修士論文概要

横浜市立大学大学院データサイエンス研究科　根本　尭

国勢調査における学歴分布では「不詳」が、2010 年は 13.1%、2020 年には 14.9%を占めている。更に、国勢調査における学歴不詳率を地域別にみると、2020 年において東京都の学歴不詳率は 26.9%にものぼる。国や地方公共団体は、学歴分布を正確に把握できないため、不正確な統計結果を基に政策を立案することになり、政策の効果的な施行が困難になる可能性がある。そのため、国勢調査における学歴の「不詳」データを補完し、正確な学歴分布を地域別に把握することは、効果的な教育政策の立案と実行にとって必要である。

そこで、本研究では、学歴不詳率が高い 2010 年および 2020 年の国勢調査において、「不詳」データを適切な学歴に分類し、より正確な学歴分布を推定することを目的とする。これまで国勢調査においては、年齢や性別の不詳の補完は行われてきたが、学歴に関して補完は実施されていない。

国勢調査における年齢や性別の不詳の補完には、按分法が従来用いられてきた。しかし、按分法では、当該変数の分布を簡易的に推定するに留まり、推定結果の精度には限界がある。そこで、本研究では国勢調査における学歴分布の推定に学校基本調査を併用することを提案する。学校基本調査は学生の学歴に関する正確な情報を提供する調査である。学校基本調査の併用によって、従来の按分法よりも妥当な学歴分布の推定が可能となると考えられる。

実際、後述のとおり、学校基本調査の併用によって、全国レベルでは従来よりも精度の高い学歴分布推定が可能となった。ただし、学校基本調査を用いても、都道府県別の学歴分布推定は困難である。そこで、本研究ではさらに、都道府県別学歴分布を推定するため、按分法及び個票データを用いた方法を試みることとする。

本研究は、まず学校基本調査を用いて学歴分布を推定する。学校基本調査と国勢調査を組み合わせて、推定対象時点の 10 年前に「15 歳から 24 歳」と「25 歳以上」の 2 つの年齢層に分けて別々に学歴分布を推定する。「25 歳以上」は、推定対象時点から 10 年前の国勢調査から、各年齢層における死亡率を考慮して推定を行う。「15 歳から 24 歳」は、推定対象時点の 10 年前に「卒業」と「在学中」に分かれるので、別々に推定する。「卒業」の人は、推定対象時点から 10 年前の国勢調査から各年齢層における死亡率を考慮して推定を行う。「在学中」の人は、学校基本調査における「卒業後の状況調査」を用いて、卒業後の進路を学歴とする。最後に、2 つの学歴分布を合算して、全体の学歴分布を推定する。

$$X_i = X_i^{(15\sim24)} + X_i^{(25+)} \quad (1)$$

次に、2 つの手法で都道府県別の学歴分布の推定を試みる。1 つめの手法として、按分法を用いた推定法が挙げられる。按分法は、国勢調査データを学歴、性別、年齢および都道府県別に細分化し、細分化された各グループの中で、「不詳」を除いた総数に対する各カテゴリの割合に基づき、「不詳」データを各カテゴリに割り当てる。なお、按分法は「全体按分法」と「層別按分法」の 2 種類ある。全体按分法は、全体の公表結果の中で「不詳」を按分

する方法である。層別按分法は、学歴、性別、年齢等で細分化されたデータ内で「不詳」を按分する方法である。本稿は、層別按分法を「按分法」と略記する。

$$X_i' = \sum_g (X_{ig} + U_g \times \frac{X_{ig}}{S_g}) \qquad (2)$$

2 つめの手法として、個票データを用いた推定方法が挙げられる。学歴不詳の個票データを「未回収」と「無記入」に分類して推定する。まず、未回収データは、グループgの未回収者数U_gに国勢調査における学歴不詳の学歴iの割合$r_{unknown,i,g}$をかけることで、未回収データにおける学歴i 全体の推定値X_i'を求める(式(3)参照)。学歴不詳の学歴iの割合$r_{unknown,i,g}$は、学校基本調査から推定した学歴分布と学歴不詳以外の学歴分布から算出した（式(4)参照)。

$$X_i' = \sum_g (U_{ig} \times r_{unknown,i,g}) \qquad (3)$$

$$r_{school,i,g} = (1-u) \cdot r_{known,i,g} + u \cdot r_{unknown,i,g} \qquad (4)$$

次に、無記入データは、回帰代入法とホットデック法を用いる。いずれの手法も国勢調査及び社会生活基本調査の匿名個票データを使用する。回帰代入法は、定数項β_{0i}及び説明変数X_n（性別、年齢、配偶者、世帯の種類)に対応する回帰係数β_{ni}に基づき、特定の学歴カテゴリiに該当する確率$P(Y=i|X)$を求める(式(5)参照)。その後、得られた各学歴カテゴリiの確率の中で、最も高い確率のカテゴリを学歴として補完する。

$$P(Y=i|X) = \frac{exp(\beta_{0i} + \beta_{1i}X_1 + \beta_{2i}X_2 + \cdots + \beta_{ni}X_n)}{\sum_{i=1}^{4} exp(\beta_{0i} + \beta_{1i}X_1 + \beta_{2i}X_2 + \cdots + \beta_{ni}X_n)} \qquad (5)$$

ホットデック法は、性別、年齢、都道府県及び世帯の種類を補助変数として、匿名個票データをグループ分けする。その後、グループ内のドナー候補が持つ学歴カテゴリの分布を基に、確率的に補完値を割り当てる。その際に、学歴不詳データの学歴iの割合$r_{unknown,i,g}$を重み付けに使用する。グループg内での学歴カテゴリiの確率$P(Y=i|g)$は、学歴iの重み$w_{i,g}$に学歴iのドナー数n_iをかけ、全ての学歴カテゴリの重み付きドナー数の合計で割ることで算出される（式(6)(7)参照)。

$$w_{i,g} = \frac{r_{unknown,i,g}}{r_{known,i,g}} \qquad (6)$$

$$P(Y=i|g) = \frac{w_{i,g} \cdot n_i}{\sum_{i=1}^{4} w_{i,g} \cdot n_i} \qquad (7)$$

まず、学校基本調査を用いることで、全国レベルでは従来の全体按分法と比べて、より妥当な学歴分布を推定することが可能となった。推定結果を具体的に見てみると、学校基本調査を用いて推定した大学・大学院の割合は、全体按分法による大学・大学院の割合と比較して、2010 年および 2020 年ともに低い結果となった。一方で、他の学歴層の割合は、全体按分法による他の学歴層の割合よりも、2010 年および 2020 年ともに高い結果となった。以上の結果は、学歴不詳者は学歴の低い層で発生しやすいという佐野・多田・山本（2015)の

指摘とも一致している。

次に、都道府県別の学歴分布の推定を試みた 2 つの方法の中では、個票データを用いた方法が最も妥当である結果が得られた。精度の比較にあたっては以下の手順を採用した。まず 2 つ方法それぞれで、推定した都道府県別の学歴分布を合算し、全国における学歴分布を求めた。次に、求められた学歴分布と、学校基本調査を用いて推定した学歴分布との間の差の絶対値の平均(MAE)を比較した。その結果、回帰代入法やホットデック法といった個票データを用いた推定法は、按分法を用いた方法よりも MAE が小さい結果となった。更に、個票データを用いた推定の中では、国勢調査を用いたホットデック法が学校基本調査を用いた推定値に最も近い結果を示した。

本研究の貢献としては、国勢調査における学歴分布の推定にあたって、学校基本調査を併用した新たな方法を提案し、従来の按分法と比べ、より妥当な推定が可能となった点が挙げられる。更に、国や地方公共団体がより正確な学歴の数値を把握し、効果的な教育政策の立案と実行に繋げられる可能性が高まった点も本研究の貢献の 1 つである。

今後の課題は、都道府県別での学歴分布推定が全国レベルに比べて正確性を欠くことである。学校基本調査を用いた推定法は性別・年齢別の推定に限定されるため、本研究のホットデック法では都道府県別に重み付けを求めることができなかった。都道府県別の学歴分布をより正確に推定するために、新たな手法の検討が必要である。