，Box-Cox 変換（Box-Cox transformation）や移動対数変換（shifted-log transformation）などを施してもよい．あるいは，分布の尖度まで考慮した変換として，Johnson の S_U 分布族（Johnson, 1949）などの柔軟な確率分布モデルを用いた変換を用いてもよい（White et al., 2011）．当然ながら，これらの変換を用いた場合，生成された補完値の組は，逆変換によってもとのスケールに戻して用いる必要がある．

3.1.2. 予測平均マッチング

z が連続変数であるもとでの補完値の生成方法として，もうひとつの代表的な方法に，予測平均マッチング（predictive mean matching）がある．線形回帰モデルによる補完値の生成方法は，正規性・線形性の仮定が妥当なもとでは性質のよい補完値を与えるが，実際にはこれらの仮定が誤っているようなとき（z と $\mathbf{x}$ の間に非線形な関数関係があったり，誤差分布に正規分布の仮定が成り立たない場合），不適切な補完値を与えてしまうことがある．また，サンプルサイズが小さいときには，$\hat{\boldsymbol{\beta}}, \hat{\sigma}^2$ の推定値の精度が低く，真の構造から大きく乖離したモデルから補完値が生

成される可能性もある．

予測平均マッチングは，ad hoc な方法ではあるが，3.1.1 節の要領で得られた回帰式をもとにして，実際に観測された z の観測データの中から「相応しい値」をランダムに選びとるという方法（いわゆる，*hot-deck* な補完値の生成方法）である．具体的には，まず，3.1.1 節の要領によって生成された $\boldsymbol{\beta}^*$ を用いて，欠測データ z_i の予測値の候補を $\boldsymbol{\beta}^{*T}\mathbf{x}_i$ と設定する．このとき，この予測値の候補からの z の観測値が得られている n_{obs} 人の回帰式に基づく点予測値の距離 $\left|\hat{\boldsymbol{\beta}}^T\mathbf{x}_h - \boldsymbol{\beta}^{*T}\mathbf{x}_i\right|$ $(h = 1, 2, \cdots, n_{obs})$ を計算し，この中で，距離の近い K 人を選択する（K は解析者が定める正の整数値）．なお，一般性を失うことなく，$\mathbf{x}_i$ の添え字集合のうち，はじめの n_{obs} 人を z_i が観測された対象者としている．予測平均マッチングは，この K 人の中からランダムに 1 名の対象者を選び，これを z_i の補完値とするという単純な方法である．

予測平均マッチングは，このように ad hoc な方法ではあるが，シミュレーション実験などによる経験的な評価では，相対的に優れた性能を示すことが知られている（例えば，Marshall, Altman and Holder, 2010）．K の設定については，5-10 程度をデフォルトに採用しているソフトウェアが多いが，Morris, White and Royston (2014)，White et al. (2011) のシミュレーション実験の報告によると，$K = 3$ のもとで良い性能が認められたとされている．

3.2. 2 値変数

z が 2 値変数である場合に最もよく用いられるのが，ロジスティック回帰モデル

$$\text{logit}\,\Pr(z = 1|\mathbf{x};\boldsymbol{\beta}) = \boldsymbol{\beta}^T\mathbf{x} \tag{7}$$

による補完値の生成方法である．原理は，3.1.1 節の線形回帰モデルと同様である．いま，$(\mathbf{x}, z)$ の完全データからの $\boldsymbol{\beta}$ の推定値を $\hat{\boldsymbol{\beta}}$ とし，その共分散行列の推定値を $\hat{\mathbf{V}}$ とする．このとき，$\boldsymbol{\beta}$ の事後分布からの近似的なサンプル $\boldsymbol{\beta}^*$ は，多変量正規分布 $\text{MVN}(\hat{\boldsymbol{\beta}}, \hat{\mathbf{V}})$ からのサンプリングによって得ることができる (Rubin, 1987)．この $\boldsymbol{\beta}^*$ を用いて，z_i の補完値は $\text{Bernoulli}\,(p_i^*)$, $p_i^* = [1 + \exp(-\boldsymbol{\beta}^*\mathbf{x}_i)]^{-1}$ から生成すればよい．

3.3. 順序を持たないカテゴリカル変数

z が順序を持たないカテゴリカル変数（L 水準, $L \geq 3$）である場合には，多項ロジスティック回帰モデルを用いることができる．L 水準のうち，特定の参照水準（第 1 水準）とそれぞれの水準を比較するロジスティック回帰モデルを，以下のように仮定する．

$$\Pr(z = l|\mathbf{x};\boldsymbol{\beta}) = \left[\sum_{l'=1}^{L}\exp(\boldsymbol{\beta}_{l'}^T\mathbf{x})\right]^{-1}\exp(\boldsymbol{\beta}_l^T\mathbf{x}) \tag{8}$$

ここで，$\boldsymbol{\beta}_l$ は q 次のベクトルであり，$\boldsymbol{\beta}_1 = \mathbf{0}$ である．モデル中のパラメータ $\boldsymbol{\beta} = (\boldsymbol{\beta}_2^T, ..., \boldsymbol{\beta}_L^T)^T$ は，$k(L-1)$ 次のベクトルである．3.2 節と同様，$\boldsymbol{\beta}$ の最尤推定値を $\hat{\boldsymbol{\beta}}$，その共分散行列の推定値を $\hat{\mathbf{V}}$ とすると，事後分布からの近似的なサンプル $\boldsymbol{\beta}^*$ は，$\text{MVN}(\hat{\boldsymbol{\beta}}, \hat{\mathbf{V}})$ からのサンプリングによって得ることができる．この $\boldsymbol{\beta}^*$ を用いて，z_i の補完値は $p_{il}^* = \Pr(z_i = l|\mathbf{x}_i;\boldsymbol{\beta}^*)$ をパラメータに持つ多項分布から生成すればよい．

この多項ロジスティック回帰モデルを用いた補完値の生成方法は，z が連続変数の場合にも，形式的に順序を持たないカテゴリカル変数と見なして適用することができる（*hot-deck* な代入法の

一種）．同様に，次節の比例オッズモデルを用いることもできる．ソフトウェアによっては，z のカテゴリ数に上限が設けられていることがあるが，適当な数に z の値を丸めれば，それほど深刻な問題にはならない．

3.4. 順序を持つカテゴリカル変数

z が順序を持つカテゴリカル変数（L 水準, $L \geq 3$）にも，前項の多項ロジスティック回帰モデルを用いることができる．また，以下に述べる比例オッズモデルを用いることもできる．比例オッズモデルは，順序属性を持つカテゴリ間を比較する 2 項確率モデルに，以下の制約を置いた，ロジスティック回帰モデルを拡張したモデルである．

$$\operatorname{logit} \Pr(z \leq l|\mathbf{x};\boldsymbol{\beta},\boldsymbol{\zeta}) = \zeta_l - \boldsymbol{\beta}^T\mathbf{x} \tag{9}$$

多項ロジッスティック回帰モデルとは異なり，$\mathbf{x}$ についての線形予測子 $\boldsymbol{\beta}^T\mathbf{x}$ はすべての累積カテゴリの対比において共通であると仮定していることが，比例オッズモデルの特徴である．累積カテゴリの対比間でのベースライン確率の差を表すパラメータ $\boldsymbol{\zeta} = (\zeta_1, \cdots, \zeta_{L-1})^T$ の最尤推定値 $\hat{\boldsymbol{\zeta}}$ は，$\boldsymbol{\beta}$ と同時に尤度関数を最大化することによって求めることができる．前節までと同様，$(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\zeta}})$ の推定された漸近分布からのサンプリングを行うことにより，近似的な事後予測分布からのサンプル $(\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*)$ を得ることができる．したがって，z_i の補完値は，

$$p^*_{il} = \Pr(z_i \leq l|\mathbf{x}_i;\boldsymbol{\beta}^*,\boldsymbol{\zeta}^*) - \Pr(z_i \leq l-1|\mathbf{x}_i;\boldsymbol{\beta}^*,\boldsymbol{\zeta}^*) \tag{10}$$

を確率に持つ多項分布から生成すればよい．

整数値をとる変数については，過分散構造を持つポアソン回帰モデルを用いた代入法を用いることもできる．

他にも，マルコフ連鎖モンテカルロ法を用いることによって，事後予測分布 $p(z|\mathbf{x},\boldsymbol{\theta})$ からの乱数生成を直接行う方法や（この場合，事後分布 $p(\boldsymbol{\theta}|\mathbf{x},z)$ からサンプリングを行うことができるため，それによって，直接，$\boldsymbol{\theta}$ についてのモンテカルロ推測を行ってもよい．いわゆる Tanner and Wong (1987) のデータ拡大アルゴリズムである），ベイズ流ブートストラップ法を用いたアプローチなど，いくつかのアプローチが提案されている．詳細について関心のある読者は，SAS Institute Inc. (2015)，Schafer (1997, 1999) をご参照いただきたい．

4. 連鎖方程式に基づく補完値の生成

4.1. 事例：卵巣がんの予後因子研究 (Clark et al., 2001)

多くの統計ソフトウェアに付属している多重代入法のパッケージには，前節に示したような補完値の生成モジュールが実装されている．しかし，これらの方法の多くは，先述の通り，「単一の変数に欠測があり，それ以外の変数には欠測がひとつもないこと」を原則としたものとなっている．繰り返しになるが，ほとんどの調査・実験研究において，欠測が 1 つの変数にのみ都合よく生じるという状況はなく，実際には，複数の変数にまたがって，個人ごとに異なるパターンで欠測は生じる．このような場合，当然ながら，3 節に示した補完値の生成方法を単純に適用することはできず，多重代入法による推測を行うことはできない．

表 1. **Clark et al. (2001)** の卵巣がんデータの欠測データ（n = 1189）.

変数名	変数	変数の型	カテゴリ数	欠測（%）
age	Age (years)	連続変数	—	0.0
figo	FIGO stage	カテゴリカル変数	4	1.8
grade	Grade of tumor	カテゴリカル変数	3	11.6
histol	Histology	カテゴリカル変数	7	0.0
ascites	Presence/absence of ascites	カテゴリカル変数	2	5.4
ps	Performance status	カテゴリカル変数	4	42.7
resdis	Residual disease	カテゴリカル変数	3	6.8
ca125	CA125 (a cancer antigen)	連続変数	—	36.7
alp	Alkaline phosphatase	連続変数	—	33.1
alb	Albumin	連続変数	—	33.0

具体例として，Clark et al. (2001) による卵巣がんの予後因子研究のデータセットを用いて，問題を整理することとしよう．この研究は，1984 年 1 月 1 日から 1999 年 12 月 31 日の間に，The Western General Hospital (Edinburgh, Scotland) において卵巣がんと診断された 1,189 名の患者を対象とした後ろ向きコホート研究であり，卵巣がんの生存予後に関連する要因の評価と，予後モデルを構築することを目的とした解析が行われている．表 1 に，Clark and Altman (2003) によって検討された 10 種類の変数の集計結果を示している．ca125 と alp のデータは，歪んだ分布をとっていたため，あらかじめ対数変換を施した上で解析に用いられている．age と histol のデータには欠測がなく，すべての対象者にデータが観測されていた．一方，ps, ca125, alp, alb は，それぞれ 43.0%, 37.0%, 33.3%, 33.0% の対象者において欠測していた．延べ 11,890 の変数のデータのうち，2,045 (17.2%) が欠測しており，831 人（69.6%）の患者は，少なくとも 1 つの変数において欠測が認められていた．237 人（19.8%）の患者において 4 つ以上の変数に欠測があったが，7 つ以上の変数で欠測が認められたのは 4 名のみ（0.4%）であった．欠測データのうち，1739 (85.0%) のものは，alb, alp, ca125, ps の欠測によるものであった．

それぞれの変数の欠測メカニズムについて，ps, ascites, resdis, alp について，カテゴリごと（欠測を起こした対象者については，別に 1 つのカテゴリを設けている）の Kaplan-Meier 曲線を図 1 に示している（Clark and Altman (2003) の Fig. 1 を一部改変）．図中の P 値は，欠測を起こした対象者とそうでない対象者での生存関数の差異についてのログランク検定の P 値である．ps については，少なくとも有意差は出ておらず，これらの対象者間で明確な差異はない．しかし，ascites, resdis については，欠測を起こした患者サブグループのほうが有意に生存予後が悪く，一定の傾向があるように考えられる．これらを無視した解析を行ってしまうと，コホート全体の生存予後は過大評価されてしまうことが示唆される．一方，alp については，その逆のことが示唆される．当然ながら，これらの評価は，個々の変数ごとに周辺的に行われたものであるが，少なくとも，欠測メカニズムを考慮した適当な解析を行う必要があるといえるだろう．

4.2. 連鎖方程式を用いた補完値の生成

先述の通り，3 節に示したような補完値の生成方法では，前節の事例のように複数の変数に個人ごとに異なるパターンで欠測が起こったデータセットの解析を行うことは難しい．MICE は，このような条件下でも，多重代入法による解析を可能とするための方法として開発された．

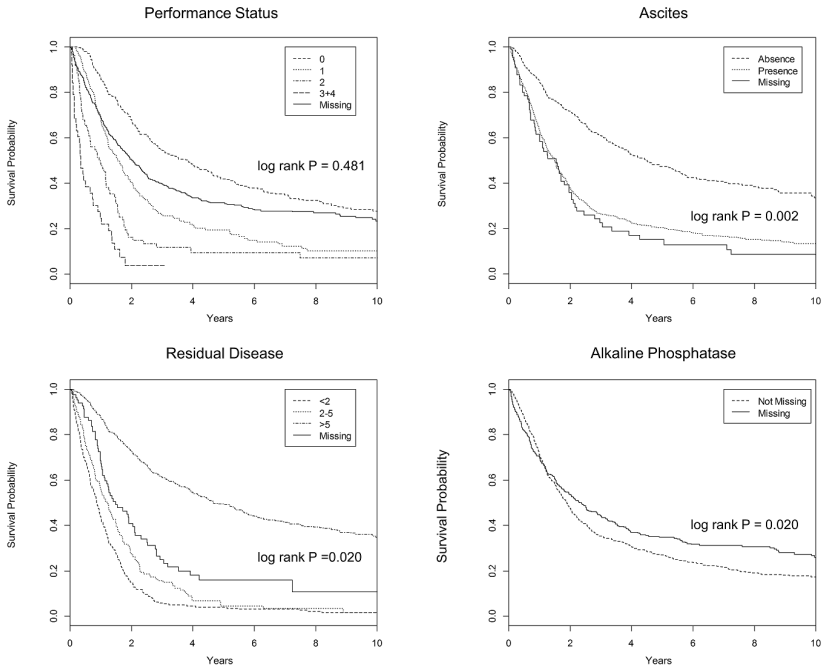

図 1. ps, ascites, resdis, alp における **Kaplan-Meier** 曲線．図中の **P** 値は，それぞれの共変量が観測された対象者と欠測した対象者の **2** グループ間での生存時間分布の差異についての **log rank** 検定の結果.

ここでは，解析対象のデータセットが $x_1, x_2, \cdots, x_q$ という変数で構成されているものとする．MICE の基本的な原理は，個々の変数に対して，3 節に示したような補完値の生成モデルをそれぞれ構築し，それらをもとに，順繰りに補完値を生成していくというものである．例えば，x_1 についての補完値を生成する際には，$x_2, ..., x_q$（もしくは，その部分集合）を説明変数とした回帰モデルによって欠測値の予測モデルを構築する．同様に，x_2 については残りの $(q-1)$ 個の変数を説明変数にしたモデルを構築し，$x_3, \cdots, x_q$ についても同様に，それら以外の $(q-1)$ 個の説明変数による予測モデルを構築する．このように，それぞれの変数の条件付き分布を残りの変数によって互いに規定することができるという仮定は，相互条件付き識別性（fully conditional specification [FCS]; van Buuren et al., 2006) といわれる．

当然ながら，それぞれの補完値の生成モデルに対して，3 節の方法によって補完値を生成するためには，説明変数となる $(q-1)$ 個の変数に完全データが得られていることが前提となる．MICE は，以下のような手順によって，補完値を代入した擬似的な完全データを，q 個の変数ごとのモデルに対して，逐次的に更新していくことによって，説明変数に欠測がある場合にも，3 節のアルゴリズムを適用できるようにするという方法である．

Step 1.（初期値の設定）

まず，便宜的に q 個の変数におけるすべての欠測データに，適当な初期値を代入し，擬似的な完全データを作成する（適当なリサンプリングなどによる補完値でよい）.

Step 2.（連鎖方程式による補完値の更新）

以下の Step 2-1, 2-2,⋯,2-q のサイクルを連鎖的に繰り返す.

Step 2-1.（x_1 についての補完値の更新）

ここまでのステップで生成された補完値のデータセットによって，擬似的な完全データを構成することができる．このデータセットにおいて，オリジナルデータで x_1 に欠測が起こった対象者の x_1 を「欠測」に戻し，3 節に示した補完値の生成方法で，$x_2, ..., x_q$ の（擬似的な）完全データから x_1 の補完値を生成する．生成された x_1 の補完値のデータセットを，x_1 の欠測データに置き換え，”current” の値として更新する．

Step 2-2.（x_2 についての補完値の更新）

Step 2-1 と同様，”current ” の擬似的な完全データに対して，オリジナルデータで x_2 に欠測が起こった対象者の x_2 を「欠測」に戻し，同じ要領で，補完値を生成する．これを，x_2 の欠測データに置き換え，”current” の値として更新する．

...

Step 2-q.（x_q についての補完値の更新）

Step 2-1, 2-2,... と同様，”current” の擬似的な完全データに対して，オリジナルデータで x_q に欠測が起こった対象者の x_q を「欠測」に戻し，同じ要領で，補完値を生成する．これを，x_q の欠測データに置き換え，”current” の値として更新する．

MICE は，この Step 2-1, 2-2,...,2-q のプロセスを連鎖的に繰り返していくことによって，逐次的に補完値のデータセットを更新していくアルゴリズムである．一般的に，Step-1 の初期値は，ラフな値に設定されることが多いため，マルコフ連鎖モンテカルロ法と同様，はじめの数回のサイクル（例えば，10-20 回程度）で得られた値は，burn-in 期間のものとして切り捨てて，それ以降の m 回のサイクルから得られた補完値を多重代入法における補完値として用いる．アルゴリズムそのものは，まさに Gibbs サンプリングに似たものとなっているが，Step 2-1, 2-2,⋯,2-q のプロセスでのそれぞれの補完値の生成が，観測データが与えられたもとでの事後予測分布からのサンプリングになっていれば，これは正確に Gibbs サンプリングに一致することになる．

しかしながら，一般的に，個々の補完値の生成では，3 節に示したような比較的簡便なパラメトリックモデルが用いられることが多く，残念ながら，これが，厳密な意味での Gibbs サンプリングに一致することはない（予測平均マッチングにおいては，近似的な意味でも事後予測分布からのサンプリングとしての理論的正当化はできない）．近年の研究により，MICE による推測の妥当性が厳密に成り立つための条件も与えられているが（Liu et al., 2013），かなり制約の強い条件となっている．しかしながら，シミュレーション等による経験的な評価においては，MICE は，他の補完値の生成方法に比べて良好な性能を示すことが多いことが知られており（Marshall et al., 2010），その実用性も相まってか，近年の *Lancet* 誌，*New England Journal of Medicine* 誌のレビューでは，欠測データの取り扱いに最も多く用いられている手法は MICE であるという報告もある（Rezvan, Lee and Simpson, 2015）．

MICE のもうひとつの特徴として，サイクルごとの補完値の生成には，3 節の方法をそのまま適用することができるため，異なる種類の変数（連続変数，2 値変数，順序属性を持つ・持たないカテゴリカル変数）が混在している場合にも，特別な措置をとることなく，既存のアルゴリズムをそのまま適用すればよいという点がある．

5. 補完値の生成モデルの構築

5.1. 変数の選択について

前節までは，MICE における補完値の生成方法についての解説を行ってきたが，多重代入法による解析の本質的な目的は，代入後のデータセットを解析・統合して得られる最終的な推測において，妥当かつ有効な結果を得ることである．ここでは，補完値の生成モデルにおける変数選択において，バイアスを避け，精度を向上させるための 2 つのポイントについてまとめる．以降では，メインの解析で用いるモデルのことを「解析モデル」，補完値の生成のための予測モデルのことを「補完モデル」と述べる．

5.1.1. 解析モデルにおける共変量とアウトカム

解析モデルにおけるバイアスを避けるために，補完モデルには，解析モデルで用いるすべての変数を含める必要がある（Schafer, 1997）．特に，解析モデルの説明変数に欠測があり，その補完値を生成する際には，解析モデルの結果変数を補完モデルに含める必要がある（Moons et al., 2006）．もしも生成された補完値を，いくつかの異なる解析モデルによる解析に共通に用いるのであれば（例えば，副次的な解析や感度解析において），すべての解析モデルに含まれる変数を網羅するように補完モデルを構築する必要がある．

生存時間解析のモデルにおいては，結果変数は，イベントまでの時間 t と，イベントの有無を表す指示変数 d によって構成される．補完モデルの説明変数が t を含むときに，これらの解析モデルの結果変数を含めるためのアプローチとして，$\{t,\log t,d\}$, $\{d,\log t\}$, もしくは $\{d,t\}$ の組を補完モデルに含めるというアプローチが考えられている（Barzi and Woodward, 2004; Clark and Altman, 2003; van Buuren, Boshuizen and Knook, 1999）．White and Royston (2009) は，解析モデルが Cox の比例ハザードモデルであるときに，この問題についての検討を行っており，$\{d,\log t\}$ を用いると，回帰係数の推定値が帰無仮説の方向にバイアスが入ることを示している．一変量モデルで，2 値の共変量を 1 つのみ扱う場合には，補完モデルは，d と累積ベースラインハザード関数 $H_0(t)$ を含むモデルとなる．実際には，$H_0(t)$ は未知であるため，Nelson-Aalen 推定量などによる適当な推定値によって近似したものを用いればよい．

5.1.2. 欠測を含む変数と相関を持つ共変量

補完モデルに，欠測を含む変数と相関を持つ共変量を含めることには，大きく 2 つの理由がある．第 1 に，MAR の仮定をより確からしいものとし，バイアスを減じるためである．多重代入法による，妥当な推測を行うための MAR の仮定は,「補完モデルに含まれた観測データで条件付けられたもとで，欠測が観測されていないデータに依存しない」というものである．したがって，補完モデルには，欠測を起こした変数を予測する変数と，欠測の有無に関する指示変数を予測する変数の両方をすべて含めるべきである．

第 2 に，より適切な補完値を生成させることにより，最終的な解析モデルについての推定値の標準誤差を減じることである．例えば，ランダム化臨床試験において，一般的に，ランダム化の後に得られる変数（割り付けられた治療へのコンプライアンスに関する変数や，治療とアウトカムの間の中間変数，解析モデルのアウトカム以外のアウトカム変数など）は，介入の効果を推定

表 2. 卵巣がんデータにおける alb と age, grade の線形回帰分析の結果. 括弧内は標準誤差の推定値.

	n	$\hat{\beta}_{\text{age}}$	$\hat{\beta}_{\text{gradeII}}$	$\hat{\beta}_{\text{gradeIII}}$
Complete Cases	727	−0.139 (0.017)	−0.303 (0.770)	−1.837 (0.714)
MICE (ALL)				
$m = 5$	1189	−0.135 (0.013)	−0.317 (0.522)	−1.749 (0.532)
$m = 200$	1189	−0.136 (0.015)	−0.397 (0.781)	−1.877 (0.755)
MICE (RES)†				
$m = 5$	797	−0.139 (0.017)	−0.291 (0.756)	−1.917 (0.722)
$m = 200$	797	−0.140 (0.017)	−0.321 (0.770)	−1.884 (0.718)

† MICE (RES): alb が欠測している対象者を除外した対象者集団において, MICE を行った結果.

するための解析モデルには加えるべきではないとされるが，これらの補助的な変数は，補完値の生成においては有用であることがある．例えば，主要エンドポイントの変数が欠測した参加者に対して，それと相応の相関を持つことが知られている他のアウトカム変数が測定されているとき，これらの観測された変数を補完モデルに加えることにより，補完モデルにおける欠測値の予測の精度を向上させることができる．これらは，最終的な推測の精度を向上させることが期待でき，また，MAR の仮定をより確からしいものとすることが期待できるだろう．

実践的には，補完モデルには，有意な関連が認められた変数のみを含める，あるいは，一定以上の関連を示した変数のみを含めるといったような基準が用いられることもあるだろう．ステップワイズ法などの変数選択のアルゴリズムも，完全ケースに対象を限定して適用することは形式的には可能である．モデル選択のための ad hoc なアプローチとして，少数（もしくは単純に 1 組）の補完値を代入したデータセットに対して，統計的なモデル選択の方法を適用するというアプローチもあり得る．

解析モデルよりも補完モデルのほうが多くの変数を含むときの多重代入法の難点については，多くの文献で議論がされてきた（Fay, 1992; Meng, 1994; Rubin, 1996）．これらの議論は，Rubin の公式による分散の推定における理論的な正当性についてのもので，推定量そのもののバイアスについての議論はない．実用上の問題に関するとりあえずの結論として，Schafer (1997) は，これらの問題はそれほど重要ではないとしている．

6. 代入後のデータの解析

6.1. アウトカムが欠測した対象者の取り扱い

解析モデルにおける結果変数が欠測した対象者は，最終的な推定値にノイズを加えるだけなので，解析から除外されるべきであるという議論は以前からある（Little, 1992; Von Hippel, 2007）．事例として，表 2 に，4.1 節の卵巣がんの臨床研究において，alb を結果変数，age, grade を説明変数とした，単純な線形回帰モデルの解析結果を示している．3 通りの結果を示しており，(i) 完全ケース解析，(ii) すべての対象者に対しての多重代入法，(iii) alb が欠測していない対象者に限定して行った多重代入法の結果を示している．代入回数を $m = 200$ とした場合の解析において，(ii), (iii) を比較すると，回帰係数の推定値は比較的類似しているが，標準誤差は，まず，解析 (ii) の $m = 5$ の設定が極端に小さく，$m = 200$ の設定とは大きく乖離している．解析 (iii) では

それほど乖離は大きくないが，これはモンテカルロ誤差による解析結果の不安定性と解釈すべきであろう．一方，(ii), (iii) を比較すると，点推定値はほとんど同じであるが，標準誤差の推定値は少しだけ (iii) のほうが小さく，確かに，結果変数が欠測した対象者を加えることで若干のノイズが加わるということがわかる．

方法 (ii) は，補完モデルに，前節のランダム化の後に得られる補助的な変数が用いられるような場合には，有用であるといえるだろう．結果変数の補完値の生成に，付加的な情報を加えることができるためである．実際のところ，方法 (ii) を用いる価値があるのは，それらの補助変数が結果変数と強い相関を持つ場合（標準誤差を小さくするため），あるいは，結果変数と結果変数が欠測する確率と相関を持つ場合（バイアスを減じるため）の 2 つのケースにまとめられるだろう．

6.2. Rubin の公式による推定値の統合

さて，回帰係数の推定値に着目することとしよう．もちろん，それぞれの代入されたデータセットにおいて計算されるさまざまな統計量にも関心はある．一般的に，解析モデルのパラメータの関数の推定量となっている統計量については，Rubin の公式によって，妥当に統合を行うことができる．統計量の分布の正規性の近似をよくするためには，適当な変換が必要とされるかもしれない（Molenberghs and Kenward, 2007）．一方，”strength of evidence” のような指標のように，パラメータの推定量でない統計量に関しては，Rubin の公式による統合を行うことはできない．概ねのところ，サンプルサイズによって系統的に値が変わるような統計量は，Rubin の公式を用いた統合を行うことはできない（White et al., 2011）．

6.3. 解析モデルにおけるモデル構築

一般的に，統計解析では，変数選択や残差診断，交互作用や非線形な関連性についての検定のように，モデル構築とモデルチェックのプロセスが必要とされる．多くの場合，欠測データが含まれるデータセットについても，完全ケースにデータを限定して，フォーマルに古典的な方法による解析が行われる．当然ながら，MCAR のようなメカニズムが仮定できない場合，このような解析は推奨されない．例えば，欠測メカニズムが MCAR でない場合，バイアスの入った回帰係数の推定値によって重要でない変数が選ばれるかもしれない．また，完全データにデータセットが限定されると，情報量の不足により，検出力が低下し，重要な共変量が検出されない可能性がある（Wood, White and Royston, 2008）．以下に，多重代入法を用いた解析におけるモデル構築をどのように行うべきかについての推奨事項をまとめる．

6.3.1. 仮説検定

一般的に，古典的なモデル構築の手法は，仮説検定を必要とする．2 節で解説した，単変量および多変量の $\boldsymbol{\theta} = \mathbf{0}$ の検定を行う Wald 統計量や，Meng and Rubin (1992) によって開発された近似的な尤度比検定などがある．尤度比検定は，同時に検定を行うパラメータが複数ある場合には有用であるだろう．古典的な統計理論においては，尤度比検定はしばしば完全データの解析においては好ましいとされるが，多重代入法を用いた解析における仮説検定ではそのような理論的根拠は得られておらず，特にどちらの方法が推奨されるということはない．一般的には，計算の簡便性から，Wald 検定がよく用いられるが，標準的な統計ソフトウェアでは，Meng and Rubin (1992) の尤度比検定も実装されているものがほとんどである．

6.3.2. 変数選択

古典的な変数選択の方法は，変数増加法，変数減少法，ステップワイズ法などのアルゴリズムによって行うことができるが，多重代入法を用いた解析にはそのままでは適用することができない．それぞれの変数選択のアルゴリズムでは，すべての代入後のデータセットに解析モデルを当てはめること（MI Step 2），そして，代入後のデータセットにおける推定値の統合（MI Step 3）を考慮する必要がある．これらのプロセスにおいて重要なのは，完全データの解析と同じように第 1 種の過誤確率を名目水準以下に保持できなくてはいけないということである（Wood et al., 2008）．しかし，これは次のようなさまざまな条件下で必ずしも可能ではないかもしれない．(i) 大規模なデータセットを対象とするとき，(ii) 反復回数 m が大きいとき，(iii) 複数のアウトカムに関心があるとき，(iv) 多くの変数や交互作用項が評価されるようなとき，などである．

プラグマティックな代替法としては，多重代入を行ったデータセットを，$m \times n$ の長さを持つ単一のデータセットであるものとして，変数選択のアルゴリズムをこのひとつのデータセットに対して行うことである．変数選択のアルゴリズムにおいて，ある共変量 x を含めるか含めないかを評価するとき，それぞれの観測データは，$(1 - f_x)/m$ の重みを持つことになる．f_x は，x における欠測データの占める割合である（Wood et al., 2008）．このアプローチは，Wood et al. (2008) によって，よい近似になることが示されており，また，多変量の部分多項式関数のような関数形の選択においても有用であることが示されている（Royston and Sauerbrei, 2008）．

6.4. モデル評価

データ解析においては，モデル評価のプロセスが含まれることになる．古典的なモデル評価の方法の多くは，それぞれの代入後のデータセットに対して適用することができる．例えば，線形回帰分析では，残差プロットなどをそれぞれの代入後のデータセットに対して作成することができる．このような個々のモデル評価の結果は，補完モデルにおいても，解析モデルにおいても，問題を特定するのに役に立つことがある．例えば，少数の代入データセットにおいて起こる極端な外れ値などは，補完モデルにおける問題が原因で起こるかもしれない．一方，すべての代入データセットにおいて一貫して起こるような問題は，解析モデルにおける問題が原因であるかもしれない．

7. 事例の解析：卵巣がんの予後因子研究 (Clark et al., 2001)

Clark and Altman (2003) では，Clark et al. (2001) の卵巣がんの予後因子研究のデータセットに対して，MICE による多変量 Cox 回帰モデルによる分析についての詳細な事例研究がなされている．ここでは，Clark and Altman (2003) の事例に倣って，MICE による予後因子解析を行うものとする．

1189 名のすべての患者に対して，追跡に関する情報が利用可能である．追跡期間中に，842 名（70.8%）の患者に死亡が確認された．残りの 347 名（29.2%）の患者における追跡期間の中央値は，1665 日（29-5852 日）であった．コホートの中での 5 年生存率は，29.6% (95%CI: 26.8-32.5%) であった．潜在的な予後因子は，表 1 に示した通りである．まず，多変量 Cox 回帰モデルによる完全データ解析（complete-case analysis）の結果を表 4 に示している．解析対象者は 362 名（イベント数 248）である．米国学術評議会の学術調査報告 (2010) で強調されているように，MCAR

表 3. R パッケージ mice における補完値の生成オプション.

名称	方法	変数の型
pmm	予測平均マッチング	連続変数
norm	線形回帰モデル（ベイズ流）	連続変数
norm.nob	線形回帰モデル（非ベイズ流）	連続変数
mean	平均代入法	連続変数
2L.norm	2 水準線形モデル	連続変数
logreg	ロジスティック回帰モデル	カテゴリカル変数（2 水準）
polyreg	多項ロジットモデル	カテゴリカル変数（3 水準以上）
polr	順序ロジットモデル	順序変数（3 水準以上）
lda	線形判別解析	カテゴリカル変数
sample	観測データからのランダムサンプル	任意

† それぞれの変数の型において，デフォルトは，pmm, logreg, polyreg, polr となっている.

(missing completely at random) の仮定は極めてあり得ない仮定であるため，得られた結果にはバイアスが生じている可能性がある．もちろん，単純な情報量の損失としても，7 割前後の情報が除かれたもとでの解析となっている．

MICE では，ここまで述べてきたとおり，複数の変数に欠測が生じた場合にも，多重代入法による解析を行うことができる．ここでは，R によるパッケージ mice（van Buuren and Groothuis-Oudshoorn, 2011）を用いた解析について解説を行う．表 3 に，mice に実装されている補完値の生成オプションの一覧を示している．mice による解析では，特に指定をしなければ，表 3 中のデフォルトの手法が用いられる．生成する補完値の組は，m によって指定することができる．デフォルトでは 5 回とされているが，2 節で述べた通り，モンテカルロ推測の手法として，相応の回数は必要である．ここでは $m = 200$ とした．補完値の生成モデルにおける，それぞれの回帰関数にモデル化する変数の組は，predictorMatrix によって指定できる．対象となるデータセットの列番号を {0, 1} によって指定する行列によって，モデル化する変数の組を選ぶことができる．擬似乱数によって生成される補完値の組を再現するためには，seed によって乱数のシードを指定しておけばよい．その他のオプションについては，van Buuren and Groothuis-Oudshoorn (2011) をご参照いただきたい．本節で示す解析プログラムは，著者のホームページ（http://normanh.skr.jp/materials.html）に公開されている．R における補完値の生成コードとその出力は以下の通りである（一部改変；tgce はここでの解析データのオブジェクト名である）.

```
> imp.tgce <- mice(tgce, m=200, predictorMatrix=predmt1, seed=34871)

Multiply imputed data set
Call:
mice(data = tgce, m = 200, predictorMatrix = predmt1, seed = 34871)
Number of multiple imputations:  200
Missing cells per column:
    patno       age      figo     grade    histol   ascites        ps    resdis
        0         0        21       139         0        65       511        81
 log_ca125   log_alp       alb         d         t
       440       396       392         0         0
Imputation methods:
    patno       age      figo     grade    histol   ascites        ps    resdis
       ""        "" "polyreg" "polyreg"        ""  "logreg" "polyreg" "polyreg"
 log_ca125   log_alp       alb         d         t
     "pmm"     "pmm"     "pmm"        ""        ""
```

```
VisitSequence:
     figo     grade   ascites        ps    resdis log_ca125   log_alp       alb
        3         4         6         7         8         9        10        11
PredictorMatrix:
          patno age figo grade histol ascites ps resdis log_ca125 log_alp alb d t
patno         0   0    0     0      0       0  0      0         0       0   0 0 0
age           0   0    0     0      0       0  0      0         0       0   0 0 0
figo          0   1    0     1      1       1  1      1         1       1   1 0 0
grade         0   1    1     0      1       1  1      1         1       1   1 0 0
histol        0   0    0     0      0       0  0      0         0       0   0 0 0
ascites       0   1    1     1      1       0  1      1         1       1   1 0 0
ps            0   1    1     1      1       1  0      1         1       1   1 0 0
resdis        0   1    1     1      1       1  1      0         1       1   1 0 0
log_ca125     0   1    1     1      1       1  1      1         0       1   1 0 0
log_alp       0   1    1     1      1       1  1      1         1       0   1 0 0
alb           0   1    1     1      1       1  1      1         1       1   0 0 0
d             0   0    0     0      0       0  0      0         0       0   0 0 0
t             0   0    0     0      0       0  0      0         0       0   0 0 0
Random generator seed value:  34871
```

それぞれの変数の欠測値に関する集計や，補完値の生成に用いた方法などが出力されている．mice の出力には，補完値を埋め込んだ m 組の擬似的な完全データセットが含まれている．

mice の出力オブジェクトに対して，以下のように coxph などの解析モデルを指定することで，それぞれの擬似的な完全データセットの解析，および，Rubin の公式による統合解析を行うことができる．

```
> ph2 <- with(imp.tgce, coxph(Surv(t, d) ~ age + figo + grade + histol + ascites
 + ps + resdis + log_ca125 + log_alp +alb))
> pool2 <- pool(ph2)
> round(summary(pool2),3)
             est    se       t          df Pr(>|t|)  lo 95  hi 95 nmis   fmi lambda
age        0.022 0.004   5.872  51990.211    0.000  0.015  0.029    0 0.060  0.060
figo2      0.638 0.165   3.876 105081.069    0.000  0.315  0.960   NA 0.041  0.041
figo3      1.164 0.142   8.172  69875.332    0.000  0.885  1.444   NA 0.051  0.051
figo4      1.216 0.174   7.003  29625.104    0.000  0.876  1.557   NA 0.081  0.081
grade2     0.386 0.176   2.201  10580.048    0.028  0.042  0.730   NA 0.136  0.136
grade3     0.427 0.170   2.510  11531.042    0.012  0.094  0.760   NA 0.131  0.130
histol2    0.182 0.203   0.897  19772.466    0.370 -0.216  0.579   NA 0.099  0.099
histol3   -0.031 0.165  -0.187  58218.824    0.852 -0.355  0.293   NA 0.057  0.057
histol4    0.338 0.143   2.370  43500.649    0.018  0.059  0.618   NA 0.066  0.066
histol5    0.798 0.190   4.196  29762.582    0.000  0.425  1.170   NA 0.080  0.080
histol6   -0.203 0.105  -1.940  81417.840    0.052 -0.409  0.002   NA 0.047  0.047
histol7    0.341 0.268   1.273   5534.986    0.203 -0.184  0.867   NA 0.189  0.189
ascites2   0.312 0.090   3.461  13311.486    0.001  0.136  0.489   NA 0.121  0.121
ps2        0.098 0.104   0.950   3151.690    0.342 -0.105  0.301   NA 0.251  0.251
ps3        0.186 0.150   1.240   1953.247    0.215 -0.108  0.480   NA 0.319  0.319
ps4        0.627 0.221   2.832    846.655    0.005  0.192  1.061   NA 0.486  0.484
resdis2   -0.105 0.113  -0.926   8489.322    0.354 -0.328  0.117   NA 0.153  0.152
resdis3   -0.635 0.113  -5.598   7746.761    0.000 -0.857 -0.412   NA 0.160  0.160
log_ca125  0.024 0.031   0.764   3140.107    0.445 -0.038  0.086  440 0.252  0.251
log_alp    0.303 0.100   3.037   1563.266    0.002  0.107  0.499  396 0.357  0.356
alb       -0.018 0.010  -1.833   2681.536    0.067 -0.038  0.001  392 0.272  0.272
```

est, se が，回帰パラメータの推定値と標準誤差である．t は，これらから計算することができる擬似 Wald 式の検定統計量，df は参照分布となる t 分布の自由度である．自由度の計算には，Barnard and Rubin (1999) による方法がデフォルトとなっている．lo 95, hi 95 は，それぞれこの参照分布から算出される 95%信頼区間の下限・上限である．nmis は欠測データの数（現在のバージョンでは，カテゴリカル変数については NA となるエラーがあるようである），fmi は欠測

表 4.　Cox 回帰モデルによる多変量解析の結果.

	Complete-Case Analysis (n=358, # deaths=245)				MICE (n=1189, # deaths=842)			
	HR	95%CI		P-value	HR	95%CI		P-value
Age (years)	1.02	1.00	1.03	0.017	1.02	1.01	1.03	< 0.001
FIGO stage								
I	1.00				1.00			
II	2.39	1.14	5.02	0.021	1.89	1.37	2.61	< 0.001
III	3.98	2.10	7.54	< 0.001	3.20	2.42	4.24	< 0.001
IV	6.07	2.91	12.67	< 0.001	3.38	2.40	4.74	< 0.001
Grade								
I	1.00				1.00			
II	1.27	0.64	2.50	0.494	1.47	1.04	2.08	0.028
III	1.12	0.59	2.13	0.723	1.53	1.10	2.14	0.012
Histology								
Serous papillary	1.00				1.00			
Adenocarcinoma	1.58	0.38	6.54	0.531	1.20	0.81	1.78	0.370
Endometrioid	0.95	0.51	1.76	0.876	0.97	0.70	1.34	0.852
Clear cell	1.70	0.99	2.91	0.055	1.40	1.06	1.86	0.018
Mixed mesodermal	3.94	1.76	8.84	0.001	2.22	1.53	3.22	< 0.001
Mucinous	0.75	0.53	1.05	0.095	0.82	0.66	1.00	0.052
Undifferentiated	1.45	0.61	3.48	0.403	1.41	1.83	2.38	0.203
Ascites								
Absence	1.00				1.00			
Presence	1.35	0.97	1.87	0.074	1.37	1.15	1.63	0.001
Performance status								
0	1.00				1.00			
1	1.15	0.84	1.57	0.383	1.10	0.90	1.35	0.342
2	0.91	0.56	1.48	0.704	1.20	0.90	1.62	0.215
3+4	2.07	0.83	5.15	0.119	1.87	1.21	2.89	0.005
Residual disease								
> 5cm	1.00				1.00			
2-5cm	1.03	0.70	1.52	0.885	0.90	0.72	1.12	0.354
< 2cm	0.61	0.42	0.89	0.010	0.53	0.42	0.66	< 0.001
Log CA125	1.06	0.96	1.17	0.238	1.02	0.96	1.09	0.445
Log alkaline phos.	1.67	1.18	2.36	0.004	1.35	1.11	1.65	0.002
Albmin	0.99	0.96	1.02	0.583	0.98	0.96	1.00	0.067

情報量の割合（fraction of missing information），`lambda` は $\hat{\mathbf{V}}(\hat{\boldsymbol{\theta}}_{\mathrm{IM}})$ のうち，補完間分散 $\mathbf{B}_{\mathrm{IM}}$ が寄与する割合 $\lambda = (\mathbf{B}_{\mathrm{IM}} + \mathbf{B}_{\mathrm{IM}}/m)/\hat{\mathbf{V}}(\hat{\boldsymbol{\theta}}_{\mathrm{IM}})$ を示している．

表 4 に，MICE による解析の結果を提示している．いずれの解析においても，`age`, `figo`, `histol` (`mixed mesodermal` vs. `serous`), `resdis`, `log alp` が有意になっている．尤度比検定によって，`histol` のすべての変数を同時に検定しても，やはり有意差が出る（$P < 0.05$）．MICE による解析では，これに加えて，`grade`, `histol (clear cell vs.serous)`, `ascites`, `ps` (3+4 vs. 1) が有意になっている．いずれの変数においても，95%信頼区間は MICE による解析のほうが狭くなっており，より多くの情報量を用いた有効な解析ができていることがわかる．また，ハザード

比の点推定値そのものについても，概ね異なる値が得られている．MICE による解析のほうが，絶対値として全体的に小さめのハザード比の推定値が得られているが（一部の変数については図 1 から示唆されたように），MCAR の妥当性が疑わしい条件下であるため，こちらの結果のほうが少なくとも真実には近いと思われる．実際のところは，米国学術評議会の学術調査報告 (2010) で強調されているように，MNAR の仮定のもとでの感度解析などまで詳細に行われることが望ましいが，これらの手法の詳細については，土居ら (2017) をご参照いただきたい．

8. 結びに代えて

本稿では，不完全データの解析において，近年，実践でも広く普及しつつある連鎖方程式を用いた多重代入法について，特に，White et al. (2011), Royston and White (2011) をもとにした解説を行った．多くの調査・実験研究において起こる欠測は，複数の変数にまたがって，しかも個人ごとに異なるパターンで欠測が起こるのが一般的であり，このような条件下で利用可能な手法は，現状ではそれほど多くはない．直接尤度法や，ベイズ流のデータ拡大法などは，理論上は適用可能であるが，識別可能性を担保するためのモデル化には，ケースバイケースでかなり複雑な検討が必要とされる．セミパラメトリック理論に基づく方法（本特集号における今井 (2017)）も，その理論・計算ツールともに，実用化に向けては未だ発展の途上であるといえるだろう．これらの方法に対して，MICE は，このような条件下で，比較的容易に実装することができる多重代入法による解析手法であり，多くの統計ソフトウェアで，非専門家にも扱いやすい計算モジュールの整備・拡充が進められた方法であるといえる．

先述の通り，その理論的正当性については，厳密な意味では必ずしも保証されないことが示されているが（Liu et al., 2013），実践的には，想定した補完値の生成モデルの構造が概ね正しければ，良好な性能を示すことが示されており（Marshall et al., 2010），今後の理論的研究がさらに進められるにつれて，より緩やかな条件下で，これらの経験的な評価を裏打ちする理論的根拠が得られていく可能性もあるだろう．また今後，セミパラメトリック理論に基づく統計手法を含め，この他にも，新規で有用な方法論が開発・整備されていく可能性があるが，MICE は，その中でも，非専門家にも理解しやすい比較的簡便で有効な方法論として，さまざまな応用分野において広く用いられていくことが予想される．

謝　辞

本研究は，日本学術振興会科学研究費補助金（課題番号：15K15954）の助成を受けて行われました．

参　考　文　献

Barnard, J., and Rubin, D. B. (1999): Small-sample degrees of freedom with multiple imputation. *Biometrika* **86**, 948–955.

Barzi, F., and Woodward, M. (2004): Imputations of missing values in practice: results from imputations of serum cholesterol in 28 cohort studies. *American Journal of Epidemiology* **160**, 34–45.

Berglund, P., and Heeringa, S. (2014): *Multiple Imputation of Missing Data Using SAS*. Cary, NC: SAS Institute Inc.

Carpenter, J., and Kenward, M. G. (2013): *Multiple Imputation and Its Application*. Chichester: Wiley.

Clark, T. G., and Altman, D. G. (2003): Developing a prognostic model in the presence of missing data: an ovarian cancer case study. *Journal of Clinical Epidemiology* **56**, 28–37.

Clark, T. G., Stewart, M. E., Altman, D. G., Gabra, H., and Smyth, J. F. (2001): A prognostic model for ovarian cancer. *British Journal of Cancer* **85**, 944-952.

Fay, R. E. (1992): When are inferences from multiple imputation valid? In *Proceedings of the Survey Research Methods Sections*, pp. 227–232. Alexandria: American Statistical Association.

Gelfand, A. E., and Smith, A. F. M. (1990): Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85**, 398–409.

Johnson, N. L. (1949): Systems of frequency curves generated by methods of translation. *Biometrika* **36**, 149–176.

Little, R. J. A. (1992): Regression with missing X's: a review. *Journal of the American Statistical Association* **87**, 1227–1273.

Little, R. J. A., and Rubin, D. B. (2001): *Statistical Analysis with Missing Data*. New York: John Wiley and Sons.

Liu, J., Gelman, A., Hill, J., Su, Y.-S., Kropko, J. (2013). On the stationary distribution of iterative imputations. *Biometrika* **101**, 151–173.

Marshall, A., Altman, D. G., and Holder, R. L. (2010): Comparison of imputation methods for handling missing covariate data when fitting a Cox proportional hazards model: a resampling study. *BMC Medical Research Methodology* **10**, 112.

Meng, X. L. (1994): Multiple-imputation inferences with uncongenial sources of input. *Statistical Science* **9**, 538–558.

Meng, X. L., and Rubin, D. B. (1992): Performing likelihood ratio tests with multiply-imputed data sets. *Biometrika* **79**, 103–111.

Molenberghs, G., and Kenward, M. (2007): *Missing Data in Clinical Studies*. Chichester: Wiley.

Moons, K. G., Donders, R. A., Stijnen, T., and Harrell, F. E., Jr. (2006): Using the outcome for imputation of missing predictor values was preferred. *Journal of Clinical Epidemiology* **59**, 1092–1101.

Morris, T. P., White, I. R., and Royston, P. (2014): Tuning multiple imputation by predictive mean matching and local residual draws. *BMC Medical Research Methodology* **14**, 75.

National Research Council. (2010): *The Prevention and Treatment of Missing Data in Clinical Trials*. Washington, D. C.: National Academies Press.

Rezvan, H. P., Lee, K. J., and Simpson, J. A. (2015). The rise of multiple imputation: a review of the reporting and implementation of the method in medical research. *BMC Medical Research Methodology* **15**, 30.

Robins, J. M., and Wang, N. (2000): Inference for imputation estimators. *Biometrika* **87**, 113–124.

Royston, P., and Sauerbrei, W. (2008): *Multivariable Model-building: A Pragmatic Approach to Regression Analysis Based on Fractional Polynomials for Modelling Continuous Variables*. Chichester: Wiley.

Royston, P., and White, I. R. (2011): Multiple Imputation by Chained Equations (MICE): Implementation in Stata. *Journal of Statistical Software* **45**, Issue 4.

Rubin, D. B. (1987): *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley.

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

SAS Institute Inc. (2015): *SAS/STAT 14.1 User's Guide*. Cary: SAS Institute Inc.

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

Schafer, J. L. (1999): Multiple imputation: a primer. *Statistical Methods in Medical Research* **8**, 3–15.

Tanner, M. A., and Wong, W. H. (1987): The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association* **82**, 528–550.

van Buuren, S., Boshuizen, H. C., and Knook, D. L. (1999): Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine* **18**, 681–694.

van Buuren, S., Brand, J. P. L., Groothuis-Oudshoorn, K., and Rubin, D. B. (2006): Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation* **76**, 1049–1064.

van Buuren, S., and Groothuis-Oudshoorn, K. (2011): mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software* **45**, 1–67.

Von Hippel, P. T. (2007): Regression with missing Ys: an improved strategy for analyzing multiply imputed data. *Sociological Methodology* **37**, 83–117.

Wang, N., and Robins, J. M. (1998): Large sample theory for parametric multiple imputation procedures. *Biometrika* **85**, 935–948.

White, I. R., and Royston, P. (2009): Imputing missing covariate values for the Cox model. *Statistics in Medicine* **28**, 1982–1998.

White, I. R., Royston, P., and Wood, A. M. (2011): Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine* **30**, 377–399.

Wood, A. M., White, I. R., and Royston, P. (2008): How should variable selection be performed with multiply imputed data. *Statistics in Medicine* **27**, 3227–3246.

今井匠 (2017): 不完全データの解析におけるセミパラメトリック推測の方法論．応用統計学 **XX**, XXX–XXX.

土居正明，大江基貴，高橋文博，藤原正和 (2017): MNAR のもとでの統計解析の方法．応用統計学 **XX**, XXX–XXX.

丸尾和司，五所正彦 (2017): National Research Council による調査報告書の解説． 応用統計学 **XX**, XXX–XXX.

（2016 年 12 月 24 日受付　2017 年 1 月 20 日最終修正　1 月 23 日採択）

著者連絡先：〒 190–8562 東京都立川市緑町 10–3
統計数理研究所
野間久史（Tel. 050–5533–8440）
E-mail: noma@ism.ac.jp

Japanese J. Appl. Statist. 34 (1) (2005), 1–20

Multiple Imputation by Chained Equation

Hisashi Noma

The Institute of Statistical Mathematics

Abstract

In most observational and experimental studies, missing data certainly happens and adequate treatments are required to prevent bias and loss of efficiency of the statistical inference. However, the missing generally occurs in multiple variables with different patterns in individual subjects. Although valid statistical inference methods are needed in these situations, most existing methods require complicated statistical models and computations. The multiple imputation by chained equation (MICE) is an effective method that can be applied in these situations, and has been widely used in many observational and experimental studies in recent years. Also, many useful statistical packages have been developed for standard statistical software recently. In this article, we provide a gentle tutorial on the MICE methodology with concrete applications to an ovarian cancer clinical study (Clark and Altman, 2003; *J. Clin. Epidemiol.* 56, 28-37).

Key words: missing data, multiple imputation, chained equation, model building, model evaluation

E-mail address: noma@ism.ac.jp

Received December 24, 2016; Received in final form January 20, 2017; Accepted January 23, 2017.