統計研究彙報　第 71 号　2014 年 3 月　（39～82）

様々な多重代入法アルゴリズムの比較
～大規模経済系データを用いた分析～

高橋　将宜†

伊藤　孝之††

Comparison of Competing Algorithms of Multiple Imputation
-Analysis Using Large-Scale Economic Data-

TAKAHASHI, Masayoshi
ITO, Takayuki

欠測データの対処法として Rubin (1978, 1987)によって提唱された多重代入法(Multiple Imputation)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。しかし、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り、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。完全条件付指定(FCS: Fully Conditional Specification)と EMB (Expectation-Maximization with Bootstrapping)アルゴリズムである。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。

キーワード : 経済調査、経理項目、欠測値 (欠損値)、補定、多重代入法 (Multiple Imputation)、R、EMB、FCS、MCMC、Amelia、MICE、NORM、SAS、SOLAS、SPSS

Using Bayesian statistics, Rubin (1978, 1987) proposed multiple imputation as a method to handle missing data, which was based on Markov chain Monte Carlo (MCMC). However, the implementation of random sampling from a posterior distribution is a difficult matter; thus, there is a room for debate in terms of computational algorithms. Recently, two other competing algorithms have been proposed. One is Fully Conditional Specification (FCS) and the other is the Expectation-Maximization with Bootstrapping (EMB) algorithm. As of this writing, it is unknown which of these algorithms outperforms the others under what circumstances. In this paper, we show the mechanisms of the various multiple imputation algorithms, and test which algorithm is superior, using the dataset of the Economic Census for Business Activity and the simulated datasets.

Keywords: Economic Survey, Accounting Item, Missing Value, Imputation, Multiple Imputation, R, EMB, FCS, MCMC, Amelia, MICE, NORM, SAS, SOLAS, SPSS

原稿受理日　平成 25 年 12 月 26 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

††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

はじめに[1]

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おそれがある。実データには、必ずと言ってよいほど欠測が発生する。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)[2]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に美しく完成されたものであるが、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。

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[3]。

第 1 節では、本稿で用いた記号について説明し、欠測に関する 3 つの主な前提について概説する。第 2 節では、欠測値補定の論理を具体的に示し、多重代入法の理論的なメカニズムを概説する。第 3 節では、3 つの多重代入法アルゴリズムを示し、それらを応用したコンピュータソフトウェアを導入する。第 4 節では、経済センサス‐活動調査のデータとシミュレーションデータを用い、3 つの多重代入法アルゴリズムの優劣を検証する。第 5 節では、経済センサス‐活動調査の速報データを用いて、多重代入法のデモンストレーションを行う。第 6 節では、多重代入法の擬似データ数（M）の決定方法を検証する。第 7 節において、結

[1] 本稿は、第 114 回研究報告会（総務省統計研修所）、第 59 回 ISI 世界統計大会（中国、香港）、2013 年度統計関連学会連合大会（大阪大学）、第 57 回経済統計学会全国研究大会（静岡大学）、2013 年度科学研究費シンポジウム（金沢大学）における報告に加筆・修正したものである。各学会における参加者の方々からは、有益なご指摘やコメントをいただいた。また、渡辺美智子先生（慶應義塾大学）、坂下信之課長（統計センター統計技術研究課）、野呂竜夫総括研究員（統計センター統計技術研究課）には、本研究の様々な段階において数々の助言や指摘をいただいた。コメントをいただいた方々には、ここに深く感謝の意を表したい。ただし、本稿にあり得るべき誤りはすべて執筆者に属する。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。

[2] 「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。

[3] 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、多重代入法の最新事情を反映したものである。また、先行文献においては、MCMC、FCS、EMB の 3 アルゴリズムを比較したものはなく、その点で欠測値補定の文献に貢献するものである。

語で締めくくる。また、補論 1 では、ISI 世界統計大会及び統計関連学会連合大会における欠測値補定に関する最新の動向を紹介する。補論 2 では、多重代入法の理論的枠組みとして使用されているベイズ統計学の骨子を参考情報として示す。

1　記号と欠測に関する 3 つの前提

本節では、1.1 項において、本研究で用いた記号と前提条件について簡潔に説明を行う。1.2 項において、欠測発生メカニズムに関する 3 つの前提について、簡単に触れる。

1.1　本稿で用いた記号

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。

1.2　欠測のメカニズムの前提

不完全データの分析では、欠測のメカニズムの種類に応じて、対象としているパラメータに関する不偏推定量が存在するか否かが決まる。よって、欠測のメカニズムの想定は重要な事項となる(岩崎, 2002, p.7; Marti and Chavance, 2011)。

欠測メカニズムとしては、主に 3 種類が提唱されている(Little and Rubin, 2002, pp.11-12, pp.312-313)。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ず、無視することができない (Little and Rubin, 2002)。これら欠測メカニズムの詳細な説明については、高橋, 伊藤 (2013, pp.20-25)を参照されたい。

1.3　まとめ

本研究では、上述の記号と欠測の発生メカニズムの前提を用いて議論を進める。とりわけ、欠測のメカニズムについては、主に MAR（及び MCAR）の前提に立って議論をする。

2　欠測値補定のメカニズムと多重代入法の理論

2.1 項では、データセット内に欠測値が含まれている場合の対処方法について概観し、従来の単一代入法(Single Imputation)による補定方法の欠点を示す。2.2 項において多重代入法のメカニズムを紹介する。

2.1　欠測値と補定

表 2.1a のデータセットには、9 人の身長、年齢、国籍、性別、体重に関するデータが記録されているが、ID 9 の人の身長の値が欠測している。

表 2.1a（例示用データ）

ID	身長	年齢	国籍	性別	体重
1	174	31	米国	男	62
2	161	45	米国	女	48
3	158	24	日本	女	42
4	163	52	米国	女	58
5	172	29	日本	男	70
6	153	38	日本	女	46
7	178	28	米国	男	70
8	170	44	日本	男	63
9	欠測	40	日本	男	69

通常、欠測値の対処法として頻繁に使用されるリストワイズ除去法では、未知の欠測値を含む行(ID 9 の行)を削除し、データセットを擬似的に長方形にすることで、統計分析を可能としているが、欠測を含まない変数（年齢、国籍、性別、体重）の貴重な情報も捨て去ってしまうことになる。もし欠測を含む変数が何であるのかさえ分からず、データセット内に他の補助変数の情報もない状況であれば、欠測値は－∞から∞までのどの値を取るのか、全く見当もつかないことになり、まさしく欠測値は未知であるため、リストワイズも致し方ない。

しかし、表 2.1a のデータでは、欠測を含む変数が「身長」であることが分かっている。つまり、欠測値は、完全に未知ではないのである。また、年齢変数を見ると、成人のデータであることが分かる。ギネス記録によれば、成人人類の身長は、約 55cm から 272cm の範囲に入ると推定できる。これだけの情報があるだけでも、「－∞～∞」という途方もない範囲から、「55cm～272cm」という有限の範囲に候補を狭めることができている。

次に、身長の値が欠測している人(ID 9)の他の変数の情報を見ると、この人は、日本人成人男性であることが分かる。日本人成人男性の身長は、平均約 170cm、標準偏差 5.8 程度で近似的に正規分布していると考えられている。したがって、ID 9 の身長は、ほぼ 100%に近い確率で 140cm 以上 200cm 以内の身長であると推定できる。これにより、最小値を 55cm から 140cm まで上げることができ、最大値を 272cm から 200cm まで下げ、推定値の幅を狭めることに成功している。さらに、日本人成人男性の身長と体重の相関データでは、身長 178cm の人の平均体重が約 69kg になるため、体重 69kg の ID 9 の身長は 178cm ぐらいと推定できるであろう。

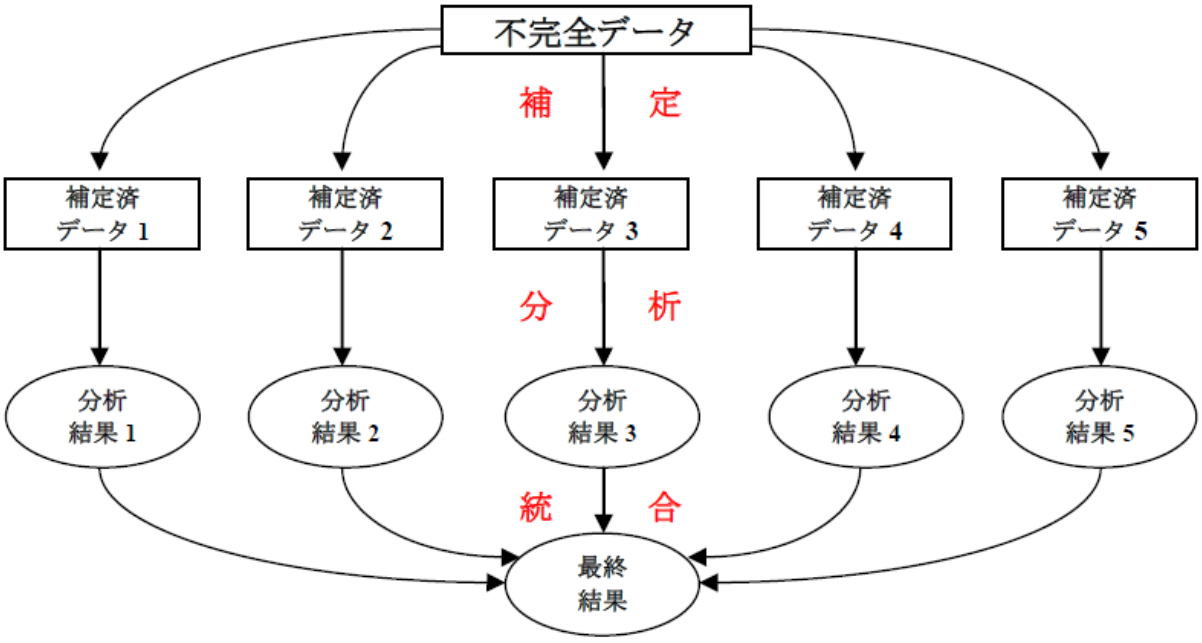
このように、論理や実データに基づいて欠測値の取り得る範囲を狭めていき、本来ならば未知であるはずの欠測値を推定して合理的な値に置き換える作業のことを補定(imputation)と呼ぶ(de Waal *et al.*, 2011, p.224)。しかし、体重 69kg の日本人成人男性のすべてが身長 178cm であるとは信じ難い。おおよそ 178cm の周辺の値であると思われるが、中には 178cm 以上の人もいれば、178cm 未満の人もいるだろう。合理的に 178cm ぐらいだと推定はできるものの、厳密に 1 つの値を特定することはできない。

欠測値は観測されず未知のため、補定値には常に不確実性がつきまとう。単一代入法では不確実性に対処できないため、次節で見るとおり多重代入法の理論が提唱されてきた。

2.2　多重代入法概論

本項では、多重代入法の基本的なメカニズムを簡潔に示す(Rubin, 1987; King *et al.*, 2001; 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[4]を構築し、この事後分布からの無作為抽出を行うことで、欠測値を M 個($M > 1$)のシミュレーション値に置き換える。その結果、補定にまつわる不確実性を反映させた M 個の補定済データセットが生成される。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 2.1 に示す[5]。

図 2.1: 多重代入法の模式図



[4] 事前分布と事後分布とは、ベイズ統計学における専門用語である。詳しくは、補論 2 を参照されたい。

[5] 図 2.1 において、「M 個のデータセットを別々に分析してから統合」という流れは非常に重要なポイントである。もし、このステップを逆にして、「M 個のデータセットを統合してから分析」した場合、多重代入法の利点が失われ、本質的に単一代入法と同じになってしまうからである。

多重代入法により生成した M 個の補定済データセットを別々に使用して、t 検定や回帰分析などの統計分析を行い、以下のとおり推定値を統合し、点推定値を算出する。$\hat{\theta}_m$をパラメータθの m 番目の補定済データセットに基づいた推定値とする。統合した点推定値$\bar{\theta}_M$は式(1)のとおりである。

$$\bar{\theta}_M = \frac{1}{M}\sum_{m=1}^{M}\hat{\theta}_m \tag{1}$$

$\bar{\theta}_M$の分散T_Mは、式(2)のとおりである[6]。$\bar{v}_M$を補定内分散の平均とする。$\tilde{v}_M$を補定間分散の平均とする。つまり、$\bar{\theta}_M$の分散は、補定内分散$\bar{v}_M$と補定間分散$\tilde{v}_M$を考慮に入れたものである。

$$T_M = \bar{v}_M + \left(1+\frac{1}{M}\right)\tilde{v}_M = \frac{1}{M}\sum_{m=1}^{M} v_m + \left(1+\frac{1}{M}\right)\left[\frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{\theta}_m - \bar{\theta}_M\right)^2\right] \tag{2}$$

欠測値を補定する際に多変量正規分布を想定（1.1 項参照）しているので、補定モデルは線形である。変数 j の i 行の観測値Y_{ij}が欠測しているとする。$Y_{i,-j}$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(3)より算出した補定値であり、~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。

$$\tilde{Y}_{ij} = Y_{i,-j}\tilde{\beta} + \tilde{\varepsilon}_i \tag{3}$$

表 2.1b のデータセットは、表 2.1a に多重代入法による補定済データセット（補定 1 から補定 5）を付置したものである。すなわち、$\widetilde{身長}_i = \tilde{\beta}_0 + \tilde{\beta}_1 年齢_i + \tilde{\beta}_2 国籍_i + \tilde{\beta}_3 性別_i + \tilde{\beta}_4 体重_i + \tilde{\varepsilon}_i$として算出したものである。

表 2.1b（例示用データ：欠測値補定済）

ID	身長	年齢	国籍	性別	体重	補定 1	補定 2	補定 3	補定 4	補定 5
1	174	31	米国	男	62	174	174	174	174	174
2	161	45	米国	女	48	161	161	161	161	161
3	158	24	日本	女	42	158	158	158	158	158
4	163	52	米国	女	58	163	163	163	163	163
5	172	29	日本	男	70	172	172	172	172	172
6	153	38	日本	女	46	153	153	153	153	153
7	178	28	米国	男	70	178	178	178	178	178
8	170	44	日本	男	63	170	170	170	170	170
9	欠測	40	日本	男	69	**184.8**	**174.6**	**178.3**	**177.0**	**173.0**

[6] なお、$(1+1/M)$は、M のサイズが無限でないために発生するシミュレーションエラーを調整する項である。M が無限大の場合、$\lim_{M\to\infty}\left(1+\frac{1}{M}\right)\tilde{v}_M = \tilde{v}_M$である。また、$\hat{\theta}_m$の分散$\mathrm{var}(\hat{\theta}_m)$の推定値を$v_m$とする。

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[7]。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(4)のとおりとなる。

$$L(\mu, \Sigma|D) \propto \prod_{i=1}^{n} N(Y_i|\mu, \Sigma) \tag{4}$$

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。μとΣが完全には既知ではなく[8]、βの推定に関して確信を持つことができない。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。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(5)となる。

$$L(\mu, \Sigma|Y_{obs}) \propto \prod_{i=1}^{n} N\left(Y_{i,obs}|\mu_{i,obs}, \Sigma_{i,obs}\right) \tag{5}$$

式(3)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(5)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; de Waal *et al.*, 2011, p.270)。

2.3　まとめ

本節では、欠測値補定の論理を示し、事前情報とデータを利用して確率を更新するベイズ統計学のメカニズムに基づいて、観測データを事前情報とし、事後分布を構築する多重代入法のメカニズムを示した。しかし、計算上の問題として、伝統的な手法によって式(5)を算出することが難しく、この事後分布からどのようにしてμとΣの無作為抽出を行うかという問題がある。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。

---

[7] たとえば、X と Y の単回帰$(Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i)$の場合、傾きの回帰係数は、$\hat{\beta}_1 = \frac{\sum(X_i-\bar{X})(Y_i-\bar{Y})}{\sum(X_i-\bar{X})^2} = \frac{\sum(X_i-\bar{X})(Y_i-\bar{Y})/(n-1)}{\sum(X_i-\bar{X})^2/(n-1)} = \frac{\mathrm{cov}(X,Y)}{\mathrm{var}(X)}$であり、切片の回帰係数は$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X}$である。つまり、平均ベクトルμと分散・共分散行列Σが既知であれば、回帰係数の算出を行うことができる。

[8] 表 2.1 の身長の場合、$\mu = \frac{174+161+158+163+172+153+178+170+身長_9}{9} = \frac{1329+身長_9}{9}$となり、これ以上、簡略化することができない。リストワイズ除去法により、$\mu_{obs} = \frac{174+161+158+163+172+153+178+170}{8} = 166.125$と算出されるが、$\mu_{obs} = 身長_9$という特殊条件の場合を除くと、$\mu \neq \mu_{obs}$である。

3　多重代入法アルゴリズムとコンピュータソフトウェア

1970 年代後半には、ハーバード大学統計学科の Rubin (1978)により多重代入法の理論が提唱されていた。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。しかし、事後分布からの無作為抽出の実装は難しく、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 (Non-Parametric Bootstrapping)を応用した EMB アルゴリズムである。

本節では、これら様々な多重代入法アルゴリズムのメカニズムを示し、それらを応用したコンピュータソフトウェアを紹介する。なお、本節では、ソフトウェアの核となる使用法についてのみ言及する。更に詳しい使用法については、Schafer (2008)、SPSS Inc. (2009)、SAS Institute Inc. (2011)、Statistical Solutions (2011)、Honaker, King, and Blackwell (2013)、van Buuren and Groothuis-Oudshoom (2013)を参照されたい。また、R に関する入門的解説については、金 (2007, 第 1 章～第 2 章)及び青木 (2009, 第 1 章～第 2 章)が分かりやすいので、参照されたい。本稿で用いたプログラムは、Amelia II Ver. 1.6.1 (R 3.0.0)、MICE 2.15（R 3.0.0）、Norm 3.0.0（R 2.9.2）、SAS 8.2、SOLAS 4.01、SPSS 18 である。

3.1 マルコフ連鎖モンテカルロ法 (MCMC): データ拡大法 (DA)

Rubin (1987)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)。本項では、MCMC のメカニズムを示し、それに基づくソフトウェアプログラムを紹介する。

3.1.1　MCMC のメカニズム

モンテカルロ法は、シミュレーション手法の 1 つであり、シリーズと呼ばれる一連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、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が、シリーズ内の現在の位置θ_tにのみ依存するという確率過程(stochastic process)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(Gill, 2008)。

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(Little and Rubin, 2002)：

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。

つまり、$P(Y_{mis}|Y_{obs},\theta_0)$に基づき、$Y_{mis}^{(1)}$を生成し、$P\left(\theta|Y_{obs},Y_{mis}^{(1)}\right)$に基づいて、$\theta_1$を生成する。次に、$P(Y_{mis}|Y_{obs},\theta_1)$に基づき、$Y_{mis}^{(2)}$を生成し、$P\left(\theta|Y_{obs},Y_{mis}^{(2)}\right)$に基づいて、$\theta_2$を生成する。この作業を収束するまで繰り返すのである。

3.1.2　MCMC を用いたコンピュータソフトウェアプログラム

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 (Schafer, 2008)及び SAS PROC MI (SAS Institute Inc., 2011)である。

3.1.2.1　R パッケージ Norm

Norm は、ペンシルベニア州立大学の Joseph L. Schafer (1997, 2008)により開発されたプログラムである。Schafer は Rubin の直弟子[9]であり、Norm は Rubin のオリジナルの多重代入法を最も忠実に再現していると言える。なお、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。

まず、Schafer のウェブサイト[10]より norm 3.0.0.zip のファイルをローカルに保存し、R を起動させ、「パッケージ→ローカルにある zip ファイルからのパッケージのインストール」をクリックして、Norm 3.0.0 をインストールする。その後、以下の要領でデータを読み込み、`library` 関数を用いて Norm を起動させる。なお、インストールは初回のみ行い、次回以降はデータの読み込みと `library` 関数による Norm の起動から作業を行えばよい。

[9] Joseph L. Schafer は、ハーバード大学において、Donald B. Rubin を指導教官として多重代入法アルゴリズムに関する論文により統計学の博士号を取得した(Schafer, 1992)。
[10] http://sites.stat.psu.edu/~jls/norm3/（2013 年 12 月 26 日アクセス）

```
setwd("D:/フォルダ名")
data<-read.csv("データセット名.csv",header=T)
attach(data)
library(norm)
```

次に、補論 2 で議論するとおり、初期値の設定が問題となるが、ここでは初期値として emNorm 関数を用いて期待値最大化法(EM)[11]の値を使用する。また、多重代入法は乱数を発生させて欠測値補定を行うため、再現性を保つために set.seed 関数によりシードを設定する。

```
emResult<-emNorm(data)
set.seed(数字)
```

下準備が整ったので、いよいよ MCMC による多重代入を行う。mcmcNorm は、マルコフ連鎖モンテカルロ法により多重代入法を行う関数である。iter=の右辺はマルコフ連鎖の繰返数であり、impute.every の右辺は上記繰返しのうち、指定の数字ごとのデータを保存する。下記の例では、繰り返し回数 5000 のうち、1000 ごとのデータを保存しているので、5000 / 1000 = 5 個の多重代入済データセットが作成されている(M = 5)。summary 関数を用いて、結果を表示する。

```
mcmcResult<-mcmcNorm(emResult, iter=5000, impute.every=1000)
summary(mcmcResult)
```

生成した補定値の結果は、mcmcResult$imp.list[[数]][,数]の形で保存されている。[数]は m 番目の補定値を表しており、[,数]は変数の番号を表している。つまり、m = 2 の 3 列目の変数の補定値は、mcmcResult$imp.list[[2]][,3]として保存されている。したがって、平均値を参照したい場合には、mean 関数を、標準偏差を参照したい場合には、sd 関数を以下のように用いる。また、回帰分析を行う場合は、以下のとおり lm 関数を用いればよい。

```
mean(mcmcResult$imp.list[[数]][,数])
sd(mcmcResult$imp.list[[数]][,数])
summary(lm(mcmcResult$imp.list[[数]][,数]~mcmcResult$imp.list[[数]][,
数]+mcmcResult$imp.list[[数]][,数]))
```

この作業を M 回繰り返し、式(1)と式(2)を用いて統合すれば、分析終了である。

[11] EM については、3.3.1 項を参照されたい。

3.1.2.2　SAS MI プロシージャ

SAS とは、Statistical Analysis System の略であり、様々な統計分析を行える汎用的商用統計プログラムとして、幅広い分野で使用されている。本項では、その機能の中でも特に欠測値補定に関し、多重代入法の使用方法について解説をする(SAS Institute Inc., 2011; Yuan, 2011)。SAS Proc MI 9.3 では、実験的に FCS（3.2 項参照）をオプションとして選べるようになっているが、今回の検証では、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。

SAS の処理は大まかに DATA ステップと PROC ステップから構成されている。DATA ステップにおいて使用するデータを読み込み、SAS 用のデータセットに加工する。次に、PROC ステップにおいて、SAS プロシージャと呼ばれる様々な統計解析機能を利用して統計分析を行う。なお、SAS の基本的な操作方法については、SAS インスティチュートジャパン (1999) などの入門書を参照されたい。

まず、data ステートメントによりデータセットに任意の名前（ここでは、sasmi）を付ける。infile ステートメントを用いてデータの保存箇所を指定し、input ステートメントにより変数に名前を付けて読み込み、run ステートメントを使用して実行する。

```
data sasmi;
infile 'D:/フォルダ名/データ名.csv' dlm=',';
input 変数 1 変数 2 変数 3;
run;
```

データが正しく読み込まれているかを確認するには、以下の print プロシージャを用いる。データセットすべてを表示すると見づらくなるので、obs = 10 と指定することで、読み込んだデータの最初の 10 行のみを表示する。

```
proc print data=sasmi (obs=10);
run;
```

多重代入法を実行するには、mi プロシージャを使用する。nimpute=の右辺は多重代入済データセット数の *M*、data=の右辺は多重代入を施す入力データセット名であり、seed=の右辺は再現性を保つためにシード値を指定する。out=の右辺は、補定値を格納した出力データの任意の名称である。mcmc ステートメントでアルゴリズムを MCMC に指定し、run ステートメントを使用して実行する。なお、SAS における欠測値はドット(.)で表され、NA や空欄はエラーの原因となることに注意されたい。

```
proc mi nimpute=数字 data=sasmi seed=数字 out=outmi;
mcmc;
run;
```

補定済データセットの基本統計量を見るには、means プロシージャを用いる。data=の右辺は、上記の多重代入で出力したデータセット名（ここでは outmi）を指定する。var ステートメントにより、基本統計量を見る変数名（ここでは変数 1）を指定する。

```
proc means data=outmi;
var 変数 1;
by _Imputation_;
run;
```

補定済データセットを用いた回帰分析を行うには、reg プロシージャを以下のように用いる。data=の右辺は、上記の多重代入で出力したデータセット名（ここでは outmi）を指定する。outest=の右辺は、パラメータの推定値や補定番号を識別するための_Imputation_変数を含む出力データの名前を指定する。covout ステートメントは、outest データセットのパラメータ推定値の分散・共分散行列を出力する。

```
proc reg data=outmi outest=outreg covout;
model 変数 1= 変数 2 変数 3;
by _Imputation_;
run;
```

以上のように出力した回帰分析の結果は、M の数だけ算出されているため、以下の mianalyze プロシージャを用いて 1 つの結果に統合する。data=の右辺は、上記の回帰分析の outest=outreg で出力したデータセットを使用する。modeleffects ステートメントでは、切片とそれぞれの変数の係数及び標準誤差を出力するように指定する。

```
proc mianalyze data=outreg;
modeleffects Intercept 変数 2 変数 3;
run;
```

3.2 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、JM 手法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)。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。本項では、FCS のメカニズムを示し、それに基づくソフトウェアプログラムを紹介する。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。

3.2.1　FCS のメカニズム

このアルゴリズムでは、一連の条件付密度$P(Y_j|Y_{-j},R,\lambda_j)$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(1)～(6)のとおりである(van Buuren, 2012)。

(1) $P(Y_{j,mis}|Y_{j,obs},Y_{-j},R)$

データセット内には、観測値として、欠測を含む変数の観測されている部分$Y_{j,obs}$と欠測を含まない変数Y_{-j}がある。これらデータセット内の観測値と回答指示行列 R を条件として、各々の変数Y_jの補定モデルを構築する。

(2) 初期値$\tilde{Y}_{j,0}$を設定

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。

(3) 繰り返し

繰り返し回数$t=1,\dots,T$まで繰り返す。また、変数$j=1,\dots,p$まで繰り返す。

(4) $\tilde{Y}_{-j,t}=(\tilde{Y}_{1,t},\dots,\tilde{Y}_{j-1,t},\tilde{Y}_{j+1,t-1},\dots,\tilde{Y}_{p,t-1})$

$\tilde{Y}_{-j,t}$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。つまり、1 から$j-1$までの変数については、t 番目の繰り返しまでの補定値が得られている（$\tilde{Y}_{1,t},\dots,\tilde{Y}_{j-1,t}$）。また、$t$ の時点でj番目の変数の補定値を算出しようとしているので、$j+1$からpまでの変数については、$t-1$番目の繰り返しまでの補定値が得られている（$\tilde{Y}_{j+1,t-1},\dots,\tilde{Y}_{p,t-1}$）。

(5) $\tilde{\lambda}_{j,t}\sim P(\lambda_{j,t}|Y_{j,obs},\tilde{Y}_{-j,t},R)$

$Y_{j,obs}$は観測値であり、$\tilde{Y}_{-j,t}$は上記で算出した補定値（t 番目の繰り返しの時点）であり、Rは回答メカニズムである。これらを条件として、補定モデルの未知のパラメータλを抽出する。

(6) $\tilde{Y}_{j,t}\sim P(Y_{j,mis}|Y_{j,obs},\tilde{Y}_{-j,t},R,\tilde{\lambda}_{j,t})$

抽出された補定モデルのパラメータλを条件に追加し、補定値$\tilde{Y}_{j,t}$の抽出を行う。

3.2.2 FCS を用いたソフトウェアプログラム

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 (van Buuren and Groothuis-Oudshoorn, 2011)、PASW Missing Values (SPSS Inc., 2009)、SOLAS (Statistical Solutions, 2011)である。

3.2.2.1 R パッケージ MICE

MICE は、オランダのユトレヒト大学の Stef van Buuren (2012)を中心としたチームにより開発された非常にフレキシブルな多重代入法プログラムである。

まず、CRAN のウェブサイト[12]より mice 2.15.zip のファイルをローカルに保存し、R を起動させ、「パッケージ→ローカルにある zip ファイルからのパッケージのインストール」をクリックして、MICE 2.15 をインストールする。その後、以下の要領でデータを読み込み、library 関数を用いて MICE を起動させる。なお、インストールは初回のみ行い、次回以降はデータの読み込みと library 関数による MICE の起動から作業を行えばよい。

```
setwd("D:/フォルダ名")
data<-read.csv("データ名.csv",header=T)
attach(data)
library(mice)
```

mice 関数を用いて多重代入を行う。data は多重代入を施すデータ名、m=の右辺は多重代入済データセット数、seed=の右辺はシード値の設定、meth="norm"は補定モデルとしてベイズ線形回帰を使用することを意味しており、これ以外にも様々なモデルを指定することができる(van Buuren and Groothuis-Oudshoom, 2013, p.42)。

```
imp<-mice(data, m=数字, seed=数字, meth="norm")
```

上記で生成した補定済データは imp の名称で保存されており、with 関数を用いて、下記のとおり回帰分析を行うことができる。統合後の結果を見るには、pool 関数を用いる。

```
fit<-with(imp,lm(変数 1~変数 2+変数 3))
summary(pool(fit))
```

[12] http://cran.r-project.org/web/packages/mice/index.html（2013 年 12 月 26 日アクセス）。2013 年 12 月 26 日現在における最新版は、mice 2.18.zip である。

さらに、生成した補定済データセットを csv ファイルとして保存するには、以下のようにすればよい。

```
dataimp<-complete(imp, action="broad", include=FALSE)
dataimpdf<-data.frame(dataimp)
write.csv(dataimpdf,"micedata.csv")
```

3.2.2.2　SPSS Missing Values (PASW Missing Values)[13]

SPSS とは、Statistical Package for Social Science の略であり、様々な統計分析を行える汎用的商用統計プログラムとして、とりわけ、人文・社会科学の分野で使用されている。本項では、その機能の中でも特に欠測値補定に関し、多重代入法の使用方法について解説をする (SPSS Inc., 2009; 岩崎, 2002, pp.326-331)。なお、SPSS の基本的な操作方法については、石村, 石村 (2007)などの入門書を参照されたい。

使用するデータを SPSS に読み込んだ上で、まず、乱数の設定を行う。「変換」タブから「乱数ジェネレータ」を選び、「アクティブジェネレータを設定」を開いて「Mersenne Twister (M)」にチェックを付ける。その後、「アクティブジェネレータの初期化」を開いて、「出発点」にチェックを入れ、「固定値」にチェックをし、具体的な値を入力し、「OK」をクリックしてシード値を設定する。

多重代入を行うには、「分析」タブから「多重代入」を選び、「欠損データ値を代入」を選択する。その後、「変数」タブから「モデル内の変数」を開き、「代入(M) = 数字」として、擬似データ数の M を設定し、「新しいデータセット」として任意の名前を付ける。その後、「方法」タブをクリックし、「ユーザー指定(c)」をクリックし、「OK」をクリックすれば、多重代入を開始する。

多重代入済データセットを用いた出力結果は以下のとおり表示する。まず、平均値と標準偏差といった基本統計量を表示するには、「分析」タブから「記述統計」を選び、「記述統計」を開いて、出力したい変数を選んで「OK」をクリックする。回帰分析を行うには、「分析」タブから「回帰」を選び、「線型」を開く。その後、「変数の選択」で「従属変数」と「独立変数」を選び、「OK」をクリックすればよい。

なお、上記のとおり、SPSS はタブをクリックするだけで簡単に使用することができるが、本格的に操作をしたい場合には、シンタックスに直接記述して操作することが望ましい。SPSS における多重代入法のシンタックスは以下のとおりである。まず、SET RNG にて、乱数のシード値を指定する。

```
SET RNG=MT MTINDEX=数字.
```

[13] 2009 年 9 月リリースのバージョン 17.0.3 から 2010 年 9 月のバージョン 18.0.3 まで、ソフトウェアの名称が PASW に変更されていたが、2010 年 8 月のバージョン 19.0 から再び SPSS の名称に戻っている。本研究では、PASW 18.0 を用いた。

多重代入を行うデータセットを spssmi として指定する。多重代入法にて使用する変数を「変数１　変数２　変数３」として指定する。IMPUTE METHOD=にて補定手法を指定する（ここでは FCS を用いた)。MAXITER=にて、繰り返し回数を指定する。NIMPUTATIONS=にて、M 数を指定し、多重代入を実行する。

```
*Impute Missing Data Values.
DATASET DECLARE spssmi.
MULTIPLE IMPUTATION 変数１　変数２　変数３
  /IMPUTE METHOD=FCS MAXITER=数字 NIMPUTATIONS=数字 SCALEMODEL=LINEAR
INTERACTIONS=NONE SINGULAR=1E-012 MAXPCTMISSING=NONE
  /MISSINGSUMMARIES NONE
  /IMPUTATIONSUMMARIES NONE
  /OUTFILE IMPUTATIONS=spssmi
```

多重代入済データセットを分析するには、まず、DATASET ACTIVATE として多重代入済データセット spssmi を起動する。DESCRIPTIVES VARIABLES=を用いて、基本統計量を出力したい変数を指定する。また、REGRESSION を用いて多重代入済データセットを用いた回帰分析を行う。

```
DATASET ACTIVATE spssmi.
DESCRIPTIVES VARIABLES=変数１
  /STATISTICS=MEAN STDDEV.

REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT turnover
  /METHOD=ENTER 変数２　変数３.

DATASET ACTIVATE データセット.
DATASET CLOSE spssmi.
```

3.2.2.3　SOLAS[14]

SOLAS は、欠測値の処理に特化したプログラムとして開発されたものである(Statistical Solutions, 2011)。なお、SOLAS の操作方法については、渡辺, 山口 (2000, pp.213-247)及び岩崎 (2002, pp.321-325)も参照されたい。SOLAS 4.01 における多重代入法は、以下の手順によって行うことができる。

Analyze メニューから、Multiple Imputation を選び、モデルとして Predictive Model Based Method を選択する。Base Setup タブにおいて、補定する変数を指定し、説明変数として使用する変数を指定する。また、補定済データセット数 M も、ここで指定する。Advanced Options タブを利用して、Randomization の Main Seed Value において、シード値を設定する。OK ボタンをクリックすれば、多重代入が開始される。なお、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)。

作成された M 個の補定済データセットを利用して、Descriptive Statistics のタブを利用して基本統計量を算出したり、Regression タブを利用して回帰分析を行うことができる。SOLAS においては、M 個の各々のデータセットを使用した分析結果と、それらを統合した結果の両方が自動的に出力され、実用的に便利である。

3.3 EMB アルゴリズム

最新のアルゴリズムとして、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(Honaker and King, 2010)。本項では、EMB のメカニズムを示し、それに基づくソフトウェアプログラムを紹介する。

3.3.1　EMB のメカニズム

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

[14] 本研究のために、SOLAS 4.01 を無償提供していただいた Statistical Solutions 社に感謝の意を表する。

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[15]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(3)を用いて欠測値の補定を行う(Congdon, 2006; Honaker and King, 2010)。MCMC や FCS とは異なり、ブートストラップ手法ではコレスキー分解[16]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。

3.3.2 EMB を用いたソフトウェアプログラム：R パッケージ Amelia

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 である(Honaker, King, and Blackwell, 2011)。Amelia は、ハーバード大学の Gary King (2001)を中心としたチームにより開発され、計算処理能力が高いと期待される多重代入法プログラムである[17]。Amelia II の詳しい使用方法については、高橋, 伊藤 (2013, pp.47-49) 及び Honaker, King, and Blackwell (2013)を参照されたい。

まず、Cran のウェブサイト[18]より Amelia 1.6.1.zip のファイルをローカルに保存し、R を起動させ、「パッケージ→ローカルにある zip ファイルからのパッケージのインストール」をクリックして、Amelia 1.6.1 をインストールする。その後、以下の要領でデータを読み込み、`library` 関数を用いて Amelia を起動させる。なお、インストールは初回のみ行い、次回以降はデータの読み込みと `library` 関数による Amelia の起動から作業を行えばよい。

```
setwd("D:/フォルダ名")
data<-read.csv("データ名.csv",header=T)
attach(data)
library(Amelia)
```

`set.seed` 関数を用いてシード値を設定し、`amelia` 関数を用いて多重代入を行う。`data` は多重代入を施すデータ名、`m=`の右辺は多重代入済データセット数である。多重代入の結果

---

[15] EMB アルゴリズムの詳細なメカニズムについては、高橋, 伊藤 (2013, pp.41-44)を参照されたい。

[16] 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、$A = HH'$に分解できることを意味する。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)。

[17] 2001 年に開発された当初の Amelia は、期待値最大化法(EM)に importance sampling (is)を応用した EMis を搭載していた(King *et al.*, 2001, pp.55-56)。2010 年には、さらなる効率化を目指し、EMB を搭載した Amelia II として生まれ変わった(Honaker and King, 2010, pp.564-565)。

[18] http://cran.r-project.org/web/packages/Amelia/index.html（2013 年 12 月 26 日アクセス）。2013 年 12 月 26 日現在における最新版は、Amelia 1.7.2.zip である。

は a.out に格納されており、write.amelia 関数を用いて csv ファイル（ファイル名：outdata）として出力することができる。また、Amelia で出力したデータの簡便な保存方法については、高橋, 伊藤 (2013, pp.82-83)を参照されたい。

```
set.seed(数字)
a.out<-amelia(data, m = 数字)
write.amelia(obj= a.out, file.stem = "outdata", orig.data = F,separate
= F)
```

多重代入済データセットを用いた統計分析を行うには、require 関数によりパッケージ Zelig を起動して以下のとおり使用すればよい (Imai, King, and Lau, 2008)。

```
require("Zelig")
z.out<-zelig(変数１~ 変数２+ 変数３, data = a.out$imputations, model = "ls",
cite = F)
summary(z.out)
```

3.4　まとめ

MCMC は、Rubin が構想したベイズ統計学に基づく多重代入法の理論に沿っており、正統的な系譜であると言えるが、多変量正規分布を仮定しており、また全変数の補定を一括して行わなければならないなど、制約も多い。一方、FCS は、必ずしも多変量正規分布を仮定しておらず、変数ごとに補定モデルを構築するため、非常にフレキシブルであるものの、Rubin の構想した多重代入法を具現化しているかどうかについて、シミュレーションによるサポートしか存在せず、理論的な根拠は必ずしも堅固ではない(van Buuren, 2012, p.249)。また、EMB は、ブートストラップによりコレスキー分解を回避しており、計算効率が高いと期待されるが、こちらも多変量正規分布を仮定しており、その前提が満たせない場合の性能には必ずしも保証はない。EM にブートストラップを応用している EMB アルゴリズムは、頻度論的な見地から十分な理論的根拠はあるものの、ベイズ統計学に基づく Rubin のオリジナルの多重代入法の観点からは議論があり得る[19]。

このように、3 つのアルゴリズムは、一長一短であり、必ずしもいずれのアルゴリズムが優れているかは、理論的に明白ではない。したがって、次節では、実データとシミュレーションデータを用いて、これら 3 つのアルゴリズムの性能を比較検証する。

[19] ただし、事後分布からの無作為抽出とブートストラップによる抽出は、漸近的に近似することが確認されている(Honaker and King, 2010, p.565)。ベイズ手法とリサンプリング手法についての議論は、Little and Rubin (2002, pp.89-90)も参照されたい。

4　多重代入法アルゴリズムの比較検証の結果

本節では、様々な文脈における多重代入アルゴリズム（及びそれらを用いたソフトウェア）の性能検証を行った。4.1 項では EDINET 情報[20]に基づくシミュレーションデータを用いて生成した大規模経済系データにおける多重代入の検証をした。4.2 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、大規模な経済の実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。

4.1 大規模シミュレーションデータの多重代入

4.1.1 基本統計量と欠測発生メカニズム

自然対数に変換[21]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 4.1 のとおりである。

表 4.1（基本統計量）

	第 1 四分位	中央値	平均値	第 3 四分位	標準偏差
売上高	8.998	10.110	10.110	11.230	1.656
資産	9.210	10.300	10.300	11.390	1.617
資本金	7.097	8.127	8.126	9.156	1.529
売上原価	8.533	9.746	9.747	10.960	1.800
事業従事者	4.221	5.053	5.054	5.888	1.237

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 として欠測を人工的に発生させた。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 万件；資産：5%＝5 万件；資本金：5%＝5 万件；売上原価：5%＝5 万件；事業従事者：1%＝1 万件。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%）。図 4.1 は欠測のない真の売上高のヒストグラムであり、図 4.2 は欠測値を含む売上高のヒストグラムである。

[20] 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://disclosure.edinet-fsa.go.jp/（2013 年 12 月 26 日アクセス）

[21] 自然対数に変換すると単位が解釈不能になるのではないかといった懸念があるが、実際に補定値を使用する際には、再変換を行って元の尺度に戻す必要がある。詳細については高橋, 伊藤 (2013, pp.80-81)を参照されたい。

図 4.1：売上高（真値）のヒストグラム

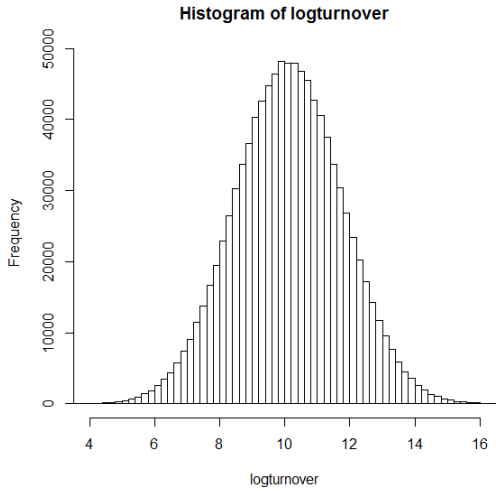
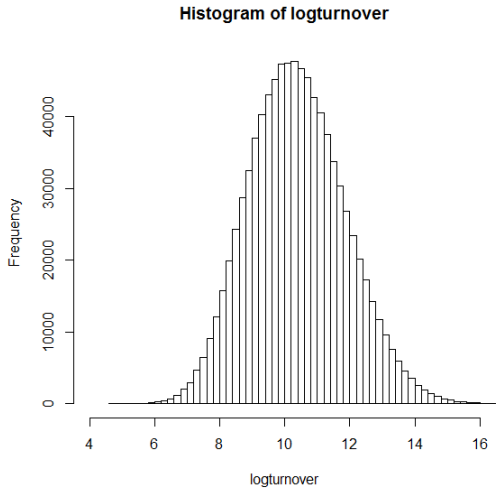


図 4.2：売上高（欠測あり）のヒストグラム



4.1.2 分析結果：シミュレーションデータ

それぞれの多重代入法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、多重代入済データセットを用いて、式(6)における$\hat{\alpha}$と$\hat{\beta}$の推定を行った。

$$\widehat{\text{売上高}}_i = \hat{\alpha} + \hat{\beta}\text{資本金}_i \tag{6}$$

結果は、表 4.2 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。Amelia、MICE、SAS、SPSS 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、欠測値を含むユニットを単純に除去するよりも、多重代入を行う方がよいことが分かる。

表 4.2：分析結果（シード値 1223、$M = 5$）

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

注：推定値は、Rubin の手法（式(1)と式(2)）により統合した。

Amelia、MICE、SAS、SPSS 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（4.2 項参照）。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ためである。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い最終的に分析することができなかった[22]。

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[23]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 4.3 として掲載している。図 4.3 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 4.3：売上高（縦軸）と資本金（横軸）の散布図

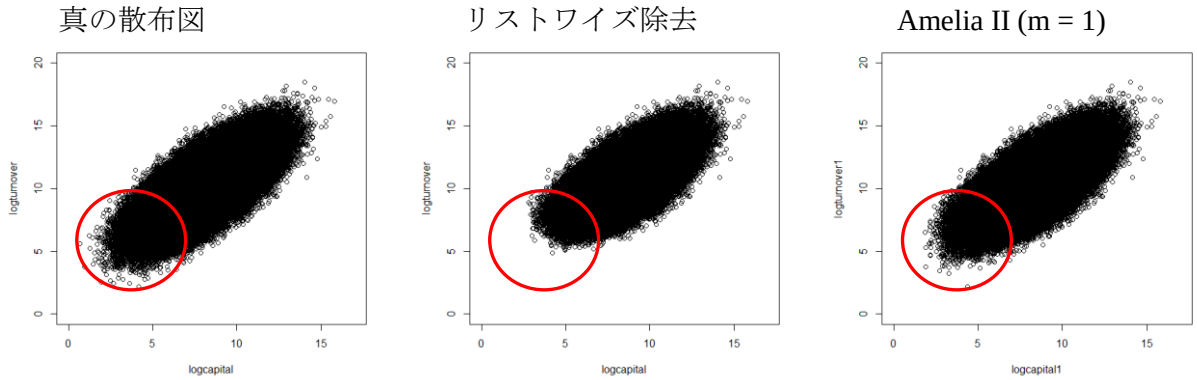


表 4.3 は、計算効率の検証を行った結果である[24]。前述したとおり、NORM と SOLAS では、大規模データセットを扱うことができなかった。Amelia と SAS は極めて速く大規模データセットを処理することができた。MICE と SPSS も、大規模データセットを扱うことはできるが、処理に多大な時間がかかった。今回は検証のために *M* を 5 に限ったが、実際には 20 ほどが推奨されるため（6 節参照）、MICE では最大 3 時間以上もの時間を要する可能性がある。

表 4.3（計算効率）

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

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。上記の結果には、データセットの読み込み時間やデータ分析の時間は含んでいない。

[22] 1 万×5 変量の小規模データセットに関して、Amelia は 5 秒、MICE は 34 秒、NORM は 36 秒、SOLAS は 3 分 14 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。

[23] 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。

[24] 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。

4.2 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

4.2.1 経済センサスとは

経済センサスは、事業所及び企業の経済活動の状態を明らかにし、我が国における包括的な産業構造を明らかにするとともに、事業所・企業を対象とする各種統計調査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 年に初めて実施された。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 年に初めて実施された[25]。

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により独自集計したものである。また、分析結果の評価等は、著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ことにも注意されたい。

4.2.2 基本統計量

分析には、2012 年 2 月に実施された経済センサス‐活動調査の速報データ（産業大分類 I）の単独事業所（個人経営以外）を用いた。なお、産業大分類 I とは、卸売・小売業のことを意味する。データセットの観測数は、277,263 である。データセットの基本統計量は、表 4.4 に示すとおりである。各々の変数において、平均値と中央値が大幅に異なっており、平均値から第 1 四分位までの距離と第 3 四分位までの距離も大幅に異なっている。このことから、いずれの変数も正規分布していないと考えることができる。

表 4.4（完全データの基本統計量、生データ）

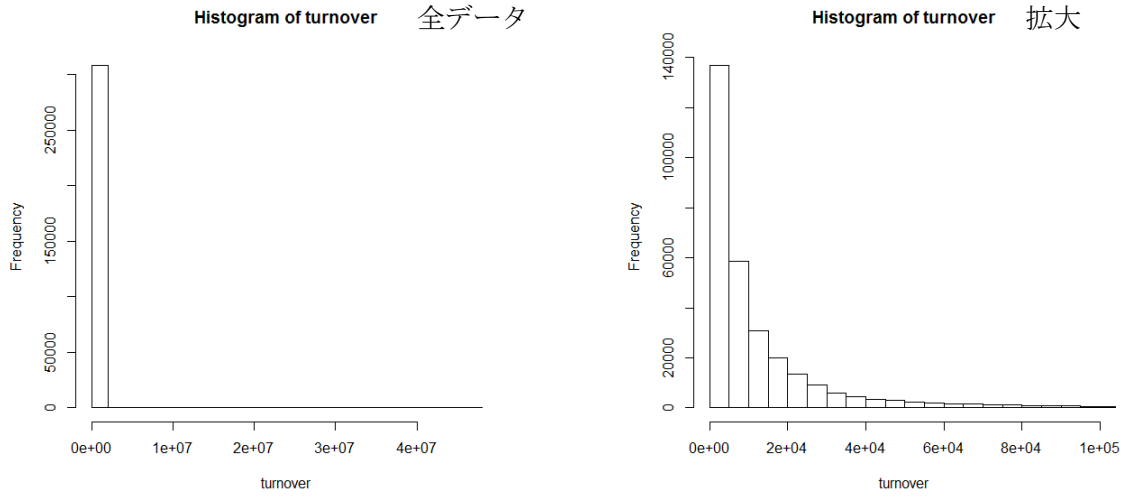
	第 1 四分位	中央値	平均値	第 3 四分位	標準偏差
売上高	2400	6200	25068	16157	256606
資本金	300	500	1334	1000	92118
事業従事者	2	4	7	8	22

注：数値は小数点以下を四捨五入している。売上高と資本金の単位は百万円である。事業従事者の単位は人である。

売上高データの分布は、図 4.4 のヒストグラムに示すとおり、経済データ特有の歪みを有しており正規分布していないことが視覚的に分かる。そこで、分布の歪みを矯正するために、自然対数に変換して分析を行うこととする。

[25] http://www.stat.go.jp/data/e-census/guide/about/purpose.htm（2013 年 12 月 26 日アクセス）

図 4.4：売上高のヒストグラム



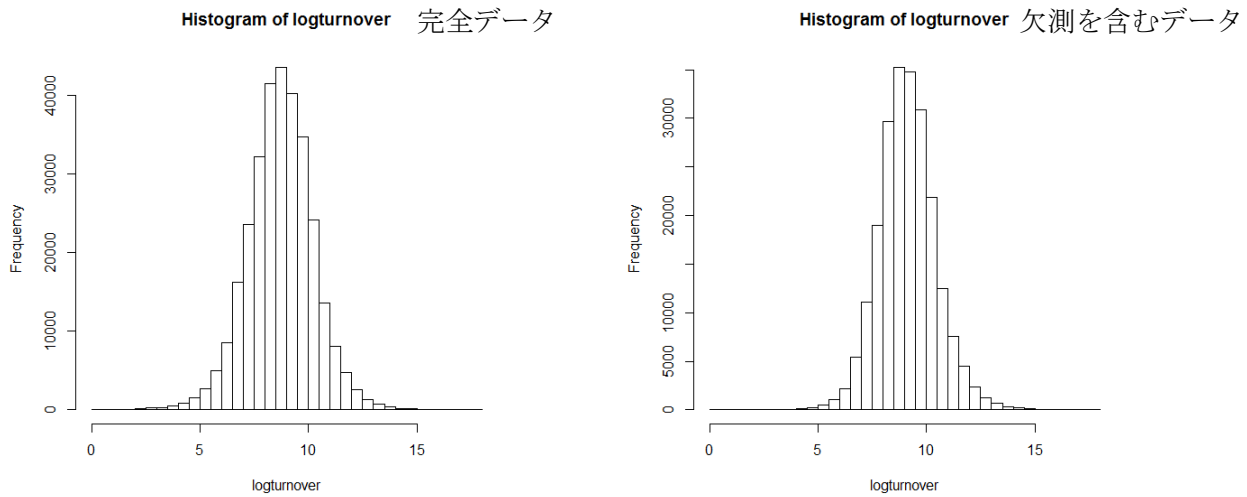
自然対数に変換したデータセットの基本統計量は、表 4.5 に示すとおりである。各々の変数において、平均値と中央値がほぼ同一となり、平均値から第 1 四分位までの距離と第 3 四分位までの距離もほぼ同一である。

表 4.5（完全データの基本統計量、自然対数）

	第 1 四分位	中央値	平均値	第 3 四分位	標準偏差
売上高	7.783	8.732	8.720	9.690	1.527
資本金	5.700	6.210	6.320	6.910	0.909
事業従事者	0.693	1.386	1.497	2.079	0.895

欠測の発生メカニズムは、MAR に基づき、売上高（自然対数）データの 20%（55,500 個）を人工的に欠測させた。また、資本金（自然対数）データの 5%（13,600 個）を無作為に人工的に欠測させ（MCAR）、事業従事者（自然対数）には欠測を発生させていない（欠測率 0%）。売上高（自然対数）の分布は、図 4.5 に示すとおり、近似的に正規分布していると言える。

図 4.5：売上高（自然対数）のヒストグラム



4.2.3 分析結果：経済センサス - 活動調査の速報データ

分析結果（100 個のシードの平均）は、表 4.6 に示すとおりである。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。リストワイズは、リストワイズ除去法を用いた分析結果である。Amelia、MICE、SAS、SOLAS、SPSS では、すべての出力結果（売上高の平均値、売上高の標準偏差、回帰係数と *t* 値[26]）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、多重代入を行う方がよいことが分かる。Norm では、27 万×3 変量のデータセットを回すことができなかった。

表 4.6（多重代入法結果：$M = 5$）

	平均値	標準偏差	傾きの係数	傾きの *t* 値
真値	8.7636	1.5099	1.2075	534.2876
リストワイズ	9.1326	1.3330	1.1431	408.1007
AMELIA	8.7820	1.4597	1.1818	428.9757
MICE	8.7819	1.4598	1.1820	420.2365
NORM	NA	NA	NA	NA
SAS	8.7819	1.4598	1.1819	421.9047
SOLAS	8.7810	1.4605	1.1830	443.6289
SPSS	8.7818	1.4599	1.1820	414.1378

注：推定値は、Rubin の手法（式(1)と式(2)）により統合した。

表 4.7～4.10 は、表 4.6 における 100 個のシードの結果について、Welch の二標本の平均に関する *t* 検定を用い、各ソフトウェア間の優劣を比較した結果（*p* 値）である。概ね、SOLAS と他のソフトウェアの間に有意な差が見られた。なお、多重性の問題を考慮する必要があり、Bonferrorni の修正を用いた[27]。

表 4.7 は、平均値の比較結果である。SOLAS と他のソフトウェアの間で、95%水準で有意な差が見られ、SOLAS が優れていた($p = 0.0000$)[28]。Amelia と SPSS の間にも、95%水準で有意な差があるものの、Bonferrorni の修正を用いた場合、有意差は消失する結果となった($p = 0.0396$)。表 4.8 は、標準偏差の比較結果である。SOLAS と他のソフトウェアの間で、95%水準で有意な差が見られ、SOLAS が優れていた($p = 0.0000$)。

表 4.7（平均値を比較した *p* 値）

	AMELIA	MICE	SAS	SOLAS	SPSS
AMELIA					
MICE	0.1273				
SAS	0.4669	0.4133			
SOLAS	0.0000	0.0000	0.0000		
SPSS	0.0396	0.5761	0.1705	0.0000	

表 4.8（標準偏差を比較した *p* 値）

	AMELIA	MICE	SAS	SOLAS	SPSS
AMELIA					
MICE	0.3161				
SAS	0.4296	0.8101			
SOLAS	0.0000	0.0000	0.0000		
SPSS	0.0588	0.3834	0.2558	0.0000	

[26] 回帰係数は$\log\left(\widehat{\text{売上高}}_i\right) = \hat{\alpha} + \hat{\beta}\log(\text{事業従事者}_i)$の$\hat{\beta}$であり、*t* 値は$\hat{\beta}$の *t* 値である。

[27] Bonferroni の修正とは、有意水準調整型の多重比較法であり、「有意水準を比較する組合せの数で割る」という方法である(栗原, 2011, p.155)。今回は、検定の組合せ数が 10 なので、Bonferroni の修正後の 95%有意水準では、*p* 値が 0.005 未満の場合に有意と見なせる。

[28] なお、*p* 値は小数点第 4 位で四捨五入している。

表 4.9 は、回帰係数（傾き）の比較結果である。SOLAS と他のソフトウェアの間で、95%水準で有意な差が見られ、SOLAS が優れていた($p = 0.0000$)。表 4.10 は、回帰係数の t 値の比較結果である。SOLAS と MICE、SAS、SPSS の間で、95%水準で有意な差が見られ、SOLAS が優れていた($p = 0.0000$)。Amelia と SOLAS の間には、95%水準で有意な差が見られなかった($p = 0.0612$)。

表 4.9（傾きの係数を比較した p 値）

	AMELIA	MICE	SAS	SOLAS	SPSS
AMELIA					
MICE	0.3369				
SAS	0.7519	0.5114			
SOLAS	0.0000	0.0000	0.0000		
SPSS	0.1007	0.4887	0.1773	0.0000	

表 4.10（傾きの t 値を比較した p 値）

	AMELIA	MICE	SAS	SOLAS	SPSS
AMELIA					
MICE	0.2892				
SAS	0.3862	0.8355			
SOLAS	0.0612	0.0029	0.0051		
SPSS	0.1093	0.5058	0.3925	0.0010	

表 4.11 は、計算効率の検証を行った結果である[29]。Amelia と SAS の処理速度は極めて速かった。MICE と SPSS の処理時間は、Amelia と SAS の数倍かかった。SOLAS は、Amelia と SAS の 15 倍以上の時間を要した。SOLAS は、100 万×5 変量データセットの処理は行えなかったが、27 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 4.11（計算効率）

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

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。上記の結果には、データセットの読み込み時間やデータ分析の時間は含んでいない。

以上のとおり、SOLAS と他のプログラムとの間に統計上の差が認められるものの、その差は微小であり、実質的な差はほとんどないと言える。

4.3 まとめ

本節では、様々な多重代入法アルゴリズムに基づくソフトウェア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 SOLAS が優位であった。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。Amelia と SAS 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。

[29] 各 PC のスペックは、表 4.3 と同じである。

5　経済センサス - 活動調査の速報データを用いた多重代入のデモンストレーション

前節までに検証したとおり、補定値の精度に関しては、各アルゴリズム間に大きな差はなく、計算効率の点で EMB を搭載した Amelia II に軍配が上がった。そこで、本節では、経済センサス - 活動調査の産業大分類 D（建設業、データサイズ 255,587）の単独事業所（個人経営以外）のデータ（自然対数変換）を実際に Amelia II の多重代入（M = 5）によって補定してみることとする[30]。対象とする変数は、売上高、資本金、事業従事者の 3 つであり、主に売上高の欠測値補定のあり方について、基本統計量（中央値、標準偏差、合計値）、処理時間、補定の診断結果を見ていくこととする。

5.1 欠測値補定

欠測値補定を行わずに算出した基本統計量（自然対数）は、表 5.1 のとおりである。中央値は 8.6651 であり、標準偏差は 1.2760 である。売上高の合計値は、2,090,959 であった。

表 5.1（補定前、自然対数）

	中央値	標準偏差	合計値
生データ	8.6651	1.2760	**2090959**

Amelia II によって多重代入を行ったデータセットをもとに算出した基本統計量（自然対数）は、表 5.2 のとおりである。多重代入(M = 5)を行うのに要した時間は 50 秒であった[31]。

表 5.2（補定後、自然対数）

	中央値	標準偏差	合計値
補定 1	8.6385	1.2782	2199977
補定 2	8.6383	1.2777	2199927
補定 3	8.6387	1.2782	2200127
補定 4	8.6391	1.2776	2200072
補定 5	8.6385	1.2779	2200005
統合		1.2778	**2200022**

以上のとおり、欠測値を補定せずに算出した売上高の合計値は 2,090,959 であったが、補定をした結果、売上高の合計値（自然対数）は 2,200,022 と推定される[32]。欠測数がわずかであるため、表 5.1 と表 5.2 には劇的な差はないが、中央値、標準偏差、合計値に影響が見られる。

---

[30] なお、海外の公的統計において多重代入法を活用している事例として、アメリカ全国保健統計センター (NCHS: National Center for Health Statistics)の国民健康調査 (NHIS: National Health Interview Survey)が挙げられる。http://www.cdc.gov/nchs/nhis/2010imputedincome.htm（2013 年 12 月 26 日アクセス）

[31] 検証に用いた PC の性能：Windows 8 を搭載したデスクパソコンであり、プロセッサは Intel Core 2 Duo CPU E8600、メモリ(RAM)は 2.00 GB、システムの種類は 64 ビットオペレーティングシステムである。

[32] 実際には、補定値を自然対数から生データの尺度に戻す必要があり、その際に補正項を追加する必要がある。詳しくは、高橋, 伊藤 (2013, pp.80-81)を参照されたい。

5.2 診断結果

実務において欠測値補定を行う際には、真値が観測されていないため、補定値の精度を直接的に検証することができない。そこで、近年では、補定値の間接的な検証方法が提唱されている(Abayomi, Gelman, and Levy, 2008; Honaker, King, and Blackwell, 2011)。これらの診断手法についての詳細は、高橋, 伊藤 (2013, pp.64-70)を参照されたい。

密度の比較(Comparing Densities)とは、観測値と補定値との密度の比較のことである。もし 2 つの密度が大幅に分離していたり、形状が異常である場合には、補定モデルの妥当性が疑われる。図 5.1 は密度の比較の結果である。赤線は補定値の密度を表しており、黒線は観測値の密度を表している。2 つの密度の重なり具合は完全ではないものの、概ね同じ値を中心に近似的に正規分布しており、問題はないと合理的に判断できる。密度の比較を行うのに要した時間は 1 秒であった。

欠測地図(Missingness Map)とは、データセット内の欠測パターンを視覚化する手法である。図 5.2 は事業従事者を昇順に並べ替えた欠測地図の結果である。ここから、事業従事者が少ないほど売上高に欠測が発生しやすいことが視覚的に分かり、欠測のメカニズムは MAR だと推定できる。欠測地図を作成するのに要した時間は 2 秒であった。

図 5.1：密度の比較

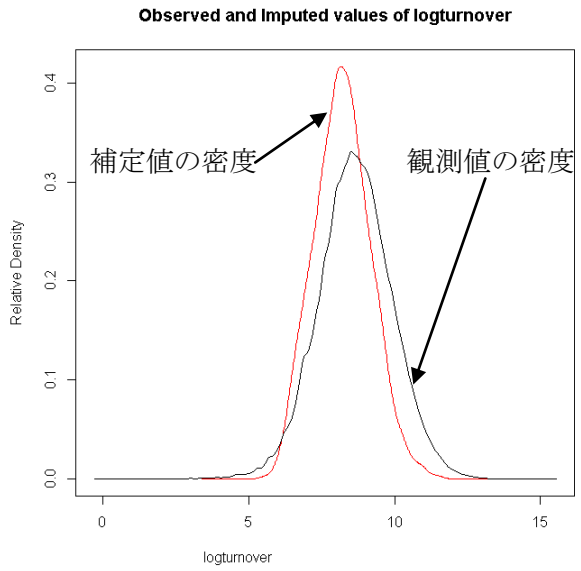
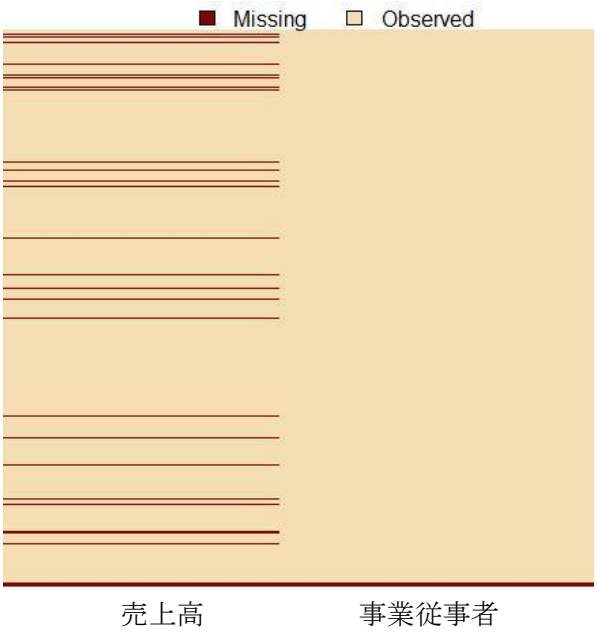


図 5.2：欠測地図（事業従事者の昇順）



過剰補定(Overimpute)とは、各々の観測値を人工的に 1 つずつ欠測させ、補定モデルを当てはめて数百回の補定を行い、90%信頼区間を図示する手法である。y = x 線が信頼区間に含まれていれば、補定モデルの信頼性が高いと言える。図 5.3 は過剰補定の結果である。y = x 線は概ね 90%信頼区間に含まれており、補定モデルの信頼性に問題はないと考えられる。過剰補定は、各々の観測値 1 つ 1 つに対して数百回の補定を行うため、要した時間は 21 分 20 秒であり、やや時間がかかった。

過散布初期値(Overdispersed Starting Values)とは、複数の EM アルゴリズムの初期値を設定し、同一の値に収束するかどうかを図示する手法である。EM アルゴリズムでは、局所的最大値が大局的最大値であるとは限らないため、複数の初期値が同一の値に収束するかどうかによって妥当性を検証する必要がある。図 5.4 は過散布初期値の結果であり、20 個の初期値すべてが同一の値に収束しており、EM アルゴリズムの収束に問題はなかったと結論付けられる。過散布初期値の計算を行うのに要した時間は 1 分 28 秒であった。

図 5.3：過剰補定

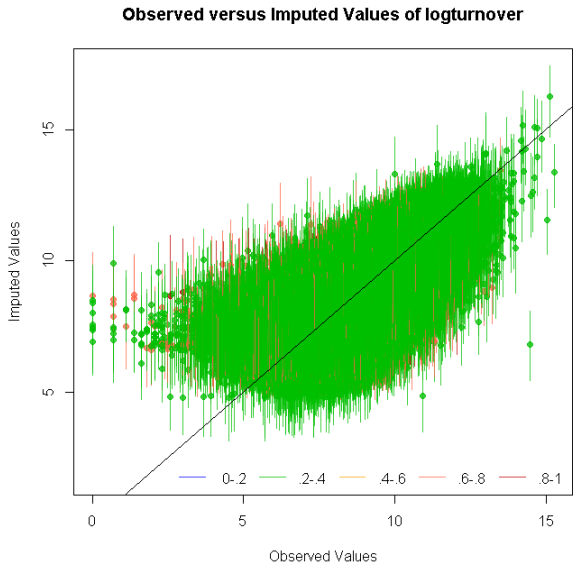
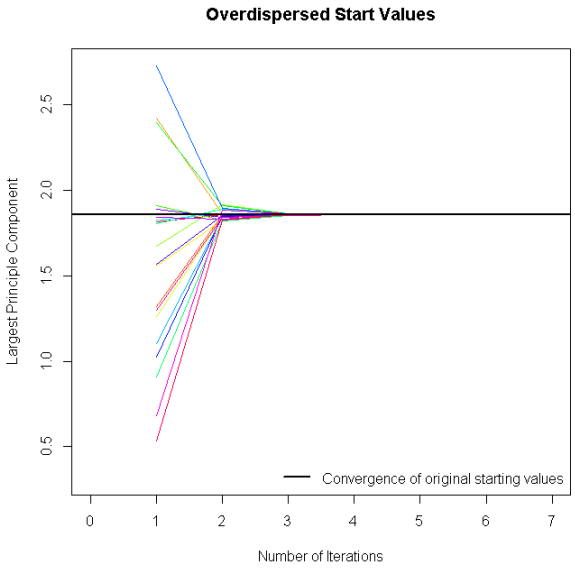


図 5.4：過散布初期値



5.3 まとめ

本節では、Amelia II による多重代入法の実務への適用可能性を探った。欠測値によるバイアスを是正し、診断手法によって補定モデルの妥当性を検証することで、公的経済統計における欠測値補定の一助を成すことができることを示した。

6　多重代入法の M 数

2 節では、紙面の都合上、$M = 5$ として例を示した。しかし、実際には M をいくつに設定すればよいのだろうか？　一般的に、シミュレーションでは、数百以上の副標本($M > 100$)を生成する必要があり、コンピュータの能力が許す限り多くの繰り返しを行うべきだと考えられるが、元来、Rubin (1987)によると、多重代入法の M は非常に小さい数字で十分だとされている。一方、近年では、6.2 節に示すとおり、M 数に関して Rubin (1987)への反論が展開されているものの、十分な結論を得るにいたっていない。大規模データセットの多重代入という文脈では、M のサイズに応じてコンピュータの限界処理能力に達してしまう可能性がある[33]。よって、本節では、多重代入データセット数 M の適切なサイズについて検証を行った。

6.1　M 数に関する議論：Rubin による相対効率の定義

Rubin (1987, p.114)によると、無限の M の代わりに有限の M を使用した場合の漸近的相対効率(ARE: Asymptotic Relative Efficiency)は、式(7)のとおり定義されている。ここで、δは欠測率を表している($0 \leq \delta \leq 1$)。ARE は%であり、単位は標準偏差である。M が無限大の場合、式(7)の極限値は 100%となり、効率性が最大に達していることになる。

$$ARE = \left(\sqrt{1 + \frac{\delta}{M}}\right)^{-1} \times 100 \qquad (7)$$

表 6.1 は、欠測率 10%($\delta = 0.1$)から 90%($\delta = 0.9$)までのデータにおいて、M を増加させた場合に、無限大の M と比較した効率性の結果を表している。この結果から、M を 5 に設定することで、欠測率が 30%あったとしても、97.13%の相対効率を達成できており、M を 10 に設定することで、仮に欠測率が 50%であったとしても、相対効率は 97.59%を達成しているとされる。

表 6.1（M と相対効率）

M	$\delta = 0.1$	$\delta = 0.2$	$\delta = 0.3$	$\delta = 0.4$	$\delta = 0.5$	$\delta = 0.6$	$\delta = 0.7$	$\delta = 0.8$	$\delta = 0.9$
1	95.35	91.29	87.71	84.52	81.65	79.06	76.70	74.54	72.55
5	99.01	98.06	97.13	96.23	95.35	94.49	93.66	92.85	92.06
10	99.50	99.01	98.53	98.06	97.59	97.13	96.67	96.23	95.78
∞	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

式(7)こそが、多くの文献(von Hippel, 2005; Congdon, 2006; Honaker and King, 2010; de Waal *et al.*, 2011; van Buuren, 2012)において、M は 5〜10 程度でよいとされる根拠なのである。

[33] たとえば、100 万観測数、10 変数のデータセットにおいて、M を 1000 に設定した場合、観測数と変数の数は、合計で 100 億に達してしまい、通常の PC では処理が困難だと考えられる。

6.2　*M* 数に関する議論：Rubin への反論

近年、コンピュータの処理速度が向上するにつれて、5～10 といった従来の *M* 数ではなく、できる限り多くの数を使用することが望ましいという提唱がされるようになってきた。Hershberger and Fisher (2003)では、単純無作為抽出の理論に基づき、*M* 数を推定すべき要因と考え、数百の *M* が要請されると結論付けた。Carpenter and Kenward (2007)及び野間, 田中 (2012)においても、*M* のサイズは、数十～数百程度が望ましいとされている。Bodner (2008)は、必要な *M* のサイズは欠測率と有意水準に応じて変更することを示した。たとえば、95%の有意水準において、欠測率 10%ならば *M* は 6 で十分であるが、欠測率が 30%であれば必要な *M* 数は 24 となり、欠測率が 70%を超えると必要な *M* 数は 114 となる。

6.3 シミュレーションデータを用いた多重代入データセット数 *M* の検証

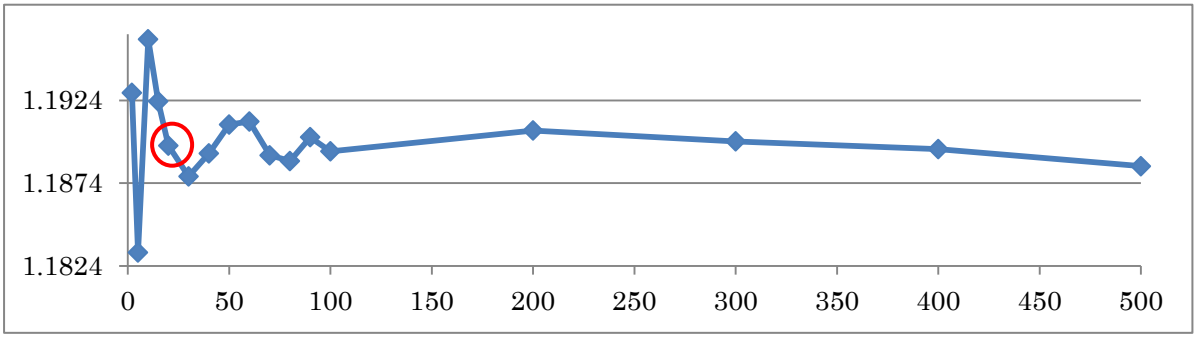
自然対数変換した経済センサス - 活動調査の速報データの情報（平均値、分散・共分散など）を基に、多変量正規分布によって観測数 1000、3 変量のシミュレーションデータセットを生成した。シミュレーションデータの基になったデータは、4.2 項と同じく、産業大分類 I の単独事業所（個人経営以外）を用いた。データセットの基本統計量は、表 6.2 のとおりである。Amelia、MICE、Norm において、表 6.2 のシミュレーションデータセット（欠測率 = 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%）に多重代入（*M* = 2, 5, 10, 15, 20, 30, 40, 50, 100, 200, 250, 300, 400, 500）を施し、多重代入済データセットを用いて、$\log\left(\widehat{\text{売上高}}_i\right) = \hat{\alpha} + \hat{\beta}\log(\text{事業従事者}_i)$ における$\hat{\beta}$とその標準誤差の推定を行った。

表 6.2（基本統計量）

	第 1 四分位	中央値	平均値	第 3 四分位	標準偏差
売上高	7.761	8.837	8.764	9.752	1.510
事業従事者	0.946	1.528	1.514	2.077	0.891
資本金	5.727	6.331	6.323	6.921	0.907

図 6.1 は Amelia による結果を図示しており、欠測率 20%の場合、*M* が 20 を超えると、係数の推定値はほぼ安定し始めていることが分かる。

図 6.1：欠測率 20%、*M* = 2～500（縦軸は係数の推定値、横軸は M の数）



また、欠測率を 50%に固定し、1000 個のシードを用いて上記の作業を繰り返して得られた係数の分布は、図 6.2 のとおりである。図 6.2 では、横軸に M 数を取り、縦軸に係数を取っている。M 数を増やせば増やすほど、箱ひげ図が小さくなっていき、推定値のばらつきが抑えられることが分かる。M 数が 10 以下の場合、最大値と最小値が、それぞれ、過大または過小になる可能性があり、得られた結果が偶発的に不正確になるおそれがある。一方、M = 50 における Amelia の箱ひげ図では、95%信頼区間が(1.188, 1.220)となり、その範囲は、わずか 0.032 となっている。Mice と Norm においても、ほぼ同様の状態となっている（なお、Norm では、M = 500 の分析は処理できなかった）。

図 6.2：係数の分布（欠測率 50%、シード数 1000、欠測率 50%）

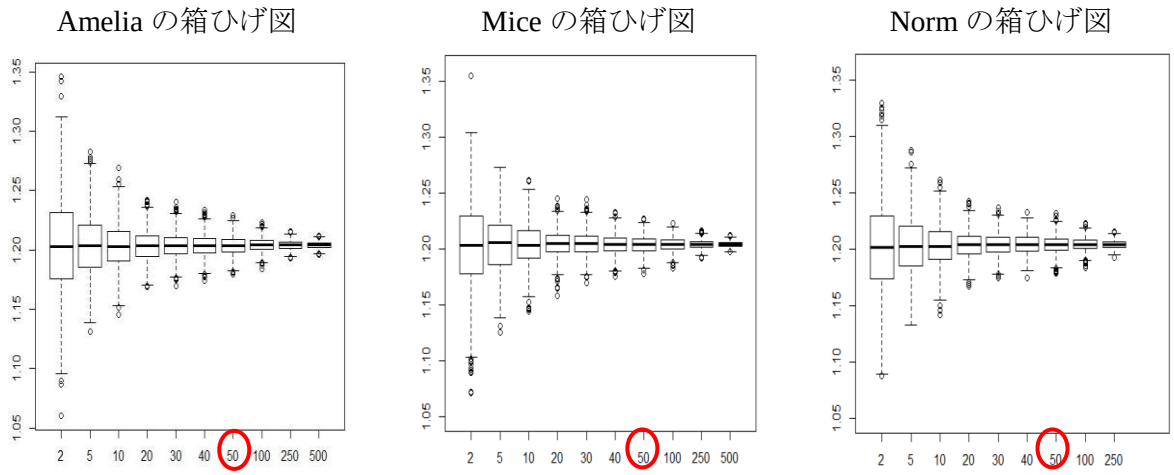


図 6.3 は、図 6.2 における 95%の信頼区間(C.I.)の長さを図示したものであり、M = 50 を超えると、95%の信頼区間の長さはほぼ一定になっており、これを超えて得られる相対効率は非常に低いことが分かる。また、3 つのアルゴリズム間において特に差異は見受けられない。

図 6.3：95%信頼区間の幅（欠測率 50%）

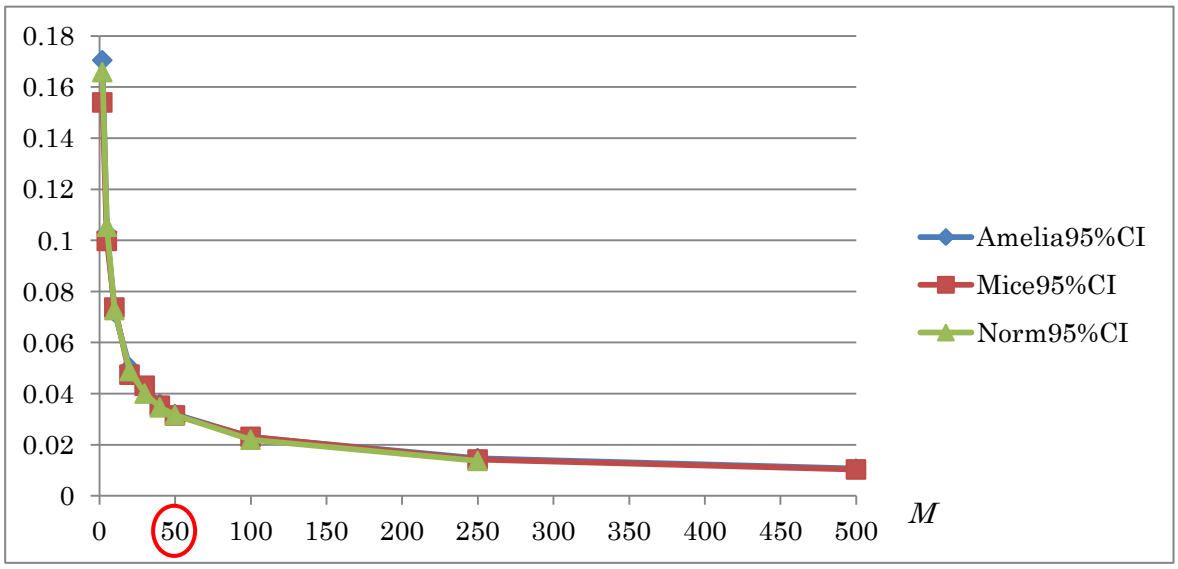
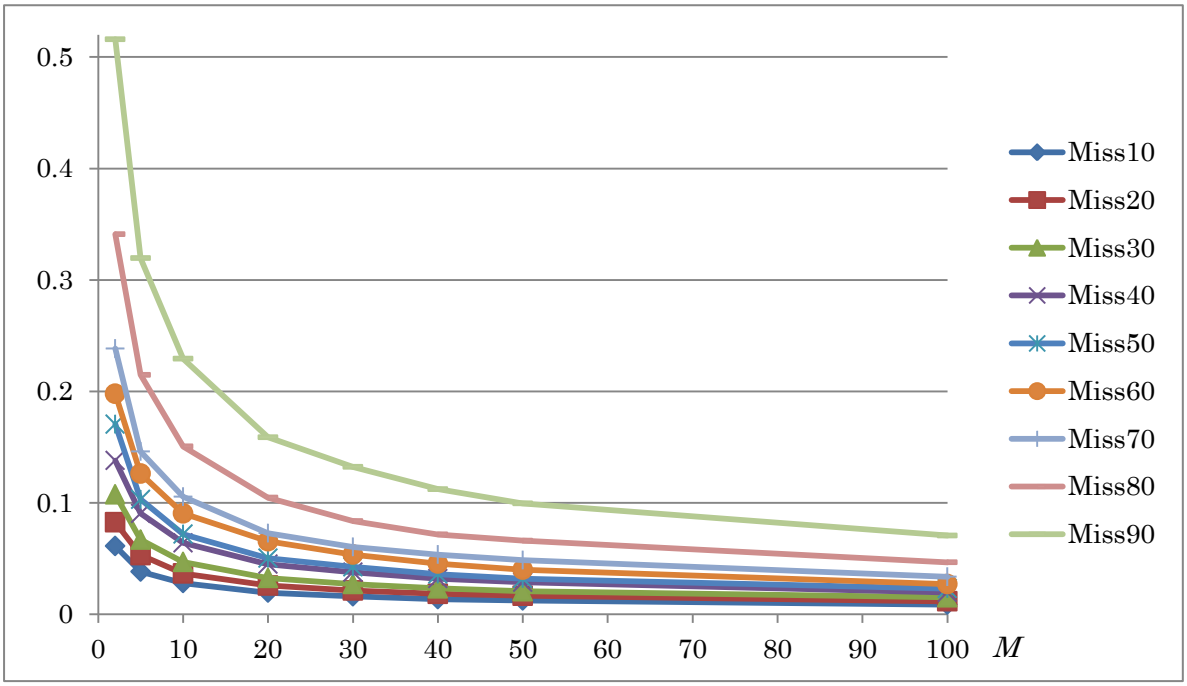


図 6.4 は、Amelia において、欠測率を 10%～90%まで変化させた場合、*M* 数に応じた係数の信頼区間の幅の変化を図示したものである。これによると、欠測率が上がるにつれて、信頼区間の幅が長くなり、それに応じて *M* 数を大きくするべきであることが分かる。

図 6.4：係数の 95%信頼区間の幅（Amelia）



6.4 まとめ

本節では、多重代入法の擬似データ数 *M* の適切なサイズについて検証を行った。概ね 5～10 では少なすぎ、20～50 程度が適切だと考えられる。欠測率に応じて、20%未満ならば *M* = 20、20%～30%ならば *M* = 30、30%～40%ならば *M* = 40、40%～50%ならば *M* = 50 といった具合に設定することが適切であろう。欠測率に関わらず、*M* = 100 を超えて得られるものは非常に少ない。また、標本サイズや欠測パターンにも依存するが、欠測率が 50%を超え始めると、たとえ *M* 数を数百まで拡大したとしても、補定値の精度を保証できなくなるおそれがある。

このように、*M* のサイズは数十程度が最適[34]だと思われるが、通常のシミュレーションと多重代入法では、主に、欠測情報の量に大きな違いがある。通常のシミュレーションでは、全データをシミュレーション値として生成するため、全情報が欠測していると言える。補定においては観測値をシミュレーション値に置き換える必要はなく、データ内の一部のみが欠測している。結果、繰り返し回数が少なくてもよいと考えられる(Honaker and King, 2010)。

[34] 精度が十分にあり、計算負荷が大きくないという意味である。

7　結語

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、わずかながらに SOLAS が優れていたものの、概ね、アルゴリズム間に決定的な差はなかったと言える。一方、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。Amelia と SAS 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。Norm は 27 万×3 変量のデータセットを分析することができず、大規模データセットの多重代入には向いていないことが分かった。データ数が 1 万に満たない小規模なデータセットの多重代入には、既存のアルゴリズムのいずれを用いても問題はないと考えられるが、数十万以上の観測値を持つ大規模なデータセットの多重代入には、Amelia または SAS が有用であると結論付けられる。

基盤としての R は、大規模データセットに向いていないことを鑑みれば、アルゴリズムとしての EMB の計算効率の高さが伺える。一方、SAS は、基盤として大規模なデータセットの処理を得意としている。SAS において EMB を実装すれば、より大きなデータセットをより速く処理できるようになると期待される。

また、多重代入擬似データ数 M については、概ね 5〜10 では少なすぎることが検証の結果示された。通常のシミュレーションと同様に、M のサイズは大きければ大きいほどよいが、50 を超えて得られる相対効率は極めて小さいことも分かった。実務においては、欠測率に応じて、M を設定することが適切である。

補論 1：欠測値補定に関する最新の研究動向

あらゆる実データにおいて、必ずと言っていいほど、欠測値は氾濫している。したがって、学会においても、補定に関する研究論文が盛んに公開されている。本節では、2013年のISI (International Statistical Institute)世界統計大会及び統計関連学会連合大会において発表された論文の中から、欠測値補定に関する最先端の研究論文4篇を簡潔に紹介する。

Handling Nonignorable Nonresponse Using Generalized Calibration with Latent Variables (Ranalli, Matei, and Neri, 2013)

概要：本報告では、対象変数の有限母集団合計値や平均値の推定が関心事である場合におけるユニット非回答の対処法を取り扱う。キャリブレーション(calibration：調整、校正)とは、推定段階において、補助情報を含めることによりユニット非回答に対処する汎用的な手法である。本報告で提示した汎用キャリブレーション手法は、通常の重み付け手法とは異なり、非回答の主原因となっている変数が回答者に関してのみ既知である場合であっても、非回答によるバイアスを補正することができる。非回答が無視できない(nonignorable)場合には、この特性はとりわけ有用である。実際に、この種の非回答を補正するために、回答者に関してのみ既知である対象変数を操作変数として利用することができる。潜在変数モデルを用いることで、顕在変数から構成概念を抽出することができ、この抽出した構成概念を汎用キャリブレーション手法における操作変数として使用する。本研究で提案している手法をシミュレーションデータと Italian Survey of Households' Income and Wealth データを利用して検証した。

所感：欠測値は、ユニット非回答と項目非回答の 2 種類に大別される。項目非回答に関する研究は盛んに行われているが、ユニット非回答の研究は稀である。また、1.2 項で議論したとおり、欠測には無視できる欠測(ignorable)と無視できない欠測(nonignorable)とがあり、後者への対処法も確立されてはいない。本研究は、このように二重の意味で興味深い。

Fractional Hot Deck Imputation for Multivariate Missing Data (Kim and Fuller, 2013)

概要：ホットデック補定は、標本調査における項目非回答の対処法として非常によく使用されているものであり、分数ホットデック補定(fractional hot deck imputation)は、ホットデック補定を効率的に行うために考案された手法である。しかし、任意の欠測パターンにおける多変量欠測データへのホットデック手法の応用は、非常に難しいものとして知られている。補定済データセット内の共分散構造を保持することが難しいからである。今回の報告では、分数ホットデック補定を多変量欠測データに拡張する。分数ホットデック補定では、対象となる項目の同時分布を、離散近似によってノンパラメトリックに推定する。離散へと変換することは、補定のセルを作成する役割を果たす。分数補定では、最初に、欠測項目のセルの補定を行い、その後、各々の補定セル内の実測値の補定を行う。キャリブレーションによる重み付けによって、補定分散を減少させる。シミュレーションデータを用い、この手法を検証した。

所感：欠測値補定の文脈において、ポスト多重代入法となる可能性があり、極めて重要になると思われる先端的研究である。今後の進展に注視するとともに、実務への適用可能性を検討したいと考えている。

Balanced *k*-nearest Neighbor Imputation (Hasler and Tille, 2013)

概要：ランダム補定は、確率的補定や撹乱的補定とも呼ばれ、補定値の分布を維持しやすいため、項目非回答への対処法として頻繁に使われる。ランダム補定の手法の中でも、ランダムホットデックには、補定値は実際の観測値であるという重要な特性がある。本研究で提案したランダムホットデック補定の新手法は、ランダムではあるものの、他のランダム補定手法と比較して、補定分散を減少させることができるという特性がある。この手法を安定的な k 最近隣補定法と名づける。この手法では、まず、レシピエント（補定されるべき欠測値）の近隣からドナー（補定値を提供する観測値）を選ぶ。各々の非回答者に関して、k 個の近隣値からランダムにドナーを選ぶ。次に、補定プロセスにおいて、補助変数の合計推定値を保持する。

所感：項目非回答への対処法として頻繁に使われるホットデック補定を発展させ、補定値の分布を維持しながら、補定分散を最小化することに成功しており興味深い。

Missing Data Analysis with Mixture Missing Mechanisms (森川, 山本, 狩野, 2013)

概要：従来、欠測値を含むデータ解析としての尤度解析では、観測変数に加え、変数の値が欠測しているとき R = 0 とし、変数の値が観測されているとき R = 1 とする二値の欠測指示行列を用いて、観測値と欠測指示行列の同時分布をモデリングしてきた。しかし、実際の場面では、複数の欠測メカニズムが混合していることが普通である。従来、そのような欠測原因については、不明としたり、もしくは把握できていたとしても、適切な解析方法がないために無視し、あたかも欠測原因は 1 つであるかのように扱って解析してきた。本報告では、複数の欠測メカニズムが混合する状況についての理論的枠組みを構築し、統計的推測法を与えた。複数の欠測原因が混合しているモデルの尤度は、単一原因のときの尤度と同じ形になるため、単一原因の場合の議論を拡張することにより、最尤推定量の強一致性等の様々な有用な結果を得られると期待される。

所感：複数の欠測メカニズムが混合している場合の欠測値対処法は確立されておらず、そういった意味で非常に意義深く、今後の展開に期待したい。

補論 2：ベイズ統計学概論

ハーバード大学統計学科の Rubin (1987)により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ていた。現在は、様々な派生系の多重代入法アルゴリズムが並存しているが、いずれも、ベイズ統計学の精神を引き継いでいるため、本節において、参考までにベイズ統計学の概略に触れる。

A2.1　確率の解釈：頻度論と主観論

確率の解釈方法として、頻度的解釈と主観的解釈の 2 種類がある。事象 A は事象 S に含まれる事象であり、事象 S が起きる回数を N、事象 A が起きる回数を n と表した場合、頻度論的な確率の解釈では、A の発生確率$P(A)$は式(8)として定義できる。

$$P(A) = \lim_{N\to\infty} \frac{n}{N} \tag{8}$$

つまり、頻度論における確率とは、長期的に繰り返し行われる非決定論的な結果の性質として、極限における相対頻度と解釈される。したがって、頻度論的な確率の解釈は、何度も繰り返し試行できる場合には直感的に分かりやすくてよいが、繰り返し起こり得ない一度限りの事象を分析するには不適当である。一方、主観的確率とは、信念の度合いとも呼ばれ、確率は、様々な状況下に応じて個人的に定義されるものとされる。つまり、主観的解釈における$P(A)$とは、事象 A が真であるという信念の度合いを表しており、$P(A)$が 1 に近づくほど信念の度合いが強くなり、0 に近づくほど信念の度合いが低いのである(矢野, 2012)。ベイズ統計学は、この主観的確率に基づいて構成されており、観測できないデータと未知のパラメータとの区別をせず、即座に利用可能な数値（目に見えるもの）と確率的に記述されるべき数値（信念：主観的確率）から世界が構成されていると考える(Gill, 2008)。

A2.2　確率の更新と条件付き確率

新しい情報を入手する度に、確率を更新することこそが、ベイズ統計学の基本的なメカニズムであり、データから学んで信念を更新するプロセスを定式化していると言える。そのために、ベイズ統計学では、「条件付け」という概念が重要な役割を果たす。もし事象 B が発生するかしないかによって、事象 A の発生確率が影響を受けるとしたら、A は B を条件としていると言える。この場合、式(9)による条件付き確率として定式化される。つまり、B を条件とした A の発生確率$P(A|B)$は、A と B が同時に起きる確率$P(A \cap B)$を B の発生確率$P(B)$で割ったものである。

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{9}$$

A2.3　ベイズの定理

ベイズの定理は、以下のとおり、条件付き確率の式から導出できる(Maddala, 2001; DeGroot and Schervish, 2002)。まず、式(10)を式(11)に変換し、式(11)を式(9)の右辺の分子に代入すると、式(12)となる。この式(12)こそがベイズの定理である。

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \tag{10}$$

$$P(A \cap B) = P(B|A) \times P(A) \tag{11}$$

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)} \tag{12}$$

仮に事象 A が起きる場合をA_1とし、起きない場合をA_2とする。また、事象 B が起きる場合をBとする。式(12)は、事象 B が起きる場合に事象 A が起きる確率$P(A_1|B)$、つまり、式(13)として再定義される。分子は事象 A が起きた場合に事象 B が起きる確率$P(B|A_1)$に事象 A が起きる確率$P(A_1)$を掛けたものとなる。分母である$P(B)$は、事象 A が起きた場合に事象 B が起きる確率$P(B|A_1)$に事象 A が起きる確率$P(A_1)$を掛けたものと事象 A が起きなかった場合に事象 B が起きる確率$P(B|A_2)$に事象 A が起きない確率$P(A_2)$を掛けたものを足し合わせたものである。

$$P(A_1|B) = \frac{P(B|A_1) \times P(A_1)}{P(B|A_1) \times P(A_1) + P(B|A_2) \times P(A_2)} \tag{13}$$

A2.4　ベイズの分析：具体例

全人口の 5%が肺病を患っているとしよう。肺病患者のうち 90%が喫煙者であり、肺病患者以外の 25%が喫煙者だとする。全人口の中から無作為に選んだ人が喫煙者である場合、肺病患者である確率を求めてみる(Weiss, 2005)。記号を以下のとおり定義する。

K：抽出された人は喫煙者である
H_1：抽出された人は肺病の患者ではない
H_2：抽出された人は肺病の患者である

上述より、全人口の 5%が肺病を患っているので、$P(H_2) = 0.05$である。肺病患者のうち 90%が喫煙者なので、$P(K|H_2) = 0.90$である。肺病患者以外の 25%が喫煙者なので、$P(K|H_1) = 0.25$である。最後に、全人口のうち肺病を患っていない人の確率は、H_2の補集合になるので、

$P(H_1) = 1 - P(H_2) = 1 - 0.05 = 0.95$となる。これらの数値を式(13)のベイズの定理に代入すると、無作為に抽出された人が喫煙者であると分かっている場合、肺病患者である確率は 0.159 となる。

$$P(H_2|K) = \frac{P(K|H_2)P(H_2)}{P(K|H_1)P(H_1) + P(K|H_2)P(H_2)} = \frac{0.90 * 0.05}{0.25 * 0.95 + 0.90 * 0.05} = 0.159$$

すなわち、ある人が喫煙者であるかどうかが分からない状態では、この人が肺病を患っている確率は 0.05 であったが、新たに喫煙に関する情報が手に入ったことで、確率が更新され、この人が肺病を患っている確率が 0.159 に上がったのである。このように、新しい情報が見つかる度に繰り返し計算を行って、確率を更新することこそが、ベイズ統計学の根幹を成すメカニズムなのである。

A2.5　条件付き確率とベイズの定理

このように、ベイズ統計学は条件付き確率と密接な関係にあるが、ベイズの定理と条件付き確率の違いは事前確率という概念の存在にある。ベイズ統計学では、データを観察する前の「確率的に記述されるべき数値」のことを事前分布と呼び、データを観察した後の「確率的に記述されるべき数値」のことを事後分布と呼ぶ。先ほどの肺病の例で見たとおり、ベイズ統計学における情報更新の手法は、$P(A|B) = P(B|A) \times P(A)/P(B)$の式で表されるベイズの定理を用いて行われる。この式の構成要素にはそれぞれ名称があり、右辺の$P(B|A)$を尤度、$P(A)$を事前確率、$P(B)$を規格化定数と呼び、左辺の$P(A|B)$を事後確率と呼ぶ(矢野, 2012; 小暮, 2013)。

肺病の例では、無作為に抽出された人が肺病患者である確率は$P(H_2) = 0.05$であるとされていた。この確率は、無作為に抽出された人が喫煙者であるかどうかを考慮していないため、事前確率と呼ばれる。その後、無作為に抽出された人が喫煙者であると分かり、この追加情報をもとに、肺病患者であるかどうかについての確率を更新した。つまり、選ばれた人が喫煙者であった場合に、その人が肺病患者である条件付き確率$P(H_2|K) = 0.159$を計算したのである。この更新された確率のことを事後確率と呼ぶ。

A2.6　事前確率の設定

今回の例では、事前確率を 0.05 と設定して議論を進めた。このように、統計学の試験問題においては、通常、事前確率は問題文の中に明記してあるものであるが、現実の分析では、多くの場合、事前確率ははっきりとしておらず、事前確率なしではベイズの定理を使用することもできない。事前確率をどのように設定するべきかは、ベイズ統計学において大きな議論の対象となるところだが(岩崎, 2002, pp.88-89; Congdon, 2006, pp.3-5; Gill, 2008, pp.135-185)、恣意的に設定した具体的な数値を一時的に使用し、ベイズの定理で強引に事後確率を計算す

るのも 1 つの方法である(矢野, 2012, p.116)。

仮に、上記の肺病の例に恣意的な事前確率を用いたとしよう。具体的な確率の数値は、先ほどの 0.159 と異なるものの、事前情報と比べて、事後情報において患者が肺病を患っていると感じる信念の度合いが高まるという構図は同じになる。実際の分析では、事前分布自体が統計モデルの構成要素と考えられ、パラメータの特性に関する仮説の 1 つとして、モデル作成のプロセスにおいて、どのような事前分布が妥当であるかを検討する必要がある。

式(12)のベイズの定理は、データに基づいて事前分布を事後分布に更新する公式である。すなわち、ベイズの定理の左辺における条件付き確率は、事前確率である$P(A)$についてすでに観測したことや信じていることと、尤度$P(B|A)$として新たに観測した情報とのバランスを取っていることを意味している。このように、事前確率とは、人類の英知として蓄積された知識に基づいて追加情報を提供しており、事後確率は、事前分布と尤度の合算した情報源に基づいていると理解でき、ベイズ統計学の分析は、伝統的な頻度論における尤度のみに基づく分析よりも多くの情報を利用していると考えることができる。

A2.7　補足事項

なお、通常、ベイズ統計学においては、事象 B はすでに起こった現象であり、確率 1 で生起するため、無意味な要素であると考える。先ほど述べたとおり、ベイズ統計学においては、観測される数値は固定的であり、観測されない数値には確率的な記述が割り当てられる。したがって、$P(B)$を無視し、$P(A|B) \propto P(B|A) \times P(A)$と比例の形式で記述することが多い。

最後に、ベイズ統計学の名称は、トーマス・ベイズによるエッセイ(Bayes, 1753)にちなんで命名されたものであるが、このエッセイにおいてベイズの定理の核となる概念は導入されたものの、ベイズの定理を一般化し確立したのは、ベイズ本人ではなく、ピエール＝シモン・ラプラスによる論文(Laplace, 1774)だと指摘されている(Gill, 2008, p.10)。

参考文献（英語）

[1] Abayomi, Kobi, Andrew Gelman, and Marc Levy. (2008). "Diagnostics for Multivariate Imputations," *Applied Statistics* vol.57, no.3, pp.273-291.

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

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

[4] Bayes, Thomas. (1753). "An Essay Towards Solving a Problem in the Doctrine of Chances," *Philosophical Transactions of the Royal Society of London* vol.53, pp.370-418.

[5] Bodner, Todd E. (2008). "What Improves with Increased Missing Data Imputations?," *Structural Equation Modeling* vol.15, pp.651-675.

[6] Carpenter, James R. and Michael G. Kenward. (2007). *Missing Data in Clinical Trials—A Practical Guide*. Birmingham: UK National Health Service, National Co-ordinating Centre for Research on Methodology.

[7] Congdon, Peter. (2006). *Bayesian Statistical Modelling*, Second Edition. West Sussex: John Wiley & Sons Ltd.

[8] DeGroot, Morris H. and Mark J. Schervish. (2002). *Probability and Statistics*. Boston: Addison-Wesley.

[9] de Waal, Ton, Jeroen Pannekoek, and Sander Scholtus. (2011). *Handbook of Statistical Data Editing and Imputation*. Hoboken, NJ: John Wiley & Sons.

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

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

[12] Hasler, Caren and Yves Tillé. (2013). "Balanced *k*-nearest Neighbor Imputation," *The 59th World Statistics Congress of the International Statistical Institute*, Hong Kong, China.

[13] Hershberger, Scott L. and Dennis G. Fisher. (2003). "A Note on Determining the Number of Imputations for Missing Data," *Structural Equation Modeling* vol.10, no.4, pp.648-650.

[14] 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, pp.561–581.

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

[16] Honaker, James, Gary King, and Matthew Blackwell. (2013). *Package 'Amelia'*. http://cran.r-project.org/web/packages/Amelia/Amelia.pdf. (Accessed December 26, 2013).

[17] 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, pp.79-90.

[18] 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.

[19] Imai, Kosuke, Gary King, and Olivia Lau. (2008). "Toward A Common Framework for Statistical Analysis and Development," *Journal of Computational and Graphical Statistics* vol.17, no.4, pp.1-22.

[20] Kim, Jae-kwang and Wayne A. Fuller. (2013). "Fractional Hot Deck Imputation for Multivariate Missing Data, *The 59th World Statistics Congress of the International Statistical Institute*, Hong Kong, China.

[21] 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, pp.49-69.

[22] Laplace, Pierre-Simon. (1774). "Mémoire sur la Probabilité des Causes par les Événemens," *Mémoires de l'Académie Royale des Sciences Presentés par Divers Savans* vol.6, pp.621-656. (English Translation: Laplace, Pierre-Simon. (1986). "Memoir on the Probability of the Causes of Events," *Statistical Science* vol.1, no.3, pp.359-378.)

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

[24] 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, pp.277-287.

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

[26] Maddala, Gangadharrao S. (2001). *Introduction to Econometrics*, third edition. Chichester: John Wiley & Sons, Ltd.

[27] Marti, Helena and Michel Chavance. (2011). "Multiple Imputation Analysis of Case-Cohort Studies," *Statistics in Medicine* vol.30, no.13, pp.1595-1607.

[28] Ranalli, M. Giovanna, Alina Matei, and Andrea Neri. (2013). "Handling Nonignorable Nonresponse Using Generalized Calibration with Latent Variables," *The 59th World Statistics Congress of the International Statistical Institute*, Hong Kong, China.

[29] 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*, pp.20–34.

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

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

[32] Schafer, Joseph L. (1992). "Algorithms for Multiple Imputation and Posterior Simulation From Incomplete Multivariate Data with Ignorable Nonresponse," Ph.D. Dissertation, Harvard University, Cambridge, MI.

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

[34] 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. http://sites.stat.psu.edu/~jls/norm3/normVersion3.pdf. (Accessed December 26, 2013).

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

[36] Statistical Solutions. (2011). *SOLAS Version 4.0 Imputation User Manual*. http://www.solasmissingdata.com/wp-content/uploads/2011/05/Solas-4-Manual.pdf. (Accessed December 26, 2013).

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

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

[39] van Buuren, Stef and Karin Groothuis-Oudshoom. (2013). *Package ‘mice’*. http://cran.r-project.org/web/packages/mice/mice.pdf. (Accessed December 26, 2013).

[40] von Hippel, Paul T. (2005). “How Many Imputations Are Needed? A Comment on Hershberger and Fisher (2003),” *Structural Equation Modeling* vol.12, no.2, pp.334-335.

[41] Weiss, Neil A. (2005). *Introductory Statistics*, seventh edition. Boston: Pearson.

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

[43] Yuan, Yang. (2011). “Multiple Imputation Using SAS Software,” *Journal of Statistical Sotfware* vol.45, no.6.

参考文献（日本語）

[44] 青木繁伸. (2009). 『Rによる統計解析』, 東京：オーム社.

[45] 石村貞夫, 石村光資郎. (2007). 『SPSSでやさしく学ぶ統計解析（第3版）』, 東京：東京図書.

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

[47] 金明哲. (2007). 『Rによるデータサイエンス：データ解析の基礎から最新手法まで』, 東京：森北出版.

[48] 栗原伸一. (2011). 『入門統計学―検定から多変量解析・実験計画法まで―』, 東京：オーム社.

[49] 小暮厚之. (2013). 「ベイズ計量経済分析入門―より柔軟なモデリングの実践に向けて」, 『経済セミナー』no.673, pp.37-42.

[50] SAS インスティチュートジャパン. (1999). 『はじめよう SAS システム～実習データ FD 付き～』, 東京：株式会社 SAS インスティチュートジャパン.

[51] 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.

[52] 野間久史, 田中司朗. (2012).「Multiple Imputation 法による 2 段階ケースコントロール研究の解析」, 『応用統計学』vol.41, no.2, pp.79-95.

[53] 森川耕輔, 山本倫生, 狩野裕. (2013). “Missing Data Analysis with Mixture Missing Mechanisms,” 2013 年度統計関連学会連合大会, 大阪大学.

[54] 矢野浩一. (2012). 「ベイズ推定へようこそ！」, 『経済セミナー』no.664, pp.113-121.

[55] 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.