研究レポート　No.1　～傾向スコアによる調査データ補正～

2011 年 4 月 12 日　株式会社アイズファクトリー　http://bodais.jp/

概要

近年、インターネットの普及により低コストでの大規模調査が実施できる反面、こうしたネット調査データは、代表性が担保されずデータ分布の偏りを含んでいるとの問題点が指摘されている。偏りのあるネット調査データを従来の無作為抽出調査の結果に近づける有力な方法として、傾向スコア（propensity score）が注目を集めている。本レポートでは、この技術に関する解説と具体的な適用事例を文献［2］を引用して紹介する。

1. はじめに

近年のインターネットの普及により、低コストのネット調査によるデータ取得が進んでいる。一方、2005 年の個人情報保護法の施行により、住民基本台帳からの無作為抽出による調査が困難な現状にある。インターネット調査は、代表性が担保されないなどデータの偏りが問題となるが、現状では、インターネット調査に頼らざるを得ない。近年、偏りのある調査データを従来の無作為抽出による調査結果に近づける有力な方法として、傾向スコア（propensity score）が注目を集めている。以下、本手法について文献［2］を引用して解説する。

2. 傾向スコア調整法によるデータ補正

傾向スコアとは、Rosenbaum と Rubin が 1983 年に発表した概念である［1］。具体的には、傾向スコアは、補正のための共変量を用いて計算され、データ分布の異なる 2 つのデータ群が有る場合に、任意のサンプルが「一方のデータ群に割り当てられる確率」を表わす。インターネット調査の場合は、あるサンプルが「インターネット調査に回答する確率」となる。

傾向スコアによる調査データの補正は以下の手順に従って行われる。

(1) インターネット調査と、無作為抽出による既存調査との比較から、補正のために最適な共変量を探索。
(2) 発見された共変量を用いて傾向スコアを計算。
(3) この逆数による重み付平均よりネット調査の設問の回答（%）の補正値を算出する。

この手順をまとめたのが図 1 である。訪問調査とインターネット調査の間にある、複数の共変量によるデータ分布の違いを傾向スコアにより一元化できるのが特長である。

補正のための共変量の探索は、文献［3］で提案された「共変量選択法」を用いる。

《共変量選択法》

・個人内で変動が少なく、かつインターネット調査と既存調査（訪問調査）で継続的に質問できる可能性が大きい項目を選ぶ
・既存調査（訪問調査）とインターネット調査間で差のある項目を選ぶ
・補正したい項目を共変量に回帰させた時の偏回帰係数が 2 群とも同じ方向になるものを選ぶ、特に標準偏回帰係数の絶対値の大きいものを選ぶ
・上記の基準で選択された共変量のセットから、さらに二乗誤差の和を減少させるように共変量を減らす

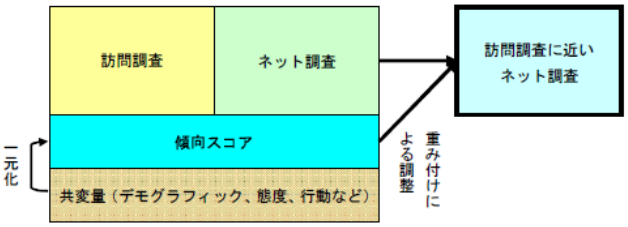

図 1. 傾向スコアによる補正

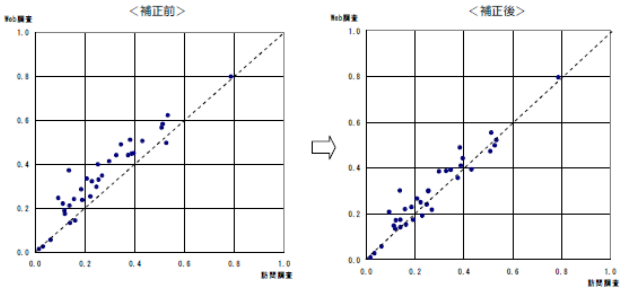

図 2. インターネット調査の補正前後での分布比較

実際の補正例を見てみよう。文献［2］の調査例では、「金融商品の保有」（12 項目）、「よく読む新聞の記事」（34 項目）の計 46 項目（全て 2 値変数）の補正を行った。共変量には、「年齢」「性別」「職業」「最終学歴」「世帯収入」などのデモグラフィック変数と、「読書頻度」「ネット利用時間」「旅行」などライフスタイルに関連する項目も含めた約 20 項目を使った。図 2 は、補正対象の変数についてカテゴリーごとの比率を、訪問調査での値を横軸、インターネット調査での値を縦軸にとってプロットした結果である。補正前の図では、インターネット調査の回答と訪問調査の回答が大きく異なり、データ点が対角線より上に分布している。補正後の図では、傾向スコアによる補正によって誤差が縮小し、データ点が対角線上に分布しており、補正が良好にいっているのが分かる。

3. まとめ

本レポートでは、偏りのあるデータを無作為抽出データに近づける有力な手法である傾向スコアについて、文献［2］を引用して紹介した。本レポートの事例の他にも、ダイレクトメールの顧客リストについて顧客ごとの予測入会率を算出する場合がある。この場合、ある年度の顧客リストとその入会実績とから予測モデルを構築して次年度の顧客リストに適用する。その際に、モデル構築用の当年度データと予測用の次年度データのデータ分布の違いを補正する手法として傾向スコアが利用できる。この他にも文献［2］では、2000 年のアメリカ大統領選挙（ブッシュ vs ゴア）の予測における成功事例が紹介されている。

4. 参考文献

[1] Rosenbaum & Rubin（1983）：Biometrika, 70 , 41－55.
[2] 星野・楠木・松本・森本（2008）：S-Plus ユーザ抄録原稿
[3] 星野・前田（2006）：統計数理、第 54 巻 1 号、191－206.