

人間総合研究センター主催 「人間科学研究交流会—Current Topics in Human Sciences—」記録

第
37
回

話題提供者：人間科学学術院 講師 中西 啓喜
演 題：学力調査から見る日本の教育社会調査の現状と課題
—教育における“ビッグデータ”とは何か？
開催日時：2018年10月10日, 18:00～19:00
開催場所：100号館第1会議室

1. 社会調査データの諸課題

「現代はビッグデータの時代である」というとチープな言い回しになるが、社会科学においても確かにその趨勢は見られる。社会学ないし教育社会学は、特定のコミュニティを対象にデータ収集した研究である「計量的モノグラフ」を、歴史的には許容どころか重視さえてきた。しかしそれでも、現在では可能な限り“ビッグ”なデータによる研究を目指す努力は確かに見られる。とりわけ、教育は「発達のプロセス」であるため、本来ならば追跡的研究が望ましいのだが、様々な困難さゆえにその知見は限定的なものに留まっている。本交流会では、自身の研究成果の報告を踏まえて、教育社会学におけるデータ科学の現状と展望について検討した。

柴田里程 (2015) による“ビッグデータ”の定義ないし特徴を確認すると、まず「規模の大きな」データという側面が指摘される。加えて、例えば、衛星から連続的に送られてくる各種センサの示す値や画像・レーダーデータ、ゲノムデータ、原子炉などの複雑な実験データインターネットを流れるパケットデータ、電力のスマートグリッドデータなどのように、「扱いの大変な」あるいは「重要な」という意味・ニュアンスを含むという (柴田 2015, p. 3)。

それでは、社会学ないし教育社会学における“ビッグデータ”は存在するか。結論から述べるのであれば、先の柴田の定義にマッチするような“ビッグデータ”はほぼ存在しない。少なくとも、日本社会学会ジャーナル『社会学評論』、日本教育社会学会ジャーナル『教育社会学研究』で“ビッグデータ”にキーワードにヒットする論文は皆無である (2019年1月10日現在)。

社会科学分野の量的データのクオリティは多くの場合、①ランダム・サンプリング (無作為抽出) と②サンプルサイズの2点によって特徴づけられる。つまり、社会科学における「豊かなデータ」とは、“サーベイデータ”と呼ばれるものが該当する。それゆえに、データ収集はディシプリンの中で高度に発達されてきた調査法に基づいて行われる

ため、パケットデータやスマートグリッドデータのような複雑さはデータ収集の段階で回避されていることになる。ところが、当然、社会学・教育社会学においてもデータに纏わる課題はいくつもある。

第一に、日本の社会調査データは分析ケース数が少ない。社会階層や不平等、社会移動、職業、教育、社会意識などを把握するために1955年以来10年に一度行われている日本の代表的な社会調査である「社会階層と社会移動全国調査 (SSM調査: Social Stratification and social Mobility)」は、毎回のサンプルサイズは約4,000である。しかし、2015年に実施されたSSM調査の有効票は7,817であり、サンプルサイズを大きくするべきという趨勢が読み取れる。また、近年のSSMデータ分析の流行のひとつに、過去のSSMデータをマージしたプーリングデータを用いる研究者も散見される。例えば、古田和久 (2018) は、1965年から2015年データの男性のみのデータをマージした分析を行っているが、それでも分析ケース数は10,613に留まる (女性が調査対象となったのが1985年調査からだったため)。このようにサンプルサイズの小ささは日本の社会調査データの課題のひとつである。

第二の課題は、分析手法の高度化にともないデータが複雑化していることである。人々はあるカテゴリーのまとまりの中で生活している。例えば、我々は国、県、市、学校などに属しながら生活しており、調査もこうしたカテゴリーごとにサンプリングされることが多い。こうしたデータをClustered Dataと呼ぶが、Clustered Dataは分析サンプルが独立しておらず、その結果、標準誤差を小さく見積もってしまい正しい効果の有無を推定できない可能性が生じるのである。そこで、それに対応する手法が必要となる。

もうひとつのデータの複雑化は、同じ子どもの身長を追跡的に測定したといったようなLongitudinal Dataである。Longitudinal Dataは、特に、特定の方法で収集されたデータを特定のメソッドで分析することを前提にするこ

「人間科学研究交流会」報告

とも多く、こうしたデータ収集と分析のパターンは徐々に増加傾向だとされている。

しかし、その一方で日本ではLongitudinal Dataによる知見は限られている。理由は、①個人を追跡的に調査するために発生する個人情報取得などの研究倫理の側面、②研究費の捻出の困難さに関する側面、③長期調査に耐えるプロジェクトの維持という人的資源に関する側面などが指摘できる。先にも述べた通り、教育は「発達のプロセス」であるため、本来ならば追跡的研究が望ましいのだが、教育に関する追跡的研究はそれほど多くはない。

2. Longitudinal Dataを分析する意義

Longitudinal Dataを用いた分析のひとつの長所として、偏りの小さい推定値を得やすいという点がある。本来、従属変数と独立変数との関係をバイアスなく推定するためには、独立変数以外に従属変数と関連がある要素はすべて統制しなければならない (Allison 2009、中澤 2012)。

例えば、たくさん勉強する児童生徒ほど学力が高いかどうかについて回帰分析によって推定値を得ようとする。その際に、教育社会学では家庭環境、性別などの生得的要因を回帰式に投入し、可能な限り「正しい」推定値を得ようとする。

しかし、通常の回帰分析では、投入する独立変数以外の要因はすべて「誤差項」として一括りに扱うため、分析結果は、生得的な能力や性格といった「観測不可能な個体特有の効果」を除去していない推定値が得られたにすぎない。そこで、「観測不可能な個体特有の効果」を取り除くために採用されるベーシックな手法は、計量経済モデルにおける固定効果モデル (Fixed Effect Model) である。これにより、誤差項を「観測不可能な個体特有の効果」と「その他の誤差項」に分割しつつ、「観測不可能な個体特有の効果」を取り除いた推定が可能となる (図1)。

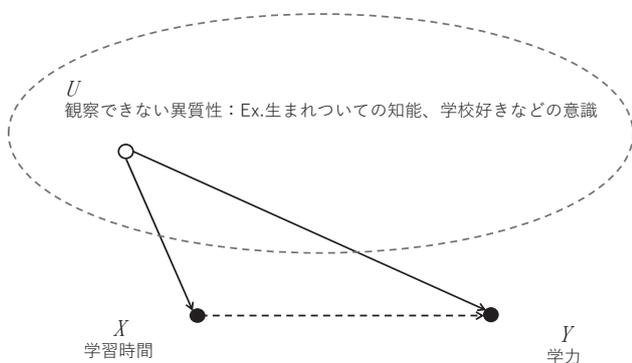


図1. 固定効果モデルの利点のイメージ

3. 議論: 社会科学・教育分野におけるビッグデータの利用と課題

こうした教育社会学の趨勢に最も俊敏に対応している組織のひとつがベネッセコーポレーションだといえる。Longitudinal Dataについては「子どもの生活と学びに関する親子調査」を実施し、“ビッグデータ”についてはベネッセ教育総合研究所「ビッグデータを活用した教育研究の取り組み」(https://berd.benesse.jp/special/bigdata/about_pj.php 2018年10月9日取得)を公表している。

問題は、教育データの分析結果の還元先はどこであるべきか、という点である。ベネッセという教育産業が展開する研究であるがゆえに、当然ベネッセを利用する学習者への還元は想定しているだろう。むしろそれは企業としては批判されるようなことではない。しかし、“サーベイデータ”として収集されたデータであるならば、筆者としては可能な限り社会に対する貢献を期待したい。これはベネッセの一連の教育調査のみに留まる問題ではない。例えば、川口俊明 (2018) は、文部科学省による全国学力・学習状況調査の目的が、「教育政策に活かす」と「児童生徒の指導に活かす」ことが混濁してしまい、「教育政策に活かさない」データが蓄積されていることを危惧している。こうしたデータの還元先についても今後検討課題となるだろう。

〈文献〉

Allison, Paul, D., 2009, *Fixed Effects Regression Models*, SAGE Publications, Inc.
 古田和久、2018、「高学歴化と職業構成の変容に関する基礎分析」古田和久編『2015年SSM調査報告書4 教育I』、pp.1-20。
 川口俊明、2018、「PISAから私たちは何を学べるのか?—「学力調査の設計」という視点から」『シノドス』https://synodos.jp/education/21994、2019年1月10日取得。
 中澤渉、2012、「なぜパネル・データを分析するのが必要なのか—パネル・データ分析の特性の紹介」『理論と方法』27(1)、pp.23-40。
 柴田里程、2015、『データ分析とデータサイエンス』近代科学社。