欠測値補完に関する調査研究報告書
【詳細版】

平成 29 年３月

内閣府経済社会総合研究所
景　気　統　計　部

本報告書は、平成 28 年 9 月～11 月にかけて内閣府経済社会総合研究所景気統計部において開催した「欠測値補完に関する調査研究」研究会における議論に基づき、作成したものである。

本報告書の作成に当たっては、「欠測値補完に関する調査研究」研究会座長の星野 崇宏 慶應義塾大学経済学部・大学院経済学研究科教授、同研究会委員の土屋 隆裕 情報・システム研究機構統計数理研究所データ科学研究系教授及び元山 斉 青山学院大学経済学部准教授から貴重な御意見やコメントを頂いた。

目次

はじめに

正確な景気判断や適切な経済財政運営を行う観点から、経済財政諮問会議等において、経済統計を始めとする政府統計の改善に関する議論が高まっている。特に、統計調査の精度に影響する欠測値の対応等の統計的手法に関する横断的な課題については、各統計調査の実施主体による取組の一層の強化が求められている。

統計調査の実施に当たり、一部の調査客体から回答が得られない場合や未回答項目が生じた場合、調査実施主体としては回答を督促し回答率を向上させ、データの欠測を最小限にする努力が必要である。それでも一部の回答が得られないために欠測を伴うデータに基づいて集計せざるを得ない場合、当該結果から得られる推定値を可能な限り正しい値に近付けるため、欠測が生じるしくみや統計調査の目的・性質に応じた適切な対応が求められる。しかし、欠測値への対応として、政府統計で広範に用いられる単一代入法（観測されたデータから推定した平均値を欠測値に代入する平均値代入や、観測データに基づいて推定された回帰モデルの理論値を欠測値に代入する回帰代入等）では、推定値の標準誤差を過小評価する可能性がある。また、統計調査のなかには、単一代入法のひとつとして、調査客体の過去の回答値を利用した横置き代入法（LOCF）を採用しているものもあるが、当該手法は経済環境が大きく変動する局面において足下の実態から離れた推定結果をもたらすという欠点がある。

本報告書は、調査客体の無回答や無記入によるデータの欠測に起因する推定の誤差に注目し、その統計的処理方法の主要なものを整理する。また、それぞれの手法を、内閣府において実施・公表している「機械受注統計調査」のデータに適用し、分析結果を示す。これらを踏まえ、欠測が生じるしくみや統計調査の目的・性質に応じて、欠測を含むデータの適切な処理方法を選択するための考え方を示すことで、政府横断的課題である欠測データ処理の改善、しいては公的統計の精度向上につなげることを目指す。

1. 欠測データに伴う問題

統計調査においては、無回答や無記入により、調査客体又は調査項目の一部について情報を得られないことがある。このような統計調査から作成されるデータは、本来観測・記録されるべき値の一部が観測・記録されておらず、「不完全データ（incomplete data）」と呼ばれる。不完全データを用いて推定を行う場合、観測された値のみを用いて推定を行うことが最も直截な方法（「完全ケース分析」と呼ばれる。第 2.1 節参照）であり、実際に広く行われているが、このような推定には、「欠測バイアス」と「推定精度の低下」という２つの問題がある。

〇欠測バイアス

不完全データでは、標本設計において意図された目標母集団の代表性が損なわれている可能性があるため、推定にバイアスを伴うおそれがある。データの一部が観測されないことによって、推定に生じるバイアスを「欠測バイアス」と呼ぶ*（欠測バイアスは選択バイアスの一種である。「選択バイアス」とは、調査客体の意思決定等によって標本抽出（特に処置群の標本抽出）が偏ることで推定にもたらされるバイアスである。）*。

例として、個人の所得額の平均値を推定するための統計調査を考える。仮に回答者の大半が学生や無業者であり、無回答者の大半が残業の多い高所得者である場合、当該調査から推定される所得額の平均値は、真の値である目標母集団の所得額の平均値を下回ることが考えられる。このとき下方の欠測バイアスが生じている。

平均値の推定における欠測バイアスは、次のように、より一般化して示すことができる。まず、統計調査の目的を、目標母集団Uの平均値μの推定とする。ここでは、目標母集団Uが、標本に含まれた場合には必ず回答する調査客体の集合U_Rと、必ず回答しない調査客体の集合U_Mに分割できるとする*（これは、母集団の要素ごとに標本に含まれた場合の回答・無回答の成否が事前に決まっているという仮定である。他方で、母集団の要素ごとに、標本に含まれた場合の回答確率が事前に与えられており、これに従って要素ごとの回答・無回答の成否が確率的に決まると仮定することもできる。前者の仮定を「確定的無回答（deterministic view of nonresponses）」と呼び、後者の仮定を「確率的無回答（stochastic view of nonresponses）」と呼ぶ。確定的無回答は確率的無回答の特殊形である。）*。目標母集団Uのなかで回答者集合U_Rに含まれる要素の割合をπ_Rとし、無回答者集合U_Mに含まれる要素の割合をπ_Mとする（$\pi_R + \pi_M = 1$）。目標母集団Uの平均値μは、回答者集合U_Rの部分母集団平均μ_Rと、無回

答者集合U_Mの部分母集団平均μ_Mとの部分集合構成比による加重平均に等しい（$\mu = \pi_R\mu_R + \pi_M\mu_M$）。標本抽出の結果得られる標本は、調査実施後に回答者と無回答者の部分標本に分割される。*標本Sのなかの回答者の部分標本の値を用いて平均値を計算した場合（この処理方法を「完全ケース分析」と呼ぶ。第2.1節参照）、それは、条件$\mu_R = \mu_M$が成り立つという特別な場合でない限り、母集団平均μの不偏推定量ではない（不偏推定量とは、当該推定量の期待値が推定対象となる母集団特性値と等しい推定量である。すなわちバイアスのない推定量である）。*回答者の部分標本のみを用いて算出した平均値$\bar{y}_R$の推定バイアスは、$\mathrm{Bias}(\bar{y}_R) = \mu_R - \mu = (1-\pi_R)(\mu_R - \mu_M)$である（*一般的に、母集団特性値$\theta$を推定目標とする推定量$\hat{\theta}$の推定バイアス$Bias(\hat{\theta}) \equiv E(\hat{\theta}) - \theta$と定義される*）。数式から明らかなとおり、(1)目標母集団Uにおける回答者の割合π_Rが小さいほど、あるいは(2) 回答者部分母集団U_Rと無回答者部分母集団U_Mの（値$\mu_R - \mu_M$で測られる）異質性が大きいほど、欠測バイアスは大きくなる。第１の点は、欠測率が大きいほど欠測バイアスは大きいということである。第２の点をより一般化して表現すると、「回答の成否と当該変数の間の相互依存性が強いほど欠測バイアスは大きくなる」といえるが、この点については、第 1.1.1 節や第 1.1 節補論で明らかにする。

母集団平均の推定の場合と同様に、母集団分散の推定における欠測バイアスを示すことも比較的に容易である。回答者集合の部分母集団分散をσ_R^2とし、無回答者集合の部分母集団分散をσ_M^2とする。目標母集団Uの分散は$\sigma^2 = \pi_R\sigma_R^2 + \pi_M\sigma_M^2 + \pi_R\pi_M(\mu_R - \mu_M)^2$である。目標母集団$U$からの無作為抽出標本$S$で、観測された値のみを用いて計算した標本不偏分散$s_R^2$の期待値は$\sigma_R^2$であるから、欠測バイアスは次式で与えられる。

$$Bias(s_R^2) = E(s_R^2 - \sigma^2) = \sigma_R^2 - \sigma^2 = (1-\pi_R)(\sigma_R^2 - \sigma_M^2) - \pi_R(1-\pi_R)(\mu_R - \mu_M)^2$$

回答者集合と無回答者集合が２次モーメントまで同質、つまり$\mu_R = \mu_M$かつ$\sigma_R^2 = \sigma_M^2$であれば欠測バイアスは生じないが、そうでない限り欠測バイアスが推定に伴う。数式から分かるとおり、標本における欠測率（の期待値$\pi_M = 1 - \pi_R$）と欠測バイアスの大きさの間には単調な関係を一般的に指摘することはできない。また、この場合は特に、回答者部分母集団と無回答者部分母集団の間の平均値の違い$\mu_R - \mu_M$が下方バイアスの効果をもつ。

〇推定精度の低下

不完全データを用いた推定に伴う第２の問題は、不完全データでは本来得られるべき情報の一部が失われているために、推定の精度が低下することである。たとえば、平均値の推定における欠測による推定精度の低下については、標本サ

イズが縮小している分だけ、標本平均の標準誤差が増加するので、推定精度の低下の程度が分かる。

標本平均の標準誤差の、欠測がない場合の推定量$\widehat{se}_{Comp}$と欠測がある場合の推定量$\widehat{se}_{Incomp}$はそれぞれ次式で与えられる。

$$\widehat{se}_{Comp} = \sqrt{\frac{\sum_{i\in S}(y_i - \bar{y})^2}{n(n-1)}} = \sqrt{\frac{s^2}{n}}$$

$$\widehat{se}_{Incomp} = \sqrt{\frac{\sum_{i\in S\cup U_R}(y_i - \bar{y}_R)^2}{n_R(n_R-1)}} = \sqrt{\frac{s_R^2}{n_R}}$$

ただしここで、nは標本サイズ、n_Rは欠測が生じた場合に回答が得られたレコードの数、$\bar{y}$及びs^2はそれぞれ順に欠測が生じない場合の標本平均及び標本不偏分散、$\bar{y}_R$及びs_R^2はそれぞれ順に欠測が生じた場合の回答者に関する標本平均及び標本不偏分散である。したがって、標本不偏分散s_R^2に欠測バイアスが生じていない場合でも、欠測がある場合の標本平均の標準誤差$\widehat{se}_{Incomp}$は欠測がない場合の標本平均の標準誤差$\widehat{se}_{Comp}$よりも$\sqrt{n/n_R}$倍だけ大きくなり、それだけ推定の精度が低下している。

当然ながら、第１の問題（欠測バイアス）の方が、第２の問題（推定精度の低下）よりも重要であり、優先的に対処することが求められる（*MSE（平均２乗誤差）を最小化する推定は、推定量のバイアス（の２乗）と分散の合計を最小化しているので、バイアスの問題と精度の問題を同列に扱っているといえる。しかし、バイアスのある推定は精度がどれほど高くても意味がないと考えることもできる*）。したがって、欠測データの統計的処理は、欠測バイアスの問題を解決することを第１の目標とし、この目的を果たす限りにおいて、第２の目標である推定精度の改善を目指す。ただし不完全データの統計的処理にはさまざまな手法があり、分析対象となる不完全データの性質や統計調査の目的に応じて適切な手法を用いることが重要である。

1.1 欠測データ処理方法の適性を決める諸条件

欠測を含むデータ、すなわち不完全データに基づく推定において、欠測バイアスを緩和、ないし除去する統計的手法にはいくつかあるが、それらの手法ごとの適性を決める条件を列挙すると次のとおりである。

(1) 欠測データメカニズム
(2) 補助的な変数の利用可能性
(3) 推定目標
    (3.1) 推定対象となる母集団特性値のモーメント次数
    (3.2) 点推定か区間推定かの別
(4) 欠測パターンと欠測率

このなかで最も重要なのは、(1)欠測データメカニズムである。これについては、第 1.1.1 節で概要を示し、第 1.1 節補論でより詳しく説明する。また、(2)補助的な変数の利用可能性は、(1)欠測データメカニズムと関連している。第 1.1.1 節及び第 1.1 節補論では、その関連性も適宜指摘する。その他の条件(3)推定目標及び(4)欠測パターンと欠測率については、第 1.2 節で概説する。

1.1.1 欠測データメカニズムと欠測データ処理方法の適性

欠測データの統計的処理法の適性に影響する諸条件のなかで、最も重要なのは、「欠測データメカニズム」である。「欠測データメカニズム」は、簡単にいえば「欠測の生じるしくみ」である。欠測データメカニズムには３種類があり、欠測の生じる確率的メカニズムの違いによって区別されるが、ここではまず３種類のそれぞれを直感的に説明する。厳密な説明としては、欠測データメカニズムの種類ごとに定義とその含意を第 1.1 節補論に示す。

○完全にランダムな欠測（missing completely at random: MCAR）

変数の欠測する確率が、当該変数の値及び他の観測されている変数の値に依存しない場合のことである。

たとえば、調査対象者が硬貨を投げて、表が出るか裏が出るかに応じて、調査に協力するか否かを決めているとする。*(この場合の欠測による標本の縮小は、当初の標本設計に非復元単純無作為抽出の段階をひとつだけ追加した多段抽出と等価である。欠測によってもたらされる効果は、すべての調査客体の抽出ウェイトに同じ値 1/2 を乗じることに等しい。)* このとき、観測されたデータの標本は、目標母集団の縮図としての性格を失ってはいないとみることができるので、観測された値のみを用いた推定に欠測バイアスは生じない。

上述の例は現実的ではないにしても、それに近いことが実際に起こり得ないわけではない。たとえば、「調査対象者の身長」を調査項目とする標本調査で、無回答者の大部分が、調査票を送付してから回収締め切りまでの期間に住居を不在にしていた者であったとする。この場合、「調査対象者の身長」と住居長期不在の事象とは独立であると考えられる（ある一定期間に住居不在となる確率は、当該者の身長に依存しない）ため、長期不在を理由とする「調査対象者の身長」の欠測（無回答）は、MCAR に極めて近いと考えられる。

MCAR は強い仮定であり、現実的には MCAR が妥当であると考えられる事象は極めて少ない。

〇ランダムな欠測（missing at random: MAR）

変数の欠測する確率が、当該変数の観測された値及び他の観測されている変数の値には依存するが、当該変数の欠測となった値には依存しない場合のことである。

たとえば、目標母集団の平均値を推定する統計調査で、回答者の大半が学生や無業者であり、無回答者の大半が有業者である場合、所得という調査変数の値が欠測する確率は、調査対象者の就業状態という変数の値に依存している。この場合、所得が観測される標本は、学生や無業者に偏ってしまうため、目標母集団の平均所得の推定には、無業者側への下方バイアスが生じる。

ただしこの場合、就業状態がすべての調査対象について観測されていれば、所得が観測されている部分標本の学生や無業者側への偏りを補正することが可能である。すなわち、「有業者は学生や無業者よりも所得が観測されにくい」という追加的な情報と就業状態の分布の情報を平均所得の推定に利用することで、欠測バイアスを緩和することができる。単純な例として、「学生や無業者は必ず所得額を回答するが、有業者は 50%の確率でしか所得額を回答しない」とすると、標本を学生や無業者と有業者とに分割すれば、それぞれの母集団の部分集合の縮図が再現される。部分標本ごとに観測された値のみを用いて標本平均を計算し、欠測を含む標本全体の就業状態構成比でそれらを加重平均すればバイアスのない推定となる*（ただし、この単純化された例では、全所得階層の有業者において、確率的に２人に１人の割合でしか所得額を回答しないという前提が重要な役割をもつ。また、回答者の割合自体については、標本から計算される割合が母集団割合の不偏推定となる）*。つまり、有業者の値に学生・無業者の値の２倍のウェイトを付けて加重平均を算出するという方法がバイアスのない推定方法のひとつとなる。

〇ランダムでない欠測（missing not at random: MNAR）

変数の欠測する確率が、その変数自体の値に依存する場合のことである。

たとえば、資産保有額の母集団平均を推定するための標本調査において、低中位資産額階級と比べて上位資産額階級は資産保有額の情報を秘匿する傾向が強いとする。このような場合、標本が低中位資産階級に偏る（欠測バイアスが生じる)。この場合は MAR と異なり、バイアスの問題を緩和するのは容易ではない。標本が低中位資産階級に偏っていること自体は分かっていても、資産保有額の情報が低中位資産階級の部分しか得られていないため、MAR の場合に示したような偏りの補正を実行することはできない。この場合は、欠測が生じるしくみをモデル化する必要がある。MNAR の下では、MCAR や MAR の場合と異なり、手持ちの情報だけではバイアスのない推定を行うことができないため、“モデルの力を借りる”必要がある。

MNAR と MAR の違いは欠測確率が欠測する変数の値に依存するか否かという点であるが、これは関連する他の変数の利用可能性に関係している。ここで、説明の便宜上、上位資産階級は北部地域に住む住民が大部分を占め、低中位資産階級は南部地域に住む住民によって構成されているという仮想的な経済を考える。上述の資産保有額に関する標本調査の例では、調査客体の居住地域の情報は調査項目として収集されていないと考えている。仮に標本に含まれるすべての調査客体について居住地域に関する情報が得られていれば、上述の説明とは状況が違ってくる。すなわち、上位資産階級が低中位資産階級と比べて資産保有額の欠測を生じやすいということの裏返しとして、北部地域に住む住民は南部地域に住む住民よりも資産保有額の欠測を生じやすいといえる。居住地域という関連する変数が利用できない場合は、保有資産額の欠測する確率が保有資産額自体の値に依存している（すなわち MNAR である）と言わざるを得ないのに対して、居住地域という関連する変数が利用できる場合は、保有資産額の欠測する確率が居住地域に依存しており、とりわけ、居住地域で条件付ければ、保有資産額の欠測する確率が保有資産額自体の値に依存しない（すなわち MAR である）ということができる。欠測確率と欠測する変数自体の値との間に相関があっても、条件付けることで、欠測する変数自体の値に対する欠測確率の依存性を消去できるような補助変数が、すべての調査客体について観測されていれば、それは MAR であるといえる。逆にそのような補助変数が理念的に存在しても、すべての調査客体について観測されていなければ（すなわちデータとして利用可能でなければ)、それは MNAR と異ならない（※より正確な議論は第 1.1 節補論を参照)。

1.1.2 図による解説

欠測データの統計的処理は、MCAR、MAR、MNAR の順に難しくなる。このことを図１－１～１－３に基づいて説明する。具体的なイメージをつかみやすくするために、ここでは世帯が保有する金融資産の額を対象とする。

○MCAR

図１－１は、MCAR の場合を示したものである。図１－１（イ）は、正しく設計された標本抽出に従って得られた標本で、仮に金融資産保有額 Y の値がすべての調査客体について観測されるとした場合の、金融資産保有額 Y のヒストグラムである。現実には無回答により、一部の調査客体について金融資産保有額 Y の値が観測されない。図１－１（ロ）は、図１－１（イ）の標本で実際に無回答による欠測が発生した場合の、回答者と無回答者とを区別した金融資産保有額 Y の合成ヒストグラムである。灰色部分が欠測値、白色部分が観測値をそれぞれ表す。図１－１（ロ）における回答者と無回答者とを分けて、それぞれについての金融資産保有額 Y のヒストグラムを示したものが図１－１（ハ）及び（ニ）である。図１－１（イ）および（ハ）に示された点線は、それぞれの観測されたヒストグラムから推定される金融資産保有額 Y の分布を表す。図１－１（イ）では、標本設計が正しい限り、真の分布を偏りなく推定できる。

図１－１（ロ）をみると、金融資産保有額 Y の値による階級区分ごとの回答率が等しいことが分かる。この特徴が本例題における MCAR の条件を反映している。この場合、回答率が金融資産保有額 Y の値に依存していない。そのため、図１－１（ハ）及び（ニ）のヒストグラムはいずれも真の姿（図１－１（イ））と比べて左右に偏ることなく、図１－１（イ）のヒストグラムを縦軸方向に定率で縮小したものとなっている。そして、図１－１（ハ）の点線に示すとおり、観測された値のみを用いた分布の推定は、欠測がなければ正しく推定される分布（図１－１（イ）の点線）と互いに縦方向に定率倍した関係となる。つまり MCAR の場合、標本サイズが縮小する（推定の精度が落ちる）だけで欠測バイアスは生じない。

○MNAR

一方、図１－２は、MNAR の場合に生じる推定上の問題を同様に示したものである。図１－２（ロ）をみると、金融資産保有額 Y の値による階級区分ごとの回答率が互いに大きく異なっていることが分かる。この特徴が本例題における MNAR の条件を反映している。金融資産保有額 Y の値が大きい階層ほど回答率が低いため、図１－２（ハ）のヒストグラムは左側に、図１－２（ニ）のヒス

トグラムは右側にそれぞれ偏る。そして、図１－２（ハ）の点線に示すとおり、観測された値のみを用いた分布の推定は、欠測がなければ正しく推定される分布（図１－２（イ）の点線）と比べて、低位資産階層側に偏ることになる。分布のこの偏りが、あらゆる推定量の欠測バイアスの源泉である。

○MAR

図１－３は、MAR の場合を示したものである。図１－３（イ）及び（ロ）は、MNAR の場合の図１－２（イ）及び（ロ）と全く同じである。MAR が MNAR と異なるのは、推定の欠測バイアスを緩和するために活用できる補助的な変数が観測されているという点である。図１－３は MAR のもとでの欠測バイアス問題への対処を示す。MAR の場合は、他の観測された情報を用いて、金融資産保有額 Y の値に基づいて分けられた（図では７つの）階層をさらに細分化できる。たとえば、世帯主の就業状態を有業か無業かの２値変数 X として、その値に基づいて各資産階層を２つに分け、それぞれのグループで観測値と欠測値を区別したものが図１－３（ニ）である。欠測が生じなかった場合の金融資産保有額 Y と、就業状態 X の同時分布の情報を図１－３（ハ）に示す。つまり図１－３（ニ）は、図１－３（ロ）と（ハ）の情報を統合したものである。図１－３（ハ）及び（ニ）では、７つの資産階層が、就業状態に基づいてそれぞれ左右に分かれており、ヒストグラムの各棒（各保有資産階層）の左側を無業世帯主、右側を有業世帯主とする。図１－３（ハ）によると、金融資産保有額が低い階層では、無業者世帯の割合が高く、金融資産保有額が高い階層では、有業者世帯の割合が高くなっている。また、図１－３（ニ）によると、無業者世帯の方が有業者世帯よりも回答率が高くなっている。

ここでの設定としては、図１－３（ロ）から（ニ）へ変形しなければ、すなわち、保有金融資産階層の世帯主就業状態による細分化を行わなければ、保有金融資産階層ごとの回答率の分布は、MNAR を表す図１－２（ロ）のヒストグラムと異なるところがない、という点が重要である。つまり、欠測の起こり方は、一見すると MNAR と同様に、金融資産保有額 Y の値が大きい階層ほど回答率が低い。しかしこの場合、それは見せかけの関係であり、世帯主就業状態 X の値で条件付けることにより、金融資産保有額の観測確率が、金融資産保有額の値自体に依存しない部分を取り出すことができる。そのことを示したのが、図１－３（ホ）及び（ヘ）である。図１－３（ホ）及び（ヘ）は、図１－３（ニ）を世帯主の就業所帯に応じて分割したものである。図１－３（ホ）及び（ヘ）は、順に無業者世帯及び有業者世帯それぞれの回答者と、無回答者とを区別した金融資産保有額 Y の合成ヒストグラムである。図１－３（ホ）及び（ヘ）のそれぞれでは、回答率が金融資産保有額 Y の値に依存していない。つまり世帯主の就業

状態で条件付けたとき、MCAR と同様の欠測状況が出現している。

図１－３（ト）及び（チ）は、順に図１－３（ホ）及び（ヘ）それぞれの観測データに基づいて金融資産保有額 Y の条件付分布を推定した結果である。図１－１の MCAR の場合と同様の理由で、世帯主の就業状態による条件付分布は偏りなく推定できる。最後に、図１－３（ト）無業者世帯の金融資産保有額分布の推定結果及び（チ）有業者世帯の金融資産保有額分布の推定結果に、図１－３（ハ）金融資産保有額 Y と世帯主就業状態 X の同時分布の情報を合わせることができれば、全体の金融資産保有額 Y の分布を偏りなく推定できる。そのことを表したのが図１－３（リ）である。

MAR に関するこの例題ではいくつかの好条件が重なっているため最終的に推定から欠測バイアスが完全に除かれるが、実際には興味の対象となる変数 Y と欠測バイアスの除去に利用可能な補助変数 X の同時分布（図１－３（ハ）の情報）が分かっているとは限らず、また、条件付けることで完全に MCAR の性質が現れるような補助変数が存在するとは限らない（図１－３（ホ）及び（ヘ）のような状況になるとは限らない）ので、MAR の下では欠測バイアスの完全な除去ではなく緩和が目指される。いずれにせよ、MAR の場合は、観測された情報に基づいて条件付けることで推定の欠測バイアスを緩和することができる。

MNAR に似た欠測の発生状況でありながら、適当な補助変数で条件付けることにより、MCAR に似た状況を部分的に取り出すことができる、換言すれば、観測された情報によって適当に層化することで、層ごとに欠測の起こりやすさが欠測した変数の値に依存しないようにできるのが、MAR である。

◇まとめ

まとめると、第１に、MCAR のもとでは、観測された情報のみを用いた推定に欠測バイアスは生じない。第２に、MAR のもとでは、観測された情報に基づいて条件付ける（層化する）ことで、条件（層）ごとに欠測バイアスを緩和できる。第３に、MNAR のもとでは、欠測バイアスを緩和できるような条件付け（層化）は、観測された情報の中には存在しない。第 2.7 節でみるように、MNAR に対しては、欠測データメカニズムのモデル化を行うことで、欠測バイアスの問題に対処する。

最後に、注意すべき点として、欠測データメカニズムのいずれが成立しているか（特に MNAR と MAR のいずれであるか）は、不完全データからは知ることができない。これは、図１－１～１－３のパネル（ロ）及び、図１－３のパネル（ニ）～（ヘ）に示した情報がデータからは知ることができないからである。このことが欠測データの統計的処理を限界付ける事実であることを、第 1.3 節で触れる。

図１－１　MCAR

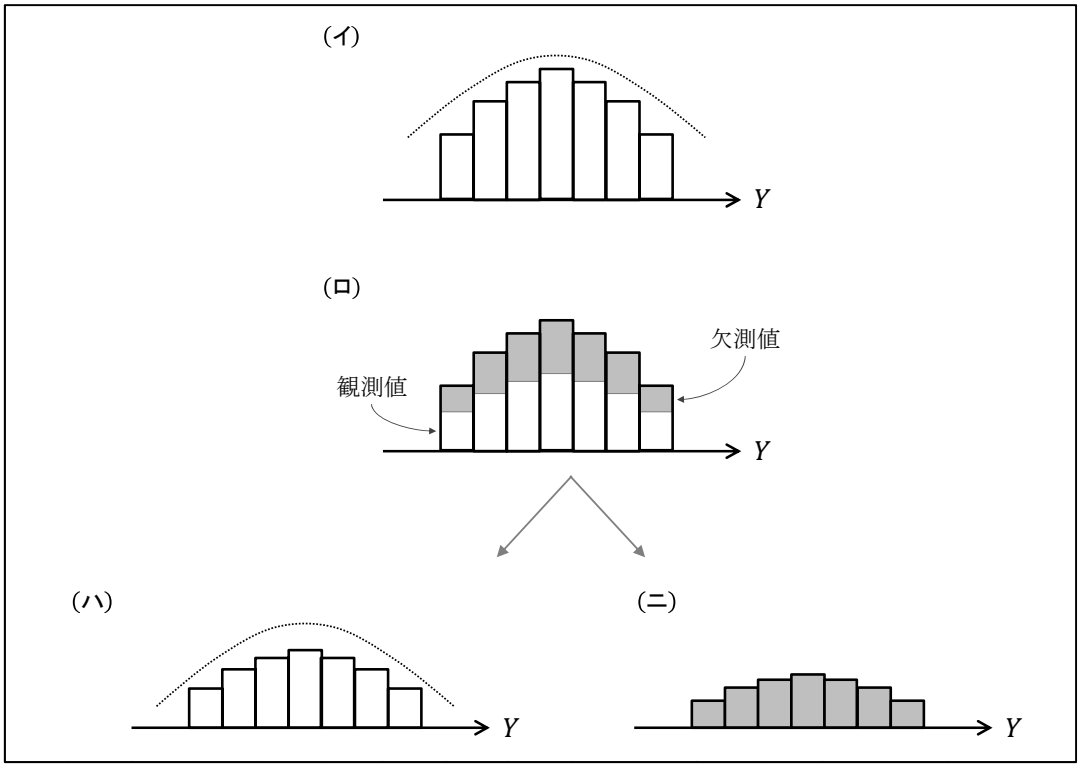


図１－２　MNAR

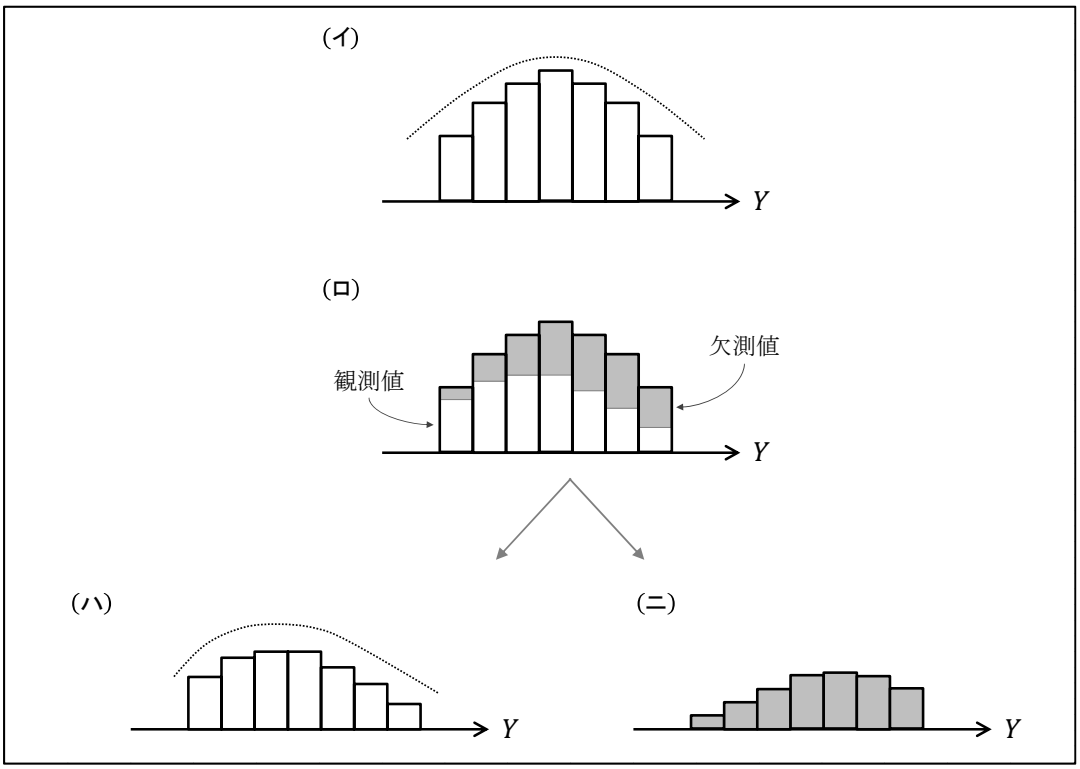
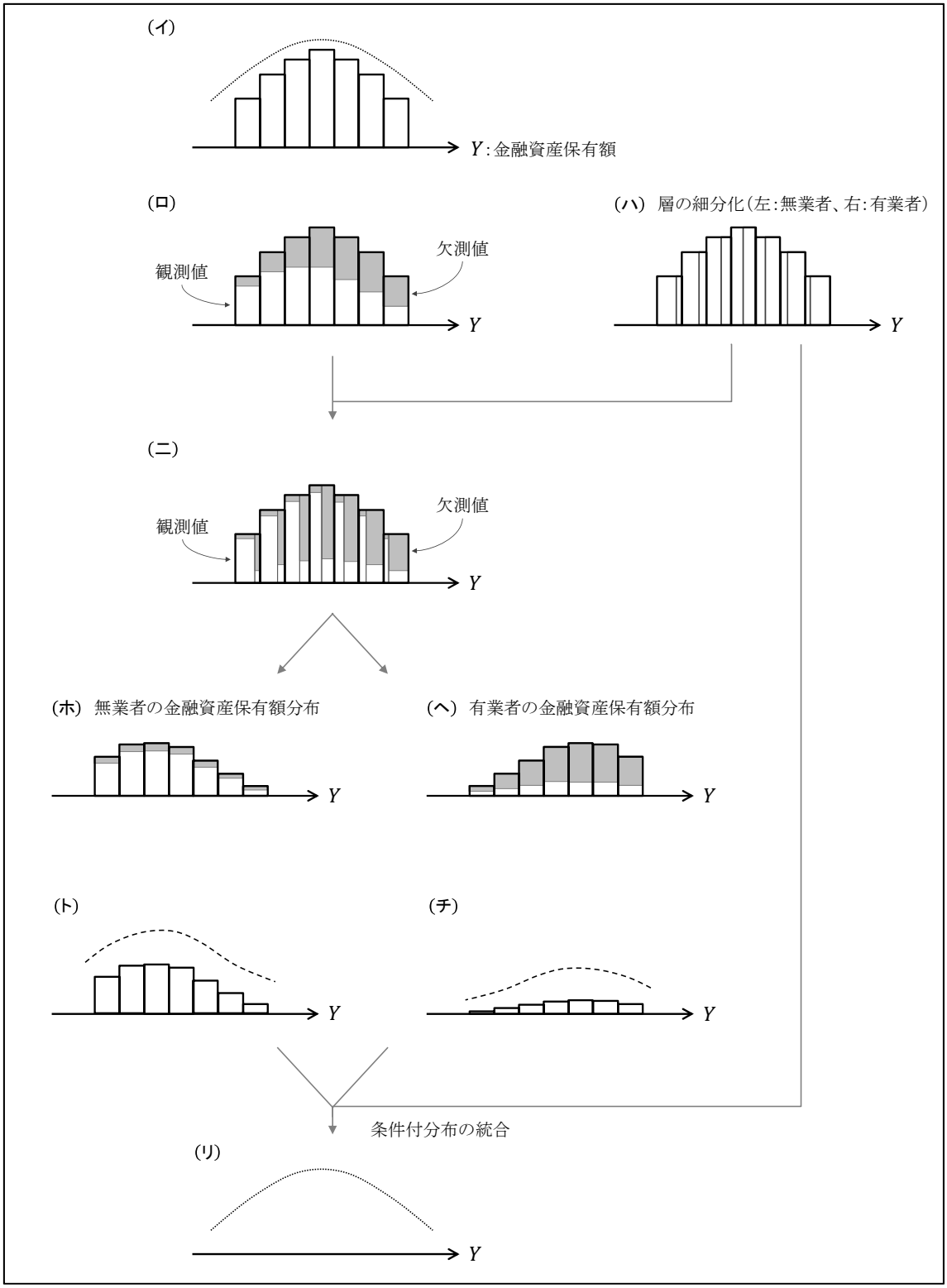


図１－３　MAR


(イ)
Y:金融資産保有額
(ロ)
観測値
欠測値
Y
(ハ) 層の細分化(左:無業者、右:有業者)
Y
(ニ)
観測値
欠測値
Y
(ホ) 無業者の金融資産保有額分布
Y
(ヘ) 有業者の金融資産保有額分布
Y
(ト)
Y
(チ)
Y
条件付分布の統合
(リ)
Y

【補論：欠測データメカニズムの詳細】

以下では、興味の対象となる変数Y_iとその観測指標R_iに対して、補助変数をベクトルX_iで表す。また、興味の対象となる変数Y_iの観測された部分（すなわちレコードiの（狭義）欠測データ）を抽象的な記号Y_i^O、欠測となった部分（すなわちレコードiの観測データ）を抽象的な記号Y_i^Mで表す。図２－１の例に基づけば、興味の対象となる変数の完全データY_iは中央の表全体であるのに対して、観測データY_i^Oは中央表の白色部分であり、欠測データY_i^Mは中央表の灰色部分である。興味の対象となる変数の完全データY_iがベクトルという形式をもつのに対して、観測データY_i^O及び欠測データY_i^Mはいわゆるベクトル空間の要素ではない。情報としては、興味の対象となる変数の完全データY_iは観測データY_i^Oと欠測データY_i^Mを統合したものである。

第 1.1.1 節で概説したとおり、欠測が生じるしくみすなわち欠測データメカニズムには、MCAR（完全にランダムな欠測）、MAR（ランダムな欠測）及びMNAR（ランダムでない欠測）の３種類がある。ここでは、欠測データメカニズムの種類ごとに定義とその含意を示す。

◇MCAR（完全にランダムな欠測）

MCAR（完全にランダムな欠測）の定義には２通りがある。

MCAR の定義１：
不完全データ(Y_i, R_i, X_i)に対して次式が成り立つとき、この不完全データに生じる欠測はMCAR であるという。

$$f(R_i|Y_i, X_i) = f(R_i) \tag{1-1}$$

この定義の同値条件として次の２つがある。

$$f(R_i|Y_i, X_i) = f(R_i) \quad \Leftrightarrow \quad R_i \perp Y_i, X_i \tag{1-2}$$

$$f(R_i|Y_i, X_i) = f(R_i) \quad \Leftrightarrow \quad f(Y_i, X_i|R_i) = f(Y_i, X_i) \tag{1-3}$$

第１の同値条件(1-2)式によると、観測指標（ないし欠測指標）と、完全データ及び補助変数とは、互いに独立である。また、第２の同値条件(1-3)式によると、完全データ及

び補助変数の分布は欠測パターンによらず同一である。

第 1.1.1 節で挙げた MCAR の極端な例では、調査対象者が硬貨を投げて表が出れば回答し裏が出れば回答しないことに決めていたが、無作為に選定された調査対象者の調査項目の値と硬貨の裏表の結果は独立である。また、調査項目たとえば身長の値の分布は、回答者グループと無回答者グループと母集団とで互いに等しい。

第 1.1.1 節図１－１の MCAR の例では、$R_i = 1$と$R_i = 0$のいずれであるかにかかわりなく、金融資産保有額Y_iの条件付分布は母集団分布に等しい（図１－１（イ）、（ハ）及び（ニ）参照）。そのことを式で表すと次式となる。

$$f(Y_i|R_i = 1) = f(Y_i|R_i = 0) = f(Y_i) \tag{1-4}$$

$f(Y_i|R_i = 1)$は図１－１（ハ）のヒストグラムの背後にある分布、$f(Y_i|R_i = 0)$は図１－１（ニ）のヒストグラムの背後にある分布、$f(Y_i)$は図１－１（イ）のヒストグラムの背後にある分布に、それぞれ対応する。

MCAR の定義としてもうひとつ別のものがある。

MCAR の定義２：
不完全データ(Y_i, R_i, X_i)に対して次式が成り立つとき、この不完全データに生じる欠測は MCAR であるという。

$$f(R_i|Y_i, X_i) = f(R_i|X_i) \tag{1-5}$$

定義１の場合と同様に、定義２の同値条件として次の２つがある。

$$f(R_i|Y_i, X_i) = f(R_i|X_i) \quad \Leftrightarrow \quad R_i \perp Y_i|X_i \tag{1-6}$$

$$f(R_i|Y_i, X_i) = f(R_i|X_i) \quad \Leftrightarrow \quad f(Y_i|R_i, X_i) = f(Y_i|X_i) \tag{1-7}$$

第１の同値条件(1-6)式によると、適当な補助変数で条件付けたときに観測指標（ないし欠測指標）と完全データは互いに（条件付）独立である。また、第２の同値条件(1-7)式によると、適当な補助変数の値で層化したとき、完全データの分布は欠測パターンによらず層内で同一である。

興味の対象となる変数の数が１である場合、つまりY_iがスカラーである場合、MCAR の第２の定義は後述の MAR の定義と等しくなるので、以下では、MCAR については第

1 の定義(1-1)に従うこととする。

条件$f(Y_i, X_i|R_i = 1) = f(Y_i, X_i)$（条件$R_i = 1$ に対する(1-3)式）が成り立つことから、MCAR の下では推定に欠測バイアスは生じない。すなわち、欠測のないデータ（不完全データの完全データ）に対して適用したときに不偏推定（あるいは一致推定）となる推定は、MCAR の不完全データに対して適用しても不偏推定（あるいは一致推定）となる。

◇MAR（ランダムな欠測）

MAR の定義：
不完全データ(Y_i, R_i, X_i)に対して次式が成り立つとき、この不完全データに生じる欠測は MAR であるという。

$$f(R_i|Y_i, X_i) = f\left(R_i\middle|Y_i^O, X_i\right) \tag{1-8}$$

MAR の同値条件として次がある。

$$f(R_i|Y_i, X_i) = f\left(R_i\middle|Y_i^O, X_i\right) \quad \Leftrightarrow \quad f\left(Y_i^M\middle|R_i, Y_i^O, X_i\right) = f\left(Y_i^M\middle|Y_i^O, X_i\right) \tag{1-9}$$

ここで、観測データY_i^O及び欠測データY_i^Mがベクトルではない点に注意を要する（本補論冒頭参照）。この点を見逃すと、MAR の同値条件として条件付独立性$R_i \perp Y_i^M\middle|Y_i^O, X_i$が形式的には導かれてしまうが、それは偽である。正しくは、観測指標ベクトルR_iの値がひとたび決まれば観測データY_i^O及び欠測データY_i^Mの形状（正確には次元）も決まるので、観測指標ベクトルR_iの値で条件付けたときに限り観測データY_i^O及び欠測データY_i^Mをベクトルで表現できるという関係にある。定義式(1-8)においては、R_iの値に応じて条件付分布$f\left(R_i\middle|Y_i^O, X_i\right)$の条件部分の次元数が決まる。同値条件(1-9)式においては、R_iの値に応じて条件付分布$f\left(Y_i^M\middle|R_i, Y_i^O, X_i\right)$の定義域部分及び条件部分の次元数が決まる。

第 1.1.2 節図 1 － 3 で挙げた MAR のやや仮想的な例では、目的となる変数Y_iは世帯の金融資産保有額、補助変数X_iは世帯主の就業状態であり、無業者に限れば回答・無回答の別は金融資産保有額と互いに独立であり、同様に、有業者に限っても回答・無回答の別は所得額と互いに独立である。同じことであるが、図 1 － 3（ホ）及び（ヘ）をみると、世帯主の就業状態の値に基づいて標本を分割すれば（すなわち補助変数の値で条件付ければ）、回答世帯と無回答世帯とで金融資産保有額の分布は等しい（互いのヒストグラムが縦方向に定率倍した関係になっている）。そのことを式で表すと次式となる。

$$f\left(Y_i\middle|R_i = 1, X_i = \text{無業者}\right) = f\left(Y_i\middle|R_i = 0, X_i = \text{無業者}\right) = f\left(Y_i\middle|X_i = \text{無業者}\right) \tag{1-10}$$

$$f(Y_i|R_i = 1, X_i = \text{有業者}) = f(Y_i|R_i = 0, X_i = \text{有業者}) = f(Y_i|X_i = \text{有業者}) \quad (1\text{-}11)$$

(1-10)式で、$f(Y_i|R_i = 1, X_i = \text{無業者})$は図１－３（ホ）のヒストグラムの白色部分（すなわち図１－３（ト）のヒストグラム）の背後にある分布、$f(Y_i|R_i = 0, X_i = \text{無業者})$は図１－３（ホ）のヒストグラムの灰色部分（図１－３では省略しているが図１－１や１－２のパネル（ニ）のヒストグラムに相当するもの）の背後にある分布、$f(Y_i|X_i = \text{無業者})$は（図１－３では示していないが）無業者の部分母集団における金融資産保有額Y_iの分布に、それぞれ対応する。有業者に関する(1-11)式についても同様である。

◇MNAR（ランダムでない欠測）

MNAR の定義： 欠測データメカニズムがMCAR でもMAR でもないとき、MNAR であるという。

ここでは慣例に従ってMCAR、MAR 及びMNAR を互いに排反な条件として定義したが、MCAR をMAR の特殊形として、またMAR をMNAR の特殊形として定義することもできる。以下では、排反関係で狭義に定義した場合と、包含関係で広義に定義した場合の両方を文脈に応じて使い分ける。

◇欠測データメカニズムと補助変数

第1.1 節冒頭でも触れたが、欠測データメカニズムと補助変数の利用可能性の間には関係があると考えることもできる。欠測データメカニズムは、不完全データのデータ生成過程に関して定義される事前概念であるため、唯一の真のデータ生成過程と対になっている。しかしながら、利用可能な補助変数の範囲に注目して次のように考えることができる。極端な例として、関係式$Y_i = h(W_i)$を満たす変数W_iが補助変数として不完全データに追加されれば、条件$f(R_i|Y_i, X_i) \neq f(R_i|Y_i^O, X_i)$であっても明らかに条件$f(R_i|Y_i, X_i, W_i) = f(R_i|X_i, W_i)$が成り立つ。これほど極端でなくても、目的となる変数Y_iの適当な代理変数が補助変数のなかに含まれているか否かによって、当該不完全データをMAR の下で生成したものであるのかMNAR の下で生成したものであるのかの別が決まるということができる。

1.2 統計調査ごとの目的・性質と欠測データ処理方法の適性

第 1.1 節冒頭で列挙したとおり、欠測データメカニズム以外にも欠測データ処理法の適性に影響を与える条件がある。それらは個別統計調査ごとの目的及び性質にかかわるものである。

公的統計において推定対象となるのは、多くの場合、母集団平均、母集団総計、母集団割合等の１次モーメントである。（一般的に、確率変数Yに対して期待値$E(Y^h)$を「確率変数Yのh次モーメント」と呼ぶ。通常１次モーメントと２次モーメントが興味の対象となることが多く、１次モーメントは平均値、２次モーメントのうち、平均値からの乖離は分散とよばれる。平均値は当該変数の「代表的な値」、分散は当該変数の「ばらつきの程度」の尺度である。）推定目標となる母集団特性のモーメント次数は、欠測データ処理法の適性を決める条件のひとつである。欠測に伴う分布のゆがみの補正結果は、手法ごとに異なるため、特に推定目標が１次モーメントであるか、１次より大きいモーメントであるかが手法の適性を大きく左右する。推定目標が１次モーメントである場合、補正後の分布が真の分布と対称性に関して同等となるような手法は、すべて推定に欠測バイアスをもたらさないといえる。

次に、推定目的が点推定にとどまるものか、区間推定ないし統計的仮説検定にも及ぶものであるかということも、欠測データ処理法の適性を決める。当然ながら、点推定のみを目的とする分析の方が、適切な処理法の選択肢の範囲が広い。欠測に伴う分布のゆがみの補正結果が、欠測バイアスを除去ないし緩和するものであっても、補正後の分布のばらつきが真の分布よりも小さくなる場合が多い。このような場合は、推定値の標準誤差が過小評価されるため、当該手法は区間推定ないし統計的仮説検定には適さないといえる。社会的な認識の大勢としては、公的統計の目的は点推定にとどまると考えられ、区間推定ないし統計的仮説検定には適さない手法の多くが従来用いられてきた。

公的統計においては、統計調査の目的よりもデータの性質の方が、欠測データ処理法の適性により大きく影響すると考えられる。データの性質としては主に、(1)どのような補助変数（欠測が生じるしくみをモデル化する際に説明変数となり得る変数）が利用可能か、(2)目的となる変数を適当に加工したときにパラメトリックな分布（正規分布やポワソン分布など）で近似できるか、(3)調査項目の欠測可能性について先験的な知見があるか、という点が重要である。

第１の点については、公的統計における補助変数は、多くの場合、フレーム（標本抽出を行うための調査客体リストすなわち母集団データベース）に情報として含まれる調査客体属性である。統計調査の結果として不完全データが与えられたとき、適当な補助変数で標本を層化し、層ごとの回答率を確認することは有

益である。層ごとの回答率に顕著な傾向がみられる場合、MAR を仮定した推定において、当該補助変数を利用することができる。第 1.1.2 節図１－３の例では、世帯ごとの金融資産保有額の欠測に関して、世帯主の就業状態が有用な補助変数となっている。このような補助変数が利用可能か否かによって、データが MAR か MNAR のどちらであるかが決まると考えることもできる。

第２の点については、欠測データ処理の手法の中で、多重代入法や尤度法のようなモデル依存性の高い方法による場合、目的となる変数を加工して定型的な分布に可能な限り近づけることで、推定の一致性を保証する適切なモデルの特定化が可能となる。たとえば、個人の所得や企業の規模のように裾の長い分布を示す変数は対数変換することで、正規分布に近い分布を示す変数を得ることができる。

第３の点については、変数の欠測可能性に関する先験的な知見が、データからは検証することのできない分析の前提を裏付けるものとして、重要な役割を果たす。特に、欠測可能性に関する先験的知見は、有意な補助変数の選択に役立つだけでなく、欠測データメカニズムのモデルの定式化にも示唆を与える。たとえば調査協力に対する謝礼は、調査対象主体が回答と無回答を選択する意思決定における誘因となり、回答群の選択原理として作用する。一定額の謝礼から得られる効用が、調査対象主体ごとに異なるためである。たとえば、謝礼から得られる追加的な効用は、通常、所得の高い主体にとっては小さいが、所得の低い主体にとっては大きいと考えられ、所得の小さい主体ほど回答確率が大きい。この場合、所得額が欠測に有意な補助変数であることが分かるだけでなく、行動原理に関する知見から調査対象主体の意思決定モデルまでもが導かれる。尤度法（第 2.6 節）では、こうした先験的な知見に基づいて調査客体の意思決定をモデル化することで欠測バイアスの問題に対処することができる。

また、変数の分布に関する先験的知見が欠測データメカニズムの識別に役立つこともある。たとえば、興味の対象となる変数 Y が対称な分布に従うと分かっている状況の下で、利用可能な情報でどのように条件付けても当該変数の観測された値の分布が非対称となる場合、欠測データメカニズムは MNAR であることが示唆される。この例では、通常データに含まれる情報のみでは識別できない欠測データメカニズムが、「変数 Y は対称な分布に従う」というデータ外の追加的な情報によって識別できるようになっている。

最後に、データが示す欠測パターン及び欠測率も欠測データ処理法の適性に影響を与える。複数の変数にわたって欠測が生じる場合、欠測パターンは複雑なものになりうる。欠測データ処理法のなかには欠測が発生するプロセスを明示的にモデル化するものもあり、複雑な欠測パターンを示すデータには分析を困難にするという理由で適さない場合がある。また、欠測率が極端に小さい場合は、欠測発生プロセスのモデルを推定する

手法で、情報量の制約が問題となる。ただし、これらはIPW 法及び尤度法を用いる場合に関する事項であり、現行の公的統計で採用されている単一代入法などの手法においてはあまり問題とならない。

1.3 欠測データ処理の限界

第 1.1.1 節で概要を示し、また第 1.1 節補論で詳細を示したとおり、「欠測データメカニズム」は、MCAR（missing completely at random: 完全にランダムな欠測）、MAR（missing at random: ランダムな欠測）、MNAR（missing not at random: ランダムでない欠測）の３種類に分類される。欠測データの統計的処理においては、MCAR は例外的であり、実践的には MAR と MNAR を想定しなければならない。

欠測データの統計的処理は、「欠測データ処理の適性は、欠測データメカニズムに応じて決まるが、欠測データメカニズム自体はデータによって検証できない」という事実によって限界づけられる。具体的には、不完全データに含まれる情報だけでは、MAR と MNAR のどちらの条件が成立しているかを検証できないのである。第 1.1.2 節図１－２及び１－３において、適当な補助変数を用いてパネル（ロ）からパネル（ニ）を作成できれば MAR であり、作成できなければ MNAR であるが、データに含まれる情報だけでは、図１－２の MNAR ではパネル（ロ）のヒストグラムを描くことができないし、図１－３の MAR では、パネル（ハ）の情報が得られていても、パネル（ロ）ひいては（ニ）のヒストグラムを描くことができないのである。MNAR ではないという仮定の下では、MCAR と MAR を比較する検定は可能であるが、MNAR ではないという仮定自体が検証できないという限界は残る。

この限界の下で最大限可能なことは、欠測が生じるしくみに関するあらゆる事態を網羅的に想定して、それらの想定ごとに適切な分析を実行し、結果を比較することである。これは「感度分析」と呼ばれるものである。不完全データが与えられたとき、その欠測データメカニズムが先験的に明らかでない限りは、いくつかの想定を組み合わせて感度分析をすることが望ましい。

換言すれば、不完全データに基づく推定からは条件付の結論しか導くことができない。欠測データメカニズムがMAR である場合を想定した推定結果とMNAR である場合を想定した推定結果が大きく異ならなければ、幸運にも頑健な結果を得たということができる。

2. 欠測データの統計的処理

上述のとおり、欠測データが与えられたとき、どのような手法が適切かを決める諸条件のうち、最も重要なものは「欠測データメカニズム」である*(ただし、他の条件とは異なり、欠測データメカニズムをデータから検証することはできないことも指摘した(第 1.3 節))*。第 1.1.1 節では、欠測データメカニズムの種類として MCAR、MAR 及び MNAR の3つがあり、欠測データ処理に伴う問題がこの順に難しくなることを直感的に説明した。本節では、主要な用語と概念を説明したうえで、第 2.1 節～第 2.6 節で欠測データの統計的処理法の主要なものを説明する。

○用語と概念

不完全データの観測されなかった値を「欠測値(missing values)」と呼ぶ。一方、観測された値は特に「観測値(observed values)」と呼ばれる。通常欠測値は、何らかの固定された値をとると考えられる。たとえば、A さんが、あるアンケート調査の調査客体に選ばれ、調査項目のひとつである年齢を回答しなかった場合、この調査から作成された不完全データにおいて、A さんの年齢は欠測値である。しかし、それは観測・記録されていないだけであって、たとえば 30 歳という真の値は存在する。このように、ある不完全データに対して、そのすべての欠測値に値が観測されていれば得られるはずのデータというものが仮想的に存在する。これを当該不完全データの「完全データ(complete data)」と呼ぶ。

不完全データは、その完全データの情報の一部に覆いをかけることによって得られたものとみなすこともできるが、ここでやや意外なことに、不完全データは対応する完全データよりも厳密に少ない情報しか含んでいないというわけではない。不完全データについては、調査実施前には、どの調査項目も潜在的に欠測する可能性があり、調査実施後には、事前の欠測可能性に関しても(事後的な)データが得られたと考えることができる。不完全データには、「(事後に)どの値が観測されどの値が観測されなかったか」という情報が含まれている。これに対して、完全データが観測された場合は「(事後に)すべての値が観測されている」という情報が得られたことになるが、この情報は欠測可能性を推定するうえでは役に立たない情報である。つまり、幸運にも欠測が生じることなく完全データが得られた場合は、(事後の)データから事前の欠測可能性を推定するのに利用可能な情報は皆無であるが、運悪く欠測が生じて不完全データが得られた場合は、(事後の)データから事前の欠測可能性を推定するのに利用可能な情報が得られたことになる。このように、不完全データはその完全データと比べて、欠測値の情報は失われているものの、事前の欠測可能性に関する情報が追加的に含まれている。実は、MNAR の欠測データメカニズムに対する欠測データ処理の要は、不完全データがその完全デ

ータと比べて追加的に有する欠測可能性に関する情報をいかに活用できるかにかかっている。念のため付言すると、欠測可能性自体には興味はないので、不完全データよりもそれに対応する完全データを得ることの方がはるかに望ましいことはもちろんである。

不完全データがその完全データと比べて追加的に有する欠測可能性に関する情報は、「観測指標」又は「欠測指標」で表すことができる。調査客体iの調査項目jを変数Y_{ij}で表す。変数Y_{ij}の値が観測されていれば値1をとり観測されなければ値0をとる2値変数R_{ij}を変数Y_{ij}の「観測指標（observation indicators）」と呼ぶ。「欠測指標（missing indicators）」は、目的の変数が観測されなければ値1をとり観測されていれば値0をとる2値変数である。明らかに観測指標と欠測指標のいずれか一方のみを用いれば十分であり、以下では観測指標に統一する。

図2－0　不完全データの構造

不完全データ

	年齢	身長(cm)	体重(kg)
Aさん	-	170	65
Bさん	28	161	-
Cさん	41	-	-
Dさん	-	171	-

ハイフン（-）は無回答を表す

（左の不完全データの）完全データ

	年齢	身長(cm)	体重(kg)
Aさん	30	170	65
Bさん	28	161	58
Cさん	41	166	60
Dさん	55	171	67

（左端の不完全データの）観測指標

	年齢	身長(cm)	体重(kg)
Aさん	0	1	1
Bさん	1	1	0
Cさん	1	0	0
Dさん	0	1	0

ある不完全データが与えられたとき、それは対応する完全データと観測指標の組合せ(Y_i, R_i)で表すことができる。このような不完全データの表現の具体例を図2－0に示す。作成された不完全データは実際には左端の表のようにみえるが、その背後には中央の表に示すとおりの現実があって、これが当該不完全データの完全データである。完全データの表の中で背景灰色の値が欠測値である。完全データのどの部分が観測されているかを示すのが、右端の表に示す観測指標のデータである。このように不完全データは、それに対応する完全データと欠測指標データの組合せによって表現できる。

一般的に、何らかのデータが得られたとき、そのデータの背後に確率空間を想定することができ、得られたデータはその確率空間におけるひとつの試行の結果とみることができる。背後に存在する確率空間を「データ生成過程（data generating process）」と呼ぶ。さて、不完全データ(Y_i, R_i)の値を発生させるデータ生成過程は、完全データ(Y_i)の値を発生させるデータ生成過程と観測指標(R_i)の値を発生させるデータ生成過程の2つによって構成される。上述のとおり、仮に完全データが観測されていればそのデータには、完全データのデータ生成過程に関する推定に資する情報は含まれているが、欠測指標のデータ生成過程に関する推定に資する情報は全く含まれていない。他方、不完全データには、完全データのデータ生成過程に関する推定に資する情報はその一部が損なわれつつ含まれており、さらに観測指標のデータ生成過程に関する推定に資する情報も含まれている。そして、統計調査の目的は完全データのデータ生成過程に関する推定であるのだから、観測指標のデータ生成過程に関する推定に資する情報は何の役にも立たないかというと、そうではない。第 2.6 節でみるとおり、欠測デー

タ処理が最も困難な MNAR に対しては、この情報を活用することで欠測バイアスの問題に対処する。

既述のとおり「欠測データメカニズム(missing data mechanism)」とは、平易にいえば「欠測の生じるしくみ」のことであるが、より正確には、観測指標R_iの(条件付)確率分布の完全データY_iの値に対する依存性のことである。この依存性には3種類があり、観測指標R_iの(条件付)確率分布が、完全データY_iの値に全く依存しない MCAR、少なくとも欠測値には依存しない MAR、そして欠測値にも依存する MNAR に分けられる。MCAR と MAR では、不完全データのデータ生成過程を構成する2つの要素すなわち完全データのデータ生成過程及び観測指標のデータ生成過程が条件付で互いに独立となり、他方 MNAR では、観測値で条件付ける限り両者は互いに独立とはならない。別の言い方をすると、完全データと欠測指標の(条件付)同時分布が、MCAR 及び MAR では欠測データの(条件付)分布と観測指標の(条件付)分布に分離できるが、MNAR では分離できない(これは、MCAR 及び MAR が「無視可能な欠測」と呼ばれることと関係している)。MNAR に対する欠測データ処理法では、完全データと観測指標の同時分布をモデル化することが求められる。

欠測データの統計的処理においては、第 2 節の例題にみるとおり、不完全データの標本を層化するために用いることのできる変数が重要な役割を果たす。不完全データのすべてのレコード（※）で値が観測されていて、標本分割の条件付け（層化）に利用可能な変数を特に「補助変数（covariates)」と呼ぶ。欠測データメカニズムは、不完全データのデータ生成過程について定義されるが、補助変数の利用可能性によって再定義できる。その場合、特に実践的には、補助変数の利用可能性が MAR と MNAR の分かれ目を決するとみることができる（第 1.1.1 節 MNAR の説明参照)。このため、欠測データの統計的処理の適性は、適当な補助変数の利用可能性に大きく依存するといえる。

※一般的に、統計調査のデータは「調査客体×調査項目」という２つの次元をもつ。このようなデータは、個人、世帯、企業といった調査客体を行、調査項目を列とする行列で表現される。つまりこの行列の第i行第j列の要素は、第i番目の調査客体の第j番目の調査項目の値を表す。個々の行を「レコード」と呼ぶ。

2.1 完全ケース分析

分析に用いる変数のすべての値が観測されている調査客体のみを用いて分析を行うことを、「完全ケース分析（complete case analysis）」と呼ぶ。完全ケース分析では、分析に用いる変数の少なくとも1つが欠測となっている調査客体を、分析対象から除外する。この操作を、「リストワイズ削除（list-wise deletion）」と呼ぶ。

ひとつの分析がいくつかの分析に分解できるとき、分解された個々の分析ごとに完全ケース分析を行うことを、「利用可能ケース分析（available case analysis）」と呼ぶ。利用可能ケース分析では、分解された個々の分析ごとに、用いる変数の少なくとも1つが欠測となっている調査客体を分析対象から除外しており、この操作を「ペアワイズ削除（pair-wise deletion）」と呼ぶ。

図2－1－1は、エンゲル係数の推定について、完全ケース分析と利用可能ケース分析の実行例を示したものである。表(A)は、仮に欠測が生じなければ得られるはずの完全データを示す。完全データによると、消費支出の標本総計は 292 万円で、食料品支出の標本総計は 65.7 万円なので、総体のエンゲル係数は 22.5%である。実際には欠測が生じ、表(A)の背景灰色で示した値は欠測値である。

完全ケース分析による総エンゲル係数の推定を表(B)に示す。消費支出と食料品支出のどちらか一方でも欠測となっている調査客体は、分析対象から除外するので、id=1, 2, 3, 4, 5, 6 の6世帯が削除される。残された6世帯については、消費支出の総計が 187 万円で、食料品支出の標本総計が 38.8 万円なので、総体のエンゲル係数は 20.7%と推定される。

利用可能ケース分析による総エンゲル係数の推定を表(C)に示す。利用可能ケース分析による総エンゲル係数の推定は、完全ケース分析による総消費支出（あるいは平均消費支出）の推定と、完全ケース分析による総食料品支出（あるいは平均食料品支出）の推定から成っている。平均消費支出の推定では、id=1, 3, 5 の3世帯が分析対象から除外される。残された9世帯については、消費支出の平均が 241 万円/9 世帯である。他方、平均食料品支出の推定では、id=2, 4, 6 の3世帯が分析対象から除外される。残された9世帯については、食料品支出の平均が 52.6 万円/9 世帯である。これら2つの推定結果を合わせて、総体のエンゲル係数は 21.8%と推定される。

図2－1－1　完全ケース分析と利用可能ケース分析

(A) 完全データ

家計id	消費支出（万円）	食料品への支出（万円）
1	16	5.5
2	16	4.3
3	17	4.4
4	18	4.5
5	18	3.9
6	20	4.3
7	22	4.8
8	26	5.5
9	28	5.8
10	33	6.7
11	36	7.6
12	42	8.4
合計	292	65.7

エンゲル係数 $= \frac{65.7}{292} = 22.5\%$

(B) 完全ケース分析

家計id	消費支出（万円）	食料品への支出（万円）
1	削除	
2	削除	
3	削除	
4	削除	
5	削除	
6	削除	
7	22	4.8
8	26	5.5
9	28	5.8
10	33	6.7
11	36	7.6
12	42	8.4
合計	187	38.8

エンゲル係数 $= \frac{38.8}{187} = 20.7\%$

(C) 利用可能分析

家計id	消費支出（万円）
1	削除
2	16
3	削除
4	18
5	削除
6	20
7	22
8	26
9	28
10	33
11	36
12	42
合計	241

家計id	食料品への支出（万円）
1	5.5
2	削除
3	4.4
4	削除
5	3.9
6	削除
7	4.8
8	5.5
9	5.8
10	6.7
11	7.6
12	8.4
合計	52.6

エンゲル係数 $= \frac{52.6}{241} = 21.8\%$

利用可能ケース分析は、完全ケース分析よりも削除するレコードの数が少なくなる分だけ、推定精度の低下が抑制されるが、変数ごとに分析対象が異なるため、変数相互間の関係性にゆがみがもたらされるという問題がある。図2－1－2は、図2－1－1と比べて他の条件は一定として、欠測パターンが異なる場合に、完全ケース分析と利用可能ケース分析の実行結果を示したものである。先の図2－1－1では、消費支出についても食料品支出についても、支出額の小さい世帯で欠測が生じやすい状況であった。これに対して、図2－1－2では、消費支出については、支出額の大きい世帯で欠測が生じ、食料品支出については、支出額の小さい世帯で欠測が生じている。つまり、図2－1－2の利用可能ケース分析では、消費支出の平均値は支出額が相対的に小さい世帯の組合せで算出され、食料品支出の平均値は支出額が相対的に大きい世帯の組合せで算出されている。このため、図2－1－2の利用可能ケース分析の結果は、総エンゲル係数の推定値が 28.5%と実際よりもかなり大きい値となっている。エンゲル係数の分母は支出規模の大きい世帯群で計算され、分子は支出規模の小さい世帯群で計算されていることの表れである。

図2－1－2 利用可能ケース分析の問題

(A) 完全データ

家計id	消費支出（万円）	食料品への支出（万円）
1	16	5.5
2	16	4.3
3	17	4.4
4	18	4.5
5	18	3.9
6	20	4.3
7	22	4.8
8	26	5.5
9	28	5.8
10	33	6.7
11	36	7.6
12	42	8.4
合計	292	65.7

エンゲル係数 $= \frac{65.7}{292} = 22.5\%$

(B) 完全ケース分析

家計id	消費支出（万円）	食料品への支出（万円）
1	削除	
2	削除	
3	削除	
4	18	4.5
5	18	3.9
6	20	4.3
7	22	4.8
8	26	5.5
9	28	5.8
10	削除	
11	削除	
12	削除	
合計	132	28.8

エンゲル係数 $= \frac{28.8}{132} = 21.8\%$

(C) 利用可能分析

家計id	消費支出（万円）
1	16
2	16
3	17
4	18
5	18
6	20
7	22
8	26
9	28
10	削除
11	削除
12	削除
合計	181

家計id	食料品への支出（万円）
1	削除
2	削除
3	削除
4	4.5
5	3.9
6	4.3
7	4.8
8	5.5
9	5.8
10	6.7
11	7.6
12	8.4
合計	51.5

エンゲル係数 $= \frac{51.5}{181} = 28.5\%$

完全ケース分析は、最も簡単な欠測データ対処法であり、多くの統計処理ソフトウェアで、不完全データを分析する場合には原則的に適用される。しかし、完全ケース分析では、欠測データメカニズムが MCAR でない場合、リストワイズ削除によって分析対象標本が偏ることで、推定にバイアスが生じる。また、回答率が十分に大きくない限り、リストワイズ削除によって、データから失われる情報の量は相当大きい。この失われた情報のために推定の精度は低下する。

本章冒頭で述べたとおり、欠測バイアスとは、不完全データに完全ケース分析を実施したときに推定に生じるバイアスのことである。欠測データの問題を考えるうえで、完全ケース分析が出発点となる。完全ケース分析とは異なる手法でバイアスのない推定を行うのが、欠測データの統計的処理である。

2.2 単一代入法

不完全データのすべての欠測値を、それぞれ適当な規則に基づいて決められた値で置き換えることによって、あたかも欠測のないデータを作成することができる。このように作成されるデータを「疑似完全データ（pseudo-complete data）」と呼ぶ。欠測値に代わる値として代入される値を「代入値（imputed values）」と呼ぶ。疑似完全データを1つ作成し、それに対して分析を適用する方法が「単一代入法（single imputation methods）」である。

前節で説明した完全ケース分析は、欠測を含むレコードを分析対象から除いており、これらの削除されたレコードに含まれる情報が無駄になっているといえる。そこで、欠測を含むレコードに含まれる情報を何らか有効活用できないかという動機が働き、その第1歩として単一代入法を位置付けることができる。単一代入法は、後述の多重代入

法、尤度法、及び IPW 法と比べて、統計的な正当性が弱いものの、MAR のもとでは1次モーメントの点推定に関する限り、欠測バイアスのない推定が可能であり、また処理手順が容易であるため、公的統計で広く用いられてきた。

単一代入法は、代入値の決め方に関していくつかの種類がある。その主要なものは以下のとおりである。

○平均値代入（mean imputations）・層化平均値代入（stratified mean imputations）

観測値の平均値を代入値とする。不完全データを補助変数で層化して、代入値となる平均値を層ごとに計算する方法は、特に「層化平均値代入（stratified mean imputations）」と呼ぶ。*層化平均値代入は、層ごとの回答率により抽出ウェイトを調整する方法と同じ点推定の結果をもたらす。*

○回帰代入（regression imputations）

欠測が生じた変数を従属変数とし、補助変数を独立変数とする回帰モデルを、観測レコードのすべてを用いて推定し、推定された回帰モデルの理論値を代入値とする。

回帰モデルの関数形の特定（線形か非線形か、どの補助変数を独立変数とするか等）及び推定方法（OLS、GLS、MLE、GMM 等）に関して選択の余地がある。
目的の変数が連続変数である場合は、線形回帰モデルを OLS により推定する方法がよく用いられる。

○確率的回帰代入（stochastic regression imputations）

回帰代入の代入値に、推定された分布から乱数発生させた誤差項を加えた値を代入値とする。誤差項を代入値へ加算するのは、回帰代入では捉えることのできない欠測値のばらつきを捉えることを意図している。

欠測する変数が連続変数である場合は、たとえば正規分布線形回帰モデルが考えられる。この場合、誤差項が従う正規分布の平均は0であり（線形回帰モデルに定数項を含む）、分散には完全ケース分析の推定値を用いる。欠測する変数が離散変数である場合は、たとえば多項ロジットモデルが考えられる。この場合、推定された確率に従って代入する離散値を確率的に決める。

○マッチング代入（matching imputations）

欠測値をもつレコードと、すべての変数が観測されているレコードの間で、互いに補助変数の値が類似しているものを対応付け、後者の観測値を前者の欠測値に代入する。補助変数の値の類似性は、適当に定義された距離によって測る。このマッチングの方法を、「最近傍マッチング（nearest neighbor matching）」と呼ぶ。

マッチング代入は OLS による回帰代入と関係がある。Angrist and Pischke (2008)が示すとお

り、マッチングと OLS 回帰のそれぞれによる平均処置効果の推定量は、両者とも補助変数の値で条件付けた処置効果の加重平均という形式をもつ。両者間では加重平均のウェイトに差異がある。

補助変数の値の類似性は、補助変数のベクトル空間における距離によって測る。用いる距離に応じて種類がある。最近傍マッチングにおいて計測されるレコードiとレコードjの間の距離$d(i,j)$の定義としては以下のものが挙げられる。

一般化2次ノルム：

$$d(x_i, x_j) = \{(x_i - x_j)' Q^{-1} (x_i - x_j)\}^{\frac{1}{2}}$$

ここで、行列Qによって種類が決まる。もっとも簡単な距離はユークリッド距離であり、$Q = I$である。行列Qが次式である場合は、「マハラノビス距離（Mahalanobis distance）」と呼ばれる。

$$Q = \frac{(X - \bar{x}' 1_n)' W (X - \bar{x}' 1_n)}{\sum_{i \in S} w_i - 1}$$

ただし、ここでw_iはレコードiの標本抽出ウェイト、行列Wは標本抽出ウェイトの対角行列、$\bar{x} \equiv \sum_{i \in S} w_i x_i / \sum_{i \in S} w_i$である。つまり、行列$W$は補助変数の標本分散共分散行列$Q = \widehat{Var(x)}$である。用いる変数の中には個体間での変動が大きいものもあれば小さいものもありまた、互いに相関の強い組合せもあれば弱い組合せもある。

マハラノビス距離は、分散の大きい変数や相関の強い変数の組合せの距離に対する効果を割り引くという考え方に基づいている。分散の大きい補助変数の距離に対する効果を割り引くという点に関しては、たとえば、身長と体重の2つの補助変数でユークリッド距離を計算すると、身長の単位を cm として体重の単位を kg とする場合と、身長の単位を m として体重の単位を同じく kg とする場合では、前者の場合で後者の場合と比べて身長の分散が大きくなる。両者の場合で距離の値が異なるだけでなく、距離の大小関係による順序も異なる。これに対してマハラノビス距離では、補助変数の標本分散共分散行列の逆行列対角成分によって単位の効果が除去される。

また、相関の強い補助変数の組が距離に与える効果を補正するという点に関しては、たとえば、個人を単位とするデータで補助変数に「身長」と「体重」と「肩からかかとまでの長さ」という3つの変数が採用されている場合と、「身長」と「体重」という2つの変数が採用されている場合では、前者の場合には正の相関が極めて強い2つの身長関連変数が体重の距離に対する効果を後者の場合と比べて弱くする。これに対してマハラノビス距離では、補助変数の標本分散共分散行列の逆行列非対角成分によって多重共線的な変数の組合せの効果を補正する。もちろん補助変数の組合せは分析者が決めるものなので、上記3変数のような無意味な組合せは採用されない。相関の強い補助変数の組が距離に与える効果を補正する必要がないと認

められるときは、行列 Q に用いる補助変数の標本分散共分散行列の非対角成分をすべて値 0 に置換えればよい。

最大値ノルム:

$$d(i,j) = \max_k \left| x_{ki} - x_{kj} \right|$$

上式の距離を用いる場合は、補助変数の単位を統一しなければならない。方法としては、変数ごとに標本標準偏差で割って基準化することや、標本内における大小関係の順位に変換することが考えられる。

予測平均(predictive mean)間距離:

$$d(i,j) = \left\{ \hat{y}(x_i) - \hat{y}(x_j) \right\}^2$$

ここで変数$\hat{y}(x_i)$は、一部に欠測が生じている変数Yから補助変数Xへの回帰モデルのレコードiにおける理論値である。

マッチングで結び付けられる相手の数の決め方に関しても種類がある。あらかじめ決められた値 k について、近いものから順に k 個の相手を結びつけ、それらのレコードの値の平均値を代入値とする場合は、「k-最近傍マッチング(k-nearest neighbor matching)」と呼ばれる。これに対して、あらかじめ決められた値 c について、距離が値 c を下回る相手を結びつけ、それらのレコードの値の平均値を代入値とする場合は、「キャリパーマッチング(caliper matching)」と呼ばれる。

補助変数の値によって、調査客体間の距離を測定する最近傍マッチングに対して、補助変数の値で条件付けた観測確率の推定値によって、距離を測定する方法が「傾向スコアマッチング(propensity score matching)」である。

補助変数X_iの値で条件付けた観測確率$p(X_i) \equiv Pr(R_i = 1|X_i)$は、観測指標$R_i$(目的となる変数$Y_i$が観測された場合に値1をとる2値変数)の補助変数$X_i$による「傾向スコア(propensity score)」と呼ばれる。傾向スコア$P(R = 1|X = x)$を言葉で表すと、「補助変数Xが特定の値xをとる場合に、目的となる変数Yが観測される確率」である。補助変数に何を用いるかによって傾向スコアの値は変わる。第 1.1.1 節 MAR の具体例(目的となる変数Yは所得額で補助変数Xは就業状態)では、「無業者は必ず回答し、有業者は 50%の確率で回答する」ので、無業者の傾向スコアの値は 1 であり($Pr(R_i = 1|X_i = \text{無業者}) = 1$)、有業者の傾向スコアの値は 0.5 である($Pr(R_i = 1|X_i = \text{有業者}) = 0.5$)。第 1.1.1 節 MCAR の極端な具体例(硬貨を投げて表が出れば回答し、裏が出れば回答しない)では、いかなる補助変数を用いても、すべての調査対象について傾向スコアの値は 0.5 である。

通常、観測指標の傾向スコアの値は知られていないので、データから推定する必要がある。傾向スコアの推定値は、観測指標を従属変数とし、補助変数を独立変数とする2項モデルの推定によって得られる。

傾向スコアマッチング代入法は、傾向スコアの推定値のみを補助変数とする最近傍マッチング代入法であるといえる。

補助変数に離散変数が含まれる場合、最近傍マッチングでは離散の補助変数の数が増えると指数的にカテゴリの数が増える(たとえば、性別と所在都道府県の2つの離散変数によって 2×47=94 のカテゴリが区別される。)ので、観測レコードと欠測レコードが各カテゴリに均等に分布しなければマッチングごとの近接性のばらつきが大きくなり、近距離マッチングと遠距離マッチングが区別されずに混在することになる。これに対して傾向スコアマッチングでは、傾向スコア(の推定値)というひとつの変量で距離を測る(すなわち傾向スコアの差(の絶対値)が距離となる)ので、遠距離マッチングの問題は緩和される。また、MAR の下では、傾向スコアの値のみで条件付けることで観測確率が欠測値に依存しなくなる(※ $f(R|Y,X)=f(R|X) \Rightarrow f\left(R\middle|Y,p(X)\right)=f\left(R\middle|p(X)\right)$ が成り立つ)ため、傾向スコアの値に基づく層化平均値代入でも欠測バイアスを緩和することができる。

○LOCF（last observation carried forward）・LVCF（last value carried forward）

同一の標本について複数時点にわたって変数の値を観測・記録することで得られるデータを「パネルデータ」と呼ぶ。パネルデータが作成される統計調査では、直近の観測値を代入値とすることが可能である。この方法は LOCF あるいは LVCF と呼ばれる。

2.2.1 各単一代入法の処理手順

各単一代入法の処理内容を理解するための例を、図2－2－1～2－2－8に示す。これらの図では、同一の不完全データに対して、それぞれの手法を適用している。例示に用いた不完全データは人工的に作成したものであり、レコード数が 20、変数の数が3(欠測指標を含めると4)である。実感をもたせるために例として、調査客体単位を個人とし、3つの変数は、変数yを今月末の対前月末体重変化分(kg)、変数x_1を前月末の対前々月末体重変化分(kg)、変数x_2を今月の対前月1日当たり運動量変化分(時間/日)とする。今月末の体重変化分yに欠測が生じ、前月末の体重変化分x_1及び今月の運動量変化分x_2には欠測が生じないとする。

○完全ケース分析

図2－2－1は、完全ケース分析の実行例を示したものである。20 人中7人で今月末の体重変化分yが観測されておらず、淡灰色で示している。これらの7人を除き、残った 13 人のみを用いて分析を行うのが完全ケース分析である。データセット2列目の変

数y^*は、体重変化分yの真の値であり、欠測となった 7 人の値も入っているが、実際は観測されていない。

ここで、人工的に作成した不完全データの性質をいくつか指摘しておく。第1に、今月末の体重変化分（の真の値）y^*と前月末の体重変化分x_1の間には正の相関、今月末の体重変化分（の真の値）y^*今月の運動量変化分x_2の間には負の相関がある。つまり、すべての調査客体について値が観測されているx_1とx_2は、y^*の値を予測するのに有用な情報を含んでいる。完全ケース分析は、このような有用な情報を一切用いておらず無駄にしているということができる。この無駄に対する「もったいない」という意識が、補助変数を用いた層化平均値代入法、回帰代入法、マッチング代入法等の手法の動機となっている。

第2に、図においてレコードは体重変化分y^*に関して昇順で並んでいるので、体重変化分y^*の値が大きい個人ほどその値が観測されていないという傾向がみてとれる。事実、この不完全データは MNAR の確率モデルにもとづいて作成している。この図では、体重変化分の真の値y^*をみせているため、欠測データメカニズムが MNAR であることが分かるが、実際には真の値は分からないので MNAR であるか否かを不完全データの情報から知ることはできない。

○平均値代入法

図2－2－2は、平均値代入法の実行例を示したものである。今月の体重変化分yについて値が観測されているレコードのみで平均値を求め、その値を 7 人分の欠測レコードの代入値とする。上述のとおり、この不完全データは MNAR 条件のもとで作成しているため、代入値となる平均値は、yの値が比較的小さい人たちの平均値であり、その小さい値を、実はyの値が大きい人たちに代入している様子が分かる。もしこの不完全データを MCAR 条件のもとで作成していれば、代入値となる平均値の算出では、体重変化分yの値が小さい人たちの側にも大きい人たちの側にも偏ることはなく、真の平均値に近い値が代入値となる。最後に、この平均値代入は、完全ケース分析と同様に、欠測データに含まれる有用な情報を一切用いていないことが分かる。

○層化平均値代入法

図2－2－3は、層化平均値代入法の実行例を示したものである。ここでは、運動量変化分x_2の値の4分位により標本を4分割している。変数 class_x_2は、x_2の値が第何分位に含まれるかを表している。4分割された部分標本のそれぞれについて、観測されているレコードのみで平均値を求め、その値を欠測値に代入する。

図の右側には4つの部分標本が示されている。第1の部分標本は、前月と比べて今月の運動量を大きく減らした人たちのグループであり、運動不足により今月は前月よりも体重が増えている可能性が高い。第2の部分標本は、運動量を減らしたものの、第1のグループほどではない人たちのグループであり、運動不足により体重が増えている

かもしれないが、その程度は第1のグループほどではないことが期待される。第3の部分標本は、運動量を大きくは変化させなかった人たちであり、体重にも大きな変化はない可能性が高い。第4の部分標本は、運動量を増やした人たちであり、体重は減っていると予想される。そして、たまたま第4の部分標本では、今月の体重変化分yの値が欠測となっている人はいないため、運動量変化分にもとづく上述の予想を確かめることができる。確かに、第4の部分標本では概ね体重が減少している。

層化は、似た者同士を同じグループにまとめるので、欠測値への代入値が似た者同士で似た値となる。このとき、何に関して似ているかが重要である。欠測する変数と相関の高い変数に関して似ているほうが、代入値は真の値に近いものとなる。運動量変化分x_2や前月の体重変化分x_1は、今月の体重変化分（の真の値）y^*と相関があるものの例だが、たとえば、「各人の連絡先電話番号の最後の3桁の数字」といった変数の値で層化すると、体重変化分とは無関係なことに関して似た者同士に似た値を代入することになる。欠測となった7人が、体重変化分に関して意味のある区別をされずに分割されるので、たとえばid=19と20の人を互いに似た者同士とはせずに、id=3と20の人を互いに似た者同士とする余地がでてくる。

この例では、今月の運動量変化分x_2の値に基づいて層化しているが、もちろん、前月の体重変化分x_1を層化に用いてもよいし、また前月の体重変化分x_1と運動量変化分x_2の両方を層化に用いてもよい。理屈としては、MNAR に対しては、欠測を生じる変数の真の値y^*に対する予測力が最も大きい変数ないし変数の組合せを層化に用いるのがよい。*ただし、実際にはどの変数ないし変数の組合せで真の値y^*に対する予測力が大きいかはデータから知ることはできない。この例では、1期前の体重変化分x_1と運動量変化分x_2の両方を組み合わせた方が今期の体重変化分（の真の値）y^*の予測力が大きいことが常識的に判断できる（図ではそうしていないが）。このように、アプリオリな知見の活用も欠測データ処理には重要である。*最後に、層化平均値代入法では、完全ケース分析や平均値代入法で無駄にされていた情報が活用されているといえる。

〇回帰代入法

図2－2－4は、回帰代入法の実行例を示したものである。まず不完全データに対して、今月の体重変化分yを被説明変数とし、前月の体重変化分x_1及び運動量変化分x_2を説明変数とする回帰分析を（完全ケース分析により）行う。そこで推定された回帰モデル（図の例では線形回帰モデル）にもとづいて、欠測を出した7人についてx_1及びx_2の値からyの理論値を代入値とする。層化平均値代入法と同様に回帰代入法では、完全ケース分析や平均値代入法では無駄にされていた情報（x_1及びx_2）が、代入値を回帰モデルの理論値として算出する段階で活用されている。

図中右上部分のグラフは、横軸に回帰分析の理論値、縦軸に観測値をとった散布図である。灰色の点は今月の体重変化分yが観測された13人を表し、黒色の点は（の

真の値）y^*が観測されなかった7人を表している。この7人については、縦軸の座標が分からないので、回帰分析の理論値を観測値の代わりに縦軸座標としている。完全ケースの 13 人については、縦軸は観測値を表し、欠測となった7人に関しては、縦軸は横軸と同じく理論値であるから、欠測となった7人の点はグラフ中の点線で示した45度線上に位置する。回帰モデルのパラメータの推定値は、灰色の点について 45 度線からの垂直距離の平方和を最小化する値である。こうして得られたパラメータ推定値に基づいて欠測となった7人の体重変化分yの理論値が計算される、つまり黒色点の 45 度線上での位置が決まる。

回帰代入法も似た者同士には似た値を代入値とするという点で、上述の層化平均値代入法と同じである。回帰代入の場合、何に関する類似性かといえば、説明変数に関する類似性である。*もっとも、層化平均値代入法は、回帰代入法の特殊例とみることもできる。すなわち、層への所属を表すダミー変数を説明変数とする線形回帰モデルによる回帰代入法が層化平均値代入である。*

〇確率的回帰代入法

図2－2－5は、確率的回帰代入法の実行例を示したものである。図2－2－6の回帰代入法と同様に、まず不完全データに対して、今月の体重変化分yを被説明変数とし、前月の体重変化分x_1及び運動量変化分x_2を説明変数とする回帰分析（完全ケース分析により）を行う。そこで推定された回帰モデル（図の例では線形回帰モデル）にもとづいて、欠測を出した7人についてx_1及びx_2の値からyの理論値を求める。確率的回帰代入法では、推定された回帰モデルの誤差項が従う分布から乱数を発生させ、得られた値を回帰モデルの理論値に加えたものを代入値とする。この点が、回帰代入法と異なる。

図中右上部分のグラフを図2－2－6のものと比較すると、欠測となった 7 人を表す黒色点は45 度線からランダムに乖離する。各黒色点の45 度線からの垂直距離がそれぞれに実現した誤差項の値である。この図によると、id=16 の人には正で比較的大きな誤差項が発生しており、id=20 の人には負で比較的大きな誤差項が発生している。つまりid=16 の人には理論的に予測されるよりも大きな体重の増加をあてがい、id=20 の人には理論的に予測されるよりも小さな体重の増加をあてがっているのであるが、このような個々人の代入値にみられる理論値からの乖離自体は無作為に乱数として発生させたものである。それでも、標本全体でこれらの個別の効果が、大数の法則により、互いに打消し合うので、確率的回帰代入法により得られた疑似完全データによる1次モーメントの推定は回帰代入法の結果と漸近的に等価である。それではなぜ誤差項を代入値に加算するのかというと、それは回帰代入法にともなう推定精度の過大評価及び1次よりも大きいモーメントの推定における欠測バイアスが、誤差項を加算することによって緩和されるからである。

〇マッチング代入法（最近傍マッチング）

図2－2－6は、最近傍マッチング代入法の実行例を示したものである。最近傍マッチングでは、変数yが欠測している調査客体i、及び変数yが観測されている調査客体jとの間の類似性を、調査客体iの補助変数の値x_i及び調査客体jの補助変数の値x_jの間の距離によって測り、距離が十分小さいもの同士を結び合わせる。そして、結び合わされた観測レコードと欠測レコードで、観測レコードの値を欠測レコードの欠測値に代入する。マッチング代入法でも、完全ケース分析や平均値代入法では無駄にされていた情報（x_1及びx_2）が、回答者と無回答者の間の距離を算出する段階で活用されている。

前月の体重変化分x_1と運動量変化分x_2という2つの補助変数があるので、2次元実数ベクトル空間で、今月の体重変化分yの値が欠測となった7人と、欠測とはならなかった 13 人の間の距離を総当たりで測る。つまり、7×13＝91 通りの組み合わせについて両者間の距離を計算する。図の例では距離概念の定義としてマハラノビス距離を用いている。欠測となった7人のそれぞれに組み合わせの相手となる候補が 13 人いるが、そのなかで最も距離の近い者と組み合わされる。計算の結果、たとえば id=3 の無回答者には id=8 の回答者が最も近い者、つまり最も似ている者として結び合わされている。両者間の距離はマハラノビス距離で 0.67 である。結ばれたもの同士の距離で両者の類似性を測るので、たとえば、id=16 の無回答者と id=12 の回答者の類似性よりも、id=13 の無回答者と id=14 の回答者の類似性のほうが大きいということも分かる（距離の値 1.44 と 0.43 の比較）。

マッチング代入法で、欠測を出したレコードと結びついて代入値を提供するレコードを「ドナー」と呼ぶ。id=14 の回答者が id=10 の無回答者にも id=13 の無回答者にも結ばれているが、このように、同じ回答者が複数の無回答者のドナーとなることがある。また、ここでの例のようにひとりの無回答者のドナーとなる回答者の数はひとりと限る必要はなく、複数のドナーを結びつける場合は、それらの平均値を代入値とするなどの処理も考えることができる。

〇マッチング代入法（傾向スコアマッチング）

図2－2－7は、傾向スコアマッチング代入法の実行例を示したものである。傾向スコアマッチング代入法では、まず標本全体を用いて傾向スコアを推定する。傾向スコアは、補助変数の値によって条件付けた観測確率である。標本全体、すなわち 20 人分すべてのレコードを用いて観測確率を2つの補助変数（前月の体重変化分x_1と運動量変化分x_2）で説明する2項回帰モデルを推定し、全員についてそれぞれの傾向スコアを得ることができる。

たとえば、図2－2－7によると、id=1 の人の傾向スコアは 0.99 であるが、これは、id=1 の人の前月の体重変化分x_1（約 0.66kg 減）と運動量変化分x_2（約 0.60 単位増）から推定した結果、id=1 の人は約 99%の確率で、今月の体重変化分（の真の値）y*を回答する、ということを意味している。そして実際、id=1 の人は回答している。一方、id=3 の人は、約 52%の確率で体重変化分（の真の値）y*を回答するとの推定結果であったにもかかわらず（それほど低くはない確率で回答する条件を備えていた人であったが）、結果的には回答してくれなかった。

次に、7人の無回答者と 13 人の回答者の間で傾向スコアの差の絶対値を総当たりで計算し、7人の無回答者それぞれについてその値が最も小さい回答者を結びつける。たとえば id=3 の無回答者は、回答者の中では id=14 の人と傾向スコアの値が最も近く、その差は約 18%ポイントである。つまり傾向スコアマッチング代入法では、観測確率に関して似た者同士を結び合わせている。無回答者の欠測値には、結ばれた相手の回答者の値を代入値とする。

○LOCF

図2－2－8は、LOCF の実行例を示したものである。LOCF は、他の単一代入の手法と異なり、パネルデータにのみ適用できる手法である。この例では、補助変数の中に欠測を生じる変数yの1期前の値x_1がすべての人について観測されているため、欠測値には 1 期前の値を代入値とする。たとえば、id=3 の調査客体は、今月の体重変化分yは観測されていないが、前月の体重変化分x_1が観測されているので、LOCF ではその値（0.59kg）を今月の体重変化分yに代入する。

今月の体重変化分の値が大きい人ほど欠測となっているので、欠測バイアスは今月の体重変化分の値が小さい側へのバイアスとなるが、LOCF がこの欠測バイアスを緩和するのは、前月と今月の体重変化分が正の相関をもつときだけである。これは他の単一代入法にはない LOCF 独自の性質である。

◇まとめ

各単一代入法を比較すると、第1に、層化平均値代入法、回帰代入法、確率的回帰代入法及びマッチング代入法は、完全ケース分析や平均値代入法では無駄に捨てられていた情報の活用が図られている。パネルデータにおいて、直近に観測された値を代入値とする LOCF も、完全ケース等と比べて情報の活用が図られている。

第2に、層化平均値代入法、回帰代入法、確率的回帰代入法及びマッチング代入法は、「似た者同士は似た値をとる」という発想になじんでいる。そこでは、補助変数に関して類似していれば欠測する変数に関しても類似しているはずであるという推定原理が働いている。この推定原理は明らかに、補助変数と欠測する変数の相関が高い

ほど正しい推定を導く。LOCF については、「似た者同士」の相手が自分自身ということになるが、他の単一代入法にはない LOCF 独自の問題（上述）を招かないためには今期と前期が似た状況であることが重要である。

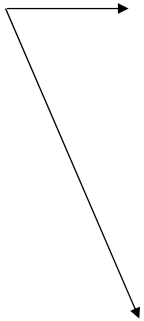
第3に、確率的回帰代入法は回帰代入法にともなう推定精度の過大評価及び1次よりも大きいモーメントの推定における欠測バイアスを緩和する。この点については、後でさらに詳しくみる。

欠測レコードの情報を無駄にせず、また補助変数の説明力を活用するという点は、層化平均値代入法、回帰代入法、確率的回帰代入法及びマッチング代入法の完全ケース分析及び単純な平均値代入法に対する優位性である。

図２－２－１ 完全ケース分析の処理手順

不完全データ

id	y*	y	missing	x_1	x_2
1	-1.49	-1.49	0	-0.66	0.60
2	-1.25	-1.25	0	0.30	1.93
3	-0.83		1	0.59	0.31
4	-0.39	-0.39	0	-0.42	-0.51
5	-0.28	-0.28	0	-1.69	0.09
6	-0.26	-0.26	0	-0.84	0.35
7	-0.21	-0.21	0	-1.01	-0.41
8	-0.18	-0.18	0	0.06	0.45
9	0.10	0.10	0	-0.44	-0.44
10	0.15		1	0.18	-0.40
11	0.18	0.18	0	-0.01	0.57
12	0.80	0.80	0	0.70	-0.83
13	0.81		1	0.33	-0.01
14	0.81	0.81	0	0.61	-0.28
15	0.86	0.86	0	-0.66	-1.40
16	1.05		1	1.78	-0.68
17	1.09	1.09	0	0.00	0.20
18	1.33		1	0.16	-1.86
19	1.41		1	0.62	-1.61
20	1.43		1	0.97	-1.12



分析に用いる情報

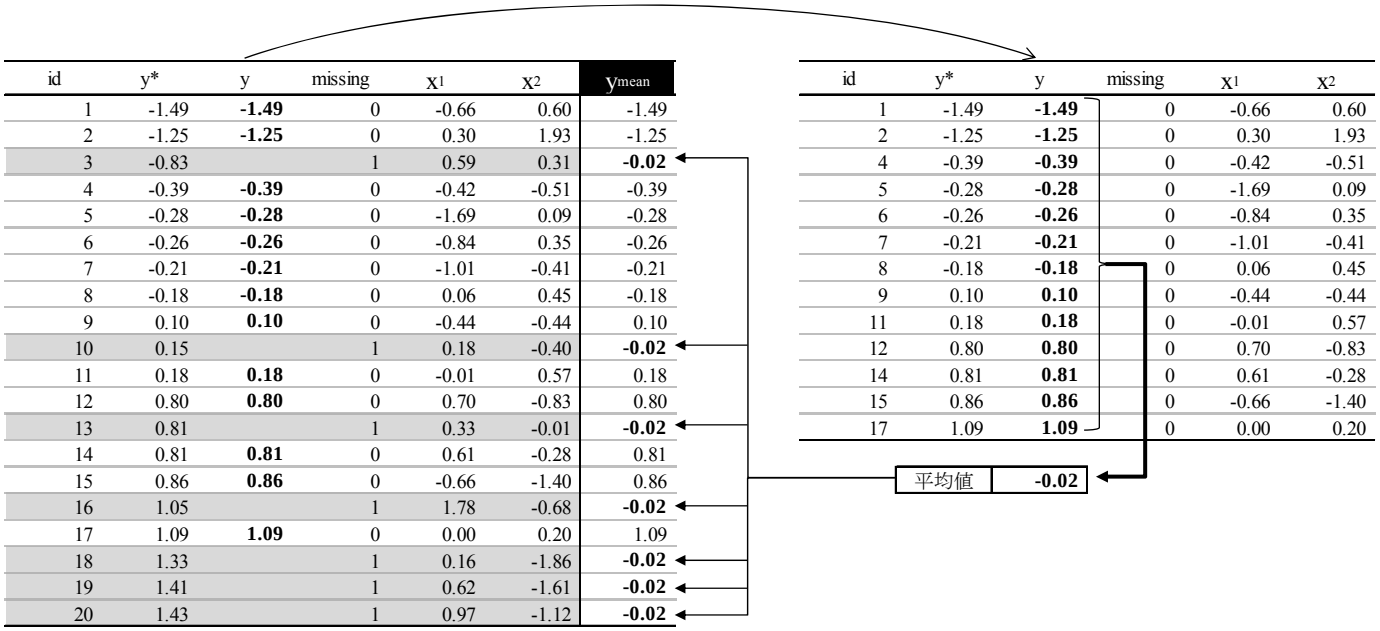
id	y*	y	missing	x_1	x_2
1	-1.49	-1.49	0	-0.66	0.60
2	-1.25	-1.25	0	0.30	1.93
4	-0.39	-0.39	0	-0.42	-0.51
5	-0.28	-0.28	0	-1.69	0.09
6	-0.26	-0.26	0	-0.84	0.35
7	-0.21	-0.21	0	-1.01	-0.41
8	-0.18	-0.18	0	0.06	0.45
9	0.10	0.10	0	-0.44	-0.44
11	0.18	0.18	0	-0.01	0.57
12	0.80	0.80	0	0.70	-0.83
14	0.81	0.81	0	0.61	-0.28
15	0.86	0.86	0	-0.66	-1.40
17	1.09	1.09	0	0.00	0.20

分析には用いない情報

id	y*	y	missing	x_1	x_2
3	-0.83		1	0.59	0.31
10	0.15		1	0.18	-0.40
13	0.81		1	0.33	-0.01
16	1.05		1	1.78	-0.68
18	1.33		1	0.16	-1.86
19	1.41		1	0.62	-1.61
20	1.43		1	0.97	-1.12

y*: **真の値**、y: **観測データ**、missing: **欠測指標**、(x_1, x_2): **補助変数**

図２－２－２ 平均値代入法の処理手順



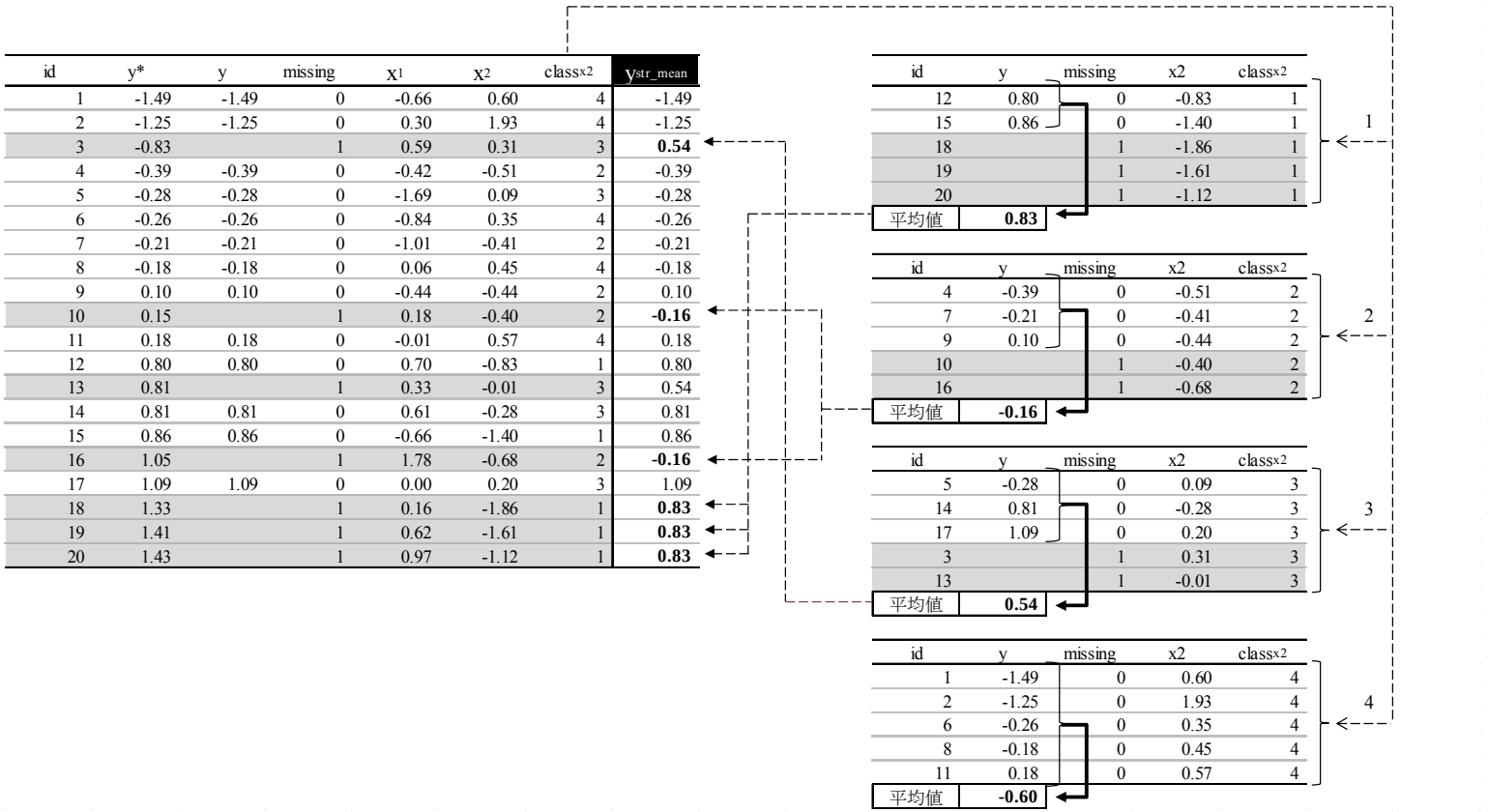
id	y*	y	missing	x_1	x_2	y_{mean}
1	-1.49	**-1.49**	0	-0.66	0.60	-1.49
2	-1.25	**-1.25**	0	0.30	1.93	-1.25
3	-0.83		1	0.59	0.31	**-0.02**
4	-0.39	**-0.39**	0	-0.42	-0.51	-0.39
5	-0.28	**-0.28**	0	-1.69	0.09	-0.28
6	-0.26	**-0.26**	0	-0.84	0.35	-0.26
7	-0.21	**-0.21**	0	-1.01	-0.41	-0.21
8	-0.18	**-0.18**	0	0.06	0.45	-0.18
9	0.10	**0.10**	0	-0.44	-0.44	0.10
10	0.15		1	0.18	-0.40	**-0.02**
11	0.18	**0.18**	0	-0.01	0.57	0.18
12	0.80	**0.80**	0	0.70	-0.83	0.80
13	0.81		1	0.33	-0.01	**-0.02**
14	0.81	**0.81**	0	0.61	-0.28	0.81
15	0.86	**0.86**	0	-0.66	-1.40	0.86
16	1.05		1	1.78	-0.68	**-0.02**
17	1.09	**1.09**	0	0.00	0.20	1.09
18	1.33		1	0.16	-1.86	**-0.02**
19	1.41		1	0.62	-1.61	**-0.02**
20	1.43		1	0.97	-1.12	**-0.02**

id	y*	y	missing	x_1	x_2
1	-1.49	**-1.49**	0	-0.66	0.60
2	-1.25	**-1.25**	0	0.30	1.93
4	-0.39	**-0.39**	0	-0.42	-0.51
5	-0.28	**-0.28**	0	-1.69	0.09
6	-0.26	**-0.26**	0	-0.84	0.35
7	-0.21	**-0.21**	0	-1.01	-0.41
8	-0.18	**-0.18**	0	0.06	0.45
9	0.10	**0.10**	0	-0.44	-0.44
11	0.18	**0.18**	0	-0.01	0.57
12	0.80	**0.80**	0	0.70	-0.83
14	0.81	**0.81**	0	0.61	-0.28
15	0.86	**0.86**	0	-0.66	-1.40
17	1.09	**1.09**	0	0.00	0.20

平均値	**-0.02**

y*: 真の値、y: 観測データ、missing: 欠測指標、(x_1, x_2): 補助変数、y_{mean}: 平均値代入による代入値

図２－２－３ 層化平均値代入法の処理手順



id	y*	y	missing	x_1	x_2	$class_{x2}$	y_{str_mean}
1	-1.49	-1.49	0	-0.66	0.60	4	-1.49
2	-1.25	-1.25	0	0.30	1.93	4	-1.25
3	-0.83		1	0.59	0.31	3	**0.54**
4	-0.39	-0.39	0	-0.42	-0.51	2	-0.39
5	-0.28	-0.28	0	-1.69	0.09	3	-0.28
6	-0.26	-0.26	0	-0.84	0.35	4	-0.26
7	-0.21	-0.21	0	-1.01	-0.41	2	-0.21
8	-0.18	-0.18	0	0.06	0.45	4	-0.18
9	0.10	0.10	0	-0.44	-0.44	2	0.10
10	0.15		1	0.18	-0.40	2	**-0.16**
11	0.18	0.18	0	-0.01	0.57	4	0.18
12	0.80	0.80	0	0.70	-0.83	1	0.80
13	0.81		1	0.33	-0.01	3	0.54
14	0.81	0.81	0	0.61	-0.28	3	0.81
15	0.86	0.86	0	-0.66	-1.40	1	0.86
16	1.05		1	1.78	-0.68	2	**-0.16**
17	1.09	1.09	0	0.00	0.20	3	1.09
18	1.33		1	0.16	-1.86	1	**0.83**
19	1.41		1	0.62	-1.61	1	**0.83**
20	1.43		1	0.97	-1.12	1	**0.83**

id	y	missing	x2	$class_{x2}$	
12	0.80	0	-0.83	1	
15	0.86	0	-1.40	1	1
18		1	-1.86	1	
19		1	-1.61	1	
20		1	-1.12	1	
平均値	**0.83**				

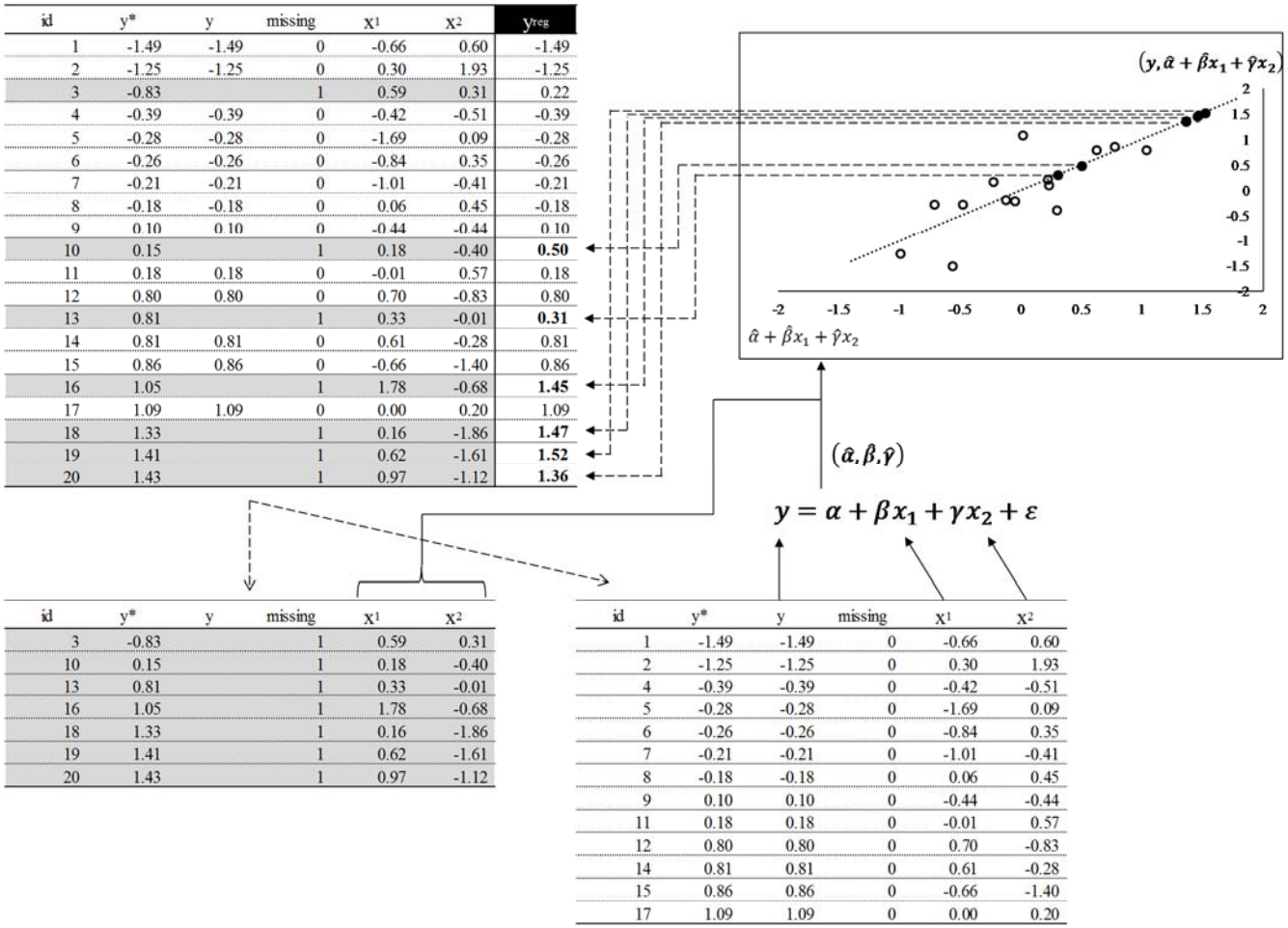
id	y	missing	x2	$class_{x2}$	
4	-0.39	0	-0.51	2	
7	-0.21	0	-0.41	2	2
9	0.10	0	-0.44	2	
10		1	-0.40	2	
16		1	-0.68	2	
平均値	**-0.16**				

id	y	missing	x2	$class_{x2}$	
5	-0.28	0	0.09	3	
14	0.81	0	-0.28	3	3
17	1.09	0	0.20	3	
3		1	0.31	3	
13		1	-0.01	3	
平均値	**0.54**				

id	y	missing	x2	$class_{x2}$	
1	-1.49	0	0.60	4	
2	-1.25	0	1.93	4	4
6	-0.26	0	0.35	4	
8	-0.18	0	0.45	4	
11	0.18	0	0.57	4	
平均値	**-0.60**				

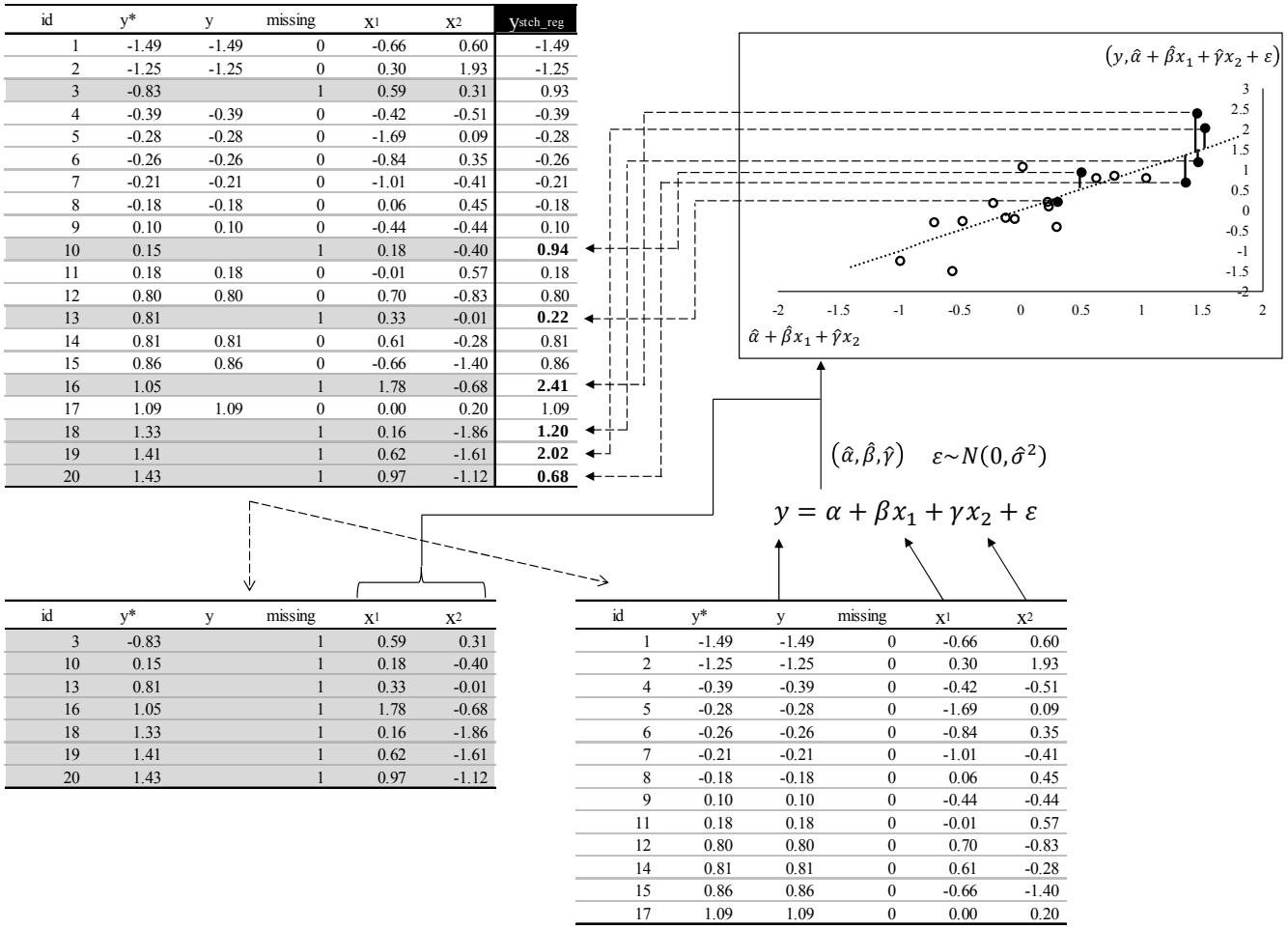
y*: 真の値、y: 観測データ、missing: 欠測指標、(x_1, x_2): 補助変数、$class_{x2}$: 補助変数 x_2 の４分位階層、y_{str_mean}: 補助変数 x_2 にもとづく層化平均値代入による代入値

図２－２－４ 回帰代入法の処理手順



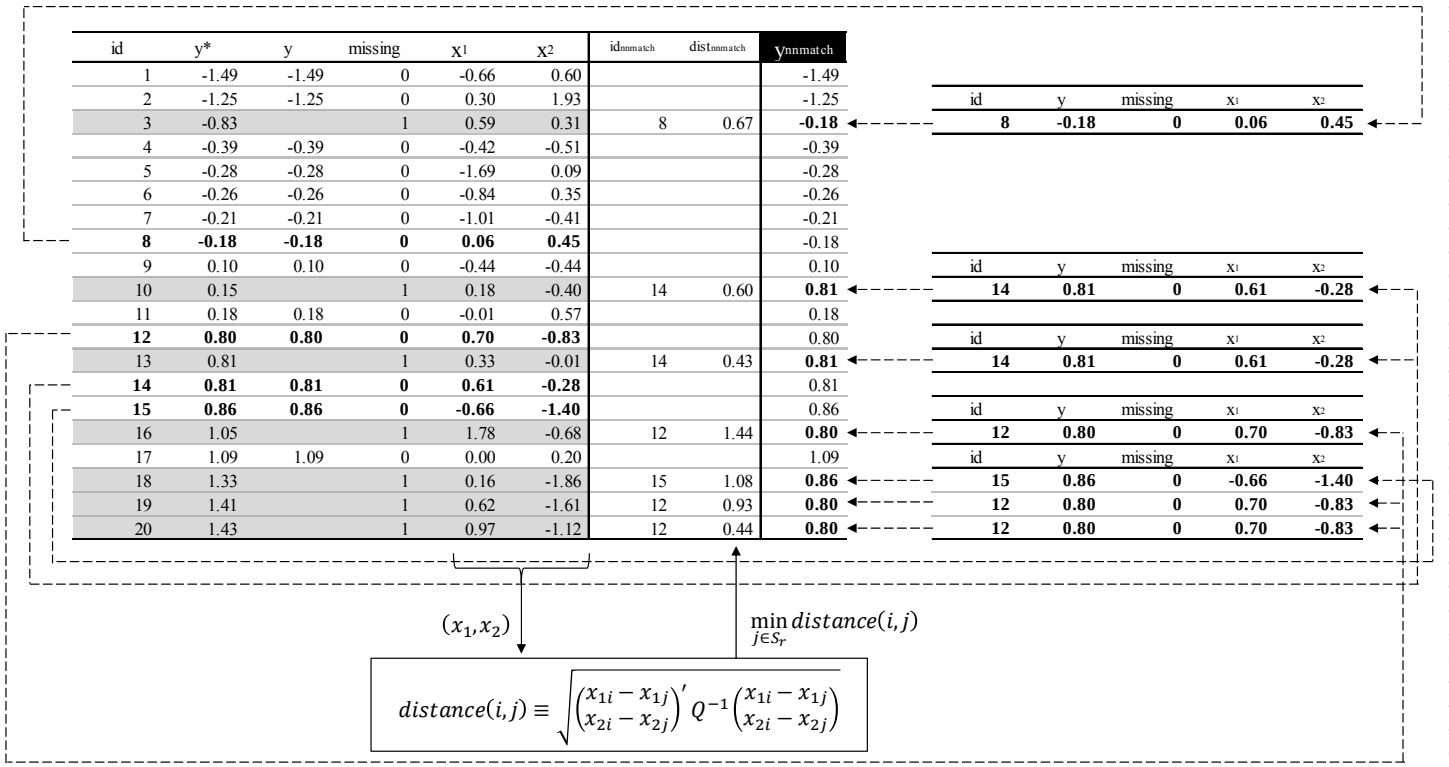
y*: 真の値、y: 観測データ、missing: 欠測指標、(x_1, x_2): 補助変数、y_{reg}: 回帰代入による代入値

図２－２－５ 確率的回帰代入法の処理手順



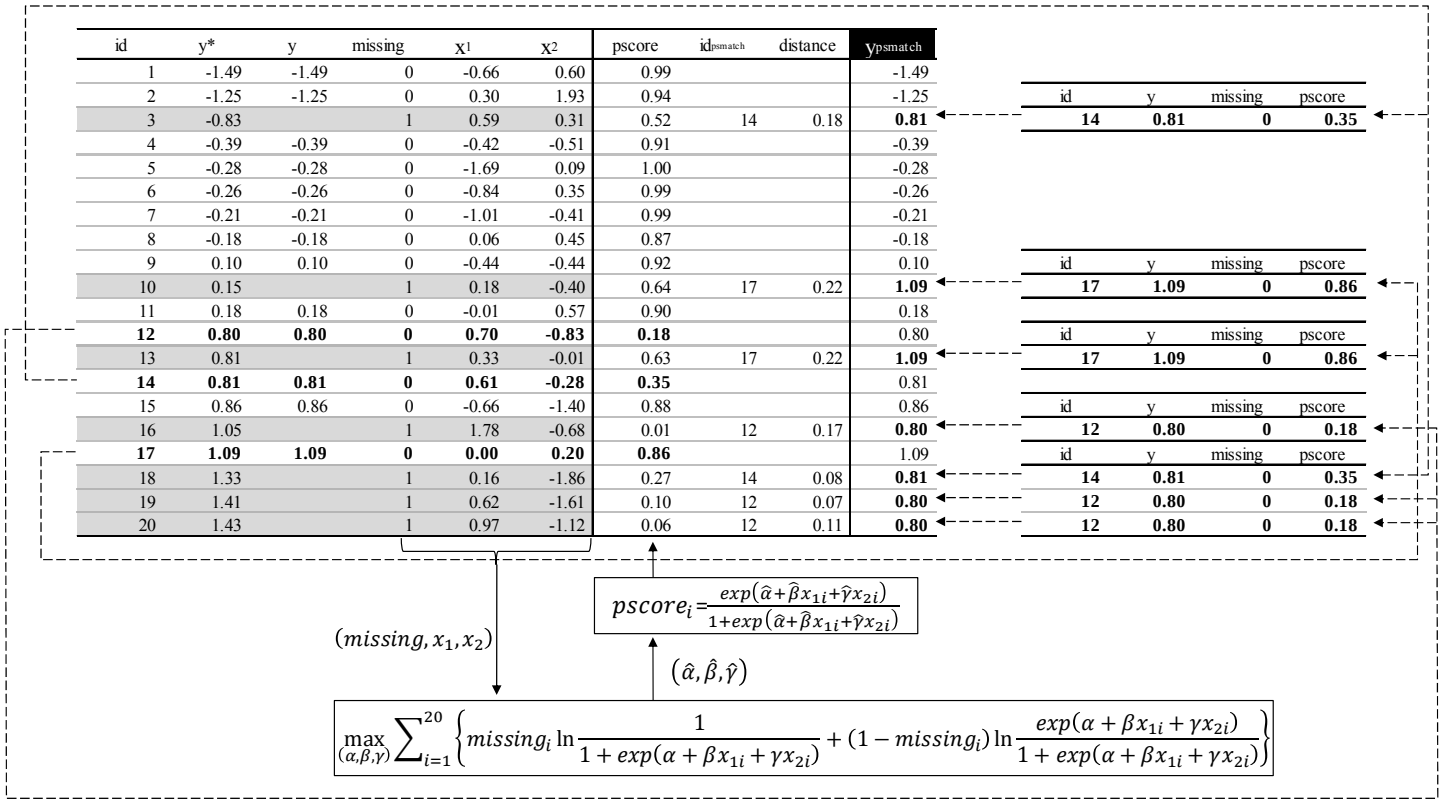
y*: 真の値、y: 観測データ、missing: 欠測指標、(x_1, x_2): 補助変数、y_{stch_reg}: 確率的回帰代入による代入値

図２－２－６ 最近傍マッチング代入法の処理手順（図は 1 対 1 マッチング）



y*: 真の値、y: 観測データ、missing: 欠測指標、(x_1, x_2): 補助変数、$id_{nnmatch}$: 最近傍マッチングによって合された相手レコードの id、$dist_{nnmatch}$: マッチングの相手との間の距離（行列 Q により距離概念を定義する）、$y_{nnmatch}$: 最近傍マッチング代入による代入値

図２－２－７ 傾向スコアマッチング代入法の処理手順（図は 1 対 1 マッチング）



y*: 真の値、y: 観測データ、missing: 欠測指標、(x_1, x_2): 補助変数、pscore: 傾向スコア、$id_{psmatch}$: 傾向スコアマッチングによって合された相手レコードの id、distance: マッチングの相手との間の距離（傾向スコアの差）、$y_{psmatch}$: 傾向スコアマッチング代入による代入値（数式はロジットモデル）

図２－２－８ LOCF の処理手順

id	y*	y	missing	x_1	x_2	y_{LOCF}
1	-1.49	-1.49	0	-0.66	0.60	-1.49
2	-1.25	-1.25	0	0.30	1.93	-1.25
3	-0.83		1	**0.59**	0.31	**0.59**
4	-0.39	-0.39	0	-0.42	-0.51	-0.39
5	-0.28	-0.28	0	-1.69	0.09	-0.28
6	-0.26	-0.26	0	-0.84	0.35	-0.26
7	-0.21	-0.21	0	-1.01	-0.41	-0.21
8	-0.18	-0.18	0	0.06	0.45	-0.18
9	0.10	0.10	0	-0.44	-0.44	0.10
10	0.15		1	**0.18**	-0.40	**0.18**
11	0.18	0.18	0	-0.01	0.57	0.18
12	0.80	0.80	0	0.70	-0.83	0.80
13	0.81		1	**0.33**	-0.01	**0.33**
14	0.81	0.81	0	0.61	-0.28	0.81
15	0.86	0.86	0	-0.66	-1.40	0.86
16	1.05		1	**1.78**	-0.68	**1.78**
17	1.09	1.09	0	0.00	0.20	1.09
18	1.33		1	**0.16**	-1.86	**0.16**
19	1.41		1	**0.62**	-1.61	**0.62**
20	1.43		1	**0.97**	-1.12	**0.97**

y*: 真の値、y: 観測データ、missing: 欠測指標、(x_1, x_2): 補助変数、y_{LOCF}: LOCF による代入値

2.2.2 各単一代入法の特徴

次に、各単一代入法の特徴を理解するために、不完全データの数値例を図２－３－１に示す。目標母集団から単純無作為抽出した 1,000 人の標本について、各調査客体の体重の前月差の値を２時点にわたって収集する調査を考える。第１時点の体重の前月差（kg）を W1、第２時点の体重の前月差（kg）を W2 とする（実は図２－２－１～８で用いた数値例のデータを発生させたモデルと同じである）。幸運にも第１時点の値はすべての調査客体について観測されたが、第２時点の値は一部の調査客体について欠測が生じたとする。この調査から作成される不完全データに、各単一代入法を適用する（図では参考までに、第 2.5 節で取り上げる多重代入法の適用についても示している。）。この場合、第２時点の体重前月差 W2 が目的の変数、第１時点の体重前月差 W1 が利用可能な補助変数となる。ここでは、目的となる変数と補助変数との間に正の相関がある場合を考える。

パネル(A)の上図は、第２時点の体重前月差 W2 が観測されなかった調査客体についても、真の値が得られたときに作成される W2 のヒストグラムである。ヒストグラムでは実際に観測されたデータの部分は淡灰色、欠測となったデータの部分は濃灰色で示し区別している。パネル(A)の下図は、第２時点と第１時点の体重前月差の散布図である。散布図では、W2 が実際に観測された調査客体は記号〇、欠測となった調査客体は記号△で示し区別している。誰もパネル(A)のような真の姿を知ることはできない。

パネル(A)をみると、第２時点の体重前月差が大きい調査客体ほど欠測する割合が高い。この点だけを考えると、欠測の割合が欠測する変数に依存しているので、欠測データメカニズムは MNAR である。しかしここでは、第２時点の体重前月差が、補助変数として利用可能な第１時点の体重前月差と正の相関を示している。このため、この補助変数で層化すれば、層ごとの欠測割合は欠測する変数の値に依存しない。つまり、条件付けに用いることで、欠測する変数の値と欠測確率との相関を消すことができるという性質をもつ変数（第１時点の体重前月差）が補助変数として利用可能なので、欠測データメカニズムは MAR とみなせる。

〇平均値代入法

パネル(A)に示す不完全データ（淡灰色部分）に平均値代入法を適用した結果を、パネル(B)に示す。第２時点の体重前月差 W2 が大きい調査客体ほど欠測する割合が高いため、W2 が小さい側へ偏った部分標本によって計算される平均値が代入値となる。このため疑似完全データの標本平均による母集団平均の推定

には下方バイアスが生じる。

また、パネル(A)のヒストグラムで濃灰色に示されたレコードは、観測されたW2 の平均値という一点に集められ、パネル(B)のヒストグラムのように高頻度帯（点）を形成する。これが、１次よりも大きい母集団モーメントの推定における過小バイアス及び平均値代入法による推定精度の過大評価の原因となる。

パネル(B)の散布図を真の姿であるパネル(A)の散布図と比べると、欠測となった調査客体について第２時点の体重前月差を一律に一定値（観測値の平均値）とすることの副作用がみてとれる。真の姿では、この目標母集団は、第１時点と第２時点の体重前月差に比較的高い正の相関がある（パネル(A)の散布図）。しかし、W2 の値が大きい調査客体でより多くの欠測が発生したため、平均値代入法による疑似完全データでは、W1 が最も大きい階層に属する調査客体で、W2 の代入値が実際よりも小さい値となっている。平均値代入法では、W2 の分布をゆがめるだけでなく、W2 と W1 との関係性までもゆがめてしまう。

〇層化平均値代入法

パネル(B)にみるような平均値代入法の問題点は、標本を適当な補助変数により層化することで緩和される。パネル(A)の不完全データに層化平均値代入法を適用した結果を、パネル(C)に示す。これは、第１時点の体重前月差 W1 の４分位点を境に標本を４層分割し、各層ごとに層内の観測データを用いて平均値代入を行ったものである。層が４つに分かれたため、パネル(C)とパネル(B)のヒストグラムを比べると、パネル(B)にみられる高頻度点は１つから４つに分散している。この分散の効果は、層の数を増やすほどより大きくなる。

〇回帰代入法

補助変数による標本の層化からさらに進んで、補助変数の値ごとにモデルに基づく推定値を代入する方法が回帰代入法である。回帰代入法の結果を、パネル(D)に示す。この例では、W2 が観測されたレコードについて、W2 を W1 へ回帰するモデルを推定し、推定されたモデルに基づく欠測レコードの理論値を代入値としている。平均値代入の場合にヒストグラムに出現していた高頻度点の問題は、解消しているようにみえる。

散布図については、パネル(B)や(C)よりは真の姿に多少近づいているようにみえるとはいえ、欠測データに対応するレコードが、疑似完全データでは回帰直線上に固定されており、回帰直線からの乖離が無視されていることが分かる。少なくとも、このばらつきが取り除かれている分だけでも、推定精度が過大評価されることになる。そこでこの点を考慮に入れた手法として、次のパネル(E)に示す確率的回帰代入法を考えることができる。

〇確率的回帰代入法

パネル(E)には、確率的回帰代入法の結果を示す。ここでは、パネル(D)の代入値に乱数発生させた誤差項を加えている。誤差項は正規分布に従い、その標準偏差は回帰推定の残差から推定した値を用いている。パネル(D)と比べて、パネル(E)では回帰直線からランダムに乖離するので、代入値のばらつきが大きくなり、ヒストグラムも散布図も真の姿により近づいているようにみえる。このように、誤差項を代入値に加算することで、代入を施した変数の分布のばらつきが維持されるため、確率的回帰代入は１次よりも大きいモーメントの推定について回帰代入よりも優れている。

〇マッチング代入法

パネル(F)及び(G)には、それぞれ最近傍マッチング代入法及び傾向スコアマッチング代入法の結果を示す。マッチング代入の結果は、「（補助変数に関して）似た者同士は（欠測値でも）似た値をとる」という回帰代入の性質を共有しつつ、回帰代入の結果が回帰直線上に集中するのと比べて、マッチング代入の結果はばらつきを保っている。しかし一部に平均値代入の結果でみられた高頻度点がみられる。欠測が起こる変数 W2 と補助変数 W1 に正の相関がある条件の下で、W1 と W2 が共に大きい値をとる領域（散布図の右上側領域）で欠測率が高くなっており、この領域では代入値を求める欠測レコードに対して、代入値を与えてくれるマッチングの相手が希少になっている。このため当該領域では、特定の観測レコードの値が代入値として頻繁に利用され、散布図及びヒストグラムにみられるとおり、代入値の高頻度点が生じることになる。

〇LOCF

パネル(H)には、LOCF の結果を示す。すなわち、欠測した W2 には W1 の値を代入している。この例では、W1 と W2 の相関が比較的に高いため、LOCF によって作成されるヒストグラムも真の姿に近いとみえる。系列相関が負となるような例（後述の図２－３－２）だと、逆の結果をもたらす。**すなわち、ヒストグラムの欠測部分は LOCF によって反転する（大きい欠測値には小さい代入値、小さい欠測値には大きい代入値）。また、相関図は当然ながら、欠測データの部分は 45 度線上に固定される。**

LOCF は、回帰代入法をパネルデータに適応した場合の特殊形である。パネル(H)LOCF に示す結果は、パネル(D)回帰代入法において定数項 0 及び補助変数の係数 1 という極めて厳しい線形制約を課したものとみることもできる。

〇補助変数との相関が負の場合

図２－３－１では、第１時点の体重前月差 W1 と第２時点の体重前月差 W2 が正の相関関係にある場合の例となっているが、W1 と W2 が負の相関関係にある場合の例を図２－３－２に示す。補助変数との相関が正である場合と負である場合で疑似完全データの分布を比較すると、LOCF を用いた場合に違いがみられる。他の単一代入法については、W1 と W2 との相関の正負にかかわらず図２－３－１で指摘した点が成り立つ。つまり、目的となる変数との相関が正であれ負であれ、補助変数が観測確率に説明力をもてば、その補助変数を利用した単一代入法によって MAR の下での推定における欠測バイアスを緩和できる。

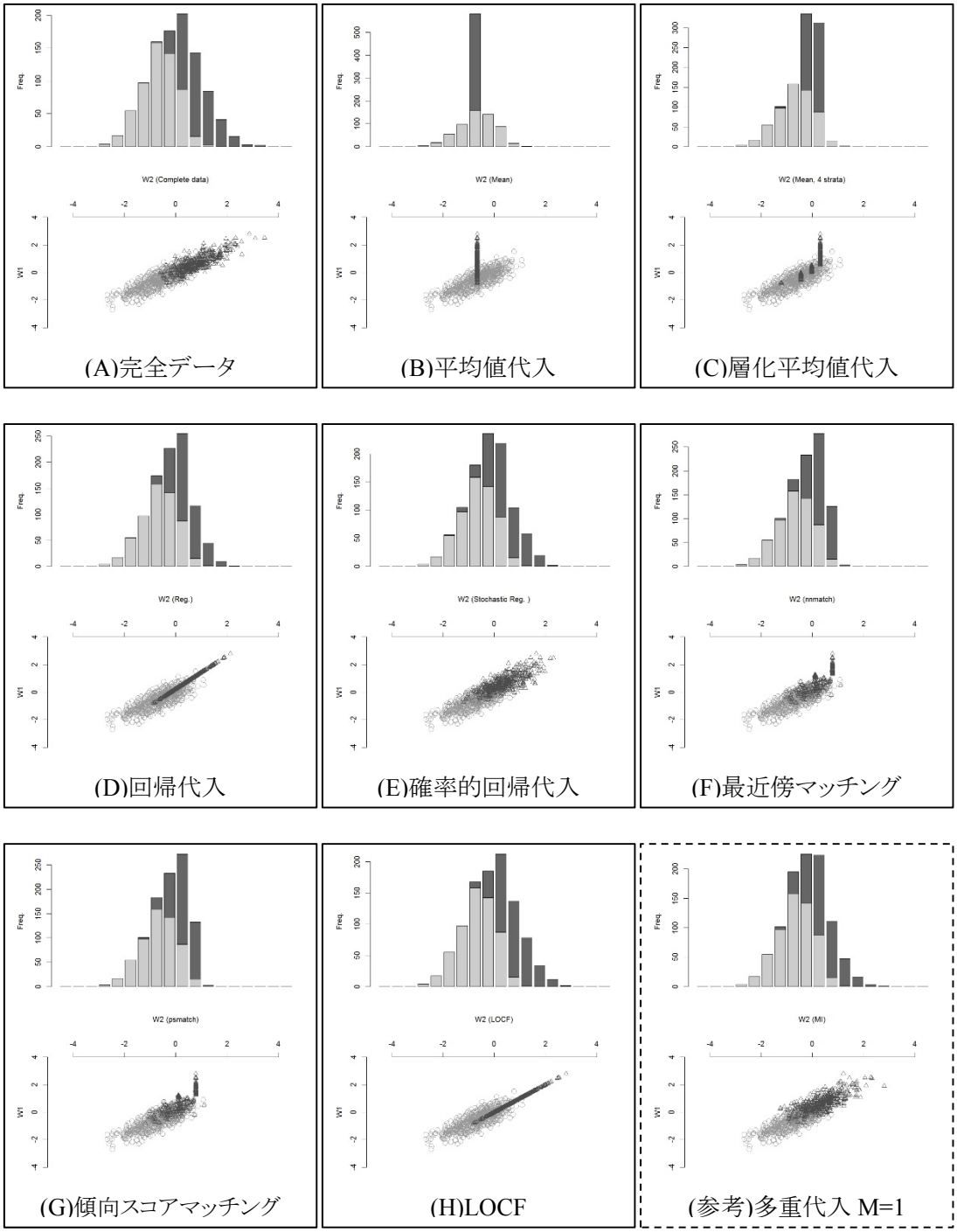
LOCF については、W1 と W2 が負の相関関係にある場合、欠測レコードで代入値と欠測値とでレコード間の大小関係が反転するため、図２－３－２パネル(A)の完全データの分布にみられる欠測率と当該変数の値との正の相関が、同パネル(H)の疑似完全データではみられなくなっている。この点で LOCF は他の手法と異なり、負の系列相関や一時的共通ショックを特徴とする変数に適用すると推定のバイアスをより大きくする。

〇補助変数との相関がない場合

図２－３－３は、第１時点の体重前月差 W1 と第２時点の体重前月差 W2 に相関がない場合について、各単一代入法の性質を示した。この場合、層化平均値代入法（パネル(C)）は平均値代入法（パネル(B)）と似たような結果をもたらす。MNAR のもとで W1 が W2 と無相関であれば（正確には、補助変数が欠測確率に説明力をもたなければ）、当該補助変数 W1 による層化のメリットはない。回帰代入（パネル(D)）もまた平均値代入法（パネル(B)）と似たような結果をもたらす。補助変数に欠測を説明する力がなければ回帰代入にもメリットはない。事実、補助変数と目的となる変数が無相関の場合、層化平均値代入法及び回帰代入法は、平均値代入法と同値である。したがってこの場合の確率的回帰代入の結果は、平均値代入の結果に誤差項のばらつきを与えたに過ぎないものとなる。マッチング代入法（パネル(F)及び(G)）も、代入値にもっともらしさを与えるマッチングではなくランダム・マッチングとなり、結果は確率的回帰代入と同様にほぼ意味のない代入となる。LOCF（パネル(H)）も（LOCF が非常に厳しい制約下での回帰代入法であると考えると）回帰代入法と同様に、真の姿（パネル(A)）を再現できない。

用いる補助変数が欠測確率に説明力をもたなければ、どの単一代入法も欠測バイアスを緩和できない。それればかりではなく、作成される疑似完全データにおいて、変数相互間の関係が代入によってゆがめられるという害をもたらすので、このように不適当な補助変数を用いた単一代入法は避けなければならない。

図2－3－1　単一代入法の数値例(1)補助変数と正の相関



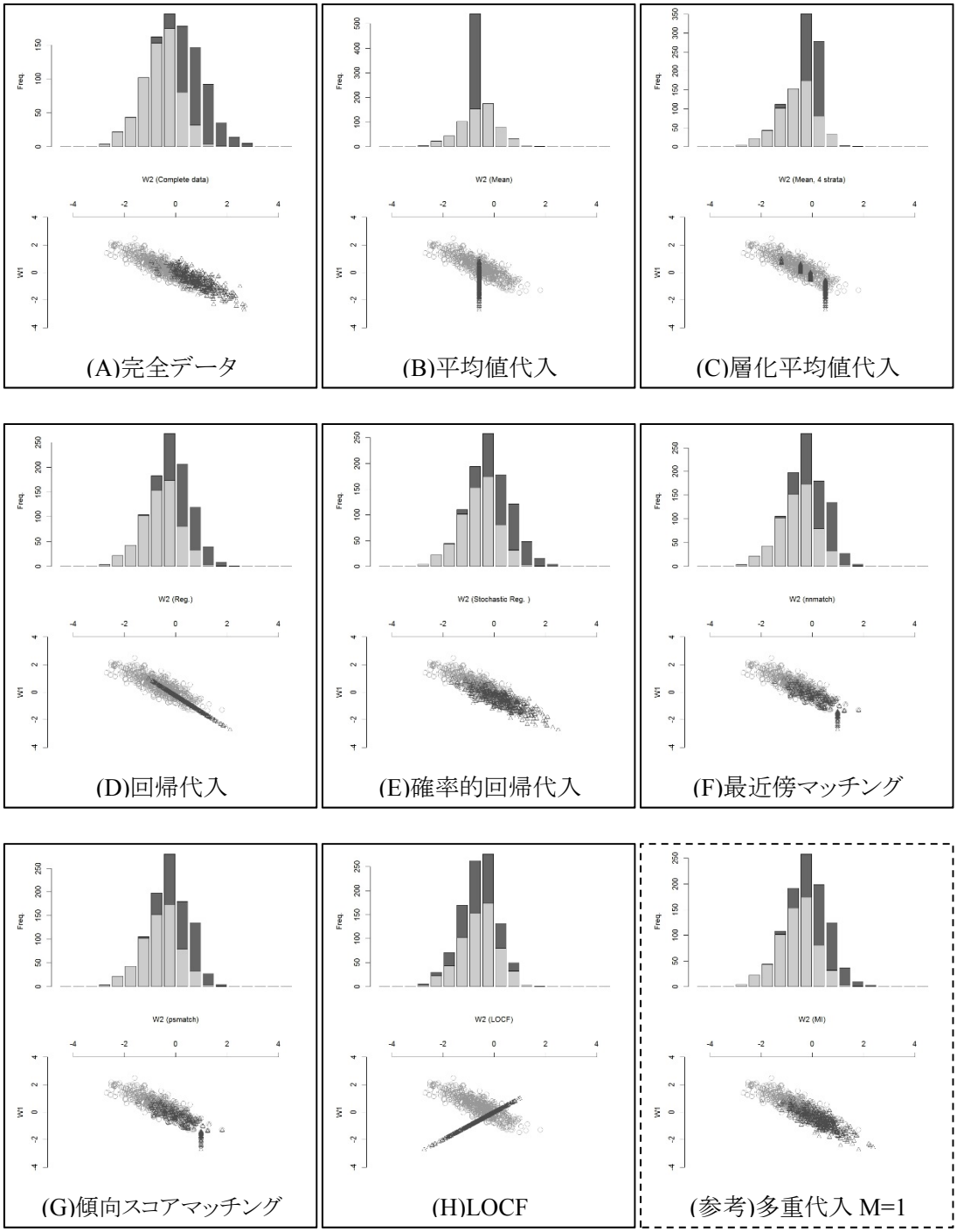
(A)完全データ　(B)平均値代入　(C)層化平均値代入

(D)回帰代入　(E)確率的回帰代入　(F)最近傍マッチング

(G)傾向スコアマッチング　(H)LOCF　(参考)多重代入 M=1

各パネルの上図は W2 のヒストグラム、下図は W2 及び W1 の散布図。ヒストグラムの淡灰色は観測データ、濃灰色は欠測データの欠測値又は代入値をそれぞれ表す。散布図の○は観測データ、△は欠測データの欠測値又は代入値をそれぞれ表す。

図2－3－2　単一代入法の数値例(2)補助変数と負の相関



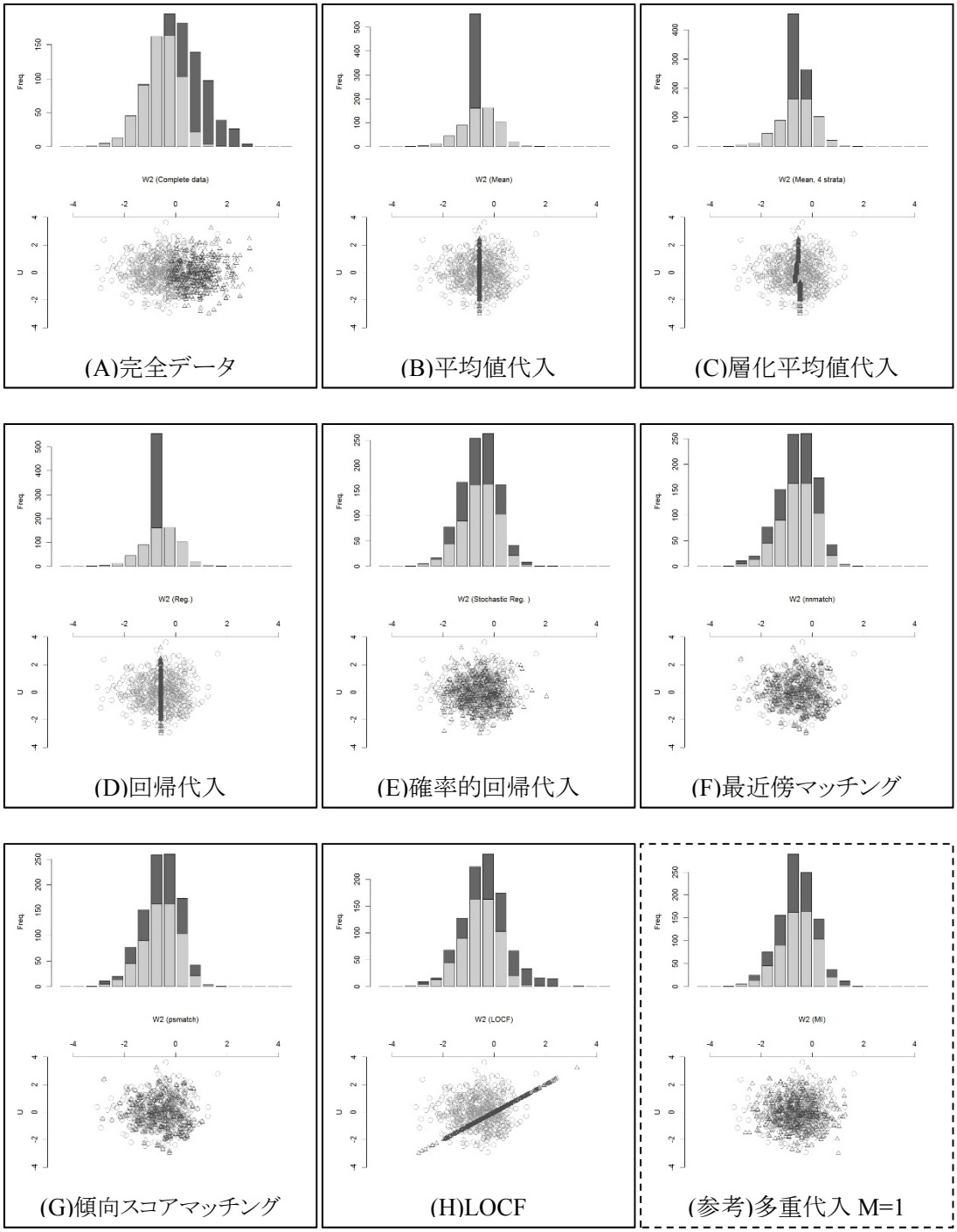
(A)完全データ　(B)平均値代入　(C)層化平均値代入

(D)回帰代入　(E)確率的回帰代入　(F)最近傍マッチング

(G)傾向スコアマッチング　(H)LOCF　(参考)多重代入 M=1

各パネルの上図は W2 のヒストグラム、下図は W2 及び W1 の散布図。ヒストグラムの淡灰色は観測データ、濃灰色は欠測データの欠測値又は代入値をそれぞれ表す。散布図の○は観測データ、△は欠測データの欠測値又は代入値をそれぞれ表す。

図2－3－3　単一代入法の数値例(3)補助変数と無相関



(A)完全データ　(B)平均値代入　(C)層化平均値代入

(D)回帰代入　(E)確率的回帰代入　(F)最近傍マッチング

(G)傾向スコアマッチング　(H)LOCF　(参考)多重代入 M=1

各パネルの上図は W2 のヒストグラム、下図は W2 及び W1 の散布図。ヒストグラムの淡灰色は観測データ、濃灰色は欠測データの欠測値又は代入値をそれぞれ表す。散布図の○は観測データ、△は欠測データの欠測値又は代入値をそれぞれ表す。

◇まとめ

平均値代入法は、MCAR の下での１次モーメントの点推定に限り、欠測バイアスを緩和する。MAR 及び MNAR の下では、欠測バイアスを緩和しない。そして無条件に、次の問題を伴う。第１に、１次よりも大きい母集団モーメントの推定には過小バイアスをもたらす。第２に、推定精度を過大評価する。第３に、変数間の関係性をゆがめる。

層化平均値代入、回帰代入法、確率的回帰代入法及びマッチング代入法は、１次モーメントの点推定であれば、MAR の下でも適切な補助変数により欠測バイアスが緩和される。また、平均値代入法と比べて、１次よりも大きいモーメントの推定におけるバイアス、推定精度の過大評価、及び変数間の関係性のゆがみの程度は抑制される。特に確率的代入法は、１次よりも大きいモーメントの推定におけるバイアス及び推定精度の過大評価をさらに抑制する。

単一代入法における欠測バイアスの緩和は、欠測する変数、欠測確率、及び補助変数相互の関係性の中に含まれる情報を活用することで可能となっている。このため、用いる補助変数の欠測に対する説明力が弱くなるほど、層化平均値代入法等の平均値代入法に対する優位性は小さくなる。

LOCF は、当期の値と直近の観測値とに正の相関がある限りにおいて、欠測バイアスを緩和できる。とりわけ LOCF の適性は、他の単一代入法と異なり、欠測データメカニズムよりも、欠測する変数の系列相関によって決まる。

【数学補論①：推定目標のモーメント次数とバイアス】

これまで各単一代入法の特徴を明らかにしてきたが、そのなかで1次よりも大きいモーメントの推定に各手法がどの程度有効かという点に関連して、推定目標のモーメント次数と推定バイアスの関係を、図2－4により視覚的に説明する。図最上部の「完全データ」は、12 の調査客体の興味の対象となる変数Yの分布を示したものである。変数Yは0以上の値をとる変数であるとする。記号○は値が観測されている調査客体を、記号Δは値が観測されていない調査客体を表している。 9の調査客体については変数Yの値が観測されており、残りの3の調査客体については変数Yの値が観測されていない。完全データとしては、 4の調査客体が$Y_i = y_1$であり、別の4の調査客体が$Y_i = y_2$であり、残りの4の調査客体が$Y_i = y_3$である。

図中上から2番目の「疑似完全データ」は、上に完全データとして示した不完全データに平均値代入法を適用した結果である。ここではy_2がy_1とy_3の平均値に等しいとしているので、3の調査客体の欠測値には値y_2が代入される。図では、代入値を与えられた3の調査客体は、点線の丸記号で表している。

図中下3つのグラフは、上から順に、1次よりも大きいモーメント、1次モーメント、1次よりも小さいモーメントの推定のしくみを示したものである。横軸は変数Yで、グラフは関数Y^hの曲線である。グラフは、 1次よりも大きいモーメント（$h > 1$）の場合は下に凸の曲線、1次モーメント（$h = 1$）の場合は45 度線、1次よりも小さいモーメント（$h < 1$）の場合は下に凹の曲線である。各グラフで、変数Yのとる値に対応する曲線上の点を頂点とする多角形（この場合は三角形）は、モーメントの推定量が取り得る値の範囲を表している。1次モーメントの場合、この領域はつぶれて直線の閉区間となる。一般的に、モーメントの推定値は、この多角形（1次モーメントの場合は直線閉区間）の内分点であり、各頂点の対応する値をとる調査客体単位の頻度がウェイトとして働く。いわば、多角形において各頂点に位置するレコードの重みで綱引きが行われて、推定値が決まる。

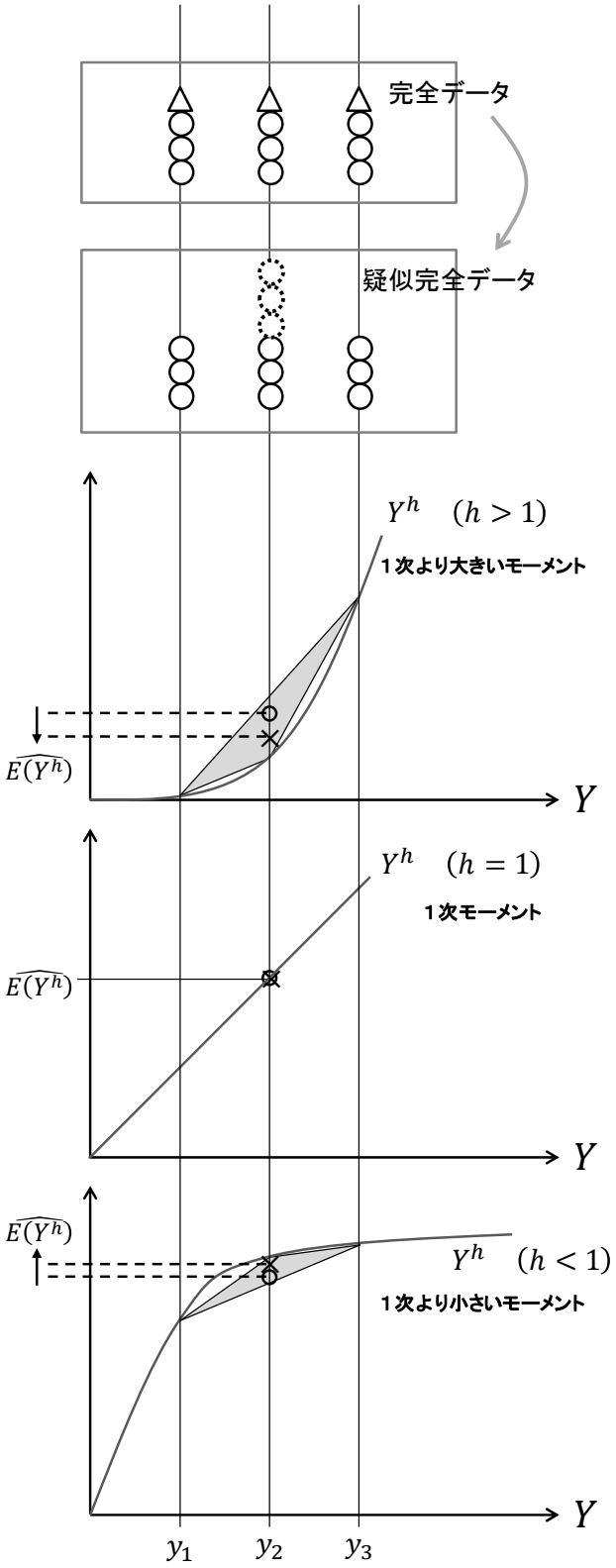
図中下3つのグラフで、記号○は完全データが得られた場合の推定値であり、記号×は疑似完全データに基づく推定値である。疑似完全データでは、3の調査客体が$Y_i = y_1$であり、別の6の調査客体が$Y_i = y_2$であり、残りの3の調査客体が$Y_i = y_3$である。完全データではすべての点に同数の調査客体が存在するので、各点の引力は互いに等しいが、疑似完全データでは$Y_i = y_2$の点の引力が他の点の引力の2倍である。このため、疑似完全データによる推定値を表す記号×は、完全データの推定値を表す記号○と比べて、$Y_i = y_2$の点により近い位置をとる。

これが、1次よりも大きいモーメントの推定の場合は推定値の下方バイアスとなって表れ、1次よりも小さいモーメントの推定の場合は推定値の上方バイアスとなって表れ

る。1 次モーメントの推定の場合は、推定量の取り得る値の範囲が直線閉区間であるため上記の作用は働かず、推定にバイアスはない。

ここでは、図を見やすくするために欠測データメカニズムは MCAR とし、単一代入法としては平均値代入法を用いて、モーメント次数とバイアスの関係を説明したが、同様の効果は他の欠測データメカニズムの下でも他の単一代入法でも働く。重要な点は、モーメント次数の乗数関数が描く曲線が、1 次よりも大きい次数では下に凸で、1 次よりも小さい次数では下に凹、1 次では 45 度線となるということである。疑似完全データの分布が完全データの分布からずれることで、推定値の取り得る値の領域における内分比がゆがみ、推定値をずらしてしまう効果が一般的に作用するのである。1 次モーメントの場合は、推定値の取り得る値の領域が直線閉区間であるため、疑似完全データの分布と完全データの分布が対称性に関して同等であれば、バイアスをもたらさない。

図２－４　推定目標のモーメント次数とバイアス



【数学補論②：単一代入法における推定精度の過大評価】

単一代入法によって作成された疑似完全データから得られる推定量$\hat{\theta}^{SI}$の分散$Var(\hat{\theta}^{SI})$は、一般的に次式で表される。

$$Var(\hat{\theta}^{SI}) = E\left(Var(\hat{\theta}^{SI}|S_R, S_M)\right) + Var\left(E(\hat{\theta}^{SI}|S_R, S_M)\right)$$

(2-2-1)

非確率的代入の場合は、回答標本S_R及び無回答標本S_Mが与えられれば推定量$\hat{\theta}^{SI}$の値は確定するから$Var(\hat{\theta}^{SI}|S_R, S_M) = 0$となり、右辺第1項は0である。

疑似完全データが完全データであると**みなした**場合の推定量$\hat{\theta}^{SI}$の分散を$Var^{\times}(\hat{\theta}^{SI})$とする。単一代入法の疑似完全データが完全データであると錯覚すると、推定量$\hat{\theta}^{SI}$の分散として$Var(\hat{\theta}^{SI})$ではなく$Var^{\times}(\hat{\theta}^{SI})$を推定することになる。$Var^{\times}(\hat{\theta}^{SI})$は、次式で定義される。

$$Var^{\times}(\hat{\theta}^{SI}) \equiv E\left(Var(\hat{\theta}^{SI}|S)\right) + Var\left(E(\hat{\theta}^{SI}|S)\right)$$

(2-2-2)

非確率的代入の場合は、標本Sが与えられれば推定量$\hat{\theta}^{SI}$の値は確定すると**みなされる**から$Var(\hat{\theta}^{SI}|S) = 0$となり、右辺第1項は0である。

さて、疑似完全データの標本$S = S_R \cup S_M$に対しては$E(\hat{\theta}^{SI}|S_R, S_M) = E(\hat{\theta}^{SI}|S)$となる一方で$S$の値の場合の数よりも$(S_R, S_M)$の値の場合の数のほうが多いことからも分かるように、$Var\left(E(\hat{\theta}^{SI}|S_R, S_M)\right) > Var\left(E(\hat{\theta}^{SI}|S)\right)$である。したがって、$Var(\hat{\theta}^{SI}) > Var^{\times}(\hat{\theta}^{SI})$となり、非確率的代入の場合は推定精度が過大評価されることが分かる。

確率的単一代入法の場合は、$Var(\hat{\theta}^{SI})$の(2-2-1)式右辺第1項が0とはならない。しかし、疑似完全データが完全データであると**みなした**場合は誤差項も確率変数とはみなされないので、$Var^{\times}(\hat{\theta}^{SI})$の(2-2-2)式右辺第1項は依然として0となって、推定精度は過大評価されたままである（このようなみなしは確率的代入を無意味にする）。これに対して、確率的代入の意を汲んで特別に当該第1項を正しく評価できれば、代入モデルが正しい限りにおいて推定精度は正しく評価されるが、一般的な推定量$\hat{\theta}^{SI}$について(2-2-1)式右辺を算出することは容易ではない。

単一代入法が推定精度を過大評価するという問題への対応として、(1)ブートストラップやジャックナイフなどのリサンプリングを用いた推定精度評価の手法と(2)多重代入法がある。本報告書では、(1)リサンプリングの手法の説明は割愛し（土屋（2009）などの教科書を参照）、(2)多重代入法を第2.5節で説明する。

2.3 キャリブレーション推定法

キャリブレーション推定法は、補助変数の標本における値、及び母集団における周辺分布の情報に基づいて「ウェイト」の調整を行う一般的な推定方法である。次節のIPW 法も、ウェイトの調整によって欠測バイアスを緩和する手法であり、直感的な説明はキャリブレーション推定法と共通する部分が大きい。両者で異なる点として、第1に、補助変数に関する母集団特性の情報を、キャリブレーション推定法では用いるが、IPW 法では用いない。第2に、欠測バイアス以外のバイアス(後述)も、キャリブレーション推定法では補正しているが、IPW 法は欠測バイアスのみを補正する。本節ではまず、「ウェイト」について説明したうえで、キャリブレーション推定法の概要を示す。また、キャリブレーション推定の特殊形である事後層化推定について、ウェイト調整によって標本の偏りを補正する考え方の直感的な理解を目指す。

一般的に標本調査におけるウェイトとは、標本に含まれた調査客体のそれぞれが、目標母集団の要素何単位分を代表しているか、を表す尺度である。ウェイトは、目標母集団の要素のそれぞれが標本に含まれる確率(「包含確率」と呼ばれる)の逆数に等しい。1%の確率で標本に含まれる調査客体は、目標母集団の要素 100 単位分(すなわち当該調査客体と他の 99 の調査対象候補)を代表し、20%の確率で標本に含まれる調査客体は、目標母集団の要素 5 単位分(すなわち当該調査客体と他の 4 の調査対象候補)を代表している(標本における母集団代表性の尺度としての包含確率の逆数の妥当性は、Horvitz-Thompson 推定量の不偏性と関連している。Horvitz-Thompson 推定量の不偏性については補論参照)。単純無作為抽出では、すべての調査対象候補が等確率で抽出されるので、すべての調査客体でウェイトの値は等しい。特に非復元単純無作為抽出の場合、すべての調査対象候補について包含確率の値はn/N(標本サイズ/母集団サイズ)、従ってウェイトの値はN/n(母集団サイズ/標本サイズ)である。(復元単純無作為抽出であれば、包含確率は$1-((\mathrm{N}-1)/\mathrm{N})^{\mathrm{n}}$、従ってウェイトの値は$\{1-((\mathrm{N}-1)/\mathrm{N})^{\mathrm{n}}\}^{-1}$である。)このように、標本抽出デザインが決まれば包含確率も決まるので、ウェイトも決まる。ウェイトとしての包含確率の逆数を、特に「抽出ウェイト」と呼ぶ。回答率 100%の標本に対しては、抽出ウェイトを用いることで偏りのない推定が可能である(Horvitz-Thompson 推定量の不偏性に関する次節の説明参照)。標本調査で欠測が生じた場合は、いわば母集団から標本へと選出された代表に欠員が生じているので、回答標本に対して抽出ウェイトを用いた推定は、母集団をあまねく反映していないことになる(都議会で都議に欠員が生じると、当該選挙区の意見が都政に反映されなくなるイメージ)。そこで、キャリブレーション推定法及び IPW 法では、回答標本におけるウェイトを調整することで、回答標本の偏りを補正することが意図される。ここで注意すべきは、ウェイト調整による補正では、欠測となった調査客体が代表している母集団の要素に類似した他の要素を代表する調査客体が、回答標

本の中になお残されていなければならないという点である（「サポート問題（support problem）」と呼ばれる）。これがウェイト調整を行う手法の前提となる。

なお、ここでは欠測バイアスを緩和する統計的処理法として、キャリブレーション推定法をとりあげているが、キャリブレーション推定法は、欠測バイアスに限らず、より一般的な標本の偏りを補正する手法である。本書では、「標本の偏り」としては、欠測が生じた場合に分析の対象となる回答標本の偏り、すなわち欠測バイアスのみを考えているが、一般的に、欠測が生じない条件下でも標本の偏りは生じる。それは、「運の悪い標本抽出結果」という形で事後的に生じるものである。単純無作為抽出法は、事前の意味では母集団の縮図となる標本を抽出するが、偏った標本が抽出される確率は0ではない（※確率比例抽出、層化抽出、多段抽出などは、このような悪運を抑制する標本抽出デザインといえる）。キャリブレーション推定法は、欠測バイアスへの対応としても利用できるが、欠測が生じない場合にも利用され、その場合は、事後的に（運悪く）標本が偏ることへの対応となっている。

キャリブレーション推定法は、補助変数について推定値が母集団特性値の真の値に等しくなるようなウェイトを用いた推定法である。ただし、補助変数について推定値が母集団特性値の真の値に等しくなるようなウェイトは一意ではない。キャリブレーション推定法では、通常無数に存在するウェイトの候補のなかで、抽出ウェイトに最も近いものを採用する。補助変数に関して、あるウェイトを用いた推定値がその母集団特性値に等しいことを表す条件式を、当該ウェイトの「キャリブレーション方程式（calibration equation）」と呼び、キャリブレーション方程式を満たしかつ抽出ウェイトからの距離を最小化するウェイトを、「キャリブレーションウェイト（calibration weight）」と呼ぶ。キャリブレーションウェイトを用いた推定が、キャリブレーション推定である（正確には補論参照）。キャリブレーションウェイトを算出するためには、用いる補助変数の母集団特性値が知られていなければならない。

キャリブレーション推定法の要点を理解するために、キャリブレーション推定法の特殊形のひとつである「層サイズによる事後層化推定」の考え方を、図2－5－1に示す。目的となる変数Yに欠測が生じ、補助変数Xの値はすべての調査客体で観測されている。図2－5－1のXY平面上の散布図は、仮に目的となる変数Yの値がすべての調査客体で観測される（すなわち完全データが観測される）場合に得られるものであり、観測データのレコードを記号○、欠測データのレコードを記号△で表す（記号○についてはX座標とY座標が両方とも知られているが、記号△についてはX座標のみが知られている）。図2－5－1では、完全データにおいて目的となる変数Yと補助変数Xの間に高い正の相関がある状況を考える。

図2－5－1（イ）に示す3つのパネルそれぞれの上方側のグラフは、目的となる変数Yについて、真の分布及び欠測によってゆがめられた分布を示したものである。目的となる変数Yの分布で、灰色の領域は、変数Yが観測されないレコードの、変数Yの値ご

との頻度を示す（図1－1～1－3のヒストグラム参照）。このグラフによると、欠測の起こりやすさが欠測する変数Yの値に依存しているので MNAR である（変数Yとの相関が高い補助変数Xが利用可能でなければ）。特に、変数Yの値が大きいほど欠測が起こりやすくなっている。このことは、散布図からも確認できる。すなわち、右側の領域ほど記号△で表される欠測レコードの割合が高くなっている。

図2－5－1（イ）に示す3つのパネルそれぞれの左側方のグラフは、補助変数Xの真の分布、及び目的となる変数Yが観測されているという条件による補助変数Xの条件付分布、を示したものである。補助変数X自体は欠測が生じない変数なので、上側方の目的となる変数Yに関する分布のグラフとの相違点に注意を要する。左側方の補助変数Xに関するグラフでは、灰色の領域は、（補助変数Xではなく）目的となる変数Yが観測されないレコードの、補助変数Xの値ごとの相対的頻度を示す。

ここで、補助変数Xの値に基づいて不完全データの標本を層化することができる。層の数をKとする。図2－5－1（イ）及び（ロ）のそれぞれに示す3つのパネルは左から順に第 1 層、第k層$(k = 2, 3, \cdots, K-1)$及び最後の第K層に注目した場合を示したものである。補助変数Xはすべて観測されているので、任意の第k層において（つまりどの層においても）、観測レコードと欠測レコードの構成比を知ることができる。たとえば、図の第k層では、観測レコード7件（7個の記号○）と欠測レコード6件（6個の記号△）から成っている（第k層の回答者数$n_k^R = 7$及び無回答者数$n_k^M = 6$）。この層では、観測レコード 1 件を(7+6)/7 倍に膨らませることで、（観測レコードのみを用いて）補助変数Xの分布を完全データのものに一致させることができる。さらに、母集団について任意の第k層のサイズN_kが知られていれば、第k層の観測レコード 1 件を（母集団第k層のサイズ/母集団サイズ）/（回答標本第k層のサイズ/標本サイズ）の倍率で膨らませることで、（観測レコードのみを用いて）補助変数Xの分布を母集団のものに一致させることができる。この場合のキャリブレーション方程式は、$\sum_{i \in S_k^R} w_i^C = N_k$である（ただし$S_k^R$は回答標本の第$k$層である）。

もっとも補助変数Xはすべて観測されているので、補助変数Xに関する推定が目的であればことさらに上記の処理をする必要はない。上記の処理を目的となる変数Yについて実行できればよいが、それは不可能である。そこで、目的となる変数Yのかわりに補助変数Xについて実行するのである。欠測は、目的となる変数Yの回答標本における分布をゆがませるが、それと連動して補助変数Xの回答標本における分布もゆがませる。この連動性に着目すると、補助変数Xの回答標本における分布のゆがみを補正すれば、それに連動して変数Yの回答標本における分布のゆがみも補正されていることが期待される。これが、キャリブレーション方程式を制約条件とすることの動機となっている。

補助変数Xの次元で行われる補正が、目的となる変数Yの次元でどのような効果を

もつかを示したのが、図2－5－1(ロ)である。ただし上述の通り、キャリブレーション推定法には欠測バイアス以外のバイアス(ここでは「運の悪い標本抽出」によるバイアスのみ)も同時に補正する機能があり、ここでは欠測バイアスを補正する機能のみを示したいので、ウェイト補正を欠測バイアスに対応する部分とそれ以外のバイアスに対応する部分の2つに分解し、前者の効果のみを図示する。単純無作為抽出の抽出ウェイトであれば、母集団について任意の第k層のサイズN_kの値を用いて、第k層の観測レコード 1 件を(母集団第k層のサイズ/母集団サイズ)/(回答標本第k層のサイズ/標本サイズ)の倍率で膨らませるが、この倍率を項(標本第k層のサイズ)/(回答標本第k層のサイズ)と項(標本サイズ/標本第k層のサイズ)×(母集団第k層のサイズ/母集団サイズ)の積としてみると、前者は欠測バイアスの補正、後者は「運の悪い標本抽出」によるバイアスの補正となっている。第k層に属する記号○(すなわち観測レコード)を(7+6)/7 倍に膨らませるのであるから、(ロ)上側方の補正前分布では、矢印で示したとおりの垂直方向の拡張が第k層に属する観測レコードの各点で起こる。これが、事後層化第k層における補正の効果である。同様の補正が他のすべての層においても行われ、目的となる変数Yの分布のゆがみが補正される。第 1 層では、3 件の観測値と 0 件の欠測値があるから、観測値のウェイトは(3+0)/3 倍に調整される。第K層では、1 件の観測値と 3 件の欠測値があるから、観測値のウェイトは(1+3)/1 倍に調整される。

図2－5－1に示した層サイズによる事後層化推定法の例では、補助変数Xの値に基づく任意の第k層において(つまりどの層においても)、観測レコード数(記号○の数)と欠測レコード数(記号△の数)の合計に占める観測レコード数(記号○の数)の割合の逆数を抽出ウェイトに乗じたものをウェイトとして、目的となる変数Yの分布を推定することで、欠測バイアスが除かれる。事後層化推定法は、目的となる変数Yと補助変数Xとの相関が高いほど、目的となる変数Yの分布の推定における欠測バイアスをより多く取り除くことができる。図2－5－1に示した原理は、ウェイト調整に関する一般的な原理であり、IPW 法でも同様にはたらいている。

次に、目的となる変数Yと補助変数Xとの相関が低いときには欠測バイアスを緩和する効果が小さくなることを、図2－5－2により示す。図2－5－2は、目的となる変数Yと補助変数Xとに相関がないときに事後層化推定法を適用した場合である。記号や配色の意味は、図2－5－1と同様である。図2－5－1との違いは、目的となる変数Yと補助変数Xとの相関がないという点だけである。図2－5－1及び図2－5－2に示された点線の楕円形は、変数Yと変数X間の全データの散布図における散らばり方(正確には同時分布の等量線)を表している。欠測レコード(記号○)と観測レコード(記号△)の散布図上の位置は、この点線の楕円形によって確率的に制約される。図2－5－1では楕円形が細長いので、レコードのY軸座標がひとたび決まれば、とり得るX軸座標の範囲はかなり狭まる(逆も然りである)が、図2－5－2では楕円形が幅広いので、レコードのY軸座標が決まっても、とり得るX軸座標の範囲は依然として広い(逆も然りであ

る)。図2－5－1及び図2－5－2のいずれにおいても、等しく目的となる変数Yの値が大きいほど欠測が起こりやすくなっているので、記号△で表されるレコードは、Y軸の右方により多く集中する。その結果、楕円形が細長い図2－5－1では、必然的にX軸方向の下方により多く集中することになるが、楕円形が幅広い図2－5－2では、X軸方向に関しては均一に分布してしまう。このため図2－5－2では、すべての層において欠測レコードの割合が等しくなっている。図2－5－2では、すべての層で観測レコードのウェイトに等しい値を掛けるので、実質的にウェイトの調整がなされないことになる。

補助変数の目的となる変数に対する相関の有無によるウェイト調整結果の違いは、図2－5－1(ロ)と図2－5－2(ロ)を比較することでも分かる。相関のある図2－5－1(ロ)で分布の左側が拡張されているのと比べて、相関のない図2－5－2(ロ)では分布のやや右寄りが拡張されており、ゆがみが正しく補正されていない。この図ではひとつの層(第 k 層)におけるウェイト調整の効果を示しているが、他の層についても同様の効果がみられる。

ここで注目すべき点は、図2－5－2(イ)及び(ロ)左側方のグラフである。目的となる変数Yと補助変数Xとに高い相関がある図2－5－1の場合は、欠測の起こりやすさと補助変数Xとの間に相関があったものの、目的となる変数Yと補助変数Xとに相関がない図2－5－2の場合は、欠測の起こりやすさと補助変数Xとの間に相関がない。これは、もともと欠測の起こりやすさが目的となる変数Yに依存しているので、目的となる変数Yとの相関が高い補助変数は、欠測の起こりやすさとも相関が高く、目的となる変数Yとの相関がない補助変数は、欠測の起こりやすさとも相関がないということの表れである。図2－5－1では変数Yの欠測・観測別分布と変数Xの欠測・観測別分布に連動性があったが、図2－5－2では解消している。回答者に関する変数Xの分布を真の分布へ向けて補正することで、間接的に変数Yの分布を補正するキャリブレーション推定法では、目的となる変数Yと補助変数Xの連動性(相関)が重要である。

このように、補助変数が目的となる変数と相関をもたなければ(より正確には、補助変数が欠測確率に対して説明力をもたなければ)、事後層化推定法は欠測バイアスを緩和できない。この点は、単一代入法、多重代入法、IPW 法などの他の欠測データ処理法についても当てはまる。

キャリブレーション推定法を実行するためには、補助変数の(層別)母集団総計を知っている必要がある。この点が、他の手法と比べてキャリブレーション推定法を実行する上での大きな制約となる。政府統計においては、母集団データベースの活用が期待される所以である。

図2－5－1　層サイズによる事後層化推定の考え方

Y：目的となる変数、X：補助変数、○：観測レコード、△：欠測レコード

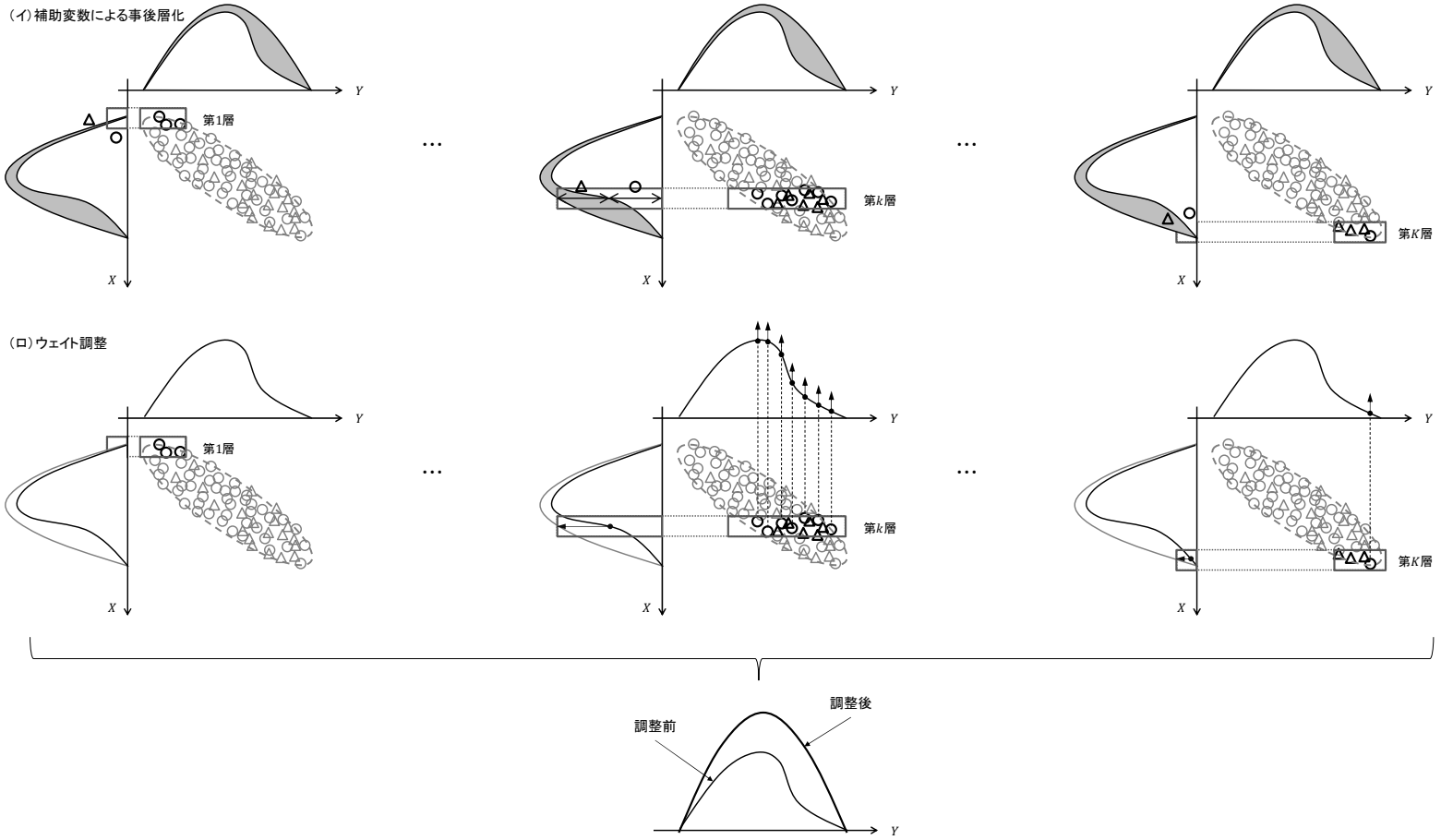
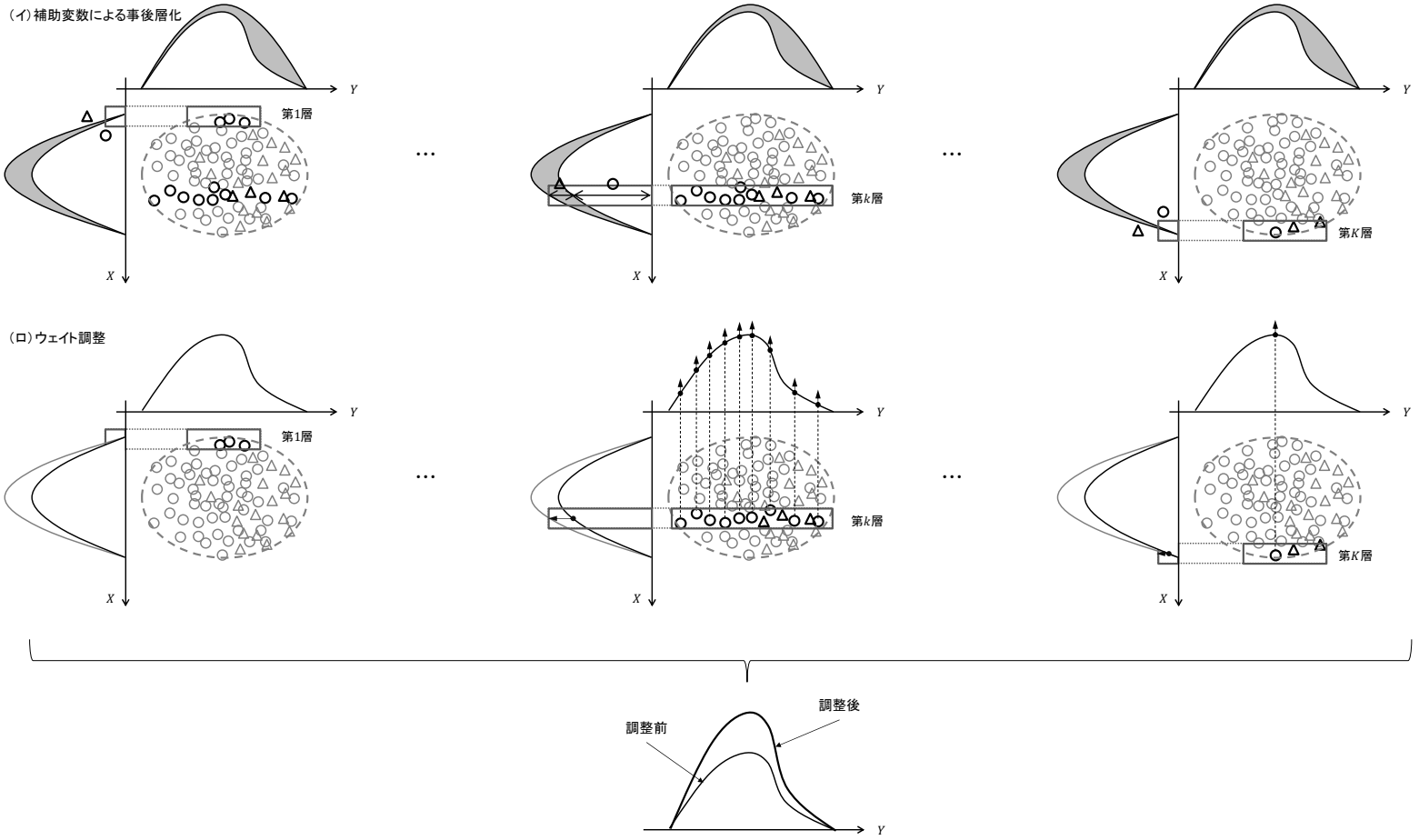


図2－5－2　目的となる変数Yと補助変数Xに相関がない場合の事後層化推定

Y:目的となる変数、X:補助変数、○:観測レコード、△:欠測レコード



【補論①：数式を用いたキャリブレーション推定法の説明】

不完全データの標本S、補助変数x、の母集団総計τ_xに対して、調整後ウェイト$w_i^C\ (i \in S)$に関する下の方程式を補助変数xの「キャリブレーション方程式」と呼ぶ。

$$\sum_{i \in S} w_i^C x_i = \tau_x$$

(2-3-1)

キャリブレーション方程式は、調整後ウェイト$w_i^C\ (i \in S)$が満たすべき条件であり、補助変数の数だけある。つまり調整後ウェイト$w_i^C\ (i \in S)$による補助変数の母集団総計の線形推定量は、真の値に一致する。

キャリブレーション推定法は、キャリブレーション方程式を制約条件として、標本Sの抽出デザインによる抽出ウェイト$w_i\ (i \in S)$と調整後ウェイト$w_i^C\ (i \in S)$の間の距離を最小化する。具体的には、次式の最小化問題を解く。

$$\min_{\{w_i^C\}_{i \in S}} \sum_{i \in S} w_i G\left(w_i^C, w_i\right)$$
$$s.t.\ \ \forall x, \sum_{i \in S} w_i^C x_i = \tau_x$$

(2-3-2)

距離関数$G\left(w_i^C, w_i\right)$の特定化に応じて種類がある。代表的なものとして、次に示す線形関数、乗法関数、ロジット関数がある。

◇線形関数

$$G\left(w_i^C, w_i\right) = \frac{1}{2}\left(\frac{w_i^C}{w_i} - 1\right)^2$$

◇乗法関数

$$G\left(w_i^C, w_i\right) = \frac{w_i^C}{w_i} \ln \frac{w_i^C}{w_i} - \frac{w_i^C}{w_i} + 1$$

◇ロジット関数（より一般的な関数は土屋（2009）参照）

$$G\left(w_i^C, w_i\right) = \left(\frac{w_i^C}{w_i} - 1\right) \ln\left(\frac{w_i^C}{w_i} - 1\right) - \frac{w_i^C}{w_i}$$

図２－５－１で説明した「層サイズによる事後層化推定法」は、(2-3-2)式において、X_iの値が第k層に属することを表す指標$D_i^k \equiv 1\left[X_i \in \text{第 k 層}\right]$を補助変数とし、距離関数に乗法関数を用いたキャリブレーション推定法である。つまり、層サイズによる事後層化推定法のキャリブレーションウェイトは次式の最適化問題の解である。

$$\min_{\{w_i^C\}_{i\in S}} \sum_{i\in S} w_i^C \ln\frac{w_i^C}{w_i} - w_i^C + w_i$$
$$s.t.\ \ \forall k, \sum_{i\in S_k} w_i^C = N_k$$

第kキャリブレーション方程式のラグランジェ乗数λ_k（$k = 1, \cdots, K$）に対して、最適解の必要条件は次式である。

$$\ln\frac{w_i^C}{w_i} + 1 - 1 + \lambda_{k(i)} = 0$$

ただし、$k(i)$は調査客体iが所属する事後層の番号である。同値条件$w_i^C = w_i e^{-\lambda_{k(i)}}$をキャリブレーション方程式に代入すると$\sum_{i\in S_k} w_i e^{-\lambda_{k(i)}} = e^{-\lambda_k}\sum_{i\in S_k} w_i = N_k$であるから、次式を得る。

$$w_i^C = \frac{w_i}{\sum_{i\in S_k} w_i} N_k$$

非復元無作為抽出であれば、$w_i = n/N$であるから、キャリブレーションウェイトは次式となる。

$$w_i^C = \frac{N_k}{n_k}$$

【補論②：Horvitz-Thompson 推定量】

標本S、変数Y_iの値y_i、包含確率π_iに対して、変数Y_iの母集団総計τ_yの Horvitz-Thompson 推定量（以下 HT 推定量）$\hat{\tau}_y^{HT}$は次式で定義される。

$$\hat{\tau}_y^{HT} \equiv \sum_{i\in S} \frac{y_i}{\pi_i}$$

HT 推定量は、包含確率の逆数をウェイトとした標本総計である。HT 推定量の不偏性は次式のとおりに示される。

$$E\left(\hat{\tau}_y^{HT}\right) = E\left(\sum_{i\in S}\frac{y_i}{\pi_i}\right) = E\left(\sum_{i\in U}\frac{y_i}{\pi_i}1[i\in S]\right) = \sum_{i\in U}\frac{y_i}{\pi_i}E(1[i\in S])$$
$$= \sum_{i\in U}\frac{y_i}{\pi_i}\pi_i = \tau_y$$

HT 推定量の分散は次式のとおりである。

$$Var\left(\hat{\tau}_y^{HT}\right) = Var\left(\sum_{i\in U}\frac{y_i}{\pi_i}1[i\in S]\right) = \sum_{i\in U}\sum_{j\in U} Cov(1[i\in S], 1[j\in S])\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}$$
$$= \sum_{i\in U}\sum_{j\in U}\left(\pi_i\pi_j - \pi_{ij}\right)\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}$$

2.4 IPW 法

標本調査において欠測が生じない場合は、包含確率（母集団の各要素が標本に含まれる確率）の逆数を調査客体ごとのウェイトとすることで、偏りのない推定ができる（Horvitz-Thompson 推定量の不偏性）。この原理を欠測が生じる場合へ拡張した推定手法が、IPW（inverse probability weighting）法である。

拡張は次のように考えることで容易となる。欠測が生じない場合は標本抽出という試行によって分析対象となる標本が得られ、欠測が生じる場合は標本抽出並びに「回答の成否」という2つの試行が合成された試行によって分析対象となる回答標本が得られる。つまり、欠測が生じない場合における試行としての標本抽出を、上述の合成試行に置換えて考えればよい。このことから、欠測が生じる場合は、包含確率の代わりに「母集団の各要素が標本に含まれかつ回答する確率」の逆数を調査客体ごとのウェイトとすることで偏りのない推定となる。

ここで、包含確率の値は標本抽出デザインに応じて決まるためその真の値が知られているのに対して、一般的に母集団の各要素について「標本に含まれかつ回答する確率」の真の値を知ることはできずデータから推定しなければならないということが問題として現れる。IPW 法の適性は、調査客体ごとの「標本に含まれかつ回答する確率」の真の値を正しく推定できるか否かにかかっている。「標本に含まれかつ回答する確率」は非常に緩やかな条件の下で包含確率と「回答する確率」に分解でき、「回答する確率」は別の2つの条件の下で不完全データから正しく推定できる。IPW 法でウェイトに用いる確率の分解及び回答確率の推定における仮定を順にみていく。

まず、一般的に、任意の調査客体の「標本に含まれかつ回答する確率」は当該調査客体の「標本に含まれた場合に回答する確率」と当該調査客体の包含確率との積に等しい。また、標本抽出と回答成否の事象が互いに母集団の要素の属性による条件付独立でありかつ母集団の要素ごとの属性が固定されている場合、任意の調査客体の「標本に含まれた場合に回答する確率」は、単に当該調査客体の「回答する確率」に等しい。まとめると、標本抽出と回答成否の事象が互いに母集団の要素の属性による条件付独立でありかつ母集団の要素ごとの属性が固定されているという条件の下では、任意の調査客体の「標本に含まれかつ回答する確率」は、当該調査客体の「回答する確率」と当該調査客体の包含確率との積に等しい（「数式を使った説明」参照）。

2つの条件（1）標本抽出と回答の有無が互いに条件付独立であること及び（2）母集団の要素ごとの属性が固定されていることは、厳しいものではない。特に、無作為抽出による測定誤差のない標本調査の場合はこれらの条件が成立している。IPW 法ではこれら2つの緩い条件を前提として任意の調査客体の「回答する確率」の値を推定し、それに標本抽出デザインによって決まる包含確率の値を乗じることで、調査客体ごとの「標本に含まれかつ回答する確率」を求める。

次に、不完全データから調査客体ごとの「回答する確率」の値を正しく推定できるためには、さらに別の条件が成立していなければならない。第1に、IPW 法においては、任意の調査客体の「回答する確率」を当該調査客体の属性の関数としてモデル化し、そのモデルのパラメータを推定することで、任意の調査客体の「回答する確率」を推定する。つまり、IPW 法では、回答確率（観測確率）のモデルが正しく特定化されていなければならない。モデルの特定化に誤りがあると、調査客体ごとの「回答する確率」の推定値に誤設定バイアスが伴うからである。第2に、回答確率（観測確率）のモデルは不完全データから推定できなければならない。すなわち、IPW 法では欠測データメカニズムは MAR でなければならない（欠測データメカニズムが MNAR だと、回答確率が観測されない値に依存するため、回答確率モデルは推定できない）。これらの2つの条件（1）観測確率モデルが正しく特定化されていること、及び（2）欠測データメカニズムが MAR であることは、IPW 法における重要な仮定である。

MAR の下では、IPW 法の回答確率モデルにおいて、任意の調査客体の「回答する確率」は、適当な補助変数の値で条件付けた回答確率、すなわち回答の傾向スコアに他ならない。つまり、IPW 法は、MAR の仮定の下で、回答の傾向スコアを推定し、その推定値と包含確率の積の逆数をウェイトとして推定を行う。MAR の下で、傾向スコアの確率モデルが正しければ、IPW 法は欠測バイアスを緩和することができる。

図2－6は、IPW 法によって欠測バイアスが緩和される原理を示したものである。与えられた不完全データに対応する完全データについて、目的となる変数Yと補助変数Xの散布図を最下部に示す。目的となる変数Yの値が観測されている調査客体を記号○、観測されていない調査客体を記号△で表している。つまり、記号○はX座標とY座標の両方が知られているが、記号△はX座標しか知られていない。散布図の上には、完全データにおける変数Yのヒストグラムを示す。灰色部分は欠測値、白色部分は観測値に対応する。ヒストグラムの上には、データから推定される変数Yの分布の形状を示す。2つの曲線のうち、上は仮に完全データが観測された場合の推定分布であり、下は回答標本から推定される分布である。ヒストグラムとの対応を分かりやすくするために、2つの分布の縮尺はそろえていない。2つの曲線に挟まれる領域の垂直距離によって、変数Yの値ごとの欠測率が表される。図2－6のデータ例では、変数Yの値が大きいほど欠測率が高くなる。これだけだと、欠測確率が欠測する変数の値に依存する、すなわち MNAR となるが、同時に図2－6のデータ例では、目的となる変数Yと補助変数Xの相関が大きい。そこで、欠測率は補助変数にも依存し、特に、補助変数だけで説明できれば MAR となる。ここでは、MAR であるとする。

図2－6左下のグラフは、散布図に示されるデータに基づいて傾向スコアを推定したときの結果を表したものである。通常傾向スコアは、観測指標Rの補助変数Xによる2項回帰モデル（ロジットモデルやプロビットモデル）によって推定される。点線が推定さ

れた傾向スコアを表す。この例では、補助変数Xの値が大きいほど傾向スコアは小さくなる。

既に述べたとおり IPW 法は、MAR の仮定の下で推定された傾向スコアの逆数を、抽出ウェイト（包含確率の逆数）に乗じることでウェイトの調整を行う。欠測によって変数Yの分布に生じたゆがみが、傾向スコアによるウェイト調整で補正される効果をみるために、補助変数Xの値ごとのウェイト調整を分けて示す。図では、補助変数Xの値が標本における最も小さい値である場合（$X = \underline{x}$）、最も大きい値の場合（$X = \overline{x}$）及び中間値の場合（$X = x$）、の3通りについて示している。このため、同じ散布図、ヒストグラム及び分布が3つ並んでいる。左から順に、最小値、中間値、最大値の場合である。

左端の散布図において強調表示された記号〇で表された調査客体は、補助変数Xの値が最も小さい。当該調査客体の傾向スコアの推定値は、左端のグラフによると、9/10 である。当該調査客体は、事前には 90%の確率で変数Yの値が観測されるという性質をもっている。変数Yに関する IPW 法の推定において、当該調査客体のウェイトは、推定された傾向スコアの逆数である 10/9 倍に調整される。このように調整されたウェイトによると、当該調査客体は、回答標本において、自身と回答標本に含まれなかった他の要素 1/9 単位分を代表するものとして扱われる。散布図の上のヒストグラムにある上向きの矢印は、この調整によって当該調査客体のウェイトが 10/9 倍に増加することを表している。

中央の散布図において強調表示された記号○で表された調査客体は、補助変数Xが値xをとる。これら調査客体の傾向スコアの推定値は、左端のグラフによると、1/2 である。当該調査客体は、事前には 50%の確率で変数Yの値が観測されるという性質をもっており、実際に当該調査客体の変数の値Yは観測された。変数Yに関する IPW 法の推定において、当該調査客体のウェイトは、推定された傾向スコアの逆数である 2 倍に調整される。このように調整されたウェイトによると、当該調査客体は、回答標本において、自身と回答標本に含まれなかった他の要素 1 単位分を代表するものとして扱われる。中央のヒストグラムにある上向きの矢印は、この調整によって、当該調査客体のウェイトが 2 倍に増加することを表している。

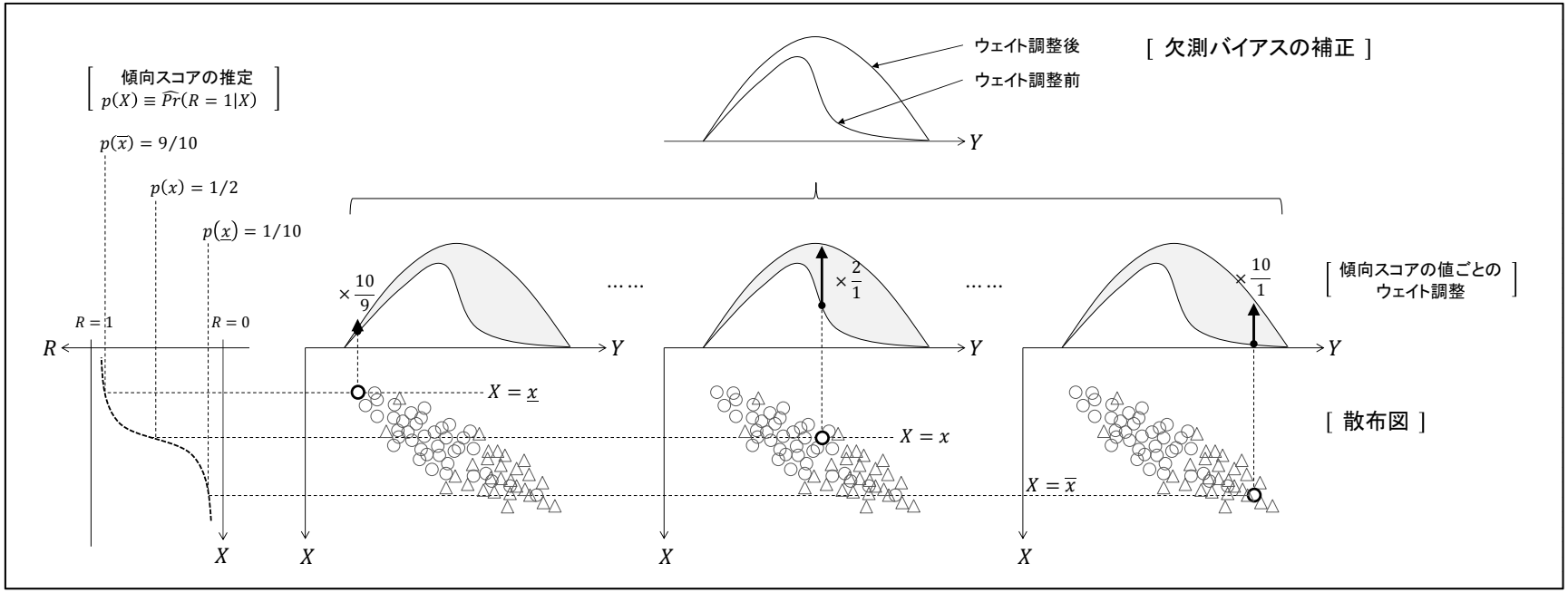
右端の散布図において、強調表示された記号○で表された調査客体は、補助変数Xの値が最も大きい。当該調査客体の傾向スコアの推定値は、左端のグラフによると、1/10 である。当該調査客体は、事前には 10%の確率で変数Yの値が観測されるという性質をもっている。変数Yに関する IPW 法の推定において、当該調査客体のウェイトは、推定された傾向スコアの逆数である 10 倍に調整される。このように調整されたウェイトによると、当該調査客体は、回答標本において、自身と回答標本に含まれなかった他の要素 9 単位分を代表するものとして扱われる。右端のヒストグラムにある上向きの矢印は、この調整によって当該調査客体のウェイトが 10 倍に増加することを表している。

回答標本のすべての調査客体について推定された傾向スコアによりウェイトを調整した結果をまとめると、図2－6最上部に示す分布のように、推定の欠測バイアスが緩和される。緩和の程度は、目的となる変数Yと補助変数Xとの（正負を問わない）相関の強さに依存する。相関係数の絶対値が1であれば、欠測バイアスは完全に除去される。相関係数の値が0であれば欠測バイアスは全く緩和されない。このことは、図2－6中央の散布図及びヒストグラムから理解することができる。相関が強ければ、散布図で水平に並んだ記号○及び△の相互の距離が小さくなる。記号△と当該記号△を代理する記号○との水平距離が小さいと、変数Yに関する代理の妥当性は強くなる。逆に相関が弱ければ、記号△と当該記号△を代理する記号○との水平距離が大きくかい離し、変数Yに関して互いに大きく異なる記号△を記号○が代理することになる。

IPW 法は、MAR の仮定の下で、傾向スコアによりウェイトを調整することで、欠測バイアスを緩和する。MAR の仮定は、回答の傾向スコアを不完全データから正しく推定するための条件である。このほか IPW 法では、傾向スコアを推定するための回答確率モデルの特定化が正しくなければならない。これら2つの条件（1）MAR 及び（2）正しい回答確率モデルは、実務においては、「欠測確率を十分に説明できる補助変数が利用可能である」という条件として考えることができる。

IPW 法に伴う実際問題として、補助変数に関して十分に広範な属性の調査客体で回答が成立していなければ、回答標本のなかに、推定される傾向スコアの値が非常に小さな調査客体が生じてしまい、最終的な推定結果が極端な値となってしまうことがある。これは、ロジットモデルやプロビットモデルのようなパラメトリックなモデル化の限界とみることもでき、また、偶々IPW 法に不向きな不完全データが得られたことの結果とみることもできる。IPW 法を実施する際は、調整後のウェイトが極端な値となっていないかを確認する必要がある。

図2－6　IPW 法の要点


傾向スコアの推定
$p(X) \equiv \widehat{Pr}(R = 1|X)$
$p(\overline{x}) = 9/10$
$p(x) = 1/2$
$p(\underline{x}) = 1/10$
$R = 1$
$R = 0$
R
X
ウェイト調整後
ウェイト調整前
[欠測バイアスの補正]
Y
$\times \frac{10}{9}$
$\times \frac{2}{1}$
$\times \frac{10}{1}$
……
$X = \underline{x}$
$X = x$
$X = \overline{x}$
[傾向スコアの値ごとの
ウェイト調整]
[散布図]

【補論：数式を用いた説明】

○母集団の要素が標本に含まれかつ回答する確率について

一般的に、変数Y_iの観測指標を2値変数R_iで表したとき、母集団の要素iが標本Sに含まれかつ回答する確率$Pr(i \in S, R_i = 1)$は次式のとおりに分解できる。

$$Pr(i \in S, R_i = 1) = \int \cdots \int Pr\left(i \in S, R_i = 1 \middle| \begin{pmatrix} Y_i \\ X_i \end{pmatrix} = \begin{pmatrix} y \\ x \end{pmatrix}\right) f_i \begin{pmatrix} y \\ x \end{pmatrix} dx\, dy$$

母集団の要素iの属性(Y_i, X_i)が固定されている場合は、次式が成り立つ。

$$Pr(i \in S, R_i = 1) = Pr\left(i \in S, R_i = 1 \middle| \begin{pmatrix} Y_i \\ X_i \end{pmatrix} = \begin{pmatrix} y_i \\ x_i \end{pmatrix}\right)$$

一般的に、上式右辺については次式が成り立つ。

$$Pr\left(i \in S, R_i = 1 \middle| \begin{pmatrix} Y_i \\ X_i \end{pmatrix} = \begin{pmatrix} y_i \\ x_i \end{pmatrix}\right)$$
$$= Pr\left(R_i = 1 \middle| i \in S, \begin{pmatrix} Y_i \\ X_i \end{pmatrix} = \begin{pmatrix} y_i \\ x_i \end{pmatrix}\right) Pr\left(i \in S \middle| \begin{pmatrix} Y_i \\ X_i \end{pmatrix} = \begin{pmatrix} y_i \\ x_i \end{pmatrix}\right)$$

母集団の要素iの属性(Y_i, X_i)が固定されている場合は、上式右辺第2項は包含確率π_iに等しい。

$$Pr\left(i \in S \middle| \begin{pmatrix} Y_i \\ X_i \end{pmatrix} = \begin{pmatrix} y_i \\ x_i \end{pmatrix}\right) = \pi_i$$

標本抽出と回答の有無は独立であるとすると、次式が成り立つ。

$$Pr\left(R_i = 1 \middle| i \in S, \begin{pmatrix} Y_i \\ X_i \end{pmatrix} = \begin{pmatrix} y_i \\ x_i \end{pmatrix}\right) = Pr\left(R_i = 1 \middle| \begin{pmatrix} Y_i \\ X_i \end{pmatrix} = \begin{pmatrix} y_i \\ x_i \end{pmatrix}\right)$$

ランダムな欠測（MAR）の下では、上式右辺については次式が成り立つ。

$$Pr\left(R_i = 1 \middle| \begin{pmatrix} Y_i \\ X_i \end{pmatrix} = \begin{pmatrix} y_i \\ x_i \end{pmatrix}\right) = Pr(R_i = 1 | X_i = x_i)$$

まとめると、母集団の要素iの属性(Y_i, X_i)が固定されており、また、標本抽出と回答の有無は独立であるとすると（この条件は厳しいものではない）、MAR の下で次式が成り立つ。

$$Pr(i \in S, R_i = 1) = Pr(R_i = 1 | X_i = x_i) \times \pi_i$$

上式右辺第1項は、回答の傾向スコアである。つまり、MAR の下では、適当な補助変数による回答の傾向スコアと包含確率の積の逆数をウェイトとすれば偏りのない推定ができる。

○Horvitz-Thompson 推定量から IPW 法への拡張

標本Sにおいて事後に完全ケースとなったレコードiが事前に完全ケース分析の対

象となる部分標本に含まれる確率すなわちレコードiが観測される確率の逆数を$q_i \equiv P(R_i = 1|i \in S)^{-1}$とする。目標母集団の要素$i$が標本$S$に含まれる確率の逆数を乗じたものを$w_i \equiv \{P(R_i = 1|i \in S) \cdot P(i \in S)\}^{-1} = P(R_i = 1, i \in S)^{-1}$とする。完全ケースの不偏推定方程式をウェイト$w_i$による加重平均とする推定方法が、IPW（inverse probability weighting）である。IPW の典型的な例として、ウェイトw_iによる Horvitz-Thompson 推定量（HT 推定量）がある。標本調査理論における HT 推定量では、標本抽出デザインによって目標母集団要素の包含確率が決まるので、ウェイトは既知である。一方欠測データ処理としての IPW では、ウェイトはデータから推定するしかない。

レコードiが観測される確率の逆数w_iで重み付けるという処理は、観測データをいわば膨らませることで欠測データ部分の欠けを補っていると考えることができる。このとき、完全ケースとなったレコードiがw_i倍に膨らんでいる。一方で、（定数1という変数についての HT 推定量が目標母集団の要素数の不偏推定量であることから、）分析対象となるすべての完全ケースについてウェイトw_iを合計した値の期待値は、目標母集団の要素数に等しい。このため、完全ケースのレコードから成る部分標本を母集団のサイズにまで膨らませているといえる。さらに、与えられた不完全データを全数調査の不完全データとみなすことで、IPW では暗示的に欠測値補完がなされていると考えることができる。欠測値を含む$w_i - 1$個のレコードに、完全ケースとなったレコードiの値を代入している。

IPW による推定は、ウェイトが正しい限り一致性を示し、欠測データメカニズムを問わず欠測バイアスの問題を解決するといえる。不完全データにおける変数y_iの総計の HH 推定量$\hat{\tau}_{Incomp}^{HH}$は次式で定義される。

$$\hat{\tau}_{Incomp}^{HH} \equiv \sum_{i \in S_r} w_i y_i = \sum_{i \in U} w_i \{P(R_i = 1|i \in S) \cdot P(i \in S)\} y_i \tag{2-4-1}$$

HH 推定量$\hat{\tau}_{Incomp}^{HH}$の期待値は次式で与えられる。

$$E\left(\hat{\tau}_{Incomp}^{HH}\right) = \sum_{i \in U} w_i E(I_i R_i) y_i = \sum_{i \in U} w_i \{E(I_i)E(R_i) + Cov(I_i R_i)\} y_i \tag{2-4-2}$$

上の式から次のことが分かる。適当に層化されたすべての部分標本において条件$Cov(I_i R_i) = 0$が成り立つ場合、ウェイトw_iの推定では、標本への包含確率と標本に含まれた場合の回答確率を分けて計算ないし推定してよい。一方、いかなる部分標本の分割を行ってもすべての部分標本で条件$Cov(I_i R_i) = 0$が成立するようにはできない場合、ウェイトw_iの推定では、標本への包含確率と標本に含まれた場合の回答確率を分けて計算ないし推定してはならない。

「推定方程式（estimation equation）」による推定法（「M 推定（M-estimation）」と呼ばれることもある）へ IPW 法を適用することで一般化される。欠測が生じない場合の推定

方程式を$\sum_{i\in S} m(Y_i, X_i, \theta) = 0$としたとき、不完全データに対する次の推定方程式による推定を特に IPWCC（invers probability weighting complete case analysis）法と呼ぶ。

$$\sum_{i\in S} \frac{R_i 1[i \in S]}{P(R_i = 1, i \in S)} m(Y_i, X_i, \theta) = 0 \tag{2-4-3}$$

上の式が不偏推定方程式となるためには次式が成り立たなければならない。

$$E\left[\frac{R_i 1[i \in S]}{P(R_i = 1, i \in S)} m(Y_i, X_i, \theta)\right] = 0 \tag{2-4-4}$$

(2-4-4)式の左辺は、次式のとおり変形できる。

$$E\left[\frac{R_i 1[i \in S]}{P(R_i = 1, i \in S)} m(Y_i, X_i, \theta)\right] = E[m(Y_i, X_i, \theta)] + \frac{Cov\big(R_i 1[i \in S], m(Y_i, X_i, \theta)\big)}{P(R_i = 1, i \in S)} \tag{2-4-5}$$

欠測が生じない場合の推定方程式$\sum_{i\in S} m(Y_i, X_i, \theta) = 0$が不偏方程式であれば条件$E[m(Y_i, X_i, \theta)] = 0$が成り立つから、(2-4-4)式は条件$Cov\big(R_i 1[i \in S], m(Y_i, X_i, \theta)\big) = 0$に等しい。

○MAR の下での傾向スコアの有用性

補助変数Xによる回答の傾向スコアは次式で定義される。

$$p(X_i) \equiv Pr(R_i = 1|X_i)$$

MAR のもとでは、傾向スコア$p(X_i)$で条件付けたとき変数Y_iとその観測指標R_iは独立である。

$$f(R_i|Y_i, X_i) = f(R_i|X_i) \quad \Rightarrow \quad f\big(R_i\big|Y_i, p(X_i)\big) = f\big(R_i\big|p(X_i)\big)$$

つまり、MAR のもとでは、傾向スコア$p(X_i)$の値に基づく変数Y_iの条件付分布は、欠測パターンに関わらず同一である。

$$f\big(R_i\big|Y_i, p(X_i)\big) = f\big(R_i\big|p(X_i)\big) \quad \Leftrightarrow \quad f\big(Y_i\big|R_i, p(X_i)\big) = f\big(Y_i\big|p(X_i)\big)$$

この事実は、補助変数の数が多い場合に役に立つ。補助変数の数が多い場合、補助変数の値で条件付ける際の区分が自明ではない。たとえば、補助変数が身長、体重、年齢、性別、住所の5つである場合、身長、体重、年齢のそれぞれが中央値未満か以上か、性別は男女のいずれか、住所は北海道、東北、北陸、関東、中部、近畿、中国・四国、九州・沖縄のいずれか、に応じて条件付けることもできれば、身長、体重、年齢のそれぞれが 4 分位階層のいずれか、性別は男女のいずれか、居住都道府県は 47 都道府県のいずれか、に応じて条件付けることもできる。これに対して、傾向スコアの値で条件付ければ、条件付けの変数が 1 次元に集約されることで処理が容易になる。ただし、条件付けの区分が細かいほうがより正確に等質回答群（回答確率が互いに等しい要素の集合）に対応させられるが、細かすぎると、各層サイズが小さくなり層ご

との推定の精度は低下する点に注意を要する。

2.5 多重代入法

多重代入法は、確率的代入の考え方に基づいて、疑似完全データを複数作成する手法であるが、単一代入法である確率的回帰代入を繰り返し互いに独立に実行するものとは異なる。欠測値に関わる不確実性としては、第１に、欠測した値の背後にあるデータ生成過程自体に関する不確実性と、第２に、欠測値（の真の値）がある特定のデータ生成過程から発生するときの不確実性の２つが区別できる。単一代入法である確率的回帰代入を繰り返し互いに独立に実行するだけでは、第２の不確実性に対応することはできても、第１の不確実性を捉えることはできない。多重代入法は、２つの不確実性に対応した代入法であるといえる。
多重代入法の考え方を理解するためには、「分散分解」を理解しなければならない。一般的に次の命題が成り立つ。

ある条件による条件付分散は、当該条件を情報として包摂する条件による条件付分散の当該条件による条件付期待値と、当該条件を包摂する条件による条件付期待値の当該条件による条件付分散の和に等しい（確率変数(A,B,C)について$Var(A|B)=E[Var(A|B,C)|B]+Var(E[A|B,C]|B)$）

この法則は、特に「分散分解」と呼ばれる。

分散分解の関係式によって、推定精度の評価における単一代入法の問題点を示す。一般的に推定量$\hat{\theta}$（標本平均でも標本分散でも何でもよい）による推定精度は、推定量$\hat{\theta}$の分散$Var(\hat{\theta})$（あるいはその平方根である標準誤差）によって評価できる。観測データを与件としたときの欠測データの条件付分布を「事後予測分布」と呼ぶ。欠測値を代入値で置換えることによって作成される疑似完全データに、所定の推定処理を実行する手法（すなわち代入法）においては、疑似完全データの代入データに不確実性が内在している。その不確実性は、次の3つに区別できる。

(1)事後予測分布を与件としたときの欠測データ生成に関する不確実性
(2)事後予測分布自体の不確実性
(3)事後予測分布の推定に関する不確実性

疑似完全データを完全データとみなして推定結果を解釈することは、これらの不確実性を捨象していることになる。

標本調査における所定の推定量$\hat{\theta}^*$は、完全データが得られた場合にのみ値を求めることができる。推定量$\hat{\theta}^*$は、観測データ(Y^O,X)と欠測データ(Y^M)の関数である。

$$\hat{\theta}^{*} = \hat{\theta}^{*}\begin{pmatrix}\text{観測データ}(Y^{O},X)\\ \text{欠測データ}(Y^{M})\end{pmatrix}$$

非確率的単一代入法の代入値は観測値の関数であるから、非確率的単一代入法によって作成される疑似完全データの代入データ$\overline{Y^{M}}_{DSI}$は観測データ(Y^{O},X)の関数である。この関数は、「代入モデル」と呼ばれ、当該関数のパラメータβ及びその推定量$\hat{\beta}$に対して、次式で表す。

$$\text{代入データ}\left(\overline{Y^{M}}_{DSI}\right) = g^{DSI}\left(\text{観測データ}(Y^{O},X)\middle|\text{パラメータの推定値}(\hat{\beta})\right)$$

非確率的単一代入法による推定量$\hat{\theta}^{DSI}$は、疑似完全データを完全データとみなして算出する推定量$\hat{\theta}^{*}$である。従って、推定量$\hat{\theta}^{DSI}$は、観測データの関数である。

$$\hat{\theta}^{DSI} = \hat{\theta}^{*}\begin{pmatrix}\text{観測データ}(Y^{O},X)\\ \text{代入データ}(\overline{Y^{M}}_{DSI})\end{pmatrix}$$

$$= \hat{\theta}^{*}\begin{pmatrix}\text{観測データ}(Y^{O},X)\\ g^{DSI}\left(\text{観測データ}(Y^{O},X)\middle|\text{パラメータの推定値}(\hat{\beta})\right)\end{pmatrix}$$

他方、確率的回帰代入法の代入値は観測値及び回帰モデルの誤差項の関数であるから、確率的回帰代入法によって作成される疑似完全データの代入データ$\overline{Y^{M}}_{SSI}$は観測データ(Y^{O},X)及び回帰モデルの誤差項ε^{SSI}の関数である。ただし、観測データと代入データの関係のうち、観測データから推定される回帰モデルのパラメータ$\hat{\beta} = \hat{\beta}\left(\text{観測データ}(Y^{O},X)\right)$を介した部分を明示して、代入モデルを次式で表す。

$$\text{代入データ}\left(\overline{Y^{M}}_{SSI}\right) = g^{SSI}\left(\begin{pmatrix}\text{観測データ}(Y^{O},X)\\ \text{誤差項}(\varepsilon^{SSI})\end{pmatrix}\middle|\text{パラメータの推定値}(\hat{\beta})\right)$$

確率的回帰代入法による推定量$\hat{\theta}^{SSI}$もまた、疑似完全データを完全データとみなして算出する推定量$\hat{\theta}^{*}$である。従って、推定量$\hat{\theta}^{SSI}$は、観測データと推定された回帰モデルの誤差項の関数である。

$$\hat{\theta}^{SSI} = \hat{\theta}^{*}\begin{pmatrix}\text{観測データ}(Y^{O},X)\\ \text{代入データ}(\overline{Y^{M}}_{SSI})\end{pmatrix}$$

$$= \hat{\theta}^{*}\left(\begin{pmatrix}\text{観測データ}(Y^{O},X)\\ g^{SSI}\left(\begin{pmatrix}\text{観測データ}(Y^{O},X)\\ \text{誤差項}(\varepsilon^{SSI})\end{pmatrix}\middle|\text{パラメータの推定値}(\hat{\beta})\right)\end{pmatrix}\right)$$

標本調査における所定の推定量$\hat{\theta}^{*}$の分散について、分散分解により、次式が成り立つ。

$$Var(\hat{\theta}^*) = E\left[Var\left(\hat{\theta}^*\begin{pmatrix}\text{観測データ}(Y^O,X)\\ \text{欠測データ}(Y^M)\end{pmatrix}\middle|\text{観測データ}(Y^O,X)\right)\right]$$
$$+ Var\left(E\left[\hat{\theta}^*\begin{pmatrix}\text{観測データ}(Y^O,X)\\ \text{欠測データ}(Y^M)\end{pmatrix}\middle|\text{観測データ}(Y^O,X)\right]\right)$$

(2-5-1)

事後予測分布を与件としたときの欠測データ生成に関する不確実性は、(2-5-1)式右辺第1項の期待値の中の条件付分散及び同第2項の分散の中の条件付期待値によって捉えられる。事後予測分布自体の不確実性は、(2-5-1)式右辺第1項の期待値及び同第2項の分散によって捉えられる。(2-5-1)式は、仮に完全データが得られたとしたときの所定の推定量の分散であるから、事後予測分布の推定に関する不確実性は存在しない。

非確率的単一代入法においては、疑似完全データを完全データとみなしているので、推定量$\hat{\theta}^{DSI}$を、欠測の生じない標本調査における所定の推定量$\hat{\theta}^*$とみなしていることになる。そのことはさらに、推定量$\hat{\theta}^{DSI}$の分散が推定量$\hat{\theta}^*$の分散であると錯覚することにつながる。非確率的単一代入法による推定量$\hat{\theta}^{DSI}$の分散については、分散分解により、次式が成り立つ。

$$\begin{aligned}Var(\hat{\theta}^{DSI}) &= \left\{\begin{array}{l} E\left[Var\left(\hat{\theta}^*\begin{pmatrix}\text{観測データ}(Y^O,X)\\ g^{DSI}\left(\text{観測データ}(Y^O,X)\middle|\text{パラメータの推定値}(\hat{\beta})\right)\end{pmatrix}\middle|\text{観測データ}(Y^O,X)\right)\right] \\ +Var\left(E\left[\hat{\theta}^*\begin{pmatrix}\text{観測データ}(Y^O,X)\\ g^{DSI}\left(\text{観測データ}(Y^O,X)\middle|\text{パラメータの推定値}(\hat{\beta})\right)\end{pmatrix}\middle|\text{観測データ}(Y^O,X)\right]\right)\end{array}\right\} \\ &= E[0] + Var\left(\hat{\theta}^*\begin{pmatrix}\text{観測データ}(Y^O,X)\\ g^{DSI}\left(\text{観測データ}(Y^O,X)\middle|\text{パラメータの推定値}(\hat{\beta})\right)\end{pmatrix}\right) \\ &= Var\left(\hat{\theta}^*\begin{pmatrix}\text{観測データ}(Y^O,X)\\ g^{DSI}\left(\text{観測データ}(Y^O,X)\middle|\text{パラメータの推定値}(\hat{\beta})\right)\end{pmatrix}\right)\end{aligned}$$

(2-5-2)

(2-5-2)式第1等号右辺第1項は、観測データを与件とした推定量$\hat{\theta}^{DSI}$の条件付分散の期待値である。推定量$\hat{\theta}^{DSI}$は観測データの関数であるため、観測データを与件とするということはその関数である推定量$\hat{\theta}^{DSI}$も与件となることを意味する。従って、観測データを与件とした推定量$\hat{\theta}^{DSI}$の条件付分散自体は値が0である。このことから、(2-5-2)式第1等号右辺第1項の値は0である。

先述の通り、非確率的単一代入法では、(2-5-2)式の値を(2-5-1)式の値とみなしているのであるが、(2-5-2)式と真の姿である(2-5-1)式とは2つの点で異なる。第1に、分散分解の第1項が(2-5-2)式では0となる。この項は、観測データを与件としたときの欠

測データの不確実性（事後予測分布自体の不確実性の一部）に由来する推定量の変動（の期待値）であり、非確率的単一代入はこれを捉えていない。第2に、(2-5-1)式右辺第2項と(2-5-1)式を比較すると、前者では観測データを与件としたときの推定量の条件付期待値において欠測データに関する積分又は積算の演算がなされているが、後者では欠測データが代入データに置換えられているため、観測データを与件としたときの欠測データの不確実性が代入モデルを介した観測データに由来する不確実性に置換えられている。つまり、この項で捉えられるべき事後予測分布自体の不確実性の部分が捉えられていない。先述の通り(2-5-1)右辺第1項で捉えられるべき部分も捉えられていないので、非確率的単一代入法では、全体として事後予測分布自体の不確実性が一切捉えられていないことが分かる。

(2-5-2)式右辺を(2-5-1)式右辺第2項と比較すると、前者では推定量が観測データのみの関数とみなされているので、後者で捉えられている事後予測分布を与件としたときの欠測データ生成に関する不確実性の部分と観測データの標本誤差が、前者では観測データ部分の標本誤差に由来する変動に置換えられている。この分だけ、推定精度が過大評価される。

また、(2-5-2)式右辺では、パラメータの推定に関わる不確実性も生じている。これは、(2-5-1)式右辺には存在しなかったものである。非確率的単一代入の実施者は、代入データを欠測データの真の値であるとみなしており、従って、事後分布のパラメータの推定値を真の値とみなしているため、(2-5-2)式右辺の分散を算出する際は、パラメータを確率変数とはみなさない。このように事後予測分布の推定に関する不確実性を適切に評価しないことも、推定精度を過大評価させる効果をもつ。

確率的回帰代入法による推定量$\hat{\theta}^{SSI}$の分散については、分散分解により、次式が成り立つ。

$$
\begin{aligned}
&Var\left(\hat{\theta}^{SSI}\right) \\
&= E\left[Var\left(\hat{\theta}^{*}\left(g^{SSI}\left(\begin{array}{c}\text{観測データ}(Y^{O},X)\\ \text{誤差項}(\varepsilon^{SSI})\end{array}\middle|\ \text{パラメータの推定値}(\hat{\beta})\right)\right)\middle|\ \text{観測データ}(Y^{O},X)\right)\right] \\
&+ Var\left(E\left[\hat{\theta}^{*}\left(g^{SSI}\left(\begin{array}{c}\text{観測データ}(Y^{O},X)\\ \text{誤差項}(\varepsilon^{SSI})\end{array}\middle|\ \text{パラメータの推定値}(\hat{\beta})\right)\right)\middle|\ \text{観測データ}\right.\right. \\
&\left.\left.(Y^{O},X)\right]\right)
\end{aligned}
\qquad (2\text{-}5\text{-}3)
$$

(2-5-2)式と(2-5-3)式を比べると、確率的回帰代入法では非確率的単一代入法の問題点が回避されていることが分かる。(2-5-3)式右辺第1項については、観測データを与件としたときの推定量$\hat{\theta}^{SSI}$の条件付分散の期待値であり、推定量$\hat{\theta}^{SSI}$は観測データだけでなく回帰モデルの誤差項にも依存するので、観測データを与件としたときの推定量$\hat{\theta}^{SSI}$の条件付分散自体は回帰モデルの誤差項に由来する変動を反映して0より大きくなる。このため、確率的回帰代入法では、非確率的単一代入法の場合のように分散分解の第1項は0とはならない。(2-5-3)式右辺第2項については、回帰モデルの誤差項に関する積分又は積算の演算が、代入モデルの誤差項によって表現される欠測データ生成に関する不確実性を表しており、所定の推定量$\hat{\theta}^{*}$の分散の分散分解第2項における欠測データに関する積分又は積算の演算が(観測データを与件としたときの)欠測データ生成に関する不確実性を表していることに対応している。ただし、この誤差項が本来の欠測データに由来する不確実性を過不足なくとらえているかは自明ではない。また、事後予測分布の推定に関する不確実性を推定精度に評価させる問題は、非確率的単一代入法の場合と同様である。

まとめると、非確率的単一代入法では、疑似完全データを完全データとみなすことによって、第１に、欠測データに関する推定の不確実性が無視され、第２に、観測データを所与としたときの欠測データの条件付に関わる不確実性が無視されていることになる。

これに対して、確率的回帰代入法のような確率的な手法による推定量$\hat{\theta}^{SSI}$は、観測データだけでなく代入モデルの確率項にも依存する（$\hat{\theta}^{SSI} = \hat{\theta}^{*}\left(\text{観測データ}, \text{代入データ}\right) = \hat{\theta}^{*}\left(\text{観測データ}, h^{SSI}(\text{観測データ}, \text{確率項})\right) = \hat{\theta}^{SSI}\left(\text{観測データ}, \text{確率項}\right)$)。つまり、不完全データに対して、欠測データに関する推定の不確実性を代入モデルの確率項で捉えようとしている。(2-5-2) 式の「欠測データ(Y^{M})」を「確率項(ε)」で置換えた式も一般的に成立する。しかし、一般的な推定量$\hat{\theta}$に関して、(2-5-2) 式を計算することは容易ではないという問題がある。また、確率的回帰代入法においては、推定された回帰モデルから生成する欠測データの、乱数としての不確実性は捉えられているが、回帰モデルの推定自体に伴う不確実性、すなわち観測データを所与としたときの欠測データの条件付分布に関わる不確実性は捉えられていない。

第 2.2 節で述べたとおり、単一代入法は、MAR の下であれば、１次モーメントに関する点推定については欠測バイアスを緩和できる。しかし、MAR の下でも、標準誤差や１次超のモーメントの推定については下方バイアスを伴う（このバイアスは欠測バイアスではなく、処理に由来するバイアスである)。これに対して、多重代入法は、(2-5-1) 及び (2-5-2) 式に示した分散分解に基づいて、またデータ生成の不確実性のみならずデータ生成過程自体に関する不確実性も考

慮に入れて推定精度の評価を可能にする手法である。

○多重代入のたとえ話

多重代入法の正確な説明は本節後半部に示し、本節前半ではまず直感的な理解を目指す。分かりやすい図解がないので、やや散文的になるが、たとえ話で説明する。多重代入法は、図２－７に示すような処理である。

図２－７　多重代入法のたとえ話

多重代入法のたとえ話	実際
1. 不完全データをよくみる	・事後予測分布$f(Y^M\|Y^O,X)=\int f(Y^M,\delta\|Y^O,X)d\delta=\int f(Y^M\|Y^O,X,\delta)f(\delta\|Y^O,X)d\delta$をモデル化
2. （いかにも背後の完全データを生成しそうな）サイコロをひとつ作る（**不確実性1**）	・分布$f(\delta\|Y^O,X)$からの乱数発生で値$\delta^{(h)}$を得る
3. 作ったサイコロを振る（**不確実性2**）	・値$\delta^{(h)}$で評価した分布$f\left(Y^M\middle\|Y^O,X,\delta^{(h)}\right)$からの乱数発生で値$Y^M_{(h)}$を得る
4. 出た目を代入値としてひとつの疑似完全データができる	・疑似完全データ$\left(Y^O,Y^M_{(h)},X\right)$を得る
2～4をH回繰り返す	
5. H個の疑似完全データのそれぞれに分析を適用	・H個の疑似推定結果$\left(\hat{\theta}^{(h)},\hat{V}^{(h)}\right)_{h=1,\cdots,H}$を得る
6. H個の分析結果をRubin則に従って統合	$\hat{\theta}_{MI}=\frac{1}{H}\sum_{h=1}^{H}\hat{\theta}^{(h)}$ $\hat{V}_{MI}=W+(1+1/H)B$ $W=\frac{1}{H}\sum_{h=1}^{H}\hat{V}^{(h)},\quad B=\frac{1}{H-1}\sum_{h=1}^{H}\left(\hat{\theta}^{(h)}-\hat{\theta}_{MI}\right)^2$

不完全データをよく眺めたうえで、その不完全データの背後にある完全データを生み出しそうな“サイコロ”をひとつ作る。ここで“サイコロ”といっているのは、データ生成過程のことである。ここで、“サイコロ”（データ生成過程）というものに関して２通りの考え方がある。第１は、データの背後には真の“サイコロ”（データ生成過程）がただひとつ存在しており、データからそれについて推定しなければならないという世界観である。第２は、“サイコロ”（データ生成過程）は、いわゆる「可能存在」であり、データに基づく限りで許される範囲における可能性の広がり、としてのみ捉えうるという世界観である。図２－７多重代入法のたとえ話における「“サイコロ”をひとつ作る」というステップは、後者の世界観で理解される。“サイコロ”の可能性の広がりが、データ生成過程に関する不確実性に対応する。図２－７の第２のステップで、いろいろな可能性のなかから無作為に選び出されたひとつの“サイコロ”を、図２－７の第３のステップで振る。この「“サイコロ”を振る」というステップが、欠測値（の真の値）が、ある特定のデータ生成過程から発生するときの不確実性に対応する。図２－７の第

４のステップで、特定の偶然性をもつ疑似完全データがひとつ作成される。このような疑似完全データを、互いに独立に複数作成することで、疑似完全データの標本が得られる。図２－７の第６ステップの具体的な内容については、本節補論を参照のこと。

○多重代入法の実行例

図２－８は、第 2.2.1 節の図２－２－１～２－２－８で用いた人工的な不完全データに対して、多重代入法の実行例を示したものである。詳細は本節後半の数式を用いた説明にまわすとして、ここでは図２－７に示した多重代入法のたとえ話では捨象されていた、補助変数の役割に焦点を当てる。

まず、図２－８中の①及び②の処理は、それぞれ“サイコロ”に関する不確実性、及びデータ発生の不確実性に対応している（図２－７の「“サイコロ”を作る」ステップが図２－８の①、図２－７の「“サイコロ”を振る」ステップが図２－８の②にそれぞれ対応している）。特に、欠測データの事後分布を特定するパラメータδ_h^*が、第h疑似完全データ作成用の“サイコロ”に対応する。ここで欠測が生じていないレコードの補助変数(x_{1i}, x_{2i})は、観測データy_i^Oとともに、“サイコロ”作成における投入要素となっている（補助変数には欠測は生じないが、図２－８では、目的となる変数に欠測が生じているレコードの補助変数(x_1^O, x_2^O)と目的となる変数に欠測が生じているレコードの補助変数(x_1^M, x_2^M)を区別している）。

図２－７の第２のステップで「不完全データをみる」のは、第２のステップで「サイコロをひとつ作る」ための情報収集であるが、そこでは欠測が生じていないレコード$(y_i, x_{1i}, x_{2i})_{i\in s^R}$だけが対象となっている。欠測が生じていないレコードを考慮して“サイコロ”を作るのであるが、出来上がった“サイコロ”を振るときには欠測が生じているレコードの補助変数の情報が利用される。たとえて言えば、“サイコロの振り方”は欠測が生じているレコードの補助変数(x_1^M, x_2^M)の値に依存する（現実世界のサイコロは、強く振るか弱く振るか、角度をつけるかといった振り方によって無作為性が変化するとは考えられないが、ここでは説明の便宜上振り方に応じて目の出方が変わってくる“サイコロ”を考えている。そもそも正多面体のサイコロを考えているわけでもない）。

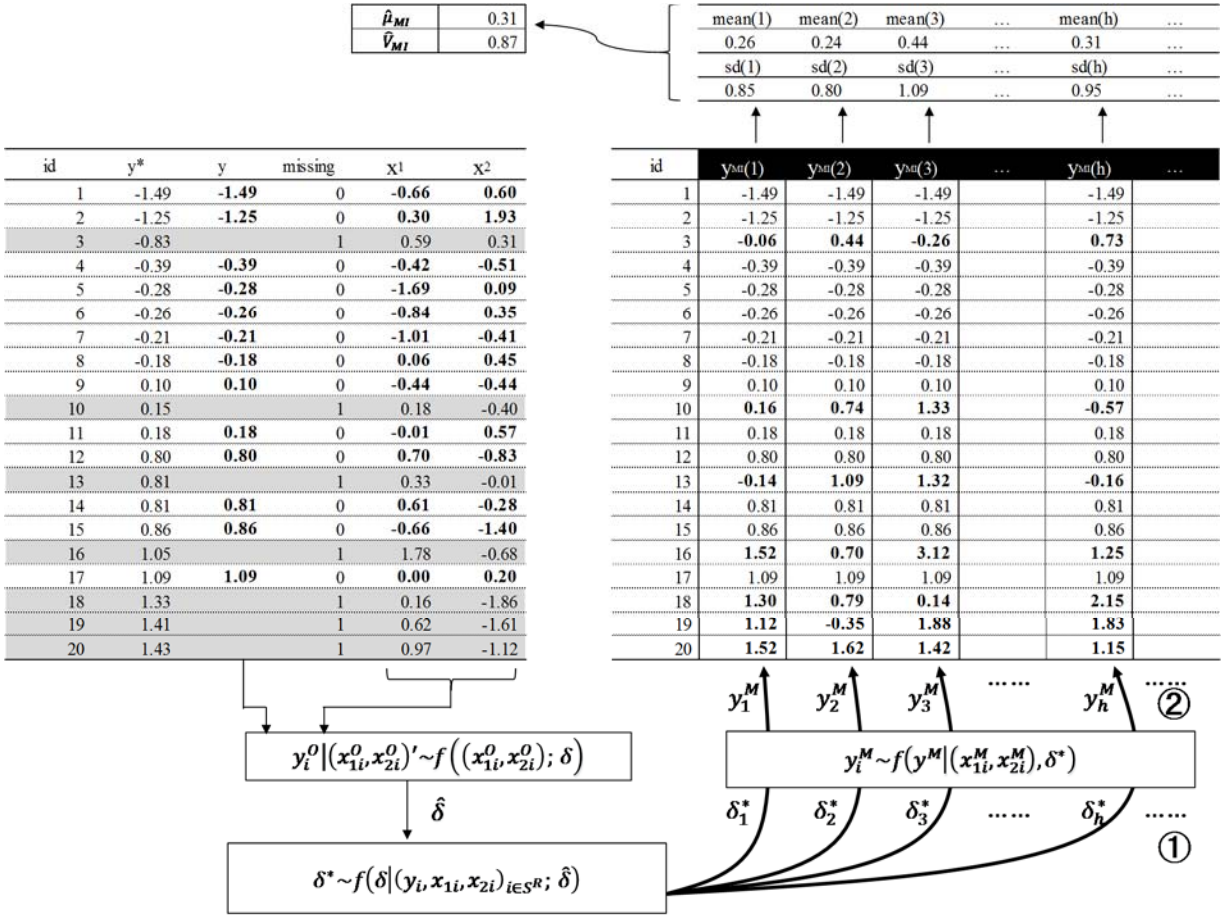
まとめると、多重代入法では、データ生成過程の可能性の広がりのなかから、互いに独立に複数の偶然性を取り出し、それらの個々の偶然性のそれぞれに対して解析を適用し、それらの個々の結果を統合することで、データ生成過程自体に関する不確実性と、データ発生に関わる不確実性を捉えた推定を行う。その際、補助変数に含まれる情報のうち、欠測値の推定に資する情報が活用される（この点は、単一代入法も同じである）。それらの情報は不確実性の範囲を狭めるといえる。欠測が生じていないレコードの情報は、データ生成過程自体の可能性の広

がりに制約を課す役割を果たし、欠測が生じているレコードの補助変数の値は、データ発生の不確実性の範囲を狭める役割を果たす。

○図２－４－３の補足説明

第 2.2.2 節の図２－３－１～２－３－３には、単一代入法に加えて多重代入法の実行結果の一例を示す。ただしこの例示では、多重代入法によって作成される疑似完全データの数は１であり、本来多重代入法が想定する適用方法ではない。ここでは、多重代入法における代入値自体の特徴を確認するためにこの図を示す。結果は、確率的回帰代入と同様のものであることが分かる。補助変数が正負を問わず目的となる変数との相関を示す場合（図２－３－１及び２－３－２）は、（図から読み取れる情報に関する限り）真の姿（各図のパネル（A））に似た代入結果となるが、補助変数が目的となる変数と無相関である場合（図２－３－３）は、代入結果は真の姿から大きく乖離する。多重代入法においても、用いる補助変数が目的となる変数と無相関である場合は、欠測バイアスを緩和する効果が期待できない。

図２－８　多重代入法の処理手順



y*: 真の値、y: 観測データ、missing: 欠測指標、(x_1, x_2): 補助変数、y_{MI}: 多重代入法による代入値

数式を用いた説明

多重代入法は、処理の手順としては次の３つの段階からなる。

第１段階：　代入ステージ

適当な補完方法（後述）により算出した値を欠測値に代入することで得られるデータを、互いに独立にH個作成する。欠測値への代入によって作成されたデータ「疑似完全データ」と呼ぶ。

第２段階：　解析ステージ

代入ステージで得られたそれぞれの疑似完全データに対して適当な推定方法を適用し、疑似完全データごとに推定値とその分散・共分散行列推定値を得る。第h疑似完全データの推定値及びその分散・共分散行列推定値をそれぞれ$\hat{\theta}^{(h)}$及び$\hat{V}^{(h)}$とする。

第３段階：　統合ステージ

解析ステージで得られたH個の推定値及びH個の分散・共分散行列推定値から、次式で定義される最終的な推定量$\hat{\theta}_{MI}$とその分散・共分散行列推定値$\hat{V}_{MI}$を計算する。

$$\hat{\theta}_{MI} \equiv \frac{1}{H}\sum_{h=1}^{H}\hat{\theta}^{(h)} \tag{2-5-4}$$

$$\hat{V}_{MI} \equiv W + \left(1+\frac{1}{H}\right)B$$

$$W \equiv \frac{1}{H}\sum_{h=1}^{H}\hat{V}^{(h)}$$

$$B \equiv \frac{1}{H-1}\sum_{h=1}^{H}\left(\hat{\theta}^{(h)} - \hat{\theta}_{MI}\right)^2 \tag{2-5-5}$$

多重代入法の上述の手順は、代入ステージの補完方法を適当に決めたとき、不完全データのなかの利用可能な情報を与件としたときのパラメータの事後分布$f(\theta|Y^O,R,X)$に基づいてベイズ推定をしているとみることができる（以下では特定のレコードを表す添え字iを省略する）。パラメータの事後分布$f(\theta|Y^O,R,X)$は、次式のとおりに分解できる。

$$f(\theta|Y^O,R,X) = \int f(\theta,Y^M|Y^O,R,X)dY^M = \int f(\theta|Y^O,Y^M,R,X)f(Y^M|Y^O,R,X)dY^M \tag{2-5-6}$$

多重代入法では、$f(Y^M|Y^O,R,X)=f(Y^M|Y^O,X)$すなわち MAR が仮定される。このとき(2-5-6)式は、次式となる。

$$f(\theta|Y^O,R,X)=\int f(\theta|Y^O,Y^M,R,X)f(Y^M|Y^O,X)dY^M$$

(2-5-7)

事後分布$f(\theta|Y^O,Y^M,R,X)$の関数形が特定化されれば、事後予測分布$f(Y^M|Y^O,X)$から互いに独立にH個の欠測変数Y^Mの値$\{Y^M_{(h)}\}_{h=1}^{H}$を発生させ（後述）、それぞれの値で事後分布$f(\theta|Y^O,Y^M,R,X)$を評価した値$f\left(\theta\middle|Y^O,Y^M_{(h)},R,X\right)$を計算すれば、次式のとおり、それらの平均値で事後分布$f(\theta|Y^O,R,X)$の近似値を求めることができる。

$$f(\theta|Y^O,R,X)\cong\frac{1}{H}\sum_{h=1}^{H}f\left(\theta\middle|Y^O,Y^M_{(h)},R,X\right)$$

(2-5-8)

(2-5-8)式は、(2-5-7)式のモンテ・カルロ法による数値計算式とみることもできる。

事後予測分布$f(Y^M|Y^O,X)$から代入値を発生させる点で、多重代入法の代入ステージにおける欠測値補完方法は、単一代入で用いられる方法とは異なる。事後予測分布$f(Y^M|Y^O,X)$は、次式のとおりに分解できる。

$$f(Y^M|Y^O,X)=\int f(Y^M,\delta|Y^O,X)d\delta=\int f(Y^M|Y^O,X,\delta)f(\delta|Y^O,X)d\delta$$

(2-5-9)

観測データと補助変数の情報の下での欠測値の推定に関する不確実性は、(2-5-9)式のパラメータδの推定に関する不確実性と、当初の手持ちの情報に加えてパラメータδの値が与えられた下での欠測値の発生に関する不確実性とによって構成される。別の言い方をすれば、分布に関する不確実性と分布からの発生に関する不確実性の２つの部分に分けられる。処理の手順としては、分布$f(\delta|Y^O,X)$から発生させた値$\delta^{(h)}$で評価した分布$f\left(Y^M\middle|Y^O,X,\delta^{(h)}\right)$から値$Y^M_{(h)}$を発生させることで、代入値を得る。

統合ステージにおける推定値の統合方法（(2-5-4)式及び(2-5-5)式）は、Rubin's rule と呼ばれる。（2-5-4）式によると、多重代入法の推定値は、疑似完全データごとの推定値の平均値である。（2-5-5）式は、多重代入法による推定の誤差が２つの部分に分けられることを示している。第１項Wは、疑似完全データごとの分散・共分散行列推定値の平均値であり、疑似完全データ内における推定の不確実性の部分を捉えている。これは、疑似完全データ内分散・共分散（within-imputation variance-covariance）と呼ばれる。第２項は、疑似完全データごとの推定値の分散

Bに調整項$(1+1/m)$を乗じたものであり、疑似完全データ間の推定値のばらつきすなわち疑似完全データ自体の不確実性の部分を捉えている。これは、疑似完全データ間分散・共分散（between-imputation variance-covariance）と呼ばれる。欠測データの占める割合が高い不完全データほど疑似完全データ間の違いが大きくなるため、疑似完全データ間分散は大きくなる。Rubin's rule は、繰り返し期待値の公式によって導かれる以下の２つの式にもとづいてモンテ・カルロ法による近似計算を行っているとみることができる。

$$E(\theta|Y^O,R,X)=E(E(\theta|Y^O,Y^M,R,X)|Y^O,R,X) \tag{2-5-10}$$

$$Var(\theta|Y^O,R,X) = E(Var(\theta|Y^O,Y^M,R,X)|Y^O,R,X)+\mathrm{Var}(E(\theta|Y^O,Y^M,R,X)|Y^O,R,X) \tag{2-5-11}$$

上の２式の各右辺で、内側の期待値オペレーションは(2-5-9)式のモンテ・カルロ法による数値計算すなわち代入ステージで発生させた疑似完全データごとの推定値計算に対応し、外側の期待値オペレーションは(2-5-7)式をウェイトに用いた計算に対応する。

多重代入法による有意水準αの区間推定は、次式で与えられる。

$$P\left(\hat{\theta}_{MI,j}\in\left[\theta-t_{\nu,1-\alpha}\sqrt{\hat{V}_{MI,jj}},\ \theta+t_{\nu,1-\alpha}\sqrt{\hat{V}_{MI,jj}}\right]\middle|Y^O,R,X\right)\cong 1-\alpha \tag{2-5-12}$$

ただし、上式において自由度 ν は次式で与えられる。

$$\nu=(H-1)\left[1+\frac{\bar{V}}{(1+1/H)B}\right]^2 \tag{2-5-13}$$

ベイズ推定であれば、次式となる。

$$P\left(\theta\in\left[\hat{\theta}_{MI,j}-t_{\nu,1-\alpha}\sqrt{\hat{V}_{MI,jj}},\ \hat{\theta}_{MI,j}+t_{\nu,1-\alpha}\sqrt{\hat{V}_{MI,jj}}\right]\middle|Y^O,R,X\right)\cong 1-\alpha \tag{2-5-14}$$

2.6 尤度法

欠測データメカニズムが MAR である場合は、補助変数の情報を活用することで、推定における欠測バイアスを緩和することができる。これに対して、欠測データメカニズムが MNAR である場合は、補助変数の活用だけでは欠測バイアスを緩和できない（すなわち、欠測バイアスの緩和に資する補助変数が利用できない）。そこで、不完全データの背後にあるデータ生成過程をモデル化することで、欠測バイアスの緩和を図

る方法として、尤度法がある。

不完全データの分析手法としての尤度法は、通常の最尤推定法を、欠測の生じるデータへ拡張したものである。最尤推定法では、データ生成過程をモデル化することで、データが発生する確率を導出し、データ発生確率をデータ生成過程のパラメータの関数とみなす。この関数は「尤度関数」と呼ばれる。最尤推定法は、与えられたデータの尤度関数を最大化するパラメータを推定量とする推定方法である。最尤推定法は、「発生したデータは最も高い確率で発生したものであろう」という推定原理に基づいている。最尤推定量が、一致性（標本サイズを増加させていくと、推定量が真の値に確率的に収束するという性質）、漸近正規性（標本サイズを増加させていくと、推定量の分布が正規分布に収束するという性質）、漸近効率性（標本サイズを増加させていくと、推定量の分散が理論的下限に収束するという性質）という望ましい性質をもつための十分条件が知られており、それらの条件のうち、モデル化が正しいという条件以外は緩やかな条件である。

不完全データの分析手法としての尤度法が、欠測の生じないデータに対する通常の最尤推定法と異なる点として、次の2つが挙げられる。第1に、不完全データの尤度関数は、データ生成過程のパラメータの関数であるだけではなく、欠測値の関数でもある。第2に、不完全データでは、欠測パターン自体がデータの構成要素となる。つまり、不完全データは、完全データとは異なる次元の情報を追加的に含んでいるため、不完全データのデータ生成過程ないし尤度関数は、完全データのデータ生成過程ないし尤度関数とは異なる次元の引数をもつ。

第1の点については、不完全データの尤度法では、尤度関数を欠測データに関して積分又は積算することによって、最尤推定法における最大化の目的関数を導出する。尤度関数を欠測データに関して積分又は積算するという処理は、「発生したデータは最も高い確率で発生したものであろう」という、最尤推定法の推定原理に即したものである。このことを、簡単な例によって示す。

硬貨を投げて表が出れば値 1、裏が出れば値 0 をとる2値変数を考える。2枚の硬貨、たとえば百円玉と五十円玉のそれぞれに、この2値変数を定義し、百円玉に対しては2値変数A、五十円玉に対しては2値変数Bとする。また2つの2値変数A及びBの和を変数Cとする（$C \equiv A + B$）。3つの変数A、B及びCのうち、任意の2つの値が分かれば残りの値も分かるので、任意の2つの変数として、変数B及びCに注目する。当該百円玉で表が出る確率をα、当該五十円玉で表が出る確率をβとすると、2つの変数として変数B及びCの同時分布は、$\Pr(B = 0, C = 0) = (1 - \alpha)(1 - \beta)$、$\Pr(B = 0, C = 1) = \alpha(1 - \beta)$、$\Pr(B = 1, C = 1) = (1 - \alpha)\beta$ 及び $\Pr(B = 1, C = 2) = \alpha\beta$（また、$\Pr(B = 0, C = 2) = \Pr(B = 1, C = 0) = 0$）である。2つの硬貨を100回投げて、各回の変数$B$及び$C$のデータを収集したとする。100回のうち30回は$(B, C) = (0, 0)$、20回は$(B, C) = (0, 1)$、25回は$(B, C) = (1, 1)$、25回は$(B, C) = (1, 2)$というデータが得ら

れたとする。この欠測が生じていないデータに対する対数尤度関数（尤度関数の対数値）$\ln L^*$ は、$\ln L^* = 30\ln(1-\alpha)(1-\beta) + 20\ln\alpha(1-\beta) + 25\ln(1-\alpha)\beta + 25\ln\alpha\beta$であるから、対数尤度関数を最大化する解の必要条件は、$\partial \ln L^*/\partial\alpha = -30/(1-\alpha) + 20/\alpha - 25/(1-\alpha) + 25/\alpha = 0$及び$\partial \ln L^*/\partial\beta = -30/(1-\beta) - 20/(1-\beta) + 25/\beta + 25/\beta = 0$となり、最尤推定の結果は、$(\hat{\alpha}, \hat{\beta}) = (0.45, 0.5)$である。

次に、どうしたわけか変数Cの値が、一部の回について観測されなかった場合を考える。上記の試行結果で、4通りのパターンのそれぞれで、5回分について変数Cの値が観測されていないとする。この場合、100回のうち25回は$(B, C) = (0, 0)$、15回は$(B, C) = (0, 1)$、20回は$(B, C) = (1, 1)$、20回は$(B, C) = (1, 2)$、10回は$(B, C) = (0, NA)$、10回は$(B, C) = (1, NA)$という結果である（ちなみに、ここで変数Cではなく変数Bに欠測が生じていれば、定義上$C = 0$ならば$B = 0$であり、$C = 2$ならば$B = 1$であるから、事実上欠測を減らすことができる。このように欠測を減らすために利用できるデータ以外の情報源を「expert knowledge」と呼ぶ）。尤度を求める際、この欠測を含む20回分については、確率測度を欠測値に関して積算する。計算としては、$(B, C) = (0, NA)$となった10回分の各々については、変数Cが値 0、1、2 のそれぞれをとる可能性を考慮して、値$\Pr(B = 0, C = 0) + \Pr(B = 0, C = 1) + \Pr(B = 0, C = 2) = (1-\alpha)(1-\beta) + \alpha(1-\beta) + 0 = 1 - \beta$、また、$(B, C) = (1, NA)$となった10回分の各々については、変数$C$が値 0、1、2 のそれぞれをとる可能性を考慮して、値$\Pr(B = 1, C = 0) + \Pr(B = 1, C = 1) + \Pr(B = 1, C = 2) = 0 + (1-\alpha)\beta + \alpha\beta = \beta$が、それぞれの欠測値に関して積算された尤度となる。$(B, C) = (0, NA)$となった10回分のそれぞれの尤度$1 - \beta$は、単に当該五十円玉で裏が出る確率であり、また、$(B, C) = (1, NA)$となった10回分のそれぞれの尤度βは、単に当該五十円玉で表が出る確率である。つまり、欠測値に関する積算（連続変数の場合は積分）という処理は、観測された変数のみの分布に基づいて尤度を求めることにほかならない。この欠測が生じているデータに対する対数尤度関数（尤度関数の対数値）$\ln L$は、$\ln L = 25\ln(1-\alpha)(1-\beta) + 15\ln\alpha(1-\beta) + 20\ln(1-\alpha)\beta + 20\ln\alpha\beta + 10\ln(1-\beta) + 10\ln\beta$であるから、対数尤度関数を最大化する解の必要条件は、$\partial \ln L/\partial\alpha = -25/(1-\alpha) + 15/\alpha - 20/(1-\alpha) + 20/\alpha = 0$ 及び $\partial \ln L/\partial\beta = -25/(1-\beta) - 15/(1-\beta) + 20/\beta + 20/\beta - 10/(1-\beta) + 10/\beta = 0$となり、最尤推定の結果は$(\hat{\alpha}, \hat{\beta}) = (0.4375, 0.5)$である。欠測値に関して積算（連続変数の場合は積分）した尤度関数は、「観測データ尤度（observed-data likelihood）」関数と呼ばれる。

第2の点は説明を要する。不完全データとそれに対応する完全データの相違を理解するためには、一見逆説的な次の事実が重要である。すなわち、不完全データは、それに対応する完全データの情報の一部に覆いを掛けたものに等しいが、不完全データにはそれに対応する、完全データには含まれない情報が追加的に含まれている。

その追加的に含まれる情報とは、「覆いの掛けられ方に関する情報」、つまり欠測パターンに関する情報である。覆いの掛けられていない完全データでは、どのように覆いが掛けられる可能性が高いか、あるいはそもそも覆いが掛けられる可能性があるのか、ということ（つまり欠測パターンの確率分布）に関して、推定する手掛かりとなる情報は一切含まれていない。

不完全データの分析手法としての尤度法では、MNAR の欠測データメカニズムに対して、不完全データに追加的に含まれた「覆いの掛けられ方に関する情報」を、欠測パターンの分布に関する推定に資する情報として有効に活用する。もちろん統計調査の目的は、興味の対象となる変数の分布に関する推定であって、欠測パターンの分布に関する推定ではない。それでも本来の目的のために「欠測パターンの分布に関する推定に資する情報」が活用できるのは、欠測データメカニズムとして MNAR を想定するからである。MNAR のもとでは、欠測パターンがどのように発生するかということと、興味の対象となる変数の値がどのような値をとるかということとの間に相互依存関係があるため、欠測パターンがどのように発生するかということを推定する上で役に立つ情報は、興味の対象となる変数の値がどのように発生するかということを推定する上でも役に立つのである。

ここで、MAR と MNAR のそれぞれで活用する情報を比較すると、次のようになる。欠測データメカニズムが MAR である場合は、補助変数の情報を活用することで、興味の対象となる変数の値と欠測パターンとの間の相互依存性を取り除くことができるので、補助変数の活用だけで興味の対象となる変数に関する推定から欠測バイアスは除かれる。他方、欠測データメカニズムが MNAR である場合は、補助変数の値で条件付けてもなお興味の対象となる変数の値と欠測パターンとの間に相互依存関係が残るので、その残された相互依存関係をモデル化したうえで、補助変数の情報だけではなく上述の「欠測パターンの分布に関する推定に資する情報」を活用することで興味の対象となる変数に関する推定から欠測バイアスは除かれる。

次に、上で説明した不完全データの尤度法が、通常の最尤推定法と異なる2つの点（通常の尤度関数が欠測値の関数となること及び欠測パターン自体がデータとなること）を踏まえて、不完全データの尤度関数を導く考え方を説明する。不完全データは、それに対応する完全データの情報の一部に覆いを掛けたものに等しいので、不完全データのデータ生成過程は、（1）それに対応する完全データのデータ生成過程と（2）完全データに覆いを掛ける確率的過程という2つのデータ生成過程が合成されたものとみることができる。前者を「興味の対象となるデータ生成過程」と呼ぶことにして、後者は「欠測データメカニズム」に他ならない。このようにみた不完全データのデータ生成過程を、「全データのデータ生成過程（generating process of full-data）」と呼び、全データのデータ生成過程から導かれる尤度関数を「全データ尤度（full-data likelihood）」関数と呼ぶ。

全データ尤度関数は、補助変数Xの値を所与としたときの興味の対象となる変数Yとその観測指標Rの条件付同時分布の確率密度（質量）関数に等しい。従って、尤度法におけるモデル化は、当該同時分布の特定化である。この同時分布自体を特定化することは可能であるが、その場合、欠測データに関する積分又は積算という処理によって、最尤推定における最大化の目的関数となる観測データ尤度関数が複雑になることに注意を要する。全データのデータ生成過程は、興味の対象となるデータ生成過程と欠測データメカニズムを合成したデータ生成過程であるとみなしたとき、全データ尤度関数は、それを構成する2つのデータ生成過程のそれぞれに対応する尤度関数に分解することができる。興味の対象となるデータ生成過程と欠測データメカニズムをそれぞれモデル化して、それぞれのモデルから導かれる尤度関数に全データ尤度関数を分解する場合のモデルは、「選択モデル」と呼ばれる（補論参照）。MNAR の下では、興味の対象となるデータ生成過程と欠測データメカニズムの間に相互依存関係がある。選択モデルによる全データのデータ生成過程のモデル化においてこの相互依存関係を表す方法の一例として、興味の対象となるデータ生成過程のモデルの誤差項と欠測データメカニズムのモデルの誤差項の同時分布を特定化するという仕方がある。この場合、2つの誤差項の相関が0であれば MAR のモデルとなる。

○Heckman の選択モデル

欠測バイアスへの対処としての尤度法の好例として（ただし無回答による欠測ではなく、値が原理的に観測されないことによる欠測ではあるが）、Heckman の選択モデルによる賃金関数の推定がある。一般的に、労働者ごとに労働市場で提示される賃金は、労働者の学歴、職歴、年齢といった属性の関数である。この関数を特に「賃金関数」と呼ぶ。標本調査によって若年女性の賃金関数を推定したい場合、標本に選ばれた調査客体ごとに、学歴、職歴、年齢といった属性と、労働市場で提示される賃金の値をデータとして収集しなければならないが、若年女性のすべてが実際に労働市場に参加しているわけではないので、一部の調査客体については「労働市場で提示される賃金」（以下「提示賃金」）は観測されない。提示賃金の欠測は、無回答によるものではなく、原理的な観測不能性によるものである。提示賃金が観測されている調査客体のデータだけを用いて賃金関数を推定した場合、推定結果は、「若年女性の賃金関数」に関するものではなく、「働いている若年女性の賃金関数」に関するものである。

ここで、標準的なミクロ経済学理論、つまり、労働によって所得を得て消費と余暇から効用水準が決まる家計による最適化問題の解として、提示賃金が留保賃金を上回る場合に働き（労働供給が正となり）、上回らない場合は働かない（労働供給は0となる）という行動が導かれる。留保賃金は経済主体の効用関数によって決まるので、モデル化する場合は、留保賃金を当該経済主体の効用関数の決定要因（たとえば家族構成、不労所得、資産水準など）の関数とみなす。まとめると、当該標本調査のデータ生成

過程のモデルは次式で表される。

$$\text{提示賃金} = h(\text{学歴}, \text{職歴}, \text{年齢}, \cdots) + \text{賃金関数の誤差項}$$

$$\text{留保賃金} = g(\text{家族構成}, \text{不労所得}, \text{資産}, \cdots) + \text{留保賃金の誤差項}$$

$$\text{提示賃金の観測指標} = \begin{cases} 1 & (\text{提示賃金} > \text{留保賃金}) \\ 0 & (\text{提示賃金} \leq \text{留保賃金}) \end{cases}$$

このモデルの誤差項にパラメトリックな分布を仮定することで尤度関数が導かれ、最尤推定を行うことができる。

Heckmanの選択モデルによる賃金関数の推定では、提示賃金の観測の成否が、労働市場への参加の有無によって決まるが、通常の統計調査における無回答による欠測についても応用できる。その場合、無回答に関する意思決定の理論モデルがあれば、欠測データメカニズムに理論的な基礎付けが得られたことになる。標準的な経済学の原理によれば、「回答することから得られる便益≦回答することの機会費用」という条件が、無回答となる必要十分条件となる。回答することから得られる便益は、社会的規範や調査協力への謝礼が考えられる（現実には前者の方が大きい割合を占めている）。回答することの機械費用は、回答する時間や労力である。回答の便益と費用は調査客体ごとに異なり、例えば調査客体が企業であれば純便益（便益－機会費用）は企業規模や業種等の属性の関数であり、調査客体が個人であれば純便益は所得や年齢等の属性の関数と考えられる。この関数形を適当に決めれば、上述の賃金関数の場合と同様に、欠測データメカニズムのモデルが得られる。ただし実際に尤度法を適用する場合には、調査客体の行動モデルを明示的に考えずに選択モデルを便宜的に用いることもあり、それは理論的な基礎付けを欠くことになる。

【補論：欠測と識別問題】

欠測の有無を問わず尤度法を用いる場合に注意すべき点として、モデルのパラメータの識別問題がある。一般的に、母集団特性値の推定における識別問題とは推定値を導く方程式の解が不定となることである。最尤推定の場合は、尤度関数がパラメータに関して凹関数でないときに推定値が不定となる。通常はモデルを構造化する（さまざまな仮定をおく）ことで識別問題を解く。欠測を含むデータにおける識別問題は、欠測の生じなかったデータにおける識別問題よりも難しくなる。それは、完全データが得られていれば識別可能なモデルでも、不完全データに適用すれば識別不可能になることを考えれば当然である。そのことを、欠測データ処理の尤度法の簡単な例によって示す。

十分大きな目標母集団において忙しい人の割合μを推定することを目的とした標本調査を考える。興味の対象となる変数Y_iは調査客体iが忙しい場合に値1をとる2値変数である。説明のための単純化として補助変数はないとする。忙しい人が調査対象に選ばれたときに回答する確率をφ_1、忙しくない人が調査対象に選ばれたときに回答する確率をφ_0とする（先験的には忙しい人の方が調査に回答しない確率が高い（$\varphi_1 < \varphi_0$））。単純無作為抽出で選ばれた1個の調査客体が忙しい人である確率は2値変数Yの母集団平均μに等しい。

調査の結果、調査対象n人からなる標本Sのなかで、「忙しい」と回答した人がn_1人、「忙しくない」と回答した人がn_0人、回答してくれなかった人がm人であった（$n_1 + n_0 + m = n$）。忙しい回答者の尤度は、母集団から無作為に1人選んだときにその人が「忙しい」と回答する確率$\mu\varphi_1$である。忙しくない回答者の尤度は、母集団から無作為に1人選んだときにその人が「忙しくない」と回答する確率$(1-\mu)\varphi_0$である。回答してくれなかった人の尤度は、母集団から無作為に1人選んだときにその人が回答してくれない確率$\mu(1-\varphi_1) + (1-\mu)(1-\varphi_0)$である。したがって、調査の結果得られた不完全データの対数尤度$\ln L(\mu, \varphi_1, \varphi_0)$は次式のとおりである。

$$\ln L(\mu, \varphi_1, \varphi_0) = n_1 \ln \mu\varphi_1 + n_0 \ln(1-\mu)\varphi_0 + m \ln \begin{Bmatrix} \mu(1-\varphi_1) \\ +(1-\mu)(1-\varphi_0) \end{Bmatrix} \tag{2-8-1}$$

不完全データの対数尤度(2-8-1)式から得られる条件は次の3式であるが、3式のうちの1式は他の2式から導かれるので、冗長（redundant）である。

$$\begin{cases} \mu\varphi_1 = n_1/n \\ (1-\mu)\varphi_0 = n_0/n \\ \mu(1-\varphi_1) + (1-\mu)(1-\varphi_0) = m/n \end{cases}$$

したがって、3つのパラメータ$(\mu, \varphi_1, \varphi_0)$は識別されない。

m人の無回答者のなかで、実は忙しい人がm_1人で、実は忙しくない人がm_0人であるとする（$m_1 + m_0 = m$）。この情報は、当然ながら、不完全データから知ることはできない。仮に、超能力的に完全データを知ることができれば、対数尤度は次式となる（そもそもそんな超能力があれば尤度推定などする必要はないが…）。

$$\ln L(\mu, \varphi_1, \varphi_0) = n_1 \ln \mu\varphi_1 + n_0 \ln(1-\mu)\varphi_0 + \begin{Bmatrix} m_1 \ln \mu(1-\varphi_1) \\ +m_0 \ln(1-\mu)(1-\varphi_0) \end{Bmatrix} \tag{2-8-2}$$

(2-8-1)式と(2-8-2)式を比べてみると、右辺第3項が異なる。不完全データの対数尤度(2-8-1)式では無回答者がひとつのカテゴリを形成し、忙しさと回答可能性の同時確率が忙しさに関して積算されている。これに対して、仮想的な完全データの対数尤度(2-8-2)式では無回答者がさらに忙しさに関して分割されている。

対数尤度(2-8-2)式から得られる条件は次の4式である。4式のうちの1式は他の3式から導かれるので、冗長である。

$$\begin{cases} \mu\varphi_1 = n_1/n \\ (1-\mu)\varphi_0 = n_0/n \\ \mu(1-\varphi_1) = m_1/n \\ (1-\mu)(1-\varphi_0) = m_0/n \end{cases}$$

この場合、3つのパラメータ$(\mu, \varphi_1, \varphi_0)$は識別される。最尤推定のようなパラメトリック推定における識別性の喪失を欠測の影響として指摘することができる。

不完全データの対数尤度(2-8-1)式における識別問題への対処としては、たとえば、1 より大きい定数aを決め、制約条件$\varphi_1 = a\varphi_0$のような仮定をおくことが考えられる。（また、定数aの値をさまざまに変えてそれぞれの推定結果を比較すればひとつの感度分析となる。）

数式を用いた説明

全データ尤度（full-data likelihood）は、興味の対象となる変数Y及び補助変数Xを所与としたときの観測指標Rの条件付確率質量関数と、補助変数Xを所与としたときの興味の対象となる変数Yの条件付確率密度関数との積で表される。

$$f\left(Y_i, R_i = r^{(k)} \middle| X_i, \zeta\right) = f\left(R_i = r^{(k)} \middle| Y_i, X_i, \varphi\right) f(Y_i | X_i, \theta) \tag{2-8-3}$$

従って全データ尤度関数のパラメータζは、興味の対象となる変数Y及び補助変数Xを所与としたときの観測指標Rの条件付確率質量関数のパラメータφ及び補助変数Xを所与としたときの興味の対象となる変数Yの条件付確率密度関数のパラメータθの関数である。特に乗法分離性から、パラメータζはパラメータφ及びθの組(φ, θ)である。

$$\zeta = \zeta(\varphi, \theta) = (\varphi, \theta)$$

このことから、全データ尤度関数のパラメータζが識別できるということは、欠測データメカニズムのパラメータφ及び興味の対象となるデータ生成過程のパラメータθも識別できるということを意味する。

観測データ尤度（observed-data likelihood）は、全データ尤度の欠測変数$Y^{(-k)}$に関する定積分として定義される。

$$f\left(Y_i^{(k)}, R_i = r^{(k)} \middle| X_i, \tilde{\zeta}\right) = \int f\left(Y_i, R_i = r^{(k)} \middle| X_i, \zeta\right) dY_i^{(-k)} \tag{2-8-4}$$

従って観測データ尤度のパラメータ$\tilde{\zeta}$は全データ尤度のパラメータζの関数である。

$$\tilde{\zeta} = \tilde{\zeta}(\zeta)$$

ここで、先述の通り全データ尤度のパラメータζは興味の対象となる変数Y及び補助変数Xを所与としたときの観測指標Rの条件付確率質量関数のパラメータφ及び補助変数Xを所与としたときの興味の対象となる変数Yの条件付確率密度関数のパラメータθの組(φ, θ)であるから、まとめると、観測データ尤度のパラメータ$\tilde{\zeta}$はパラメータφ及びθの組(φ, θ)の関数である。

$$\tilde{\zeta} = \tilde{\zeta}(\zeta) = \tilde{\zeta}(\varphi, \theta)$$

このことから、観測データ尤度関数のパラメータ$\tilde{\zeta}$が識別できても、欠測データメカニズムのパラメータφ及び興味の対象となるデータ生成過程のパラメータθも識別できるとは限らない。

MAR の仮定の下では、興味の対象となる変数Y及び補助変数Xを所与としたときの観測指標Rの条件付確率質量関数が興味の対象となる変数のうちの観測されている変数$Y^{(k)}$及び補助変数Xを所与としたときの観測指標Rの条件付確率質量関数に等しい（$f\left(R_i = r^{(k)} \middle| Y_i, X_i, \varphi\right) = f\left(R_i = r^{(k)} \middle| Y_i^{(k)}, X_i, \varphi\right)$）から、全データ尤度の欠測変数$Y^{(-k)}$に関する定積分としての観測データ尤度は、興味の対象となる変数Y及び補助変数Xを所与としたときの観測指標Rの条件付確率質量関数と補助変数Xを所与としたときの観測されている変数$Y^{(k)}$の条件付確率密度関数の積で表すことができる。

$$\begin{aligned} f\left(Y_i^{(k)}, R_i = r^{(k)} \middle| X_i, \tilde{\zeta}\right) &= \int f\left(Y_i, R_i = r^{(k)} \middle| X_i, \zeta\right) dY_i^{(-k)} \\ &= f\left(R_i = r^{(k)} \middle| Y_i^{(k)}, X_i, \varphi\right) \int f(Y_i | X_i, \theta) dY_i^{(-k)} \\ &= f\left(R_i = r^{(k)} \middle| Y_i^{(k)}, X_i, \varphi\right) f\left(Y_i^{(k)} \middle| X_i, \tilde{\theta}\right) \end{aligned} \tag{2-8-5}$$

従って MAR の仮定の下では、観測データ尤度のパラメータ$\tilde{\zeta}$は、興味の対象となる

変数Y及び補助変数Xを所与としたときの観測指標Rの条件付確率質量関数のパラメータφ及び補助変数Xを所与としたときの興味の対象となる変数$Y_i^{(k)}$の条件付確率密度関数のパラメータ$\tilde{\theta}$の組$(\varphi, \tilde{\theta})$である

$$\tilde{\zeta} = \tilde{\zeta}\left(\varphi, \tilde{\theta}(\theta)\right) = \left(\varphi, \tilde{\theta}(\theta)\right)$$

このことから、パラメータ$\tilde{\theta}$とパラメータθが1対1対応であれば、MARの下では、観測データ尤度関数のパラメータ$\tilde{\zeta}$が識別できるということは、欠測データメカニズムのパラメータφ及び興味の対象となるデータ生成過程のパラメータθも識別できるということを意味する。ここで、欠測データメカニズムのパラメータφと興味の対象となるパラメータθの定義域が分離しているという条件($\{(\varphi, \theta)\} = \Phi \times \Theta$)が加われば、補助変数$X$を所与としたときの興味の対象となる変数$Y_i^{(k)}$の条件付確率密度関数の部分のみを目的関数とする最尤推定によって欠測バイアスを緩和することができる。この場合は、欠測データメカニズムをモデル化する必要がない。このことから、MAR と MCAR は「無視可能な欠測データメカニズム(ignorable missing data mechanism)」と呼ばれる。

ここで、(2-8-5)式に類似した式が MAR の仮定によらなくても導出できることに注意を要する。観測データ尤度は、興味の対象となる変数のうちの観測されている変数$Y^{(k)}$及び補助変数Xを所与としたときの観測指標Rの条件付確率質量関数と、補助変数Xを所与としたときの観測されている変数$Y^{(k)}$の条件付確率密度関数との積で表される。

$$f\left(Y_i^{(k)}, R_i = r^{(k)} \middle| X_i, \tilde{\zeta}\right) = f\left(R_i = r^{(k)} \middle| Y_i^{(k)}, X_i, \xi\right) f\left(Y_i^{(k)} \middle| X_i, \tilde{\theta}\right)$$

(2-8-6)

従って観測データ尤度関数のパラメータ$\tilde{\zeta}$は、興味の対象となる変数のうちの観測されている変数$Y^{(k)}$及び補助変数Xを所与としたときの観測指標Rの条件付確率質量関数のパラメータξ及び補助変数Xを所与としたときの観測されている変数$Y^{(k)}$の条件付確率密度関数のパラメータ$\tilde{\theta}$の関数である。特に、パラメータ$\tilde{\zeta}$はパラメータξ及び$\tilde{\theta}$の組$(\xi, \tilde{\theta})$である。

$$\tilde{\zeta} = \tilde{\zeta}(\xi, \tilde{\theta}) = (\xi, \tilde{\theta})$$

ここで、一般的に条件付分布$f(A|B)$と条件付分布$f(A|B, C)$の間には関係式$f(A|B) = \int f(A|B, C) f(C|B) dC$が成り立つから、興味の対象となる変数のうちの観測されている変数$Y^{(k)}$及び補助変数Xを所与としたときの観測指標Rの条件付確率質量関数は、興味の対象となる変数Y及び補助変数Xを所与としたときの観測指標Rの条件付確率質量関数と事後予測分布の積の欠測変数に関する定積分に等しい。

$$f\left(R_i = r^{(k)} \middle| Y_i^{(k)}, X_i, \xi\right) = \int f\left(R_i = r^{(k)} \middle| Y_i, X_i, \varphi\right) f\left(Y_i^{(-k)} \middle| Y_i^{(k)}, X_i, \delta\right) dY_i^{(-k)}$$

また、一般的に条件付分布$f(A|B,C)$と条件付分布$f(A,B|C)$の間には関係式$f(A|B,C)=f(A,B|C)/\int f(A,B|C)dA$が成り立つから、事後分布のパラメータ$\delta$は補助変数$X$を所与としたときの興味の対象となる変数$Y$の条件付確率密度関数のパラメータ$\theta$の関数である。

$$\delta=\delta(\theta)$$

従って、興味の対象となる変数のうちの観測されている変数$Y^{(k)}$及び補助変数Xを所与としたときの観測指標Rの条件付確率質量関数のパラメータξは、興味の対象となる変数Y及び補助変数Xを所与としたときの観測指標Rの条件付確率質量関数のパラメータφ及び補助変数Xを所与としたときの興味の対象となる変数Yの条件付確率密度関数のパラメータθの関数である。

$$\xi=\xi(\varphi,\delta)=\xi\big(\varphi,\delta(\theta)\big)$$

観測データ尤度のパラメータ$\tilde{\zeta}$がパラメータξ及び$\tilde{\theta}$の関数であり、パラメータξはパラメータφ及びθの関数であり、またパラメータ$\tilde{\theta}$はパラメータθの関数である（変数$Y^{(k)}$の変数Xによる条件付分布は変数Yの変数Xによる条件付分布の欠測変数$Y^{(-k)}$に関する定積分である）ことから、観測データ尤度のパラメータ$\tilde{\zeta}$はパラメータφ及びθの関数である。

$$\tilde{\zeta}=\tilde{\zeta}\big(\xi,\tilde{\theta}\big)=\big(\xi,\tilde{\theta}\big)=\Big(\xi\big(\varphi,\delta(\theta)\big),\tilde{\theta}(\theta)\Big)$$

記号	定義
Y_i	調査客体iについての興味の対象となる変数Yのベクトル
R_i	変数Y_iの観測指標ベクトル
$r^{(k)}$	第k欠測パターンを表す観測指標ベクトルの値
$Y_i^{(k)}$	第k欠測パターンにおける変数Y_iの観測される部分
$Y_i^{(-k)}$	第k欠測パターンにおける変数Y_iの観測されない部分
X_i	調査客体iについての補助変数Xのベクトル
ζ	全データ(Y,R)の補助変数Xによる条件付分布のパラメータ
$\tilde{\zeta}$	観測データ$\big(Y^{(k)},R=r^{(k)}\big)$の補助変数$X$による条件付分布のパラメータ
φ	観測指標Rの変数(Y,X)による条件付分布のパラメータ
θ	変数Yの変数Xによる条件付分布のパラメータ
$\tilde{\theta}$	変数$Y^{(k)}$の変数Xによる条件付分布のパラメータ
ξ	観測指標$R=r^{(k)}$の変数$\big(Y^{(k)},X\big)$による条件付分布のパラメータ
δ	変数$Y^{(-k)}$の変数$\big(Y^{(k)},X\big)$による条件付分布（事後予測分布）のパラメータ

○欠測データメカニズムと分布のモデル化の種類

MNARの下では、条件付確率質量関数$f(R_i|Y_i, X_i)$が、MCARやMARの場合とは異なり、これ以上簡単な形には変形できないので、この関数を直接扱う（モデル化）する。完全データの同時分布ないし条件付分布の推定におけるモデル化に当たっては、分布の分解方法に応じてモデルの種類が区別される。ベイズ公式により、完全データの条件付分布について次式が一般的に成り立つ。

$$f(Y_i, R_i|X_i) = f(R_i|Y_i, X_i)f(Y_i|X_i) = f(Y_i|R_i, X_i)f(R_i|X_i) \tag{2-8-7}$$

(2-8-7)式の第1等号の表現形式に基づくモデル化は「選択モデル（selection model）」と呼ばれる。(2-8-7)式の第2等号の表現形式に基づくモデル化は「パターン混合モデル（pattern mixture model）」と呼ばれる。欠測データメカニズムごとの条件は、選択モデルにおいて直接的にモデル化に反映できることが分かる（選択モデルを構成する条件付確率質量関数$f(R_i|Y_i, X_i)$の部分で欠測データメカニズムが定義されているため）。パターン混合モデルにおいて欠測データメカニズムごとの条件をモデル化に反映させる場合は、欠測データメカニズムごとの定義の同値条件（(1-9)式等）の形式で行う。

分布の分解をさらに一般化すると、次のとおりである。変数b_iを、見えない固有効果（unobserved fixed effect）や潜在変数（latent variable）といった、調査客体単位iに固有の確率変数とする。変数b_iは原理的に観測することができない。ベイズ公式により、完全データの条件付分布について次式が一般的に成り立つ。

$$\begin{aligned} f(Y_i, R_i, b_i|X_i) &= f(R_i|Y_i, b_i, X_i)f(Y_i|b_i, X_i)f(b_i|X_i) \\ &= f(Y_i|R_i, b_i, X_i)f(R_i|b_i, X_i)f(b_i|X_i) \end{aligned} \tag{2-8-8}$$

(2-8-8)式は、調査客体単位iについての目的となる変数Y_i、観測指標R_i及び調査客体単位iに固有の確率変数b_iの補助変数X_iによる条件付分布$f(Y_i, R_i, b_i|X_i)$が2通りに分解できることを表している。(2-8-8)式において、調査客体単位iに固有の確率変数b_iを定数としたとき、第1等号は選択モデル、第2等号はパターン混合モデルである。また、調査客体単位iに固有の確率変数b_iが条件$Y_i \perp R_i|b_i, X_i$を満たすものであれば、第1等号と第2等号は同じ表現形式となり、「パラメータ共有モデル（shared parameter model）」と呼ばれる。

○Heckmanの選択モデルの詳細

興味の対象となる変数Y_i、観測指標R_i、補助変数(X_i, Z_i)に対して、Heckman の選択モデルは次式で表される。

$$\begin{cases} Y_i = X_i'\beta + \varepsilon_i \\ R_i = 1[Z_i'\gamma + \xi_i > 0] \\ \begin{pmatrix}\varepsilon_i \\ \xi_i\end{pmatrix} \sim N\left[\begin{pmatrix}0 \\ 0\end{pmatrix}, \begin{pmatrix}\sigma^2 & \rho\sigma \\ \rho\sigma & 1\end{pmatrix}\right] \end{cases}$$

(2-8-9)

上記モデル(2-8-9)の第1式は、興味の対象となる変数Y_iの補助変数X_iへの回帰モデルである。回帰の部分は、補助変数X_iで条件付けたときの変数Y_iの条件付期待値$E(Y_i|X_i)$を興味の対象として、条件付期待値$E(Y_i|X_i)$の線形近似$E(Y_i|X_i) \approx X_i'\beta$と解釈することもできる。政府統計では、属性別の平均値が推定対象となることがよくあるが、その場合補助変数としては属性ダミーを用いればよい。全体の平均値が推定対象であれば、補助変数は定数項のみとすればよい。

モデル(2-8-9)の第2式は、欠測データメカニズムのモデルの本体部分である。右辺は指標関数$1[\cdot]$であり、カギ括弧内の条件が真であれば値 1 をとる2値変数である。つまり、条件$Z_i'\gamma + \xi_i > 0$が成り立てば観測指標$R_i = 1$であり（変数Y_iは観測される）、条件$Z_i'\gamma + \xi_i \leq 0$が成り立てば観測指標$R_i = 0$である（変数$Y_i$は観測されない）。このモデルでは、変数$Z_i'\gamma + \xi_i$の値がある閾値（一般性を失わずに0とされる）を超えるか超えないかで、観測されるか欠測となるかが決まっていると仮定されている。変数$Z_i'\gamma + \xi_i$は2つの部分から構成されており、第1項$Z_i'\gamma$は補助変数Z_iの線形関数であり、補助変数Z_iは観測の成否に関連しそうな変数である。パラメータγは。欠測データメカニズムのパラメータ（第 2.7 節(2-8-3)式のパラメータφに相当）であり、当然観測されないが、識別される限りにおいて推定される（このモデルでは識別される）。第2項ξ_iは、欠測データメカニズムの誤差項であり、誰にも観測されない。

モデル(2-8-9)の第3式は、興味の対象となる変数Y_iの回帰モデルにおける誤差項ε_iと、欠測データメカニズムの誤差項ξ_iの同時分布である。欠測データメカニズムの誤差項ξ_iの分散は、一般性を失うことなく値 1 に基準化される。誤差項ε_iと誤差項ξ_iの相関係数ρが0であれば、興味の対象となる変数Y_iが観測されるか否かは変数Y_i自体の値に依存しない。相関係数ρが0とは異なる値であれば、興味の対象となる変数Y_iが観測されるか否かは変数Y_i自体の値に依存する。特に、相関係数ρが正（負）であれば変数Y_iは小さい（大きい）値のときに欠測となりやすい。このモデルでは、欠測データメカニズムが、MAR であることを条件$\rho = 0$によって、MNAR であることを条件$\rho \neq 0$によってそれぞれ表している。ここで注意すべき点は、相関係数ρが有意に0と異なるかを検定することによって欠測データメカニズムが MNAR であるか MAR であるかを検証できるのは、モデル(2-8-9)が正しい限りにおいてである。第 1.3 節で強調したとおり、一般的には欠測データメカニズムをデータから検証することはできない。モデル(2-8-9)の第3式に関して最後に、誤差項の正規性が比較的強い仮定である点にも注意すべきである。

○離散変量の選択モデル

調査客体ごとの離散変量の値を2時点分収集する統計調査を考える。調査客体単位iの第1時点及び第2時点の当該離散変数をそれぞれ順にY_{1i}及びY_{2i}とする。当該離散変量が取り得る値はJ種類あるものとし、値を整数$1,\cdots,J$で表す。単純化のため、第1時点の当該離散変量Y_{1i}の値はすべての調査客体について観測され、第2時点の当該離散変量Y_{2i}に欠測が発生するものとする。

2時点分とも当該変量の値が観測された調査客体の数をn^Rとする。それらのなかで、値$(Y_1=y_1,Y_2=y_2)$であるものの数を$n_{y_1y_2}$とし、値$Y_1=y_1$であるものの数を$n_{y_1,+}$とする。$n_{y_1,+}\equiv\sum_{y_2}n_{y_1y_2}$である。第2時点の当該変量$Y_{2i}$の値が観測されなかった調査客体の数を$n^M$とする。それらのなかで、値$(Y_1=y_1,Y_2=y_2)$であるものの数を$\tilde{m}_{y_1y_2}$とし、値$Y_1=y_1$であるものの数を$m_{y_1,+}$とする。$m_{y_1,+}\equiv\sum_{y_2}\tilde{m}_{y_1y_2}$である。当然ながら、$\tilde{m}_{y_1y_2}$の値は分析者に知られていない。以上をまとめると、下表のように示すことができる。背景淡灰色の部分は、値を知ることができないことを表している。

		Y_2			Total	Y_2			Dropouts
		1	⋯	J		1	⋯	J	
Y_1	1	n_{11}	⋯	n_{1J}	$n_{1,+}$	$\tilde{m}_{11}$	⋯	$\tilde{m}_{1J}$	$m_{1,+}$
	⋮	⋮		⋮	⋮	⋮		⋮	⋮
	J	n_{J1}	⋯	n_{JJ}	$n_{J,+}$	$\tilde{m}_{J1}$	⋯	$\tilde{m}_{JJ}$	$m_{J,+}$
Total					n^R				n^M

完全データが値$(Y_1=y_1,Y_2=y_2)$である場合の条件付観測確率を$\varphi_{y_1y_2}$とする。

$$\varphi_{y_1y_2}\equiv P(R_i=1|Y_1=y_1,Y_2=y_2) \tag{2-8-10}$$

完全データの同時分布のパラメータを次式のとおりに定義する。

$$\begin{cases}\pi_{y_1}\equiv P(Y_1=y_1)\\ \pi_{y_2|y_1}\equiv P(Y_2=y_2|Y_1=y_1)\end{cases} \tag{2-8-11}$$

第2時点の当該変量Y_{2i}の値が観測された調査客体の部分標本をS_rとし、第2時点の当該変量Y_{2i}の値が観測されなかった調査客体の部分標本をS_mと観測データ尤度Lは次式で与えられる。

$$\ln L=\sum_{i\in S_r}\left(\ln\pi_{y_{1i}}+\ln\pi_{y_{2i}|y_{1i}}+\ln\varphi_{y_1y_2}\right)+\sum_{i\in S_m}\left(\ln\pi_{y_{1i}}+\ln\sum_{y_2}\pi_{y_{2i}|y_{1i}}\left(1-\varphi_{y_1y_2}\right)\right)$$

3. 感度分析

第 2.2～2.6 節で、欠測を含むデータの統計的処理方法として、単一代入法、キャリブレーション推定法、IPW 法、多重代入法、及び尤度法を説明した。これらの手法の適性を決める諸条件のうち最も重要なものは、欠測データメカニズムである。MAR の条件下では、適切な補助変数の利用によって欠測バイアスが緩和される。単一代入法、キャリブレーション推定法、IPW 法、及び多重代入法は、MAR の下で、補助変数の利用により欠測バイアスを緩和する手法である。他方、MNAR の条件下では、補助変数の利用だけでなく、欠測データメカニズムのモデル化によって欠測バイアスの緩和を図る。欠測データメカニズムを明示的に（選択モデル）、あるいは非明示的に（パターン混合モデル）モデル化し、モデルから導かれる観測データ尤度により最尤推定を行う方法が、不完全データ分析手法としての尤度法である。

第 1.3 節で述べたとおり、手法ごとの適性を決める条件である欠測データメカニズムは、分析者が手にしている不完全データからは検証不可能である。換言すれば、手法ごとの推定結果から導かれる結論は、検証不可能な前提に基づいている。そこで、検証不可能な前提条件を変化させたときに、推定結果がどのように変化するかを確認する作業が必要になる。この作業は、「感度分析」と呼ばれる。感度分析によって、前提条件を常識的な範囲で変化させても推定結果に大きな変化が生じなければ、頑健な結論を導くことができる。あるいは逆に、推定結果に大きな変化をもたらさないような前提条件の領域を特定することで、導かれる結論の頑健性の程度を知ることができる。

感度分析は明確に定義される概念ではない（※たとえば、季節調整法「X-12-ARIMA」の事後診断における安定性分析も、時系列末端で中心化移動平均を算出するために必要な時系列値を欠測値とみなせば、広義の感度分析と考えることができる）が、典型的な感度分析では、MAR を特殊形として含む MNAR モデルに基づいて、MAR 及び MNAR のそれぞれを前提条件とした推定結果を比較する。選択モデルによる感度分析の具体例を次に示す。

例１．選択モデルによる感度分析の例

調査項目として、来期に景気はよくなると思うかを、「はい」か「いいえ」で回答してもらう調査を、1 千人の調査客体から成る同じ標本で2時点にわたって実施する。調査客体iの第1時点における回答Y_{1i}及び第2時点における回答Y_{2i}は2値変数で、「はい」の場合は値1、「いいえ」の場合は値0をとる。単純化のため、第1時点はすべての調査客体が回答し、第2時点では脱落が生じ一部の調査客体が無回答であるとする。観測指標は、2値変数R_iによって定義される。第2時点の回答が得られていれば観測指標の値は1（$R_i = 1$）であり、第2時点の回答が得られていなければ観測指標の値は0

$(R_i = 0)$である。調査の結果、表３－１に示した通りの不完全データが得られたとする。

表3－1　不完全データ

		2時点とも回答$(R = 1)$		脱落$(R = 0)$	計
		第2時点			
		はい	いいえ		
第1時点	はい	50	150	350	550
	いいえ	200	200	50	450
計		250	350	400	1000

この標本調査の目的は、来期景気見通しをよしとする経済主体の割合を推定することである。第1時点は欠測が生じていないから、標本設計が正しい限り、来期景気見通しをよしとする経済主体の割合を偏りなく推定できる。第2時点は欠測が生じているため、推定に欠測バイアスが伴う可能性がある。ここでは、第2時点における来期景気見通しをよしとする経済主体の割合を、選択モデルを用いて推定する。

選択モデルの選択方程式を、次式の通りロジットモデルで定式化する。

$$\ln \frac{P\left(R = 1 \middle| Y_1,\, Y_2\right)}{1 - P\left(R = 1 \middle| Y_1,\, Y_2\right)} = \varphi_0 + \varphi_1 Y_1 + \varphi_2 Y_2$$

対数尤度は次式の通りである。

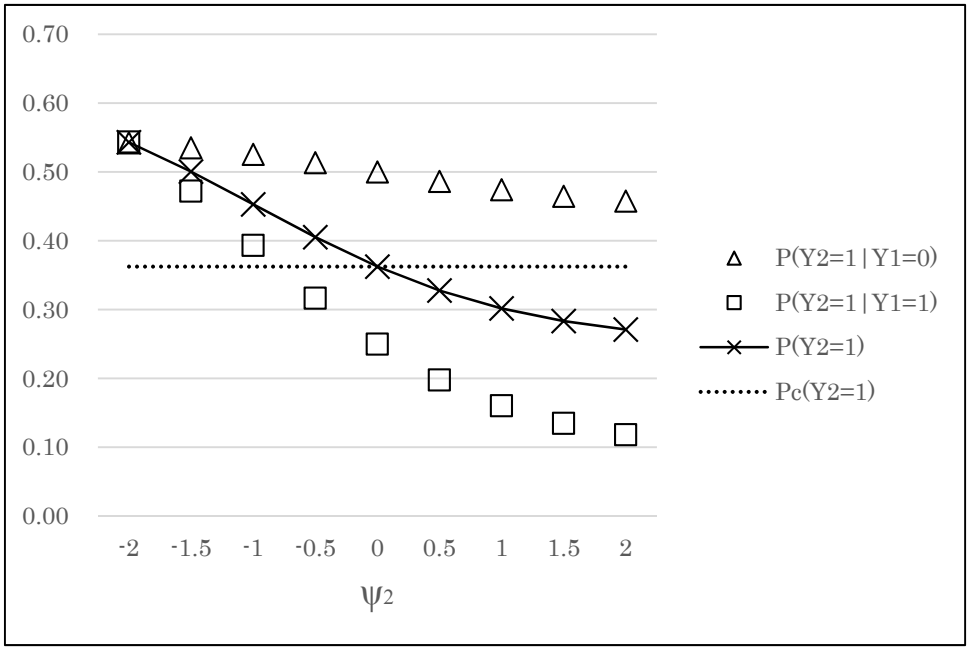
$$\ln L = \left\{ \begin{array}{l} \sum_{(y_1, y_2)} n_{y_1 y_2} \left(\ln \pi_{y_1 y_2} + \ln \dfrac{exp(\varphi_0 + \varphi_1 y_1 + \varphi_2 y_2)}{1 + exp(\varphi_0 + \varphi_1 y_1 + \varphi_2 y_2)} \right) \\ + \sum_{y_1} m_{y_1 +} \ln \left(\sum_{y_2} \pi_{y_1 y_2} \dfrac{1}{1 + exp(\varphi_0 + \varphi_1 y_1 + \varphi_2 y_2)} \right) \end{array} \right\}$$

ただし、$n_{y_1 y_2}$は$(Y_{1i} = y_1, Y_{2i} = y_2)$となる調査客体の数、$m_{y_1 +}$は$(Y_{1i} = y_1, Y_{2i} = NA.)$となる調査客体の数である。このモデルでは、パラメータφ_2の値が0であれば欠測データメカニズムはMARとなり、0でなければMNARとなる。従って、欠測データメカニズムがMARとMNARのいずれであるかという検証不可能な条件に、推定結果がどのように依存しているかを確認するためには、パラメータφ_2の値を0近傍で変化させ、それぞれの値に対する推定結果を比較すればよい。

図３－2は、パラメータφ_2の値を0近傍で変化させ、それぞれの値に対する推定結果を示したものである。パラメータφ_2の値を-2 から 2 までの範囲で変化させている。これはオッズ比で倍率 0.14～7.4 の範囲に相当する。第2時点において来期の景気見通しをよいとする経済主体の割合$P(Y_2 = 1)$を実線で連結された記号×で表す。また、推定目標である第２時点の来期景気見通しをよいとする経済主体の割合につい

て、母集団特性値の完全ケース分析による推定値（図の凡例では系列P_C）を点線で示す。パラメータφ_2の値が0である場合の推定値は、完全ケース分析の結果と一致する。欠測データメカニズムを無視した場合、第2時点の来期景気見通しをよしとする経済主体の割合の推定値は、中央の破線が示す値 0.36 である。この値が、欠測データメカニズムの条件を変えたときにどれほど変わり得るかを、感度分析によって確認できる。

図3－2　2時点にわたる2値変数データに関する感度分析の例



パラメータφ_2の値を-2 から 2 まで変化させたときに、第2時点において来期の景気見通しをよいとする経済主体の割合$P(Y_2 = 1)$は、0.54から0.27までの範囲を動き、頑健な結論は得られない。パラメータφ_2の値が小さいと、第2時点に来期の景気見通しをよしとする人ほど脱落しやすいというモデルになるので、脱落した 400 人のなかで第2時点の来期景気見通しをよしとする人の割合が大きく予測される。逆に、パラメータφ_2の値が大きいと、第2時点に来期の景気見通しをよしとしない人ほど脱落しやすいというモデルになるので、脱落した400 人のなかで第2時点の来期景気見通しをよしとする人の割合が小さく予測される。このため、第2時点において来期の景気見通しをよいとする経済主体の割合は、パラメータφ_2の値に対して右下がりの曲線を描く。

第2時点において来期の景気見通しをよいとする経済主体の割合$P(Y_2 = 1)$は、第1時点において来期の景気見通しをよくないとした経済主体のうち、第2時点の来期景気見通しをよいとする経済主体の割合$P(Y_2 = 1|Y_1 = 0)$と、第1時点において来期の景気見通しをよいとした経済主体のうち、第2時点の来期景気見通しもよいとする経済

主体の割合$P(Y_2 = 1|Y_1 = 1)$との加重平均である。図3－2では、第1時点において来期の景気見通しをよくないとした経済主体のうち、第2時点の来期景気見通しをよいとする経済主体の割合$P(Y_2 = 1|Y_1 = 0)$を記号△、第1時点において来期の景気見通しをよいとした経済主体のうち、第2時点の来期景気見通しもよいとする経済主体の割合$P(Y_2 = 1|Y_1 = 1)$を記号□で、それぞれ表す。分析の結果によると、パラメータφ_2の値を-2 から 2 まで変化させたときに、第1時点において来期の景気見通しをよくないとした経済主体のうち、第2時点の来期景気見通しをよいとする経済主体の割合$P(Y_2 = 1|Y_1 = 0)$は、0.54 から 0.46 までの範囲を動き、比較的頑健である。他方、第1時点において来期の景気見通しをよいとした経済主体のうち、第2時点の来期景気見通しもよいとする経済主体の割合$P(Y_2 = 1|Y_1 = 1)$は、0.54 から 0.12 までの範囲を動き、頑健ではない。これは、第1時点において来期の景気見通しをよくないとした経済主体 450 人のうち、脱落した者は 50 人であるのに対して、第1時点において来期の景気見通しをよいとした経済主体 550 人のうち脱落した者は 350 人にものぼり、両者の部分標本で欠測率が大きく異なることによる。当然ながら、欠測率が低いほど、感度分析の結果は頑健となる。

第2時点の来期景気見通しをよしとする人の割合は、欠測データメカニズムを無視した推定の結果によると 0.36 であるが、MNAR の可能性を考慮すると値が大きく変わる。通常の統計調査では 0.36 という推定値しか公表されないが、実はこの推定結果は欠測データメカニズムに関する仮定に強く依存していることが、感度分析から明らかとなる。

例２．パターン混合モデルによる感度分析の例

興味の対象となる変数を調査客体の身長H及び体重Wとする。身長にも体重にも欠測が生じ得るとする（項目単位の欠測）。身長の観測指標R_Hと体重の観測指標R_Wの値を与件としたときの身長と体重の条件付同時分布を特定化することで、パターン混合モデルが得られる。ここでは、欠測パターンごとの身長と体重の同時分布として正規分布を仮定する。

$$\begin{pmatrix} H \\ W \end{pmatrix} \Big| \begin{pmatrix} R_H = r_H \\ R_W = r_W \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_{H(r_H,r_W)} \\ \mu_{W(r_H,r_W)} \end{pmatrix}, \begin{pmatrix} \sigma^2_{H(r_H,r_W)} & \sigma_{HW(r_H,r_W)} \\ \sigma_{HW(r_H,r_W)} & \sigma^2_{W(r_H,r_W)} \end{pmatrix} \right)$$

このパターン混合モデルでは、未知のパラメータの数は 5×4＝20 である（欠測パターンごとの平均値ベクトル及び分散共分散行列の要素数＝5、欠測パターン数＝4）。しかし、識別可能なパラメータの数は 5＋2＋2＋0＝9 である（身長と体重の両方が観測される第１欠測パターンでは５つのパラメータ（$\mu_{H(1,1)}$，$\mu_{W(1,1)}$，$\sigma^2_{H(1,1)}$，$\sigma^2_{W(1,1)}$，$\sigma_{HW(1,1)}$）、身長のみが観測される第２欠測パターンでは２つのパラメータ（$\mu_{H(1,0)}$，$\sigma^2_{H(1,0)}$）、体重のみが観測される第３欠測パターンでは２つのパラメータ（$\mu_{W(0,1)}$，

$\sigma^2_{W(0,1)}$）が識別され、身長と体重の両方が観測されない欠測パターンでは識別されるパラメータはない）。

パターン混合モデルでは、過小識別への対処として、モデルに制約を課す。たとえば、分散共分散行列は欠測パターンによらず同一であるという制約により、パラメータの数は 20 から 11 に減る。これら 11 のうち識別可能なパラメータの数は 7 である。分散共分散が欠測パターンごとに同一であるとする制約は、強い条件ではあるものの、欠測データメカニズムが MNAR であることを妨げるものではない。さらに、平均値パラメータに関して制約条件$\pi_{H(0,0)} = \pi_{H(0,1)}$及び$\pi_{W(0,0)} = \pi_{W(1,0)}$を課す。これは、先の分散共分散同一制約とともに、4つある欠測パターンを実質的に身長と体重のそれぞれについて2つずつの欠測パターンに縮約する役割を果たす。この制約を追加することで、未知のパラメータの数は 9 となる。これら 9 のうち識別可能なものは 7 である。パターン混合モデルでは、識別のための制約条件を適当に組み替えて、制約条件の組合せごとに推定結果を比較することで、感度分析が行われる。

例３．欠測率の仮説検定に対する感度分析

第 1 節に示した母集団平均の推定における欠測バイアスの例にもとづいて、パターン混合モデルによる感度分析のごく簡単な例を示す。母集団の中で、調査対象に選ばれた場合には必ず回答する者と必ず回答しない者があらかじめ決まっているとする（これを「無回答に対する決定論的な見方（deterministic view of nonresponse）」という）。これらの集合を、それぞれ回答者母集団、無回答者母集団と呼ぶことにする。回答者母集団の母集団全体に占める割合をπ_Rとし、無回答者母集団の母集団全体に占める割合をπ_Nとする。定義上$\pi_R + \pi_N = 1$である。興味の対象となる変数の母集団平均をμとし、回答者母集団の平均をμ_Rとし、無回答者母集団の平均をμ_Nとする。定義から、$\mu = \pi_R\mu_R + \pi_N\mu_N = (1 - \pi_N)\mu_R + \pi_N\mu_N$である。ここで、統計調査実施主体の分析者が、楽観的に MCAR を仮定して完全ケース分析により母集団平均μに関する次式の仮説検定を行うとする。

$$\begin{cases} H_0: & \mu = \mu_0 \\ H_1: & \mu \neq \mu_0 \end{cases} \tag{3-1}$$

完全ケース分析の標本平均$\hat{\mu}_{Incomp}$については$E(\hat{\mu}_{Incomp}) = \mu_R$であるから、分析者の意図する仮説検定(3-1)に反して、実際は次式の仮説検定における帰無仮説モデルに直面していることになる。

$$\begin{cases} H_0': & \mu_R = \mu_0 \\ H_1': & \mu_R \neq \mu_0 \end{cases} \tag{3-2}$$

ここで、分析者の意図する仮説検定(3-2)を同値変形して次式を得る。

$$\begin{cases} H_0: & \mu_R = \dfrac{\mu_0 - \pi_N \mu_N}{1 - \pi_N} \\ H_1: & \mu_R \neq \dfrac{\mu_0 - \pi_N \mu_N}{1 - \pi_N} \end{cases} \tag{3-3}$$

分析者は、実際にはH_0'モデルに基づいているのに、帰無仮説H_0を検定しているつもりになっている。このときに、帰無仮説H_0を本当は棄却すべきでないのに、分析者が棄却と判断する確率β、及び帰無仮説H_0を本当は棄却すべきなのに、分析者が棄却せずと判断する確率γを示したのが図3－3である（※β及びγは、それぞれ「第1種の過誤の確率」及び「第2種の過誤の確率」のように聞こえるがそうではない点に注意せよ。）。

図3－3　検定の欠測率に対する感度分析の概要

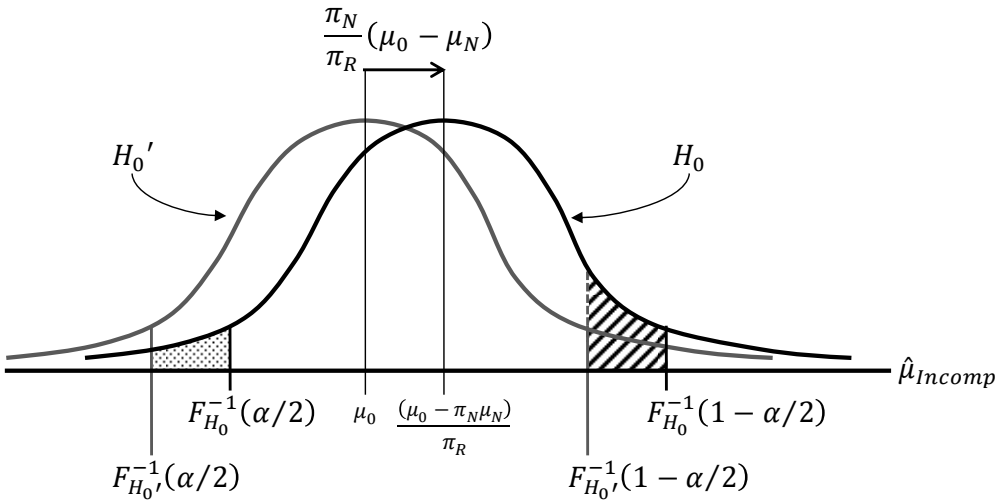


図3－3に示す通り、意図した帰無仮説H_0の下で標本平均$\hat{\mu}_{Incomp}$の従う分布（図の確率密度関数曲線H_0）は、意図されない帰無仮説H_0'の下で標本平均$\hat{\mu}_{Incomp}$の従う分布（図の確率密度関数曲線H_0'）を横軸方向に$\pi_N(\mu_0 - \mu_N)/(1 - \pi_N)$だけ平行移動したものに等しい。帰無仮説$H_0$及び$H_0'$の下での標本平均$\hat{\mu}_{Incomp}$の分布関数を、それぞれ$F_{H_0}$及び$F_{H_0'}$と表す。仮説検定の有意水準を$\alpha$としたとき、推定値$\hat{\mu}_{Incomp}$の値が2つの点$F_{H_0'}^{-1}(1-\alpha/2)$と$F_{H_0}^{-1}(1-\alpha/2)$に挟まれた区間に生じた場合、分析者は帰無仮説$H_0$を棄却すべきでないのに棄却と判断する（ただし$F_{H_0}^{-1}$及び$F_{H_0'}^{-1}$は、それぞれ順に確率分布関数$F_{H_0}$及び$F_{H_0'}$の逆関数である）。したがって、図3－3の右方に斜線で示した領域の面積が、棄却すべきでないのに棄却と判断する確率βに等しい。同様に、推定値$\hat{\mu}_{Incomp}$の値が2つの点$F_{H_0'}^{-1}(\alpha/2)$と$F_{H_0}^{-1}(\alpha/2)$に挟まれた区間に生じた場合、

分析者は帰無仮説H_0を棄却すべきであるのに棄却せずと判断する。したがって、図3－3の左方に点描で示した領域の面積が、棄却すべきなのに棄却せずと判断する確率γに等しい。βとγは次式で与えられる。

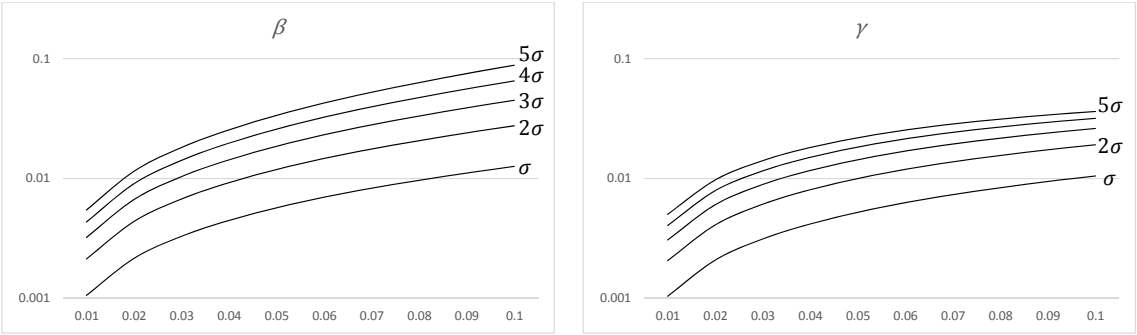
$$\beta = 1 - \frac{\alpha}{2} - F_{H_0}\left(F_{H_0'}^{-1}\left(1 - \frac{\alpha}{2}\right)\right) = 1 - \frac{\alpha}{2} - F_{H_0}\left(F^{-1}\left(1 - \frac{\alpha}{2}\middle|\mu_0\right)\middle|\frac{\mu_0 - \pi_n\mu_n}{1 - \pi_n}\right) \tag{3-4}$$

$$\gamma = \frac{\alpha}{2} - F_{H_0}\left(F_{H_0'}^{-1}\left(\frac{\alpha}{2}\right)\right) = \frac{\alpha}{2} - F_{H_0}\left(F^{-1}\left(\frac{\alpha}{2}\middle|\mu_0\right)\middle|\frac{\mu_0 - \pi_n\mu_n}{1 - \pi_n}\right) \tag{3-5}$$

分布関数Fを正規分布、検定の有意水準αを10%とした場合の感度分析の結果を、図3－4に示す。一般性を失わずに、$\mu_0 = 0$としている。図3－4のグラフは、パラメータμ_nの値を$\sigma, 2\sigma, \cdots, 5\sigma$の範囲で動かしたときの、欠測率の関数としての$\beta$値及び$\gamma$値をパラメータ$\mu_n$の値ごとに示している。

興味の対象となる変数が、先験的に正規性を有することが分かっており、また、回答群と無回答群で平均の差が常識的に、たとえば 2 標準偏差以上異なることは考えにくい場合に、不完全データで有意水準 10%の平均値の検定を行うことを考える。この場合、図3－4によると、欠測率が 4%以下であれば、棄却すべきでないのに棄却と判断する確率βは 1%以下に抑えられ、また、欠測率が 5%以下であれば、棄却すべきなのに棄却せずと判断する確率γは 1%以下に抑えられることが分かる。

図3－4　検定の欠測率に対する感度分析



4. 機械受注統計調査データを用いた分析

本節では、第2節で紹介した各手法を内閣府の「機械受注統計調査」のデータへ適用する。機械受注統計調査は、「機械等製造業者の受注した設備用機械類について、毎月の受注実績を調査したものであり、調査対象は主要機械等製造業者、調査時点は毎月月末である」(内閣府『機械受注統計調査報告』より)。1987 年 4 月実績以降、調査対象企業数は 280 社となっている。主要な調査項目には、発注部門別、受注機種別の受注額がある。

2006 年 4 月から 2014 年 3 月までの期間を分析対象期間とする。この期間の無回答発生状況は、無回答が生じた場合でも1～3社にとどまる月が多く、無回答率は概して低い。機械受注統計調査では、無回答への対応として LOCF を用いている。

機械受注統計調査の調査項目である「受注額合計」を、興味の対象とする。当該統計調査はパネルデータであることから、前回調査以前のデータが補助変数として利用できる。前回調査の発注者の経済部門(「製造業」、「非製造業」、「官公需」、「外需」等)及び受注機種(「原動機」、「重電機」、「電子・通信機械」、「産業機械」等)、前回調査以前の「受注額合計」の調査客体ごと平均値及び変動係数を、前回調査以前のデータからの補助変数として用いる。このほか、調査客体企業の所在都道府県も補助変数として利用可能である。

LOCF、層化平均値代入法、層別合計伸び率による代入、回帰代入法、確率的回帰代入法、最近傍マッチング代入法、傾向スコアマッチング代入法、IPW 法、多重代入法、及び Heckman の選択モデルによる尤度法を機械受注統計調査に適用する。このうち「層別合計伸び率による代入」以外は、第 2 節で説明した。「層別合計伸び率による代入」は、まずデータを補助変数によって層化し、前月調査も受注額が観測されている調査客体に限って層ごとに前月及び当月調査の受注額総計を求める。そして両者の値から算出される層別受注額総計の伸び率を、各層に含まれる当月無回答企業の前月値に乗じた値を、当月受注額の代入値とする。これは、個別企業の受注額伸び率という変数について、層ごとの加重平均値を代入値としているので、層化平均値代入法の亜種である。回帰代入法、確率的回帰代入法、多重代入法、及び選択モデルによる尤度法については、受注額の対数値に処理を適用した。各手法の詳細は表4－0の通りである。

表4－0　機械受注統計調査に適用した各処理法の詳細

層化平均値代入法: 前回調査以前の「受注額合計」の調査客体ごと平均値及び変動係数のそれぞれについて4分位に分割し、合計16層に層化。各層で観測値の平均値を代入。

層別合計伸び率による代入：

前回調査以前の「受注額合計」の調査客体ごと平均値及び変動係数のそれぞれについて4分位に分割し、合計16層に層化。各層で前月及び当月で受注額を回答している調査客体にそれぞれの月の限り受注額総計を求め、その伸び率を当月無回答客体の前月値に乗じた値を代入値とする。

回帰代入法・確率的回帰代入法：

前回調査の発注者の経済部門ダミー及び受注機種ダミー、前回調査以前の「受注額合計」の調査客体ごと平均値及び変動係数を補助変数とする回帰モデル。

最近傍マッチング代入法：

前回調査の発注者の経済部門ダミー及び受注機種ダミー、前回調査以前の「受注額合計」の調査客体ごと平均値及び変動係数を補助変数とし、マハラノビス距離による1対1マッチング。

傾向スコアマッチング代入法：

次式の2項ロジットモデルにより回答傾向スコアを推定し、その推定値に関して最近傍マッチング代入法を実行。

$$\mathrm{logit}\frac{\text{回答傾向スコア}_{it}}{1-\text{回答傾向スコア}_{it}}=\beta'\begin{pmatrix}(\text{受注額 within 平均})_{i,t-1}\\(\text{受注額 within 変動係数})_{i,t-1}\\(\text{発注部門ダミー})_{i,t}\\(\text{受注機種ダミー})_{i,t}\\(\text{所在都道府県ダミー})_{i}\\(\text{調査月ダミー})_{t}\end{pmatrix}$$

IPW 法：

上記傾向スコアマッチング代入法における回答傾向スコアの逆数をウェイトとする。

多重代入法：

1. 回帰モデル$(\text{受注額})_{it}\,|\,(\text{補助変数})_{it}\sim N\left(\beta'(\text{補助変数})_{it},\sigma^2\right)$を完全ケース分析により推定し推定値$\hat{\beta}$及び$\hat{\sigma}^2$を得る。
2. 以下の乱数を得る。

$$\tilde{\sigma}^2=\frac{\hat{\sigma}^2(\text{完全ケースの数}-\text{パラメータの数})}{\text{自由度}(\text{完全ケースの数}-\text{パラメータの数})\text{の}\chi^2\text{分布から発生させた乱数}}$$

$$\tilde{\beta}|\tilde{\sigma}^2\sim N\left(\hat{\beta},\tilde{\sigma}^2(X_O{}'X_O)^{-1}\right)$$

ただし、X_Oは完全ケースの補助変数データ行列

3. 以下の乱数を代入値とする。

$$(\text{受注額})_{it}\sim N(X_M\tilde{\beta},\tilde{\sigma}^2I)$$

ただし、X_Mは欠測が生じたレコードの補助変数データ行列

4. 2と3のステップを10回繰り返し、10個の疑似完全データを作成する。

ただし、$\left(\text{補助変数}\right)_{it} = \begin{pmatrix} \left(\text{受注額 within 平均}\right)_{i,t-1} \\ \left(\text{受注額 within 変動係数}\right)_{i,t-1} \\ \left(\text{発注部門ダミー}\right)_{i,t} \\ \left(\text{受注機種ダミー}\right)_{i,t} \\ \left(\text{所在都道府県ダミー}\right)_{i} \end{pmatrix}$

選択モデルによる尤度法:

下記の選択モデルを最尤推定し、調査時点毎に受注額 between 平均の推定値$\hat{\mu}_t$に調査対象企業数を乗じた値を受注額総計の推定値$\hat{\tau}_t$とする。

$$\begin{cases} \left(\text{受注額}\right)_{it} = \left(\text{受注額 between 平均}\mu\right)_{t} + \left(\text{誤差項}\varepsilon\right)_{it} \\ \left(\text{観測指標}\right)_{it} = 1\left[\gamma' \begin{pmatrix} \left(\text{受注額 within 平均}\right)_{i,t-1} \\ \left(\text{受注額 within 変動係数}\right)_{i,t-1} \\ \left(\text{発注部門ダミー}\right)_{i,t} \\ \left(\text{受注機種ダミー}\right)_{i,t} \\ \left(\text{所在都道府県ダミー}\right)_{i} \end{pmatrix} + \left(\text{誤差項}\eta\right)_{it}\right] \\ \begin{pmatrix} \varepsilon \\ \eta \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1 \\ \rho\sigma_1 & 1 \end{pmatrix}\right) \end{cases}$$

手法ごとの推定結果を図4－1に示す。手法ごとに調査客体約 280 社の受注額総計の推定値が得られるが、実際には無回答率が小さい（0～3%）ことから、手法間で推定値に大きな差は出ない。そこで、手法ごとの推定値の代わりに、図4－1では、手法ごとの「補完率」を示す。補完率は次式で定義される。

$$\text{補完率} = \frac{\text{推定値} - \text{観測値総計}}{\text{観測値総計}}$$

図4－1上段左端のグラフはLOCF、層化平均値代入法、及び層別合計伸び率による代入について、同上段中央のグラフは回帰代入法、確率的回帰代入法、最近傍マッチング代入法、及び傾向スコアマッチング代入法について、同上段右端のグラフはIPW 法、及び多重代入法について、同下段のグラフは選択モデルについて、補完率の時系列を示している。すべての手法の結果を1つのグラフに表示すると、判別できなくなるので、このように分けて示している。補完率の値が 0 の月は、全ての調査対象企業が回答した月である。選択モデル以外の手法では、多くの月で補完率が 0.1%未満であり、高い月でも 1～数%である点が共通している。*選択モデルは、他の手法と比べて独自性の強い結果である。選択モデルは、尤度法であり文字通り代入しているわけではないので、推定結果が示す負値の「補完率」は、負の代入値を意味するものでは*

ない。選択モデルから推定される受注額総計が観測値合計を下回っている。選択方程式の推定結果(本報告書では提示していない)によると、モデルが正しい限りにおいて回答確率と受注額との間にわずかながら正の相関がある(受注額を対数変換して、推定目標の方程式の誤差項と選択方程式の誤差項の相関係数パラメータの値は$\hat{\rho} = 0.09$である。ただし有意ではない)。つまり、無回答企業の受注額は比較的に小さいとの推定が働いているので、全体の平均値は押し下げられ、総額の推定値も押し下げられる。(注:機械受注統計調査の場合、推定目標は総計であるため、選択モデルによる尤度法の適用においては2通りの考え方がある。第1は、推定方程式のパラメータとしての平均の推定値に調査客体数を乗じる方法である。第2は、推定された選択モデルの理論値を欠測値に代入する方法である。尤度法の本来の考え方からは、第1の方法が適当である。)

次に、機械受注統計調査の観測データを完全データとみなして、そのデータに欠測を確率的に発生させることで、上記各手法のパフォーマンスを比較するシミュレーションを行う。欠測を発生させるモデルは、MNAR と MAR の2通りを用いる。MNAR のモデルとしては、企業iが回答する確率を企業iの受注額と企業iの所在地に依存させ、MAR のモデルとしては、企業iが回答する確率を企業iの所在地のみに依存させる。また、それぞれの欠測データメカニズムに対して、無回答企業の割合の期待値(以下「欠測率」)が 10%の場合と 20%の場合を考える。

欠測データメカニズムと欠測率の組合せ各々に対して、繰り返し互いに独立に不完全データを生成させ、それぞれの不完全データに各手法を適用する。繰返しの試行回数は 100 とした。各手法について、試行ごとの推定値及び真の値(完全データとみなしている観測データの受注額総計)から RRMSE(相対平方平均自乗誤差)を算出する。RRMSE は次式で定義される。

$$RRMSE \equiv \frac{\sqrt{\frac{1}{100}\sum_{k=1}^{100}\left(\text{第}k\text{試行の推定値} - \text{真値}\right)^2}}{\text{真値}}$$

RRMSE は、推定バイアス及び推定量の分散の増加関数である(MSE(平均自乗誤差)は推定バイアスの自乗と推定量の分散の和に等しい)ため、RRMSE の値が小さい手法ほどパフォーマンスがよいといえる。RRMSE でパフォーマンスを測る場合、確率的回帰代入法は回帰代入法に必ず劣後することが分かっているので、確率的回帰代入法を試す必要はないが、参考として含めた。

シミュレーションの結果を、図4－2－1～4に示す。LOCF と他の手法の比較を容易にするため、LOCF と比較対象となる他の手法の RRMSE を1つのグラフに重ねて表示している。LOCF の RRMSE は、2008 年後半から 2009 年にかけて高まっている。これは、リーマンショックのようなマクロショックに対して LOCF がぜい弱なためである。そ

れでも、LOCF は、層別合計伸び率による代入、回帰代入法、及び最近傍マッチング代入法とともに最もパフォーマンスの良い手法に属している。

図4－2－1～4に認められる他の点として、確率的回帰代入法は回帰代入法よりも RRMSE が大きい。これは理論的に予想されることである。また、傾向スコアマッチング代入法、IPW 法、及び選択モデルによる尤度法が、他の手法に比べて劣っている。機械受注統計調査では回答率が十分に高いために、回答傾向スコアの推定にサポート問題が生じているためである（用いる補助変数の範囲を制限し、調査月効果の pooled 2項回帰モデルとしたが、推定結果は調査月効果が支配的である）。回答率が十分高い調査には、傾向スコアの推定を伴う手法は適さない。最後に、選択モデルによる尤度法については、MNAR のモデル化における同時正規性（第 2.6 節参照）の仮定によって RRMSE がより大きくなっていると考えられる。これらの特徴は、欠測データメカニズム及び欠測率によらず、全ての結果に共通して認められる。

図4－1　推定結果

縦軸は補完率[＝(推定値－観測値総計)÷観測値総計]

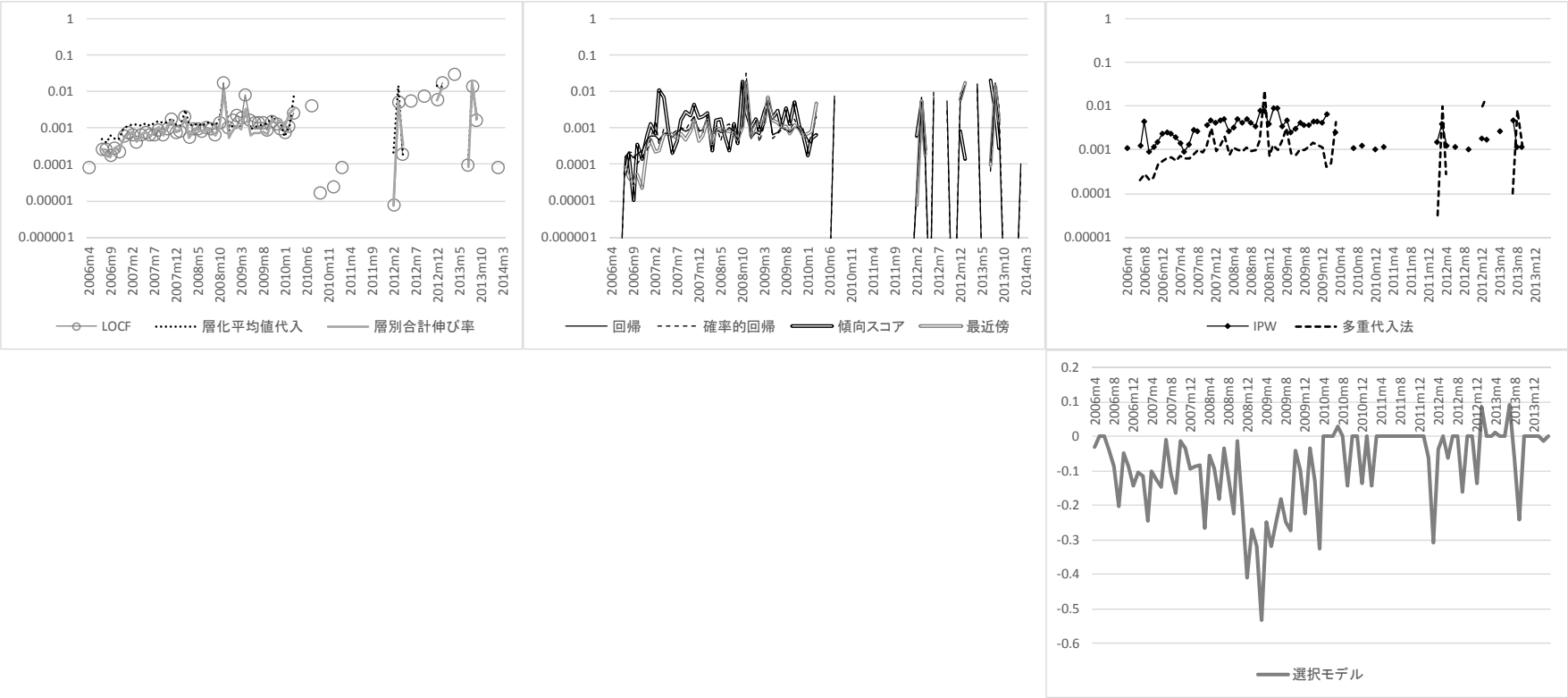


図4－2－1　シミュレーション結果（MNAR、欠測率 10%）：　縦軸は RRMSE

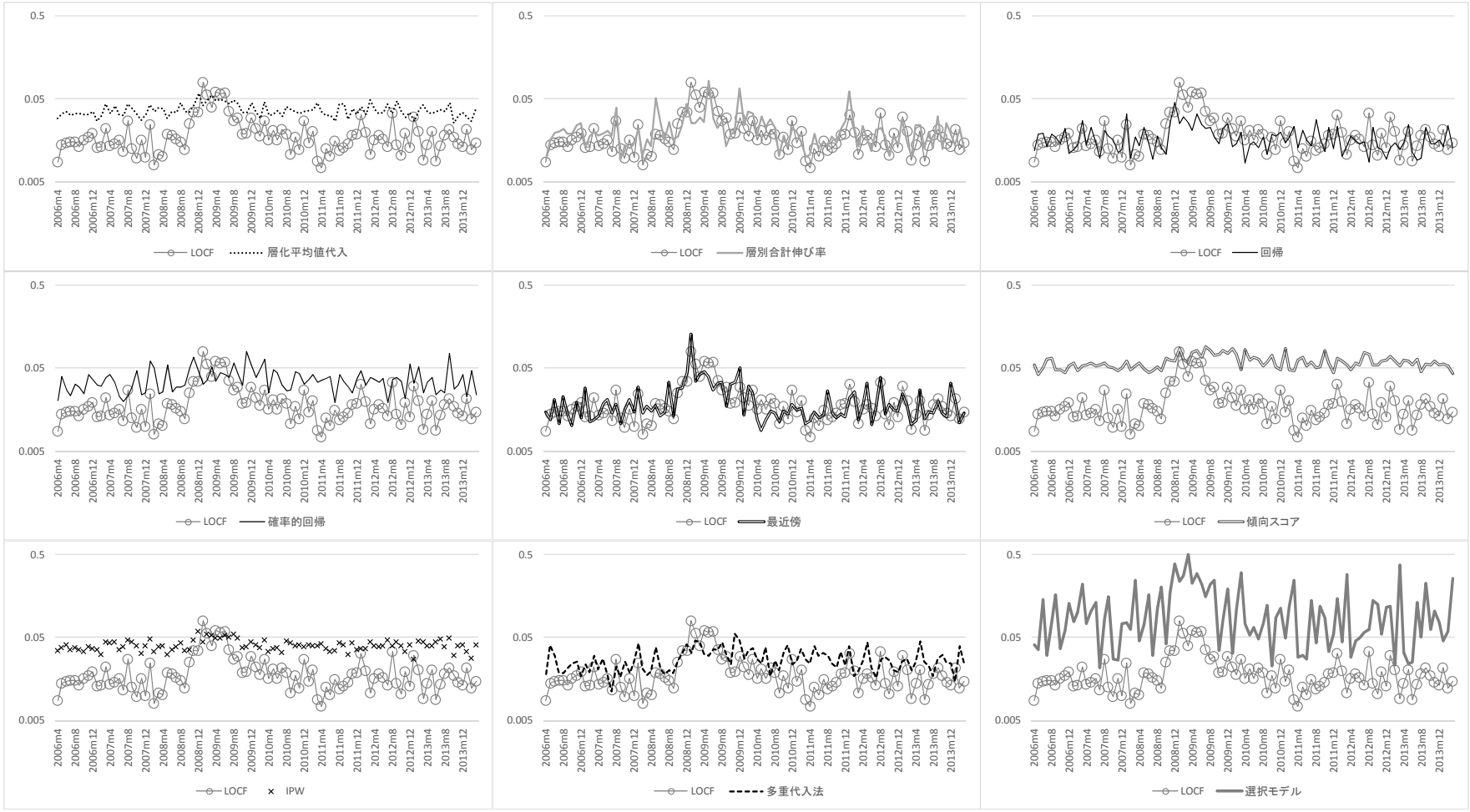
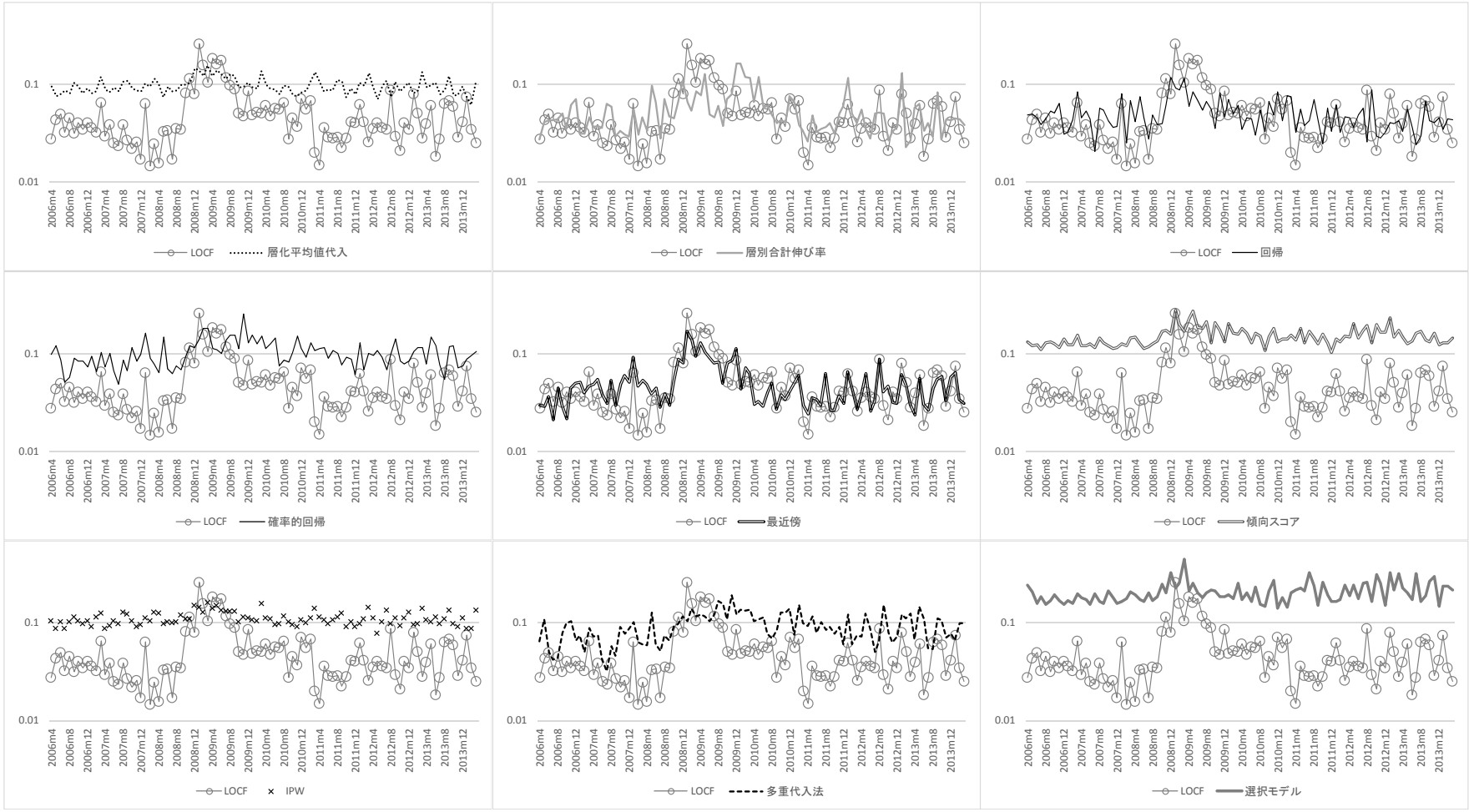


図4－2－2　シミュレーション結果（MNAR、欠測率 20%）：　縦軸は RRMSE

図4－2－3　シミュレーション結果（M AR、欠測率 10%）：　縦軸は RRMSE

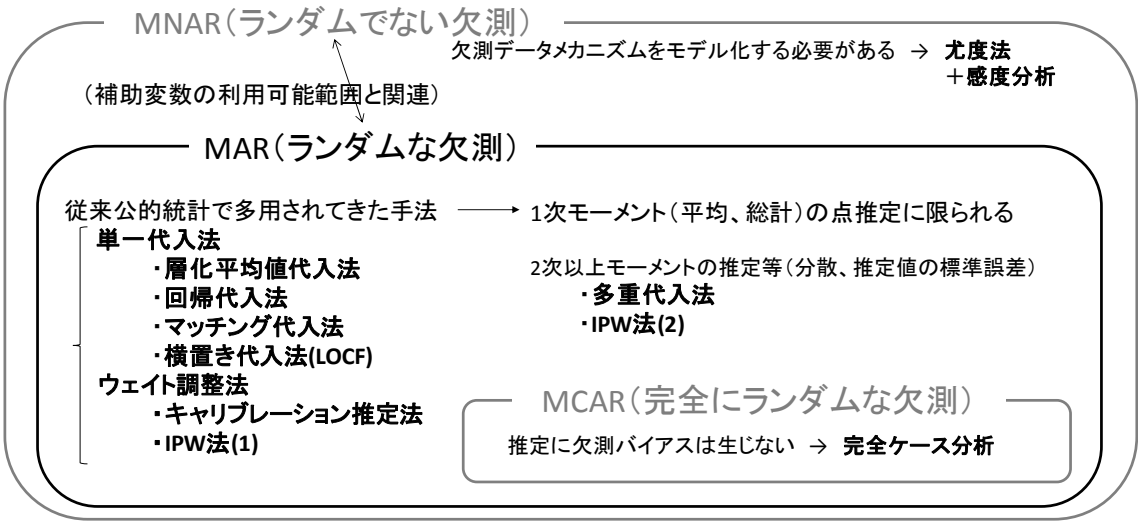
図4－2－4　シミュレーション結果（MAR、欠測率 20%）：　縦軸は RRMSE



5. まとめ

欠測を含むデータを用いた推定には、欠測バイアスと推定精度の低下という2つの問題が伴うことを第 1 節冒頭で述べた。推定精度の低下は、標本サイズの縮小による標準誤差の増大であり、この点に関する限り、対処は比較的簡単である。予め標本設計の段階で、見込まれる回答率を考慮して、標本サイズを大きめに設定しておけばよい。これに対して、欠測バイアスへの対処はより困難である。第 1 節から第 3 節では、欠測バイアスを緩和するための統計的処理方法と、それらの手法ごとの適性を決める諸条件について解説した。そのまとめを、図5－1に示す。

図5－1



不完全データの統計的処理法の適性を決める条件のうち最も重要なものは、欠測データメカニズムである。欠測データメカニズムは、MCAR（完全にランダムな欠測）、MAR（ランダムな欠測）、及びMNAR（ランダムでない欠測）の3種類に分類され、それらの間には図5－1に示した通りの包含関係がある。

MCAR の下では、推定に欠測バイアスは生じない。このため、完全ケース分析で問題ない。ただし、実際には、MCAR が成立することは非常に稀であると考えた方がよい。

MAR の下では、欠測バイアスが生じるが、適切な補助変数を利用することで緩和できる。単一代入法、キャリブレーション推定法、IPW 法、及び多重代入法によって欠測バイアスは緩和される。単一代入法及びウェイト調整法（図5－1では、回答標本を層化し、層ごとに回答率を回答傾向スコアの推定値とし、その逆数を抽出ウェイトに乗じる調整方法を、「IPW 法(1)」とし、一般的な IPW 法は「IPW 法(2)」とした）は、従来公的

統計で多用されてきた手法である。これらの手法は、1次モーメントの点推定に限って欠測バイアスを緩和する。2次モーメント以上の推定が統計調査の目的に含まれる場合は、多重代入法や IPW 法の利用が求められる。多重代入法は推定精度の評価における簡便性に利点があり、IPW 法はセミパラメトリック推定としてモデル化に対する頑健性に利点がある。

MNAR の下では、欠測バイアスが生じ、補助変数の利用だけでは緩和できない（あるいはバイアスの緩和に資する補助変数が観測されていない）ので、モデルの力を借りてバイアスを補正する。興味の対象となる変数のデータ生成過程だけではなく欠測データメカニズムもモデル化し、モデルの特定が正しい限りにおいて効率的な推定となる尤度法を用いる。欠測データメカニズムが MAR と MNAR のどちらであるかを検証することは不可能であることから、両方の可能性を考慮して、それぞれの条件を想定した推定結果を比較する作業、すなわち感度分析が求められる（第 1.3 節及び第 3 節）。欠測データメカニズムの条件を変化させても、推定結果に大きな変化が生じなければ、幸いにも、比較的頑健な結論を不完全データから得たことになる。

最後に、統計調査の実務において重要な点として、（1）補助変数の利用可能性と（2）理論モデルの役割を指摘する。第1に、理屈としては MAR と MNAR を分けるものは欠測に関するデータ生成過程であるが、実践的には、MAR と MNAR の境界を決めるものは、適切な補助変数の利用可能性である。適切な補助変数とは、それで条件付けることにより、興味の対象となる変数の条件付分布が欠測パターンごとに異ならなくなるような補助変数である。適切な補助変数が利用可能でない（観測されていない）ために MNAR を想定せざるを得ないという状況は十分考えられる。このため、母集団データベース等のフレームの整備拡充や柔軟な運用が、統計調査における不完全データの統計的処理には重要となる。第2に、MNAR への対応としてモデルの力を借りる場合、モデルの誤設定バイアスと欠測バイアスの間のトレードオフに直面する。このため、用いるモデルは、調査客体の行動原理を捉えた理論モデルから導かれることが望ましい。そのことによって、モデルのパラメータの解釈が明確になるだけでなく、誤設定バイアスの危険性が緩和される。統計調査実施者は、日常的に調査客体の行動原理に十分な関心を払うことが求められる。

【補論：最小編集箇所原則に基づく編集（Fellegi-Holt 法）】

代入法によって作成された疑似完全データはもとより、統計調査から得られた観測データにおいてさえも、データに含まれる変数の値相互間で論理的な矛盾が生じることがある。たとえば、世帯を調査単位とするデータで、父親の年齢が子の年齢を下回る場合や、企業を調査単位とするデータで、負債と自己資本の合計が資産に一致しない場合などである。誤記入や悪意の回答などによって、観測データにもこのような論理矛盾が生じる。観測データや疑似完全データにおける論理矛盾を解消する処理を、「編集（editing）」と呼ぶ。本節では、不完全データの統計的処理に関連する周辺的事項として、編集の概要を説明する。

編集では、データに含まれる変数の値の一部を別の値で置換えることで、論理矛盾を解消する。このとき、どの変数の値を、どのような値で置換えるかという問題に直面する。上述の例では、父親の年齢と子の年齢のどちらを直すべきか、あるいは、負債、自己資本、資産のいずれの項目を直すべきか、またそれぞれの場合にどの値に直すべきかという問題である。この問題に対しては、「最小編集箇所原則」と呼ばれる原則が提示されている。最小編集箇所原則とは、編集によって修正する値の数は最小限にとどめるべきであるとする原則である。

編集において、「父親の年齢＞子の年齢」、「負債＋自己資本＝資産」などの論理的に満たされるべき条件は、「編集規則（edit rules）」と呼ばれる。上述の例では、「母親の年齢＞子の年齢」、「負債＝固定負債＋流動負債」などのように、編集規則に含まれる条件式は多くある。最小編集箇所原則に則れば、編集規則の制約下で編集箇所の数を最小化するという最適化問題を解くことで、修正すべき変数が決まる。当該最適化問題は、特に「ELP（error localization problem）」と呼ばれ、ELP の解として編集箇所を決定する方法は、「Fellegi-Holt 法」と呼ばれる。上述の世帯調査の例で、父親の年齢＝25 歳、子の年齢＝26 歳、母親の年齢＝24 歳というデータであれば、「父親の年齢＞子の年齢」という編集規則の条件を満たすために、父親の年齢を編集すると、「母親の年齢＞子の年齢」という編集規則の条件を満たすためには、母親の年齢（もしくは子の年齢）も編集しなければならないが、子の年齢を編集すれば「父親の年齢＞子の年齢」及び「母親の年齢＞子の年齢」という編集規則の2つの条件が同時に満たされる。従ってこの場合は、子の年齢を直すのが望ましい。

最小編集箇所原則に依る場合、編集箇所の数ではなく、編集箇所の重み付きの数を最小化するという一般化が可能である。この一般化により、誤記入や秘匿の生じやすい変数と生じにくい変数を異なる扱いにすることができる。すなわち、変数の信頼性を表す尺度として信頼ウェイト（confidence weights）を定義し、信頼ウェイトで重み付けした編集箇所数を最小化する。

Fellegi-Holt 法によって特定された編集箇所に、どのような値を代入するかという問題に対しては、通常マッチング代入法（編集の文脈では特に「hot-deck」と呼ばれる）が用いられる。すなわち、疑似完全データないし観測データを、編集規則を満たさないレコード群と編集規則を満たすレコード群に分割し、両者間で補助変数を用いてマッチングを行う。ただし、編集後のデータが編集規則を満たすとは限らない。ひとつの編集過程（この場合は代入）で編集規則が満たされなければ、編集規則が満たされるようになるまで別の編集過程を追加する必要がある。

Fellegi-Holt 法による編集は、最小編集箇所原則に基づいて代入個所を特定するが、より一般化された編集として Scholtus (2014)の一般化 ELP がある。一般化 ELP では、編集過程を代入及び線形変換の集合と考える。上記の世帯調査の例で、調査項目を父親の年齢、母親の年齢、第1子の年齢の3項目とすると、これら3項目それぞれに関する代入、父親の年齢に定数を加える処理、母親の年齢に定数を加える処理、第1子の年齢から定数を引く処理といった編集過程の要素が考えられる。一般化 ELP では、適当に定義された編集過程に対して、編集過程の要素ごとにウェイトの値を定め、当初のデータから出発して編集規則を満たすデータに至る編集過程のうち、ウェイトの総計を最小化する編集過程の経路を選択する。編集過程の要素のウェイトが、当該編集過程の逆写像としての誤記入等が発生する確率の自然対数値に－1を乗じたものに等しければ、一般化 ELP による編集は、誤記入等発生のデータ生成過程に基づく最尤推定法の近似演算として解釈できる（Scholtus 2014）。

Fellegi-Holt 法による編集は、システム化されて公的統計で用いられている。実際に運用されている編集システムの例として、カナダ政府の編集システム「Banff」の概要を表6－1に示す。編集システム「Banff」は、9の機能を有する。表6－1で「Errorloc」と呼ばれる処理が、編集箇所を決定する。表6－1で「Donorimputation」、「Estimator」、及び「Massimputation」と呼ばれる処理が、マッチング代入や回帰代入を行う。そのほかの処理は、編集規則に関する処理、外れ値処理、診断等である。

表6－1　カナダ政府の編集システム「Banff」の機能（Kozak 2005）

Procedure Verifyedits 編集規則の整合性チェック、重複統合、端点算出、及び帰結制約（implied edits）表示。
Procedure Editstats レコードごとに、編集規則に対する真偽判定。値は、「pass（真）」、「miss」、「fail（偽）」の3つ。欠測データにより真偽判定できないレコードは値「miss」をとる。次の5つの表が出力される。 1. 個別制約条件別のレコードごと真偽判定 2. 制約条件数別のレコードごと真偽判定の分布 3. 制約条件全体に対する真偽判定結果別レコード数 4. 個別制約条件に対する真偽判定結果別関連項目延数 5. 制約条件全体に対する真偽判定結果別関連項目延数 当該表の使いみち: 偽の割合を高くするような制約条件は除外する。 制約条件が ELP の最適化に悪影響を与えていないか確認できる。 編集・代入のステップごとに当該表を出力することで各処理の効果を評価できる。
Procedure Outlier Hidiroglou-Berthelot 法により外れ値を特定。外れ値ではないものの欠測値補完に利用できないほどの振れ幅をもつような値も特定。
Procedure Errorloc ELP を解いて代入すべきレコード・項目を特定。
Procedure Deterministic 代入が必要なレコード・項目のうち、編集規則によって値が確定するものにその値を代入。
Procedure Donorimputation 代入が必要なレコード・項目にマッチング代入を実行。
Procedure Estimator 標本から推定された線形回帰モデルによって生成される値または標本推定値を代入。
Procedure Prorate 内訳項目の合計値が合計項目の値に一致することを表す等号条件を与えたときに、それらの条件が成立するように、等号条件ごとに含まれる変数値にスケール変換を実施。
Procedure Massimputation 層化抽出された標本に関して、1層目には含まれるが2層目には含まれない抽出単位に対して2層目の調査項目のマッチング代入を実行。

参考文献

阿部貴行(2016)『欠測データの統計解析』朝倉書店
高井啓二・星野崇宏・野間久史(2016)『欠測データの統計科学』岩波書店
高橋将宜・阿部穂日・野呂竜夫(2015)「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」独立行政法人統計センター製表技術参考資料 30.
土屋隆裕(2009)『概説　標本調査法』朝倉書店
星野崇宏(2009)『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
Kozak, Robert, 2005. “The BANFF System for Automated Editing and Imputation.” Proceedings of the Survey Methods Section, SSC Annual Meeting, June 2005.
Little, Roderick J. A., and Donald B. Rubin, 2002. Statistical Analysis with Missing Data. Second edition, John Wiley & Sons.
Molenberghs, Geert, and Michael G. Kenward, 2007. Missing Data in Clinical Studies. John Wiley & Sons Ltd.
Molenberghs, Geert, Garrett Fitzmaurice, Michael G. Kenward, Anastasios Tsiatis, and Geert Verbeke (Editors), 2014. Handbook of Missing Data Methodology. Chapman & Hall / CRC Press.
Scholtus, Sander, 2014. “A Generalised Fellegi-Holt Paradigm for Automatic Editing.” United Nations Economic Commission for Europe, Working Paper.