様々な多重代入法アルゴリズムの比較*

高橋　将宜**, 伊藤　孝之***

要旨

多重代入法には様々な計算アルゴリズムが存在するが、いずれのアルゴリズムが、どのような状況において優れているのかは不明である。本稿では、3 種類の多重代入法アルゴリズムのメカニズムを示し、公的経済統計における欠測値の補定に関して、これらのアルゴリズム間の相対的優位性を比較検証した。

序論

データが欠測している場合、利用可能なデータサイズが縮小し、効率性が低下する。さらに、観測値と欠測値との間に体系的な差異が存在する場合、統計分析の結果に偏りが発生する恐れがある。したがって、実際の統計分析においては、何らかの形で欠測値に対処することがほとんど常に必須なことであり、欠測データの対処法として多重代入法(Multiple Imputation)[1]が提唱されてきた(Rubin, 1987)。

理想的な欠測値への対処法は、欠測値を含む不完全データが、欠測値のない完全データと同一になる方法であるが、このような目標は、いかなる補定法を用いても達成できない。つまり、調査票を丹念に設計し、緻密にデータを収集することこそが、欠測値への最善の対処法である。しかし、いったん調査が終わると、それ以上のデータを収集できない段階となり、ここで統計的手法に基づく欠測値補定法が重要になってくる。多重代入法は、不完全データを用いた統計分析が、完全データによる統計分析と同様に、統計的に妥当になる欠測値対処法である。多重代入法の理論的概念はシンプルで、発案されてから数十年の時が経過しているが、事後分布からの無作為抽出の実装は難しく、計算アルゴリズムに関しては議論の余地がある。

1980 年代に Donald B. Rubin によって提唱されたオリジナルの多重代入法の理論は、ベイズ統計学の枠組みで構築され、マルコフ連鎖モンテカルロ法(MCMC: Markov chain Monte Carlo)に基づいていた。近年、MCMC の代替法として 2 つのアルゴリズムが提唱されている。そのうちの 1 つは、完全条件付指定(FCS: Fully Conditional Specification)である。2 つ目の代替アルゴリズムは、伝

---

* 本稿は、2013 年度統計関連学会連合大会における報告資料として、現在遂行中の研究内容を抜粋・要約したものである。なお、本研究の分析結果は、総務省・経済産業省『平成 24 年経済センサス‐活動調査』の速報結果の調査票情報を著者が独自集計したものであり、速報段階の結果であることに留意されたい。また、本稿の内容は、執筆者の個人的見解を示すものであり、機関の見解を示すものではない。（原稿提出日：2013 年 7 月 9 日）

** 独立行政法人統計センター統計情報・技術部統計技術研究課上級研究員

*** 独立行政法人統計センター製表部管理企画課経済センサス業務推進室（統計技術研究課併任）統計専門職

[1] 「多重代入法」とは、Multiple Imputation の訳である。総務省統計局及び統計センターでは、Imputation の訳語として「補定」を用いているが、Multiple Imputation の訳語としては「多重代入法」の呼び名が一般的に流通している(高橋, 伊藤, 2013, p.20)。よって、本稿においても、「多重代入法」の用語を用いる。

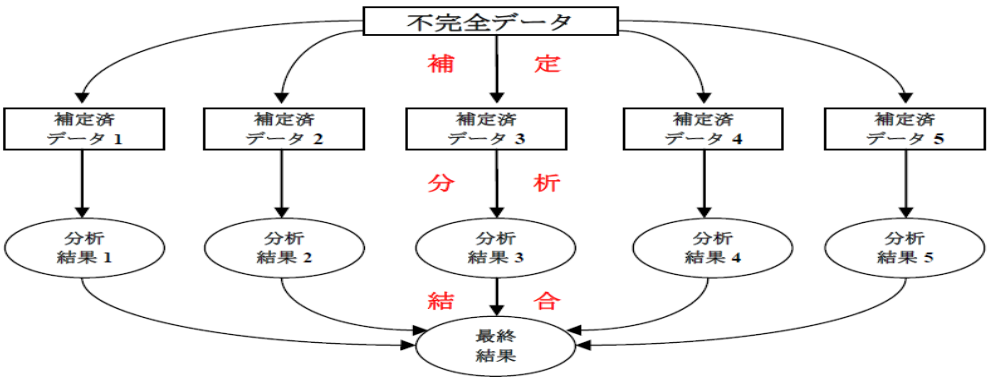
統的な期待値最大化法(EM: Expectation-Maximization)にノンパラメトリック・ブートストラップ法を応用した EMB アルゴリズムである。

したがって、多重代入法と一口に言っても、ソフトウェアに実装されているアルゴリズムには様々な方法があり、現時点において、いずれのアルゴリズムがどのような状況において優れているのかは不明である。本稿では、様々な多重代入法アルゴリズムのメカニズムを示し、経済センサス - 活動調査の速報データとシミュレーションデータを用い、公的経済統計における欠測値補定に関して、いずれのアルゴリズムが優れているかを検証する。各々のアルゴリズムは、真値との比較、計算効率などの点で評価を行う[2]。

1. 多重代入法の理論

多重代入法の理論は、Rubin (1978)によって初めて提唱された。本節では、Rubin による多重代入法の基本的なメカニズムを簡潔に示す(Rubin, 1987; Schafer, 1999; King *et al.*, 2001; Takahashi and Ito, 2012; 高橋, 伊藤, 2013)。多重代入法では、観測データを条件として、欠測データの事後分布を構築し、この事後分布からの無作為抽出を行うことで、補定にまつわる不確実性を反映させた M 個($M > 1$)の補定済データセットを生成することにより、欠測値を M 個のシミュレーション値に置き換える。これら M 個の補定済データセットを別々に使用して統計分析を行い、しかるべき手法により結果を統合し、点推定値を算出する。$M = 5$ の多重代入法の概要を図 1.1 に示す。

図 1.1: 多重代入法の模式図



欠測値を補定する際に多変量正規分布を想定しているので、補定モデルは線形である。Y_{ij}が欠測しているとする[3]。$Y_{i,-j}$は、変数Y_jを除く i 行のすべての観測値である。$\tilde{Y}_{ij}$は、式(1)より算出し

[2] 様々な多重代入法アルゴリズムを検証した最初の論文として、Allison (2000)及び Horton and Lipsitz (2001)を挙げられる。また、2000 年代における多重代入法の発展について、Allison (2002)、Horton and Kleinman (2007)、Lin (2010)も参照されたい。本稿は、日進月歩の速度で発展し続けている多重代入法に関して、最新の事情を反映したものである。

[3] 本研究で用いた記号は、以下のとおりである。Dを$n \times p$のデータセットとする（n = 標本サイズ、p = 変数の数)。もしデータが欠測していなければ、Dは平均値ベクトルμと分散・共分散行列Σで多変量正規分布しているとする。つまり、$D \sim N_p(\mu, \Sigma)$である。i を観測値のインデックスとし、$i = 1, \ldots, n$とする。j を変数のインデックスとし、$j = 1, \ldots, p$とする。$D = \{Y_1, \ldots, Y_p\}$とし、Y_jは D の j 番目の列とし、Y_{-j}はY_jの補集合とする。つまり、D 内のY_j以外のすべての列である。Rを回答指示行列(Response Indicator Matrix)とする。DとRの次元は同じであり、D が観測されるとき $R = 1$ である。D が観測されないとき $R = 0$ である。また、Y_{obs}を観測データとし、Y_{mis}を欠測データとする。つまり、$D = \{Y_{obs}, Y_{mis}\}$である。

た補定値であり、~は適切な事後分布からの無作為抽出を表す。また、βは回帰係数、εは根本的（根源的）な不確実性を表す。

$$\tilde{Y}_{ij} = Y_{i,-j}\tilde{\beta} + \tilde{\varepsilon}_i \tag{1}$$

回帰係数の算出に必要な情報は、平均値、分散、共分散の情報であり、これらはすべてμとΣに含まれている。したがって、もしμとΣが完全に既知であるならば、Y_jに基づいて真の回帰係数βを決定的に算出することができ、欠測値も決定的に補定することができる。この場合、完全データの尤度関数は、式(2)のとおりとなる。

$$L(\mu, \Sigma|D) \propto \prod_{i=1}^{n} N(Y_i|\mu, \Sigma) \tag{2}$$

残念ながら、ほとんどのデータセットには、ほぼ常に欠測値が含まれている。そこで、観測データY_{obs}の尤度を形成する際に、MAR を想定する[4]。Dの i 行の観測値を$Y_{i,obs}$と定義し、$\mu_{i,obs}$をμのサブベクトルとし、$\Sigma_{i,obs}$をΣのサブ行列とする。周辺分布は正規であるので、観測データY_{obs}の尤度関数は式(3)となる。

$$L(\mu, \Sigma|Y_{obs}) \propto \prod_{i=1}^{n} N\left(Y_{i,obs}|\mu_{i,obs}, \Sigma_{i,obs}\right) \tag{3}$$

μとΣは完全に既知ではないため、βの推定に関して確信を持つことができない。式(1)における$\tilde{\beta}$は、通常の最小二乗法におけるβの推定値$\hat{\beta}$とは異なり、こういった推定不確実性が存在していることを意味している。しかし、伝統的な手法により、式(3)を算出して、事後分布からμとΣの無作為抽出を行うことは難しい(Allison, 2002)。こういった問題を解決するために、次節で説明するとおり、様々な計算アルゴリズムが提唱されているが、これらのアルゴリズム間の相対的な優劣は、はっきりと分かっていない。

2. 多重代入法アルゴリズムとコンピュータソフトウェア

2.1 マルコフ連鎖モンテカルロ法 (MCMC): データ拡大法 (DA)

Rubin によって提唱された元来の多重代入法は、ベイズ統計学の枠組みで構築され、マルコフ連鎖モンテカルロ法(MCMC)に基づいていた(Rubin, 1987; Schafer, 1997; Little and Rubin, 2002; 岩崎, 2002; Gill, 2008)。モンテカルロ法は、シミュレーション手法の 1 つであり、一連（シリーズ）のシミュレーション値を何らかの確率分布に基づいて生成するものである。マルコフ連鎖は、確率過程であり、t の時点におけるシリーズ内の位置から別の位置へ移動する確率は、シリーズ内の現在の位置θ_tにのみ依存するものである。したがって、前期までの値$\theta_0, \ldots, \theta_{t-1}$から条件付で独立となる。MCMC の基本的なメカニズムは、もしこの連鎖が無限に長く繰り返されたならば、対象

[4] 欠測メカニズムとしては、主に 3 種類が提唱されている。1 つ目の欠測メカニズムは MCAR (**M**issing **C**ompletely **A**t **R**andom)であり、欠測の発生確率は観測データとは関係なく、完全に無作為に発生している：$P(R|D) = P(R)$。2 つ目の欠測メカニズムは MAR (**M**issing **A**t **R**andom)であり、欠測の発生確率は観測データを条件とした場合、無作為に発生している：$P(R|D) = P(R|Y_{obs})$。3 つ目の欠測メカニズムは NI (**N**on**I**gnorable)であり、欠測の発生確率はデータから独立ではなく、$P(R|D)$を単純化することはできない (Little and Rubin, 2002)。

となる事後分布を見つけることができるという点にある。したがって、連鎖を繰り返し行うことにより、これらの値の基本統計量を生成することができる。MCMC の驚異的なメカニズムは、こうして得られた分布からのシミュレーション値の各々が系列的に相関があるにもかかわらず、最終的に、周辺分布からの独立した抽出値と見なせるという点である。

データ拡大法(DA: Data Augmentation)は、MCMC の計算アルゴリズムである。Augmentation とは、「拡大」を意味する英語であるが、DA 法では、データの欠測している箇所に適当な値（初期値θ_0）を付置することで擬似的にデータを「拡大」して一時的な完全データを作成し、ここから繰り返し手法を用いて推定値を徐々に改善していく方法である。そういった意味で、DA 法はマルコフ連鎖を形成している。データ拡大法の基本的なメカニズムは、初期値θ_0から、観測データを条件として生成した欠測値の分布から補定値を生成し(I-Step: Imputation Step)、事後分布からパラメータ値を生成し(P-Step: Posterior Step)、収束するまでこれら 2 つのステップを繰り返すものである：

I-Step: $P(Y_{mis}|Y_{obs},\theta_t)$に基づいて、$Y_{mis}^{(t+1)}$を生成する

P-Step: $P\left(\theta|Y_{obs},Y_{mis}^{(t+1)}\right)$に基づいて、$\theta_{t+1}$を生成する

ここでθは未知のパラメータであり、t は繰り返し回数を意味する。

コンピュータソフトウェアプログラムでは、この種のアルゴリズムは、ジョイントモデル手法(JM: Joint Modeling)によって可能となっており、ここでは、欠測データの多変量分布の条件付分布から補定値を生成している(van Buuren and Groothuis-Oudshoorn, 2011)。もし真の同時分布が、多変量正規によって近似できるならば、統計分析も妥当なものになると保証できる(Drechsler, 2009)。このアルゴリズムを使用しているソフトウェアは、R パッケージ Norm 3.0.0 (Schafer, 2008)[5]及び SAS PROC MI 9.3 (SAS Institute Inc., 2011)[6]である。

2.2 完全条件付指定 (FCS): 連鎖方程式 (Chained Equations)

完全条件付指定(FCS: Fully Conditional Specification)は、MCMC の代替法として提唱されているアルゴリズムであり、この手法では、多変量欠測データの補定を変数ごとに行う(van Buuren and Groothuis-Oudshoorn, 2011; van Buuren, 2012)。つまり、各々の不完全な変数に対して補定モデルを構築し、それぞれの変数に対して補定値を繰り返し作成する。このアルゴリズムでは、一連の条件付密度$P\left(Y_j|Y_{-j},R,\lambda_j\right)$を介して多変量分布$P(Y,R|\theta)$を指定する。そして、$Y_{-j}$と R を条件として、Y_jを補定する。ここで、λは補定モデルの未知のパラメータである。まず、周辺分布を利用して、単純無作為抽出を行う。次に、条件付で指定した補定モデルを使用して、補定を繰り返す。条件付で指定する補定モデルには多くの種類があるが、最も有力なものは MICE (Multivariate Imputation by Chained Equations)アルゴリズムである。MICE とは、「連鎖方程式による多変量補定」という意味であり、メカニズムは以下のとおりである。

[5] Norm 3.0.0 は、R 2.9.2 以前の基盤でのみ動作する点に注意が必要である。
[6] SAS 9.3 では、実験的に FCS（2.2 項参照）をオプションとして選べるようになっているが、今回の実験では、このオプションは使用せず、MCMC のオプションのみを使用した。

データセット内の観測値と回答指示行列 R に基づいて、各々の変数Y_jの補定モデルを構築する：$P(Y_{j,mis}|Y_{j,obs}, Y_{-j}, R)$。その後、各々の変数に対し、観測値$Y_{j,obs}$からの無作為抽出により補定の初期値$\tilde{Y}_{j,0}$を設定する。このプロセスを$t = 1, \dots, T$まで繰り返す。また、このプロセスを$j = 1, \dots, p$まで繰り返す。$\tilde{Y}_{-j,t} = (\tilde{Y}_{1,t}, \dots, \tilde{Y}_{j-1,t}, \tilde{Y}_{j+1,t-1}, \dots, \tilde{Y}_{p,t-1})$は、$Y_j$を除く t 番目の繰り返しの時点における完全データである。観測値、補定値（t 番目の繰り返しの時点）、回答メカニズムを条件として、補定モデルの未知のパラメータを抽出する：$\tilde{\lambda}_{j,t} \sim P(\lambda_{j,t}|Y_{j,obs}, \tilde{Y}_{-j,t}, R)$。その後、補定値の抽出を行う：$\tilde{Y}_{j,t} \sim P(Y_{j,mis}|Y_{j,obs}, \tilde{Y}_{-j,t}, R, \tilde{\lambda}_{j,t})$。

JM と比較して、FCS には、適切な多変量分布が存在していなくても補定が可能であるという利点がある。このアルゴリズムを使用しているソフトウェアは、R パッケージ MICE 2.13 (van Buuren and Groothuis-Oudshoorn, 2011)、PASW Missing Values 18 (SPSS Inc., 2009)、SOLAS 4.01 (Statistical Solutions, 2011)[7]である。

2.3 EMB アルゴリズム

近年では、EMB (Expectation-Maximization with Bootstrapping)アルゴリズムが提唱されており、これは、伝統的な期待値最大化法(EM: Expectation-Maximization)にノンパラメトリック・ブートストラップ法を応用したものである。

EM アルゴリズムでは、まず始めに何らかの分布を想定して、平均値と分散の初期値を設定する。これらの初期値を使用して、モデル尤度の期待値を計算し、尤度を最大化し、これらの期待値を最大化するモデルパラメータを推定し、分布の更新を行う。値が収束するまで期待値ステップ(Expectation Step: E-Step)と最大化ステップ(Maximization Step: M-Step)を繰り返す。収束した値は、最尤推定値(Maximum Likelihood Estimate)であることが知られている（渡辺, 山口, 2000）。形式的には、期待値最大化法は、以下のとおり要約できる(Schafer, 1997; Little and Rubin, 2002)。初期値θ_0から始め、以下の 2 つのステップを繰り返す：

E-Step: $Q(\theta|\theta_t) = \int l(\theta|Y)\, P(Y_{mis}|Y_{obs}; \theta_t) dY_{mis}$

ここで $l(\theta|Y)$は対数尤度である

M-Step: $\theta_{t+1} = \arg\max_\theta Q(\theta|\theta_t)$ をθに関して最大化する

ノンパラメトリック・ブートストラップ法では、観測された標本データを擬似的に母集団として扱う。つまり、標本サイズ n の観測された標本データから、標本サイズ n の副標本(subsample)の無作為な復元抽出（重複を許す抽出）を行う(Wooldridge, 2002)。

これら 2 つのアルゴリズムを組み合わせることで、EMB アルゴリズムのメカニズムは以下のとおりとなる。ある不完全データ（標本サイズ= n）において、q 個の値が観測され、$n - q$ 個の値が欠測しているとする。まず、ブートストラップ法により、この不完全データから、標本サイズ n のブートストラップ副標本の抽出を M 回行う。次に、これら M 個のブートストラップ副標本の各々に EM アルゴリズムを適用し、μとΣの点推定値を M 個算出し、M 個の式(1)を用いて欠測値

[7] SOLAS は FCS の例であるが、繰り返しを行わない点に注意されたい(van Buuren and Groothuis-Oudshoorn, 2011)。本研究のために、SOLAS 4.01 を無償提供していただいたことに関して、Statistical Solutions に感謝の意を表する。

の補定を行う(Congdon, 2006; Honaker and King, 2010)。上述した 2 つのアルゴリズムとは異なり、ブートストラップ手法では、コレスキー分解[8]を行う必要はなく、χ^2分布からの抽出を行う必要もない(van Buuren, 2012)。したがって、計算の面で効率性が高いと期待される。

このアルゴリズムを使用しているソフトウェアは、R パッケージ Amelia II (version 1.6.1)である(Honaker, King, and Blackwell, 2011)。

3. 分析結果

3.1 項では EDINET 情報[9]に基づくシミュレーションデータを用い、3.2 項では 2012 年 2 月に我が国で初めて実施された経済センサス - 活動調査の速報データを用い、経済データにおける様々な多重代入法アルゴリズムの優劣を比較検討した。

3.1 巨大シミュレーションデータの多重代入

自然対数に変換した EDINET データの情報（平均値、分散・共分散など）をもとに、多変量正規分布によって観測数 100 万、5 変量のシミュレーションデータセットを生成した。データセットの基本統計量は、表 3.1 のとおりである。

表 3.1

	最小値	第 1 四分位	中央値	平均値	第 3 四分位	最大値	標準偏差
売上高	2.201	8.998	10.110	10.110	11.230	18.480	1.656
資産	2.584	9.210	10.300	10.300	11.390	18.370	1.617
資本金	0.691	7.097	8.127	8.126	9.156	15.780	1.529
売上原価	1.367	8.533	9.746	9.747	10.960	18.800	1.800
従事者数	0.000	4.221	5.053	5.054	5.888	11.080	1.237

3.1.1 欠測発生メカニズム

欠測を含む変数を y とし、欠測発生メカニズムを規定する変数を $X\{x1, x2, x3, x4\}$ とする。また、標準正規乱数をe_iとする。$\hat{y}_i = \hat{\alpha} + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4$の回帰分析を行い、予測値$\hat{y}_i$を算出する。その後、$\hat{y}_i$に標準正規乱数$e_i$を足したものを超変数とし、この値に応じてデータセットを昇順に並び替え、MAR により欠測を発生させた。各々の変数における欠測率は以下のとおりである：売上高：10%＝10 万個；資産：5%＝5 万個；資本金：5%＝5 万個；売上原価：5%＝5 万個；事業従事者数：1%＝1 万個。合計 500 万レコードのうち、26 万レコードが欠測している。また、100 万ユニットのうち、12 万 7,453 ユニットに欠測値が含まれている（12.7%）。

---

[8] コレスキー分解(Cholesky Decomposition)とは、もし A が正定値対称行列$(A = A')$であるならば、$A = HH'$に分解でき、ここで行列 H は対角線上に正の要素を持つ下三角行列である (Leon, 2006)。

[9] EDINET とは、**E**lectronic **D**isclosure for **I**nvestors' **NET**work の略であり、金融庁によって管理されている「金融商品取引法に基づく有価証券報告書等の開示書類に関する電子開示システム」のことである(金融庁, 2011)。これは、提出された書類をインターネット上で閲覧を可能とするシステムである。http://info.edinet-fsa.go.jp. (2012 年 7 月 9 日アクセス)

3.1.2 予備的結果

それぞれのプログラムにおいて、上記のデータセットに多重代入（$M = 5$）を施し、多重代入済みデータセットを用いて、式(4)における$\hat{\alpha}$と$\hat{\beta}$の推定を行った。

$$\log\left(\widehat{\text{売上高}}_i\right) = \hat{\alpha} + \hat{\beta}\log(\text{資本金}_i) \tag{4}$$

結果は、表 3.2 のとおりである。真値は、欠測のない完全なデータセットを用いた分析結果である。List-Wise は、リストワイズ除去法を用いた分析結果である。SAS、MICE、SPSS、Amelia では、すべての出力結果（回帰係数、標準誤差、*t* 値）が、リストワイズ除去法と比べて真値に近づいている。したがって、欠測を含むユニットを単純に除去するよりも、多重代入を行なう方がよいことが分かる。

表 3.2：分析結果（シード値 1223、$M = 5$）

	真値	List-Wise	Norm	SAS	MICE	SOLAS	SPSS	Amelia[10]
$\hat{\alpha}$	3.7260	4.5959	NA	3.9623	3.9900	NA	4.0120	3.9505
s.e.($\hat{\alpha}$)	0.0062	0.0071	NA	0.0069	0.0066	NA	0.0070	0.0069
t($\hat{\alpha}$)	604.5123	650.9793	NA	576.4200	605.1471	NA	598.8190	576.1718
$\hat{\beta}$	0.7862	0.6973	NA	0.7598	0.7568	NA	0.7530	0.7613
s.e.($\hat{\beta}$)	0.0007	0.0008	NA	0.0008	0.0008	NA	0.0010	0.0008
t($\hat{\beta}$)	1054.7180	839.0746	NA	927.7656	960.7229	NA	938.9850	930.9872
n	1000000	872547	NA	998848	1000000	NA	998514	998848
欠測率	0.0000	12.7453	NA	0.1152	0.0000	NA	0.1486	0.1152

注：5 つの係数と標準誤差は、Rubin の手法 (高橋, 伊藤, 2013, pp.36-37) により統合した。

SAS、MICE、SPSS、Amelia の間では、わずかながら、MICE による結果が優れていたが、今回の結果は、シード値 1223 のみに基づくものであり、シードによる結果への影響を考慮する必要がある（現在進行中）。また、全変数が欠測しているレコード（ユニット）について、MICE のみ補定を行えるため、MICE の *n* は 100 万となっている。補定モデルからのシミュレーション値を無作為に生成していると考えられる。Norm では、100 万×5 変量のデータセットを回すことができなかった。SOLAS においては、分析自体は行えるものの、メモリ不足のため、途中でエラーとなってしまった。今後の要検討事項である[11]。

多重代入法による補定値の分布を確認するために、参考として、Amelia による多重代入済データセット(m = 1)の散布図を、真の散布図及びリストワイズ除去による散布図と並列して、図 3.1

[10] Amelia の計算結果を自動的に統合する別プログラムである Zelig では、大規模データセットを扱うことができない。今回の実験では、高橋, 伊藤（2013, p.82）で公開したコードを使用して対応した。

[11] 1 万×5 変量のデータセットに関して、SOLAS は 3 分 14 秒、NORM は 36 秒、Amelia は 5 秒、MICE は 34 秒の処理時間を要した。また、1 万×5 変量のデータセットの補定値の精度に関して、プログラム間に優劣は見られなかった。したがって、小規模データセットの多重代入については、いずれのプログラムを使用しても大差はないと言える。

として掲載している。図 3.1 では、左下に欠測値が偏っているが、補定を行うことにより、分布の復元を真値に近づけることに成功していることが分かる。

図 3.1：売上高（縦軸）と資本金（横軸）の散布図（自然対数）

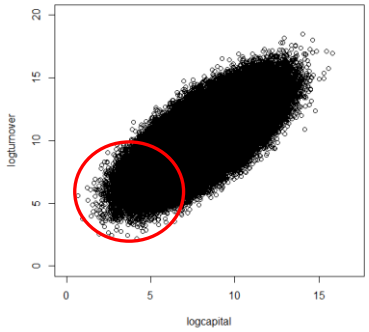
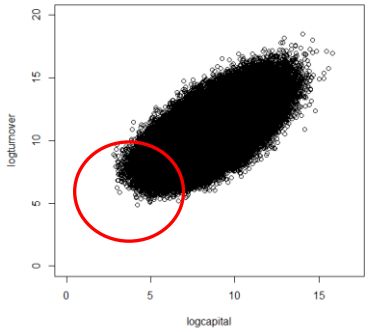
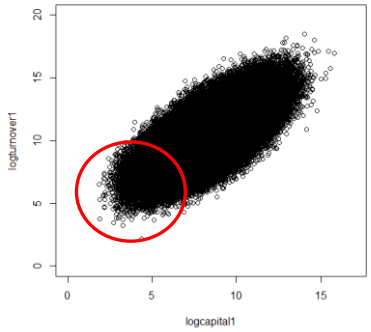
真の散布図	リストワイズ除去	Amelia II (m = 1)[12]
		

表 3.3 は、計算効率の検証を行った結果である[13]。前述したとおり、NORM と SOLAS では、巨大データセットを扱うことができなかった。SAS と Amelia は極めて速く巨大データセットを処理することができた。MICE と SPSS も、巨大データセットを扱うことはできるが、処理に多大な時間がかかった。SAS は、そもそも巨大なデータセットに向いている基盤として知られている。一方、R は、巨大なデータセットにはあまり向いていないとされている。したがって、Amelia で使用されている EMB アルゴリズムを SAS の基盤で実装すれば、さらなる計算効率の向上が見込まれる。今回は検証のために M を 5 に限ったが、実際には 20 以上が推奨されるため（高橋, 伊藤, 2013, p.46）、MICE では、最大 3 時間以上もの時間を要する可能性がある。

表 3.3：シミュレーションデータ（$M = 5$）

	NORM	SAS	MICE	SOLAS	SPSS	AMELIA
PC1	動作せず	NA	48 分 16 秒	動作せず	NA	5 分 30 秒
PC2	NA	NA	28 分 21 秒	NA	21 分 35 秒	3 分 41 秒
PC3	NA	4 分 33 秒	40 分 56 秒	NA	NA	4 分 38 秒

注：報告値は、多重代入（$M = 5$）を行うのに要した時間である。「動作せず」は、プログラムがフリーズして機能しないことを意味する。NA は、当該の PC で分析を行わなかったことを意味する。繰り返し回数の最大値は 20 に設定した。上記の結果にデータセットの読み込み時間やデータ分析の時間は含まず、純粋に多重代入の計算を行う時間のみである。

[12] Amelia II の散布図は合計 5 枚あるが、任意の 1 枚を表示している。他の 4 枚もほぼ同様の図である。

[13] 使用したパソコンの性能は、以下のとおりである。PC1 は、Windows Vista を搭載したノートパソコンであり、プロセッサは Intel Core 2 Duo CPU T9400、メモリ(RAM)は 2.00 GB、システムの種類は 32 ビットオペレーティングシステムである。PC2 は、Windows Vista を搭載したデスクトップパソコンであり、プロセッサは Intel Core 2 Duo CPU E8400、メモリ(RAM)は 2.00 GB、システムの種類は 32 ビットオペレーティングシステムである。PC3 は、Windows 7 を搭載したデスクトップパソコンであり、プロセッサは Intel Core i5 CPU 670、メモリ(RAM)は 4.00 GB、システムの種類は 32 ビットオペレーティングシステムである。

3.2 経済センサスの実データを用いた多重代入法アルゴリズムの比較

3.2.1 経済センサスとは

経済センサスは、日本全国の事業所及び企業の経済活動を調査対象とし、我が国における包括的な産業構造を明らかにし、各種経済統計のための母集団情報を整備することを目的としている。経済センサス‐基礎調査は、事業所・企業の基本的構造を明らかにするもので、平成 21 年に実施された。経済センサス‐活動調査は、事業所・企業の経済活動の状況を明らかにするもので、事業所・企業の名称や所在地だけではなく、経営組織、従業者数、売上金額といった様々な情報を収集するために平成 24 年に実施された[14]。

得られた調査結果は、総務省統計局のウェブサイト[15]に順次、公表されており、本稿の研究段階（2013 年 7 月）の時点では、速報集計のみ公開されているが、2013 年 8 月以降、確報集計も公表されていく予定である。本研究の分析結果は、総務省・経済産業省『平成 24 年経済センサス‐活動調査』の速報結果の調査票情報を著者が独自集計したものであり、速報段階の結果であることに留意されたい。

3.2.2 経済センサス‐活動調査の速報データを用いた多重代入法アルゴリズムの比較

経済センサス‐活動調査の速報データを用いて、様々な多重代入法アルゴリズムの処理速度を下記のとおり分析した。経済センサス‐活動調査には、約 580 万事業所・企業のデータが含まれているが、今回の検証では、産業 I（卸売業、小売業）の単独事業所（個人経営以外）のデータ（331,953 事業所）を用いた。使用した変数は、売上（収入）金額、売上原価、資本金又は出資金・基金の額、従業者数の 4 つである。各々の変数には、約 1%から 15%程度の欠測値が含まれている。生データの分布には経済データ特有の偏りがあるため、自然対数に変換し正規分布を近似した。

今回は、予備的実験として、処理速度の検証のみを行った。結果は表 3.6 に示すとおりである。Amelia と SAS の処理速度は極めて速かった。MICE と SPSS の処理速度は、Amelia と SAS の数倍かかっており、シミュレーションデータによる分析と一致している。SOLAS は、100 万×5 変量データセットの処理は行えなかったが、30 万×4 変量の実データセットの処理を行うことはできた。NORM は、今回もフリーズし、処理を行うことができなかった。

表 3.6：経済センサスのデータ(M = 5)

	NORM	SAS	MICE	SOLAS	SPSS	AMELIA
PC1	動作せず	NA	10 分 35 秒	22 分 15 秒	NA	1 分 24 秒
PC2	NA	NA	7 分 18 秒	NA	4 分 2 秒	55 秒
PC3	NA	1 分 15 秒	9 分 17 秒	NA	NA	1 分 14 秒

注：報告値は、多重代入（M = 5）を行うのに要した時間である。「動作せず」は、プログラムがフリーズして機能しないことを意味する。NA は、当該の PC で分析を行わなかったことを意味する。繰り返し回数の最大値は 20 に設定した。上記の結果にデータセットの読み込み時間やデータ分析の時間は含まず、純粋に多重代入の計算を行う時間のみである。

[14] http://www.stat.go.jp/data/e-census/guide/about/purpose.htm（2013 年 7 月 9 日アクセス）
[15] http://www.stat.go.jp/data/kouhyou/e-stat_e-census2012.xml（2013 年 7 月 9 日アクセス）

4. 結語と将来の課題

本稿では、様々な多重代入法アルゴリズムのメカニズムを示し、それらの性能を比較検証した。補定の精度という点では、いずれのアルゴリズムにも決定的な差はなかったが、わずかながらMICE が優位であった。今回の結果は、1 つのシード値にのみ基づくものであり、ランダムな影響を排除するために複数のシード値を用いて比較検証をする予定である。また、計算効率という点では、アルゴリズム間に大きな差が見られた。SAS と Amelia は、シミュレーションデータにおいても、経済センサス - 活動調査の速報データにおいても、十分な性能を発揮することが分かった。

参考文献

[1] Allison, Paul D. (2000). "Multiple Imputation for Missing Data: A Cautionary Tale," *Sociological Methods and Research* vol.28, no.3: 301-309.

[2] Allison, Paul D. (2002). *Missing Data*. CA: Sage Publications.

[3] Drechsler, Jörg. (2009). "Far From Normal - Multiple Imputation of Missing Values in a German Establishment Survey," *Work Session on Statistical Data Editing, UNECE*, Neuchâtel, Switzerland, October 5-7, 2009.

[4] Gill, Jeff. (2008). *Bayesian Methods—A Social Sciences Approach*, Second Edition. London: Chapman & Hall/CRC.

[5] Honaker, James and Gary King. (2010). "What to do About Missing Values in Time Series Cross-Section Data," *American Journal of Political Science* vol.54, no.2: 561–581.

[6] Honaker, James, Gary King, and Matthew Blackwell. (2011). "Amelia II: A Program for Missing Data," *Journal of Statistical Software* vol.45, no.7.

[7] Horton, Nicholas J. and Ken P. Kleinman. (2007). "Much Ado About Nothing: A Comparison of Missing Data Methods and Software to Fit Incomplete Data Regression Models," *The American Statistician* vol.61, no.1: 79-90.

[8] Horton, Nicholas J. and Stuart R. Lipsitz. (2001). "Multiple Imputation in Practice: Comparison of Sotfware Packages for Regression Models with Missing Variables," *The American Statistician* vol.55, no.3: 244-254.

[9] 岩崎学. (2002). 『不完全データの統計解析』. 東京：エコノミスト社.

[10] King, Gary, James Honaker, Anne Joseph, and Kenneth Scheve. (2001). "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation," *American Political Science Review* vol.95, no.1: 49-69.

[11] Leon, Steven J. (2006). *Linear Algebra with Applications*, Seventh Edition. Upper Saddle River, NJ: Pearson/Prentice Hall.

[12] Lin, Ting Hsiang. (2010). "A Comparison of Multiple Imputation with EM Algorithm and MCMC Method for Quality of Life Missing Data," *Quality & Quantity* vol.44, no.2: 277-287.

[13] Little, Roderick J. A. and Donald B. Rubin. (2002). *Statistical Analysis with Missing Data*, Second Edition. New Jersey: John Wiley & Sons.

[14] Rubin, Donald B. (1978). "Multiple Imputations in Sample Surveys - A Phenomenological Bayesian Approach to Nonresponse," *Proceedings of the Survey Research Methods Section, American Statistical Association*: 20–34.

[15] Rubin, Donald B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons.

[16] SAS Institute Inc. (2011). *SAS/STAT 9.3 User's Guide*. Cary, NC: SAS Institute Inc.

[17] Schafer, Joseph L. (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall/CRC.

[18] Schafer, Joseph L. (1999). "Multiple Imputation: A Primer," *Statistical Methods in Medical Research* vol.8: 3-15.

[19] Schafer, Joseph L. (2008). *NORM: Analysis of Incomplete Multivariate Data under a Normal Model, Version 3*. Software Package for R. University Park, PA: The Methodology Center, the Pennsylvania State University.

[20] SPSS Inc. (2009). *PASW Missing Values 18*. Chicago, IL: SPSS Inc.

[21] Statistical Solutions. (2011). *SOLAS Version 4.0 Imputation User Manual*. http://www.solasmissingdata.com/wp-content/uploads/2011/05/Solas-4-Manual.pdf. (Accessed on July 9, 2013).

[22] Takahashi, Masayoshi and Takayuki Ito. (2012). "Multiple Imputation of Turnover in EDINET Data: Toward the Improvement of Imputation for the Economic Census," *Work Session on Statistical Data Editing, UNECE*, Oslo, Norway, September 24-26, 2012.

[23] 高橋将宜, 伊藤孝之. (2013).「経済調査における売上高の欠測値補定方法について～多重代入法による精度の評価～」,『統計研究彙報』第 70 号 no.2, 総務省統計研修所, pp.19-86.

[24] van Buuren, Stef and Karin Groothuis-Oudshoorn. (2011). "mice: Multivariate Imputation by Chained Equations in R," *Journal of Statistical Software* vol.45, no.3.

[25] van Buuren, Stef. (2012). *Flexible Imputation of Missing Data*. London: Chapman & Hall/CRC.

[26] 渡辺美智子, 山口和範　編著. (2000). 『EMアルゴリズムと不完全データの諸問題』. 東京：多賀出版.

[27] Wooldridge, Jeffrey M. (2002). *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.