集計マクロレベル情報とミクロレベルデータを融合した広告効果推定法の開発と応用

常勤研究者の部

代表研究者 星野崇宏
慶應義塾大学
経済学部経済学科
教授

共同研究者
宮崎慧
関西大学
商学部
准教授

中川宏道
中村学園大学
流通科学部
准教授

竹内真登
東北学院大学
経営学部
助教

猪狩良介
日本学術振興会
特別研究員

加藤諒
日本学術振興会
特別研究員

本研究では、広告効果研究においてこれまで別個に用いられ、不整合を引き起こすことさえあったマクロレベルの広告効果モデルとミクロレベルの広告効果モデルを一体として運用し新たな知見を与える新しい広告効果モデルを開発し、実務に応用する共同研究を行うことで妥当性検証を行うことにある。

近年ビッグデータの入手可能性と R や Python などフリーの解析ソフトウェア・パッケージの充実、機械学習的なアプローチの普及の中で、後述するようにいわゆるブラックボックス型の解析が非常に盛んにおこなわれている。特に集計時系列情報をもとにした広告効果の推定においては（上記のような構造を明示しない代わりに識別性の問題を解決するために恣意的な仮

定を置くローカルトレンドモデルなどの）状態空間モデルがよく利用され（例えばGoogleが発表したCausal impactモデル,Brodersenら015)、一方ミクロレベルの広告効果推定ではコンバージョン有無の予測などにランダムフォレストなどの機械学習などが利用されている。これらの解析法に比べて、消費者の認知や意思決定・企業間の広告費の読みあいや価格競争などのメカニズムを仮定した構造推定はその説得性という点でマーケティング分野においても示唆に富む考え方である。一般にブラックボックス型のモデリングの問題点は、単に解釈が難しいだけではなく、得られた結果を実際のアクションに落とし込むことが困難であるという点がある。

政府の政策決定、企業の経営判断など、実際の意思決定に資するモデリングや解析は、ブラックボックス型のものではなく、実際に経済現象や経営事象において生起しているメカニズムに基づいたものであることが望まれる。具体的には広告効果の分析から広告費の策定や期間最適配分を行うためには、投入されればされるほど弾力性(影響力)は小さくなるという摩耗効果が存在したり、広告が時間経過に伴って忘却されるいった消費者の広告認知や記憶に関する効果、広告閲覧と購買ニーズの発生が異なるといった時間のずれがあることを考慮した解析を行う必要がある。一方、マクロ的な要因も存在する。企業の多くは対売り上げの広告費の比率を一定に収めるといった予算策定をしていることが知られて、景気変動に伴う広告費変動が存在し得る。また、売り上げが過去に比べて予想より少なく、かつ潤沢な留保がある場合には積極的に広告出稿をするといった売り上げや財務指標と広告費の双方向的な関連が存在する。

これに加えて、同時に複数の競合企業が互いの広告投入量を考慮しながら出稿費を決定する場合が多く、このような観点からはゲーム理論的な状況設定を考慮する必要がある。また、価格販促と広告費が一定の予算制約内に収まるという制約関係も存在する。このように考えた際に、広告に関する投入と認知、購買などの関係はメカニズムとしては企業側の要因と消費者側の要因の２つのメカニズム要因が存在しており、それらの結果として広告投入の売り上げへの関数が規定される。

さらに言えば、実際のデータから背後のメカニズムを理解するための障壁となるものとして、データに様々なバイアスが影響を与えることである。広告効果であれば、広告出稿以外の要因が広告投入に影響を与える場合に、それらの要因をすべて説明変数に投入しないと正しい回帰係数の推定ができない。これ

は一般的に説明要因である変数を無視して解析を行う場合には、その変数が誤差に含まれることにより、誤差と説明変数に相関が生じて、正しい係数の推定ができないという内生性バイアスの問題である。具体的な例を挙げれば、TVCMなどは比較的多額の広告費を投入しないと出稿すらできず、また 3 か月から 6 か月前に枠の購入を決定するなどといった形で事前に出稿を決める必要があるため、企業も代理店も当該商品やサービスが最もニーズがある前後に広告費を集中させがちである。一方屋外広告は比較的長期にかつ比較的少額から出稿できるために、年間を通じて広告費がニーズの集中時期以外にも分散される。

この問題はブラックボックス型の広告効果推定においてしばしば生じる「負の係数問題」と密接に関連している。これは欧米でも昔から知られていた問題であり複数メディアへの広告出稿費で売り上げを説明するような回帰分析を実行すると、どれかのメディアの係数が負になることが多い、という問題であり、内生性問題による説明が可能である。

また、集計レベルの時系列のデータでは個人レベルの意思決定の結果が集計されてしまうことに伴う推定値のバイアスが生じることが知られている。

では個人レベルの広告への反応を見るデータセットのみを利用すればよいかといえば、個人レベルのデータセットは往々にして特定の対象に偏ったデータになることが多い。

このように、「理論的には広告の投入とそれへの反応のメカニズムはミクロレベルとマクロレベルで理解できる」が、「集計時系列データしか存在しない場合には集計バイアスが、説明変数が過小な場合は内生性バイアスが、ミクロデータの偏りがある場合には選択バイアスが、それぞれ働く」ため、単純な解析では真の広告効果推定は可能ではなく、これらのバイアスを無視した集計時系列データに対するブラックボックス型の解析では正しい広告効果推定はできない。

本報告書の研究はこれについて統計学的な方法論の発展によりこれを可能にする手法を開発し、いくつかの例に適用するものである。

この議論を進めるために、まず通常のマルチソースのミクロデータ間の統計的データ融合と、ミクロデータとマクロデータの統計的データ融合についてその異同を議論する。

統計的データ融合は、異なる情報源から得られるデータ（これをマルチソースデータと呼ぶ）を、一つの情報源から得られるデータ（これをシングルソースデータと呼ぶ）に統合するための統計分析手法を示す。シングルソースデー

タとマルチソースデータのイメージを図１に示す。シングルソースデータ(図1－a)では、分析に用いる変数の全てが同じ対象者から得られており、項目 A と B の関係性を直接把握することができる。項目 A は広告接触変数、項目 B を購買実績と考えれば、項目 A と B の相関係数を算出したり、回帰モデル等を利用可能である。一方でマルチソースデータは、関心のある変数が別々の対象者から分割して得られているデータである。同じ対象者から項目 A と B は同時に得られていないため、通常はこれらの関係性を把握することはできない。そこで、統計的にマルチソースデータを解析して両者の関係性を把握することを考える。

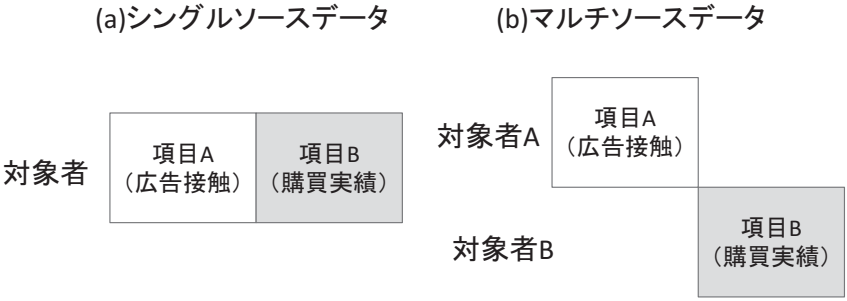

図１．シングルソースとマルチソースデータの概念図

統計的データ融合の概念を図２に示す。統計的データ融合では、マルチソースデータにおいて、変数が得られていない網掛け部分を欠測データとしてみなして合理的に埋めることを考える。そのために、統計的データ融合では、異なるデータ間に共通する「糊しろ」としての変数を用意する。この「糊しろ」は「共変量」と呼ばれ、変数群 A・B 双方に影響するデモグラフィック属性やサイコグラフィック属性等が用いられる。共変量を有効に活用することで、欠測値を統計学的に補完し、図２の状態から、変数群 A と変数群 B の関係性を把握する。データ融合の統計学的性質およびマーケティング分野での事例は、星野(2009)に示されているが、データ融合自体はニールセンなどの調査会社や、主にヨーロッパでの広告効果測定の実務で Data Integration という広義の複数データの活用法の一種として利用されてきた(例えば全米広告調査協会,2003)。

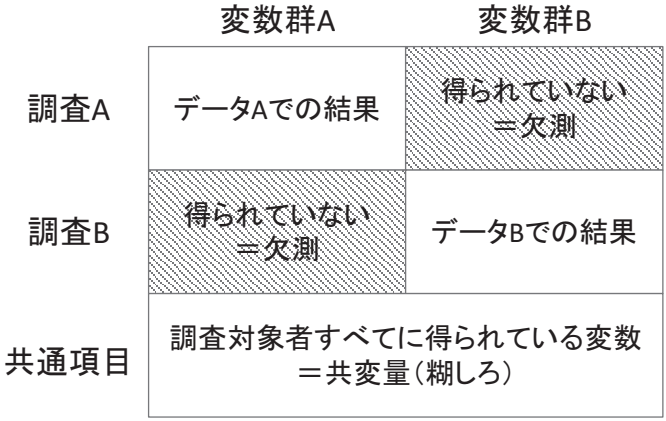

図２．統計的データ融合の概念図

さて星野(2009)によると、従来用いられてきた統計的データ融合は、(1)マッチング、(2)潜在変数モデリング、(3)回帰モデルの利用、の３つに分類される。

星野（2009）ではこれらの方法論の問題点として予測精度の低さとモデリングの柔軟性の欠如を指摘しており、セミパラメトリックを仮定した手法の利用を示唆している。

さらに近年では、一部のシングルソースデータを利用したモデルや、ミクロデータにおけるデータ融合にマクロデータを補助情報として用いる方法等が近年では提案されている。例えば、Gilura and McCulloch(2013)は一部の完全データが得られている状況下におけるマルチレベルデータの統計的データ融合を提案している。また、Feit et al (2013)もマクロデータとミクロデータに対するデータ融合を提案している。これらの方法論は一見、１章で述べたマクロレベルとミクロレベルの広告効果推定に直接適用できるように見えるが、実際には非常に限定されたモデルとデータ取得状況にのみ利用可能であり、一般の広告効果推定の枠組みでは利用できない。また、これらの先行研究で提案された手法ではマクロレベルの情報に対してミクロレベルの情報が一般には特殊な対象者・消費者から得られているという「選択バイアス」の問題を全く考慮できていない。そこで本研究では１章で議論したように、マクロレベルの情報に対してミクロレベルの情報が一般には特殊な対象者・消費者から得られているという「選択バイアス」の問題を踏まえたマクロレベルとミクロレベルのデータを融合する方法論を開発する。

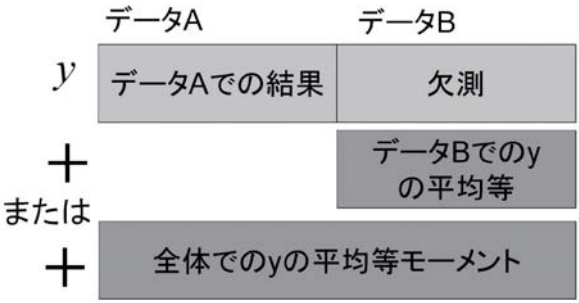

図３：ミクロレベルとマクロレベルの融合について

ここで図３において関心のある変数群をｙとする。上記の標本全体（データＡとＢ）で無作為抽出と仮定しよう。ここで

ケース１）データＡが無作為抽出であり、マクロ情報で推定の精度を高めたい

ケース２）データＡは無作為抽出ではなく、マクロ情報でバイアスを除去したい

の２つのケースが考えらえる。通常はケース２であり、ｙの平均や分散、または特定の階級幅での比率などモーメントだけ代表性ある調査から得られているような状況が考えらえる。

ケース１の状況において推定を行うための解析法としてよく知られているのはImben and Lancaster(1994)の一般化モーメント法を用いた方法などが挙げられる。Feit et al (2013)はマクロデータとミクロデータという、レベルの全く異なるデータに対するデータ融合を提案しているが、ここではミクロデータ側のモデルを階層モデルとしてそのモデルの尤度を設定し、マクロデータがミクロデータの集計データになるように尤度を周辺化するという手法をとっている。但し、この場合モデル仮定が非常に強く、また周辺化に当たって膨大な次元の数値積分を実施する必要がある。

これらの方法の問題点としては、ケース２の場合の推定は行えないことにある。これまで議論してきたように、一般にケース２の場合にはこれまで行われてきたのは上記のようなマクロ情報とミクロ情報を融合する方法というよりは、例えばｙやｘには含まれない補助変数について母集団分布（マクロ側の分布）が分かっているときに、ミクロデータ側での補助変数の分布が重みづけをする

ことで母集団分布と等しくなるような「重み」を計算するキャリブレーション推定（土屋、2009）や傾向スコアを用いた逆確率推定（星野、2009）などである。但しこれらの方法はあくまで選択バイアスを補正するための方法であり、関心のある変数 y 側の時系列情報との融合を行うといったものに利用できる方法ではない。

次に本研究の研究目的のために GMM タイプの目的関数のある準ベイズ推定(Chernozhukov and Hong, 2003)についての拡張を行う。ここでの目的はミクロデータ由来のデータについては尤度を用いること、加えて選択バイアスを考慮するモデリングを行うことである。

θ を r 次元ベクトルのパラメーターとする。ここで、準ベイズ事後分布は式(1)である。

$$q(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{exp\{L_n(\boldsymbol{\theta})\}p(\boldsymbol{\theta})}{\int_{\boldsymbol{\Theta}} exp\{L_n(\boldsymbol{\theta})\}p(\boldsymbol{\theta})d\boldsymbol{\theta}} \propto exp\{L_n(\boldsymbol{\theta})\}p(\boldsymbol{\theta}), \qquad (1)$$

但し、$p(\theta)$は θ に対する事前分布であり、Θ は θ のパラメーター空間である。そして $Ln(\theta)$は GMM や M-estimators、対数尤度関数の代わりに経験尤度関数などの目的関数を示している。(Chernozhukov and Hong, 2003; Hoshino, 2008; Yin, 2009; Yang and He, 2012)

このとき準ベイズ事後平均は次のように表される。

$$\hat{\boldsymbol{\theta}} = \int_{\boldsymbol{\Theta}} \boldsymbol{\theta} q(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta} = \int_{\boldsymbol{\Theta}} \boldsymbol{\theta}\left(\frac{exp\{L_n(\boldsymbol{\theta})\}p(\boldsymbol{\theta})}{\int_{\boldsymbol{\Theta}} exp\{L_n(\boldsymbol{\theta})\}p(\boldsymbol{\theta})d\boldsymbol{\theta}}\right)d\boldsymbol{\theta}.$$

ここで緩い正則条件のもとでは、準ベイズ事後平均は一貫しており漸近的に正規分布に従うということ知られている。(Kim, 2002; Chernozhukov and Hong, 2003; Yin, 2009; Yang and He, 2012)

GMM タイプの目的関数は次のように定義される。

$$L_n(\boldsymbol{\theta}) = -\frac{n}{2}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{m}(\boldsymbol{y}_i|\boldsymbol{\theta})\right)^T \boldsymbol{\Omega}_n^{-1}(\boldsymbol{\theta})\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{m}(\boldsymbol{y}_i|\boldsymbol{\theta})\right),$$

ここでは、$m(y_i|\theta)$が $E[m(y|\theta)]= 0$ となるようなモーメント制約であり、$\Omega n(\theta)$は最適重み行列である。

$$\boldsymbol{\Omega}_n(\boldsymbol{\theta}) = E\left[\boldsymbol{m}(\boldsymbol{y}|\boldsymbol{\theta})\boldsymbol{m}(\boldsymbol{y}|\boldsymbol{\theta})^T\right].$$

Yin (2009) and Li and Jiang (2016)はモーメント制約であるm(yi|θ)に一般化推定方程式GEEを用いている。この方法は、同一被験者の観測が相関しているような縦断データに応用することが出来る。しかし各パラメーターへの制約を除いて理論制約のような追加的な制約がもしも存在するならば、GEE を用いたこの方法では各パラメーターの制約と m(yi|θ)内の追加的な制約をどちらも同時に組み込まなければならない。そのような場合には潜在変数モデルやノンパラメトリック、セミパラメトリックベイズモデルのような柔軟性のあるモデルには適用できなくなる。

この論文では柔軟性のあるモデルを組めるようにするため、モーメント制約m(yi|θ)を２つに分ける。１つが尤度関数でありもう１つが追加的なモーメント制約である。

$$\boldsymbol{m}(\boldsymbol{y}|\boldsymbol{\theta}) = \left(\frac{\partial}{\partial \boldsymbol{\theta}^T}\log p(\boldsymbol{y}|\boldsymbol{\theta}) \quad \boldsymbol{m}^{*T}(\boldsymbol{y}|\boldsymbol{\theta})\right)^T.$$

この公式からいくつかの柔軟性をもった計算結果が導かれ、ランダム効果やセミパラメトリックモデルのような柔軟性のあるモデルを構築することが可能になる。MCMC を実行した際の追加的なモーメント制約 m*T(y|θ)とは関係のない他のパラメーターをえがくとき、尤度関数(と事前分布)のみから簡単にサンプルをえがくことが出来る。加えて、式(2)に潜在変数を簡単に含むことが出来る。(詳しくはHoshino and Igari (2017)を参照)

この論文では、θを無作為に標本抽出するためパラメーターベクトルθに準ベイズジョイント事後分布を適用している。

$$q(\boldsymbol{\theta}|\boldsymbol{y})_{QB*} = \frac{\{\prod_{i=1}^n p(\boldsymbol{y}_i|\boldsymbol{\theta})\} \times \exp\left[Q_n^*(\boldsymbol{\theta})\right] \times p(\boldsymbol{\theta})}{\int \{\prod_{i=1}^n p(\boldsymbol{y}_i|\boldsymbol{\theta})\} \times \exp\left[Q_n^*(\boldsymbol{\theta})\right] \times p(\boldsymbol{\theta}) d\boldsymbol{\theta}} \tag{3}$$

$$Q_n^*(\boldsymbol{\theta}) = -\frac{n}{2}\left[\frac{1}{n}\sum_{i=1}^n \boldsymbol{m}^*(\boldsymbol{y}_i|\boldsymbol{\theta})\right]^T \boldsymbol{\Omega}_n^{*-1}(\boldsymbol{\theta})\left[\frac{1}{n}\sum_{i=1}^n \boldsymbol{m}^*(\boldsymbol{y}_i|\boldsymbol{\theta})\right],$$

但しΩ* n(θ)はE[m*(y|θ)m*(y|θ)T]がサンプルサイズが多いと収束する先の行列である。

ここで、モーメントm*(y|θ)の外部情報を条件として、以下の準ベイズ事後分布を倍にした尤度に比例していることに留意する。

$$q(\boldsymbol{\theta}|\boldsymbol{m}^*)_{QB*} = \frac{\exp\left[Q_n^*(\boldsymbol{\theta})\right] \times p(\boldsymbol{\theta})}{\int \{\exp\left[Q_n^*(\boldsymbol{\theta})\right] \times p(\boldsymbol{\theta}) d\boldsymbol{\theta}}.$$

Hoshino and Igari (2017)では、この推定量の一貫性と漸近的な特性の証明を行っている。

さらに、セミパラメトリック準ベイズ推論において、先ほどの式(3)をDPM形式である式(4)と入れ替える。

$$\boldsymbol{m}(\boldsymbol{y}|\boldsymbol{\theta}) = \left(\frac{\partial}{\partial \boldsymbol{\theta}^T} \log \sum_{k=1}^{\infty} p(\boldsymbol{y}|\boldsymbol{\theta}_k, \boldsymbol{q}=k)p(\boldsymbol{q}=k) \quad \boldsymbol{m}^{*T}(\boldsymbol{y}|\boldsymbol{\theta})\right)^T. \tag{4}$$

この場合、m*(y|θ)がDPM形式でのモーメント制約である。

ここでは、S次元の補助情報y* = (y* 1, ..., y* S)Tが利用できると考える。モーメント制約であるm*(yi|θ)は式(5)が成り立つことで定義される。

$$\boldsymbol{m}^*(y_i|\boldsymbol{\theta}) = \begin{bmatrix} I_i^1[y_1^* - E[y_i|\boldsymbol{\theta}]] \\ \cdots \\ I_i^S[y_S^* - E[y_i|\boldsymbol{\theta}]] \end{bmatrix}, \tag{5}$$

この場合、被験者 i がグループ s(例えば性別や年齢層などのグループ)に属するときにIs i = 1となる。そして期待値はE[yi|θ] =Σ∞ k=1 πkE[yi|θk]となる。

結果としてのセミパラメトリック準ベイズ事後分布は、下の式で表現できる。

$$q(\boldsymbol{\theta}, \boldsymbol{q}|\boldsymbol{y})_{SQB*} = \frac{\{\prod_{i=1}^n \sum_{k=1}^{\infty} p(\boldsymbol{y}_i|\boldsymbol{\theta}_k, q_i=k)p(q_i=k)\} \times \exp\left[Q_n^*(\boldsymbol{\theta})\right] \times p(\boldsymbol{\theta})}{\int \{\prod_{i=1}^n \sum_{k=1}^{\infty} p(\boldsymbol{y}_i|\boldsymbol{\theta}_k, q_i=k)p(q_i=k)\} \times \exp\left[Q_n^*(\boldsymbol{\theta})\right] \times p(\boldsymbol{\theta})d\boldsymbol{\theta}}.$$

さて、ここで選択バイアスを考慮するモデリングを導入する。y を従属変数ベクトル、x を独立変数ベクトル、z を欠測指標ベクトルとする。従属変数ベクトルである y は無視できない欠測を含んでいる。ここで考えている選択モデルは以下である。

$$p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\lambda})p(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{\gamma}),$$

ここでp(z|y, γ)は欠測メカニズムを表している選択モデルであり、p(y|x, λ)は線形回帰や比例ハザードモデル、またはポアソン回帰モデルなどのパラメトリックモデルである。

yiが欠測していて、それ自体の変数の値に対応しているような選択モデルについて考えていく。zi = 1のときにyiは観測される、zi =0のときにyiは欠測

していると定義する。サンプル選択モデルを含む選択モデルにおいて、もし欠測メカニズムが誤って指定されたら推定結果にはかなりのバイアスが生じるということは広く知られている。そこで、誤指定を防ぐために選択メカニズム p(z|y, γ)のなかで DPM を用いることにする。

$$p(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{\gamma}) = \sum_{k=1}^{\infty} \pi_k p(\boldsymbol{z}|\boldsymbol{y},\boldsymbol{\gamma}_k).$$

DPM は柔軟性があるが、実際のモデルの識別はしばしば弱いことがあり、外部情報からのモーメント制約を用いている。
さて、結果としてのセミパラメトリック選択モデルの尤度関数は以下のように表現できる。

$$\begin{aligned} L &= \prod_{i=1}^{n} \left\{ p(y_i|\boldsymbol{x_i},\boldsymbol{\lambda}) p(z_i|y_i,\gamma) \right\}^{z_i} \left\{ \int \{ p(y_i|\boldsymbol{x_i},\boldsymbol{\lambda}) p(z_i|y_i,\gamma) \} dy_i \right\}^{1-z_i} \\ &= \prod_{i=1}^{n} \left\{ p(y_i|\boldsymbol{x_i},\boldsymbol{\lambda}) \sum_{k=1}^{\infty} \pi_k p(z_i|y_i,\gamma_k) \right\}^{z_i} \left\{ \int \{ p(y_i|\boldsymbol{x_i},\boldsymbol{\lambda}) \sum_{k=1}^{\infty} \pi_k p(z_i|y_i,\gamma_k) \} dy_i \right\}^{1-z_i}. \end{aligned}$$

ここでS次元の集団レベルの情報である y* = (y* 1, ..., y* S)T は、真の分布のパラメーターを識別するようになっているとする。このとき集団レベルの情報である m*(yi| λ)からのモーメント制約は次のように定義される。

$$\boldsymbol{m}^*(y_i|\boldsymbol{\lambda}) = \begin{bmatrix} I_i^1 [y_1^* - E[y_i|\boldsymbol{x}_i,\boldsymbol{\lambda}]] \\ \cdots \\ I_i^S [y_S^* - E[y_i|\boldsymbol{x}_i,\boldsymbol{\lambda}]] \end{bmatrix},$$

ここでは、被験者 i が性別や年齢層などのグループである s に属する場合に Is i = 1 となる。潜在変数である ymiss i や DPM である qi の構成要素は期待値 E[yi|xi, λ]、blocked ギブスサンプラー、モンテカルロ統合には含まれず、モーメント制約の計算では必要とされない。

モーメント制約の目的関数は下記のように表現できる

$$Q_n^*(\boldsymbol{\lambda}) = -\frac{n}{2} \Big(\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{m}^*(y_i|\boldsymbol{\lambda}) \Big)^T \boldsymbol{\Omega}_n^{*-1}(\boldsymbol{\lambda}) \Big(\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{m}^*(y_i|\boldsymbol{\lambda}) \Big),$$

ここでは Ω* n(λ) = E[m*(y| λ)m*(y| λ)T]である。
以上から、準ベイズ事後分布は以下のように表現できる。

$$
\begin{aligned}
q(\boldsymbol{\lambda}, \boldsymbol{\gamma}|\boldsymbol{y}, \boldsymbol{z}) &\propto L \times exp\{Q_n^*(\boldsymbol{\lambda})\} \times p(\boldsymbol{\lambda}) \times \prod_{k=1}^{M} p(\gamma_k) \\
&= \prod_{i=1}^{n} \left\{ p(y_i|\boldsymbol{x_i}, \boldsymbol{\lambda}) \sum_{k=1}^{M} \pi_k p(z_i|y_i, \gamma_k) \right\}^{z_i} \left\{ \int \{ p(y_i|\boldsymbol{x_i}, \boldsymbol{\lambda}) \sum_{k=1}^{M} \pi_k p(z_i|y_i, \gamma_k) \} dy_i \right\}^{1-z_i} \\
&\quad \times exp\left\{ -\frac{n}{2} \left(\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{m}^*(y_i|\boldsymbol{\lambda}) \right)^T \boldsymbol{\Omega}_n^{*-1}(\boldsymbol{\lambda}) \left(\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{m}^*(y_i|\boldsymbol{\lambda}) \right) \right\} \\
&\quad \times p(\boldsymbol{\lambda}) \times \prod_{k=1}^{M} p(\gamma_k),
\end{aligned}
$$

上記の準ベイズ事後分布からの母数推定、具体的にはマルコフ連鎖モンテカルロ法の実装については、報告書本文をご覧いただきたい。

提案手法を用いた解析例として、競合企業での購買有無を外部マクロ情報として考慮した購買間隔への広告販促の効果推定を複数実施したが、特に以下ではいくつかの解析例を紹介する。

マーケティング実務では、研究者はしばしば消費者が最後の購買から 30 日以上商品を購買していないとき分析でデータを削除するという仮定を置くことがあり、これは打ち切り指標が欠測しているのと同じだと考えられている。これらのデータ利用はマーケティングにおける実践ビジネスの場ではよくあることであり、RF 分析と呼ばれている（図４）。

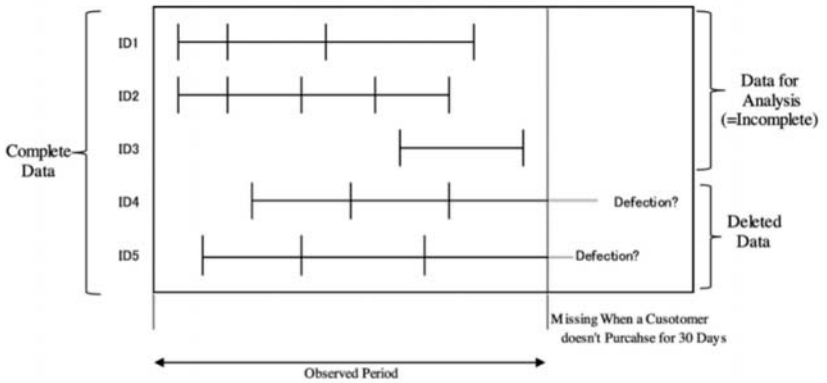

図４：RF 分析の模式図

RF 分析の場では、R が最終購買の発生を示している。しかしながら、分析から除外された（図４での Deleted Data）消費者も、将来的には商品の購買を続けるかもしれない。ここで、この分析の結果は偏った推定や間違った解釈の原因となる可能性がある。本研究で実証分析として、インテージ(株)からいただいた Syndicated Consumer Index (SCI)データを使用する。SCI データは日本のマーケティング分野での購買パネルデータの業界標準である。SCI は購買の発生、購買された商品、消費者により購買された商品の数、商品の量と値段、そして日付付きの購買された店名を記録している。分析では、消費者が最終購買日から 30 日以上購買をしていなかったら観察されなくなるとする。従ってこのデータセットは不完全なものであり、結果として強いバイアスのある結果が生みだされる。分析では、ティッシュやトイレットペーパーという名前を製品カテゴリーとして使っており、2015 年の１月から５月の分のデータを使用している。N は 11579 である。

図５に観察期間のヒストグラムを載せてある。

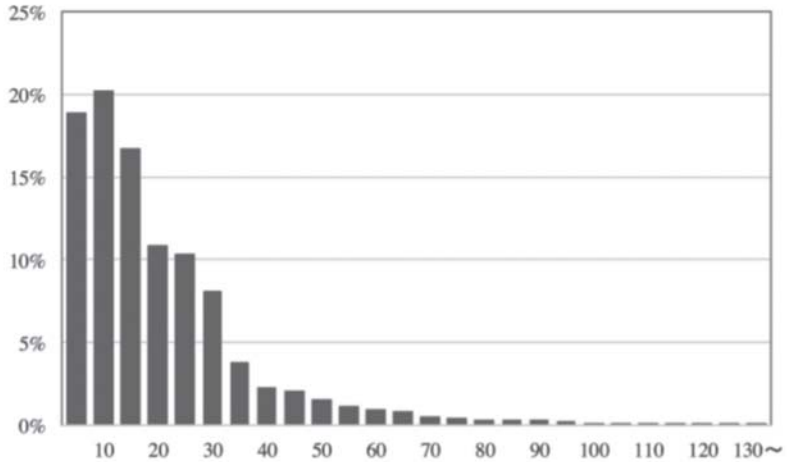

図５：観測された間隔の分布

実証分析で用いているティッシュ/トイレットペーパー製品は、平均して 20 日ごとに購買がなされている。もし研究者が長い期間、例えば 2, 3 年にわたって購買発生を捉えたデータを使うことが出来るなら、打ち切り指標が欠測して

いることによるバイアスは最小に抑えられるだろう。しかしもし耐久財のような購買ごとのインターバルが長いような商品の記録を分析したとしたら、たとえ長期間のパネルデータを利用してもそのバイアスは無視してしまうには大きくなる。

従って、提案手法を用いて、打ち切り期間がわからないような購買データを扱うことが妥当である。

最終購買日からの日にちを独立変数T、そして共変量xを値段、性別(男性は1)、そして世帯人数の4種類を定義する。そして、値段は標準化し、購買発生時の価格が通常の価格と等しい場合に1とされる。ここで価格の係数は負になるべきである。なぜなら価格割引が利用できる際に消費者は商品を購買する可能性が高いからである。また性別(男性)の係数は負になるべきである。なぜなら女性の消費者のほうが多くの場合日用品などをより頻繁に購買を行うからである。一方で、世帯人数の係数は正であるべきである。なぜなら大家族の消費量は、体系的に1人や2人世帯のそれよりも多いからである。最後に、補助情報であるt*を定義する。具体的にはSCIの全データから計算される購買間隔の分布を補助情報とする。ここで、xのために計8つの群を使用する。価格の範囲を3つ(0.9以下、0.9から1、1以上)、男性or女性の性別、そして子供の有無である(表1)。

表1：補助情報

		Auxiliary Information
Total		23.97
Price	under 0.9	22.81
	from 0.9 to 1	24.16
	over 1	24.83
Gernder	male	26.46
	female	23.36
Child	yes	22.62
	no	25.06

ここではSCIの全データから補助情報を計算したが、同様に政府統計やマーケティングリサーチ会社からしばしば提供される外部シンジケート調査を用い

ることでもこの情報を計算することが出来ることに注意しておきたい。

実データ分析においては、２つのモデルからパラメーターを推定する。それは通常のベイズモデルと提案モデルの２つである。この２つのモデルの中で、収束後に 20000 回 MCMC を回し、Geweke 法により各モデルの収束を確認した。この推定された結果は表２に示している。

表２：推定結果

		Normal Bayes Mean		Proposed Mean	
α	shape parameter	1.144	(1.129, 1.159)	1.017	(1.0142, 1.020)
β_0	intercept	-3.455	(-3.609,-3.304)	-3.107	(-3.117,-3.097)
β_1	price	-0.140	(-0.274,-0.001)	-0.313	(-0.323,-0.302)
β_1	gender(male=1)	0.084	(0.040, 0.128)	-0.098	(-0.100,-0.096)
β_1	family size	0.060	(0.045, 0.075)	0.064	(0.062, 0.065)

ここでは、事後平均と 95%信頼区間が記されている。結果によると、この２つのモデルの係数は異なっており、特に価格と性別でそうであった。通常のベイズモデルは価格の効果を過小評価しており、加えて性別の係数はかなり正の値であった。一方、提案モデルの性別の係数はかなり大きな負の値であった。通常のベイズモデルからの結果は先行研究と矛盾しており、マーケティングの実践の場での経験的な知識とも矛盾している。しかし提案モデルからの結果はその知識や経験に合っている。このように、不完全データからの生存時間分析は偏った推定を導くことを研究者ならびに実務家は考慮しなくてはならない。

もう一つ解析例を例示する。IDPOS データ、位置情報、広告出稿としてのチラシデータ、商圏情報データなど複数のデータソースを融合した解析を提案した方法論を応用して実施した解析例を示す。

本研究で提案したマクロデータとミクロデータの融合解析の方法論が機能するかどうかを確認することが目的であるため、あえて得られている完全なミクロデータのうち一部の店舗のデータのみミクロデータが利用可能であり、それ以外の店舗については店舗の集計データのみ利用可能であるという形式にしたデータセットを利用している。具体的には某スーパーチェーンから提供いただ

いた IDPOS データと各店舗の商圏情報を利用した解析を行う。本研究では利用可能な 208 店舗の 52 週のデータのうち、約 3 割である６０店舗のデータについてはミクロレベルで、それ以外の 148 店舗のデータについてはマクロレベルで利用可能なデータを利用した。但しミクロレベルデータについては当該期間において各店舗でビールを少なくとも 5 回以上購入した ID について、各店舗 100 人を抽出するサンプリングを実施している。つまりサンプルサイズは全店舗利用可能な場合には 208 店舗×100 人×52 週のデータとなっている。

またここでは、非常に購入頻度の高いビールの特定 SKU について、ミクロレベルでは個人IDごとでの週次の購入数量をゼロ過剰ポアソン回帰モデルで特定化し、その説明変数には先週の当該 SKU 購買有無、その SKU の「当該店舗での期間最大売価－その時点での売価」、月次ダミー、個人の固定効果を利用する。

ここでの関心はその SKU の「当該店舗期間最大売価－その時点での売価」、つまり値引き販促の効果が店舗によってどの程度異なるか、である。より具体的には値引き販促の係数について、店舗間でどのように異なるかに関心があるものとし、その店舗が存在する商圏のいくつかの変数が含まれるベクトル v を用いて、この係数が以下の変量効果モデルに従うと仮定する。

$$\beta_s^P = \eta_0 + v_s^t \eta + \varepsilon_s , \quad \varepsilon_s \sim N(0, \sigma_e^2)$$

より具体的には v の要素は 2 変数で 2㎞ 商圏内での競合店舗の数と世帯収入 1000 万円高所得者の比率である。但し、解釈の容易性のためそれぞれ店舗全体で計算した値を標準化した値を用いることとした。

一方、マクロレベルのデータとしては w 週での店舗 s の平均購買数量 $\tilde{y}_{ws}$ を変数として、これの時系列データを利用する。但し 208 店舗から一部の 60 店舗に店舗 s が選ばれるかどうかは、v と $\tilde{y}_{ws}$ の年間平均を説明変数とするロジスティック回帰モデルによって決定した。具体的には、より競合店舗が多く、世帯収入が高く、平均購買数量が高い店舗が 60 店舗に選ばれやすいという形でミクロデータにバイアスが生じるように設定した。

解析としては、ここでは (1) 選択された 60 店舗のみで計算した値引弾力性、(2)3 章で提案された、マクロ情報も加味した解析による値引弾力性、ならびに (3)208 店舗全体で計算した値引弾力性、の比較を行う。(1) と (3) の違いがあれば、まず通常のミクロデータをもちいた解析では選択バイアスが存在すること、

および(2)が(3)に近い値を出していれば、提案した手法を用いることでミクロデータとマクロデータの融合の結果選択バイアスを除去しながら解析を行うことが可能であることが示されたことになる。

一番の関心である平均価格での価格弾力性、および競合店舗数と高所得世帯比率の価格弾力性に与える係数を計算したところ表3のようになった。

	方法1		方法2		方法3	
	推定値	s.e.	推定値	s.e.	推定値	s,e,
平均での弾力性	2.7535		3.0883		3.1203	
高所得者比率	−0.0776	0.0133	−0.0986	0.0110	−0.0979	0.0083
競合の数	0.0872	0.0128	0.1309	0.0109	0.1325	0.0081

表3：価格弾力性、および競合店舗数と高所得世帯比率の価格弾力性への係数

関心の対象である値引弾力性の値が店舗レベル変数にどれだけ依存するかを可視化する目的で、横軸を「競合店舗数と高所得世帯比率の標準化得点」、縦軸を当該 SKU の期間平均売価における値引弾力性の係数とする図を方法(1),(2),(3)について計算した。但し値引弾力性はポアソン回帰の係数そのものではないため、例えば中川、星野,2017 の式17を参照されたい。

表3からは、60店舗だけでの値引弾力性は大きく過小評価されていること、それだけではなく店舗自体もより均質的に選択がされているため、高所得者比率や競合の数の値引弾力性への影響が過小評価されている。方法1に比べて全店舗を利用した方法3であると、弾力性の値自体高く、また両変数の影響は1.5倍程度に拡大する。

提案手法は全店舗の週次での店舗レベルの価格変化と購買数量の関係を時系列データとしたものをマクロ情報として利用しており、方法3にかなり近い値を推定していることが分かる。

さらに、一情報、チラシ情報、IDPOS からの購買履歴情報という異なるソースからのデータを利用した解析例を説明する。本報告書の目的はマクロデータとミクロデータのデータ融合手法の開発とマーケティング実務、広告効果推定への応用であるため、6.1 と同様に解析の目的はこれら3つの情報をシングルソースデータとして利用した解析をメインとするのではなく、位置情報とチラシ情報はあくまでその一部だけミクロデータが利用できる設定にして、どの程度提案手法が推定値や解析結果を復元できるか、という提案手法の妥当性検証に

利用している。

具体的には我々の研究グループでは当該チェーンの一部である 10 店舗について 2017 年 8 月から 10 月の 3 か月間にわたって以下の情報を取得した。

(1) 10 店舗の IDPOS データ

(2) 10 店舗とその競合店舗と考えられる 70 店舗の合計 80 店舗について、スマートフォンから取得できる位置情報

(3) 上記 80 店舗の飲料に関するチラシ情報

これまでマーケティング・サイエンスや消費者行動研究で利用されてきた通常の形態の IDPOS データの限界として、競合他店への来店や購買が不明であるということが挙げられる。例えば自社で値引きやチラシ広告を実施したことで、どの程度競合他社を頻繁に利用している消費者が来店するか、あるいは逆に自社で日常的に購買している顧客が他社のチラシ広告・値引きにどの程度影響を受けるかについては小売企業もメーカーも大きな関心を持っているが、実際にこれらについて定量的に把握することができなかった。唯一この種の解析の可能性のあるデータソースはインテージ SCI やマクロミル QPR といったどの店舗であっても特定期間の購買履歴であれば収集するスキャナーパネルデータであるが、「それぞれ全国で 5 万人や 3 万人といった規模であり、特定の商圏では実際に解析ができない（今回の商圏についても各店舗十数人程度と予想されまったく解析に足る規模ではない）」「店舗の販促情報やチラシ広告などの情報は取得されていない」といった問題点がある。

そこで、我々研究グループでは上記に記載した 3 つのデータソースからの情報取得を行い、位置情報と IDPOS を店舗への滞在時間とレジ通過時間の情報を利用することでマッチングし、IDPOS 上の顧客の一部については他店への来店有無が分かるというデータ形式に変換した。

但しこれらの位置情報データの取得とマッチングは通常のマーケティング実務場面で利用することは難しく、またせいぜい一般に利用され始めているのは NTT ドコモのモバイル空間統計のような集計レベルの位置情報データである。

そこで個人レベルで IDPOS データと位置情報データマッチングしたデータの解析結果を“真値”とし、本研究で提案されたミクロレベルのデータ（IDPOS データ）とマクロレベルのデータ（チラシの出稿および複数店舗利用についての集計データ）を用いた解析によってどの程度その真値に近い推定が可能かを評価することとした。

またここで IDPOS を利用させていただいたスーパーチェーンは EDLP 型の価格政策を採用しており、生鮮日配品を除く多くの商品で価格変動はほとんどない。

具体的な解析としてはまずチラシデータと位置情報だけで解析可能ないくつかの基礎的分析をおこなった。集計レベルの直観的分析として、小売企業が良く行うであろう解析として、チラシ販促時の顧客数が販促を行わなかったときに比べて何パーセント上昇したか、およびチラシの内生性を考慮して曜日ダミーを入れた解析、シェアの分析、店舗間の利用行動の推移確率行列についての解析を行う。

次に、3 か月の日次データ（92 日間）での自店舗への来店および清涼飲料水カテゴリーの購買有無を説明するためのプロビット回帰モデルを考える。

まずは競合他社のチラシ販促が自店舗への来店に与える影響を考えるための分析として、説明変数をチラシデータから取得できる「競合店舗のチラシ販促有無」IDPOS データから取得できる「個人の変量効果」および「曜日ダミー・休日ダミー」とする。ここで個人の店舗ロイヤルティは個人の変量効果に吸収されると考えることが可能である。

また、本来の関心は「競合店舗のチラシ販促有無」が「自店舗へ来店するかしないか」ではなく、自社顧客が「他店に来店するかしないか」に関心がある。自店舗に来ないとしても、競合店に来店しなければ、他店へのスイッチがなく、自社で将来購買する可能性が高いからであり、小売りにとっては特段問題がない。

そこで、ここでは従属変数である「他店への来店か、自店への来店か、当日どの店舗にも来店しないか」の 3 カテゴリーの名義プロビット回帰モデルを想定し、それを「自社のチラシ」「他店のチラシ」「個人の変量効果」および「曜日ダミー・休日ダミー」で説明するモデルを考える。今回位置情報のデータが利用でき、IDPOS とのマッチングも行われたため、上記の 3 カテゴリーの名義プロビット回帰モデルを推定することができる。

しかし、IDPOS データだけでは「他店への来店」か「当日どの店舗にも来店しない」かが識別できないため、従属変数が正確に測定できない、統計学的には測定誤差のあるデータ、あるいは変数内誤差のあるデータ（例えば Carroll ら 2006）とみなせる。この状況で外部データとして日次の位置情報データを ID レベルでの複数店舗の来店分布として集計して利用することを考える。具体的に

は前節同様、方法１として IDPOS での単純な自社来店有無を目的変数としたプロビット回帰モデルから、他社のチラシ出稿の弾力性を推定する。方法２として 3 章以下で提案された手法を用いて、
また方法３は位置情報と IDPOS をマッチングさせたデータからの解析結果であり、方法１に比べて方法２がどの程度方法３に近い精度の解析結果を与えることが可能なのかが手法の妥当性検証となる。

さらに、自社に来店した顧客に限定して、自店での清涼飲料水カテゴリーでの購買有無を「清涼飲料水カテゴリーへのチラシ販促により他店に来店したかどうか」「個人の変量効果」「曜日ダミー・休日ダミー」「当該カテゴリーの平均的な価格販促額（6.1 同様に SKU ごとの期間最大売価と日次の価格の差分のカテゴリー平均)」で説明する解析を実施する。
但し、「清涼飲料水カテゴリーへのチラシ販促により他店に来店したかどうか」自体については位置情報を取得しないとわからず、あくまでも IDPOS とチラシ情報からは「競合店舗の清涼飲料水カテゴリーへのチラシ販促有無」しかわからない。

これについても上記同様、方法１～３の比較を行う。方法１と方法 3 はそれぞれ「IDPOS を使ったナイーブな推定」と「位置情報と IDPOS をマッチングした完全データでの推定」であり、一方方法２で利用する外部情報としてはチラシデータと位置情報データと結合することで「競合店舗の清涼飲料水カテゴリーへのチラシ販促によって競合店舗に来店したかどうか」の集計分布である。

まずチラシデータと位置情報だけで解析可能ないくつかの基礎的分析の結果として、位置情報の各店舗での検出数の単純な時系列解析を行う。具体的には以下のモデルを考える。

$$y_{jt} = \beta y_{jt-1} + \gamma w_{jt} + \lambda z_{jt} + \omega^t d_t + e_{jt}$$

ここで yjt は j 店舗での t 日の位置情報の検出数であり、wは自社チラシダミー、z は他社チラシダミー、d は曜日ダミーベクトルであり、モデル適合の観点から AR1 モデルで十分であった但しここでは日次で 92 日中 80 日以上出稿する 6 店舗は除外した解析を実施した。

また、自社チラシの平均での弾力性を計算したところ、95%信頼区間は[0.1385, 0.2770]であった。

次に店舗間のスイッチング行動をマルコフ連鎖モデルを用いて推定する。これはIDで紐づいた消費者の推移を追う位置情報でないとできないが、この分析だけであればIDPOSは不要であり、位置情報とチラシ情報についての外部情報の分析とみなすことが可能である。また自社側の10店舗を10エリアとみなすと、これらエリアごとに競合店舗数が4〜11と異なるため別々に推移確率行列を作成する。

具体的には推移確率行列のｉ行ｊ列要素(店舗ｊからｉに推移する)を

$$p_{ijt} = \frac{\exp(\mu_{ij} + \omega d_t + \gamma w_{it} + \lambda z_{jt})}{\sum_{i=1}^{K} \exp(\mu_{ij} + \omega d_t + \gamma w_{it} + \lambda z_{jt})}$$

但しｄは曜日ダミー、ｗやｚは店舗ｉと店舗ｊがｔ日の朝チラシを出すと１となるダミーである。

ここで、簡単のため10地域でチラシの係数が共通というモデルを計算したところ、チラシ販促の平均弾力性の95%信頼区間は[0.0893, 0.1521]と低いことがわかった。

次に上記で説明した２つの解析を行った結果を示す。具体的には「他店への来店か、自店への来店か、当日どの店舗にも来店しないか」の3カテゴリーのうち、他店への来店を説明する名義プロビットの係数から弾力性を計算している。但し方法１では識別性の問題から、単に「自店に来店しない」ことを説明する2値の回帰を実施している。さらに、自社に来店した顧客に限定して、自店での清涼飲料水カテゴリーでの購買有無と「清涼飲料水カテゴリーへのチラシ販促により他店に来店したかどうか」との関係についての解析を実施した。

		方法1		方法2		方法3	
		推定値	s.e.	推定値	s.e.	推定値	s,e,
他社来店への効果	自社チラシ	−0.1392	0.0402	−0.1487	0.0387	−0.1602	0.0339
	他社チラシ	0.4834	0.0744	0.6274	0.0725	0.6938	0.0662
購買への効果	自社チラシ	0.0872	0.3333	0.1355	0.2792	0.1399	0.2374
	他社チラシ	−0.7983	0.0071	−0.8539	0.0630	−0.8904	0.0558

表３：３つの方法での解析結果の違い

上記の解析結果からは、それぞれ方法３のように位置情報をマッチングする

ことで、見た目のチラシ販促の効果(方法１)よりもその効果が大きいことが分かる。また方法２のようにチラシ販促データと位置情報をあくまで外部情報として利用した場合(方法２)であっても、方法３にかなり近い精度の解析結果を導出することが可能となることが分かる。

特にこの節の解析結果は測定誤差モデルにおいての外部情報を利用する解析の一つとして位置付けることができ、マーケティング実務のみならず統計学的にも新規的な方法であると評価できると考える。