製表技術参考資料 30

ISSN　1348－7345
ISSN－L 1348－7329

公的統計における欠測値補定の研究：多重代入法と単一代入法

NSTAC

Working Paper No.30

平成 27 年 6 月

独立行政法人　統計センター

製表技術参考資料は、独立行政法人 統計センターの職員がその業務に関連して行った製表技術に関する研究の結果を紹介するためのものである。

ただし、本資料に示された見解は、執筆者の個人的見解である。

目　次

公的統計における欠測値補定の研究：多重代入法と単一代入法*

高橋将宜†、阿部穂日††、野呂竜夫†††

要　旨

日本国内におけるすべての事業所・企業を対象とし、経理項目を網羅的に調査する経済センサス‐活動調査が、2012 年 2 月に我が国で初めて実施された。経済センサスは、経済統計における母集団情報を整備する基幹統計と位置づけられる重要な統計調査である。しかし、調査データでは、すべての項目についてデータが得られる保証はなく、欠測により発生する偏りに対処しなければならない。そこで、独立行政法人統計センター（以下、「統計センター」という。）では、多重代入法を始めとする欠測値補定の研究を実施してきた。本稿は、我が国の公的統計における統計作成に資する材料として、その研究成果を記録として残すものである。また、統計実務者にとっての手引書としての性格上、専門的な数式を一切用いず、数値例と図表によって欠測値補定についての直感的な解説を試みたものである（数式は脚注に記載している）。本稿では、確定的回帰補定、確率的回帰補定、比率補定といった単一代入法と EMB アルゴリズムによる多重代入法に関して、各々の補定手法のメカニズム及びそれぞれの手法の長所と短所を紹介している。また、統計分析のためのフリーソフトウェア環境 R による実行方法も示した。擬似的に欠測値を発生させた平成 24 年経済センサス‐活動調査のデータ等を使用して、欠測に起因する誤差の数値評価を多重代入法によって行う。

本稿の構成は以下のとおりである。第 1 章では、本研究の経緯と背景について簡潔に説明する。第 2 章では、欠測データの問題点を示し、欠測メカニズムについて解説する。第 3 章では、単一代入法の手法を詳説する。第 4 章では、EMB アルゴリズムによる多重代入法を詳説する。第 5 章では、諸外国における多重代入法の研究事例及び導入の事例に関する調査結果を報告する。第 6 章では、平成 24 年経済センサス‐活動調査の確報データに多重代入法を適用した例を示す。第 7 章では、平成 24 年経済センサス‐活動調査の確報データを用い、多重代入法の有用性を検証する。第 8 章では、結語と今後の展望にて締めくくる。

* 渡辺美智子慶應義塾大学大学院教授には、本研究の様々な段階においてご助言・ご指摘をいただいた。また、坂下信之統計センター統計技術研究課長及び小林良行総務省統計研修所教授には、本稿のすべての原稿に目を通していただいた。ここに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著者に属する。なお、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。本研究の分析結果は、統計法第 33 条の規定に基づき、総務省・経済産業省『平成 24 年経済センサス‐活動調査』の確報結果の調査票情報を二次利用することにより、統計センターにおいて独自集計したものである。各章の主な執筆担当は以下のとおりである：
高橋（1 章、2 章、3 章、4 章、5 章、6 章、8 章、付録 1、付録 2、付録 4、コラム 1）
阿部（7 章、付録 3）
野呂（1 章、8 章、コラム 2）

† （独）統計センター統計情報・技術部統計技術研究課上級研究員（東洋大学経済学部非常勤講師）
†† 一橋大学経済研究所附属社会科学統計情報研究センター助教（統計技術研究課元研究員）
††† 内閣府大臣官房政府広報室世論調査専門官（統計技術研究課元総括研究員）

公的統計における欠測値補定の研究：多重代入法と単一代入法

高橋将宜、阿部穂日、野呂竜夫

1　研究の経緯及び背景

経済センサスは、我が国の全産業における事業所・企業を対象とする全数調査である。平成21年７月には事業所及び企業の活動の状態を把握することに重点を置いた第１回の経済センサス‐基礎調査が総務省により実施され、平成24年２月には経理項目など経済活動の実態の把握に重点を置いた経済センサス‐活動調査が総務省と経済産業省の共同で実施された。

第II期「公的統計の整備に関する基本的な計画[1]」（平成26年３月25日閣議決定）にあるように、事業所単位における経理項目を把握することは一般的に困難であることが知られている。経済センサス‐活動調査は、売上高などの経理項目を調査事項としており、これらの項目では一定期間（１年間）継続的に計測する必要があることから無記入が発生しやすいと想定された。そこで、産業に広く共通して適用可能な欠測値の統計的処理手法を新たに検討する必要性が高まり、統計センター製表部（現：統計編成部）を通じ総務省統計局から依頼を受けて、平成22年度から統計センター研究主幹（現：統計技術研究課）において、事業所・企業調査における補定方法[2]の研究を開始した。

平成22年度から23年度にかけて、経済センサス‐活動調査第２次試験調査のデータ（平成22年２月実施）を用い検証を行った。まず、売上高を補定対象変数（被説明変数）とし、従業者数を補助変数[3]（説明変数）とする回帰による４種類の補定モデルを構築した[4]。また、売上高合計と従業者数合計の比率を係数とするモデル（「比率の当てはめ」、本稿では「比率補定」という。）を構築した。検証には、これら５種類のモデルを用い、売上金額の総和を二段階で比較した[5]。検証の結果、５つのモデルの中で平方根線形回帰と比率補定の当てはまりが良く、これら２つのうちで比率補定の方が、リスクが小さいという結論を得た。検証結果は製表技術参考資料18にまとめた(伊藤, 野呂, 阿部, 土井, 2012)。ただし、第２次試

---

[1] 下記のウェブサイトにて、pdf 版を閲覧できる。http://www.soumu.go.jp/main_content/000283567.pdf（2015 年 6 月 1 日閲覧）

[2] 「補定」とは、“imputation”の訳語である。データの得られなかった箇所に何らかの適切な値を入れる方法であり、「代入」、「補完」、「補填」、「決め付け」など、様々な訳語がある。

[3] 通常、補助変数(auxiliary variable)は、「調査前に値が分かっている変数」(土屋, 2009, p.6)を意味する。しかし、欠測値補定では、補定モデルが変数間の因果関係を表していないため、「被説明変数」を「補定の対象変数」(target variable)、「説明変数」を「補助変数」と呼ぶ(de Waal *et al.*, 2011, p.228)。

[4] ４種類の回帰モデルは、切片のない回帰モデル、自然対数線形回帰（残差分散の指数による補正）、自然対数線形回帰（非線形最小二乗法による補正）、平方根線形回帰である。また、切片のない回帰モデルは$y = \hat{\beta}x$であり、自然対数線形回帰は$log(y) = \hat{\alpha} + \hat{\beta}log(x)$であり、平方根線形回帰は$\sqrt{y} = \hat{\alpha} + \hat{\beta}\sqrt{x}$である。

[5] 第一段階の検証の目的は、記入値と補定値の乖離率を判断基準とすることで、母集団推計値を適切に推定できるモデルを選定することである。つまり、補定モデルの全般的な当てはまりの良さを検証した。その結果、平方根線形回帰法と比率補定の２種のモデルの全般的な当てはまりが良いことが分かった。第二段階の検証の目的は、当てはまりの良い２つのモデルのうち、大きなリスクを生まないモデルを選定することである。すなわち、複数の欠測データを用いた検証の中で、売上金額総和の真値との最大の乖離が小さいモデルの方が、リスクが小さいモデルと判断した。

験調査のデータではデータ数が少なかった。また、比率補定のリスクが小さいことは分かったが、最適モデルと断定するには検証が限定的であった。

そこで、平成 24 年度には、欠測のメカニズムや欠測への対処法[6]における問題点を整理するとともに、Rubin (1978, 1987)が提唱した多重代入法[7]による補定を含めて、事業所・企業などを対象とする経済系調査における経理項目の欠測値補定の研究を開始した。金融庁の管理する EDINET データを用いた検証にて、多重代入法と単一代入法[8]との比較を行い、多重代入法の優位性を確認した(高橋, 伊藤, 2013)。平成 25 年度には、経済センサス‐活動調査の速報データを用い、複数の多重代入法アルゴリズムの優劣を比較し、EMB アルゴリズム（本稿第 4 章参照）による多重代入法が優位であるとの結論を得た(高橋, 伊藤, 2014)。平成 26 年度には、多重代入法を自動エディティング手法として活用する方法(Takahashi, 2014a)や補定の診断手法として活用する方法(Takahashi, 2014b)を提唱した。また、諸外国の公的統計における多重代入法の活用例を調査した(高橋, 2014)。

これまでに調査の実施時期に応じて行ってきた研究成果は、表 1.1 に記すとおりである。

表 1.1：調査の実施時期と研究の実施時期

	経済センサスの調査	統計技術研究課における研究
H21 年度	・経済センサス‐基礎調査実施 ・経済センサス‐活動調査第２次試験調査実施	
H22 年度		・伊藤, 野呂, 阿部, 土井(2012)
H23 年度	・経済センサス‐活動調査実施	
H24 年度		・高橋, 伊藤(2013)
H25 年度		・高橋, 伊藤(2014)
H26 年度	・経済センサス‐基礎調査実施	・高橋(2014) ・Takahashi (2014a) ・Takahashi (2014b)
H27 年度		・本稿刊行予定
H28 年度	・経済センサス‐活動調査実施予定	

本報告は、多重代入法と単一代入法に関する一連の研究成果を我が国の公的統計における統計作成に資する材料として記録に残すものである。

[6] 欠測への対処法の中で本稿において扱わないものとして、平均値補定、コールドデック補定、ホットデック補定がある。詳細は、高橋, 伊藤(2013, pp.27-30)を参照されたい。

[7] 「多重代入法」は、“multiple imputation”の訳語であり、他に「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」など様々に訳されることもある。本研究では、平成 24 年度の研究以来「多重代入法」で統一して表記しているので、本稿においてもこれを用いることとする。多重代入法は、補定を複数回実施して得られた複数の推定値の統合を行う方法であり、代入、分析、及び統合の三つのステージが行われる。完全データセットを複数用意しておき、解析自体は分析者の関心に任せるといった利点がある(星野, 2009, pp.220-222)。

[8] 「単一代入法」は、“single imputation”の訳語であり、データが欠測している場合に、一つの値を代入することで補完して、一つの完全データを作成する方法である。

2　欠測値の問題点と欠測のメカニズム

本章では、調査誤差と無回答[9]の種類について概観し、欠測[10]の発生メカニズムに関する前提や影響を説明する。2.1 節では、標本誤差と非標本誤差からなる調査誤差について概観する。2.2 節では、無回答の種類として、全項目無回答と一部項目無回答について、数値を用いて例証する。2.3 節では、欠測の発生メカニズムに関して、数値例と散布図を用いて視覚的に確認をしていく。2.4 節では、欠測への対処の概略を説明する。本章の目的は、欠測値による偏りを是正できる条件を示すことである。

2.1 調査の種類と誤差の種類

一般に、分析者が関心を持ち、分析の対象とする全体を母集団と呼ぶ。母集団には 2 種類あり、1 つは理論的で無限の大きさと仮定するもので無限母集団と呼ばれ、もう 1 つはある一時点における人数といった有限の大きさと考えられる有限母集団である(迫田, 高橋, 渡辺, 2014, pp.63-64)。公的統計では、通常、「2015 年 10 月 1 日現在における日本国内の在住者」といった具合で、有限母集団を想定している。この母集団からすべてのユニット（構成要素の単位）を選び出し、測定を行ってデータ化したものが全数調査によるデータである。また、母集団からその一部分を選び出し、測定してデータ化したものが標本調査によるデータである(熊原, 渡辺, 2012, pp.146-148)。

標本調査では、母集団全体を調べなかったことにより標本誤差が発生するが、標本が無作為に抽出された場合には、標本誤差を明確に数値で評価することができる。公的統計においても、家計調査年報などにおいて標本誤差を公表し、数値で評価できるようにしている(総務省統計局, 2013)。一方、母集団を測定した際、回答者による記入ミスや回答拒否などが生じることがある。こういった測定誤差は非標本誤差の一種であるが、非標本誤差を数値によって見積もることは難しい。また、全数調査においても、標本調査においても、非標本誤差は発生し得る(松田, 伴, 美添, 2000, p.55; 土屋, 2009, p.17)。

次節で説明するとおり、無回答にともなう欠測は、標本誤差として処理できる場合と非標本誤差として処理しなければならない場合とがある。

---

[9] 無回答とは、nonresponse の訳語であり、文字どおり、「回答が無い」ことを意味している。例えば、「あなたは何歳ですか？」と聞かれ、回答者が無言の場合、これは無回答である。類似の概念として「非回答」と「未回答」がある。非回答は「回答に非ず（あらず）」と読み下すことができる。「あなたは何歳ですか？」と聞かれ、「私は男です。」と答えた場合、明らかに回答として成立しておらず、非回答である。未回答は「未だ回答せず」と読み下すことができ、現時点では回答されていない状態を表す。ただし、3 つの概念の違いは必ずしも明確ではなく重複がある。無言も回答として成立していないため、無回答は非回答の一種である。また、即答していないだけでこれから回答する可能性があれば、無言は未回答である。調査終了時点まで未回答であり続けたものが無回答だと考えることができる。非回答は、無回答と誤答を含めた概念だと考えられる。欠測値補定の文献では、nonresponse の訳語として「無回答」の用語が使用されることが一般的である(松田, 伴, 美添, 2000, p.55; 岩崎, 2002, p.1; 星野, 2009, p.26; 土屋, 2009, p.198)。

[10] 欠測とは、missing の訳語であり、データが得られていない状態を指す。また、欠損や欠落とも言うが、本稿では欠測の用語を用いる。

2.2 無回答の種類とその影響

上述したとおり、調査データには、測定誤差にともなう欠測が発生するおそれがある。欠測は、構成要素単位で発生する全項目無回答[11]と変数単位で発生する一部項目無回答[12]の 2 種類に大別される。表 2.1 は、観測数 15 人、変数 2 個のシミュレーションによる模擬データである。合計で 30 個のデータが記録されているが、灰色セルは無回答により欠測していることを表し、白抜き数字は本来得られるはずの真値を表すものとする[13]。

表 2.1：シミュレーションによる調査データの例(n = 15)

ID	収入	年齢
1	543	51
2	272	24
3	797	59
4	239	26
5	415	35
6	371	34
7	650	54
8	495	47
9	553	56
10	710	55
11	421	38
12	410	42
13	386	40
14	280	29
15	514	49

注：灰色セルは無回答により欠測していることを表し、白抜き数字は本来得られるはずの真値を表すものとする。

全項目無回答とは、調査客体から調査票が回収できなかったり、回収された調査票が白紙であったりする場合を意味する。表 2.1 の ID1 がこれにあたる。全項目無回答は、無作為なサブサンプリング[14]の構造を持つことが多く、その場合の誤差は標本誤差に吸収され、標準誤差を用いることで数値評価でき、偏りは発生しない(King *et al.,* 2001, p.49)[15]。一方、

[11] Unit nonresponse
[12] Item nonresponse：単に項目無回答とも呼ぶ。
[13] 観測数 15 人の小規模なデータを用いているのは、簡単のためである。実際の補定作業では、通常の統計分析と同様に、標本サイズは大きい方が望ましい。
[14] Politis (1998)及び De Bin *et al.* (2014)は、サブサンプリングを重複なしの非復元抽出法($n > n_s$)、リサンプリングを重複ありの復元抽出法($n = n_r$)と定義している。ここで、元データの標本サイズはn、サブサンプリングの標本サイズはn_s、リサンプリングの標本サイズはn_rである。この定義では、通常のジャックナイフはサブサンプリングであり、通常のブートストラップはリサンプリングだが、ジャックナイフもブートストラップもリサンプリングと呼ぶことも多い(Shao and Tu, 1995, p.12)。日本語では、サブサンプルを副標本、リサンプルを再標本と呼ぶ(松田, 伴, 美添, 2000, p.361, p.368)。しかし、日本語の「副標本法」は、必ずしも Politis (1998)の言うサブサンプリングと同じ概念ではない(土屋, 2009, pp.185-190)。
[15] ただし、全項目無回答が偏りを発生させると疑われる場合には、再調査などによって対応することも多い。全項目無回答への対処については、松田, 伴, 美添(2000, pp.56-62)が詳しいので、そちらを参照されたい。また、Heckman の選択モデルによって偏りに対処する手法も知られており、星野(2009, pp.146-155)が詳しいので、そちらを参照されたい。本稿が対象としている欠測は一部項目無回答の方であるため、全項目無回答については、これ以上の言及はしない。

一部項目無回答とは、調査客体から回収された調査票に回答があったものの、いくつかの質問に回答がない状態である。表 2.1 の ID3 の収入の値や ID7 の年齢の値がこれにあたる。

欠測値を処理しない場合に発生する最初の問題は、統計的計算処理が不可能になることである。例えば、表 2.1 の収入の平均値（算術平均）を算出したいとしよう。もし 15 人全員のデータが観測されていれば、収入の平均値は 470.4 万円である。しかし、灰色セルの白抜き数字が欠測しているとすれば、収入の平均値は$(5716 + 収入_1 + 収入_3)/15$となり、これ以上の計算ができない。平均値すら算出できないということは、不完全データ[16]では標準偏差や相関係数などほとんどの統計処理を行うことができないことが分かるであろう。

そこで、多くの統計ソフトウェアでは、リストワイズ除去[17]により、欠測値を含む行を削除し、データを擬似的に「完全」な状態にした上で分析を行うことが多い。表 2.2 は、表 2.1 にリストワイズ除去を施したものである。

表 2.2：欠測を除去したデータ(n = 12)

ID	収入	年齢
2	272	24
4	239	26
5	415	35
6	371	34
8	495	47
9	553	56
10	710	55
11	421	38
12	410	42
13	386	40
14	280	29
15	514	49

リストワイズ除去を行えば、一見するとデータが「完全」な状態となるため、例えば、収入の平均値は$5066/12 = 422.2$万円となり、統計処理を施すことができる。しかし、この値は、12 人の収入の平均値であって、元々のデータにおける 15 人の収入の平均値（470.4 万円）ではない。この 2 つの値は、一般的に一致しない[18]。また、表 2.1 の場合、収入の欠測率は 13.3% (= 2/15)であり、年齢の欠測率も 13.3% (= 2/15)である。一方、調査項目に一部でも無回答（欠測値）が含まれている場合に欠測値を処理しないと、ユニットの行全体を脱落させることとなってしまい、標本サイズが急速に縮小する。結果、表 2.2 では、データ全

[16] 調査計画どおりに得られたデータを完全データ(complete data)と呼び、そうではないデータを不完全データ(incomplete data)と呼ぶ。本稿では、主に、欠測していないデータを完全データ、欠測しているデータを不完全データと称している(渡辺, 山口, 2000, p.1, p.31)。

[17] 英語では、list-wise deletion と言う。また、完全データ分析(complete-case analysis)やケースワイズ除去(case-wise deletion)とも言う(Baraldi and Enders, 2010, p.10)。ただし、「完全データ分析」と言った場合の「完全データ」は、脚注 16 で言うところの真の意味での「完全データ」ではないことに注意されたい。

[18] 真の平均値をμとし、リストワイズ除去による平均値をμ_{obs}とする。$(収入_1 + 収入_3 + 収入_7)/3 = \mu_{obs}$が成り立つ特殊条件の場合を除くと、一般に$\mu \neq \mu_{obs}$である。

体の欠測率が 20% (= 6/30)に増加している。ID3 の年齢の情報や ID7 の収入の情報が活かされておらず、データ資源が無駄になっていることが分かる。

2.3 欠測のメカニズムとその影響

不完全データの統計解析では、Little and Rubin (2002, pp.11-19, pp.312-315)によって提唱されている 3 つの欠測のメカニズムを考慮に入れて分析を行う[19]。どのようなメカニズムで欠測が発生しているかによって、データ分析への影響が変わるため、欠測のメカニズムを考慮に入れることが大事なのである(岩崎, 2002, pp.7-10)。詳細については、高橋, 伊藤(2013a, pp.20-25)も参照されたい。

1 つ目のメカニズムは、欠測が完全に無作為なケースである。これを Missing Completely At Random の略で MCAR と呼ぶ。これは、ある値の欠測する確率が、そのユニットのデータと無関係であることを意味する[20]。例えば、調査票を受け取った人がサイコロを転がして、1〜5 が出たら回答し、6 が出たら回答しないとする。この場合、欠測は完全に無作為であると考えられる[21]。つまり、欠測データは、完全データからの無作為なサブサンプルと見なすことができる(Allison, 2002, p.3)。標本調査における無作為抽出は、まさしくこのメカニズムを利用して、母集団から大多数の調査客体を完全に無作為な形で欠測させていると捉えることができる(渡辺, 山口, 2000, p.3)。このように、欠測が MCAR である場合には、標本誤差の範囲で数値化して欠測による誤差を評価できる。

2 つ目のメカニズムは、欠測が条件付きで無作為なケースである。これを Missing At Random の略で MAR と呼ぶ。これは、データを条件とした欠測の条件付き確率が、観測データを条件とした欠測の条件付き確率に一致することを意味する[22]。つまり、観測データを条件として、欠測確率の分布が非観測データから独立している。例えば、年齢の高い人になるほど収入について答えない確率が高くなり、データ内に年齢に関する情報が含まれていれば、収入の欠測は年齢を条件として無作為であると言える。もし欠測データが MAR である場合、欠測に対処しない分析は偏っているおそれがある。また、後述するとおり、この偏りは、補助変数を用いた補定によって是正することができる。なお、MCAR は MAR の特殊ケースであり、MCAR の場合に存在し得る偏りも補定により改善できる。

3 つ目のメカニズムは、欠測が無視できないケースである。これを Non-Ignorable の略で NI と呼ぶ[23]。これは、ある値の欠測する確率がその変数の値自体に依存しており、かつ、

---

[19] 本節は、Little and Rubin (2002)をもとに、King *et al.* (2001)、Allison (2002, pp.3-5)、van Buuren (2012, pp.6-7, pp.31-33)、Carpenter and Kenward (2013, pp.10-21)の考え方に依拠している。

[20] 数式では、$P(\mathbf{K}|\mathbf{D}) = P(\mathbf{K})$である。記号の意味は、付録 1 を参照されたい。

[21] 現実的には、特に理由もなく「たまたま」回答しなかったという状態である。

[22] 数式では、$P(\mathbf{K}|\mathbf{D}) = P(\mathbf{K}|\mathbf{D}_{\mathbf{obs}})$である。記号の意味は、付録 1 を参照されたい。

[23] 厳密には、欠測データメカニズムが無視可能(ignorable)であるためには、MAR であり、かつ、欠測発生に関するパラメータと推測の目的である母数の事前分布がお互いに無関係(distinctness)であるという 2 つの条件が満たされる必要がある(Little and Rubin, 2002, pp.119-120)。しかし、通常、実用上の目的では、MAR の条件が満たされれば欠測データモデルを無視可能と見なすことが多い(Allison, 2002, p.5; van Buuren, 2012, p.33)。すなわち、実用上の意味で NI とは NMAR(Not Missing At Random)のことである。

観測データを条件にしてもこの関係を崩すことができないことを意味する[24]。例えば、収入の高い人になるほど収入について答えない確率が高くなり、データ内に収入の欠測確率を予測できる情報が含まれていなければ、収入の欠測は無視できない。もし欠測データが NI である場合、必ずしも補定によって欠測による偏りを是正できるとは限らない。個別の欠測データに応じた対処法を採用する必要がある。

そこで、実際に MAR とはどのようなものか、表 2.3 を例にして説明する。表 2.3 では、収入 500 万円あたりを境にして、収入の高い人と低い人の組に分けられそうである。収入 500 万円以上の場合、6 人中 3 人が回答しておらず欠測確率が 50%である。一方、収入 500 万円未満の場合、9 人中 2 人が回答しておらず欠測確率が 22%である。つまり、収入の値が上がるにつれて、収入の欠測確率も上昇している。もし表 2.3 のデータ内にある情報から収入の欠測確率を予測することができなければ、欠測は無視できない NI ということになる。しかし、実際には、年齢を見ると 50 歳以上の 5 人中 3 人が欠測しており欠測確率は 60%であり、50 歳未満の 10 人中 2 人が欠測しており欠測確率は 20%である。つまり、年齢の観測データを条件として、収入における欠測の確率を予測でき、欠測は条件付きで無作為に発生する MAR と考えられる。なお、小規模データのため分かりにくいが、MAR とは、年齢の各カテゴリー内において、どの値が欠測するかは 60%と 20%の確率で無作為になっていることを意味する。ゆえに、収入の欠測は「条件付きで無作為」である。

表 2.3：調査データにおける MAR の例

ID	収入	年齢
1	543	51
2	272	24
3	797	59
4	239	26
5	415	35
6	371	34
7	650	54
8	495	47
9	553	56
10	710	55
11	421	38
12	410	42
13	386	40
14	280	29
15	514	49

注：灰色セルは無回答により欠測していることを表し、白抜き数字は本来得られるはずの真値を表すものとする。

MAR は「欠測する値に依存しない欠測」と説明されることがあるが、この説明は正確ではない。Carpenter and Kenward (2013)が述べるように、「MAR とは、ある個体の変数を

[24] 数式では、$P(\mathbf{K}|\mathbf{D}) \neq P(\mathbf{K}|\mathbf{D}_{\mathbf{obs}})$である。記号の意味は、付録 1 を参照されたい。

観測する確率がその変数の値から独立しているということを意味しているわけではない。完全に正反対であって、MAR では、ある変数を観測する可能性は、その値に依存することがある。しかし、重要なことは、観測データを条件として、この依存性が除去されるということである」(p.12)。簡単に考えてみよう。調査票を手にしたとき、「私は年齢が高いから、収入について答えないことにしよう」と考える人がいるとは想像しにくい。そうではなく、「私は収入が高いから、収入について答えないことにしよう」と考える人がいるのは容易に想像できる。すなわち、収入の欠測は、収入の値に依存して発生すると想像できる。そのように発生した収入の欠測データにおいて、年齢を条件としたときに収入の欠測確率を予測できるかどうかが鍵である。つまり、同じ年齢の値をとる人に対して、収入の欠測の可能性が収入自体の値に影響されていない場合、MAR と呼ぶのである(渡辺, 山口, 2000, pp.7-8)。MAR では、欠測が、欠測する値に無条件で依存しないわけではなく、条件付きで依存しないのである[25]。

図 2.1 は完全データにおける年齢と収入の散布図であり、図 2.2 は不完全データにおける年齢と収入の散布図である。年齢の値が高くなるにつれて、収入の欠測率が高くなるように設定されていることが分かる。

図 2.1：完全データ

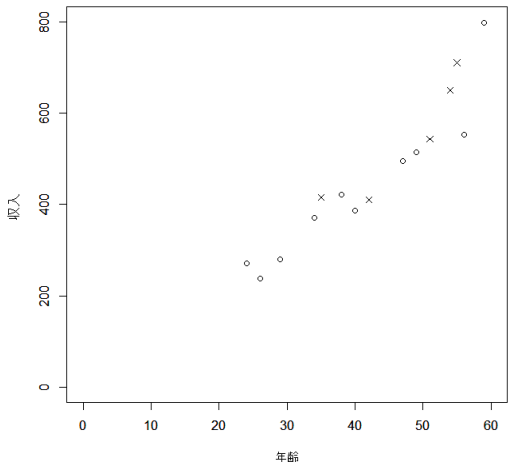
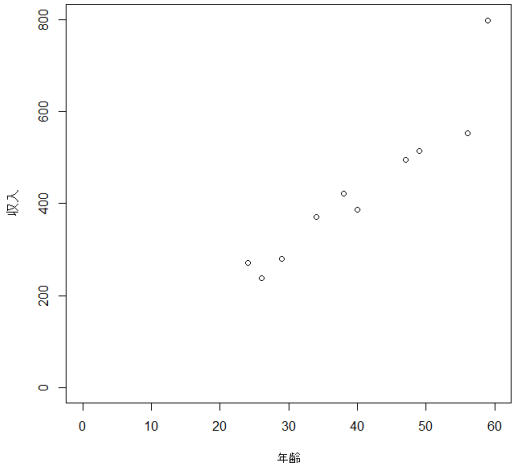


図 2.2：不完全データ



注 1：完全データの散布図では、欠測値を×印で表している。
注 2：後に利用するため、散布図の左下は(0,0)の座標を含む形で作成している。

2.4 欠測への対処

本節では、表 2.3 のデータを例に欠測への対処法を概観する。表 2.4 に示すとおり、欠測していない場合の収入の平均値は 470.4 万円だが、欠測をともなう場合の収入の平均値は 432.8 万円で、欠測値の影響で真値を過小評価して偏りが発生している。また、標準偏差や標準誤差にも影響が出ていることが見て取れる。

[25] ただし、Seaman *et al.* (2013)にあるとおり、MAR の定義には、realized MAR と everywhere MAR の 2 種類あると考えられ、研究者によって MAR の解釈が異なる場合があることに注意されたい。

表 2.4：完全データと不完全データの統計量

	真値	リストワイズ
平均値	470.4	432.8
標準偏差	161.7	166.8
標準誤差	41.8	52.8
観測数	15	10

松田, 伴, 美添(2000, pp.70-73)にあるとおり、欠測への対処は二段階に分けて考えることができる。データ収集段階における対処とデータ収集後における対処である。データ収集段階における未回答としての欠測値は、再訪問や再調査を実施し、実測値の回収に努めることが望ましい。一方、データ収集後における無回答としての欠測値は、統計的に処理する必要があり、その方法の 1 つとして補定による処理が挙げられる。その場合、本稿で説明するとおり、単一代入法（単一補定法）や多重代入法（多重補定法）など、いくつかの手法が考案されている。補定以外の無回答への対処法については、土屋(2009, pp.198-207)を参照されたい。

前節で説明したとおり、年齢を条件とすれば MAR となり、年齢を補助変数として用いる補定によって、収入における欠測データの偏りを是正できると期待される。詳しくは第 3 章で説明するが、収入の補定値の平均値は 463.5 万円となり、欠測による偏りの大部分を是正できる。実際の場面では、MAR の前提が満たされているかを観測データから直接的に検証できないが、多くの補助変数を含めることにより MAR の前提が満たされると期待できる(King *et al.*, 2001, p.51; Baraldi and Enders, 2010, p.27; van Buuren and Groothuis-Oudshoorn, 2011, p.22)。また、4.5 節で説明するとおり、欠測地図を作成したり、密度の比較を行ったりすることにより、間接的に MAR の前提を可視化して検証することも可能である[26]。

本稿は、データ収集後における一部項目無回答としての欠測値処理を論じるものである。欠測値を非標本誤差として処理する場合、多重代入法を用いることで、補定による推定誤差の数値評価を行うことができる(松田, 伴, 美添, 2000, pp.365-366; 高橋, 2015)[27]。

[26] 欠測地図と密度の比較について、詳しくは 4.5 節を参照されたい。また、Honaker *et al.* (2011)及び高橋, 伊藤(2013a, pp.64-74)も参照されたい。
[27] この数値評価は、補定モデルを条件とするものである。

3　単一代入法による補定手法

本章では、欠測による影響を是正するための補定手法を紹介する。前章から引き続き、表 2.3 にて用いたデータを例にする[28]。前述したとおり、欠測メカニズムが MAR ならば、欠測は補助変数を条件として無作為であるため、補助変数を利用した補定によって偏りを是正できる（MAR の検証は 4.5 節参照）。補定手法には、平均値補定、コールドデック補定[29]、ホットデック補定[30]など様々な方法があり、それぞれに長所と短所がある。詳しくは、高橋, 伊藤(2013a, pp.27-30)を参照されたい。本章では、特に、回帰モデルに基づく補定手法に関して、数式を用いず、図と数値による例を用いて詳説する。高橋, 伊藤(2013a, pp.30-33)も合わせて参照されたい。本章で説明する手法は、総称して単一代入法と呼ばれる。

3.1 節にて確定的回帰補定、3.2 節にて確率的回帰補定、3.3 節にて比率補定を扱う。各々の節において、散布図を用いたメカニズムの説明、補定の数値結果による例示、統計量を用いた長所と短所の比較、補定を実行するための R コードを紹介する。3.4 節にて、それぞれの手法の長所と短所をまとめる。

3.1 確定的回帰補定

3.1.1 確定的回帰補定のメカニズム

回帰補定とは、名前のとおり、回帰モデルから算出した予測値を用いて欠測値の代わりとするものである[31]。中でも、本項で説明するものは、確定的回帰補定[32]と呼ばれる。図 3.1 は収入と年齢の散布図であり、観測データをもとに算出した回帰直線（補定モデル）が示されている。白丸は観測値であり、×印は欠測値を表す。図 3.2 は確定的回帰補定による補定済みデータにおける散布図である。白丸は観測値であり、黒丸は補定値を表す。

これらの補定値は、最小二乗法[33]による予測値である。ガウス・マルコフの前提[34]と MAR の前提がすべて満たされた場合、最小二乗法による予測値は最良の線形不偏推定量による

---

[28] 本章と次章では、単一代入法と多重代入法の理論的な考え方を小規模なシミュレーションデータを用いて解説している。よって、簡単のため、データの標本サイズは 15 と小さく、検証の目的には本来ふさわしくない。本章の目的は、手法の説明であり、検証を目的としてはいない点に注意されたい。手法の検証は、本稿第 6 章と第 7 章を参照されたい。

[29] データ外情報における欠測変数の報告値を補定値とする方法をコールドデック(cold deck)と呼ぶ。

[30] データ内情報における補助変数を用いて選んだ観測値を補定値とする方法をホットデック(hot deck)と呼ぶ。補定対象のデータをレシピエントと呼び、補定値を提供する観測データをドナーと呼ぶ(Andridge and Little, 2010, p.40)。

[31] $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$を補定値とする。

[32] Deterministic Regression Imputation

[33] 最小二乗法とは、Ordinary Least Squares (OLS)の訳語で、被説明変数の実測値と回帰式から算出される予測値との差を二乗して足し上げたものが最小になるように回帰直線を引く手法である。詳しくは、迫田, 高橋, 渡辺(2014, pp.118-123)を参照されたい。

[34] ガウス・マルコフの前提とは、下記の 5 つである：母集団モデルにおいて、y はx とε に$y = \beta_0 + \beta_1 x + \varepsilon$ としてパラメータ線形で関係していること；母集団からサイズn の標本を無作為抽出していること；xに変動があり、説明変数間に完全な線形関係がないこと； x を条件とした場合、誤差項ε の期待値は 0 であること；x を条件とした場合、誤差項ε の分散は均一であること。この 5 つの前提のもとで、最小二乗法による回帰係数は、母集団パラメータの最良線形不偏推定量(Best Linear Unbiased Estimator: BLUE)である。詳しくは、Wooldridge (2009, pp.84-104)なども参照されたい。

補定値である。しかし、確定的回帰補定による補定値は、すべて、回帰線の真上に乗っていることが視覚的に分かる。すなわち、誤差が考慮されておらず、原データに存在していた真値のばらつきを封じていることが分かるであろう。

図 3.1：モデルと原データ

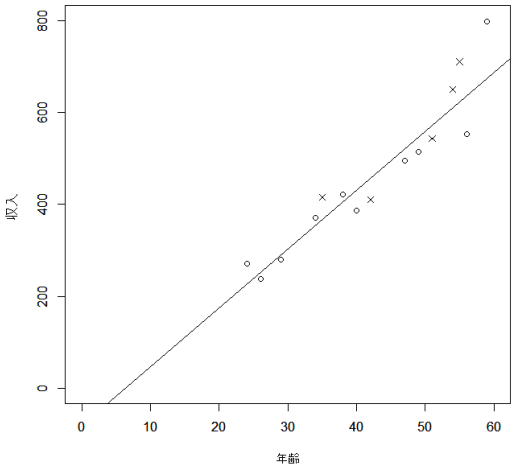
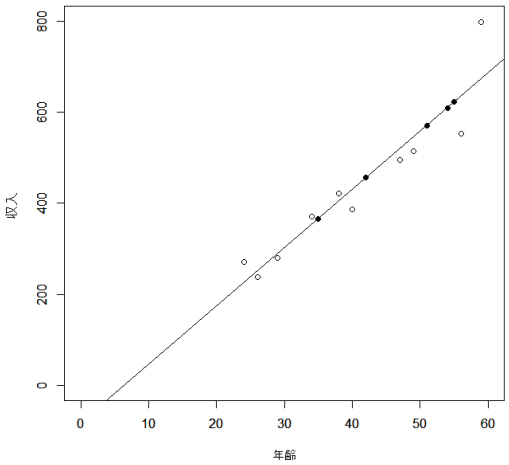


図 3.2：モデルと補定済みデータ



注 1：切片 ＝ －80.8、傾き ＝ 12.8（回帰係数は観測データのみから算出し、図 3.1 と図 3.2 で共通）
注 2：○は観測値、×は欠測値、●は補定値を表す。

3.1.2 確定的回帰補定の例

表 3.1 は、上記の方法による確定的回帰補定の結果である。すなわち、観測されたデータをもとに回帰分析を行って回帰係数を算出し、得られた予測値を欠測値の代わりとして採用したものである。表 3.1 は、図 3.2 の元となるデータである。

表 3.1：確定的回帰補定による補定済みデータの例

ID	収入	年齢	補定値	差の二乗
1	543	51	*571*	784
2	272	24	272	0
3	797	59	797	0
4	239	26	239	0
5	415	35	*366*	2401
6	371	34	371	0
7	650	54	*609*	1681
8	495	47	495	0
9	553	56	553	0
10	710	55	*622*	7744
11	421	38	421	0
12	410	42	*456*	2116
13	386	40	386	0
14	280	29	280	0
15	514	49	514	0
平均	433	43	464	2945.2

注：灰色セルは無回答により欠測していることを表し、白抜き数字は本来得られるはずの真値を表すものとする。イタリック（斜体）は補定値を表す。差の二乗は、収入と補定値との差を二乗したものである。

表 3.1 をよく見てみると、ID1 の収入の真の値は 543 万円だが、補定値は 571 万円となっている。543 – 571 = –28 が、欠測によって発生した誤差である。同様に計算すると、ID5 の場合の欠測によって発生した誤差は 415 – 366 = 49 である。これらを二乗して平均すると 2945.2 となり、欠測による誤差がどれだけあるかが分かる。すなわち、補定によって偏りを是正できるとは言っても、個別の値 1 つ 1 つを完全に復元できているわけではない。

なお、欠測値補定における目標は、個別の値を完全に復元することよりも、母集団パラメータの正確な推定を可能にすることである。母集団パラメータ（母数）とは、平均値や標準偏差といった母集団分布の形状を決める値のことである。

3.1.3 確定的回帰補定の長所と短所

表 3.2 は、表 3.1 を用いて算出した基本統計量(確定回帰)である。比較対象として、真値とリストワイズ除去による値を掲載している。

表 3.2：確定的補定済みデータの統計量

	真値	リストワイズ	確定回帰
平均値	470.4	432.8	463.5
標準偏差	161.7	166.8	152.9
標準誤差	41.8	52.8	39.5
観測数	15	10	15

確定的回帰補定による平均値(463.5 万円)は、前提が満たされた場合には不偏推定量であり、実際に補定を行うことで真値(470.4 万円)の復元へと近づいている。MAR の前提が近似的なものであり、完全な復元は達成できないが、リストワイズ除去(432.8 万円)と比べて、非常に良い値になっていることが伺える。

しかし、図 3.2 の散布図から明らかだったように、原データに存在していたばらつきが復元されておらず、確定的回帰補定による標準偏差(152.9)は真値(161.7)を過小推定している。また、それにともなって、確定的回帰補定による標準誤差(39.5)も真値(41.8)を過小推定している。

確定的補定は、その名のとおり乱数を用いない手法であるため、シード値を設定する必要がなく、毎回、同じ結果を得ることができる。繰り返し計算なども行わないため、計算負荷は非常に低い。また、今回は単回帰モデルを用いたが、データ内に利用できる補助変数が複数あれば、重回帰モデルとして複数の補助変数を用いて補定の精度を向上させることができる。前述したとおり、できるだけ多くの補助変数を利用できれば、MAR の前提が満たされる可能性が高くなると期待でき(King *et al.,* 2001)、この点は重要である。

本節で示した結果は、シミュレーション（模擬実験）によるデモであるため、真値が分かっている。よって、真値と補定値との差を直接計算して、欠測による推定誤差を算出できた。しかし、現実には、真値は不明であり、確定的回帰補定による単一代入法では、欠測による推定誤差を評価することができない。

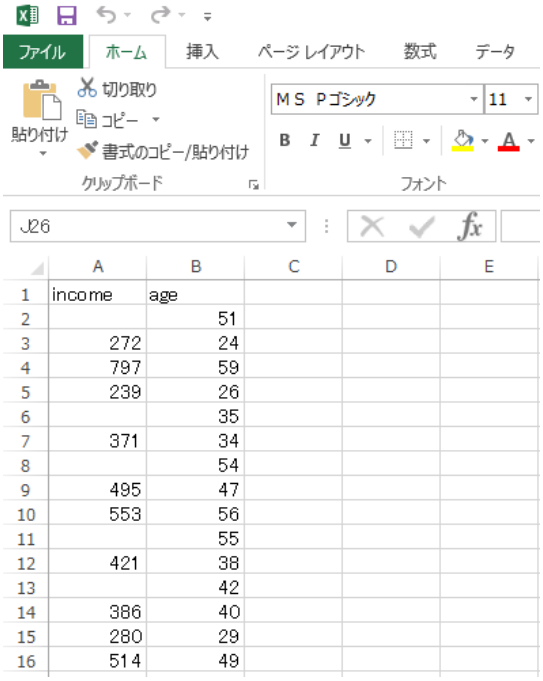
3.1.4 確定的回帰補定の R コード

下記のコードにて、使用する「データ名」と補定する「変数名」を指定し、R に貼り付ければ、確定的回帰補定による補定済みデータセット(dsidata.csv)を作成することができる。R の基本的な使用方法については、青木(2009, pp.1-10)を参照されたい。

```
data<-read.csv("データ名.csv",header=T)  #「データ名」を指定
attach(data)
model<-lm(data[,1]~data[,2:ncol(data)],
          data=data)                     #回帰モデルの指定
imp1<-predict(model,data)                #予測値を算出
imp<-data[,1]
for(i in 1:nrow(data)){                  #欠測値を補定値に置換
 if (is.na(imp[i])=="TRUE"){
 imp[i]<-imp1[i]
 }else{
 imp[i]<-imp[i]}}
dsidata<-data.frame(data,imp)
write.csv(dsidata,"dsidata.csv")         #データを出力
```

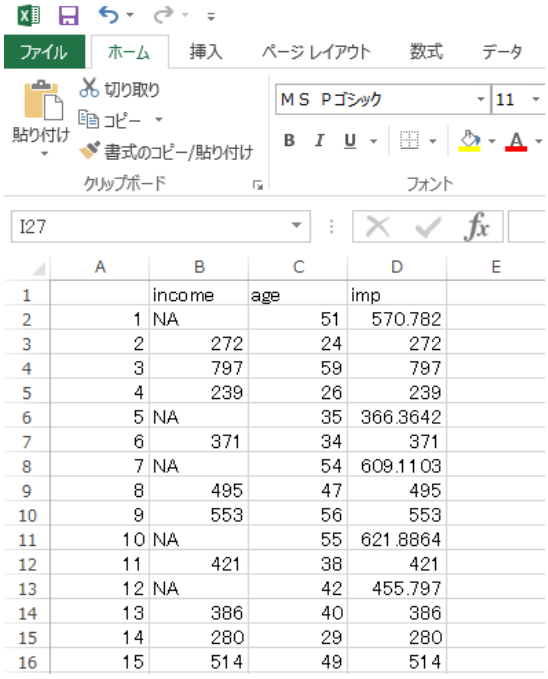
なお、このコードでは、補助変数の数はいくつあってもよいが、図 3.3 にあるとおり補定の対象となる欠測値を含む変数はデータ内の 1 列目に格納されていることを前提としている。また、図 3.4 にあるとおり、補定済み変数は imp として出力される。3.2.4 節と 3.3.4 節のコードにおいても同様である。

図 3.3：入力データの例



	A	B	C	D	E
1	income	age			
2		51			
3	272	24			
4	797	59			
5	239	26			
6		35			
7	371	34			
8		54			
9	495	47			
10	553	56			
11		55			
12	421	38			
13		42			
14	386	40			
15	280	29			
16	514	49			

図 3.4：出力データの例



	A	B	C	D	E
1		income	age	imp	
2	1	NA	51	570.782	
3	2	272	24	272	
4	3	797	59	797	
5	4	239	26	239	
6	5	NA	35	366.3642	
7	6	371	34	371	
8	7	NA	54	609.1103	
9	8	495	47	495	
10	9	553	56	553	
11	10	NA	55	621.8864	
12	11	421	38	421	
13	12	NA	42	455.797	
14	13	386	40	386	
15	14	280	29	280	
16	15	514	49	514	

3.2 確率的回帰補定

3.2.1 確率的回帰補定のメカニズム

確率的回帰補定[35]は、確定的回帰補定と同じく、回帰モデルにより予測値を算出し欠測値の代わりとして使用する。2 つの手法の違いは、確率的回帰補定では各々の補定値に無作為な乱数による誤差項を追加する点である。

図 3.1 は先ほどと同じく、収入と年齢の散布図であり、観測データをもとに算出した回帰直線（補定モデル）が示されている。白丸は観測値であり、×印は欠測値を表す。図 3.5 は確率的回帰補定による補定済みデータにおける散布図である。白丸は観測値であり、黒丸は補定値を表す。回帰線は図 3.2 と同一だが、補定値は図 3.2 とは異なり回帰線の上下にばらついていることが視覚的に分かる。

図 3.1：モデルと原データ

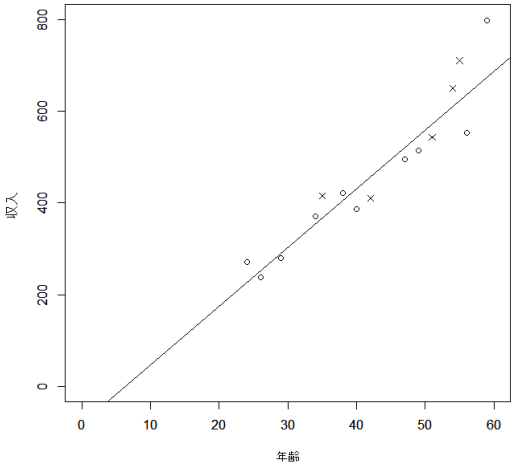
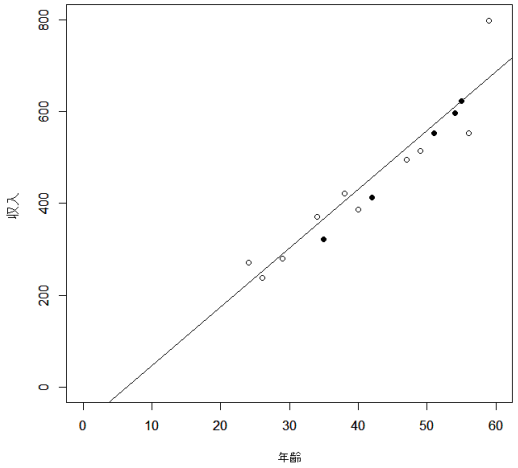


図 3.5：モデルと補定済みデータ



注 1：切片 ＝ －80.8、傾き ＝12.8（回帰係数は観測データのみから算出し、図 3.1 と図 3.5 で共通）
注 2：○は観測値、×は欠測値、●は補定値を表す。

回帰線は先ほどの確定的回帰補定の場合と同一であり、この線から得られた予測値は最小二乗法による予測値であり、ガウス・マルコフの前提と MAR の前提がすべて満たされた場合、不偏推定量である。しかし、先ほどの確定的回帰補定による補定値が回帰線の真上に乗っており、原データのばらつきを再現できていなかったのに対して、確率的回帰補定では観測されたデータをもとに回帰分析を行って予測値を算出するだけはなく、乱数に基づく誤差項を追加する。Hu *et al.* (2001, p.15)に示されているとおり、誤差項には 3 種類の方法[36]が提唱されているが、通常は、「平均値 ＝0、分散 ＝ 回帰モデルの残差分散」の正規分布から無作為に抽出する(Allison; 2002, p.29; de Waal *et al.*, 2011, p.259)。

---

[35] 撹乱的回帰補定とも言う。また、英語では、Stochastic Regression Imputation または Random Regression Imputation と言う。

[36] 3 種類の方法とは、(1)平均値を 0 とし、観測データから得られた分散を用いた正規乱数を用いる方法；(2)平均値を 0 とし、観測データをもとに算出した回帰モデルの残差分散を用いた正規乱数；(3)平均値を 0 とし、補助変数の値が似通っている観測値の残差を用いる方法の 3 種類である。

3.2.2 確率的回帰補定の例

表 3.3 は、上記の方法による確率的回帰補定の結果である。すなわち、観測されたデータをもとに回帰分析を行って回帰係数を算出し、得られた予測値に上述の方法で誤差項を付置し、欠測値の代わりとして採用したものである[37]。表 3.3 は、図 3.5 の元となるデータである。

表 3.3：確率的回帰補定による補定済みデータの例

ID	収入	年齢	補定値	差の二乗
1	543	51	*552*	81
2	272	24	272	0
3	797	59	797	0
4	239	26	239	0
5	415	35	*322*	8649
6	371	34	371	0
7	650	54	*597*	2809
8	495	47	495	0
9	553	56	553	0
10	710	55	*623*	7569
11	421	38	421	0
12	410	42	*481*	5041
13	386	40	386	0
14	280	29	280	0
15	514	49	514	0
平均	433	43	460	4829.8

注：灰色セルは無回答により欠測していることを表し、白抜き数字は本来得られるはずの真値を表すものとする。イタリック（斜体）は補定値を表す。差の二乗は、収入と補定値との差を二乗したものである。

表 3.3 をよく見てみると、ID1 の収入の真の値は 543 万円だが、補定値は 552 万円となっている。543 – 552 = –9 が、欠測によって発生した誤差である。同様に計算すると、ID5 の場合の欠測によって発生した誤差は 415 – 322 = 93 である。先ほどと同様に、これらを二乗して平均すると、欠測による誤差が 4829.8 あることが分かる。前述したとおり、補定によって偏りを是正できるとは言っても、個別の値 1 つ 1 つを完全に復元できているわけではない。欠測値補定における目標は、個別の値を完全に復元することよりも、母集団パラメータの正確な推定を可能にすることである。

3.2.3 確率的回帰補定の長所と短所

表 3.4 は、表 3.3 を用いて算出した基本統計量(確率)である。なお、シード値[38]による影響を示すため、3 つの任意のシード値による結果を掲載している(確率 1, 確率 2, 確率 3)。

---

[37] $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$を算出し、残差$\hat{u}_i = Y_i - \hat{Y}_i$を算出する。残差分散を利用して、誤差項$e_i \sim N(0, \sigma_{\hat{u}_i})$を発生させ、$\check{Y}_i = \hat{Y}_i + e_i$を補定値とする。

[38] シード値とは、擬似乱数を生成する際に使用する元となる値のことである。この値を設定することで、擬似乱数に再現性を持たせることができる。

また、比較対象として、真値、リストワイズ除去による値(LW)、確定回帰による値を掲載している。

表 3.4：確率的補定済みデータの統計量

	真値	LW	確定回帰	確率 1	確率 2	確率 3
平均値	470.4	432.8	463.5	460.2	463.3	465.8
標準偏差	161.7	166.8	152.9	153.8	155.0	153.1
標準誤差	41.8	52.8	39.5	39.7	40.0	39.5
観測数	15	10	15	15	15	15

注：LW はリストワイズ除去(List-Wise Deletion)により欠測を除去した値を表す。また、確率的回帰補定については、シード値を 3 回変えて実行した。

確率的補定による収入の平均値(460.2 万円, 463.3 万円, 465.8 万円)は、確定的補定の場合と同じく、前提が満たされた場合には不偏推定量であり、実際に補定を行うことで、リストワイズ除去(432.8 万円)と比較して、真値(470.4 万円)の復元へと近づいている。どれだけ復元されているかは、乱数の影響もあり一概には言えないが、通常は確定的補定(463.5 万円)よりも精度は劣り、リストワイズ除去(432.8 万円)よりは良い。ただし、シード値を 1 つしか利用できないため、確定的補定(463.5 万円)と比較して、確率回帰 1 の結果(460.2 万円)は偶発的に悪い結果となっている。一方、確率回帰 3 の結果(465.8 万円)は、偶発的に良い結果になっている。シード値を変更することによる影響が大きいことが見て取れる。

確率的補定は、その名のとおり乱数を用いた手法であるため、毎回、異なった結果が算出される。結果の再現性を確保するにはシード値を設定する必要があるが、補定値の精度は、どのシード値を選んだかに依存する。どのシード値による結果が良いかは、事前には分からない点が難点である。また、確定的補定と比較すると、乱数を発生させる分だけ計算負荷が高くなるが、通常の PC において影響が出るほどではない。

誤差項を追加したことにより、原データに存在していたばらつきを復元し、確定的補定の標準偏差(152.9)と比較して、確率的補定の標準偏差の推定結果(153.8, 155.0, 153.1)は真値(161.7)に近づき改善されている。また、それにともなって、確率的補定による標準誤差(39.7, 40.0, 39.5)の推定も改善されている(真値 = 41.8)。

確定的回帰補定も確率的回帰補定も、いずれの手法も不偏推定量であるため、平均して、同じ結果が得られる。ただし、多くの場合、誤差項の影響によって確率的回帰補定の方が点推定値に関する精度は自然と低くなる(de Waal *et al.*, 2011, p.231)。なお、確定的補定の場合と同じく、データ内に利用できる補助変数が複数あれば、重回帰モデルとして複数の補助変数を用いて、補定の精度を向上させることができる。

また、本節で示した検証方法は、真値が既知であることを前提としている。今回の結果は、シミュレーションによるデモであるため、真値が分かっている。よって、真値と補定値との差を直接計算して、欠測による推定誤差を算出できた。しかし、現実には、真値は不明であり、確率的回帰補定による単一代入法では、欠測による推定誤差を評価することができない。

3.2.4 確率的回帰補定の R コード

下記のコードにて、使用する「データ名」と補定する「変数名」を指定し、R に貼り付ければ、補定済みデータセット(ssidata.csv)を作成することができる。

```
data<-read.csv("データ名.csv",header=T)          #「データ名」を指定
attach(data)
model<-lm(data[,1]~data[,2:ncol(data)],
         data=data)                              #回帰モデルの指定
imp1<-predict(model,data)                        #予測値の算出
imp<-data[,1]
resid1<-imp-imp1                                 #残差の算出
set.seed(1223)                                   #シード値を指定
e<-rnorm(nrow(data),0,sd(resid1,na.rm=TRUE))     #誤差項の算出
imp2<-imp1+e                                     #補定値の算出
for(i in 1:nrow(data)){                          #欠測値を補定値に置換
 if (is.na(imp[i])=="TRUE"){
 imp[i]<-imp2[i]
 }else{
 imp[i]<-imp[i]}}
ssidata<-data.frame(data,imp)
write.csv(ssidata,"ssidata.csv")                 #データの出力
```

なお、このコードでは、補助変数の数はいくつあってもよいが、補定の対象となる欠測を含む変数は、データ内の 1 列目に格納されていることを前提としている。R の基本的な使用方法については、青木(2009, pp.1-10)を参照されたい。入出力データについては、本稿 3.1.4 節も参照されたい。

3.3 比率補定

3.3.1 比率補定のメカニズム

比率補定[39]とは、切片のない単回帰分析の構造をしたモデルをしており、補助変数と補定対象変数の比率によって回帰モデルの傾きを推定し、得られた予測値を補定値として採用するものである[40]。比推定量として知られるモデルである(土屋, 2009, pp.70-80)。

図 3.6 は、収入と年齢の散布図だが、先ほどまでとは異なり、観測データをもとに算出した原点(0, 0)を通る比率による回帰直線（補定モデル）が示されている。白丸は観測値であり、×印は欠測値を表す。図 3.7 は比率補定による補定済みデータにおける散布図である。白丸は観測値であり、黒丸は補定値を表す。図 3.2 と同じく、補定値は回帰線の真上に乗っていることが視覚的に分かる[41]。

---

[39] Ratio Imputation

[40] $\hat{Y}_i = \hat{\omega} X_i$を補定値とする。なお、$\hat{\omega} = \bar{Y}_{obs} / \bar{X}_{obs}$である。

[41] なお、3.2 節同様に誤差項を追加することで、比率補定を確率的補定にすることもできる(Hu *et al.*, 2001, pp.15-16)。

図 3.6：モデルと原データ

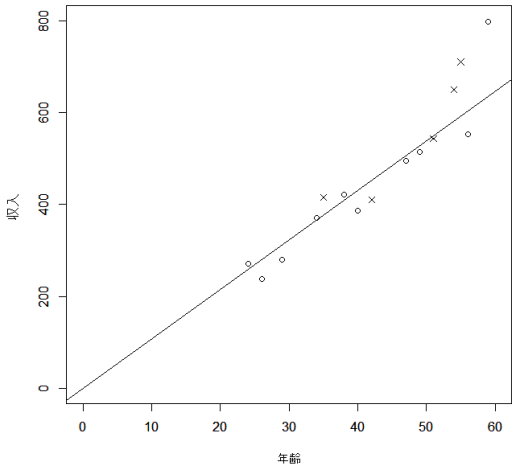
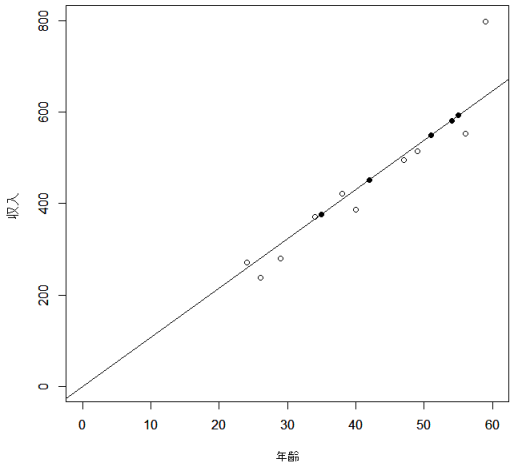


図 3.7：モデルと補定済みデータ



注 1：切片 = 0、傾き = 10.8（回帰係数は観測データのみから算出し、図 3.6 と図 3.7 で共通）
注 2：○は観測値、×は欠測値、●は補定値を表す。

回帰線は、先ほどの確定的回帰補定の場合と異なり、収入の平均値と年齢の平均値の比率である。したがって、この線から得られた予測値は、最小二乗法による予測値ではない。しかし、推定方法は非常に簡易的であるものの、モデルの前提が満たされている場合には、不偏推定量であることが知られている (Shao, 2000, p.79; Liang *et al.*, 2008, p.2)[42]。比率補定の前提とは、「補助変数の値が 0 のとき、補定対象変数の値が 0 になること」と「2 つの変数が 1 対 1 で比例関係にあること」である。すなわち、通常の回帰分析で言うところの「切片が 0 であること」と重回帰分析における「制御変数（統制変数）が必要ないこと」が前提となっている。例えば、経済データにおける今期の売上高が欠測しており、前期の売上高が観測されている場合には、比率補定の前提が満たされると考えられ、比率による補定を行うことが望ましい[43]。一方、経済データにおける売上高が欠測しており、従業者数が観測されている場合には、比率補定の前提が必ずしも満たされている保証がなく、比率による補定を行うことが望ましいかどうか、議論の余地がある。

なお、比率補定には 2 種類ある。1 つ 1 つの観測値同士の比率を計算してその平均を使用する場合と、1 つ 1 つの変数の平均値を計算してその比率を使用する場合である(Liang *et al.*, 2008, pp.2-4)。前者は比率の平均による補定であり、後者は平均の比率による補定である[44]。モデルの前提が満たされている場合には、両者ともに不偏推定量であるが、後者の平均の比率による補定の方が外れ値に対して比較的頑健な手法であるため、一般的に比率補定と言った場合には、平均の比率による補定のことを意味する (Office for National

[42] 数学的に厳密に定義すると、母集団モデルが$y_i = \omega x_i + \sqrt{x_i}\varepsilon_i$であり、$\omega$は傾きを表す回帰係数であり、$\varepsilon_i$は平均 0 と未知の分散を持ち、$x_i$から独立した誤差項である。このモデルのもとで、$\bar{Y}_{obs}/\bar{X}_{obs}$は、$\omega$の不偏推定量である(Shao, 2000, p.79)。

[43] 比率補定は、欠測変数の報告値と補助変数の情報の両方を用いる手法と見なすことができるため、ウォームデック(warm deck)と呼ばれることがある(Shao, 2000, p.80)。コールドデックとホットデックについては、脚注 29 と 30 を参照されたい。

[44] つまり、比率の平均による補定は$y = \frac{1}{n}\sum_i^n \frac{y_i}{x_i}$であり、平均の比率による補定は$y_i = \frac{\bar{y}}{\bar{x}}x_i$である。

Statistics, 2014)。

比率補定は、前提や手法が簡易的であり、散布図などの視覚情報による目視確認が容易であるため、諸外国における公的統計の実務で頻繁に使用される。例えば、米国センサス局(Thompson and Williams, 2003; Bechtel *et al.*, 2011)、英国国家統計局(Office for National Statistics, 2014)、オランダ統計局(de Waal *et al.*, 2011, pp.244-246)などにおいて、実際に使用されている補定手法である。

3.3.2 比率補定の例

表 3.5 は、上記の方法による比率補定の結果である。すなわち、観測されたデータをもとにいったんリストワイズ除去を行った上で収入と年齢の平均値をそれぞれ算出し、その比率を計算し、年齢の値に乗じたものを欠測値の代わりとして採用したものである。表 3.5 は、図 3.7 の元となるデータである。

表 3.5：比率補定による補定済みデータの例

ID	収入	年齢	補定値	差の二乗
1	543	51	*549*	36
2	272	24	272	0
3	797	59	797	0
4	239	26	239	0
5	415	35	*377*	1444
6	371	34	371	0
7	650	54	*581*	4761
8	495	47	495	0
9	553	56	553	0
10	710	55	*592*	13924
11	421	38	421	0
12	410	42	*452*	1764
13	386	40	386	0
14	280	29	280	0
15	514	49	514	0
平均	433	43	459	4385.8

注：灰色セルは無回答により欠測していることを表し、白抜き数字は本来得られるはずの真値を表すものとする。イタリック（斜体）は補定値を表す。差の二乗は、収入と補定値との差を二乗したものである。

表 3.5 をよく見てみると、ID1 の収入の真の値は 543 万円だが、補定値は 549 万円となっている。543 – 549 = －6 が、欠測によって発生した非標本誤差である。同様に計算すると、ID5 の場合の欠測によって発生した誤差は 415 – 377 = 38 である。先ほどと同様に、これらを二乗して平均すると、欠測による誤差が 4385.8 あることが分かる。前述したとおり、補定によって偏りを是正できるとは言っても、個別の値 1 つ 1 つを完全に復元できているわけではない。欠測値補定における目標は、個別の値を完全に復元することよりも、母集団パラメータの正確な推定を可能にすることである。

3.3.3 比率補定の長所と短所

表 3.6 は、表 3.5 を用いて算出した基本統計量(比率)である。比較対象として、真値、リストワイズ除去による値(LW)、確定的回帰補定による値(確定)、確率的回帰補定による値(確率)を掲載している。

表 3.6：比率補定済みデータの統計量

	真値	LW	確定	確率 1	確率 2	確率 3	比率
平均値	470.4	432.8	463.5	460.2	463.3	465.8	458.6
標準偏差	161.7	166.8	152.9	153.8	155.0	153.1	147.6
標準誤差	41.8	52.8	39.5	39.7	40.0	39.5	38.1
観測数	15	10	15	15	15	15	15

注： LW はリストワイズ除去(List-Wise Deletion)により欠測を除去した値、確定は確定的回帰補定による値、確率は確率的回帰補定による値、比率は比率補定による値を表す。また、確率的回帰補定については、シード値を 3 回変えて実行した。

比率補定による収入の平均値は 458.6 万円で、リストワイズ除去による値(432.8 万円)と比べて真値(470.4 万円)に近く、良い値になっている。手法は簡易的だが、欠測値を放置するよりも比率補定によって欠測値を処理する方がよいことが分かる。また、確定的回帰補定同様に、比率補定は乱数を用いない手法であるため、シード値を設定する必要がなく、毎回、同じ結果を得ることができる。繰り返し計算も行わず、回帰係数の算出も簡易的であるため、計算負荷は非常に低い。裏を返せば、確定的補定と同様に原データに存在していたばらつきを復元できず、比率補定による標準偏差(147.6)は真値(161.7)を過小推定している。それにともなって、比率補定による標準誤差(38.1)も過小なものとなっている(真値 = 41.8)。

比率補定は、そのモデル構造が切片のない単回帰モデルとして簡易的な構造をしているのが特徴である。これは、上述したとおり、実務家にとって使いやすいという長所があるが、諸刃の剣である。補助変数が 0 の値を取るとき、補定の対象としている変数が確実に 0 となることが既知であれば、比率補定の前提が満たされており、不偏推定量となるためよいモデルである。しかし、補助変数が 0 の値を取るとき、補定の対象としている変数が確実に 0 となることが既知ではない場合には、切片自体を推定するべきパラメータとして含めることができる回帰補定モデルの方が優れている可能性がある(de Waal *et al.*, 2011, p.245)[45]。例えば、年齢が 0 歳のとき収入は確かに 0 円だと考えることは妥当であっても、実際にそのようなデータが存在すること自体があり得ず、切片を強制的に 0 にすることが妥当かどうか、それ自体が検証の対象となり得る。大卒者の収入に限ったデータであれば、20 歳未満の人の収入はすべて 0 になるとも考えられる。そうであれば、理論上の切片の値は 0 ではない可能性がある。

また、データ内に利用できる補助変数が複数あったとしても、重回帰モデルとして複数の補助変数を同時に用いることができない。よって、1 つの変数によって MAR の前提を満た

[45] これは、一般的な回帰モデルにおける切片のあるモデルと切片のないモデルにおいても同様である(Wooldridge, 2009, p.59)。

すことができると確信できる場合のみ、ふさわしいモデルである(Hu *et al.*, 2001, p.10)。

本節で示した検証方法は、真値が既知であることを前提としている。今回の結果は、シミュレーションによるデモであるため、真値が分かっている。よって、真値と補定値との差を直接計算して、欠測による推定誤差を算出できた。しかし、現実には、真値は不明であり、比率補定による単一代入法では、欠測による推定誤差を評価することができない。

3.3.4 比率補定の R コード

下記のコードにて、使用する「データ名」を指定し、R に貼り付ければ、補定済みデータセット(ridata.csv)を作成することができる。なお、このコードでは、補定の対象となる欠測を含む変数がデータ内の 1 列目、補助変数がデータ内の 2 列目に格納されていることを前提としている。R の基本的な使用方法については、青木(2009, pp.1-10)を参照されたい。入出力データについては、本稿 3.1.4 節も参照されたい。

```
data<-read.csv("データ名.csv",header=T)          #「データ名」を指定
attach(data)
imp<-data[,1]
data2<-na.omit(data)                             #リストワイズ除去
Ratio<-mean(data2[,1])/mean(data2[,2])           #比率の推定
imp1<-Ratio*data[,2]                             #補定値の算出
for(i in 1:nrow(data)){                          #欠測値を補定値に置換
 if (is.na(imp[i])=="TRUE"){
 imp[i]<-imp1[i]
 }else{
 imp[i]<-imp[i]}}
ridata<-data.frame(data,imp)
write.csv(ridata,"ridata.csv")                   #データの出力
```

3.4 単一代入法の長所と短所のまとめ

本章で示したとおり、MAR の前提が正しいとき、確かに補定によって欠測値の偏りを是正できる。各手法の長所と短所は、表 3.7 に要約しているとおりである[46]。

表 3.7：補定手法の長所と短所一覧

	確定的回帰補定	確率的回帰補定	比率補定
平均値	○	△	○
標準偏差	×	○	×
標準誤差	×	△	×
複数の補助変数	○	○	×
シード値による影響	○	×	○
計算負荷	○	○	○
欠測による推定誤差	×	×	×

注：○=良い、△=場合による、×=悪い。なお、長所と短所は、前提条件が満たされた場合の理想状態を想定したものである。

[46] 本章の説明とともに、Baraldi and Enders (2010, pp.9-15)も合わせて参照されたい。

表 3.8 には、3 つの補定手法により算出した補定値を再掲している。ここから分かるとおり、使用する補定モデルに応じて、補定値は変化する。これは、補定が推定行為であるために、前提としたモデルに応じて推定値が変化するためである。すなわち、たとえ単一代入法を用いていたとしても、実際には補定値は 1 つの値に定まったわけではなく、1 つの補定値の背後には様々な可能性が隠れているのである。ここから、欠測値を補定するには、複数の値を算出する必要があると直感的に分かるであろう。

表 3.8：複数の補定手法による補定値の例

ID	収入	確定的回帰補定	確率的回帰補定	比率補定
1	543	*571*	*552*	*549*
2	272	272	272	272
3	797	797	797	797
4	239	239	239	239
5	415	*366*	*322*	*377*
6	371	371	371	371
7	650	*609*	*597*	*581*
8	495	495	495	495
9	553	553	553	553
10	710	*622*	*623*	*592*
11	421	421	421	421
12	410	*456*	*481*	*452*
13	386	386	386	386
14	280	280	280	280
15	514	514	514	514
平均	433	464	460	459

注：灰色セルは無回答により欠測していることを表し、白抜き数字は本来得られるはずの真値を表すものとする。イタリック（斜体）は補定値を表す。表 3.1、表 3.3、表 3.5 を統合したものである。

本章で示したとおり、欠測のメカニズムが MAR であれば、欠測による偏りを補定によって是正することができる。しかし、補定を行ったとしても、欠測値に対応する真値が完全に復元された訳ではなく不確実な部分が残っており、欠測値を補定したことによる誤差が存在していた。本章では、手法の説明のために真値の分かっているシミュレーションデータを用い、真値と補定値との差を直接計算して欠測値補定に起因する誤差の評価を行うことができた。一方、実際に欠測値補定を行う際には、欠測値に対応する真値は常に不明である。そもそも、真値が分かっているならば補定を行う必要がない。本章で示した単一代入法と呼ばれる手法では、欠測による推定誤差を評価することができない。そこで、本稿では、欠測値に対応する真値が分からないとしても、多重代入法を用いることで、このような欠測による誤差の評価を行えることを示す。

コラム 1：ベイズ統計学の基本的概念

多重代入法は、伝統的な頻度論に基づく統計学ではなく、ベイズ統計学の枠組みを用いて開発されてきた手法である。よって、ここで、ベイズ統計学の基礎知識について簡単に触れておく。

伝統的な統計学における頻度論的確率は、長期的に繰り返し行われる非決定論的な結果の性質として、極限における相対頻度である。例えば、コインを投げて表が出る確率が 1/2 というのは、無限の回数を試行した場合、表の出る確率が 1/2 になるということである。一方、ベイズ統計学における主観的確率は、信念の度合いとも呼ばれ、様々な状況下に応じて個人的に定義される。この場合、状況に応じて、コインを投げて表が出る確率は、必ずしも 1/2 と定義されるとは限らない。

新しく入手された情報に応じて確率を更新することこそが、ベイズ統計学の基本的なメカニズムであり、データから学んで信念を更新するプロセスを定式化している。そのために、ベイズ統計学では、「条件付け」という概念が重要な役割を果たす。もし事象 B が発生するかしないかによって事象 A の発生確率が影響を受けるとしたら、A は B を条件としていると言う。例えば、喫煙するかどうかによって肺に病気を患う確率が影響を受けることは容易に想像できる。この場合、喫煙（事象 B）によって、肺病（事象 A）の発生確率が影響を受けるので、肺病は喫煙を条件としていると言える。

ベイズの定理は、この条件付き確率から導出できる。ゆえに、ベイズ統計学は条件付き確率と密接な関係があるが、ベイズの定理と条件付き確率の違いは、事前分布という概念の存在にある。ベイズ統計学では、データを観察する前のパラメータ分布のことを事前分布と呼び、データを観測した後のパラメータの条件付き分布のことを事後分布と呼ぶ。一般的に、事前分布は人類の英知として蓄積された知識に基づいて追加情報を提供しており、事後分布は事前分布と尤度（ゆうど）の合算した情報源に基づいていると理解できる。なお、尤度とは、観測データが与えられたときの母集団パラメータの尤もらしさ（もっともらしさ）の度合いを表す（尤度については、4.1.2 項も参照されたい）。

すなわち、ベイズの定理とは、データに基づいて事前分布を事後分布に更新するメカニズムである。ベイズの定理における条件付き確率は、事前分布についてすでに観測したことや信じていることと、尤度として新たに観測した情報とのバランスを取っており、ベイズ統計学の分析は、伝統的な頻度論における尤度のみに基づく分析よりも多くの情報を利用していると考えることができる。なお、ベイズの定理は、得られた結果から原因を探る計算式と解釈することができる。

出典：高橋, 伊藤(2014, pp.75-78)

4　多重代入法のアルゴリズムとソフトウェア

前章で述べたとおり、補定によって偏りを是正できる。しかし、補定は、個別の値を完全に復元するわけではない。欠測値補定における目標は、個別の値を完全に復元することではなく、母集団パラメータの正確な推定を可能にすることである。そのためには、欠測値を補定したことによる誤差を適切に評価する必要がある。前章で説明した単一代入法ではこの目的を達成できず、多重代入法が解決策として提案されてきた(Rubin, 1987)。

多重代入法とは、欠測データの分布から複数の補定値を無作為に抽出するものである。しかし、欠測データは、定義上、観測されないため、欠測データの分布自体は不明である。そこで、観測されているデータから欠測値の事後分布を推定し、そこから無作為抽出を行う。なお、多重代入法は、「単一代入法を複数回実行したもの」と説明されることがあるが、正確ではない。確定的単一代入法を複数回実行しても、得られる結果はすべて同一である。また、確率的単一代入法を複数回実行した場合、確かに異なった値が算出されるが、本章で示すような回帰パラメータの推定を考慮に入れることができない（本稿 4.9 節参照）。

第 3 章で示した回帰モデルによる補定を行うには個別の値は必要なく、変数の平均値と分散共分散の情報が利用できれば回帰係数の算出ができる[47]。よって、観測データから欠測値の事後分布を推定し、そこから平均値と分散共分散の無作為抽出を行うことにより、回帰パラメータの推定を考慮に入れた補定を実行できる。これが多重代入法である。

本章では、4.1 節にて、期待値最大化法にブートストラップを応用した EMB[48]アルゴリズムによる多重代入法を解説する[49]。4.2 節では、EMB アルゴリズムを用いた多重代入法の数値例を示す。4.3 節において、基本統計量の比較によって多重代入法の長所と短所を示す。4.4 節は、多重代入法を実行するための R コードを与える。4.5 節では、MAR の前提の検証方法を提示する。4.6 節では、事前情報の設定によって、どのように多重代入法の結果を改善できるか示す。4.7 節にて、EMB 以外の多重代入法アルゴリズム（マルコフ連鎖モンテカルロ法と完全条件付指定）を簡潔に紹介する。4.8 節にて、統計センターにおけるこれまでの多重代入法の研究成果を示す。4.9 節にて、多重代入法と多重化した単一代入法との

---

[47] 二変量の場合について、簡潔に記す。母集団モデルは、$Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i$である。傾き$\beta_2$の最小二乗法による推定値は$\hat{\beta}_2 = \frac{\sum(X_i-\bar{X})(Y_i-\bar{Y})}{\sum(X_i-\bar{X})^2} = \frac{\sum(X_i-\bar{X})(Y_i-\bar{Y})/(n-1)}{\sum(X_i-\bar{X})^2/(n-1)} = \frac{\mathrm{cov}(X,Y)}{\mathrm{var}(X)}$として、$X$ の分散及び X と Y の共分散に基づいて算出することができる。また、切片β_1の最小二乗法による推定値は$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2\bar{X}$であり、$X$ と Y の平均値に基づいて算出することができる。すなわち、変数の平均値ベクトル$\boldsymbol{\mu}$と分散共分散行列$\boldsymbol{\Sigma}$の情報が利用できれば、回帰係数の算出ができる。よって、もし$\boldsymbol{\mu}$と$\boldsymbol{\Sigma}$が完全に既知ならば、Y_jに基づいて真の回帰係数βを決定的に算出することができる。この場合、完全データの尤度関数は$L(\boldsymbol{\mu},\boldsymbol{\Sigma}|\mathbf{D}) \propto \prod_{i=1}^{n} N(Y_i|\boldsymbol{\mu},\boldsymbol{\Sigma})$である。しかし、値が欠測している不完全データにおいては、$\boldsymbol{\mu}$と$\boldsymbol{\Sigma}$が既知ではなく、βの推定に関して確信を持つことができない。そこで、観測データ$\mathbf{D}_{\mathbf{obs}}$の尤度を形成する際に、MAR を想定する。$\mathbf{D}$の i 行の観測値を$\mathbf{D}_{\mathbf{i,obs}}$と定義し、$\boldsymbol{\mu}_{i,obs}$を$\boldsymbol{\mu}$のサブベクトルとし、$\boldsymbol{\Sigma}_{i,obs}$を$\boldsymbol{\Sigma}$のサブ行列とする。周辺分布は正規なので、観測データ$\mathbf{D}_{\mathbf{obs}}$の尤度関数は$L(\boldsymbol{\mu},\boldsymbol{\Sigma}|\mathbf{D}_{\mathbf{obs}}) \propto \prod_{i=1}^{n} N\left(\mathbf{D}_{\mathbf{i,obs}}|\boldsymbol{\mu}_{\mathbf{i,obs}},\boldsymbol{\Sigma}_{\mathbf{i,obs}}\right)$である。ここから無作為抽出を行うのである(高橋, 伊藤, 2013a, pp.38-40; 高橋, 伊藤, 2014, p.45)。

[48] Expectation-Maximization with Bootstrapping

[49] Rubin (1978, 1987)により提唱されたオリジナルの多重代入法は、4.7 節で紹介するマルコフ連鎖モンテカルロ法(MCMC)に基づくものである。MCMC による多重代入法については、高橋, 伊藤(2014, pp.46-50)及び野間, 田中, 田中, 和泉(2012, pp.108-111)を参照されたい。

違いについて例証する。4.10 節にて、多重代入法を用いる 8 つの利点をまとめた。

4.1 EMB アルゴリズムによる多重代入法のメカニズム

本稿では、R パッケージ Amelia を推奨している（4.8 節参照）。Amelia は、期待値最大化法とブートストラップを組み合わせた EMB アルゴリズムによる汎用的多重代入法プログラムである(Honaker and King, 2010; Honaker *et al.*, 2011)。本節では、始めに多重代入法の散布図と結果表を提示し、概念の導入を図った後、ブートストラップ法と期待値最大化法について詳説し、次節の EMB アルゴリズムによる多重代入法への導入とする。

図 4.1 は、収入と年齢の散布図だが、前章とは異なり、観測データをもとに EMB アルゴリズムにより算出した 3 本の回帰直線（補定モデル）が示されている[50]。つまり、$M = 3$ の多重代入法のモデルを示している。なお、M とは、多重代入済みデータの数のことを意味する。また、白丸は観測値であり、×印は欠測値を表す。図 4.2 は多重代入法による補定済みデータにおける散布図である。白丸は観測値であり、黒四角は補定 1 回目の補定値、黒丸は補定 2 回目の補定値、黒三角は補定 3 回目の補定値を表す。確率的回帰補定による図 3.3 と同じく、補定値は回帰線の上下にばらついていることが視覚的に分かる。

図 4.1：モデルと原データ

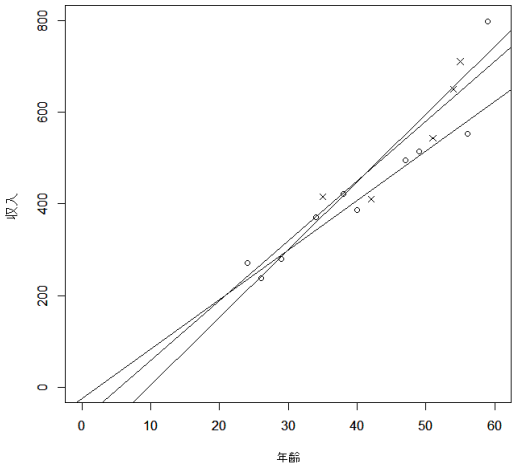
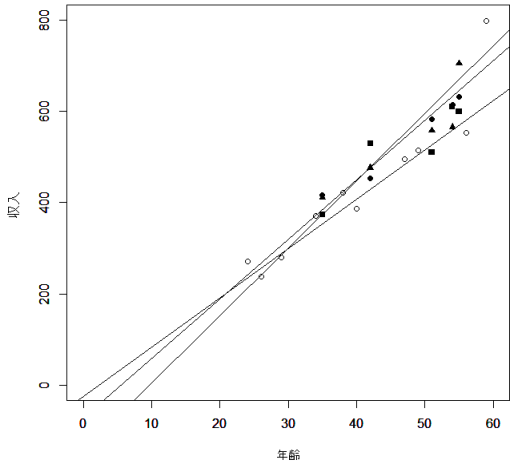


図 4.2：モデルと補定済みデータ



注 1：切片 1 = －28.0、傾き 1 = 10.9；切片 2 = －66.0、傾き 2 = 12.8；切片 3 = －144.7、傾き 3 = 14.9（回帰係数は観測データのみから算出し、図 4.1 と図 4.2 で共通）
注 2：○は観測値、×は欠測値、■、●、▲は、それぞれ 1～3 回目の補定値を表す。
注 3：実際に多重代入を行った図だが、説明のために線の重ならない 3 つのパターンを選んだ。

すなわち、多重代入法とは、回帰パラメータの推定に関するばらつきと個別の値のばらつきの両方を同時に考慮した手法なのである。回帰線は、先ほどの単一代入法による回帰補定の場合と同じく、最小二乗法による予測値である。しかし、確定的回帰補定や確率的回帰補

[50] Amelia には補定モデルの回帰線を算出する機能が搭載されていないため、図 4.1 と図 4.2 は EMB アルゴリズムによる多重代入法を自前でプログラムした結果である。2015 年現在、コードは非公開だが、デバック後、公開の予定である。

定と異なる点は、ブートストラップ法（4.1.1 項参照）が間に入っているため、補定 1～M までの複数の回帰係数が得られる点である。また、それらの係数は EM アルゴリズム（4.1.2 項参照）によって改善されている。回帰補定の場合と同じく、ガウス・マルコフの前提と MAR の前提条件がすべて満たされた場合、これらの予測値は不偏推定量である。さらに、多重代入法では、確率的回帰補定と同じく、観測されたデータをもとに回帰分析を行って予測値を算出するだけではなく、乱数に基づく誤差項も追加している[51]。

4.1.1 ブートストラップ法

ブートストラップ[52]とは、リサンプリング手法の一種である。リサンプリングは、手元にある標本データから再度、標本抽出を行うことを意味する[53]。このように標本データから得られた下位の標本データのことを再標本と呼ぶ。R パッケージ Amelia で用いられているブートストラップは、ノンパラメトリック[54]なブートストラップである(Honaker *et al.*, 2015)。ノンパラメトリック・ブートストラップでは、手元の標本データ(サイズ n)を擬似的に母集団と考え、そこから同サイズ(サイズ n)の再標本の復元抽出を M 回実行する(Shao and Tu, 1995; Horowitz, 2001)。復元抽出とは、重複を許す抽出のことである。本項では、ノンパラメトリック・ブートストラップの直感的なメカニズムについて、数値例を用いて説明する。技術的な詳細は、小西, 越智, 大森 (2008, pp.5-68)を参照されたい[55]。

表 4.1 を例に具体的に見ていく。原データは、第 3 章で用いたデータの一部である。簡単のため、欠測していない完全データを例にする。また、表 4.1 では、説明のため、意図的に ID の数を 6 つに限定している。したがって、サイコロを用意することでブートストラップを手作業で再現し、実感することができる。例えば、サイコロを振って 2 の目が出たら、原データの ID2 の情報（収入＝272、年齢＝24）を再標本 1 に記録する。再びサイコロを振って、再度 2 の目が出たとしても重複を許すので、原データの ID2 の情報（収入＝272、年齢＝24）を再標本 1 に記録する。再びサイコロを振って、3 の目が出たら、原データの ID3 の情報（収入＝797、年齢＝59）を再標本 1 に記録する。この作業を 6 回繰り返して得られたデータが再標本 1 となる。さらにこの作業を 3 回繰り返すことで、表 4.1 のようなデータセットが作成できる。

ただし、表 4.1 の再標本データは、ID 番号順に並べ替えなおしたものであり、実際の作業では ID 番号が無作為に選ばれるため、順番は上記のように整然としたものとはならない。しかし、変数のペア同士の値は一致している必要がある。つまり、収入の値と年齢の値を別

---

[51] $\tilde{Y}_i$は、$\tilde{Y}_i = \tilde{\beta}_0 + \tilde{\beta}_1 X_i + \tilde{\varepsilon}_i$より算出したシミュレーション値であり、~は適切な事後分布からの無作為抽出を表す。また、βは回帰係数、εは根本的（根源的）な不確実性を表す（高橋, 伊藤, 2014, p.44）。

[52] ブートストラップとは、ブーツ(boot)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。ここから、英語では「自力でやりとげる」という意味で使用され、コンピュータ用語では「外部情報なしで作動する自己完結したプログラム」のことを意味する。

[53] リサンプリングとサブサンプリングの違いについては、脚注 14 を参照されたい。

[54] 母集団の分布を仮定せずに分析を行う手法のことである(迫田, 高橋, 渡辺, 2014, p.185)。

[55] データ$Y_1, \ldots, Y_n$が独立にまた同一の分布 F に従っているとし、この分布を$\hat{F}(y) = \frac{1}{n}\sum_{i=1}^{n} I(Y_i \leq y)$で推定する。ここで、$I(Y)$は集合$Y$の指標関数である。この$\hat{F}(y)$に基づいて、ブートストラップ再標本を生成する。

個に無作為抽出してはならないということである。また、各々の再標本データの作成は、無作為なサイコロの目の結果に依存するので、実行するたびに異なる結果が得られることになる。実際の運用では、もちろん手作業でサイコロを転がすわけではなく、コンピュータ上において擬似乱数を用い、シード値を設定することで無作為抽出の結果を再現できる。

表 4.1：原データとブートストラップデータの例 1

原データ			再標本 1			再標本 2			再標本 3		
ID	収入	年齢	ID	収入	年齢	ID	収入	年齢	ID	収入	年齢
1	543	51	2	272	24	1	543	51	3	797	59
2	272	24	2	272	24	2	272	24	3	797	59
3	797	59	3	797	59	2	272	24	5	415	35
4	239	26	4	239	26	4	239	26	5	415	35
5	415	35	6	371	34	4	239	26	5	415	35
6	371	34	6	371	34	6	371	34	6	371	34

注：簡単のため、観測数を 6 に制限している。

表 4.2 を用いて、完全データにおける一変量のノンパラメトリック・ブートストラップを例示する。これはブートストラップを 1,000 回実行した例である（再標本 1～再標本 1000）。

表 4.2：原データとブートストラップデータの例 2

原データ		再標本 1		再標本 2		…	再標本 1000	
ID	収入	ID	収入	ID	収入	…	ID	収入
1	543	10	710	2	272	…	13	386
2	272	2	272	3	797	…	9	553
3	797	5	415	10	710	…	7	650
4	239	5	415	9	553	…	5	415
5	415	3	797	10	710	…	11	421
6	371	13	386	6	371	…	1	543
7	650	3	797	8	495	…	8	495
8	495	11	421	13	386	…	6	371
9	553	13	386	8	495	…	6	371
10	710	2	272	1	543	…	10	710
11	421	12	410	9	553	…	13	386
12	410	2	272	14	280	…	4	239
13	386	8	495	4	239	…	15	514
14	280	7	650	11	421	…	5	415
15	514	4	239	11	421	…	7	650
$\bar{y}$	470.4		462.5		471.7			468.9

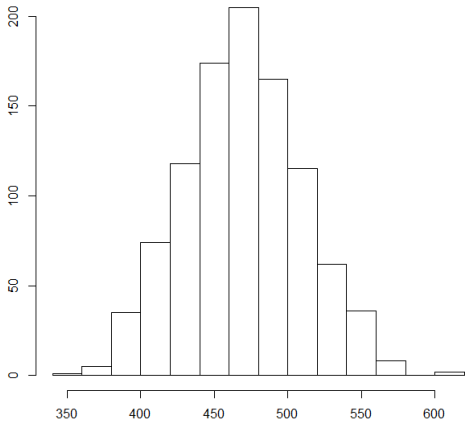
注：$\bar{y}$は収入の平均値を表す。

原データの標本平均は、3 章でも見たとおり 470.4 である。また、標準誤差は、標準偏差を標本サイズの平方根で割ったものであり、41.8 である。この 2 つの情報を用いることで標本誤差を評価することができる(熊原, 渡辺, 2012, pp.166-171)。一方、再標本 1 の収入の平均値は 462.5 であり、再標本 2 の収入の平均値は 471.7 であり、…、再標本 1000 の収入

の平均値は 468.9 である。これら 1,000 個の再標本平均値は、平均 469.8、標準偏差 40.2 で正規分布を近似している（図 4.3）。

1,000 個の再標本平均値の標準偏差(40.2)は、原データの標準誤差(41.8)とほぼ一致することが分かる。このように、ブートストラップを用いることで、標本平均値の経験分布を構築し、標本誤差の評価を行うことができるのである。

図 4.3：ブートストラップ再標本平均値の経験分布



一変量におけるノンパラメトリック・ブートストラップの R コードは、以下のとおり入力すれば実行できる。

```
data<-read.csv("データ名.csv",header=T)          #「データ名」を指定
attach(data)
set.seed(1223)                                   #シード値の設定
m<-1000                                          #ブートストラップ再標本数
databoot<-matrix(NA,nrow(data),m)
for(i in 1:m){                                   #ブートストラップ
 databoot[,i]<-sample(data[,1],replace=TRUE)}
 meanboot<-matrix(NA,1,m)
for(i in 1:m){                                   #再標本平均値の算出
 meanboot[1,i]<-mean(databoot[,i])}
mean(meanboot)                                   #平均値の統合
sd(meanboot)                                     #平均値の標準偏差
hist(meanboot)                                   #平均値の経験分布
```

表 4.3 は、欠測を含む不完全データの例である。表 4.1 と同様、サイコロを使ってブートストラップを手作業で再現できる。このようにして作成したブートストラップ再標本により、欠測値を補定する際に生じる推定誤差を評価するのである。

表 4.3：原データとブートストラップデータの例 3

原データ			再標本 1			再標本 2			再標本 3		
ID	収入	年齢	ID	収入	年齢	ID	収入	年齢	ID	収入	年齢
1	543	51	2	272	24	1	NA	51	3	797	59
2	272	24	2	272	24	2	272	24	3	797	59
3	797	59	3	797	59	2	272	24	5	NA	35
4	239	26	4	239	26	4	239	26	5	NA	35
5	415	35	6	371	34	4	239	26	5	NA	35
6	371	34	6	371	34	6	371	34	6	371	34

注：灰色セルは無回答により欠測していることを表し、白抜き数字は本来得られるはずの真値を表すものとする。灰色セルの白抜き NA は欠測を表す。簡単のため、観測数を 6 に制限している。

4.1.2 EM アルゴリズム

我々の目的は、第 3 章で実行したように、回帰モデルを用いて欠測値を補定することである。上述したとおり、回帰モデルを構築するには、平均値、分散、共分散の情報が必要である。平均値は、これまで見てきたとおり、収入や年齢の平均値のことである。分散とは、収入や年齢の値が、それぞれ個別に、どれだけばらついているかを表す指標であり、標準偏差の二乗である。共分散とは、収入の値と年齢の値がどれだけ一緒に動いているか、つまり、どれだけ線形の関係があるかを表す指標である。なお、平均値、分散、共分散などのことをパラメータと呼ぶ。

表 4.3 の例では、再標本 1 のデータには欠測値が含まれていない。この場合、得られた再標本データをそのまま利用して、収入を補定対象変数とし、年齢を補助変数として、回帰モデルを構築し、得られた回帰係数を使って補定をすればよい。一方、再標本 2 と再標本 3 には欠測値が含まれている。この場合、再標本データから得られる平均値、分散、共分散の値には欠測による偏りが反映されてしまう。したがって、不完全データを完全データにするためには、平均値、分散、共分散といった分布に関する情報（パラメータ）が必要となるが、これらの情報を推定するために不完全データを使うという鶏と卵の問題が発生する。そこで、EM[56]アルゴリズムによる繰り返し手法によって推定を行うことが推奨される。

EM アルゴリズムを理解するには、まず、最尤（さいゆう）推定法[57]について知っておくことが前提となる。最尤推定値とは、実際に観測された標本データを観測する確からしさ（尤度）が最大となるパラメータ推定値(Long, 1997, p.26)であり、漸近（ぜんきん）正規性、不変性、一致性、漸近効率性というナイスな特性を持つ[58]。正規性とは、標本抽出を繰り返した場合に、推定値が正規分布となることを意味する。不変性とは、母数θの最尤推定量$\hat{\theta}$があり、この母数θの関数$g(\theta)$がある場合、最尤推定量$\hat{\theta}$の関数$g(\hat{\theta})$は$g(\theta)$の最尤推定量であることを意味する。一致性とは、推定値が大標本において近似的に不偏であることを意味する。効率性とは、標準誤差が、他のいかなる一致推定量によるものと同等かそれ以下で

[56] Expectation-Maximization
[57] Maximum Likelihood Estimation (MLE)
[58] 英語では、Normality（正規性）、Invariance（不変性）、Consistency（一致性）、Efficiency（効率性）の頭文字をとり、「すばらしい」という意味と掛けて NICE な特性を持つと言う。

あることを意味する。漸近性は、有限な標本サイズにおいては近似的な特性であるが、標本サイズが大きくなるにつれて改善することを意味する(Allison, 2002, pp.13-14; Greene, 2003, pp.472-483)。

日本人男性の身長を例に考えてみる。真のデータは、平均μcm、標準偏差 5.8 で正規分布していると仮定しよう。つまり、正規分布の母集団パラメータのうち平均値が不明だとする。そして、5 人の男性の身長の標本データ(166.4, 171.1, 165.2, 179.3, 171.9)があるとする。すなわち、この場合の最尤推定とは、真のデータが平均μcm、標準偏差 5.8 で正規分布[59]しており、標本データ(166.4, 171.1, 165.2, 179.3, 171.9)が得られた場合、μがどの値であれば、この標本データを手に入れる可能性が最も高くなるかという問題である[60]。最尤法を視覚化するために、図 4.4～図 4.9 に示すように、$\hat{\mu}$を 3 つの値に変化させた場合の分布を書いてみると分かりやすい。ただし、図 4.4～図 4.9 は最尤法の直感的なイメージであって、厳密な理論は脚注 60 を参照されたい。

図 4.4 では標本データの 5 つすべてが分布の右半分に位置しており、図 4.5 では標本平均も分布の右すそに位置している。図 4.6 では標本データの 5 つの値すべてが分布の中央付近にまんべんなく位置しており、図 4.7 では標本平均は分布のほぼ中心に位置している。図 4.8 では標本データの 5 つとも分布の左半分に偏っており、図 4.9 では標本平均は分布の左すそに位置している。したがって、これら 3 つの分布から考えた場合、今回の標本データは、図 4.6（図 4.7）の分布から生成された可能性が最も高そうである[61]。つまり、μは 170 と推定される。これが最尤推定法の直感的な概念であり、実際に一変量の最尤推定値は標本平均そのもの、つまり、170.78cm である。また、元々のシミュレーションデータは、平均 170cm、標準偏差 5.8 の正規分布から生成されたものである。

このように、最尤推定量は好ましい特性を持つわけだが、欠測データにそのまま適用することは困難である。そこで、繰り返し手法を用いて最尤推定量を算出する方法として EM アルゴリズムが提唱されてきた(Allison, 2002, pp.17-20)。

---

[59] 今回の例では、分布は正規と仮定しているが、最尤推定における分布は正規に限られるものではない。

[60] 一般論として、もし y が標準偏差 1 の正規分布に従うとすれば、y の確率密度関数は、$f(y_i|\mu,\sigma=1)=\frac{1}{\sqrt{2\pi}}exp(\frac{-(y_i-\mu)^2}{2})$である。ここでは、$\mu$が不明なので、尤度関数は$L(\mu|y_i,\sigma=1)=f(y_i|\mu,\sigma=1)$である。$n$個の独立した観測値に関する尤度は、各々の尤度の積、すなわち、$L(\mu|\boldsymbol{y},\sigma=1)=\prod_{i=1}^{n}L(\mu|y_i,\sigma=1)=\prod_{i=1}^{n}f(y_i|\mu,\sigma=1)$である。対数尤度は、$\ln L(\mu|\boldsymbol{y},\sigma=1)=\sum_{i=1}^{n}lnL(\mu|y_i,\sigma=1)=\sum_{i=1}^{n}lnf(y_i|\mu,\sigma=1)$である(Long, 1997, pp.27-28)。$L(\mu|y_i,\sigma=1)$は、$(2\pi)^{-\frac{n}{2}}\exp\left\{-\frac{n}{2}\left(\mu-\frac{1}{n}\sum y_i\right)^2-\frac{1}{n}(\sum y_i)^2+\sum y_i^2\right\}$であるから、$\exp\left\{-\frac{n}{2}\left(\mu-\frac{1}{n}\sum y_i\right)^2\right\}$に比例する（定数倍）。最尤推定値は、この式を最大化する$\hat{\mu}=\bar{y}$（標本平均）である。実際に、横軸μに関するグラフは標本平均値で対称であり、標本平均値においてピーク（最大値）となる。

[61] 実際には、無数のケースを確認する。

平均 $\hat{\mu}$ = 160cm、標準偏差 5.8 の正規分布

図 4.4：標本データの値

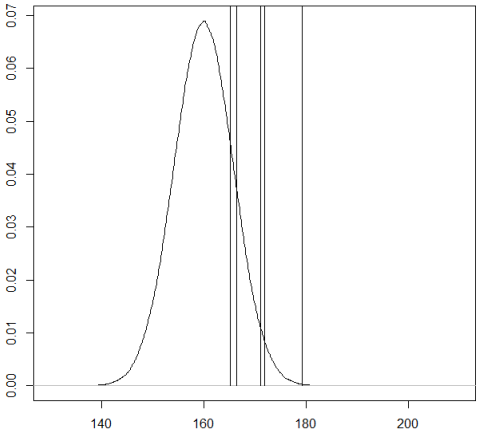
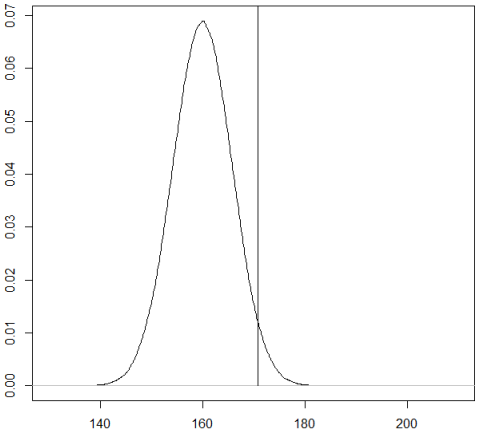


図 4.5：標本平均



平均 $\hat{\mu}$ = 170cm、標準偏差 5.8 の正規分布

図 4.6：標本データの値

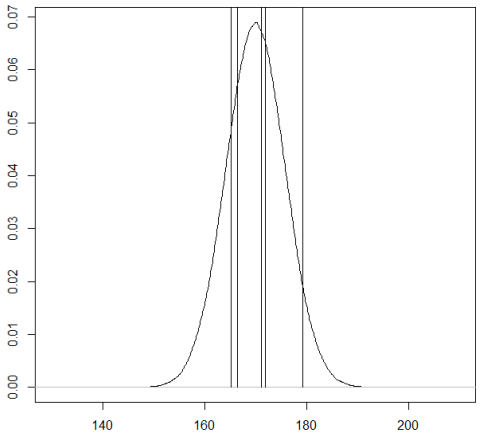
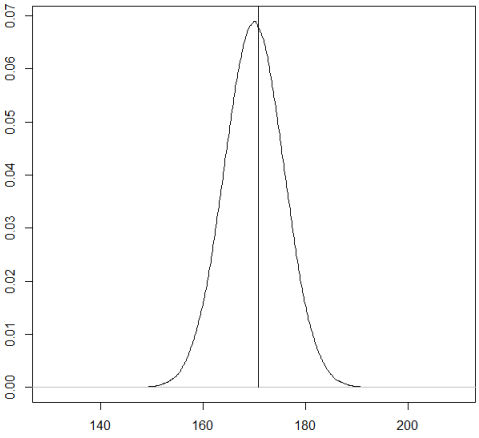


図 4.7：標本平均



平均 $\hat{\mu}$ = 180cm、標準偏差 5.8 の正規分布

図 4.8：標本データの値

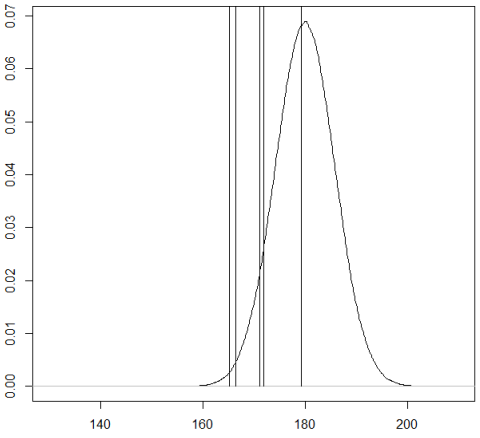
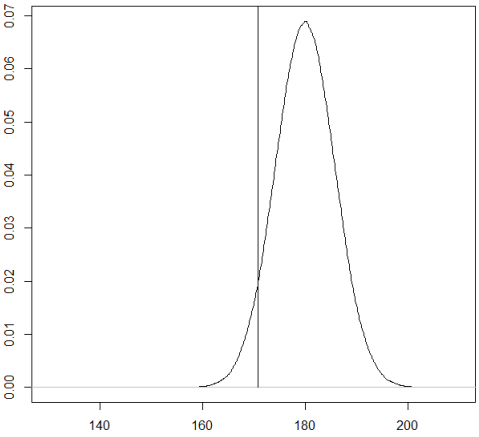


図 4.9：標本平均



EM アルゴリズムとは、Expectation-Maximization の略で、日本語では期待値最大化法と呼ばれ、期待値ステップと最大化ステップの 2 つのステップから成り立つ。平均値、分散、共分散といったパラメータの初期値を適当に選び、期待値ステップにおいて観測データとパラメータ推定値を条件として完全データの十分統計量[62]の期待値を推定する。また、最大化ステップでは、推定された完全データの十分統計量の値をもとにパラメータ推定値を更新する。そして、アルゴリズムを繰り返し実行し、更新前後のパラメータ推定値の差がなくなったと判断された段階で収束したものとする(Hu *et al.*, 2001, p.11)。つまり、EM アルゴリズムでは、期待値ステップにおいて、モデルパラメータを条件として欠測データの確率分布を推測し、最大化ステップでこれらの情報を用いてモデルパラメータを更新する[63]。よって、EM アルゴリズムとは、データ内の欠測値を対数尤度における欠測部分の期待値に置き換え、尤度の最大化を行い、これらの期待値ステップと最大化ステップの繰り返し計算を通して、最尤推定値を求めるものである(渡辺, 山口, 2000, p.34)[64]。技術的な詳細は、小西, 越智, 大森(2008, pp.71-141)を参照されたい。以下では、EM アルゴリズムの直感的なメカニズムについて数値例を使って説明する。岩崎(2002, pp.288-290)の例も合わせて参照されたい。

先ほどと同じく、真のデータは平均 170cm、標準偏差 5.8 で正規分布していると仮定しよう。また、母集団パラメータのうち平均値が不明だとする。取り消し線の引いた値が欠測しているとし、5 人の男性の身長の標本データ(166.4, 171.1, 165.2, 179.3, ~~171.9~~)があるとしよう。つまり、先ほどとは異なり、最後の 1 人の値が欠測している。よって、手元に利用できる標本データは、4 人の男性の身長の標本データ(166.4, 171.1, 165.2, 179.3)である。

実際の計算は以下のとおり行われる。期待値ステップにおいて、観測データとパラメータの初期値を条件とし、パラメータの条件付き期待値を算出する。観測データとは、上記の標本データ(166.4, 171.1, 165.2, 179.3)のことであり、パラメータの初期値は適当に選んだ値である。ここでは 150 としよう。まず、観測データの値を合計する。次に、完全データの標本サイズと不完全データの標本サイズの差を計算し、初期値に乗じる。この 2 つを足し合わせた値が期待値ステップで求める値である。すなわち、(166.4 + 171.1 + 165.2 + 179.3) + (5－4) * 150 = 832 である[65]。最大化ステップにおいて、期待値ステップで計算したパラメータの条件付き期待値を用いて、パラメータを更新する。上記の 832 を完全データの標

---

[62] 十分統計量とは、個別のデータの値が分からなくても、母集団パラメータの推定を十分に行うことができる統計量を意味する。詳しくは、岩崎(2002, p.74)を参照されたい。

[63] 期待値ステップの名前は、確率分布全体を構築する必要はなく、十分統計量の期待値を算出すればよいことに由来する。同じく、最大化ステップの名前は、モデルの更新はデータの対数尤度の期待値を最大化することに由来する(Do and Batzoglou, 2008, pp.897-898)。

[64] 初期値θ_0から始め、期待値ステップにおいて、$Q(\theta|\theta_t) = \int l(\theta|Y)\, P(Y_{mis}|Y_{obs};\theta_t)dY_{mis}$とする。なお、ここで $l(\theta|Y)$は対数尤度である。最大化ステップにおいて、$\theta_{t+1} = \arg\max_{\theta} Q(\theta|\theta_t)$ をθに関して最大化する。そして、これら 2 つのステップを収束するまで繰り返す(高橋, 伊藤, 2014, p.55)。

[65] 数式では、$E[m_1|\theta_t, Y_{obs}] = \sum_{i=1}^{m} y_i + (n-m)\mu_t$となる。岩崎(2002, p.288)も参照されたい。

本サイズ 5 で割ることにより求められる。すなわち、166.4 である[66]。

再び、期待値ステップに戻る。先ほどは初期値として適当に選んだ 150 という値を用いた。今度は、上記の最大化ステップで得られた 166.4 を用いる。すなわち、計算結果は(166.4 + 171.1 + 165.2 + 179.3) + (5－4) * 166.4 = 848.4 となる。さらに、再度、最大化ステップにおいて、上記の 848.4 を完全データの標本サイズ 5 で割る。すなわち、169.68 である。

この作業を収束するまで実行する。収束とは、最大化ステップで得られた推定値が変化しなくなる状態のことである。上述したとおり、このようにして収束した値は最尤推定値となることが知られている。一変量の場合は、単なるリストワイズ除去による平均値と同一となり、面白みに欠けるが、共変量を用いることで推定値を大幅に改善することができる[67]。一変量における EM アルゴリズムの R コードは、以下のとおり入力すれば実行できる。

```
x1<-c(166.4, 171.1, 165.2, 179.3, 171.9)              #完全データ
x2<-c(166.4, 171.1, 165.2, 179.3)                     #不完全データ
mu<-150                                               #初期値
expect<-NA
for(i in 1:100){
 expect[i]<-sum(x2)+(length(x1)-length(x2))*mu[i]     #E ステップ
 mu[i+1]<-1/length(x1)*expect[i]                      #M ステップ
 if(mu[i+1]-mu[i]<0.0001){                            #収束判定
 break}}
mu[length(mu)]                                        #最尤推定値
```

このように、EM アルゴリズムでは、観測情報を条件として欠測データの分布を推定し、繰り返し計算によって平均値、分散、共分散といった情報を改善するのである。なお、多変量についての直感的な解説は、Allison (2002, pp.19-20)を参照されたい。

4.2 EMB アルゴリズムによる多重代入法の例

表 4.4 は、EMB アルゴリズムを搭載した R パッケージ Amelia による多重代入法の結果である。簡単のため、多重代入済みデータ数（M 数）は 3 個に制限している。表 4.4 をよく見ると、ID1 の補定値は、「518, 604, 530」となっており、補定を行うたびに異なった値を算出していることが分かる。補定では、年齢から収入の値を推測しており、事実が判明したわけではない。これら補定値のばらつきの大きさは、未知の欠測値に対する補定モデルの精度を示している。

[66] 数式では、$\mu_{t+1} = \frac{1}{n}[\sum_{i=1}^{m} y_i + (n-m)\mu_t]$となる。岩崎(2002, p.289)も参照されたい。

[67] 二変量の場合の EM アルゴリズムは、回帰モデルによる予測と一致する(Allison, 2002, pp.19-20)。

表 4.4：多重代入法による補定済みデータの例($M = 3$)

ID	収入	年齢	補定値 1	補定値 2	補定値 3
1	543	51	*518*	*604*	*530*
2	272	24	272	272	272
3	797	59	797	797	797
4	239	26	239	239	239
5	415	35	*340*	*350*	*389*
6	371	34	371	371	371
7	650	54	*536*	*550*	*584*
8	495	47	495	495	495
9	553	56	553	553	553
10	710	55	*520*	*650*	*597*
11	421	38	421	421	421
12	410	42	*432*	*465*	*434*
13	386	40	386	386	386
14	280	29	280	280	280
15	514	49	514	514	514
$\bar{y}$	433	43	445	463	458

補定間のばらつき

補定間の標準偏差 9.34

注：灰色セルは無回答により欠測していることを表し、白抜き数字は本来得られるはずの真値を表すものとする。イタリック（斜体）は補定値を表す。$\bar{y}$は収入の平均値を表す。

収入の平均値も 445 万円、463 万円、458 万円として変動しており、これら 3 つの値の標準偏差を計算すると、9.34 となる。これを補定間の標準偏差(BSD)[68]と呼び、欠測値を推定した結果、収入の平均値がどれだけばらついているかを表す指標となる(Honaker *et al.*, 2011, p.23)。補定モデルの精度が高い場合、補定 1～M までの補定値は、ほぼ同じ値が算出され、補定間の標準偏差も小さくなる。一方、補定モデルの精度が低い場合、補定 1～M までの補定値は非常に異なった値が算出され、補定間の標準偏差も大きくなる。

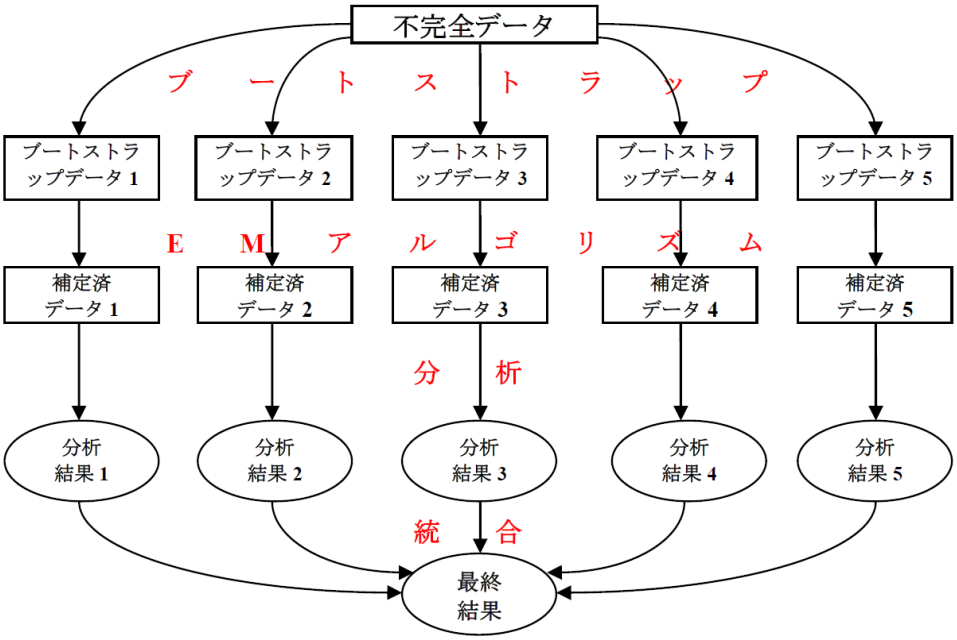
また、補定 1 データにおける収入の標準偏差は 142.0、補定 2 データにおける収入の標準偏差は 154.6、補定 3 データにおける収入の標準偏差は 147.1 である。これを補定内標準偏差(WSD)[69]と呼び、収入という変数自体にどれだけのばらつきがあるかを表す。確率的補定が捉えることができているのは、この部分だけである。

R パッケージ 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 4.10 のとおりである(Honaker *et al.*, 2011, p.4)。まず、4.1.1 節で説明したノンパラメトリック・ブートストラップにより、5 つの再標本を生成する。それぞれのブートストラップ再標本に、4.1.2 節で説明した EM アルゴリズムを適用し、平均値、分散、共分散の値を改善して回帰モデルを構築し、欠測値のシミュレーション値を算出する。さらに、それぞれのシミュレーション値に、確率的補定の場合と同じように誤差項を付したものを補定値とする。その結果、補定済みデータセットが 5 個作成される。統計分析において、それぞれのデータセットを別々に使用して、しかるべき手法により統合して最終結果と

[68] Between-imputation Standard Deviation
[69] Within-imputation Standard Deviation

する[70]。

図 4.10：EMB アルゴリズムによる多重代入法の模式図($M = 5$)



なお、ブートストラップ再標本に EM アルゴリズムを適用して得られた最尤推定値は、ベイズ統計における事後分布からの無作為抽出による推定値と漸近的に等価であり、Rubin (1987, pp.118-119)の言う適切な補定[71]と見なすことができる(Little and Rubin, 2002, pp.216-217)。

4.3 多重代入法の長所と短所

表 4.5 は、上記の手法により算出した基本統計量(多重)である。なお、確率的補定の場合と同様に、シード値による影響を示すため、3 つの任意のシード値による結果を掲載している(多重 1, 多重 2, 多重 3)。

比較対象として、真値、リストワイズ除去による値(LW)、確定的回帰補定による値(確定)、確率的回帰補定による値(確率)、比率補定による値(比率)も掲載している。表 4.5 の多重代入法による結果は、1,000 個の多重代入済みデータによる結果を統合したものである。平均値と標準偏差の統合は、Rubin (1987, p.76)に示されているとおり、M 個の統計量の算術平均により算出すればよい (Marshall *et al.*, 2009)。本章の説明とともに、Baraldi and Enders

[70] 多重代入法により生成した M 個の補定済データセットを別々に使用して、t 検定や回帰分析などの統計分析を行い、以下のとおり推定値を統合し、点推定値を算出する。$\hat{\theta}_m$をパラメータθの m 番目の補定済データセットに基づいた推定値とする。統合した点推定値$\bar{\theta}_M$は、$\bar{\theta}_M = \frac{1}{M}\sum_{m=1}^{M}\hat{\theta}_m$である。$\bar{\theta}_M$の分散$T_M$は、$T_M = \bar{W}_M + \left(1+\frac{1}{M}\right)\bar{B}_M = \frac{1}{M}\sum_{m=1}^{M} var(\hat{\theta}_m) + \left(1+\frac{1}{M}\right)\left[\frac{1}{M-1}\sum_{m=1}^{M}(\hat{\theta}_m - \bar{\theta}_M)^2\right]$のとおりである。なお、$\bar{W}_M$を補定内分散の平均とし、$\bar{B}_M$を補定間分散の平均とする。つまり、$\bar{\theta}_M$の分散は、補定内分散$\bar{W}_M$と補定間分散$\bar{B}_M$考慮に入れたものである（高橋, 伊藤, 2014, p.44）。

[71] Proper imputation

(2010, pp.15-18)も合わせて参照されたい。

表 4.5：多重代入済みデータの統計量

	真値	LW	確定	確率 1	確率 2	確率 3	比率	多重 1	多重 2	多重 3
平均値	470.4	432.8	463.5	460.2	463.3	465.8	458.6	462.1	461.6	461.9
標準偏差	161.7	166.8	152.9	153.8	155.0	153.1	147.6	-	-	-
標準誤差	41.8	52.8	39.5	39.7	40.0	39.5	38.1	-	-	-
WSD	-	-	-	-	-	-	-	155.2	154.3	154.8
BSD	-	-	-	-	-	-	-	12.3	11.8	11.9
TSE	-	-	-	-	-	-	-	40.2	40.0	40.1
CI UL	-	-	-	-	-	-	-	486.8	485.3	485.8
CI LL	-	-	-	-	-	-	-	437.4	438.0	438.0
観測数	15	10	15	15	15	15	15	15	15	15

注 1：- は該当値がないことを表す。
注 2：LW はリストワイズ除去(List-Wise Deletion)により欠測を除去した値、確定は確定的回帰補定による値、確率は確率的回帰補定による値、比率は比率補定による値、多重は多重代入法による値を表す。
注 3：WSD は補定内における標準偏差、BSD は補定間における標準偏差、TSE は全体の標準誤差である。
注 4：CI は欠測に起因する誤差に関する信頼区間（95%水準）であり、UL は上限、LL は下限を表す。
注 5：確率的回帰補定と多重代入法については、シード値を 3 回変えて実行した。（ただし、確率的補定と多重代入法では同一のシード値を使用した）。多重代入法では、各々のシード値に対して、$M = 1000$ を実行した。

多重代入法による平均値(462.1, 461.6, 461.9)は、確定的補定の場合と同じく前提が満たされた場合には不偏推定量である。実際に補定を行うことで、リストワイズによる値(432.8)と比べて、真値(470.4)の復元へと近づいている。今回は小標本であるため、確定的補定の値(463.5)と乖離があるが、大標本において前提がすべて満たされている場合、$M = \infty$の多重代入による結果は、確定的補定の結果と一致する特性がある[72]。

確率的補定と同様に誤差項を追加したことにより、原データに存在していたばらつきを復元し、確定的補定による標準偏差(152.9)と比較して、多重代入法による標準偏差(WSD)の推定結果(155.2, 154.3, 154.8)は改善している(真値 = 161.7)。また、補定内標準偏差(WSD)と補定間標準偏差(BSD)を同時に考慮に入れたことで、標準誤差(TSE)の推定値(40.2, 40.0, 40.1)も改善している(真値 = 41.8)。確定的補定や確率的補定の場合と同じく、データ内に利用できる補助変数が複数あれば、重回帰モデルとして複数の補助変数を用いて、補定の精度を向上させることができる。

多重代入法を用いる最大の利点は、欠測に起因する誤差の評価を行える点である。BSD（補定間の標準偏差）は、1,000 個の平均値がどれだけばらついているかを表している。例えば、多重 1 の結果では、平均値の点推定値が 462.1 であると分かっただけではなく、95%水準で±24.6 (= 12.3×2)の欠測に起因する誤差があることが示されている。95%信頼区間で表すと、C.I. (437.4, 486.8)である。他の補定手法では標準誤差を用いることで標本誤差がどれだけあるかを示すことはできるものの、真値が分からない限り、欠測に起因する非標

[72] 、モデルの前提が正しく MAR の前提も正しい場合、$M = \infty$の多重代入法と単一代入法の結果はいずれも不偏であり、点推定値の結果に関してほぼ同一の結果となる(Donders *et al.*, 2006, p.1089)。

本誤差の評価を行うことができない。多重代入法では、補定間の標準偏差(BSD)を確認することで、集計結果に対する欠測値の影響を数値で評価することができる[73]。Amelia には補定モデルの回帰線を算出する機能が搭載されていないため、図 4.11 は EMB アルゴリズムによる多重代入法を自前でプログラムし $M = 1000$ の多重代入モデルを図示したものである。2015 年現在、コードは非公開だが、デバック後、公開の予定である。ここから、回帰モデルのパラメータ推定値は、欠測の影響を大きく受けていることが分かる。

図 4.11：多重代入モデル($M = 1000$)

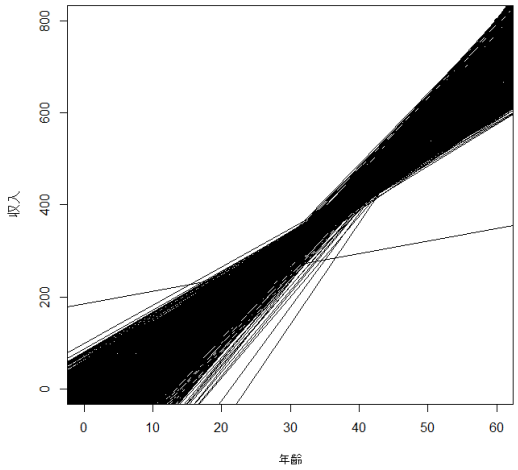


表 4.6 は、多重代入モデルの回帰係数の基本統計量である。3.1.1 項における確定的回帰補定における係数は、切片 = －80.8、傾き = 12.8 であった。多重代入法における係数は、平均して、切片 = －75.2、傾き = 12.6 であった。上述したとおり、大標本において前提がすべて満たされている場合、$M = \infty$の多重代入による結果は、確定的補定の結果と一致する特性があり、今回の結果も、ほぼ一致していることが分かる。一方、切片も、傾きも、ばらつきが大きく、欠測に起因する誤差が非常に大きいことが分かる。このばらつきの大きさは、表 4.5 の補定間標準偏差(BSD)に反映されている。

表 4.6：多重代入モデルの回帰係数

	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
切片	－517.5	－125.0	－75.6	－75.2	－11.1	185.9	80.9
傾き	2.7	10.5	12.7	12.6	14.1	22.0	2.2

[73] 経済データのように、元のデータが正規分布ではなくても、ブートストラップ再標本に基づいてパラメータ推定値の標本分布を経験的に近似することができるため、上記の手法に則って平均値に関する欠測に起因する誤差の評価を行うことができると考えられる。

多重代入法は、確率的補定と同様に乱数を用いた手法であるため、毎回、異なった結果が算出される。もちろん、シード値を設定すれば、同一のシード値のもとで再現性を確保できる。補定値の精度は、どのシード値を選んだかに依存する可能性があり、どのシード値による結果が良いかは、事前には分からない。しかし、確率的補定による結果がシード値ごとに大幅に変化するのに対して、多重代入法による結果はすでに多数の結果を統合しているため、シード値を変化させても、比較的安定した結果を得ることができる。図 4.12 は 1,000 個のシード値を用いた多重代入法による平均値の分布であり、図 4.13 は 1,000 個のシード値を用いた確率的単一代入法による平均値の分布である。ここから、シードを変えることで、確率的単一代入法による結果は大きく変化するが、多重代入法による結果にはほとんど影響が出ないことが分かる。

図 4.12：多重代入法による平均値

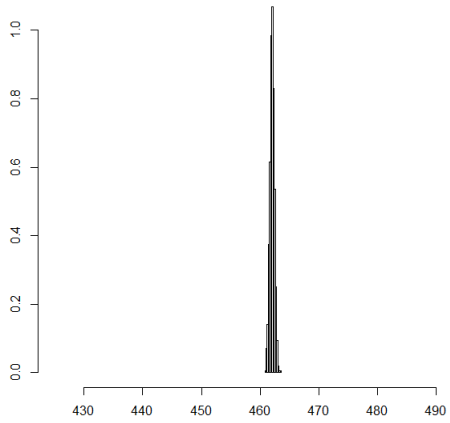
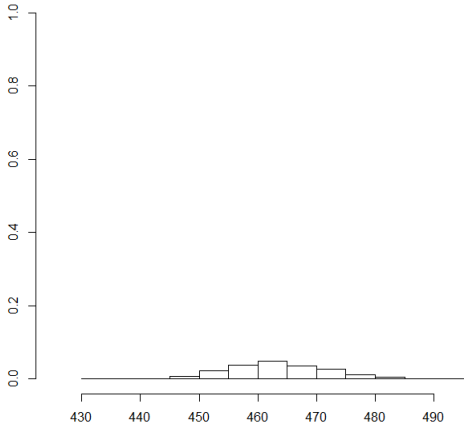


図 4.13：確率的単一代入法による平均値



多重代入法の最大の短所は、計算負荷の大きさである。乱数を発生させるだけではなく、複数のデータセットを同時に保持しながら分析を行うため、大規模なデータの補定を行う場合には、コンピュータ上に十分な容量を確保する必要がある。また、コンピュータの仕様によっては、分析に時間がかかるおそれもある。ただし、この点についても、かつては実用上の問題となることがあったが、2015 年現在における家庭用 PC のレベル（コラム 2 参照）であれば、100 万単位の観測数のデータを問題なく多重代入することができる。また、多重代入済みデータ数（M数）については、付録 2 も参照されたい。

4.4 多重代入法の R コード

下記のコードにて、使用する「データ名」と補定済みデータの「数」を指定し、R に貼り付ければ、補定済みデータセット(midata.csv)を作成することができる（Amelia を組み込んだ R コードについては、本稿の付録 3 も参照されたい）。なお、このコードでは、補助変数の数はいくつあってもよいが、補定の対象となる欠測を含む変数は、データ内の 1 列目に格納されていることを前提としている。つまり、入力データは図 3.3 と同一である。その

他、R パッケージ Amelia の使用方法について、詳細は、Honaker *et al.* (2011)、Honaker *et al.* (2015)、高橋, 伊藤(2013a, pp47-49, 82-83)を参照されたい。また、Amelia を使用するには、事前に CRAN からパッケージのインストールを行う必要がある。R におけるパッケージのインストールについては青木(2009, pp.7-8)を、R の基本的な使用方法については青木(2009, pp.1-70, pp.279-307)を参照されたい。

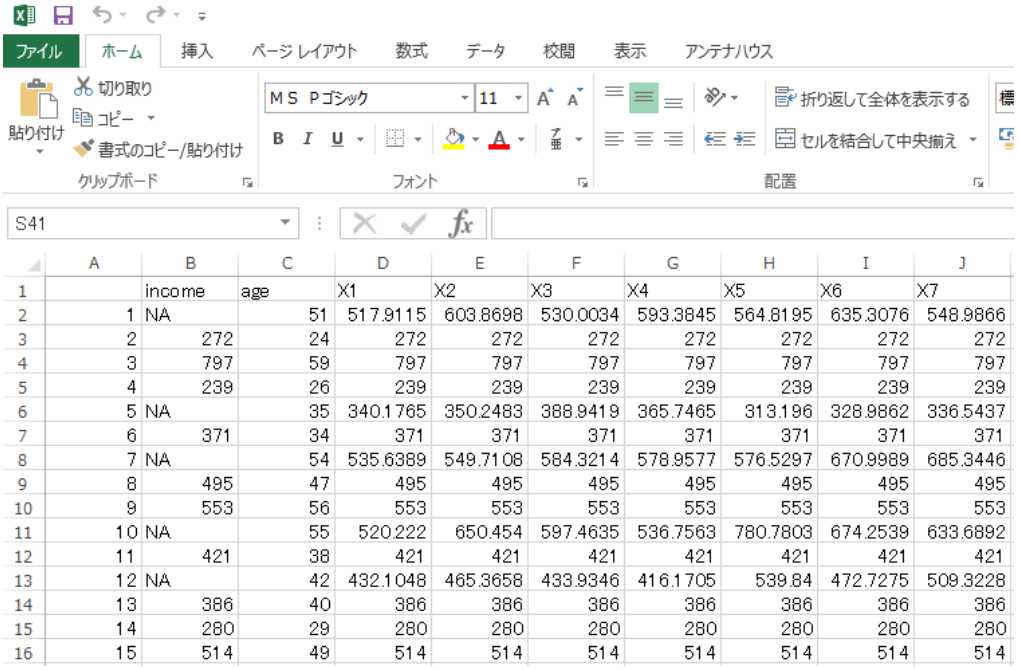
```
data<-read.csv("データ名.csv",header=T)  #「データ名」を指定
attach(data)
m<-数                                    #補定済みデータの「数」を指定
library(Amelia)                          #Amelia を起動
mat<-matrix(NA,nrow(data),m)
set.seed(1223)                           #シード値を指定
a.out<-amelia(data,m=m)                  #多重代入法を実行
for(i in 1:m){                           #補定値の格納
 ximp<-a.out$imputations[i]
 ximp<-data.frame(ximp)
 ximp2<-ximp[1]
  for(ii in 1:nrow(data)){
   mat[ii,i]<-ximp2[ii,]}}
midata<-data.frame(data,mat)             #データの出力
write.csv(midata,"midata.csv")
```

経済系データのように、分布が対数正規である場合には、下記のように多重代入モデルにおいて、自然対数変換を実行することにより対応できる。

```
a.out<-amelia(data,m=m,                  #多重代入法を実行
 logs=c(“変数名”,”変数名”))               #対数変換する「変数名」を指定
```

図 4.14 は出力データである。A 列は ID 番号、B 列は補定対象の欠測変数、C 列は補助変数である。複数の補助変数を用いた場合、D 列以降に順次記載される。X1 は補定 1 回目の補定値、X2 は補定 2 回目の補定値である。図 4.14 では紙面の都合上、X7 までしか掲載していないが、M の数だけ Xn が記載される。今回の場合、一番右の変数は X1000 である。

図 4.14：出力データの例



	A	B	C	D	E	F	G	H	I	J
1		income	age	X1	X2	X3	X4	X5	X6	X7
2	1	NA	51	517.9115	603.8698	530.0034	593.3845	564.8195	635.3076	548.9866
3	2	272	24	272	272	272	272	272	272	272
4	3	797	59	797	797	797	797	797	797	797
5	4	239	26	239	239	239	239	239	239	239
6	5	NA	35	340.1765	350.2483	388.9419	365.7465	313.196	328.9862	336.5437
7	6	371	34	371	371	371	371	371	371	371
8	7	NA	54	535.6389	549.7108	584.3214	578.9577	576.5297	670.9989	685.3446
9	8	495	47	495	495	495	495	495	495	495
10	9	553	56	553	553	553	553	553	553	553
11	10	NA	55	520.222	650.454	597.4635	536.7563	780.7803	674.2539	633.6892
12	11	421	38	421	421	421	421	421	421	421
13	12	NA	42	432.1048	465.3658	433.9346	416.1705	539.84	472.7275	509.3228
14	13	386	40	386	386	386	386	386	386	386
15	14	280	29	280	280	280	280	280	280	280
16	15	514	49	514	514	514	514	514	514	514

また、平均値や標準偏差などの基本統計量を算出するには、上記のコードに続いて、以下のとおり入力すればよい。

```
mean<-c(NA,m)                                  #平均値（M 個）の算出
for(k in 1:m){
 mean[k]<-mean(mat[,k])}
var<-c(NA,m)                                   #分散（M 個）の算出
for(l in 1:m){
 var[l]<-var(mat[,l])}
W<-sum(var^2)/m                                #補定内分散の算出
WSD<-sqrt(W)                                   #補定内標準偏差の算出
B<-(sd(mean))^2                                #補定間分散の算出
BSD<-sd(mean)                                  #補定間標準偏差の算出
T<-W+(1+1/m)*B                                 #全体の分散の算出
TSE<-sqrt(T)/sqrt(nrow(data))                  #全体の標準誤差の算出
mean(mean)                                     #M 個の平均値の統合
WSD                                            #補定内標準偏差
BSD                                            #補定間標準偏差
TSE                                            #全体の標準誤差
```

さらに、Amelia により生成した多重代入済みデータを用いた統計分析を行うには、R パッケージ Zelig を用いればよい。下記では、重回帰分析を行う方法を示す。他の手法への適用方法は、Imai *et al.* (2008)を参照されたい。Amelia の場合と同じく、事前に CRAN からパッケージのインストールを行う必要がある。

```
library(Amelia)                                          #Amelia を起動
set.seed(1223)                                           #シード値を指定
a.out<-amelia(data,m=数)                                 #多重代入法を実行
library(Zelig)                                           #Zelig を起動
z.out<-zelig(変数１ ~ 変数２ + … + 変数 p,
 data=a.out$imputations,model="ls", cite=FALSE)          #ls は線形回帰
summary(z.out)
```

4.5 MAR の検証方法

本稿第 2 章と第 3 章で論じたとおり、欠測メカニズムが MAR であるならば、補助変数を用いた補定によって不完全データの偏りを是正することができる。しかし、欠測データは、定義上、観測されないため、MAR の前提を観測データから直接的に検証することはできない。ただし、これは、欠測データ解析に特有の問題ではない。一般的に推測統計では、一定の前提のもとに統計分析を行っており、こういった前提を観測データから直接的に検証することは不可能である。例えば、回帰分析におけるガウス・マルコフの前提（脚注 34 参照）は、母集団に関する前提であるが、実際にこの前提が母集団データにおいて妥当であるかどうかは、手元にある標本データから直接的に検証することはできない。しかしながら、標本データにおける残差を分析することにより、間接的に診断を行えることが知られている(Fox, 1991)。欠測値補定においても、こういった回帰診断の手法を用いて、補定モデルの妥当性を検証することは重要である(van Buuren, 2012, p.146)。

さらに、Abayomi *et al.* (2008)は、欠測メカニズムの前提と補定モデルとを結びつける間接的な診断方法の提案を行った。Abayomi *et al.* (2008, p.280)は、欠測値が未知なる真の分布に即しているかどうかを検定することはできないが、補定モデルの当てはまりは、常に検証されるべきものでありその場合、観測データに照らして、補定モデルの検証を行うことが自然だと考えられると主張している。

本稿で使用した R パッケージ Amelia においても、4 つの診断手法が組み込まれている(Honaker *et al.*, 2011, pp.25-35)。これらの手法については、高橋, 伊藤(2013, pp.64-74)において詳説した。本節では、MCAR と MAR の欠測メカニズムにより欠測を発生させたシミュレーションデータを用いて、欠測データメカニズムの検証方法を提示する。このデータは、これまで例示に用いてきた収入と年齢の小規模なシミュレーションデータを拡大させて、観測数を増やしたものである。

表 4.7 は、完全データと不完全データの基本統計量である。MCAR の欠測は、乱数により発生させた。MAR における欠測は、ロジスティック回帰分析を用いて発生させた(van Buuren, 2012, pp.31-32; Spies *et al.*, 2014)[74]。

[74] 欠測発生モデルは、$y_i = \frac{1}{1+exp\{-(\hat{\beta}_0+\hat{\beta}_1 x_i)\}}$である。ここで、$y_i$は補定対象変数であり、$x_i$は補助変数である。具体的には、まず、補定対象変数を 0-1 の二値変数として記録しなおす。ここで、0 は欠測値、1 は観測値を表す。次に、補助変数を用いたロジスティック回帰分析により、二値変数の補定対象変数が 0 となる確率を推定する。この結果をもとに、同一の確率カテゴリー内において、乱数を用いて欠測を発生させた。

表 4.7：基本統計量

	最小	第 1 四分位	中央値	平均値	第 3 四分位	最大	標準偏差
収入 1	92	329	414	410	493	682	108.2
収入 2	103	324	410	406	492	682	108.4
収入 3	92	303	373	371	444	622	99.3
年齢	20	33	39	38	44	49	7.0

注 1：収入 1 は完全データ、収入 2 は MCAR による欠測データ、収入 3 は MAR による欠測データ
注 2：n = 697 (完全データ); n = 483 (不完全データ); 欠測率 30.7%

収入 1 は、欠測していない収入の完全データであり、平均値は 410、標準偏差は 108.2 である。観測数は 697 である。

収入 2 は、MCAR により欠測を発生させた収入の不完全データであり、平均値は 406、標準偏差は 108.4 である。観測数は 483、欠測率は 30.7%である。欠測メカニズムが MCAR の場合、理論上、偏りは存在しないと期待されるが、平均値にわずかながら偏りがある。シミュレーションデータであり、欠測メカニズムが MCAR であることが分かっているので、多重代入法(M = 100)により補定を行うことで、平均値は 409.3、標準偏差は 108.4 となり、偏りを是正できる。

収入 3 は、MAR により欠測を発生させた収入の不完全データであり、平均値は 371、標準偏差は 99.3 である。観測数は 483、欠測率は 30.7%である。欠測メカニズムが MAR の場合、理論上、偏りが存在すると期待され、実際に平均値及び標準偏差に偏りが見られる。シミュレーションデータであり、欠測メカニズムが MAR であることが分かっているので、多重代入法(M = 100)により補定を行うことで、平均値は 411.2、標準偏差は 109.1 となり、偏りを是正できる。

真値が不明である場合に、欠測データが MCAR、MAR、NI のいずれであるかをどのように検証すればよいかが問題となるが、下記のとおり、多重代入を実行した後、`plot` 関数、`overimpute` 関数、`disperse` 関数、`missmap` 関数を用いて診断を行う。なお、`plot` 関数と `overimpute` 関数では、検証したい「変数名」を指定する。

```
set.seed(1223)                          #シード値の設定
library(Amelia)                         #Amelia を起動
a.out<-amelia(data, m=100)              #多重代入法を実行
plot(a.out,which.vars="変数名")         #密度の比較：「変数名」を指定
overimpute(a.out,var="変数名")          #過剰補定：「変数名」を指定
disperse(a.out,dims=1,m=1000)           #過散布初期値
odrdata<-order(data[,2])                #欠測地図のためにデータを並べ替え
 data2<-data[odrdata,]
 missmap(data2,rank.order=FALSE,        #欠測地図
 y.cex=0,col=c(1,2))
```

図 4.15 と図 4.16 は、`plot` 関数を用いた密度の比較の結果である。この図では、観測値の密度と補定値の密度の比較を行っている。もし欠測メカニズムが MCAR であるならば、

2 つの密度はほぼ重なると考えられる。一方、MAR の場合、観測データと欠測データとの間には、体系的な差があると考えられるので、今回の図のように、2 つの密度は重ならず、一定の方向に偏りがあることが見受けられる。

図 4.15：密度の比較(MCAR)

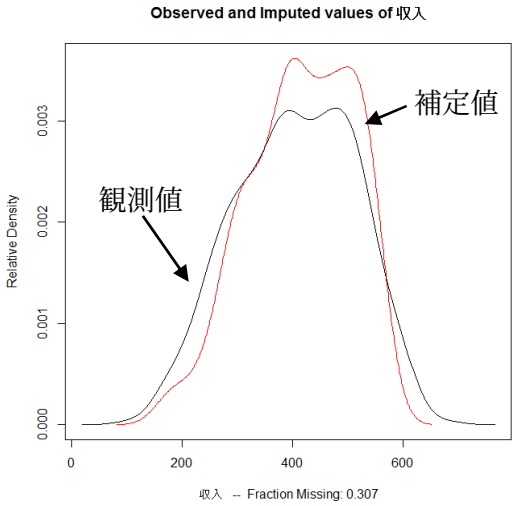
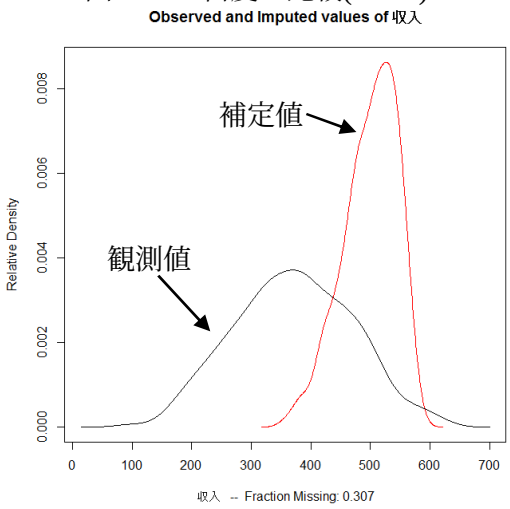


図 4.16：密度の比較(MAR)



Abayomi *et al.* (2008)及び Honaker *et al.* (2011)のいずれにおいても、密度の比較を行う際には、*M* 個の補定値を統合した平均を使用している。しかし、この方法によって図示される結果は、本質的に単一代入法の結果と変わらず、補定間のばらつきが密度に反映されない。そこで、EMB アルゴリズムによる多重代入法を独自にプログラムし、*M* 個の補定値を個別に図示するコードを開発した（2015 年現在、コードは非公開だが、デバック後、公開の予定）。100 回の多重代入法を実行した結果、図 4.17 の MCAR のケースでは、観測データは補定データの中に完全に含まれて一致している。図 4.18 の MAR のケースでは、観測データとすべての補定データとの間に一定の差が見られる。

図 4.17：密度の比較(MCAR)

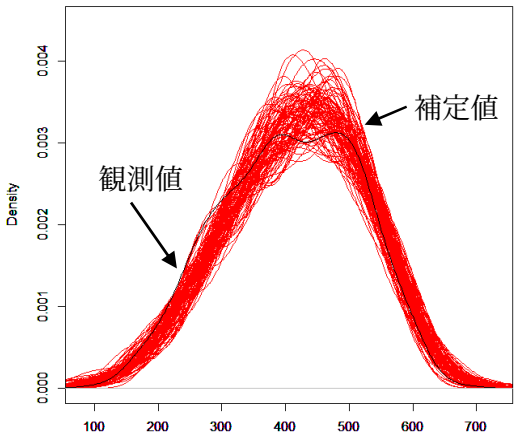


図 4.18：密度の比較(MAR)

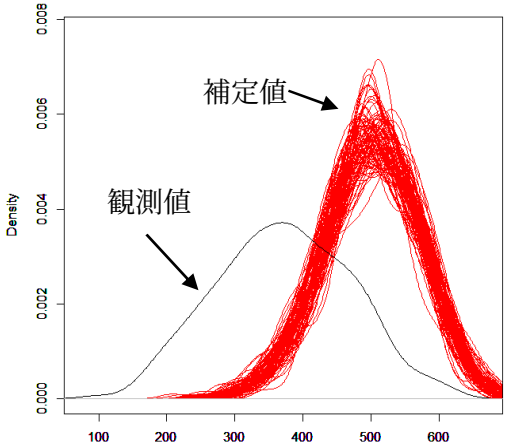


図 4.19 と図 4.20 は、`missmap` 関数を用いた欠測地図である。年齢順にデータを昇順に

並べ替えている。このように、観測データをいろいろな順序に並べ替え、不完全データにパターンが見られるかどうかを視覚的に確認することができる。MCAR の場合、年齢の並び順に関わらず、収入の欠測は全体的にばらついていることが視覚的に分かる。一方、MAR の場合、年齢が高くなるほど、収入の欠測率が上がっていく様子が分かる。

図 4.19：欠測地図(MCAR)

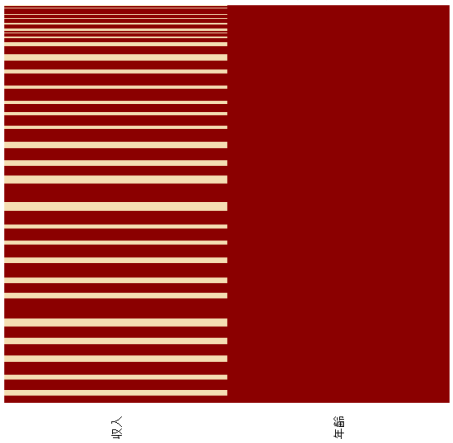


図 4.20：欠測地図(MAR)

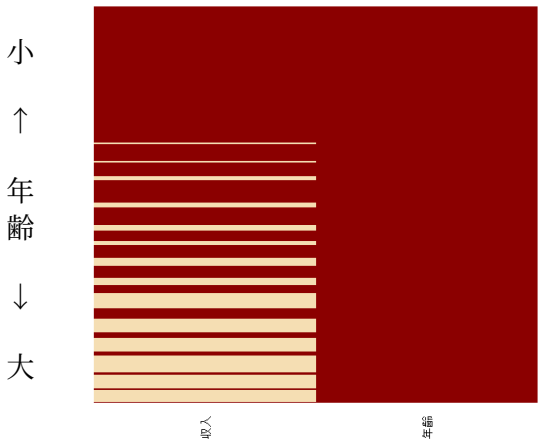


図 4.21 と図 4.22 は、過剰補定の図である。この機能は、Amelia に特有のもので、補定モデルを構築した後に、観測データを 1 つずつ人工的に欠測させ、数百回の多重代入による 90%信頼区間を図示したものである。過剰補定に関する詳細な理論は、Blackwell *et al.* (2015)を参照されたい。この図の横軸は観測データ、縦軸は補定データである。この図には、y = x となる 45 度線が付されており、補定モデルの当てはまりがよい場合には、90%信頼区間の中に y = x の 45 度線が含まれると想定される。また、引数として横軸を xlim=c(数, 数)と指定し、y = x の線を視覚的にも 45 度にすることが望ましい。

図 4.21：過剰補定(MCAR)

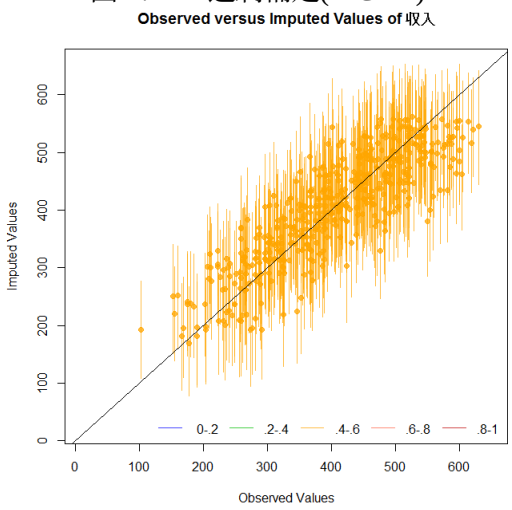
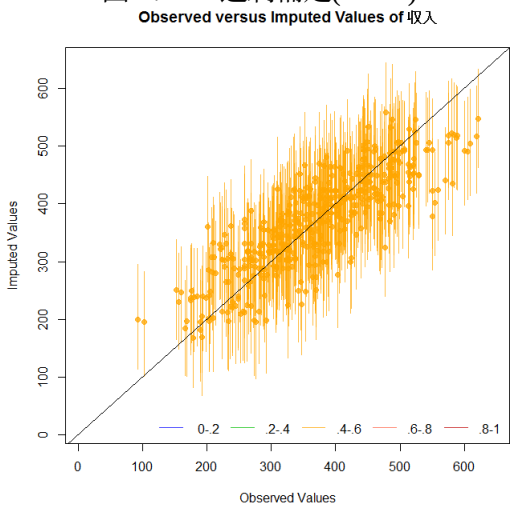


図 4.22：過剰補定(MAR)



最後に、MAR の検証と直接の関係はないが、4.1.2 項で見たとおり、EMB アルゴリズム

による多重代入法では、何らかの初期値を設定する必要がある。EM アルゴリズムは、異なる初期値に対して比較的頑健であり、尤度関数が単峰である場合には、極大値に収束する。しかし、尤度関数が単峰ではない場合、EM アルゴリズムは局所的最大値に収束する場合があるため、複数の初期値が同一の値に収束するかどうかを判定する必要がある(渡辺, 山口, 2000, pp.39-40)。図 4.23 と図 4.24 は、過散布初期値の図である。今回の場合、1,000 個の初期値を使用したすべての結果が、同一の結果に収束していることが視覚的に分かる。

図 4.23：過散布初期値(MCAR)

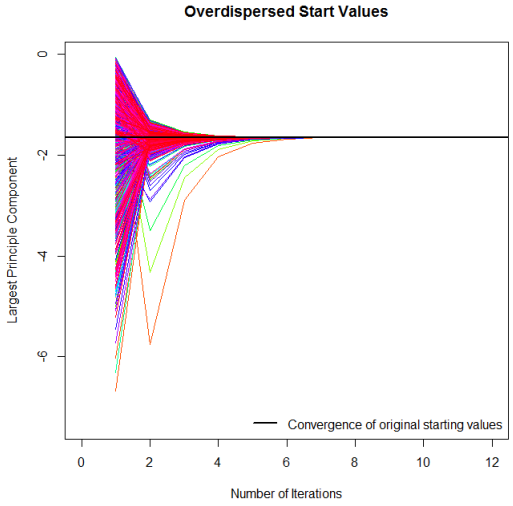
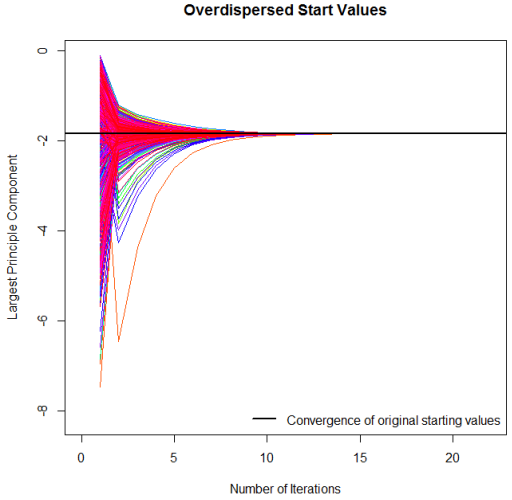


図 4.24：過散布初期値(MAR)



4.6 事前分布の導入による補定の改善

コラム 1 で述べたとおり、多重代入法はベイズ統計学の枠組みで構築されたものであり、事前分布と尤度を合算させることにより事後分布を構成する。つまり、事前分布を明示的に指定するとき、多重代入法の真価が発揮される。したがって、本節では、事前分布の活用方法を紹介する。EMB アルゴリズムにおける事前分布は、期待値ステップにおいて導入され、期待値ステップにおける補定値を通じて間接的に最大化ステップに影響を与えるものである。EM アルゴリズムは一般的に最尤推定法だと見なされるが、欠測データを処理する場合、最大化ステップにおける事前分布はベイズの事前分布と見なすことができる(Honaker and King, 2010, p.570)。

4.6.1 観測値に関する事前分布

事前分布とは、過去の研究やデータ、学術上や一般の常識、個人の経験などに基づく追加情報である。ある変数の値が欠測していたとしても、おおよその値が分かっているということは多い。例えば、日常生活において、他人の身長の値を正確に知っていることは稀だが、目測により「A さんは隣に立ったとき、自分よりも 3 から 5cm ぐらい背が高い」といった具合で、ある程度の情報が分かるかもしれない。自分の身長が 170cm だとすれば、A さんの身長の事前分布は、95%有意水準において、173cm から 175cm である。この分布の中心

が平均値、つまり 174cm である。標準偏差×2 の中に 95%の観測値が含まれることを考えれば、標準偏差は 1/2、つまり 0.5 と指定できる。このような事前の情報があるとき、個別の欠測データのセルに関して、一般的なモデルパラメータではなくベイズ事前分布として事前情報を Amelia に取り入れることができる(Honaker *et al.*, 2011, pp.20-23)。なお、観測値に関する事前分布は、Blackwell *et al.* (2015)の提唱する過剰多重代入法(multiple overimputation)と原理的には同じものである。

事前分布の活用は、通常のベイズ分析と同様のロジックにしたがって行われる。つまり、補定はモデルと事前分布との重み付け平均であり、この重みはデータと事前分布との相対的な優劣によって決定されるものである。すなわち、モデルの予測力が高い場合には事前分布の重みが下がり、モデルの予測力が低い場合には事前分布の重みが上がるということである。なお、個別の観測値に関する事前分布とは、欠測データセルの分布に関する分析者の信念を表すものである。これは、上述の例のように、平均値と標準偏差という形で取り入れることができる。

Amelia において観測値に関する事前分布を入力するには、4 列からなる事前分布行列を構築する。この行列の各々の行は、観測値または変数に関する事前分布を表す。また、行列の 1 列目は観測値の行番号であり、行列の 2 列目は観測値の列番号である。行列の 3 列目と 4 列目は、欠測値の事前分布における平均値と標準偏差である。

例えば、表 4.8 にあるとおり、収入の変数における 7 番目の観測値について、個人的な会話から 600 万円台であることが分かっているとしよう。また、収入の変数における 10 番目の観測値についても、個人的な会話から 700 万円前後であることが分かっているとしよう。

表 4.8：事前分布を含む例

ID	収入	年齢	事前分布
1	543	51	なし
2	272	24	272
3	797	59	797
4	239	26	239
5	415	35	なし
6	371	34	371
7	650	54	600~700
8	495	47	495
9	553	56	553
10	710	55	680~720
11	421	38	421
12	410	42	なし
13	386	40	386
14	280	29	280
15	514	49	514

注：灰色セルは無回答による欠測、白抜き数字は本来得られるはずの真値を表す。

ID7 の収入の値は、600 万円から 700 万円の間なので、中心を 650 万円とし、上下に 50 万円のばらつきがある。標準偏差×2 の中に 95%の観測値が含まれることを考えれば、事前分布の標準偏差は 50/2 = 25 と指定できる。すなわち、平均値 650 万円、標準偏差 25 という事前分布として理解することができる。入力する情報は、以下のとおり pmat1 を 1×4 行列として構築すればよい。7 は観測値がデータの 7 行目にあることを意味する。1 は売上高の変数がデータの 1 列目にあることを意味する。650 と 25 は上記で説明したとおりである。Amelia には、priors=pmat1 として入力すればよい。

```
pmat1<-matrix(c(7,                                  #観測値の番号
                1,                                  #変数の番号
                650,                                #事前分布の平均値
                25),                                #事前分布の標準偏差
                nrow=1, ncol=4)                     #1×4 の事前分布行列
a.out<-amelia(data,m=m,priors=pmat1)                #多重代入法の実行
```

ID10 の収入の値について、「700 万円前後」という情報があるとし、これを 680 万円から 720 万円と解釈すれば、平均値 700 万円、標準偏差 10 の事前分布と指定できる。以下のとおり pmat2 を 2×4 行列として構築する。1 列目の情報は、上記と同様に ID7 に関する事前分布である。2 列目の情報が、新たに加えた ID 10 に関するものである。

```
pmat2<-matrix(c(7,      10,                         #観測値の番号
                1,       1,                         #変数の番号
                650,  700,                          #事前分布の平均値
                25,    10),                         #事前分布の標準偏差
                nrow=2, ncol=4)                     #2×4 の事前分布行列
a.out<-amelia(data,m=m,priors=pmat2)                #多重代入法の実行
```

表 4.9 は、観測値に関する事前分布を考慮した場合と考慮しない場合とを比較した結果である。事前情報が増えるにしたがって、平均値も標準偏差も改善し、真値に近づいていることが分かる。より多くの観測値について信頼に足る事前の情報が増えることで、補定の精度を向上させられる可能性があると期待できる。今後、経済センサスのデータが蓄積することにより、こういった事前分布を活用できるようになり、多重代入法の有益性が向上すると言える。さらに 3 個以上の観測値に関して事前情報がある場合には、同様の手順にて、事前分布行列に 3 列目、4 列目を追加すればよい。

表 4.9：観測値に関する事前分布の影響

	真値	LW	多重代入法 事前分布なし	多重代入法 事前分布あり 観測値 7	多重代入法 事前分布あり 観測値 7&10
平均値	470.4	432.8	462.1	462.9	471.4
標準偏差	161.7	166.8	-	-	-
標準誤差	41.8	52.8	-	-	-
WSD	-	-	155.2	156.2	164.0
BSD	-	-	12.3	10.7	7.4
TSE	-	-	40.2	40.4	42.4
CI UL	-	-	486.8	484.3	486.3
CI LL	-	-	437.4	441.6	456.6
観測数	15	10	15	15	15

注 1：- は該当値がないことを表す。
注 2：LW はリストワイズ除去(List-Wise Deletion)により欠測を除去した値を表す。
注 3：WSD は補定内における標準偏差、BSD は補定間における標準偏差、TSE は全体の標準誤差である。
注 4：CI は欠測に起因する誤差に関する信頼区間（95%水準）であり、UL は上限、LL は下限を表す。
注 5：多重代入法では、各々のシード値に対して、$M = 1000$ を実行した。

4.6.2 変数の値に関する事前分布

通常、多くの変数において、値の取り得る範囲は事前に分かっていることが多い。例えば、身長の値は物理的な長さである以上、論理的に負になることはなく、また、成人男性の身長は 50cm 以上、280cm 未満であることが経験的に分かっている。経済データでは、論理的に売上高や従業者数が負の値になることはない。このように、ある変数の取り得る値の範囲が論理的あるいは経験的に既知であるならば、それを事前分布の一種としてモデルに組み込むことも可能である。こういった問題は、一般的に、適切な変数変換により処理できることが多いが、Amelia では切断正規分布モデルからの無作為抽出を行うことで、補定値を既知の範囲内に収めることができる(Honaker *et al.*, 2011, pp.23-25)。

これまで用いてきた収入と年齢のデータでは、確定的回帰補定モデル（3.1 節）において年齢が 0 歳の場合、補定値は－80.8 万円となる。仮に、表 4.10 のように、ID5 の年齢が 0 歳だったとしよう。0 歳にして収入が 415 万円という例は、非現実的だが、例示用データにおいて負の補定値を算出するための人為的な設定である。

表 4.10：年齢 = 0 を含む例

ID	収入	年齢
1	543	51
2	272	24
3	797	59
4	239	26
5	415	0
6	371	34
7	650	54
8	495	47
9	553	56
10	710	55
11	421	38
12	410	42
13	386	40
14	280	29
15	514	49

注：灰色セルは無回答による欠測、白抜き数字は真値を表す。

収入は、金額データであるため、論理的に非負のデータ、すなわち 0 以上の値である。よって、収入の最小値は 0 円と設定することが論理的に可能である。Amelia において変数の値の範囲を指定するには、3 列からなる行列を構築する。最初の列は、データ内における変数の列番号を意味する。2 つ目の列は、範囲の最小値を表す。3 つ目の列は、範囲の最大値を表す。入力する行列は bds のようにすればよい。下記の例では、1 は収入の変数がデータ内の 1 列目にあることを意味し、0 は収入の最小可能値が 0 であることを意味している。797 は観測データ内における最大値である。また、Amelia には、bounds=bds として入力すればよい。

```
bds <- matrix(c(1,0,797), nrow =1 , ncol = 3)
a.out <- amelia(data, m = m, bounds = bds)
```

また、仮に、このデータが東京都にて週 40 時間労働で勤務する人たちに関するものだとしよう。東京都における最低賃金は時給 888 円である。よって、年間収入は最小でも 170 万円以上あると考えられる。もし最低賃金を事前分布として考慮するならば、下記のとおり入力すればよい。

```
bds2 <- matrix(c(1,170,797), nrow =1 , ncol = 3)
a.out <- amelia(data, m = m, bounds = bds2)
```

表 4.11 は、収入の補定値の最小可能値を設定しなかった場合、0 と設定した場合、170 と設定した場合の結果を示している。もし事前分布を用いなかった場合、ID5 の補定値は、平均－78.2 万円、標準偏差 90.4、最大値 135.4 万円、最小値－513.2 万円であった。しかし、事前分布を設定することにより、モデル自体の切片を強制的に 0 とすることなく、補定値を正の値に収めることができる。その結果、補定値の精度も向上している。

表 4.11：変数の値の範囲に関する事前分布の影響

	真値	LW	多重代入法 事前分布なし	多重代入法 事前分布あり 最小値 0	多重代入法 事前分布あり 最小値 170
平均値	470.4	432.8	432.1	439.1	449.4
標準偏差	161.7	166.8	-	-	-
標準誤差	41.8	52.8	-	-	-
WSD	-	-	209.0	191.6	171.0
BSD	-	-	8.9	11.0	11.1
TSE	-	-	54.0	49.6	44.2
CI UL	-	-	449.9	461.0	471.6
CI LL	-	-	414.2	417.1	427.3
観測数	15	10	15	15	15

注 1：- は該当値がないことを表す。
注 2：LW はリストワイズ除去(List-Wise Deletion)により欠測を除去した値を表す。
注 3：WSD は補定内における標準偏差、BSD は補定間における標準偏差、TSE は全体の標準誤差である。
注 4：CI は欠測に起因する誤差に関する信頼区間（95%水準）であり、UL は上限、LL は下限を表す。
注 5：多重代入法では、各々のシード値に対して、$M = 1000$ を実行した。

今回は、例示目的のため、あえて ID5 の年齢を 0 歳に設定した。実際にこのようなデータがある場合には、補助変数である年齢の値が誤っている可能性を探るべきである。補助変数の値が正常な範囲であるにも関わらず、補定値が補定対象変数の常識的な範囲外になった場合に、今回のような措置を講ずるべきである。

また、変数の値の取るべき範囲を強制的に指定することには注意が必要である。Honaker *et al.* (2011, p.23)は、変数の値の範囲を満たすべきなのは M 個の補定値の平均であって、M 個の補定における個別の補定値は範囲の外にあってもよいと言う。その理由は、補定値が範囲外に出るということ自体が、補定における真の不確実性を反映しているからである。ただし、M 個の補定値の平均が範囲外に出た場合、補定モデル自体の信頼性が疑わしく、モデルを再構築するべきである。一方、van Buuren and Groothuis-Oudshoorn (2011, p.11) によると、補定値は、マイナスのカウントや妊娠した父親のように、明らかに不可能な値とならないようにするべきだという主張もある。ゆえに、範囲を強制的に指定することには、議論の余地がある。

4.6.3 リッジ事前分布

経済センサス‐活動調査の全データは全数調査であり数百万の観測数を持つ。統計分析に用いるデータは、均一なグループであることが望ましいため、補定においても、都道府県や産業分類などにより層分けし、均一なグループ内で行われるべきものと考えられる。その結果、各グループは、小規模なものとなる可能性がある。本章では、データの層分けについて深く追求せず、都道府県や産業分類といった常識的な範囲での分け方を使用したが、どの

ような層分けがベストであるかは事前には分からず、データの蓄積を待ちながら検討するべき課題だと言える。

層の分け方次第では、欠測率が非常に高かったり、観測数が非常に少なかったりするグループが発生するおそれがある。そういった場合には、EM アルゴリズムは非常に不安定となり、補定の結果は補定モデルの指定の仕方に大きく依存することとなる。

こういった場合には、リッジ事前分布[75]を追加する策が考えられる(Honaker *et al.*, 2011, pp.19-20)。リッジ事前分布とは、各変数の平均値と分散はそのままにしつつ、変数間の共分散をゼロに近づけることで、モデルの安定性を達成しようというものである(Schafer, 1997, pp.155-157)。この手法は、現存するデータと同じ平均値と分散を持ち、共分散が 0 となる人工的な観測値を追加していると考えることができる。多くのベイズ手法における分析と同様に、リッジ事前分布は、偏りの増加を犠牲としつつ分散を減らすことで効率性を向上させようとするものである。このトレードオフの関係において、偏りの不利益が効率性の利益を上回らないように調整することが肝要である。

Amelia では、`empri` の右辺に数値を指定することで導入できる。データ全体の 1%としたい場合には `empri=0.01*nrow(data)`とし、データ全体の 2%としたい場合には `empri=0.02*nrow(data)`などとする。

```
a.out <- amelia(data, m = m, empri = 0.01*nrow(data))
```

引き続き、収入と年齢のデータを用いる。データ数は 15 個、観測数は 10 個しかない。仮に、補助変数の二乗や三乗を含めた多項式を用いたい場合、モデルが不安定になるおそれがある。今回は、例として、補助変数に年齢、年齢の二乗、年齢の三乗、年齢の四乗を用いたモデルを採用した。すなわち、観測数 10 に対し、推定すべきパラメータが 5 つあり、非常に不安定なモデルである。

リッジ事前分布を利用した場合の結果は、表 4.12 に示すとおりである。リッジ事前分布を用いることで、補定モデルを安定させることができ、平均値及び標準偏差の値が改善している。しかし、事前分布の値を 0.1 から 0.2 に大きくすると、標準偏差の値は改善したが、平均値は悪化した。リッジ事前分布をどの値に設定するのがベストであるか、理論的にあらかじめ分かるわけではない。よって、データが蓄積するにしたがって、事前情報としてのリッジ事前分布の設定に関する情報が得られれば、小規模データにおける補定モデルを安定させることができると期待される。

[75] Ridge prior

表 4.12：リッジ事前分布の影響

	真値	LW	多重代入 事前分布なし	多重代入 事前分布あり 0.1	多重代入 事前分布あり 0.2
平均値	470.4	432.8	452.4	456.9	455.4
標準偏差	161.7	166.8	-	-	-
標準誤差	41.8	52.8	-	-	-
WSD	-	-	201.2	155.8	157.9
BSD	-	-	62.9	10.9	14.4
TSE	-	-	54.5	40.3	40.9
CI UL	-	-	578.1	478.6	484.2
CI LL	-	-	326.7	435.1	426.6
観測数	15	10	15	15	15

注 1：- は該当値がないことを表す。
注 2：LW はリストワイズ除去(List-Wise Deletion)により欠測を除去した値を表す。
注 3：WSD は補定内における標準偏差、BSD は補定間における標準偏差、TSE は全体の標準誤差である。
注 4：CI は欠測に起因する誤差に関する信頼区間（95%水準）であり、UL は上限、LL は下限を表す。
注 5：多重代入法では、各々のシード値に対して、$M = 50$ を実行した。

4.7 他の多重代入法アルゴリズムとソフトウェア

多重代入法は、ベイズ統計学における情報更新のメカニズムを利用して、観測データを条件として欠測値の事後分布を構築し補定を行うものである。上述したとおり、複数の補定済みデータセットを構築するには、平均値、分散、共分散といったパラメータを複数回推定する必要があるが、その点に関して、いくつかのアルゴリズムが提唱されている。

Rubin (1987)の提唱した従来の多重代入法は、ベイズ統計学の代表的なアルゴリズムであるマルコフ連鎖モンテカルロ法[76]に基づいていた。R パッケージ Norm や SAS の MI プロシージャなどがこれを採用している(SAS Institute Inc., 2011; Fox, 2015)。また、ユトレヒト大学の van Buuren (2012)により、連鎖方程式による完全条件付指定[77]アルゴリズムも提唱されており、R パッケージ mice、SPSS、SOLAS などがこれを採用している(SPSS Inc., 2009; Statistical Solutions, 2011; van Buuren and Groothuis-Oudshoorn, 2011; van Buuren and Groothuis-Oudshoorn, 2015)。これらのアルゴリズムとソフトウェアについての詳細は、高橋, 伊藤(2014, pp.46-55)を参照されたい。

4.8 これまでの研究成果

経済データにおける欠測値補定方法を検証するために、金融庁によって管理されている EDINET データを用いて確定的単一代入法と多重代入法による精度の評価を行った。その初期の研究成果は、2012 年 9 月に北海道大学にて開催された統計関連学会連合大会(高橋, 伊藤, 2012a)とノルウェーにて開催された国連欧州経済委員会(UNECE)の統計的データエディティングに関するワークセッション(Takahashi and Ito, 2012)にて報告を行った。ま

[76] Markov chain Monte Carlo (MCMC)
[77] Fully Conditional Specification (FCS)

た、確定的単一代入法、確率的単一代入法、多重代入法による精度の評価を行った初期の研究成果は、2012 年 11 月に奈良教育大学にて開催された科学研究費シンポジウム(高橋, 伊藤, 2012b)にて報告を行った。これらの学会等で得られた知見を反映し、加筆修正した最終的な研究成果は、2013 年 3 月刊行の統計研究彙報第 70 号に掲載されている(高橋, 伊藤, 2013a)。

また、3 つの多重代入法アルゴリズムと 6 つのソフトウェアのパフォーマンスについて、EDINET データを用いて検証を行った初期の研究成果は、2013 年 8 月に香港で開催された ISI 世界統計大会(Takahashi and Ito, 2013)及び 2013 年 9 月に大阪大学にて開催された統計関連学会連合大会(高橋, 伊藤, 2013b)にて報告を行った。また、3 つのアルゴリズムのパフォーマンスについて、平成 24 年経済センサス‐活動調査の速報データを用いて検証を行った初期の研究成果は、2013 年 11 月に金沢大学で開催された科学研究費シンポジウム (高橋, 伊藤, 2013c)にて報告を行った。これらの学会等で得られた知見を反映し、加筆修正した最終的な研究成果は、2014 年 3 月刊行の統計研究彙報第 71 号に掲載されており(高橋, 伊藤, 2014)、アルゴリズムについては、Amelia の EMB が最良だと分かった。その成果をもとに、本稿は Amelia による EMB を中心に論じている。

4.9 多重代入法と多重化単一代入法の違い

前述したとおり、多重代入法による補定値は単一代入法を複数回実行したものではなく、欠測データの事後分布から無作為抽出したパラメータ推定値を用いたシミュレーション値である。本節では、多重代入法と多重化単一代入法の違いを散布図により例証する。van Buuren (2012, p.55)も合わせて参照されたい。なお、本稿で言う多重化単一代入法とは、複数のシード値を用いて確率的単一代入法を複数回実行したものである。

図 4.2 は、4.1 節における「モデルと補定済みデータ」の散布図と同一であり、多重代入法による補定モデルと補定値を図示している。図 4.25 は、図 3.5 における「モデルと補定済みデータ」をもとにシード値を 3 回変えて実行した確率的回帰補定の結果を図示したものである。

図 4.2：多重代入法（注 1）

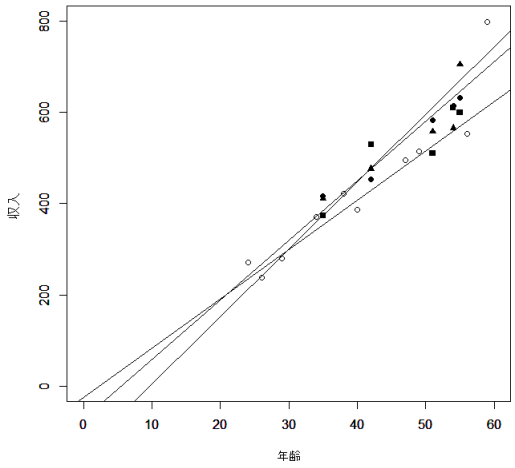
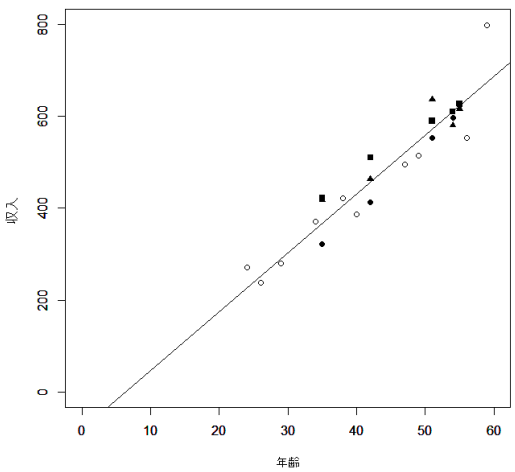


図 4.25：多重化単一代入法（注 2）



注 1：切片 1 = －28.0、傾き 1 = 10.9；切片 2 = －66.0、傾き 2 = 12.8；切片 3 = －144.7、傾き 3 = 14.9
注 2：切片 = －80.8、傾き = 12.8
注 3：○は観測値、■、●、▲は、それぞれ 1～3 回目の補定値を表す。

これらの図から明らかなように、多重代入法では、複数の回帰モデルにより回帰係数の安定性を評価できる。しかし、多重化単一代入法では、回帰モデルは 1 つだけであり、回帰係数の安定性を評価できない。つまり、多重代入法は、単一代入法を複数回実行したものではないのである。

4.10 多重代入法の 8 つの利点

表 4.13 は、多重代入法を用いる 8 つの利点をまとめたものであり、どの手法に対して優位であるかを示している。総合的に、単一代入法はこれら 8 つの利点を同時に達成することができず、多重代入法を用いる方が有益である。

表 4.13：多重代入法の 8 つの利点

利点	下記に対して優位
1. 平均値の推定	確率的単一代入法
2. 標準偏差の推定	確定的単一代入法
3. 標準誤差の推定	単一代入法全般
4. 複数の補助変数の同時活用	比率補定
5. シード値の影響	確率的単一代入法
6. 事前情報の活用	単一代入法全般
7. 対数正規分布データの補定	確定的単一代入法
8. 補定間の分散による診断	単一代入法全般

利点 1 の平均値の推定に関して、$M = 100$ の多重代入法による平均値は確定的単一代入法の値とほぼ一致し、どちらも不偏推定量である。一方、確率的単一代入法による推定値は、撹乱項の影響を受けるため、精度が低い。利点 2 の標準偏差の推定に関して、撹乱項を導入している多重代入法と確率的単一代入法が推奨され、確定的単一代入法では過小推定とな

る。利点 3 の標準誤差の推定に関して、推定不確実性と根本的不確実性を同時に考慮に入れることができるのは、多重代入法だけである。利点 4 の複数の補助変数の同時活用は、多重代入法に限らず回帰による補定モデル全般に当てはまる利点であり、特に比率補定の大きな欠点である。利点 5 のシード値の影響は、確率的単一代入法の大きな欠点である。M = 100 の多重代入法の結果は、シード値の影響をほとんど受けない。当然ながら、確定的単一代入法もシード値の影響は受けない。利点 6 の事前情報の活用は、ベイズ手法としての多重代入法の特徴であり、頻度論的手法としての単一代入法にはない利点である。データがより多く蓄積することで、多重代入法と単一代入法との優劣が広がると考えられる。利点 7 の対数正規分布データの補定は、付録 4 で論じているが、多重代入法と確率的単一代入法の副次的な利点である。利点 8 の補定間の分散による診断は、多重代入法に固有の機能であり、補定モデルを前提とした上で、欠測に起因する誤差を数値評価できる。

コラム 2：コンピュータの歴史

ここまで見てきたとおり、多重代入法では補定済みデータを複数生成するため、演算量が多くなると懸念されることがある。そこで、本コラムでは、コンピュータの処理速度の歴史について概観する。

Rubin が多重代入法を提唱した 1970 年代は、パーソナルコンピュータが発売され始めた時期であり、大容量で高速な演算は大型コンピュータ（汎用コンピュータ）によってのみ可能であった。1969 年に初めて人類が月面着陸したアポロ計画におけるコンピュータ(AGC)は、約 2MHz の演算能力であった。また、総理府統計局（現総務省統計局）は、国勢調査の集計のため 1961 年に IBM705（主記憶容量 40KB）を導入した。

パーソナルコンピュータ元年とも言われる 1977 年には、スティーブ・ジョブズらにより Apple II が発売された。この時期からマイクロチップを搭載したコンピュータが作成され、コンピュータの高速化、小型化が進展した。1980 年代には、大規模集積回路(LSI)が開発され、卓上パーソナルコンピュータで高速の演算処理が行えるようになった。先の Apple II を発売したアップルコンピュータ社は、1984 年に Macintosh (128K)を発売し、我が国においても普及した。

1990 年代に入ると、インターネットが普及し始め、1995 年にマイクロソフト社から発売されたオペレーティングシステム(OS) Windows 95 の登場により、ソフトウェアの開発が大幅に進展し、個人へのコンピュータの普及が急進した。1990 年代のパーソナルコンピュータ普及後、マイクロプロセッサの高速化とハードディスクなどのストレージの大容量化及び小型化が進んだ。2.5 インチハードディスクドライブ(HDD)が搭載されるようになり、ノート型が主流となった。

1990 年代後半から 2000 年代前半にかけては、CPU の高速化が進展し、1995 年に CPU クロック周波数は 100MHz 程度であったのが、2000 年には 1 GHz に達し、2004 年には 3.8GHz と高速化した。2000 年代後半にはクロック周波数は改善せずマルチコア CPU が主流を示すようになった。2001 年に Windows XP が、2009 年に Windows 7 が発売され、ハードディスク容量 1 TB のものが安価に入手可能となった。

このように、コンピュータは、年々、高速化・大容量化・小型化・低価格化しており、2015 年現在では、家庭用のパーソナルコンピュータでも大規模な演算を行えるようになっている。すなわち、現在では、多重代入法をストレスなく実行できる環境が整っている。

出典 1：川崎(2010)

出典 2：情報処理学会のウェブサイト[78]

[78] http://museum.ipsj.or.jp/computer/personal/index.html（2015 年 6 月 1 日閲覧）

5　諸外国の公的統計における多重代入法の研究と適用事例

ここまで見てきたとおり、調査データにおける欠測への対処は常に問題となるものであり、特に多重代入法による対処が推奨される。しかし、我が国の公的統計に多重代入法が導入されたことはない。そこで、諸外国における補定に関する研究と適用の事例について調査を行った(高橋, 2014)。5.1 節では米国の公的統計における欠測への対処法について、5.2 節ではドイツの公的統計における欠測への対処法について、5.3 節ではスイスの公的統計における欠測への対処法について、それぞれ紹介する。

5.1 米国における欠測値補定

米国の経済センサスの補定は、経済調査に関する統計手法と研究部門[79]において行われている。補定は、エディティングと一体化して同時に実施している。すなわち、エディット規則に基づいて、補定すべき項目を検出し、1 つ目の項目に複数のモデルを当てはめ、最も当てはまりのよいモデルを用いて補定を実施する。その後、再び 2 つ目の項目へと戻り、複数のモデルを当てはめ、最も当てはまりのよいモデルを用いて補定を実施する。このプロセスの繰り返しである。

基本的には、収集しているデータは、売上高や従業員数など、マイナス値になる項目はない。また、エディット規則を満たすことを優先しており、補定においてもマイナス値が算出されることはあり得ない。基本的な考え方として、比率、内容チェック、レンジチェック、バランスチェックの 4 点を重視している。また、欠測値の補定だけではなく、観測値であったとしてもエディット規則に適合していない値は、上記の手順にしたがってエディティング及び補定を実施して訂正している。

特に補定に関しては、基本的に IRS[80]から提供を受けた税務データを用いて、コールドデック[81]によって実測値を確保するように努めている。実測値を入手できない欠測値に関しては、統計モデルを利用して補定を行っている。具体的には、Plain Vanilla というシステムを用いて比率補定を多用しているが、比率補定以外にも、複数のモデルを指定でき、実際にどのモデルを用いるかは実務専門家により決定される(羽渕, 上田, 小高, 高橋, 小泉, 2012)。つまり、欠測値を補定する行為は、ダイナミックな営みであり、事前にモデルを完全に固定することはできない。テストデータを用いて事前に入念な準備をした上で、実際のデータの状況に即して臨機応変にモデルの構築と再構築を行わなければならないのである。

米国センサス局の実務では単一代入法を用いているが、多重代入法の研究も行っている(García *et al.*, 2014)。所得及びプログラム参加調査[82]では、現在、確定的ホットデック[83]を用いて補定を行っているが、ホットデックではカテゴリー数の増加によってドナー数の減

---

79 Office of Statistical Methods and Research for Economic Programs
80 Internal Revenue Service：内国歳入庁 = 国税庁
81 コールドデックについては、脚注 29 を参照されたい。
82 Survey of Income and Program Participation (SIPP)：「所得及び社会保障受給調査」の意
83 ホットデックについては、脚注 30 を参照されたい。

少につながり、同一のドナーから複数の欠測値を補定する確率が高くなる。逆に、セル数を分割することでドナー数は増えるが、ドナーとレシピエントの特徴を捉えにくくなるといったことが問題点として指摘されている。米国センサス局の統計研究技術センター[84]では、TEA という独自ソフトウェアを用いて、ランダムホットデックと逐次抽出型回帰による多重代入法(SRMI)[85]に基づいた 2 つの手法を検証している。検証の結果、SRMI には、データセット内のどのような変数でもモデルに組み込める可能性があることが利点として挙げられているが、計算負荷が高いということが分かった。

結果、計算負荷といった実務上の問題のため、米国の経済センサスは単一代入法を用いており、多重代入法を導入する予定はない。しかし、SRMI は、本質的に柔軟性に優れるが計算効率で劣る完全条件付指定(FCS：4.7 節参照)と同じアルゴリズムであり、EMB アルゴリズムを用いることで計算負荷の問題は回避できると考えられる(高橋, 伊藤, 2013a)。また、コラム 2 にあるとおり、近年のコンピュータの性能を鑑みれば、計算負荷は事実上、問題ではなくなりつつある。

米国の公的統計では、全米保健統計センターの国民健康調査において、世帯所得と個人収入の欠測値補定に多重代入法を適用している(National Center for Health Statistics, 2013)。国民健康調査は、様々な収入のレベルに応じた健康状態を確認し、収入と健康状態の関係について研究するためのデータを提供することを目的としている。しかし、世帯収入と個人収入の欠測率が高いため、多重代入法を用いて欠測値の処理をしている。1997 年から 2015 年までの調査票、データセット、関連文書は、米国疾病予防管理センター[86]のウェブサイト[87]に公開されており、$M = 5$ の多重代入済みデータもここで公開されている。

5.2 ドイツにおける欠測値補定

ドイツ連邦統計局の数理統計手法部門[88]では、2010 年の農業センサスデータを用いた多重代入法の研究を行っている(Spies *et al.*, 2014)。母集団サイズは 14,213 の農場であり、補定の対象としている変数は「水の消費量」である。欠測数が 2,088 あり、エラーが 213 個あるため、合計で 2,301 のレコードに欠測が発生している。すなわち、欠測率は約 16.2%である。使用した補助変数は、灌漑用地のサイズ、気候上の水分平衡、使用可能な農場のキャパシティー、灌漑手法、水源の種類である。MAR を前提として、ベイズ統計学に基づく 2 つの補定方法の検証を行った。1 つ目はランダムホットデックであり、2 つ目は予測平均値マッチング[89]である。暫定的な結論として、ホットデックでは極端な値が補定値として採用される可能性があるが、予測平均値マッチングでは現実的な補定値が採用される可能性が高いということが分かった。なお、このモデルは、Rubin (1987, p.168)に記載されているモ

[84] Center for Statistical Research and Methodology
[85] Sequential Regression Multiple Imputation (SRMI)
[86] Centers for Disease Control and Prevention (CDC)
[87] http://www.cdc.gov/nchs/nhis/quest_data_related_1997_forward.htm（2015 年 6 月 1 日閲覧）
[88] Department for Mathematical-Statistical Methods
[89] Predictive Mean Matching (PMM)

デルである。

ドイツ連邦統計局では、この研究と並行して、2010 年農業センサスにおいて多重代入法を初めて採用した[90]。予備的な導入であったため非常に限られた範囲での適用だが、数理統計手法部門において推定値の算出を行い、結果を農業部門に提出し、そこから欧州統計局(Eurostat)に結果が転送された。実務上の課題として、試験調査データを用いて補定モデルを準備したが、実データに応用する際になって、モデルの当てはまりがよくないことが判明し、大幅な改良を施す必要があった。統計作成の段階では、大幅な変更を実施するのに十分な時間的余裕がないため、予期しない問題が発生しても素早く正確に対応できるように、多重代入法の手法について実務者があらかじめ精通している必要がある。米国の場合と同様に、補定モデルを事前に完全に固定できるわけではない。

また、調査誤差について、農業センサスは全数調査であるが、全数調査であっても欠測データにより誤差が発生する。すなわち、2.1 節で記したとおり、完全データとしての全数調査における標本誤差は 0 だが、不完全データとしての全数調査には欠測に起因する非標本誤差が存在する。このような種類の誤差を考慮するには、多重代入法により補定間の標準偏差を算出しなければならない。このように、理論的に健全な分散に関する推定値を入手できることが、多重代入法の主な長所である。

一方、欠測値の問題は調査実施後の問題であり、多重代入法を用いることによって分散に関する情報を得ることで、データ品質が非常に低いということが判明した場合に、どのような対処をするべきかという実務上の問題がある。そもそも、こういったことは、従来の単一代入法による補定手法による実務では無視される傾向にあった問題である。つまり、欠測による非標本誤差が大きかったとしても、それを数値化して評価できなかったために、そのような誤差はないものとして処理してきたということである。

なお、ドイツ連邦統計局では、集計値の公表に際して、変動係数 0.15 未満を基準としている(Spies *et al.*, 2014, p.8)。変動係数とは、相対標準偏差とも呼ばれ、標準偏差を平均値で割ったものである(熊原, 渡辺, 2012, pp.97-98)。すなわち、表 4.5 における補定間標準偏差(BSD)の値と補定内標準偏差(WSD)の値を統合した全体の標準偏差を平均値で割ったものが変動係数である。この値が 0.15 を上回る集計値は公表に適さず、モデルの変更などを行って対処する。

5.3 スイスにおける欠測値補定

スイス連邦統計局では、2010 年及び 2013 年の所得と生活状況に関する統計調査[91]における欠測値補定に多重代入法を用いた(Swiss Federal Statistical Office, 2014)。この調査は、世帯を対象として、貧困や生活状況に関する調査を行うことを目的としている。所得の

[90] 2010 年ドイツ農業センサスにおける多重代入法の適用に関する記述内容は、ドイツ連邦統計局(Sarah Giessing: Head of Section, Mathematical-Statistical Methods of Data Editing and Imputation)への筆者の独自取材に基づいている（2014 年 8 月 12 日；2014 年 12 月 4 日）。

[91] Statistics on Income and Living Conditions (SILC)

欠測値は、SAS のマクロである IVEware を用いた多重代入法により補定されている。補定値には、フラグが立てられており、どの値が補定されたかが分かるようになっている。

スイス連邦統計局[92]によると、IVEware を用いて補定を行った主な理由は、非常によい補助変数があり高精度な回帰モデルを作ることのできる環境があったからである。補定の対象とした変数は対数変換した所得であり、補助変数は社会保障の支払額である。また、SAS の既存の機能を用いることで、多変量回帰補定を行うことが可能であり、使用方法も簡便である。実際のところ、補定による分散の推定が行えることは、非常によい副産物であった。複数の多重代入済みデータセットをどう扱うか、また、これらをユーザーに提供すべきかどうかが実務上の大きな問題になり、多重代入済みデータセットが複数算出されることで、データのボリュームも実務上の問題になる。よって、最終的に維持するデータは、多重代入法による補定済みデータの平均 1 つのみである。なお、スイス連邦統計局では、2000 年の国勢調査においても多重代入法を使用したことがあるとのことであった。

92 2010 年の所得と生活状況に関する統計調査における多重代入法の適用に関する記述内容は、スイス連邦統計局(Daniel Kilchmann: Methodologist)への筆者の独自取材に基づいている（2014 年 7 月 7 日；2014 年 8 月 18 日）。

6　平成 24 年経済センサス‐活動調査のデータへの多重代入法の適用例

第 3 章と第 4 章では、単一代入法と多重代入法の理論的な考え方を小規模なシミュレーションデータを用いて解説した。本章では、経済センサス‐活動調査の実データを用い、多重代入法の有用性を示す。6.1 節では、平成 24 年経済センサス‐活動調査における産業大分類 J（金融業，保険業）のデータの中から特に東京都のデータに注目し、MAR の欠測を発生させた上で多重代入法の検証を行う。このように検証することで、真値と補定値との比較を行うことができる。6.2 節では、同じく平成 24 年経済センサス‐活動調査における産業大分類 J の東京都のデータを用い、実際の欠測値を多重代入法により補定する。

なお、経済センサス‐活動調査の全産業を用い、確定的回帰補定、確率的回帰補定、比率補定、多重代入法の比較検証を行った結果は、本稿第 7 章を参照されたい。また、確定的回帰補定、確率的回帰補定、多重代入法による精度の比較検証については、高橋，伊藤(2013a)も参照されたい。

6.1 多重代入法の検証例：産業大分類 J（東京都）のデータ

ここまで、欠測値補定における多重代入法の有用性について述べてきた。本節では、実際に平成 24 年経済センサス‐活動調査の確報データに多重代入法を用いた例を示す。すべてのデータを同時に用いることは現実的ではないため、産業大分類 J の東京都のデータ（単独事業所）を用いた。売上（収入）金額（以下、売上高）の欠測値を補定する補助変数としては、費用総額（以下、費用）、資本金額（以下、資本金）、従業者合計男女（以下、従業者）、常用雇用者数（以下、雇用者）、事業従事者数男女（以下、事業従事者）などが候補として考えられる。

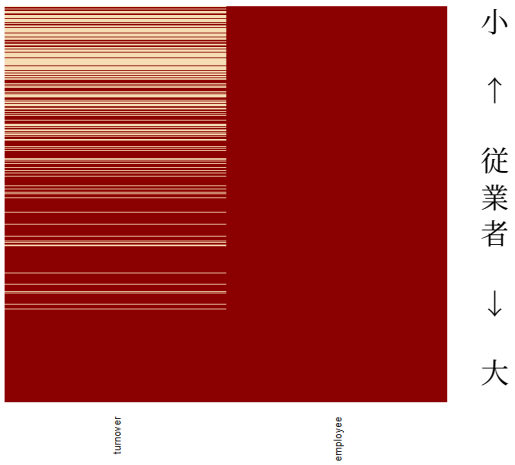
ここでは詳細は省くが、従業者以外の補助変数から補定できる売上高の欠測値の数は極めて少ない。売上高の欠測指標を 100 とした場合、費用の欠測指標は 99.9、資本金の欠測指標は 106.8、従業者の欠測指標は 77.6、常用雇用者の欠測指標は 109.1、事業従事者の欠測指標は 91.4 である。欠測指標が 100 を超える変数については、補定対象である売上高よりも欠測率が高く、補助変数として有用ではない。これらの変数の中で、従業者の欠測率は極めて低い。そこで、本節では、もし補助変数として従業者を用いて欠測値補定をした場合、どのような結果が得られるかを示す。また、多重代入法により、得られた推定結果の信頼性の検証を行う。このようにすることで、従来ではパラメトリックモデルによって補定することがはばかられたデータに関して補定を行うことができ、製表業務の大幅な効率化が期待できる。

6.1.1 データの基本統計量

欠測値補定の精度を評価するため、欠測値を含む行を除去したデータを完全データとして利用し、ロジスティック回帰分析を用いて MAR の原則によりデータを欠測させる(van

Buuren, 2012, pp.31-32; Spies *et al.*, 2014)[93]。具体的には、従業者数が少なくなるほど売上高の欠測率が上がる形で欠測を発生させた。図 6.1 は、従業者(employee)を昇順に並び替えた欠測地図である。欠測地図では、白く表されている部分がデータ内の欠測箇所であり、欠測のパターンを視覚的に確認できる。従業者が少なくなればなるほど売上高(turnover)の欠測率が高くなっており、欠測は MAR になるように設定してあることが分かる。上述したとおり、これは検証のために発生させた人工的な欠測の結果であり、実データにおける欠測と同一ではない。欠測地図については、高橋, 伊藤(2013a, pp.65-66)も参照されたい。

図 6.1：欠測地図



完全データ及び不完全データの基本統計量は、表 6.1 に示すとおりである。MAR の原則に基づいて欠測を発生させたため、完全データにおける売上高の平均値の真値(101302)と比べて、不完全データにおける売上高の平均値(131710)は過大推定になっている様子が伺える。この偏りをどれだけ是正できるか、また、欠測に起因する誤差がどれだけ発生しているかを本節で示す。

表 6.1：データの基本統計量

	第 1 四分位	中央値	平均値	第 3 四分位	標準偏差
売上高（完全）	922	2733	101302	9688	967628
売上高（不完全）	1332	3612	131710	13980	1113129
従業者	2	3	14	7	93

注：欠測値の数は 547、完全データにおける観測数は 2203、不完全データにおける観測数は 1656 である。経済センサス‐活動調査の実データを用いた分析のため、開示リスクを回避する意図で、最小値と最大値は伏せている。売上高の単位は 100 万円、従業者の単位は人である。

また、表 6.1 を見ると、中央値から第 1 四分位と第 3 四分位までの距離は均等ではなく、各々の変数の分布は正規ではないと思われる。対数データの基本統計量は、表 6.2 のとおりであり、中央値から第 1 四分位及び第 3 四分位までの距離が是正されている様子が分かる。

93 具体的な方法は、脚注 74 も参照されたい。

表 6.2：自然対数変換後の基本統計量

	第 1 四分位	中央値	平均値	第 3 四分位	標準偏差
売上高（完全）	6.827	7.913	8.111	9.179	2.184
売上高（不完全）	7.195	8.194	8.515	9.546	2.160
従業者	0.693	1.099	1.413	1.946	1.131

注：欠測値の数は 547、完全データにおける観測数は 2203、不完全データにおける観測数は 1656 である。経済センサス‐活動調査の実データを用いた分析のため、開示リスクを回避する意図で、最小値と最大値は伏せている。

図 6.2 は完全データにおける売上高（生データ）のヒストグラムであり、図 6.3 は完全データにおける売上高（自然対数）のヒストグラムである。図 6.2 では 0 の付近に多くの観測値が集まり、徐々に減っており、経済データに典型的な対数正規分布の形を示している。図 6.3 では、釣鐘型の分布になっており、自然対数に変換することで正規分布を近似している。

図 6.2：売上高（生データ：完全）

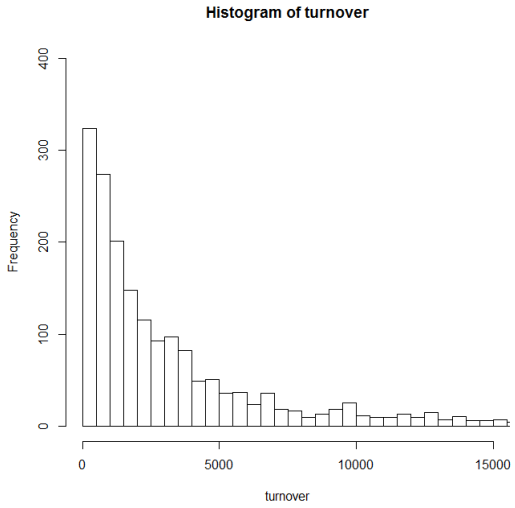


図 6.3：売上高（対数データ：完全）

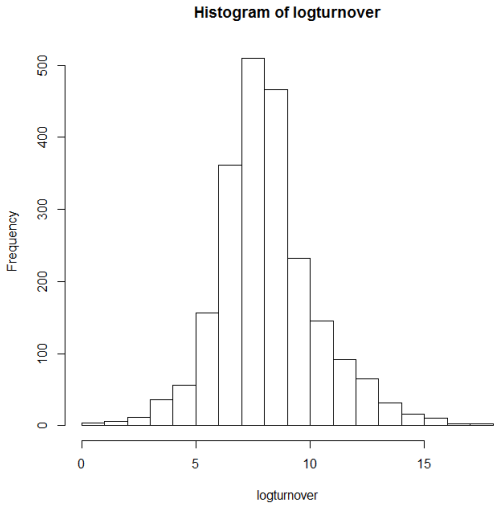


図 6.4 は不完全データにおける売上高（生データ）のヒストグラムであり、図 6.5 は不完全データにおける売上高（自然対数）のヒストグラムである。おおむね、売上高の低い値に欠測が多く発生している。また、完全データの場合と同様に生データは対数正規分布の形をしており、自然対数に変換することで正規分布を近似している。

図 6.4：売上高（生データ：不完全）

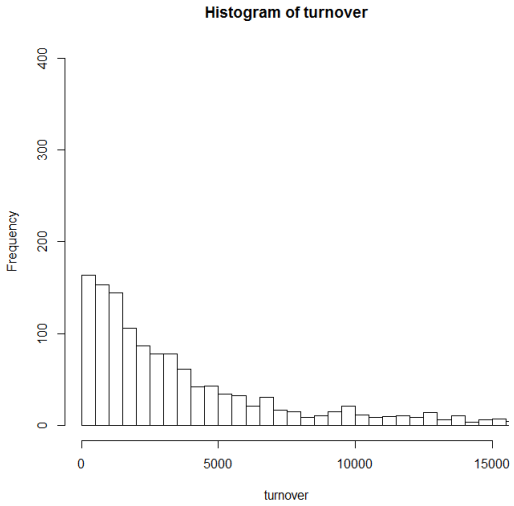


図 6.5：売上高（対数データ：不完全）

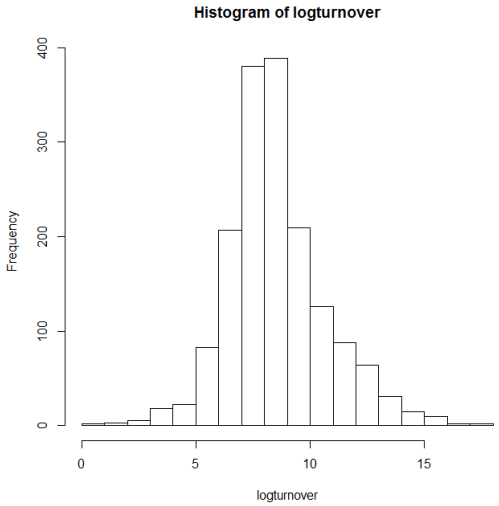


図 6.6 は従業者（生データ）のヒストグラムであり、図 6.7 は従業者（自然対数）のヒストグラムである。図 6.6 では 0 の付近に多くの観測値が集まり、徐々に減っており、対数正規分布の形を示している。図 6.7 では、自然対数に変換することで正規分布に近づいているが、右すそが長く、近似は完全ではない。

図 6.6：従業者（生データ）

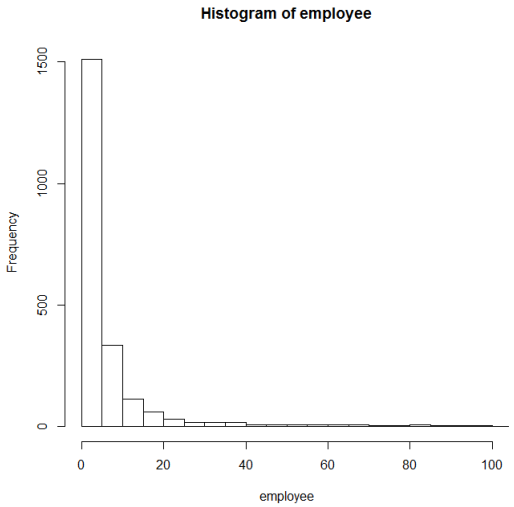


図 6.7：従業者（対数データ）

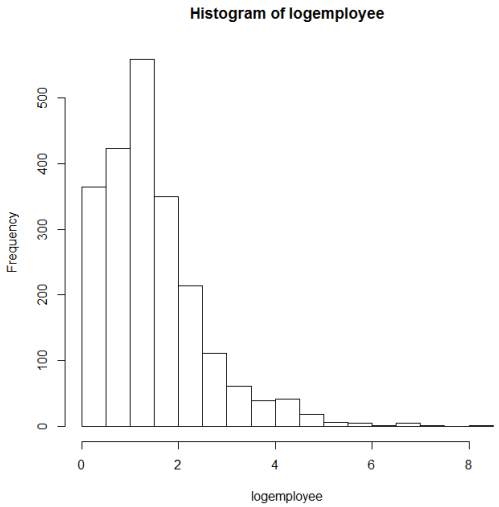


表 6.3 に歪度と尖度を示す。正規分布の場合、歪度は 0 であり、尖度は 3 である。売上高も従業者も、生データの場合、歪度は 0 よりも大きく、尖度も 3 よりも大きい。自然対数に変換した売上高の歪度と尖度は、正規分布を近似していることが分かる。つまり、自然対数に変換したデータの周辺分布は近似的に正規分布だと言える。自然対数に変換した従業者の歪度は 1.298 でやや歪みが残っており、尖度は 5.863 でやや尖っている。生データより正規分布に近くなっているが、近似は完全ではない。

表 6.3：歪度と尖度

	歪度 (正規分布 = 0)	尖度 (正規分布= 3)
売上高（生データ）	21.381	549.895
売上高（自然対数）	0.485	4.362
従業者（生データ）	24.049	745.592
従業者（自然対数）	1.298	5.863

周辺分布が正規であることと同時分布が正規であることは、同義ではないため、R パッケージ mvnormtest によりシャピロ-ウィルク検定(Jarek, 2015)を実行し、多変量正規分布の検定を行った。この検定における帰無仮説は、「データは多変量正規分布」であり、統計量 W が 1 に近いほど正規性を示す。すなわち、帰無仮説を棄却しなければデータが多変量正規分布だと想定でき、帰無仮説を棄却すればデータが多変量正規分布だと想定できない(Shapiro and Wilk, 1965)。結果は表 6.4 に示すとおりである。今回のデータでは、生データも自然対数変換後のデータも、いずれも p-値が 0.000 であり帰無仮説を棄却できる。したがって、いずれの場合もデータは多変量正規分布ではない。ただし、自然対数に変換することで、W の値は 1 に近づいており自然対数変換後のデータは多変量正規分布に近づいていることが分かる。

表 6.4：シャピロ-ウィルク多変量正規分布検定

	W	p
生データ	0.090	0.000
自然対数	0.945	0.000

図 6.8 は生データにおける従業者と売上高の散布図であり、図 6.9 は対数データにおける従業者と売上高の散布図である。

図 6.8：生データの散布図

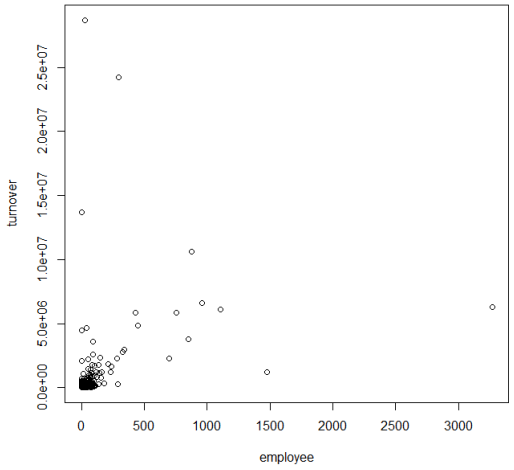
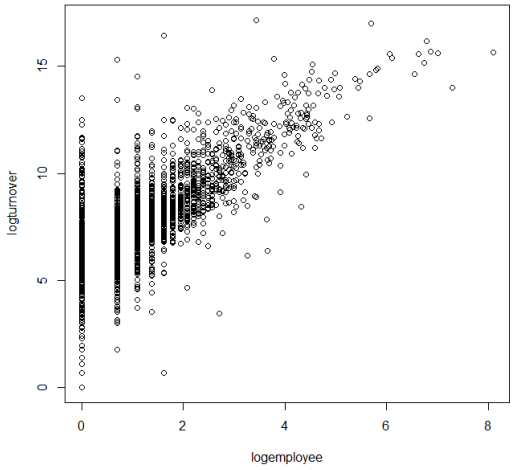


図 6.9：対数データの散布図



6.1.2 多重代入法による欠測値補定の結果

6.1.1 節では正規性の検定を行ったが、その結果だけでは、自然対数モデルが最適であるとの明確な結論は得られなかった。そこで、6.1.1 節で示したデータを用いて多重代入法を実行した。結果は表 6.5 のとおりである。

表 6.5：分析結果

	完全データ	欠測データ	1 次線形モデル	自然対数モデル
平均値	101302	131710	114182	99813
標準偏差	967628	1113129	-	-
WSD	-	-	1098932	966625
BSD	-	-	12910	112
観測数	2203	1656	2203	2203

注：- は該当値がないことを表す。多重代入の *M* 数は 100 に設定した。WSD（補定内標準偏差）と BSD（補定間標準偏差）については、4.2 節も参照されたい。

BSD（補定間の標準偏差）を見ることで、補定モデルの信頼度を評価できる。1 次線形モデルによる平均値(114182)は、真値(101302)との乖離が大きいだけではなく、BSD の値も 12910 と大きく、モデルの信頼性が低い。一方、自然対数モデルによる平均値(99813)は、真値との乖離が小さいだけではなく、BSD の値も 112 と小さく、モデルの信頼性が高い。よって、自然対数モデルによる多重代入法を行うことで、欠測による偏りを大幅に是正できていることが分かる[94]。もちろん、実際の現場では、真値は常に不明である。そういった場面でも、BSD の値を見ることで間接的に補定モデルの信頼性を評価できる。

なお、単一代入法による確定的回帰補定の結果は、平均値 116599（1 次線形モデル）と平均値 99802（自然対数モデル）である。しかし、真値が分からない限り、単一代入法では結果の評価が行えない。

今回の検証では、多重代入済みデータ数(*M* 数)を 100 に設定したが、一般的な家庭用 PC を用いた場合、分析に必要な時間はわずか数分である（本稿付録 2 も参照されたい）。

図 6.10 と図 6.11 は、平均値の経験分布である。1 次線形モデルでは、BSD の値が 12910 と非常に大きく図 6.10 では平均値が大きくばらついている。一方、自然対数モデルでは、BSD の値が 112 と非常に小さく図 6.11 では平均値が一定の値を示している。

[94] 自然対数による補定値を生データのスケールに戻す方法については、付録 4 を参照されたい。

平均値のヒストグラム（経験分布）

図 6.10：1 次線形モデル

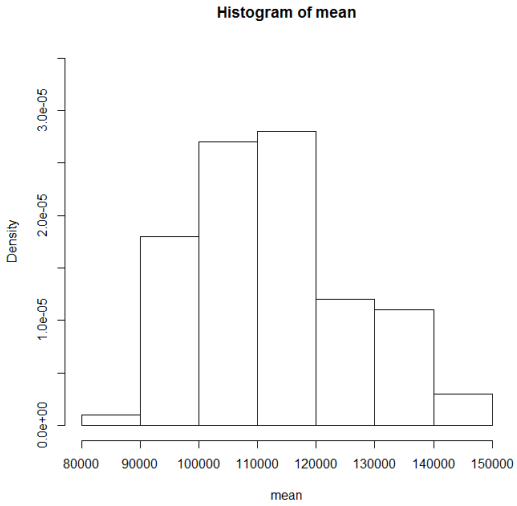
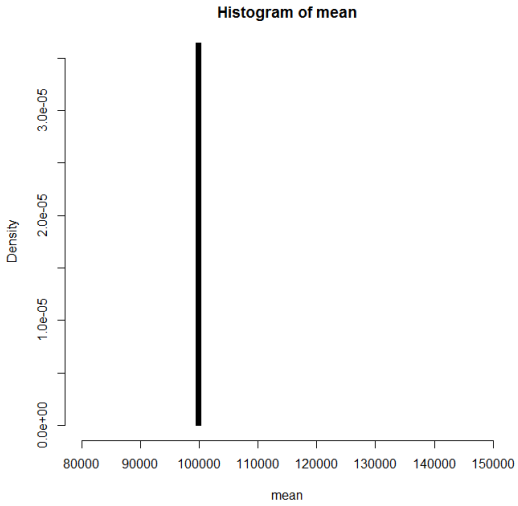


図 6.11：自然対数モデル



また、補定モデルの回帰線を図 6.12 と図 6.13 のように描き出すことで、視覚的にモデルの安定性を評価することもできる[95]。図 6.12 の 1 次線形モデルによる回帰線は、大きく上下に揺らいでおり、モデルが不安定であることが分かる。一方、図 6.13 の自然対数モデルによる回帰線は、ほぼ一定のラインを保っており、非常に安定したモデルだと分かる。

回帰直線の信頼度(M = 100)

図 6.12：1 次線形モデル

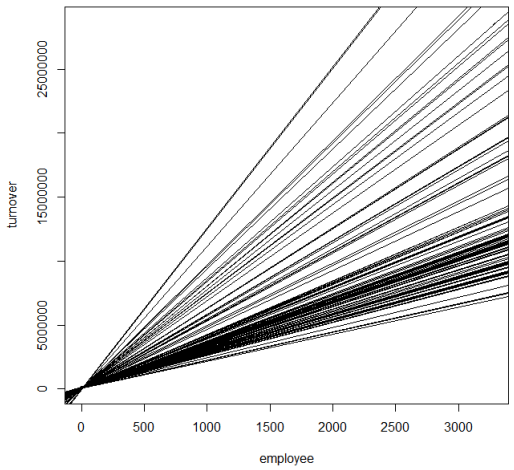
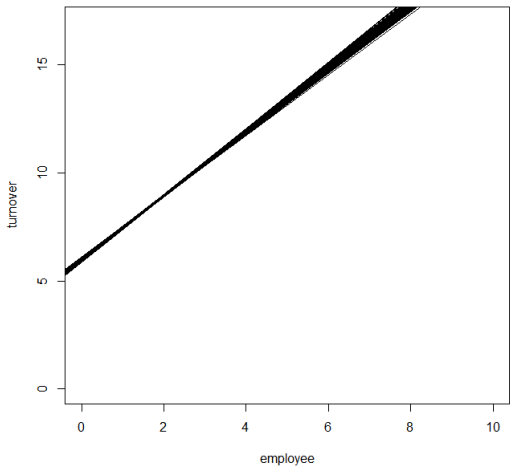


図 6.13：自然対数モデル



このように、補定を行うだけではなく、補定の診断評価を行うことができる点が、多重代入法を用いる大きな利点である(Takahashi, 2014b)。同様の方法で、平方根モデルなど、他

[95] Amelia には補定モデルの回帰線を算出する機能が搭載されていないため、図 6.12 と図 6.13 は EMB アルゴリズムによる多重代入法を自前でプログラムした結果である。

のモデルの検証にも応用できる。さらに、密度の比較と過剰補定など、他の診断手法については、4.5 節及び高橋, 伊藤(2013a, pp.64-74)を参照されたい。

6.2 多重代入法の実行例：産業大分類 J（東京都）のデータ

前節から引き続き、産業大分類 J（金融業，保険業）の東京都のデータ（単独事業所）を用いた。本節では、欠測を人工的に発生させるのではなく、実際の欠測値を補定する。なお、先ほど述べたとおり、売上高の欠測指標を 100 とした場合、費用の欠測指標は 99.9 であり、費用を補助変数としたモデルから補定できる売上高の欠測値の数は多くない。しかし、費用と売上高の相関係数は約 0.99 と極めて高く、精度の高い補定ができると期待される[96]。そこで、少ないながらも精度の高い補定ができると予想される「費用→売上高」モデルについて、実データにおける欠測値を多重代入法により補定することで、その精度の評価を行う。

実際の欠測値を補定しているため、本節では、真値は不明である。しかし、多重代入法などの診断手法を駆使することで、補定モデルの精度評価を行うことができる。まず始めに、費用を昇順に並べた欠測地図（図 6.14）を利用して、欠測のパターンを確認する。費用が少なくなればなるほど、売上高の欠測率が高くなる様子が伺え、欠測のメカニズムは MAR であると示唆されている[97]。

図 6.14：欠測地図

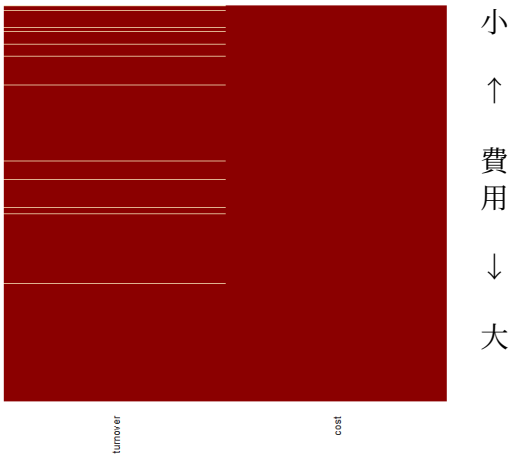


表 6.6 が結果である。図 6.10 の欠測地図から分かるとおり、不完全データにおける売上高の平均値は、真値を過大推定していると思われる。1 次線形回帰モデルと自然対数モデルによる売上高の平均値は、それぞれ、81564 と 81361 である。特に自然対数モデルの BSD（補定間標準偏差）は極めて小さく、推定精度が高いことが分かる。

96 相関の高い変数同士では、高い精度の補定ができると期待できる。しかし、同時欠測の確率も高く、補助変数としての総合的な有用性は低い可能性がある。

97 この図だけで欠測メカニズムが MAR であるという確証が得られるわけではないが、今回の場合、MAR の条件と矛盾しない結果が得られている。MAR の検定については、4.5 節も参照されたい。

表 6.6：分析結果

	不完全データ	1 次線形モデル	自然対数モデル
平均値	83520	81564	81361
標準偏差	930923	-	-
WSD	-	918547	918361
BSD	-	426	39
CI UL	-	82415	81439
CI LL	-	80712	81282

注：- は該当値がないことを表す。多重代入の M 数は 100 に設定した。WSD（補定内標準偏差）と BSD（補定間標準偏差）については、4.2 節も参照されたい。CI UL は 95%信頼区間の上限、CI LL は 95%信頼区間の下限を表す。

表 6.7 は、1 次線形モデルの回帰係数の基本統計量である。表 6.7 から、傾きの 95%信頼区間は C.I.(0.951, 1.075)で、ほぼ 1.000 であり、1 対 1 で比例関係にあることが示唆されている。しかし、切片の 95%信頼区間は、C.I.(2857.3, 12444.4)であり、生データの場合、欠測値を考慮した場合のモデルにおいて切片は 0 ではないと考えられ、比率補定は適切でない可能性を示唆している。

表 6.7：1 次線形モデルの回帰係数

	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
切片	968.792	5970.280	7548.130	7650.850	9109.730	16454.900	2396.767
傾き	0.976	0.995	1.003	1.013	1.018	1.175	0.031

表 6.8 は、自然対数モデルの回帰係数の基本統計量である。表 6.8 から、傾きの 95%信頼区間は C.I.(0.987, 1.019)で、ほぼ 1.000 であり、1 対 1 で比例関係にあることが示唆されている。また、切片の 95%信頼区間は、C.I.(−0.101, 0.171)である。ゆえに、対数変換データの場合、欠測値を考慮した場合のモデルにおいて、切片は 0 であると考えられ、比率補定は適切である可能性を示唆している。

表 6.8：対数モデルの回帰係数

	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
切片	−0.105	−0.015	0.038	0.035	0.080	0.226	0.068
傾き	0.982	0.998	1.004	1.003	1.008	1.019	0.008

また、補定モデルの回帰線を図 6.15 と図 6.16 に描き出している[98]。前節で使用した従業者とは異なり、費用の場合、図 6.15 の 1 次線形モデルによる回帰線も比較的安定していることが分かる。しかし、図 6.16 の自然対数モデルによる回帰線は、さらに安定的な一定のラインを保っており、こちらの方が優れたモデルであることが示唆されている。

[98] Amelia には補定モデルの回帰線を算出する機能が搭載されていないため、図 6.15 と図 6.16 は EMB アルゴリズムによる多重代入法を自前でプログラムした結果である。

回帰直線の信頼度($M = 100$)

図 6.15：1 次線形モデル

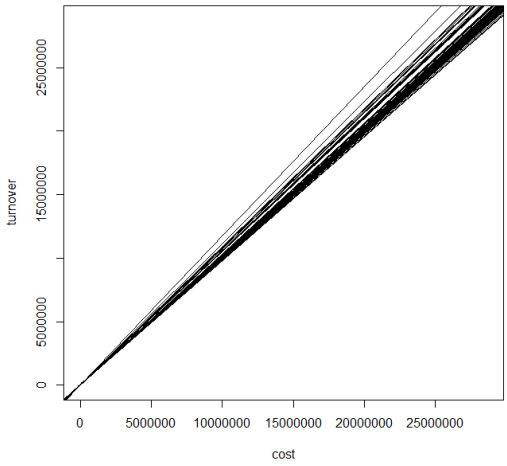
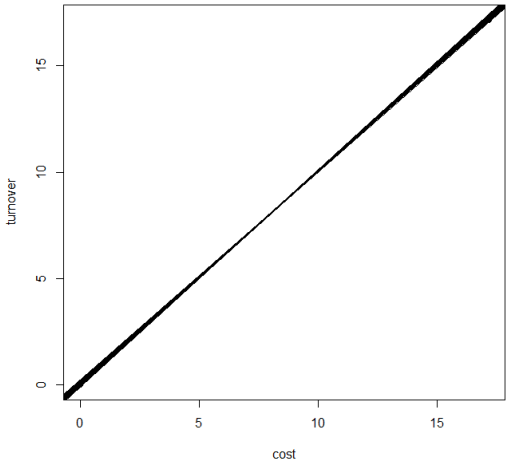


図 6.16：自然対数モデル



ただし、これは、必ずしも費用の方が従業者よりも優れた補助変数であるということを意味しない。というのも、「費用→売上高」で補定できる欠測値の数は少なく、「従業者→売上高」で補定できる欠測値の数が多いため、自ずと後者の方が補定による誤差が大きくなるからである。

図 6.17 は密度の比較の結果である。図 6.14 の欠測地図から推測されるとおり、補定値の密度は小さい値に偏っており、欠測メカニズムは MAR であることが示唆される。図 6.18 の過剰補定から、補定モデルの当てはまりは非常によいと結論付けることができる。

自然対数モデルの診断結果

図 6.17：密度の比較

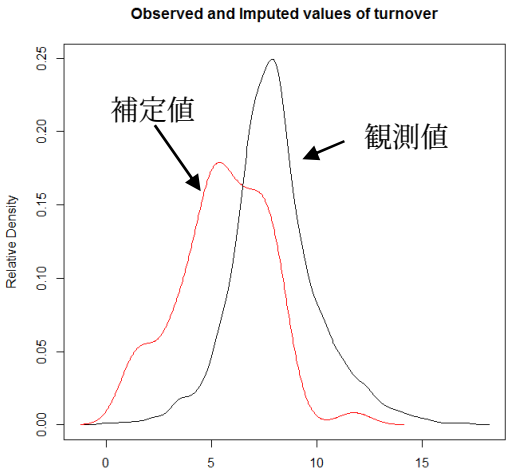
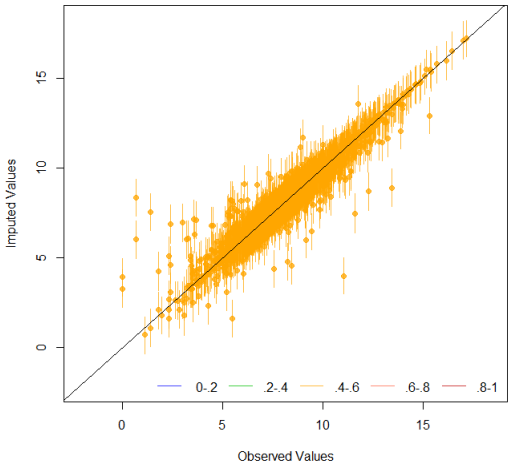


図 6.18：過剰補定



7　平成 24 年経済センサス‐活動調査のデータを用いた多重代入法の検証

平成 24 年経済センサス‐活動調査のデータを用い、確定的単一代入法、確率的単一代入法、比率補定及び多重代入法により、費用から売上の補定を行い、その結果を集計した。7.1 節では、使用したデータの基本統計量（自然対数）を示し、検証方法の方針を説明する。7.2 節では、4 つの補定手法による補定の結果を示す。

7.1 データの基本統計量と検証方法

平成 24 年経済センサス‐活動調査では、「単独事業所調査票」、「企業調査票」、「事業所調査票」、「産業共通調査票」の４種類の調査票が用いられた。ここで、単独事業所調査票は単独事業所（他の場所に同一経営の本所（本社・本店）や支所（支社・支店）を持たない事業所）の調査に用いられ、企業調査票・事業所調査票は同一経営の複数の事業所を持つ事業所（それぞれ本所・支所）の調査に用いられた。産業共通調査票は新規事業所の調査に用いられ、単独事業所、本所、支所の別を記入する欄があった。これらの中で、今回の検証では、単独事業所調査票と産業共通調査票のうち単独事業所のデータを検証の対象とした。また、今回の検証における補定は産業中分類別に行うこととした。

表 7.1 は自然対数変換を行った売上の基本統計量（産業中分類別[99]）である。表 7.1 から、中分類 02 の売上の歪度が－1.145 でやや歪んでおり、中分類 50、72、85 の売上の尖度が、それぞれ、5.248、4.816、4.906 でやや尖っているが、全体的にはどの中分類においても、売上の歪度はほぼ 0 であり尖度はほぼ 3 である。正規分布の場合、歪度は 0 であり尖度は 3 である。よって、売上を自然対数に変換することで、周辺分布において正規分布を近似していることが分かる。

表 7.2 は自然対数変換を行った費用の基本統計量（産業中分類別）である。表 7.2 から、中分類 02 の費用の歪度が－1.086 でやや歪んでおり、中分類 50 と 84 の費用の尖度が、それぞれ、5.089、4.822 でやや尖っているが、全体的にはどの中分類においても、歪度はほぼ 0 であり尖度はほぼ 3 である。よって、費用を自然対数に変換することで、周辺分布において正規分布を近似していることが分かる。

表 7.1 と表 7.2 に示したデータを用い、費用を補助変数として確定的単一代入法、確率的単一代入法、比率補定及び多重代入法により産業中分類別に売上の補定を行った。また、実務上の演算時間を考慮し、多重代入法の補定済みデータ数は 20 で十分だと判断した[100]。産業中分類別に売上の補定値を合計し、多重代入法による補定値については補定値合計の平均・標準偏差・変動係数・信頼区間を算出した。

---

99 ここでは表を掲載可能な大きさにする都合から、産業中分類を符号で表している。産業中分類の名前については以下の「平成 24 年経済センサス‐活動調査 産業分類一覧」を参照されたい。
http://www.stat.go.jp/data/e-census/2012/kakuho/bunrui.htm （2015 年 6 月 1 日閲覧）

100 「付録 2 多重代入法による補定済みデータ数」も参照されたい。

表 7.1：完全データ 売上（対数変換） 基本統計量

中分類	第1四分位	中央値	平均	第3四分位	標準偏差	歪度	尖度
01	7.382	8.281	8.227	9.125	1.427	-0.338	4.127
02	7.202	8.443	7.994	9.263	1.914	-1.145	4.245
03	8.253	9.024	8.979	9.890	1.418	-0.582	4.553
04	7.549	8.572	8.474	9.545	1.561	-0.436	3.452
39	6.875	8.102	8.096	9.306	1.799	-0.139	3.554
40	6.299	7.604	7.665	9.059	2.218	-0.011	3.541
50	8.259	9.253	9.263	10.203	1.501	0.519	5.248
51	7.448	8.563	8.555	9.729	1.689	-0.182	3.244
52	7.678	8.815	8.892	10.100	1.818	0.056	3.272
53	7.755	8.882	8.856	9.989	1.718	-0.118	3.546
54	7.819	8.838	8.822	9.857	1.586	-0.115	3.836
55	7.256	8.371	8.382	9.514	1.717	-0.026	3.508
57	5.704	6.659	6.548	7.496	1.438	-0.367	3.721
58	6.201	7.151	7.226	8.288	1.673	-0.150	3.127
59	6.782	7.741	7.663	8.630	1.443	-0.327	3.697
60	6.174	7.244	7.243	8.473	1.676	-0.280	3.159
61	5.704	6.824	6.956	8.187	1.871	0.117	3.243
68	6.397	7.440	7.523	8.648	1.748	0.023	3.580
69	5.714	6.685	6.676	7.601	1.460	0.115	3.606
70	6.492	7.650	7.630	8.846	1.778	-0.121	3.475
71	6.511	7.790	7.859	9.265	2.180	0.003	3.325
72	6.460	7.375	7.252	8.169	1.439	-0.532	4.816
73	7.556	8.689	8.634	9.785	1.735	-0.213	3.728
74	6.397	7.243	7.263	8.189	1.412	-0.143	3.897
75	6.009	6.963	7.088	8.162	1.729	0.082	3.422
76	5.858	6.576	6.537	7.244	1.156	-0.221	4.096
77	6.277	7.208	7.167	8.143	1.499	-0.263	3.843
78	5.030	5.784	5.772	6.503	1.200	0.006	4.201
79	5.298	6.449	6.557	7.788	1.924	0.152	3.050
80	5.886	7.093	7.427	8.863	2.156	0.322	2.797
82	4.394	5.485	5.500	6.518	1.620	0.243	3.497
83	6.879	8.173	7.871	8.981	1.652	-0.556	3.721
84	6.046	7.433	7.723	9.670	2.338	-0.025	2.451
85	7.601	8.639	8.506	9.413	1.534	-0.219	4.906
87	6.602	7.759	7.823	9.075	1.891	-0.115	3.523
88	7.780	8.706	8.639	9.582	1.409	-0.376	3.836
89	6.600	7.399	7.399	8.268	1.244	-0.187	3.476
90	5.914	6.908	6.976	8.119	1.674	-0.143	3.457
91	7.676	8.854	8.666	9.788	1.576	-0.583	3.681
92	6.873	8.013	7.955	9.063	1.667	-0.238	3.624
95	6.381	7.974	7.840	9.376	2.131	-0.117	2.810
@Z	7.380	8.426	8.372	9.460	1.624	-0.382	3.891
G1	6.739	7.648	7.744	8.793	1.682	-0.141	3.708
G2	6.276	7.317	7.350	8.527	1.843	-0.226	3.751
I1	7.591	8.700	8.694	9.908	1.753	-0.251	3.330
I2	5.886	6.908	6.940	7.953	1.644	-0.028	3.585
K1	6.215	7.028	7.061	7.875	1.334	0.203	3.615
LZ	6.217	7.155	7.256	8.353	1.642	-0.034	3.706
M2	5.991	6.802	6.870	7.740	1.407	0.037	3.643
NZ	5.298	6.358	6.389	7.449	1.774	0.083	3.368
PZ	6.601	7.637	7.509	8.422	1.426	-0.296	2.986
R1	6.040	6.952	7.025	8.042	1.700	0.061	3.682
R2	6.330	7.438	7.365	8.484	1.676	-0.263	3.457

表 7.2：完全データ 費用（対数変換） 基本統計量

中分類	第1四分位	中央値	平均	第3四分位	標準偏差	歪度	尖度
01	7.345	8.243	8.190	9.090	1.416	-0.289	3.922
02	7.153	8.343	7.938	9.203	1.885	-1.086	4.057
03	8.216	8.987	8.950	9.824	1.374	-0.491	4.353
04	7.485	8.567	8.443	9.478	1.531	-0.384	3.424
39	6.876	8.088	8.092	9.286	1.759	-0.043	3.337
40	6.310	7.576	7.682	8.991	2.092	0.185	3.393
50	8.186	9.231	9.214	10.169	1.530	0.452	5.089
51	7.325	8.492	8.462	9.680	1.743	-0.215	3.201
52	7.496	8.731	8.756	10.056	1.927	-0.056	3.224
53	7.601	8.810	8.739	9.941	1.801	-0.199	3.455
54	7.669	8.761	8.731	9.810	1.631	-0.107	3.600
55	7.090	8.293	8.267	9.469	1.789	-0.071	3.382
57	5.375	6.397	6.320	7.337	1.522	-0.335	3.633
58	5.858	6.908	6.968	8.127	1.785	-0.187	3.091
59	6.450	7.523	7.428	8.517	1.587	-0.366	3.541
60	5.852	7.009	7.015	8.341	1.794	-0.278	3.068
61	5.455	6.633	6.748	8.084	1.990	0.054	3.163
68	6.234	7.327	7.408	8.537	1.749	0.100	3.378
69	5.106	6.372	6.207	7.428	1.817	-0.343	3.405
70	6.312	7.550	7.506	8.783	1.841	-0.164	3.333
71	6.480	7.817	7.886	9.273	2.114	0.124	3.141
72	6.023	7.010	6.947	7.926	1.473	-0.256	3.962
73	7.477	8.655	8.598	9.760	1.724	-0.115	3.302
74	6.045	7.065	7.058	8.096	1.528	-0.128	3.433
75	5.793	6.819	6.921	8.074	1.807	0.029	3.425
76	5.517	6.254	6.248	6.987	1.230	-0.140	3.996
77	5.914	6.918	6.904	7.962	1.592	-0.183	3.550
78	4.344	5.209	5.195	6.082	1.451	-0.077	3.947
79	4.804	6.204	6.220	7.678	2.178	-0.067	2.980
80	5.642	6.994	7.262	8.831	2.268	0.240	2.697
82	3.829	5.094	5.040	6.261	1.904	0.035	3.221
83	6.397	7.777	7.484	8.704	1.821	-0.568	3.745
84	5.946	7.429	7.681	9.551	2.292	0.070	2.297
85	7.550	8.558	8.441	9.336	1.500	-0.131	4.822
87	6.659	7.744	7.829	8.989	1.772	0.121	3.191
88	7.632	8.631	8.521	9.518	1.475	-0.473	3.775
89	6.215	7.115	7.109	8.115	1.411	-0.258	3.341
90	5.398	6.676	6.652	8.014	1.890	-0.219	3.179
91	7.650	8.833	8.630	9.761	1.592	-0.649	3.884
92	6.774	7.963	7.853	9.026	1.749	-0.388	3.801
95	6.351	7.901	7.814	9.343	2.120	-0.138	2.916
@Z	7.313	8.269	8.269	9.324	1.646	-0.366	3.979
G1	6.598	7.603	7.622	8.740	1.724	-0.084	3.342
G2	6.082	7.249	7.264	8.488	1.828	-0.081	3.481
I1	7.244	8.505	8.476	9.781	1.870	-0.273	3.183
I2	5.438	6.567	6.591	7.734	1.807	-0.004	3.174
K1	5.712	6.767	6.702	7.709	1.594	-0.205	3.688
LZ	5.919	7.024	7.086	8.254	1.717	0.001	3.475
M2	5.635	6.544	6.572	7.509	1.543	-0.013	3.724
NZ	4.736	6.045	6.011	7.273	2.028	-0.091	3.197
PZ	6.351	7.577	7.366	8.351	1.536	-0.523	3.441
R1	6.021	6.908	7.004	7.935	1.664	0.134	3.652
R2	5.986	7.242	7.122	8.384	1.845	-0.349	3.373

例として、産業中分類 01、02 の完全データについて、売上（自然対数変換）と費用（自然対数変換）のヒストグラムを示す（図 7.1～7.4）。産業 01 は歪みが少なく、産業 02 は多少左に歪んでいるが、おおむね山なりの分布をしている。

図 7.1：産業 01 完全データ 売上（対数変換）

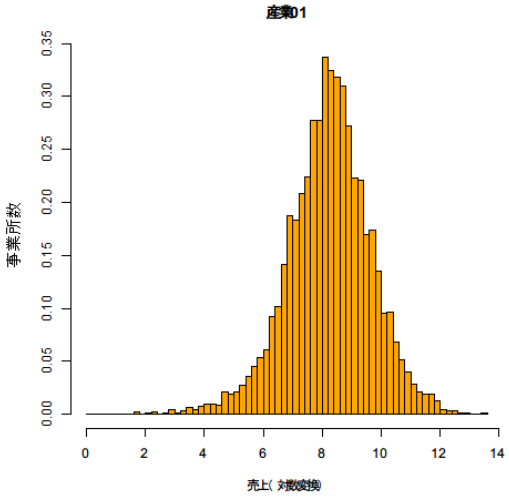


図 7.2：産業 01 完全データ 費用（対数変換）

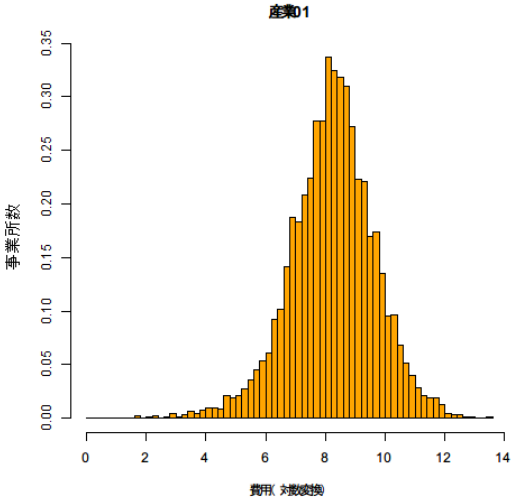


図 7.3：産業 02 完全データ 売上（対数変換）

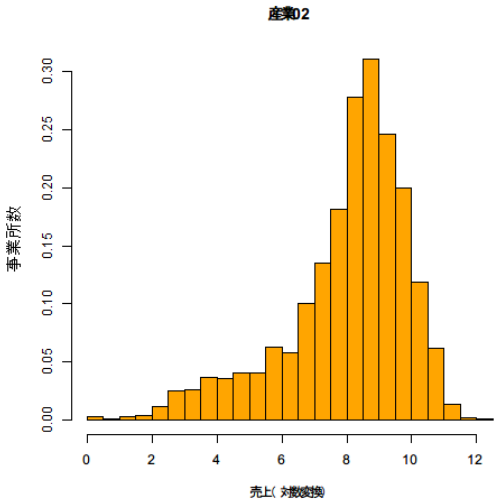
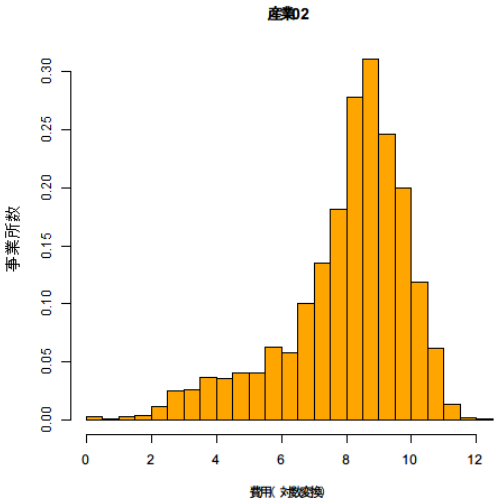


図 7.4：産業 02 完全データ 費用（対数変換）



7.2 検証の結果

補定方法の違いにより売上補定値の合計が変化する。単一代入法による売上補定値の合計を産業中分類別にまとめたものが表 7.3 である。表 7.3 を見ると、補定方法別の補定値合計はおおむね同じ大きさとなっており、著しく外れた方法はない。しかし単一代入法では、補定値を一つしか出力しないため、その補定値の信頼性を評価することができない。

一方、多重代入法では補定済みデータ数が複数（ここでは前述のとおり 20）個出力されるが、補定済みデータごとに売上補定値の合計を計算し、補定値合計の平均・標準偏差・変動係数・95%信頼区間を産業別にまとめたものが表 7.4 である。多重代入法は補定値を複数作成するため、その信頼性を評価することができる。表 7.4 を見ると、検証の対象とした 53 の産業中分類のうち、変動係数が 0.150 未満となった産業中分類は 20 あった[101]。これら変動係数が小さい産業中分類については、補定値の信頼性が高いと考えられる。

変動係数が大きい産業中分類について、該当する産業中分類に含まれる事業所のデータを確認したところ、外れ値が存在していることが分かった。後述するように、そういった事業所を補定対象から外すことで、変動係数を小さくすることができる。

このような事業所については、費用記入値が正しいとしても、この事業所の売上を推計モデルにより補定することは望ましくない。モデルのパラメータが少し変化しただけでも、補定値は大きく変動し、補定値の信頼性は低下するからである。結果的に集計表の品質が低下することとなるため、このような事業所は推計モデルによらない補定を検討すべきである。

事前に全産業の散布図を描くなどして、推計モデルによる補定を行うことで集計表の品質が低下する可能性のある事業所を確認することは、実務上、現実的ではない。そこで、推計モデルによる補定と補定値の誤差の情報を同時に得つつ、誤差が大きく補定値の信頼度が低い場合は別の方法で補定を行い、集計表の品質が大きく低下しないようにすることが現実的だと考える。しかしながら、単一代入法では誤差の情報を得ることはできない。

多重代入法は複数のモデルにより補定値を複数（今回の検証では 20）作成するため、補定を行いながら補定された売上の合計に期待される値の分布を推定することができる。また他の手法による補定を行う場合でも、予め多重代入法で補定値の分布を推定しておき、推計モデルによる補定から除くべき事業所を明らかにしておくことで、集計表の品質を向上させることができる。

また多重代入法による複数のデータセットを活用することにより、集計表レベルでの補定の影響を評価し、更なる品質向上を目指すことが可能となる。

---

101 変動係数の活用方法については、本稿 5.2 節「ドイツにおける欠測値補定」も参照されたい。

表 7.3：単一代入法による売上高補定値の合計

産業 中分類	確定的単一代入法 による売上補定値合計	確率的単一代入法 による売上補定値合計	比率補定 による売上補定値合計
01	153,835	178,229	141,280
02	52,188	44,053	46,952
03	8,601	10,195	8,221
04	58,261	80,957	55,691
39	700,841	742,195	670,975
40	9,160	7,933	8,330
50	660	606	541
51	483,872	476,326	437,651
52	595,923	597,054	511,197
53	1,326,533	1,305,005	1,308,991
54	798,979	690,496	733,143
55	1,125,322	1,127,524	998,297
57	262,833	151,794	322,931
58	806,877	600,910	803,161
59	336,762	270,676	311,309
60	1,816,535	1,759,317	2,250,341
61	46,446	39,681	36,842
68	326,144	258,691	246,307
69	675,942	659,446	544,723
70	66,248	54,808	57,733
71	167,386	138,534	151,183
72	684,594	654,523	689,201
73	16,740	13,601	15,510
74	359,800	413,185	397,982
75	96,977	101,593	90,273
76	4,284,079	4,179,100	6,206,900
77	40,904	38,250	34,482
78	313,319	295,437	276,342
79	197,534	203,155	205,989
80	88,357	66,027	71,371
82	170,809	143,820	282,306
83	729,525	724,617	620,379
84	1,596	1,604	1,406
85	554,336	497,592	537,907
87	3,259	3,128	3,168
88	128,676	105,029	133,817
89	311,633	314,098	396,040
90	50,379	42,678	44,876
91	36,530	30,371	32,184
92	137,540	139,818	115,919
95	662	584	445
@Z	62,072	56,055	53,925
G1	140,648	199,093	140,244
G2	33,252	29,179	26,012
I1	7,672,792	6,661,721	6,077,359
I2	1,437,430	1,337,379	1,345,113
K1	47,261	53,141	41,758
LZ	704,156	551,849	584,635
M2	74,290	110,831	70,317
NZ	75,590	57,136	66,904
PZ	21,341	18,644	18,932
R1	13,530	11,229	12,289
R2	1,496,136	1,412,228	1,570,888

表 7.4：多重代入法による売上高補定値合計の評価

産業中分類	多重代入法による売上補定値合計 平均	多重代入法による売上補定値合計 標準偏差	変動係数	多重代入法 95%信頼区間	
01	147,806	16,411.8	0.111	115,639	179,973
02	48,337	17,980.7	0.372	13,095	83,580
03	8,485	2,741.5	0.323	3,112	13,859
04	61,902	19,911.4	0.322	22,876	100,929
39	702,465	169,034.6	0.241	371,157	1,033,773
40	8,935	1,863.2	0.209	5,283	12,587
50	638	160.3	0.251	324	952
51	459,302	81,624.3	0.178	299,318	619,285
52	580,714	93,977.2	0.162	396,518	764,909
53	1,227,053	176,556.5	0.144	881,003	1,573,104
54	752,331	44,268.6	0.059	665,564	839,097
55	1,031,589	71,925.9	0.070	890,615	1,172,564
57	274,280	60,716.3	0.221	155,276	393,284
58	713,635	118,026.9	0.165	482,302	944,968
59	312,606	34,833.3	0.111	244,333	380,879
60	1,730,633	347,009.5	0.201	1,050,494	2,410,771
61	45,339	10,998.0	0.243	23,783	66,895
68	295,043	21,242.1	0.072	253,409	336,678
69	645,186	33,304.5	0.052	579,909	710,463
70	62,444	7,628.5	0.122	47,492	77,396
71	164,342	36,700.6	0.223	92,409	236,275
72	680,629	77,977.9	0.115	527,792	833,466
73	15,607	1,785.0	0.114	12,109	19,106
74	353,878	66,939.0	0.189	222,677	485,078
75	103,503	30,104.4	0.291	44,499	162,508
76	4,267,860	274,337.4	0.064	3,730,158	4,805,561
77	38,177	5,756.8	0.151	26,894	49,460
78	309,047	17,840.4	0.058	274,079	344,014
79	178,597	31,884.1	0.179	116,104	241,090
80	87,780	14,963.3	0.170	58,451	117,108
82	167,727	42,205.8	0.252	85,004	250,451
83	691,211	64,738.3	0.094	564,324	818,098
84	1,652	859.0	0.520	-31	3,336
85	552,744	57,330.3	0.104	440,376	665,111
87	3,606	851.6	0.236	1,937	5,275
88	136,105	26,909.1	0.198	83,363	188,847
89	295,794	47,706.0	0.161	202,290	389,297
90	50,822	7,338.3	0.144	36,439	65,205
91	34,813	4,192.6	0.120	26,595	43,030
92	130,260	13,807.3	0.106	103,198	157,322
95	564	158.7	0.281	253	875
@Z	60,306	15,982.8	0.265	28,980	91,633
G1	139,081	32,281.4	0.232	75,810	202,353
G2	31,712	8,395.6	0.265	15,257	48,167
I1	6,605,669	434,094.3	0.066	5,754,845	7,456,494
I2	1,375,337	88,069.6	0.064	1,202,721	1,547,954
K1	41,591	7,679.2	0.185	26,540	56,642
LZ	609,464	106,209.1	0.174	401,294	817,633
M2	65,906	12,054.6	0.183	42,279	89,533
NZ	70,543	23,952.0	0.340	23,597	117,488
PZ	21,212	5,266.1	0.248	10,890	31,533
R1	12,992	1,977.0	0.152	9,118	16,867
R2	1,524,680	117,107.9	0.077	1,295,149	1,754,212

8　結語と今後の展望

統計センター統計技術研究課では、経済センサス‐活動調査の創設を機に経理項目の欠測値補定における統計学的処理方法を検討した。本稿では、欠測のメカニズムに始まり、欠測値補定方法の中でも特に我が国の公的統計において利用されていない多重代入法について、単一代入法との相違を中心に簡単な例を使用して手法を説明した。多重代入法は諸外国において研究が進められており、諸外国の公的統計における多重代入法の適用例とともに経済センサス‐活動調査の実データに利用した場合の結果を記載した。特に、補定値はあくまで推定値という観点から、誤差を有するという点を意識し、補定に起因する誤差評価を行えることが多重代入法の利点であることを示した。

定義上、欠測値に対応する真値は常に不明であるため、従来は欠測値補定の精度評価は無視される傾向があった。しかし、Abayomi *et al.* (2008)により欠測値補定の間接的な診断手法が提唱されて以来、補定モデルや補定値の妥当性を検証することが推奨され始め、各種の診断手法が提唱されてきている(Honaker *et al.*, 2011, pp.25-35)。

「公的統計の品質保証に関するガイドライン」[102]（各府省統計主管課長等会議申合せ、平成 23 年 4 月 8 日改定）によると、「公的統計の品質に関する自己評価結果や客観的な評価結果の活用を通じた公的統計の見直し・効率化を推進する」(p.1)とあり、「公的統計の品質表示事項」として「誤差の範囲等の結果精度に関する情報（回収率、有効回答率及びその計算方法等）」と指摘されている(p.6)。多重代入法による集計結果に対する欠測値の影響の数値評価は、この目的に合致していると言える。

さらに、欠測値補定以外にも、応用方法として、混淆正規分布モデルなどによって検出したエラーデータを除去し多重代入法により補定する自動エディティングの手法(Takahashi, 2014a)として使用したり、公開ミクロデータにおける合成データ生成手法として多重代入法を使用(Reiter, 2009)したりすることが考えられる。

コラム 1 で述べたとおり、多重代入法はベイズ統計学の枠組みで構築されたものであり、事前情報と尤度を合算させ事後分布を構築することでその真価が発揮される。本稿では、4.6 節にて事前分布の利用方法について簡単に触れた。日本の経済センサスは、まだ始まったばかりである。今後、データが蓄積していくにしたがって、本稿で提唱している多重代入法による欠測誤差の評価も、有益に行うことができるようになるであろう。

よって、今後の展望としては、調査を重ねてデータが蓄積され、検証に使用できるデータが拡充した上で、さらなる検証を行い、実務への導入につながっていくことを期待する。その際に、欠測値補定の手引書として、本稿が資するものとなれば幸いである。

---

[102] 下記のウェブサイトにて、pdf 版を閲覧できる。http://www.stat.go.jp/index/seido/pdf/3-4.pdf（2015 年 6 月 1 日閲覧）

参考文献（英語）

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

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

[3] Andridge, Rebecca R. and Roderick J. A. Little. (2010). “A Review of Hot Deck Imputation for Survey Non-response,” *International Statistical Review* vol.78, no.1, pp.40-64.

[4] Baraldi, Amanda N. and Craig K. Enders. (2010). “An Introduction to Modern Missing Data Analyses,” *Journal of School Psychology* vol.48, no.1, pp.5-37.

[5] Bechtel, Laura, Yarissa Gonzalez, Matthew Nelson, and Roberta Gibson. (2011). “Assessing Several Hot Deck Imputation Methods Using Simulated Data from Several Economic Programs,” *Proceedings of the Survey Research Methods Section*, American Statistical Association, pp.5022-5036.

[6] Blackwell, Matthew, James Honaker, and Gary King. (2015). “A Unified Apraoch to Measurement Error and Missing Data: Details and Extensions,” *Sociological Methods and Research*, forthcoming.

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

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

[9] Carpenter, James R. and Michael G. Kenward. (2013). *Multiple Imputation and its Application*. A John Wiley & Sons Publication.

[10] De Bin, Riccardo, Silke Janitza, Willi Sauerbrei, and Anne-Laure Boulesteix. (2014). “Subsampling Versus Bootstrapping in Resampling-Based Model Selection for Multivariable Regression,” *Technical Report* no.171 (Department of Statistics, University of Munich).

[11] DeGroot, Morris H. and Mark J. Schervish. (2002). *Probability and Statistics*, 3rd edition. Addison-Wesley.

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

[13] Do, Chuong B. and Serafim Batzoglou. (2008). “What is the Expectation Maximization Algorithm?,” *Nature Biotechnology* vol.26, no.8, pp.897-899.

[14] Donders, A. Rogier T., Geert J. M. G. van der Heijden, Theo Stijnen, and Karel G. M. Moons. (2006). “Review: A Gentle Introduction to Imputation of Missing Values,” *Journal of Clinical Epidemiology* vol.59, pp.1087-1091.

[15] Fox, John. (1991). *Regression Diagnostics*. Sage Publications.

[16] García, María, Chandra Erdman, and Ben Klemens. (2014). “Multiple Imputation Methods for Imputing Earnings in the Survey of Income and Program Participation,” *Work Session on Statistical Data Editing, United Nations Economic Commission for Europe*, Paris, France, April 28-30 2014.
Available online at: http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2014/mtg1/Topic_2_USA_Garcia.pdf

[17] Graham, John W., Allison E. Olchowski, and Tamika D. Gilreath. (2007). “How Many Imputations are Really Needed? Some Practical Clarifications of Multiple Imputation Theory,” *Prevention Science* vol.8, no.3, pp.206-213.

[18] Greene, William H. (2003). *Econometric Analysis*, fifth edition. Prentice Hall.

[19] Gujarati, Damodar N. (2003). *Basic Econometrics*, fourth edition. McGraw-Hill.

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

[21] Honaker, James, Gary King, and Matthew Blackwell. (2011). “Amelia II: A Program for Missing Data,” *Journal of Statistical Software* vol.45, no.7, pp.1-47.

[22] Horowitz, Joel L. (2001). “The Bootstrap,” in *Handbook of Econometrics*, vol.5, edited by James J. Heckman and Edward Leamer. Elsevier.

[23] Hu, Ming-xiu, Sameena Salvucci, and Ralph Lee. (2001). *A Study of Imputation Algorithms.* Working Paper No. 2001–17. U.S. Department of Education. National Center for Education Statistics.
Available online at: http://nces.ed.gov/pubs2001/200117.pdf

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

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

[26] Liang, Hua, Haiyan Su, and Guohua Zou. (2008). “Confidence Intervals for A Common Mean with Missing Data with Applications in AIDS Study,” *Computational Statistics & Data Analysis* vol.53, no.2, pp.546-553.

[27] Little, Roderick J. A. and Donald B. Rubin. (2002). *Statistical Analysis with Missing Data*, second edition. John Wiley & Sons.

[28] Long, J. Scott. (1997). *Regression Models for Categorical and Limited Dependent Variables*. Sage Publications.

[29] Marshall, Andrea, Douglas G. Altman, Roger L. Holder, and Patrick Royston. (2009). “Combining Estimates of Interest in Prognostic Modelling Studies after Multiple Imputation: Current Practice and Guidelines,” *BMC Medical Research Methodology* vol.9, no.57.

[30] National Center for Health Statistics. (2013). “Multiple Imputation of Family Income and Personal Earnings in the National Health Interview Survey: Methods and Examples”
Available online at: http://www.cdc.gov/nchs/data/nhis/tecdoc12.pdf

[31] Office for National Statistics. (2014). “Change to Imputation Method used for the Turnover Question in Monthly Business Surveys,” Guidance and Methodology: Retail Sales.
Available online at: http://www.ons.gov.uk/ons/guide-method/method-quality/specific/economy/retail-sales/index.html

[32] Politis, Dimitris N. (1998). “Computer-Intensive Methods in Statistical Analysis,” *Signal Processing Magazine, IEEE* vol.15, no.1, pp.39-55.

[33] Reiter, Jerome P. (2009). “Multiple Imputation for Disclosure Limitation: Future Research and Challenges,” *Journal of Privacy and Confidentiality* vol.1, no.2, pp.223-233.

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

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

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

[37] Seaman, Shaun, John Galati, Dan Jackson, and John Carlin. (2013). “What Is Meant by ‘Missing at Random’?,” *Statistical Science* vol.28, no.2, pp.257-268.

[38] Shao, Jun. (2000). “Cold Deck and Ratio Imputation,” *Survey Methodology* vol.26, no.1, pp79-85.

[39] Shao, Jun and Dongsheng Tu. (1995). *The Jackknife and Bootstrap*. Springer.

[40] Shapiro, Samuel S. and Martin B. Wilk. (1965). “An Analysis of Variance Test for Normality (Complete Samples),” *Biometrika* vol.52, no.3/4, pp.591-611.

[41] Spies, Lydia, Sven Schmiedel, and Katrin Schmidt. (2014). “Simulating Multiple Imputation of Water Consumption in the German Agricultural Census 2010,” *Work Session on Statistical Data Editing, United Nations Economic Commission for Europe*, Paris, France, April 28-30 2014.

Available online at: http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2014/mtg1/Topic_2_Germany.pdf

[42] Swiss Federal Statistical Office. (2014). “Quality Report, v1: Swiss SILC Survey, 2013”
Available online at: http://www.bfs.admin.ch/bfs/portal/en/index/themen/20/22/lexi.Document.189555.pdf

[43] 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.
Available online at: http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2012/35_Japan.pdf

[44] Takahashi, Masayoshi and Takayuki Ito. (2013). “Multiple Imputation of Missing Values in Economic Surveys: Comparison of Competing Algorithms,” *Proceedings of The 59th World Statistics Congress of the International Statistical Institute (ISI)*, Hong Kong, China, pp.3240-3245.

[45] Takahashi, Masayoshi. (2014a). “An Assessment of Automatic Editing via the Contamination Model and Multiple Imputation,” *Work Session on Statistical Data Editing, United Nations Economic Commission for Europe*, Paris, France.
Available online at: http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2014/mtg1/Topic_1_Japan.pdf

[46] Takahashi, Masayoshi. (2014b). “Diagnosing the Imputation of Missing Values in Official Economic Statistics via Multiple Imputation: Unveiling the Invisible Missing Values,” *International Association for Official Statistics (IAOS) Conference 2014*, Da Nang, Vietnam.
Available online at: https://iaos2014.gso.gov.vn/document/taka1.p1.v1.pdf

[47] Thompson, Katherine J. and Quatracia Williams. (2003). “Developing Imputation Models for the Services Sectors Portion of the Economic Census,” *Federal Committee on Statistical Methodology Research Conference*, November 17-19, 2003, Arlington, Virginia, U.S.A.
Available online at: https://fcsm.sites.usa.gov/files/2014/05/2003FCSM_Thompson_Williams.pdf

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

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

pp.1-67.

[50] Wooldridge, Jeffrey M. (2009). *Introductory Econometrics: A Modern Approach*, 4th edition. South-Western.

参考文献（日本語）

[51] 青木繁伸. (2009).『R による統計解析』, オーム社.

[52] 伊藤孝之, 野呂竜夫, 阿部穂日, 土井満喜. (2012). 「平成 24 年経済センサス‐活動調査の経理項目補定方法の研究【第 1 部】売上（収入）金額の補定方法と外れ値除外方法の比較・検討」, 『製表技術参考資料』no.18, pp.1-69.

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

[54] 川崎茂. (2010).「コンピュータの半世紀 ― 国勢調査を支える情報技術」,『統計 Today』no.27. http://www.stat.go.jp/info/today/027.htm

[55] 熊原啓作, 渡辺美智子. (2012).『身近な統計』, 一般財団法人放送大学教育振興会.

[56] 小西貞則, 越智義道, 大森裕浩. (2008).『計算統計学の方法：ブートストラップ・EM アルゴリズム・MCMC』, 朝倉書店.

[57] 迫田宇広, 高橋将宜, 渡辺美智子. (2014).『問題解決力向上のための統計学基礎：Excel によるデータサイエンススキル』, 一般財団法人日本統計協会.

[58] 総務省統計局. (2013).「家計調査年報（家計収支編）平成 25 年（2013 年）家計調査の概要」. http://www.stat.go.jp/data/kakei/2013np/gaiyou.htm

[59] 高橋将宜. (2014).「公的統計における多重代入法の利活用方法の可能性～諸外国における適用を例に～」, 『経済統計学会　第 58 回（2014 年度）全国研究大会報告要旨集』, pp.81-82.

[60] 高橋将宜. (2015).「欠測値補定における非標本誤差の数値評価」, 第 130 回研究報告会（総務省統計研修所）.

[61] 高橋将宜, 伊藤孝之. (2012a).「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～」, 『2012 年度統計関連学会連合大会講演報告集』, p.174.

[62] 高橋将宜, 伊藤孝之. (2012b).「経済調査における経理項目の欠測値補定方法～EMB アルゴリズムによる多重代入法～」, 『2012 年度科学研究費シンポジウム講演予稿集～統計科学の基礎的理論とその応用～』, pp.1-10.

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

[64] 高橋将宜, 伊藤孝之. (2013b).「様々な多重代入法アルゴリズムの比較」, 『2013 年度統計関連学会連合大会講演報告集』, p.42.

[65] 高橋将宜, 伊藤孝之. (2013c). 「大規模経済系データにおける様々な多重代入法アルゴリズムの検証」, 2013 年度科学研究費シンポジウム～統計科学の新展開～（金沢大学）.

http://stat.w3.kanazawa-u.ac.jp/sympo/30930.pdf
[66] 高橋将宜, 伊藤孝之. (2014).「様々な多重代入法アルゴリズムの比較～大規模経済系データを用いた分析～」,『統計研究彙報』第 71 号, no.3, pp.39-82.
[67] 土屋隆裕. (2009).『概説　標本調査法』, 朝倉書店.
[68] 野間久史, 田中司朗, 田中佐智子, 和泉志津恵. (2012).「Multiple Imputation 法によるネステッドケースコントロール研究, ケースコホート研究の解析」, 『計量生物学』vol.33, no.2, pp.101-124.
[69] 羽渕達志, 上田聖, 小高敦, 高橋将宜, 小泉英希. (2012).「平成 24 年度米国センサス局出張報告」, 独立行政法人統計センター.
[70] 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』, 岩波書店.
[71] 松田芳郎, 伴金美, 美添泰人. (2000). 『ミクロ統計の集計解析と技法』, 講座ミクロ統計分析第 2 巻, 日本評論社.
[72] 渡辺美智子, 山口和範. (2000). 『EM アルゴリズムと不完全データの諸問題』, 多賀出版.

参考文献（ソフトウェアマニュアル）

[73] Fox, John. (2015). *Package 'Norm'*.
Available online at: http://cran.r-project.org/web/packages/norm/norm.pdf
[74] Honaker, James, Gary King, and Matthew Blackwell. (2015). *Package 'Amelia'*.
Available online at: http://cran.r-project.org/web/packages/Amelia/Amelia.pdf
[75] Jarek, Slawomir. (2015). *Package 'mvnormtest'*.
Available online at: http://cran.r-project.org/web/packages/mvnormtest/mvnormtest.pdf
[76] SAS Institute Inc. (2011). *SAS/STAT 9.3 User's Guide*. SAS Institute Inc.
Available online at: http://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/viewer.htm
[77] SPSS Inc. (2009). *PASW Missing Values 18*. SPSS Inc.
Available online at: http://www.unt.edu/rss/class/Jon/SPSS_SC/Manuals/v18/PASW Missing Values 18.pdf
[78] Statistical Solutions. (2011). *SOLAS Version 4.0 Imputation User Manual*.
Available online at: http://www.solasmissingdata.com/wp-content/uploads/2011/05/Solas-4-Manual.pdf
[79] van Buuren, Stef and Karin Groothuis-Oudshoorn. (2015). *Package 'mice'*.
Available online at: http://cran.r-project.org/web/packages/mice/mice.pdf

注：すべてのウェブサイトは、2015 年 6 月 1 日現在において閲覧確認したものである。

付録 1 本稿で用いた記号

本研究で用いた記号は、以下のとおりである。$\mathbf{D}$を$n \times p$のデータセットとする。nは標本サイズであり、pは変数の数である。もしデータが欠測していなければ、$\mathbf{D}$は平均値ベクトル$\boldsymbol{\mu}$と分散共分散行列$\boldsymbol{\Sigma}$で多変量正規分布しているとする。つまり、$\mathbf{D} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$である。

iを観測値のインデックスとし、$i = 1, \ldots, n$とする。jを変数のインデックスとし、$j = 1, \ldots, p$とする。$\mathbf{D} = \{Y_1, \ldots, Y_p\}$とし、$\mathbf{Y_j}$は$\mathbf{D}$の$j$番目の列とし、$\mathbf{Y_{-j}}$は$\mathbf{Y_j}$の補集合とする（$\mathbf{Y_{-j}}$は$\mathbf{D}$内の$\mathbf{Y_j}$以外のすべての列である）。特に、本稿では、簡単のため、二変数間の補定について記述しているので、Y_1は補定の対象とする欠測変数（補定対象変数）Yとし、Y_2は補助変数Xとしている。よって、$\mathbf{D} = \{Y_i, X_i\}$である。

$\mathbf{K}$を回答指示行列とする。$\mathbf{D}$と$\mathbf{K}$の次元は同じである。$\mathbf{D}$が観測されるとき$\mathbf{K} = 1$であり、$\mathbf{D}$が観測されないとき$\mathbf{K} = 0$である。また、$\mathbf{D_{obs}}$を観測データとし、$\mathbf{D_{mis}}$を欠測データとする。つまり、$\mathbf{D} = \{\mathbf{D_{obs}}, \mathbf{D_{mis}}\}$である。

βは母集団における回帰パラメータであり、$\hat{\beta}$は実測データにおける最小二乗法による回帰推定値である。また、ωは比率による傾きのパラメータを表し、$\hat{\omega}$は実測データにおける比率による傾きを表す。$\tilde{\beta}$は多重代入法による回帰推定値である。

付録 2 多重代入法による補定済みデータ数

多重代入法は乱数を利用したシミュレーション手法の一種である。従来、Rubin (1987, p.114)によれば、多重代入法による補定済みデータ数（M数）は、5〜10 程度で十分だとされてきた。実際、通常のシミュレーションでは全データをシミュレーション値として生成するため、全情報が欠測していると考えることができ、補定では観測値をシミュレーション値に置き換える必要はなく、データ内の一部のみが欠測しているため、繰り返し回数が少なくてもよいと考えることができる(Honaker and King, 2010)。

かつて、コンピュータの性能が低かった時代には、多重代入済みデータ数（M数）を少なくすることに大きな利益があった。しかし、現代のコンピュータでは、大量の演算を高速に行うことが可能となっている（コラム 2 参照）。あらゆるシミュレーション手法と同様に、多重代入法による補定済みデータ数（M数）も多ければ多いほどよい。とは言え、Mのサイズが大きくなればなるほど、演算時間がかかり、データの容量も大きくなる事実自体は変わらない。つまり、「十分な回数での繰り返しを行うことが望ましい」(野間, 田中, 田中, 和泉, 2012, p.107)が、実務上、Mのサイズをいくつに設定すれば「十分な回数」であるかという問題が生じる。よって、本付録では、経済センサス‐活動調査のデータを元にしたシミュレーションデータを用いて、下記の条件にて補定済みデータ数の簡易的な検証を行う[103]。

多変量正規乱数を用いたシミュレーションデータにおいて、売上高を模した変数に MCAR で欠測を発生させ、費用を模した変数を用いて多重代入法を行った。欠測率は 10%、20%、40%の 3 種類である。データの観測数は、100、1,000、10,000、100,000、500,000 の 5 種類を用意した。使用したコンピュータのスペックは、以下のとおり、ごくありふれたコンピュータである。プロセッサ：Intel® Xeon® CPU E5-2690 v2 @ 3.00GHz 3.00 GHz (2 プロセッサ)；実装メモリ(RAM)：4.00 GB；システムの種類：64 ビットオペレーティングシステム。より高性能なコンピュータを用いることで、多重代入に必要な時間を削減することができるのは、言うまでもない。

表 A.2.1 は、補定済みデータ数（M数）を 5、10、20、100、1000 にした場合に多重代入を行うのにどれだけの時間がかかるかを示したものである。なお、下記の時間は、純粋に多重代入を行う時間のみであり、R を起動させたり、データを読み込んだり、Amelia を起動したりする時間は含んでいない。

標本サイズ 100 のデータにおいて、M = 100 の多重代入を行うのに必要な時間は、わずか 1 秒未満である。M= 1000 の多重代入を行うのに必要な時間は、10 秒未満である。標本サイズ 1,000 のデータにおいて、M = 20 の多重代入を行うのに必要な時間は、わずか 1 秒未満である。M = 1000 の多重代入を行うのに必要な時間は、約 30 秒である。標本サイズ 1 万のデータにおいて、M = 20 の多重代入を行うのに必要な時間は、わずか 5 秒である。M= 1000 の多重代入を行うのに必要な時間は、約 5 分である。よって、標本サイズ 1 万未

[103] 本付録の分析結果は、予備的なものである。高橋, 伊藤(2014, pp.68-71)も参照されたい。また、Carpenter and Kenward (2007)、Graham *et al.* (2007)、Bodner (2008)も合わせて参照されたい。

満のデータでは、たとえ M を 1000 に設定したとしても、わずかな時間で演算が終えられることが分かる。

標本サイズ 10 万のデータにおいて、M = 20 の多重代入を行うのに必要な時間は、約 1 分である。M = 100 の多重代入を行うのに必要な時間は、5 分未満である。M = 1000 の多重代入を行うのに必要な時間は、約 45 分である。標本サイズ 50 万のデータにおいて、M = 20 の多重代入を行うのに必要な時間は、約 5 分である。M = 100 の多重代入を行うのに必要な時間は、約 25 分である。M = 1000 の多重代入を行うのに必要な時間は、約 3 時間である。標本サイズが 10 万を超え始めると、M を 1000 に設定するのは現実的ではないが、50 万観測数のデータにおける M = 100 の多重代入は 30 分未満で実行できる。

表 A.2.1：多重代入の実行に要する時間

n	欠測率	M = 5	M = 10	M = 20	M = 100	M = 1000
100	10%	0.10	0.14	0.20	0.75	8.13
	20%	0.06	0.18	0.28	0.80	7.83
	40%	0.08	0.17	0.24	0.85	8.15
1000	10%	0.25	0.36	0.77	3.27	32.44
	20%	0.25	0.36	0.72	3.32	31.18
	40%	0.22	0.42	0.70	3.20	32.29
10000	10%	1.46	2.85	5.42	27.88	294.28
	20%	1.49	2.83	5.64	29.30	288.97
	40%	1.48	3.00	5.83	28.95	309.03
100000	10%	14.22	27.71	54.31	268.93	2761.87
	20%	16.08	28.39	56.43	267.14	2729.72
	40%	14.25	27.56	53.83	263.11	2717.39
500000	10%	74.25	144.28	289.17	1444.12	11767.20
	20%	72.17	141.60	282.20	1411.28	11699.05
	40%	77.62	156.89	307.22	1531.56	10325.04

注：nは標本サイズ、Mは多重代入済みデータ数、報告値は proc.time のユーザタイム（秒）である。

そこで、M数をいくつに設定すればよいかという疑問が生じる。表 A.2.2 は、標本サイズ 1000 のデータにおいて、補定済みデータ数（M数）を 5、10、20、30、40、50、60、70、80、90、100、1000 にした場合、平均値と BSD（補定間標準偏差）にどれだけの影響が出るかを示したものである。表 A.3.2 から、おおむね、Mが 20 になると平均値と BSD が安定し始め、Mが 50 以上あれば平均値と BSD は非常に安定している。Mが 10 未満の場合、平均値と BSD の値が偶発的に小さすぎたり、大きすぎたりするため、欠測による推定誤差の評価を適切に行うことができないおそれがある。欠測率に応じて、M を増加させるのがよいと考えられる。もし実務上において演算時間が問題となるならば、Mを 20 程度に設定すればよいであろう。観測数 50 万のデータにおいて M = 20 の多重代入を行うのに必要な時間は、わずか 5 分程度であり、業務に大きな支障が出るとは考えにくい。

表 A.2.2：多重代入済みデータ数と補定の精度

欠測率	*M*	平均値	BSD	CI UL	CI LL
10%	5	18897.93	283.01	19463.95	18331.92
	10	18854.73	230.21	19315.15	18394.31
	20	18830.58	201.61	19233.80	18427.36
	30	18823.39	168.21	19159.81	18486.96
	40	18834.12	165.06	19164.25	18504.00
	50	18836.85	153.95	19144.74	18528.95
	60	18834.59	147.76	19130.12	18539.07
	70	18830.56	145.11	19120.78	18540.33
	80	18835.23	146.89	19129.01	18541.44
	90	18846.53	152.67	19151.88	18541.19
	100	18843.38	150.29	19143.97	18542.80
	1000	18845.54	141.23	19127.99	18563.09
20%	5	18951.60	164.37	19280.34	18622.87
	10	18884.25	225.58	19335.42	18433.09
	20	18898.08	229.08	19356.24	18439.93
	30	18906.69	222.96	19352.62	18460.77
	40	18964.46	230.09	19424.65	18504.28
	50	18975.48	240.75	19456.97	18493.98
	60	18970.72	225.10	19420.92	18520.53
	70	18958.99	221.16	19401.32	18516.66
	80	18957.73	222.19	19402.11	18513.34
	90	18957.03	222.52	19402.06	18512.00
	100	18956.83	224.06	19404.94	18508.71
	1000	18960.00	227.57	19415.14	18504.86
30%	5	18776.19	92.64	18961.47	18590.90
	10	18940.24	279.66	19499.55	18380.92
	20	18837.87	333.44	19504.74	18171.00
	30	18838.59	318.79	19476.16	18201.02
	40	18859.80	347.39	19554.58	18165.01
	50	18847.86	341.01	19529.88	18165.84
	60	18846.44	343.35	19533.15	18159.74
	70	18853.51	332.85	19519.20	18187.81
	80	18840.61	326.01	19492.63	18188.59
	90	18844.34	314.06	19472.46	18216.21
	100	18842.78	312.55	19467.88	18217.69
	1000	18852.76	316.47	19485.70	18219.82
40%	5	19057.23	414.25	19885.72	18228.74
	10	19058.00	613.55	20285.11	17830.89
	20	19130.06	514.76	20159.58	18100.54
	30	19063.40	464.97	19993.34	18133.46
	40	19130.69	460.82	20052.33	18209.05
	50	19082.38	443.96	19970.31	18194.46
	60	19071.39	418.47	19908.33	18234.46
	70	19098.52	422.14	19942.81	18254.23
	80	19087.68	410.52	19908.73	18266.63
	90	19068.00	405.63	19879.25	18256.74
	100	19051.18	401.74	19854.65	18247.71
	1000	19006.92	408.72	19824.37	18189.48

注：n = 1000

付録 3 Amelia を組み込んだ R コードの例

Amelia パッケージによる多重代入法を行う R コードの例として、平成 26 年中の実験で用いたコードの簡約版を載せる。R の基本的な使用方法については、青木(2009, pp.1-70, pp.279-307)を参照されたい。

入力データは、表 A.3.1 のように、層分けや不要な列の削除は終わっているものとする。すなわち入力データは補定を行う層ごとに別々のデータとして保存され、事業所の識別に用いる ID、補定されるべき欠測値を含む補定対象変数及び補定に用いる補助変数を残し、それ以外のデータは削除されているとする。平成 26 年の実験では、入力データを csv 形式で保存していた。ここでは各入力データのレイアウトは以下のとおりとし、補定対象変数の欠測値はブランクとする。ここでd_iは補定対象変数、i_iは補助変数、添字iは事業所の ID を表す。なお、補助変数はいくつあってもよい。

表 A.3.1：入力データのレイアウト

事業所 ID	補定対象変数	補助変数
事業所 1		i_1
事業所 2	d_2	i_2
事業所 3		i_3
事業所 4	d_4	i_4
事業所 5	d_5	i_5
・ ・ ・	・ ・ ・	・ ・ ・
事業所 n	d_n	i_n

表 A.3.2 を用いて、多重代入法の実行方法を説明する。今、ある入力データのパス名を `H:/入力データ/補定前.csv` とし、`Amelia` パッケージに含まれる `amelia` 関数を用いて多重代入法による補定を行い、その結果を `H:/出力データ/補定済.csv` というファイルに出力したいとする。これを行う R コードは以下のとおり。なお#以降の斜体での記入はコメントであり、コードの左にある数字は後述の説明に用いる行番号である。

まず 2 行目と 4 行目でそれぞれ、入力データのパスを変数 `in_file_name` に、出力データのパスを変数 `out_file_name` に格納する。パスは適宜変更でき、入力ファイルと出力ファイルを保存するフォルダを別のものにすることもできる。パスは文字列なので引用符で囲む。なお R において¥はエスケープ文字のため、フォルダ階層の区切りに¥を使うとエラーが発生する。これを防ぐため区切り文字には / （スラッシュ）を用いる。

7 行目ではブートストラップ再標本の数を指定する（4.1.1 節、付録 2 参照）。ここでは、20 としている。

表 A.3.2：多重代入法を行う R コードの例

```
 1  # 入力ファイルのパス。適宜変更のこと。
 2  in_file_name  <- "H:/入力データ/補定前.csv"
 3  # 出力ファイルのパス。適宜変更のこと。
 4  out_file_name <- "H:/出力データ/補定済.csv"
 5
 6  # ブートストラップ再標本の数を指定
 7  m <- 20
 8
 9  # 擬似乱数のシード値を指定。任意の数字に変えてもよい。
10  set.seed(1223)
11
12  # データをファイルから読込
13  a.data <- read.csv(in_file_name, header = TRUE)
14
15  # Ameliaパッケージの読込
16  require(Amelia)
17
18  # Ameliaによる多重代入
19  a.out <- amelia(a.data,              # data:多重代入を施すデータセット
20                  m = m,               # m:ブートストラップ再標本の数
21                  logs = c("補定対象変数",
22                           "補助変数"), # logs:対数変換する変数名
23                  idvars = "事業所ID") # idvars:補定に用いない名目変数
24
25  # 多重代入による補定値列をデータフレームへ格納
26  impdata <- data.frame(lapply(a.out$imputations, "[", 2))
27  # 列の名前を整える
28  colnames(impdata) <- paste(names(a.out$imputations),
29                                  "補定対象変数",
30                                  sep = ".")
31
32  # Ameliaへ入力したデータの後ろに補定値を追加
33  ameliadata <- data.frame(a.data, impdata)
34
35  # ファイルへ出力
36  write.csv(ameliadata, file = out_file_name, row.names = FALSE)
```

10 行目では擬似乱数のシード値を指定する。同じシード値からは毎回同じ擬似乱数列が生成されるため、結果を再現することができる。

13 行目では変数 in_file_name に格納されたパス（ここでは H:/入力データ/補定前.csv）にある csv データを読み込み、変数 a.data へ格納する。

16 行目は Amelia パッケージを読み込んでいる。なお R ではパッケージの読み込みを行う関数として library 関数と require 関数があるが、require 関数は指定したパッケー

ジが既に読み込まれている場合は何もしない仕様となっている。

19 から 23 行目で amelia 関数により多重代入を行い、結果を変数 a.out へ格納する。R は空白や改行を無視するため、一つの行を分割し読みやすくしている。まず 19 行目、amelia(の直後に多重代入を行うデータセット（ここでは a.data）を指定する。20 行目の引数 m = の後の数字はブートストラップ再標本の数を指定する（ここでは 7 行目で指定したとおり m = 20 を指定）。21 行目と 22 行目では引数 logs = の後に対数変換する変数名（ここでは補定対象変数と補助変数）を指定する。変数名はベクトルの形で与える。R では c 関数で囲むことでベクトルを作成できる。また変数名は文字列なので引用符で囲む。23 行目では引数 idvars = の後に補定に用いない名目変数（ここでは事業所 ID）を指定している。文字列なので引用符で囲む。

26 行目では、得られた結果から補定値列だけを取り出し、データフレーム impdata へ格納する。補定値列はブートストラップ再標本の数（ここでは 20）だけ出力される。28 から 30 行目はデータフレーム impdata の列名を整えている。

33 行目では入力データ（ここでは a.out）とデータフレーム impdata を横に結合し、出力すべきデータフレーム ameliadata へ格納する。36 行目ではデータフレーム ameliadata を、変数 out_file_name に格納されたパス（ここでは H:/出力データ/補定済.csv）へ、csv ファイルとして出力する。

表 A.3.2 のコードを R へ貼り付けると、上記のとおりの処理が行われ、出力データとして補定済.csv が出力される。出力データのデータレイアウトは表 A.3.3 のとおり。ここで$impm_i$は多重代入法による補定対象変数の補定値を、mはブートストラップ再標本の番号を、添字iは事業所の ID を表す。

表 A.3.3：出力データのレイアウト

事業所 ID	補定対象変数	補助変数	補定対象変数 .imp1	補定対象変数 .imp2	・・・	補定対象変数 .impm
事業所 1		i_1	$imp1_1$	$imp2_1$	・・・	$impm_1$
事業所 2	d_2	i_2	$imp1_2$	$imp2_2$	・・・	$impm_2$
事業所 3		i_3	$imp1_3$	$imp2_3$	・・・	$impm_3$
事業所 4	d_4	i_4	$imp1_4$	$imp2_4$	・・・	$impm_4$
事業所 5	d_5	i_5	$imp1_5$	$imp2_5$	・・・	$impm_5$
・ ・ ・	・ ・ ・	・ ・ ・	・ ・ ・	・ ・ ・	・・・	・ ・ ・
事業所 n	d_n	i_n	$imp1_n$	$imp2_n$	・・・	$impm_n$

なお、amelia 関数の引数のうち、主なものは表 A.3.4 のとおりである。表 A.3.2 の例に含まれる引数も再掲する。詳細は amelia 関数のヘルプを参照されたい。関数のヘルプは?関数名で表示できる（ここでは?amelia。事前に Amelia パッケージを読み込んでおく必要がある）。

表 A.3.4：amelia 関数の主な引数

m	デフォルトは５。ブートストラップ再標本の数。
p2s	デフォルトは１。２にすると処理中に出力されるメッセージが詳細なものになる。
idvars	列番号または変数名のベクトルを指定する。その変数は識別符号として扱われる。分析からは除かれるが、補定後に出力されるデータセットには含まれている。
ts	時系列データにおいて列番号または変数名を指定する。その変数は時点として扱われる。
cs	列名または変数名を指定する。その変数はクロスセクション変数として扱われる。
logs	列名または変数名のベクトルを指定する。その変数は対数変換が必要なデータとして扱われる。
sqrts	列名または変数名のベクトルを指定する。その変数は平方根変換される。指定された列のデータに負の数が含まれてはならない。
lgstc	列名または変数名のベクトルを指定する。その変数はロジスティック関数で変換される。指定された列のデータは 0 以上 1 以下でなくてはならない。
noms	列名または変数名のベクトルを指定する。その変数は名義変数として扱われる。
ords	列名または変数名のベクトルを指定する。その変数は順序変数として扱われる。

付録 4 対数正規分布データの補定

本稿で示したとおり、経済データの分析を行う場合には、対数変換を行うことが多い。しかし、回帰補定を行う場合、$\widehat{\log(y_i)}$の算出は最終的な目標ではない。補定モデルの補定対象変数は$\log(y_i)$であるが、実際の補定値はy_iの推定値でなければならない。つまり、補定対象変数の生データの値である。数学的には対数を指数に変換すれば元に戻るわけだが、回帰分析においては、事はそう簡単ではない(Wooldridge, 2009, pp.210-214)。例えば、y_iは補定の対象となる変数、x_iは補助変数、ε_iは誤差項であるとしよう。回帰モデルが式(1)である場合、真のモデルは式(2)である。これが意味するのは、生データにおける回帰モデルは式(3)であり、生データにおける真のモデルは式(4)である(Gujarati, 2003, p.564)。

$$\widehat{\log(y_i)} = \log\left(\hat{\beta}_0\right) + \hat{\beta}_1 \log(x_i) \tag{1}$$

$$\log(y_i) = \log(\beta_0) + \beta_1 \log(x_i) + \varepsilon_i \tag{2}$$

$$\hat{y}_i = \hat{\beta}_0 x_i^{\hat{\beta}_1} \tag{3}$$

$$y_i = \beta_0 x_i^{\beta_1} \exp(\varepsilon_i) \tag{4}$$

さらに、μを平均値、σ^2を分散とすれば、対数正規分布の変数 Yの期待値は、式(5)である(DeGroot and Schervish, 2002, p.278)。

$$E(Y) = \exp\left(\mu + \frac{\sigma^2}{2}\right) = \exp(\mu)\exp\left(\frac{\sigma^2}{2}\right) \tag{5}$$

これらのことから、単純に$\widehat{\log(y_i)}$を指数変換しただけでは、y_iの期待値を体系的に過小推定してしまうことが分かる。その誤差は、式(5)にあるとおり、$\exp(\sigma^2/2)$である。よって、この誤差を補正するために、式(6)のλ_0が必要となる。

$$\hat{y}_i = \lambda_0 \exp\left(\widehat{\log(y_i)}\right) \tag{6}$$

λ_0の値は$\exp(\sigma^2/2)$だが、σ^2は不明であるため、λ_0の値も不明である。よって、λ_0を推定する必要がある。少なくとも過小推定を補正したいので、$\lambda_0 > 1$であることははっきりしているが、残念ながらλ_0の不偏推定量は存在しない。補正項の候補として、回帰の標準誤差$\hat{\sigma}^2$を用いた式(7)の$\hat{\lambda}_0$を使用することが考えられる。式(7)の$\hat{\lambda}_0$は、λ_0の不偏推定量ではないが一致推定量である。多くの場合、この補正項を用いることで良質な変換を行えるが、式(7)の$\hat{\lambda}_0$は誤差項の正規性を仮定している。回帰モデルにおいて、誤差項の正規性はガウス・マルコフの前提（脚注 34 参照）ではなく、大標本においては不要な前提である[104]。

[104] ただし、小標本における統計的推測を妥当なものとするためには必要な前提である。

$$\hat{\lambda}_0 = \exp\left(\frac{\hat{\sigma}^2}{2}\right) \tag{7}$$

そこで、誤差項の正規性を仮定しない式(8)の$\tilde{\lambda}_0$が候補として有力である。なお、ここで、$\hat{u}_i = \log(y_i) - \widehat{\log(y_i)}$である。これは、Duan のスミアリング推定値の特殊ケースである。この補正項も、やはり不偏推定量ではなく一致推定量に過ぎないが、最小二乗法の残差の平均は常に 0 であり、式(8)の$\tilde{\lambda}_0$は常に 1 以上になるという好ましい特性を持っている。

$$\tilde{\lambda}_0 = \frac{1}{n}\sum_{i=1}^{n} \exp(\hat{u}_i) \tag{8}$$

よって、確定的補定による補定値（自然対数）を生データのユニットに戻すには式(9)のとおりにすればよい。

$$\hat{y}_i = \tilde{\lambda}_0 \exp\left(\widehat{\log(y_i)}\right) \tag{9}$$

対数変換したデータを用いた補定モデルに誤差項を追加する確率的補定や多重代入法では、話がやや複雑になる。平均値 0、分散$\sigma^2_{\hat{u}_i}$で正規分布するe_iを指数化すると、式(5)にあるとおり、その平均値はもはや 0 ではない。すなわち、確率的補定や多重代入法による補定値（自然対数）を生データのユニットに戻した場合、$\exp(\sigma^2/2)$分だけ上ぶれすることが分かる。よって、確率的補定と多重代入法による補定値（自然対数）を生データのユニットに戻すには式(10)のとおり、二重に補正すればよい。

$$\hat{y}_i = \exp\left(\frac{\sigma^2}{2}\right) \exp\left(\widehat{\log(y_i)}\right) \exp\left(\frac{-\sigma^2}{2}\right) \tag{10}$$

結果、式(10)では、$\exp(\sigma^2/2)\exp(-\sigma^2/2) = 1$という関係が成り立っている。すなわち、確率的補定と多重代入法では、自然対数から生データに補定値を変換する際に、特別な補正を行う必要がないことを意味している。これは、経済データの補定における多重代入法の副次的な利益と言える。

製 表 技 術 参 考 資 料 30

平成 27 年 6 月発行

編集・発行　　独立行政法人　統計センター

〒162－8668

東京都新宿区若松町 19－1

電　話　代表　03（ 5273 ）1200

統計情報・技術部統計技術研究課　　TEL ： 03－5273－1368

E-mail ： research@nstac.go.jp