

統計マイクロデータによる探索的研究

統計数理研究所

樺 広計 所長



研究背景・マイクロデータ活用の経緯

日本では統計学やデータ解析を専門とした人材が昔から少なかったため、統計学の専門家として、様々な分野の研究者と共同研究を行い、データ解析の支援を行ってきました。これまでも様々なデータに触れてきましたが、その中でも公的統計マイクロデータは標本が全国から無作為にサンプリングされていること・回収手続きが厳格で信頼性が高く、サンプルサイズも非常に大きいため、有用性の高いデータとして研究に利用しています。

下記に示す自殺要因に関する研究においてもマイクロデータを活用しました。本研究は公衆衛生・精神衛生を専門とされている先生方が中心で取り組まれていた厚生労働科学研究への参加がきっかけで着手することとなりました。厚生労働省国民生活基礎調査ではK6（気分障害・不安の尺度）を測定しているため、厚労科研チームで竹林由武先生（福島県立医大）らを中心に探索的データ分析を実施しました。その結果、興味深い発見が可能となると分かったため、本格的に研究を進めることとなりました。

研究概要

活用したデータと手法等

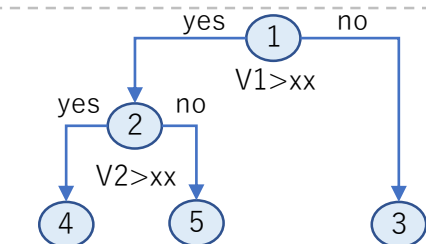
まず、国民生活基礎調査のマイクロデータを対象に、1980年代後半に開発された第2世代人工知能（分類樹）による分析を竹林由武先生が中心に行い、介護者など自殺ハイレスク群で自殺リスクを高めている要因を抽出しました。

次に、岡檀先生（統計数理研究所）が社会生活基本調査のマイクロデータと教育用標準データセット（総務省統計局：都道府県・市区町村のすがた（社会・人口統計体系）の市町村別データから作成）を用いて、個人要因と1,741自治体レベルで集計した環境要因を合わせて分類樹分析をおこない、（自殺リスクを高める要因である）介護に携わる時間を増加させる要因を分類樹形式で抽出しました。

研究結果

国民生活基礎調査マイクロデータを対象とした分析では、精神疾患者・介護者・無職者を自殺ハイレスク群としました。いずれの群においても、最も自殺リスクを高めている要因（予測要因）は主観的なストレス状態の自覚であることが分かりました。その他にも世帯種類、貯蓄高、総所得、職業分類等が自殺リスクに大きな影響を与えることを明らかとしました。また、精神疾患のリスクを高める大きな要因として脳卒中既往や低い貯蓄高等が大きいことを確認しました。

次に社会生活基本調査マイクロデータを用いた分析により、個人要因よりも環境要因、市区町村の特性よりも都道府県の特徴が、介護に携わる時間に強い影響を与えている可能性が示唆されました。



✓ 再起分割分析により、分類樹形式で予測要因を抽出

マイクロデータ活用の有用性など

上記のように、公的統計マイクロデータは、対象が広範で解析可能な変数の種類もサンプルサイズも非常に大きいデータです。一定のサンプルサイズを必要とする人工知能など探索的分析も可能な有用性の高いデータなのです。社会生活基本調査、国民生活基礎調査などのマイクロデータを用いないとできない地域研究は他にも多々存在すると考えています。一方で、自治体別特徴の集計分析を行うと、小さな自治体ではサンプルサイズが小さくなり現状では公開不可能となるケースもありました。より地域政策分析に活用しやすい標本設計についても検討が進むことを期待しております。

経済学等の実証研究では既に公的統計マイクロデータの認知度が高い一方、環境、医療等認知度を高める余力のある分野も数多くあります。公的統計マイクロデータの認知度向上、データサイエンスの専門家育成推進、オンサイト施設で利用できる統計マイクロデータの種類・量の増加等に伴い、様々な分野で統計マイクロデータの利活用が進んでいくのではないのでしょうか。

【利活用事例】

①Takebayashi, Kubota and Tsubaki (2016) 『Risk profiles for severe mental health problem: classification and regression tree analysis』 Proc. COMPSTAT 2016, pp.291-302.

②岡、岡本、久保田、竹林、谷道、樺（2019）『社会生活基本調査マイクロデータを利用した介護高負担要因の探索的解析』 「官民オープンデータ利活用の動向及び人材育成の取組」研究会