社会調査の偏り補正に関する覚え書き

―傾向スコア分析について―

吉　村　治　正*

Propensity Score Analysis for Web Survey Data Adjustment: A Memorandum.

Harumasa YOSHIMURA

要　旨

傾向スコア分析を用いたウェブ調査の偏り補正を試行した。回答率低下への有効な対処がない現状で、社会調査には適切なデータ補正の方法の導入が不可欠となっている。この点、旧来的なレーキングなどの重み付けに比べると、傾向スコア分析は適用範囲も広く補正の効果も著しい。しかしながら、主に検証事例の不足から、補正に必要な変数の予測がなかなかつかないという問題を抱えている。傾向スコア分析は今後の社会調査には必須の技法となるため、一日も早い研究事例の蓄積が求められる。

キーワード：傾向スコア分析、ウェブ調査、偏り

Ⅰ．はじめに

本稿は、社会調査における傾向スコア分析の方法と手順を確認し、その適用を事例として報告する。といっても、本稿は統計学としての貢献を目指すものではない。数理統計学に立脚した専門研究としては、星野らによる卓越した研究書（星野 2009；高井・星野・野間 2016）があり、筆者としてはこれらと肩をならべようなどとは思っていない。本稿が意図するのは、あくまでも覚え書き、初歩の統計学知識をなんとか所有する程度の筆者が、今後社会調査を行う上で必須となる傾向スコア分析の基礎的な発想と手順を整理しまとめておくことである。

日本の社会調査（social survey）の方法論は、訪問面接法による高い回答率の確保を根幹としてきた。しかしながら、戦後70年の社会状況の変化の中で、訪問面接法の実施も、それによる高回答率の確保も困難を極めるようになってきた。調査員の涙ぐましい努力にかかわらず、回答率の低下には歯止めがかからず、今日の訪問面接法の回答率は郵送法や留め置き法を下回ることがめずらしくない。これは日本だけでなく海外でも同じである。したがって、このような状況で必要とされるのは、訪問面接法に代わる代替的調査法の開発と、回答率が低下したことで生じた偏りを補正する技術の向上である（松田 2015； 埴淵・他 2015）。しかしながら、これに関しては日

平成29年 9 月 7 日受理　*社会学研究科社会学専攻　教授

本の社会学者は腰が重い。一次集計結果を提示する際も重み付けによる補正（weighting）を行っているものはきわめて珍しいし、補正を目的とした社会調査の研究もほとんど目にしない。だが、こうした現状は一日も早く脱する必要がある。本稿で取り上げる傾向スコア分析も、本邦の社会学者は無関心としかいいようのない状態であるが、これからは頻繁に利用され、社会調査の常識となっていくと予想される。いや、そうならねば社会学者は社会調査に対する発言権を失ってしまうのである。

社会調査データに傾向スコア分析を用いる場合、大きく二つの使い方がある。第一の使い方は、既知の標本に対して社会調査を実施し、回答者と非回答者を比較する。通常の社会調査では得られない非回答者からの情報を用いることで、非回答が発生した場合の影響を予測しようとする。たとえば Olson（2006）は、裁判所の協力を得て離婚訴訟の記録を入手し、その訴訟の当事者に社会調査への協力を依頼し、回答者と非回答者の比較を行った。その結果、回答者のうち非回答者に近いと判定された人（傾向スコアの低い人）の方が不正確な回答をするとは断定できず、したがって回答率を高めても測定誤差が増大するとは必ずしも予想しえないという結論に至った[1)]。Fricker & Tourangeau（2010）も同様に、縦断的（longitudinal）デザインの Current Population Survey の調査対象者の最初の 2 回の調査の回答内容を用い、その後の 6 回の調査依頼に対して非回答となった人の回答内容を予測しようとした。その結論は、Olson と同じく、非回答になりやすい人は不正確な回答をするとは言い得ないというものであった。ただし、Fricker らは、非回答になりやすい人は項目非回答の発生率が高いことも指摘している。

もう一つの使い方は、旧来型の社会調査とウェブ調査を並行して行い、傾向スコア分析を用いてウェブ調査の結果を補正しようとするものである（Lee 2006a; Lee 2006b; Schonlau et.al. 2009）。ウェブ調査の場合、旧来の社会調査法と異なり、標本抽出という手順を踏まないのが一般的である。ウェブ調査では、インターネット上のホームページなどでの広告や通販サイトでの登録メールアドレスを通じて事業者が協力者を募り、募集に応えた人をプールしておき、このプールされたモニターないしはパネル（panel）[2)] に対して社会調査の依頼を行う。したがってウェブ調査の場合、インターネットを使わない人が回答者になることがない（網羅誤差が大きい）だけでなく、そもそも自発的にモニターあるいはパネルに応募する人々がどのような人々なのか、応募しない人たちとどのように異なっているのかが全く分からない（selection error もしくは frame error）。これは旧来の社会調査法では想定されていなかった事態である。そのためにウェブ調査の場合は、どのような理由でどのような項目にどのような偏りが生じているのかを予想できない。そこで、ウェブ調査では、旧来的な社会調査の結果と対比させて偏りの出方を把握し、適切な補正の方法を考案していく必要がある。このために傾向スコア分析を用いようというわけである。

Ⅱ．傾向スコア分析の基本論理

傾向スコア分析（propensity score analysis）は、主に心理学や疫学などで採用される実験的研究（experimental research）の場面を想定して考案された。実験的研究では、被験者を二つのグループに分け、一つのグループ（処置群 = treatment group）の被検者に対しては疫学的（新薬

の投与など）・心理学的（映像を見せるなど）処置を行い、処置を行わないもう一つのグループ（統制群 = control group）と比較して、症状の改善や行動の変化に差がみられるかを検証する。この際に前提となるのが、グループの割り振りがランダムであること、つまり処置群に振り分けられた被験者と統制群に振り分けられた被験者との間に差が存在しないことである。処置を行う前は同じだったのに処置を行った後は差が生じたことが確認されることで、行った処置が症状や行動を変える効果があったとみなされるわけである。

しかしながら、このランダムな割り振りは、実際の場面ではきわめて困難もしくは不可能なことが少なくない。たとえば疫学を例にあげれば、ガンであれインフルエンザであれ、開発中の新薬の効果を検証するような場合、処置群に症状の重い患者が集まり、統制群は症状の軽い患者から構成されるようなことが生じる。これは様々な理由によるが、その一つに、実験の実施に携わる調査者が、重度の患者に対しては新薬の効果に一縷の望みを託そうとし、軽度の患者であれば新薬の予想外の副作用が生じることを恐れる、そうした感情を意識的・無意識的に反映させてしまうという指摘がされている。ところが新薬を投与された処置群の患者が重度の患者ばかりだと、たとえ新薬が効果を発揮しても、処置後の症状は処置群の方が重いという事態が生じる。これでは実験の結果を受け入れることはできない。

被検者のランダムな振り分けを妨げる要因は少なくない。調査者自身が多少とも無意識のうちに感情を働かせる、被検者が自ら実験の状況を変えるような行動をとる、実験の内容によりランダムな被検者の振り分けが倫理的に不可能とされる、など。こうした要因を交絡要因（confounding factor）という。そして、この交絡要因が働くことで実験の結果に偏りが生じる。したがって、この交絡要因の影響をなんらかの方法でコントロールしてやる必要がある。実験の状況設定にもよるが、交絡要因は容易に特定できる場合もあるし、特定が難しい場合もある。特定が容易な例としては、処置群に比較的年齢の高い被検者が集まり、この年齢の違いが交絡要因となって働いていると推測される場合がある。このような場合、同じ年齢の被検者を選んで比較する、あるいは年齢に対し重み付け（weight）を与えるといった方法で交絡要因の影響をコントロールすることができる。では、複数の変数が交絡要因を構成している場合はどうすればいいか。これが傾向スコア分析の出発点となる。

傾向スコア分析の手順は、まず観察された変数を用いて交絡要因の特定とその影響を測定し、次いでこの交絡要因の影響を個々のケースに対して重み付けを行うことで調整する、という二段階になっている（Lee 2006b; Leite 2017）。まず、この第一段階となる交絡要因の影響の測定について、もっとも平易な統計モデルを用いて表現すると、以下のようになる（Leite 2017）。標本が処置群と統制群に振り分けられており、その標本の個々について観察された変数が共変量を構成している場合、

$$\mathrm{logit}(Z_i = 1 \mid X) = \beta_0 + \beta_1 X_1 + \cdots \beta_k X_k$$

ただし、

$$\mathrm{logit}(Z_i = 1 \mid X) = log\left(\frac{P(Z_i = 1)}{1 - P(Z_i = 1)}\right)$$

$Z_i = 0$ ：統制群

$Z_i = 1$ ：処置群

すぐに理解できるように、これはロジスティック重回帰分析のモデルである。つまり、観察された変数 X_1 から X_k を独立変数として、個々のケースが統制群に属するか処置群に属するかをロジスティック重回帰分析によって予測しようというわけである。年齢なり性別なり血圧なり、処置群と統制群との間に差が存在すれば、その独立変数が有意に現れる。有意でなければ処置群と統制群の間に差がなく、したがって交絡要因としては作用していないということになる。そして、この回帰方程式における期待値（$e_i(X)$）、つまり個々のケースが処置群になる確率を傾向スコア（propensity score）と呼ぶ。したがって傾向スコアとは、「観察された共変量が与えられたときの処置群に割り当てられる条件つき確率」（Leite 2017, 5）と定義される。

次に個々のケースに与えられた傾向スコアから重み付け値（weight）を算出し、これを補正に用いる。重み付け値の算出方法はいくつか提唱されている（Leite 2017）。もっとも一般的なものを紹介していくと、

<1>　ATT（average treatment effect on the treated）は、

$$W_i = Z_i + (1 - Z_i)\frac{e_i(X)}{1 - e_i(X)}$$

と定義される。この $\frac{e_i(X)}{1 - e_i(X)}$ が、先のロジスティック回帰分析の従属変数（指数関数変換を行ったもの）である。したがって、もし統制群（$Z_i = 0$）であれば、

$$W_i = 0 + (1 - 0)\frac{e_i(X)}{1 - e_i(X)} = \frac{e_i(X)}{1 - e_i(X)}$$

もしも処置群（$Z_i = 1$）であれば

$$W_i = 1 + (1 - 1)\frac{e_i(X)}{1 - e_i(X)} = 1$$

となる。つまりATTは、統制群に重み付けを行い、処置群に共変量をそろえる方法である。

<2>　ATE（average treatment effect）は、

$$W_i = \frac{Z_i}{e_i(X)} + \frac{1 - Z_i}{1 - e_i(X)}$$

と定義される。この時、統制群（$Z_i = 0$）であれば、

$$W_i = \frac{0}{e_i(X)} + \frac{1 - 0}{1 - e_i(X)} = \frac{1}{1 - e_i(X)}$$

処置群（$Z_i = 1$）であれば、

$$W_i = \frac{1}{e_i(X)} + \frac{1 - 1}{1 - e_i(X)} = \frac{1}{e_i(X)}$$

となる。要するに、ATE は、処置群であれば傾向スコアの逆数、統制群であれば 1 から傾向スコアを引いた数の逆数が重み付け値となる。

ATT も ATE も、傾向スコア分析の重み付け値としては、もっとも計算が簡単なものである。ただし、それだけに注意すべき点も多い。例えば ATE は概念的にも非常にシンプルでわかりやすい半面で、データの性質によっては極端に大きい、あるいは小さい値をとるケースが現れるこ

とがある（Leite 2017）。これは外れ値を生み出しやすい社会調査データの場合は特に注意を要する点で、この点を考慮した場合にアルターネイティブになるのが次の層化重み付け法である。

<3>　層化重み付け法（stratified weighting）

Lee & Valliant（2009）によれば、層化重み付けの手順は、まず処置群も統制群も含め、すべてのケースを傾向スコアの高低によってクラスター化する。一般的には５つ程度のクラスターに分ける（五分位化）ことが多い。各クラスターに入るケースの数は同じだが、クラスターごとに処置群と統制群の割合は異なる。この割合を均一化するためには、

$$fc \equiv \frac{n_c^R / n^R}{n_c^W / n^W}$$

n_c^R　：統制群のうち、そのクラスターに入った数

n^R　：統制群のケースの総数

n_c^W　：処置群のうち、そのクラスターに入った数

n^W　：処置群のケースの総数

をクラスターごとに算出し、それを処置群の重み付け値として与えればよい。（統制群の重み付け値は１を与える）。こうすることで、処置群と統制群の共変量は等しくなる。つまり層化重み付け法は、ATEとは逆に処置群に重みをつける。

Ⅲ．事　例

傾向スコア分析の手順を実際のデータを用いて再確認しよう。この事例でデータとして用いるのは、2016年12月に筆者と共同研究者が行った「外国との付き合い方に関する意識調査」[3)]と題する、関西六県に居住する25歳以上69歳以下の日本人男女を対象とした意識調査である。この調査の目的は、若年者を中心に広がっているとされる「右傾化」といわれる現象の分析にあった。インターネット上では、外国や外国人に対する排外的な言論や他者に対する攻撃的な意見が強く現れるといわれる。いわゆる「ネット右翼」と呼ばれる現象である。この理由は、インターネット上では検索エンジンの働きにより自分と同様な意見・考え方の情報ばかりに接することになり、その結果、意見が極端な方向に強化されてしまうから（Sunsteen 2007；パリサー 2012）といわれている。この理論に従えば、インターネットのヘビーユーザーほど右翼的、つまり中国や韓国に対して批判的で愛国主義的になるはずである。ところが、ウェブ調査にモニター登録する人は、そもそもインターネットのヘビーユーザーが多いと予想される。すると、モニター登録型のウェブ調査で右傾化や愛国心といった事柄を把握しようとすると、偏りが生じる危険性が高い。ここで傾向スコア分析の登場となる。つまり、このリスクに対処するために、旧来型の社会調査とモニター登録型のウェブ調査を並行して実施し、回答者の基礎属性やインターネット利用の状況が等しくなるように共変量のバランスを取り直した上で、愛国心などの右翼的傾向をあらためて検討する必要性が生じる。

この目的のため、まず一般的な登録モニター型のウェブ調査業者に依頼し、上記条件に合致する調査対象者からの回答を収集してもらった（合計400名）。これを「処置群」とおいた。これと

同時に、住民基本台帳から無作為に抽出した調査対象者に対して、業者に委託したものと全く同じ内容の質問を全く同じ回答方式（ウェブ回答）で依頼し、回答を得た。こちらが旧来型の社会調査に準じるもの、いわゆる「統制群」ということになる。調査の実施過程の詳細については省略するが、業者委託のデータについては標本抽出を行っていないので回答率は不明、住基台帳からサンプリングしたデータについては、依頼総数1200件に対し有効回答数274件つまり22.8％となった[4)]。

まず、この二つの調査の回答者を比べてみよう。表１にまとめたとおり、登録モニターからの回答（モニター標本）と住基台帳からの回答（住基標本）とでは、年齢（モニター標本の方がやや高齢）、未婚者割合（モニター標本の方が未婚者が多い）、世帯所得（モニター標本の方が低い）などの項目について差が現れている。またインターネットの利用時間についても、休日の利用時間でモニター標本の方がはっきりと長くなっている。つまりモニター標本は、インターネットのヘビーユーザーに偏っていることがわかる[5)]。

表１　住基標本とモニター標本の回答者

平均年齢	Mean	St. Dev	F
住基標本	46.49	11.26	9.84**
モニター標本	49.26	11.23	
女性割合	Mean	St. Dev	F
住基標本	0.48	0.50	3.71
モニター標本	0.41	0.49	
未婚者割合	Mean	St. Dev	F
住基標本	0.16	0.37	9.90**
モニター標本	0.27	0.44	
教育年数	Mean	St. Dev	F
住基標本	14.77	1.92	6.23*
モニター標本	14.38	2.04	
世帯所得	Mean	St. Dev	F
住基標本	641.6	410.0	18.29**
モニター標本	505.4	397.1	
平日Web時間	Mean	St. Dev	F
住基標本	2.88	2.56	1.70
モニター標本	3.13	2.38	
休日Web時間	Mean	St. Dev	F
住基標本	2.46	1.87	35.00**
モニター標本	3.66	2.84	
Web活動	Mean	St. Dev	F
住基標本	18.20	5.18	15.23**
モニター標本	16.40	6.19	

* p <0.05
** p <0.01

では、実際に傾向スコア分析を行ってみる。まず、第一段階のロジスティック回帰分析の変数設定から始めると、ウェブ調査の偏り補正を目的とする場合、一般的に補正対象となるウェブ調査を１、比較のための調査もしくは準拠となる調査を０とした変数を従属変数とおく（Lee 2006b）。共変量つまり独立変数となる変数は、交絡要因として作用している可能性のある変数、したがっ

てコントロールする必要のある変数となる。ただし、共変量としてどの変数を含めるかについては、研究者ごとに意外なほど意見が異なる。Lee（2006b）は、共変量は回答者の基本的属性に限定すべきと主張するが、Schonlau et.al.（2009）は、基本的属性だけでなく行動や態度に関する項目も含めるべきと主張する。共変量となる変数を増やすことの利点は、独立変数が増えることで決定係数が上がり、そのために補正の正確さが向上するという点にある。逆に難点といえるのが、処置変数が何であるかが不明瞭になる点である。ここでは基本的属性にあたる年齢・性別・婚姻状態・教育年数・就労状態・世帯所得とインターネットに関する行動特性変数である平日ウェブ利用時間・休日ウェブ利用時間・ウェブ活動スコアを共変量つまり独立変数とした[6)]。

表 2　ロジスティック回帰分析アウトプット

	オッズ比	標準誤差	z	P>|z|
年齢	1.190	0.078	2.64	0.008**
年齢二乗	0.998	0.001	−2.45	0.014*
女性 d	0.662	0.131	−2.08	0.037*
未婚 d	1.666	0.421	2.02	0.043*
教育年数	0.945	0.045	−1.21	0.228
就労 d	0.644	0.160	−1.77	0.076
世帯所得	0.999	0.000	−2.59	0.010**
平日Web時間	0.885	0.048	−2.24	0.025*
休日Web時間	1.424	0.087	5.76	0.000**
Web活動	0.755	0.056	-3.79	0.000**
Web活動二乗	1.006	0.002	2.91	0.004**
（定数）	1.113	2.033	0.06	0.953

* p <0.05
** p <0.01

表 2 がその結果である。このモデルの尤度比カイ二乗統計量は110.05、p<0.01、疑似決定係数は0.130であった。つまり決定係数が大きいわけではないが、統計学的にはモデルが成立していることがわかる。言い換えれば、年齢、性別、婚姻状態、世帯所得、平日および休日のウェブ利用時間、およびウェブ活動スコアに関して、モニター標本（処置群）と住基標本（統制群）との間に差が存在する。

次に、この結果から層化重み付け法を用いて、補正のための重み付け値を算出する。ロジスティック回帰分析によって、個々のケースについて傾向スコアが算出された。これを五分位化したときのケース件数が表 3 である。

表 3　層化重み付け（5 層）のケース件数ならびに重み付け値

五分位	住基標本	モニタ標本	（計）	重み付け値
1	88	41	129	3.539
2	58	70	128	1.366
3	45	83	128	0.894
4	35	93	128	0.621
5	16	112	128	0.236
（計）	242	399	641	

第一五分位から見ていくと、統制群（住基標本）は全体で242件（$n^R=242$）で、このうち第一五分位に入るのが88件（$n_c^R=88$）、さらに処置群（モニター標本）は全体で399件（$n^W=399$）、第一五分位に入るのが41件（$n_c^W=41$）なので、

$$f_1=(88/242)/(41/399)=3.539$$

同様に第五五分位まで算出し、これらの値を処置群（モニター標本）の重み付け値とする。なお、住基標本は準拠となるデータであるので、こちらの重み付け値は一律で１となる。

これを用いて、右傾化に関する態度測定項目の点数を補正前と補正後で比較したのが表４である。この調査では中国および韓国に対する好感度を「親しみを感じる」・「国家元首が好き」・「外交関係改善の必要性がある」・「芸能文化歴史に興味がある」・「その国から来た知人友人がいる」・「旅行などで行ってみたい」の６項目で測定している。愛国心については辻（2008）にしたがって６項目を設定、その合計点で測定した。

表４　右傾化測定項目の平均値、傾向スコアによる補正前と補正後

	補正前			補正後		
中国への好感度	Mean	St. Dev	F	Mean	St. Dev	F
住基標本	−4.00	4.67	38.49**	−4.12	4.51	12.34**
モニター標本	−6.28	4.68		−5.48	4.85	
韓国への好感度	Mean	St. Dev	F	Mean	St. Dev	F
住基標本	−3.41	4.91	39.97**	−3.39	4.88	10.92**
モニター標本	−5.87	4.96		−4.79	5.34	
愛国心	Mean	St. Dev	F	Mean	St. Dev	F
住基標本	5.76	4.71	13.04**	5.75	4.79	10.06**
モニター標本	4.35	5.09		4.47	5.04	

**$p<0.01$

住基台帳から抽出した回答とウェブ調査事業者の登録モニターからの回答を補正しない状態で比べると、登録モニターからの回答者の方が中韓に対してネガティブで愛国心も弱い。もしもこの態度の差が回答者の基本的属性やインターネットの利用の仕方の違いによって生じたものであるとすれば、傾向スコアを用いて補正すれば、差が消えるはずである。だが、補正後の平均値の違いを見ると、やはりモニター標本の方が中韓に対して否定的で愛国心に対しても批判的という結果が出ている。点数差を見る限り、中韓に対する好感度は補正によって多少は向上している様子が見られるが、まだ住基標本との差が顕著に残っている。愛国心についてはほとんど変化がない。つまり傾向スコア分析によって基本的属性やインターネット活動の特性の違いがもたらす影響をコントロールしても、右傾化に関連する態度の違いは残存していることになる。

Ⅳ．考　察

傾向スコアを用いてモニター登録型のウェブ調査の「偏り」を補正できるか。補正の技術としての傾向スコア分析は、きわめて高いポテンシャルを示している。しかしながら、本稿でとりあ

げた事例に関する限り、まだ不十分な補正と言わざるを得ない。本稿の事例で補正が不十分に終わった理由として考えられるのは、第一に統制すべき共変量が適切にモデルに含まれていなかったという可能性である。このロジスティック回帰分析の疑似決定係数は0.13。これは説明力としては低いとみなさなければならない。決定係数が高いほど予測が正確になり適切な補正ができるようになるのだから、属性やウェブ活動に関する質問項目を再確認し見過ごしていた変数をモデルに含める、連続量の変数をカテゴリー化してみる、二次関数・三次関数でのフィットを試みるなど、決定係数を高めるようなモデルを構築する必要がある。もしも決定係数があがることで、この中韓への好感度や愛国心の差が消えていけば、ウェブ調査の偏りを適切に補正することができているということになる。

もしもモデルを見直しても態度に関する差が残るとしたら、あらかじめ予想して測定された変数以外に好感度や愛国心に影響をおよぼす要因があるということになる。これが第二の可能性である。わかりやすく言うと、ウェブ調査のモニター登録をする人が、インターネットのヘビーユーザーであるというだけではなく、なにか他の人々とは異なる特定の行動特性あるいは心性を持っていて、それが態度測定の場面で影響をおよぼしているという可能性である。こちらの場合であれば、研究事例を蓄積していくことで、この特性が少しずつ明らかになっていくことが期待できる。

個別の事例を離れ、ウェブ調査の偏りの補正という全体的な文脈に立ち戻ると、傾向スコア分析の有用性に驚かされると同時に、実用面での難しさも目につく。第一の難点は、本事例でも見られたように、高い決定係数が得られないことが少なくないという点である。決定係数が低いと補正の精度も低下する。一回や二回程度、散発的に行った社会調査では、決定係数を高めるためにどのような変数が必要であるかがわからず、結果として補正も不十分なものにとどまらざるを得ない。

第二の難しさとは、傾向スコア分析は常に準拠すべきデータを必要とするという点である。レーキング（raking）などの旧来的な重み付けの方法は、母集団に関する周辺分布がわかっていれば適用できた。これは便利であると同時に、年齢や性別などの限られた変数についてのみ補正が可能ということでもあった。これに対し、傾向スコア分析は基本的属性だけでなく行動特性や態度なども補正の基準とすることができる。これについては、補正の精度が飛躍的に高まることを期待できるのだが、それは同時に、これらの行動特性や態度などについて、周辺分布だけでない情報を必要とするということでもある。つまり傾向スコア分析は、常に比較対象となるデータ、「正解」とみなし得るデータが存在することで初めて適用可能となる。ウェブ調査の場合でいえば、ウェブ調査の結果が偏っていて補正したいと望むのであれば、ウェブ調査と異なる方法（訪問面接法や郵送法など）で得られた、同じ内容で偏りのないデータが必要となるのである。これは、ある意味で矛盾する点である。つまり、根本的な問題として、旧来的な社会調査の実施が困難になっているからこそウェブ調査への移行が期待されるのであり、常に旧来的な方法に基づく社会調査の実施を伴わねばならないということであれば、そもそもウェブ調査を実施する必要性はない。

この問題への対処であるが、比較実験を繰り返すことで補正に有効な変数が次第に特定され、

補正に必要な係数（重み付け値）が経験的に安定して予測できるようになれば、やがて準拠データを収集する必要はなくなっていく。ウェブ調査で旧来的な社会調査法と同じデータが得られることになれば、困難さを抱えた旧来的な社会調査法からウェブ調査への移行が進む。これは社会調査の一つの方向性といってよい。だが、そのためには、補正のための予測が安定してくることが必須となる。結局のところ、ウェブ調査への移行に必要なのは、なによりも事例の蓄積と経験的なノウハウの構築ということになる。

これまでは社会調査のデータ補正といえば、性別や年齢など、ごく限られた属性についてしか行い得ず、しかもその補正がどこまでうまくいっているかもはっきりしないことが多かった。これに比べると傾向スコアによる補正は、基礎属性だけでなく行動の特性や場合によっては心的特性なども用いることが可能となる。この点、補正の幅は飛躍的に広がったといってよい。

注

1）回答率とデータの正確さとの関係については別稿で論じる予定だが、この点も覚え書きとして付記しておく。回答率が上がるほど収集された情報が正確になると考えるのは社会学者であり、心理学者や社会心理学者は回答率が上がるとむしろ情報の信ぴょう性は低下すると考える。これは、社会学者と心理学者でイメージしている非標本誤差の種類が異なることから生じる見解の相違といってよい。社会学者は回答率の高低を非回答誤差の問題、つまり回答している人と回答していない人が異なっていることで生じる偏りとみなす。そのため、回答率が上がる（回答していない人が相対的に減る）ことで収集した情報は正確になると考える。これに対し心理学者の場合、関心のない人に無理に回答を求めてもいい加減な回答、不正確な回答しか得られず、結果として情報の正確さが低下すると考える。つまり心理学者は回答率が上がることで測定誤差が増大するリスクを重視する。それ故、社会学者は回答率を高めることに血眼になり、心理学者は高い回答率に疑いのまなざしを向ける。単純に、回答率は高いほうがいい、というものではない。このあたりの研究事例については、吉村（2017）を参照されたい。

2）一般的に社会調査の専門用語でパネル（panel）という場合、縦断的（longitudinal）調査における個々のフェーズ（調査の時点）を指す。これに対してウェブ調査の場合、調査対象となった個々人をパネルと呼ぶ（Collegaro et.al. 2014）。標本（sample）とよぶことはほとんどない。標本抽出していないものを標本と呼べるかといわれると確かにその通りだが、用語法に混乱の印象があるのは否めない。ともあれ、こうした用語法の変化がどのような理由で生じたかは、はっきりしないが、おそらく2010年前後にはこうした用語法が定着してきたと思われる。例えばウェブ調査について比較的早期に著された Best & Krueger（2004）では sample という用語を用いているが、10年後の Tourangeau et.al.（2013）では panel もしくは volunteer web panel という表現で一貫させている。

3）この調査は、平成27～29年度文科省科研費基盤研究（C）：課題番号JP15K03827『社会学的知見に基づくWeb調査の代表性の分析』（代表研究者：吉村治正）および大川情報通信基金2015年度研究助成『インターネット調査の偏向性の研究』（助成番号15-23、代表者：吉村治正）の助成を受けて行われた。

4）依頼者1200名の内訳は、宛先不明６件、拒否連絡226件、郵送回答（希望者のみ：有効回答件数に数えず）80件、無効回答13件、無反応602件、そして有効回答件数273件であった。

5）表中のウェブ活動というのは、メールの送受信・ネットニュースの閲覧・ネットサーフィン・ネット通販・オンラインゲーム・SNS・動画鑑賞・ブログの開設や更新・他人のブログへの書き込みの9項目についての利用頻度をスコア化し合計したもので、得点が高いほどインターネットでの活動に深くかかわっ

ていることを示している。

6）ダミー変数の定義は以下の通り。女性ｄ（女性＝１、男性＝０）、未婚ｄ（未婚者＝１、既婚・離死別＝０）、就労ｄ（就労中＝１、不就労＝０）。なお、世帯所得の単位は万円、平日および休日ウェブ時間の単位は時間。

参考文献

Best, Samuel J. & Brian S. Krueger. 2004. *Internet Data Collection.* SAGE University Paper.

Collegaro, Mario, Reg Baker, Jelke Bethlehem, Anja S. Goritz, Jon A. Kroznick & Paul J. Lavrakas, eds. 2014. *Online Panel Research: A Data Quality Perspective.* Wiley.

Fricker, Scot & Roger Tourangeau. 2010.“Examining the Relationship between Nonresponse Propensity and Data Quality in Two National Household Surveys.” *Public Opinion Quarterly,* 74(5): 934-55.

Lee, Sunghee. 2006a. “An Evaluation of Nonresponse and Coverage Errors in a Prerecruited Probability Web Panel Survey.” *Social Science Computer Review,* 24:460-475.

Lee, Sunghee. 2006b.“Propensity Score Adjustment as a Weighting Scheme for Volunteer Panel Web Surveys.” *Journal of Official Statistics,* 22(2): 329-349.

Lee, Sunghee & Richard Valliant. 2009.“Estimation for Volunteer Panel Web Surveys Using Propensity Score Adjustment and Calibration Adjustment.” *Sociological Methods & Research,* 37(3): 319-343.

Leite, Walter. 2017. *Practical Propensity Score Methods Using R.* SAGE.

Olson, Kristen. 2006.“Survey Participation Nonresponse Bias, Measurement Error Bias, and Total Bias.” *Public Opinion Quarterly,* 70(5): 737-758.

Schonlau, Matthias, Arthur van Soest, Arie Kapteyn, & Mick Couper. 2009.“Selection Bias in Web Surveys and the Use of Propensity Scores.” *Sociological Methods and Research,* 37(3): 291-318.

Tourangeau, Roger, Fredrick G. Conrad & Mick Couper. 2013. *The Science of Web Surveys.* Oxford University Press.

埴淵智哉・村中亮夫・安藤雅登 2015.「インターネット調査におけるデータ収集の課題」、『E-journal GEO』10(1)：81-98.

星野崇博 2009.『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

井上哲浩・日本マーケティングサイエンス学会編 2007.『Webマーケティングの科学－リサーチとネットワーク』、千倉書房

カッツ, ミシェル 2013『医学的介入の研究デザインと統計－ランダム化/非ランダム化研究から傾向スコア、捜査変数法まで』、木原雅子・木原正博訳、メディカル・サイエンス・インターナショナル社（Katz, Mitchell H. 2010. *Evaluating Clinical and Public Health Interventions: A Practical Guide to Study Design and Statistics.* Cambridge University Press.）

松田映二 2015.「インターネット調査の新しい可能性」、『政策と調査』9：5-18.

大隅昇・前田忠彦 2007・2008.「インターネット調査の抱える課題」(1・2)、『日本世論調査協会会報』100：58-70＆101：79-94.

パリサー，イーライ 2012.『閉じこもるインターネット』、井口耕二訳、早川書房（Pariser, Eli. 2011. *The Filter Bubble.* Penguin Press.）

高井啓二・星野崇宏・野間久史　2016.『欠測データの統計科学』、岩波書店

辻大介 2008.『インターネットにおける右傾化現象に関する実証研究：調査結果概要告書』

吉村治正 2017.『社会調査における非標本誤差』、東信堂

Summary

This is a memorandum regarding a method of survey data adjustment. Facing to unremitting growth of refusal and consequent decline of response rates, survey researchers have urgent demand for some reliable adjustment methods. Propensity score analysis (PSA), a statistical procedure to reduce Web survey biases through controlling covariates, is for this purpose practiced and examined its usability. The result is remarkable; the PSA shows a quite high potential comparing to conventional cell-weighting, but simultaneously suggests a difficulty to build effective models in actual situations. Further implications are discussed.

Key words：propensity score analysis, Web survey, bias