【研究論文】（『統計学』第114号 2018年3月）

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 ―

高橋将宜*

要旨

「官民データ活用推進基本法」の施行により，公的統計の調査によって収集されたデータの二次利用が促進され，匿名データ（ミクロデータ）として利用・分析できる環境が整ってきた。しかし，調査票を活用してデータを収集する公的統計調査では，完全な形でデータが得られることはまれであるため，公的統計における欠測値は代入法によって処理されている。本稿では，データの使用者側の視点から，代入済みの匿名データを利用した実証分析を行う際に，欠測値が代入されていることによってどのような影響があるか論じる。具体的には，本稿は，全国消費実態調査の匿名データを用いて，バイアスを考慮した推定手法の適用可能性を模索する。また，全国消費実態調査の匿名データから，サブサンプリングによってシミュレーション分析を行い，リストワイズ除去，単一代入法，多重代入法などの欠測値処理の仕方によって，分析結果の精度にどのような影響が出るかを検証する。

キーワード

多重代入法，単一代入法，欠測データ，公的統計，匿名データ

1．はじめに

20世紀半ばまでの実証研究ではマクロ集計値による分析が主流だったが，近年ではミクロレベルの個体行動に関する分析の需要が増えてきている（坂田，2006，p.31）。供給側についても，2016年12月に「官民データ活用推進基本法」が施行され，公的統計における二次利用が促進されている。国勢調査，労働力調査，住宅・土地統計調査，全国消費実態調査，就業構造基本調査など，公的統計の調査によって収集されたデータは，独立行政法人統計センターを通じて，匿名データ（ミクロデータ）として利用できる[1)]。以前は，都道府県や市区町村を単位としたマクロな集計値からしか分析が行えなかった社会・経済現象について，世帯や企業といった調査単位からのミクロレベルの分析が可能となっている。

しかしながら，調査票によってデータを収集する公的統計調査では，データが完全な状態で得られることはまれである。観測データを条件とした場合に欠測が無作為なMAR[2)]（Missing At Random）であれば，欠測値を何らかの値に置き換える代入法（imputation）によって欠測値を処理することができる[3)]。よって，諸外国を含めて，公的統計では欠測値の対処方法として代入法が採用されている（de Waal et al., 2011；野村総合研究所，2013）。特に，欠測値の処理について，集計値を算出する目的の調査データには確定的な単一代入

* 正会員，東京外国語大学経営戦略情報本部
e-mail：mtakahashi@tufs.ac.jp

法（deterministic single imputation）がふさわしく，公開を前提としたミクロデータには多重代入法（multiple imputation）がふさわしいことが示されている（高橋，2017，p.77）。これは，データ提供側の欠測対処法に関して論じたものである。

現在，匿名データとして提供されているミクロデータでは，欠測がどのように処理されているか明示的ではない部分があり，分析の際には注意を要する。実際に，2011年から2016年までの6年間に全国消費実態調査の匿名データを用いた実証研究（12件）を検討したところ，欠測を適切に処理しているものは1件もなかった。そこで，本研究では，データの使用者側の欠測値処理について論じる。

本稿は，全国消費実態調査の匿名データを用いた個体行動に関する実証分析を通じて，バイアスを考慮した推定手法の適用可能性について，匿名データによる計量分析手法のさらなる展開を模索していくものである。本稿では，世帯や住居に関する事項といった属性ごとに，家計上の収入と支出，年間収入及び貯蓄に関して，どのような差異があるか実証的に分析する研究を想定している。このような実証分析を行う際に，全国消費実態調査の欠測は，単一代入法によって処理されていると推認されるため，分析結果に影響が出る可能性がある。よって，本稿では，そのような影響を考慮した分析方法について考察する。

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全国消費実態調査（平成16年：2004年）の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点に注意されたい。

本稿第2章では，全国消費実態調査の匿名データにおける変数の特徴と欠測値処理の状況について論じる。第3章では，欠測値処理の方法によって分析結果に影響が出る例として，二人世帯の母子家庭に関する分析を扱う。第4章では，全国消費実態調査の匿名データからサブサンプリング（subsampling）によるシミュレーション分析を行い，欠測値処理の方法が実証分析の結果に与える影響を検証する。第5章において締めくくりとする。

2．全国消費実態調査の匿名データ

全国消費実態調査は，「家計の実態を所得，消費，資産の三面から総合的にとらえようとするもの」[4]である。2004年調査の匿名データは，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が，本研究では二人以上世帯を対象とし，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していない。本研究で使用したデータの標本サイズは43,861である。

2.1　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。食料（以下，食費）を被説明変数とし，就業人員，住宅延べ床面積（以下，住宅面積），年齢5歳階級（以下，年齢），年間収入，消費支出（10区分分類），非消費支出，実支出以外の支出（以下，実支出以外），繰越金，貯蓄現在高を説明変数として分析を行う。なお，実際の分析では，住宅面積について結果を見やすくするため，100で割って$10m^2$単位とした。実収入，実収入以外の収入，繰入金，支出総額，実支出，消費支出，通信は，分析モデルには含めないが代入モデルには含める補助変数（高橋・渡辺，2017，p.16）として使用している。なお，消費支出の10区分分類は，食費，住居，光熱・水道，家具・家事用品（以下，家具・家事），被服及び履物（以下，被服・履物），保健医療，交通・通信，教育，教養娯楽，その他の消費支出（以下，その他消費）であるが，教育については，0が欠測を表すかどうか不明なため，本研究では使用

表１　変数の一覧

変数番号	変数名	変数の種類	変数番号	変数名	変数の種類
V0456	食料	被説明変数	V0598	非消費支出	説明変数
			V0609	実支出以外の支出	説明変数
V0018	就業人員	説明変数	V0622	繰越金	説明変数
V0029	住宅延べ床面積	説明変数	V0671	貯蓄現在高	説明変数
V0042	年齢5歳階級	説明変数			
V0399	年間収入	説明変数	V0401	実収入	補助変数
V0498	住居	説明変数	V0439	実収入以外の収入	補助変数
V0504	光熱・水道	説明変数	V0452	繰入金	補助変数
V0509	家具・家事用品	説明変数	V0453	支出総額	補助変数
V0519	被服及び履物	説明変数	V0454	実支出	補助変数
V0537	保健医療	説明変数	V0455	消費支出	補助変数
V0542	交通・通信	説明変数	V0551	通信	補助変数
V0557	教養娯楽	説明変数			
V0567	その他の消費支出	説明変数	V0553	教育	使用しない

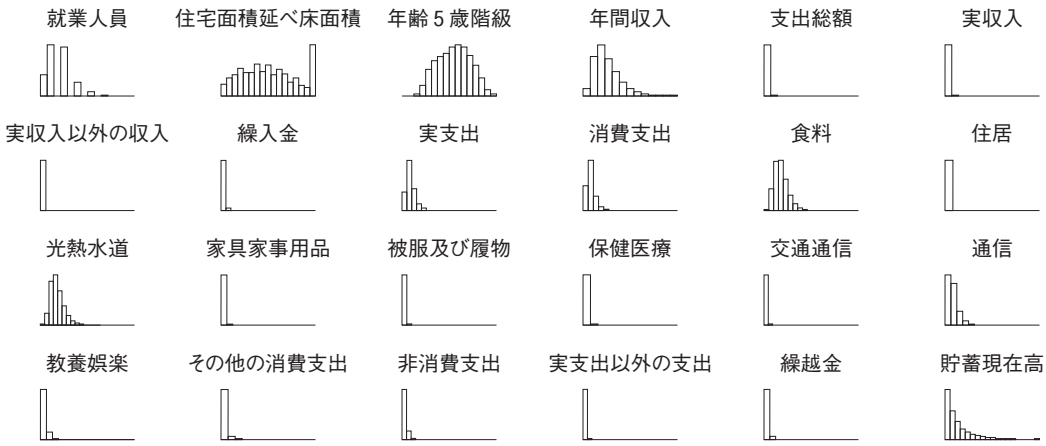

図１　生データの分布

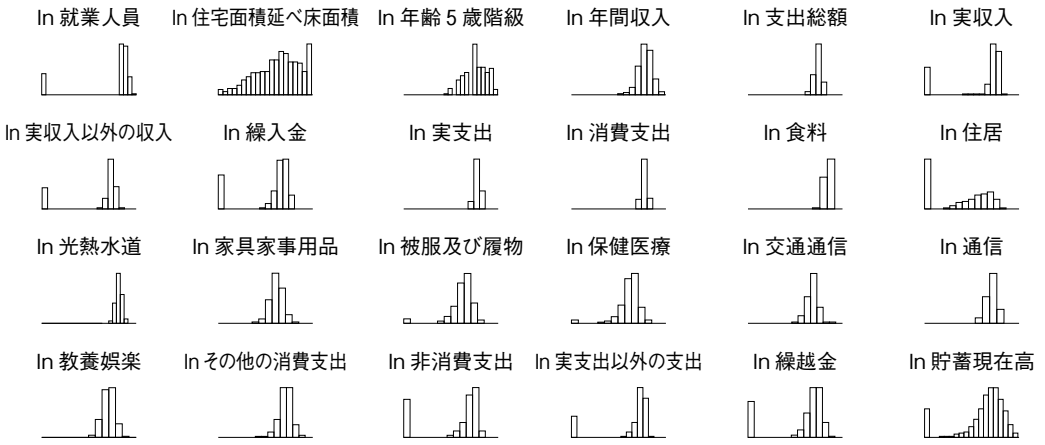

図２　自然対数変換後の分布（lnは自然対数を表す）

しない。

図1と図2のヒストグラム[5]から，就業人員は生データのままとし，それ以外の変数は自然対数に変換して使用した。なお，最小値が0の変数には，微小な値を足して自然対数に変換した。しかし，本稿の分析結果が示すとおり，多くの変数において0は欠測値を表していると推認できるため，実際には，0を欠測させた上で多重代入法によって処理をすることが望ましい。

2.2 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説明変数は年間収入である。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。よって，1もしくはブランクの場合，年間収入の値が欠測していることがわかる。43,861個の観測数のうち，3,024個が欠測しており，欠測率は6.9％である。全国消費実態調査では，年間収入が不詳の世帯については，「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」している[6]。なお，就業人員(V0018)と有業人員(V0391)は同じデータである。

説明変数の中で，貯蓄現在高についても，欠測を明示的に特定することができる。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_貯蓄」(V0010)という変数があり，ここで「1＝貯蓄に不詳あり」，「0＝貯蓄に不詳なし」，「ブランク＝年収票無し」を表している。よって，年間収入と同様に，1もしくはブランクの場合，貯蓄現在高の値が欠測していることがわかる。43,861個の観測数のうち，3,825個が欠測しており，欠測率は8.7％である。貯蓄現在高の欠測値は，0に置き換えられている。

なお，この欠測率は，米国において収入と所得について調査した公的統計よりも極めて低い。たとえば，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％である(Schenker et al., 2006, p.925)。実際に全国消費実態調査において，上記の基準で除去されるもの以外にも欠測が発生していたかどうかは定かではないが，本研究では，上記で特定できた値のみを欠測値とみなすものとする。

その他の変数における欠測値は，明示的にフラグなどは立っていないが，食費や光熱・水道代などの生活費が毎月0円とは考えられないため，0として処理されている値は，もともとは欠測値だったと推定される。ただし，教育費は，子育てをしていない世帯の場合，0が欠測を表しているかどうか不明なため，本研究では使用しない。住居費については，「住宅ローンの有無」(V0395)が1または「家賃・地代を支払っている世帯の割合」(V0396)が1のとき，0は欠測を表すと確定できる。

1つの欠測値に対して代入値は1つしか含まれていないため，全国消費実態調査の匿名データにおける欠測値は，単一代入法によって処理されていることがわかる。回帰分析などにおいて，統計的推測を行う際には，標準誤差を過小評価してしまうおそれがあり，妥当な統計的推測が行えない可能性がある(高橋・渡辺，2017，p.71)。

2.3 全国消費実態調査の匿名データにおけるトップコーディング

匿名データでは，秘匿の目的で，リサンプリング，識別情報の削除等，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，リコーディングが施されている[7]。中でも，トップコーディングは極端に大きな

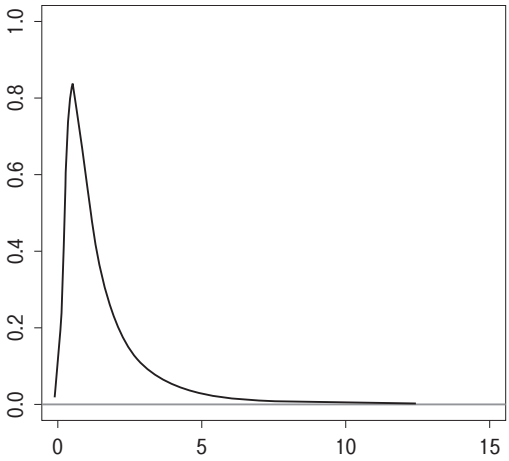

図３　トップコーディング前のイメージ図

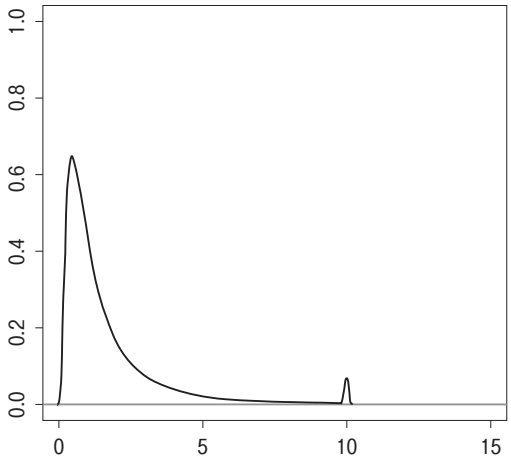

図４　トップコーディング後のイメージ図

値に関して上限値を設けて図３の分布を図４のように切断している。このようにトップコーディングされている場合，真値は10以上のどこかに存在することだけはわかっている。

住宅面積について，200m^2以上は200m^2にトップコーディングされており，43,861個の観測数のうち，4,646個の観測値が該当する。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，230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，268個の観測値が該当する。貯蓄現在高について，9500万円以上は9500万円にトップコーディングされており，369個の観測値が該当する。特に，住宅面積のトップコーディング率は10％を超えており，図１のヒストグラムからも影響が大きいと判断できるため，特別な対処が必要である。

トップコーディングが実施されている場合，可能な最小の値のみが報告されているため，トップコーディングされている値をいったん欠測させた上で，観測値に関するベイズの事前分布[8]を活用しながら多重代入法によって対処することができる（Honaker & King, 2010; Honaker et al., 2011, pp.20-23）。たとえば，図３と図４の例では，10よりも大きいという情報がわかっているので，この情報を事前分布として活用できる。つまり，観測データを条件として，10よりも大きい部分からトップコーディングされた値の事後予測分布を構築し，ここから無作為な抽出を行う。今回のデータでは，住宅面積の上位10％ほどがトップコーディングされているため，この変数のトップコーディングは，欠測値として処理し直すこととする。

2.4　全国消費実態調査の匿名データを用いた先行研究

表２は，2011年から2016年までの６年間に全国消費実態調査の匿名データを用いた実証研究の一覧である。この中で，吉川・天野・島田（2011），バラスet al.（2012），木村（2012），魚住（2014），上村・足立・金田（2016）は，欠測値に言及しておらず，どのように処理したか不明である。増田（2015）は，被説明変数の０を処理する方法としてTOBITモデルを用いているが，０の値が欠測値を代入したものであれば，欠測値が適切に処理されていないおそれがある。それ以外の研究では，不詳や０はリストワイズ除去によって処理されている。説明変数における欠測が完全に無作為なMCAR[9]（Missing Com-

表２　先行研究における欠測値の処理方法（2011年～2016年）

著者	処理方法
吉川・天野・島田（2011）	不明
平山（2011）	リストワイズ除去
バラスet al.（2012）	不明
花岡（2012）	比例配分，リストワイズ除去
木村（2012）	不明
猿山et al.（2013）	リストワイズ除去
Higa（2013）	リストワイズ除去
渡辺（2013）	リストワイズ除去
魚住（2014）	不明
増田（2015）	TOBITモデル（被説明変数の０対策）
田村・松林（2015）	リストワイズ除去
上村・足立・金田（2016）	不明

注１：http://www.nstac.go.jp/services/jisseki-zensho.html
注２：実証研究のみを対象とし，教育目的のものは除外した。

pletely At Random）でなければ，一般的に，リストワイズ除去による分析結果には偏りが生じるおそれがある（高橋・渡辺，2017，p.23）。

３．実証分析の例

本章では，全国消費実態調査の匿名データを用いて，二人世帯の母子家庭[10]について，食費を被説明変数とした重回帰分析を実行する。この例を通じて，欠測の処理方法の違いにより分析結果にどのような差が出るか例証することを目的とする。比較した欠測処理方法は，提供されたままのデータをそのまま用いた「元のデータ」，欠測値と推定されるセルを含む行全体を除去した「リストワイズ」，RパッケージAmeliaを用いてEMB（Expectation-Maximization with Bootstrapping）アルゴリズムによって実行した「多重代入法[11]」である。

被説明変数の食費は，前章で見たとおり，自然対数に変換している。分析モデルは，(1)式であり，帰無仮説$H_0：\beta_1=0$の検証を行う（ここでlnは自然対数を表す）。エンゲルの法則では，食費は収入の高低に関わらず一定であると考えられ，生活水準の高い世帯ではエンゲル係数が低いとされた。もしそうであれば，帰無仮説は棄却されないはずである。一方，現代社会では，収入が増えれば高級な食材を購入することができるようになり，収入は食費に対して影響があるとも考えられる。もしそうであれば，帰無仮説は棄却されるはずである。食費以外の消費支出行動を統制した上で年間収入の食費に与える影響を分析した結果は，表３のとおりである。

$$\begin{aligned}\ln(\text{食費}) = {} & \beta_0 + \beta_1 \ln(\text{年間収入}) + \beta_2 \text{就業人員} \\ & + \beta_3 \ln(\text{住宅面積}) + \beta_4 \ln(\text{年齢}) \\ & + \beta_5 \ln(\text{住居}) + \beta_6 \ln(\text{光熱・水道}) \\ & + \beta_7 \ln(\text{家具・家事}) + \beta_8 \ln(\text{被服・履物}) \\ & + \beta_9 \ln(\text{保健医療}) + \beta_{10} \ln(\text{交通・通信}) \\ & + \beta_{11} \ln(\text{教養娯楽}) + \beta_{12} \ln(\text{その他消費}) \\ & + \beta_{13} \ln(\text{非消費支出}) + \beta_{14} \ln(\text{実支出以外}) \\ & + \beta_{15} \ln(\text{繰越金}) + \beta_{16} \ln(\text{貯蓄現在高}) + \varepsilon_i\end{aligned} \quad (1)$$

自然対数に変換しているので，係数の解釈はパーセント変化を表す（Wooldridge, 2009, pp.189-192）。元のデータを用いた分析では，他の変数の値が一定の場合，年間収入が１％増加すると，食費は0.036％増加するが，この結果は５％水準で統計的に有意ではない。リストワイズ除去を用いた分析では，他の変数の値が一定の場合，年間収入が１％増加すると，食費は0.037％増加するが，この結果も５％水準で統計的に有意ではない。すなわち，

表３　分析結果（二人世帯の母子家庭）

	元のデータ			リストワイズ			多重代入法（$M=100$）		
	係数	標準誤差	p値	係数	標準誤差	p値	係数	標準誤差	p値
切片	4.534	0.684	0.000	4.536	0.684	0.000	4.398	0.442	0.000
ln（年間収入）	**0.036**	**0.069**	**0.597**	**0.037**	**0.069**	**0.592**	**0.125**	**0.035**	**0.000**
就業人員	0.014	0.092	0.883	0.014	0.092	0.883	0.013	0.052	0.797
ln（住宅面積）	− 0.019	0.067	0.771	− 0.019	0.067	0.774	− 0.036	0.048	0.453
ln（年齢）	0.607	0.161	0.000	0.608	0.161	0.000	0.602	0.105	0.000
ln（住居）	0.001	0.007	0.942	0.000	0.007	0.949	0.002	0.005	0.743
ln（光熱・水道）	0.185	0.048	0.000	0.185	0.048	0.000	0.199	0.035	0.000
ln（家具・家事）	0.126	0.032	0.000	0.126	0.032	0.000	0.082	0.020	0.000
ln（被服・履物）	0.065	0.027	0.016	0.065	0.027	0.016	0.065	0.018	0.000
ln（保健医療）	− 0.015	0.024	0.540	− 0.014	0.024	0.542	− 0.010	0.017	0.549
ln（交通・通信）	− 0.000	0.043	0.995	− 0.000	0.043	0.994	0.004	0.029	0.134
ln（教養娯楽）	0.062	0.031	0.047	0.062	0.031	0.047	0.055	0.021	0.009
ln（その他消費）	0.037	0.033	0.259	0.037	0.033	0.259	0.012	0.020	0.544
ln（非消費支出）	− 0.010	0.027	0.724	− 0.010	0.027	0.726	− 0.018	0.019	0.338
ln（実支出以外）	− 0.002	0.039	0.969	− 0.002	0.039	0.967	0.006	0.020	0.757
ln（繰越金）	0.040	0.025	0.108	0.040	0.025	0.108	0.015	0.017	0.361
ln（貯蓄現在高）	0.027	0.020	0.191	0.026	0.020	0.197	0.019	0.014	0.183
決定係数	0.586			0.586			0.606		
n	299			168			299		

注：被説明変数は，ln（食費）である。lnは自然対数を表す。

表４　感度分析の結果

	感度パラメータ				
	− 0.494	− 0.247	0.000	0.247	0.494
係数	0.128	0.130	0.130	0.128	0.124
p値	0.000	0.000	0.000	0.000	0.000
決定係数	0.602	0.603	0.602	0.602	0.601

欠測値を処理しない場合，年間収入は食費に影響を与えていないと結論付けられる。一方，多重代入法を用いた分析では，他の変数の値が一定の場合，年間収入が１％増加すると，食費は0.125％増加し，この結果は５％水準で統計的に有意である。もし年間収入が２倍になると，食費は12.5％増加すると解釈できる。このように，欠測値をどう処理したかによって，分析結果の解釈が劇的に変わりうることが示唆されている[12]。

表４は，欠測が無作為ではないNMAR[13]（Not Missing At Random）の仮定の下，感度分析（sensitivity analysis）を実行したものである。感度分析とは，もし真の欠測メカニズムがNMARである場合，MARを仮定した分析結果にどのような影響が出るかを検証するものである。いずれの感度パラメータの下でも，結果はほぼ同じであり，結論は頑健であることが確認されている。感度分析の具体的な実行方法については，高橋・渡辺（2017，pp.155-163）を参照されたい。

４．実データを用いたシミュレーション

前章では一例を取り上げたに過ぎないが，

表５　職業分類（二人以上世帯）

職業符号番号	職業	標本サイズ
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

本章では全国消費実態調査の匿名データからサブサンプリングによるシミュレーションを実行する。表５は，職業分類とそれぞれの職業に分類される標本のサイズを示している。

この中から，職業符号番号１（常用労務作業者），職業符号番号３（民間職員），職業符号番号４（官公職員１），職業符号番号５（官公職員２），職業符号番号12（無職）を分析のベースとなるデータとして用いる。なお，職業符号番号４（官公職員１）と職業符号番号５（官公職員２）は，１つのグループとして扱った。

職業符号番号２（臨時及び日々雇労務作業者），職業符号番号11（その他），職業符号番号13（家族従業者）は，小標本であり，サブサンプリングの分析に適していないため使用していない。職業符号番号６（商人及び職人），職業符号番号７（個人経営者），職業符号番号８（農林漁業従業者），職業符号番号９（法人経営者），職業符号番号10（自由業者）は，収入総額，実収入，実収入以外，繰入金，非消費支出，繰越金のデータが存在しないため，本研究では使用していない。

4.1　シミュレーション設計

4.1.1　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。一方，ブートストラップでは，標本サイズ n の観測データから復元抽出によってサイズ n の再標本（resample）を無作為抽出するが，これは真のモデルに近いと期待される推定モデルから得られたサイズ n の標本である。すなわち，Politis et al.（2001, p.1106）は，サブサンプリングの標本は真の母集団からの正しい標本（サイズは正しくない）である一方，リサンプリングの標本は真の母集団からの正しくない標本（サイズは正しい）の可能性があると指摘している。

そこで，本稿では，表６の職業のデータから欠測値を含む行を除去して得られた観測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。

表６　擬似母集団の一覧

職業符号番号	職業	標本サイズ
1	常用労務者	6235
3	民間職員	7655
4・5	官公職員	3552
12	無職	7254

サブサンプリング分析では，分析者が任意で副標本サイズ b を決めなければならない。しかし，b が n に近すぎると，すべての副標本統計量 $\hat{\theta}_b$ は標本統計量 $\hat{\theta}_n$ とほぼ変わらなくなり，副標本分布が過度に狭くなって，信頼区間を過小推定してしまう。一方，b が小さすぎると，信頼区間を過小推定または過大推定することになる。Politis et al.（2001）は，具体的な副標本サイズの決め方を示していないが，Di Zio & Guarnera（2013, pp.548-549）及び栗原（2015，p.7）では，約1/5から1/6のサイズが採用されているため，本稿も前例にならって，表６の擬似母集団から1,000回の非復元抽出によるサブサンプリングにて副標本サイズ $n/5$ の副標本データを生成する。

4.1.2　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３種類の方法で欠測を発生させた[14]。MARでは，被説明変数（食費）を条件として，乱数に基づいて，各々の説明変数を欠測させた。MCARでは，乱数のみに基づいて，各々の説明変数を欠測させた。NMARでは，各々の説明変数自体を条件として，乱数に基づいて，各々の説明変数を欠測させた。MARとNMARは，種類の差ではなく程度の差であり（Graham, 2009, p.567；高橋・渡辺，2017，p.21），これは弱MARとして理解できる[15]。各変数の欠測率は，実データにおける各変数の欠測率に対応させ，それぞれ１％から14％である。データ全体の欠測率は，約４割である。

4.2　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

第３章と同様に，(1)式の重回帰モデルにおける β_1 を推測の対象とする[16]。(2)式の偏り（bias），(3)式の二乗平均平方根誤差（RMSE: root mean squared error），(4)式の比率の標準誤差に基づく95％信頼区間のカバー率の範囲[17]を評価方法として使用する。

$$\begin{aligned}\ln(\text{食費}) = {} & \beta_0 + \beta_1 \ln(\text{年間収入}) + \beta_2 \text{就業人員} \\ & + \beta_3 \ln(\text{住宅面積}) + \beta_4 \ln(\text{年齢}) \\ & + \beta_5 \ln(\text{住居}) + \beta_6 \ln(\text{光熱・水道}) \\ & + \beta_7 \ln(\text{家具・家事}) + \beta_8 \ln(\text{被服・履物}) \\ & + \beta_9 \ln(\text{保健医療}) + \beta_{10} \ln(\text{交通・通信}) \\ & + \beta_{11} \ln(\text{教養娯楽}) + \beta_{12} \ln(\text{その他消費}) \\ & + \beta_{13} \ln(\text{非消費支出}) + \beta_{14} \ln(\text{実支出以外}) \\ & + \beta_{15} \ln(\text{繰越金}) + \beta_{16} \ln(\text{貯蓄現在高}) + \varepsilon_i\end{aligned} \tag{1}$$

$$\text{Bias}(\hat{\beta}) = E(\hat{\beta}) - \beta \tag{2}$$

$$\text{RMSE}(\hat{\beta}) = \sqrt{E(\hat{\beta} - \beta)^2} \tag{3}$$

$$\text{SE}(\pi) = \sqrt{\frac{\pi(1-\pi)}{s}} \tag{4}$$

これらの評価手法を用いて，完全データ（欠測を人工的に発生させる前のデータ），リストワイズ除去によるデータ（欠測値を含む行全体を除去），０置き換えデータ（欠測値を０で置き換え），単一代入法によるデータ（欠測値を重回帰モデルによる単一の予測値で置き換え），多重代入法によるデータ（Rパッケージ Amelia II を用いてEMBアルゴリズムによって多重代入法を実行）の評価を行う。

4.3　シミュレーションによる検証結果

表７はMARの検証結果，表８はMCARの検証結果，表９はNMAR（弱MAR）の検証結果である。いずれの分析においても，多重代入済みデータセット数は100に設定している。太字の結果は，パフォーマンスの悪い結果を示している。

多重代入法による結果は，すべての場合に

表7　MARの検証結果

評価方法	手法	職業1	職業3	職業4・5	職業12
偏り（＜0.005）	完全データ	0.001	0.000	0.001	0.000
	リストワイズ	**－0.019**	**－0.027**	**－0.014**	－0.004
	0置き換え	**－0.163**	**－0.172**	**－0.186**	**－0.131**
	単一代入法	**0.009**	**0.008**	**0.013**	**0.009**
	多重代入法	－0.001	－0.003	0.002	0.001
RMSE	完全データ	0.030	0.028	0.049	0.023
	リストワイズ	**0.040**	**0.043**	**0.062**	**0.028**
	0置き換え	**0.163**	**0.172**	**0.186**	**0.131**
	単一代入法	0.035	0.032	0.057	**0.028**
	多重代入法	0.031	0.029	0.051	0.024
95%カバー率（93.6~96.4）	完全データ	94.6	96.0	95.0	**92.6**
	リストワイズ	**92.3**	**87.8**	94.8	**92.2**
	0置き換え	**0.0**	**0.0**	**0.0**	**0.0**
	単一代入法	**92.1**	**93.3**	**92.1**	**89.1**
	多重代入法	95.0	95.8	94.8	**93.0**

表8　MARの検証結果

評価方法	手法	職業1	職業3	職業4・5	職業12
偏り（＜0.005）	完全データ	0.003	0.001	0.001	0.002
	リストワイズ	0.004	0.001	0.002	0.003
	0置き換え	**－0.128**	**－0.138**	**－0.153**	**－0.100**
	単一代入法	**0.013**	**0.010**	**0.013**	**0.011**
	多重代入法	0.003	0.001	0.003	0.002
RMSE	完全データ	0.029	0.027	0.049	0.024
	リストワイズ	**0.042**	**0.039**	**0.073**	**0.035**
	0置き換え	**0.129**	**0.138**	**0.153**	**0.100**
	単一代入法	0.036	0.033	0.058	0.030
	多重代入法	0.031	0.029	0.052	0.025
95%カバー率（93.6~96.4）	完全データ	95.8	95.7	94.3	**92.5**
	リストワイズ	95.5	95.2	93.7	**89.9**
	0置き換え	**0.0**	**0.0**	**0.0**	**0.0**
	単一代入法	**91.9**	**92.9**	**91.2**	**86.9**
	多重代入法	95.4	95.4	94.3	**92.2**

表9　NMAR（弱MAR）の検証結果

評価方法	手法	職業1	職業3	職業4・5	職業12
偏り（＜0.005）	完全データ	－0.002	－0.001	0.000	0.001
	リストワイズ	**－0.005**	－0.004	0.004	**0.013**
	0置き換え	**－0.138**	**－0.151**	**－0.160**	**－0.105**
	単一代入法	**0.008**	**0.011**	**0.010**	**0.010**
	多重代入法	－0.002	0.000	－0.000	0.002
RMSE	完全データ	0.030	0.027	0.048	0.024
	リストワイズ	**0.043**	**0.040**	**0.071**	**0.035**
	0置き換え	**0.138**	**0.151**	**0.160**	**0.105**
	単一代入法	0.034	0.033	0.056	0.029
	多重代入法	0.031	0.029	0.051	0.025
95%カバー率（93.6~96.4）	完全データ	96.1	96.5	96.0	**93.3**
	リストワイズ	94.4	94.8	94.0	**88.5**
	0置き換え	**0.0**	**0.0**	**0.0**	**0.0**
	単一代入法	**93.3**	93.7	**93.5**	**87.7**
	多重代入法	95.6	96.1	95.5	**93.1**

おいて偏っていない。また，RMSEを基準とした場合，最もパフォーマンスがよい。職業12（無職）を除いて，すべての場合において，95％信頼区間のカバー率も93.6～96.4％の範囲内に入っている。職業12の検証では，無職のため年間収入が極端に少ない世帯が外れ値として存在しており，完全データの結果にも悪影響が出ている。

リストワイズ除去は，MCARの場合は偏っていないものの，MARの場合に偏りが著しい。また，RMSEを基準とした場合，すべての場合において2番目にパフォーマンスが悪い。MARの場合，95％信頼区間のカバー率は，93.6～96.4％の範囲内に入っていない。

単一代入法は，RMSEを基準とした場合は次点であるものの，偏りが大きい。単一代入法は，標準誤差が過小となっているだけではなく，被説明変数を用いて欠測値の予測を行ったにも関わらず，その事実を反映していないため，被説明変数の情報を二重で活用していることによって偏りが発生している（van Buuren, 2012, p.62）。また，95％信頼区間のカバー率は，93.6～96.4％の範囲内に入っていない。

欠測値を0で置き換える手法は，著しく偏っており，すべての基準において最もパフォーマンスが悪い。匿名データにおける0の値は，注意深く精査して，欠測値と推認されるものについては，適切に処理する必要がある。

したがって，偏り，効率性，信頼区間のカバー率といったすべての基準から，匿名データの欠測値は，多重代入法によって処理し直した上で分析を実行することが望ましい。

5．結語

本稿では，全国消費実態調査の匿名データを用いて，欠測データを使用する際の注意点と対処法について論じた。匿名データは，集計値ベースの個票データから作成されているため，欠測値は，合計値などを集計することを前提として処理されている。すなわち，確定的単一代入法によって処理されている。本稿は，回帰分析などの統計的推測を行う場合，単一代入法によって処理されたデータをそのまま用いたり，欠測値を除去したりするだけでは十分ではないおそれがあることを示した。この問題は，匿名データを使用する分析者が，多重代入法により欠測値を処理し直すことで解決できることを示し，トップコーディングの問題もベイズの事前分布を活用することで，多重代入法により解決できることも示した。

謝辞

本稿は，経済統計学会第61回全国研究大会（2017年9月）の企画セッション「政府統計ミクロデータの作成技法に関する諸問題」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，2名の査読者から有益なコメントをいただき本稿を改善することができた。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）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

2）MARでは，観測データを条件とした場合の欠測確率が，データを条件とした欠測確率に一致する。たとえば，大学入試と大学での成績の関連を考えるとき，大学入試の合格者については入学後の大学の成績の欠測確率は0％であるが（簡単のため，休学と退学はないものとする），大学入試の不合格者については入学後の大学の成績の欠測確率は100％である。大学入試の成績という観測データを条件とした場合，合格者と不合格者のそれぞれのグループ内では，一定の確率で欠測が発生しており，このような状況をMARとよぶ。

3）本稿では，多重代入法による欠測データを用いた統計分析について扱っている。詳しくは，高橋・渡辺（2017）を参照されたい。また，モデルに基づく尤度解析法による欠測データを用いた統計分析については，阿部（2016，pp.62-92）及び高井・星野・野間（2016，pp.23-101）を参照されたい。

4）総務省統計局（2004a）「調査のねらい」を参照されたい。
http://www.stat.go.jp/data/zensho/2004/pdf/01nerai.pdf

5）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。

6）総務省統計局（2004b）「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

7）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

8）観測値に関する事前分布とは，特定の世帯についての情報を活用するものである。なお，観測値に関する事前分布は，Schafer（1997，pp.155-157）によって導入されているリッジ事前分布とは異なる。リッジ事前分布では変数間の共分散を0に近づけることでモデルの安定性を得ようとするものであるが，観測値に関する事前分布は各々の観測値に関する平均値と標準偏差を指定することで，切断分布からの代入を行うものである。詳細は，高橋・渡辺（2017，pp.164-166）を参照されたい。

9）MCARでは，ある値の欠測する確率がその値と無関係である。たとえば，ルーレットをまわして偶数が出たら回答し，奇数が出たら回答しないとすれば，ある値が欠測する確率はその値と無関係であり，完全に無作為なMCARである。

10）この章の目的は，具体的な分析において結論が変わり得ることを示すことである。その意味では，「二人世帯の母子家庭」以外のどのような例でもよいが，シングルマザーの子育てにおいて，食費という生活の中で最も重要な要素の1つに関して，母親の収入がどのような影響を及ぼすかということは，政策的なテーマとしても重要であると考え，この例を採用した。

11）本稿では，Rパッケージ Amelia II を用いたが，このアルゴリズムは適切な多重代入法（proper multiple imputation）である（Takahashi, 2017）。また，すべての変数が量的変数の場合，データ拡大法（data augmentation）によるRパッケージ norm と完全条件付き指定（fully conditional specification）によるRパッケージ mice のいずれを用いてもよい（高橋・渡辺，2017，pp.69-71）。一方，質的な変数を含む場合は，Rパッケージ mice がよく使用される。Rパッケージ mice による多重代入法については，野間（2017）も参照されたい。

12）この例では，年間収入の影響力は劇的に変化しているが，他の統制変数については被服・履物と教養娯楽が1％水準で結果が変わる以外の影響はない。しかし，二人世帯高齢者の分析など，統制変数の結果が大きく変わる例もあり，どの変数の値がどのような影響を受けるかは，一律に決まるわけではないことに注意が必要である。

13）NMARでは，ある値の欠測確率がその値自体に依存しており，かつ，観測データを条件としても欠測を無視できない。たとえば，身長に関するデータにおいて，身長が低い人ほど欠測が多く発生する場合，データ内に身長の欠測確率を予測できる情報がなければ，ある人の身長の値の欠測確率がその人の身長の値自体に依存しており，かつ，観測データを条件としても欠測を無視できず，NMARである。

14）3種類の欠測メカニズムについては，高橋・渡辺（2017，pp.15-21）を参照されたい。

15）ここでいう弱MARとは，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，上述したMARとNMARの定義の中間に位置する状態である。たとえば，身長に関するデータにおいて，身長が低い人ほど欠測が多く発生する場合，体重などの情報が完全に観測されていれば，身長の欠測確率をある程度の精度で予測できると考えられる。本稿では，この状態を弱MARとよんでいる。

16）β_1の真値は，0.142（常用労務作業者），0.149（民間職員），0.161（官公職員），0.108（無職）である。

17）ここで，πは比率，sはシミュレーション回数を表している。95％信頼区間のカバー率とは，名目で95％の信頼区間が真のパラメータの値を捕らえることができた割合のことである。なお，1,000回のシミュレーションにおける95％信頼区間のカバー率の範囲は，$\sqrt{0.95\times0.05/1000}\approx0.007=0.7\%$であるため，93.6％から96.4％の範囲（$95\pm2\times0.7$）に入っていれば，統計的に正しい結果といえる（高橋・渡辺，2017，p.22）。

参考文献

阿部貴行（2016）『欠測データの統計解析』，朝倉書店.

上村敏之・足立秦美・金田隆幸（2016）「女性の労働供給と保育料軽減政策」，『経済学論究』第69巻，第4号，pp.17-39.

魚住龍史（2014）「SASによる匿名データ分析：パック旅行費支出と世帯情報の関連の検討」，公的統計のミクロデータ等を用いた研究の新展開，統計数理研究所.

木村和範（2012）「所得格差変動の年齢階級別要因分解：全国消費実態調査ミクロデータを用いて」，『季刊北海学園大学経済論集』第59巻，第4号，pp.1-37.

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』第108号，pp.1-14.

坂田幸繁（2006）「個票データと統計利用」，『統計学』第90号，pp.31-42.

猿山純夫・服部哲也・松岡秀明・落合勝昭（2013）「農業保護はどの程度家計負担を増やしているか：個票データを用いた主要6品目の影響推計」，『JCER Discussion Paper』第140号，pp.1-24.

高井啓二・星野崇宏・野間久史（2016）『欠測データの統計科学：医学と社会科学への応用』，岩波書店.

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.

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.

田村英朗・松林洋一（2015）「所得不確実性と家計消費：「全国消費実態調査」に基づく計量分析」，『神戸大学経済学研究科Discussion Paper』第1516号，pp.1-23.

ディミトリス＝バラス・ダニー＝ドーリング・中谷友樹・ヘレナ＝タンストール・花岡和聖（2012）「英国と日本における社会格差：2つの島嶼経済・社会の比較研究に向けて」，『季刊社会保障研究』第48巻，第1号，pp.46-61.

野間久史（2017）「連鎖方程式による多重代入法」，『応用統計学』第46巻，第2号，pp.67-86.

野村総合研究所（2013）「統計データの補完推計に関する調査：報告書」，平成24年度内閣府大臣官房統計委員会担当室請負調査.

花岡和聖（2012）「公的統計「匿名データ」を用いた小地域単位での地理空間分析の可能性：空間的マイクロシミュレーションによる地理的な合成ミクロデータの生成」，『人文地理』第64巻，第3号，pp.195-211.

平山洋介（2011）「持家取得における既婚女性の就業の役割」，『日本建築学会計画系論文集』第76巻，第663号，pp.983-992.

増田幹人（2015）「子ども数と教育費負担との関係」，『季刊社会保障研究』第51巻，第2号，pp.223-232.

吉川直樹・天野耕二・島田幸司（2011）「人口・世帯構造変化を考慮した日本における食料消費に伴う環境負荷のシナリオ分析」，『環境情報科学論文集』第25巻，pp.125-130.

渡辺久里子（2013）「等価尺度の推計と比較：消費上の尺度・制度的尺度・OECD尺度」，『季刊社会保障研究』第48巻，第4号，pp.436-446.

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

Di Zio, M. & Guarnera, U.（2013）, "A Contamination Model for Selective Editing," *Journal of Official Statistics* Vol. 29, No. 4, pp.539-555.

Graham, J.W.（2009）, "Missing Data Analysis: Making It Work in the Real World," *Annual Review of Psychology* Vol. 60, pp.549-576.

Higa, K.（2013）, "Estimating Upward Bias in the Japanese CPI Using Engel's Law," *Global COE Hi-Stat Discussion Paper Series* No. 295, pp.1-22.

Honaker, J. & King, G.（2010）, "What to do about Missing Values in Time Series Cross-Section Data," *American Journal of Political Science* Vol. 54, No. 2, pp.561-581.

Honaker, J., King, G. & Blackwell, M.（2011）, "Amelia II: A Program for Missing Data," *Journal of Statistical Software* Vol. 45, No. 7, pp.1-47.

Politis, D.N., Romano, J.P., & Wolf, M.（2001）, "On the Asymptotic Theory of Subsampling," *Statistica Sinica* Vol. 11, No. 4, pp.1105-1124.

Schafer, J.L.（1997）, *Analysis of Incomplete Multivariate Data*, Chapman & Hall/CRC.

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.

Takahashi, M.（2017）, "Statistical Inference in Missing Data by MCMC and Non-MCMC Multiple Imputation Algorithms: Assessing the Effects of Between-Imputation Iterations," *Data Science Journal* Vol. 16, No. 37, pp.1-17.

van Buuren, S.（2012）, *Flexible Imputation of Missing Data*, Chapman & Hall/CRC.

Wooldridge, J.M.（2009）, *Introductory Econometrics: A Modern Approach*（4th edition）, South-Western.

The Improvement of Analyses based on Anonymized Microdata by Multiple Imputation: An Illustration using the Anonymized Microdata of the National Survey of Family Income and Expenditure

Masayoshi TAKAHASHI*

Summary

Since the "Basic Act on the Advancement of Public and Private Sector Data Utilization" came into force, the secondary use of the official statistics data has been advanced in such a way that the anonymized microdata are now available for academic analyses. However, in official statistics, where data are collected through survey questionnaires, it is rare to obtain complete data; thus, imputation is utilized in order to tackle the issue of missing values in official statistics. From a data user's perspective, this article discusses how missing data would affect the conclusions made in the analyses using imputed microdata. Specifically, this article seeks an estimation method to take bias into account, utilizing the anonymized microdata of the National Survey of Family Income and Expenditure. Furthermore, this article examines the impact, on the analyses, of missing data treatments such as listwise deletion, single imputation, and multiple imputation, by way of subsampling simulations based on the anonymized microdata of the National Survey of Family Income and Expenditure.

Key Words

Multiple imputation, single imputation, missing data, official statistics, anonymized microdata

* IR Office, Tokyo University of Foreign Studies

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

2016年9月より，新しい規程にもとづいて，「研究論文」と「報告論文」が設定されました。皆様からの積極的な投稿をお待ちしております。

1．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については，下記の［注記2］をご確認下さい。
2．次号以降の発行予定日は，
　第115号：2018年9月30日，第116号：2019年3月31日です。
3．投稿に際しては，新規程にもとづく「投稿規程」，「執筆要綱」，および「査読要領」などをご熟読願います。最新版は，学会の公式ウェブサイトをご参照下さい。
4．原稿は編集委員長（下記メールアドレス）宛にお送り下さい。
5．原稿はPDF形式のファイルとして提出して下さい。また，紙媒体での提出も旧規程に準拠して受け付けます。紙媒体の送付先は編集委員長宛にお願いします（住所は会員名簿をご参照下さい）。
6．原則として，すべての投稿原稿が査読の対象となります。
7．投稿から発刊までに要する期間は，通常3ヶ月以上を要します。投稿にあたっては十分に留意して下さい。

編集委員会，投稿応募についての問い合わせは，
下記メールアドレス宛に連絡下さい。
また，編集委員長へのメールアドレスも下記になります。

editorial@jsest.jp

来年度（2018年度）の編集委員は，つぎのとおりです。
編集委員長　水野谷武志（北海学園大学）
副委員長　池田　伸（立命館大学）
編集委員　小林良行（総務省統計研究研修所）
松川太一郎（鹿児島大学）
山田　満（東北・関東支部所属）

［注記1］『統計学』の定期刊行に努めておりますので，できるかぎり早期のご投稿をお願いします。115号（2018年9月30日発行予定）への掲載を想定した場合，「研究論文」と「報告論文」の原稿は，2018年6月初旬を目途として，それまでにご投稿ください。
［注記2］「研究論文」と「報告論文」は，別個に査読し，区分を変更しません。区分につきましては自分で申告して投稿しますが，この点ご留意ください。
［注記3］書評，資料および海外統計事情等について，執筆，推薦，および依頼等をお考えの会員がおられましたら，企画や思いつきの段階で結構ですので，できるだけ早い段階で，編集委員会にご一報下さい。

以上

編集後記

研究成果を投稿くださいました皆様，査読に関わってくださいました皆様に心よりお礼申し上げます。さて年度変わって次号115号より，水野谷編集委員長のもとで本誌が作成されます。編集委員会では機関誌『統計学』を充実させていくために，皆様からの率直な意見と，研究成果の積極的な投稿をお待ちしています。今後ともよろしくお願い申し上げます。（藤井輝明　記）

執 筆 者 紹 介

長屋政勝　（京都大学名誉教授）　　高橋将宜　（東京外国語大学経営戦略情報本部）

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　No.114

2018年3月31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　西村善博

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. 114 2018 March

Articles

Activities of the Society

JAPAN SOCIETY OF ECONOMIC STATISTICS