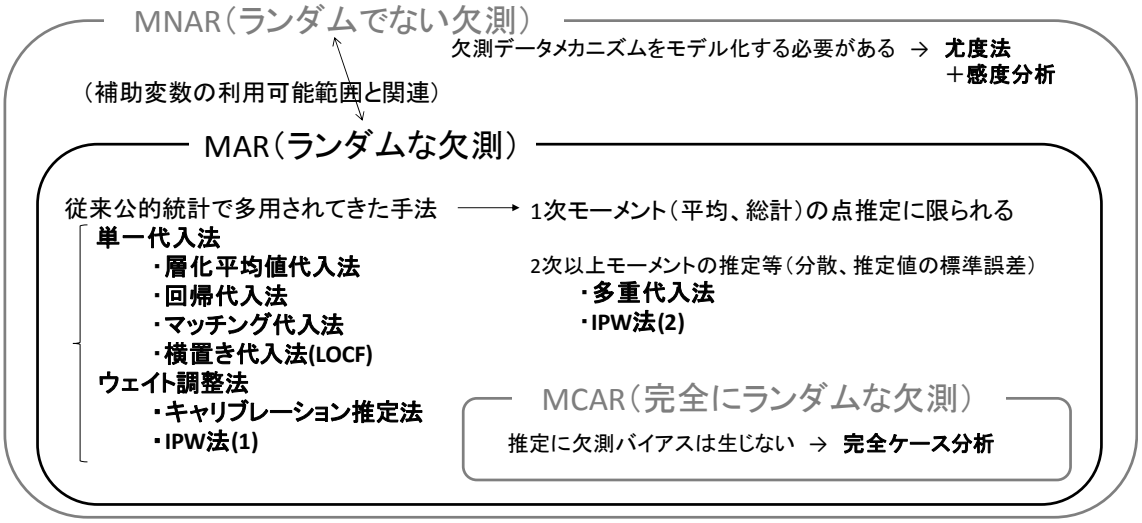
5. まとめ

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。