【『統計学』創刊60周年記念特集論文
特集A：標本設計情報とミクロデータ解析の実際】
（『統計学』第117号 2019年9月）

サンプルセレクションバイアス補正方法の比較検証

― 社会生活基本調査ミクロデータを利用して ―

栗原由紀子*

要旨

本研究は，公的統計ミクロデータの利活用を目指して，ターゲットとする公的統計調査と同種の調査票を使い，近接する時期にWeb調査を実施するケースを想定し，Web調査におけるサンプルセレクションバイアスの補正方法について検証した。その結果，まず傾向スコアとキャリブレーションを比較したとき，補助標本のサイズが同じであればMSEに顕著な差はみられなかった。また，キャリブレーションにおいては，周辺度数を利用する方法（レイキング比推定量）とクロス度数を利用する方法（一般化回帰推定量）のいずれでも大きな差はみられなかった。さらに，バイアス発生要因となった変数が特定できない場合，共変量の組合せによってはMSEが上昇するケースも観測されたことから，条件付き独立性に関する指標を用いて，適切な共変量の組合せを確認する必要があることが示唆された。

キーワード

傾向スコア，キャリブレーション，サンプルセレクションバイアス，条件付き独立性

1．はじめに

近年，公的統計のミクロデータの研究利用が広まり，実証研究の自由度が高まりつつある。しかしながら，公的統計の中には，いわゆる大規模標本調査のようにサンプルサイズは極めて大きいにも関わらず調査実施時期が数年周期で実施されるために，循環的・季節的変化に関しては断続的にしか捉えられない設計のものが少なくない。また，公的統計には設定されていない項目ではあるが，より詳細な実態把握を目指すには，新規の調査項目の追加が必要となるケースもある。すなわち，調査未実施の期間の情報や追加的な調査項目の情報を得るには，既存の公的統計のみでは限界がある。

公的統計のミクロデータを基礎に置きながら，新たに追加情報を捕捉する方法としては，比較的，安価かつ容易に調査が可能となるWeb調査の利用が考えられる。しかしながら，Web調査によって得られたデータにはいくつかの問題が内在する。とくに，Web調査では，登録ユーザを調査対象者とする調査方式が多いことから，サンプルセレクションバイアスの発生に関する問題が指摘されている。

星野・前田（2006）および星野（2010：169-190）では，三ヵ年分の訪問調査とWeb調査の

* 正会員，立命館大学経済学部

データを用いて傾向スコア[1)]やレイキングを用いた場合について検討しており，傾向スコアを用いた補正およびその共変量選択に関する簡便法などを提示している。また，基本属性などに関する母集団情報が利用できれば，レイキングを含めてキャリブレーションによる補正も可能である。これまでの研究では，実際の調査データを用いてバイアスの程度をコントロールしながらMSEの程度を計測する方法は採用されておらず，また，条件付き独立性の成否に関する指標とMSEとの関連も捉えられていない。

本稿では，公的統計ミクロデータの利活用を目指して，公的統計と同種の調査票を使い，公的統計調査とほぼ近接する時期にWeb調査を実施するケースを想定し，Web調査におけるサンプルセレクションバイアスの補正方法について検討する。具体的には，社会生活基本調査の匿名データを仮想母集団とし，サンプルセレクションバイアスの程度をコントロールしながら標本抽出実験を行う。サンプルセレクションバイアスの補正方法には，傾向スコアとキャリブレーションを用い，これら補正方法や補正に使用する共変量の組合せによって，補正の程度に相違があるかどうかを比較検証する。

2．検証の枠組み

2.1　検証方法の概要

本研究は，以下の手順により検証を行う（図1）。

Step 1：社会生活基本調査の匿名データを仮想母集団とする。

Step 2：仮想母集団から検証のための統計量（以下，目標統計量と呼称する）を算出し，これを仮想母数とする。

Step 3：全体の抽出率を一定としつつ，サンプルセレクションバイアス（SSB：Sample Selection Bias）の程度をコントロールする抽出法を用いて，仮想母集団から標本（Web調査標本と想定）を抽出する。このような方法で抽出した標本のことを，本稿ではサンプルセレクションバイアス標本（SSB標本）と呼称する。

Step 4：上記の抽出標本を基に，補正をしない標本統計量，IPW推定（IPW：Inverse Probability Weight）による標本統計量，およびキャリブレーション推定による標本統計量をそれぞれ算出する。

Step 5：Step 3からStep 4を100回くり返し，仮想母数と標本統計量を用いてバイアスやMSEを算出し，これを検証用統計量とする。

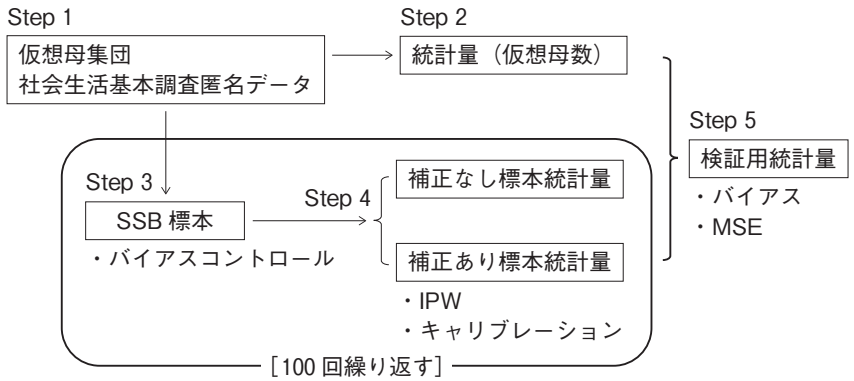

図1　検証方法の概要

Step 6：セレクションバイアスをコントロールする抽出法（抽出パターン）を10通り用意し，Step 3からStep 5をくり返し計測する。さらに，同様の実験を，部分母集団を設定したケース（仮想部分母集団と呼称）についても行う。

2.2　検証用のデータセット

検証に用いるデータは，2006年社会生活基本調査の匿名データである。このデータにおいて，二人以上世帯の女性に関する月曜日から金曜日までの40322ケースを仮想母集団（以下,「女性」とも記す）として利用し，ここからSSB標本サイズ5000を抽出する。

表1には，検証に使用するデータセットの変数を整理している。目標統計量は2次活動時間[2)]の母平均の推定量とする。バイアス発生要因をX_1変数とし，仮想母集団からサンプルセレクションバイアスの程度を調整しながらSSB標本を抽出する際にはX_1変数のみを用いる[3)]。以下では，このような変数をサンプルセレクションバイアス調整変数（SSB調整変数）とよぶことにする。

補正のための共変量としては，X_1からX_6までの6つの変数を用いている。また，仮想部分母集団としては，女性の有業者（$X_2=1$）と女性の地方居住者（$X_3=2$）を選定している。これら変数の仮想母集団に関する基本統計量と相関係数・クラーメルのVは，表2と表3にそれぞれ示している[4)]。

2.3　SSB標本の抽出方法

SSB標本の抽出方法を以下のように整理する。まず，SSB標本全体の抽出率を$f^*=0.124$と設定する。すなわち，仮想母集団サイズ$N=40322$に対してSSB標本サイズ$n=5000$となるように抽出を行う[5)]。

ここで，X_1のカテゴリー1に属するSSB標本の抽出率を$f(X_1=1)$と標記すれば，カテゴ

表1　目標変数と共変量

変数名	記号	カテゴリカルデータ・数量データ
目標変数 2次活動時間	Y	数量データ（分）
共変量・因子		
配偶関係	X_1	1：既婚　2：既婚以外
就業状況	X_2	1：有業　2：有業以外
居住地	X_3	1：都市　2：地方居住者
年齢	X_4	1：20～39歳　2：40～59歳　3：60歳以上
年収	X_5	1：400万円未満　2：400～799万円　3：800万円以上
学歴	X_6	1：小学・中学卒　2：高校・旧制中卒 3：短大・高専および大学・大学院卒

表2　仮想母集団（女性）の基本統計量（N＝40322）

Y	25%点：315　中央値：510　75%点：660 平均値：482.5　標準偏差：228.2		
X_1	1：75.4%　2：24.6%	X_4	1：27.5%　2：38.5%　3：33.9%
X_2	1：58.0%　2：42.0%	X_5	1：38.1%　2：39.2%　3：22.7%
X_3	1：70.6%　2：29.4%	X_6	1：21.8%　2：50.1%　3：28.1%

（注）　Yは分単位の数値，X_1からX_6はカテゴリー別の構成比（%）を示しており，四捨五入のため合計が100%にならないケースがある。

表3　仮想母集団と仮想部分母集団の相関係数・クラーメルのV

(a)　女性（N＝40322）

	Y	X_1	X_2	X_3	X_4	X_5
X_1	−0.109					
X_2	−0.518	0.091				
X_3	−0.001	0.015	0.051			
X_4	−0.366	0.291	0.412	0.044		
X_5	0.119	0.043	0.129	0.111	0.160	
X_6	0.257	0.076	0.212	0.119	0.365	0.191

(b)　女性・地方居住者（N＝11849）

	Y	X_1	X_2	X_3	X_4	X_5
X_1	−0.058					
X_2	−0.483	0.172				
X_3	NA	NA	NA			
X_4	−0.358	0.304	0.370	NA		
X_5	0.129	0.059	0.143	NA	0.192	
X_6	0.249	0.089	0.179	NA	0.335	0.197

(c)　女性・有業者（N＝23371）

	Y	X_1	X_2	X_3	X_4	X_5
X_1	−0.089					
X_2	NA	NA				
X_3	−0.001	0.033	NA			
X_4	−0.121	0.413	NA	0.039		
X_5	0.069	0.089	NA	0.121	0.134	
X_6	0.112	0.139	NA	0.113	0.348	0.190

（注）　YとX_1～X_6との値は相関係数，X_1～X_6どうしの値はクラーメルのVの値を示している。また，NAは仮想部分母集団に使用した変数であるために相関係数などは観測されないケースを意味している。

表4　SSB比率とSSB調整変数のカテゴリー別抽出率（女性，N＝40322）

$p(X_1=1)$	0.10	0.19	0.28	0.37	0.46	0.54	0.63	0.72	0.81	0.90
$f(X_1=1)$	0.02	0.03	0.05	0.06	0.08	0.09	0.10	0.12	0.13	0.15
$f(X_1=2)$	0.45	0.41	0.36	0.32	0.28	0.23	0.19	0.14	0.10	0.05

（注）　数値は小数第3位を四捨五入したものである。

リー別のSSB標本の度数は以下のように求められる。

$$n(X_1=1)=N(X_1=1)f(X_1=1) \quad (1)$$

$$n(X_1=2)=n-n(X_1=1) \quad (2)$$

抽出の際には，$f(X_1=1)$を引数としてセレクションバイアスをコントロールしつつ，各カテゴリーの層内ではランダムサンプリングとなるように設定する。ただし，各カテゴリーのケースの数が500以上となるように，カテゴリー1の抽出率の下限と上限をf^L，f^Uとしてそれぞれ定める。

$$f^L(X_1=1)=\frac{500}{N(X_1=1)}$$

$$f^U(X_1=1)=\frac{n-500}{N(X_1=1)} \quad (3)$$

これにより，カテゴリー2についても上限と

下限が定まる。

さらに，抽出後のカテゴリー1の構成比は，以下のように示される。

$$p(X_1=1)=\frac{n(X_1=1)}{n} \tag{4}$$

本稿では，SSBの程度を示す統計量として$p(X_1=1)$を用い，これをサンプルセレクションバイアス比率（SSB比率）と呼称する。当然，ランダムサンプルとなるのは，カテゴリー1の比率が以下の式を満たす場合となる。

$$p(X_1=1)^*=\frac{n(X_1=1)}{n}=\frac{N(X_1=1)}{N} \tag{5}$$

実際には，SSB比率は10通り用意しており，表4にはSSB比率と各カテゴリーの抽出率を整理している。とくにSSB比率が0.72の時にランダムサンプリングに近いSSB標本が得られており，この比率から乖離するに従いバイアスの程度は大きくなる。

なお，部分母集団によっては適切な共変量などが異なるケースが想定されるため，女性の地方居住者や有業者といった仮想部分母集団についてもそれぞれ同様の検証を行う。その際，SSB標本全体の抽出率は，女性の地方居住者で0.42（仮想部分母集団サイズ11849），女性の有業者で0.21（仮想部分母集団サイズ23371）となる。

2.4　サンプルセレクションバイアス補正方法

⑴　IPW推定法

サンプルセレクションバイアスの補正のために傾向スコアによるIPW推定法を用いる場合には，「強く無視できる割り当て条件」の成立が必要となる（星野（2010：43-45））。これを，本研究の枠組みで整理すれば，SSB標本または補正用の補助標本のいずれかを示す割り当て変数をd（SSB標本$d=1$，補助標本$d=0$），目標変数をy（SSB標本の目標変数y_1，補助標本の目標変数y_0を合わせたもの），共変量を$\boldsymbol{x}$（SSB標本と補助標本を合わせたもの）とするとき，以下の関係が成立しているものとする。

$$y \perp d \mid \boldsymbol{x}$$

これは，割り当てdは共変量$\boldsymbol{x}$にのみ依存し，目標変数yには依存しないことを意味する。なお，補助標本とは，SSB標本とは異なる情報として，たとえば母集団情報や他調査に基づいて得られたデータセットなども想定している。

このような条件のもとで，IPWは以下のように推定される。まず，SSB標本の確率（傾向スコア）をロジスティック回帰モデルにより算出する。なお，iは要素の番号を意味する。

$$P(d_i=1|\boldsymbol{x}_i)=e_i=\frac{\exp(\boldsymbol{x}_i'\beta)}{1+\exp(\boldsymbol{x}_i'\beta)} \tag{6}$$

次に，この傾向スコアを用いたウェイトw_i^{IP}を，(7)式を用いて算出する。

$$w_i^{IP}=\frac{d_i}{e_i} \tag{7}$$

そこで，IPWを用いた目標統計量は次のように求められる[6)]。

$$\hat{E}(\bar{y}|d_i=1)=\frac{1}{\sum_s w_i^{IP}}\sum_s w_i^{IP}y_i \tag{8}$$

本研究では，割り当て変数$d=0$に該当する補助標本として，仮想母集団から5000ケースをランダムに抽出して用いる場合と，仮想母集団をそのまま補助標本として用いる場合との2つのケースを用意し，これらの違いをIPW.SmallモデルとIPW.Fullモデルとして区別する。また，共変量の組合せによっても結果が異なる可能性があることから，これらの違いも捉えられるように表5のようにいくつかの組合せについて検証を行う。

⑵　キャリブレーション

キャリブレーションは，共変量$x_{i,j}$（$j=1,\cdots,J$は共変量の番号）について，(9)式のようにウェイト付き統計量$\sum_s w_i^c x_{i,j}$と母集団統計量$\sum_U x_{i,j}$が一致するという条件を満たしつつ，

表5　IPW推定と補助標本

モデル名	補助標本のサイズ	補正に利用する共変量の組合せ
IPW.Small	SSB標本サイズ（5000）	1変数　X_1，X_2，X_3 2変数　(X_1，X_2)，(X_1，X_3)，(X_2，X_3) 3変数　(X_1，X_2，X_3)
IPW.Full	仮想母集団サイズ（40322）	

（注）　仮想部分母集団の女性の地方居住者のケースについてはX_1とX_2の組合せ，女性の有業者のケースについてはX_1とX_3の組合せを共変量として適用する。また，表中の変数組合せの「,」は傾向スコア計算時に交互作用項を導入していないことを示している。

表6　CLB補正と補助標本

モデル名	補正手法	補助標本のサイズ	補正に利用する共変量の組合せ
CLB.Marginal	乗法関数 （レイキング比推定量）	仮想母集団サイズ （40322）	周辺度数 1変数　X_1，X_2，X_3 2変数　(X_1，X_2)，(X_1，X_3)，(X_2，X_3) 3変数　(X_1，X_2，X_3)
CLB.Cross	線形関数 （一般化回帰推定量）	仮想母集団サイズ （40322）	クロス度数 2変数　(X_1*X_2)，(X_1*X_3)，(X_2*X_3) 3変数　($X_1*X_2*X_3$)

（注）　仮想部分母集団の女性の地方居住者のケースについてはX_1とX_2の組合せ，女性の有業者のケースについてはX_1とX_3の組合せを共変量として適用する。また，表中の変数組合せの「,」は周辺度数，「*」はクロス度数を用いることを示している。

⑽式のようにキャリブレーションウェイトw_i^cと既存のウェイトw_iの距離関数$G(w_i^c, w_i)$が最小となるようにw_i^cを求める方法である（土屋（2009：130-134），Valliant, R., Dever, and F. Kreuter（2013：349-395））。

$$\sum_s w_i^c x_{i,j} = \sum_U x_{i,j}, \quad j=1, \cdots, J \tag{9}$$

$$\underset{0 \le w_i^c}{\arg\min} \sum_s G(w_i^c, w_i) \tag{10}$$

距離関数には，レイキング比推定量[7]を算出する際に用いられる⑾式の乗法関数や，一般化回帰推定量[8]を算出する際に用いられる⑿式の線形関数などが挙げられる。

$$G(w_i^c, w_i) = w_i^c log_e\left(\frac{w_i^c}{w_i}\right) - w_i^c + w_i \tag{11}$$

$$G(w_i^c, w_i) = \frac{(w_i^c - w_i)^2}{2w_i} \tag{12}$$

キャリブレーションウェイトを用いる時，目標統計量のキャリブレーション推定値は以下のように求められる[9]。

$$\hat{\bar{y}} = \frac{1}{\Sigma_s w_i^c} \sum_s w_i^c y_i \tag{13}$$

本研究では，共変量の周辺度数情報を用いる場合を想定したレイキング比推定，および共変量のクロス度数情報を用いる場合を想定した一般化回帰推定を行い，それぞれCLB.Marginalモデルおよび CLB.Crossモデルと呼称して，周辺度数かクロス度数かの情報量の違いによるMSEの相違を捕捉する。

表6には，キャリブレーションの2種類のモデルを整理しており，表6中の「補正に利用する共変量の組合せ」において，各変数の周辺度数を共変量として用いた場合には（変数名，変数名）と表記し，クロス度数を共変量として用いた場合には（変数名＊変数名）と表記している（以下同様）。

2.5　検証用の統計量

⑴　条件付き独立性の成否

IPW推定を行う前提として，標本の割り当てが共変量のみに依存し，目標変数には依存

しないことが不可欠であり，これを2.4節では「強く無視できる割り当て条件」と呼称した。このような条件の成否を数値で捉えるために，条件付き独立性（CIA：Conditional Independence Assumption）を仮定し，本稿では⑭式として表すことにする。

$$f(Y, d|\boldsymbol{X}) = f(Y|\boldsymbol{X})f(d|\boldsymbol{X}) \tag{14}$$

なお，$\boldsymbol{X}$は共変量の任意の組合せを示している。CIAの成否の判断には，⑮式に基づく指標を用い，これをCID（Conditional Independence and Dependence Index）と呼称する（栗原（2015））。

$$CID = Cor(E(Y|\boldsymbol{X}), E(d|\boldsymbol{X})) \tag{15}$$

本稿では，Yの$\boldsymbol{X}$への回帰残差ε_Y（重回帰モデル），およびdの$\boldsymbol{X}$への回帰残差ε_d（ロジスティックモデル）を求め，それら残差の相関係数をCIDの推定値とする。

$$CID = Cor(\varepsilon_Y, \varepsilon_d) \tag{16}$$

CIDがゼロに近い場合には，CIAが成立した状況にあり，ゼロから乖離している場合には，CIAの成立が確認できないものと判断する。

さらにCIDについて，6つの共変量の全ての組合せ（$k=1, \cdots, K$）別に，各SSB比率（$b=1, \cdots, 10$）に対して抽出回（$t=1, \cdots, 100$）ごとに算出したとき，これを$CID_{k,b,t}$と記す。併せて，共変量の組合せに伴うCIDの相違を評価するためにDFCID（Difference of CID）を用いる。これは，共変量を用いずに無情報（定数項のみの回帰）で算出した$\widetilde{\mathrm{CID}}$を基準とし，基準値からの絶対値の距離を求めたものである[10]。また，頑健性を捕捉するために最もバイアスの大きいSSB比率0.1に関するDFCIDを用いて，その抽出回tに関する平均値をMCIDとして算出する。

$$DFCID_{k,b,t} = |CID_{k,b,t}| - |\widetilde{CID}_{b,t}| \tag{17}$$

$$MCID_{k(b=0.1)} = \sum_t \frac{DFCID_{k,t(b=0.1)}}{100} \tag{18}$$

DFCIDまたはMCIDがゼロに近い場合には，無情報の場合と共変量を用いた場合とでCIDに差がみられないことを意味する。これがマイナスの値としてゼロから乖離する場合には共変量を用いたことによるCIDの改善が観測されたケースとなり，反対にプラスの値としてゼロから乖離する場合には共変量を用いたことによるCIDの悪化が観測されたケースを意味する。

⑵　バイアスとMSE

本検証では，2次活動時間の仮想母平均$\bar{Y}^*$の推定を目的とする。このとき，⑻式または⒀式を用いて，共変量の組合せk，SSB比率b，抽出回t別に目標統計量$\bar{Y}_{k,b,t}$の推定値が算出される。そこで，期待値，分散，およびバイアスは以下のように求まる。

$$\hat{E}(\hat{\bar{Y}}_{k,b}) = \frac{1}{T}\sum_t \bar{Y}_{k,b,t} \tag{19}$$

$$\hat{V}(\hat{\bar{Y}}_{k,b}) = \frac{1}{T}\sum_t (\bar{Y}_{k,b,t} - \hat{E}(\hat{\bar{Y}}_{k,b}))^2 \tag{20}$$

$$\widehat{\mathrm{Bias}}(\hat{\bar{Y}}_{k,b}) = \hat{E}(\hat{\bar{Y}}_{k,b}) - \bar{Y}^* \tag{21}$$

このとき，MSEは㉒式により算出される。

$$\widehat{\mathrm{MSE}}(\hat{\bar{Y}}_{k,b}) = \hat{V}(\hat{\bar{Y}}_{k,b}) + \widehat{\mathrm{Bias}}(\hat{\bar{Y}}_{k,b})^2 \tag{22}$$

本稿では，MSEを用いてサンプルセレクションバイアスと補正方法との関係を捉える。

なお，図2には，補正をしない時の推定値のバイアスをSSB比率別に求めた結果を示している。仮想母集団および仮想部分母集団のいずれの場合についても，SSB比率0.7付近でランダムサンプリングとなるよう設計しているため，SSB比率が0.7から乖離するに従いバイアスをもつ様子が確認できる。バイアスの程度を比較すると，女性の有業者グループ，女性グループ，女性の地方居住者グループの順に大きい。

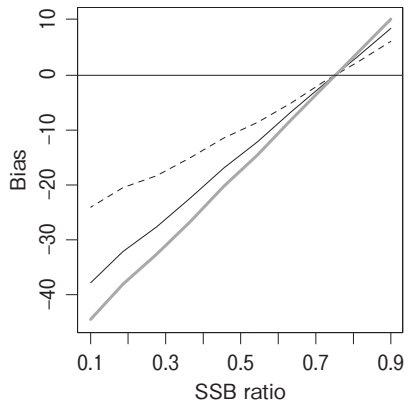

図２　SSB比率とバイアス傾向（補正なしの推定値）

（注）　実線が女性，破線が女性の有業者，太い灰色線が女性の地方居住者を示している。

３．検証結果

3.1　仮想母集団に関する結果

図３は，共変量の組合せ別に，SSB比率によってMSEとCIDがどのように推移するかを示したものである。まず，SSB調整変数であるX_1を共変量とした(a)の結果を確認すると，補正なしモデルではSSB比率に応じてMSEが高くなるのに対して，X_1を共変量としたとき，当然のことながら，いずれのモデルにおいてもMSEの改善が観測された（「改善」や「悪化」とは，補正なしモデルと比較してMSEが低いか高いかを意味する。以下同様）。ただし，改善の程度は補助標本のサイズにより異なり，仮想母集団サイズを補助標本として用いたIPW.Fullモデル，CLB.Marginalモデル，CLB.Crossモデルであれば，SSB比率に依らずMSEは低い水準を推移し，これら３つのモデルの間には大きな違いはみられない。

これに対して，補助標本サイズがSSB標本サイズと同じであるIPW.Smallモデルの場合には，MSEは改善するが仮想母集団サイズを利用した他のモデルほど改善の程度は大きくない。また，SSB調整変数X_1に加えて，X_2またはX_3を共変量とした場合の図３(d)，(f)，(g)についても，同様の傾向が示されている。これら共変量の組合せに関するCIDを確認すると，SSB比率のいずれのケースについてもほぼゼロ付近を推移している。

一方で，図３(c)のようにX_3のみを共変量とした場合，いずれのモデルでもMSEの改善はみられない。このとき，SSB比率が0.7から乖離するに従い，CIDもゼロから乖離する傾向がみられる。さらに，図３(b)または(e)のように，X_2を共変量とした場合には，MSEの悪化がみられ，補正に用いるべきではない共変量の存在が確認された。CIDをみると，SSB比率とともにゼロから大きく乖離する傾向が捉えられている。

3.2　仮想部分母集団に関する結果

仮想部分母集団を女性の地方居住者や有業者とした場合にも，仮想母集団（女性）に関する結果とほぼ同様の傾向が示されている。すなわち，SSB調整変数X_1を共変量として利用すれば（図４(a)，(c)，図５(a)，(c)），MSEは改善され，とくに仮想母集団サイズを用いたIPW.Fullモデル，CLB.Marginalモデル，およびCLB.Crossモデルにおいて改善の程度は大きい。これらのケースでは，CIDもゼロ付近を推移している。

これに対して，SSB調整変数を共変量に使用していない図４(b)と図５(b)を比較すると，仮想部分母集団による違いが表れている。女性の有業者に関する図５(b)ではMSEは改善も悪化もみられないが，女性の地方居住者に関する図４(b)ではMSEの悪化がみられる。CIDは，両方ともSSB比率に応じてゼロから乖離する傾向が示されているが，乖離傾向は女性の地方居住者に関する図４(b)のほうが大きい。

以上のことから，MSEの改善が期待できるのはCIDがゼロ付近に分布する場合に限るものと考えられる。CIDのゼロからの乖離が観測されるとき，乖離の程度が小さい場合にはMSEは不変であるが，乖離の程度が大きい場

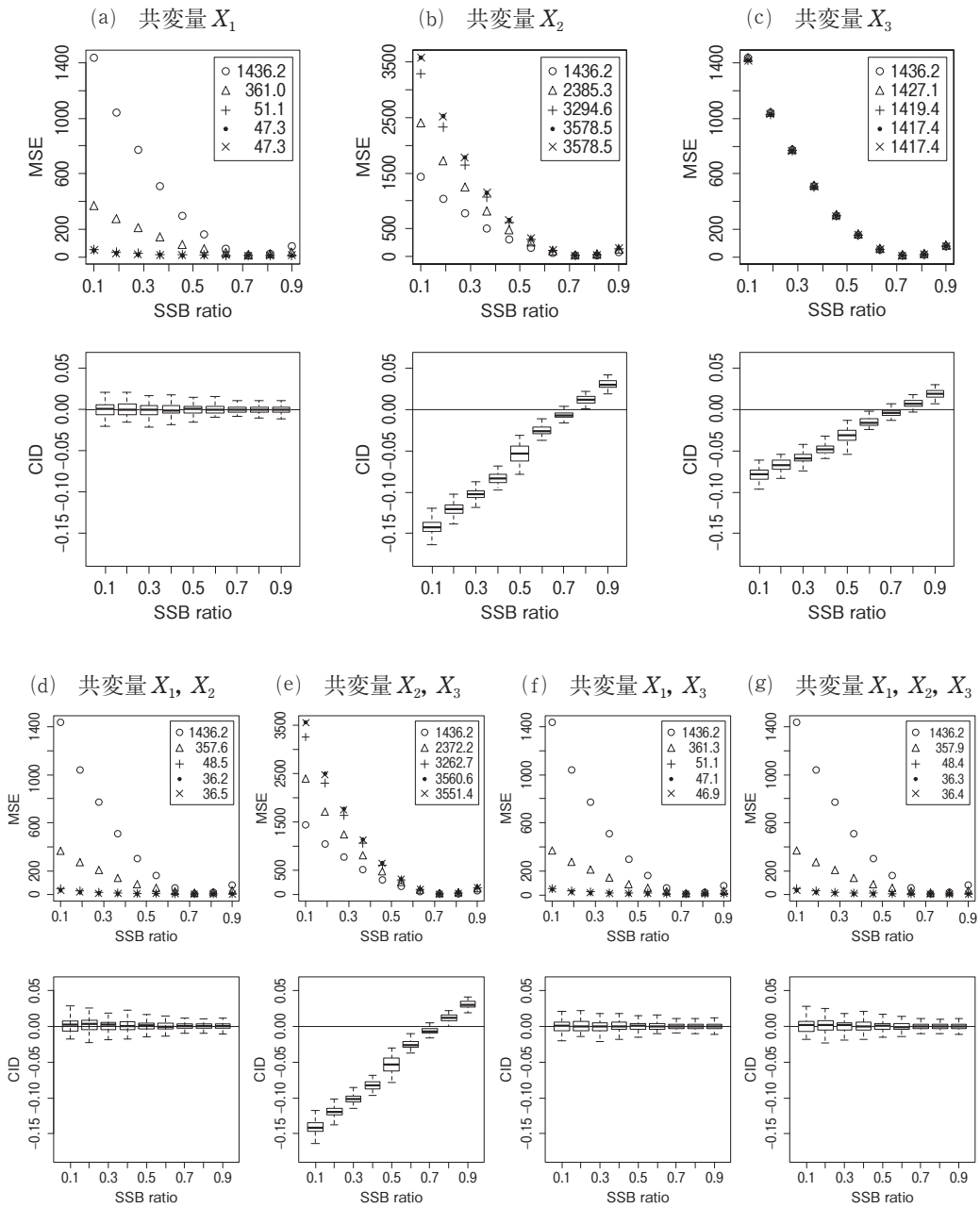

図３　共変量の組合せ別，MSEとCIDの分布（女性）

（注）　図中のマーカー，○，△，＋，●，×は，それぞれ補正なしモデル，IPW.Small，IPW.Full，CLB.Marginal，CLB.Crossによる推定結果を示している。また，凡例の数値はSSB比率0.1の場合のMSEの値である。

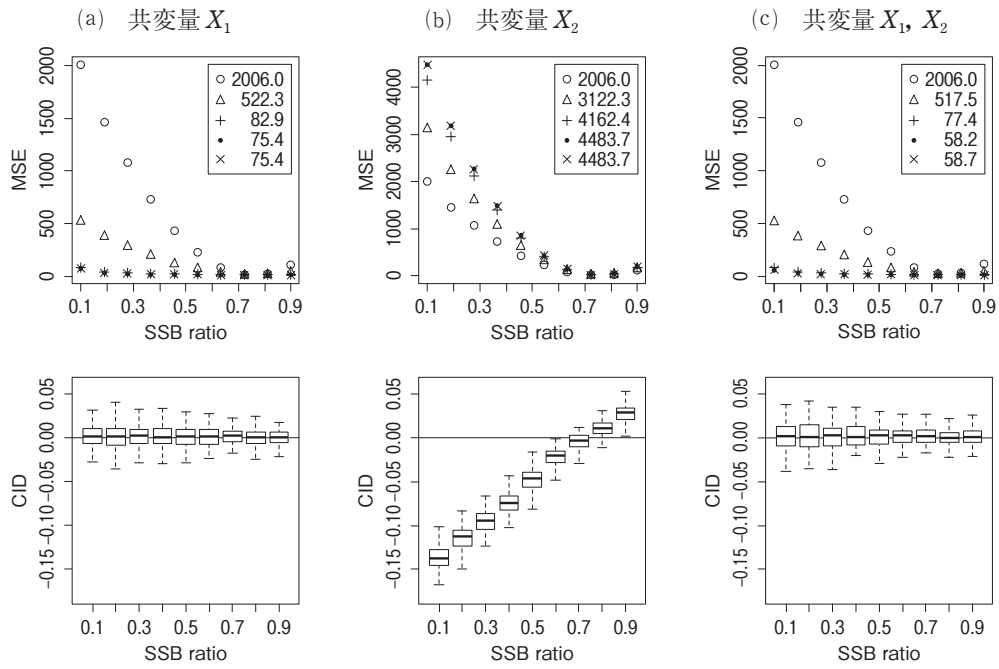

図4　共変量の組合せ別，MSEとCIDの分布（女性・地方居住者）

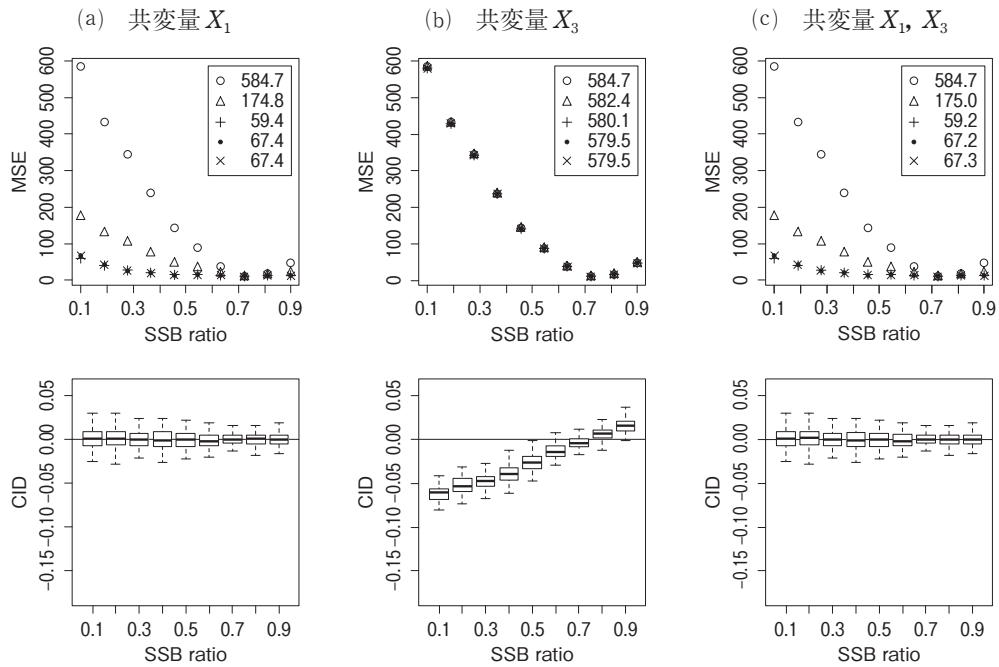

図5　共変量の組合せ別，MSEとCIDの分布（女性・有業者）

（注）　図中のマーカー，○，△，＋，●，×は，それぞれ補正なしモデル，IPW.Small，IPW.Full，CLB.Marginal，CLB.Crossによる推定結果を示している。また，凡例の数値はSSB比率0.1の場合のMSEの値である。

合にはMSEは悪化する可能性があることが推察される。次節では，これらCIDの大きさとMSEとの関係を詳細に検討していく。

3.3　共変量の選択

表7には，6つの共変量を用いたときの全ての組合せに関するMCIDの結果を整理している。2.5節で定義したようにMCIDとは，DFCID（基準$\widetilde{\mathrm{CID}}$と共変量適用時のCIDとの絶対値の差）について，SSB比率0.1における標本抽出回平均を求めたものである。表内の「※」には，MCIDの改善・不変・悪化の傾向を確認するために，代表例として選んだ共変量の組合せとMCIDの値を示しており[11]，図6，7，8には，それら代表例についてMSEとDFCIDの結果を示している。

まず，女性については，120通りの組合せの中で，MCIDの改善がみられたのは，SSB調整変数X_1を共変量として含む組合せ（63通り）である。これに対して，X_1を含まずX_2を含む組合せ（31通り）の場合，MCIDは悪化している。ただし，X_1とX_2を除いた共変量の組合せ（26通り）のときには，MCIDに大きな変化はみられなかった。代表例を示した図6からも，DFCIDがマイナスで分布しているときにはMSEは改善（図6(a)），DFCIDがプラスで分布しているときにはMSEは悪化しており（図6(c)），さらに，DFCIDがゼロ付近にある時にMSEは同水準にある様子が示されている（図6(b)）。以上から，共変量の組合せによってはMSEの不変や改善のみならず，悪化させるケースもあるため，共変量の選択は極めて重要な問題と考えられる。

女性の地方居住者を仮想部分母集団とした場合については，仮想母集団（女性）の結果とほぼ同様の傾向が示されている（表7，図7）。一方で，女性の有業者を仮想部分母集団としたとき，SSB調整変数を共変量に用いない場合であっても，図8(b)のようにX_4を含む共変量の組合せであれば，DFCIDの改善が観測された。これにより，SSB調整変数の代替変数として機能する共変量の存在も示唆された。

なお，図9には6変数すべてを共変量とした結果が示されている。3変数までを共変量とした結果では，キャリブレーションであれば周辺度数とクロス度数のいずれを用いてもMSEに大きな差はみられなかった。しかしながら，図9の凡例数字で確認できるように6変数を共変量とした場合，周辺度数を利用したMSEの方がクロス度数を利用したMSEよりも若干高めの水準にある。すなわち，利用する共変量は同じであっても，情報量（周辺

表7　全ての共変量の組合せに基づくMCIDの結果

分析対象	改善	不変	悪化
【女性】 6変数を用いた120通りの組合せ	［−0.072，−0.071］内の63通り（X_1を含む全ての組合せ） ※X_1（−0.072）	［−0.001，0.007］内の26通り（X_1とX_2を除く全ての組合せ） ※X_5（−0.001）	［0.032，0.064］内の31通り（X_1を除きX_2を含む全ての組合せ） ※X_2（0.064）
【女性・地方居住者】 X_3を除く57通りの組合せ	［−0.031，−0.029］内の31通り（X_1を含む全ての組合せ） ※X_1（−0.031）	［−0.001，0.017］内の11通り（X_1とX_2を除く全ての組合せ） ※X_5（−0.001）	［0.055，0.095］内の15通り（X_1を除きX_2を含む全ての組合せ） ※X_2（0.095）
【女性・有業者】 X_2を除く57通りの組合せ	［−0.052，−0.051］内の31通り（X_1を含む全ての組合せ） ※X_1（−0.052）	［−0.009，0.000］内の15通り（X_1を除き，X_4を含む組合せ，またはX_6を含まない全ての組合せ） ※$X_4 * X_5$（−0.009）	［0.009，0.010］内の11通り（X_1とX_4を除き，X_6を含む全ての組合せ） ※X_6（0.01）

（注）　MCIDは，SSB比率0.1に関するDFCIDの標本抽出回平均を示している。2つ以上の変数を用いる場合には，周辺度数とクロス度数の相違も考慮しているため，単なる組合せの総数ではない。

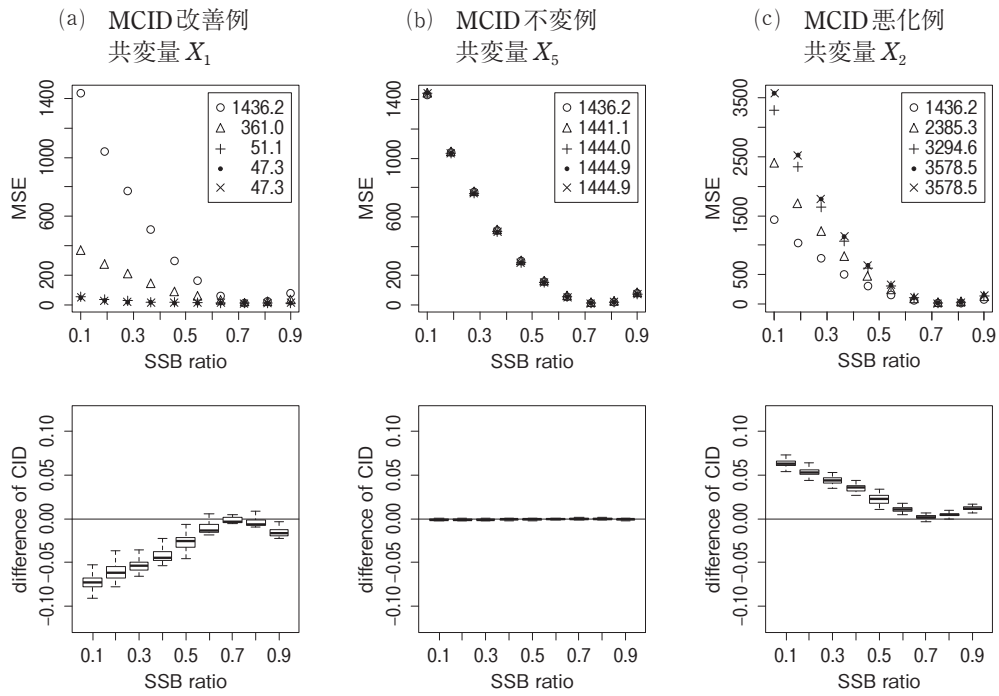

図6　MCIDケース別，MSEとCIDの分布（女性）

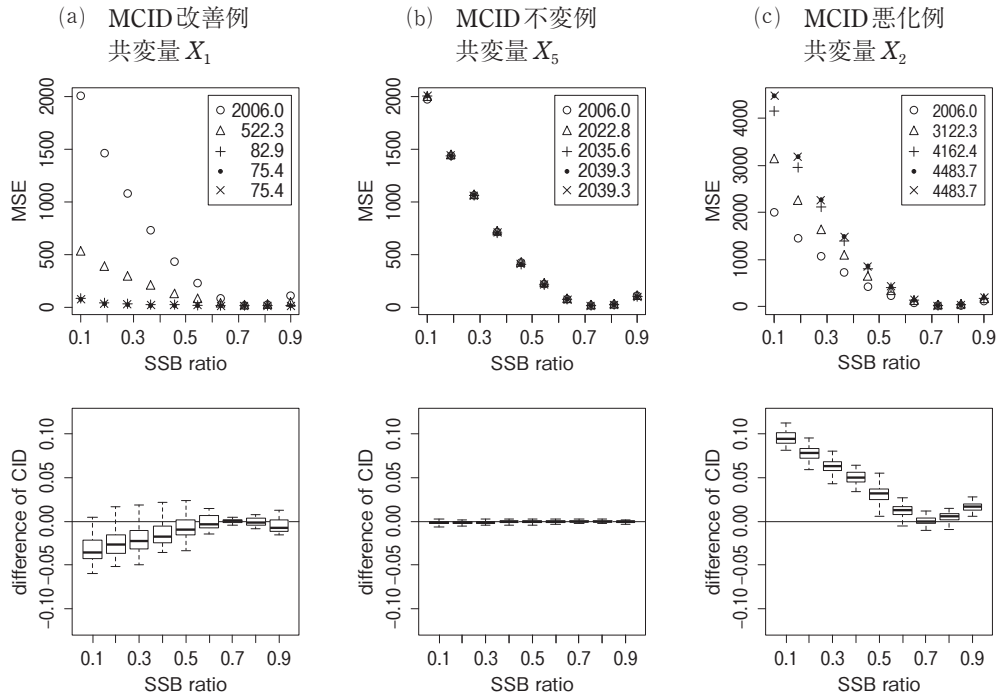

図7　MCIDケース別，MSEとCIDの分布（女性・地方居住者）

（注）　図中のマーカー，○，△，＋，●，×は，それぞれ補正なしモデル，IPW.Small，IPW.Full，CLB.Marginal，CLB.Crossによる推定結果を示している。また，凡例の数値はSSB比率0.1の場合のMSEの値である。

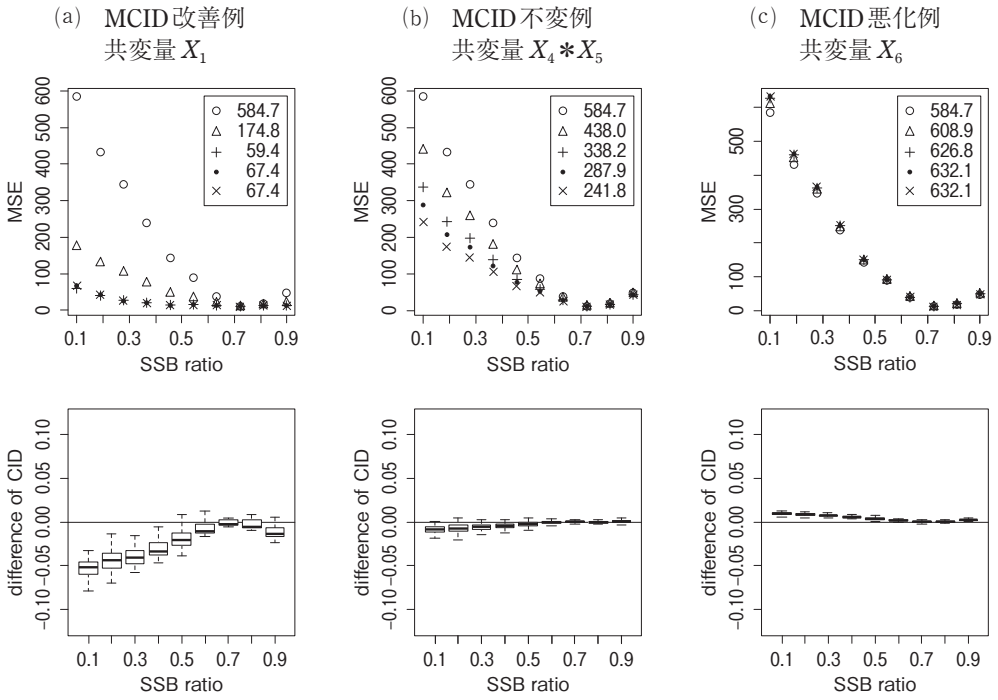

図８　MCIDケース別，MSEとCIDの分布（女性・有業者）

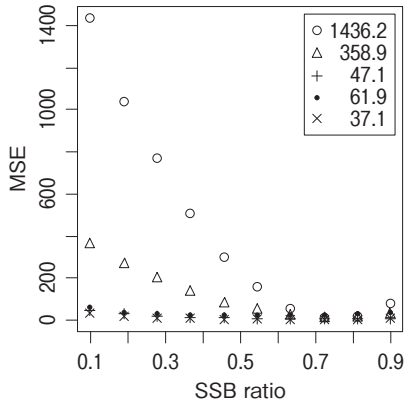

図９　６つの共変量を用いた結果（女性）

（注）　図８，図９ともに，マーカー，○，△，＋，●，×は，それぞれ補正なしモデル，IPW.Small，IPW.Full，CLB.Marginal，CLB.Crossによる推定結果を示している。また，凡例の数値はSSB比率0.1の場合のMSEの値である。

度数かクロス度数か）によって，推定精度が異なる可能性があることが推察される。

４．おわりに

本研究は，公的統計ミクロデータの利活用を目指して，公的統計をベースとして新規追加情報をWeb調査により捕捉する際のサンプルセレクションバイアスの補正方法について検討を行った。具体的には，社会生活基本調査のミクロデータを仮想母集団とし，標本抽出実験により，サンプルセレクションバイアスの補正方法とMSEの関係や，補正に用い

る共変量とCIDとの関係について明らかにした。

まず，補助標本のサイズは大きい（母集団サイズにより近い）ほうがMSEは改善するが，サイズが同じであればIPWとキャリブレーションのいずれでも結果に大きな差はみられなかった。IPWには個票が必要であるが，キャリブレーションでは共変量に関する母集団集計値があれば推定できることから，補正に用いるデータの利用条件によって補正方法の選択が可能である。

次に，キャリブレーションにおいて，共変量の周辺度数のみを利用する方法（レイキング比推定量）とクロス度数を利用する方法（一般化回帰推定量）とでは，いずれを用いても大きな差は観測されなかった。ただし，変数の組合せによっては，周辺度数による補正はその改善の度合いがやや低いケースがあったため，補助標本の利用条件として可能であれば，クロス集計値による補正がより適切と考えられる。

さらに，補正に使用する共変量の組合せとしては，バイアス発生要因，あるいはそれと類似した情報を有する変数を共変量として使用する場合にはMSEの改善は確認されたが，適切な共変量が使用されない場合にはMSEの悪化も観測された。実際のWeb調査においてはバイアス発生要因を特定することは困難であるため，少なくとも条件付き従属性を捉えたCIDのような指標を用いて，改善の可能性のある共変量の組合せであることを確認することが不可欠である。

近年，Web調査は，迅速かつ適時に問題関心である社会実態を観測する有力な調査手段を与えているが，他方で常にサンプルセレクションバイアス発生の問題がつきまとう。しかしながら，既存の公的統計調査をふまえてWeb調査の設計を行うことで，補正に必要となる条件付き独立性の成否の確認や共変量の入手は可能となる。

本稿では，条件付き独立性を確認するための指標としてCIDに関する基準値からの差を用いて一定の傾向を捉えたが，この指標に関する使用条件の一般化などについては，より詳細な検討が不可欠である。また，公的統計の調査設計を拡張的に利用するために，Web調査の設計をどのように立案すれば，より効率的で精度の高い結果が得られるかに関しても，詳細な検討が求められる。これらについては，今後の課題としたい。

謝辞

本研究はJSPS科研費（課題番号16K20894）の助成を受けたものです。また，本分析は，一橋大学経済研究所附属社会科学情報研究センターから社会生活基本調査（平成18年度分）の匿名データの提供を受けたものです。本分析結果は，総務省が公表する統計とは関係ありません。

注

1）傾向スコアの基本概念についてはRosenbaum, P.R. and Rubin, D.B.（1983）などを参照のこと。

2）2次活動とは，「仕事や家事など社会生活を営む上で義務的な性格の強い活動」（総務省統計局の社会生活基本調査より引用）に分類されるものであり，実際には「通勤・通学」，「仕事」，「学業」，「家事」，「介護・看護」，「育児」，「買い物」が2次活動に分類される。

3）本研究では，分析者が調査設計者となりえるWeb調査の特性を踏まえて，バイアス発生要因を，目標変数ではなく共変量（説明変数）に割り当てたケース（外生的標本設計）を前提としている。すなわち，ランダムな欠測（MAR：Missing at random）を有するデータセットを作り出して抽出実験を

行っている。ランダムな欠測については，星野（2010：27-29）または岩崎（2002：182-206）を参照のこと。

4）クラーメルのV算出の際には，Rのvcdパッケージassocstats関数を用いている。

5）社会生活基本調査の標本設計では，層化2段により世帯単位で抽出しているが，本稿で用いた匿名データには実際の抽出に係る情報が付与されていないことから層化などは行わずに，全体の抽出率を0.124としてSSB比率に基づいて個人単位で抽出を行っている。なお，社会生活基本調査ミクロデータに固有の標本設計を踏まえたウェイト補正に関する先行研究としては，栗原（2010）および栗原・坂田（2014）が挙げられる。

6）星野（2010：69）を参考にIPW計算式を整理し，また星野（2010：229）に掲載されている統計ソフトRのコードを参考に推定している。

7）レイキング比推定は，事後的に母集団の共変量に関する周辺度数と一致するようにウェイトを用いて補正を行う方法である。これによれば，母集団に関する詳細なクロス集計表が入手できない場合でも，周辺情報までは補正できる。レイキングの推定法はDeming, W.E. and Stephan, F.F.（1940）によりIterative Proportional Fitting法が提示され，現在では，キャリブレーションの枠組みで乗法関数を用いたキャリブレーションウェイトにより推定することができる。

8）一般化回帰推定量は，共変量を用いて得られた回帰係数を用いて補正を行う方法である。

9）統計ソフトRのsurvey packagesの関数calibrateとsvyglmを利用している。キャリブレーションウェイト計算時にはウェイトの値の範囲を設けていないが，実際の計算結果としてマイナスの値になったり，極端に大きな値や小さな値になる事例はないことは確認済みである。

10）CIDはゼロに近いほどCIAの成立が期待できるため，ゼロからの距離として基準値からの乖離を計測するために，絶対値の差を求めている。

11）代表例の選択基準としては，SSB比率が最も低い0.1のケースについて，極端に数値が変化する箇所で3区分し，改善区分と不変区分ではMCIDが最小値となる組合せ，悪化区分ではMCIDが最大値となる組合せを用いている。なお，MCIDの値は四捨五入による小数第3位までを用いており，MCIDの最大値または最小値が複数ある場合には，最も共変量の数が少ない組合せを代表例として採用した。

参考文献

［1］ 岩崎学（2002）『不完全データの統計解析』エコノミスト社.

［2］ 栗原由紀子（2010）「社会生活基本調査ミクロデータにおける平日平均統計量と標本誤差の計測」『統計学』（経済統計学会）第99号，pp.20-35.

［3］ 栗原由紀子（2015）「統計的マッチングにおける推定精度とキー変数選択の効果 ― 法人企業統計調査ミクロデータを対象として ― 」『統計学』（経済統計学会）第108号，pp.1-15.

［4］ 栗原由紀子・坂田幸繁（2014）「ミクロデータ分析における調査ウェイトの補正効果 ― 社会生活基本調査・匿名データの利用に向けて ― 」『弘前大学人文学部人文社会論叢（社会科学編）』（弘前大学人文学部）第31号，pp.93-113.

［5］ 土屋隆裕（2009）『概説　標本調査法』朝倉書店.

［6］ 星野崇宏（2010）『調査観察データの統計科学』岩波書店.

［7］ 星野崇宏・前田忠彦（2006）「傾向スコアを用いた補正法の有意抽出による標本調査への応用と共変量の選択法の提案」,『統計数理』，第54巻第1号，pp.191-206.

［8］ Deming, W.E. and Stephan, F.F.（1940）, “On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known”, *The Annals of Mathematical Statistics*, 11, pp.427-444.

［9］ Rosenbaum, P.R. and Rubin, D.B.（1983）, “The central role of the propensity score in observational studies for causal effects”, *Biometrika*, 70, Issue 1, pp.41-55.

［10］ Valliant, R., J.A. Dever, and F. Kreuter（2013）, *Practical Tools for Designing and Weighting Survey Samples*, Springer.

Verification of the Adjustment Methods for Sample Selection Bias Using Microdata of the Survey on Time Use and Leisure Activities

Yukiko KURIHARA*

Summary

To promote the utilization of official statistics microdata, this research aims to verify the adjustment methods for sample selection bias using a sampling experiment in which the microdata of the Survey on Time Use and Leisure Activities are used as the virtual population data set. The three major results of the study are follows: First, the difference of MSE is not observed between the propensity score and calibration if the auxiliary sample sizes are the same. Second, when using the calibration method, the MSEs do not significantly differ from the usage of the marginal frequency (raking estimator) or the cross frequency (generalized regression estimator). Third, if the causal variables of the sample selection bias cannot be identified, the deterioration of MSEs is observed in several combinations of covariates so that confirming the establishment of the conditional independent assumption is necessary before the bias adjustment utilizing covariates.

Key Words

Propensity score, Calibration, Sample selection bias, Conditional independent assumption

* College of Economics, Ritsumeikan University

編集委員会からのお知らせ
機関誌『統計学』の編集・発行について

編集委員会

機関誌『統計学』への投稿を募集しています。

1．原稿は編集委員長宛に送付して下さい（下記メールアドレス）。

2．投稿は常時受け付けています。

なお，書評，資料および海外統計事情等の分類の記事については調整が必要になることもありますので念のため事前に編集委員長に照会して下さるようお願いします。

3．次号以降の発行予定日は，

第118号：2020年3月31日，第119号：2020年9月30日です。

なお，投稿から掲載が決まるまでに要する期間は通常3ヶ月以上を要します。

4．原則として，すべての投稿原稿が審査の対象となります。投稿に際しては，「投稿規程」，「執筆要綱」，および「査読要領」の確認をお願いします。最新版は，本学会の公式ウェブサイト（http://www.jsest.jp/）を参照して下さい。

投稿，編集委員会についての問い合わせや執筆の推薦その他とも，下記編集委員長のメールアドレス宛に送付して下さい。

editorial@jsest.jp

以上

編集後記

投稿していただきました執筆者のみなさま，そしてお忙しい中快く論文の審査をお引き受けいただきました査読者のみなさまに改めてお礼申し上げます。また，『統計学』創刊60周年記念事業委員会は本誌第112号に続き特集の編集ありがとうございました。（池田伸　記）

執 筆 者 紹 介

栗原由紀子　（立命館大学経済学部）　　平井太規　（神戸学院大学現代社会学部）
西村善博　　（大分大学経済学部）　　村上雅俊　（阪南大学経済学部）

支部名	事務局		
北　海　道 ……	062-8605	札幌市豊平区旭町 4-1-40 北海学園大学経済学部 （011-841-1161）	水野谷武志
東北・関東 ……	192-0393	八王子市東中野 742-1 中央大学経済学部 （042-674-3406）	伊藤伸介
関　　西 ……	640-8510	和歌山市栄谷 930 和歌山大学観光学部 （073-457-8557）	大井達雄
九　　州 ……	870-1192	大分市大字旦野原 700 大分大学経済学部 （097-554-7706）	西村善博

『統計学』編集委員

委 員 長　池田　伸（関西，立命館大学）
副委員長　小林良行（東北・関東，総務省統計研究研修所）
委　　員　水野谷武志（北海道，北海学園大学），山田　満（東北・関東），
　　松川太一郎（九州，鹿児島大学）

『統計学』60周年記念事業委員会

委 員 長　大井達雄（和歌山大学）
副委員長　水野谷武志（北海学園大学）
委　　員　池田　伸（立命館大学），伊藤伸介（中央大学），
　　杉橋やよい（専修大学），村上雅俊（阪南大学），
　　金子治平（会長，神戸大学），上藤一郎（常任理事長，静岡大学）

統　計　学　No.117

2019年9月30日　発行

発 行 所　経　済　統　計　学　会
〒112-0013　東京都文京区音羽1-6-9
音羽リスマチック株式会社
TEL/FAX　03（3945）3227
E-mail：office@jsest.jp
http://www.jsest.jp/

発 行 人　代表者　金　子　治　平

発 売 所　音羽リスマチック株式会社
〒112-0013　東京都文京区音羽1-6-9
TEL/FAX　03（3945）3227
E-mail：otorisu@jupiter.ocn.ne.jp
代表者　遠　藤　誠

昭和情報プロセス㈱印刷

©経済統計学会

Statistics

No. 117 2019 September

Japan Society of Economic Statistics