統計的データ融合の諸手法と応用例について

慶應義塾大学経済学部・理化学研究所 API センター　　星野　崇宏

統計的データ融合は、異なる情報源から得られるデータ（これをマルチソースデータと呼ぶ）を、一つの情報源から得られるデータ（これをシングルソースデータと呼ぶ）に統合するための統計手法の総称であるシングルソースデータでは、分析に用いる変数の全てが同じ対象者から得られており、すべての変数間の関係性を直接把握することができる。一方でマルチソースデータは、関心のある変数が別々の対象者から分割して得られているデータである。同じ対象者からすべての変数は同時に得られていないため、通常はこれらの関係性を把握することはできない。そこで、統計的にマルチソースデータを解析して両者の関係性を把握することを考える。

統計的データ融合では、マルチソースデータにおいて、変数が得られていない部分を欠測データとしてみなして合理的に埋めることを考える。そのために、統計的データ融合では、異なるデータ間に共通する「糊しろ」としての変数を用意する。この「糊しろ」は「共変量」と呼ばれ、欠測の起きえる変数群に影響する属性変数等が用いられる。共変量の情報を有効に活用することで調査 A の対象者には変数群 B が欠測し、調査 B の対象者には変数群 A が欠測するような状況下においては本来識別できないはずの変数群 A と変数群 B の相関を一定の仮定の下で推測することが目的となる。データ融合の統計学的性質およびマーケティング分野での事例は、星野(2009)に示されているが、データ融合自体はニールセンなどの調査会社や、主にヨーロッパでの広告効果測定の実務で Data Integration という広義の複数データの活用法の一種として利用されてきた(例えば全米広告調査協会,2003)。また経済学では疑似パネルデータや data combination という名称で同様の研究が行われてきた（例えば Ridder and Mofiitt,2007)。

本発表では、一般のマルチソースデータ間の統計的データ融合だけではなく、ミクロデータとマクロデータが混在する場合でのデータ融合手法についてレビューを行い、マーケティング、医療データベース、政府統計等での応用例(例えば Igari andHoshino,2018)について紹介する予定である。

参考文献

Advertising Research Foundation. (2003), ARF guidelines for data integration, ARF, New York.

Hoshino, T. (2013). “Semiparametric Bayesian estimation for marginal parametric potential outcome modeling: Application to causal inference”. Journal of the American Statistical Association, 108(504), 1189-1204.

Igari, R. and Hoshino, T.(2018) Bayesian Data Combination Approach for Repeated Durations under Unobserved Missing Indicators. Computational Statistics & Data Analysis, 126, 150-166.

Ridder, G., and Moffitt, R. (2007) “The Econometrics of Data Combination”, Handbook of Econometrics 6B, 5470-5547.

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.