複数代替案による合成選択肢を含むデータにも適用可能な離散選択モデルの推定特性

江田　裕貴[1]・倉内　慎也[2]

[1]学生会員　愛媛大学大学院　理工学研究科（〒790-8577 愛媛県松山市文京町3番）
E-mail: eda.yuki.12@cee.ehime-u.ac.jp

[2]正会員　愛媛大学准教授　理工学研究科（〒790-8577 愛媛県松山市文京町3番）
E-mail: kurauchi@cee.ehime-u.ac.jp

「公共交通」などの合成選択肢のサービスレベルを規定する場合，「鉄道」や「バス」などそれに属す代替案のうち所要時間等の条件が最も望ましい代替案の属性値をサービスレベルとする「決め打ち」を行うことが一般的である．また，被験者のモビリティから利用可能性を考慮することは，分析に利用されるデータに偏りが生じるリスクを孕んでいる．それに対し，著者らが以前提案した上位選択モデルを適用することで解決を図ることができると考えた．そこで，本研究ではシミュレーションデータを用いて「決め打ち」と上位選択モデルの推定精度を検証した．その結果，上位選択モデルでは各要素選択肢を確率的に表し特定ミスが生じにくくなっているため，決め打ちと比較してパラメータを精確に推定できることを確認した．

Key Words : *discrete choice model, elemental alternatives, aggregate alternatives, best selected model*

1. はじめに

交通需要予測では，複数の要素選択肢により構成される合成選択肢を用いて分析を行う場合が多々ある．ここで定義される要素選択肢，合成選択肢は，それ以上分解することが不可能な代替案，複数の要素選択肢によって構成される代替案をそれぞれ指すものとする．合成選択肢は，基本的に分析コストを削減するために使用されるほか，アンケート調査等においては要素選択肢レベルでの選択結果が得られている場合と，そうでない場合の双方でも扱われる．代表例として，前者では，PT調査データを用いた交通手段選択分析において，「鉄道」や「バス」という要素選択肢を「公共交通」という合成選択肢にまとめて集計やモデル化を行うことが挙げられる．一方，後者の例としては，自動車の購買行動調査が挙げられ，被験者が必ずしも保有車両のグレードまで把握していると限らないことから，車種という合成選択肢での回答を要請することがある．いずれのケースにおいても，被験者集団の特徴や傾向を知ることを目的とした集計分析においては特に問題はない．しかし，モデル化のように所要時間や費用といった交通サービスレベル（LOS）と選択行動との関係を分析する場合を考えてみよう．当然ながら合成選択肢を構成している各々の要素選択肢のLOSは異なるため，合成選択肢のLOSを設定するためには何らかの仮定をおく必要がある．

これに対し実務においては，所要時間等の特定の属性の値が最も望ましい要素選択肢を合成選択肢として特定したり（以下，「決め打ち」と呼称），LOSや個人属性等に基づき利用可能性の低い要素選択肢を排除する，等の対応が頻繁になされる．前者については，一般に，IIAが成り立つ多項ロジットモデルを用いてモデル化を行った場合，選択肢をランダムにサンプリングする限りにおいてはパラメータ推定値の一致性は保証される[1]．しかし，例えば所要時間に基づいて決め打ちを行うと，所要時間の重要度が極めて大きいとの仮定を暗においていることになるため，パラメータ推定値にバイアスが生ずるものと考えられる．後者では，個人のモビリティにより利用可能な要素選択肢を意図的に絞り込むため，被験者が直面している代替案集合を誤って特定してしまう，利用可能性が高いと判断した要素選択肢の影響を過大に評価してしまう，といった問題が生じてしまう．これらの問題に対する改善策として，著者らが提案した複数の代替案の選択を扱うことのできる上位選択モデル[2]の適用が有効であると思われる．上位選択モデルは，後述するように，合成選択肢のような詳細な選好関係が不明なデータに対し，想定される各順位付けパターンをランク

ロジットモデル等で表現し周辺和をとり，その出現確率を定式化したものである．

以上のような問題意識のもと，本研究では，LOSデータは要素選択肢レベルで与えられるものの，モデル推定に際して必要となる選択データは合成選択肢レベルであるような状況を対象に，シミュレーションデータを用いて，決め打ちによって多項ロジットモデルを適用した場合と上位選択モデルを適用した場合におけるパラメータ推定精度を比較することを目的とする．

2. 上位選択モデル

(1) 上位選択モデルの概要

アンケート調査等においては，「鉄道・バス・自動車・自転車・タクシーという 4 つの代替案の中からよく利用する交通手段を 2 つ選んでください」というように，代替案集合の中から望ましさに照らし合わせて複数の代替案を回答してもらう形式（以下，「上位選択形式」と呼称）が用いられることがある．しかし，離散選択モデルが複数の代替案の中から 1 つの代替案を選ぶことを前提としたモデルであるのに対し，上位選択形式は複数の代替案の選択を許容しており，その詳細な選好関係が不明であることからそのまま適用することができない．そのため，収集されたデータは単に集計されるだけで，効用関数の特定には活用されない．上位選択モデルは，想定される各順位付けパターンをランクロジットモデル等で表現した上で，その周辺和をとることにより回答の出現確率を定式化したものである．

(2) 上位選択モデルの一般式

まず，個人 n が J 個の代替案から望ましさの順に T 番目までの代替案を回答した場合を考える．すると，その回答ベクトル $\mathbf{I}_n$（$=\{i_{1n},\cdots,i_{tn},\cdots,i_{Tn}\}$）の出現確率は，効用関数の誤差項にガンベル分布を仮定したランクロジットモデルの場合，次式で表すことができる．

$$P_n(\mathbf{I}_n)=\prod_{t=1}^{T}\prod_{i\in G_{tn}}\left[\frac{\exp(V_{itn})}{\sum_{m\in G_{tn}}\exp(V_{mn})}\right]^{d_{itn}} \tag{2a}$$

ここに，i_{tn}：個人nがt番目に望ましいと回答した代替案，G_{tn}：個人 n が t 番目に望ましい代替案を選ぶときに直面している代替案の集合，d_{itn}：個人nにとってt番目に望ましい代替案がiであった場合は1，そうでなければ0となるようなダミー変数，V_{itn}：個人 n の t 番目に望ましいと回答した代替案 i に対する効用の確定項，V_{mn}：個人nの代替案mに対する効用の確定項，を表している．

すると，個人 n が J 個の代替案から望ましさに照らし合わせて T 個の代替案を選んだ場合の回答の出現確率は，式(2a)の周辺和をとることで，次式のように与えられる．

$$P(C_{Tn})=\sum_{G_{Tn}\in R_{Tn}}P_n(\mathbf{I_n}) \tag{2b}$$

ここに，C_{Tn}：個人 n によって選ばれた上位 T 個の代替案の組み合わせ，R_{Tn}：個人 n が選んだ上位 T 個の代替案の組み合わせと合致するような順位付けパターン G_{Tn}の集合（ただし $G_{Tn}\in R_{Tn}$），である．

このように，上位選択モデルでは，選択代替案が要素選択肢レベルでは不明である場合に対しても，全ての要素選択肢の LOS を明示的に考慮した上で，確率論的に厳密にモデル化がなされている点に特長がある．しかし，既存研究で指摘されているように，対数尤度関数のヘッセ行列が正定値行列となることが保証されず，数値解法を用いたパラメータ推定値は局所的最適解となる可能性が生じる，選択される代替案数が増加すると順位付けの組み合わせが増大し推定精度が低下する，といった問題点に留意すべきである．

(3) 合成選択肢への適用方法

上位選択モデルを用いると，合成選択肢の出現確率は，それを構成する要素選択肢の出現確率の和によって表すことができる．J個の要素選択肢が存在しているとすると，合成選択肢の出現確率は次式で表される．

$$P(S_{Tn})=\sum_{i\in C}P_n(i) \tag{2d}$$

ここに，i：合成選択肢S_{Tn}を構成する要素選択肢，C：J個の要素選択肢により構成される代替案集合，S_{Tn}：任意のT個の要素選択肢により構成される合成選択肢の組み合わせ，を表す．

一方で，個人nがJ個の代替案から望まない$J-T$個の代替案を選んだ場合を考えよう．すると，その回答の出現確率は，望ましいと回答されない出現確率に等しい．したがって，式(2b)より，次の関係式が導かれる．

$$P(W_{(J-T)n})=1-P(C_{Tn}) \tag{2e}$$

ここに，$W_{(J-T)n}$：個人nによって選ばれた下位$J-T$個の代替案の組み合わせ，である．

ここから，合成選択肢の出現確率はそれを構成しない要素選択肢が出現しない確率で表せるため，

$$P(S_{Tn})=1-\sum_{j\in\{C-S_{Tn}\}}P_n(j) \tag{2f}$$

ここに，j：合成選択肢S_{Tn}を構成しない要素選択肢，である．

つまり，合成選択肢の出現確率は，それを構成する要

素選択肢の出現確率の和，あるいはそれ以外の要素選択肢が出現しない確率により表すことができ，以下の関係式が成り立つ．

$$\sum_{i\in C} P_n(i) = 1 - \sum_{j\in\{C-S_{Tn}\}} P_n(j)$$

3. モデルの推定特性

シミュレーションにより，合成選択肢を含まない状態で多項ロジットモデルを適用した場合（以下，多項ロジットモデル）と，合成選択肢を含む人工的な選択データを作成した上で，決め打ちを行った上で多項ロジットモデルを適用した場合（以下，モデル1）と，上位選択モデルを適用した場合（以下，モデル2）の双方について，効用関数の推定精度を検証する．

(1) シミュレーションデータの作成

シミュレーションデータの作成手順として，はじめに効用パラメータの真値を設定した上で，2つの説明変数および誤差項に乱数を発生させ，各代替案の効用値を計算する．今回は，簡易的に分析を行うために説明変数を2つ用意している．次に，複数の要素選択肢をまとめた合成選択肢を構成した上で，効用値が最も高い選択肢を抽出した選択データを作成する．その後，効用パラメータを未知として扱い，選択データと説明変数を用いて各モデルにより推定されたパラメータと真値との差やモデルの適合度について比較を行った．適合度に関して，推定パラメータを用いて計算される真の選好順位の出現確率に対応するAICを用いた．なお，AICはその値が小さくなるほど適合度が高いことを示している．今回は，サンプル数を300，要素選択肢を4とした．また，乱数の影響を考慮するため，誤差項に0~1の異なる乱数を発生させた10個のデータセットを用いて分析を行っている．効用関数を以下の通り設定した．

$$U_{in} = \alpha_i + \beta_1 x_{1in} + \beta_2 x_{2in} + \varepsilon_{in} \quad (3a)$$

ここに，α_i：代替案 i の定数項（代替案1では0に固定），β_k：k 番目の説明変数の限界効用，x_{kin}：個人 n の代替案 i に対する k 番目の説明変数，ε_{in}：誤差項，を表す．

(2) パラメータ推定精度

表-1にモデル1・2によるパラメータ推定値，**図-1**に説明変数間のパラメータ比（β_1/β_2）を示す．ここで，データ1とは要素選択肢1・2を合成選択肢A，データ2では要素選択肢1・2，要素選択肢3・4をそれぞれ合成選択肢A, Bとした上で選択データを作成している．なお，モデル2を適用する場合，合成選択肢に含まれる要素選択肢の定数項を個別に設定すると，決め打ちを行っていることと同義になってしまうため，水色の枠で囲ったように合成選択肢に含まれる要素選択肢に関しては共通の定数項を設定している．加えて，もともと定数項が存在しないため真値の信頼水準を考慮していない．今回，モデル1で決め打ちを行った説明変数はx_{1in}である．

表-1より，モデル1ではほとんどのパラメータが真値から有意に異なった値に過小推定されていることがわかる．これは，ロジットモデルのパラメータ推定値には誤

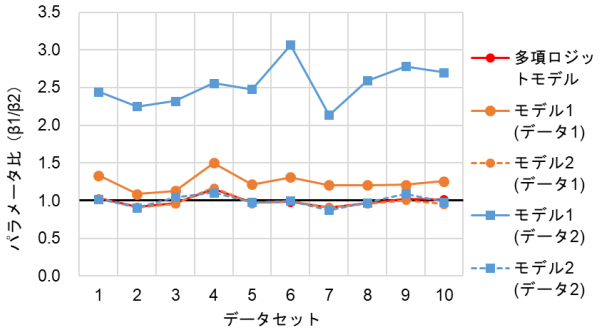

図-1 説明変数間のパラメータ比（β_1/β_2）

表-1 パラメータ推定値

		真値	モデル1										モデル2									
			1	2	3	4	5	6	7	8	9	10	1	2	3	4	5	6	7	8	9	10
多項ロジットモデル	α_2	1	0.461	1.270**	0.639*	0.848**	0.648**	1.152**	0.767**	0.635*	1.614**	1.064**										
	α_3	2	1.547**	2.318**	2.331**	1.808**	1.635**	1.968**	2.210**	2.162**	2.345**	2.228**										
	α_4	3	2.041**	3.968**	2.721**	2.780**	3.106**	2.764**	2.808**	2.949**	3.460**	3.229**										
	β_1	10	9.791**	10.639**	10.309**	10.147**	9.116**	8.718**	12.156**	9.760**	9.445**	9.672**										
	β_2	10	9.497**	11.626**	10.690**	8.783**	9.372**	8.884**	13.308**	10.114**	9.219**	9.506**										
データ1	α_2	1	0.385	-0.323	0.913**	0.552*	-0.123	0.600**	0.373	-0.225	0.594*	0.003										
	α_3	2	1.212**	0.854**	2.113**	1.343**	0.970**	1.417**	1.252**	1.262**	1.610**	1.172**	1.293**	1.488**	1.903**	1.282**	1.231**	1.328**	1.660**	1.728**	1.460**	1.636**
	α_4	3	1.621**	1.923**	2.508**	2.122**	2.166**	2.030**	1.599**	1.773**	2.514**	1.982**	1.790**	3.051**	2.302**	2.213**	2.662**	2.093**	2.198**	2.477**	2.491**	2.663**
	β_1	10	7.946**	7.353**	9.001**	8.671**	7.797**	7.315**	7.959**	7.493**	8.427**	7.548**	9.642**	9.834**	9.974**	9.544**	8.831**	8.253**	11.194**	9.291**	8.927**	9.311**
	β_2	10	5.950**	6.757**	7.972**	5.772**	6.415**	5.580**	6.589**	6.222**	6.972**	6.006**	9.361**	10.800**	10.398**	8.141**	9.145**	8.283**	12.255**	9.679**	8.863**	9.743**
データ2	α_2	1	0.273	-0.120	0.460	0.422	0.002	0.634**	0.425	-0.097	0.586**	-0.091										
	α_3	2	1.429**	0.947**	1.719**	1.311**	1.493**	1.716**	1.300**	1.459**	2.037**	1.213**	1.533**	1.800**	2.160**	1.630**	1.818**	1.674**	1.968**	2.123**	1.992**	2.133**
	α_4	3	1.093**	1.182**	1.774**	1.606**	1.398**	1.756**	1.126**	1.699**	1.871**	1.551**										
	β_1	10	7.469**	6.578**	9.064**	8.366**	7.510**	8.405**	7.561**	8.306**	8.808**	8.414**	9.350**	7.399**	10.575**	8.492**	7.904**	7.848**	10.810**	9.437**	9.323**	9.303**
	β_2	10	3.057**	2.922**	3.905**	3.271**	3.032**	2.741**	3.541**	3.204**	3.165**	3.116**	9.144**	8.166**	10.128**	7.689**	8.094**	7.853**	12.342**	9.685**	8.553**	9.484**

**：推定値 5%有意，*：推定値 10%有意，青字：真値の信頼水準 95%以下，赤字：真値の信頼水準 90%以下

差項の分散が逆数的に影響することを踏まえると，決め打ちにより誤った要素選択肢を合成選択肢として特定してしまった結果，誤差項の分散が大きくなったためであると考えられる．加えて，パラメータ比（β_1/β_2）も真値である1から大きく乖離していることが分かり，決め打ちでは，時間価値のような重要な指標を誤って推計する危険性が高いと言える．一方，モデル2では，一部のデータセットではバイアスが見受けられるが，パラメータ推定精度は比較的良好であり，パラメータ比（β_1/β_2）も1とほぼ一致した値となっている．ここから，上位選択モデルでは合成選択肢を含んでいても時間価値等の重要な指標を正しく推定できると言えよう．

(3) モデルの適合度

図-2にモデルの適合度を示す．これより，データ1・2共にモデル2の方が適合度が高くなっていることがわかる．特に合成選択肢を2つ含むデータ2では，その差は顕著である．さらに，モデル2は元データと同程度の再現性を有していることが分かる．これは，合成選択肢を含んでいてもモデル2は全ての要素選択肢の効用関数を明示的に考慮した上でモデル化を行っているためであると考えられる．

(4) 合成選択肢の特定ミス数

図-3に，決め打ちを行った際に特定化に失敗した合成選択肢の数を示す．ここで，1→2は実際の合成選択肢は要素選択肢1であるが要素選択肢2を誤って特定化してしまったサンプル数を意味している．図より，特に4→3の値が大きい．また，各データセットのサンプル数が300であることを考えるとデータ1では約10%，データ2では20%のサンプルで合成選択肢の特定ミスが起こっていることが分かる．その点，上位選択モデルでは各要素選択肢を確率的に表し特定ミスが生じにくくなっているため，パラメータが精確に推定値されていたと考えられる．

4. まとめ

本研究では，LOSデータは要素選択肢レベルで与えられるものの，モデル推定に際して必要となる選択データは合成選択肢レベルであるような状況を対象に，決め打ちによる多項ロジットモデルを適用した場合と上位選択モデルを適用した場合におけるパラメータ推定精度を比較した．その結果，決め打ちを行った場合には，パラメータ推定値に重大なバイアスが生ずる危険性があることを確認した．一方，上位選択モデルを用いた場合には，

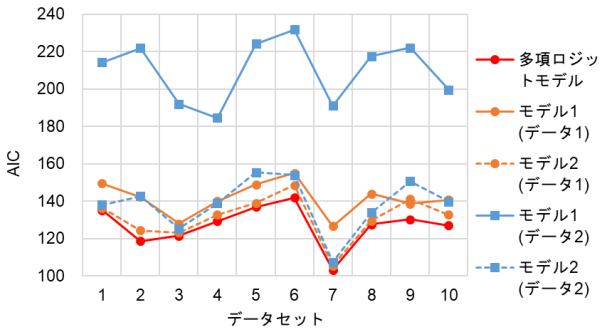

図-2 AICの値

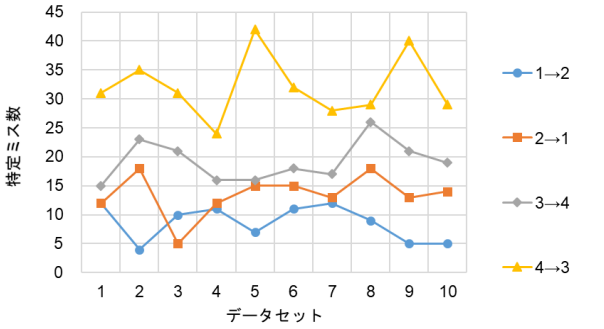

図-3 合成選択肢の特定ミス数

要素選択肢の選択を直接的にモデル化した場合と比較しても推定精度に大差ないことが明らかとなった．

近年では，ITの進展により，交通需要の分析に必要な多様なデータベースや検索エンジンが整備され，要素選択肢レベルでのLOSデータが比較的容易に作成できるようになっている．一方で，個人情報保護の観点等から，選択行動が要素選択肢としては観測できないような状況は少なからず存在するものと考えられる．加えて，上位選択モデルを用いれば，選択行動が合成選択肢として観測されているような既存データとの比較や融合利用も可能である．パラメータ同定の観点から，合成選択肢に含まれる個々の要素選択肢の定数項を推定することはできない等の課題はあるが，データの有効活用という点で今後様々な事例への適用が望まれる．

参考文献

1) Ben-Akiva, M. and Lerman, S.: Discrete Choice Analysis: Theory and Application to Travel Demand, The MIT Press, 1985.

2) 江田裕貴，倉内慎也：複数代替案の選択を考慮した離散選択モデルの開発とその基本特性に関する研究，土木計画学研究・講演集，Vol.54（CD-ROM），2016.

(2017. 7. 31 受付)

A STUDY ON IDENTIFIABILLITY OF DISCRETE CHOICE MODEL WITH AGGREGATE ALTERNATIVES

Yuki EDA, Shinya KURAUCHI