公的調査における欠測値代入補完と
代入者としての政府の役割

慶應義塾大学経済学部・大学院経済学研究科教授
星野崇宏

本報告書では委員会での議論と内閣府の専門職員の方々およびインテージリサーチのご尽力により、今後社会的にますます重要となる公的調査での欠測の適切な処理の基礎的な議論が丁寧に行われていると考える。

ここでは、特に欠測値の代入アプローチの重要性、および代入者としての政府の役割について議論したい。

欠測データ解析については本報告書に記載されているように完全にランダムな欠測(MCAR)・ランダムな欠測(MAR)等の欠測メカニズムの分類、単調欠測かどうかや項目単位の欠測か(Item nonresponse)ユニット単位の欠測か(Unit nonresponse)といった欠測データの形式による分類などがある。

但し欠測のあるデータを利用・公表するという観点からは、欠測値を代入補完するかしないかという観点は非常に重要である。すでに紹介された様々な解析法は、筆者が指摘しているように（高井・星野・野間,2016）

【1】欠測をある基準で無視する
- ・完全ケース解析(Complete case analysis) 観測確率の逆数で重みを付ける場合を含む
- ・利用可能なデータを用いた解析(Available case analysis)
- ・“ランダムな欠測”の下で尤度の中で欠測に関わる部分を無視

⇒観測データの尤度（直接尤度）を用いた解析

【2】欠測値に値を代入補完する
- ・単一代入

  平均値代入／回帰代入／確率的回帰代入／Hot Deck など
- ・多重代入

【3】モデリングを行う
- ・“ランダムでない欠測”での欠測のモデリングを行った解析

と大別できるが、特に欠測値の代入補完をする場合には「欠測値の代入者」と「代入によって得られた疑似完全データを利用する解析者」が異なり得ることがいくつかの前提のもとに許容されるという点が非常に重要である。

近年では多重代入法は医学データ解析でも非常によく利用されるようになったが、その多くの場合では多重代入は適切ではない。そもそも代入者の利用する代入モデルと解析者の利用する解析モ

デルが同じである場合には、それらのモデルの母数推定は観測データ尤度の最大化による最尤推定などを行う方が効率的であり、つまり欠測値を代入補完する必要はない。

社会科学分野、特に公的統計でなぜ多重代入が開発されたかと言えば、代入者の代入モデルが解析者の解析モデルと異なる場合を考えたいからである。具体的には、代入者の持つ情報が解析者よりも豊かである場合があり、また代入モデルと解析モデルが異なることが一般的である（図表 1-1 参照)。

公的統計の場合であれば、政府や自治体は回答者や企業など回答主体について、その居住地や税務情報など公的機関が所有するデータを活用することが可能であるが、個人や企業の特定を避けるためにこれ自体を解析者に個票レベルで与えることはできない。また、解析モデルにおいては関心のない変数を代入モデルにおいては利用する必要がある場合がある。

このような変数の例としては住所情報が挙げられる。都心から離れた住宅地であれば通勤時間が長くなるために、在宅時間が短くなり、訪問調査では未回収が多くなるなどといった場合には、解析者に対しては詳細な住所情報の変数を開示はできないため解析モデルには含まれないが、欠測値の代入においては利用すべき重要な補助変数になる。代入補完すべき変数が収入であれば、住所情報は家賃や地価に関連するため、代入値を決定するうえで非常に重要であるだけでなく、この変数が欠測発生確率を規定していると考えられるため、この変数を除外して欠測値を代入するということは、解析モデルにおいて MAR の仮定が成立しないために得られる推定値にバイアスを生じさせる可能性が高い。

図表 1-1 代入モデルと解析モデルの関係（高井・星野・野間 2016 の第 4 章より抜粋）

モデルのタイプ	統合した推定量の 一致制の必要条件	注意点
(a)代入モデル⊂解析モデル	代入モデルが真のモデルを含むこと	代入モデルを解析モデルにも利用した方が推定量の分散が小さくなる
(b)融和性のあるモデル	代入モデル＝解析モデルが真のモデルを含むこと	多重代入法ではなく観測データの尤度の最大化を利用する方がよい
(c)代入モデル⊃解析モデル	解析モデルは代入モデルの周辺モデルと見なせる、あるいは解析モデルが真のモデルを含むこと	解析モデルの観測データの尤度の最大化はバイアスが生じる可能性があるため多重代入がよい
(d)代入モデルと解析モデルが入れ子構造でない	代入モデルが真のモデルを含み、解析ステージで完全データからの母数の一致推定ができること	解析ステージの一致制の根拠が明確ではない

公的調査自体が GDP 等のマクロ統計指標作成等、政府の施策立案に資する情報として以外に国民全般が活用できる公共財であり、正確な情報の収集と提供は我が国の行政のみならず民間の効率的なビジネス実施にも有用である。

政府が公開する統計情報や個票データにおいても欠測が一定以上存在する場合には、個人情報や機微情報・税務情報などを保有する政府こそが適切な代入モデルを用いて欠測値を代入補完することが可能であり、また多重代入アプローチは他の解析法と異なり「代入者と解析者が異なる」「代入モデルと解析モデルが異なる」ことを許容し、これらの情報を削除して提供することが可能な方法論である。

加えて、ここ 1，2 年の間に公的統計の正確性に関する議論が非常に盛んになっていることに対応し、企業活動等から自動的に生成される大規模データであるいわゆるビッグデータも利用した新しい統計指標の作成が議論されているが、その際には民間からの一定程度の粒度でのデータ提供が重要となる（例えば星野,2017)。この場合においても、公的統計作成のために必要な部分のみ欠測のない形で代入されたデータの提供を受けることが可能になれば、政府は企業に対して必ずしも個票データの提供は要求しなくても済むという場合が存在する。

このように考えると、公共財としての公的統計調査のデータを整備し補完することの重要性を考えると、代入アプローチは統計学的な観点での妥当性を有する上で実務的な有効性を持つという点で、欧米諸国同様に国内でも今後ますます利用されていくべきであると考える。

（参考文献）

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

星野崇宏(2009)『調査観察データの統計科学：因果推論・選択バイアス・データ融合』岩波書店

星野崇宏(2016)「統計的因果効果の基礎：特に傾向スコアと操作変数法を用いて」『岩波データサイエンス』3, 62-90.

星野崇宏(2017)「統計調査と指標のバイアスの理解と対処：欠測データの枠組みによる統一的理解」『統計』2017 年 1 月号, 20-26.