【研究論文】（『統計学』第119号 2020年9月）

傾向スコアマッチングの適用による新たな多変量型の比率代入法：全国消費実態調査の匿名データを用いた検証

高橋将宜*

要旨

公的経済統計における欠測値の処理方法を改善することは，重要な学術的かつ社会的課題である。これまで，わが国および諸外国における公的経済統計では，欠測値を処理する方法として比率代入法（ratio imputation）を頻繁に活用してきた。しかし，比率代入モデルは切片なしの単回帰モデルであるため，データ内に多数の共変量があっても，それらの情報を活用できない。そこで，本研究では，傾向スコア（propensity score）を活用することで，多数の共変量のバランシングを行い，マッチングをすることによって比率代入法の精度向上を目指す。全国消費実態調査の匿名データを使ったサブサンプリングによるシミュレーションやモンテカルロ・シミュレーションを通じて，提案手法と伝統的な手法との優劣を比較検証する。

キーワード

欠測データ，比率代入法，傾向スコアマッチング，公的統計，匿名データ

1．はじめに

欠測データの研究には，学術的に重要な意義があるだけではなく，公的統計の実務にも応用可能なものとして社会貢献できるという点で，実際的に重要な意義もある。2020年6月2日には，「公的統計の整備に関する基本的な計画」の変更点が閣議決定され，「法人企業統計調査における欠測値の補完方法の改善」や「事業所・企業や各種法人等に係る統計調査を実施するに当たり，欠測値の補完や集計の充実等を検討」することとされている[1)]。本研究の目的は，任意の経済データに欠測が発生しているとき，変数の合計値や平均値を集計するための適切な方法を学術的に追究することである。その結果として，本研究の成果により，公的統計における欠測値処理の実務の改善にも寄与するものである。

諸外国においても，わが国においても，公的経済統計における欠測値処理方法として，比率代入法（ratio imputation）がよく用いられる（高橋，2017）。しかしながら，比率代入法は切片なしの単回帰モデルであるため，データ内に多数の共変量があっても，それらの情報を活用できない。本研究では，比率代入モデルに組み込むことのできない共変量の集合 X の情報を，傾向スコアマッチング（propensity score matching）を用いて活用する新たな手法を提案する。傾向スコアによって多数の共変量のバランシングを行ってマッチングをした上で，比率代入法を用いることで，精度の向上を目指す。

一般的に，傾向スコアは，「処置の割付けを表すダミー変数 Z および観測された共変量

* 正会員，長崎大学情報データ科学部
e-mail：m-takahashi@nagasaki-u.ac.jp

の集合Xに対し，Xが与えられたときに個体が処置に割付けられる確率$e(X)=P(Z=1|X)$」（岩崎，2015，p.96）として定義される。本研究では，「処置の割付」を「欠測の状態」と置き換えることで，傾向スコアを活用する。すなわち，本研究における傾向スコアは，欠測を表すダミー変数Zおよび観測された共変量の集合Xに対し，Xが与えられたときに個体が欠測する確率$e(X)=P(Z=1|X)$である。傾向スコアについては，「特にXの次元が大きい場合には，その情報が1次元の$e(X)$に集約されるため，実際のデータ解析上きわめて有用である」（岩崎，2015，p.97）ことが指摘されており，単回帰モデルとしての比率代入法の欠点を補うことができると期待される。

本研究では，全国消費実態調査の匿名データ[2)]を使って，サブサンプリングによるシミュレーションを行い，提案手法を伝統的な手法と比較検証する。また，手法の優劣に関する一般性を担保するために，モンテカルロ・シミュレーションによるエビデンスも示す。よって，本研究の成果は，全国消費実態調査の欠測値だけではなく，経済センサス－活動調査の欠測値処理など，不均一分散の兆候を示す公的経済統計全般に幅広く応用が可能である。ただし，経済センサス－活動調査のミクロデータ自体は利用できないため，本稿では，実データとして全国消費実態調査（2004年）の匿名データを用いる。

本稿の構成は以下のとおりである。2節では，分析に用いた全国消費実態調査の匿名データの特徴を解説する。3節では，従来より公的経済統計の欠測値処理に使用されている比率代入法の長所と短所について論じる。4節では，傾向スコアを応用することで比率代入法に多変量の情報を組み入れる方法を提案する。5節では，全国消費実態調査の匿名データを用いたノンパラメトリックなサブサンプリング分析により，欠測値処理手法の優劣を比較検証する。6節では，パラメトリックなモンテカルロ・シミュレーションを実行することで，各手法の優劣に関するエビデンスを補強する。7節において，本研究での知見をまとめる。

2．全国消費実態調査の匿名データ

本稿では，全国消費実態調査（2004年）の匿名データを用いた。この匿名データは，二人以上世帯（約4.4万レコード）と単身世帯（約0.4万レコード）に分けて提供されている。本研究では，単身世帯は標本サイズが小さく後述するサブサンプリングによる分析に適さないため対象とせず，二人以上世帯を対象とした。この節では，全国消費実態調査（2004年）の匿名データについて，本研究において重要となる点について言及する。

2.1　全国消費実態調査の匿名データにおける欠測値とその処理方法

本稿における主要な変数は，年間収入（V0399）である。この変数は「年収・貯蓄等調査票」により調査した年間収入に基づいている。また，全国消費実態調査の匿名データには，「調査票等の有無_年収票_不詳_年間収入」（V0009）という変数があり，ここで「1＝年間収入不詳あり」，「0＝年間収入不詳なし」，「ブランク＝年収票無し」を表している。年間収入が不詳の世帯について，全国消費実態調査では，「世帯主の職業，消費支出額，世帯主の年齢，有業人員により年間収入を推計」（総務省統計局，2004）している[3)]。つまり，推計式は，(2.1)式のとおり，重回帰モデルと考えられる。

$$\widehat{\text{年間収入}}_i=\hat{\beta}_0+\hat{\beta}_1\text{世帯主の職業}_i+\hat{\beta}_2\text{消費支出}_i+\hat{\beta}_3\text{世帯主の年齢}_i+\hat{\beta}_4\text{有業人員}_i \quad (2.1)$$

一般的に，(2.1)式のβ_jは，通常の最小二乗法（OLS：Ordinary Least Squares）によって推定されることが多い。ガウス・マルコフの

仮定が満たされるとき，通常の最小二乗法による$\hat{\beta}_j$は最良線形不偏推定量（BLUE：Best Linear Unbiased Estimator）である。なお，ガウス・マルコフの仮定とは，パラメータに関する線形性，無作為抽出，完全な多重共線性がないこと，誤差項の条件付き期待値＝0，均一分散である（Wooldridge, 2009, pp.84-94）。この点については，3節で重要となる。

2.2　本研究で使用した変数

本研究で使用した変数の一覧は，表1に示すとおりである。これらの変数は，(2.1)式に登場している変数である。年間収入の平均値を集計対象とし，特に，年間収入に欠測が発生している状況において，どのようにして年間収入の平均値を集計できるかを議論する。実際の調査においても，「年齢階級別の年間収入の平均値」や「貯蓄年収比＝年間収入の平均値に対する貯蓄現在高の比率」などを分析しており，年間収入の平均値を適切に集計することは重要である[4)]。

就業人員（V0018）は，匿名データでは0～6までの値で記録されているが，すべてに1を足した。結果として，0→1人を意味し，6→7人を意味するため，理解しやすい。また，このようにすることで，対数変換しても問題は起こらなくなる。なお，匿名データでは，秘匿の目的で，8人以上の世帯は削除されている。就業人員は，間隔尺度の量的変数である。なお，就業人員と有業人員（V0391）は同じデータである。

年齢5歳階級（V0042）は，1～18までのカテゴリを持つ変数で，1は0～4歳，2は5～9歳を意味し，以後，5歳ずつの均等な間隔で，17は80～84歳，18は85歳以上を意味している。よって，年齢5歳階級は間隔尺度の量的変数である。

職業符号（V0048）は，表2のとおり，13種

表1　変数の一覧

変数番号	変数名	平均値	標準偏差	第1四分位数	中央値	第3四分位数
V0018	就業人員	2.421	0.967	2.000	2.000	3.000
V0042	年齢5歳階級	11.214	2.799	9.00	11.00	13.00
V0048	職業符号	NA	NA	NA	NA	NA
V0399	年間収入	656.978	364.523	397.000	580.000	838.000
V0455	消費支出	31.590	19.290	19.921	27.222	37.473

注：NA＝該当なし（質的データのため）

表2　職業符号（二人以上世帯）

職業符号	職業	標本サイズ	使用○/×
1	常用労務作業者	9322	○
2	臨時及び日々雇労務作業者	247	×
3	民間職員	11069	○
4	官公職員1	947	○
5	官公職員2	3984	○
6	商人及び職人	4375	○
7	個人経営者	463	×
8	農林漁業従業者	1623	×
9	法人経営者	1312	×
10	自由業者	578	×
11	その他	86	×
12	無職	9852	○
13	家族従業者	3	×

注：○は分析に使用した職業符号を表し，×は分析に使用しなかった職業符号を表す。

類の職業からなる1～13までのカテゴリを持つ変数で，名義尺度の質的変数である。ただし，カテゴリ2，7，8，9，10，11，13は観測数が少なく，5節のサブサンプリングによる分析に適さないため，本稿では用いない。また，カテゴリ4の観測数は947で少ないため，カテゴリ4（官公職員1）とカテゴリ5（官公職員2）を統合して1つのカテゴリとした。

年間収入（V0399）は，比率尺度の量的変数であり，単位は万円/年である。なお，本稿の研究目的は欠測値の処理であるため，2.1項の情報から考えて，年間収入の欠測値を代入処理したと考えられる1753個の値は分析から除外している。本研究の目的は，2004年の調査において欠測値を代入処理した値自体を復元することではないためである。

消費支出（V0455）は，比率尺度の量的変数である。もともとの匿名データでは，単位は円/月であるが，年間収入と単位を揃えるため，万円/月にコーディングし直した。

なお，もともとの匿名データの観測数は43,861であるが，職業符号の統廃合と年間収入の欠測値の代入値を除去したため，本研究で使用したデータの観測数は36,936である。

主な変数の分布は，図1のヒストグラムのとおりである。データの秘匿という観点から，ヒストグラムの軸は意図的に表示していないが，年間収入，消費支出，就業人員は右にすその長い分布をしていることが分かる。

年間収入に欠測が発生しているときに，消費支出の観測データから年間収入の値を予測したい。そこで，図2は年間収入と消費支出の関係を示した散布図である。ここから，年間収入と消費支出は正の相関関係があることが見て取れるが，同時に，消費支出が大きくなればなるほど年間収入の分散が大きくなっており，不均一分散の兆候も見て取れる。実際に，年間収入を被説明変数とし，消費支出

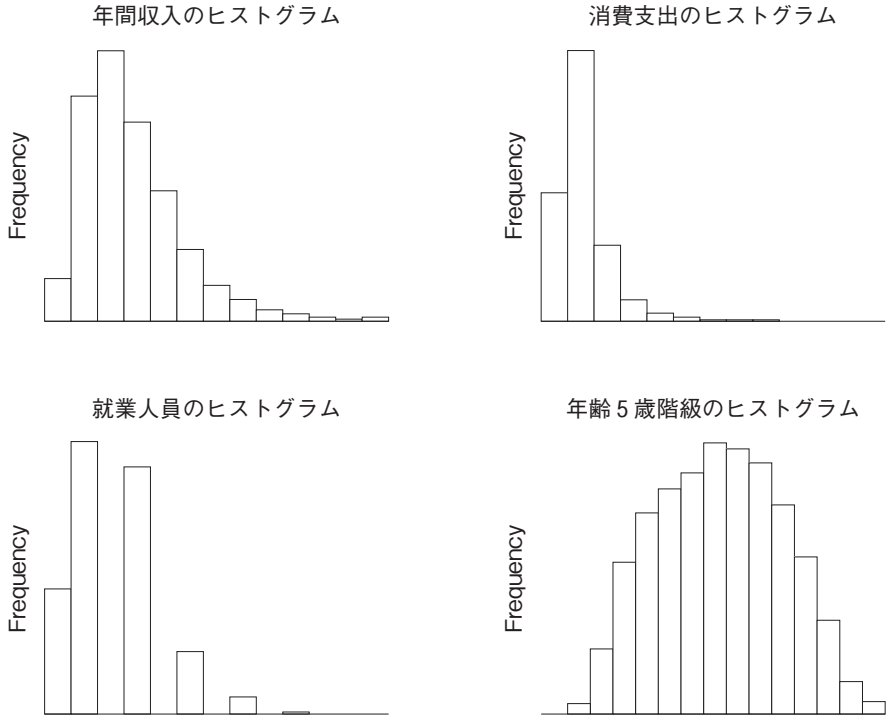

図1　生データの分布

注：データの秘匿という観点から，ヒストグラムの軸は意図的に表示していない。

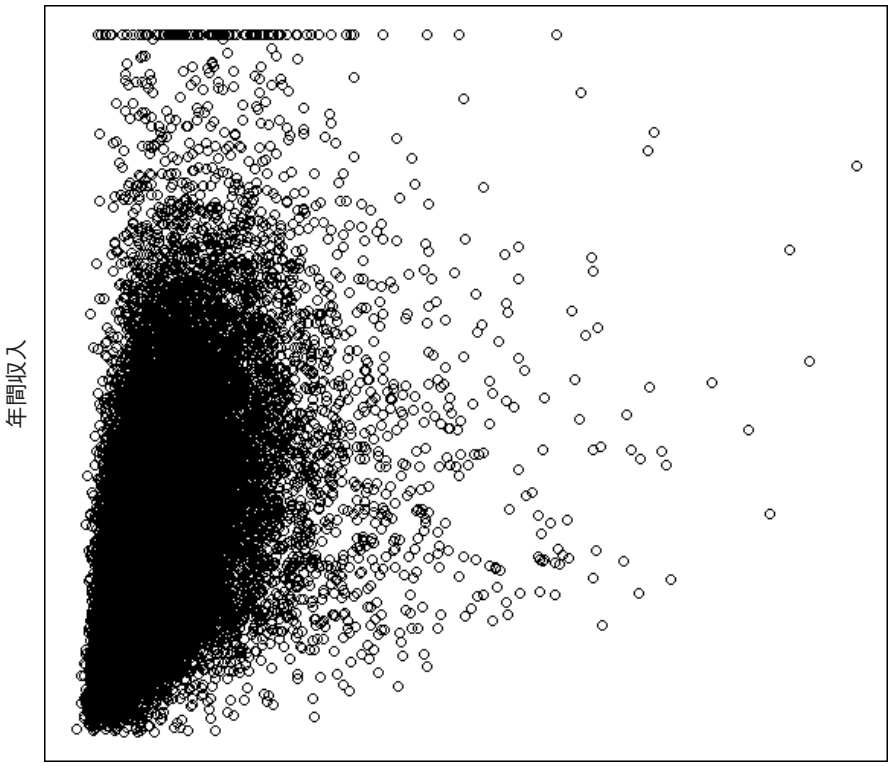

図２　年間収入と消費支出の散布図

注：データの秘匿という観点から，散布図の軸は意図的に表示していない。

を説明変数として回帰モデルを作り，ブリューシュ・ペイガン（Breusch-Pagan）検定[5]を実行すると，BP＝2859.3，df＝1であり，帰無仮説「誤差項の分散は均一である」を５％の有意水準で棄却する（$p=0.000$）。つまり，誤差項の分散は不均一と考えられる。この点は，欠測値を代入処理するモデリングにおいて重要になるため，３節において詳述する。

なお，匿名データには秘匿処理が施されている。リサンプリング，識別情報の削除，特異なレコードの削除，トップコーディングとボトムコーディング，リコーディングである[6]。中でも，トップコーディングは極端に大きな値に関して上限値を設けて，上限値よりも上の値を切断している。このようにトップコーディングされている場合，真値は上限値以上のどこかに存在することだけは分かっている。年齢について，85歳以上は85歳にトップコーディングされており，203個の観測値が該当する。年間収入について，2500万円以上は2500万円にトップコーディングされており，98個の観測値が該当する。

３．比率代入法の長所と短所

$D=(Y_i, X_{ji})$ を考える（$i=1,\cdots,n$；$j=1,\cdots,p-1$）。つまり，n 個の観測数，p 個の変数からなるデータ D である。特に，Y_i を被説明変数とし，X_{ji} を $p-1$ 個の説明変数としよう。具体的には，(2.1)式と同様に，$p=5$ 個の重回帰モデル(3.1)式を考える。

$$Y_i=\beta_0+\beta_1X_{1i}+\beta_2X_{2i}+\beta_3X_{3i}+\beta_4X_{4i}+\varepsilon_i \quad (3.1)$$

(2.1)式と同様に，(3.1)式の β_j は，通常の最小二乗法（OLS）によって推定されることが

多いが，均一分散の仮定が満たされないとき，通常の最小二乗法（OLS）による$\hat{\beta}_j$は最良線形不偏推定量（BLUE）ではない。

年間収入や消費支出などの経済データの分布は，図1で見たとおり，右にすそが長い。このような変数では，図2で見たとおり，回帰モデルを構築した場合，誤差項の分散が不均一となることが多い。つまり，誤差項εの期待値は0だが，分散は$\sigma^2X_i^{2\theta}$といった具合にX_iの値に比例して不均一である。

このような不均一分散が疑われる変数に欠測が発生している場合，諸外国の公的経済統計では，重回帰モデルではなく，比率代入法（ratio imputation）を用いることによって，経済データの欠測値に対処していることが知られている（de Waal et al., 2011, p.244；高橋, 2017）。比率代入法とは，(3.2)式のように，切片なしの単回帰モデルの構造をしている。その長所は，誤差項ε_iの分散が不均一の場合でも，最良線形不偏推定量（BLUE）となる好ましい性質を持っている。具体的には，重み付き最小二乗法（WLS：Weighted Least Squares）のフレームワークを使うと，$\theta=0.0$のとき(3.3)式の通常の最小二乗法による$\hat{\beta}_1$がBLUEであり，$\theta=0.5$のとき(3.4)式の平均値の比率による$\hat{\beta}_1$がBLUEであり，$\theta=1.0$のとき(3.5)式の比率の平均値による$\hat{\beta}_1$がBLUEである（Takahashi et al., 2017）。また，これらは最尤推定量（MLE：Maximum Likelihood Estimator）である（Little & Rubin, 2020, p.149）。なお，Σを取る範囲は$i=1, \cdots, n$である。

$$Y_i=\beta_1X_i+\varepsilon_i$$
$$\varepsilon_i\sim N(0, \sigma^2X_i^{2\theta}) \tag{3.2}$$

$$\hat{\beta}_1=\frac{\Sigma X_iY_i}{\Sigma X_i^2} \tag{3.3}$$

$$\hat{\beta}_1=\frac{\Sigma Y_i/n}{\Sigma X_i/n}=\frac{\bar{Y}}{\bar{X}} \tag{3.4}$$

$$\hat{\beta}_1=\frac{1}{n}\sum\frac{Y_i}{X_i} \tag{3.5}$$

したがって，年間収入の合計値の算出を目的とした場合，重回帰式のβは，不均一分散の影響により，BLUEではない。よって，比率代入法を用いる必要がある。

一方，比率代入法の最大の短所として，このモデルは(3.6)式のような単回帰モデルの形しか取ることができないことが挙げられる。つまり，(2.1)式のように他の共変量（世帯主の職業，世帯主の年齢，有業人員）の情報がデータ内にあったとしても，それらの情報を十分に活用することができない。

$$\widehat{\text{年間収入}}_i=\hat{\beta}_1\text{消費支出}_i \tag{3.6}$$

4．本研究の提案手法

前節で見たとおり，経済データの特性を考えると，比率代入法は好ましい特性を持っているものの，単回帰モデルの形でしか使用できないという構造的な制約がある。そこで，この比率代入法の欠点を克服するために，本研究では，比率代入モデルに組み込むことのできない共変量の集合Xの情報を，傾向スコアマッチングを用いて活用することを提案する。

前節から引き続き，$D=(Y_i, X_{ji})$を考える（$i=1, \cdots, n$；$j=1, \cdots, p-1$）。つまり，表3のとおり，観測数がn個であり，p個の変数からなるデータDである。特に，Y_iに欠測が発生しており，X_{ji}を$p-1$個の変数としよう。太枠で囲んだ部分が，$n\times(p-1)$の共変量の集合Xである。すなわち，表3において，太枠で囲んだ$p-1$個の変数からなる情報を傾向スコアで縮約して活用する[7)]。

一般的に，傾向スコアは，「処置の割付けを表すダミー変数Zおよび観測された共変量の集合Xに対し，Xが与えられたときに個体が処置に割付けられる確率$e(X)=P(Z=1|X)$」（岩崎，2015，p.96）と定義される。本研究では，この定義を以下のように拡張する。「処置の割付」を「欠測の状態」と置き換える

表 3　データ D（シミュレーションによる乱数）

世帯番号	Y_i	$X_{1,i}$	$X_{2,i}$	…	$X_{p-1,i}$
1	463	44	2	…	13
2	850	35	4	…	11
3	1006	50	3	…	10
4	538	36	2	…	8
5		49	3	…	10
⋮	⋮	⋮	⋮	⋱	⋮
$n-1$	193	22	1	…	17
n	454	20	3	…	12

注：空欄はデータが欠測していることを表す。添字 i は世帯番号を表す。

表 4　データの具体例（シミュレーションによる乱数）

世帯番号	年間収入$_i$	欠測指示行列$_i$	消費支出$_i$	年齢 5 歳階級$_i$
1	463	0	44	13
2	850	0	35	11
3	1006	0	50	10
4	538	0	36	8
5		1	49	10

注：空欄はデータが欠測していることを表す。添字 i は世帯番号を表す。

ことで，傾向スコアを活用する。つまり，本研究における傾向スコアは，欠測を表すダミー変数 Z および観測された共変量の集合 X に対し，X が与えられたときに個体が欠測する確率 $\mathrm{e}(X)=\mathrm{P}(Z=1|X)$ である（阿部，2016，p.103）。

このようにして傾向スコアを用いる利点としては，「特に X の次元が大きい場合には，その情報が 1 次元の $\mathrm{e}(X)$ に集約されるため，実際のデータ解析上きわめて有用である」（岩崎，2015，p.97）ことが指摘されており，単回帰モデルとしての比率代入法の欠点を補うことができると期待される。ただし，傾向スコアを用いる際には，「処置によって影響を受けた変数はモデルに含めるべきではないというコンセンサス」（岩崎，2015，p.105）があり，比率代入法の対象となる欠測変数 Y_i 自体は，傾向スコアの推定に使ってはいけないことに注意が必要である。

傾向スコアによりマッチングした後，比率代入法のパラメータ推定を行い，欠測値を予測する[8)]。年間収入に欠測が発生しており，消費支出との平均値の比率による代入法を用いる場面を考える。このとき，表 4 のとおり，データ内に世帯主の年齢の情報もあるとしよう。

具体的なアルゴリズムは，以下のとおりである。欠測指示行列より，表 4 には 2 つのグループがある。欠測指示行列 = 0 のグループでは，年間収入の値が観測されている。一方，欠測指示行列 = 1 のグループでは，年間収入の値が欠測している。世帯番号 5 の年間収入の値が欠測している。消費支出と世帯主の年齢より，世帯番号 1～4 の中で世帯番号 5 に最も近いのは，世帯番号 3 である。よって，(4.1) 式のとおり，世帯番号 3 の消費支出と年間収入の比率を世帯番号 5 の消費支出の値に掛けた値を，世帯番号 5 の年間収入の欠測値に代入する。

$$\widehat{\text{年間収入}}_5 = \frac{1006}{50}\text{消費支出}_5$$

$$= \frac{1006}{50}49 = 985.88 \qquad (4.1)$$

実際には，傾向スコアの推定は，(4.2)式のとおり，ロジスティック回帰モデルによって行う（阿部，2016，p.103）。ここで，Kを欠測指示行列（0＝欠測，1＝観測）とする[9]。

$$\text{logit}(\text{p}(K_i = 1)) = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \cdots + \beta_{p-1} X_{p-1,i} \qquad (4.2)$$

傾向スコアが得られたら，処置の状況（本研究では欠測）に割り当てられる尤度の近いものどうしで構成される新たな標本を作る。これがマッチングの基本的な考え方である（Guo & Fraser, 2015, p.145）。マッチングの方法はさまざまあるが，本研究では最近隣法（nearest neighbor）を用いる。各々の処置群のユニットに対して，一度に一つずつ，対照群から最も近いユニットを選ぶ貪欲マッチング（greedy matching）である（Ho et al., 2011, p.7）。他のマッチング手法に関しては，Guo & Fraser（2015, pp.145-152）および岩崎（2015, pp.119-120）を参照されたい[10]。

5．実データを用いたサブサンプリングによる実証

本節では，全国消費実態調査の匿名データからノンパラメトリックなサブサンプリングによるシミュレーションを実行する。

5.1 シミュレーション設計

5.1.1 サブサンプリング

Nを母集団サイズ，nを標本サイズ，bを副標本（subsample）サイズとしよう（$N>n>b$）。観測データは，サイズNの母集団から無作為抽出されたサイズnの標本とする。この観測データを擬似母集団として，そこから非復元抽出によってサイズbの副標本を無作為抽出した場合，この副標本は真のモデルから得られたサイズbの標本とみなすことができる。これをサブサンプリングという（Politis et al., 2001, p.1106）。本節では，全国消費実態調査の二人以上世帯（観測数36,936）の匿名データを擬似母集団として扱い，そこからサブサンプリングによって得られた副標本を用いて分析を行う。本稿では，副標本サイズbを1000に設定し，サブサンプリングを1万回実行する[11]。

5.1.2 欠測の発生方法

上記のルールによって得られた各々の副標本において，以下の方法で欠測を発生させた。Schenker et al.（2006, p.925）より，1997年から2004年までのNational Health Interview Surveyにおける収入と所得の欠測率はいずれも平均して約30％だったことを考慮して，欠測率は約30％に設定している。

- MCAR[12]：一様乱数［0, 1］の値に応じて，年間収入の欠測を約30％発生させた。
- MAR[13]（強MAR）：消費支出が中央値よりも大きい場合に年間収入の欠測率を約50％，消費支出が中央値よりも小さい場合に年間収入の欠測率を約10％とし，全体として約30％の欠測を発生させた。
- NMAR[14]（弱MAR）：年間収入が中央値よりも大きい場合に年間収入の欠測率を約50％，年間収入が中央値よりも小さい場合に年間収入の欠測率を約10％とし，全体として約30％の欠測を発生させた。

5.2 シミュレーションにおける推測の対象と結果の評価方法

2節で説明したとおり，年間収入の平均値をθとして，集計対象とする。(5.1)式の偏り（bias）および(5.2)式の二乗平均平方根誤差（RMSE：root mean squared error）を評価方法として使用する（Gujarati, 2003, pp.899-901；Carsey & Harden, 2014, pp.87-89）。偏り（bias）

はゼロであれば，推定量が平均して母数に一致することを意味するため，ゼロに近いほど良い。一方，推定量の性能を比較するときには，偏りと効率性（efficiency）にトレードオフが発生することがある。このとき，推定量の偏りだけでなく，効率性も評価できる指標が二乗平均平方根誤差（RMSE）である。偏りがゼロに近く，RMSEもゼロに近いものが，理想的な推定量である。なお，効率性のみを示した指標として，SE（標準誤差：推定量の標準偏差）も示す。

$$\mathrm{Bias}(\hat{\theta}) = \mathrm{E}(\hat{\theta}) - \theta \tag{5.1}$$

$$\mathrm{RMSE}(\hat{\theta}) = \sqrt{\mathrm{E}(\hat{\theta} - \theta)^2} \tag{5.2}$$

これらの評価手法を用いて，以下の6つの結果を比較して精度の評価を行う。

完全データは，観測数36,936の匿名データからサブサンプリングによって得られたサイズ1000の副標本である。欠測は発生していないため，1万回のサブサンプリングの平均値は不偏と期待される。つまり，この結果が，精度の上限に相当する。

リストワイズ除去によるデータは，上記の完全データに5.1.2で説明した欠測を発生させて，欠測値を含む行全体を除去したデータである。欠測による偏りは一切の処理をされていないため，1万回のサブサンプリングの平均値の期待値は偏っていると期待される。つまり，この結果が，精度の下限に相当する。

通常の比率代入法は，上記の完全データに5.1.2で説明した欠測を発生させたデータにおいて，(3.4)式の平均値の比率を用いた(3.6)式の比率代入法によって欠測値を処理したものである。

傾向スコアを用いた比率代入法は，上記の完全データに5.1.2で説明した欠測を発生させたデータにおいて，4節で提案した手法によって欠測値を処理したものである。

重回帰モデル（線形）は，上記の完全データに5.1.2で説明した欠測を発生させたデータにおいて，(2.1)式の重回帰モデルによって欠測値を処理したものである。

重回帰モデル（対数変換）は，上記の完全データに5.1.2で説明した欠測を発生させたデータにおいて，(5.3)式の重回帰モデルによって欠測値を処理して得られた年間収入の予測値を指数変換したものである。

$$\begin{aligned}\log(\widehat{\text{年間収入}}_i) = {} & \hat{\beta}_0 + \hat{\beta}_1 \text{世帯主の職業}_i \\ & + \hat{\beta}_2 \log(\text{消費支出}_i) + \hat{\beta}_3 \text{世帯主の年齢}_i \\ & + \hat{\beta}_4 \log(\text{就業人員}_i)\end{aligned} \tag{5.3}$$

なお，世帯主の職業は名義尺度の質的変数なので，重回帰モデル（線形）および重回帰モデル（対数変換）では一連のダミー変数としてモデリングした。

5.3 サブサンプリングによるシミュレーション結果

表5は検証結果である。3種類の欠測メカニズムごとに，それぞれの手法の偏り（bias）と二乗平均平方根誤差（RMSE）を報告している。いずれも0に近いほど良い性能を示す。

MCARでは，完全データ，リストワイズ，比率代入法（通常），比率代入法（傾向スコア），重回帰モデル（線形）のいずれも偏りはない。これは，欠測の発生メカニズムが完全に無作為であるという性質上，推定結果にも偏りは発生しないと期待されるとおりである。しかしながら，MCARですら，重回帰モデル（対数変換）には偏りがある。Wooldridge（2009, pp.210-214）にあるとおり，対数変換したモデルの予測値を指数変換した場合，誤差項の期待値が0にならないため偏りが発生するからである。ここから，公的統計における欠測値処理では，対数変換した重回帰モデルが使われない理由が分かる。

MAR（強MAR）では，完全データには偏りがなく，リストワイズが最も偏っている。完全データを除くと，提案手法の比率代入法（傾向スコア）が最も偏りが少ない。一方，

表 5　サブサンプリングによるシミュレーションの結果

欠測メカニズム	手法	Bias	SE	RMSE
MCAR	完全データ	0.145	11.492	11.492
	リストワイズ	0.127	13.694	13.694
	比率（通常）	0.282	14.049	14.052
	比率（傾向スコア）	0.505	14.124	14.133
	重回帰（線形）	0.233	12.879	12.880
	重回帰（対数変換）	−12.888	12.563	17.997
MAR（強MAR）	完全データ	−0.126	11.501	11.501
	リストワイズ	−45.187	13.147	47.061
	比率（通常）	31.444	15.743	35.165
	比率（傾向スコア）	5.996	14.525	15.713
	重回帰（線形）	−7.599	13.759	15.717
	重回帰（対数変換）	−14.557	13.279	19.703
NMAR（弱MAR）	完全データ	−0.017	11.553	11.552
	リストワイズ	−77.388	12.881	78.453
	比率（通常）	−37.089	14.155	39.698
	比率（傾向スコア）	−35.593	14.313	38.363
	重回帰（線形）	−39.924	13.011	41.991
	重回帰（対数変換）	−52.574	12.733	54.094

RMSE基準では，比率代入法（傾向スコア）と重回帰モデル（線形）の性能はほぼ同等である（RMSEは小さい方が優れていることを表す）。結論としては，RMSE基準では，比率代入法（傾向スコア）と重回帰モデル（線形）の性能はほぼ同等だが，偏りの少なさという点で比率代入法（傾向スコア）が勝る。MAR（強MAR）の場合の各手法の偏りを具体的に確認しよう。完全データにおける年収の平均値と比較して，リストワイズでは約45.2万円低く，比率代入法（通常）では約31.4万円高く，比率代入法（傾向スコア）では約6.0万円高く，重回帰モデル（線形）では約7.6万円低く，重回帰モデル（対数変換）では約14.6万円低い。

NMAR（弱MAR）では，完全データには偏りがなく，リストワイズが最も偏っている。完全データを除くと，提案手法の比率代入法（傾向スコア）が最も偏りが少ない。RMSEで判断しても，同様である。NMAR（弱MAR）の場合の各手法の偏りを具体的に確認しよう。完全データにおける年収の平均値と比較して，リストワイズでは約77.4万円低く，比率代入法（通常）では約37.1万円低く，比率代入法（傾向スコア）では約35.6万円低く，重回帰モデル（線形）では約39.9万円低く，重回帰モデル（対数変換）では約52.6万円低い。

なお，MAR（強MAR）では消費支出が中央値よりも大きい場合に，NMAR（弱MAR）では年間収入が中央値よりも大きい場合に，年間収入の欠測率が約50％としている。MAR（強MAR）では消費支出が中央値よりも小さい場合に，NMAR（弱MAR）では年間収入が中央値よりも小さい場合に，年間収入の欠測率を約10％としている。よって，偏りのほとんどが負となっている。

6．モンテカルロ・シミュレーションによるエビデンス

前節では，全国消費実態調査の匿名データからノンパラメトリックなサブサンプリングによる実証を行った。本節では，パラメトリックなモンテカルロ・シミュレーションによるエビデンスも提示することで，提案手法

の優位性に関する主張を補強する[15]。各変数は，(6.1)式と(6.2)式のとおりに生成した。

$$
\begin{aligned}
&Y_i = \beta_0 X_{1i}^{\beta_1} e^{\varepsilon_{1i}} \\
&\beta_0 = 1 \\
&\beta_1 = 1.5 \\
&X_{1i} \sim LN(1.5, 0.5) \\
&\varepsilon_{1i} \sim N(0, 0.3\sqrt{X_{1i}})
\end{aligned} \qquad (6.1)
$$

$$
\begin{aligned}
&X_{ji} = \beta_0 Y_i^{\beta_j} e^{\varepsilon_{ji}} \\
&\beta_2 = 0.4 \\
&\beta_3 = 0.3 \\
&\beta_4 = 0.3 \\
&\varepsilon_{ji} \sim N(0, 1) \\
&j = 2, 3, 4
\end{aligned} \qquad (6.2)
$$

(6.1)式は，指数回帰モデルである。図1で見たとおり，年間収入や消費支出といった変数は，右に裾が長い分布をしている。このような変数は，自然対数に変換すると正規分布に従う対数正規分布と考えられる(Gujarati, 2003, p.175)。経済学における具体例としては，コブ・ダグラス型生産関数を参照されたい(浅野・中村，2009，p.97)。ε_{1i}の分散は$0.3\sqrt{X_{1i}}$であり，X_{1i}の値に応じて大きくなる不均一分散となっている。つまり，通常の最小二乗法による回帰モデルの仮定は満たされていない。(6.2)式も同様だが，ここで，X_{ji}はX_1と直接の関係はないものの，Y_iを通じて間接的な関係を持っている。つまり，分析モデルに含める必要性はないが，代入モデルに含めることで欠測値処理の精度が高められる補助変数の役割を担っている。

Y_iの欠測発生方法は，5.1.2に準じてMCAR，MAR(強MAR)，NMAR(弱MAR)として，欠測率も約30%とした。1万回のモンテカルロ・シミュレーションを実行する。評価方法も，5.2項に準じて，(5.1)式の偏り(bias)および(5.2)式の二乗平均平方根誤差(RMSE)を評価方法として使用する。

6.1　モンテカルロ・シミュレーションの結果

表6は検証結果である。3種類の欠測メカニズムごとに，それぞれの手法の偏り(bias)と二乗平均平方根誤差(RMSE)を報告している。

表6　モンテカルロ・シミュレーションの結果

欠測メカニズム	手法	Bias	SE	RMSE
MCAR	完全データ	−0.016	1.870	1.870
	リストワイズ	−0.030	2.205	2.205
	比率(通常)	−0.032	2.170	2.170
	比率(傾向スコア)	−0.050	2.132	2.132
	重回帰(線形)	−0.035	2.211	2.211
	重回帰(対数変換)	−1.416	1.656	2.179
MAR(強MAR)	完全データ	−0.001	2.117	2.117
	リストワイズ	−3.201	2.003	3.776
	比率(通常)	−1.410	2.213	2.624
	比率(傾向スコア)	−0.328	2.252	2.276
	重回帰(線形)	0.730	4.860	4.914
	重回帰(対数変換)	−2.293	1.552	2.769
NMAR(弱MAR)	完全データ	−0.022	1.685	1.685
	リストワイズ	−3.736	1.740	4.121
	比率(通常)	−2.502	1.851	3.113
	比率(傾向スコア)	−1.559	1.805	2.385
	重回帰(線形)	−0.698	3.122	3.199
	重回帰(対数変換)	−3.507	1.360	3.761

サブサンプリングの場合と同様に，MCARでは，完全データ，リストワイズ，比率代入法（通常），比率代入法（傾向スコア），重回帰モデル（線形）のいずれも偏りはないが，MCARですら，重回帰モデル（対数変換）には偏りがある。

MAR（強MAR）では，完全データには偏りがなく，リストワイズが最も偏っている。完全データを除くと，提案手法の比率代入法（傾向スコア）が最も偏りが少ない。効率性も加味したRMSE基準でも比率代入法（傾向スコア）の性能が最も良い。

NMAR（弱MAR）では，完全データには偏りがなく，リストワイズが最も偏っている。完全データを除くと，重回帰モデル（線形）が最も偏りが少なく，提案手法の比率代入法（傾向スコア）がそれに続く。一方，比率代入法（傾向スコア）の方がRMSE基準において重回帰モデル（線形）よりも優れている。偏りと効率性のトレードオフにおいてどちらを優先すべきかは悩ましい問題だが，RMSEは2つの不偏ではない推定量の優劣を示す（Gujarati, 2003, p.902）。よって，RMSE基準より，総合的には，重回帰モデル（線形）よりも比率代入法（傾向スコア）の方が優れていると判断できる。

特に，比率代入法（傾向スコア）と重回帰モデル（線形）のBiasとRMSEの結果に関して，MARとNMARで結果が逆転している点について詳しく見てみよう。NMARの場合の重回帰モデル（線形）のBiasが小さいことは，実際には良くない兆候を示していることを説明する。表6において，リストワイズのBias（偏り）に注目する。MARのとき，Bias＝－3.201であり，NMARのとき，Bias＝－3.736である。つまり，NMARの方がMARよりも偏りが大きい。

本稿におけるMARとは，Xの値が中央値よりも大きい場合にYの欠測率を約50%，Xの値が中央値よりも小さい場合にYの欠測率を約10%という意味であった。また，本稿におけるNMARとは，Yの値が中央値よりも大きい場合にYの欠測率を約50%，Yの値が中央値よりも小さい場合にYの欠測率を約10%という意味であった。

つまり，いずれの場合もYの大きい方の値に欠測が多く発生しており，リストワイズの結果は真値を過小推定している点では同じである。しかし，NMARのとき，欠測はYの値そのものに直接依存しているため，NMARの過小推定の度合いの方がMARのときよりも大きい。表6より，完全データの精度が最も良く，リストワイズの精度が最も悪い。複数の代入法によって，リストワイズにおける過小推定をいかにして完全データの値に近づけるかという作業をしている。

MARのとき，比率代入法（傾向スコア）のBiasは－0.328で，重回帰モデル（線形）のBiasは0.730である。このとき，比率代入法（傾向スコア）のBiasの方が，絶対値の意味で小さい。さらに，重回帰モデル（線形）のBiasはプラスの値となっており，これは，もともとマイナスだったリストワイズの値を大きく補正しすぎたことを意味している。また，SE（標準誤差）を見比べると，比率代入法（傾向スコア）のSEは2.252で，重回帰モデル（線形）のSEは4.860で，比率代入法（傾向スコア）の方が効率性が良い。

これを踏まえて，NMARの結果を改めて確認する。比率代入法（傾向スコア）のBiasは－1.559で，重回帰モデル（線形）のBiasは－0.698である。一見すると，重回帰モデル（線形）の方が比率代入法（傾向スコア）より偏りが小さいように見える。では，重回帰モデル（線形）の方が比率代入法（傾向スコア）よりも優れているのであろうか？

1万回のモンテカルロ・シミュレーションにおいて，比率代入法（傾向スコア）による平均値の最大値は57.38であるが，重回帰モデル（線形）による平均値の最大値は129.56で

ある。Takahashi, Iwasaki, and Tsubaki (2017) が示したとおり，重回帰モデル（線形）は不均一分散が発生しているとき最良線形不偏推定量（BLUE）ではないが，比率代入法は最良線形不偏推定量である。改めて表 6 の SE（標準誤差）を見比べると，比率代入法（傾向スコア）の SE は 1.805 であり，重回帰モデル（線形）の SE は 3.122 である。比率代入法（傾向スコア）の方が効率性が良い。それは，不均一分散が発生しているとき，重回帰モデル（線形）は最良線形不偏推定量ではないからである。

ゆえに，重回帰モデル（線形）では，大きく外れたケースが期待値を大きい方向に引っ張っている。結果として，MAR では期待値が過大になっていた。同様に，NMAR でも，期待値が過大になったが，リストワイズの値から分かるとおり，欠測データの偏りがマイナス側に大きかったため，皮肉な結果として，重回帰モデル（線形）の過大推定と欠測データのリストワイズの過小推定がほぼ相殺して，期待値は真値の付近にあった。しかし，これは，重回帰モデル（線形）の性能が良いことを意味していない。なぜなら，重回帰モデル（線形）の SE が大きいことは，ある 1 つの調査における集計値が真値から大きく外れることを意味しているからである。

そこで，表 6 をもう一度確認すると，偏りと効率性の双方を評価する RMSE を基準とすると，重回帰モデル（線形）の評価は著しく悪いことが示されている。つまり，総合的に，比率代入法（傾向スコア）の方が重回帰モデル（線形）よりも優れているのである。

7．結語

本研究では，傾向スコアを応用して比率代入法を多変量に拡張する方法について論じた。本稿では，統計実務における集計を想定して，母平均推定のための欠測値処理方法の優劣について，全国消費実態調査の匿名データによるサブサンプリングおよびモンテカルロ・シミュレーションを行った。結果は，表 7 にまとめたとおりであり，従来の手法よりも優れていることが示されている。

表 7　結果のまとめ

手法	Bias	RMSE
完全データ	◎	◎
リストワイズ	×	×
比率（通常）	△	△
比率（傾向スコア）	○	○
重回帰（線形）	△	△
重回帰（対数変換）	×	×

注：記号の意味は，以下のとおりである。
◎「最も優れる」，○「優れる」，△「可」，×「不可」

すべての条件下において完全データが最も優れているが，実際の調査では完全データは利用できない。また，ほぼすべての条件下において，リストワイズが最も悪い結果である。よって，何らかの方法で欠測値に対処する必要がある。しかしながら，重回帰モデル（対数変換）には大きな偏りが確認される。したがって，採用しうる手法は，比率代入法（通常），比率代入法（傾向スコア），重回帰モデル（線形）のいずれかである。

すべての条件下において，提案手法の比率代入法（傾向スコア）は，比率代入法（通常）よりも優れている。また，6 節の NMAR のシミュレーションにおいて，従来の手法である重回帰モデル（線形）は比率代入法（傾向スコア）よりも一見すると偏りが少ないように見えるが，RMSE 基準より，提案手法の比率代入法（傾向スコア）の方が優れており，重回帰モデル（線形）は RMSE 基準でリストワイズよりも悪い性能となる可能性があることが示された。つまり，総合的に，提案手法である比率代入法（傾向スコア）が最も優れている。

分析に使う回帰モデルと傾向スコアに使うモデルの両方が正しく指定されていれば，傾向スコアを使った分析と線形回帰モデルを

使った分析の結果は似通ったものになる（Austin, 2011, p.461）。しかし，回帰分析では変数Yと変数Xの関数関係を適切にモデリングしなければならないが，傾向スコアではその必要はないことが指摘されている（星野；2009，p.67；安井，2020，pp.109-110）。すなわち，線形回帰モデルを使う代入法では，指定した代入モデルが変数Yと変数Xの関数関係を適切にモデリングしていなければならない。一方，傾向スコアを比率代入法に適用した手法では，その必要がない。したがって，傾向スコアを活用した提案手法の方が，モデルの誤設定に強いため，良い結果につながると考えられる[16)]。

謝辞

本稿は，経済統計学会第62回全国研究大会（2018年9月）のセッションGにおける報告に加筆・修正したものである。経済統計学会の参加者の方々から有益なコメントをいただいた。また，匿名の2名の査読者からは，本稿の改善に資する有益なコメントを多数いただいた。ここに深く感謝の意を表したい。ただし，本稿にあり得べき誤りはすべて執筆者に属する。なお，本研究の内容は，統計法に基づいて独立行政法人統計センターから匿名データの提供を受けたもので，分析結果は匿名データを基に筆者が独自に作成・加工したものであり，行政機関等が作成・公表している統計等とは異なる。

注

1）なお，「公的統計の整備に関する基本的な計画」の変更については，総務省のウェブサイトを確認されたい。https://www.soumu.go.jp/menu_news/s-news/01toukatsu01_02000176.html

2）匿名データについては，独立行政法人統計センターの「公的統計のミクロデータ利用」を参照されたい。http://www.nstac.go.jp/services/archives.html

3）総務省統計局（2004）「平成16年全国消費実態調査　用語の解説」を参照されたい。http://www.stat.go.jp/data/zensho/2004/kaisetsu.htm#4

4）総務省統計局（2015）「平成26年全国消費実態調査：二人以上の世帯の家計収支及び貯蓄・負債に関する結果」を参照されたい。https://www.stat.go.jp/data/zensho/2014/pdf/gaiyo3.pdf

5）ブリューシュ・ペイガン検定については，高橋・渡辺（2017，pp.101-102）を参照されたい。

6）秘匿については，独立行政法人統計センターの「匿名データの利用に関するFAQ（回答）」を参照されたい。http://www.nstac.go.jp/services/faq-a-anonymity.html

7）補助変数を同時に欠測している場合，マッチングを行うことはできない。しかし，これは重回帰モデルの場合にも予測値を計算できないことになるため，本稿では，補助変数を同時に欠測しているケースは対象としていない。

8）ドナーを探してくる範囲を狭く設定している場合，ドナーとレシピエントは似通ったものとなり好ましいが，適当な近傍ケースが得られないことがあり得る。この場合，ドナーを探す範囲を広く設定すれば，いつかは必ずドナーが見つかるが，ドナーとレシピエントの距離が遠くなる可能性がある（高橋・渡辺，2017，p.119）。マッチングの精度が代入法の精度にどの程度の影響を与えるか，確認する必要がある。この点は，提案手法の短所である。しかしながら，重回帰モデルによる代入法においても，ガウス・マルコフの仮定が満たされているかどうか，残差の分析によって確認する必要がある。よって，実務的な意味で，どちらの手法が煩雑かは一概には言えない。

9）本研究では，傾向スコアの推定に際して，RパッケージMatchItを用いた（Ho et al., 2011）。なお，本研究で用いたRのバージョンは，R 3.6.3である。

10）本研究の初期段階において，貪欲マッチングの代わりに最適マッチング（optimal matching）も試したが，パフォーマンスはあまり変わらなかった。また，最適マッチングでは収束しないケースも

あったことから，貪欲マッチングを採用した。

11）サブサンプリングを用いたデータエディティング手法の検証については，Di Zio & Guarnera（2013）も参照されたい。

12）MCARは，Missing Completely At Randomの略で，欠測の発生メカニズムが完全に無作為であることを意味する。これは，3つの欠測メカニズムの中で最も強い仮定である。欠測メカニズムの詳細は，高橋・渡辺（2017，pp.15-17）を参照されたい。

13）MARは，Missing At Randomの略で，直訳では無作為な欠測であるが，実際の意味は観測データに条件付けた場合に無作為な欠測である。これは，3つの欠測メカニズムの中で2番目に強い仮定である。

14）NMARは，Not Missing At Randomの略で，無作為ではない欠測である。なお，MARとNMARは，種類の差ではなく程度の差である（Graham, 2009, p.567；高橋・渡辺，2017，p.21）。よって，このNMARは，弱MARとして理解できる。ある値の欠測確率がその値自体に依存しているものの，観測データを条件とした場合，欠測をある程度まで無視できる状態である（高橋，2018）。これは，3つの欠測メカニズムの中で最も弱い仮定である。

15）モンテカルロ・シミュレーションの設計についてはCarsey & Harden（2014）を，日本語によるモンテカルロ法の解説は大野・井川（2015，pp.36-53）を，imputation手法の評価としてのシミュレーションの設計についてはvan Buuren（2018, pp.51-53）を，モンテカルロ・シミュレーションの実例についてはKing et al.（2001, pp.59-62）をそれぞれ参照されたい。

16）査読者より「本来欠測は割付けというより何らか意図的な行為（NMAR）とすると，属性の交絡のみならずサンプルセレクションの側面があるので，属性のバランス処理の方法で優位のPSでもそのような欠測値問題全体の枠組みについて限界がある」との指摘があった。しかしながら，脚注14で指摘したとおり，MARとNMARは，種類の差ではなく程度の差である。ゆえに，代入法において使用できる補助変数をできる限り多くすることで，MARの仮定を満たす確率が向上することが知られている。これを包括的分析法（inclusive analysis strategy）という（Enders, 2010, p.6；高橋・渡辺，2017，p.21）。その意味でも，傾向スコアによって補助変数の情報を代入モデルにできる限り多く取り入れることによって，欠測による偏りを是正できると考えられる。

参考文献

［1］ 浅野晢・中村二朗（2009）『計量経済学』，（第2版），有斐閣.

［2］ 阿部貴行（2016）『欠測データの統計解析』，朝倉書店.

［3］ 岩崎学（2015）『統計的因果推論』，朝倉書店.

［4］ 大野薫・井川孝之（2015）『モンテカルロ法入門』，一般財団法人金融財政事情研究会.

［5］ 高橋将宜（2017）「諸外国の公的統計における欠測値の対処法：集計値ベースと公開型ミクロデータの代入法」，『統計学』第112号，pp.65-83.

［6］ 高橋将宜（2018）「多重代入法による匿名データの解析特性の改善について：全国消費実態調査を例に」，『統計学』第114号，pp.15-29.

［7］ 高橋将宜・渡辺美智子（2017）『欠測データ処理：Rによる単一代入法と多重代入法』，共立出版.

［8］ 星野崇宏（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.

［9］ 安井翔太（2020）『効果検証入門：正しい比較のための因果推論/計量経済学の基礎』，技術評論社.

［10］ Austin, P.C.（2011）“An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies”, *Multivariate Behavioral Research* Vol. 46, No. 3, pp.399-424.

［11］ Carsey, T.M. & Harden, J.J.（2014）*Monte Carlo Simulation and Resampling Methods for Social Science*, Sage Publications, Inc.

［12］ de Waal, T., Pannekoek, J., & Scholtus, S.（2011）*Handbook of Statistical Data Editing and Imputation*，John Wiley & Sons.

［13］ Di Zio, M. & Guarnera, U.（2013）“A Contamination Model for Selective Editing”, *Journal of Official Statistics* Vol. 29, No. 4, pp.539-555.
［14］ Enders, C.K.（2010）*Applied Missing Data Analysis*, The Guilford Press.
［15］ Graham, J.W.（2009）“Missing Data Analysis: Making It Work in the Real World”, *Annual Review of Psychology* Vol. 60, pp.549-576.
［16］ Gujarati, D.N.（2003）*Basic Econometrics*, fourth edition, McGraw-Hill.
［17］ Guo, S. & Fraser, M.W.（2015）*Propensity Score Analysis: Statistical Methods and Applications*, second edition, Sage Publications.
［18］ Ho, D.E., Imai, K., King, G., & Stuart, E.A.（2011）“MatchIt: Nonparametric Preprocessing for Parametric Causal Inference”, *Journal of Statistical Software* Vol. 42, No. 8, pp.1-28.
［19］ King, G., Honaker, J., Joseph, A., & Scheve, K.（2001）“Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation”, *American Political Science Review* Vol. 95, No. 1, pp.49-69.
［20］ Little, R.J.A., & Rubin, D.B.（2020）*Statistical Analysis with Missing Data*, third edition, John Wiley & Sons.
［21］ Politis, D.N., Romano, J.P., & Wolf, M.（2001）“On the Asymptotic Theory of Subsampling”, *Statistica Sinica* Vol. 11, No. 4, pp.1105-1124.
［22］ Schenker, N., Raghunathan, T.E., Chiu, P.-L., Makuc, D.M., Zhang, G., & Cohen, A.J.（2006）“Multiple Imputation of Missing Income Data in the National Health Interview Survey”, *Journal of the American Statistical Association* Vol. 101, No. 475, pp.924-933.
［23］ Takahashi, M., Iwasaki, M., & Tsubaki, H.（2017）“Imputing the Mean of a Heteroskedastic Log-Normal Missing Variable: A Unified Approach to Ratio Imputation”, *Statistical Journal of the IAOS* Vol. 33, No. 3, pp.763-776.
［24］ van Buuren, S.（2018）*Flexible Imputation of Missing Data*, second edition, Chapman & Hall/CRC.
［25］ Wooldridge, J.M.（2009）*Introductory Econometrics: A Modern Approach*, 4th edition, South-Western.

A New Multivariate-type Ratio Imputation Model by Propensity Score Matching: Evidence from the Anonymized Microdata of the National Survey of Family Income and Expenditure

Masayoshi TAKAHASHI*

Summary

It is academically and socially important to improve the imputation method in official economic statistics. To this date, ratio imputation has been often utilized as a way of dealing with missing values in official economic statistics around the world. However, the ratio imputation model is a bivariate regression model without intercept; thus, it cannot incorporate information from many covariates in data. This article aims to improve the ratio imputation model by applying propensity score matching, so as to balance many covariates. This article examines the performance of the proposed method and other traditional methods, by subsampling simulation based on the Anonymized Microdata of the National Survey of Family Income and Expenditure and by Monte Carlo simulation.

Key Words

Missing data, ratio imputation, propensity score matching, official statistics, anonymized microdata

* School of Information and Data Sciences, Nagasaki University

機関誌『統計学』投稿規程

経済統計学会（以下，本会）会則第3条に定める事業として，『統計学』（電子媒体を含む。以下，本誌）は原則として年に2回（9月，3月）発行される。本誌の編集は「経済統計学会編集委員会規程」（以下，委員会規程）にもとづき，編集委員会が行う。投稿は一般投稿と編集委員会による執筆依頼によるものとし，いずれの場合も原則として，本投稿規程にしたがって処理される。

1．総則

1-1　投稿者

会員（資格停止会員を除く）は本誌に投稿することができる。

1-2　非会員の投稿

(1)　原稿が複数の執筆者による場合，筆頭執筆者は本会会員でなければならない。

(2)　常任理事会と協議の上，編集委員会は非会員に投稿を依頼することができる。

(3)　本誌に投稿する非会員は，本投稿規程に同意したものとみなす。

1-3　未発表

投稿は未発表ないし他に公表予定のない原稿に限る。

1-4　投稿の採否

投稿の採否は，審査の結果にもとづき，編集委員会が決定する。その際，編集委員会は原稿の訂正を求めることがある。

1-5　執筆要綱

原稿作成には本会執筆要綱にしたがう。

2．記事の分類

2-1　研究論文

以下のいずれかに該当するもの。

(a)　統計およびそれに関連した分野において，新知見を含む会員の独創的な研究成果をまとめたもの。

(b)　学術的な新規性を有し，今後の研究の発展可能性を期待できるもので，速やかな成果の公表を目的とするもの。

2-2　報告論文

研究論文に準じる内容で，研究成果の速やかな報告をとくに目的とする。

2-3　書評

統計関連図書や会員の著書などの紹介・批評。

2-4　資料

各種統計の紹介・解題や会員が行った調査や統計についての記録など。

2-5　フォーラム

本会の運営方法や統計，統計学の諸問題にたいする意見・批判・反論など。

2-6　海外統計事情

諸外国の統計や学会などについての報告。

2-7　その他

全国研究大会・会員総会記事，支部だより，その他本会の目的を達成するために有益と

思われる記事。

3．原稿の提出

3-1　投稿

原稿の投稿は常時受け付ける。

3-2　原稿の送付

原則として，原稿は執筆者情報を匿名化したPDFファイルを電子メールに添付して編集委員長へ送付する。なお，ファイルは『統計学』の印刷レイアウトに準じたPDFファイルであることが望ましい。

3-3　原稿の返却

投稿された原稿（電子媒体を含む）は，一切返却しない。

3-4　校正

著者校正は初校のみとし，大幅な変更は認めない。初校は速やかに校正し期限までに返送するものとする。

3-5　投稿などにかかわる費用

⑴　投稿料は徴収しない。

⑵　掲載原稿の全部もしくは一部について電子媒体が提出されない場合，編集委員会は製版にかかる経費を執筆者（複数の場合には筆頭執筆者）に請求することができる。

⑶　別刷は，研究論文，報告論文については30部までを無料とし，それ以外は実費を徴収する。

⑷　3-4項にもかかわらず，原稿に大幅な変更が加えられた場合，編集委員会は掲載の留保または実費の徴収などを行うことがある。

⑸　非会員を共同執筆者とする投稿原稿が掲載された場合，その投稿が編集委員会の依頼によるときを除いて，当該非会員は年会費の半額を掲載料として，本会に納入しなければならない。

3-6　掲載証明

掲載が決定した原稿の「受理証明書」は学会長が交付する。

4．著作権

4-1　本誌の著作権は本会に帰属する。

4-2　本誌に掲載された記事の発行時に会員であった執筆者もしくはその遺族がその単著記事を転載するときには，出所を明示するものとする。また，その共同執筆記事の転載を希望する場合には，他の執筆者もしくはその遺族の同意を得て，所定の書面によって本会に申し出なければならない。

4-3　前項の規定にもかかわらず，共同執筆者もしくはその遺族が所在不明のため，もしくは正当な理由によりその同意を得られない場合には，本会が承認するものとする。

4-4　執筆者もしくはその遺族以外の者が転載を希望する場合には，所定の書面によって本会に願い出て，承認を得なければならない。

4-5　4-4項にもとづく転載にあたって，本会は転載料を徴収することができる。

4-6　会員あるいは本誌に掲載された記事の発行時に会員であった執筆者が記事をウェブ転載するときには，所定の書類によって本会に申し出なければならない。なお，執筆者が所属する機関によるウェブ転載申請については，本人の転載同意書を添付するものとする。

4-7 会員以外の者，機関等によるウェブ転載申請については，前号を準用するものとする。

4-8 転載を希望する記事の発行時に，その執筆者が非会員の場合には，4-4，4-5項を準用する。

1997年7月27日制定（2001年9月18日，2004年9月12日，2006年9月16日，2007年9月15日，2009年9月5日，2012年9月13日，2016年9月12日一部改正）

機関誌『統計学』の編集・発行について

『統計学』編集委員会

みなさまからの投稿を募集しています。ぜひ研究成果の本誌上での発表をご検討ください。

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次のとおりです。
　第120号：2021年3月31日
　第121号：2021年9月30日
4．原則として，すべての投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

編集後記

投稿者のみなさま，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。編集委員会の活動にご理解ご協力ありがとうございました。　（小林良行　記）

執 筆 者 紹 介

高橋将宜　（長崎大学情報データ科学部）　　水野谷武志　（北海学園大学経済学部）
氏川恵次　（横浜国立大学大学院国際社会科学研究院）

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） mizunoya@econ.hokkai-s-u.ac.jp	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 （042-674-3421） ysakata@tamacc.chuo-u.ac.jp	坂田幸繁（代行）
関　　西 ……	580-8502	松原市天美東 5-4-33 阪南大学経済学部 （072-332-1224） m-murakami@hannan-u.ac.jp	村上雅俊
九　　州 ……	890-0065	鹿児島市郡元 1-21-30 鹿児島大学法文学部 （099-285-7601） matsukawa@leh.kagoshima-u.ac.jp	松川太一郎

統　計　学　No.119　　定価 1,760円（本体1,600円）

2020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

Statistics

No. 119 2020 September

Articles

Short Articles

JSES Activities

Japan Society of Economic Statistics