取れなかったデータをどうする？
―調査における欠測データの取扱いについて―

農業・農村領域　研究員　楠戸　建

1．はじめに

何かを調べようというとき(1)に、調べたい全ての対象者について、知りたい全ての調査内容が得られるのが理想的ですが、データが得られなかった場合の取扱いに困ったことはないでしょうか。例えばアンケート調査を行うときを考えると、アンケートへの協力をお願いして、「嫌です」と回答されると調査はそれまでになります。運よく「いいですよ」と言ってもらったとしても、今度は「うーん、これは答えたくないな」と答えてもらえないこともあるでしょう。このように、本来得られるはずであったにもかかわらず、得られなかったデータを「欠測データ」(Missing data)(2)と呼びます。欠測データとして取り扱われるものは、非常に多岐にわたり、関心のあるデータについて、本来得られるべき情報が一部でも得られない場合には、その得られなかったデータの全てが欠測データに該当します(3)。本稿では、このような欠測データへの対応法について御紹介します。

2．欠測のメカニズムと対応法

欠測データへの対応については、ガイドライン化が進んでおり、National Research Council（2010）やLittle et al.（2012）、国内では内閣府（2017）などで整理されています。

欠測が発生するメカニズムは、図1に示すように、完全にランダムな欠測（MCAR：Missing Completely At Random）、ランダムな欠測（MAR：Missing At Random）、ランダムでない欠測（MNAR：Missing Not At Random）の三つに大別されます（Little and Rubin, 2020）。欠測データへの対応法は、このうちのどれに該当するかによって異なります。以降では、内閣府（2017）を参考に、それぞれについて簡単に紹介します。

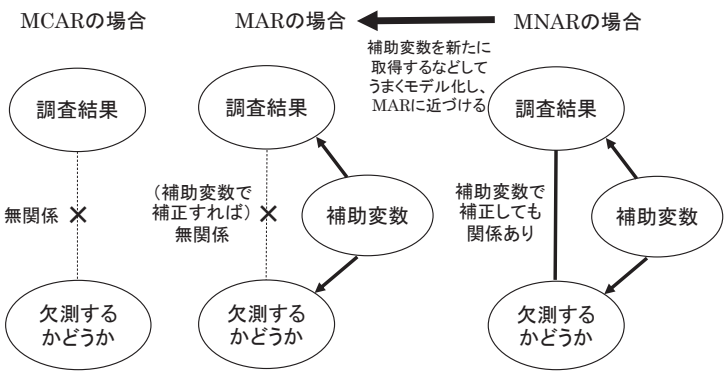

図1　欠測メカニズムごとの調査結果と欠測の関係

（1）完全にランダムな欠測の場合は無視して問題ない

欠測データメカニズムがMCARであるとは、「調査の目的変数（知りたい調査項目等）の欠測する確率が、調査の目的変数そのものの値及び他の観測されているデータの値に依存しない場合」を指します。例えば、調査対象者が硬貨を投げて、表が出れば回答し、裏なら回答しないという場合が該当します。

この場合には、使えるデータが減るという問題はありますが、観測されたデータのみを用いても調査目的の推計における偏り（欠測バイアス）は生じず、特に対処は必要になりません。

しかし、一般に回答者は非回答者と異なる特徴を持つなど、このような理想的な状況は想定しにくいと考えられます。

（2）ランダムな欠測の場合は補助変数で補正する

次に、欠測データメカニズムがMARであるとは、「調査の目的変数の欠測する確率が、調査の目的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、欠測となった調査の目的変数の値には依存しない場合」を指します。

ここでは例として、有機農産物への購入意向を把握する目的でアンケートを行うときに、購入意向に関する調査項目の一部に欠測があるときを考えましょう。調査の目的変数に該当する項目は有機農産物への購入意向、他の観測されているデータ（補助変数(4)）は、アンケートで聴取した他の項目や、調査前や調査時に付加的に取得した情報などであるとします。ここで、回答者の大半が高所得層で、非回答者の大半が中～低所得層である場合、購入意向という調査項目が欠測する確率は、調査対象者の所得という変数の値に依存しています。この場合、購入意向が観測される標本は、年収が高い層に偏ってしまいますが、所得“だけ”が欠測の有無と購入意向の両方に関連しているときには、所得という補助変数が全ての調査対象について観測されていれば、その偏りを補正することが可能です。

ランダムな欠測の場合は、このように、背景とな

る補助変数を用いて補正を行うことで対応が可能です。補正の方法としては、「傾向スコア法」や「マッチング」、「代入法」、「尤度ベースの解析」などの手法が提案されています（e.g., 星野，2009）。

（3）ランダムでない欠測の場合は、モデルを使ってランダムな欠測に近づける

最も取扱いが難しいのは、欠測データメカニズムがランダムでない欠測（MNAR）の場合で、これは「調査の目的変数の欠測する確率が、調査の目的変数自体の値に依存する場合」を指します。

有機農産物の例を再び出すと、購入意向が低い回答者がこの項目に回答しない傾向が強いという場合、得られた標本は有機農産物への購入意向が高い回答者に偏る（欠測バイアスが生じる）ことになります。先ほどのMARの場合では、所得層によって欠測がうまく補正できる場合でしたが、MNARの場合には、そもそも所得などのデータが観測できない場合や、所得だけでは欠測の有無が説明できない場合を含み、バイアスの問題を緩和するのは容易ではありません。よく「そのアンケートって結局関心のある人が答えているだけでは？」とコメントされるのは、この部分に起因するものです。

対応法としては、新たに補助変数として利用可能なデータを収集して、MARの仮定をうまく満たすようにモデル化を行うなどの対応方法が提案されています。しかし、いくらモデル化をしても、欠測しているデータそのものは観測できないため、補正がうまくいったかどうかは直接検証できないという因果推論における根本問題（Holland, 1986）と同様の問題を抱えることになります。この限界の下で最大限可能な対応法としては、欠測が生じる仕組みに関するあらゆる事態を網羅的に想定して、できるだけ多くの補助変数を取得し、それらの想定ごとに適切な分析を行った結果を比較することが推奨されています。このような分析は「感度分析」と呼ばれます。

以上の手続は、先ほどのアンケートへのコメントに対して、「では、アンケートに答えない関心のない人とは誰なのか？」と一歩進んで考えることとほぼ同じことであると言えます。

3．おわりに

欠測データは何かについて調べようとするときに常につきまとう問題です。近年の社会調査における回収率は低下傾向にあり（星野，2010；松岡・前田，2015）、欠測データの取扱いはますます重要になると予想されます。調査の回収率を上げるための努力をしても、なお発生してしまう欠測を含むデータを適切に分析するためには、調査時に背景情報を収集することが重要な点は既に述べたとおりです。SDGsにも、目標17『パートナーシップで目標を達成しよう』の中で「（前略）質が高く、タイムリーかつ信頼性のある非集計型データの入手可能性を向上させる（17. 18）」ことが掲げられ、非集計データを含めた背景情報をうまく用いて欠測メカニズムを踏まえた補正やモデル化を行い、よりバイアスのない調査結果を得ることは、意志決定の根拠としての確かさを磨き上げることに直結するものです。

欠測を含むデータの分析における手法上の発展は現在も目覚ましく進んでいます。しかし、私たちが調査するときに立ち帰らなければならないのは「何を明らかにするためにデータを取るのか」を明確にした上で、「それに影響を与える要因は何か」、「欠測したデータは調査結果にどのように影響を与えうるか」という基本であり、この基本は欠測データという視点からも重要な点であると言えるのです。

【注】

(1) このようなデータを取得する手続については、林（2017）などで解説されています。
(2) 欠測がない「完全データ」に対して、欠測が含まれるデータを「不完全データ」と呼ぶこともあります。
(3) 欠測データが具体的にどのような形で現れるかについては、星野（2009）などが参考になります。
(4) ほとんど同様の用語として、共変量（Covariate）が使われることもあります。

【文献リスト】

内閣府（2017）『欠測値補完に関する調査研究報告書【詳細版】』https://www.esri.cao.go.jp/jp/stat/report/report_all_detail.pdf（2021年11月23日参照）.

林知己夫編（2017）『社会調査ハンドブック（新装版）』朝倉書店.

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（2010）「調査不能がある場合の標本調査におけるセミパラメトリック推定と感度分析：日本人の国民性調査データへの適用」『統計数理』58（1）：3-23.

松岡亮二・前田忠彦（2015）「「日本人の国民性第13次全国調査」の欠票分析：個人・地点・調査員の特性と調査回収状況の関連」『統計数理』63（2）：229-242.

Holland, P. W.（1986）Statistics and Causal Inference. *Journal of the American Statistical Association*, 81：945-960.

Little, R. J. and Rubin, D. B.（2020）*Statistical Analysis with Missing Data*, 3nd. eds., John Wiley & Sons.

Little, R. J., D'Agostino, R., Cohen, M. L., Dickersin, K., Emerson, S. S., Farrar, J. T., Frangakis, C., Hogan, J. W., Molenberghs, G., Murphy, S. A., Neaton, J. D., Rotnitzky, A., Scharfstein, D., Shih, W. J., Siegel, J. P., and Stern, H.（2012）The prevention and treatment of missing data in clinical trials. *New England Journal of Medicine*, 367（14）：1355-1360.

National Research Council（2010）*The Prevention and Treatment of Missing Data in Clinical Trials*, National Academies Press.