統計数理(2010)
第 58 巻 第 1 号 3–23

特集「日本人の国民性調査研究— 平成期の 20 年 —」
[原著論文]

調査不能がある場合の標本調査における セミパラメトリック推定と感度分析： 日本人の国民性調査データへの適用

星野 崇宏†

(受付 2009 年 8 月 28 日；改訂 12 月 16 日；採択 2010 年 2 月 10 日)

要　　旨

社会調査や市場調査において，近年訪問調査などの従来型調査の回収率が低下してきており，調査不能による推定のバイアスが問題となっている．本論文では標本調査での調査不能を選択バイアスとして定式化し，既存の共変量調整法を利用する場合の問題点を指摘する．さらに共変量情報を十分に利用できない場合に，調査不能による推定のバイアスをどれくらい見積もれば良いかを議論するために有用であると考えられるモデルとして，「調査不能となるかどうか」にも回答値にも影響を与える隠れた共変量を潜在変数として仮定し，ディリクレ過程混合モデルによる表現を行うことで，セミパラメトリックに調査不能を調整するモデルを提案する．またこのモデルにおいて一部のパラメータを変化させることによって，「調査不能を考慮した上で，推定値にどの程度の信頼区間を考えるべきか」を考える感度分析を実施することができる．この手法を第 12 次日本人の国民性調査データに適用したところ，調査不能標本すべてが同一の回答を行うと仮定した場合には 95%信頼区間の幅が最大 50% となるのに対して，提案したモデルを用いることでせいぜい 13% 程度に抑えられることがわかった．

キーワード： ディリクレ過程混合モデル，傾向スコア，隠れた共変量，社会調査，共変量調整，選択バイアス．

1. 問題意識と目的

日本における社会調査や世論調査はこれまで，住民基本台帳や選挙人名簿から，(層別抽出などを含む)無作為抽出された対象者に対する訪問面接調査や訪問留置調査が中心であった．しかし近年このような従来型の調査での回収率の低下には著しいものがある．実際，統計数理研究所が実施している日本を代表する継続的な社会調査である「日本人の国民性調査」においても，最近の数回の調査では拒否や不在などによる調査不能の率が徐々に高まっており，第 1 次調査では調査不能率は 17%であったものが，第 12 次調査では 48%にまで達している．

このような標本調査での調査不能の問題に対して，これまでも標本調査論の研究分野では「調査不能の理由を明確にし，これを補助情報として利用する」方法(Särndal, 2005)や，調査デザインを工夫することでこれを回避するという方向で研究が行われてきた(例えば Groves et al., 2002)．

† 名古屋大学大学院 経済学研究科：〒464–8601 愛知県名古屋市千種区不老町

特に前者の方向性としては,「調査不能になるかどうか」にも「回答そのもの」にも関連する様々な変数(補助情報と呼ばれることが多いが,本論文では以降,共変量(Covariate)と呼ぶ)を同定し,回収標本と調査不能標本の間で共変量の分布が共通になるように調整を行う方法(以降,共変量調整と呼ぶ)が利用されることが多い.具体的には性別・年齢別や居住地域などのデモグラフィック変数を少数利用し,事後的に調整を行う方法として,事後層別や,レイキング法が利用されることが多い.

一般には「調査不能になるかどうか」にも「回答そのもの」にも関連する共変量は多数存在すると考えられるが,以降(第 3 節)に示すように,共変量が多数にのぼる場合にはこれらの方法を利用することができない.

一方,有意抽出に基づく調査,特にインターネット調査データを基準となる(通常は無作為抽出標本からの)調査データに近づける調整を行う方法として近年よく利用されているのが,セミパラメトリックな共変量調整法の一種である「傾向スコアを用いた重み付け推定法」である(Taylor et al., 2001; 星野, 2007; Schonlau et al., 2009).傾向スコアを用いた解析は,複数の共変量が存在する場合に「回答値の共変量への回帰関数」を設定せずに共変量調整を行う方法であり,「無作為抽出された,基準となる調査データ」を準拠集団として,「インターネット調査などの有意抽出に基づく調査データ」から「準拠集団での結果」を予測に利用する方法として,市場調査などで利用されている.

さて,無作為抽出標本の一部が調査不能になることに伴う調査データの偏りと,有意抽出に基づく調査の偏りは「欠測のあるデータからの推定の偏り」として同じ問題構造を持っている(星野, 2009, 5・6 章)ことから,理論上は傾向スコアなどの「複数の共変量を利用するセミパラメトリックな共変量調整法」を「調査不能になるかどうか」に対して利用することは可能である.しかし第 3 節に述べるように,調査不能となった対象者からはデモグラフィック変数のうち,性別・年齢別や居住地域などごく一部しか情報が得られないことが多い.

従って,傾向スコアを用いた重み付け推定法などの「複数の共変量を利用するセミパラメトリックな共変量調整法」を調査不能の問題に直接利用することは難しい.

一方,経済学では Heckman らの研究(Heckman, 1974)以降,母集団を代表しない標本から得られた推論の偏り(選択バイアスと呼ぶ)に関する理論的あるいは実証的な研究が蓄積されている.特に Heckman の一連の研究で仮定されているモデルはプロビット型のパラメトリックモデル(プロビット選択モデルと呼ばれる)であり,「(欠測する可能性のある)潜在的な回答の値」そのものによって選択されるかされないか(観測されるか欠測か)が決定される.選択バイアスの問題も有意抽出に基づく調査の偏りと同様に,「欠測のあるデータからの推定の偏り」として同じ問題構造を持っている(星野, 2009)ため,選択バイアスに対して開発された方法を標本調査における調査不能に対して利用することは可能である.但し,プロビット選択モデルの仮定が誤っている場合には推定には大きなバイアスが生じる可能性が指摘されており,仮定の少ないセミパラメトリックモデルを用いることで,よりロバストな解析が可能になると期待される.

そこで本研究では,標本調査において「潜在的な回答の値」そのものが調査不能の確率に影響を与えるモデルのうち,「回答傾向の潜在変数と共変量」が調査不能の確率に影響を与えるモデルを考える.また調査不能の確率のモデリングに従来のようなロジスティック回帰モデルではなく,仮定の弱いセミパラメトリックなモデルである有限ディリクレ過程混合モデル(Ishwaran and James, 2001)を用いて,調査不能に起因する選択バイアスを補正する方法を提案する.そして,第 12 次日本人の国民性調査データに対して適用し,回収率の低下のもとで信頼区間等の構成をどのように行うべきかを議論する.

本論文の構成は以下の通りである.第 2 節において調査不能標本が存在する場合の標本調査データを欠測のあるデータとして表現し,選択バイアスの問題として定式化する.第 3 節では

この問題に対する既存の解析手法とその問題点を提示する．第 4 節では「回答傾向の潜在変数と共変量」が調査不能の確率に影響を与えるセミパラメトリックモデル(セミパラメトリックベイズモデル，具体的には有限ディリクレ過程混合モデル)を提案し，Blocked Gibbs sampler を利用した推定法を示す．

第 5 節では第 12 次日本人の国民性調査への適用を行い，提案された推定値や信頼区間等が単純な集計と比べてどの程度変化するかを示すことで，回収率が高くはない場合に「調査不能を考慮した場合にはどの程度の幅をもって結果を解釈するべきか」について解析した具体例を示す．

2. 調査不能のある調査データの欠測データとしての表現

本論文では無限母集団を想定する．まず無作為抽出によって得られた標本のサンプルサイズを N とする．そのうち N_1 人については調査票が回収され，$N_2(=N-N_1)$ 人については調査不能であるとする．ここで関心の対象となっている項目の回答値を y とし，調査票が回収された調査対象者ならば $z=1$，調査不能ならば $z=0$ となるインディケータ変数を z とする．また，「調査不能になるか」どうかにも「回答値そのもの」にも関連する共変量を $\boldsymbol{x}$ とする．このとき，標本のレベルでのデータは図 1 のように表現することができ，$z=1$ ならば y が観測され，$z=0$ ならば y は欠測されるデータになっていると考えることができる．また，本論文では y が確率変数であると考え，その母集団分布の母数(母集団平均など)の推定に関心があるとする．

ここで，母集団平均 $E(y)$ の不偏推定値

$$\hat{E}(y)=\frac{1}{N}\sum_{i=1}^{N}y_i \tag{2.1}$$

は得られず，一方回収標本から得られる推定値

$$\bar{y}_{obs}=\frac{\sum_{i=1}^{N}z_iy_i}{\sum_{i=1}^{N}z_i} \tag{2.2}$$

は y と z が独立である場合を除き，母集団平均の不偏推定量にはならない．ここで i は第 i 回答者の値であることを表す．

母集団平均に限らず，回収標本のデータだけを用いて関心のある変数 y の分布の母数推定を行うと，推定値にはバイアスが存在する．この問題は計量経済学でよく議論される，「特定の対象者だけ標本に選択されたり，関心のある変数の観測値が得られる」選択バイアスの問題と全く同じ構造を有している．そこで選択バイアスに関する研究で利用される用語を利用して，「特定の対象者の選択」を

	回収標本(z=1)				回収不能標本(z=0)		
所属群(z)	1	1	1	1	0	0	0
対象者番号	1	2	・・・	N1	・・・	N-1	N
回答y	y_1	y_2	・・・	y_{N1}	・・・	y_{N-1}	y_N
共変量$\boldsymbol{x}$	$\boldsymbol{x}_1$	$\boldsymbol{x}_2$	・・・	$\boldsymbol{x}_{N1}$	・・・	$\boldsymbol{x}_{N-1}$	$\boldsymbol{x}_N$

図 1. 欠測データとしてみた「調査不能のある調査データ」．灰色は欠測しているデータを表す．

観測値による選択(Selection on observables)：選択されるかどうかは共変量などの観測値に依存する

観測されないものによる選択(Selection on unobservables)：選択されるかどうかは観測値以外の要因にも依存する

に分けて考えることとする．

具体的にはこの状況で「観測値による選択」が行われている場合とは

$$p(z|y,\boldsymbol{x}) = p(z|\boldsymbol{x}) \tag{2.3}$$

つまり共変量の値によって回収されるか調査不能になるかが決まるということである．一方「観測されないものによる選択」とは共変量 $\boldsymbol{x}$ だけではなく，潜在的な測定値 y そのものの値によっても回収されるかどうかが決定される場合である．

一方，統計学における欠測の議論(例えば Little and Rubin, 2002)からは，前者の場合には y の欠測は“ランダムな欠測(Missing at randon)”と言うことができる．従って，回収標本における「回答値の共変量への回帰モデル」を利用して母集団平均の一致推定量を構成できる．具体的には，式(2.3)をベイズの定理を使って書き直せば

$$p(y|\boldsymbol{x}, z=1) = p(y|\boldsymbol{x}, z=0) = p(y|\boldsymbol{x}) \tag{2.4}$$

となる．従って期待値を取って

$$E(y|z=1,\boldsymbol{x}) = E(y|\boldsymbol{x}) \tag{2.5}$$

が成立する．さらに共変量に関して期待値をとれば

$$E(y) = E_{\boldsymbol{x}}[E(y|\boldsymbol{x})] = E_{\boldsymbol{x}}[E(y|z=1,\boldsymbol{x})] = \int E(y|z=1,\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x} \tag{2.6}$$

と表現できる(ここで E_a は a について期待値を取ることを示し，$p(\boldsymbol{x})$ は共変量の密度関数である)．従って「($z=1$ での) y の $\boldsymbol{x}$ への回帰関数」を $g(\boldsymbol{x}|\boldsymbol{\beta})$ とし，$\boldsymbol{\beta}$ を回帰関数のパラメータとすると，母集団平均の推定量

$$\frac{1}{N}\sum_{i=1}^{N} g(\boldsymbol{x}_i|\boldsymbol{\beta}) \tag{2.7}$$

または

$$\frac{1}{N}\sum_{i=1}^{N} z_i y_i + (1-z_i)g(\boldsymbol{x}_i|\boldsymbol{\beta}) \tag{2.8}$$

は観測されているデータのみを用いた一致推定量である(但し $\boldsymbol{\beta}$ はその推定値で置き換えるが，一般に一致性のある推定値を利用すれば一致性は保たれる)．

一方，後者の「観測されないものによる選択」の場合には，式(2.4)は成立しない．従って式(2.7)や式(2.8)の推定量には一致性はない．

3. 既存の共変量調整について

3.1 既存の手法とその問題点

調査不能標本が存在する場合の結果の偏りを解決するための推定法を，大きく分けて以下の 5 つに分類して説明する．1 番目から 4 番目の方法は「観測値による選択」の仮定が成立する場合に一致推定値を与える方法である．また 2 の「重みづけ法及びキャリブレーション推定」

と 4 の「傾向スコアによる推定法」は類似の方法と言うこともできるが，ここでは問題点などの説明の都合上，これを分けて議論する．

1. **回帰モデルの利用**：「回答値の共変量への回帰モデル」を利用して調整を行う．具体的には式(2.7)や式(2.8)を用いて母集団平均を推定する方法である．標本調査法でいう(一般化)回帰推定量を用いてもよい．

2. **重みづけ法及びキャリブレーション推定**：1 つの共変量に着目して層別し，その共変量の周辺分布が母集団の周辺分布と等しくなるように層(さらに結果的には対象者)に重みをつけて推定を行う方法を事後層別法(Post-stratification)と呼ぶ．これを複数の共変量に対して行い，各共変量の周辺分布が母集団の周辺分布に等しくなるように重みを定める方法がレイキング法(Deming and Stephan, 1940; Ireland and Kullback, 1968)又は反復比例フィッティング(Iterative proportional fitting)である．実際には「特定の変数に注目し，まずその周辺分布が真値に合うように各セルの重みを計算し，次に別の変数の周辺分布が真値に合うように各セルの重みを計算し，, , 」という形で反復計算を実施する．

また，「共変量についての重み付き標本平均が母集団平均に一致する」という制約をつけて重みを計算する方法であるキャリブレーション推定(Deville and Särndal, 1992)は，事後層別法やレイキング法，一般化回帰推定量など様々な方法を下位に含む一般的な方法として考えることができる．

3. **代入法**：単一代入法やそれを複数繰り返す多重代入法などがあげられるが，共変量の情報を利用した代入法はマッチングとして理解することができる(星野, 2009)．

4. **傾向スコアを利用した共変量調整**：式(2.7)や式(2.8)による推定量は回帰モデルを利用しているが，共変量 $\boldsymbol{x}$ が高度であるほど，回帰関数の正しい指定が困難になる．そこで回帰関数を指定しない方法として，「共変量 $\boldsymbol{x}$ を用いて z を説明する」式(2.3)から「各回答者が回収される($z=1$ となる)確率」を求める．この確率が傾向スコアであり，具体的にはロジスティック回帰分析モデルなどの予測確率として計算できる．さて式(2.4)より

$$E_{\boldsymbol{x}}\left\{E\left(\frac{z}{E(z|\boldsymbol{x})}y\,\middle|\,\boldsymbol{x}\right)\right\}=E_{\boldsymbol{x}}\left\{E\left(\frac{z}{E(z|\boldsymbol{x})}\middle|\boldsymbol{x}\right)E(y\,|\,\boldsymbol{x})\right\}=E(y) \tag{3.1}$$

となるため，回答者 i の傾向スコアを $e_i=p(z_i=1|\boldsymbol{x})$ とすると，以下の IPW (Inverse probability weighting：逆確率重み付け)推定量

$$\frac{\sum_{i=1}^{N}\frac{z_i}{e_i}y_i}{\sum_{i=1}^{N}\frac{z_i}{e_i}} \tag{3.2}$$

は母集団平均の一致推定量になる．

5. **プロビット選択モデルの利用**：3.2 節にて詳しく紹介する．

まず 1 から 4 までの方法の問題点を示す．

回帰モデルでは回帰関数 $g(\boldsymbol{x}|\boldsymbol{\beta})$ を誤って設定すると母集団平均の推定値に大きなバイアスが生じる(星野, 2009)．特に「観測値による選択」であると考えられる程度に多くの共変量を利用する場合，回帰関数の誤設定の可能性が増大してしまう．

また事後層別法やレイキング法は計算方法の制約から，共変量に連続変数が存在するときには利用しにくいことや，各変数のカテゴリー数が多いと計算が難しくなること，共変量としてせいぜい 5・6 変数しか調整に利用できないこと，といった欠点がある．さらに，これらのモデルを用いることの問題点としてしばしば指摘されるのは「特異な対象者の重みを増して集計を行ってしまう可能性がある」ことである．たとえば 30 代男性は「仕事が忙しく，在宅率が

低い人が多い」ために調査不能率が高い．したがって集計時に計画標本での 30 代男性の比率に合わせるように調整を行うと，「仕事が忙しくなく，在宅率が高い」ために回収標本に含まれた調査対象者の意見が割り増されて集計されることになり，結果として事後層別やレイキング法による推定値は回収標本からの単純な集計値に比べて「調査不能の標本を含む計画標本全体から得られるはずの推定値」から乖離する可能性がある．

共変量情報を用いた代入法は「回収標本と調査不能標本間でのマッチング」として理解できるが，マッチングの方法の恣意性や共変量が多数にわたると生じる次元問題など（星野, 2009）が存在する．

一方，傾向スコアを利用した共変量調整については，インターネット調査の標本の偏りに起因するバイアスを調整を行う方法としてはうまく働くことが多くの研究で報告されている（例えば Taylor et al., 2001; 星野, 2007; 星野・森本, 2007）．傾向スコアを用いた調整法があまりうまく機能しなかったことを示す研究もあるが，その場合にはたまたま利用できる少数の共変量を利用して解析が行われており，「調整に利用するための共変量の候補となる変数の情報を大量に取得する」「その中から調整に利用すべき共変量を選択して解析を行う」というステップが踏まれていない．

傾向スコアを算出する際に利用する共変量の選択に関して，星野・前田（2006）は共変量選択の重要性を指摘し，具体的な共変量の選択方法を提案している．因果効果推定のために傾向スコアを利用する場合についても，Brookhart et al.（2006）がシミュレーション研究から星野・前田（2006）が提案した共変量選択法と同様の選択法が有効であることを示している．さて，傾向スコアを用いた選択バイアスに対する共変量調整の問題点として指摘されてきたのは，

(1) 傾向スコアを適用する際の条件である「強く無視できる割り当て」（Strongly Ignorable Treatment Assignment）（調査不能のある場合では式(2.3)，つまり「観測値による選択」に対応する）が成立しない可能性があること

(2) 傾向スコアを利用する際には，無作為抽出標本（または調査不能を含めた標本）において「無作為抽出標本と有意抽出標本の差異を決める共変量情報が十分得られている」必要があること（未回収標本から得られるのは性年齢や地域情報，未回収の理由程度である）

の 2 点である．但し，十分な共変量を得ることができれば，星野・前田（2006）が提案した方法で選ばれた共変量のセットを用いて(1)の問題が解決されることが期待できるため，本質的には(1)と(2)の問題は同じものである．実際，「インターネット調査を無作為抽出に基づく訪問調査に近づける調整」を行う場合には，両者において十分な共変量情報を取得するように調査票を作成し，さらに実験的な研究を積み重ねることで，「強く無視できる割り当て」条件を近似的に満たす共変量のセットを利用した，再現性のある調整が可能になる事例が報告されている（星野, 2007）．

一方，訪問調査において「調査不能標本を含めた確率抽出標本に対して調整を行う」場合では，通常は抽出台帳に記載されている情報しか共変量として利用できず，たとえば住民基本台帳であれば，「居住地域」「性別」「年齢」「同居家族の人数」程度しか得ることはできない．また実験研究によって「調査不能標本から共変量と目的となる回答どちらも得る」ことが難しいため，インターネット調査での「強く無視できる割り当て」に対応する「観測値による選択」条件を近似的に成立させる共変量を探索することができず，結果として(2)の理由から十分な調整が期待できない可能性が高い．

実際，インターネット調査に対する調整については，傾向スコアを推定するために利用した共変量が属性変数だけの場合には調整がうまく作用しないことも報告されており（星野・前田, 2006），事後層化やレイキングと同様の問題が生じる可能性がある．

一方，キャリブレーション推定量を利用すれば，理論上は「回収標本での共変量の値」と「共変量の母集団平均」から共変量調整を行うことが可能である(Särndal and Lundström, 2005)が，調査不能標本での共変量の値が分からないと，データから「観測値による選択」(式2.3)の条件を満たす共変量を選択することはできない．

本節で紹介した傾向スコアやキャリブレーション推定量などの方法は「観測値による選択」が起こっている状況で一致推定量を与える方法であるが，実際には「観測されないものによる選択」が起こっている可能性は高い．このような場合に利用されるパラメトリックモデルとして有名なものに，プロビット選択モデルがある．

3.2　プロビット選択モデルとその問題点

プロビット選択モデルは，「就業していないと観測されない」賃金が何によって規定されているのかを，就業有無を説明する要因を考慮に入れて考えたい，といった労働経済学での問題関心から提案されたモデルであるが，現在では経済学全般を含め様々な分野でよく利用されるようになっている．

具体的なモデリングは以下の通りである．y_i を調査対象者 i における y の値，$\boldsymbol{x}_{yi}$ を y を説明する独立変数ベクトルの値とする．ここでの関心の対象は，独立変数と従属変数の線形回帰モデル

$$y_i = \boldsymbol{x}_{yi}^t \boldsymbol{\beta}_y + \epsilon_{yi} \tag{3.3}$$

の偏回帰係数ベクトル $\boldsymbol{\beta}_y$ にあるとする．但し，結果変数がすべての調査対象者について観測されるわけではなく，ある特定の調査対象者でのみ観測されると考える．結果変数が観測されるかどうかは，ある共変量 $\boldsymbol{x}_y$ の値に依存すると考える．ここで δ_i を調査対象者 i の潜在的な状態変数と考え，共変量の値 $\boldsymbol{x}_{\delta i}$ がこの潜在変数に影響を与える回帰モデルを考える．つまり，

$$\delta_i = \boldsymbol{x}_{\delta i}^t \boldsymbol{\beta}_\delta + \epsilon_{\delta i} \tag{3.4}$$

とし，$\delta_i > 0$ なら y_i が観測される，と考える．ここで，式(3.4)は「各対象者の結果変数が得られるかどうか」＝「選択されるかどうか」を決定する式であることから，「選択方程式」と呼ぶことがある．

ここで ϵ_y と ϵ_δ に2変量正規分布を仮定し，母数推定時の識別性の問題から，ϵ_δ の分散を1とする．つまり

$$\begin{pmatrix} \epsilon_y \\ \epsilon_\delta \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_y^2 & \rho\sigma_y \\ \rho\sigma_y & 1 \end{pmatrix} \right) \tag{3.5}$$

と考える．経済学の応用研究では，$\boldsymbol{\beta}_y$ などの母数推定には Heckman の提案した二段階推定(Heckman, 1979)がよく利用されてきたが，2変量正規分布の条件付き分布の性質を利用すれば，尤度関数 L は

$$\begin{aligned} L &= \prod_{i:\delta_i \le 0} \Pr(\delta_i \le 0) \quad \times \prod_{i:\delta_i > 0} [\Pr(\delta_i > 0 | y_i) \Pr(y_i)] \\ &= \prod_{i:\delta_i \le 0} [1 - \Phi(\boldsymbol{x}_{\delta i}^t \boldsymbol{\beta}_\delta)] \quad \times \prod_{i:\delta_i > 0} \left[\Phi\left(\frac{1}{\sqrt{1-\rho^2}} \left\{ \boldsymbol{x}_{\delta i}^t \boldsymbol{\beta}_\delta + \frac{\rho}{\sigma_y}(y_i - \boldsymbol{x}_{yi}^t \boldsymbol{\beta}_y) \right\} \right) \right. \\ &\qquad \left. \times \frac{1}{\sigma_y} \phi\left(\frac{y_i - \boldsymbol{x}_{yi}^t \boldsymbol{\beta}_y}{\sigma_y} \right) \right] \end{aligned} \tag{3.6}$$

となり，これを最大化する最尤推定量は容易に得ることができる．$\phi(\cdot)$ と $\Phi(\cdot)$ はそれぞれ標準正規分布の確率密度関数と累積分布関数である．

ここで，欠測インディケータ z を y が観測される場合を $z=1$，観測されない場合を $z=0$ となる変数とすると，2 変量正規分布の性質から

$$p(z=0|y,\boldsymbol{x}_y,\boldsymbol{x}_\delta)=p(\delta\leq 0|y,\boldsymbol{x}_y,\boldsymbol{x}_\delta)=1-\Phi\left(\frac{1}{\sqrt{1-\rho^2}}\left\{\boldsymbol{x}_\delta^t\boldsymbol{\beta}_\delta+\frac{\rho}{\sigma_y}(y-\boldsymbol{x}_y^t\boldsymbol{\beta}_y)\right\}\right) \tag{3.7}$$

となる．ϵ_y と ϵ_δ に相関がある場合 $(\rho\neq 0)$ には，「$\delta\leq 0$ となる確率」は $(m=0$ において欠測している$)$ y に依存するため，y の欠測は“ランダムでない欠測”になる．

さて，式 (3.3) や式 (3.4) は「共変量と結果変数」，および「共変量と観測されやすさを示す潜在変数」についての線形回帰モデルである．これらの式において非線形回帰やそれ以外の様々な回帰関数を仮定することは理論的には可能である．しかし，

1. 回帰関数が正しく指定出来ない場合には推定値に大きなバイアスが生じ得る．
2. 誤差の分布が 2 変量正規分布であることを仮定しているが，誤差の分布仮定への頑健性が無い．分布仮定のチェックもできない．

といった問題点がしばしば指摘されている (星野, 2009)．

4. 選択バイアスに対するセミパラメトリックベイズモデル

本研究では，共変量によって調査不能標本に含まれるかどうかが決定される「観測値による選択」条件ではなく，「調査不能標本になるかどうかは，潜在的な回答値そのものにも依存する」と仮定する．具体的には，y が観測される確率を

$$\Pr(z=1|\boldsymbol{y},\boldsymbol{x})=e(\boldsymbol{y},\boldsymbol{x},\boldsymbol{\alpha}) \tag{4.1}$$

とすると，尤度は，

$$\prod_{i=1}^{N}\left\{p(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{\theta})p(z_i=1|\boldsymbol{y}_i,\boldsymbol{x}_i,\boldsymbol{\alpha})\right\}^{z_i}\left\{\int p(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{\theta})p(z_i=0|\boldsymbol{y}_i,\boldsymbol{x}_i,\boldsymbol{\alpha})d\boldsymbol{y}_i\right\}^{1-z_i} \tag{4.2}$$

と表すことができる．但し $\boldsymbol{\alpha}$ や $\boldsymbol{\theta}$ はパラメータである．

上記のモデルにおいて，$p(\boldsymbol{y}|\boldsymbol{x})$ が正規分布であり，また $p(z|\boldsymbol{y},\boldsymbol{x})$ がプロビット回帰分析モデルの場合がプロビット選択モデルである．但しプロビット選択モデルでは y は単変量であったが，本研究では研究目的上，複数の項目を同時に考える必要がある．

このモデルにおいて最も関心があるのは $\boldsymbol{y}$ の周辺分布の母数である．例えば $\boldsymbol{y}$ が一次元であり J 個のカテゴリーをもつ変数とすると，関心があるのは $y=j$ の周辺比率

$$p(y=j)=\int p(y=j|\boldsymbol{x},\boldsymbol{\theta})p(\boldsymbol{x})d\boldsymbol{x} \tag{4.3}$$

であり，これを推定するためには $\boldsymbol{\theta}$ を正しく推定できればよい．

さて，すでに指摘したように，プロビット選択モデルにおいてはモデルの設定が誤っていれば $\boldsymbol{y}$ の母集団平均の推定量にバイアスが生じる．

そこで Lee and Berger (2001) はディリクレ過程混合モデルを利用して，$p(z|\boldsymbol{y},\boldsymbol{x})$ に分布仮定を置かずに推定する方法を提案している．

任意の分布は正規分布など同一の分布の母数を変化させたものの混合分布を用いて表現できることが知られている (Sethuraman, 1994)．そこで $\boldsymbol{x}$ や $\boldsymbol{y}$ の関数となっている二項比率や多項比率について無限の要素数の混合分布を仮定することで，あらゆる形の回帰関数を表現できる．これがディリクレ過程混合モデルである．

但し，Lee and Berger (2001)では単変数の欠測を考えており，調査不能のような複数項目すべてが欠測する場合を想定してはいない．また，調査データにおいては回答は通常カテゴリカル変数であることから，単純に Lee and Berger (2001)を拡張しても識別性が保証されない．

そこで本研究ではまず $\boldsymbol{y}$ や z の背後に共通した潜在変数 f を仮定する．この潜在変数は回答傾向にも「回収か調査不能か」の群別にも影響を与える“隠れた共変量(Hidden covariate)”(星野, 2009)として考えることができ，Imbens (2003)による感度分析のためのモデルを多変量の回答に拡張したものとして考えることができる．

また，この f と共変量 $\boldsymbol{x}$ の相関に関心があるわけではないので，ここでは f と $\boldsymbol{x}$ は独立であると仮定する．

本研究で提案するモデルを，まずプロビット選択モデルでの式(3.3)において，y が多変量でありかつカテゴリカル変数を含む場合に拡張したものとして表現する．$\boldsymbol{y}$ がカテゴリカル変数であり，(二値，順序及び名義)プロビットモデルに従うとし，カテゴリカル変数の背後に存在する潜在的効用ベクトル $\boldsymbol{u}$ が

$$\boldsymbol{u} = \boldsymbol{B}^t \boldsymbol{x} + \boldsymbol{\lambda} f + \boldsymbol{\epsilon} \tag{4.4}$$

に従うとする．ベクトル $\boldsymbol{y}$ の中の特定の変数が二値の場合は，対応する $\boldsymbol{u}$ の要素がゼロより大きければ 1，ゼロ以下ならば 0 とする．$\boldsymbol{y}$ の要素が順序尺度水準の場合や名義尺度水準の変数の場合は順序プロビットモデルや名義プロビットモデルと同様とする．また，$\boldsymbol{\epsilon}$ は各要素が独立ではなく，相関を持つ確率変数ベクトルであるとする．

さらに，プロビット選択モデルの問題点として誤差変数の正規性の仮定や共変量についての回帰関数の仮定が指摘されることから，式(4.4)の代わりに

$$\boldsymbol{u} = \boldsymbol{\lambda} f + \boldsymbol{\epsilon}^* \tag{4.5}$$

とし，$\boldsymbol{\epsilon}^*$ の分布がディリクレ過程混合分布に従うとする．具体的には $\boldsymbol{\epsilon}^*$ の分布として

$$\boldsymbol{\epsilon}^* \sim \sum_{k=1}^{\infty} \pi_k N(\boldsymbol{B}_k \boldsymbol{x}, \boldsymbol{\Sigma}_k) \tag{4.6}$$

とする．但し $\boldsymbol{B}_k, \boldsymbol{\Sigma}_k$ は k 番目の要素に対する母数 $\boldsymbol{B}, \boldsymbol{\Sigma}$ の値である．また，回収か調査不能を表すインディケータ z については，プロビット選択モデル同様に，背後に潜在変数 δ が存在し，$\delta > 0$ ならば $z = 1$，$\delta \leq 0$ ならば $z = 0$ とする．また式(3.4)の代わりに

$$\delta = \gamma f + \eta \tag{4.7}$$

とし，η の分布として

$$\eta \sim \sum_{k=1}^{\infty} \pi_k N(\boldsymbol{\alpha}_k^t \boldsymbol{x}, \phi_k^2) \tag{4.8}$$

を仮定する．但し $\boldsymbol{\alpha}_k, \phi_k^2$ は k 番目の要素に対する母数 $\boldsymbol{\alpha}, \phi^2$ の値である．ここで γ が未知の場合にはモデルの識別性がないので，これを特定の値に固定して推定を行う．またこの値を変化させることで感度分析を行うことができる．

結果として，尤度は式(4.2)の代わりに

$$\begin{aligned} &\prod_{z_i=1}^{N} \int p(y_i | f_i, \boldsymbol{x}_i, \boldsymbol{\theta}) p(z_i = 1 | f_i, \boldsymbol{x}_i, \boldsymbol{\alpha}) p(f_i) df_i \\ &\quad \times \prod_{z_i=0}^{N} \iint p(y_i | f_i, \boldsymbol{x}_i, \boldsymbol{\theta}) p(z_i = 0 | f_i, \boldsymbol{x}_i, \boldsymbol{\alpha}) p(f_i) df_i dy_i \end{aligned} \tag{4.9}$$

$$= \prod_{z_i=1}^{N} \int p(y_i|f_i,\boldsymbol{x}_i,\boldsymbol{\theta})p(z_i=1|f_i,\boldsymbol{x}_i,\boldsymbol{\alpha})p(f_i)df_i$$
$$\times \prod_{z_i=0}^{N} \int p(z_i=0|f_i,\boldsymbol{x}_i,\boldsymbol{\alpha})p(f_i)df_i$$

と表現できる.

5. Blocked Gibbs sampler による推定

5.1 有限ディリクレ過程混合モデルの階層モデルとしての表現

Ishwaran and Zarepour (2000) は無限次元ではなく，十分に大きな要素数(通常は 10 から 20 程度)をもつ有限要素数の混合分布によって，任意の分布への十分精度の高い近似を行うことが可能であることを示している．具体的には，Y の分布が L 次元の有限ディリクレ過程事前分布 $DP_L(a,G_0)$

$$\boldsymbol{y} \sim \sum_{l=1}^{L} p_l f(\cdot|\boldsymbol{\theta}_l) \tag{5.1}$$

に従う場合を考え，L が大のときには有限ディリクレ過程混合モデルがディリクレ過程混合モデルを十分な精度で近似することを示した．さらに，Ishwaran and James (2001) はこの有限ディリクレ過程混合モデルでの母数の事後分布を求めるためのアルゴリズムとして Blocked Gibbs sampler を提案している．ディリクレ過程混合モデルでの母数の事後分布導出のためのアルゴリズムに比べて解析的な計算要素が少なくてすむことから，本研究では有限ディリクレ過程混合モデルを仮定した際の Blocked Gibbs sampler を利用した推定を行う.

ここでは Miyazaki and Hoshino (2009) と同様に，潜在変数が存在する場合の Blocked Gibbs sampler を考える.

階層ベイズモデルとしてこのモデルを表現すると

$$\begin{gathered}
\boldsymbol{y}_i = g(\boldsymbol{u}_i) \quad z=1(\delta>0) \\
\boldsymbol{u}_i|f_i,\boldsymbol{x}_i,\boldsymbol{\theta},\boldsymbol{k} \sim p(\boldsymbol{u}_i|f_i,\boldsymbol{x}_i,\boldsymbol{\theta}_{k_i}),\ \ \delta_i|f_i,\boldsymbol{x}_i,\boldsymbol{\theta},\boldsymbol{k} \sim p(\delta_i|f_i,\boldsymbol{x}_i,\boldsymbol{\theta}_{k_i}) \quad (i=1,\ldots,N), \\
k_i|\boldsymbol{\kappa} \sim \sum_{l=1}^{L} \kappa_l 1_l(\cdot) \\
\boldsymbol{\kappa} \sim p(\boldsymbol{\kappa}),\quad \boldsymbol{\theta} \sim p(\boldsymbol{\theta}|\boldsymbol{\tau}),\quad \boldsymbol{\tau} \sim p(\boldsymbol{\tau})
\end{gathered} \tag{5.2}$$

と表現できる．但し関数 g は観測値 $\boldsymbol{y}$ と潜在変数 $\boldsymbol{u}$ を対応づける非確率関数である．またサンプルサイズを N，対象者 i の混合要素への所属を表すインディケータを k_i (例えば対象者 i が第 l 要素に所属するなら $k_i=l$) とし，$1_l(\cdot)$ は $k_i=l$ なら 1，それ以外なら 0 となるインディケータを示す．母数 $\boldsymbol{\theta}_{k_i}$ は対象者 i の所属する要素 k_i に対応する母数 $\boldsymbol{\theta}$ であり，具体的には $\boldsymbol{\theta}$ には $\boldsymbol{B}$，$\boldsymbol{\Sigma}$ および $\boldsymbol{\alpha}$ が含まれる.

ここで $\boldsymbol{\kappa}=(\kappa_1,\ldots,\kappa_L)$ の事前分布は Stick-breaking 表現と呼ばれる以下の式に従うとする.

$$\begin{gathered}
\kappa_l = V_l \prod_{m=1}^{l-1} (1-V_m) \\
V_l \sim \mathrm{Beta}(a_l,b_l)
\end{gathered} \tag{5.3}$$

但し $b_l=\sum_{m=l+1}^{L} a_m$ であり，$a_l=\nu/L$ とすると ν は「大きくなるほど，多くの要素数に対象者が所属しやすくなる」，そして「大きいほど複雑なモデルを表現できる」ことを示すハイパー

パラメータであり，平滑化に関連する母数と言うことができる．また，κ_l が従う分布は一般化ディリクレ分布になる．

Blocked Gibbs sampler では通常の混合分布モデルのためのマルコフ連鎖モンテカルロ法と基本的には同じように条件付き事後分布から母数を乱数発生させればよい．

5.2　事前分布の設定と Blocked Gibbs sampler

まず事前分布を設定する．本研究では

$$
\begin{aligned}
&f \sim N(0,1), \quad \boldsymbol{\lambda} \sim N(\boldsymbol{\mu}_{\boldsymbol{\lambda}}, \sigma_{\boldsymbol{\lambda}}^2 \boldsymbol{I}), \quad \mathrm{vec}(\boldsymbol{B}) \sim N(\boldsymbol{\mu}_{\boldsymbol{B}}, \sigma_{\boldsymbol{B}}^2 \boldsymbol{I}) \\
&\boldsymbol{\alpha} \sim N(\boldsymbol{\mu}_{\boldsymbol{\alpha}}, \sigma_{\boldsymbol{\alpha}}^2 \boldsymbol{I}), \quad \phi^2 \sim \chi^{-2}(n_{\phi^2}, c), \quad \boldsymbol{\Sigma} \sim W^{-1}(n_{\boldsymbol{\Sigma}}, \boldsymbol{D})
\end{aligned} \tag{5.4}
$$

と設定する．但し，χ^{-2} は逆カイ二乗分布を，W^{-1} は逆ウィッシャート分布を表す．また，$\mathrm{vec}(A)$ は行列 A を縦につないだベクトルを表す．

本研究で提案されたモデルに対する Blocked Gibbs sampler は以下のような乱数発生を繰り返し行うことで実行できる．

1. $\boldsymbol{u}$ の発生：$\boldsymbol{u}$ については対応する $\boldsymbol{y}$ の要素の尺度水準によってサンプリング方法が異なるが，他の母数がすべて得られている場合には Albert and Chib (1993) と同じである．
2. $\boldsymbol{\lambda}, \boldsymbol{\Sigma}, \boldsymbol{B}, \boldsymbol{\alpha}, \phi$ の発生：マルコフ連鎖モンテカルロ法での各 iteration で各対象者を事前に設定した最大の要素数＝ L 個分の要素のどれかに所属させる．そしてすべての要素 (L 個) 分の $\boldsymbol{\theta}$ について，毎回必ず乱数を事後分布から発生させるが，対象者が一つも所属しない要素の $\boldsymbol{\theta}$ は事後分布ではなく事前分布から乱数を発生させればよい．

ここで，ある iteration で対象者が一つ以上所属する要素数が m であるとすると，$\boldsymbol{k}$ のユニークな値は m 種類になるが，それを $\{k_1^*, \ldots, k_m^*\}$ とする．このとき，誰も所属していない要素に対応する $L-m$ 個分の $\boldsymbol{\theta}$ は事前分布 $p(\boldsymbol{\theta}|\boldsymbol{\tau})$ から，そして対象者が一つ以上所属する要素に対応する m 個分の $\boldsymbol{\theta}$ は

$$
\begin{aligned}
p(\boldsymbol{\theta}_{k_j^*}|\boldsymbol{k}, \boldsymbol{u}, \delta, f, \boldsymbol{x}) &\propto p(\boldsymbol{\theta}_{k_j^*}|\boldsymbol{\tau}) \\
&\times \prod_{\{i:k_i=k_j^*\}} p(\boldsymbol{u}_i|k_i, f_i, \boldsymbol{x}_i, \boldsymbol{\theta}_{k_i}) p(\delta_i|k_i, f_i, \boldsymbol{x}_i, \boldsymbol{\theta}_{k_i}) \quad (j=1,\ldots,m)
\end{aligned} \tag{5.5}
$$

から発生させればよい．ここで $\boldsymbol{\lambda}$ の完全条件付き事後分布は式 (4.5) が f について混合因子分析モデルであることから Hoshino (2001) の結果をそのまま利用することができ，

$$
\begin{aligned}
\boldsymbol{\lambda}_{k_j^*}|\cdots \sim N\Bigg(&\left(\sigma_{\boldsymbol{\lambda}}^{-2}\boldsymbol{I} + \sum_{i:k_i=k_j^*}^{N} f_i^2 \boldsymbol{\Sigma}_{k_j^*}^{-1}\right)^{-1} \left[\sigma_{\boldsymbol{\lambda}}^{-2}\boldsymbol{I}\boldsymbol{\mu}_{\boldsymbol{\lambda}} + \sum_{i:k_i=k_j^*}^{N} f_i^2 \boldsymbol{\Sigma}_{k_j^*}^{-1}(\boldsymbol{u}_i - \boldsymbol{B}_{k_j^*}\boldsymbol{x}_i)\right], \\
&\left(\sigma_{\boldsymbol{\lambda}}^{-2}\boldsymbol{I} + \sum_{i:k_i=k_j^*}^{N} f_i^2 \boldsymbol{\Sigma}_{k_j^*}^{-1}\right)^{-1}\Bigg)
\end{aligned} \tag{5.6}
$$

となる．但し $|\cdots$ はデータと他の母数を所与としたことを表す．また $\boldsymbol{\Sigma}$ と $\boldsymbol{B}$ についても，各要素への所属が決定した後では $\boldsymbol{u}$ は多変量回帰分析モデルに従うため，完全条件付き事後分布はそれぞれ逆ウィッシャート分布と多変量正規分布になる (Gelman et al., 2003)．α と ϕ^2 についても，δ が重回帰分析モデルに従うため，完全条件付き事後分布はそれぞれ多変量正規分布と逆カイ二乗分布に従う．

3. f の発生：f_i を発生させるときには対象者 i がどの要素に所属しているかが事前に分かる

ために，その情報を利用することで

$$p(f_i|k_i,\boldsymbol{u}_i,\delta_i,\boldsymbol{x}_i,\boldsymbol{\theta}_{k_i}) \propto p(\boldsymbol{u}_i|k_i,f_i,\boldsymbol{x}_i,\boldsymbol{\theta}_{k_i}) \times p(\delta_i|k_i,f_i,\boldsymbol{x}_i,\boldsymbol{\theta}_{k_i}) \times p(f_i) \quad (i=1,\ldots,N) \tag{5.7}$$

から発生させる．$p(\boldsymbol{u}_i|k_i,f_i,\boldsymbol{x}_i,\boldsymbol{\theta}_{k_i})$ と $p(\delta_i|k_i,f_i,\boldsymbol{x}_i,\boldsymbol{\theta}_{k_i})$ は多変量正規分布であるため，f_i の完全条件付き事後分布も

$$f_i|\cdots \sim N\big(\big(\boldsymbol{\lambda}_{k_i}^t\boldsymbol{\Sigma}_{k_i}^{-1}\boldsymbol{\lambda}_{k_i}^t+\gamma_{k_i}^2\phi_{k_i}^{-2}+1\big)^{-1}\big[\boldsymbol{\lambda}_{k_i}^t\boldsymbol{\Sigma}_{k_i}^{-1}(\boldsymbol{u}_i-\boldsymbol{B}_{k_i}\boldsymbol{x}_i)+\gamma_{k_i}\phi_{k_i}^{-2}(\delta_i-\boldsymbol{\alpha}_{k_i}^t\boldsymbol{x}_i)\big], \big(\boldsymbol{\lambda}_{k_i}^t\boldsymbol{\Sigma}_{k_i}^{-1}\boldsymbol{\lambda}_{k_i}^t+\gamma_{k_i}^2\phi_{k_i}^{-2}+1\big)^{-1}\big) \tag{5.8}$$

と正規分布に従う．

4. $\boldsymbol{k}$ の発生：要素への所属のインディケータ $\boldsymbol{k}$ は $\sum_{k=1}^{L}p_{ki}\delta_k(\cdot)\quad(i=1,\ldots,N)$ の確率で発生させる．但し p_{ki} は

$$p_{ki}=\frac{\kappa_k p(\boldsymbol{u}_i|k_i,f_i,\boldsymbol{x}_i,\boldsymbol{\theta}_{k_i})p(\delta_i|k_i,f_i,\boldsymbol{x}_i,\boldsymbol{\theta}_{k_i})}{\sum_{k=1}^{L}\kappa_k p(\boldsymbol{u}_i|k_i,f_i,\boldsymbol{x}_i,\boldsymbol{\theta}_{k_i})p(\delta_i|k_i,f_i,\boldsymbol{x}_i,\boldsymbol{\theta}_{k_i})} \tag{5.9}$$

となる．

5. $\boldsymbol{\kappa}$ の発生：κ_k は一般化ディリクレ分布

$$\kappa_k=V_k\prod_{m=1}^{k-1}(1-V_m)$$
$$V_k\sim \mathrm{Beta}\left(a_k+M_k,b_k+\sum_{m=k+1}^{L}M_m\right)$$

から発生させる．但し M_k は第 k 要素への所属対象者数である．

5.3 母集団平均の推定と感度分析

Blocked Gibbs sampler によって得られた事後分布からの母数の乱数列を利用して，母集団平均の推定が可能である．具体的には母集団平均 $E(\boldsymbol{y})$ は

$$E(\boldsymbol{y})=\int\sum_{l=1}^{L}p_l p(y|k=l,f,\boldsymbol{x},\boldsymbol{\theta}_l)p(f)p(\boldsymbol{x})df\,d\boldsymbol{x} \tag{5.10}$$

と表現できる．従って推定値は

$$\hat{E}(\boldsymbol{y})=\frac{1}{M}\sum_{m=1}^{M}\frac{1}{N}\sum_{i=1}^{N}\prod_{l=1}^{L}\big(p(y_i|k_i^m=l,f_i^m,\boldsymbol{x}_i,\boldsymbol{\theta}_l^m)\big)^{1(k_i^m=l)} \tag{5.11}$$

となる．但し $1(k_i^m=l)$ は m 回目の iteration で k_i が l になれば 1，それ以外では 0 となるインディケータである．

すでに述べたように，本研究のモデルでは潜在的な回答傾向 f が「調査不能を規定する変数」δ へ与える影響力 γ を様々な値に変化させて推定を行うことにより，感度分析を実行することができる．また平滑化パラメータである ν を大きくすることでより複雑なモデルが設定されることになることから，γ と ν を同時に変化させることによる推定値と信頼区間の幅の変化を調べることによって「調査不能標本を考慮に入れた場合の推定値がどれくらいの変動を持ち得るのか」を調べることができる．

6. 日本人の国民性調査データに対する適用

ここでは 2008 年に実施された第 12 次日本人の国民性調査において，K 型調査票に回答した対象者，及び割り当てられていたが調査不能だった対象者のデータを利用した．サンプルサイズは回収標本 1729 人，調査不能標本 1482 人(ただし一部欠損がある 2 名を除外した)の計 3211 人であり，計画標本に対する回収率は 53.8%であった．

本研究では「男女の能力差」「しきたりに従うか」「宗教を信じるか」「人は信頼できるか」「環境の保護は重要か」など様々な内容を有する 29 の項目について「D.K. (わからない)」や「その他」を除いた回答を利用し，また図示や比較のわかりやすさのために三値以上の回答値のある項目の比率の推定値はカテゴリー数分の二値変数の比率として表示した(但し，推定値の標準誤差についてはモデルに従って推定を行っている)．

すでに 3 節で述べたように，調査不能の影響を除去するための共変量調整を行うためには，無作為抽出標本(または調査不能を含めた標本)において「無作為抽出標本と有意抽出標本の差異を決める共変量情報が十分得られている」ことが必要であるが，実際に未回収標本から得られるのは性年齢や地域情報，未回収の理由程度である．実際，第 12 次日本人の国民性調査において調査不能標本から得られた，回収標本と共通の情報としては「地域」「市群別」「性」「年齢」「住居形態」の 5 つのみである．

そこでまず，これらの 5 つの変数を用いて「調査不能かどうか」を予測する単純なロジスティック回帰分析を実施した．またここでは年齢は 11 階級の年代別に変換して利用した(詳しくは統計数理研究所 国民性調査委員会, 2009, 参照)．ロジスティック回帰モデルの推定値を用いて計算した予測確率を利用して判別を行った場合の正判別率は 63.8%，Cox & Snell の疑似決定係数は 0.095 であり，そもそもこれら 5 つの変数だけでは調査不能を予測しているとは言い難い．また，回収標本における「回答値と 5 つの変数の関係」は総じて強くなく，これらを共変量として調整を行ってもあまり良い効果は期待できないことが予想できる．但し，ここで利用する日本人の国民性調査データにおいては，真値であるところの「回収率が 100%であった場合の回答平均」はわからない．そこで，本研究では，

推定方法 1：回収標本から得られた単純比率

推定方法 2：5 つの変数を利用して計算した傾向スコアによる IPW 推定量

推定方法 3：「調査不能にも回答値にも影響を与える潜在変数」を含むモデルを利用した推定値．但しパラメトリックモデルの場合(ディリクレ過程混合モデルで $L=1$ の場合)．

推定方法 4：「調査不能にも回答値にも影響を与える隠れた共変量」を含む，ディリクレ過程混合モデルによる推定値($L=20$)

の 4 つの方法で母集団平均(比率)を推定し，これらを比較することで，単純比率や IPW 推定量がどの程度「調査不能にも回答値にも影響を与える隠れた共変量」が存在する場合の推定値と異なった値を取るのかを調べた．推定方法 2 については，「5 つの変数を用いて回収標本か調査不能標本か」を予測するロジスティック回帰分析を実施し，回収標本に割り当てられる予測確率を傾向スコアとした．また，“推定方法 3”については γ を 0.2, 0.4, 0.7 の 3 通りに変化させて解析し，“推定方法 4”については γ を 0.2, 0.4, 0.7，ν を 1, 3, 10 のそれぞれ 3 通りに変化させて解析した．ここで γ の二乗は各要素における「潜在変数 f による調査不能への割り当ての分散説明率」を表していると考えてよく，$\gamma=0.2,\ 0.4,\ 0.7$ はそれぞれ分散説明率が 4%, 16%, 49% に近似的に対応し，隠れた共変量が分散説明率を 50% 以上有することはあまり想定しえないことであることから，$\gamma=0.7$ での推定値は隠れた共変量の影響を最大限考慮した場合に対応すると考えてよい．一方，ν は事前に仮定される要素数を規定するハイパーパラ

メータとして考えることができ，クロスバリデーションなどからこの値を決定することも可能であるが，本研究では 3 つの値を用意しそれぞれの場合で推定を行った．

ここで，推定方法 3 及び 4 での推定にはマルコフ連鎖モンテカルロ法を利用しており，iteration を 35,000 回実施し，最初の 5,000 回を Burn-in Phase として除外し，30,000 回分の乱数から事後分布を得た．また ν が 1 の場合には事後の混合要素数のメディアンは 6，3 の場合は 8，10 の場合は 9 となり，ν を 10 以上に大きくしても混合要素数は変化しないと考えられるため，$\nu=10$ の場合を「もっとも複雑なモデル」とすることに問題はないと考えられる．

図 2 は縦軸に推定方法 2 で得られた推定値を，図 3 は推定方法 3 で得られた推定値を，図 4 は推定方法 4 のうち $\nu=1$ の場合の推定値を，図 5 は推定方法 4 のうち $\nu=3$ の場合の推定値を，図 6 は推定方法 4 のうち $\nu=10$ の場合の推定値を，それぞれプロットしたものである．

推定値の分散はほとんどの場合推定方法 1 が最小であり，推定方法 4 が最大になっている．そこで，各比率について，図 7 に横軸に推定方法 1 の 95% 信頼区間の幅を，縦軸に推定方法 4 の γ が 0.7，ν が 10 の場合の 95% 信頼区間の幅をプロットした．

図 7 から，「調査不能を考慮した信頼区間」は単純比率による信頼区間より幅を大きく想定する必要がある可能性があることが示唆される．但し，図 2 から図 6 で見たように，推定値自

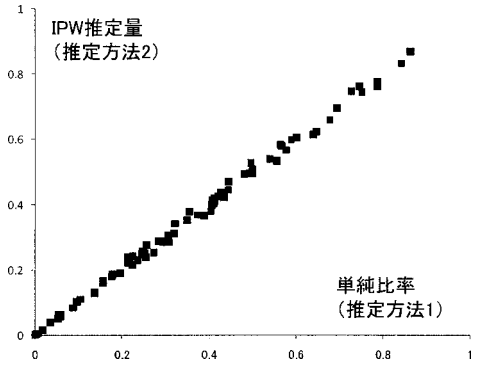

図 2. 単純比率と IPW 推定量の比較.

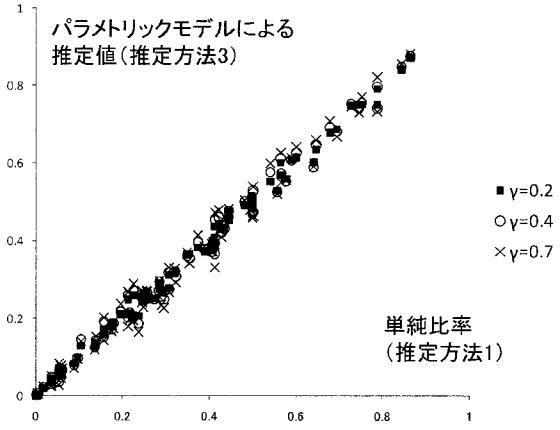

図 3. 単純比率とパラメトリックモデルによる推定量の比較.

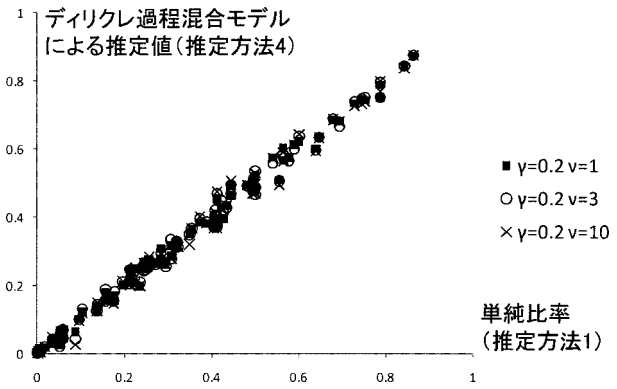

図 4. 単純比率とディリクレ過程混合モデル ($\gamma = 0.2$) による推定量の比較.

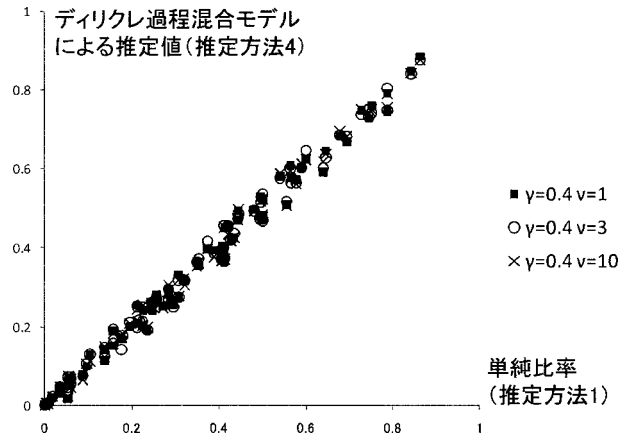

図 5. 単純比率とディリクレ過程混合モデル ($\gamma = 0.4$) による推定量の比較.

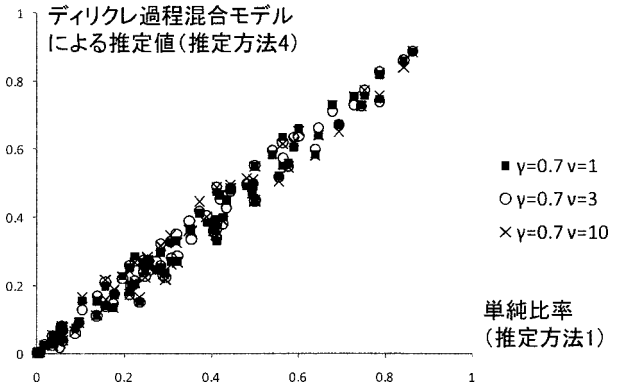

図 6. 単純比率とディリクレ過程混合モデル ($\gamma = 0.7$) による推定量の比較.

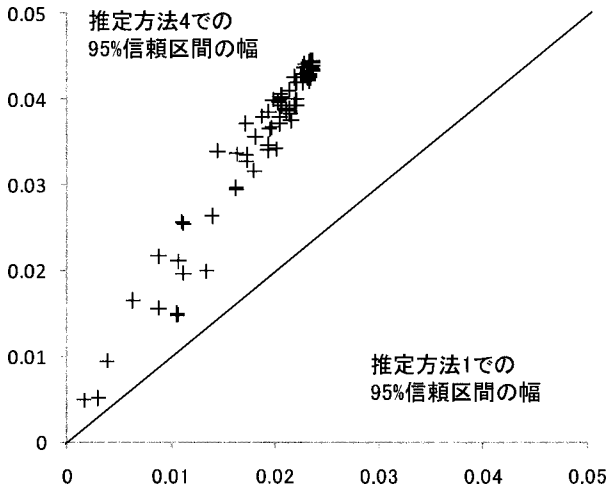

図 7. 単純比率とディリクレ過程混合モデルによる推定値の信頼区間の幅の比較.

体も単純比率による推定値と，もっとも「調査不能が隠れた共変量によって説明される割合が高く」「複雑なモデル」を想定した場合(推定方法 4 の $\gamma=0.7$，$\nu=10$ の場合)での推定値とは大きく異なる.

そこで，各母数(比率)について，推定方法 1 での推定値を $\hat{p}^1$，「推定方法 4 の $\gamma=0.7$，$\nu=10$ の場合」の推定値を $\hat{p}^4$，さらにその標準誤差を ξ とするとき，

$$\zeta=|\hat{p}^1-\hat{p}^4|+2\times 1.96\times\xi \tag{6.1}$$

を縦軸に，$\hat{p}^1$ を横軸にプロットした(図 8)．上記の量は推定方法 1 による通常の 95% 信頼区間の代わりに，「調査不能を考慮した推定値の信頼区間に 2 つの推定値の差を加味した変動幅」であり，図 8 はその量を推定方法 1 の関数としてみるとどのような変化をするかを示している．この関数関係を明確にするために，カーネル回帰分析モデルによって推定方法 1 の推定値と式(6.1)の値の関係を求め，図 8 に記入した．但しここでカーネル回帰分析のバンド幅は Silverman の “Rule-of-thumb” の方法である $1.06\times\hat{\sigma}_{\hat{p}^1}\times N^{-1/5}$ を利用している(但しここでの N は比率の数 83 である).

結果として，このデータの場合では，通常の比率の推定量の信頼区間の幅が最大になる比率 0.5 において，推定方法 1 による単純な集計値の変動の幅を最大 13% 見積もればよいということがわかる.

具体的な項目に関する推定値について紹介する．表 1 は「人は信頼できるか」「他人のためか自分のためか」「宗教を信じるか」「選挙への関心」といった質問項目について，推定方法 1，推定方法 2，推定方法 3 ($\gamma=0.4$)および推定方法 4 ($\gamma=0.7$，$\nu=10$)の場合の推定値を示した(ここでも「その他」「D.K.」については排除して結果を表示している)．「人は信頼できるか」という項目は「他人のためか自分のためか」という項目とともに米国の総合社会調査(General Social Survey; GSS, Davis, Smith & Marsden. 1972–2004)で「信頼感」尺度として利用されている項目を一部改変したものであり，一般的な信頼感を測定するために社会調査でしばしば用いられ，両者の相関は高いことが知られている．しかし，推定方法 1 (単純比率)と推定方法 2 (IPW 推定量)を比較すると，「人は信頼できるか」の項目の「信頼できる」とする比率が推定

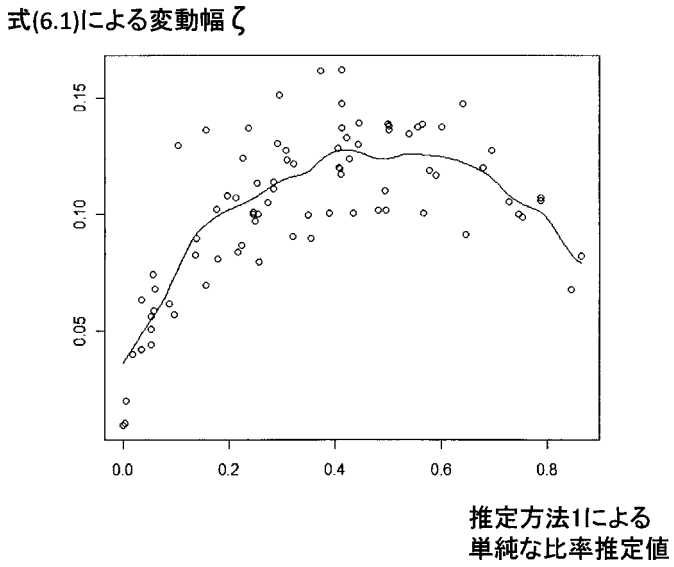

図 8.　単純比率と変動幅 ζ の関係.

方法 1 よりも推定方法 2 の方が増えているのに対して,「他人のためか自分のためか」の項目の「他人の役にたとうとしている」とする比率は逆に推定方法 2 の方が減少している．これは,すでに 3.1 節で説明したように「ごく少数の共変量を用いて調整を行う」だけでは傾向スコアによる調整を含めた共変量調整があまりうまく作用していないためと思われる.「宗教を信じるか」については「信じていない」とする率が推定方法 1 から 4 にかけて一貫して増大しているが,これは「回答していない人は宗教を信じていない」ことを示している．また「選挙への関心」についても,「回答していない人は投票しない」ということを推測していることになり,十分納得できるものであると考えられる.

また,「推定方法 1 と推定方法 4 の乖離」は一貫して「推定方法 1 と推定方法 3 の推定値の乖離」と同じ方向に拡大されている.

表 1.　「人は信頼できるか」「他人のためか自分のためか」「宗教を信じるか」「選挙への関心」についての推定結果の比較.

項目	選択肢	推定方法1	推定方法2	推定方法3	推定方法4
人は信頼できるか	1. 信頼できる	0.32182	0.34213	0.30780	0.28027
	2. 用心する	0.67818	0.65787	0.69220	0.71973
他人のためか自分のためか	1. 他人の役に	0.41060	0.40254	0.39181	0.37957
	2. 自分のことだけ	0.58940	0.59746	0.60819	0.62043
宗教を信じるか	1. 信じている	0.27241	0.25272	0.24799	0.24488
	2. 信じていない	0.72759	0.74728	0.75201	0.75512
選挙への関心	1. なにをおいても投票	0.40523	0.38109	0.37053	0.36129
	2. なるべく投票	0.48081	0.49166	0.49937	0.49374
	3. あまり投票する気にならない	0.05988	0.06362	0.06705	0.07688
	4. ほとんど投票しない	0.05407	0.06364	0.06305	0.06809

7. 結論

近年の国民のライフスタイルやワークスタイルの変化，過剰なプライバシー意識の高まりなど，調査環境の悪化が非可逆的であることを考えると，従来型調査の回収率は今後さらに低下する可能性が極めて高い．このような現状では，混合モードを用いた調査も一つの方向性として重要であるが，本研究では調査不能となった対象者のデータを予測するモデルを仮定し感度分析を行うことで「調査不能を考慮した上でどの程度の幅を見積もって推論を行えば良いか」についての方法を提案した．

以下，この方法と「調査不能となった対象者についてはどのような結果が得られるかはまったく分からない」とする「最悪のシナリオ」と比較をしてみたい．

今回の日本人の国民性調査において，K 型調査票についての回収率は 53.8% であったが，この場合，ある意見項目に対して「賛成」の比率が 50%であっても，まったくのモデル仮定を置かなければ，調査不能である 46.2% の対象者すべてが「賛成」または「反対」であるという可能性がある．このような統計的推測の面からは「最悪のシナリオ」を考えると，推定値が「賛成」の比率は $53.8 \div 2 = 26.9\%$ から $53.8 \div 2 + 46.2 = 73.1\%$ までの値をとる可能性がある．これに加えて推定値の標本変動を加味し，信頼区間を構成しようと考えれば，サンプルサイズが $N = 1729$ であっても

$$\left(0.269 - 1.96 \times \sqrt{\frac{0.269 \times (1 - 0.269)}{1729}},\quad 0.731 + 1.96 \times \sqrt{\frac{0.731 \times (1 - 0.731)}{1729}}\right) \approx (0.2481,\quad 0.7519) \tag{7.1}$$

となり信頼区間の幅が 50% を超えてしまい，折角無作為抽出による大規模な調査を実施しても，何の情報も無いに等しいことになる．

一方前節で示したように，本研究で利用したモデルからは 13% 程度の幅をもって推論を行えばよい，ということが示唆されている．

本研究で利用したモデルは，「回答値にも回収／調査不能にも影響を与える」潜在的な傾向(隠れた共変量)が存在することのみ想定した，仮定の少ないセミパラメトリックなモデルである．このようなモデルを利用することで，50% の幅をもって推論することを求める「最悪のシナリオ」に比べると十分説得的，かつ「仮定が少ない(＝その分推定量の分散が大きい)モデルを利用した推論方法を提供することには意味があると考えられる．

本研究で提示されたモデルは「隠れた共変量」を明示的に組み込んだモデルであるが，「共変量と回答」および「共変量と調査不能傾向」の回帰関係は仮定しないという意味でセミパラメトリックな推定法を利用しており，ロバストな結果を与えていると考えられる．但し本研究でこのようなモデルを利用する目的は母数推定のためというより，調査不能を考慮した上でどの程度推定値の信頼区間を見積もれば良いかという疑問に答えるためのものであり，推定値自体を利用するというより，回収標本から得られる単純推定値を解釈するために利用するべき補助的なモデルとして有用であると考えられる．

また，本研究で利用したモデルとは異なる考え方によって構成された欠測データ状況での信頼区間の構成法として，Copas and Eguchi (2005)の方法がある．彼らは欠測によってデータ発生モデルが仮定したモデルのチューブ近傍によって表現できる場合，最悪のケースであっても推定量の漸近分散を 2 倍すればよいという非常に分かりやすい法則を与えている．但しデータ発生モデルがチューブ近傍によって表されるモデルとなるかどうかなどを含め，本研究で利用した方法との関連については今後検討する必要がある．

一方，3 節においてインターネット調査の調整との関連で述べたように，共変量を注意深く選

択した場合には共変量調整を行うことで調査不能によるバイアスを減少出来る可能性があることから，本論文のような「解析手法に依存する方法」だけでなく，調査不能に関連し，かつ調査項目にも関連する要因が何であるかを今後とも研究していくという方向性（例えば土屋, 2010）も今後発展していくべきであろう．

謝　辞

匿名の査読者の方には本論文の改稿に際して参考になるコメントを頂いたことを感謝いたします．本研究は独立行政法人新エネルギー・産業技術総合開発機構（NEDO）平成 19 年度産業技術研究助成事業（研究題目「共変量情報の高度利用によるネットリサーチのバイアス除去法の開発とマーケティング製品開発への利用」：研究代表者　星野崇宏），及び独立行政法人科学技術振興機構（JST）戦略的創造研究推進事業さきがけ「知の創生と情報社会」領域（研究題目「マルチソースデータ高度利用のための統計的データ融合」：研究代表者　星野崇宏）による研究費を利用して実施しました．

参　考　文　献

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data, *Journal of the American Statistical Association*, **88**, 669–679.

Brookhart, M. A., Schneeweiss, S., Rothmanm, K. J., Glynn, R. J., Avorn, J. and Stürmer, T. (2006). Variable selection for propensity score models, *American Journal of Epidemiology*, **163**, 1149–1156.

Copas, J. B. and Eguchi, S. (2005). Local model uncertainty and incomplete data bias (with discussion), *Journal of the Royal Statistical Society, Series B*, **63**, 459–512.

Deming, W. E. and Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known, *Annals of Mathematical Statistics*, **11**, 427–444.

Deville, J. C. and Särndal, C. E. (1992). Calibration estimators in survey sampling, *Journal of the American Statistical Association*, **88**, 1013–1020.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2003). *Bayesian Data Analysis*, Chapman & Hall, New York.

Groves, R. M., Dilman, D. A., Eltinge, J. L. and Little, R. A. J. (2002). *Surveys Nonresponse*, Wiley, New York.

Heckman, J. J. (1974). Shadow prices, market wages and labor supply, *Econometrica*, **42**, 679–694.

Heckman, J. J. (1979). Sample selection bias as a specification error, *Econometrica*, **47**, 153–161.

Hoshino, T. (2001). Bayesian inference for finite mixtures in confirmatory factor analysis, *Behaviormetrika*, **28**, 37–64.

星野崇宏（2007）．インターネット調査に対する共変量調整法のマーケティング調査への適用と調整の効果の再現性の検討，行動計量学，**34**，33–48.

星野崇宏（2009）．『調査観察データの統計科学―因果推論・選択バイアス・データ融合―』，岩波書店，東京.

星野崇宏，前田忠彦（2006）．傾向スコアを用いた補正法の有意抽出による標本調査への応用と共変量の選択法の提案，統計数理，**54**，191–206.

星野崇宏，森本栄一（2007）．第 1 章　インターネット調査の偏りを補正する方法について：傾向スコアを用いた共変量調整法，『Web マーケティングの科学―リサーチとネットワーク―（井上哲浩・日本マーケティングサイエンス学会 編）』，27–59，千倉書房，東京.

Imbens, G. W. (2003). Sensitivity to exogeneity assumption in program evaluation, *American Economic Review*, **93**, 126–132.

Ireland, C. T. and Kullback, S. (1968). Contingency tables with given marginals, *Biometrika*, **55**, 179–188.

Ishwaran, H. and James, L. F. (2001). Gibbs sampling methods for stick-breaking priors, *Journal of the American Statistical Association*, **96**, 161–173.

Ishwaran, H. and Zarepour, M. (2000). Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models, *Biometrika*, **87**, 371–390.

Lee, J. and Berger, J. O. (2001). Semiparametric Bayesian analysis of selection models, *Journal of the American Statistical Association*, **96**, 1397–1409.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd ed., Wiley, New York.

Miyazaki, K. and Hoshino, T. (2009). A Bayesian semiparametric item response model with Dirichlet process priors, *Psychometrika*, **74**, 375–393.

Särndal, C-E. and Lundström, S. (2005). *Estimation in Surveys with Nonresponse*, Wiley, New York.

Schonlau, M., van Soest, A., Kapteyn, A. and Couper, M. (2009). Selection bias in web surveys and the use of propensity scores, *Sociological Methods & Research*, **37**, 291–318.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors, *Statistica Sinica*, **4**, 639–650.

Taylor, H., Bremer, J., Overmeyer, C., Siegel, J. W. and Terhanian, G. (2001). The record of internet-based opinion polls in predicting the results of 72 races in the November 2000 U.S. elections, *International Journal of Market Research*, **43**, 127–136.

統計数理研究所 国民性調査委員会(2009). 国民性の研究 第 12 次全国調査—2008 年全国調査—, 統計数理研究所研究リポート, No. 99.

土屋隆裕(2010). 調査への指向性変数を用いた調査不能バイアスの二段補正—「日本人の国民性第 12 次全国調査」への適用—, 統計数理, **58**, 25–38.

Semiparametric Estimation under Nonresponse in Survey and Sensitivity Analysis: Application to the 12th Survey of the Japanese National Character

Takahiro Hoshino

Department of Economics, Nagoya University

In recent years, the collection rate for conventional types of surveys such as visit survey with random sampling has been declining. Therefore, a solution for bias due to nonresponse needs to be developed. We formulate the bias due to nonresponse in a survey as "selection bias" in econometrics, and I point out the problems in applying covariate adjustment methods to nonresponse in surveys. In this paper, we propose a semiparametric Bayes model where a latent hidden covariate affect both the response variables and the indicator of nonresponse using Dirichlet process mixtures. By changing some portion of parameters, we can conduct a sensitivity analysis to investigate how much the confidence interval is under a high rate of nonresponse. We apply the proposed method to the 12th survey of the Japanese National Character, and found that the method provides more reasonable confidence intervals, compared to that calculated without any model assumption.

Key words: Dirichlet process mixture model, propensity score, hidden covariate, sociological surveys, covariate adjustment, selection bias.