

アンケートデータにおける選択バイアスの補正手法の選択方法についての一考察

沢田拓也[†] 西尾和恭[†] 石澤由宇輔[†] 須子統太[†]

早稲田大学社会科学部[†]

1. はじめに

アンケート調査は様々な場面で行われている。これは完全無作為抽出を前提としているものの、一般的にその実現は難しく、選択バイアスによって集計結果の信頼性が低下する 경우가多々ある。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が、どのようなアンケートデータにも対応できる万能な手法はなく、対象となるデータによって有効な補正手法が異なることが知られている[1]。

そこで本研究では、ある調査データに対して補正精度の高い補正手法を選択する方法について検討する。

2. 選択バイアスとその補正手法

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態で、単純な解析を行うことで生じる分析結果の歪みのことを指す[2]。

回答分布の推定において、選択バイアスの補正手法として代表的な回帰モデル法による補正方法について示す。

いま、 $X_i = (x_{i1}, x_{i2}, \dots, x_{id})$, ($i = 1, 2, \dots, N$) を d 次元の共変量ベクトルとする。ここで、 $x_{ij} \in X_j$ とする。共変量は例えばアンケート回答者の基本属性などを用いることができる。また、 m 択のアンケートの回答を $y_i \in \{1, 2, \dots, m\}$ とする。

回帰モデル法では、得られたアンケートデータから、共変量 X を説明変数、アンケート回答 y を目的変数ととらえ、 $p(y | X)$ を推定する。 $p(y | X)$ の推定には、決定木やロジスティック回帰など、様々な学習モデルを利用することができる。ある学習モデルを利用し推定した条件付き確率を $\hat{p}(y | X)$ とすると、回帰モデル法では次式を用いて、回答分布の推定を行う。

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

従来研究により判明していることは、 $\hat{p}(y_i | X_i)$ を推定する際に用いる学習モデルが異なると、回答分布の補正精度が変わるという事実である。また、設問によって最適な学習モデルは異なることが知られている[1]。

3. 選択バイアスの補正手法の選択アルゴリズム

本研究では、特定の設問について推定精度の高い学習モデルを選択する方法を提案する。回帰モデル法では、 $\hat{p}(y | X)$ の推定精度が高ければ高いほど、選択バイアスの補正精度が高くなると考えられる。そこで、複数の学習モデルの候補から、 $\hat{p}(y | X)$ の推定精度の高い学習モデルを選択する方法を考える。しかし、真の $p(y | X)$ 値は未知であるため、一般的に、 $\hat{p}(y | X)$ の推定精度を測ることはできない。そのため、与えられたアンケートデータを用いて各学習モデルに対し、クロスバリデーションを行うことで、近似的に推定精度の高い学習モデルを選択する手法を提案する。

4. 評価実験

4.1. 選択バイアスを含むデータの生成方法

実際のアンケートデータを使用して、提案手法の評価を行う。もとのデータには選択バイアスは含まれていないと仮定したうえで、意図的にサンプリングしたデータで、人工的に選択バイアスを発生させた。本実験では回答者が専業主婦かどうかという属性に着目し、本来の母集団から専業主婦である回答者のみをサンプリングすることで、疑似的に選択バイアスを含むデータを生成した。

使用するデータは、株式会社 ADK マーケティング・ソリューションズ生活総合調査 2017 の中の、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タである。4 択アンケートは全てで 35 項目あり、データ総数は 15042 個、選択バイアスを発生させた母集団のデータ総数

は 2163 個である。

4. 2. 実験方法

本研究では条件付き確率を表現する学習モデルとして、決定木、SVM、ロジスティック回帰の3つを用いる。各学習モデルに対し、選択バイアスを含むデータを用いてクロスバリデーションを行う。具体的なクロスバリデーションの方法は様々考えられるが、以下3種類の方法を行い、比較した。

- (A) データの10分割クロスバリデーション
- (B) データの100分割クロスバリデーション
- (C) 各選択肢のサンプル数を揃えたうえでのクロスバリデーション

なお、これらのクロスバリデーションでは共変量として、アンケート回答者の属性を問う質問項目の中の「年齢」「都道府県」「最終学歴」「年収_目安」を採用した。選択バイアスを発生させるために用いた「専業主婦」は使用していない。

(A)では、人工的に選択バイアスを発生させた2163個のデータを10分割し、9割を学習データ、1割をテストデータとする。学習データをもとに導いたモデルで、テストデータを予測する。この作業を10回繰り返す。(B)についても同様の手順である。ただしこれらのクロスバリデーションには回答分布が偏っている際の、予測モデルの学習不足という問題点があった。例えばある質問項目の回答の多くが選択肢1と選択肢2に偏っている場合、モデルが選択肢3と選択肢4の回答を一切予測しない場合がある。そこで、学習データの回答に偏りが出ないようにリサンプリングしたうえでクロスバリデーションを行ったのが(C)である。具体的な実験の流れは次の通りである。

[(C)の手順]

- i 質問項目毎に、回答数が最小の選択肢を見つける。その回答数をa個とする。
- ii 他3つの選択肢の回答群から、それぞれa個だけランダムサンプリングする。
- iii 総回答数が4*a個となった状態で4*a分割クロスバリデーションを行う。

(A), (B), (C)それぞれ、各質問項目について3つの手法でクロスバリデーションを行ったうえで、誤り率の平均値を計算し、最も誤り率の平均値が低い手法を調べる。

一方で各質問項目に関して、もとのデータで4択アンケートの各選択肢に答えた人の割合の分布と、専業主婦の回答者のみのサンプルで各選択

肢に答えた人の割合の分布を、KLダイバージェンスを用いて測る。3手法それぞれのKLダイバージェンスを算出し、その中で最も値が小さい手法を調べる。

特定の質問項目について、最も誤り率の平均値が小さい手法と、最もKLダイバージェンスの値が小さい手法が一致していれば、望ましい選択手法であるといえる。

4. 3. 実験結果

実験結果を以下の表1に示す。なお、補正精度の評価に関して、KLダイバージェンスの値が小さいほど補正精度が高いと判断している。

(A), (B)に関してはともに設問の一致数が少ないといえる。(C)では、(A), (B)に比べて一致数が増え、条件付き確率の予測精度と補正精度の関係性は向上した。

表1 予測精度の最も高い学習モデルと、補正精度の最も高い学習モデルの一致数

	(A)	(B)	(C)
一致した設問数 (全35項目)	6	5	20

5. 考察

(A)と(B)を比べてみると、設問の一致数にはほぼ差はない。クロスバリデーションを行う際のデータ分割数は、学習モデルの予測精度には影響がないと考えられる。

また(C)での一致率が最も良かった要因は、クロスバリデーションを行う際のリサンプリングが学習モデルの推定精度を向上させた点にあると考えられる。

6. まとめ

本研究では、選択バイアスの補正手法の選択アルゴリズムを提案した。

今後の課題は、回帰モデル法以外の傾向スコア法などの他の補正手法も含めた、選択アルゴリズムの提案が挙げられる。

参考文献

- [1] 西尾和恭, 岡野拓巳, 芝千尋, 戸次陸, 須子統太 “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”, 情報処理学会第81回全国大会 講演文集, vol. 4, p. 687 (2019)
- [2] 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)