

やさしい非集計分析 第5章

複数データを用いた非集計分析

M1 斉藤

目次

0. 複数データ統合的利用の必要性
1. 選択肢別調査データを用いたモデル推定法
2. 時間的ないし空間的に異なるデータを用いたモデル構築方法(モデルの移転方法)
3. 精度が既知である集計データを用いたモデル構築方法
4. 意識データと実行動データとの統合利用方法

0. 複数データ統合利用の必要性

0.1 背景

- 非集計モデルは一般に、サンプリングされたデータから構築
⇒ **情報量の不足、情報の偏り**が生じる
cf) サンプリングデータ: 家庭訪問調査など
- 解決策として
 - ① サンプル数を増大させる(非効率的)
 - ② 性質の異なる追加的なデータで補う(⇒今回の内容)

0.2 複数データの統合的利用が必要とされる例

①低頻度トリップの効率的データ収集

都市間交通、観光交通、買回り品買い物交通など

-家庭訪問調査、選択肢別調査

②再現性の向上

構築されたモデルを時間的・空間的に異なる需要の推計に用いる場合

-モデル構築に用いたデータ+需要推計を行う箇所のデータ

③意識データの活用

SPでは回答誤差など特有のバイアスが生じる

-SPデータ+RPデータ

1. 選択肢別調査データを用いた モデル推定法

1. 1 非集計分析におけるサンプリング方法

①特性値別サンプリング

⇒トリップメーカーの持つ各種の特性値ごとにサンプリングを行う方法
(居住地域、年齢、性別など)

②選択肢別サンプリング

⇒1つあるいは複数の選択肢ごとにサンプリングする方法

③特性値と選択肢による層別サンプリング

⇒①と②をクロスさせた、特性値、選択肢の組み合わせに応じたサンプリング方法

C:選択肢

	i=1	i=2	i=3
X 特性値	Z1		
	Z2		
	Z3		

1. 1 非集計分析におけるサンプリング方法

①特性値別サンプリング

⇒トリップメーカーの持つ各種の特性値ごとにサンプリングを行う方法
(居住地域、年齢、性別など)

②選択肢別サンプリング

⇒1つあるいは複数の選択肢ごとにサンプリングする方法

③特性値と選択肢による層別サンプリング

⇒①と②をクロスさせた、特性値、選択肢の組み合わせに応じたサンプリング方法

C:選択肢

	i=1	i=2	i=3
Z1			
Z2			
Z3			

X
特性値

1. 1 非集計分析におけるサンプリング方法

①特性値別サンプリング

⇒トリップメーカーの持つ各種の特性値ごとにサンプリングを行う方法
(居住地域、年齢、性別など)

②選択肢別サンプリング

⇒1つあるいは複数の選択肢ごとにサンプリングする方法

③特性値と選択肢による層別サンプリング

⇒①と②をクロスさせた、特性値、選択肢の組み合わせに応じたサンプリング方法

C:選択肢

X
特性値

	i=1	i=2	i=3
Z1			
Z2			
Z3			

1. 2 層別サンプリングのパラメータ推定法

いまG個の層を考え、各層の (i, z) ($i \in C, z \in X$)の組み合わせの集合を $(C \times X)_g$ とする

①の特性値サンプリングの場合の集合は $(X)_g$, ②の選択肢別サンプリングの場合 $(C)_g$ とあらわされ、それぞれ③の特殊ケースであることがわかる

母集団における (i, z) の確率密度関数は式(1)のようにあらわされる

$$f(i, z) = P(i | z, \theta) \cdot P(z) \quad \dots (1)$$

$P(i | z, \theta)$: ロジットモデルなど、特性値 z と未知パラメータ θ が
与えられたときに選択肢 i が選択される確率

$P(z)$: 特性値 z の周辺分布

定式化は、各層 g が選ばれたという条件下で (i, z) の組み合わせが実現する尤度を最大化することによって行う。尤度 L は以下の式(2)であらわされる。

$$L^* = \prod_{g=1}^G \prod_{n=1}^{N_g} \frac{f(i_n, z_n)}{\sum_{j \in C, y \in X} f(j, y)} \cdot H(g) \quad \dots (2)$$

N_g : 層 g のサンプル数 i_n, z_n : サンプル n が選んだ選択肢, n の特性値

$H(g)$: 層 g のサンプルに占める割合 G : 層の数

1.3 選択肢別サンプリングサンプリングデータを用いたパラメータ推定法

$$L^* = \prod_{g=1}^G \prod_{n=1}^{N_g} \frac{f(i_n, z_n)}{\sum_{j \in C, y \in X} f(j, y)} \cdot H(g) \quad \dots (2)$$

ここで、分母を変形すると以下のようになる

$$\begin{aligned} \sum_{j \in C_g, y \in X} f(j, y) &= \sum_{j \in C_g} \sum_{y \in X} P(j | y, \theta) \cdot P(y) \\ &= \sum_{j \in C_g} Q(j) = Q(i) \end{aligned}$$

よって、対数尤度関数は以下のようになる

$$L_c = \sum_{i \in C} \sum_{n=1}^{N_i} \ln(P(i | Z_n, \theta)) + \sum_{i \in C} \sum_{n=1}^{N_i} \ln(P(Z_n) \cdot \frac{H(i)}{Q(i)})$$

$Q(i)$ は θ を含む関数であるため第2項は無視しえず、従来手法では未知パラメータ θ を求めることはできない。

1. 3 選択肢別サンプリングサンプリングデータを用いたパラメータ推定法(代替手法)

①WESML推定量

サンプルと母集団のシェアの違いを重みとした推定方法

$$L_w = \sum_{i \in C} \sum_{n=1}^{N_i} \frac{Q(i)}{H(i)} \ln(P(i | Z_n, \theta))$$

$Q(i)$: 母集団の*i*番目選択肢分担率

$H(i)$: サンプルの番目選択肢分担率

②MM推定量

$$\begin{aligned} L_{MM} &= \sum_{i \in C} \sum_{n=1}^{N_i} \ln \left(\frac{P(i | Z_n, \theta) \cdot \frac{Q(i)}{H(i)}}{\sum_{j \in C} P(j | Z_n, \theta) \cdot \frac{Q(j)}{H(j)}} \right) \\ &= \sum_{i \in C} \sum_{n=1}^{N_i} \ln \left(\frac{\exp[V_{in}] \cdot \frac{Q(i)}{H(i)}}{\sum_{j \in C} \exp[V_{jn}] \cdot \frac{Q(j)}{H(j)}} \right) \\ &= \sum_{i \in C} \sum_{n=1}^{N_i} \ln \left(\frac{\exp \left[V_{in} + \ln \frac{Q(i)}{H(i)} \right]}{\sum_{j \in C} \exp \left[V_{jn} + \ln \frac{Q(j)}{H(j)} \right]} \right) \end{aligned}$$

ロジットモデルを用いたMM推定量式は、効用項に $\ln \frac{H(i)}{Q(i)}$ を加えた形

よって、定数項を含んだロジットモデルを用いる場合、通常のパラメータ推定後に推定された定数項を以下のように修正すればよい

$$\alpha'_i = \alpha_i + \ln \frac{Q(i)}{H(i)}$$

例①

- データの概略

- 選択肢別調査データ

- サンプル利用量:バス:11(36.7%)

- 自動車:19(63.3%)

- 利用総量調査データ

- Q(1)=0.3, Q(2)=0.7 (ただし、1:バス,2:自動車)

- WESML推定量を用いる場合

$$L_w = \sum_{i=1}^{N_1} \frac{0.3}{0.367} \ln \frac{e^{V_1}}{e^{V_1} + e^{V_2}} + \sum_{i=1}^{N_2} \frac{0.7}{0.633} \ln \frac{e^{V_2}}{e^{V_1} + e^{V_2}}$$

- MM推定量を用いる場合

$$\alpha'_i = \alpha_i + \ln \frac{0.3}{0.367}$$

1.4 家庭訪問調査と選択肢別調査との統合利用

- シェアが10%にも満たない選択肢に関する需要分析を効率的に行うには、選択肢別調査データの追加が有効

①定数項修正を行う方法

定数項をもつロジットモデルを採用した場合のみ適用可能
基本的な考え方は、MM推定量の場合と同じ

②尤度関数の結合による方法

通常の方法とWESML推定とを、対数尤度式を通じて結合する方法
以下のように、家庭訪問調査データを用いた対数尤度 L_H と選択肢別調査データを用いた対数尤度 L_C を和で結合し L の最大化を行う

$$L = L_H + L_C$$

2. 時間的ないし空間的に異なるデータを用いたモデル構築方法(モデルの移転方法)

2. 1 モデルの移転方法

①追加データが非集計データの場合

- 修正パラメータを推定する方法

- 移転先の行動データが不十分な場合には、移転するモデルパラメータの更新、修正にそのデータを用いることが有効

$$P_{in} = \frac{\exp[\alpha V_{in} + \beta_i]}{\sum_{j=1}^J \exp[\alpha V_{jn} + \beta_j]}$$

V_{in} : 移転元のパラメータと移転先の説明変数により算出される、個人nのI番目選択肢の選択確率

α : 効用項のスケールを調整するパラメータ

β_i : 集計シェアを修正するパラメータ

- 全パラメータを更新する方法

- 移転元、移転先のパラメータの重み平均により新たなパラメータを求める方法
ベイズ推定の考え方を用いる

2. 1 モデルの移転方法

②追加データが集計データの場合

- ここでの集計データとは、集計された各選択肢のシェア及び説明変数の集計値
- パラメータ間の相対値である時間価値等は等しい。時間や空間的な差異は定数項に大きな影響を及ぼす

$$S_i = \frac{\exp[V_{in} + \beta_i]}{\sum_{j=1}^J \exp[V_{jn} + \beta_j]}$$

上式が、I番目選択肢のシェア S_i に対する修正式である。

2.2 移転性の評価方法

追加データが非集計データの場合

- パラメータ値の差の評価

移転元のパラメータを θ_A 、移転で推計されるパラメータを θ_B とし、両パラメータ値間の差の検定を行う

- 対数尤度を用いた説明力の評価

① $-2[L_{A+B}(\theta_{A+B}) - L_A(\theta_A) - L_B(\theta_B)]$

② $-2[L_A(\theta_A) - L_B(\theta_B)]$

③ $1 - (L_A(\theta_A)/L_B(\theta_B))$

④ $[L_B(\theta_A) - L_B(0)]/[L_B(\theta_B) - L_B(0)]$

追加データが集計データの場合

- 集計値の差の評価

-推計された集計シェアの値と観測シェア間の、相対誤差、絶対誤差、RMS誤差などの指標を用いて評価

3. 精度が既知である集計データを用いた モデル構築方法

3. 1 概要

- 2との違い
 - 集計データの精度に関わる情報を合理的に更新プロセスに組み込む
- ベイズ理論を用いてパラメータを更新する

4. 意識データと実行動データとの 統合利用方法

モデル構築の仮定

意識データと実行動データの違いを誤差分散に帰着
(実行動データによるモデル誤差)
 $= \alpha \times (\text{意識データによるモデル誤差})$