令和 5 年度　未来研究ラボシステム　研究成果報告書

研究種目：　　**研究期間：令和 4 年 10 月～令和 5 年 9 月**

研究課題名：無視できない欠測値データ解析におけるセミパラメトリック漸近有効推定量の提案―二重機械学習(double machine learning)を用いた最適な推定量の構築―

ラボ長

所属：基礎工学研究科システム創成専攻・社会システム数理領域

氏名：森川 耕輔

研究成果

(概要)

標本調査では，時間的・経済的な問題から，全数調査をではなく一部の標本を抽出した上で元の母集団の統計的性質を推測する．データの抽出は無作為に行われるわけでなく，包含確率と呼ばれる標本抽出の設計者が決定する確率値に依存して抽出される．本研究では，この包含確率を最大限利用し二重機械学習の理論を用いることで，最も効率的な推定量を提案する．また，標本に欠測値（無回答）が生じる場合においても有効な推定量を提案する．

(本文)

有限母集団(finite population)の特性を調査する際，時間的・経済的な制約のため母集団全体からデータを収集することは難しい．そのため標本調査では，コストを最小限にしつつ情報量を落とさないように母集団全体から一部のデータ(sampled data)のみを抽出することを考える．図 1 でその概要を示す．有限母集団はある無限母集団からの独立同一標本と考えられるが，最終的に得られる観測データは包含確率という標本抽出の設計者が決定する確率値に依存して抽出されるため，偏った標本となる．例えば図 1 では値が極端に小さいデータは優先的に抽出されるようにデザインされた標本である．この包含確率の情報を用いれば，得られた標本が偏っていたとしてもそのバイアスを補正し興味のある無限母集団のパラメータを推定することが可能となる．古典的な方法では，Horvitz and Thompson (1952)による重み付き推定量がある．しかし，この推定量は情報の一部分を無駄にしており，有効推定量ではない．

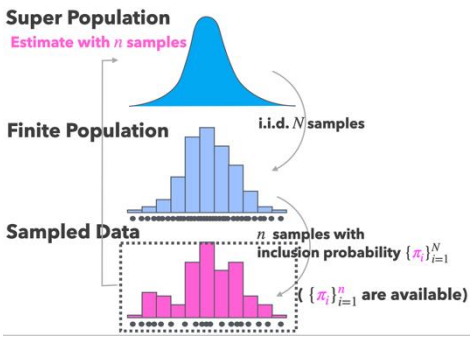

図 2. 標本調査の概要.

Morikawa et al. (2022)では，セミパラメトリック推定の理論を用いることで，標本調査における種々の問題に対する効率的な推定量を２つ提案した．１つはパラメトリック作業用モデルを用いる方法で，もう１つは二重機械学習によるノンパラメトリック作業用モデルを用いる方法である．前者の方法は，用意した作業用モデルが正しい場合はセミパラメトリック漸近下限に達する推定量であるが，後者は常に下限に達する推定量である．図 2 では推定対

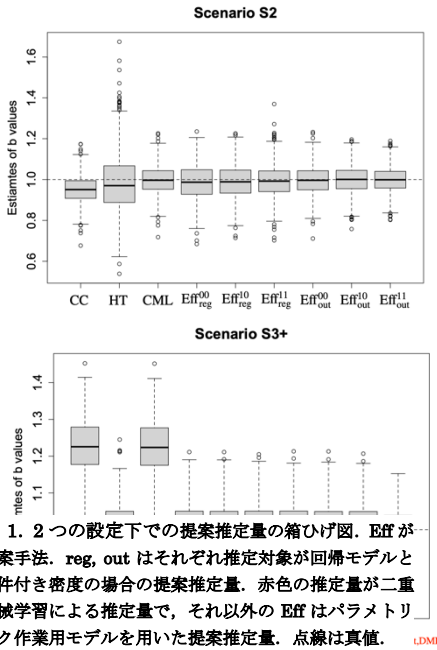

図 1. 2 つの設定下での提案推定量の箱ひげ図．Eff が提案手法．reg, out はそれぞれ推定対象が回帰モデルと条件付き密度の場合の提案推定量．赤色の推定量が二重機械学習による推定量で，それ以外の Eff はパラメトリック作業用モデルを用いた提案推定量．点線は真値．

象を回帰モデルと条件付き密度関数とした場合の回帰モデルの傾きの推定量の箱ひげ図である．上図の設定では，既存手法(CC, HT, CML)より提案手法(Eff)の推定効率が圧倒的に良いことが示されている．また，下図の設定では推定手法に二重機械学習による推定方法も示している．作業用モデルの特定が難しいような状況でも，提案手法による二重機械学習を用いることで効率的な推定量を構成可能となる．

次に Morikawa et al.(2023)では，標本に欠測値がある状況への拡張を行った．当初の想定通りデータが得られない欠測や無回答の問題は，標本調査のみならず多くの分野で避けられない重要な問題である．Morikawa et al. (2022)を拡張することで，標本の抽出（１段階目），その後無回答（2 段階目）といった 2 段階の欠測が生じたとしても効率的な統計解析を行える手法を構築した．さらに，国勢調査といった外部情報がある場合において，現在の標本と外部情報を効率的に組み合わせることが可能な推定量を提案した．

研究経費（R5 年度）の内訳

備品費	消耗品費	旅費	謝金	その他	合計
0 円	208,874 円	231,126 円	0 円	0 円	440,000 円

共同研究者等

(1)共同研究者（氏名・所属）

Jae Kwang Kim・アイオワ州立大学統計学部

寺田吉壱・システム創成専攻数理科学領域

(2)研究協力者（氏名・所属・学年（学生の場合））

別府健治・システム創成専攻社会システム数理領域・博士後期課程２年

相田　航・システム創成専攻社会システム数理領域・博士前期課程２年

発表論文等（令和 6 年 3 月 31 日現在）

〔雑誌論文〕

Morikawa, K. and Kim, J. K. (2022). Semiparametric adaptive estimation under informative sampling, submitted to *Annals of Statistics*, first revision invited. arXiv:2208.06039.

Morikawa, K., Beppu, K. and Aida, W. (2023). Efficient multiple-robust estimation for nonresponse data under informative sampling, submitted to *Biometrika*. arXiv:2311.06719.

外部資金獲得状況・申請状況

令和 5(2023)年度 基盤研究(A)「情報統計が拓く地震モデリング数理基盤」. 研究分担者.

令和 6(2024)年度 若手研究「特異な傾向スコアを用いた統計的推測：傾向スコアの境界値問題への対処」. 研究代表者.

参考となる HP 等

https://sites.google.com/site/kosukemorikawa/