大規模経済系データにおける様々な多重代入法アルゴリズムの検証

高橋　将宜[†]　伊藤　孝之[††]

要旨

多重代入法(Multiple Imputation)の計算アルゴリズムには、マルコフ連鎖モンテカルロ法(MCMC: Markov chain Monte Carlo)、完全条件付指定(FCS: Fully Conditional Specification)、EMB(Expectation-Maximization with Bootstrapping)がある。しかし、3 つのアルゴリズム間の優劣は判然としていない。また、多重代入法の擬似データ数(*M*)をいくつに設定すればよいかについても、明確な答えは見つかっていない。そこで、本稿では、大規模データセットとしての経済センサス - 活動調査の速報データを用いて、3 つのアルゴリズムの比較検証を行った。さらに、経済センサス - 活動調査の速報データに基づくシミュレーションデータを用い、多重代入法の擬似データ数(*M*)をいくつに設定すればよいかについて検証を行った。

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。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)[2]が提唱されてきた(Rubin, 1987)。

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、マルコフ連鎖モンテカルロ法(MCMC: Markov chain Monte Carlo)、完全条件付指定(FCS: Fully Conditional Specification)、EMB(Expectation-Maximization with Bootstrapping)の 3 つを有力なものとして挙げられる。現時点において、3 つのアルゴリズム間の優劣は判然としていない。また、多重代入法の擬似データ数(*M*)をいくつに設定すればよいかについて、明確な答えは見つかっていない。そこで、本稿では、大規模データセットとしての経済センサス - 活動調査の速報データを用いて、3 つのアルゴリズムの比較検証を行った。さらに、経済センサス - 活動調査の速報データに基づくシミュレーションデータを用い、多重代入法の擬似データ数(*M*)をいくつに設定すればよいかについて検証を行った。

1. 欠測値補定と多重代入法の理論

多重代入法の理論は、Rubin (1978)によって初めて提唱され、Rubin (1987)において体系化された。本節では、欠測値補定と Rubin (1978, 1987)による多重代入法の基本的なメカニズムを簡潔に示す(King *et al.*, 2001; 高橋, 伊藤, 2013)。

† 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専門職

[1] 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基に著者が独自集計したものである。また、本稿の内容は、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。

1.1　欠測値補定のメカニズム

表 1.1 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。補定 1 から補定 5 は、多重代入法による補定済データセットを意味している。

表 1.1

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**

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、全く検討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 1.1 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 55 センチから 272 センチ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、「55 センチ～272 センチ」という有限の範囲に候補を狭めることができている。

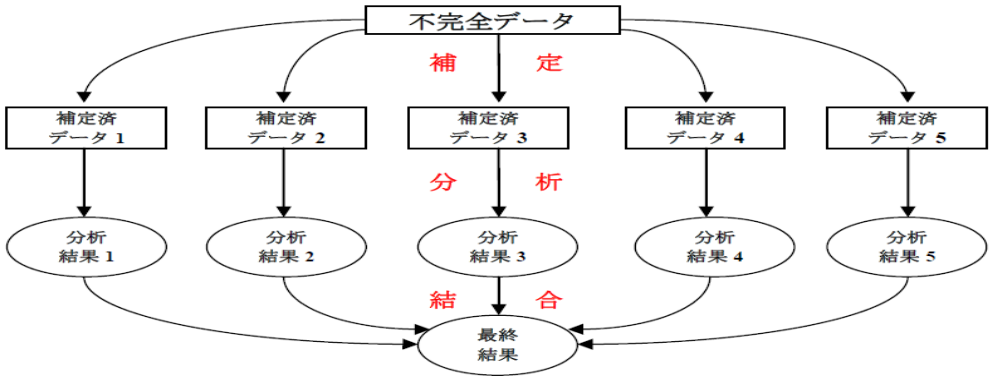
次に、身長の値が欠測している人(ID 9)の他の変数の情報を見ると、この人は、日本人成人男性であることが分かる。日本人成人男性の身長は、平均約 170 センチ、標準偏差 5.8 程度で近似的に正規分布していると考えられている。したがって、ID 9 の身長は、ほぼ 100%に近い確率で 140 センチ以上 200 センチ以内の身長であると推定できる。これにより、最小値を 55 センチから 140 センチまで上げることができ、最大値を 272 センチから 200 センチまで下げ、推定値の幅を狭めることに成功している。さらに、日本人成人男性の身長と体重の相関データでは、身長 178 センチの人の平均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

このように、論理や実証的データに基づいて、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合理的な値に置き換える作業のことを補定(imputation)と呼ぶ。しかし、体重 69 キロの日本人成人男性のすべてが、身長 178 センチであるとは信じ難い。おおよそ 178 センチの周辺の値であると思われるが、中には 178 センチ以上の人もいれば、178 センチ未満の人もいるだろう。合理的に 178 センチ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。

1.2　多重代入法のメカニズム

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。表 1.1 では、補定 1～補定 5 の部分が補定済データセットにあたり、身長の値は、184.8、174.6、178.3、177.0、173.0 となっている。数値のばらつきが大きいのは、データサイズが小さく、補定値が不安定であること（不確実性）を反映している。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 に図示）。

図 1.1: 多重代入法の模式図



多重代入法によりできあがった M 個の補定済データセットを別々に使用して、通常の統計分析（検定や回帰分析など）を行い、以下の手順にしたがって推定値を統合し、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。$\hat{\theta}_m$の分散$\mathrm{var}(\hat{\theta}_m)$の推定値を$v_m$とする。$\bar{v}_M$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り、$(1 + 1/M)$は、$M$ のサイズが有限であるために調整を施す項である[3]。

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

[3] もし M が無限大であるならば、$\lim_{M\to\infty}\left(1 + \frac{1}{M}\right)\tilde{v}_M = \tilde{v}_M$となる。

本研究では、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）[4]。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。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、式(3)のとおり、線形である。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。

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

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[5]。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。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ため、μとΣが完全には既知ではなく[6]、βの推定に関して確信を持つことができない。

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、事後分布からμとΣの無作為抽出を行うことで、こういった推定不確実性を反映している。しかし、伝統的な手法により事後分布から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。

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

本節では、主だった 3 種類の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を用いて、それらのアルゴリズムを応用したコンピュータソフトウェアの検証を行う。

2.1 マルコフ連鎖モンテカルロ法 (MCMC): データ拡大法 (Data Augmentation)

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; Little and Rubin, 2002;)。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)[7]として利用可能である。そのメカニズムは、以下のとおりである (Schafer, 1997; 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、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、確率

---

[4] 本研究で用いた記号の意味は、以下のとおりである。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。

[5] たとえば、Xと Yの単回帰$(Y_i = \beta_0 + \beta_1 X_i + u_i)$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。

[6] 表 1.1 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。

[7] 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。

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$\theta_0, \dots, \theta_{t-1}$から条件付で独立となる。また、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から算出）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。データ拡大法の基本的なメカニズムは、初期値θ_0から、I-Step (Imputation Step)において、$P(Y_{mis}|Y_{obs}, \theta_t)$に基づいて、$Y_{mis}^{(t+1)}$を生成する。つまり、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する。次に、P-Step (Posterior Step)において、$P\left(\theta|Y_{obs}, Y_{mis}^{(t+1)}\right)$に基づいて、$\theta_{t+1}$を生成する。つまり、事後分布からパラメータ値を生成する。そして、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。

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)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)である。

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(van Buuren, 2012)。MICE とは、「連鎖方程式による多変量補定」という意味であり、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：$P\left(Y_{j,mis}|Y_{j,obs}, Y_{-j}, R\right)$。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。$\tilde{Y}_{-j,t} = \left(\tilde{Y}_{1,t}, \dots, \tilde{Y}_{j-1,t}, \tilde{Y}_{j+1,t-1}, \dots, \tilde{Y}_{p,t-1}\right)$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：$\tilde{\lambda}_{j,t} \sim P\left(\lambda_{j,t}|Y_{j,obs}, \tilde{Y}_{-j,t}, R\right)$。その後、補定値の抽出を行う：$\tilde{Y}_{j,t} \sim P\left(Y_{j,mis}|Y_{j,obs}, \tilde{Y}_{-j,t}, R, \tilde{\lambda}_{j,t}\right)$。

2.3 EMB アルゴリズム

EMB アルゴリズムは、比較的新しいアルゴリズムであり、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。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)。EMB アルゴリズムのメカニズムは以下のとおりである。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復元抽出を 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)。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。

2.4 経済センサス - 活動調査の速報データを用いた分析結果

本項では、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））のデータ（観測数 277,263）を用い、上述の 3 つの多重代入法アルゴリズムの検証を行った。欠測の発生メカニズム[9]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）、事業従事者数（自然対数）には欠測を発生させていない（欠測率 0%）。分析結果は、表 2.1 に示すとおりである。

表 2.1（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	処理速度
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	1 分 24 秒
MICE	8.7819	1.4598	1.1820	420.2365	10 分 35 秒
NORM	NA	NA	NA	NA	動作せず

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 及び MICE 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[10]）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。Amelia と MICE の間では、統計的に有意な差は見られなかった[11]。また、「処理速度」は、計算効率の検証を行った結果である[12]。EMB アルゴリズムを搭載した Amelia の処理速度は極めて速かった。FCS アルゴリズムに基づく MICE の処理速度は、Amelia の数倍かかった。なお、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。

3. 多重代入法の M 数

1 節では、機械的に $M = 5$ 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

---

[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] MAR (**M**issing **A**t **R**andom)とは、観測データを条件とした場合、欠測の発生確率は無作為であることを意味する：$P(R|D) = P(R|Y_{obs})$。また、MCAR (**M**issing **C**ompletely **A**t **R**andom)とは、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することを意味している：$P(R|D) = P(R)$ (Little and Rubin, 2002)。

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

[11] 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 検定により、95%水準で検定を行った。

[12] 使用したコンピュータは、Windows Vista、プロセッサ：Intel Core 2 Duo CPU T9400、メモリ(RAM)：2.00 GB、32 ビットオペレーティングシステムを搭載した一般的なノートパソコンである。

方、近年では、3.2 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[13]。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。

3.1 M 数に関する議論：相対効率に関する Rubin の主張

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(4)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(4)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。

$$ARE = \left(\sqrt{1 + \frac{\delta}{M}}\right)^{-1} \times 100 \qquad (4)$$

表 3.1 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 50%あったとしても、95.35%の相対効率を達成できており、仮に欠測率が 90%であったとしても、相対効率は 92.06%を達成しているとされる。

表 3.1：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
15	99.67	99.34	99.01	98.69	98.37	98.06	97.75	97.44	97.13
20	99.75	99.50	99.26	99.01	98.77	98.53	98.29	98.06	97.82
∞	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

式(4)こそが、多くの文献(Schafer, 1997; King *et al.*, 2001; Allison, 2002; von Hippel, 2005; Congdon, 2006; 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。

3.2 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連れて、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)、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、必用な 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

[13] 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。

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。

3.3 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

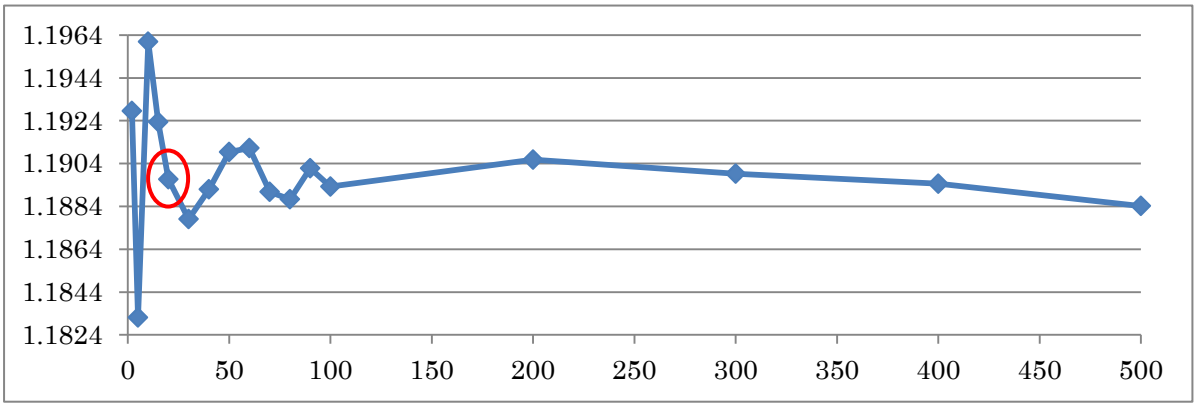
自然対数変換した経済センサス - 活動調査の速報データの情報（平均値、分散・共分散など）を基に、多変量正規分布によって観測数 1000、3 変量のシミュレーションデータセットを 3 つ生成した。シミュレーションデータの基になったデータは、2.4 項と同じく、産業大分類 I の単独事業所（個人経営以外）を用いた。データセットの基本統計量（紙面の都合上、データ 1 のみ記載）は、表 3.2 のとおりである。Amelia、MICE、Norm において、表 3.2 のシミュレーションデータセット（欠測率 = 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%）に多重代入（M = 2, 5, 10, 15, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500）を施し、多重代入済データセットを用いて、$\log\left(\widehat{\text{売上高}}_i\right) = \hat{\alpha} + \hat{\beta}\log(\text{事業従事者数}_i)$における$\hat{\beta}$とその標準誤差の推定を行った[14]。

表 3.2

	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 1	3.314	7.761	8.837	8.764	9.752	15.422	1.510
従事者数 1	0.000	0.946	1.528	1.514	2.077	3.993	0.891
資本金 1	3.699	5.727	6.331	6.323	6.921	9.088	0.907

図 3.1 では、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 3.1：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 3.2 のとおりである。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 では、95%信頼区間が(1.188, 1.220)となり、その範囲は、わず

[14] 乱数による影響を見るため、シードを 10 個設定した。紙面の都合上、Amelia におけるシミュレーションデータ 1 の分析結果の概略のみを掲載している。分析結果の詳細は、当日報告する。なお、真値は 1.2075 である。

か 0.032 となっている。

図 3.2：係数の分布（欠測率 50%、シード数 1000）

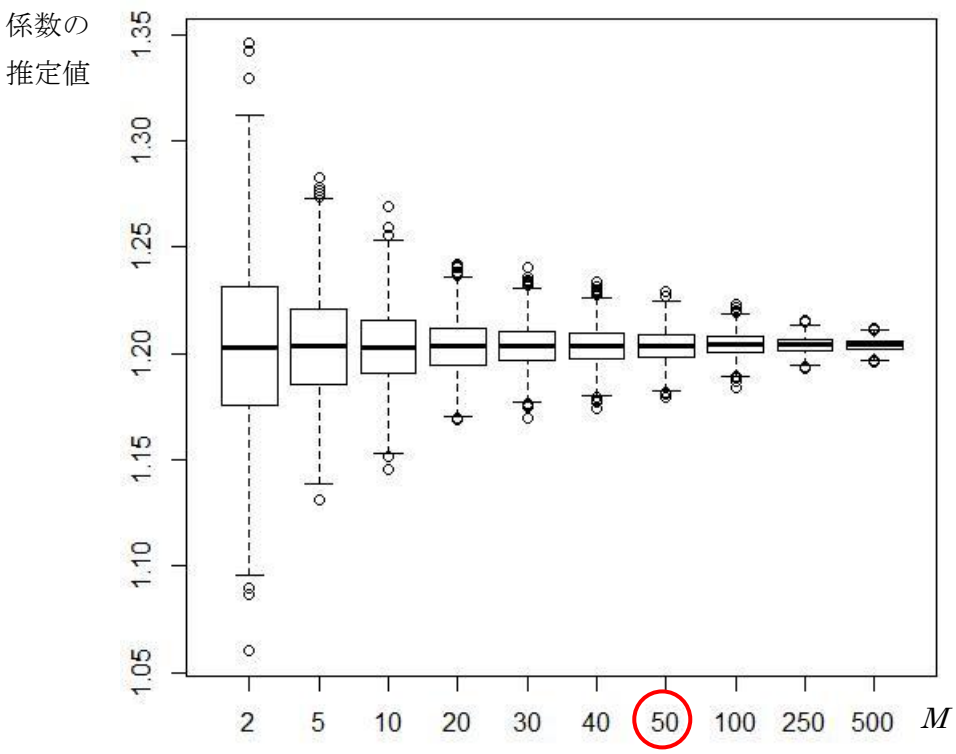
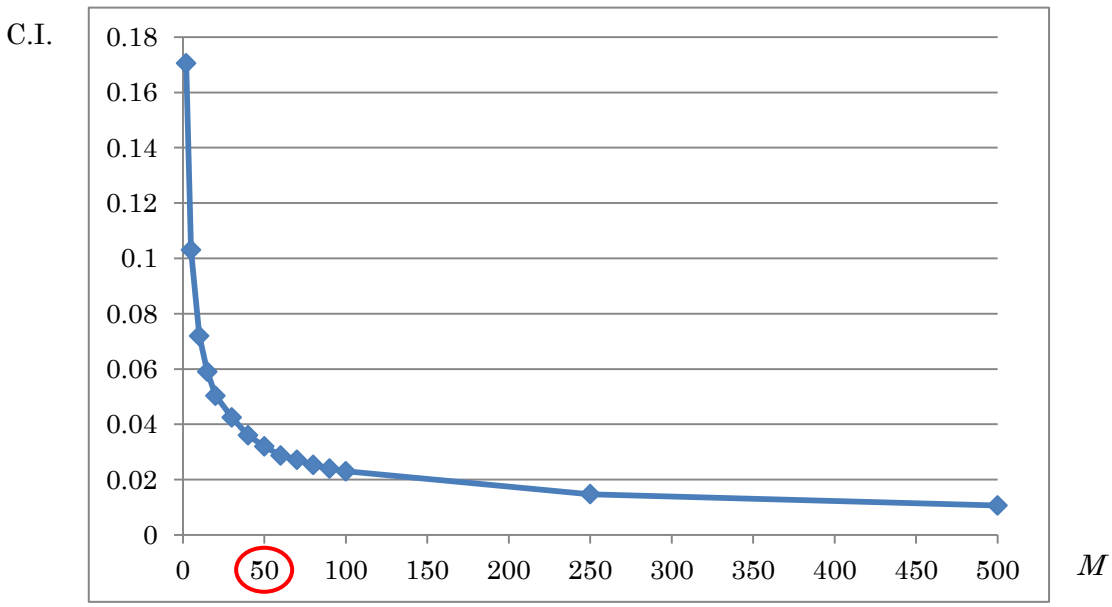


図 3.3 は、図 3.2 における 95%の信頼区間(C.I.)の長さを図示したものであり、$M = 50$ を超えると、95%の信頼区間の長さはほぼ一定になっており、これを超えて得られる相対効率は非常に低いことが分かる。

図 3.3：95%信頼区間の長さ



4. 結語

本研究では、Amelia と MICE の補定値の精度には大きな差がないことが分かった。一方、Amelia の計算処理速度は極めて速く、Norm は 27 万×3 変量のデータセットを分析することができなかった。多重代入法の擬似データ数 M については、詳細な結果は発表時に報告するが、概ね 5～10 では少なすぎ、20～50 程度が適切だと考えられる[15]。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。

参考文献

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

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

[3] 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.

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

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

[6] 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.

[7] 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.

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

[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, pp.49-69.

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

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

[13] 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.

[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*, pp.20–34.

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

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

[17] 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.

[18] 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.

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

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

[21] 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.

[15] 従来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。すなわち、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わけだが、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ため、繰り返し回数が少なくてもよいと考えられる(Honaker and King, 2010)。