

## アンケートデータにおける選択バイアス補正法に関する一考察

西尾和恭<sup>†</sup> 岡野拓巳<sup>†</sup> 芝千尋<sup>†</sup> 戸次陸<sup>†</sup> 須子統太<sup>†</sup>早稲田大学社会科学部<sup>†</sup>

## 1. はじめに

近年様々な場面で紙媒体やWebによるアンケート調査が行われている. 一般的に、アンケート調査では完全無作為抽出を行うことは困難なことが多く、選択バイアスによって集計結果の信頼性が低くなる場合が多い. そのため、従来より選択バイアスの補正手法がいくつか提案されている. しかし、どの手法がどのような場面に適しているかについては十分に示されていない. そこで、本稿では実際のアンケートデータを用いた実験により、各手法の性質を明らかにすることを目的とする.

## 2. 選択バイアス

選択バイアスとは、「本来調査の対象とする集団」から一部の対象者が選択されている、あるいは脱落している状況で、単純な解析を行うことで生じる結果の歪みのことを呼ぶ[3]. アンケート調査を例に挙げると、本来の調査対象として日本国民全体を考えた場合に、アンケート回答者の属性の分布が日本国民全体の分布と異なっている状況で、回答者のデータをそのまま解析することで生じる結果の歪みが選択バイアスである.

このような選択バイアスが生じると解析結果の信頼性が低くなる. そのため、従来選択バイアスの補正手法として、傾向スコア法や回帰モデルを利用した方法が提案されている.

## 3. 選択バイアス補正手法

代表的な選択バイアス補正手法について示す.

いま、 $X_i = (x_1, x_2, \dots, x_d) \in \mathcal{X}^d$  ( $i=1, 2, \dots, N$ ) を  $d$  次元の共変量ベクトルとし  $x_k$  ( $k=1, 2, \dots, d$ ) はアンケート回答者の基本属性を表すこととする. また、アンケート回答の有無を  $Z_i \in \{0, 1\}$  (回答有=1) とし、 $m$  択のアンケートの回答を  $y_i \in \{1, 2, \dots, m\}$  とする.

傾向スコア法・回帰モデル法どちらの手法においても、共変量ベクトルは本来の調査対象・アンケート回答集団ともに得られているものとする.

## 3.1 傾向スコア法

傾向スコア法による選択バイアスの補正法を示す. まず第  $i$  対象者がアンケートに答える確率  $e_i$  ( $0 \leq e_i \leq 1$ ) を推定する.

$$e_i = \hat{p}(Z_i = 1 | X_i) \quad (1)$$

これを第  $i$  対象者の傾向スコアという[1]. 本実験では  $e_i$  の推定にロジスティック回帰を用いる.

次に傾向スコアで重み付けした IPW 推定量[2]

$$U = \frac{\sum_{i=1}^N \frac{Z_i(1-e_i)y_i}{e_i}}{\sum_{i=1}^N \frac{(1-Z_i)e_i}{1-e_i}} \quad (2)$$

を推定する. これを本来の調査対象のアンケート回答の割合に関する推定量とする.

## 3.2 回帰モデル法

回帰モデル法による選択バイアスの補正法を示す. まずアンケート回答者のデータのみを用いて、第  $i$  対象者が各選択肢に回答する確率  $p_i(y_i | X_i)$  を推定するモデルを学習する. 本実験では条件付き確率の推定に、決定木・SVM・ロジスティック回帰を用いた.

次に全てのデータに対し、第  $i$  対象者が各選択肢に答える確率を推定し、各  $i$  に対して平均をとった、

$$\hat{p}(y) = \frac{1}{N} \sum_{i=1}^N \hat{p}(y_i | X_i) \quad (3)$$

を計算し、これを本来の調査対象のアンケート回答の割合に関する推定量とする.

## 4. 実験概要

本実験では、実際のアンケートデータに対し、擬似的に選択バイアスを発生させることで、各手法の補正の精度を調べる.

使用するデータは、株式会社アサツーディ・ケイ生活総合調査 2017、消費者の基本属性に関する質問と、消費意識に関する 4 択アンケートのデー

A Study on Selection Bias Correction Methods for  
Questionnaire Data  
Kazutaka Nishio, Takumi Okano, Chihiro Shiba, Riku  
Hetsugi, Tota Suko  
Waseda University School of Social Sciences

タ(35項目)を用いる。データ総数は15042個である。バイアスの生成は、元のアンケートデータ全体を本来の調査対象とし、そこから特定の属性のサンプルのみを抽出し、これを選択バイアスを含むアンケート回答とみなすこととする。バイアスの測り方として、元のデータで4択アンケートの各選択肢に答えた人の割合の分布と、特定のサンプルで各選択肢に答えた人の割合の分布のKLダイバージェンスを用いる。また、補正精度も元のデータの分布と各手法の分布とのKLダイバージェンスで測ることとする。

## 5. 実験結果

属性が「学生」のサンプルをバイアスを含むアンケート回答集団として分析した結果を示す。共変量は都道府県・性別・同居家族人数を用いた。

まず全体と回答集団の共変量の違いを示す。(表1) 表1から同居人数において違いがあることがわかる。

表1: 全体と回答集団の共変量の違い

| 共変量   | 全体    | 回答集団 |
|-------|-------|------|
| 合計人数  | 15042 | 1582 |
| 性別    |       |      |
| 男性    | 7459  | 817  |
| 女性    | 7583  | 765  |
| 地域別人数 |       |      |
| 関東    | 11467 | 1192 |
| 近畿    | 3575  | 390  |
| 同居人数  |       |      |
| 1     | 2344  | 239  |
| 2     | 3422  | 77   |
| 3     | 4044  | 382  |
| 4     | 3720  | 588  |
| 5人以上  | 1512  | 296  |

次に全体と回答者の各選択肢の回答割合を全35項目のうち3項目を抜粋して示す。(表2)

表2: 全体と回答集団の解答割合

|    |      | 1     | 2     | 3     | 4     |
|----|------|-------|-------|-------|-------|
| Q1 | 全体   | 0.120 | 0.436 | 0.306 | 0.147 |
|    | 回答集団 | 0.121 | 0.427 | 0.328 | 0.122 |
| Q2 | 全体   | 0.156 | 0.619 | 0.192 | 0.032 |
|    | 回答集団 | 0.154 | 0.609 | 0.200 | 0.034 |
| Q3 | 全体   | 0.088 | 0.501 | 0.353 | 0.056 |
|    | 回答集団 | 0.087 | 0.472 | 0.386 | 0.053 |

ここでQ1は全ての質問項目、全ての変数の組み合わせで補正精度が最も高かった項目、Q2は全ての質問項目において補正精度が最も低かった項目を抜粋した。

次に抜粋した質問項目に対しての各手法の補正精度を示す。(図1)

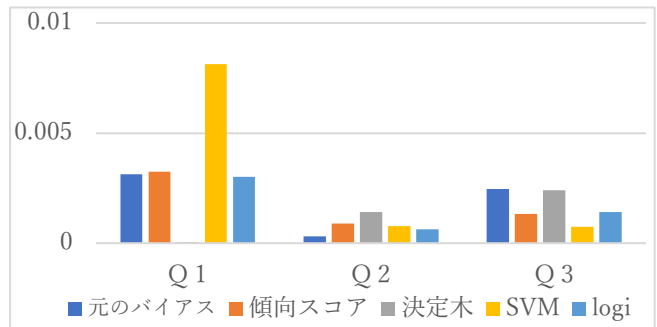


図1: 元のパイアスと各手法のKL距離

Q1・Q2に対しては都道府県と同居人数、Q3に対しては全ての共変量を用いた。

図1から傾向スコア法と回帰モデル法で補正精度に大きな違いはないことがわかる。また、表2、図1から補正精度は全体と回答集団の回答割合の違いに依存しないことが分かる。

次にアンケート35項目に対して全ての変数の組み合わせのなかで最も精度が高かった手法の数を示す。(表3)

表3: 各手法の割合

| 手法 | 補正無 | 傾向 | 決定木 | SVM | logit |
|----|-----|----|-----|-----|-------|
|    | 1   | 10 | 11  | 10  | 3     |

表3から何かしらの補正手法を用いることでバイアスが減少することがわかる。しかし、設問により最も補正できる手法が異なっており、各手法の優劣がつけられないことが分かった。

## 6. まとめ

本研究では、実際のアンケートデータに対し擬似的な選択バイアスを発生させることで、様々な選択バイアスの補正手法の精度を評価する実験を行った。実験の結果より、同一の対象者に関するアンケートでも、アンケートの設問によって選択バイアスの補正精度が最も高い手法が異なっている事がわかった。実際のアンケート調査で選択バイアスが含まれてしまう状況では、本実験のようにバイアスの含まれない正しいデータは得られず、どの手法がバイアスを最も小さくするのかは分からない。そのため、得られたデータのみから、選択バイアス補正手法を選ぶ手法が必要となるが、具体的な手法については今後の課題としたい。

## 参考文献

- [1] Rosenbaum, P. R. and Rubin, D. B. “The Central Role of the Propensity Score in Observational Studies for Causal Effects”, *Biometrika*, Vol. 70, pp. 41-55. (1983)
- [2] Rosenbaum, P. R. and Rubin, D. B. “The Bias Due to Incomplete Matchings”, *Biometrics*, Vol. 41, pp. 103-116. (1985)
- [3] 星野崇宏 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(2009)