統計研究彙報　第 70 号　2013 年 3 月　（19～86）

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～

高橋　将宜†

伊藤　孝之†

Imputing Missing Values of Turnover in Economic Surveys
-Assessment of Multiple Imputation-

TAKAHASHI, Masayoshi
ITO, Takayuki

企業の売上高といった経理項目を対象とする経済調査では、回答ユニットの形態が多岐にわたることもあり、データの広がりが大きく、かつ、記入漏れや記入誤りが生じやすい。その結果、調査から得られる情報だけでは、経済の実態を正しく反映できないものとなるおそれがある。そこで、独立行政法人統計センター（以下、「統計センター」とする）では、将来の経済調査における経理項目の結果精度の向上に資するものとして、データエディティング手法について研究を行っている。本研究では、EDINET データを使用し、個別データの補定方法として多重代入法(Multiple Imputation)を評価した。本稿では、様々な欠測値対処法とその限界を示し、代替法として多重代入法を導入し、フリーソフトウェア R の汎用多重代入法パッケージ Amelia を利用して、多重代入法による欠測値補定の精度評価を行う。

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、Amelia

There are various types of responding units in economic surveys, which collect information on accounting items, such as turnover of enterprises. In such surveys, the distribution of accounting items is vast, and missing values and errors will be frequently produced. As a result, there is a possibility that the actual condition of the economy may not be reflected, based solely on the information obtained from the survey.Therefore, we are engaging in research on data editing strategies, in order to improve the quality of the future economic surveys. In this research, we use EDINET data to evaluate multiple imputation as a method for micro-data imputation. In this paper, we describe various imputation techniques and their limitations, illustrate the mechanism and advantages of multiple imputation, and evaluate the performance of imputation models via R package Amelia (a general-purpose multiple imputation tool).

Keywords: Economic Survey, Accounting Item, Missing Value, Imputation, Multiple Imputation, R, Amelia

---

原稿受理日　平成 24 年 12 月 20 日　†　独立行政法人統計センター情報技術部統計技術研究課

はじめに[1]

個人や世帯を対象とする調査と異なり、企業の売上高といった経理項目を対象とする経済調査では、回答ユニットの形態が多岐にわたることもあり、データの広がりが大きく、かつ、記入漏れや記入誤りが生じやすい。その結果、調査から得られる情報だけでは、経済の実態を正しく反映できないものとなるおそれがある。そこで、統計センターでは、将来の経済調査において上記懸念事項を改善するために、データエディティング手法について研究を行い、経理項目の結果精度の向上に資するものとする。本研究では、EDINET データを使用し、個別データの補定方法として多重代入法[2]を評価した。本稿では、様々な欠測値対処法とその限界を示し、代替法として多重代入法を導入し、フリーソフトウェア R の汎用多重代入法パッケージ Amelia を利用して、多重代入法による欠測値補定の精度評価を行う。

第 1 節では欠測に関する 3 つの主な前提について具体例を交えながら詳説する。第 2 節では様々な欠測値の対処法を紹介し、その限界を示す。第 3 節では統計センター及び国連欧州経済委員会(UNECE: United Nations Economic Commission for Europe)のワークセッションにおける補定の先行研究を概観し、先行文献に占める本稿の貢献を示す。第 4 節では多重代入法のメカニズムを、具体例を示しながら詳説し、第 5 節では EMB(Expectation Maximization with Bootstrapping)アルゴリズムを導入する。第 6 節では多重代入法の *M* 数の決定方法を示し、第 7 節では R の汎用多重代入法パッケージ Amelia II を紹介する。第 8 節では EDINET データを用い多重代入法と単一代入法を比較し、第 9 節では多重代入法による補定の精度を診断する。第 10 節ではシミュレーションデータを用い多重代入法と単一代入法を比較する。第 11 節において結語と将来の課題で締めくくる。

1　欠測に関する 3 つの前提

本節では、データが欠測するメカニズムに関する 3 つの主な前提について概観する(Little and Rubin, 2002, pp.11-12, pp.312-313; King *et al.*, 2001, pp.50-51)。欠測値を含むデータを分析する際には、欠測のメカニズムの種類によって対象となるパラメータの不偏推定量が存在するか否かが決まるため、欠測のメカニズムを正しく想定することが不可欠である(岩崎, 2002, p.7; Marti and Chavance, 2011)。

---

[1] 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。本稿は、第 101 回研究報告会（総務省統計研修所）、2012 年度統計関連学会連合大会（北海道大学札幌キャンパス）、2012 年国連欧州経済委員会統計データエディティングに関するワークセッション（ノルウェー、オスロ）、2012 年度科学研究費シンポジウム（島根県松江市、奈良県奈良市）における報告に加筆・修正したものである。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。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。

[2] 「多重代入法」とは、Multiple Imputation の訳であり、「多重補定法」、「多重補完法」、「多重補填法」、「多重決め付け法」、「マルチプル・インピュテーション」など様々に訳し得る。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、学術的には「多重代入法」の呼び名が流通している(渡辺, 山口, 2000; 岩崎, 2002; 星野, 2009; 宮本, 安藤, 逸見, 山下, 高橋, 2012)。また、一般的にも、Google 上において（2012 年 12 月 20 日現在)、「多重代入」（7,950 件)、「多重補定」（5 件）、「多重補完」（424 件）、「多重補填」（24 件)、「多重決め付け」（89 件）、「マルチプル・インピュテーション」（8 件）のヒット数であった。よって、本稿では、最も汎用的に使用されている「多重代入法」の用語を用いる。

1.1　例示用データ

被説明変数Y及び説明変数Xを含むデータ行列をDと定義する。すなわち、$D = \{Y, X\}$である。欠測インディケータ行列をKと定義する。DとKの次元は同じであり、Dが観測される場合にはKの値が 1 であり、Dが欠測している場合にはKの値は 0 である。また、データの観測値をD_oとし、欠測値をD_kと定義する。すなわち、$D = \{D_o, D_k\}$である。

表 1.1 は、例示のために以下の手順で生成したシミュレーションデータセットであり、架空の会社における社員の月収及び年齢を表しているとする。*X1* は、平均値 40、標準偏差 6、個数 10 の正規乱数であり、便宜的に年齢を模した説明変数である。*X2* は、サイコロ投げを模した 10 個の一様乱数である。*e* は、平均値 0、標準偏差 3、個数 10 の正規乱数である。*Y* は、1 + 1**X1* + *e* によって生成された被説明変数であり、便宜的に月収（万円）を模しているとする。また、*K* は、欠測インディケータ行列であり、*Ky*、*Kx1*、*Kx2* はそれぞれ、*Y*、*X1*、*X2* のどの値が欠測しているかを示す。表 1.1 では、$D = \{Y, X1, X2\}$であり、$K = \{Ky, Kx1, Kx2\}$である。DとKの次元は、それぞれ 10×3 の行列である。参考までに、完全データの Y（月収）の平均値は 41.51、中央値は 40.20、標準偏差は 4.51 である。

表 1.1：生データ

	D			K		
id	Y (月収)	X1 (年齢)	X2 (サイコロ)	Ky	Kx1	Kx2
1	40.5	42	3	1	1	1
2	43.1	37	4	1	1	1
3	36.9	36	5	1	1	1
4	39.3	34	3	1	1	1
5	50.1	46	1	1	1	1
6	39.9	44	4	1	1	1
7	44.3	40	6	1	1	1
8	38.9	35	2	1	1	1
9	35.5	35	5	1	1	1
10	46.6	40	5	1	1	1

1.2　MCAR：欠測は完全にランダム

1 つ目の前提は、MCAR と呼ばれ、***M**issing **C**ompletely **A**t **R**andom*（欠測は完全にランダム）の略である。MCAR の状態では、$P(K|D) = P(K)$であり、KはDから独立であることを意味する。すなわち、欠測はデータ内の情報から独立して発生している。

たとえば、収入に関する調査において、サイコロを振り、1～4 が出た人は回答し、5 又は 6 が出た人は回答しないとしよう。すなわち、表 1.2 に示すとおり、*X2*（サイコロ）の値が 5 以上の場合、*Y*（月収）の値に欠測が発生すると想定する。仮に、網掛けになっている箇所は、データ内に含まれていないとしよう。欠測メカニズムに関する情報が含まれておらず、かつ、

真の欠測メカニズムはランダムなサイコロの値に依存しており、観測データ内の情報から独立であるので、表 1.2 は MCAR の典型例である。表 1.2 では、*X1*（年齢）の値を参考にして、*Y*（月収）の欠測パターン[3]を予測できない状態となっているが、真の欠測メカニズムは、サイコロ投げにより完全にランダムに決まっており、無視することができる(Ignorable)。参考までに、不完全データ（MCAR）の *Y*（月収）の平均値は 41.97、中央値は 40.20、標準偏差は 4.25 であり、完全データとほぼ一致し、大幅に偏ってはおらず、欠測を無視できることが伺える。

表 1.2：MCAR

	D			K		
id	Y (月収)	X1 (年齢)	X2 (サイコロ)	Ky	Kx1	Kx2
1	40.5	42	3	1	1	0
2	43.1	37	4	1	1	0
3	36.9	36	5	0	1	0
4	39.3	34	3	1	1	0
5	50.1	46	1	1	1	0
6	39.9	44	4	1	1	0
7	44.3	40	6	0	1	0
8	38.9	35	2	1	1	0
9	35.5	35	5	0	1	0
10	46.6	40	5	0	1	0

1.3 MAR：欠測はランダム

2 つ目の前提は、MAR と呼ばれ、***M**issing **A**t **R**andom*（欠測はランダム）の略である。MAR の状態では、$P(K|D) = P(K|D_o)$であり、KはD_kから独立であることを意味する。すなわち、あるセルの値が欠測する確率は、D_oに依存しているが、D_oを考慮した後は、D_kから独立である。MAR とは、データ内にある情報により、欠測データパターンを予測することができる状態である[4]。非常に簡潔に要約すれば、MAR とは、欠測の発生メカニズムが、欠測を含む変数の値には基づかず、データ内の他の変数の値に応じて発生している状態のことである。

たとえば、もし年齢が上がるほど収入に関して答えない傾向があり、年齢に関する項目が観測データ内に存在しているならば、収入の欠測は MAR であると言える。表 1.3 では、*X1*（年齢）の値が 40 以上になると自動的に欠測が発生するようになっている。網掛けとなっているセルは、データセット内に含まれていないとしよう。*X1*（年齢）の値を見ることにより、*Y*（月収）の欠測パターンを予測できる。つまり、*X1* の値が 40 未満のとき *Y* の値は観測され、*X1* の値が 40 以上のとき *Y* の値は欠測する。参考までに、不完全データ（MAR）の *Y*（月収）

[3] 欠測データパターンは、データ内のどの値が観測され、どの値が欠測しているかを表す。一方、欠測データメカニズムは、データ内の欠測と変数の値との関係を表す(Little and Rubin, 2002, p.4)。

[4] この予測は必ずしも因果関係に基づくものである必要はない。たとえば、一般的な年功序列では、年齢が上がれば上がるほど収入が高くなり、年齢が収入の原因であると考えられ、収入が年齢の原因とは考えられないが、補定の文脈では、収入と年齢のどちらが原因又は結果であってもかまわないのである。

の平均値は 38.74、中央値は 38.90、標準偏差は 2.88 であり、MCAR の場合と比較して、完全データから乖離しているが、説明変数 *X1* の値を考慮に入れれば、*X1* の値が 40 未満のときの *Y* の平均値は 38.74 であり、これは完全データにおける条件付き平均値と完全に一致する。すなわち、観測データを考慮に入れれば、欠測を無視できることが分かる。

表 1.3：MAR

	D			K		
id	Y (月収)	X1 (年齢)	X2 (サイコロ)	Ky	Kx1	Kx2
1	40.5	42	3	0	1	1
2	43.1	37	4	1	1	1
3	36.9	36	5	1	1	1
4	39.3	34	3	1	1	1
5	50.1	46	1	0	1	1
6	39.9	44	4	0	1	1
7	44.3	40	6	0	1	1
8	38.9	35	2	1	1	1
9	35.5	35	5	1	1	1
10	46.6	40	5	0	1	1

1.4　NI：欠測は無視できない

3 つ目の前提は、NI と呼ばれ、***NonIgnorable***（欠測は無視できない）の略である[5]。NI の状態では、$P(K|D)$は単純化することができず、KはDから独立ではないことを意味する。すなわち、欠測は、欠測を伴う変数の値に応じて発生する。

たとえば、高収入の人ほど収入に関して非回答率が高いとする。表 1.4 では、*Y*（月収）の値が 40 以上になると欠測している。網掛けとなっているセルは、データセット内に含まれていないとしよう。すると、データセット内の情報から *Y* の欠測パターンを予測することができない。*X2*（サイコロ）の値を見ても、*Y* のどの値が欠測するのか予測ができない。しかし、MCAR の場合とは異なり、網掛けとなっている *Y* の欠測値を見れば分かるとおり、*Y* の欠測は無作為ではなく、パターンが存在している。このような場合、欠測は NI であり、無視できない。参考までに、不完全データ（NI）の *Y*（月収）の平均値は 38.1、中央値は 38.90、標準偏差は 1.83 であり、完全データから大きく乖離していることが伺える。さらに、データセット内の情報から、欠測のパターンを推定できないため、欠測を無視できないことが分かる。

[5] NI は、NMAR (***N**ot **M**issing **A**t **R**andom*：ランダムに欠測していない)とも呼ばれる(Little and Rubin, 2002, p.312)。また、NI に対して、MCAR は Ignorable（欠測は無視できる）であり、MAR は条件次第で Ignorable である(Little and Rubin, 2002, p.119; 岩崎, 2002, p.5)。厳密には、欠測メカニズムが無視できるのは、MAR であり、かつ、以下の分離条件が満たされる場合である：結合パラメータスペース(θ, ψ)が、パラメータスペースθとパラメータスペースψの積である場合に、パラメータθとψは分離できる(Little and Rubin, 2002, pp.119-120)。

表 1.4：NI

	D			K		
id	Y (月収)	X1 (年齢)	X2 (サイコロ)	Ky	Kx1	Kx2
1	40.5	42	3	0	0	1
2	43.1	37	4	0	0	1
3	36.9	36	5	1	0	1
4	39.3	34	3	1	0	1
5	50.1	46	1	0	0	1
6	39.9	44	4	1	0	1
7	44.3	40	6	0	0	1
8	38.9	35	2	1	0	1
9	35.5	35	5	1	0	1
10	46.6	40	5	0	0	1

1.5 NI→MAR

表 1.4 のように、欠測を伴う変数の値に応じて欠測が発生する場合、欠測の発生メカニズムは NI となる。しかし、表 1.5 における年齢変数のように、どの人が高収入であるかを合理的に予測できそうな変数、すなわち、**確率的**に欠測を予測できる変数が含まれていれば、たとえ実際の欠測発生メカニズムが NI であったとしても、事実上の MAR と言える。

表 1.5 では、年齢が 40 歳以上の場合 4/5 の確率で月収に欠測が発生し、年齢が 40 歳未満の場合 4/5 の確率で観測されている。すなわち、データ内に存在する情報から、欠測の発生パターンを確率的に予測することが可能となっており、MAR の一例と言える。

表 1.5：NI→MAR

	D			K		
id	Y (月収)	X1 (年齢)	X2 (サイコロ)	Ky	Kx1	Kx2
1	40.5	42	3	0	1	1
2	43.1	37	4	0	1	1
3	36.9	36	5	1	1	1
4	39.3	34	3	1	1	1
5	50.1	46	1	0	1	1
6	39.9	44	4	1	1	1
7	44.3	40	6	0	1	1
8	38.9	35	2	1	1	1
9	35.5	35	5	1	1	1
10	46.6	40	5	0	1	1

1.6 まとめ

したがって、今回の例を見て分かるように、収入における欠測は MCAR、MAR、NI のいずれにもなり得る。つまり、一般的に、ある変数における欠測が MCAR であるか、MAR で

あるか、NI であるかは、一義的に決まることではなく、飽くまでも特定のデータセットの中に欠測を予測できる情報があるかどうかによって決まるのである。しかし、現実のデータセットでは、欠測値は当然のことながら欠測しているので、データセット内の情報から欠測を予測できるかどうかを当該のデータセットを用いて統計的に検定することはできない[6]。

よって、MCAR、MAR、NI は、欠測値に対応する真値を知らない限り、決して証明することのできない前提なのであり、過去の情報や外部情報、論理などを駆使することにより、前提を妥当なものとして確立しなければならない[7] (Gelman and Hill, 2006, p.531)。欠測値補定の文脈では、MAR を前提とすることが多く、MAR の前提は、必ずしもすべての欠測データにおいて現実的ではないかもしれないが、少なくとも MCAR の前提よりは現実的であり、また、NI の前提に基づいた特殊な手法よりも正確な予測値を算出できることが多いということが経験的に分かっている(Little and Rubin, 2002, p.19)。

表 1.5 の例が示すとおり、たとえ欠測の発生メカニズムが Y の値に依存していたとしても、十分に確率的に予測を行える変数がデータセット内に存在していれば、事実上の MAR であると言える。すなわち、説明変数の数を増やしたり、被説明変数と相関性の高い説明変数を使用したり、変換を行って当てはまりのよいモデルを探し出すことによって、MAR の前提は、より妥当なものに近づいていくと言える。

2　様々な欠測値対処法とその限界

どれほど注意深く調査を設計したとしても、すべてのデータを有効な回答データとして回収できることは非常にまれであり、あらゆるデータセットにおいて、欠測はほとんど常に発生すると言える。データ内に欠測値が存在するということは、利用可能なデータサイズが縮小し、効率性が下がるだけではなく、標準的な統計分析手法が適用できないことを意味する。さらに、回答者と非回答者の間に体系的な差異が存在するならば、データに偏りが存在する可能性がある(Rubin, 1987, p.1)。したがって、統計実務においては、何らかの形で欠測値に対処することが常に必要なのである。

2.1　リストワイズ除去法

最も簡単で、かつ、最もよく使用されている欠測データの対処法として、リストワイズ除去法(List-Wise Deletion)[8]が挙げられるであろう。欠測のある行はすべて捨て去ってしまう方法

[6] 3 つの前提の中で MCAR のみ統計的に検定を行うことが可能である。Little's MCAR test についての詳細は、Enders (2010, pp.19-21)を参照されたい。ただし、Little's MCAR test では、検定の検出力が低く、第二種過誤(Type II error)が起きやすい。すなわち、真の欠測メカニズムは、MCAR ではないにもかかわらず、帰無仮説（欠測メカニズム = MCAR）を棄却できず、誤って MCAR の前提を信じてしまう可能性が高いということである。しかしながら、統計検定の非対称性を考えれば、Little's MCAR test において、欠測メカニズムが MCAR ではないことを実証することはできるが、欠測メカニズムが MCAR であることを実証することは、そもそも、できないのである。

[7] 欠測の前提を直接的には検定できないが、補定モデルの妥当性を間接的に検証することは可能である。詳細に関しては、本稿 9 節を参照されたい。

[8] ケースワイズ除去法(Case-Wise Deletion)、完全ケース分析(Complete-Case Analysis)とも呼ばれる(Cranmer and Gill, 2012)。多くの統計ソフトにおいて、欠測値を含んだデータセットを分析する際に、デフォルトの設定となっている。

である。表 2.1 では、月収、年齢、性別に関して、10 人のデータが表示されているが、ID 番号 10 の月収の値が欠測している（前節に引き続き、網掛けとなっているセルは、データセット内に含まれていないとする)。

表 2.1：生データ 1

	D			K		
id	月収	年齢	性別	Ky	Kx1	Kx2
1	40.5	42	女	1	1	1
2	43.1	37	男	1	1	1
3	36.9	36	男	1	1	1
4	39.3	34	男	1	1	1
5	50.1	46	男	1	1	1
6	39.9	44	女	1	1	1
7	44.3	40	男	1	1	1
8	38.9	35	女	1	1	1
9	35.5	35	女	1	1	1
10	46.6	40	男	0	1	1

ここで、10 人の月収の平均値を求めたいとしよう。初級数学で習うとおり、数式的には、平均値$\bar{x}$は以下の式(1)により簡単に求められる。

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

そこで、(40.5 + 43.1 + 36.9 + 39.3 + 50.1 + 39.9 + 44.3 + 38.9 + 35.5 + x_{10})/10 とすればよいわけだが、欠測値が未知の値として存在しているため、(368.5 + x_{10})/10 となり、これ以上、単純化することができない。つまり、データ内に欠測値が存在している場合には、平均値を求めるといった非常に簡単な統計分析すら行うことができないのである。

ところが、表 2.1 のデータを、標準的な統計ソフトで読み込み、平均値を求めるコマンドを選択すれば、40.94 という結果が出るであろう。標準的な統計ソフトの多くでは、欠測値の存在するデータを読み込んだ場合に、表 2.2 のように、自動的にリストワイズ除去を行い、ID 番号 10 の情報をすべて捨て去り、標本サイズを 10 から 9 に縮小している。すなわち、平均値の計算では、(40.5 + 43.1 + 36.9 + 39.3 + 50.1 + 39.9 + 44.3 + 38.9 + 35.5)/9 = 40.94 となり、平均値を数値として算出しており、一見すると統計分析を問題なく行えるように思えるが、これは 10 人の平均値ではなく観測された 9 人の平均値なのである。

表 2.2：リストワイズ除去済データ（**ID10 の行を削除**）

	D			K		
id	月収	年齢	性別	Ky	Kx1	Kx2
1	40.5	42	女	1	1	1
2	43.1	37	男	1	1	1
3	36.9	36	男	1	1	1
4	39.3	34	男	1	1	1
5	50.1	46	男	1	1	1
6	39.9	44	女	1	1	1
7	44.3	40	男	1	1	1
8	38.9	35	女	1	1	1
9	35.5	35	女	1	1	1

欠測が MCAR の場合、リストワイズによる除去は、無作為抽出と同じであると考えられるので、情報量が少なくなるだけであり、偏りにはほとんど影響を与えない。しかし、情報量が少なくなることにより、推定値の精度が下がり、標準誤差が人工的に大きなものとなる(Cranmer and Gill, 2012)。さらに、現実のデータでは、MCAR を前提とする根拠は希薄である。データ内のどのセルが欠測しているかを確率的に予測できるときはいつでも、MCAR の前提は誤りなのであり、この場合、推定値に偏りが生じる(Schafer, 1999, pp.6-7; King *et al.*, 2001, pp.51-52)。また、標本調査において MCAR を前提できる場合ならばよいが、全数調査において、リストワイズ除去法を用いることは根本的に無意味である。さらに、除去した行には ID 番号 10 の年齢と性別に関する貴重な情報が記録されているにもかかわらず、それらを捨て去ってしまっており、非常にモッタイナイ手法である。

リストワイズ除去法にはこういった限界があるので、近年では、欠測データの対処法として、単一代入法(Single Imputation)が直感的な方法として頻繁に使用されている。すなわち、欠測の穴を、観測データに基づいて算出した何らかの予測値で置き換えるという方法である。この「何らかの予測値」の代表的なものとして、以下のものを挙げることができる：平均値補定；コールドデック補定；ホットデック補定；回帰補定(Little and Rubin, 2002, pp.60-61; de Waal *et al.*, 2011, p.230, pp.246-247, p.249)。

2.2　平均値補定

平均値補定は、欠測値を回答された観測値の平均値に置き換える手法である。表 2.1 の例に戻れば、ID 番号 10 の月収の補定値として、観測されている 9 人の月収の平均である 40.94 を用いるということである。したがって、10 人の月収の平均値は、(40.5 + 43.1 + 36.9 + 39.3 + 50.1 + 39.9 + 44.3 + 38.9 + 35.5 + **40.94**)/10 = **40.94** となり、数値として求めることができ、統計分析を行えるが、10 人目の情報は標本平均の算出に貢献していないことが分かる。この例から分かるとおり、平均値補定では、年齢と性別といった情報もまったく考慮していない。

また、表 2.3 は、n = 1000、平均値 100、標準偏差 10 の正規分布に基づく乱数に 10%の欠測値を人工的に無作為発生させ、平均値補定を行った結果の基本統計量である。この表を一

見すると、補定はおおむね成功したかのように見える。

表 2.3：基本統計量

	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
正データ	63.30	93.05	100.12	99.93	106.69	129.49	9.91
平均値補定	63.30	94.21	100.02	100.02	105.86	129.49	9.37

図 2.1 は正データのヒストグラムであり、図 2.2 は補定済データのヒストグラムであり、図 2.3 は、正データと補定済データの散布図である。正データのヒストグラムと比べて、平均値補定済データのヒストグラムには、横軸 100 の周辺に大幅な偏りがあることが視覚的に分かる。また、散布図には、平均値補定の値が縦一直線に並んでいる。この結果をもとに、表 2.3 を改めて見直せば、**標準偏差が過小推定**されていたことに気付くであろう。したがって、平均値補定は、一見すると容易で合理的な方法と思えるが、実際には、データの分布を大幅に歪めてしまうため、実用には適さない。

図 2.1：正データのヒストグラム

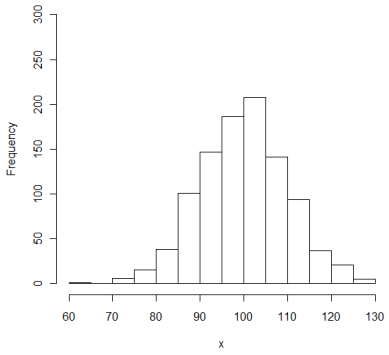


図 2.2：補定済データのヒストグラム

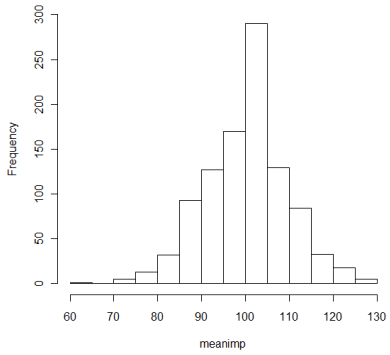
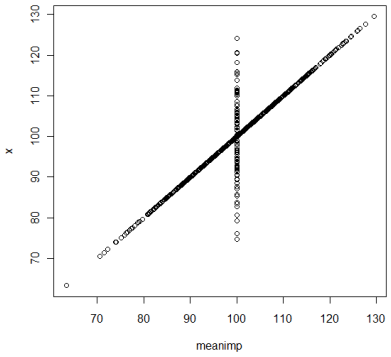


図 2.3：正データと補定済データの散布図



2.3　コールドデック補定

コールドデック(Cold Deck)補定は、欠測値を外部情報又は同一調査の前期データの値に置き換える手法である。たとえば、表 2.4 のように、ID 番号 10 の月収を知りたいとしよう。本人に直接聞いたところ、拒否されたとする。税務署には表 2.5 のような税務データが存在し、仮に表 2.5 を閲覧することができたとすれば、この値をコールドデック補定値として使用できる。コールドデック補定済データは、表 2.6 のとおりであり、このデータを用いて 10 人の収入の平均値を求めるならば、(40.5 + 43.1 + 36.9 + 39.3 + 50.1 + 39.9 + 44.3 + 38.9 + 35.5 + **46.6**)/10 = 41.51 となり、数値として算出でき、統計分析を行える。

表 2.4：生データ 2

	D	K
id	月収	Ky
1	40.5	1
2	43.1	1
3	36.9	1
4	39.3	1
5	50.1	1
6	39.9	1
7	44.3	1
8	38.9	1
9	35.5	1
10	46.6	0

表 2.5：税務情報

	D	K
id	月収	Ky
1	40.5	1
2	43.1	1
3	36.9	1
4	39.3	1
5	50.1	1
6	39.9	1
7	44.3	1
8	38.9	1
9	35.5	1
10	**46.6**	1

表 2.6：コールドデック補定済データ

	D	K
id	月収	Ky
1	40.5	1
2	43.1	1
3	36.9	1
4	39.3	1
5	50.1	1
6	39.9	1
7	44.3	1
8	38.9	1
9	35.5	1
10	**46.6**	1

コールドデック補定は、同一人物や同一ユニットの過去の値を用いるため、直感的に合理的な手法だと感じられるが、コールドデック補定値を完全な真の値と同一に扱うことは問題視されている(Little and Rubin, 2002, pp.60-61)。今回の収入の例のように、外部データと現在のデータの間に大幅な相違が予想されない場合には比較的よい手法だと言えるが、たとえば、前年の収入データの場合は、今年のデータと異なっている可能性が高いと想像されるので、**不確実性**についての議論が欠かせない。また、2012 年 12 月現在、我が国では、統計調査の代わりとして、税務情報を使用してはいけないことになっており[9]、この種のコールドデック補定は行えない。さらに、今回の調査で回答しなかった人は、前回調査や他の調査でも回答していない可能性が高いと想像でき、どこまでさかのぼっても欠測している可能性も否定できない。加えて、今回の調査と他の調査では、用語の定義、測定の尺度や精度など、様々な条件が異なっている可能性があり、互換性にも注意が必要である。

2.4　ホットデック（ドナー）補定

ホットデック(Hot Deck)補定は、欠測値をデータ内の似通った回答ユニットの値に置き換える方法である。この手法は、マッチングとも呼ばれ、変数 *Y* に欠測が存在するユニットに関して、観測データの *X* の値が似通っているユニットを探し出し、そのユニットの *Y* の値を用いるものである(Gelman and Hill, 2006, p.538)。このように、*Y* の値を提供するユニットのことをドナーと呼ぶ。月収の例で言えば、以下の表 2.7 に示すデータセットにおいて、ID 番号 10 の回答者に最も似通っているのは、ID 番号 7 である。よって、ID 番号 7 がドナーとなり、その月収の値 44.3 を ID 番号 10 の月収の補定値として使用する。すなわち、補定済データは、表 2.8 のとおりとなる。この補定済データセットに基づく 10 人の平均月収は、(40.5 + 43.1 + 36.9 + 39.3 + 50.1 + 39.9 + 44.3 + 38.9 + 35.5 + **44.3**)/10 = 41.28 となり、数値として算出することが可能であり、統計分析を行える。

[9] http://www.stat.go.jp/data/e-census/2012/qa1.htm#a11 (2012 年 12 月 20 日アクセス)

表 2.7：生データ 3

ID	月収	年齢	性別
1	40.5	42	女
2	43.1	37	男
3	36.9	36	男
4	39.3	34	男
5	50.1	46	男
6	39.9	44	女
7	**44.3**	**40**	**男**
8	38.9	35	女
9	35.5	35	女
10	46.6	**40**	**男**

表 2.8：ホットデック補定済データ

ID	月収	年齢	性別
1	40.5	42	女
2	43.1	37	男
3	36.9	36	男
4	39.3	34	男
5	50.1	46	男
6	39.9	44	女
7	**44.3**	40	男
8	38.9	35	女
9	35.5	35	女
10	**44.3**	40	男

マッチングは、ノンパラメトリックな回帰補定であり、回帰モデルを設定することが困難な状況における代替案として使用することが考えられる。最尤法(MLE: Maximum Likelihood Estimation)に基づく傾向スコア(Propensity Score)を算出し、ドナーを探索する方法が推奨されている(Gelman and Hill, 2006, p.538)。しかし、どれほど「似ている」としても、同一ユニットではない以上、欠測値に対応する真値は同一の値とは限らず、ホットデック補定値は、偶然にも近い値であるかもしれないが、偶然にも外れた値である可能性があり、コールドデックの場合と同様に、**不確実性**に関する議論が必須である[10]。また、データセット内に似通ったユニットが存在しない可能性もあり得る。

2.5 確定的回帰補定

回帰補定は、欠測値を回帰曲線より算出された予測値に置き換える手法である。今回は、ID 番号 8 及び ID 番号 9 の 2 箇所に欠測のある表 2.9 のデータを用いた例を示す。

表 2.9：生データ 4

ID	月収	年齢
1	40.5	42
2	43.1	37
3	36.9	36
4	39.3	34
5	50.1	46
6	39.9	44
7	44.3	40
8	38.9	35
9	35.5	35
10	46.6	40

表 2.10：回帰補定済データ

ID	月収	年齢
1	40.5	42
2	43.1	37
3	36.9	36
4	39.3	34
5	50.1	46
6	39.9	44
7	44.3	40
8	**39.7**	35
9	**39.7**	35
10	46.6	40

一般的に、年功序列にしたがって、年齢が高い人ほど月収も高いと合理的に想定できるので、月収を被説明変数とし、年齢を説明変数とする以下の単回帰モデルにより補定を行って

[10] この点に関し、Cranmer and Gill (2012)において多重ホットデック補定が開発されている。

みることとする。まず、観測されている ID 番号 1 から 7 までと 10 のデータを用いて、以下の式(2)のパラメータαとβを最小二乗法(OLS: Ordinary Least Squares)により推定し、これら推定されたパラメータ値を用い、補定対象となっている ID 番号 8 の年齢のデータを代入し、ID 番号 8 の月収を推定する（なお、ID 番号 9 の月収も同様の手順で推定する）。補定済データセットは、表 2.10 のとおりとなる。

$$\widehat{月収}_i = \hat{\alpha} + \hat{\beta}\, 年齢_i \tag{2}$$

$$\widehat{月収}_i = 19.104 + 0.587\, 年齢_i$$

$$\widehat{月収}_8 = 19.104 + 0.587 * 35 = 39.652$$

したがって、10 人の平均月収は、(40.5 + 43.1 + 36.9 + 39.3 + 50.1 + 39.9 + 44.3 + 39.7 + 39.7 + 44.3)/10 = 41.78 となり、数値として算出でき、統計分析を行える。回帰分析の予測値を補定値として用いることは、未知の値の予測を行う回帰分析の本来の目的に合致していると考えられるが、一般的に、回帰補定の値は、**変動（ばらつき）を過小推定**する傾向にある(de Waal *et al.*, 2011, p.231)。表 2.10 に示されているとおり、複数の欠測値が同一の X の値を持つ場合、補定値も決定論的に同一になってしまい、**補定内不確実性**を考慮に入れていないからである。つまり、ID 番号 8 の人物と ID 番号 9 の人物は、偶然にも完全に同じ月収である可能性はゼロではないが、現実的にはその可能性は低いと想像される。補定すべき真値が、ID 番号 8 と ID 番号 9 の間で異なっている可能性が高いと想定できるにもかかわらず、補定値は完全に同一となっている点が問題なのである。さらに、1.1 節で示したとおり、真のαは 1 であり、真のβも 1 であったが、欠測値が存在しているため、補定に用いるモデルの係数（$\hat{\alpha}$と$\hat{\beta}$）は、真の係数と同一ではなく、**推定不確実性（補定モデル間の不確実性）**も考慮に入れる必要がある。ゆえに、コールドデック及びホットデックの場合と同様に、不確実性に関する議論が必須である。

2.6　確率的回帰補定

不確実性を導入する手法として、単一代入法による補定値にランダムノイズを加味する確率的補定[11]が知られている(Little and Rubin, 2002, p.60; Allison, 2002, pp.28-29; 西郷, 2010, p.3)。確定的回帰補定の場合と同じく、ID 番号 8 及び ID 番号 9 の 2 箇所に欠測のある表 2.9 のデータを用い、観測されている ID 番号 1 から 7 までと 10 のデータに基づいて、式(2)により、パラメータαとβを最小二乗法(OLS)により推定する。その後、式(3)のとおり、OLS の残差を計算し、その標準偏差を算出する。式(4)に示すとおり、標準正規乱数 e に OLS の残差の標準偏差を掛け合わせ、各々の補定値に加えることで不確実性を反映させる。このようにすること

[11] この手法は、Stochastic Imputation や Random Imputation の名前で知られている。また、確率的補定と対象的に、2.5 節で紹介した手法は、確定的補定(Deterministic Imputation)と呼ばれる。

で、表 2.11 に示すとおり、ID 番号 8 と ID 番号 9 の補定値を異なったものとすることができる。

$$\widehat{月収}_i = \hat{\alpha} + \hat{\beta}\, 年齢_i \tag{2}$$

$$\hat{u}_i = 月収_i - \hat{\alpha} - \hat{\beta}\, 年齢_i \tag{3}$$

$$\widehat{月収}_i = \hat{\alpha} + \hat{\beta}\, 年齢_i + \sigma_{\hat{u}_i} e_i \tag{4}$$

$$\widehat{月収}_i = 19.104 + 0.587\, 年齢_i + \sigma_{\hat{u}_i} e_i$$

$$\widehat{月収}_8 = 19.104 + 0.587 * 35 + 3.67 * (-0.79) = 36.753$$

$$\widehat{月収}_9 = 19.104 + 0.587 * 35 + 3.67 * (-0.91) = 36.313$$

表 2.9：生データ 4

ID	月収	年齢
1	40.5	42
2	43.1	37
3	36.9	36
4	39.3	34
5	50.1	46
6	39.9	44
7	44.3	40
8	38.9	35
9	35.5	35
10	46.6	40

表 2.11：確率的補定済データ

ID	月収	年齢
1	40.5	42
2	43.1	37
3	36.9	36
4	39.3	34
5	50.1	46
6	39.9	44
7	44.3	40
8	**36.8**	35
9	**36.3**	35
10	46.6	40

確率的補定では、補定内の不確実性を反映させることはできるが、補定モデルが 1 つしか存在しないため、補定モデル間の不確実性、すなわち、推定不確実性を反映させることができない。したがって、確率的補定による補定データを用いた統計分析では、標準誤差が不正確となり、誤った統計的推論を導くおそれがある(Little and Rubin, 2002, pp.65-66; Allison, 2002, p.29; Gelman and Hill, 2006, p.542)。

2.7 まとめ

リストワイズ除去が単純に情報を捨て去ってしまうのに対し、単一代入法ではデータ内に存在している情報を活かしており、単一代入法は欠測率が 5%未満の場合には、総じて、良好な手法と言える(Schafer, 1999, p.7)。しかし、単一代入法では、本来ならば未知であるはずの欠測値を、たった 1 つの値に置き換えることによって、あたかも既知であるかのごとくに扱ってしまう点が問題である(Rubin, 1987, pp.12-13)。したがって、欠測率が高くなればなるほ

ど、単一代入法では**変動の過小推定**となり、分散や標準偏差に基づく指標に偏りが生じてしまうため、十分な手法ではない。これを補う方法として、最近では**多重代入法**が推奨されている。多重代入法では、補定内分散と補定間分散を考慮に入れることによって、単一代入法の問題を克服している。

3　先行研究に占める本研究の貢献

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、統計センターにおいても、データエディティングの一環として補定に関する研究を盛んに行ってきた。また、国連欧州経済委員会(UNECE)の統計データエディティングに関するワークセッションでは、エディティング及び補定(E&I: Editing and Imputation)として、補定に関する研究論文が盛んに公開されている。2000年から2011年までに報告された約300論文を調査したところ、多重代入法に関する論文が5篇あることが分かった。本節では、統計センター及び国連欧州経済委員会(UNECE)の統計データエディティングに関するワークセッションでの、補定に関する主な研究を簡潔に紹介し、本研究の貢献を示す。

3.1　統計センターにおける補定に関する研究

西郷 (2004)では、標本抽出メカニズムと欠測メカニズムとの関連及び補定のもとでの補定値の精度評価の議論を行っており、中でも多重代入データのためのリサンプリング法として、多重代入法の概要が簡潔に紹介されている。村田, 畠山, 磯部, 亀本 (2008)では、平成16年サービス業基本調査のデータを用いて、経理項目に対する回帰補定の改善の可能性を探求し、経済センサスの経理項目補定への応用を目指した。西郷 (2010)では、補定方法の最近の発展として、二重に保護された確率的補定を紹介している。また、伊藤 (2011)では、平成24年経済センサス‐活動調査に向けて、実務的な補定法に関する検討を行った。さらに、和田 (2012)では、繰返し加重最小二乗法(IRLS: Iteratively Reweighted Least Squares)に基づき、企業財務データにおける外れ値の影響を自動的に制御し、安定した補定値を得られるアルゴリズムを示した。

3.2　UNECE ワークセッションにおける多重代入法の報告

米国保健医療統計センター(National Center for Health Statistics)の Harris (2002)は、全米健康・栄養調査(National Health and Nutrition Examination Survey)における多重代入法の応用可能性について検討を行った。その結果、単一代入法と比較して、優れた点推定値を算出できる可能性があることが分かった。一方、オーストリア統計局の Burg (2008)は、四半期ごとに行われる労働調査における失業率の補定の研究を行った。Burg (2008)によれば、リストワイズ除去法による失業率の推定値と多重代入法による失業率の推定値との間には、有意な差は認

められなかった[12]。このように、相反する結論が出ているが、これら予備的研究を発展させる研究は、現在に至るまでの 10 年間、行われていない。

ドイツ雇用調査局(IAB: Institute for Employment Research)の Bender *et al.* (2006)は、多重代入法を従来の欠測値補定法として使用するのではなく、開示リスクの軽減法として使用した。つまり、欠測値を補定する代わりに、開示リスクの高い観測値を多重代入値に置き換えるということである。この考え方は、開示リスク軽減の文脈において革新的だと思われ、非常に興味深いが、本稿が対象としている欠測値対処法としての多重代入法とは異なる文脈である。ドイツ雇用調査局の Drechsler (2009)は、汎用ソフトウェアにおいて多重代入法を行う際の注意点を丁寧に解説している：すなわち、分布に偏りのある連続変数の補定；非連続変数の補定；非負制約；線形制約である。Honaker *et al.* (2012a, pp.16-19, pp.26-28)に示されているとおり、本稿で利用する R の多重代入法パッケージ Amelia の最新版では、こういった問題を十分に扱うことができる。また、ドイツ連邦統計局の Schmidt (2009)は、SAS の多重代入法プログラムである IVEware を使用した感想を紹介しており、SAS に関して実用的で有用な内容となっているが、現在のところ、フリーソフトウェアである R を用いた多重代入法に関する議論はされていない。

3.3 まとめ

統計センターにおいてはエディティングの一環として補定の研究を盛んに行い、実務的な手法の研究をしてきた。しかし、これまでのところ、多重代入法に関する研究は、西郷 (2004)においてわずかに触れられているものの、本格的には取り扱っていない。したがって、本稿はこうした穴を埋める目的を持つものであり、日本語における多重代入法に関する文献として、公的統計の精度向上に資するものである。

さらに、国連欧州経済委員会(UNECE)の統計データエディティングに関するワークセッションにおいても、300 ある論文の中、多重代入法を扱った論文は 5 つしかなく、さらに、上述したとおり、それら 5 論文においても多重代入法による欠測値補定の精度評価に関して十分な議論が行われたとは言えない状況である。この点に関する貢献として、2012 年統計データエディティングに関するワークセッションにおいて、本研究の縮小版の報告を行った。詳細については、Takahashi and Ito (2012)を参照されたい。

4 多重代入法（Multiple Imputation）概論

早くも 1970 年代後半には、ハーバード大学統計学科の Donald B. Rubin (1978)により多重代入法の理論が提唱されていた。しかし、単一代入法と比較して、多重代入法には数倍（= *M* 倍）の作業量が必要になるという欠点があった(Rubin, 1987, pp.17-18)。したがって、コンピュータが発展段階にあった 1980 年代から 1990 年代まで、実務家にとっての多重代入法は高嶺

[12] 2.1 節で示したとおり、リストワイズ除去法と多重代入法との間に有意な差がなかったとすれば、それは、偶然にも欠測が完全にランダム(MCAR)であった可能性が指摘できる。一般的に欠測が完全にランダムである可能性は低く、Burg (2008)の結果を一般化して考えることは難しいであろう。

の花であり、理論的に存在していることは知られていても、実用には適さないものであった。日本語において多重代入法を解説したものとして、岩崎 (2002, pp.309-314)、星野 (2009, pp.219-223)、野間, 田中 (2012, pp.82-86)が推奨できるが、いずれも紙面に限りがある。4.1 節では、Rubin の提唱した多重代入法を、具体例を交えながら詳説する(Rubin, 1987, pp.15-22, pp.75-81; Little and Rubin, 2002, pp.85-89)。4.2 節では、本研究で使用した R の多重代入法パッケージ Amelia の多重代入法モデルを詳説する(King *et al.*, 2001, pp.53-54; Honaker and King, 2010, pp.576-578; Honaker, King, and Blackwell, 2011, pp.3-4)。

4.1　Rubin の多重代入法の例示

Rubin による多重代入法は、モンテカルロ法に基づき、欠測値を M 個($M > 1$)のシミュレーション値に置き換えるものであった。各々の欠測値を M 個の値に置き換え、M 個の補定済データセットを作成する。これらの補定済データセットにおいて、観測値はすべて同一であるが、欠測値は不確実性を反映し、異なった値となっている。多重代入法とは、すなわち、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、この分布から無作為抽出を行って、複数個の補定データを生成するのである(Schafer, 1999; King *et al.*, 2001, p.53; Gill, 2008, p.324)。表 4.1 は、表 2.9 のデータを用い、$M = 3$ の多重代入法を行った実例[13]である。

表 4.1：多重代入済データセットの例

ID	月収	年齢	**補定 1**	**補定 2**	**補定 3**
1	40.5	42	40.5	40.5	40.5
2	43.1	37	43.1	43.1	43.1
3	36.9	36	36.9	36.9	36.9
4	39.3	34	39.3	39.3	39.3
5	50.1	46	50.1	50.1	50.1
6	39.9	44	39.9	39.9	39.9
7	44.3	40	44.3	44.3	44.3
8	38.9	35	**38.4**	**44.1**	**40.5**
9	35.5	35	**38.2**	**39.3**	**43.7**
10	46.6	40	46.6	46.6	46.6

ここで注目すべきは、ID 番号 8 の月収に関する補定値は、1 回目(= **38.4**)、2 回目(= **44.1**)、3 回目(= **40.5**)のそれぞれにおいて異なった値となっており、補定間の不確実性が補定間分散として考慮に入れられている。また、ID 番号 8 と ID 番号 9 の補定値は、単一代入法では 39.7 と同一になってしまっていたが、多重代入法では異なった値となっており、補定内不確実性が補定内分散として考慮に入れられている。

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用いて分析を行う。つまり、各々の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

[13] これらの数値は、年齢を説明変数とし、月収を被説明変数として、実際に Amelia を用いて多重代入法を行った結果である。ただし、現実には、多重代入法を行うには、観測数 8 個では少ないことを断っておく。表 4.1 は、多重代入法を視覚的に例示するための具体例として使用しており、この文脈では、補定の精度を度外視している点をご了承願いたい。

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(5)のとおりであり、$\hat{\theta}_m$の単純な算術平均である。

$$\bar{\theta}_M = \frac{1}{M}\sum_{m=1}^{M}\hat{\theta}_m \tag{5}$$

すなわち、多重代入法では、観測値をもとに欠測値のシミュレーション値を算出し、乱数によるノイズを加えながら同一手順を複数回繰り返し、これらの複数の推定値の平均値を補定値として使用する(Shadish, Cook, and Campbell, 2002, p.337)。M 個の補定値の平均を取ることにより、単一代入法と比較して、推定値の効率性を増加させることができる[14]。

表 4.1 を使用して例示すれば、ID 番号 8 の月収の補定の点推定値は(38.4 + 44.1 + 40.5)/3 = **41.0** ということになる。一方、ID 番号 9 の月収の補定の点推定値は(38.2 + 39.3 + 43.7)/3 = **40.4** ということになる。

仮に、表 4.1 のデータを用いて、年齢を説明変数として月収を説明する単回帰モデルに興味があるとしよう。回帰分析の結果は表 4.2 のとおりである。ここで、補定 1 モデルは補定 1 データセットを用いた回帰分析の結果であり、補定 2 モデルは補定 2 データセットを用いた回帰分析の結果であり、補定 3 モデルは補定 3 データセットを用いた回帰分析の結果である。

表 4.2：多重代入済データを用いた回帰分析の例

	補定 1 モデル		補定 2 モデル		補定 3 モデル	
切片	16.140	(10.884)	23.569	(11.662)	24.380	(11.432)
傾き	0.656	(0.278)	0.483	(0.298)	0.464	(0.292)
n		10		10		10

注：報告値は係数(標準誤差)の順である。

したがって、統合モデルの切片は(16.140 + 23.569 + 24.380)/3 = **21.363** であり、統合モデルの傾きは(0.656 + 0.483 + 0.464)/3 = **0.534** である。

また、多重代入法による推定値の分散は 2 つの部分から成り立つ。まず、$\hat{\theta}_m$の分散var$(\hat{\theta}_m)$の推定値をv_mとする。補定内分散の平均$\bar{v}_M$は、式(6)のとおりであり、v_mの単純な算術平均である。補定内分散は、使用しているデータが有限の標本であり、無限の母集団ではないために発生する通常の統計的な変動（ばらつき）の指標と同じである。

$$\bar{v}_M = \frac{1}{M}\sum_{m=1}^{M} v_m \tag{6}$$

[14] 単一の標本推定値と比較して、モンテカルロやブートストラップなどによる副標本(sub-sample)に基づく推定値は、副標本の数が無限大に近づくにつれて、偏りが少なくなり、一致推定値となるからである。

補定間分散の平均$\tilde{v}_M$は、式(7)のとおりである。$\bar{\theta}_M$を算出する際に自由度を 1 つ失っているため、$\tilde{v}_M$の算出では$M-1$によって自由度を調整している。補定間分散は、標本データ内に欠測値が存在しているという事実を考慮したものである。

$$\tilde{v}_M = \frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{\theta}_m - \bar{\theta}_M\right)^2 \tag{7}$$

$\bar{\theta}_M$の分散T_Mは、式(8)のとおりである。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。ここで、$(1+1/M)$は、M のサイズが有限であるために調整を施す項である[15]。すなわち、$\tilde{v}_M/M$は、式(5)が有限のサイズの M に基づいているために起きるシミュレーションエラーである。

$$T_M = \bar{v}_M + \left(1+\frac{1}{M}\right)\tilde{v}_M \tag{8}$$

表 4.2 の例に戻ると、統合モデルの傾き **0.534** に対応する標準誤差は以下のとおり求められる。まず、式(6)に、各々の回帰係数の分散を代入し、$\bar{v}_M$を算出する。ここで、回帰係数の分散は、標準誤差の二乗であることに注意して$\bar{v}_M$を算出する。

$$\bar{v}_M = \frac{0.278^2 + 0.298^2 + 0.292^2}{3} = 0.084$$

次に、式(7)に各々の回帰係数と式(5)より求めた統合後の回帰係数の点推定値との差の二乗の和を代入する。

$$\tilde{v}_M = \frac{(0.656-0.534)^2 + (0.483-0.534)^2 + (0.464-0.534)^2}{3-1} = 0.011$$

最後に、$\bar{v}_M$の値と$\tilde{v}_M$の値を式(8)に代入し、T_Mは回帰係数の分散なので、平方根を取ることにより、回帰係数の標準誤差が算出できる。

$$\sqrt{T_M} = \sqrt{0.084 + \left(1+\frac{1}{3}\right)0.011} = \sqrt{0.099} = 0.315$$

このようにして算出した回帰係数の標準誤差の点推定値は、単純な算術平均([0.278 + 0.298

[15] 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。

+ 0.292]/3) = 0.289 とは異なる値となることに注意したい。

4.2 多重代入法モデル

理論上、多重代入法においては、欠測のメカニズムとして、MAR を前提とすることは必須ではないが(Schafer, 1999, p.8)、現実的には、本研究で使用するものを含めて、多くのアルゴリズムにおいて MAR が前提とされている。また、当然、単純に単一代入法を M 回繰り返しても、同じ補定値が M 個算出されるだけであり、実際に多重代入法を行うには、適切な事後確率分布を構築し、M 個の異なった補定値を算出しなければならない。この目的のために何らかの統計モデルを想定する必要があり、想定され得る統計モデルは多岐にわたるが、最も汎用性が高いとされているのは多変量正規分布モデルである。

Dを$n \times p$データセットとする（nは標本サイズ、pは変数の数とする）。具体的には、もしDが $n = 3, p = 2$ のデータセットであれば、式(9)のとおりとなる。

$$D = \begin{bmatrix} Y_1 & X_1 \\ Y_2 & X_2 \\ Y_3 & X_3 \end{bmatrix} \tag{9}$$

Dは、もし欠測値がないならば、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。すなわち、$D \sim N_p(\mu, \Sigma)$ということであり標本サイズ 3 個の二変量データセットの文脈では、μは式(10)のとおりである。

$$\mu = [\mu_Y \quad \mu_X] \tag{10}$$

ここで、

$$\mu_Y = \bar{Y} = (Y_1 + Y_2 + Y_3)/3$$
$$\mu_X = \bar{X} = (X_1 + X_2 + X_3)/3$$

また、Σは式(11)のとおりである。

$$\Sigma = \begin{bmatrix} \mathrm{var}(Y) & \mathrm{cov}(Y,X) \\ \mathrm{cov}(X,Y) & \mathrm{var}(X) \end{bmatrix} \tag{11}$$

ここで、

$$\mathrm{var}(Y) = E(Y - \mu_Y)^2 = \frac{\sum(Y_i - \bar{Y})^2}{N}$$

$$\mathrm{var}(X) = E(X - \mu_X)^2 = \frac{\sum(X_i - \bar{X})^2}{N}$$

$$\mathrm{cov}(Y,X) = \mathrm{cov}(X,Y) = E\{(X-\mu_X)(Y-\mu_Y)\} = \frac{\sum(X_i-\bar{X})(Y_i-\bar{Y})}{N}$$

多変量正規分布を想定しているので、欠測値は線形的に補定される[16]。したがって、回帰分析と同様のやり方で、補定値を算出する。観測値 i 及び変数 j の値、すなわちD_{ij}が欠測しているとしよう。$\tilde{D}_{ij}$を観測値 i 及び変数 j のシミュレーション値、すなわち補定値とする。$D_{i,-j}$を、変数 j を除く i 行のすべての観測値とする。**補定値$\tilde{D}_{ij}$は、式(12)により算出**する。ここで、～は適切な事後分布からの無作為抽出であることを示している。したがって、$\tilde{D}_{ij}$の無作為抽出値は、観測される他の変数$D_{i,-j}$の 1 次関数である。**$\tilde{\beta}$は推定不確実性**、つまり補定間の不確実性を表している。**$\tilde{\varepsilon}_i$は根本的な不確実性**、すなわち補定内の不確実性を表している[17]。

$$\tilde{D}_{ij} = D_{i,-j}\tilde{\beta} + \tilde{\varepsilon}_i \tag{12}$$

上述したとおり、欠測値は線形的に補定される。$\tilde{\beta}$をどのようにして推定するかを示すために、通常の最小二乗法(OLS)による回帰分析におけるβの推定過程を参考として示す。行列形式における回帰モデルは式(13)であり、行列形式における回帰係数の算出は式(14)により行われる。

$$\underset{(n\times 1)}{\mathbf{y}} = \underset{(n\times k)}{\mathbf{X}}\ \underset{(k\times 1)}{\boldsymbol{\beta}} + \underset{(n\times 1)}{\mathbf{u}} \tag{13}$$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{14}$$

ここで、二変量の文脈[18]であれば、上記の式(13)と(14)は、それぞれ式(15)、(16)、(17)のとおりとなる。

$$Y_i = \beta_1 + \beta_2 X_i + u_i \tag{15}$$

$$\hat{\beta}_2 = \frac{\sum(X_i-\bar{X})(Y_i-\bar{Y})}{\sum(X_i-\bar{X})^2} \tag{16}$$

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2\bar{X} \tag{17}$$

したがって、式(11)に示した情報Σを利用して、回帰係数$\hat{\beta}_2$を以下の式(18)のとおり、X の分散

[16] ただし、回帰分析におけるのと同様に、対数変換や指数変換などを行って変数を変換させることにより、非線形のデータ分布にもおおむね対応できる。

[17] Amelia における多重代入法で欠測補定時に加味される不確実性$\tilde{\varepsilon}_i$の特性について、不確実性$\tilde{\varepsilon}_i$のみに影響される欠測補定値がどのような分布となるか確認した。$\tilde{\varepsilon}_i$は、ランダムなガウスノイズ（正規分布に基づく乱数）であると結論付けられることが分かった。

[18] 多変量の文脈に関しては、竹村，谷口(2003, p.24)を参照されたい。

及び X と Y の共分散に基づいて算出することができる。

$$\hat{\beta}_2 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})/(n-1)}{\sum(X_i - \bar{X})^2/(n-1)} = \frac{\mathrm{cov}(X,Y)}{\mathrm{var}(X)} \tag{18}$$

また、式(10)に示した情報μを利用して、回帰係数$\hat{\beta}_1$を以下の式(19)のとおり、X と Y の平均値に基づいて算出することができる。

$$\hat{\beta}_1 = \mu_1 - \hat{\beta}_2 \mu_2 \tag{19}$$

すなわち、回帰係数を算出するために必要な情報は、平均値及び分散・共分散の情報であり、これらの情報はμとΣにすべて含まれていることが分かる。つまり、補定の文脈において、もしμとΣが完全に分かっているならば、D_jに基づく真の回帰係数βを決定的に算出することができ、欠測値を決定的に補定することができるのである。

ここで、データが完全である場合の尤度関数は式(20)のとおりとなる。すなわち、完全データDが与えられたときの平均値ベクトルμと分散共分散行列Σの尤度は、平均値ベクトルμと分散共分散行列Σが与えられたときの各々のデータの正規分布の積に比例する。

$$L(\mu, \Sigma | D) \propto \prod_{i=1}^{n} N(D_i | \mu, \Sigma) \tag{20}$$

しかし、現実にはデータ内に欠測値が存在している。観測データD_oの尤度を形成する際に、MAR を前提とする。つまり、1.3 節で見たとおり、$P(K|D) = P(K|D_o)$である。Dの i 行の観測値を$D_{i,o}$とし、$\mu_{i,o}$をμのサブベクトルとし、$\Sigma_{i,o}$をΣのサブ行列とする。また、$\mu_{i,o}$及び$\Sigma_{i,o}$は、i ごとに変化しない。周辺密度は正規であるので、観測データD_oの尤度は式(21)のとおりである。すなわち、観測データD_oが与えられたときの平均値ベクトルμと分散共分散行列Σの尤度は、平均値サブベクトル$\mu_{i,o}$と分散共分散サブ行列$\Sigma_{i,o}$が与えられたときの各々の観測データの正規分布の積に比例する。

$$L(\mu, \Sigma | D_o) \propto \prod_{i=1}^{n} N\left(D_{i,o} | \mu_{i,o}, \Sigma_{i,o}\right) \tag{21}$$

4.3　まとめ

μとΣは完全には分からないためにβを確実に知ることができない。$\tilde{\beta}$は、この推定不確実性が存在することを表し、これは補定間の不確実性を表している。さらに、$\tilde{\varepsilon}_i$は根本的な不確実性を表し、補定内の不確実性を表している。これは、Σがゼロ行列ではないためである。つまり、現実世界に根本的に存在している不確実性である。すなわち、多重代入値$\tilde{D}_{ij}$のランダム性は、βを確実に知ることができない推定不確実性と$\tilde{\varepsilon}_i$による根本的な不確実性の 2 つから成り立つのである。

算出上の問題として、伝統的な手法によって式(21)を算出することができず、したがって、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。これを解決する手段として、次節で説明する EMB アルゴリズムを用いる。

5　EMB(Expectation Maximization with Bootstrapping)アルゴリズム

Rubin (1978)による多重代入法は、モンテカルロ法に基づくものであり、既存の多くの多重代入法プログラムは、ベイズ統計学の枠組みの中で、マルコフ連鎖モンテカルロ法(MCMC: Markov Chain Monte Carlo)に基づいていることが多い[19]。一方、本研究で使用する R の多重代入法パッケージ Amelia では、ブートストラップを応用した EM アルゴリズムの一種を用いている。本節では EM アルゴリズム及びブートストラップを概観し、EMB アルゴリズムについて説明を行う[20]。

5.1　Expectation Maximization (EM) アルゴリズム

調査データのすべてが得られていないとき、得られなかった部分を含む全データを不完全データと呼ぶ。欠測値を含むデータは、まさしく不完全データの最たるものであり、現実に存在するほとんどのデータは不完全データであると言える。不完全データを完全なものにするためには、平均や分散といった分布に関する情報が必要となるが、その平均や分散を推定するために不完全データを使うことになり、鶏と卵の問題となってしまい、解析的には解決することが困難である。これを解決するために、何らかの手段により初期値を定め、そこから繰り返し法を用いて推定を行う方法が提唱されてきた。その代表例が EM (Expectation Maximization：期待値最大化)アルゴリズムである。

EM アルゴリズムでは、Expectation ステップにおいて期待値計算を行い、Maximization ステップにおいて尤度最大化計算を行う。これは、不完全データに基づいて最尤推定値(MLE)を導く一般的なアルゴリズムである。通常、EM アルゴリズムでは、ある種の分布を仮定して、平均値や分散の初期値を仮に設定する。この初期値に基づいてモデル尤度の期待値を計算し、尤度の最大化計算を行い、得られた期待値を最大化するパラメータを推定し、分布を

[19] マルコフ連鎖モンテカルロ法に基づく多重代入法については、野間, 田中 (2012, pp.84-85)を参照されたい。
[20] EMB アルゴリズムの理論的根拠については、Rubin (1987, pp.192-195)及び Little and Rubin (2002, p.216)を参照されたい。

更新する。期待値計算及び最大化計算を繰り返し、最終的に収束した値が最尤推定値(MLE)である(渡辺, 山口, 2000, pp.32-35; Gill, 2008, p.309)。こうして繰り返すことで収束した値は、局所的最大値であることが証明されている。しかし、複数の峰のある分布では、局所的最大値が大局的最大値であるとは限らないので、複数の初期値から EM アルゴリズムを行い、複数の収束した値の中から最も大きいものを選ぶ必要がある[21](中村, 小西, 1998, p.168; 渡辺, 山口, 2000, p.40)。

5.2 モンテカルロ・シミュレーション vs.ブートストラップ・リサンプリング

θを母集団パラメータとする。モンテカルロ法では、まずシミュレーションの対象となる母集団の情報を決める。すなわち、母集団におけるθの値、母集団の分布、標本サイズ n を設定し、そこから標本サイズ n の標本を無作為に抽出し、θの推定値を算出し、この作業を M 回繰り返す。$\hat{\theta}_m$を m 回目のシミュレーションで得られたθの推定値だとしよう（ここで、$m = 1, \dots, M$である）。こうして得られた標本平均を使用して$\mathrm{E}(\hat{\theta})$を、また標本分散を使用して$\mathrm{var}(\hat{\theta})$を推定できる。しかし、典型的なモンテカルロ法では、母集団分布の仮定をしなければならず、その仮定の根拠はしばしば希薄である。また、モンテカルロ法により得られたシミュレーション結果は、この特定の分布にのみ当てはまるものであり、それ以上の結論を導き出すことはできず、分布の仮定が誤っていた場合には、得られた結論も誤りとなる。すなわち、モンテカルロ法は、特定の分布において、漸近理論の近似値がどのようになるかを調べるためには有用だが、特定の標本が与えられた場合の推計に関しては有用性が低い(Wooldridge, 2002, pp.377-378)。

ブートストラップは、非常に一般的なリサンプリング手法であり、漸近理論の近似に対する代替法として使用できる。その目的は、1 次漸近論に頼ることなく、$\hat{\theta}$の分布を推定することである。ブートストラップには、様々な種類があるが、本節では、Amelia に応用されているノンパラメトリック・ブートストラップ[22]を紹介する(Wooldridge, 2002, p.379; Shao and Tu, 1995, pp.9-15; Shao, 2002, pp.309-310; DeGroot and Schervish, 2002, pp.753-763)。ノンパラメトリック・ブートストラップでは、観測された標本を擬似的に母集団として扱う。すなわち、手元にある標本サイズ n の標本データから、標本サイズ n の副標本(sub-sample)の**無作為な復元抽出**（重複を許す抽出）を M 回繰り返す。たとえば、N = 100 万、平均 100、標準偏差 16 の母集団があり、そこから生成された表 5.1 のような n = 10 の標本データがあるとしよう。標本平均値は 104.4 である。

[21] EM アルゴリズムが大局解に収束したかどうかを検証する手法については、本稿 9 節における過散布初期値(Overdispersed Starting Values)を参照されたい。

[22] ブートストラップ(bootstrap)とは、もともと、ブーツ（boot：長靴）を履く際に、ブーツの口に付いているストラップ（つまみ）を引っ張ることで、他人の力を借りず自力でブーツを履くことを意味する。すなわち、理論や外部の情報に頼らず、手元にあるデータのみで自力で解析を行うということである。パラメトリック・ブートストラップというものも存在するが、以上のとおり、ブートストラップの本来的な意味は、理論的に分布やモデルを仮定せず、手元のデータのみによって自力で解析を行うという意味で、ノンパラメトリックが本来の姿である。

表 5.1：標本データ

ID	1	2	3	4	5	6	7	8	9	10
観測値	107	108	125	75	74	109	111	113	123	97

ノンパラメトリック・ブートストラップでは、表 5.1 のデータセットから標本サイズ 10 個の副標本を無作為に復元抽出し、それを M 回繰り返す。たとえば、表 5.2 は $M = 2$ のブートストラップ副標本の例である。ここで、いくつかの観測値は同一副標本内において重複して抽出されている。

表 5.2：ブートストラップ副標本データの例

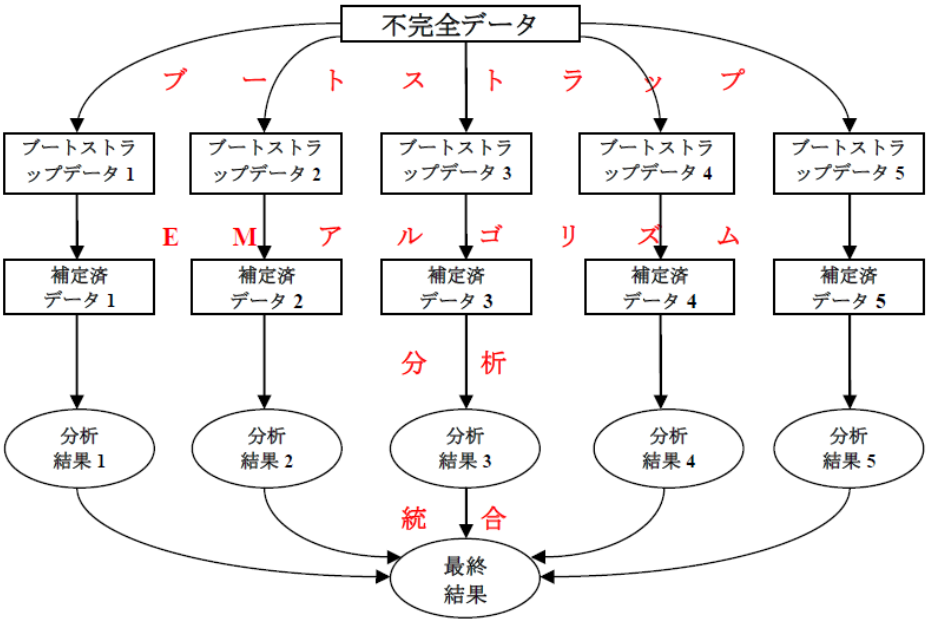
ID	1	2	3	4	5	6	7	8	9	10
副標本 1	75	107	108	75	109	125	97	109	108	113
副標本 2	109	125	109	108	107	97	113	97	111	113

5.3　EMB アルゴリズム

一般に、ベイズ・ブートストラップによる多重代入法では、MAR を前提とし、以下の手順で補定を行う。変数yの標本サイズを n とし、q 個の値が観測され（$y_1, \ldots, y_q$ = 観測値）、$n - q$ 個の値が欠測しているとしよう。欠測データを埋めるために、ノンパラメトリック・ブートストラップの手順に従い、$y_1, \ldots, y_q$から q 個の値を、無作為に復元抽出する。次に、これら q 個の値の中から、補定値$y_{q+1}^*, \ldots, y_n^*$を復元抽出する(Congdon, 2006, p.504)。

Amelia における EMB アルゴリズムを使用した $M = 5$ の多重代入法を概念的に図示すれば、図 5.1 のとおりである。

図 5.1：EMB アルゴリズムを用いた多重代入法の概念図



出典：Honaker, King, and Blackwell (2011, p.4)

まず、観測値(q 個)と欠測値($n-q$ 個)のある不完全データ(標本サイズ n)が存在する。この不完全データをもとに、ノンパラメトリック・ブートストラップにより、標本サイズ n のブートストラップ副標本データを M 個（ここでは 5 個）作成する。たとえば、表 5.3 のような $n=10$ の標本データがあるとしよう（表 5.1 に欠測値 = NA を 1 つ加えたものである）。

表 5.3：標本データ

ID	1	2	3	4	5	6	7	8	9	10
観測値	107	108	125	75	74	109	111	113	123	**NA**

表 5.3 のデータセットから標本サイズ 10 個の副標本を無作為に復元抽出する。ここでは、$M=2$ の場合を考える。ノンパラメトリック・ブートストラップにより得られた副標本は、表 5.4 のとおりだと仮定しよう。

表 5.4：ノンパラメトリック・ブートストラップ副標本

ID	1	2	3	4	5	6	7	8	9	10
副標本 1	123	107	74	74	107	75	111	74	125	123
副標本 2	**NA**	75	108	75	123	123	107	113	109	75

ここで、副標本 1 には、偶然にも欠測値が含まれなかった。この場合は、EM アルゴリズムによりμとΣの点推定値を算出する必要はなく、副標本平均及び副標本分散を伝統的な手法で計算すればよい。しかし、多くの欠測値補定の文脈では、標本データに多数の欠測値があり、副標本 2 のように、副標本データにも欠測値が含まれる可能性が高い。この場合は、伝統的な手法では副標本平均及び副標本分散を算出できないため、得られた副標本データをもとに EM アルゴリズムを行い、μとΣの点推定値を算出する。

図 5.1 に戻ると、5 つのブートストラップ副標本データに EM アルゴリズムを適用し、5 つのμとΣの点推定値に基づいて 5 つの$\tilde{\beta}$を算出し、5 つの式(12)を用いて補定を行い、5 つの補定済データセットを作成する。5 つの補定済データセットを別々に統計分析し、式(5)にしたがって結果を統合し最終結果とする(Honaker and King, 2010, p.565)。

5.4 まとめ

EMB アルゴリズムによる多重代入法では、不完全データをもとに、標本サイズ n のブートストラップ副標本データを M 個作成する。これら M 個のブートストラップ副標本データに EM アルゴリズムを適用し、M 個のμとΣの点推定値を算出し、M 個の式(12)を算出して補定を行い、M 個の補定済データセットを作成する。M 個の補定済データセットを別々に統計分析し、式(5)を用い、結果を統合し最終結果とする。

6　多重代入法の M 数と相対効率

通常のブートストラップにおいては、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であるが、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。この違いは、主に、欠測情報の量にかかわっている。すなわち、通常のブートストラップによる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からである(Honaker and King, 2010, p.565)。

多くの文献において、欠測率が極端に高くない限り、M は 5～10 程度でよいとされている(King *et al.*, 2001, p.53; Gelman and Hill, 2006, p.542; 野間, 田中, 2012, p.84)が、その根拠は示されていないことが多い。また、欠測率が極端に高いとはどの程度なのかも示されていない。したがって、本節では、欠測率が何%のときに、M がどれぐらいあればよいかを数値で示す。

6.1　相対効率：式

有限の M と無限大の M との漸近的相対効率(ARE: Asymptotic Relative Efficiency)は式(22)のとおりである(Rubin, 1987, p.114)。ここで、δは欠測率を表す($0 \le \delta \le 1$)。ARE は%であり、単位は標準偏差である(Schafer 1999, p.7)。M が無限大の場合、式(22)の極限値は 100%となり、効率性が最大に達していることを表す。

$$ARE = \left(\sqrt{1+\frac{\delta}{M}}\right)^{-1} \times 100 \qquad (22)$$

6.2　相対効率：表

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 1 から 50 まで増やした場合に、無限大の M と比較した効率性の結果を表している。たとえば、欠測率 10%($\delta = 0.1$)、M = 5 の場合、相対効率は 99.01%となり、M が無限大の場合と比較して、補定推定値の標準偏差は約 0.99%大きい。M を 20 に増やした場合、相対効率は 99.75%となり、M が無限大の場合と比較して、補定推定値の標準偏差は約 0.25%大きいだけである。また、欠測率 50%($\delta = 0.5$)、M = 5 の場合、相対効率は 95.35%となり、M が無限大の場合と比較して、補定推定値の標準偏差は約 4.65%大きい。M を 20 に増やした場合、相対効率は 98.77%となり、M が無限大の場合と比較して、補定推定値の標準偏差は約 1.23%大きいだけである。

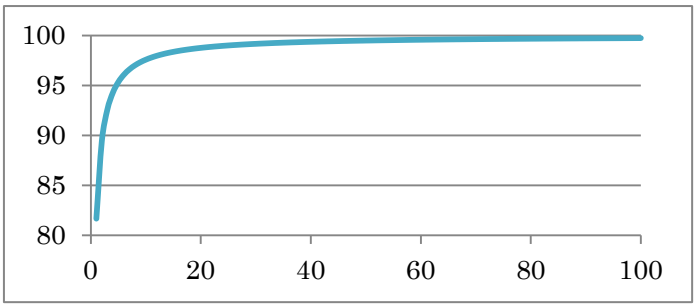
表 6.1：M と相対効率

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
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
30	99.83	99.67	99.50	99.34	99.18	99.01	98.85	98.69	98.53
40	99.88	99.75	99.63	99.50	99.38	99.26	99.14	99.01	98.89
50	99.90	99.80	99.70	99.60	99.50	99.41	99.31	99.21	99.11
∞	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

6.3　相対効率：図

図 6.1 は、欠測率 50%において、横軸に M を 1 から 100 まで増やした場合に、縦軸に得られる相対効率を図示している。

図 6.1：M のサイズと相対効率



6.4　まとめ

M が大きくなればなるほど、ARE は 100%に近づくため、理論的には、M は大きければ大きいほどよいと言える。しかし、M が大きくなればなるほど、実務上の煩雑さが増える。上記のとおり、欠測率 50%の場合、$M = 20$ を超えて得られる相対効率は極めてゼロに近い。本稿では、売上高の欠測率を最大で 50%と想定しており、M は 20 で十分であると結論付けた。

7　R パッケージ Amelia II[23]

7.1　Amelia[24]とは

4 節で見たとおり、早くも 1978 年には、Donald Rubin によって多重代入法の理論が提唱されていた。しかし、その理論上の美点とは裏腹に、計算上の制約が強く、多重代入法は長らく実務に応用されてこなかった。1990 年代後半のある社会科学の分野では、約 94%もの論文において、欠測値の対処法としてリストワイズ除去法が用いられていた。社会科学におけるこうした実情を踏まえて、ハーバード大学の Gary King を中心とするチームにより、汎用多重代入法プログラム Amelia I の開発が行われた(King *et al*., 2001)。以後、10 年の歳月が流れ、Amelia は様々な社会科学の応用分野において使用されてきたが、時系列横断データなどの巨大データセットにおける多重代入に対応するため、新たに EMB アルゴリズムを実装した Amelia II として生まれ変わった(Honaker and King, 2010)。

Amelia II では、主に以下の 2 つの前提を設けている(Honaker, King, and Blackwell, 2011, p.3)。1 つ目は、理論上の真の完全データは、多変量正規で分布しているというものである。多変量正規分布は、データの真の分布の大まかな近似としてよく使用されるものであり、また多くの変数は、変換を行うことにより、正規分布の前提を現実的なものとすることが可能である。2 つ目の前提は、欠測はランダム(MAR)又は完全にランダム(MCAR)を想定している。したがって、欠測が NI である場合には、個別の対処法が必要である[25]。

4 節で見たとおり、多重代入法を行うには、適切な事後分布からμとΣを無作為抽出する必要がある。一般に、μとΣの無作為抽出が難しい理由として、μとΣの要素は$p(p+3)/2$個あり、変数の数 p に応じて急速に増大することが挙げられる。R のパッケージを含めて様々な多重代入法プログラム[26]が存在するが、多くの既存ソフトウェアで採用されているアルゴリズムでは、巨大データセットを扱うことができない。EM アルゴリズムにブートストラップを応用した EMB アルゴリズムを実装する Amelia II では、32,000 観測値、240 変数（合計 720 万観測値-変数）の実データセットの補定を行うことができる。総パラメータ数は 29,160 個であり、これは 4 億 2516 万 7380 個の個別要素を含む 29160×29160 分散共分散行列を反転できることを意味する。したがって、このアルゴリズムで扱える限界サイズは、利用可能なメモリーサイズにのみ依拠している(Honaker and King, 2010, pp.564-565; Honaker, King, and Blackwell, 2011, p.4)。

---

[23] 本稿では、R 2.15.0 を用いた。また、本稿で用いた Amelia II は、Version 1.6.1 である。Amelia II は、下記ウェブサイトから無料でダウンロードし、R に実装して使用可能である：
http://gking.harvard.edu/amelia/ 又は http://cran.r-project.org/web/packages/Amelia/ (Accessed on December 20, 2012)

[24] Amelia とは、20 世紀初頭に活躍し、女性飛行士として史上初めて大西洋単独飛行に成功した Amelia Earhart という米国人女性飛行士の名に因んで命名された。Amelia Earhart は、1937 年 7 月、赤道上世界一周飛行の最中に行方不明となり、その行方の真実はいまだに謎とされている。彼女の華やかな経歴とともに、米国では現在でも伝説視されている女性飛行士である。

[25] 欠測データへの個別の対応については、King *et al*. (2001, pp.65-66)を参照されたい。また、1.5 節において指摘したとおり、MAR と NI の実務上の差は、質の問題と言うよりも程度の問題である。

[26] Amelia 以外の多重代入法プログラムについての概論は、Yucel (2011)、Schmidt (2009)、Drechsler (2009)を参照されたい。日本語の文献では、岩崎 (2002, 11 章)を参照されたい。また、渡辺, 山口 (2000, 付章)には、SOLAS についての詳しい記述があるので、参考にされたい。

EDINET データは数千件の情報しかないので、Amelia II において、実際にどの程度のサイズのデータまで対応できるかを検証するために、EDINET と同様のデータセットを下記の要領でシミュレーションによって生成した。*X* は平均 5.195、標準偏差 1.195 の正規乱数であり、*e* は平均 0、標準偏差 1 の正規乱数であり、*Y* は 6.225 + 0.769**X1* + *e* によって生成される変数である。*X* は事業従事者変数を模しており、*Y* は売上高変数を模している。*Y* と *X* の相関性は約 0.68 であり、観測数は 100 万である。この二変量データセットに、人工的に 50%の欠測を発生させ、*M* = 20 の多重代入のシミュレーションを行った。

その結果、約 24 分で多重代入法の処理を行えることが分かった[27]。現在、日本で行われている最大の経済調査は、経済センサスであり、平成 21 年経済センサス - 基礎調査によると、対象企業の数は約 600 万となっている（総務省, 2011, p.2）。産業や都道府県などで層化してエディティングを行うことを考慮に入れれば、二変量の 100 万データセットの補定処理が行えることは、十分な処理能力があると言える。

7.2 Amelia II の使用法概論[28]

前節で紹介したウェブサイトより、Amelia II をすでにダウンロードし、実装したことを前提として、Amelia II の使用法について簡潔に説明を行う。まず、`setwd` 関数を用い、使用するデータが格納されているフォルダを指定し、`read.csv` 関数などを用いてデータを読み込み、`attach` 関数によりデータを付置する。そして、`library` 関数により Amelia II を起動する。また、`set.seed` 関数により、シード（初期値）の設定を行う。Amelia II は、ブートストラップを用いて乱数[29]を発生させているため、シードを設定しない場合、毎回、異なる補定値を生成してしまう点に注意が必要である。

```
setwd("D:/My Documents/フォルダ名")
data<-read.csv("データ名.csv",header=TRUE)
attach(data)
library(Amelia)
set.seed(1223)
```

[27] 検証に用いたパソコンは、Windows Vista を搭載した一般的なノートパソコンであり、以下のとおりの性能となっている。プロセッサ: Intel Core 2 Duo CPU T9400; メモリ: 2.00 GB; システムの種類: 32 ビットオペレーティングシステム。

[28] 本節では、Amelia II の核となる使用法についてのみ言及する。更に詳しい Amelia II の使用法については、Honaker, King, and Blackwell(2012a)及び Honaker, King, and Blackwell(2012b)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。

[29] 乱数とは、本来、生成されるたびに異なる数値となるべきものである。しかし、科学的分析においては再現性が重要であり、そのためにシードの設定を行い、乱数の再現性を保つことが多い。だが、シードを設定し、同じシードを繰り返し使い続けるならば、それはもはや乱数ではないため、シミュレーション結果へのシード選択の影響を十分に考慮に入れなければならない。すなわち、ある特定のシードによって偏りのあるシミュレーションデータが生成されるならば、そのシード番号を用いることは有益ではない。Amelia においては正規分布を想定しているため、正規分布を正確に生成できるシード番号がよいシードと言える。R において、標本サイズ 1000 の正規分布データをシード番号 1 から 5000 まで作成し、歪度、尖度、Jarque-Bera テスト、カーネル密度などを検証した結果、以下のシード番号が推奨できることが分かった : 43, 393, 864, 1223, 1403, 1712, 1992, 2725, 2748, 2902。本稿の結果は、シード番号 1223 を用いたものである。

多重代入法を行うには、amelia 関数を用いる。ここで、a.out は多重代入法の結果を格納する任意の変数名である。data は使用しているデータ名、m=の右辺は多重代入法により生成するデータセットの数であり、既定では 5 個となっている。logs=の右辺には、対数変換したい変数名を指定するが、変換しない場合には、logs=を含める必要はない。

```
a.out <- amelia(data, m = 5, logs=c("変数 1","変数 2"))
```

上記で作成した 5 個の補定済データセットを以下の write.amelia 関数を用いて csv ファイルとして書き出して保存し、分析などを行うことができる。さらに簡便に多重代入済データセットを一括保存する方法を付録 (pp.82-83)に示すので参考にされたい。

```
write.amelia(obj = a.out, file.stem = "outdata")
```

4.1 節のように、出力した csv ファイルを用いて、多重代入済データセットを手作業により統合することも可能だが、R には、幸いなことに Zelig というパッケージが用意されている (Imai, King, and Lau, 2008; Honaker, King, and Blackwell, 2011, pp.35-36)。まず、require 関数を用い Zelig を起動させる。次に、zelig 関数を用いて統計分析を行う。ここで、z.out は、結果を格納する任意の変数名であり、変数 1 は被説明変数名、変数 2 は説明変数名、data=の右辺は使用する補定済データセットを格納した変数名である。Amelia II では、補定済データセットは a.out$imputations の名前で格納されている。また、model=の右辺は統計分析に使用するモデル名で、ls は最小二乗法による回帰分析を表している。

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

統合モデルの結果を表示したい場合は、下記のとおり summary 関数を使用する。一方、個別のデータセットを用いた分析を表示したい場合は、print 関数を用いる。表示したい個別データの番号を subset=の右辺に指定する。

```
summary(z.out)
print(summary(z.out),subset=1:5)
```

7.3　まとめ

Amelia II は無料の汎用多重代入法のツールであり、巨大データセットも扱うことができ、簡潔なコマンドで多重代入法を行うことができるため、実用面での活躍に期待が持てる。

8　EDINET 売上高の補定の検証結果

本研究では、EDINET データを利用し、多重代入値と単一代入値を、それぞれ、売上高の真値と比較した。単一代入法としては、2.5 節で紹介した確定的回帰補定と 2.6 節で紹介した確率的回帰補定の 2 種類を使用した。また、分布の復元を評価する方法として、散布図による視覚的アプローチと欠測値補定データの標準偏差を使用した。

8.1　データセット

本稿で用いた 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。今回使用したデータの対象となっているのは、2011 年 3 月 31 日に決算を迎える上場企業 3,587 社である[30]。本研究では、「日本標準産業分類」にしたがって、EDINET の全データを下記のとおり産業に分類した：産業 E（製造業)、産業 I（卸売業・小売業)、産業 D（建設業)、産業 G（情報通信業)、産業 L（学術研究・専門技術サービス業）[31]。EDINET の売上高に欠測値がないため、「真値」を知ることができ、補定の精度を的確に評価できる点が有益である[32]。

本稿では、2 つの変数を使用した。1 つ目は、売上高（単位＝百万円）であり、補定の対象となる被説明変数である。売上高変数に人工的に欠測値を発生させて、実験を行う。もう 1 つは、事業従事者数（単位＝人）であり、説明変数である。直感的に、事業従事者数が多くなれば、売上高も大きくなると考えられる。使用したモデルは、1 次多項式と自然対数変換である。生データの基本統計量は、表 8.1 に示すとおりである。

表 8.1：　基本統計量（生データ）

変数	データ数	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(E)	1222	67	10060	23690	119300	66000	8243000	413242
従事者数(E)	1222	3	81	169	419	386	20950	1072
売上高(I)	571	47	12500	31250	144300	88830	8981000	577050
従事者数(I)	571	7	63	133	273	256	7683	557
売上高(D)	158	230	18420	44800	112200	110200	1154000	202486
従事者数(D)	158	6	100	183	394	349	5874	733
売上高(G)	276	20	2340	6908	55450	17010	3373000	309069
従事者数(G)	276	7	76	168	454	433	9783	929
売上高(L)	191	9	960	4482	26520	12420	1397000	110531
従事者数(L)	191	1	25	59	164	133	6284	508

[30] 提出日は、2011 年 6 月 30 日からさかのぼり 1 年以内の企業である。

[31] http://www.stat.go.jp/index/seido/sangyo/19-3.htm（2012 年 12 月 20 日アクセス）

[32] ここで言う「真値」とは、企業が報告した値のことであり、虚偽報告は想定していない。今回の実験では、EDINET に明らかなエラーが存在している場合、エラーを除去して実験を行った。産業 E のデータ数は 1224 だが、内 2 つは重複分であり除去した。産業 I にも重複分が 1 件あり除去した。事業従事者変数には、1165 件の欠測値が存在しているため、これらの企業はデータセットから除外した。

まず、上記のデータが正規分布の前提を満たしているかどうかを確認する。完全な正規分布は、歪度(S: Skewness) = 0、尖度(K: Kurtosis) = 3 となり、歪度と尖度は、それぞれ、式(23)と(24)のとおり求められる(Gujarati, 2003, p.886, p.890; Greene, 2003, pp.848-849)。ここで、μは平均値を表し、σは標準偏差を表す。また、$E(X-\mu)^2$は二次積率である分散(σ^2)であり、$E(X-\mu)^3$は三次積率であり、$E(X-\mu)^4$は四次積率である。

$$S = \frac{E(X-\mu)^3}{\left[\sqrt{E(X-\mu)^2}\right]^3} = \frac{E(X-\mu)^3}{\left[\sqrt{\sigma^2}\right]^3} = \frac{E(X-\mu)^3}{\sigma^3} \tag{23}$$

$$K = \frac{E(X-\mu)^4}{[E(X-\mu)^2]^2} = \frac{E(X-\mu)^4}{[\sigma^2]^2} = \frac{E(X-\mu)^4}{\sigma^4} \tag{24}$$

表 8.2 に、全産業の売上高と事業従事者変数の S（歪度）と K（尖度）を示す。すべての変数において、S（歪度）はゼロよりも大幅に大きく、K（尖度）は 3 よりも大幅に大きい。**合理的に正規分布を近似しているとは言えない**。

表 8.2：S（歪度 = 0）と K（尖度 = 3）

	売上高(E)	従事者(E)	売上高(I)	従事者(I)	売上高(D)	従事者(D)	売上高(G)	従事者(G)	売上高(L)	従事者(L)
S（歪度）	9.953	10.088	9.867	7.076	3.501	4.790	9.184	5.578	10.297	9.713
K（尖度）	148.651	148.232	122.508	72.754	15.869	30.498	91.648	45.414	125.395	113.467

参考までに、産業 E のデータを図示する。図 8.1 は売上高のヒストグラムであり、図 8.2 は事業従事者数のヒストグラムであり、図 8.3 は売上高と事業従事者数の散布図である。S（歪度）と K（尖度）の値から推測されるとおり、典型的な経理項目データと同じように、非常に偏った分布になっていることが分かる。

図 8.1：売上高（生データ）

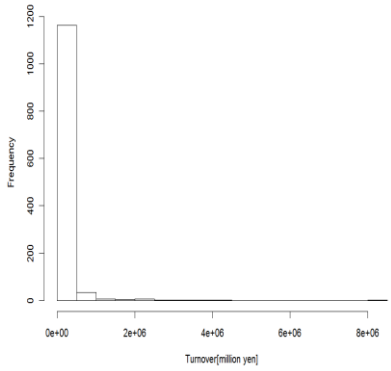


図 8.2：事業従事者数（生データ）

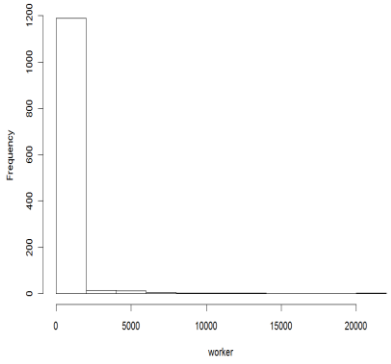
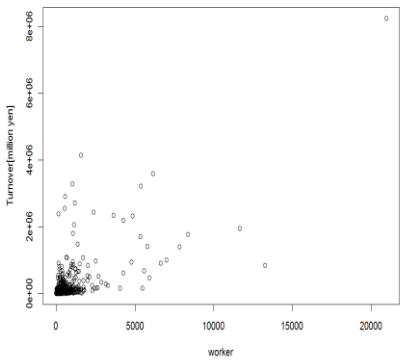


図 8.3 （生データ、r = 0.682）



自然対数変換後の基本統計量は表 8.3 に示すとおりである。

表 8.3：基本統計量（自然対数）

変数	データ数	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(E)	1222	4.204	9.216	10.070	10.220	11.100	15.920	1.553
従事者数(E)	1222	1.099	4.394	5.127	5.195	5.955	9.950	1.195
売上高(I)	571	3.850	9.433	10.350	10.400	11.390	16.010	1.582
従事者数(I)	571	1.946	4.139	4.887	4.903	5.545	8.947	1.100
売上高(D)	158	5.439	9.821	10.710	10.690	11.610	13.960	1.413
従事者数(D)	158	1.792	4.600	5.207	5.254	5.856	8.678	1.151
売上高(G)	276	3.008	7.758	8.840	8.850	9.741	15.030	1.677
従事者数(G)	276	1.946	4.327	5.124	5.206	6.071	9.188	1.309
売上高(L)	191	2.178	6.867	8.407	8.245	9.427	14.150	2.023
従事者数(L)	191	0.000	3.219	4.078	4.089	4.887	8.746	1.342

表 8.4 に、全産業の自然対数変換後の売上高と事業従事者変数の *S*（歪度）と *K*（尖度）を示す。すべての変数において、*S*（歪度）はゼロに近く、*K*（尖度）は 3～4 ほどであり、自然対数変換をしたところ、典型的な経理項目データと同じように、**合理的に正規分布を近似している**ことが分かる。

表 8.4：*S*（歪度 = 0）と *K*（尖度 = 3）

	売上高(E)	従事者(E)	売上高(I)	従事者(I)	売上高(D)	従事者(D)	売上高(G)	従事者(G)	売上高(L)	従事者(L)
S（歪度）	0.389	0.307	0.053	0.342	-0.187	0.063	0.483	0.205	-0.101	-0.014
K（尖度）	3.726	3.525	4.175	3.521	3.665	3.989	4.567	3.030	3.021	4.067

参考までに、産業 E のデータを図示する。図 8.4 は売上高のヒストグラムであり、図 8.5 は事業従事者数のヒストグラムであり、図 8.6 は売上高と事業従事者数の散布図である。比較的きれいな正規分布となっていることが分かる。

図 8.4：売上高（自然対数）

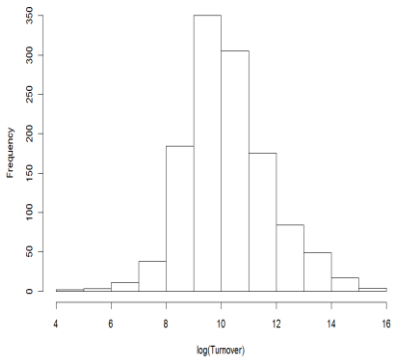


図 8.5：事業従事者数（自然対数）

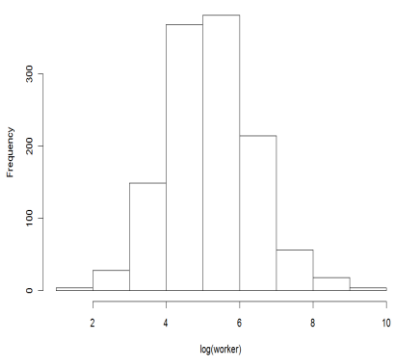
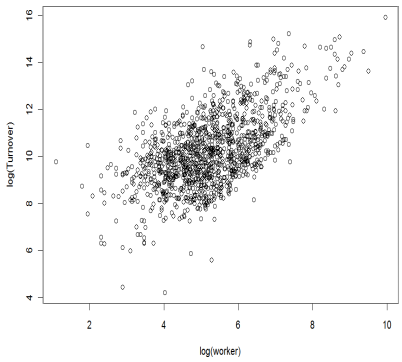


図 8.6 （自然対数、r = 0.593）



8.2　欠測メカニズム

本稿では、NI を対象外とし、MCAR と MAR を前提として、以下の 6 つの欠測メカニズムを用いた：

(1) 完全な無作為抽出(MCAR)
(2) 事業従事者数が**小**の場合に、売上高に欠測が発生(MAR)
(3) 事業従事者数が**中**の場合に、売上高に欠測が発生(MAR)
(4) 事業従事者数が**大**の場合に、売上高に欠測が発生(MAR)
(5) 事業従事者数が**大又は小**の場合に、売上高に欠測が発生(MAR)
(6) 系統抽出(MAR)

データセット内に占める欠測値の割合は、30%、40%、50%の 3 種類であり、したがって、本研究における実験では、合計で 18 種類の欠測を用意した。参考までに、図 8.7 は欠測率 50%の MCAR の散布図、図 8.8 は欠測率 50%の MAR（事業従事者＝小）の散布図、図 8.9 は欠測率 50%の MAR（事業従事者＝中）の散布図、図 8.10 は欠測率 50%の MAR（事業従事者＝大）の散布図、図 8.11 は欠測率 50%の MAR（事業従事者＝大小）の散布図、図 8.12 は欠測率 50%の MAR（系統抽出）の散布図である。

図 8.7：MCAR (50%)

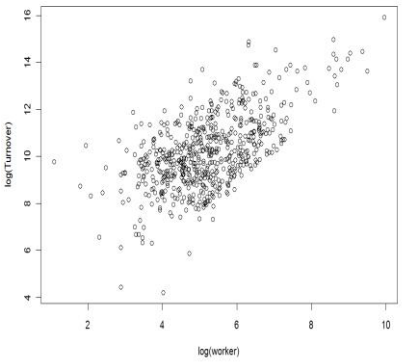


図 8.8： MAR (50%、従事者=小)

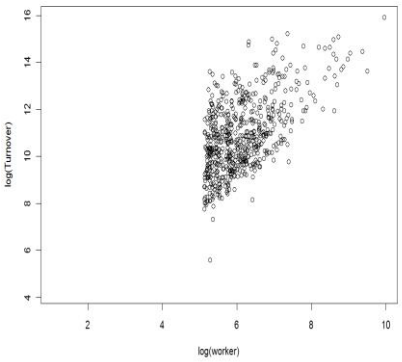


図 8.9：MAR (50%、従事者=中)

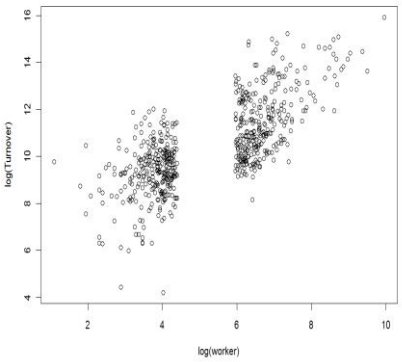


図 8.10：MAR (50%、従事者=大)

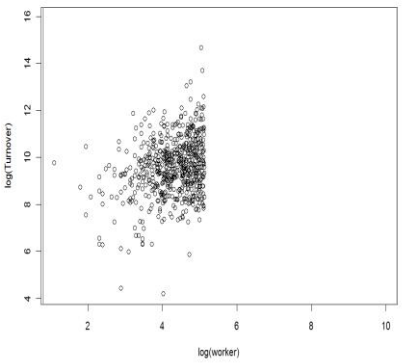


図 8.11： MAR (50%、従事者=大小)

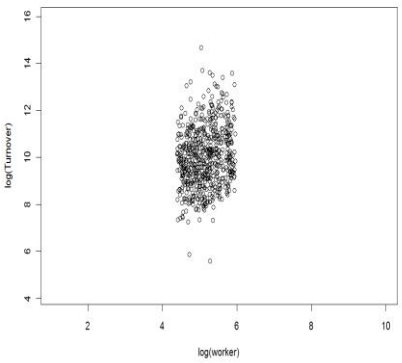
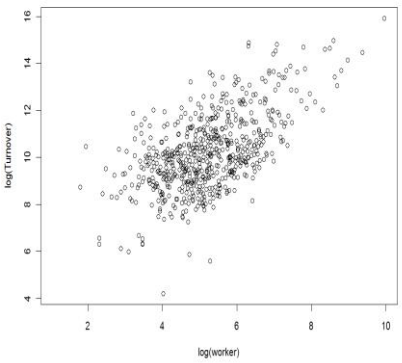


図 8.12：MAR (50%、系統抽出)



8.3　多重代入法と単一代入法の比較検証結果

本研究では、多重代入法と単一代入法の性能差を以下の要領で評価した。まず、検証 1 として 8.3.1 節において、完全データの売上高総額（真値）と欠測値補定後の売上高総額の差を比較し、点推定値の精度を評価する。次に、検証 2 として、8.3.2 節～8.3.7 節において、散布図による視覚的検証を行い、検証 3 として 8.3.8 節において欠測値補定データの標準偏差の検証を行って、真のデータに存在する変動（ばらつき）を再現できているかどうかを検証する。また、検証 4 として 8.3.9 節では、補定済データセットを用いた統計分析を行う。

8.3.1　検証 1：多重代入法と単一代入法による点推定値の精度比較（全産業）

本節では、確定的回帰補定による単一代入法、確率的回帰補定による単一代入法、多重代入法の 3 つの手法を用い、多重代入値と単一代入値を、それぞれ、売上高の真値と比較した。完全データの売上高総額（真値）と欠測値補定後の売上高総額の差を比較し、3 つの手法にランク付けを施した。本稿 6 節で述べたとおり、多重代入データセットの数、M は 20 に設定した。すなわち、多重代入値は、20 個の多重代入済データセットの平均である。

評価方法について、表 8.5 の結果を用いて例示する。表 8.5 は、実際に出力した結果の一部である。産業 E（製造業）の完全データの売上高総額（真値）は 145,785,642 であり、多重代入値と単一代入値のいずれが、この真値に近いかを検証している。表 8.5 では、系統抽出による欠測メカニズム、欠測率 50%、自然対数モデルの場合を例として挙げる。この場合、多重代入値と真値との差が 2,316,949 と最も小さく（1 位）、確定的補定値と真値との差が 7,820,137 と次いでおり（2 位）、確率的補定値と真値との差が 15,156,670 と最も大きかった（3 位）。

表 8.5

産業	データ数	完全データの売上高総額（真値）
E(製造業)	1222	145,785,642

欠測メカニズム	欠測率	モデル		欠測補定済 売上高総額	真値との差 （絶対値）
系統抽出による欠測	50%	自然対数線形回帰	確定補定	137,965,504.8	**7,820,137**
	50%	自然対数線形回帰	確率補定	130,628,972.0	**15,156,670**
	50%	自然対数線形回帰	多重代入	143,468,693.0	**2,316,949**

表 8.6 は、上記の手法を用いて、多重代入法、確定的補定、確率的補定をランク付けした結果を一覧表にまとめたものである。全 5 産業×2 モデル×6 欠測メカニズム×3 欠測率 = 180 ケースの内、確定的補定が 1 位となったケースが 46 回、確率的補定が 1 位となったケースが 67 回、多重代入法が 1 位となったケースが 67 回といった具合である。平均順位とは、(1 位×回数+2 位×回数+3 位×回数)/180 により求め、全体を通じてどの手法がよかったかを示している。表 8.6 の結果より、確率的補定による売上高総額が真値に最も近かった。

表 8.6：全結果

全産業		1 位	2 位	3 位	平均順位
	確定補定	46 回	70 回	64 回	2.100
	確率補定	67 回	57 回	56 回	**1.933**
	多重代入	67 回	53 回	60 回	1.967

しかし、8.1 節において S（歪度）と K（尖度）を検証した結果から、EDINET の生データは正規性の前提を満たしていないことが分かっている。そこで、表 8.7 では、1 次多項式と自然対数に分けて結果を表示した。正規性を満たさない 1 次多項式では多重代入法の当てはまりは悪く、正規性の前提を満たしている自然対数モデルでは多重代入法の当てはまりはよいことが分かった。したがって、以下では、対数モデルにのみ焦点を絞り、詳細を検討する。

表 8.7：モデル別

モデル		1 位	2 位	3 位	平均順位
1 次多項式	確定補定	22 回	47 回	21 回	1.989
	確率補定	35 回	24 回	31 回	**1.956**
	多重代入	33 回	19 回	38 回	2.056
自然対数	確定補定	24 回	23 回	43 回	2.211
	確率補定	32 回	33 回	25 回	1.922
	多重代入	34 回	34 回	22 回	**1.867**

表 8.8 は、産業別の結果を表示している。産業 E（製造業）、産業 D（建設業）、産業 G（情報通信業）では確率的補定の当てはまりがよかったが、産業 I（卸売業・小売業）及び産業 L（学術研究・専門技術サービス業）では多重代入法の当てはまりがよかった。いずれの産業においても、確定的補定の当てはまりはよくなかった。

表 8.8：産業別

産業		1 位	2 位	3 位	平均順位
E (n = 1222)	確定補定	5 回	2 回	11 回	2.333
	確率補定	8 回	8 回	2 回	**1.667**
	多重代入	5 回	8 回	5 回	2.000
I (n = 571)	確定補定	2 回	1 回	15 回	2.722
	確率補定	4 回	12 回	2 回	1.889
	多重代入	12 回	5 回	1 回	**1.389**
D (n = 158)	確定補定	5 回	7 回	6 回	2.056
	確率補定	7 回	7 回	4 回	**1.833**
	多重代入	6 回	4 回	8 回	2.111
G (n = 276)	確定補定	7 回	3 回	8 回	2.056
	確率補定	8 回	4 回	6 回	**1.889**
	多重代入	3 回	11 回	4 回	2.056
L (n= 191)	確定補定	5 回	10 回	3 回	1.889
	確率補定	5 回	2 回	11 回	2.333
	多重代入	8 回	6 回	4 回	**1.778**

表 8.9 は、欠測メカニズム別の結果を表示している。系統抽出、従事者大小、ランダムの欠測メカニズムにおいて確率的補定の当てはまりがよかったが、従事者小、従事者中、従事者大の欠測メカニズムにおいて多重代入法の当てはまりがよかった。現実的には、系統抽出

による欠測や完全にランダムな欠測が発生するとは予期されないので、現実的な欠測メカニズムにおいて、多重代入法の当てはまりがよかったと言える。確定的補定の当てはまりは、いずれの欠測メカニズムにおいてもよくなかった。

表 8.9：欠測メカニズム別

欠測メカニズム		1 位	2 位	3 位	平均順位
系統抽出	確定補定	4 回	7 回	4 回	2.000
	確率補定	6 回	6 回	3 回	**1.800**
	多重代入	5 回	2 回	8 回	2.200
従事者：小	確定補定	6 回	1 回	8 回	2.133
	確率補定	4 回	5 回	6 回	2.133
	多重代入	5 回	9 回	1 回	**1.733**
従事者：中	確定補定	1 回	7 回	7 回	2.400
	確率補定	4 回	5 回	6 回	2.133
	多重代入	10 回	3 回	2 回	**1.467**
従事者：大	確定補定	3 回	3 回	9 回	2.400
	確率補定	4 回	7 回	4 回	2.000
	多重代入	8 回	5 回	2 回	**1.600**
従事者：大小	確定補定	4 回	2 回	9 回	2.333
	確率補定	8 回	5 回	2 回	**1.600**
	多重代入	3 回	8 回	4 回	2.067
ランダム	確定補定	6 回	3 回	6 回	2.000
	確率補定	6 回	5 回	4 回	**1.867**
	多重代入	3 回	7 回	5 回	2.133

表 8.10 は、欠測率ごとの結果を示している。欠測率 40%の場合には確率的補定の当てはまりがよかったが、欠測率 30%及び 50%の場合には多重代入法の当てはまりがよかった。いずれの欠測率においても、確定的補定の当てはまりはよくなかった。

表 8.10：欠測率別

欠測率		1 位	2 位	3 位	平均順位
30%	確定補定	11 回	5 回	14 回	2.100
	確率補定	6 回	16 回	8 回	2.067
	多重代入	13 回	9 回	8 回	**1.833**
40%	確定補定	8 回	7 回	15 回	2.233
	確率補定	13 回	11 回	6 回	**1.767**
	多重代入	9 回	12 回	9 回	2.000
50%	確定補定	5 回	11 回	14 回	2.300
	確率補定	13 回	6 回	11 回	1.933
	多重代入	12 回	13 回	5 回	**1.767**

ここまで、多重代入値と単一代入値の点推定値の精度比較を行ったが、8.3.2 節から 8.3.7 節まで、多重代入値と単一代入値を、データの分布の復元という観点から検証する。実際には、全産業に関し、6 個の欠測メカニズムと 3 個の欠測率を用い、1 次多項式と自然対数のモデルによって検証したが、以下では紙面の都合により、産業 E に関し、50%の欠測率において、自然対数変換したデータを用いた検証結果（6 種類の欠測発生メカニズム）を例示する。

8.3.2　検証 2：MCAR の結果（産業 E、欠測率 50%、自然対数）

完全な無作為抽出の場合、真値の標準偏差（自然対数）は **1.532** である。確定的補定による単一代入値の標準偏差（自然対数）は **0.889** であり、確率的補定による単一代入値の標準偏差（自然対数）は **1.493** であり、多重代入値の標準偏差（自然対数）の平均は **1.525** である。多重代入値の標準偏差が最も復元力が高かった。

図 8.6 は、売上高と事業従事者数の真値の散布図である。図 8.13 は、確定的補定の散布図であり、図 8.14 は、確率的補定の散布図であり、図 8.15 は、多重代入の散布図（m = 10、すなわち 10 回目の多重代入のときの散布図[33]）である。図 8.13 から図 8.15 では、観測値を黒丸で表し、補定値を赤丸で表している。確率的補定と多重代入法では、真値のばらつきを比較的よく復元できていると言える。

図 8.6：真値

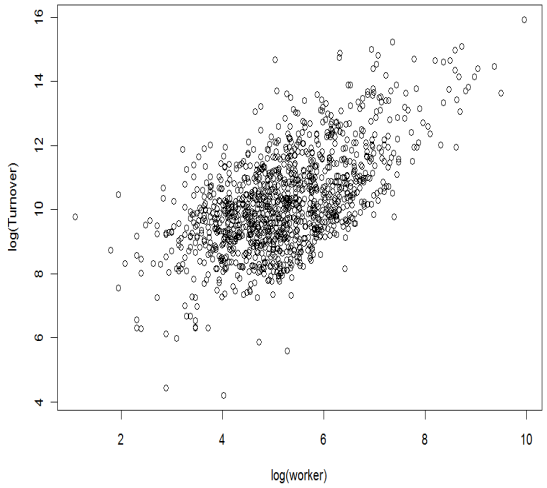


図 8.13：確定的補定

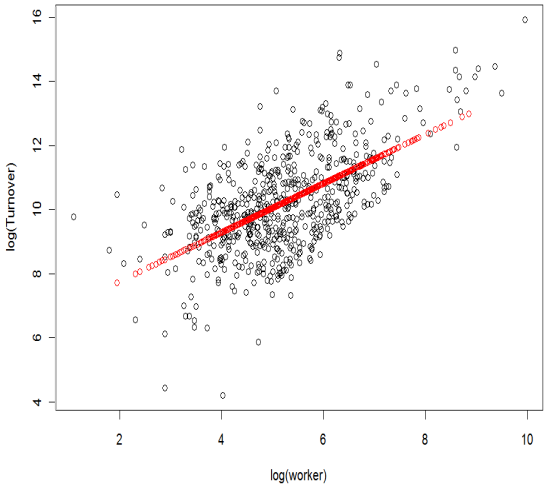


図 8.14：確率的補定

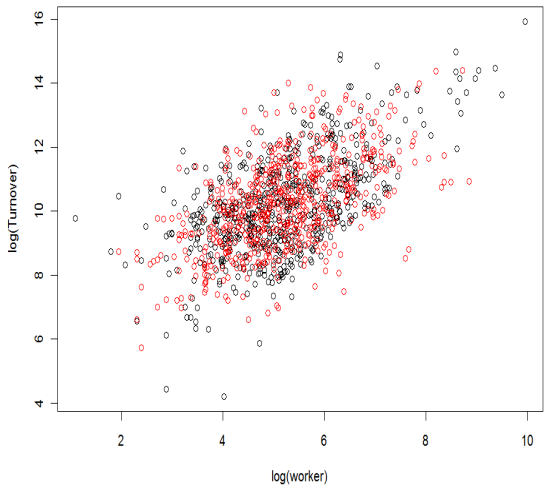
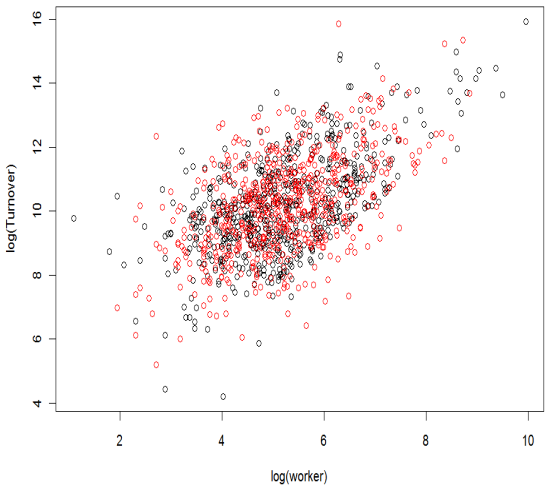


図 8.15：多重代入法(m = 10)



[33] 多重代入法の *M* を 20 に設定しているため、m = 1 から m = 20 までの 20 個の散布図が生成されている。紙面の都合上、今回は、任意の散布図を 1 つだけ選んだが、他の 19 個の散布図も、ほぼ同様となっている。

8.3.3　検証 2：MAR（従事者小）の結果（産業 E、欠測率 50%、自然対数）

MAR(事業従事者数＝小)の場合、真値の標準偏差（自然対数）は **1.240** である。確定的補定による単一代入値の標準偏差（自然対数）は **0.736** であり、確率的補定による単一代入値の標準偏差（自然対数）は **1.446** であり、多重代入値の標準偏差（自然対数）の平均は **1.426** である。多重代入値の標準偏差が最も復元力が高かった。

図 8.6 は、売上高と事業従事者数の真値の散布図である。図 8.16 は、確定的補定の散布図であり、図 8.17 は、確率的補定の散布図であり、図 8.18 は、多重代入の散布図（m = 10）である。図 8.16 から図 8.18 では、観測値を黒丸で表し、補定値を赤丸で表している。確率的補定と多重代入法では、真値のばらつきを比較的よく復元できていると言える。

図 8.6：真値

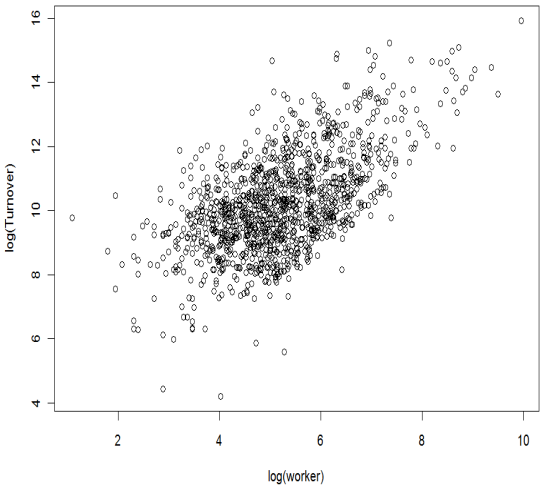


図 8.16：確定的補定

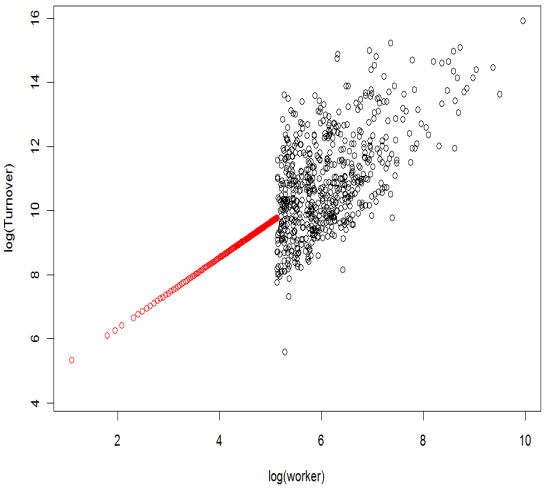


図 8.17：確率的補定

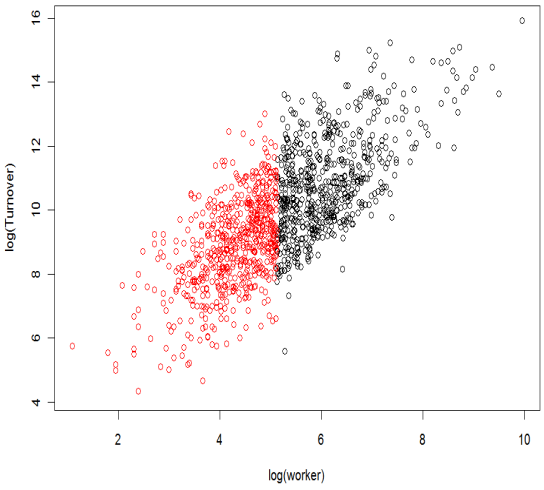
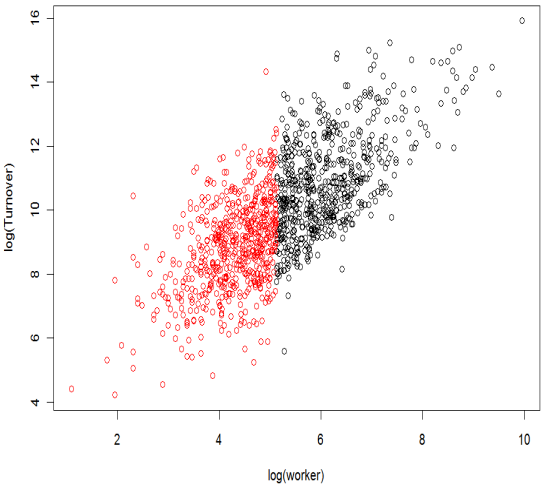


図 8.18：多重代入法(m = 10)



8.3.4　検証 2：MAR（従事者中）の結果（産業 E、欠測率 50%、自然対数）

MAR(事業従事者数＝中)の場合、真値の標準偏差（自然対数）は **1.282** である。確定的補定による単一代入値の標準偏差（自然対数）は **0.331** であり、確率的補定による単一代入値の標準偏差（自然対数）は **1.225** であり、多重代入値の標準偏差（自然対数）の平均は **1.245** である。多重代入値の標準偏差が最も復元力が高かった。

図 8.6 は、売上高と事業従事者数の真値の散布図である。図 8.19 は、確定的補定の散布図であり、図 8.20 は、確率的補定の散布図であり、図 8.21 は、多重代入の散布図（m = 10）である。図 8.19 から図 8.21 では、観測値を黒丸で表し、補定値を赤丸で表している。確率的補定と多重代入法では、真値のばらつきを比較的よく復元できていると言える。

図 8.6：真値

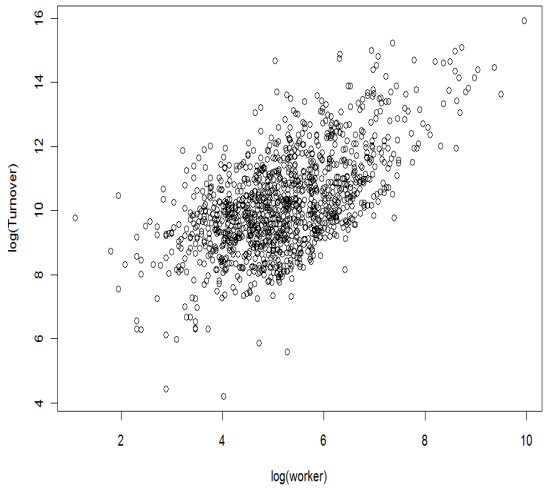


図 8.19：確定的補定

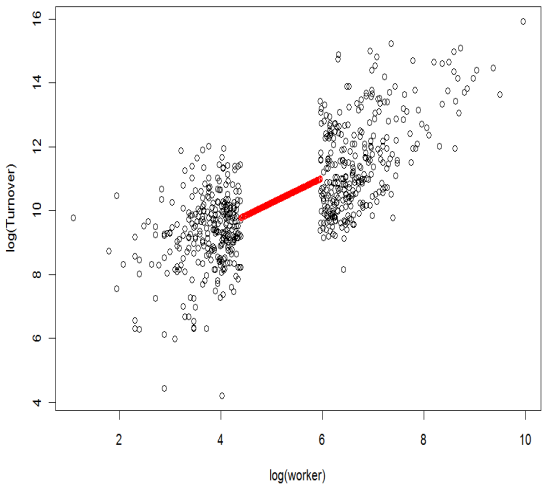


図 8.20：確率的補定

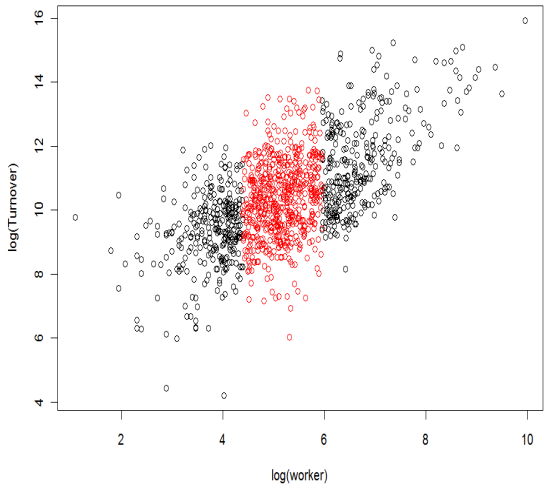
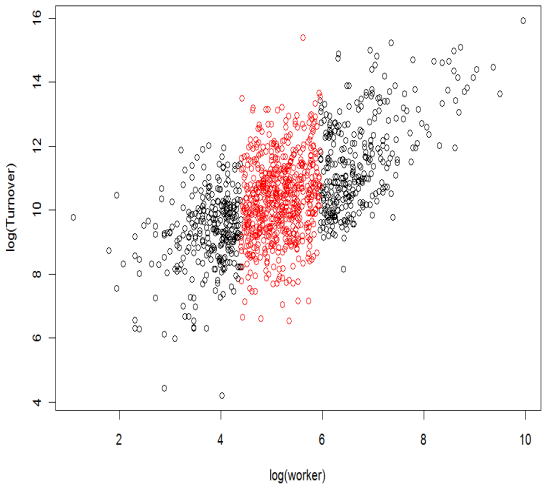


図 8.21：多重代入法(m = 10)



8.3.5　検証 2：MAR（従事者大）の結果（産業 E、欠測率 50%、自然対数）

MAR(事業従事者数＝大)の場合、真値の標準偏差（自然対数）は **1.554** である。確定的補定による単一代入値の標準偏差（自然対数）は **0.418** であり、確率的補定による単一代入値の標準偏差（自然対数）は **1.256** であり、多重代入値の標準偏差（自然対数）の平均は **1.157** である。確率的補定の標準偏差が最も復元力が高かった。

図 8.6 は、売上高と事業従事者数の真値の散布図である。図 8.22 は、確定的補定の散布図であり、図 8.23 は、確率的補定の散布図であり、図 8.24 は、多重代入の散布図（m = 10）である。図 8.22 から図 8.24 では、観測値を黒丸で表し、補定値を赤丸で表している。確率的補定と多重代入法では、真値のばらつきを比較的よく復元できていると言える。

図 8.6：真値

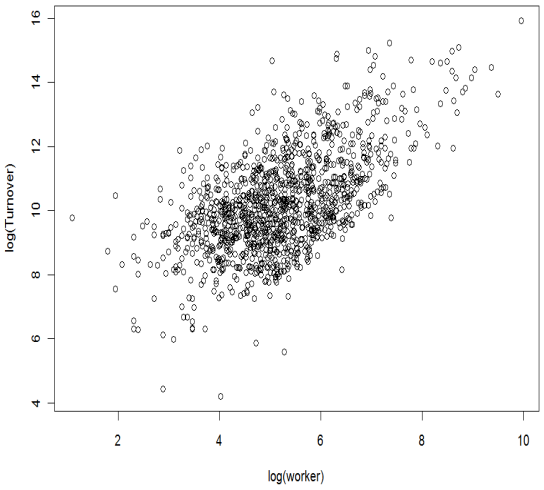


図 8.22：確定的補定

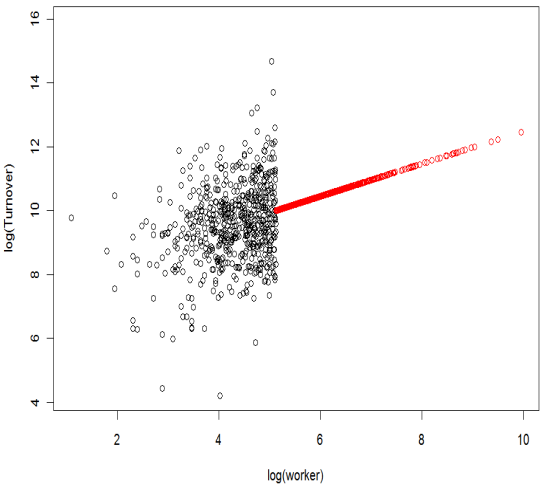


図 8.23：確率的補定

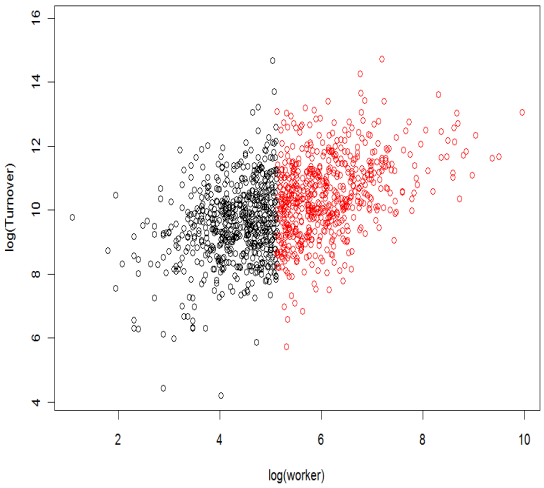
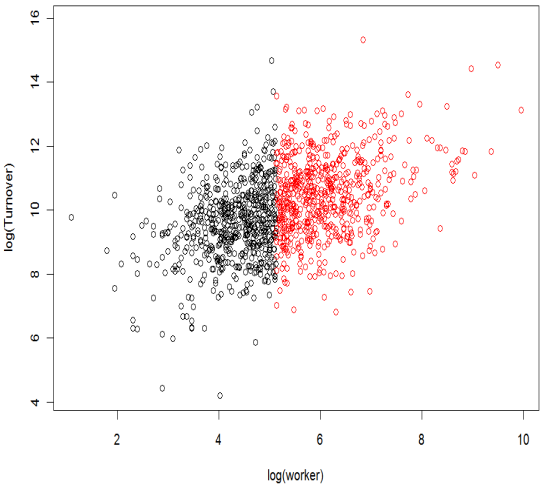


図 8.24：多重代入法(m = 10)



8.3.6　検証 2：MAR（従事者大小）の結果（産業 E、欠測率 50%、自然対数）

MAR(事業従事者数＝大小)の場合、真値の標準偏差（自然対数）は **1.760** である。確定的補定による単一代入値の標準偏差（自然対数）は **0.937** であり、確率的補定による単一代入値の標準偏差（自然対数）は **1.533** であり、多重代入値の標準偏差（自然対数）の平均は **1.599** である。多重代入値の標準偏差が最も復元力が高かった。

図 8.6 は、売上高と事業従事者数の真値の散布図である。図 8.25 は、確定的補定の散布図であり、図 8.26 は、確率的補定の散布図であり、図 8.27 は、多重代入の散布図（m = 10）である。図 8.25 から図 8.27 では、観測値を黒丸で表し、補定値を赤丸で表している。確率的補定と多重代入法では、真値のばらつきを比較的よく復元できていると言える。

図 8.6：真値

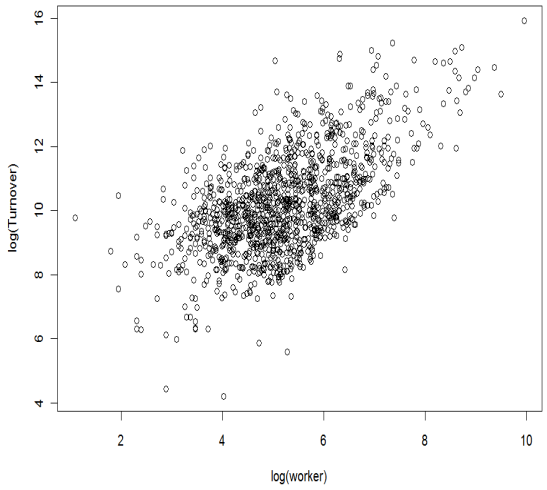


図 8.25：確定的補定

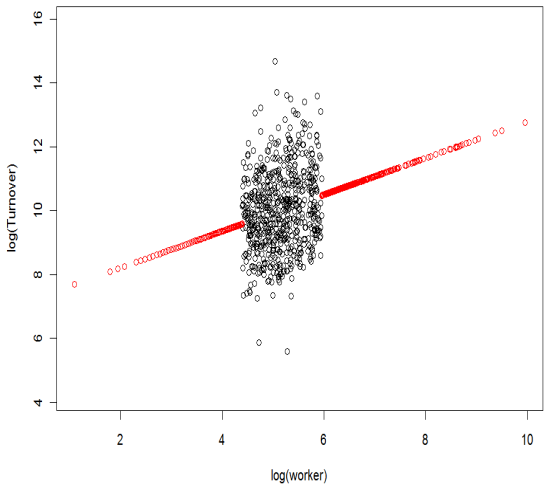


図 8.26：確率的補定

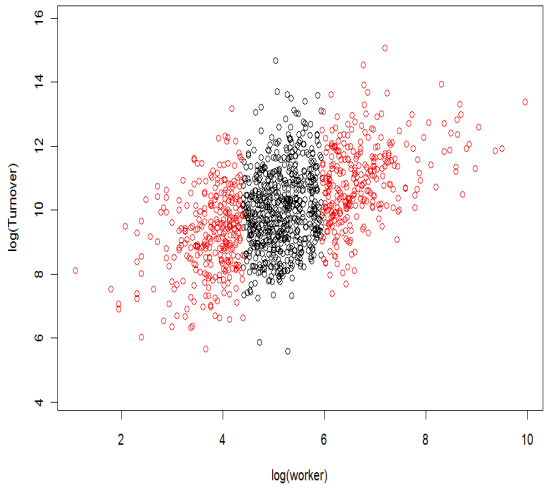
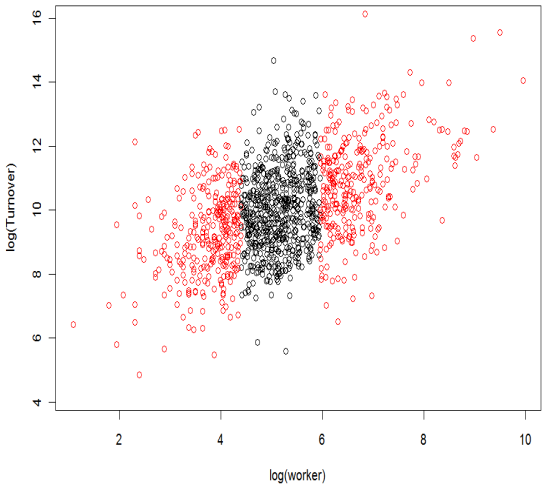


図 8.27：多重代入法(m = 10)



8.3.7　検証 2：MAR（系統抽出）の結果（産業 E、欠測率 50%、自然対数）

MAR(系統抽出)の場合、真値の標準偏差（自然対数）は **1.489** である。確定的補定による単一代入値の標準偏差（自然対数）は **0.956** であり、確率的補定による単一代入値の標準偏差（自然対数）は **1.564** であり、多重代入値の標準偏差（自然対数）の平均は **1.586** である。確率的補定の標準偏差が最も復元力が高かった。

図 8.6 は、売上高と事業従事者数の真値の散布図である。図 8.28 は、確定的補定の散布図であり、図 8.29 は、確率的補定の散布図であり、図 8.30 は、多重代入の散布図（m = 10）である。図 8.28 から図 8.30 では、観測値を黒丸で表し、補定値を赤丸で表している。確率的補定と多重代入法では、真値のばらつきを比較的よく復元できていると言える。

図 8.6：真値

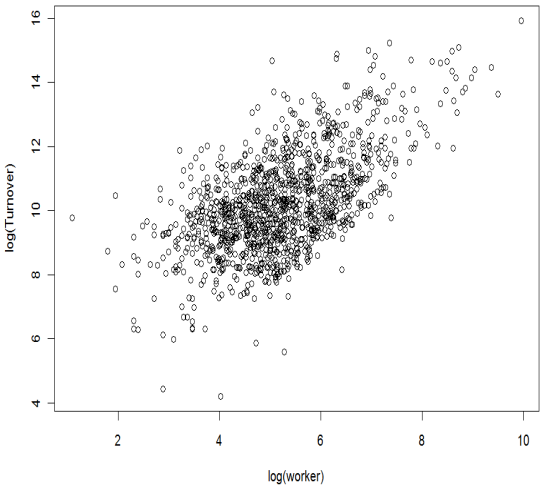


図 8.28：確定的補定

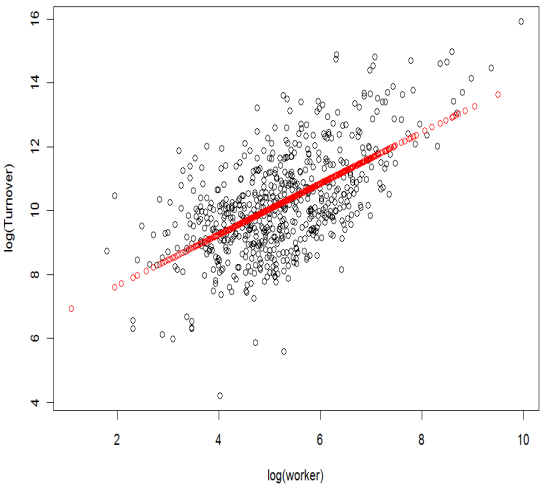


図 8.29：確率的補定

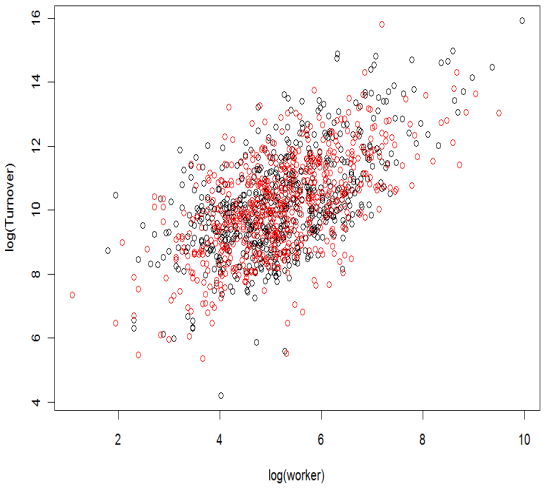
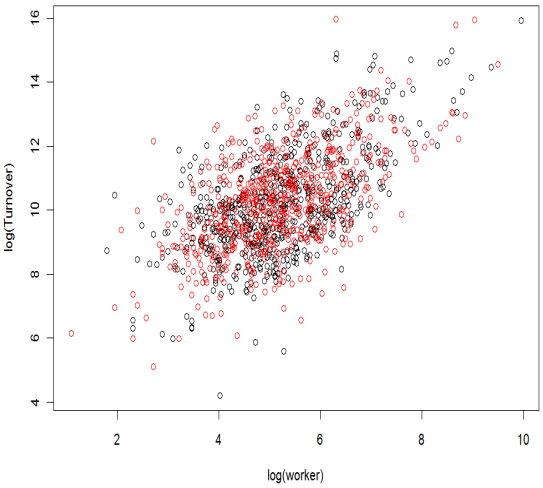


図 8.30：多重代入法(m = 10)



8.3.8　検証 3：標準偏差の精度

表 8.11 では、自然対数変換を用いた全 90 パターン（5 産業、6 欠測メカニズム、3 欠測率）において、各々の補定手法が真の標準偏差の復元にどれだけ近かったかを示している。全体的に、多重代入法による標準偏差の復元が最も優れていた。

表 8.11

全産業		1 位	2 位	3 位	平均順位
	確定補定	2 回	0 回	88 回	2.956
	確率補定	41 回	48 回	1 回	1.556
	多重代入	47 回	42 回	1 回	**1.489**

8.3.9　検証 4：統計分析の精度

統計データを収集することの 1 つの大きな目的は、得られたデータを用い、未知の母集団パラメータの推定を行うことにある。そこで、産業 E、50%欠測、従事者小、対数モデルのケースを例に、補定済データセットを用いた統計分析を行った結果を例示する。使用したモデルは、売上高を被説明変数とし、事業従事者を説明変数とする単回帰モデルである。すなわち、式(25)における$\hat{\alpha}$と$\hat{\beta}$の推定を行う。

$$\log\left(\widehat{\text{売上高}}_i\right) = \hat{\alpha} + \hat{\beta}\log\left(\text{事業従事者数}_i\right) \tag{25}$$

分析結果を表 8.12 に示す。ある変数が他の変数に与える影響は、傾きの大きさによって測られるので、ここでは傾きに注目する。真値の傾きは 0.771 であり、標準誤差は 0.030 である。多重代入済データセットを用いた分析では、傾きは 1.087 と推定された。確定補定済データセットを用いた分析では、傾きは 1.105 と推定された。確率補定済データセットを用いた分析では、傾きは 1.096 と推定された。傾きの推定値は、いずれの場合も、真値と比較して過大推定となっている。一方、多重代入済データセットを用いた分析では、傾きの標準誤差は 0.053 であった。確定補定済データセットを用いた分析では、傾きの標準誤差は 0.021 であった。確率補定済データセットを用いた分析では、傾きの標準誤差は 0.030 であった。

表 8.12

モデル	切片 (標準誤差)	傾き (標準誤差)	傾きの t 値
真値	6.219 (0.160)	**0.771** (**0.030**)	25.730
多重代入	4.225 (0.324)	**1.087** (**0.053**)	20.440
確定補定	4.116 (0.114)	**1.105** (**0.021**)	51.599
確率補定	4.185 (0.159)	**1.096** (**0.030**)	36.725

すなわち、確定補定済データセット及び確率補定済データセットを用いた統計分析では、傾きの推定値が真値とは異なっているにも関わらず、標準誤差が小さく、*t* 値が過大に有意と

なっている。一方、多重代入法では、傾きの推定値が真値と異なっていることを反映し、標準誤差が大きくなり、*t* 値が真値に近くなっている。参考までに、全産業（従事者小欠測）における *t* 値のパフォーマンス結果は、表 8.13 のとおりであった。

表 8.13

全産業		1 位	2 位	3 位	平均順位
	確定補定	0 回	0 回	15 回	3.000
	確率補定	6 回	9 回	0 回	1.600
	多重代入	9 回	6 回	0 回	**1.400**

8.4　まとめ

多重代入法による補定の点推定値は、おおむね、確定的補定よりも当てはまりがよく、標準偏差の推定においても優れており、分布の再現性も高いことが分かった。多重代入法と確率的補定の精度は、かなり拮抗しているものの、全体的には多重代入法の方が当てはまりはよく、統計推定における優位も確認された。

9　多重代入法の精度評価：補定の診断

補定モデルが正しく設定されているとき、多重代入推定量は漸近的に不偏であり、分散も正しく推定される(Marti and Chavance, 2011)。本稿 8 節では、真値と補定値の比較を直接的に行って補定モデルの評価を行った。実験目的では、これが補定モデルの精度を評価する最も妥当な手法であるが、現実には、真値は常に不明である。したがって、現実には、補定モデルの当てはまりを直接的に検証することはできない。結果として、補定の診断手法は、長らく見過ごされてきた。しかし、Abayomi, Gelman, and Levy (2008)によって、補定モデルの当てはまり及び欠測メカニズムの前提に関し、間接的な検証を行えることが示されている。

9.1　Amelia における欠測値診断手法

Amelia II においても、密度の比較(Comparing Densities)機能、過剰補定(Overimpute)機能、過散布初期値(Overdispersed Starting Values)機能、欠測地図(Missingness Map)機能が利用可能である(Honaker, King, and Blackwell, 2011, pp.25-35)。これらの診断手法を使用するには、まず、7.2 節で示したとおり、Amelia による多重代入法を以下の要領で行う。

```
library(Amelia)
set.seed(1223)
a.out <- amelia(data, m = 20)
```

密度の比較機能では、観測値と補定値の密度の比較を行う。2 つの密度がほぼ重なっている場合、補定モデルには問題がないと結論付けられるが、そもそも、欠測値と観測値は体系

的に異なっている可能性があるからこそ、補定を行うのであり、先験的に 2 つの密度が類似するとは必ずしも期待できない。もし 2 つの密度が大幅に異なっていたり、密度が異常な形状を示している場合には、補定モデルの妥当性を再検証する必要があるが、2 つの密度の形状が異なること自体は、即座に補定モデルの異常性を意味しない。密度の比較機能を使用するには、以下の plot 関数を用いる。ここで、a.out は多重代入値を格納している任意の変数名、which.vars=の右辺は密度の比較を行う変数名である。

```
plot(a.out,which.vars="変数")
```

過剰補定機能は、補定モデルの当てはまりを検証する Amelia に特有の機能である。過剰補定では、各々の観測値を、あたかも欠測しているかのように扱い、観測値を人工的に 1 つずつ欠測させ、各々の人工的欠測値の補定を数百回行い、90%信頼区間を図示する。また、同一の図上に y = x 線を図示している。もし補定モデルが完璧であるならば、すべての補定値は y = x 線上に位置することとなる。y = x 線が、信頼区間に含まれているかどうかを見ることで、補定モデルの信頼度をチェックすることができる。過剰補定機能を使用するには、以下の overimpute 関数を用いる。ここで、var=の右辺は過剰補定を行う変数名である。

```
overimpute(a.out,var="変数")
```

過散布初期値機能では、EM アルゴリズムの初期値を様々に設定し、過剰に散布させた場合に、同一の値に収束するかどうかを図示する。5.1 節で述べたとおり、複数の峰のある分布では、局所的最大値が大局的最大値であるとは限らない。したがって、複数の初期値から EM アルゴリズムを行い、複数の収束した値の中から最も大きいものを選ぶ必要がある。しかし、EM の収束過程は非常に高次元であり、図示することができないため、過散布初期値機能では、最大の主成分分析値と 2 番目に大きい主成分分析値を図示している。過散布初期値機能では、乱数により複数の初期値を設定しているので、再現性を保つために、シードの設定を行う必要がある。その後、過散布初期値機能を使用するために、以下の disperse 関数を用いる。ここで、dims=の右辺は主成分分析の次元数であり、1 は最大の主成分を、2 は 2 番目に大きな主成分を表示する。m=の右辺は初期値の数である。

```
set.seed(1223)
disperse(a.out,dims=1,m=20)
```

欠測地図機能では、データセット内の欠測パターンを視覚化することで、データセット内のどこに欠測が発生しているかを視覚的に認識することができる。この機能によって欠測発生のメカニズムが NI、MAR、MCAR のいずれであるかを断定することはできないが、上述の密度の比較機能とともに使用することで、欠測発生メカニズムを推測する一助となる。欠測地図機能を使用するには、以下の missmap 関数を用いる。ここで、欠測にパターンがある

かどうかを調べるために、説明変数を様々に並び替えて、複数の欠測地図をチェックすることを推奨する。

```
missmap(a.out)
```

以上のとおり、1 つ 1 つの診断法は間接的であり、決定的証拠ではないが、組み合わせて使用することにより、補定モデルの妥当性を推測検証することができる。

9.2 欠測の前提の診断方法に関する指針

本節では、密度の比較、欠測地図、過剰補定、過散布初期値を利用することで、真値が分からずとも、欠測の前提を推定できることを、以下のとおり生成したシミュレーションデータセットを用い、指針として示す。

```
set.seed(1223)
X1<-rnorm(1000)
X2<-rnorm(1000)
e<-rnorm(1000)
Y<-10+3*X1+e
```

各変数間の相関係数は、表 9.1 に示すとおりである。*Y* と *X1*、*Y* と *X2* の相関が重要な値となる。

表 9.1：相関

変数	X1	X2	e	Y
X1	1.0000			
X2	0.0124	1.0000		
e	-0.0383	0.0549	1.0000	
Y	**0.9445**	**0.0299**	0.2920	1.0000

MCAR（診断 1）においては、*X2*（乱数）を条件として、*Y* に欠測を発生させ、データセットには *X1* と *Y* のみを格納する。**MAR（診断 2）**においては、*X1* を条件として、*Y* に欠測を発生させ、データセットには *X1* と *Y* のみを格納する。**NI→MAR （診断 3）**においては、*Y* を条件として、*Y* に欠測を発生させ、データセットには *X1* と *Y* のみを格納する。**NI（診断 4）**においては、*Y* を条件として、*Y* に欠測を発生させ、データセットには *X2* と *Y* のみを格納する。欠測率は、約 50%である。

9.2.1　診断 1: MCAR

診断 1 では、MCAR の指針を示す。図 9.1 では、2 つの密度がほぼ完全に重なっている。すなわち、真の欠測の発生メカニズムが MCAR である場合、補定値の密度と観測値の密度は、ほぼ重なるということが分かる。また、図 9.2 において、*X1* の値をどのように並べ替えても、欠測地図における *Y* の欠測にはパターンが存在しない。すなわち、欠測のパターンは、観測データに基づかずランダムに発生している様子が分かる。一方、図 9.3 の過剰補定を見れば分かるとおり、補定モデルの当てはまりは非常によい。また、図 9.4 の過散布初期値から、すべての初期値は、同一の値に収束しており、EM アルゴリズムにも問題はなかったと推定できる。

図 9.1：密度

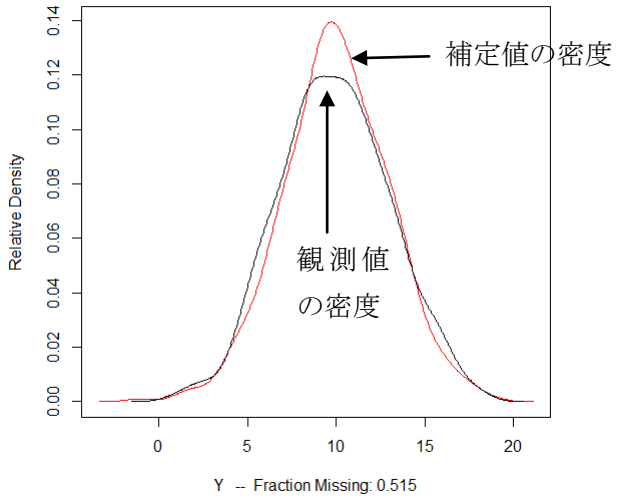


図 9.2：欠測地図

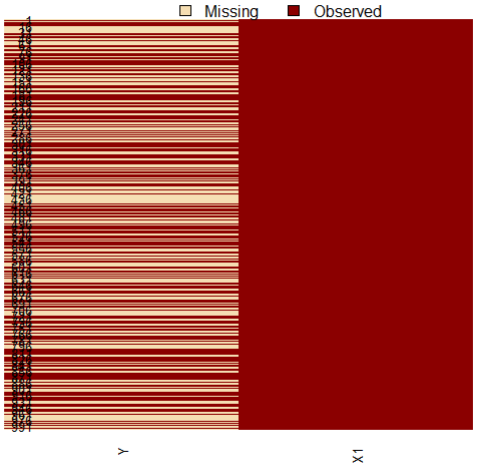


図 9.3：　過剰補定

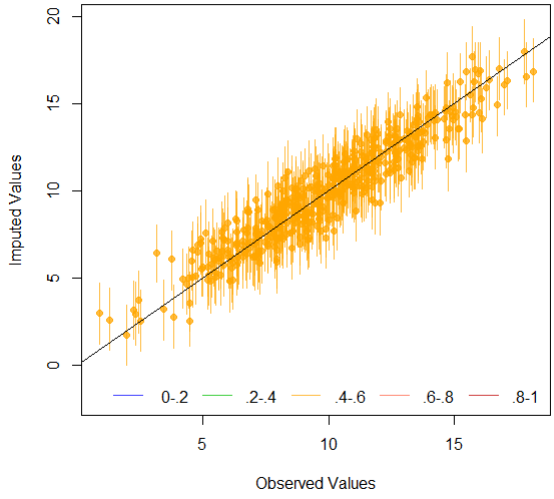
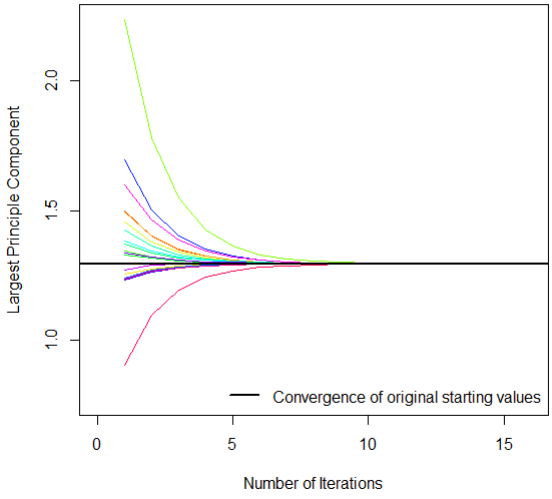


図 9.4：過散布初期値



9.2.2　診断 2: MAR

診断 2 では、MAR の指針を示す。図 9.5 では、2 つの密度が大きく分離している。すなわち、真の欠測の発生メカニズムが MAR である場合、補定値の密度と観測値の密度は、必ずしも重なり合わないことが分かる。また、図 9.6 において、*X1* の値を並べ替えたところ、欠測地図における *Y* の欠測に完全なパターンが存在している。すなわち、欠測のパターンは、観測データに基づいて発生している様子が分かる。一方、図 9.7 の過剰補定を見れば分かるとおり、補定モデルの当てはまりは比較的よい。また、図 9.8 の過散布初期値から、すべての初期値は、同一の値に収束しており、EM アルゴリズムにも問題はなかったと推定できる。

図 9.5：密度

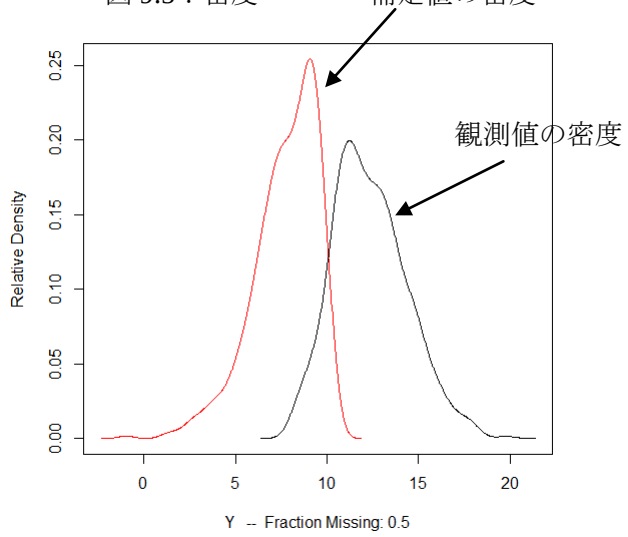


図 9.6：欠測地図

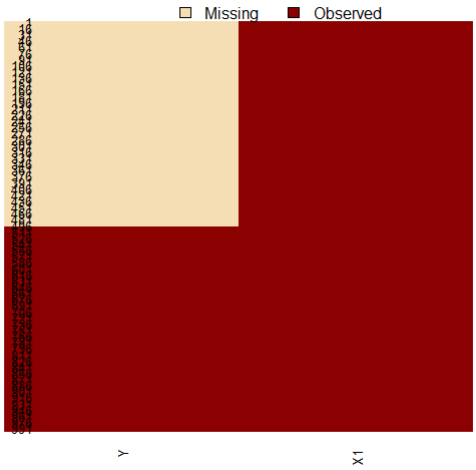


図 9.7：　過剰補定

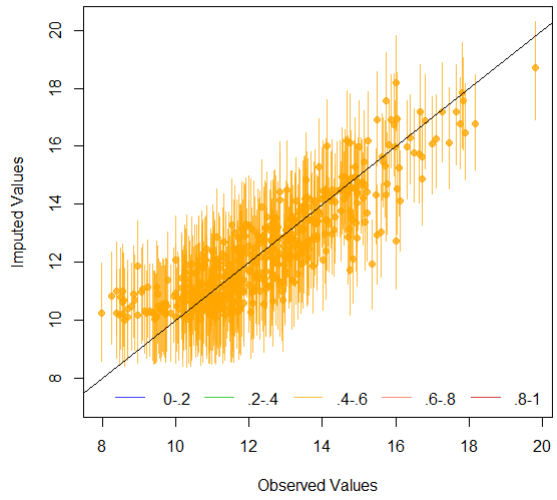
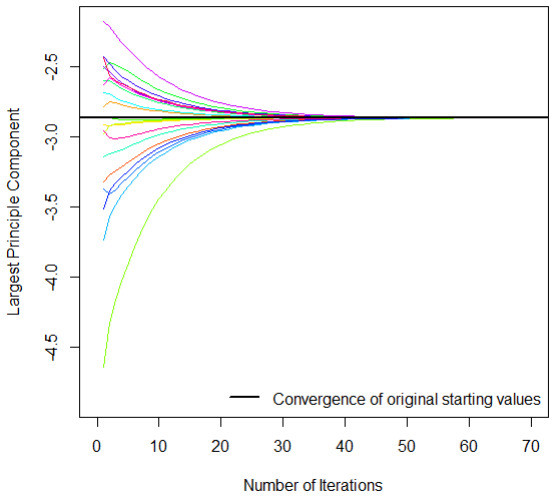


図 9.8：過散布初期値



9.2.3　診断 3: NI→MAR (r = 0.94)

診断 3 では、真の欠測発生メカニズムは NI であるが、事実上の MAR として処理できる場合の指針を示す。図 9.9 では、2 つの密度が大きく分離しており、MAR のケースと酷似している。また、図 9.10 において、*X1* の値を並べ替えたところ、欠測地図における *Y* の欠測にほぼ完全なパターンが存在している。すなわち、欠測のパターンは、観測データと大きな関連がある様子が分かる。一方、図 9.11 の過剰補定を見れば分かるとおり、補定モデルの当てはまりは比較的よい。また、図 9.12 の過散布初期値から、すべての初期値は、同一の値に収束しており、EM アルゴリズムにも問題はなかったと推定できる。本来ならば NI であるはずだが、*Y* と *X* の相関が非常に高いため、事実上の MAR として診断できる。

図 9.9：密度

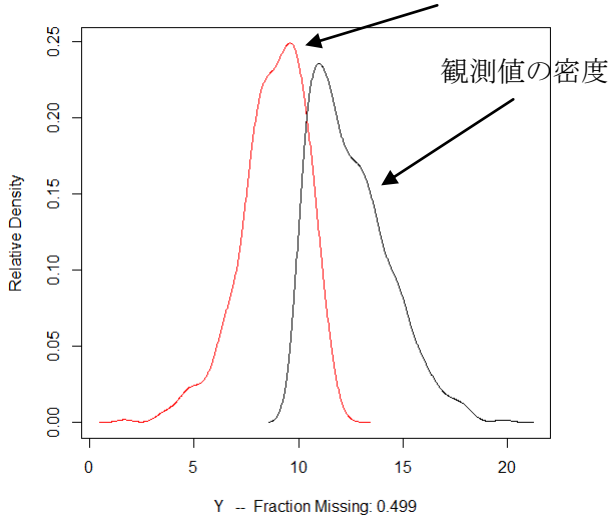


図 9.10：欠測地図

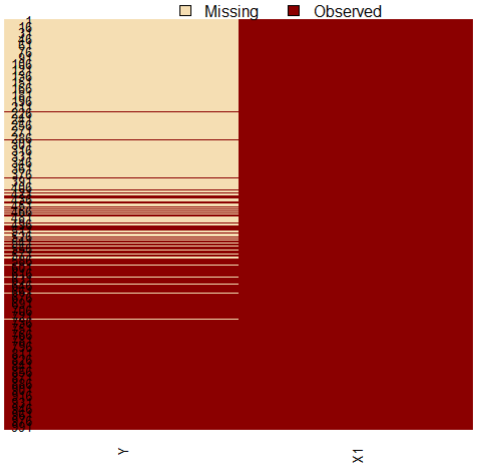


図 9.11：過剰補定

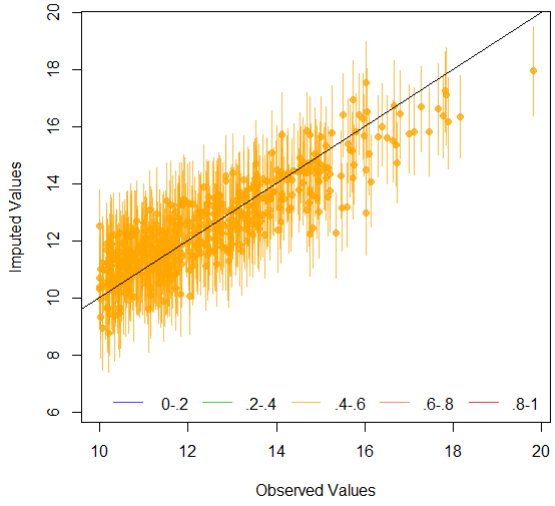
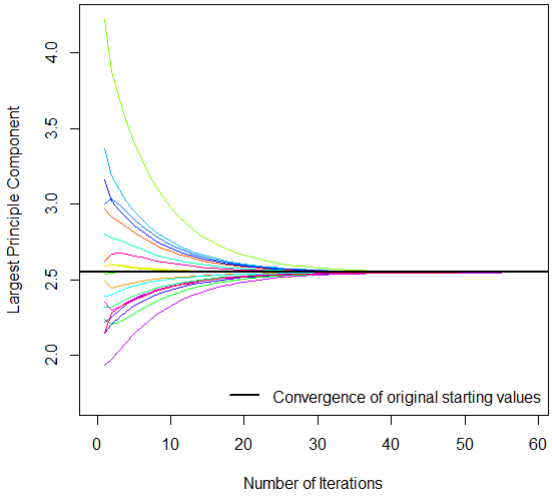


図 9.12：過散布初期値



9.2.4　診断 4: NI

診断 4 では、NI の指針を示す。図 9.13 では、2 つの密度が大きく分離している。すなわち、真の欠測の発生メカニズムが NI である場合、補定値の密度と観測値の密度は、重なり合わないことが分かる。また、図 9.14 において、*X2* の値をどのように並べ替えても、欠測地図における *Y* の欠測にはパターンが存在しない。すなわち、欠測のパターンは、観測データに基づかずに発生している様子が分かる。一方、図 9.15 の過剰補定を見れば分かるとおり、補定モデルの当てはまりは非常に悪い。これが典型的な NI の診断結果であると言える。今回の NI の場合では、図 9.16 の過散布初期値から、すべての初期値は、同一の値に収束しており、EM アルゴリズムには問題はなかったと推定できる。

図 9.13：密度

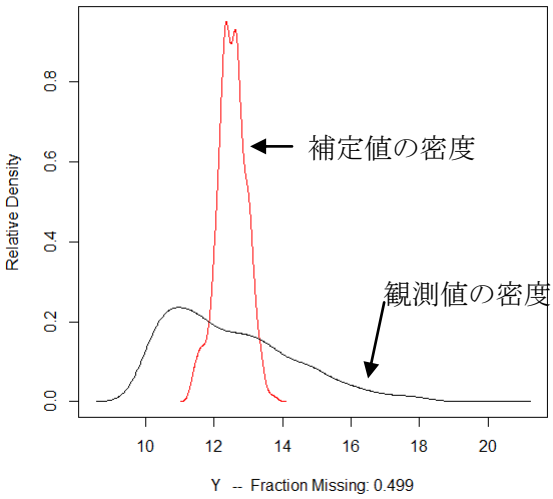


図 9.14：欠測地図

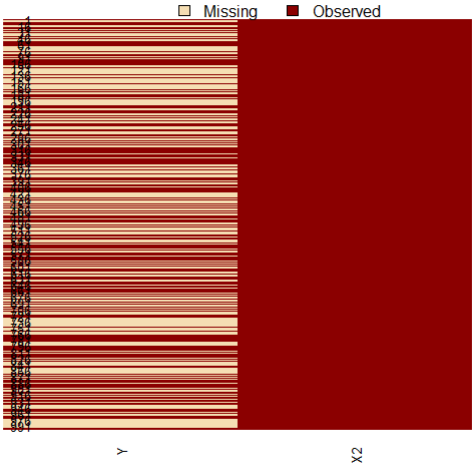


図 9.15：過剰補定

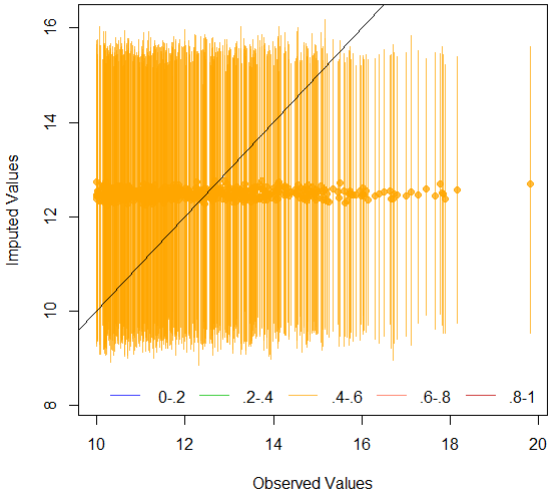
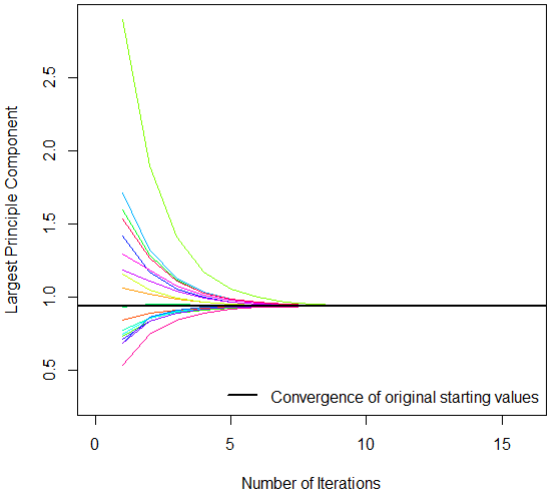


図 9.16：過散布初期値



9.3　指針の応用

本節では、9.2 節で示した指針を EDINET に応用し、欠測の発生メカニズムの推定が行えるかどうかを検証する。

9.3.1　MCAR、EDINET 産業 E、欠測率 50%、自然対数の診断

図 9.17 では、2 つの密度は合理的に酷似している。つまり、欠測のメカニズムは MCAR ではないかと推測できる。図 9.18 では、売上高の欠測にはパターンが認められない。したがって、図 9.17 と図 9.18 の情報に基づいて、真の欠測メカニズムは MCAR であると推定できる。実際には、これらの欠測値を作り出したメカニズムが分かっているので、上述の診断は正しいことが分かっている。

図 9.17：密度(MCAR)

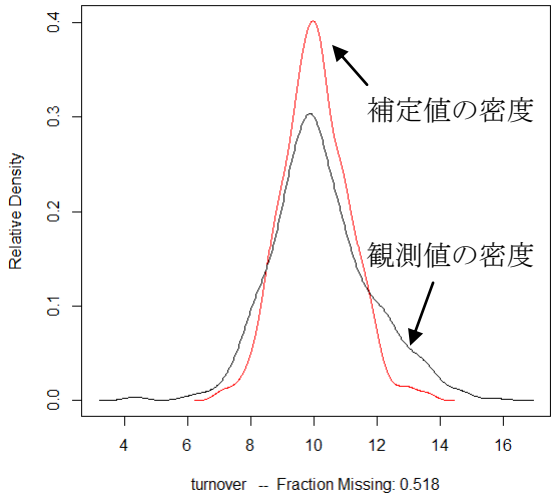
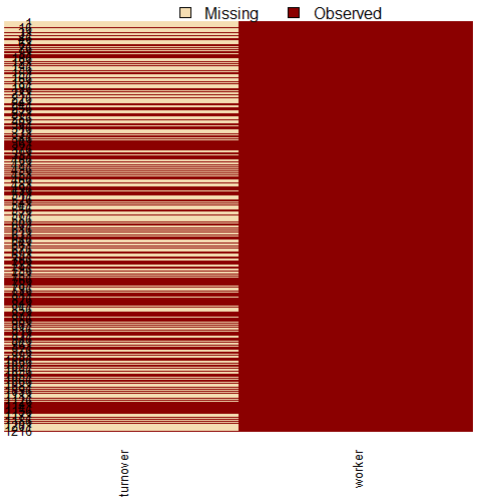


図 9.18：欠測地図(MCAR)



9.3.2　MAR（事業従事者数小）、EDINET 産業 E、欠測率 50%、自然対数の診断

図 9.19 では、2 つの密度は完全に分離しており、欠測の発生メカニズムは MCAR ではないと推定される。図 9.20 では、事業従事者数を昇順で並べており、売上高の欠測に明らかなパターンが認められ、真の欠測メカニズムは MAR と推定できる。したがって、欠測値の密度と観測値の密度は異なっているべきだと推定でき、図 9.19 には問題がないと判断できる。実際には、これらの欠測値を作り出したメカニズムが分かっているので、上述の診断は正しいことが分かっている。

図 9.19：密度(MAR)

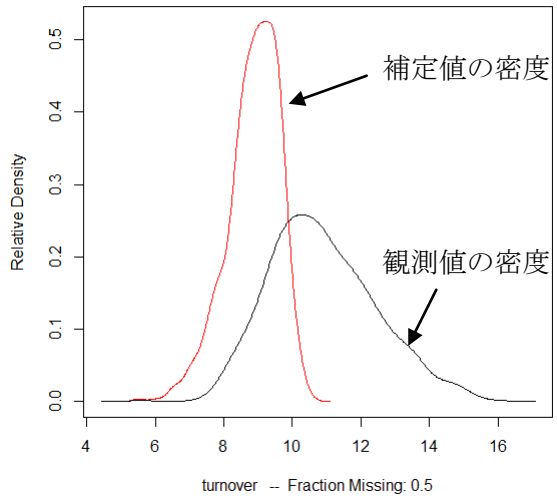
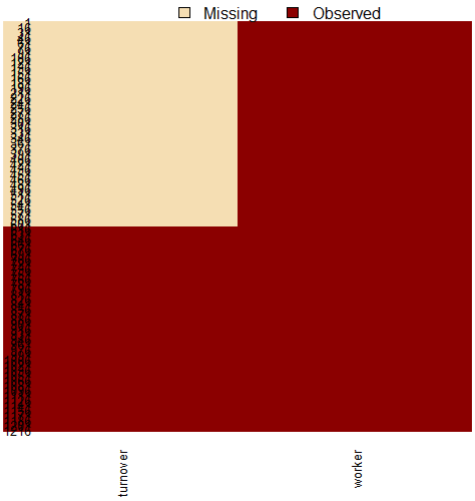


図 9.20：欠測地図(MAR)



9.3.3 MAR（事業従事者数中）、EDINET 産業 E、欠測率 50%、自然対数の診断

図 9.21 では、2 つの密度は完全に分離しており、欠測の発生メカニズムは MCAR ではないと推定される。図 9.22 では、事業従事者数を昇順で並べており、売上高の欠測に明らかなパターンが認められ、真の欠測メカニズムは MAR と推定できる。したがって、欠測値の密度と観測値の密度は異なっているべきだと推定でき、図 9.21 には問題がないと判断できる。実際には、これらの欠測値を作り出したメカニズムが分かっているので、上述の診断は正しいことが分かっている。

図 9.21：密度(MAR)

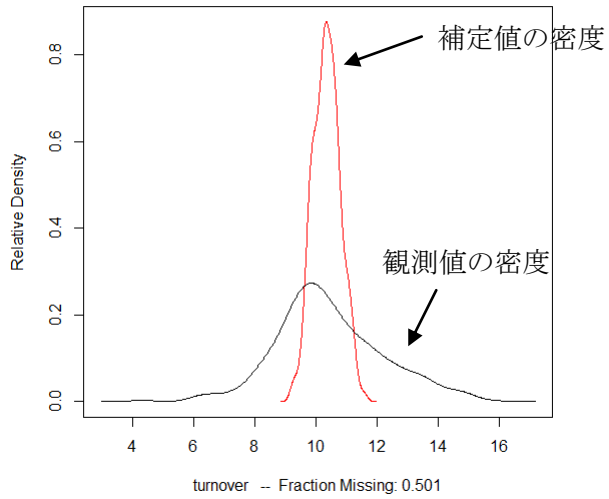
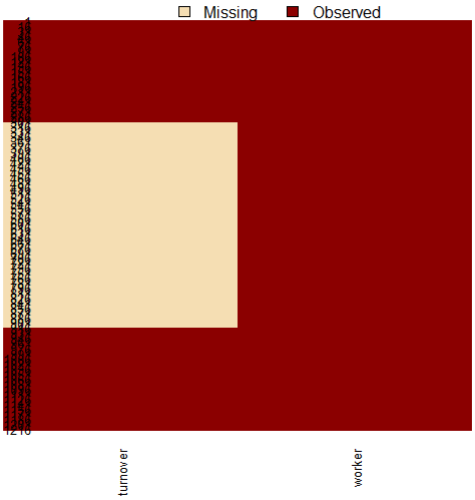


図 9.22：欠測地図(MAR)



9.3.4　MAR（事業従事者数大）、EDINET 産業 E、欠測率 50%、自然対数の診断

図 9.23 では、2 つの密度は完全に分離しており、欠測の発生メカニズムは MCAR ではないと推定される。図 9.24 では、事業従事者数を昇順で並べており、売上高の欠測に明らかなパターンが認められ、真の欠測メカニズムは MAR と推定できる。したがって、欠測値の密度と観測値の密度は異なっているべきだと推定でき、図 9.23 には問題がないと判断できる。実際には、これらの欠測値を作り出したメカニズムが分かっているので、上述の診断は正しいことが分かっている。

図 9.23：密度(MAR)

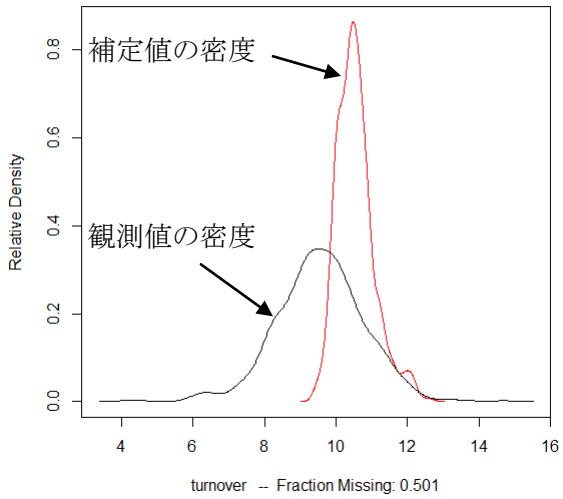
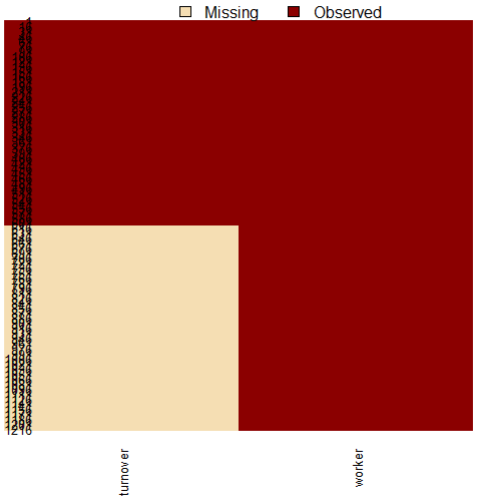


図 9.24：欠測地図(MAR)



9.3.5　MAR（事業従事者数大小）、EDINET 産業 E、欠測率 50%、自然対数の診断

図 9.25 では、2 つの密度は完全に分離しており、欠測の発生メカニズムは MCAR ではないと推定される。図 9.26 では、事業従事者数を昇順で並べており、売上高の欠測に明らかなパターンが認められ、真の欠測メカニズムは MAR と推定できる。したがって、欠測値の密度と観測値の密度は異なっているべきだと推定でき、図 9.25 には問題がないと判断できる。実際には、これらの欠測値を作り出したメカニズムが分かっているので、上述の診断は正しいことが分かっている。

図 9.25：密度(MAR)

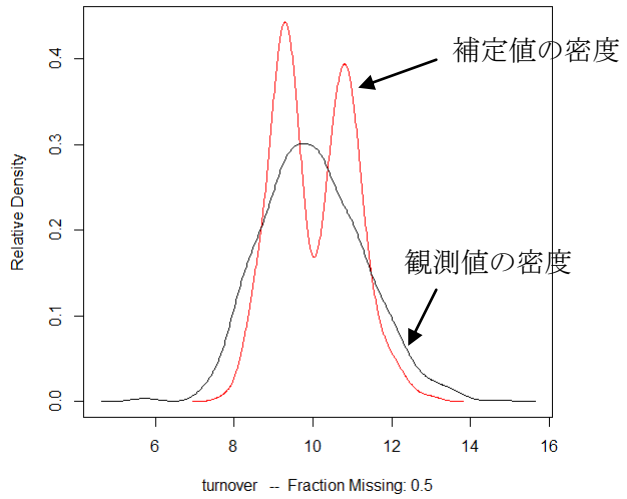
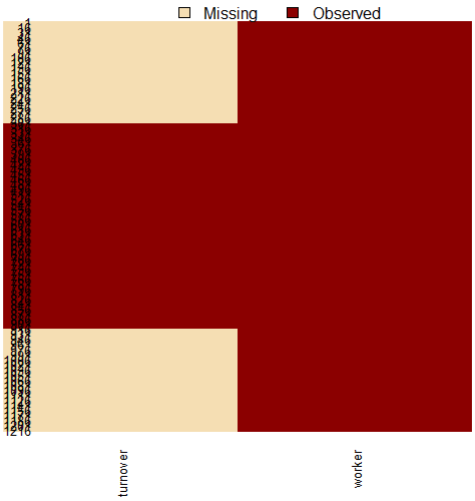


図 9.26：欠測地図(MAR)



9.3.6 MAR（系統抽出）、EDINET 産業 E、欠測率 50%、自然対数の診断

図 9.27 では、2 つの密度はほぼ重なっており、欠測の発生メカニズムは MCAR だと推定される。しかし、図 9.28 では、事業従事者数を昇順で並べたところ、一見すると売上高の欠測はランダムのように見えるが、欠測が一定間隔で発生していることが分かる。すなわち、欠測メカニズムは MAR であると推定できる。実際には、これらの欠測値を作り出したメカニズムが分かっているので、上述の診断は正しいことが分かっている。

図 9.27：密度(MAR)

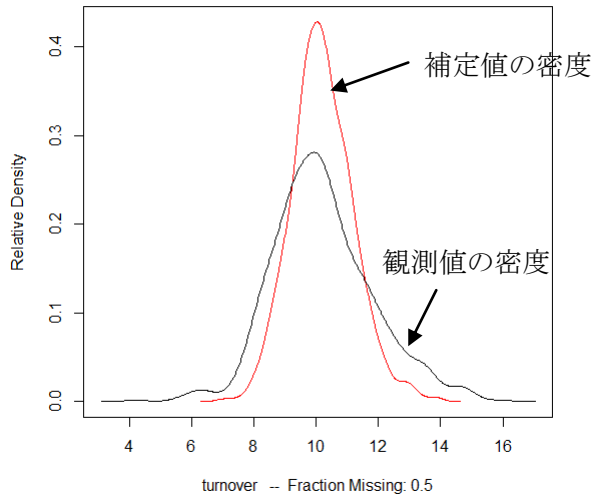
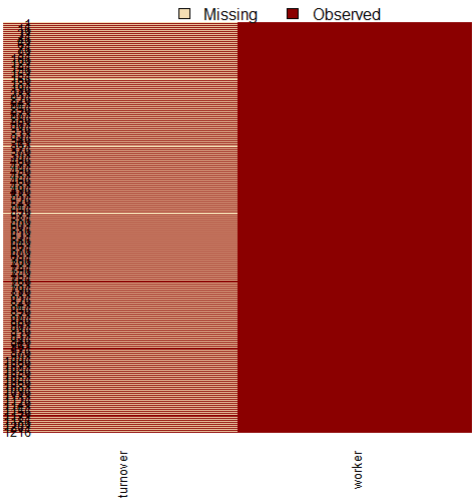


図 9.28：欠測地図(MAR)



9.4 まとめ

Amelia II において利用可能な補定の診断手法、すなわち、密度の比較機能及び欠測地図機

能を駆使することにより、真の欠測メカニズムを効率よく推定できることが分かった。また、過剰補定機能によって、補定モデルの精度を間接的に検証することもできる。さらに、過散布初期値機能により、EM アルゴリズムが大局解に収束したかどうかを推定確認できる。

10　多重代入法の精度評価：多重代入値と単一代入値の分布比較

10.1　極限における多重代入値の平均

多重代入法の M を無限大にし、無限個の補定値の平均を取ったならば、その値は何に収束するのであろうか？　EDINET の産業 E（製造業）のデータを用いて、欠測率 50%において、事業従事者が大きいデータのみを欠測させたパターンに基づき、1 次多項式で多重代入を行った。その結果を図 10.1～10.4 に示す。

図 10.1：M = 5 (乖離率 19.63%)

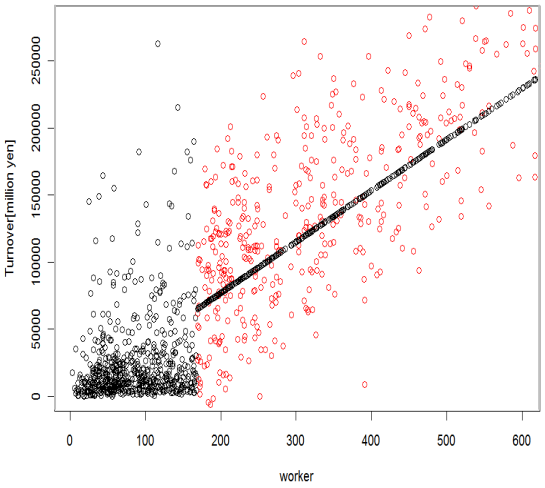


図 10.2：M = 20 (乖離率 7.23%)

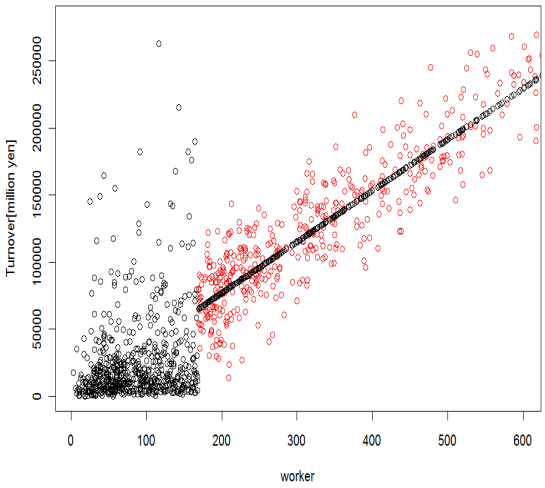


図 10.3：M = 100 (乖離率 3.72%)

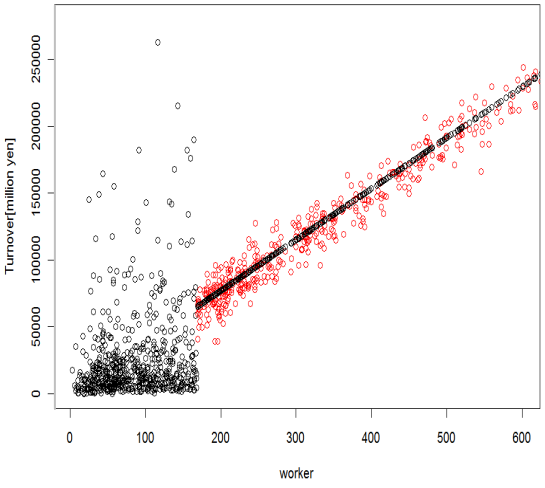


図 10.4：M = 30000 (乖離率 0.18%)

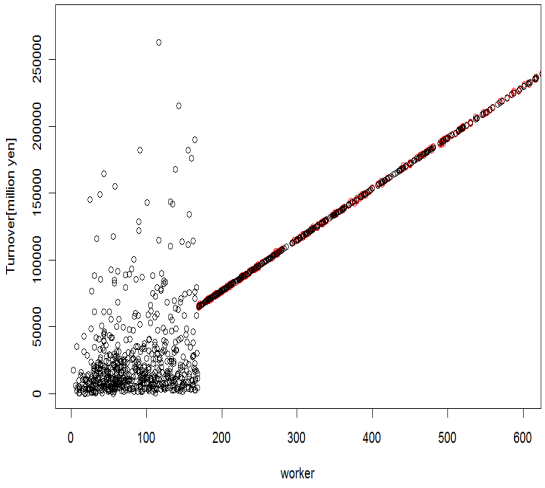


図 10.1 から図 10.4 まででは、一列を形成している黒丸が確定的補定による単一代入値であり、その周辺に散らばっている赤丸が多重代入値である。図 10.1 は $M = 5$ であり、図 10.2 は $M = 20$ であり、図 10.3 は $M = 100$ であり、図 10.4 は $M = 30000$ である。ここから分かるとおり、M のサイズが無限大に近づくにつれ、多重代入値の平均は確定的補定による単一代入値に近づくことが分かる[34]。図 10.4 では、多重代入値の平均と単一代入値は、ほぼ完全に重なり、図上において区別がつかなくなっている。

したがって、多重代入法では、単一代入値を中心とする多数の補定値を作り出していることが分かる。すなわち、無限個の補定値の平均は、単一代入値に収束するのである。それでは、なぜ単一代入法ではなく、多重代入法を使用する必要があるのだろうか？　8 節で示したとおり、多重代入値は単一代入値よりも真値に近いことが分かったが、もし極限において 2 つが同一であるならば、なぜこういった結果となったのであろうか？　本節では、下記の要領でシミュレーションを行った。x は平均値 100、標準偏差 10、標本サイズ 100 の正規乱数であり、e は、平均値 0、標準偏差 15 の正規乱数である。y は、$5+2x+e$ によって生成された x と e の 1 次関数である。上記のデータセットの y の値を人工的に 1 つ欠測させ、その補定値を以下のとおり検証した。ここで、**y の真値**は **220.5** であり、**単一代入値**は **236.7** である。$M = 1000$ の**多重代入値**の平均は **237.3** であり、真値よりも単一代入値に限りなく近い。

10.2　多重代入値の分布

表 10.1 は、$M = 1000$ の多重代入値の基本統計量である。

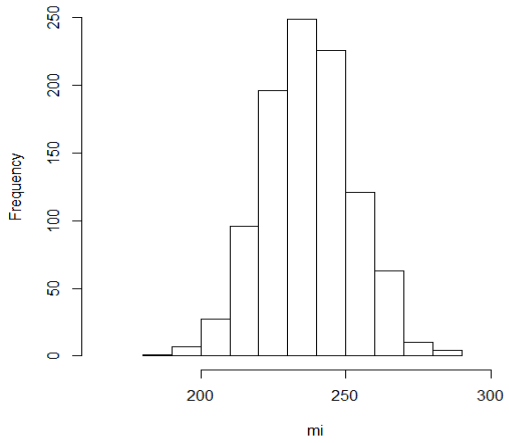
表 10.1：**多重**代入値の基本統計量

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
189.0	226.8	237.3	**237.3**	247.0	287.9	15.227

図 10.5 は、$M = 1000$ の多重代入値のヒストグラムである。図 10.5 は、平均値 237.3 を中心とする正規分布となっている。ここで、S（歪度）は 0.079 であり、K（尖度）は 2.912 であり、ほぼ完璧な正規分布であった。

[34] 乖離率 $= (\Sigma|\text{多重代入値}_i - \text{単一代入値}_i|)/\Sigma\,\text{単一代入値}_i$

図 10.5：**多重**代入値のヒストグラム



M が無限大に近づけば、多重代入値のヒストグラムは、単一代入値を中心とする正規分布となる。95%の信頼度を持って、真値は 206.8 から 267.7 までの間に位置すると推定でき、事実、*y* の真値 220.5 はこの区間に含まれている。

10.3　単一代入値の分布

表 10.2 は、欠測値を除いた 99 個の観測値をもとに単回帰によって単一代入法を行った際の回帰パラメータの推定値である。

表 10.2：単一代入法によるモデル

切片	傾き	*x* の値
-0.772 (15.760)	2.153 (0.156)	110.307

注：報告値は、係数（標準誤差）の順

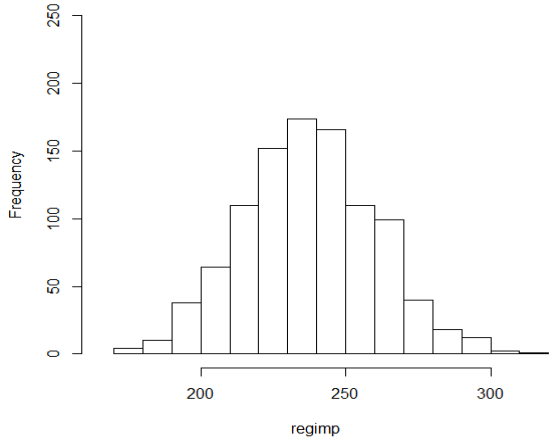
回帰パラメータの推定値及びその標準誤差を利用し、手作業によって不確実性を導入し、単一代入法を多重化する。表 10.3 は多重化した単一代入値の基本統計量である。

表 10.3：多重化した**単一**代入値の基本統計量

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
171.5	221.5	237.2	237.6	253.4	311.8	22.881

図 10.6 は多重化した単一代入値のヒストグラムである。多重代入法と比べて、95%の信頼区間が大きく(191.8, 283.4)、効率性が落ちていることが分かる。

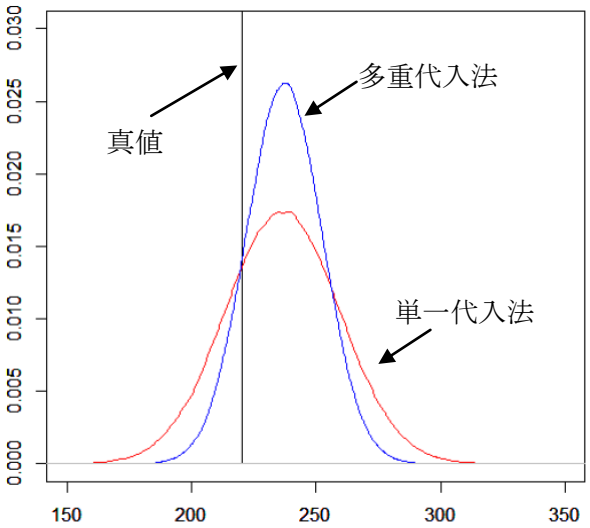
図 10.6 : 多重化単一代入値のヒストグラム



10.4　まとめ：多重代入値と単一代入値の分布比較

図 10.7 は、多重代入値の分布、多重化した単一代入値の分布、真値（縦線）を図示したものである。図 10.7 から、多重代入値の平均と単一代入値は、極限において同一となり、このシミュレーションでは、平均して、どちらの補定法も過大推定となるが、多重代入値の効率性が高いことが見て取れる。すなわち、多重代入値の過大推定は 290 より低いが、単一代入値の過大推定は 310 を超える。また、多重代入値の過小推定は、190 程度までだが、単一代入値の過小推定は 170 程度まで落ちる可能性がある。したがって、極限において多重代入値の平均と単一代入値は同一となるものの、単一代入法と比べて、多重代入法は効率性が高い補定方法なのである。

図 10.7：多重代入法と多重化単一代入法の分布及び真値



11　結語と将来の課題

本研究を通じて、多重代入法による補定の点推定値は、おおむね、単一代入法と比較して、劣るものではないことが分かった。また、多重代入法は、確定的単一代入法よりも、標準偏差の推定や分布の再現といった点で、圧倒的に優れていることも分かった。さらに、多重代入法を行うプログラムとして、R パッケージの Amelia II が有用であることも分かった。

しかし、欠測値を補定する完璧な方法は存在しない。多重代入法といえども、その例外ではなく、想定している前提が大幅に間違っていれば、補定値の精度は保証できないであろう。事実、正規性の前提を満たしていない 1 次多項式による補定の精度評価では、多重代入法のパフォーマンスはよくなかった。本研究で使用した多重代入法の診断方法は、まだ発展段階にある。多重代入データセットの精度を保証するためには、さらなる診断手法の開発が将来の課題となるであろう。こういった目的に関し、R パッケージ VIM が有用ではないかと期待される(Templ, Kowarik, and Filzmoser, 2011)。

さらに、多重代入法は、様々なアルゴリズムを用いて、様々なソフトウェアとして存在しており、Amelia だけではなく、将来的には、複数の多重代入法プログラムの優劣を比較検討したいと考えている。

また、本研究では、EDINET データにおける外れ値の存在を考慮に入れなかったが、回帰モデルの妥当性は、外れ値の影響を大きく受ける。将来の課題として、外れ値の多重代入法に与える影響を考慮に入れる必要がある。多変量外れ値検出法としては、高橋 (2012)を応用する方向性を考えている。

補論　log(y)を被説明変数として用いた回帰分析における y の予測値算出方法[35]

本研究では、自然対数変換を用いた補定を行ったが、log(y)を被説明変数として用いた回帰分析における y の予測値算出方法には、下記のとおり修正項を追加しなければならず、注意が必要である。以下、修正項の候補として 3 つの手法を紹介する(Wooldridge, 2009, pp.210-214)。変数 y を自然対数変換したものを$\log(y)$とする。式(26)のとおり、k 個の説明変数x_jに基づき、最小二乗法(OLS)によって$\hat{\beta}_j$を算出し、$\log(y)$の予測値$\widehat{\log(y)}$を算出する。

$$\widehat{\log(y)} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k \tag{26}$$

いったん$\hat{\beta}_j$が算出されれば、k 個の説明変数x_jの値を知ることで、対数変換後の y の予測値$\widehat{\log(y)}$自体は単純な値の代入により求めることができる。しかし、得られた値は対数であり、もともとの y の尺度に戻す必要がある。対数(log)は指数(exponential)を逆にしたものであるため、$\widehat{\log(y)}$を単純に指数変換すればよいと想像できる。しかし、式(27)のような単純な変換手法では、y の期待値を体系的に過小推定してしまうことが知られている。

$$\hat{y} = \exp\left(\widehat{\log(y)}\right) \tag{27}$$

そこで、式(28)のように、補正する項λ_0(lambda)を追加する必要がある。式(27)は y の期待値を体系的に過小推定してしまうことが分かっているため、λ_0は 1 よりも大きい値でなければならない。

$$\hat{y} = \lambda_0 \exp\left(\widehat{\log(y)}\right) \tag{28}$$

しかし、λ_0は不明であるため、様々な手法によって推定する必要がある。推定方法の 1 つ目の候補は式(29)である。ここで、$\tilde{\lambda}_0$は式(30)のとおり推定され、$\hat{\sigma}$は、回帰式の標準誤差(standard error of the regression)[36]である。

$$\hat{y} = \tilde{\lambda}_0 \exp\left(\widehat{\log(y)}\right) \tag{29}$$

$$\tilde{\lambda}_0 = \exp\left(\hat{\sigma}^2/2\right) \tag{30}$$

式(30)は、λ_0の不偏推定量ではないが、一致推定量である。対数変換による y の予測値を補正するλ_0には、残念ながら不偏推定量は存在しない。したがって、唯一の最良な再変換方法があるわけではない。$\hat{\sigma}^2 > 0$なので、$\tilde{\lambda}_0 > 1$となり、推定量として好ましい。しかし、式(30)は、OLS の誤差項の正規性を前提としており、OLS においては、誤差項が正規でない場合に

[35] この点に関し、統計センター統計技術研究課の和田かず美上級研究員のご指摘に感謝する。
[36] 回帰式の標準誤差は、residual standard error とも呼ばれる。

も有用な場合がある[37]。以下、誤差項の正規性を前提としない 2 つの手法を紹介する。

式(31)における$\hat{\lambda}_0$は、Duan のスミアリング推定値(Smearing Estimate)と呼ばれる方法である。ここで、$\hat{\lambda}_0$は式(32)のとおりであり、$\hat{u}_i$は式(33)のとおり OLS の残差である。$\hat{\lambda}_0$も不偏推定量ではないが、一致推定量であり、OLS の残差の標本平均は常にゼロとなるため、$\hat{\lambda}_0$は常に 1 よりも大きくなるという好ましい特性がある。また、上述したとおり、誤差項の正規性を前提としない点も好ましい。

$$\hat{y} = \hat{\lambda}_0 \exp\left(\widehat{\log(y)}\right) \tag{31}$$

$$\hat{\lambda}_0 = n^{-1} \sum_{i=1}^{n} \exp(\hat{u}_i) \tag{32}$$

$$\hat{u}_i = \log(y_i) - \widehat{\log(y_i)} \tag{33}$$

また別の手法として、λ_0を式(34)のとおり切片なしの回帰モデルに基づいて推定する方法がある。ここで、$\check{\lambda}_0$は式(35)のとおりであり、$\bar{\bar{y}}_i$は式(36)のとおりである。$\check{\lambda}_0$は、まれに 1 よりも小さな値となる場合があり、この点は好ましくないが、もしも$\check{\lambda}_0$が 1 よりも小さい場合には、誤差項uと説明変数x_jとの独立性に関する前提が守られていないことを示唆しており、使用したモデルが、そもそも、妥当ではないことを意味している。

$$\hat{y} = \check{\lambda}_0 \exp\left(\widehat{\log(y)}\right) \tag{34}$$

$$\check{\lambda}_0 = \left(\sum_{i=1}^{n} \bar{\bar{y}}_i^2\right)^{-1} \left(\sum_{i=1}^{n} \bar{\bar{y}}_i \, y_i\right) \tag{35}$$

$$\bar{\bar{y}}_i = \exp\left(\widehat{\log(y_i)}\right) \tag{36}$$

このように、対数による予測値を再変換する方法はいくつも存在する。Amelia における再変換方法は、仕様書に明示されていないため判然としていなかったが、独自に検証した結果、式(31)の$\hat{\lambda}_0$とほぼ一致していることが分かった。また、EDINET 産業 E 及び I のデータを用い、対数変換による補定を行い、下記の 5 つの手法により再変換をし、いずれの手法が最も真値に近かったかについて検証した：残差指数の平均$\hat{\lambda}_0$；残差分散の指数$\tilde{\lambda}_0$；切片なしの残差付与$\check{\lambda}_0$;非線形最小二乗法;残差付与なし。その結果を表 A.1 に示す。これらの手法の中で、$\hat{\lambda}_0$のパフォーマンスが最も優れていることが分かった。

表 A.1

順位	1	2	3	4	5
モデル	残差指数の平均$\hat{\lambda}_0$	残差分散の指数$\tilde{\lambda}_0$	切片なし残差付与$\check{\lambda}_0$	非線形最小二乗法	残差付与なし

[37] OLS の誤差項が正規分布ではない場合、OLS 推定量も正規分布せず、*t* 統計量も *t* 分布しないため、OLS の誤差項は正規分布していることが望ましい。しかし、たとえ OLS の誤差項が正規分布していなかったとしても、中心極限定理により、OLS 推定量は漸近的正規性を満たすことが知られている。すなわち、十分に大きな標本サイズにおいて、OLS 推定量は正規分布を近似するということであり、この場合、OLS の誤差項は必ずしも正規分布している必要がないことになる。詳しくは、Wooldridge (2009, pp.172-176)を参照されたい。

付録：Amelia による多重代入データセットの簡便な保存方法

7.2 節で紹介した write.amelia 関数により多重代入済データセットを出力した場合、手作業による編集が煩雑となる。そこで、下記のとおり、多重代入済データセットを簡便に一括保存できるコードを独自に開発した。コード内の赤字で記されている部分は、汎用化できない箇所なので、該当する情報を手入力する。

```
#事前準備
setwd("D:/My Documents/フォルダ名")                #フォルダ指定
data<-read.csv("データ名.csv",header=TRUE)         #データ読み込み
n <- 数                                            #多重代入法の M 数を手入力
attach(data)                                       #データ付置
set.seed(1223)                                     #シード設定
library(Amelia)                                    #Amelia 起動

#多重代入済データセット格納
a.out <- amelia(data, m = n)                       #多重代入
mat <- matrix(NA,nrow(data),n)                     #マトリックスの初期値
for( i in 1:n){
  yimp <- a.out$imputations[i]
  yimp <- data.frame(yimp)
  yimpy <- yimp[1]
  for (ii in 1:nrow(data)){mat[ii,i] <- yimpy[ii,]}
}
ameliadata <- data.frame(data,mat)

#出力（ファイル名：outdata.csv）
write.csv(ameliadata,file="outdata.csv")
```

注意点としては、補定を行いたい変数をデータ内の 1 列目に格納する。参考までに、表 F.1 にデータセットの例を示す。表 F.1 では、testy の空欄は、欠測である。したがって、欠測のある testy を 1 列目に格納している。2 列目には、欠測のない説明変数である testx を格納している。

表 F.1

testy	testx
	-1.011010
	1.148615
⋮	⋮
47.59753	18.314400
50.30039	18.845700

参考までに、$M = 3$ の多重代入により出力したファイルの例を表 F.2 に示す。ファイル内の 1 列目は通し番号となっており、2 列目の testy は欠測を含んでいる補定対象の変数（NA は欠測値を表す）、testx は説明変数、X1 は m = 1 の多重代入データセット、X2 は m = 2 の多重代入データセット、X3 は m = 3 の多重代入データセットである。

表 F.2

	testy	testx	X1	X2	X3
1	NA	-1.011010	15.38001	6.376609	-4.74943
2	NA	1.148615	8.988338	5.714070	19.97219
⋮	⋮	⋮	⋮	⋮	⋮
999	47.59753	18.31440	47.59753	47.59753	47.59753
1000	50.30039	18.84570	50.30039	50.30039	50.30039

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: 273-291.
2. Allison, Paul D. (2002). *Missing Data*. CA: Sage Publications.
3. Bender, Stefan, Jörg Drechsler, Agnes Dundler, Susanne Rässler, and Thomas Zwick. (2006). "A New Approach for Disclosure Control in the IAB Establishment Panel – Multiple Imputation for a Better Data Access," *Work Session on Statistical Data Editing, United Nations Economic Commission for Europe*, Bonn, Germany, 25-27 September 2006.
4. Burg, Thomas. (2008). "Estimation of Preliminary Unemployment Rates by Means of Multiple Imputation," *Work Session on Statistical Data Editing, United Nations Economic Commission for Europe*, Vienna, Austria, 21-23 April 2008.
5. Congdon, Peter. (2006). *Bayesian Statistical Modelling*, Second Edition. West Sussex: John Wiley & Sons Ltd.
6. Cranmer, Skyler J. and Jeff Gill. (2012). "We Have to Be Discrete About This: A Non-Parametric Imputation Technique for Missing Categorical Data," *British Journal of Political Science*, forthcoming.
7. DeGroot, Morris H. and Mark J. Schervish. (2002). *Probability and Statistics*. Boston: Addison-Wesley.
8. de Waal, Ton, Jeroen Pannekoek, and Sander Scholtus. (2011). *Handbook of Statistical Data Editing and Imputation*. Hoboken, NJ: John Wiley & Sons.
9. 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.
10. Enders, Craig K. (2010). *Applied Missing Data Analysis*. New York: Guilford Press.
11. Gelman, Andrew, and Jennifer Hill. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. New York: Cambridge University Press.
12. Gill, Jeff. (2008). *Bayesian Methods—A Social Sciences Approach*, Second Edition. London: Chapman & Hall/CRC.
13. Greene, William H. (2003). *Econometric Analysis*, Fifth Edition. New Delhi: Pearson Education, Inc.
14. Gujarati, Damodar N. (2003). *Basic Econometrics*, Fourth Edition. New York: McGraw-Hill.
15. Harris, Kenneth W. (2002). "Use of Data Editing and Multiple Imputation in Health Surveys," *Work Session on Statistical Data Editing, United Nations Economic Commission for Europe*, Helsinki, Finland, 27-29 May 2002.
16. 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.
17. Honaker, James, Gary King, and Matthew Blackwell. (2011). "Amelia II: A Program for Missing Data," *Journal of Statistical Software* vol.45, no.7.
18. Honaker, James, Gary King, and Matthew Blackwell. (2012a). *Amelia II: A Program for Missing Data* Version 1.6.1. http://cran.r-project.org/web/packages/Amelia/vignettes/amelia.pdf. (Accessed on December 20, 2012).

19. Honaker, James, Gary King, and Matthew Blackwell. (2012b). *Package 'Amelia'* Version 1.6.1. http://cran.r-project.org/web/packages/Amelia/Amelia.pdf. (Accessed on December 20, 2012).
20. 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: 1-22.
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: 49-69.
22. Little, Roderick J. A. and Donald B. Rubin. (2002). *Statistical Analysis with Missing Data*, Second Edition. New Jersey: John Wiley & Sons.
23. Marti, Helena and Michel Chavance. (2011). "Multiple Imputation Analysis of Case-Cohort Studies," *Statistics in Medicine* vol.30, no.13: 1595-1607.
24. 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.
25. Rubin, Donald B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons.
26. Schafer, Joseph L. (1999). "Multiple Imputation: A Primer," *Statistical Methods in Medical Research* vol.8: 3-15.
27. Schmidt, Katrin. (2009). "Multiple Imputation with Standard Software: First Application Experiences," *Work Session on Statistical Data Editing, United Nations Economic Commission for Europe*, Neuchâtel, Switzerland, 5-7 October 2009.
28. Shadish, William R., Thomas D. Cook, and Donald T. Campbell. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin Company.
29. Shao, Jun. (2002). "Replication Methods for Variance Estimation in Complex Surveys with Imputed Data," in *Survey Nonresponse* edited by Robert M. Groves, Don A. Dillman, John L. Eltinge, Roderick J. A. Little. New York: John Wiley & Sons, pp.303-314.
30. Shao, Jun and Dongsheng Tu. (1995). *The Jackknife and Bootstrap*. New York: Springer.
31. 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, United Nations Economic Commission for Europe*, Oslo, Norway, 24-26 September 2012.
32. Templ, Matthias, Alexander Kowarik, and Peter Filzmoser. (2011). "Imputation of Complex Data With R-Package VIM: Traditional and New Methods Based on Robust Estimation," *Work Session on Statistical Data Editing, United Nations Economic Commission for Europe*, Ljubljana, Slovenia, 9-11 May 2011.
33. Wooldridge, Jeffrey M. (2002). *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.
34. Wooldridge, Jeffrey M. (2009). *Introductory Econometrics—A Modern Approach*, Fourth Edition. Mason: South-Western.
35. Yucel, Recai M. (2011). "State of the Multiple Imputation Software," *Journal of Statistical Software* vol.45, no.1.

参考文献（日本語）

36. 青木繁伸. (2009). 『Rによる統計解析』, オーム社, 東京.
37. 伊藤孝之. (2011). 「経済センサス‐活動調査における経理項目の補定方法について」, 平成23年度第1回統計技術研究会, 独立行政法人統計センター, 平成23年12月26日.（非公開）
38. 岩崎学. (2002). 『不完全データの統計解析』, 東京, エコノミスト社.
39. 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』, 森北出版, 東京.
40. 金融庁. (2011). *EDINET-Electronic Disclosure for Investors' NETwork*. http://info.edinet-fsa.go.jp. (2012年 12 月 20 日アクセス).
41. 西郷浩. (2004). 「無回答データの補定に関する研究の動向」, 『2003年度データエディティング研究会報告』, 独立行政法人統計センター研究センター.（非公開）
42. 西郷浩. (2010). 「補定方法の最近の発展」, 『2009年度統計技術研究会報告』, 独立行政法人統計センター研究主幹.（非公開）
43. 総務省統計局統計調査部. (2011). 『平成 21 年経済センサス‐基礎調査（確報）結果の公表』. http://www.stat.go.jp/data/e-census/2009/kakuho/gaiyou/pdf/youyaku.pdf. (2012 年 12 月 20 日アクセス).
44. 高橋将宜. (2012). 「諸外国のデータエディティング及び混淆正規分布モデルによる多変量外れ値検出法についての研究」, 『製表技術参考資料17』, 独立行政法人統計センター.
45. 竹村彰通, 谷口正信. (2003). 『統計学の基礎I―線形モデルからの出発』, 東京, 岩波書店.
46. 中村永友, 小西貞則. (1998).「情報量基準に基づく多変量正規混合分布モデルのコンポーネント数の推定」, 『応用統計学』 vol.27, no.3: 165-180.
47. 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: 79-95.
48. 星野崇宏. (2009). 『調査観察データの統計科学―因果推論・選択バイアス・データ融合』, 東京, 岩波書店.
49. 村田磨理子, 畠山昌子, 磯部祥子, 亀本薫. (2008). 「サービス業基本調査における経理項目の補定法」, 『製表技術参考資料8』, 独立行政法人統計センター.
50. 宮本道子, 安藤雅和, 逸見昌之, 山下智志, 高橋淳一. (2012). 「中小企業データベースに基づく欠測を考慮した信用リスク評価について」, 2012 年度統計関連学会連合大会, 北海道大学札幌キャンパス.
51. 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.
52. 和田かず美. (2012).「多変量外れ値の検出～繰返し加重最小二乗(IRLS)法による欠測値の補定方法～」, 『統計研究彙報第69号』, 総務省統計研修所.