【論　文】

ミクロデータ分析における調査ウェイトの補正効果
―社会生活基本調査・匿名データの利用に向けて―

栗原由紀子*・坂田　幸繁**

要　旨

本稿は、社会生活基本調査・匿名データを素材に、母集団パラメータを推定する場合の調査ウェイトの取り扱いについて検討した。具体的には、ウェイトの使用・不使用、およびウェイト調整などのアプローチの相違に起因する推定バイアスを計測するとともに、バイアスが発生する場合の現実的な対処法を提示することを課題とした。

主な結果は以下３点にまとめられる。まず、社会生活基本調査ミクロデータについては、男女・年齢など分析の前提となる基本属性の違いによって生活行動や生活時間が大きく異なるため、分析対象である目標母集団の人口構成に注意を払う必要がある。標本による推定構成比が母集団構成比から大きく歪んでいる場合、得られる推定値にはバイアスが生じるリスクが高い。次に、回帰分析では、推定構成比の歪みが係数の値と有意性の両方に強く影響を与える。線形推定用乗率であっても、本来の母集団人口構成比と異なる場合には、その歪みが推定バイアスの発生要因となる。さらに、このようなリスクを回避するには、国勢調査などの補助情報を用いてウェイト調整を行い、それを回帰分析などの推定に用いることが有力な解法のひとつであることを示した。

１．はじめに

政府基幹統計における大規模標本調査に代表されるような調査では、集計計画に規定された母集団特性値を、人、時間、予算などのリソースが限られる中、目標精度で獲得できるように綿密な標

* 弘前大学人文学部講師　連絡先 yukuri@cc.hirosaki-u.ac.jp
** 中央大学経済学部教授　連絡先 ysakata@tamacc.chuo-u.ac.jp

本研究は、「政府統計データのアーカイビングシステムの構造と機能に関する国際比較研究」日本学術振興会科学研究費補助金基盤研究（B）（課題番号：22330070、研究代表者：法政大学　森博美、平成22年度～25年度）の成果の一部である。また、本研究は個票データの二次分析に基づいている。二次分析に当たっては、一橋大学経済研究所付属社会科学統計情報研究センターから社会生活基本調査（平成８年度、13年度、18年度）の匿名データの提供（申請者：中央大学・坂田幸繁）を受けたことを付記して、関係諸機関への謝辞とします。

本設計が施されている。公的統計の、いわゆるミクロデータとして提供される匿名化個票データセットにおいて、通常それは調査ウェイト（単に乗率あるいは復元乗率とも呼ぶ）という形で、秘匿処理を施された調査票情報とともに提供されている。

一般論としていえば、政府統計ミクロデータを用いて目標母集団（例えば全国）の特徴や傾向を捉えるには、少なくとも推定精度を維持する意味でウェイトの利用は不可欠のはずである。しかし、調査目的が集計計画に具体化され、その重要度・優先度の高低が標本設計に反映されるので、あらゆる利用目的（ミクロデータの二次利用）に対応できる万能の処方箋としてのウェイトが付されるわけではない。また、通常、個人や世帯統計についてはセンサス（国勢調査）人口への復元が基本であり、そこからのズレを補正する機能もウェイトは担うことができるが、ウェイトに課せられる補正内容は、集計目標に応じて様々に変わりうる。

例えば、日本の社会生活基本調査（総務省）では、地域・男女・年齢区分別推定値までを国勢調査人口と整合的に補正しているが、オランダでは曜日および季節調整のための補正に留まり、ドイツでは地域・世帯類型・社会階級まで調整されている[1)]。このような特殊な制約（目的）をウェイトは背負っているせいか、とくに推定対象や利用目的が調査計画のそれと異なるユーザに関しては、ウェイト利用を躊躇したり、逆に漫然とウェイトを適用する事例なども散見される。

ウェイトをめぐるこのような事情は、推定結果の精確性をどの程度見積もればよいのか、適切な評価を困難とし、統計的実証分析の場へのミクロデータの普及と本格利用の妨げとなる場合も考えられよう。本稿ではこのような状況の改善に資することを企図して、生活行動や時間把握を目的とする「社会生活基本調査」（総務省）のミクロデータを素材に、ミクロデータ分析においてウェイトの使用・不使用により、どの程度推定値に差が生じ、真値からの歪みがどれくらい発生するのか検証することにしたい。また可能な場合には必要な調整手法を提示し、その効果も示すことにする。さらに、回帰分析などのパラメータ推定における調査ウェイトの適用についても検討を加えることにする[2)]。

[1)] これら各国の調査情報はCenter for Time Use Reserchウェブサイトから得られる。なお、生活時間調査ガイドラインEurostat（2009）によれば、生活時間統計の推定は複雑であり、いくつかのウェイトを使い分ける必要があると述べている。一つは個体に付される抽出ウェイト、二つ目は包含確率ウェイト、三つ目は調査日を対象に付加されるダイアリー・ウェイト、最後に事後層化推定のための補助ウェイトである。特に、無回答補正のために、包含確率ウェイト、ダイアリー・ウェイト、補助ウェイトを活用すべきであると指摘している。

[2)] 社会生活基本調査本来のウェイト利用に関する先行研究には、高橋・臼井（2005）が挙げられる。2001年調査について、母集団人口への補正効果について詳細に議論したものである。結果としては、ベンチマーク人口に対して、母集団人口の推定値は、年齢別では30代までは下方にバイアスをもち、60代以上では上方にバイアスをもつ。配偶関係では、未婚が下方バイアス、有配偶が上方バイアス、そのほかにも職業別、世帯類型別などについてベンチマーク人口との相違を確認している。また、男女・年齢別人口により補正を行うことで、男女別や年齢別については母集団人口と一致するが、配偶関係別や職業別の人口ではバイアスが残ることが指摘されている。このようなバイアスは匿名データにも当然引き継がれるが、一部の標本設計情報が提供されないこともあり、さらに問題は複雑化している。

2．標本設計と乗率

2.1　標本設計

社会生活基本調査（以下では、「社会調」とも略記）は、層化２段抽出により標本抽出が行われている。その標本設計の概念図を示せば、図１(a)のようである。まず、都道府県で層化（$k = 1, \ldots, K$）し、各都道府県別に国勢調査（以下、「国調」とも略記）の人口に比例させて抽出率（C_{kg}/C_k）で国調の調査区（$g = 1, \ldots, G_k$，G_kは標本調査区数を示す）を抽出する（図中の網掛けの部分）。C_kおよびC_{kg}は、国勢調査における都道府県kの人口数および都道府県kの調査区gの人口数をそれぞれ表す。標本調査区をもとに母集団調査区へと復元するには、各標本調査区に抽出ウェイト$f^{(1)} = C_k/G_kC_{kg}$を適用すればよい。

次に、抽出された調査区内の世帯を対象として、単純無作為抽出により標本世帯を抽出する（図１(a)右端図の×印）。都道府県kの調査区gの国調世帯数M_{kg}を基礎数に、抽出率はm_{kg}/M_{kg}で与えられる。ここでm_{kg}は標本世帯数である。調査区内から抽出された標本世帯から調査区内の全世帯を復元するには抽出ウェイト$f^{(2)}=M_{kg}/m_{kg}$を用いればよい。さらに、調査日を指定する際には、標本調査区をランダムに８等分し、それぞれを調査グループとして調査曜日が割り当てられる[3)]（図１bを参照）。

図１（a）層化２段抽出の流れ（概念図）

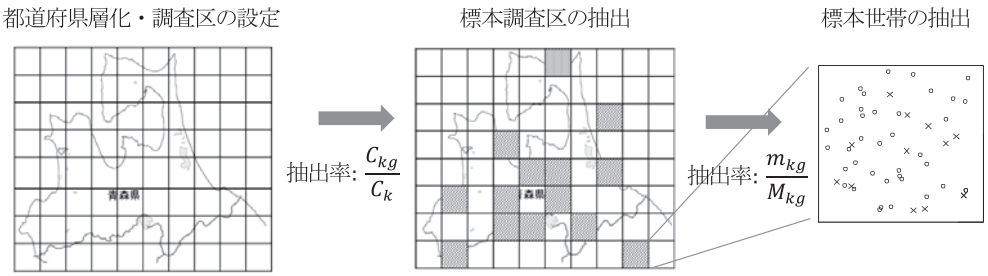
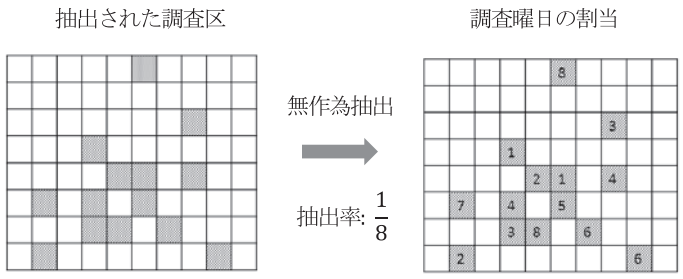


図１（b）標本調査区に対する調査曜日の割当（概念図）



調査グループ(q)	調査曜日(h)
1	土日（6,7）
2	土日（6,7）
3	日月（7,1）
4	火水（2,3）
5	木金（4,5）
6	金土（5,6）
7	土日（6,7）
8	土日（6,7）

[3)] 調査曜日の割り当てに関する推定上の問題は栗原（2010）を参照。

したがって、最終的には曜日毎に、標本世帯iの調査票データに対しては、抽出率の逆数を掛け合わせることで、(1)式のように母集団推定のためのウェイトw_iを作成できる。これは社会生活基本調査報告書において「線形推定用乗率」と呼ばれている。ただしr_{kg}は調査区の合併などを調整するための修正項である。

$$w_i = f^{(1)} \times f^{(2)} \times r_{kg} \tag{1}$$

このウェイトは世帯iについての算式であるから、一義的に世帯用の乗率である。しかし、とくに不都合がない限りは、個人を対象とする推定においても、同一世帯内の世帯員jには全て同じ抽出ウェイト（$w_j = w_i$）が与えられて計算される。

2.2 母集団特性値の推定量

第i世帯の個人jが属す都道府県（k）・男女（δ）・年齢（d）別カテゴリー・グループを（$k \cdot \delta \cdot d$）で表すことにする。(1)のように定義された個人用の線形推定用乗率w_jを用いると、各曜日に対して（$k \cdot \delta \cdot d$）グループの推定人口$\widehat{N}_{k \cdot \delta \cdot d}$、および任意の調査変数$y$（例えば行動時間）に関する総計$\hat{Y}_{k \cdot \delta \cdot d}$やその平均推定量$\hat{\mu}_{k \cdot \delta \cdot d}$が次式のように線形推定量として求められる[4)]。

$$\widehat{N}_{k \cdot \delta \cdot d} = \sum_{j \in k \cdot \delta \cdot d} w_j \tag{2}$$

$$\hat{Y}_{k \cdot \delta \cdot d} = \sum_{j \in k \cdot \delta \cdot d} w_j y_j \tag{3}$$

$$\hat{\mu}_{k \cdot \delta \cdot d} = \frac{\hat{Y}_{k \cdot \delta \cdot d}}{\widehat{N}_{k \cdot \delta \cdot d}} \tag{4}$$

すでに述べたように、社会調の調査設計は世帯抽出であり、厳密には世帯に関する統計量であればその母集団特性値が推定できる。標本設計に利用したグループレベルにおいて、標本からの推定世帯数は対応する母集団世帯数、すなわち国調世帯数に一致する。しかし、世帯用の乗率を個人用の乗率と読替える方式では当然このような一致は保証されない。例えば、個人ベースの統計量である都道府県別の性別・年齢別人口の推定値$\widehat{N}_{k \cdot \delta \cdot d}$は、国調人口$C_{k \cdot \delta \cdot d}$と一致しないのが通常である。しかし、都道府県別・性別・年齢別という基本集団レベルにおける推定値と抽出フレームとして機能する国調人口との乖離は望ましいことではなく、状況においては目標変数yの推定値の信頼性にも疑義が生じてしまう。

そこで、都道府県・男女・年齢別人口については国調人口と一致するように、ウェイトを調整した比推定用乗率w_j^*（以下では、「調整ウェイト」と呼称する）が作成される。

$$w_j^* = \frac{C_{k \cdot \delta \cdot d}}{\widehat{N}_{k \cdot \delta \cdot d}} \times w_j, \quad j \in k \cdot \delta \cdot d \tag{5}$$

[4)] この考え方は復元抽出法によるHansen-Hurvits推定量（HH推定量）によるものであり、本稿では抽出率が極めて小さい場合を想定し、このHH推定量の考え方を敷衍する。

これを用いれば、都道府県・男女・年齢別人口 $\widehat{N}^{*}_{k\cdot\delta\cdot d}$ は国勢調査人口 $C_{k\cdot\delta\cdot d}$ に一致する。当然、それらの要素のみで構成される部分母集団の推定人口数、例えば男女別人口、または年齢別人口などは国調人口のそれに一致する。

$$\widehat{N}^{*}_{k\cdot\delta\cdot d} = \sum_{j\in k\cdot\delta\cdot d} w^{*}_{j} = C_{k\cdot\delta\cdot d} \tag{6}$$

なお、任意の変数の総計 $\widehat{Y}^{*}_{k\cdot\delta\cdot d}$ を算出する場合においても、このような調整ウェイトを利用することにより国調ベースの人口数に対応した総計の推定が可能となる。

$$\widehat{Y}^{*}_{k\cdot\delta\cdot d} = \sum_{j\in k\cdot\delta\cdot d} w^{*}_{j} y_{j} = \frac{C_{k\cdot\delta\cdot d}}{\widehat{N}_{k\cdot\delta\cdot d}} \widehat{Y}_{k\cdot\delta\cdot d} \tag{7}$$

ところで、都道府県・男女・年齢別の平均値（または比率）$\hat{\mu}^{*}_{k\cdot\delta\cdot d}$ については、(8)式のようにウェイトによる人口調整効果は表れない。したがって、個別の層の部分母集団として平均値（比率）のみを推定する場合には、線形推定用乗率でも問題ないことになる。

$$\hat{\mu}^{*}_{k\cdot\delta\cdot d} = \frac{\widehat{Y}^{*}_{k\cdot\delta\cdot d}}{\widehat{N}^{*}_{k\cdot\delta\cdot d}} = \hat{\mu}_{k\cdot\delta\cdot d} \tag{8}$$

しかしながら、調整ウェイトに使用した都道府県・男女・年齢のうち、いずれかの層を部分母集団として指定せずに平均値（比率）を算出する場合には、やはりウェイトの人口調整が必要となる。例えば、部分母集団を構成する変数に都道府県（d）を使用せずに全国平均、または九州ブロックなどの地域平均として、男女・年齢別人口を対象とした推定値を算出した場合、調整ウェイトを用いれば、人口(9)式、総計(10)式、平均値(11)式のいずれの統計量についても国調人口ベースの値として推定できる。

$$\widehat{N}^{*}_{\delta\cdot d} = \sum_{j\in\delta\cdot d} w^{*}_{j} \cong C_{\delta\cdot d} \tag{9}$$

$$\widehat{Y}^{*}_{\delta\cdot d} = \sum_{j\in\delta\cdot d} w^{*}_{j} y_{j} \cong \sum\nolimits_{k} \frac{C_{k\cdot\delta\cdot d}}{\widehat{N}_{k\cdot\delta\cdot d}} \widehat{Y}_{k\cdot\delta\cdot d} \tag{10}$$

$$\hat{\mu}^{*}_{\delta\cdot d} \cong \sum\nolimits_{k} \frac{C_{k\cdot\delta\cdot d}}{C_{\delta\cdot d}} \cdot \frac{\widehat{Y}_{k\cdot\delta\cdot d}}{\widehat{N}_{k\cdot\delta\cdot d}} \tag{11}$$

とくに、比推定用乗率を用いたときには、各男女・年齢別カテゴリーの都道府県別人口構成比 $C_{k\cdot\delta\cdot d}/C_{\delta\cdot d}$ の相違を考慮して、国調ベースの都道府県別人口構成比の加重平均として男女・年齢別推定値が算出されている。しかし、線形推定用乗率による平均値(12)式では、男女・年齢別人口の内訳としての都道府県別構成比が考慮されないため、(8)式の場合とは異なり、調整ウェイトによる平均値とは一致しない（$\hat{\mu}^{*}_{\delta\cdot d} \neq \hat{\mu}_{\delta\cdot d}$）。

$$\hat{\mu}_{\delta\cdot d} = \frac{\widehat{Y}_{\delta\cdot d}}{\widehat{N}_{\delta\cdot d}} \tag{12}$$

なお、線形推定用乗率による平均値では、都道府県人口構成比が線形推定用乗率と調整ウェイトと

で近似している場合には、どちらのウェイトを使用しても結果数値に変わりはないが、推定値が背負う意味合いと性格が基本的に異なる点に注意する必要がある。

2.3 ミクロデータからの推定値

社会調の匿名化ミクロデータは、本来の個票データセットから80%抽出したリサンプリング・データであり、ウェイトはさらに抽出率の逆数を乗じたものになる。ここでは、リサンプル・ウェイトと呼称する。これを$w_j^{\#}$とおけば、すでに定義した調整ウェイトw_j^*および線形推定用乗率w_jとの関係は次のようである。

$$w_j^{\#} = w_j^* \times \frac{1}{0.8} = \frac{C_{k\cdot\delta\cdot d}}{\widehat{N}_{k\cdot\delta\cdot d}} \times w_j \times \frac{1}{0.8}, \quad j \in \delta \cdot d \tag{13}$$

ただし、匿名データでは都道府県kと調査区gを識別する変数は提供されないので、世帯について無作為抽出したものと仮定して分析を行うことになる（栗原 (2010)）。また、都道府県kを識別できないことから、匿名データでは男女・年齢別部分母集団（$\delta \cdot d$）を最小単位とした母集団推定が想定される。

リサンプル・ウェイトによる匿名データからの推定値である男女・年齢別人口$\widehat{N}_{\delta\cdot d}^{\#}$、および任意の変数の総計$\widehat{Y}_{\delta\cdot d}^{\#}$と平均値$\hat{\mu}_{\delta\cdot d}^{\#}$は、(9)～(10)式および(14)～(16)式より、国勢調査人口$C_{\delta\cdot d}$とこの人口を目標母集団とした調整ウェイトによる推定値にほぼ一致する。（$\widehat{N}_{\delta\cdot d}^{\#} \cong C_{\delta\cdot d}$, $\widehat{Y}_{\delta\cdot d}^{\#} \cong \widehat{Y}_{\delta\cdot d}^{*}$, $\hat{\mu}_{\delta\cdot d}^{\#} \cong \hat{\mu}_{\delta\cdot d}^{*}$）

$$\widehat{N}_{\delta\cdot d}^{\#} = \sum_{j\in\delta\cdot d} w_j^{\#} \cong C_{\delta\cdot d} \tag{14}$$

$$\widehat{Y}_{\delta\cdot d}^{\#} = \sum_{j\in\delta\cdot d} w_j^{\#} y_j \cong \sum\nolimits_k \frac{C_{k\cdot\delta\cdot d}}{\widehat{N}_{k\cdot\delta\cdot d}} \widehat{Y}_{k\cdot\delta\cdot d} \tag{15}$$

$$\hat{\mu}_{\delta\cdot d}^{\#} = \frac{\widehat{Y}_{\delta\cdot d}^{\#}}{\widehat{N}_{\delta\cdot d}^{\#}} \cong \sum\nolimits_k \frac{C_{k\cdot\delta\cdot d}}{C_{\delta\cdot d}} \frac{\widehat{Y}_{k\cdot\delta\cdot d}}{\widehat{N}_{k\cdot\delta\cdot d}} \tag{16}$$

とくに、匿名データでの分析は都道府県変数が利用できないため、常に全国単位での平均値を算出することになるが、都道府県レベルの人口構成比は調整されていることから、都道府県の人口構成比の歪みは全国単位の平均値に関しては補正されていることになる。なお、ウェイトを利用しない標本平均$\hat{\mu}_{h,\delta\cdot d}^{0}$は次式であり、当然ウェイトを用いた母平均推定値とは一致しない。

$$\hat{\mu}_{\delta\cdot d}^{0} = \frac{\sum_{j\in\delta\cdot d}\ y_j}{n_{\delta\cdot d}} \tag{17}$$

2.4 部分母集団の推定と乗率の利用

男女・年齢以外の層（もしくは群/カテゴリー・グループ）に関して、例えば就業の有無（a）別人口を部分母集団として推定しようとするとき、先のリサンプル・ウェイト$w_j^{\#}$を使用しても、就

業の有無別でのウェイト調整を行っていないため、男女・年齢・就業の有無別の推定人口は母集団人口とは一致しない（$\widehat{N}^{\#}_{\delta\cdot d\cdot a} \neq C_{\delta\cdot d\cdot a}$）。

$$\widehat{N}^{\#}_{\delta\cdot d\cdot a} = \sum_{j\in\delta\cdot d\cdot a} w_j^{\#} \cong \sum\nolimits_k \frac{C_{k\cdot\delta\cdot d}}{\widehat{N}_{k\cdot\delta\cdot d}} \widehat{N}_{k\cdot\delta\cdot d\cdot a} \tag{18}$$

$$\widehat{Y}^{\#}_{\delta\cdot d\cdot a} = \sum_{j\in\delta\cdot d\cdot a} w_j^{\#} y_j \cong \sum\nolimits_k \frac{C_{k\cdot\delta\cdot d}}{\widehat{N}_{k\cdot\delta\cdot d}} \widehat{Y}_{k\cdot\delta\cdot d\cdot a} \tag{19}$$

そこで、男女・年齢・就業の有無別人口が部分母集団の人口と一致するように、再度、ウェイトを調整する方法が考えられる。これを調整リサンプル・ウェイト（$w_j^{\#a}$）と呼ぶことにする[5)]。なお、リサンプル・データには都道府県の変数が含まれないため、都道府県の層までは調整できない。

$$w_j^{\#a} = \frac{C_{\delta\cdot d\cdot a}}{\widehat{N}^{\#}_{\delta\cdot d\cdot a}} \times w_j^{\#},\ j \in \delta\cdot d\cdot a \tag{20}$$

これを用いれば、男女・年齢・就業の有無別推定人口および任意のyの総計は、国勢調査の当該部分母集団に対応する数値として得られる。また、平均値（比率）については、調整リサンプル・ウェイトによる結果と単純なリサンプル・ウェイトによるそれとは理論的に一致する。平均値だけに関心がある場合には、就業の有無に関するウェイトの調整は不要となり、匿名化データに付与されているリサンプル・ウェイトによる推定値が利用できることになる。

$$\widehat{N}^{\#a}_{\delta\cdot d\cdot a} = \sum_{j\in\delta\cdot d\cdot a} w_j^{\#a} \cong C_{\delta\cdot d\cdot a} \tag{21}$$

$$\widehat{Y}^{\#a}_{\delta\cdot d\cdot a} = \sum_{j\in\delta\cdot d\cdot a} w_j^{\#a} y_j \cong \frac{C_{\delta\cdot d\cdot a}}{\widehat{N}^{\#}_{\delta\cdot d\cdot a}} \widehat{Y}^{\#}_{\delta\cdot d\cdot a} \tag{22}$$

$$\hat{\mu}^{\#a}_{\delta\cdot d\cdot a} = \frac{\widehat{Y}^{\#a}_{\delta\cdot d\cdot a}}{\widehat{N}^{\#a}_{\delta\cdot d\cdot a}} \cong \frac{\widehat{Y}^{\#}_{\delta\cdot d\cdot a}}{\widehat{N}^{\#}_{\delta\cdot d\cdot a}} = \hat{\mu}^{\#}_{h,\delta\cdot d\cdot a} \tag{23}$$

ここで注意すべきは、調整リサンプル・ウェイトを作成する際に使用した性別や年齢区分を使用せずに平均値を算出する、もしくはこれらカテゴリーの一部をプールして平均値を算出する場合である。例えば、年齢区分を使用せずに男女・就業の有無別統計量を計算する場合、調整リサンプル・ウェイトを用いれば、当然、人口は国勢調査ベースの部分母集団と一致し、総計および平均値は目標部分母集団を国調人口としたときの推定値として得られる。

しかし、平均値に関しては、男女・年齢・就業の有無別平均値であればリサンプル・ウェイトとこれを調整したリサンプル・ウェイトによる結果とは一致するが、これらの一部変数を使用せずに平均値を算出した場合には、調整リサンプル・ウェイトによる結果とは必ずしも一致しない。そのため部分母集団の分析を目標とする場合には、調整リサンプル・ウェイトの利用をまず第１に考慮する必要がある。

[5)] 実際には、社会調の調査時点と人口調整に用いる国勢調査の調査時点がずれてしまうが、今回は平成17年国勢調査人口を用いてリサンプル・ウェイトを補正している。

$$\widehat{N}^{\#a}_{\delta\cdot a} = \sum_{j\in\delta\cdot a} w_j^{\#a} \cong C_{\delta\cdot a} \tag{24}$$

$$\widehat{Y}^{\#a}_{\delta\cdot a} = \sum_{j\in\delta\cdot a} w_j^{\#a} y_j \cong \sum_d \frac{C_{\delta\cdot d\cdot a}}{\widehat{N}^{\#}_{\delta\cdot d\cdot a}} \widehat{Y}^{\#}_{\delta\cdot d\cdot a} \tag{25}$$

$$\hat{\mu}^{\#a}_{\delta\cdot a} = \frac{\widehat{Y}^{\#a}_{\delta\cdot a}}{\widehat{N}^{\#a}_{\delta\cdot a}} \cong \sum_d \frac{C_{\delta\cdot d\cdot a}}{C_{\delta\cdot a}} \cdot \frac{\widehat{Y}^{\#}_{\delta\cdot d\cdot a}}{\widehat{N}^{\#}_{\delta\cdot d\cdot a}} \tag{26}$$

3．検証のアプローチ

3.1　基本統計量の検証

検証のために使用したデータは、平成18年度社会生活基本調査の匿名データである。リサンプリング・データとして24万ケースが得られており、これをすべて使用して検証する。具体的には、ウェイトを使用しない場合、あるいは上記で定義した各種のウェイトを使用した場合に実際の推定結果に相違があるか否かをテストし、あればどのような相違が生じているのかについて検討を加える。使用するウェイトと推定値を表１に整理しておく。本稿では社会生活基本調査を検証の素材としていることから、生活時間分析における母集団特性値の基本統計量として最もよく利用される総平均時間（総時間／総人口）を対象に精査している。なお、ウェイトを使用しない推定とは、言うまでもなくサイズnの単純無作為抽出標本からの推定を想定していることと同義である。

表１　平均推定量の検証に関するウェイトの種類と表記方法

ウェイトの種類	ウェイト不使用	リサンプル・ウェイト $w_j^{\#}$	調整リサンプル・ウェイト $w_j^{\#a}$	調整リサンプル・ウェイト $w_j^{\#b}$
推定値の表記	samp	pop (w)	pop (wa)	pop (wb)
男女・年齢別	$\hat{\mu}^{0}_{\delta\cdot d}$	$\hat{\mu}^{\#}_{\delta\cdot d}$	$\hat{\mu}^{\#a}_{\delta\cdot d}$	$\hat{\mu}^{\#b}_{\delta\cdot d}$
男女・年齢・就業の有無別	$\hat{\mu}^{0}_{\delta\cdot d\cdot a}$	$\hat{\mu}^{\#}_{h\delta\cdot d\cdot a}$	$\hat{\mu}^{\#a}_{\delta\cdot d\cdot a}$	
男女・就業の有無別	$\hat{\mu}^{0}_{\delta\cdot a}$	$\hat{\mu}^{\#}_{\delta\cdot a}$	$\hat{\mu}^{\#a}_{\delta\cdot a}$	
男女・年齢・配偶関係別	$\hat{\mu}^{0}_{\delta\cdot d\cdot b}$	$\hat{\mu}^{\#}_{\delta\cdot d\cdot b}$		$\hat{\mu}^{\#b}_{\delta\cdot d\cdot b}$
男女・配偶関係別	$\hat{\mu}^{0}_{\delta\cdot b}$	$\hat{\mu}^{\#}_{\delta\cdot b}$		$\hat{\mu}^{\#b}_{\delta\cdot b}$

（注）調整リサンプル・ウェイトの$w_j^{\#a}$は男女・年齢・就業の有無まで調整、$w_j^{\#b}$は男女・年齢・配偶関係まで調整したウェイトである。

3.2　回帰分析に関する検証

目的変数をy、説明変数ベクトルを$\boldsymbol{x}$とする回帰モデル$y=\boldsymbol{x}'\boldsymbol{\beta}+e,\ e\sim N(0,\sigma^2)$を考える。一般に回帰係数ベクトル$\boldsymbol{\beta}$の最尤推定量は、尤度$L(\boldsymbol{\beta})$を最大化すればよいから、次のスコア関数（あるいは推定方程式）$sc(\boldsymbol{\beta})$をゼロとおき推定値$\widehat{\boldsymbol{\beta}}$を求めればよい。

$$sc(\boldsymbol{\beta})=\frac{\partial}{\partial\boldsymbol{\beta}}lnL(\boldsymbol{\beta})=\sum_i\frac{\boldsymbol{x_i}(y_i-\boldsymbol{x_i'\beta})}{\sigma^2}=\boldsymbol{0} \tag{27}$$

標本調査データからの回帰モデルの推定については、（実在の有限）母集団の位置付けによりデザイン・ベースとモデル・ベースという異なるアプローチがある[6)]。本稿では、前者の立場から、回帰係数は母集団において［スコア関数＝0］を満足するような母集団特性値として性格付けし、それを標本調査データから推定する。すなわち、母集団Uにおける全数レベルのスコア関数(28)に対して、調査ウェイトで母集団に戻した標本スコア関数(29)をその推定値として、それを0とするよう$\boldsymbol{\beta}$の推定値を定めればよい[7)]。

$$sc_U(\boldsymbol{\beta})=\sum_{i\in U}\frac{\boldsymbol{x_i}(y_i-\boldsymbol{x_i'\beta})}{\sigma^2} \tag{28}$$

$$\widehat{sc_U}(\boldsymbol{\beta})=\sum_{i\in s}w_i\frac{\boldsymbol{x_i}(y_i-\boldsymbol{x_i'\beta})}{\sigma^2} \tag{29}$$

ところで、回帰分析などでは、通常、抽出率の逆数として作成される線形推定用乗率を用いるが、社会調匿名データでは国調人口ベースに調整された比推定用乗率しか付与されていないため、実際にはこの乗率を使用するしかない。このような問題を含めて、使用する乗率による推定量の近似精度の相違を評価しておく必要がある。推定にウェイトを用いない場合、あるいは線形推定用乗率や調整ウェイトを回帰分析に利用した場合、どの選択がよく母集団特性値を捉えることができるのか、このような点を明らかにするために回帰推定については次のような実験を行うことにした。

①　社会生活基本調査の匿名データを母集団に設定する。

②　そこから無作為標本を抽出するが、無回答などにより回収標本に偏りがあるケースを想定するとともに、調整ウェイトの評価に紛れがないように、標本の単純構成としては母集団の人口構成比からかなりの程度乖離が生じるように設計した標本を推定実験に用いることにする。具体的には、抽出率を有配偶で45歳未満の層は0.025、有配偶で45歳以上の層は0.1、それ以外の層は0.05として標本抽出を行い、全体としては、母集団人口約24万ケースから、約１万７千ケースを抽出する。

③　この標本に対して、線形推定用乗率、男女・年齢別人口まで調整した乗率、および男女・年

[6)] 標本調査データに対する回帰係数や標準誤差の推定については、土屋(2009) pp.222–230を参照。また、有限母集団パラメータや超母集団パラメータの推定に関する議論の詳細はBinder & Roberts (2003) を参照のこと。

[7)] ここでは尤度概念は形式であり、推定された係数の解釈問題は残るが、標本調査の技術論理としては一貫している。

齢・配偶関係別人口まで調整した乗率を作成し、回帰パラメータを推定する（表２参照）。回帰分析の目的変数には仕事時間、説明変数には性別、年齢、年齢二乗、配偶関係、自由時間（第３次活動時間）を用いる。

④　結果が標本に依存しないように、①から③の手順をループで50回繰り返す。それらの推定結果について、回帰係数の信頼区間に真値が含まれる回数をカバレッジとして計測する。また、回帰係数がゼロから有意に離れているかどうかを示すｐ値について、0.05以下である回数、および0.1以下である回数をそれぞれカウントすることで、ウェイト使用の妥当性を評価する。

表２　回帰分析の検証に関するウェイトの種類と表記方法

ウェイトの種類	ウェイト不使用	線形推定用 乗率 w_i	調整 ウェイト $v_j^{\#}$	調整 ウェイト $v_j^{\#b}$
推定値の表記	samp	pop (w)	pop (v)	pop (vb)

（注）調整ウェイトの $v_j^{\#}$ は男女・年齢まで調整、$v_j^{\#b}$ は男女・年齢・配偶関係まで調整したウェイトである。

４．検証結果

４.１　標本設計の範囲内での基本統計量の推定

（1）男女・年齢階級別の推定値

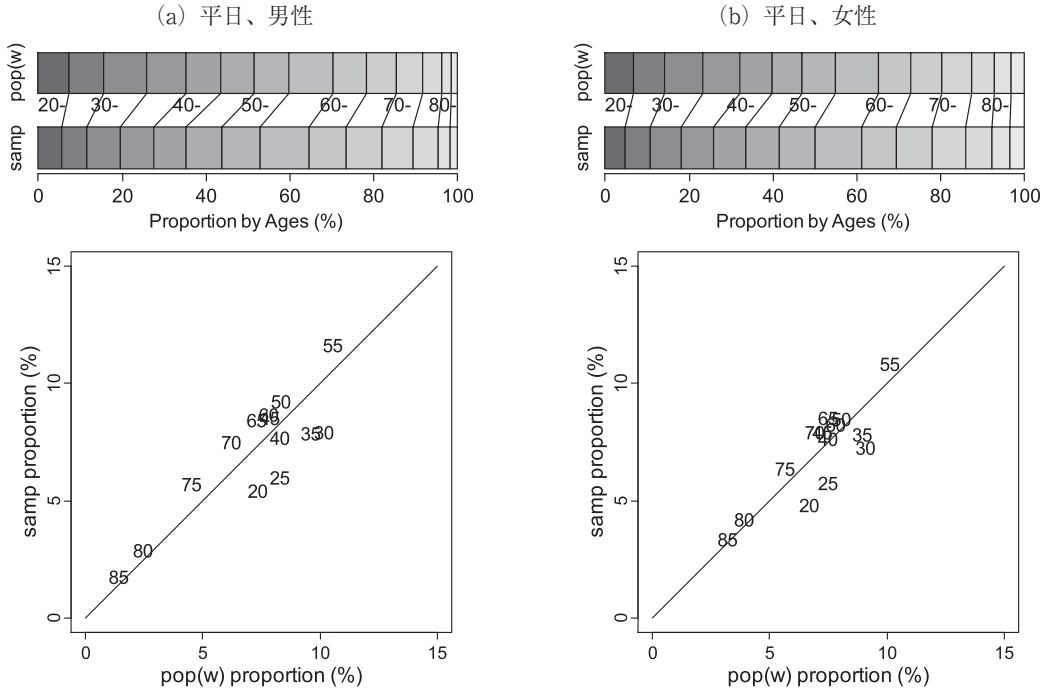
まず、匿名データに付与されているウェイトにより復元した母集団推定値pop（w）を、ウェイトを使用しない場合の推定値sampと比較してみよう。図２は、男女・年齢別人口構成比をグラフ化したものである。明らかに男性・女性ともに、ウェイトを使用しない場合には40歳以上の人口比重がかなり高めになってしまい、年齢構成において一種の上方バイアスをもつことが確認できる。

このような人口構成比の相違は、生活時間に関する各種変数の総平均時間に影響を及ぼす可能性が高い。そこで、曜日別に、仕事時間および通勤時間について、総平均時間（点推定値）とその信頼区間を算出し、その結果を整理したものが図３である。ここでは、30–34歳男性集団についてだけ、例示的に推定結果を示している。○印と実線がウェイトを使用したケースpop（w）、×印と破線がウェイトを使用しないケースsampである。ウェイトを使用したpop（w）が本来の信頼区間（の近似）と考えてよい。

仕事時間についてみると、いずれの曜日についてもsampよりpop（w）の信頼区間の幅が広くなっている。点推定値は、木曜日を除けば、ウェイトの違いにかかわらず似たような値を示している。木曜日に関しては、調整ウェイトによる結果とウェイト不使用での結果に大きな乖離がみられ、信頼区間も大きく異なる結果になった。

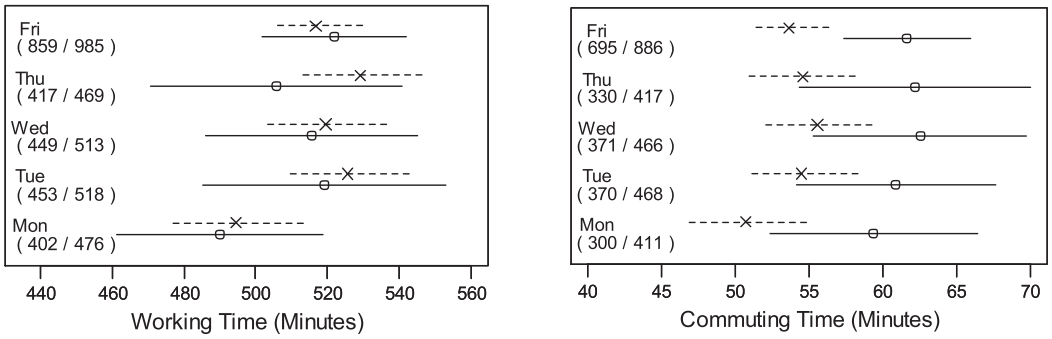
通勤時間に関しては、仕事時間よりも２つの推定値間の乖離が著しい。sampはpop（w）に対して下方バイアスをもつ。さらに、sampの信頼区間（破線）に、pop（w）の総平均時間（○印）が

図 2　男女・年齢階級別人口構成比（20歳以上）



（注）図中の数値は、年齢 5 歳階級の各区分の最初の年齢を示している。

図 3　仕事時間および通勤時間の総平均時間（30–34歳・男性）



（注）○印と実線はpop(w)、×印と破線はsampによる平均値と信頼区間をそれぞれ示す。なお、カッコ内は（行動者標本サイズ / 標本サイズ）を示す。

含まれないケースも複数観察される。とくに、金曜日については、sampとpop（w）の信頼区間が重なっておらず、sampでは真の母平均値が捉えられない可能性が高い。

このような特性は、30–34歳男性集団に限定されることではなく、行動種類や年齢層、曜日の違いにより、方向や程度の差はあれ、他の集団でも観察されるはずである。そこで、すべての男女・年齢別集団に拡げ、推定結果の特徴を次項でみていこう。

（2）男女・年齢階級別推定値の相対バイアス率

図4－1、4－2には、行動種類別総平均時間に関して、男女・年齢階級・曜日別samp推定値の相対バイアス率を全ての行動種類について整理している。ただし、バイアスはpop（w）の推定値を真値とみなし、そこからの乖離を次のような相対バイアス率で測っている[8)]。

$$\text{相対バイアス率} = \frac{\text{samp 推定値} - \text{pop（w）推定値}}{\text{pop（w）推定値}} \times 100$$

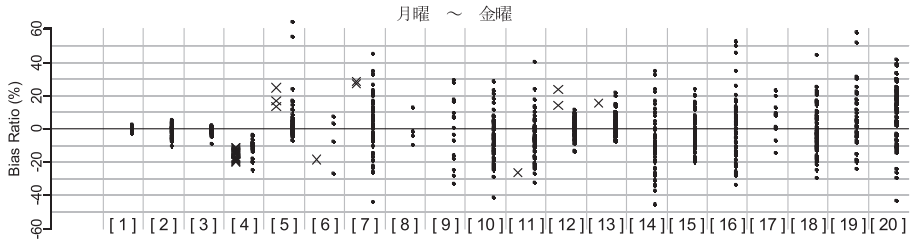
また図中のマークは、行動種類毎に年齢（13階級）別samp推定値の相対バイアス率をプロットしたものである。したがって、平日については5曜日×13年齢階級の65個、日曜は13年齢階級の13個の点がそれぞれプロットされている。そのうち、pop（w）推定値がsampの95%信頼区間に含まれていない場合の相対バイアス率は「×」（それ以外は「・」）で表示している。「×」が多ければ、推定誤差を考慮してもsamp推定値が不適切であることを意味する。なお、相対バイアス率は分母の大小に依存するため、pop（w）による行動種類別総平均時間推定値の年齢階級間の最大値（max）と最小値（min）を下段に示している。

男性（図4－1）については、平日で通勤（4）、学習（6）、移動（11）に関して下方バイアスがみられる。逆に、仕事（5）、家事（7）、テレビ（12）、休養・くろつぎ（13）の行動種類に関しては上方バイアスが観察される。また日曜日は、平日よりも、バイアスをもつ行動種類が少ない。買い物（10）、テレビ（12）は下方バイアス、仕事（5）、育児（9）、交際・つきあい（18）については上方バイアスが窺える。

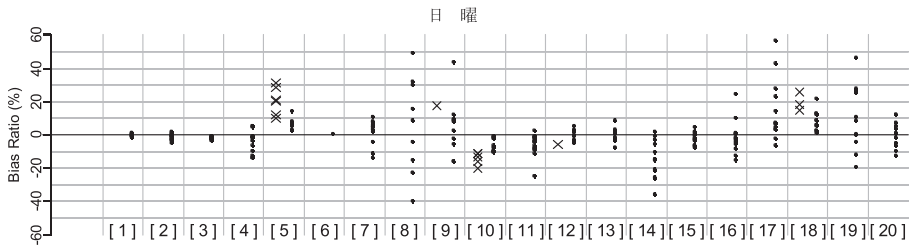
女性の年齢別推定値の相対バイアス率（図4－2）からは、平日には、食事（3）、通勤（4）、買い物（10）で下方バイアスが、仕事（5）で上方バイアスが生じている。テレビ（12）については、下方・上方バイアスの両方が観測され、年齢や曜日によりバイアス方向が異ることが分かる。さらに日曜については、男性の傾向と同様に、平日よりもバイアスをもつ行動種類が少ない。具体的には、身の回りの用事（2）、買い物（10）、休養・くつろぎ（13）では下方バイアスが、逆に、仕事（5）で上方バイアスがみられる。

8）相対バイアス率は、推定量の期待値、あるいは推定値平均をベースに定義されるべきものであるが、ここでは標本設計のベースとなる男女・年齢別統計量であり、また十分な標本サイズが確保されているので、近似指標として推定値をそのまま使用いている。

図 4－1　年齢階級・行動種類別相対バイアス率（20歳以上、男性）

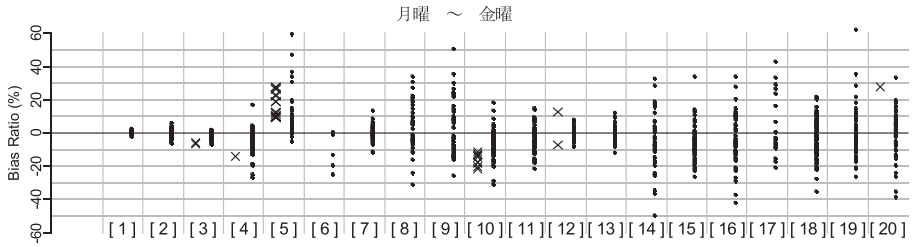


	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]	[11]	[12]	[13]	[14]	[15]	[16]	[17]	[18]	[19]	[20]
max	607	85	128	76	575	125	65	5.5	15	25	36	320	199	30	71	32	11	42	69	26
min	417	55	77	3.5	42	75	3.2	2.0	2.0	4.2	11	71	53	3.7	15	3.0	4.6	7.1	2.8	4.9

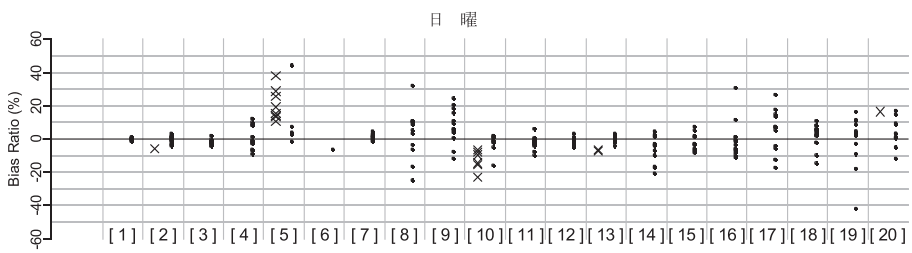


	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]	[11]	[12]	[13]	[14]	[15]	[16]	[17]	[18]	[19]	[20]
max	606	83	125	17	164	16	49	3.9	28	45	57	293	166	21	122	29	17	68	28	28
min	488	58	82	1.1	28	16	8.0	1.0	1.4	10	16	126	89	7.9	33	8.1	5.4	7.8	1.6	16

図 4－2　年齢階級・行動種類別相対バイアス率（20歳以上、女性）



	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]	[11]	[12]	[13]	[14]	[15]	[16]	[17]	[18]	[19]	[20]
max	592	101	130	63	325	80	241	14	108	45	40	266	202	22	51	20	10	44	62	38
min	396	74	83	2.3	10	56	30	1.5	3.3	7.6	10	79	63	2.1	14	3.4	2.8	11	3.1	6.5



	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]	[11]	[12]	[13]	[14]	[15]	[16]	[17]	[18]	[19]	[20]
max	606	94	123	17	158	16	217	9.3	80	58	62	249	183	18	74	15	12	61	34	24
min	452	82	90	0.6	5.3	16	33	0.9	1.1	10	9.8	114	77	3.3	28	3.5	3.1	12	1.2	15

（注）max, minは総平均時間の最大値および最小値をそれぞれ示している。pop（w）推定値が信頼区間内に含まれている場合の相対バイアス率は「×」、それ以外の相対バイアス率は「・」で示している。[　]内の数字は行動種類の符号（付表）を示す。

4.2　標本設計から外れる部分母集団の推定

(1) 人口構成比に大きな違いがない場合（平日、男性、有業者）

図5には、平日の有業者の男性についてpop(wa)に対するsampまたはpop(w)による年齢階級別人口構成比を示している。有業者の男性に関しては、pop(wa)とpop(w)との間には年齢別人口構成比に大きな違いはみられないことがわかる。これに対してsamp は他の2者の推定値と大きく異なる。pop(wa)の推定値は母集団比率にほぼ等しくなるように推定していることから、これと比較すれば、sampによる推定値は母集団比率をかなり歪める結果となることがわかる。

さて、真値に近い値を与えるpop(wa)に対するsampまたはpop(w)の相対バイアス率をプロットしたものが図6である。図4と同様の形式で作図しているが、上段2図には曜日・年齢階級別推定値の相対バイアス率を行動種類別にプロットしている（5曜日×13年齢階級の65個の点）。他方で、下段2図には年齢別ではなく、曜日別推定値の相対バイアス率が行動種類別にプロットされており、行動種類ごとに5曜日分の数値がプロットされていることに注意されたい。

明らかに、ウェイト不使用（samp）の場合には、曜日・年齢階級別の推定値に関して通勤（4）、仕事（5）、移動（11）、テレビ（12）、休養・くろつぎ（13）などでバイアスが顕著である。さらに年齢層を考慮しない曜日別の推定値に関しては、家事（7）などが加わりバイアスの発生する行動種類が異なっている。

これに対して、年齢変数を使用しても使用しなくても、pop(wa)に対するpop(w)のバイアスはほとんどみられない。就業状況にかかわりなく、年齢別構成比に大きな違いがないため、このレベルにおいてバイアスは生じていない。

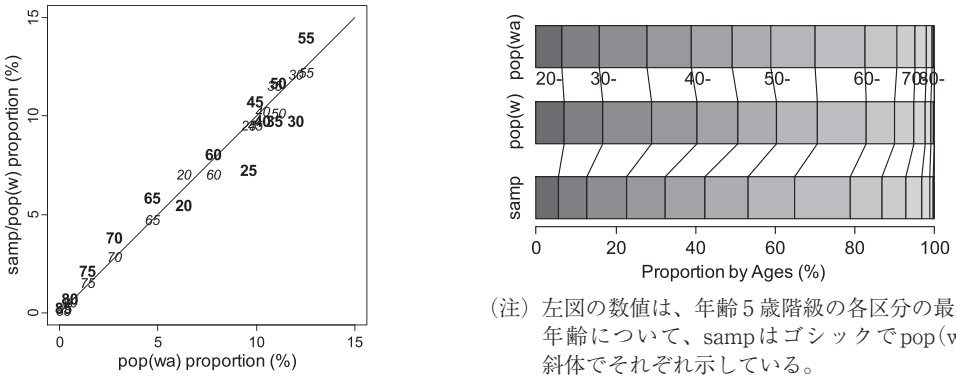
(2) 人口構成比が大きく違う場合（平日、男性、有配偶）

平日の有配偶男性の年齢階級別人口構成比（図7）が示すように、性別・年齢まで修正されているウェイトpop(w)と、性別・年齢・配偶関係まで調整したウェイトpop(wb)とでは、人口構成比に大きな相違があることがわかる。当然、sampにおいてもpop(w)やpop(wb)とは構成比が大きく異なっている。

pop(wb)を基準とした相対バイアス率のグラフが図8である。sampの推定値は、多くの行動種類でバイアスが生じる可能性が高いことを示唆している。これに対して、pop(w)の推定値に関しては、曜日・年齢階級別プロットが示すように、平均値が年齢別に算出されているため、pop(w)とpop(wb)の差はみられない。年齢別構成比に違いがあっても、年齢別の平均値であれば影響はないためである。

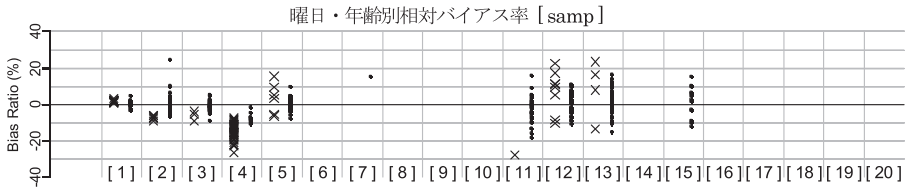
しかし、全年齢合計の曜日別プロットをみれば明らかなように、pop(w)ではpop(wb)とウェイト調整レベルが異なるため、年齢構成の相違に起因するバイアスの発生が顕著に見て取れる。人口構成比が異なるはずの属性（ここでは年齢など）に関して細分せずに平均値を算出する場合、ウェイト使用・不使用による年齢別構成比の違いがバイアス発生頻度の違いとなって強く表れている。

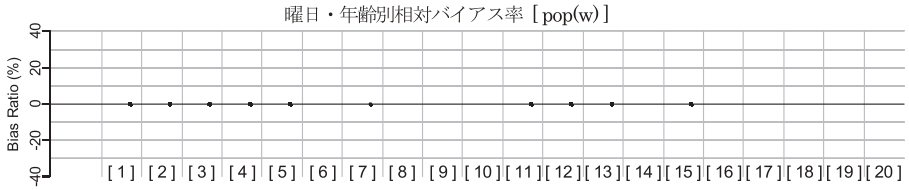
図 5　年齢階級別人口構成比（平日、20歳以上、男性、有業者）



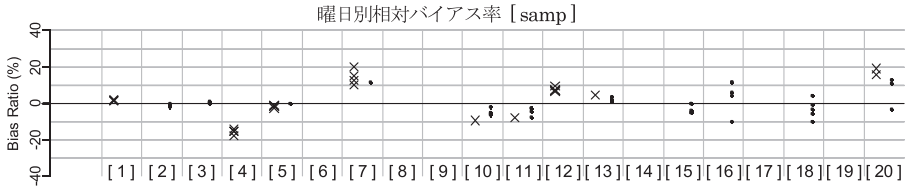
（注）左図の数値は、年齢５歳階級の各区分の最初の年齢について、sampはゴシックでpop（w）は斜体でそれぞれ示している。

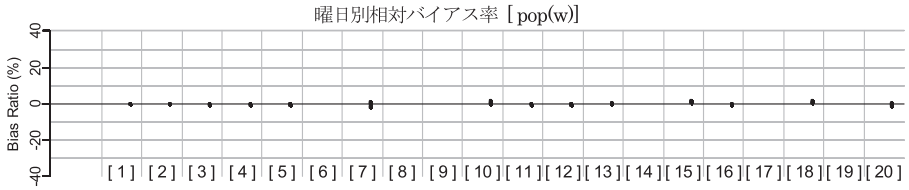
図 6　総平均時間の相対バイアス率（月曜～金曜、20歳以上、男性、有業者）





	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]	[11]	[12]	[13]	[14]	[15]	[16]	[17]	[18]	[19]	[20]
max	561	79	148	75	593	-	18	-	-	-	34	249	111	-	57	-	-	-	-	-
min	415	53	75	22	247	-	18	-	-	-	16	68	50	-	20	-	-	-	-	-

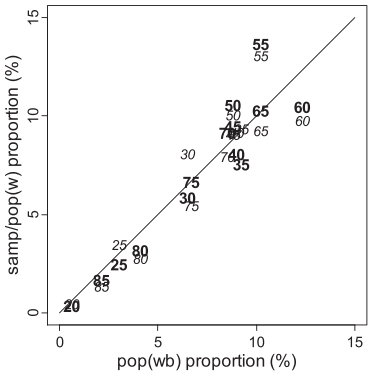


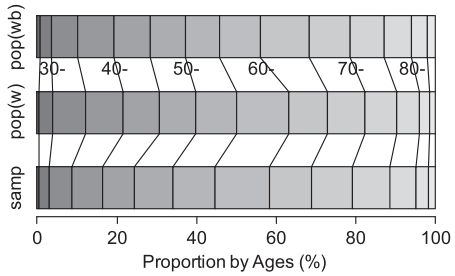


	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]	[11]	[12]	[13]	[14]	[15]	[16]	[17]	[18]	[19]	[20]
max	445	63	94	63	511	-	8.9	-	-	7.7	26	107	65	-	29	8.0	-	19	-	7.9
min	438	61	90	60	500	-	7.5	-	-	7.1	18	100	61	-	25	6.3	-	12	-	7.1

（注）max, minは、pop（wa）の総平均時間の最大値および最小値をそれぞれ示している。また、samp推定値またはpop（w）推定値が信頼区間内に含まれていない場合の相対バイアス率は「×」、それ以外の相対バイアス率は「・」で示している。［　］内の数字は行動種類の符号（付表）を示す。

図 7　年齢階級別人口構成比（平日、20歳以上、男性、有配偶）

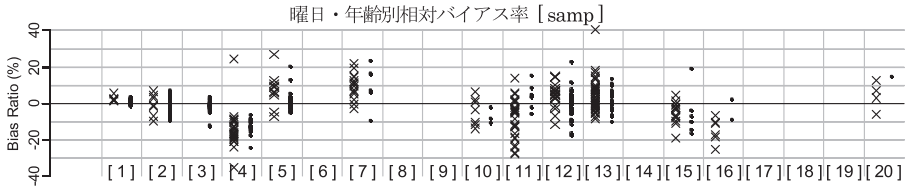




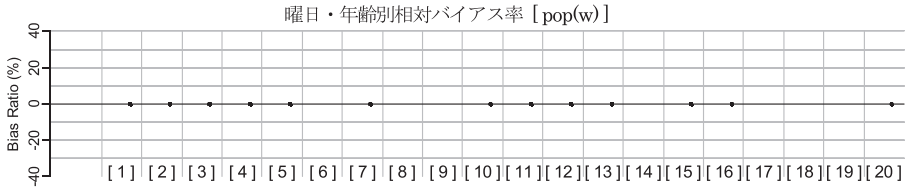
（注）左図の数値は、年齢５歳階級の各区分の最初の年齢について、sampはゴシックでpop(w)は斜体でそれぞれ示している。

図 8　総平均時間の相対バイアス率（月曜～金曜、20歳以上、男性、有配偶）

曜日・年齢別相対バイアス率［samp］

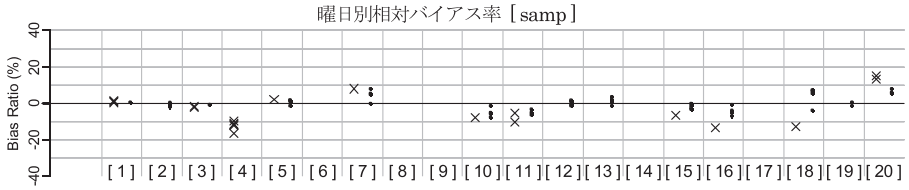


曜日・年齢別相対バイアス率［pop(w)］

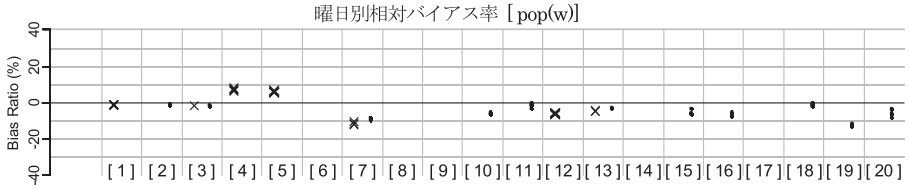


	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]	[11]	[12]	[13]	[14]	[15]	[16]	[17]	[18]	[19]	[20]
max	464	67	103	46	391	-	21	-	-	11	29	157	79	-	37	14	-	18	12	13
min	460	66	99	43	376	-	17	-	-	9.3	22	144	71	-	33	13	-	12	11	11

曜日別相対バイアス率［samp］



曜日別相対バイアス率［pop(w)］



	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]	[11]	[12]	[13]	[14]	[15]	[16]	[17]	[18]	[19]	[20]
max	610	87	130	75	609	-	54	-	-	22	37	309	212	-	72	33	-	-	-	24
min	412	44	75	18	81	-	20	-	-	13	18	51	51	-	21	18	-	-	-	16

（注）max, minは、pop(wb)の総平均時間の最大値および最小値をそれぞれ示している。また、samp推定値またはpop（w）推定値が信頼区間内に含まれていない場合の相対バイアス率は「×」、それ以外の相対バイアス率は「・」で示している。［　］内の数字は行動種類の符号（付表）を示す。

4.3　回帰係数の比較検証（シミュレーション）

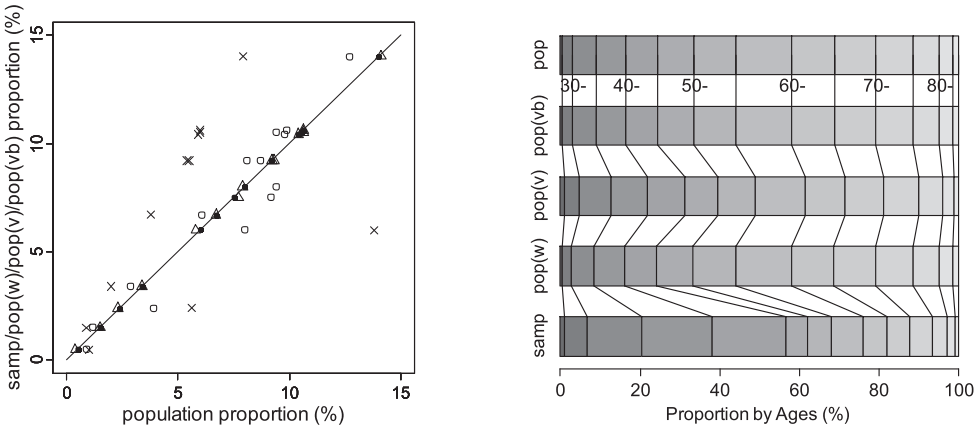
3.2節で述べたように、回帰係数の検証については、社会調ミクロデータを母集団としてそこからの抽出標本を用いて、ウェイト不使用（単純無作為抽出と想定）のsamp、年齢・配偶関係を抽出変数としたときの線形推定用乗率によるpop（w）、男女・年齢まで調整した比推定用乗率を使うpop（v）、さらに加えて配偶関係も調整した比推定用乗率によるpop（vb）という4種類の推定値を算出するという実験を50回繰り返している。評価用に使用したモデルは、仕事時間を目的変数として、性別、年齢、年齢の二乗、配偶関係、自由時間（第3次活動）で説明する線形回帰モデルである。なお、母集団特性値（真値）はpopと表記している。

これまでと同様に、年齢構成比を例に、各ウェイトによる推定量の特徴を母集団比率（pop、真値）との比較で見ておこう。図9は有配偶男性についての推定年齢構成比（50回の抽出実験の平均値）を整理している。配偶関係まで修正したpop（vb）では、当然、母集団比率と一致している。線形推定用乗率を用いたpop（w）は母集団比率に近いが、性別・年齢まで修正したpop（v）はやや乖離傾向にあり、ウェイトを用いないsampではほぼ母集団の年齢構成比率が捉えられないことがわかる。

回帰分析に関する抽出実験結果の代表例として年齢を取り上げ、その回帰係数と信頼区間の50回分の結果を図10に整理している。ウェイトを使用しないsampのパフォーマンスの悪さが一目瞭然である。さらにここから、50回の中で信頼区間に真値が含まれる回数をカバレッジとして計測したものをすべての回帰係数について計算し、表3にまとめている。また、50回の抽出実験において5％水準または1％水準で有意となった回数も示している。

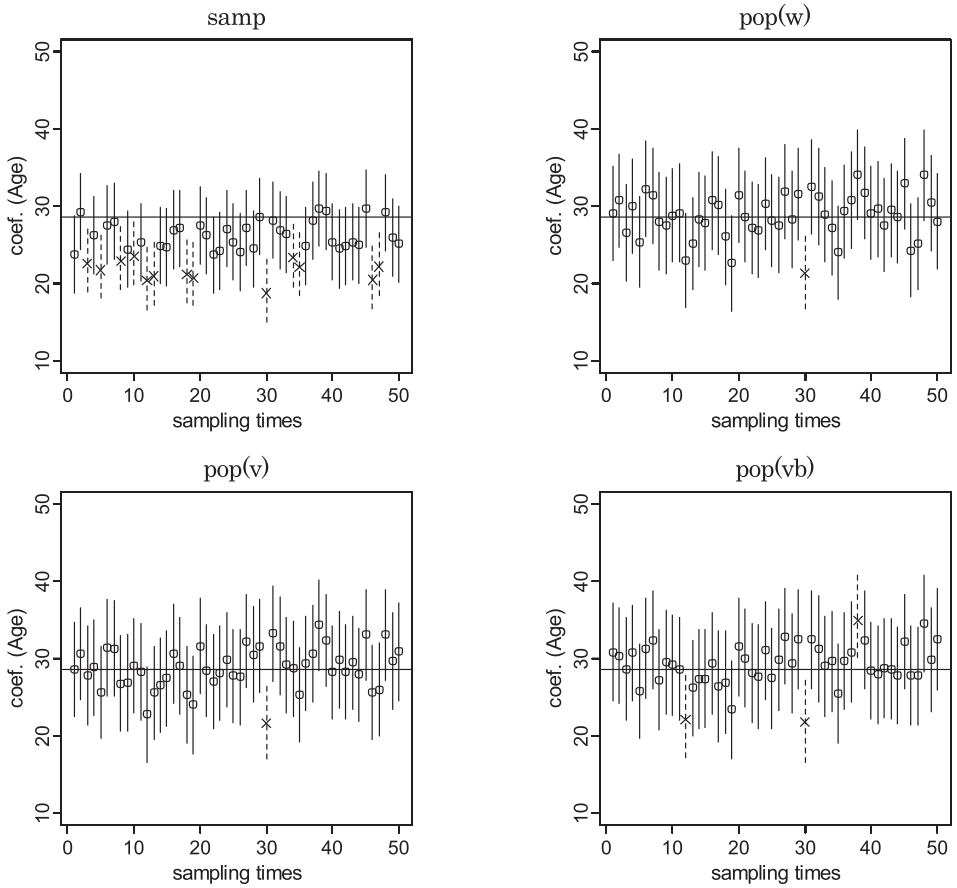
カバレッジをみると、sampでは年齢の係数で50回中37回、有配偶ダミーで3回しか信頼区間に真値を含んでいない。pop（w）、pop（v）、pop（vb）については、どちらも47回以上信頼区間内に真

図9　回帰分析用データセットの年齢階級別人口構成比（20歳以上、有配偶男性）



（注）「●」「○」「△」「×」印は、それぞれpop（vb）、pop（v）、pop（w）、sampを示している。

図10　回帰係数の信頼区間（標本抽出50回分）と真値の関係



（注）信頼区間内に真値（横実線）が含まれている場合には○印と実線で係数と信頼区間を示し、含まれていない場合には×印と破線で係数と信頼区間を示している。

表３　回帰係数の信頼区間のカバレッジと有意性（単位：回）

変数名	pop	samp			pop（w）			pop（v）			pop（vb）		
	coef.	coverage	p5	p1	coverage	p5	p1	coverage	p5	p1	coverage	p5	p1
男性	−												
女性	−162.09	44（88）	50	50	49（98）	50	50	49（98）	50	50	50（100）	50	50
年齢	28.52	37（74）	50	50	49（98）	50	50	49（98）	50	50	47（94）	50	50
年齢×年齢	−1.90	46（92）	50	50	48（96）	50	50	48（96）	50	50	48（96）	50	50
未婚	−												
有配偶	−71.99	3（6）	50	50	49（98）	50	50	48（96）	50	50	48（96）	50	50
その他	−26.94	12（24）	12	18	47（94）	46	47	47（94）	37	39	47（94）	45	47
不明	−34.84	44（88）	6	10	47（94）	6	10	48（96）	14	17	47（94）	9	11
自由時間	−0.75	45（90）	50	50	47（94）	50	50	47（94）	49	49	47（94）	50	50
定数項	570.00	37（74）	50	50	49（98）	50	50	50（100）	42	42	49（98）	50	50

（注）「−」はダミー変数の基準カテゴリーを示す。また、カッコ内は、抽出50回に対する比率を示している。表頭のp5およびp1は、50回の試行のうちで、有意水準５％または10％で有意になった回数をそれぞれ示す。

値が含まれている。sampでは人口構成比に歪みのある年齢と配偶関係に関する係数が捉えられていないことが分かる。また、係数が有意となる回数は、sampの配偶関係を示す係数で少なく、pop（w）、pop（v）、pop（vb）であれば未婚とその他（離別・死別）の係数は有意に異なるという結果が得られる。sampでは未婚とその他（離別・死別）の係数に違いはないと判断してしまう恐れがある。

今回の抽出実験では、人口構成比に大きな歪みがある場合、ウェイトを利用せずに回帰分析を行えば得られた回帰係数はバイアスをもち、有意性も正しく評価できないことが明らかとなった。さらに、一部のカテゴリカル変数（男女・年齢階級）まで調整したウェイト（社会調で付与されているウェイト）pop（v）では、やはり調整していない変数（配偶関係）に関する回帰係数、とくに「その他」の係数が有意性において低めの数値（５％水準で37回）となっている。回帰分析には線形推定用乗率を用いるべきではあるが、比推定用乗率しか与えられない場合には、使用するカテゴリカル変数（基本人口構成に関する変数）について適切に調整したウェイトpop（vb）を用いれば、線形推定用乗率の結果pop（w）に近似することが確認された。

５．おわりに

本研究では、社会生活基本調査ミクロデータを素材に、母集団パラメータを推定する場合の、調査ウェイトの取り扱いについて検討した。理論的にはウェイトの利用は当然のことであるが、実際にはウェイトといっても標本設計の抽出変数に制約され、任意の部分母集団の推定にそのまま利用できるわけではない。また匿名化ミクロデータでは標本設計に使う都道府県や調査区情報などが提供されない、あるいは世帯用の線形推定用乗率から作成された比推定用乗率だけが提供され、ユーザは選択の余地なく、それを使用せざるを得ないといった事情が加わる。このような現状を考慮して、本稿では、ウェイトの使用・不使用、調整ウェイトの作成と利用など、アプローチの違いによる推定バイアスの程度を計測し、またそのようなバイアスに対する現実的な対処法を提示することを課題とした。

その結果を整理すれば次のようである。社会調データについては、男女・年齢など分析の前提となる基本属性の違いによって行動や生活時間が大きく異なるため、平均値や比率などの基本統計量を算出する際には、分析対象である目標母集団の人口構成にとくに注意を払う必要がある。特定の属性に関して標本による推定構成比が母集団構成比から大きく歪んでいる場合、その属性変数のカテゴリー別に区分せずにプールしたまま算出した推定値についてはバイアスが生じるリスクが高い。とくに回帰分析では、推定構成比の歪みが係数の値と有意性の両方に強く影響を与える。線形推定用乗率であっても、本来の母集団人口構成比と異なる場合には、その歪みが推定値に影響を及ぼす。

平均値（比率）のような基本統計量であれ、回帰係数の推定であれ、このような問題への対処法

として３点指摘しておくことにする。

① まずは解析準備として、ウェイトの使用、もしくは不使用により人口構成比に母集団と推定値に大きな相違がないか確認する。違いがあれば、ウェイトの利用（修正含む）を前提に推定作業を計画する。

② 個人単位の分析の際には、少なくとも（提供されている）男女・年齢まで調整されたウェイトを使用する（標準誤差の推定にも注意する）。

③ 国勢調査など公的統計から基本属性別母集団人口が公表されている場合には、この情報を補助変数に利用して分析を進める（例えば、比推定、あるいは事後層化推定など）[9)]。

匿名化ミクロデータを含め、政府統計の個票データの２次利用には、標本設計などの調査法による制約から様々な工夫が必要となる。その中でも補助情報を利用した上記③のウェイト補正の考え方は有力な方法のひとつであり、比推定用乗率を補正ウェイトとして利用しようとするものである。ただ、乱用すれば過剰調整が発生し、ある標本要素に適正とはいえないウェイトが付されるリスクが高まる。そのため、実際の適用は慎重にならざるを得ない側面もある。また回帰をはじめモデル分析を目標とする場合に、比推定用乗率の使用を一般化できるほどの理論的根拠を用意できるわけではない[10)]。しかし現実のデータは多くの制約に縛られており、そこを潜り抜けて母集団という実体に迫るには、直面する問題を理解し、可能な改善を試み、その経験を蓄積し理論化するというプロセスが不可欠である。モデルベースの推定問題を含め、そのようなプロセスから得られる追加的な論点については稿を改めて論じることにしたい。

（付表） 行動種類と符号

符号	行動種類	符号	行動種類
1	睡眠	11	移動（通勤・通学を除く）
2	身の回りの用事	12	テレビ・ラジオ・新聞・雑誌
3	食事	13	休養・くつろぎ
4	通勤・通学	14	学習・研究（学業以外）
5	仕事	15	趣味・娯楽
6	学業	16	スポーツ
7	家事	17	ボランティア活動・社会参加活動
8	介護・看護	18	交際・つきあい
9	育児	19	受診・療養
10	買い物	20	その他

9) 本来は、ウェイトに補助情報を組み込むという方法は作業が煩雑となることから、キャリブレーション推定などの採用を優先的に検討することが望まれる。

10) 通常は、国調人口などの補助情報を利用して、ベース人口に合致するように線形推定用乗率を補正して作成される。この意味では比推定用乗率は基準人口への補正という機能が一義的であるが、その延長線上で、異なる単位（例えば世帯から個人など）への変換ウェイトとしても機能する。

参考文献

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

［2］ 栗原由紀子（2011），『生活時間データによる時空間共有行動の統計的解析』，博士論文.

［3］ 総務省統計局（2002），『平成12年国勢調査，調査区関係資料利用の手引』，日本統計協会.

［4］ 総務省統計局（2003），『平成18年社会生活基本調査報告』，財務省印刷局.

［5］ 高橋雅夫・臼井彩子（2005），「平成13年社会生活基本調査における標本の代表性と調査結果の推定について」，『統計研究彙報』，第62号，pp.23–70.

［6］ 土屋隆裕（2009），『概説標本調査法』，朝倉書店.

［7］ 舟尾暢男（2005），『The R Tips』，株式会社九天社.

［8］ 水野谷武志（2010），「欧州統一生活時間調査（HETUS）ガイドライン－2008年版（翻訳と解説）」，『統計研究参考資料』No.107，pp.21–23，法政大学日本統計研究所.

［9］ Binder, D.A. & G.R. Roberts（2003），“Design-based and Model-based Methods for Estimation Godel Parameters,” *Analysis of Survey Data*, Chapter 3.

［10］ Cochran, W.G.（1977），*Sampling Techniques*, Third Edition, John Wiley & Sons.

［11］ Eurostat（2009），*Harmonised European time use surveys: 2008 guidelines*, eurostat Methodologies and Working papers.

［12］ Kurihara, Y.（2010），“Estimation of Weekday Averages and Their Variance with The Resampled Data from The Survey on Time Use and Leisure Activities,” *The Annual of the Institute of Economic Research Chuo University*, No.41, The Institute of Economic Research Chuo University.

［13］ Skinner, C.J.（1989），*Analysis of Complex Surveys*, ed. C.J. Skinner, D.Holt & T.M.F. Smith, pp.23–58, John Wiley & Sons.

［14］ StataCorp.（2009），Stata Survey Data Reference Manual Release 11, pp.163–164.

［15］ Wolter, K.M.（2007），*Introduction to Variance Estimation Second Edition*, Springer.

［16］ Centre for Time Use Research のウェブサイト（http://www.timeuse.org/information/studies/）