（要旨）On the model construction for estimating response propensity scores for survey data adjustment

（調査データ補正のための回答傾向スコアを推定するモデル構築について ）

福田昌史 05S110003J

理工学研究科数学専攻

国内でかつて行われていた調査は，住民基本台帳や選挙人名簿という網羅性の高いリストからの対象者抽出が可能なことに加え，高い回答率を得ることができ，理想的な環境が整っていた．しかしこの環境は，法改正による台帳閲覧の制限拡大や，プライバシー意識の高まり，オートロックマンションの増加などによる回答率の低下によって年々悪化してきている（Synodinos & Yamada, 2000; 吉野, 2002, 鈴木, 2006; 上村, 2007; 雀田, 2008）．現在マスコミによって使われている電話調査（RDD）も，固定電話のみを対象としているために，「固定電話を持たずに携帯電話しか使わない層」の意見が調査結果に反映していないのではないかという懸念も報告されている（Keeter et al., 2007; Blumberg & Luke, 2007）．また，郵送調査による高回答率も報告されているが（松田, 2008 など），郵便を使った調査は時間がかかるという欠点があり，万能とは言えない．

その一方で，情報技術の発展によってインターネットを使った調査（Web 調査）が近年行われるようになった．Web 調査は，スピード性や低コスト，多様な属性変数，マルチメディアを使った調査画面など，魅力的な多くの利点を持つが，自発的に会員登録した集団を対象としているために回答者の構成が偏っており，例えば「日本の有権者全体」などのより大きい目標母集団に関する推論ができないという大きな欠点を持っている．

この「偏った回答者集団のデータ」から偏りを除去する（減らす）方法として「傾向スコア」を使った推定手法が使われている．傾向スコアは元々，無作為実験ができない場合の，観察データから処理効果を推定するために提案されたもので（Rosenbaum & Rubin,1983），現在はさまざまな分野に応用が広がっている．この処理効果は，処理（Treatment）を受けた人と，受けていない人の結果変数（Outcome）の平均の差のことをいう．例えば，「アルコールの乱用による所得への影響」を知りたい場合（Jones & Richmond, 2006）は，「アルコールの乱用」が処理，「所得」が結果変数に相当するが，現実にはアルコールを大量摂取する人をくじ引きで無作為に決めることができず，乱用者と非乱用者の収入の差を直接求めても意味のある結論を得ることができない．この場合，性別や年齢，教育，違法行為の頻度…など，多数の変数（共変量とも呼ぶ）を揃えた上，共通の特徴を持つ個体同士で比較する必要がある．傾向スコアは，このように比較が可能となるような共変量の値が与えられたときの，処理を受ける確率（上記の例の場合，性・年齢・教育などの変数を固定したときの，アルコールを乱用している条件付き確率）として定義される．つまり傾向スコアは共変量の値が決まれば値が決まるもので，個体それぞれが持つ値である．これまでの研究では，多数の共変量を揃えて同じ共変量の値を持った個体同士を比較することが困難な

場合でも，共変量を１次元に集約した傾向スコアの値による層別を基にした比較（Rosenbaum & Rubin, 1983）や，重み付け推定（Rosenbaum,1987）によって処理効果の推定が可能であることが示されている．

この傾向スコアが，偏りのある調査回答データの調整にも応用されている．（その用途で使われるときは「回答傾向スコア」とも呼ばれる．）例えば，Taylor(2000, 2001)や Isaksson & Försman(2004)は，Web 調査のデータに傾向スコアで補正を適用して選挙予測をした．しかし，その用途は Web 調査の補正のみにとどまらず，無回答誤差の補正や，網羅度の低いリストを使った調査の補正など，多岐にわたる．

傾向スコアを使う研究において，処理効果の推定の場面では２種類のグループ「処理群」と「非処理群」のデータが扱われていたが，調査データの調整では２種類の調査データ「偏りのある調査の回答データ」と「偏りのない調査の回答データ」を扱う．しかし，これまでの標本調査の文献では，処理効果の推定で使われてきた「処理群と非処理群の割り当て」という文脈を，そのまま「偏りのある調査の回答と，偏りのない調査の回答」に対応させて説明するだけで，厳密な議論がされていなかった．それらの説明では，両方の調査回答データが得られているところから議論がスタートしているが，どのように回答データが母集団から集まるのかを考慮する必要がある．

調査への参加は，現実には「母集団の中から対象者が選ばれ，選ばれた対象者が回答，もしくは拒否する」という２段階のプロセスを経ることによって決まるものであり，処理効果の推定の場合のようにデータをどちらかに振り分けるという考え方は適当ではない．そこで本論文では，参加集団に偏りのある調査（論文内では「Running survey」と表記）と，偏りのない，基準となる調査（「Reference survey」と表記）の２つの調査の，それぞれの対象者に選ばれる確率（抽出確率）と，選ばれたときに回答する確率（回答確率）を使って傾向スコアを新たに定義し直した．さらに，それによる重み付け推定が，これまでよく使われている事後層別補正と関連していることを示した．

傾向スコアを使った実際の研究においては，真の傾向スコアの値は未知であるため，これを推定する必要がある．その推定には通常，ロジスティック回帰モデルが使われるが，真の共変量セットも未知であるため，推定モデルに使う共変量の組み合わせをいかに見つけるかが最も重要な課題となっている．

この課題について，元来，傾向スコアは処理の割り当ての概念で説明され（Rosenbaum & Rubin, 1983），処理の割り当てに関係している変数をモデルに入れるのがよいと考えられたが，その後の，処理効果推定のための傾向スコア推定に関する研究では，「結果変数に相関のある全ての変数をモデルに含めるべき（Rubin & Thomas,1996）」，「結果変数と，処理の両方と相関のある変数を全てモデルに含めるべき（Perkins et al.,2000）」，「処理との相関がなく，結果変数と相関がある変数を含めるべき．そうすれば推定量の分散が減少する．また，逆に処理と相関があり，結果変数と相関がない変数を含めると分散が増加する（Brookhart, 2006 など）」「真のモデルによって推定した傾向スコアが最適な処理効果の推定とは限らな

い（Austin, 2007 など）」という結論が得られている．

その一方，調査データを調整することを目的とした傾向スコアの推定に適した変数の選び方に関する文献はそれほど多くなく，星野・前田(2006)が変数選択に関するガイドラインを示したほか，実際の調査データの調整に使われている変数となる質問項目の例が報告されている（Taylor, 2000, 2001; Schonlau et al., 2004, Schonlau, 2007 など）にとどまっている．

本論文では，どのような変数を推定モデルに加えれば（除去すれば）調査データの調整に効果があるのかに焦点を絞って検証した．具体的には，モデルに含める変数の候補として，結果変数と相関があるかどうか，抽出確率と相関があるかどうか，回答確率と相関があるかどうかがそれぞれ異なる 7 つのタイプの変数を考え（表１），それらの変数の全ての組み合わせからなる 127 モデルを，コンピューターシミュレーションで検証した．

（表１）効果を検証した 7 変数

No association with study variable

	No association with response	Associated with response
No association with sampling		$x_{<\cdot\cdot\theta>}$
Associated with sampling	$x_{<\cdot\pi\cdot>}$	$x_{<\cdot\pi\theta>}$

Associated with study variable

	No association with response	Associated with response
No association with sampling	$x_{<y\cdot\cdot>}$	$x_{<y\cdot\theta>}$
Associated with sampling	$x_{<y\pi\cdot>}$	$x_{<y\pi\theta>}$

シミュレーションでは，仮想的な 10 万人の母集団データを生成した．母集団を構成する各個体には，結果変数（論文内では study variable と表記），7 変数，偏りのある調査への抽出確率と回答確率，基準調査への抽出確率と回答確率の値を与えた．目標は，偏りのある調査から，傾向スコア補正を通じてバイアスを除去し，結果変数の母集団平均を正確に推定することである．まず基準調査のデータ約 5000 人を抽出し，回答確率に応じて回答者を決定する．その後，偏りのある調査データ約 1000 人を 10,000 回選択し，それぞれの回で 127 個のモデル全てを適用して傾向スコアの推定の正確さと補正の関係や，7 変数の効果を検証した．推定の正確さは，傾向スコアの推定値と真の値の間の平均二乗誤差（MSE）をモデルごとに計算し，それぞれのモデルによる改善の程度（どれだけバイアスを除去できたか）と比較した．7 変数の効果は，それぞれの変数について「含むモデル（With- model）」

と「含まないモデル（Without- model）」の２つのモデルからなるペアを 63 組作り，With モデルと Without モデルでの母集団平均推定の補正の程度と推定値の分散を比較した．

シミュレーションによって以下の結果が得られた．まず，傾向スコアの推定値が真の値に近く，推定誤差が小さいからといってデータの補正がうまくいくわけではないということが分かった．言いかえれば，既存のロジスティック回帰モデルのモデル選択基準がそのまま使えるわけではないことを意味している．そして，モデルに必要不可欠な変数は，結果変数と相関があり，同時に調査参加に関する確率（抽出確率や回答確率，もしくはその両方）とも相関がある変数だった（表中の，$x_{<y\cdot\theta>}$，$x_{<y\pi\cdot>}$，$x_{<y\pi\theta>}$）．これは，Perkins et al. (2000)に類似した結果と言える. その中で最も調査データの調整に貢献したのは, 結果変数, 抽出確率，回答確率の３つ全てに相関がある変数（$x_{<y\pi\theta>}$）だった．また，結果変数と抽出確率の２つに相関がある変数（$x_{<y\pi\cdot>}$）と，結果変数と回答確率の２つに相関がある変数（$x_{<y\cdot\theta>}$）については，後者の方がわずかに良い推定だった．また，結果変数だけに相関がある変数（$x_{<y\cdot\cdot>}$）については，モデルに含めることによる効果は限定的で，Rubin & Thomas (1996)と同様の結果をはっきりと確認することはできなかった．相関の強さなどを変化させて, さらなる検証が必要である. その他の変数, つまり結果変数とは相関がない変数（$x_{<\cdot\cdot\theta>}$，$x_{<\cdot\pi\cdot>}$，$x_{<\cdot\pi\theta>}$）については，モデルに含めるべきでないことが分かった．それらをモデルから除去することにより，推定値や推定値の分散が改善された．

以上の結果より，興味のある結果変数が複数ある場合，その中のある一つの結果変数をうまく補正できる傾向スコアモデルは，別の結果変数をうまく補正できるとは限らず，それぞれ最適なモデルが異なるといえ，万能な共変量の組み合わせは存在しないことも分かった．

本論文でのシミュレーションでは検証しなかった課題として，共変量と３変量（結果変数，抽出確率，回答確率）間の相関の強さが変わると，調整にどう影響するか，共変量同士に相関がある場合，基準調査（Reference Survey）の標本誤差を考慮した場合はどうなるか，などが挙げられる．これらの検証も今後必要となる．