【『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118号 2020年3月）

個票データの解析的利用と抽出ウェイトの役割

坂田幸繁*

要旨

標本調査からの調査票情報，あるいは個票データをモデル解析に利用する場合に，標本設計情報，とりわけ抽出ウェイトをどのように処理すべきかを検討した。そのさい，利用者サイドが有する理論的・分析的視点は統計作成者のそれとは通常異なる点に配慮し，標本調査データの本来的な母集団記述統計量的性格をモデル解析（超母集団分析）に接合しようとするデザイン一致推定量の考え方に主に依拠しながら，回帰モデルでの単純推定と加重推定方式との特性比較を単純なシミュレーションで確認した。その結果，モデルの正しさを想定することが困難な2次利用の局面ではとくに，モデル解析において抽出ウェイトによる加重推定を戦略的に重視すべきことを論点提起した。

キーワード

個票データ，2次利用，モデルパラメータ，超母集団モデル，抽出ウェイト

1．はじめに

本稿では，抽出率が異なる層化標本を事例に，公的標本統計における調査票情報，あるいは個票データセット（以下では，ミクロデータとも呼ぶ）に基づくモデル解析のための抽出ウェイト[1)]（抽出率の逆数とする）の利用について，統計利用者としての立場からの具体的な指針を検討したい。改めて指摘するまでもなく，実際のミクロデータ解析では，標本が抽出された当該の実在する有限母集団，あるいはその関心のある部分母集団の大きさを推定するような場合には抽出ウェイトによる復元が必須の作業と意識されるが[2)]，他方で多変量の統計的関連を確率モデルで表現し，その回帰係数などの推定に変数間の構造をとらえようとする場合，抽出ウェイトの取扱いには少なからぬ曖昧さや混乱がみられる。とくに構造把握に多用される回帰分析では，正しいモデルを含む関連する変数をすべて導入しておけば，標本設計も含めて様々な要因をコントロールできるので，ウェイトを考慮せずとも真の係数の推定値が得られるとの考え方もある。あるいは，任意の母集団要素に成立するモデルを探しているのだから，抽出ウェイトによる母集団の歪みの調節は不要という考え方も一つの主張である[3)]。

このような標本データの利用に関する問いに対して，わが国に先行すること1980年代にはミクロデータの利用を開始した欧米では，すでに課題の枠組みが整理され，問題への解法の提示や事例の蓄積が進められている。とりわけ1980年代に調査票情報の分析と並行

* 正会員，中央大学経済学部

しながら集中的に解析法の議論が深められており，Skinner, Holt, and Smith（1989）の“Analysis of Complex Survey”，およびKasprzyk, Duncan, Kalton and Singh（1989）の“Panel Surveys”は今日に続く標本調査データに関する方法的なフレームワークを提供し，標本調査情報の利用に関する議論の中核を示すものと位置付けられる。またPfeffermann（1993）では，本稿の主たる関心事でもあるモデル分析における抽出ウェイトの役割に関して直截的，包括的なサーベイを与えている。そして，Chambers and Skinner（eds）（2003）は先出Skinner, Holt, and Smith（1989）の現代的な更新であり，新たな展開を含む補完的労作といえる。なお，最尤法に限ってみると，Breckling, Chambers, Dorfman, Tam, and Welsh（1994），およびその発展でもあるChambers, Steel, Wang, and Welsh.（2012）は標本調査情報からの最尤推論に関する議論の現代的な拡張を試みる業績といえる。いずれにしても応用的な事例研究を含めて，議論の核心となる方法論や概念的な図式はすでに開陳されているといっても過言ではない[4)]。

しかしそれら全体の適切な整理や総括的把握には，本稿の一論考で収まるべくもなく，ミクロデータの利用をめぐってさらに追加的な考察と学会での議論を積み重ねる必要がある。本稿では，60周年記念事業の本来の趣意にそって，標本調査情報の利用に関わる本学会でのさらに進んだ議論の先触れ[5)]として，もっとも単純なモデルのパラメータ推定に関わる抽出ウェイトの取扱いの指針を，これら先行する業績の一端をシミュレーションベースで追体験することにより提起しようとしている。とりわけ，本学会では仮説演繹的な計量分析よりむしろ社会研究にみられる帰納的発見的アプローチからミクロデータを利用するケースも多い。そのさいモデルの正しさを出発点としないときの抽出ウェイトの利用の是非は，ミクロデータの統計利用者には解決すべきハードルである。

問題の本質を複雑化せず提起するために，本稿では乗率形態で抽出ウェイトが付与された標本データだけが利用可能とする[6)]。いわば，ミクロデータに基づく2次利用の形態のうちもっとも頻度の高い単独利用に限定する。標本調査の結果としての調査票情報，あるいは個票データセット以外の情報は何も利用できない状況を想定している。したがって無回答などの回答構造の組み込みや他の補助的な母集団情報の利用は考えない[7)]。むしろ標本設計に利用された層化変数やクラスター情報の一部，もしくは全部が，匿名化のため制約されたデータ環境を対象とする。

次節以降，そのときの抽出ウェイトに対する基本的なアプローチの方法と考え方をPfeffermann（1993）によるサーベイ論文のガイドラインに依拠しつつ再提起し[8)]，3，4節でシミュレーション結果を示し，その特徴を確認しながら，作成者≠利用者という本来の批判統計の視点に重きをおいて本学会としての課題を改めて考えてみたい。結論的には，上記のように限定したモデル分析において本稿は，抽出ウェイトを利用した加重推定によるアプローチの有用性を主張している。

2．モデルベースのアプローチの特徴

実在の有限母集団（サイズN，データは所与の固定数値）に対して，例えば線形回帰式$y=A+Bx+u$（uは残差）に対して最小2乗法基準であてはめたときのxとyの回帰的関係を表す要約統計量Bをその確率標本（サイズn）から推定しようとする。このようなアプローチをデザインベースのアプローチとよび，母集団要素の変数値x，yの関数であるBをセンサスパラメータという。これに対して，確率モデル$Y=\alpha+\beta X+\varepsilon$（$\varepsilon$は攪乱項で平均0，分散$\sigma^2$の正規分布）を想定し，有限母集団はこのモデルから発生した確率変数の実現値集合と考える。そしてそのような母集団から，

ある標本デザインに従って確率抽出したサンプルに基づきモデルのパラメータ α，β を推定しようとする。このケースをモデルベースのアプローチといい，ターゲットの α，β を改めてモデルパラメータと呼ぶ[9]。線形回帰モデルとロジスティック回帰モデルを具体例に想定して，その特徴を整理しよう。まずデザインベースのアプローチを簡単に整理したうえで[10]，モデルベースのアプローチを検討する。

2.1 デザインベースのアプローチとセンサスパラメータ

(1) 標本推定方程式とセンサスパラメータ

実在の有限母集団の関係する変数値を $\mathrm{y}_U=(y_1, \cdots, y_N)'$, $\mathrm{x}_U=(x_1, \cdots, x_N)'$ とする。これに最小 2 乗基準で線形回帰式をあてはめるとき，A，B は母集団要素の当該値（母集団データと呼ぶ）の関数として次のように定義される。$\bar{y}_U$，$\bar{x}_U$ は母平均である。

$$\mathrm{B}=(\Sigma_{t=1}^{N} x_t y_t - N\bar{x}_U\bar{y}_U)/(\Sigma_{t=1}^{N} x_t^2 - N\bar{x}_U^2),$$
$$\mathrm{A}=\bar{y}_U - \mathrm{B}\bar{x}_U$$

いま当該母集団からのある抽出デザインのもとでの標本要素の変数値（以下，標本データと呼ぶ）から母数A，B を推定する。これがデザインベースのアプローチであり，標本調査のテキストに示されるような母平均推定と同じく，母集団データの関数としての記述的特性値を標本データから推定する問題に帰着する。通常，ある母集団要素 t が標本 S に包含される確率 $\pi_t=P(t\in S)$ はその標本への抽出確率を表し，その逆数 $w_t=1/\pi_t$ が抽出ウェイトである。標準的な標本調査の問題であるから，抽出標本 $\mathrm{y}_S=(y_1, \cdots, y_n)'$, $\mathrm{x}_S=(x_1, \cdots, x_n)'$ からのセンサスパラメータの推定値は，標本要素に i を割り当て，

$$\hat{\mathrm{B}}=(\Sigma_{i=1}^{n} w_i x_i y_i - N\widehat{\bar{x}_U}\widehat{\bar{y}_U})/(\Sigma_{i=1}^{n} w_i x_i^2 - N\widehat{\bar{x}_U^2}),$$
$$\hat{\mathrm{A}}=\widehat{\bar{y}_U} - \hat{\mathrm{B}}\widehat{\bar{x}_U}$$

と書ける。ここで，$\widehat{\bar{x}_U}=\Sigma_{i=1}^{n} w_i x_i / \Sigma_{i=1}^{n} w_i$，$\widehat{\bar{y}_U}=\Sigma_{i=1}^{n} w_i y_i / \Sigma_{i=1}^{n} w_i$ である。

このような推定ルール（いまは最小 2 乗法）を手続き的に表現すると，母集団レベルでは次式の正規方程式のように，最小 2 乗基準で残差 2 乗和の偏微分を 0 とおくことに等しい。

$$\Sigma_{t=1}^{N}[y_t - (\mathrm{A}+\mathrm{B}x_t)]=0,$$
$$\Sigma_{t=1}^{N} x_t[y_t - (\mathrm{A}+\mathrm{B}x_t)]=0$$

ここで改めて重回帰モデルに拡張し，定数項を含めてデータ行列 x，y とパラメータベクトル B を定義すると，上記の考え方はつぎの母集団推定方程式 G(B) として再表現できる。

$$\mathrm{G}(B)=\mathrm{x}_U'\mathrm{y}_U - \mathrm{x}_U'\mathrm{x}_U B=0$$

これを標本データで表現すると，

$$\mathrm{G}_S(B)=\hat{\mathrm{G}}(B)$$
$$=\mathrm{x}_S W_S \mathrm{y}_S - \mathrm{x}_S' W_S \mathrm{x}_S B=0$$

$W_S=diag(w_1, \cdots, w_n)$ は抽出ウェイト行列である。$\mathrm{G}_S(B)$ は，母集団推定方程式 G(B) に対して加重標本推定方程式となる。センサスパラメータ B とその推定量（解）$\hat{B}_W$ は次式で与えられる（Pfeffermann 1993, p.318）。なお x_t，x_i はそれぞれ母集団要素 t，標本要素 i の説明変数ベクトルである。

$$B=(\mathrm{x}_U'\mathrm{x}_U)^{-1}\mathrm{x}_U'\mathrm{y}_U=\left(\sum_{t\in U} x_t x_t'\right)^{-1}\sum_{t\in U} x_t y_t$$
$$\hat{B}_{\mathrm{W}}=(\mathrm{x}_S' W_S \mathrm{x}_S)^{-1}\mathrm{x}_S' W_S \mathrm{y}_S$$
$$=\left(\sum_{i\in S} w_i x_i x_i'\right)^{-1}\sum_{i\in S} w_i x_i y_i$$

モデルに準じた関数式を全数データに当てはめ，そのときの全数データの関数として表現される係数（センサスパラメータ）解を標本データから推定する問題は，モデルの当てはめルールを母集団に対して表現した母集団推定方程式を，標本データによる推定式としての標本推定方程式として再設定し，その解

を求める問題として一般化される[11]。いまは線形回帰モデルと最小2乗基準ルールを例示したが，当然，これを拡大して，ロジスティック回帰などの非線形モデルの当てはめと尤度ベースの推定ルールに関しても同様に定めることができる。

(2) 擬似尤度

母集団 $t=1,\cdots,N$ の x が与えられたときの y 値が条件付き確率密度関数 $f_U(y_t;x_t,B)$ に独立に従うと仮定して，母集団データにこのモデルを当てはめたときのセンサスパラメータ B は次の母集団推定方程式の解で与えられると考えればよい。例えば $y=\{0,1\}$ としてロジスティック回帰モデル $\Pr\{y_t=1\}=\exp(x_t'B)/[1+\exp(x_t'B)]$ を考えると，母集団データに対応するこのモデルの尤度を $f_U(\mathrm{y}_U;\mathrm{x}_U,B)$ とおいてその対数尤度関数 $l(B)=\sum_{t=1}^{N}\log f_U(y_t;x_t,B)$ を最大にするように B を定めればよい。それはふつう，母集団尤度方程式から，

$$\mathrm{G}(B)=\partial l(B)/\partial B=\sum_{t=1}^{N}\partial\log f_U(y_t;x_t,B)/\partial B=0$$

の解で定義され，それを標本データから推定するには，先ほどと同じく，尤度方程式で与えられる母集団推定方程式を標本データから推定すればよい。標本推定方程式は次式である。

$$\mathrm{G}_S(B)=\sum_{i=1}^{n}w_i\partial\log f_S(y_i;x_i,B)/\partial B=0$$

このような推定量を擬似最大尤度推定量（Pseudo Maximum Likelihood Estimator）という。形式的に対数尤度の加重和を最大化するように推定値 $\hat{B}_w$ が求められる[12]。しかし，母集団の推定方程式に対応するセンサスパラメータ（母集団記述統計量）を標本から推定方程式を通して求めていることになる。y の母集団分布やモデルの分布形とは無関係に，標本抽出による確率変動に対して，ターゲットである既知の母集団データの関数（センサスパラメータ）について不偏性などの望ましい推定量の性質を満たそうとする点にデザインベースの推定の特長がある。

2.2 モデルベースのアプローチと2次利用

モデルパラメータの推定を目的とするモデルベースのアプローチにおいては，まず母集団は先のモデル $\mathrm{Y}=\alpha+\beta\mathrm{X}+\varepsilon$（$\varepsilon$は攪乱項で平均0，分散 σ^2 の正規分布）を満たす確率変数 Y, X（あるいは超母集団 superpopulation）のサイズ N の実現値集合であり，そこからある特定の抽出デザインに従って確率抽出されたサイズ n のデータが標本データ（$\mathrm{y}_S,\mathrm{x}_S$）となる。ここでの目標は，標本データからモデルパラメータ（α,β）を求めることにある。デザインベースの母集団記述統計量（DPQ；descriptive population quantity）の推定では，抽出ウェイトで母集団に戻すことが必須であったが，ここではモデルパラメータを推定したいのであるから，抽出ウェイトの利用は不可避というわけではなく，先述のように特定の母集団を超えたより一般的な法則をモデルの確率分布として求める立場もあれば，逆にそのようなモデルは実際の母集団では妥当せず，センサスにおける，例えば最小2乗解である母集団記述統計量を目標パラメータと考える立場もある。後者は，標本調査データによるモデル解析的利用には消極的ともいえる[13]。

Pfeffermann（1993）において第3の立場として提唱するのは，調査データとモデル解析を両立（妥協）させる2次利用のアプローチである。そこでは，モデルからの実現値の分布としての確率的ゆらぎ（ξ）と母集団の下での抽出デザインに起因する確率的揺らぎ（p）の2つの変動がミックスされていることに留意が必要である。このアプローチを整理するために，キーとなる2つの基本概念，CDPQ（Corresponding Descriptive Population Quantity；対応母集団記述統計量），およびDC（Design Consistency；デザイン一致性）が導入さ

れる（Pfeffermann 1993, p.320）。

CDPQ：母集団は未知パラメータβをもつモデルに従う確率変数の実現値であり，ある推定ルール（例えば最小2乗距離基準）のもとでの母集団推定方程式の解$T(N)$をβに対するCDPQと呼ぶ。冒頭の線形単回帰モデルについてはCDPQが母集団記述統計量DPQであるBと一致するケースである。当然，想定されるモデルに応じて定義されるCDPQが，一般によく利用されるタイプのDPQと一致するとは限らない。

DC：サイズNの有限母集団からの標本統計量（サイズn）$t_S(n)$について，母集団と標本がそれらの構成を維持したままサイズを増加していくとき，その極限が母集団記述統計量$T(N)$に確率収束するとき，$t_S(n)$はデザイン一致性を有するという。つまり，

$$\mathrm{plim}_{n\to\infty,N\to\infty}[t_S(n)-T(N)]=0$$

2.3　デザイン一致推定量

まずモデルパラメータの推測をめざしながら，モデルの下での最適推定量を求めるのではなく，CDPQに対するデザイン一致性をもつ推定量のクラスから推定量を求めようとする。デザイン一致推定量に絞る理由として掲げるのは，下記に示す分析のロバストネス（頑健性）である（同，p.321）。

① 母集団において想定されたモデルが正しいとき

推定ルールによってモデルに対して一致性をもつCDPQが得られているとすれば，母集団サイズが大きくなればCDPQはモデルパラメータに収束する。したがって，CDPQに関する任意のデザイン一致推定量は標本抽出pとモデル由来の分布ξとの混合分布においてモデルパラメータに対する一致性をもつ。

② 想定されたモデルが正しくないとき

モデルパラメータや最適推定量といった概念は無意味であり，解釈困難である。しかし，CDPQにはモデルの有効性とは無関係に存在する実体があり，明確な解釈を有している。例えば先の線形回帰モデルの係数A，Bは最小2乗距離基準での有限母集団におけるY値の最良線形予測である。さらに，母集団値は（Y, X）の同時ξ分布からの確率変数と仮定しているから，DPQ（A, B）は間違ったモデルではあっても線形回帰係数（α, β）の一致推定量となる。

つまり，モデルの想定が正しくとも誤っていても，経験的に意味のある推定量を提供している点にその特長をみている[14]。

一般にCDPQのデザイン一致（DC）推定量がモデルパラメータθの一致推定量であることを次式が示している（同，p.321）。

$$\begin{aligned}t_S-\theta&=(t_S-T)+(T-\theta)\\&=O_p(n^{-\frac{1}{2}})+O_\xi(N^{-\frac{1}{2}})=O_p(n^{-\frac{1}{2}})\end{aligned}$$

ただし，確率的漸近オーダー$O_p(n^{-\frac{1}{2}})$は標本抽出変動pによるもの，$O_\xi(N^{-\frac{1}{2}})$はモデルによる確率変動ξによるものである。母集団サイズは十分大きいと考えるなら，t_Sは真値に確率収束する。先の回帰モデルの例では，$t_S=\hat{B}_W$ $T=B, \theta=\beta$と読替えればよい。

また，θまわりへのt_Sの分散は次のように分解される（同，p.321）。

$$\begin{aligned}\mathrm{Var}_{p\xi}(t_S)&=E_\xi[Var_p(t_S|Y)]+Var_\xi[E_p(t_S|Y)]\\&=E_\xi[Var_p(t_S|Y)]+O(N^{-1})\end{aligned}$$

したがって標本サイズより母集団サイズがはるかに大きいような通常のケースでは，本来のξとpとの混合分布のもとでの推定量t_Sの分散は，母集団値Yを所与としたときの標本抽出誤差のξ変動に関する期待値で近似でき，それは通常の標本誤差を推定すれば，推定誤差の近似が得られることを教えている[15]。

3．デザイン一致性と無視可能性

本稿が想定する2次利用の状況は，標本

データの単独利用である。そのため母集団に関して他の利用可能な情報があったとしても，デザイン一致推定量の考え方ではそれを推定に利用することができないという欠点がある。他方で，いまの議論の文脈では推定量の選択肢は多いわけではなく，詰まるところ抽出ウェイトを利用した加重推定か，ウェイトを無視した単純推定かの選択に帰着する。このとき考慮すべき要因に，モデルに対する抽出デザインの無視可能性（ignorability），あるいはデザインの無情報性（noninformability）がある。厳密な定義と議論はRubin（1976），Little（1982），Skinner, Holt, and Smith（1989）の第6，12章に詳しいが，無回答や欠測データの調整をめぐる近年の観察データの分析方法の要点でもあり，有用な文献[16]も多く刊行されていることからここではこれ以上立ち入らない。抽出ウェイトの利用とその社会統計学的含意に関心があるので，その幾分教科書的な単純な例示で先を進めることにしたい。

3.1 簡単なシミュレーション

線形単回帰モデル（$Y=\alpha+\beta X+\varepsilon$）を想定し，標本データは単純化して，目的変数，あるいは説明変数のいずれかで2層に区分され，各層に異なる抽出率を割り当てる（層内では単純無作為抽出）。まず想定した正しいモデル（仮説的無限母集団）から抽出（ξ分布）した*i.i.d.*データN=10,000を母集団とし，それを観測データの抽出枠として，そこから適当なサイズの層化標本（p分布）を複数回抽出し，各抽出データに対してウェイト無しの推定OLSと加重推定WLSを繰り返し（1,000回），設定した真値に推定結果がどのように一致するか（不偏性，一致性）を検証する。

抽出率でコントロールすることにより，$N\rightarrow\infty$のとき$n\rightarrow\infty$となり，デザイン一致性を満たす抽出の枠組みは満たされる。またモデルと層化デザインとの関係において，説明変数であるXで層化する標本は外生的層化データ，目的変数であるYで層化した場合内生的層化データと呼ばれる。結論的に言えば，いまの例では前者は無視可能な標本，無情報な標本抽出であり，パラメータ推定に抽出デザインは無視してよい。これに対して，後者はモデルに対して無視可能でない標本抽出，無情報でない標本であり，推定に際して抽出デザインを考慮する必要がある。

3.2 結果の解説

まずモデルが正しく特定され，それが線形回帰モデルであるとき，

① 外生的層化データに対してウェイトの利用は不要であり，
② 内生的層化データに対してはウェイトの何らか調整が必要である

ことは，切断データへの直線のあてはめを思い浮かべれば直感的に明らかである。例えば，$-\infty<X<+\infty$，$-\infty<Y<+\infty$においてX＞0，あるいはY＞0で切断したデータセット（抽出率≒0）からのOLS直線は，X＞0で切断したケースでは真の直線のよい近似を与えるが，Y＞0のケースでは真値からの大きな乖離が生じる。

実際，$Y=0.0+0.6X+\varepsilon$を真の回帰モデルとして発生させた母集団（図1-1）に対して，Xを層化変数にX＜0のとき抽出率0.01で，X≧0のとき抽出率0.2で構成した標本とそのOLS直線が図1-2に示されている。同様にYを層化変数に構成したサンプルとOLS直線が図1-3である。1,000回の繰り返し抽出実験での回帰係数の推定結果（OLSによる推定値平均）は外生的標本では一致性を有し，内生的標本では一致性が成立しない（表1）。

正しいモデルのもとでXに対するYの条件付き平均関数が線形であれば，外生的層化標本ではXによる抽出率の濃度に関わりなく平均関数は非線形な変化をしないので，独立標本であればウェイトは不要である。なお，ウェイトを使用しても一致性は有するが，効

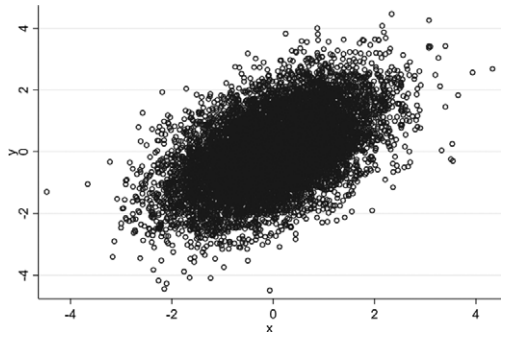

図1-1　母集団：モデルからの実現値

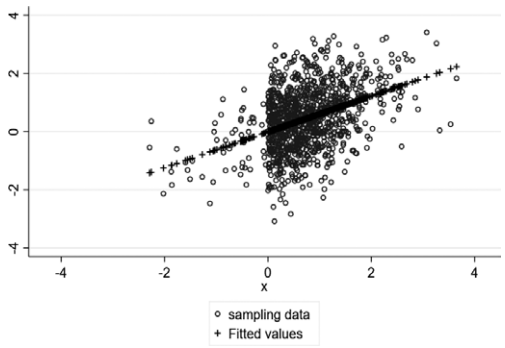

図1-2　Xによる層化データ（外生的）

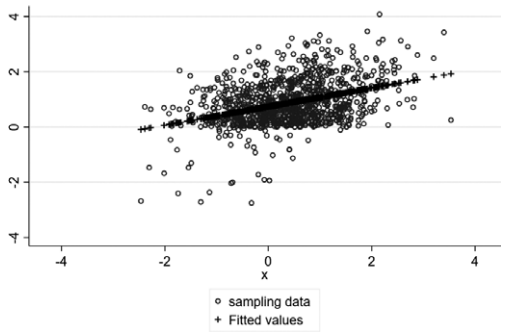

図1-3　Yによる層化データ（内生的）

率が低下する。

これに対して，内生的層化標本では，Xが与えられたときのYの条件付き分布に歪みをもたらす無情報ではないサンプリングであるため，そのままでは条件付き平均関数にバイアスを与える。OLSの推定値平均（0.34）が真の値（0.6）から大きく乖離していることが確認できる。その修正には，ウェイトを用いてYの本来の分布を再現する必要があり，抽出ウェイトによる重み付きWLS推定などの調整が行われねばならない。実際，ウェイト適用のシミュレーション結果は真値周りに推定値が分布していることがわかる。

一般に線形回帰モデルに対しては，母集団レベルでモデルが正しく想定されているとき，層化変数が外生的であれば，OLSもWLSも一致性をもつが，WLSでは分散が大きくなり推定効率が低下することが知られている[17]。WLS推定量は先出（2節）の $\hat{B}_W$ であり，デザイン一致推定量のクラスである。他方で，層化が内生的ならば（目的変数で層化するケース），ウェイトを無視した推定（例えばOLS）はバイアスをもつが，WLSは一致性をもつ。

このことは非線形モデルで尤度ベースの推定ルールを有するアプローチの場合も原則的に成立する。外生的な層化標本であれば，単純なウェイトを使わない最尤推定で一致性を有し，漸近有効性も標準的な仮定では成立する。それに対して内生的層化標本では，デザイン一致推定量の考え方が有用である。結論的には，擬似尤度による一種の重み付き対数尤度（weighted maximum likelihood）による推定が，一致推定量を与える。いずれにしても，このような特徴は母集団モデルが正しく想定されているという限定付きである。

次節では，非線形モデルを含めて，モデルの想定が誤っているケースにも拡大しながら，改めてデザイン一致推定量が提起する意義を，抽出ウェイトを利用しない単純な推定量とそれを利用する重み付き推定量との比較としてシミュレーションによって検討しよう。

補注）ロジスティック回帰モデルに関しては，内生的層化標本の場合のウェイトの利用は定数項に影響するだけで，回帰係数は通常の最尤法の推定値と変わらない。そのため例外的に，回帰係数だけが関心事であればウェイトを利用しなくとも効率よく推論は可能である[18]。（次節参照）

表1　抽出実験

①外生的層化標本　　真のモデル：Y=0.0+0.6X+ε, ε~N(0, 1)
sample（抽出率）：0.01 in X＜0 & 0.2 in X≧0　　（母集団N=10,000）

推定量	標本数	Mean	Std. Dev.	Min	Max
OLS	1,000	0.60467	0.04229	0.45620	0.74194
WLS	1,000	0.60405	0.07029	0.36674	0.81629
OLS：SRS 0.1	1,000	0.60164	0.10113	0.26935	0.93437

②内生的層化標本
sample（抽出率）：0.01 in Y＜0 & 0.2 in Y≧0　　（母集団N=10,000）

推定量	標本数	Mean	Std. Dev.	Min	Max
OLS	1,000	0.34141	0.02601	0.24863	0.41395
WLS	1,000	0.60312	0.07126	0.39170	0.89334
OLS：SRS 0.1	1,000	0.60565	0.10148	0.22685	0.96107

注）OLSは単純最小2乗推定量，WLSは抽出ウェイトによる加重最小2乗推定量推定量，OLS：SRS0.1は抽出率0.1の単純無作為抽出標本のOLS（参考系列）

4．シミュレーションからみるデザイン一致推定量の特性

デザイン一致推定量（抽出ウェイトによる加重推定）の特徴を理解するために，ここでは非線形モデルも含めて，想定したモデルが間違っているケースにも拡張してみよう[19]。前節にみた2変量の条件付きモデルに交絡のある第3の変数を導入する。

4.1　モデルと実験方法

①　検証用モデルの設定

交絡項Zをもつ単純ではあるが，一般性のあるモデル（図2）を考える。Zには，X, Yに対して内生性をもつ層化変数の役割ももたせることにする。ここでは線形回帰モデルとロジスティック回帰モデルを想定している。

［線形回帰モデル］

$Y=-0.6X+0.4Z+\varepsilon_Y$,　　$Z=\varepsilon_Z$,

$X=Z+\varepsilon_X$,　　$\varepsilon_Z, \varepsilon_X, \varepsilon_Y$~*i.i.d.* N(0, 1)

［ロジスティック回帰モデル］

Y=1 if $Y^*\geq 0$,　or Y=0 if $Y^*<0$

$Y^*=-0.6X+0.4Z+\varepsilon_{Y^*}$,　　$Z=\varepsilon_Z$,

$X=Z+\varepsilon_X$,　　$\varepsilon_Z, \varepsilon_X$~*i.i.d.* N(0, 1),

ε_{Y^*}~$L_0(0, 1)$: $L_0(0, 1)$は平均0，分散1のロジスティック分布

ここで，この検証モデルのもとでXが与えられたときのY（あるいはY^*）の条件付き分布の平均関数は$Y=-0.4X$である。真のモデルを根拠にした偽モデルの理論係数値であるので，この係数値を「擬似真値」と呼んでおく。

②　母集団（「母集団標本」と呼ぶ）を真のモデルから生成

想定した線形回帰モデルとロジスティック回帰モデル（あるいはそれらの仮説的無限母集団）からサイズN=10,000の*i.i.d.*データを単純無作為抽出で生成する。

なお，参考系列として，N=100,000とN=500,000の母集団値も生成し，漸近有効性を確認している。

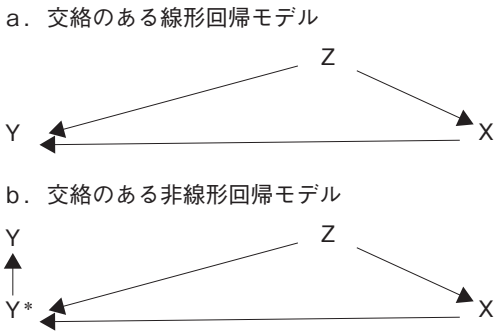

図2　検証用モデル

③　層化抽出標本（標本データ）の生成

目的変数Y，あるいは説明変数X, Zのそれぞれを層化変数とする層化抽出標本を3セット用意する。層化変数の値が負であるか，非負であるか（2値変数のケースでは0であるか，1であるか）で2層に区分し，前者には0.01，後者には0.2という異なる抽出率を割り当てる。なお，層内は単純無作為抽出とする。

④　標本データからのモデルパラメータの推定と評価指標の算出

推定すべきパラメータは，定数項bc^{**}, Xの係数bx^{**}, Zの係数bz^{**}と表す。①で提示した母集団を生成する真のモデル（0とおく）を標本データに適用した場合のモデルパラメータの推定値（$bc0^{*}$, $bx0^{*}$, $bz0^{*}$）と交絡変数Zを無視した誤ったモデル（1とおく）を適用したときの推定値（$bc1^{*}$, $bx1^{*}$）を求める。ただし，それぞれ抽出ウェイトを用いない単純推定（s）とウェイトを用いる加重推定値（w）とを計算している。したがって，これ以降例えば，間違ったモデルを適用したときのXの係数の加重推定値は bx1w，正しいモデルを適用したときのZの係数の単純推定値（ウェイトなし）は bz0s などと表記する。

さらに推定値の予測的評価指標として，下記の統計量を計算する。

a．予測誤差2乗平均；$msqr^{**}$（線形回帰モデルの場合）

$$msqr^{**}=\frac{1}{N-n}\sum_{t\notin S}(\hat{y}_t-y_t)^2$$

推定に用いた標本以外（$t\notin S$）の母集団要素の観測値 y_t を用いて，モデル予測値 $\hat{y}$ の予測誤差2乗和$msqr^{**}$を計算する。

b．平均KL（カルバック・ライブラー）情報量；mkl^{**}（ロジスティック回帰モデルの場合）

$$mlk^{**}=\frac{1}{n}\sum_{i\in S}[p_{i0}\ln(p_{i0}/\hat{p}_i)+(1-p_{i0})\ln((1-p_{i0})/(1-\hat{p}_i))]$$

推定したモデルによる予測確率のKL情報量の標本平均であり，0に近いほどよい。ただし，p_{i0}は真のモデルによる理論確率，$\hat{p}_i$は想定されたモデルによる予測確率である。

c．的中率；crc^{**}（ロジスティック回帰モデルの場合）

予測確率＞0.5のとき $\hat{y}=1$，それ以外は0と予測して，推定に用いた標本以外の母集団要素に対する的中率を計算する。

⑤　標本抽出からモデル推定の③と④のプロセスを1,000回繰り返し，そのときの推定されたパラメータの分布特性と予測評価指標a，b，cの分布特性を算出

なお，原理的には，②の母集団生成も複数回実施し，その各母集団に対して③と④を複数回実施すべきであろうが，すでに述べたように十分大きなNに対して母集団特性値はほぼ真値の近似を与えている状況であるから，本稿では母集団の複数の生成実験は省略している。

4.2　線形回帰モデルのケース

まず検証用モデルのもとで生成したN=10,000の実現値集合（母集団データ）に対する真のモデル（0）と偽のモデル（1）のもとでの推定パラメータ（センサスパラメータ）の特性をみておこう（表2-1）。想像されるように，Nが十分大きいとき真のモデルパラメータのよい近似を与えている。また偽のモデルに関しても，真の構造から生じる変数間の疑似的な連関（bx1=−0.39123）を捉えている。

標本データによる推定特性に関しては，Xで層化したケースは基本的に前節と同じであるから，ここでは交絡変数Zで層化したケース（表2-2）と目的変数Yで層化したケース（表2-3）を検討しよう。

Zで層化した標本データに関しては，正しいモデル（0）が想定されていればウェイトを

表2-1　母集団データによる推定特性（線形回帰）

真のモデル（0）の推定パラメータ（センサスパラメータ）
　真のモデル：Y=bc0+bx0·X+bz0·Z　　（N=10,000）

パラメータ	Coef.	Std. Err.	t	P>t	[95% Conf. Interval]	
bx0	−0.58635	0.01016	−57.71	0.000	−0.60626	−0.56643
bz0	0.39030	0.01440	27.11	0.000	0.36207	0.41852
bc0	−0.00004	0.01011	0.00	0.997	−0.01986	0.01978

偽のモデル（1）の推定パラメータ（センサスパラメータ）
　偽のモデル：Y=bc1+bx1·X　　（N=10,000）

パラメータ	Coef.	Std. Err.	t	P>t	[95% Conf. Interval]	
bx1	−0.39123	0.00743	−52.65	0.000	−0.40579	−0.37666
bc1	−0.00363	0.01047	−0.35	0.729	−0.02416	0.01690

注）bx1の擬似真値＝−0.4

表2-2　交絡項Zで層化した標本データの推定特性（線形回帰）

層化変数　Z　　抽出率　0.01：0.2　（Z<0：Z≥0）
（交絡因子が外生的層化変数である実際的なケース）

モデル	推定量	標本数	Mean	Std. Dev.	Std. Err.	[95% Conf. Interval]	
真	bx0s	1,000	−0.58994	0.02872	0.00091	−0.59172	−0.58816
	bz0s	1,000	0.40915	0.05005	0.00158	0.40605	0.41226
	bc0s	1,000	−0.01134	0.04039	0.00128	−0.01385	−0.00884
	msqr0s	1,000	1.02678	0.00697	0.00022	1.02635	1.02721
真	bx0w	1,000	−0.58713	0.07357	0.00233	−0.59170	−0.58256
	bz0w	1,000	0.39414	0.10200	0.00323	0.38781	0.40047
	bc0w	1,000	−0.00168	0.07130	0.00226	−0.00610	0.00275
	msqr0w	1,000	1.03797	0.01569	0.00050	1.03700	1.03894
偽	bx1s	1,000	−0.45731	0.02450	0.00078	−0.45883	−0.45579
	bc1s	1,000	0.18482	0.03428	0.00108	0.18269	0.18695
	msqr1s	1,000	1.15087	0.02032	0.00064	1.14961	1.15213
偽	bx1w	1,000	−0.39074	0.05239	0.00166	−0.39399	−0.38749
	bc1w	1,000	−0.00419	0.07490	0.00237	−0.00884	0.00046
	msqr1w	1,000	1.10784	0.01368	0.00043	1.10699	1.10869

表2-3　目的（結果）変数Yで層化した標本データの推定特性（線形回帰）

層化変数　Y　　抽出率　0.01：0.2　（Y<0：Y≥0）
（内生的層化のケース）

モデル	推定量	標本数	Mean	Std. Dev.	Std. Err.	[95% Conf. Interval]	
真	bx0s	1,000	−0.33749	0.02512	0.00079	−0.33904	−0.33593
	bz0s	1,000	0.22755	0.03215	0.00102	0.22556	0.22955
	bc0s	1,000	0.71990	0.02272	0.00072	0.71849	0.72131
	msqr0s	1,000	1.74049	0.04600	0.00146	1.73764	1.74335
真	bx0w	1,000	−0.59034	0.07271	0.00230	−0.59486	−0.58583
	bz0w	1,000	0.39508	0.10273	0.00325	0.38870	0.40146
	bc0w	1,000	0.00875	0.07285	0.00230	0.00422	0.01327
	msqr0w	1,000	1.03976	0.01902	0.00060	1.03858	1.04094
偽	bx1s	1,000	−0.21563	0.01794	0.00057	−0.21674	−0.21451
	bc1s	1,000	0.74741	0.02293	0.00073	0.74599	0.74883
	msqr1s	1,000	1.86163	0.04676	0.00148	1.85873	1.86453
偽	bx1w	1,000	−0.39292	0.05326	0.00168	−0.39622	−0.38961
	bc1w	1,000	0.00176	0.07557	0.00239	−0.00293	0.00645
	msqr1w	1,000	1.10943	0.01826	0.00058	1.10830	1.11057

使用せず（s）とも真のパラメータ値を推定可能である。また加重推定（w）値も真の値の近似を与えるが，ウェイトを使わない場合に比べ，標準誤差が過大となっている。

交絡する層化変数がモデルに導入されず，モデルが誤って想定されている場合（1）については，ウェイトを用いることで意味のある推定値（疑似真値＝−0.4）が得られる。他方で，ウェイトを使わない場合，意味のある推定値は得られず，結果の解釈は困難である。なお予測誤差2乗平均をみると，ウェイトを使わない場合（1s）よりウェイトを用いた推定（1w）の方が低めであり，予測的視点でも加重推定の良さが示されている。

それでは結果変数Yでの層化標本についてはどうであろう。表2-3にみるように，正しくモデルを想定（0）しても，ウェイトを使わない単純推定では推定値（0s）は大きなバイアスをもち，真値の良い推定量とはなっていない。これに対して加重推定量（0w）は真値の良い近似を与えており，また予測誤差msqrにおいても大幅な改善がみられる。

モデルが正しく想定されていない場合（1）でも同様であり，加重推定量（1w）が擬似真値のよい近似を与えているのに対して，単純推定（1w）では擬似真値からも大きく乖離しており，予測誤差も悪化している。

線形回帰モデルにおいて，モデルが正しく想定されているならばともかく，そうでない場合には，抽出ウェイトを利用することで真のモデル（構造）による擬似的な連関を示すセンサスパラメータを獲得できる。予測的にも与えられた推定ルールを基準にした相対優位な推定量を与えることがわかる。なお，モデルが正しいと想定されても，加重推定量は真のパラメータのよい近似を与えているが，外生的層化データでは推定誤差が過大になる点に注意を要する[20)]。

4.3　非線形モデルの場合（ロジスティック回帰モデルの事例）

ここでは尤度ベースの推定ルールでロジスティック回帰モデルを取り上げる。線形モデルの場合と同じくN=10,000の母集団データに対して，標本データからの推定特性を確認する。抽出ウェイトを考慮しない単純な最尤推定量（s）とウェイトを利用する擬似尤度推定量（w）との比較である。後者がデザイン一致推定量に対応している。

一般に尤度ベースのモデル推定においても，最小2乗距離ルールでの線形モデルと同じ特徴（前節）が成立するが，ロジスティック回帰モデルは内生的な層化標本の場合に例外的な性格を有している（3節補注参照）。しかしそれでも，デザイン一致推定量のロバストな性格がどのように発現するか確認しておきたい。なお，検証用モデル（4.1節）のもとで生成した母集団データの推定特性に関しては，本稿では線形回帰モデルとロジスティック回帰モデルにおける潜在変数モデル部分が同型であり，両者の推定特性は本質的に変わりないので割愛している。表2-1とほぼ同じ結果が得られていることだけを指摘しておく。

既に述べたように標本データは，X, Z, Yのいずれかで層化した3通りの抽出標本を用意している。表3-1はXによる層化標本の推定結果表であるが，適用モデルとしては層化変数であり直接的因果関係にある変数Xを説明変数に含む場合を整理している。表3-2は，層化変数がZで内生性をもつという実際的なケースで，それを含む真のモデルとそれを無視する偽のモデルの結果を整理している。表3-3では，結果変数Yを層化変数とする推定特性を整理しており，ロジスティック固有の特性が浮かび上がる。

Xで層化した標本データの推定特性（表3-1）からみておこう。正しくモデル（0）が想定されていれば，ウェイトを無視した推定量（0s）でも真値の良い近似を与えており，誤差

表3-1　Xで層化した標本データによる推定特性（ロジスティック回帰）

層化変数　X　　抽出率　　0.01：0.2　（X＜0：X≥0）
（外生的層化で直接的因果にあるケース）

モデル	推定量	標本数	Mean	Std. Dev.	Std. Err.	[95% Conf.	Interval]
真	bx0s	1,000	−0.56610	0.07812	0.00247	−0.57095	−0.56125
	bz0s	1,000	0.31334	0.08418	0.00266	0.30812	0.31857
	bc0s	1,000	0.01369	0.08457	0.00267	0.00844	0.01893
	mkl0s	1,000	0.00263	0.00250	0.00008	0.00247	0.00278
	crc0s	1,000	0.62071	0.00453	0.00014	0.62043	0.62099
真	bx0w	1,000	−0.60329	0.16141	0.00510	−0.61330	−0.59327
	bz0w	1,000	0.37981	0.22024	0.00697	0.36614	0.39348
	bc0w	1,000	0.01633	0.15883	0.00502	0.00647	0.02619
	mkl0w	1,000	0.00858	0.00892	0.00028	0.00802	0.00913
	crc0w	1,000	0.61368	0.01261	0.00040	0.61290	0.61446
偽	bx1s	1,000	−0.40649	0.06570	0.00208	−0.41057	−0.40242
	bc1s	1,000	0.01307	0.08467	0.00268	0.00782	0.01833
	mkl1s	1,000	0.01080	0.00237	0.00008	0.01066	0.01095
	crc1s	1,000	0.60829	0.00394	0.00013	0.60805	0.60853
偽	bx1w	1,000	−0.40421	0.11789	0.00373	−0.41153	−0.39690
	bc1w	1,000	0.01237	0.15695	0.00496	0.00263	0.02211
	mkl1w	1,000	0.01488	0.00809	0.00026	0.01438	0.01539
	crc1w	1,000	0.60301	0.01580	0.00050	0.60203	0.60399

表3-2　交絡項Zで層化した標本データの推定特性（ロジスティック回帰）

層化変数　Z　　抽出率　　0.01：0.2　（Z＜0：Z≥0）
（交絡因子が外生的層化変数である実際的なケース）

モデル	推定量	標本数	Mean	Std. Dev.	Std. Err.	[95% Conf.	Interval]
真	bx0s	1,000	−0.61734	0.06428	0.00203	−0.62133	−0.61335
	bz0s	1,000	0.33063	0.10977	0.00347	0.32381	0.33744
	bc0s	1,000	0.07476	0.08380	0.00265	0.06956	0.07996
	mkl0s	1,000	0.00370	0.00389	0.00012	0.00346	0.00394
	crc0s	1,000	0.61748	0.00579	0.00018	0.61712	0.61784
真	bx0w	1,000	−0.60686	0.16174	0.00512	−0.61690	−0.59682
	bz0w	1,000	0.39011	0.22165	0.00701	0.37635	0.40386
	bc0w	1,000	0.01606	0.15090	0.00477	0.00669	0.02542
	mkl0w	1,000	0.00820	0.00840	0.00027	0.00767	0.00872
	crc0w	1,000	0.61368	0.01016	0.00032	0.61305	0.61431
偽	bx1s	1,000	−0.50807	0.05166	0.00163	−0.51127	−0.50486
	bc1s	1,000	0.23274	0.06635	0.00210	0.22862	0.23686
	mkl1s	1,000	0.02003	0.00554	0.00018	0.01968	0.02037
	crc1s	1,000	0.60127	0.00325	0.00010	0.60107	0.60147
偽	bx1w	1,000	−0.40314	0.11451	0.00362	−0.41024	−0.39603
	bc1w	1,000	0.01463	0.14864	0.00470	0.00540	0.02385
	mkl1w	1,000	0.01436	0.00615	0.00019	0.01398	0.01475
	crc1w	1,000	0.60352	0.01060	0.00034	0.60286	0.60418

（参考）

N=100,000	Mean	Std. Dev.
bx0s	−0.59589	0.01935
bz0s	0.38643	0.03310
bc0s	0.00981	0.02790

N=500,000	Mean	Std. Dev.
bx0s	−0.60112	0.00894
bz0s	0.39932	0.01476
bc0s	−0.00117	0.01225

表3-3　結果変数Yで層化した標本データの推定特性（ロジスティック回帰）

層化変数　Y　　抽出率　0.01：0.2　（Y=0：Y=1）
（内生的層化のケース）

モデル	推定量	標本数	Mean	Std. Dev.	Std. Err.	[95% Conf. Interval]	
真	bx0s	1,000	−0.59187	0.15435	0.00488	−0.60145	−0.58230
	bz0s	1,000	0.37230	0.21526	0.00681	0.35894	0.38566
	bc0s	1,000	3.03221	0.15615	0.00494	3.02252	3.04190
	mkl0s	1,000	0.84148	0.06837	0.00216	0.83724	0.84573
	crc0s	1,000	0.44789	0.00177	0.00006	0.44778	0.44800
真	bx0w	1,000	−0.61597	0.17084	0.00540	−0.62657	−0.60537
	bz0w	1,000	0.39451	0.22920	0.00725	0.38028	0.40873
	bc0w	1,000	0.03809	0.15709	0.00497	0.02834	0.04783
	mkl0w	1,000	0.00859	0.00769	0.00024	0.00811	0.00907
	crc0w	1,000	0.61257	0.01595	0.00050	0.61158	0.61356
偽	bx1s	1,000	−0.40354	0.11100	0.00351	−0.41043	−0.39665
	bc1s	1,000	3.01845	0.15411	0.00487	3.00889	3.02801
	mkl1s	1,000	0.85310	0.06789	0.00215	0.84888	0.85731
	crc1s	1,000	0.44786	0.00175	0.00006	0.44775	0.44797
偽	bx1w	1,000	−0.40986	0.11938	0.00378	−0.41727	−0.40245
	bc1w	1,000	0.02368	0.15473	0.00489	0.01407	0.03328
	mkl1w	1,000	0.01465	0.00616	0.00020	0.01427	0.01503
	crc1w	1,000	0.60399	0.01435	0.00045	0.60310	0.60488

（参考）

N=100,000	Mean	Std. Dev.
bx0s	−0.59862	0.04953
bz0s	0.39713	0.06449
bc0s	3.00493	0.04610

N=500,000	Mean	Std. Dev.
bx0s	−0.60492	0.02133
bz0s	0.40427	0.02992
bc0s	2.99468	0.02161

も相対的に小さい。ウェイトを使った加重尤度推定量（0w）もよい近似を与えるが，精度はウェイトを使わない場合に比べ悪化している。

モデルの想定に誤りがある場合（1）でも，ウェイトの使用（1w），不使用（1s）にかかわらず擬似真値（−0.4）の良い近似を与えている。精度はウェイトありの方が低下している。また予測パフォーマンスもウェイトなしの推定量の方がよい。真のモデルも偽のモデルも層化変数Xを説明変数に含めているので，ウェイトの使用は必要なく，モデルに対して無視可能な標本抽出となっている。

内生性を有する変数Zが層化変数に使用された場合，あるいは層化変数ZがXにもYにも影響し内生性を有している場合（表3-2），モデルが正しく想定（0）されていれば，ウェイトを利用する（0w），しない（0s）にかかわらず，真のパラメータ値の近似（一致性）を与えている。ウェイトを使用する場合，ウェイトを使用しない単純推定に比べ，推定精度は低下し，予測評価指標（mkl）も悪化している。なお，一見するとウェイトを使わない場合バイアスが生じているような値をとるパラメータbz0sが観測されるが，$N\to\infty$のとき$n\to\infty$となりbz0s→真値（0.4）となるデザイン一致推定量の性質が表下部の（参考）系列からみてとれる[21]。

他方でモデルの想定に誤りがある場合（1），いまは層化変数Zがモデルにおいて無視されているとき，ウェイトを使わない推定量（1s）はバイアスをもち，この欠点はウェイトを利用する推定量（1w，擬似尤度推定量）によって修正される。パラメータbx1wの値が

示しているように擬似真値を教えてくれる。

結果変数Yによる内生的層化データの推定特性を整理したものが表3-3である。モデルが正しく想定されている場合（0），単純な最尤推定（0s）でもウェイトを使った擬似尤度推定量（0w）でも，XとZ，それぞれの回帰係数は真値の良い近似を与えている。定数項（bc0s）の推定値は異なっているが，これは抽出率で補正できることがわかっている[22]。したがってこのケースでは対数オッズへのある変数の効果（回帰係数）に限ってみるとウェイトの利用は必須ではない。ロジスティック回帰モデルに固有の特徴である。またZをモデルに含めない誤ったモデルについても，ウェイトを使っても使わなくとも，定数項を除き擬似真値が推定されている。なお，平均KL情報量，および的中率の予測評価指標で単純に比較する限りは，定数項のバイアスのため抽出ウェイトを利用する擬似尤度推定量の予測精度が高い。予測に関しては何らかのウェイト補正が必要となる。

5．結びにかえて―批判統計の解析的課題

2次利用が想定する統計の制度的背景条件には，統計の作成主体と利用主体の分離がある。とくに予算や人員のみならず，被調査者側での回答の真実性を担保する一定の強制力なり信頼関係を考慮すると，それなりのカバレッジや品質を備えた統計調査の担い手は自ずと限られる。本稿で取り上げた公的統計はその典型であり，実際には政府，あるいはそれに準じる公的機関がその中心的担い手とならざるを得ない。実際，政府は最大の統計生産者であり，また消費者（1次利用者）である。そしてこのような統計作成体制に対して，一般の利用者は外在的対象である統計表（あるいは集計された統計数字）を利用目的に合うように加工・処理するほかなく，そのための理論と技術と方法論が必要となる。これがいわば「統計利用者のための統計学」であり，批判的利用であれ積極的利用であれ，本学会の重要な研究上の柱のひとつであったはずである[23]。

それは，いわば統計表や統計数字といった集計情報の2次利用の方法を論じてきたといってよい。そこでは，統計表作成までの工程を理論的過程と技術的過程に論理的に峻別し，前者を信頼性，後者を正確性問題と位置づけ検討してきた。前者では作成者の目標や対象（社会）認識，あるいは理論規定と利用者が有するそれらとの間にある乖離が，後者では調査の社会過程として実行可能性の技術的適合性の要求程度が問題となる[24]。そしてこのような本学会での視点は，ミクロデータなど集計化されない調査票情報の2次利用に関しても共有できるはずである[25]。

調査票情報の2次利用が集計情報の利用と異なるのは，後者では集計過程で作成主体の理論規定が統計情報に組み込まれ，統計数字として実現されている点にある。そのため理論規定が統計数字に一体化しており，作成者と利用者との社会認識や理論が異なる場合には，程度の差はあれ利用上の大きな制約となる。これに対して集計過程を経ない調査票情報，あるいは個票データセットは集計前の分布情報を与えてくれるとともに，理論制約の強い集計概念に統合する前の技術的操作的調査票データが利用可能である。そして，その限りで作成者の理論規定によるデータ利用の制約からは相対的に免れているように思われるかもしれない。

作成者と利用者が理論的に異なっても，調査の現実案としては，技術的に実行可能で適合的な調査事項や調査方式が採用されなければならないため，調査票情報のレベルでは理論的な違いは薄められている。しかし，利用者は統計作成者とは異なる理論的視点を有している。その視点からは既存の調査票情報では，分析に必要な変数が調べられていない，統計的定義がずれている，調査対象に歪みが

ある，標本設計情報の秘匿や利用制約など，調査票情報レベルにおいてさえ利用者が想定する正しいモデルのもとでの分析はかなわないのが普通であろう[26]。

このような利用者の立場を批判統計の視点と呼ぶことにすれば，そのような利用者は，やむを得ず不完全な間違ったモデルでの分析を避けて，抑制的にモデル分析をあきらめ，記述統計的利用にとどまるべきであろうか。それも選択肢のひとつにはちがいないとしても，本稿の立場では，デザイン一致推定量の考え方を援用して，批判統計の視点からも積極的に例えば回帰モデルなど解析的手法を適用すればよいものとする。モデルが正しければ回帰係数は変数が与える構造的因果効果を教えてくれる，間違っていたとしても母集団記述統計量として推定ルールのもとで予測的連関を提供してくれる。とくに本稿で設定した問題の枠組みであるミクロデータの単独利用という状況では，このような意味において，デザイン一致推定量の考え方が調査票情報の積極的な解析的利用を批判統計の立場からも支えてくれるように思われる。そして，その有用性の程度はミクロデータレベルでの信頼性，正確性の具体的な議論の深化にも依存するであろう。「標本設計情報とミクロデータ解析」をめぐる60周年特集企画テーマの最終論考として，社会科学としての統計学研究をめざす本学会への提案的結びとしたい。

注

1）個票データの提供形態にもよるが，標本設計情報がすべて提供されるわけではない。しかし，母集団に戻すための計数として，例えば，復元乗率，線形推定用乗率，比推定調整率などは提供されるのが通例といえる。本稿では問題を複雑にしないため，抽出ウェイトに対応する線形推定用乗率を念頭におき論を進める。

2）Cochran（1977）に代表される標準的な標本調査論はこのための方法を提供している。

3）80年代から90年代にかけての欧米での標本設計情報の利用をめぐる論点のひとつは，ウェイトを使って復元する有限母集団の記述統計量の役割に関するものであった。このようなウェイト不要の主張は，モデル解析において記述統計的役割は不要であるか，その重要性は薄いという考え方に帰着する。Kasprzyk, Duncan, Kalton and Singh（1989）におけるHoem（p.539）やFienberg（p.570）による論争的な主張を参照されたい。

4）日本においては，ミクロデータ公開に向けての科学研究費の特定（旧称，重点）領域研究の成果としての松田・伴・美添（2000）がある。また土屋（2009）は今日的な手法を含めて広範な標本調査法を体系的に整理している。またビッグデータ利用も含めた最近の標本データに関わる展開は，Skinner and Wakefield（2017）などを参照されたい。

5）坂田（2019）ではCameron and Trivedi（2005）におけるミクロ計量経済学の方法論理の説明に依拠して，標本データの利用問題を整理している。それに対して，本稿ではむしろ，標本調査データの母集団記述統計的役割に重きを置いて論を進めている。

6）無回答をはじめとする回答構造の歪みは，標本設計のランダムネスを崩すため，利用上の大きな制約であることが指摘されていた。ここでは標本設計情報の利用に限定しているため，回答構造に歪みはないものと仮定している。本学会での標本調査，およびその解析的利用をめぐる議論に関しては坂元（1976），木村（1976），岩崎（2018），また個票データの2次利用については坂田（2006）を参照されたい。

7）母集団の補助情報や回答構造などをデータの発生構造として統一的に表現するには尤度概念が不可欠であり，完全情報尤度や標本尤度が提起され利用されている（Chambers, Steel, Wang, and Welsh 2012，2章）。しかし，一般利用者には一部の母集団情報を含む尤度関数の導出は容易というわけではなく，また本稿で設定した状況（標本データだけが利用できる）では必ずしも必須というわけではない。そのため一般的なデータの生成過程を再現する本来の尤度論的なアプローチは取り上げでい

ない。なお，記述統計的性格を有する擬似尤度については本論で取り上げている。

8）Pfeffermann（1993）では，先行研究のサーベイ論文でありながら，標本データの本来目的（母集団特性の記述）とその解析的利用（モデルパラメータの推定）が峻別され，前者が後者に活かせるのか，活かせるとすればどのような方法が可能なのか，といった視点が明確にうかがえる。その点で，これまでの本学会の解析的利用に対する批判的系譜と問題認識をかなりの程度共有している。そこには，正統派標本調査論の枠組みにおいてモデル解析への2次利用を論じる姿勢が鮮明に表れている。なお，詳細な議論についてはChambers and Skinner（eds）（2003）3章および8章参照。

9）Chambers and Skinner（eds）（2003）3章でも指摘されるように，このような図式は2相標本抽出の枠組みで捉えられる。

10）デザインベースとモデルベースのアプローチという標本からの推論のフレームワークについては，Skinner, Holt, and Smith（1989）第1章，Chambers and Skinner（eds）（2003）第2，3章，土屋（2009）第13章などを参照されたい。

11）言うまでもなく，デザインベースの推定においては，母集団要素の値，あるいは全数データは，固定値であり所与である。ある確率変数からの実現値とみなす確率的な変動は許容してはいない。その意味で，母平均などの統計量と同様にセンサスパラメータはあくまで記述的要約統計量である。

12）擬似尤度とその分散推定については，Skinner, Holt, and Smith（1989）, pp.80-84, Binder（1983），土屋（2009）第13章を参照されたい。すでに主要な統計解析ソフトウェアでは実装されており，利用者に供されている。

13）例えば，Kish and Frankel（1974）参照。

14）Chambers and Skinner（eds）（2003）第3章，pp.45-48も参照されたい。

15）デザイン一致性をめぐる推定の論理については，Chambers and Skinner（eds）（2003）第3章で定数項モデルを用いた丁寧な説明が展開されている。

16）Chambers and Skinner（eds）（2003）第2章，高井・星野・野間（2016），阿部（2016），高橋・渡辺（2017），など

17）このような推定効率をめぐる議論に関してはPfeffermann（1996）も参照されたい。

18）例えばSkinner, Holt, and Smith（1989）第9章，マダラ（2004）第8章など参照のこと。

19）確率ウェイトを使わない単純な推定とウェイトを使う推定に対してパラメータBの点推定値$\hat{B}$，$\hat{B}_w$を1000回求め，それぞれの分布特性（平均，分散など）を比較している。抽出率が少し高めだが，N, nが十分大きいことから（Chambers and Skinner（eds）（2003）3章），後者をデザイン一致推定量の近似として扱っている。

20）加重推定量についてはいくつか欠点があり，モデルの実現値としての母集団における偏りの可能性，モデルが間違っていたとして異なる層別割合で構成される母集団に対しては意味がない，あるいは層ごとに異なるパラメータ値に対して加重推定値は偏りをもつ点などが指摘されている（Pfeffermann 1993, p.329）。

21）表3-1の真のモデルのパラメータ特性（0s）に関してもバイアスと疑われそうな推定値が観測されている。これに関しても$N \to \infty$となるような実験を行うと真値に収束することが確認できる。

22）ウェイトを使わない定数項の推定値に抽出率の対数の差を加えればよい。例えばSkinner, Holt, and Smith（1989），p.199，マダラ（2004），pp.392-394参照。

23）本学会の「社会科学としての統計学」をめぐる議論に関しては，統計学第30号（経済統計研究会，1976）「社会科学としての統計学 ― 日本における成果と展望 ― 」創立20年記念号，および1986, 1996，2006年の経済統計学会編『社会科学としての統計学』第2集，第3集，第4集（産業統計研究社）の各記念号を参照されたい。

24）例えば，大屋（1995）の「付論　統計学批判考」参照。また公的統計における調査目的と統計主体との乖離（形式性，一面性）については，濱砂（2011）による考察がある。

25）個票データの利用方法論に関しては坂田（2006）参照。

26）統計調査のプロセスと同様に，調査票情報における理論制約と技術精度についてはミクロデータ利用にとって改めて検討すべき課題といえる。坂田（2006）に対するコメントにおいて岩井（2006, p.44）は，法律婚と事実婚をめぐる調査個票内の矛盾の処理を引き合いに「上記の矛盾項目も，法律婚としては14歳以下の有配偶はありえないが，事実婚ならありえるケースである。法律婚を前提に

しているのも，一定の社会認識を前提しているといえる。このように調査結果の集計過程の諸論点も，単に技術的視点に止まらず，理論的視点が内在しているといえる」と的確に指摘し，個票データの信頼性問題を提起している。

参考文献

Binder, D. A.（1983）, “On the Variances of Asymptotically Normal Estimators from Complex Surveys”, *International Statistical Review* 51, 279-292.

Breckling, J. U., R. L. Chambers, A. H. Dorfman, S. M. Tam, and A. H. Welsh（1994）, “Maximum Likelihood Inference from Sample Survey Data”, *International Statistical Review*, 62, 349-363.

Cameron, A. C., and P. K. Trivedi（2005）, *Microeconometrics: Methods and Applications*, Cambridge University Press.

Chambers, R. L. and C. J. Skinner, eds.（2003）, *Analysis of Survey Data*, Wiley.

Chambers, R. L., D. G. Steel, S. Wang, and A. H. Welsh（2012）, *Maximum Likelihood Estimation for Sample Surveys*. Taylor and Francis CRC.

Cochran, W. G.（1977）, *Sampling Techniques*, Third edition, Wiley.

Kasprzyk, D., G. Duncan, G. Kalton and M. P. Singh, eds.（1989）, *Panel Surveys*, Wiley.

Kish, L. and M. P. Frankel（1974）, “Inference from Complex Samples”, *Journal of the Royal Statistical Society*, Ser. B 36, 1-37.

Little, R. J. A.（1982）, “Models for Nonresponse in Sample Surveys”, *Journal of the American Statistical Association* 77, 237-250.

Skinner, C. J., D. Holt, and T. M. F. Smith（1989）, *Analysis of Complex Surey*, Wiley.

Skinner, C. J., and J. Wakefield（2017）, “Introduction to the Design and Analysis of Complex Survey Data”, *Statistical Science*, 32(2), 165-175.

Pfeffermann, D.（1993）, “The Role of Sampling Weights When Modeling Survey Data”, *International Statistical Review*, Vol. 61, No. 2, 317-337.

Pfeffermann, D.（1996）, “The Use of Sampling Weights for Survey Data Analysis”, *Statistical Methods in Medical Research*, Vol. 5, No. 3, 239-261.

Rubin, D. B.（1976）, “Inference and Missing Data”, *Biometrika* 53, 581-592.

阿部貴行（2016）『欠測データの統計解析』（統計解析スタンダード）朝倉書店.

岩崎俊夫（2018）『社会統計学の伝統と継承：論点と関連論文（1955-90）』お茶の水書房.

岩井浩（2006）「コメント：個票データと統計利用」『統計学』第90号，42-44.

大屋祐雪（1995）『統計情報論』，九州大学出版会.

木村和範（1976）「推計学批判」『統計学』第30号，101-120.

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

坂田幸繁（2019）「パラメータ推定と抽出ウェイトの利用：尤度を中心に」『公的統計情報：その利活用と展望』，中央大学経済研究所叢書75.

坂元慶行（1976）「標本調査」『統計学』第30号，84-93.

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・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.

土屋隆裕（2009）『概説標本調査法』，朝倉書店.

濱砂敬郎（2011）「いわゆる『公的統計』の公共的な性格について：調査目的の観点から」,『統計学』第100号，1-13.

マダラ.，G. S.（2004）『マダラ計量経済分析の方法（改訂3版）』（佐伯親良訳），エコノミスト社.

松田芳郎・伴金美・美添泰人編（2000）『ミクロ統計の集計解析と技法』，日本評論社.

【Special Section: The 60th Anniversary of the Journal】
Special Topic A: Problems in Microdata Analysis of Official Statistics Based on Probability Sampling Designs

Effects of Sampling Weights on the Secondary Analysis of Official Statistics Microdata

Yukishige SAKATA*

Summary

Applying questionnaire information and micro data of official statistics to model analysis, this study examines how to account for the survey design and sampling weights, especially when the theoretical and analytical frameworks differ between the user side of the secondary analysis and the statistical survey agency side. This study compares the features of simple, unweighted estimators with those of weighted estimators in a regression model using simulation data, based on the concept of the design-consistent estimator as adjusting the primitive characteristics of the descriptive statistics of the sampling survey data to the model analysis, such as in a super population model. Difficulties arise in hypothesizing proper models in the secondary use of official statistics.; thus, the importance of strategically weighted estimators using sampling weights in model analysis is verified.

Key Words

Microdata, Secondary Analysis, Model parameter, Superpopulaion model, Sampling weight

* Faculty of Economics, Chuo University

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119号：2020年9月30日
　第120号：2021年3月31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。
5．編集委員会は2020年4月から次の体制となります。引続きよろしくお願いします。
　2020年度編集委員会委員長　小林良行（東北・関東）
　同副委員長　村上雅俊（関西）
　同委員　水野谷武志（北海道），山田　満（東北・関東），松川太一郎（九州）

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。『統計学』創刊60周年記念事業委員会は2つの特集の編集ありがとうございました。昭和情報プロセス㈱品川様には印刷でいつもお世話になっています。（池田伸　記）

執 筆 者 紹 介

坂田幸繁　（中央大学経済学部）　　山口幸三　（総務省統計研究研修所）
武田英俊　（京都大学大学院総合生存学館）　　芳賀　寛　（中央大学経済学部）

支 部 名　　事　務　局

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.118　　定価 1,760円（本体1,600円）

2020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　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. 118 2020 March

Japan Society of Economic Statistics