基礎ゼミ第5回

MNLの導出とパラメータ推定

朝倉研究室 修士1年 小泉大哉

目次

- 離散選択モデルによる予測
- モデルの集計方法
- 二項選択モデル
- 二項ロジットモデル(BNL)の導出および推定
- 多項ロジットモデル(MNL)の導出および推定
- まとめ

離散選択モデルによる予測

- **交通行動に離散選択モデルを導入する目的**

①個人の行動をシステマティックに理解
→推定された効用のパラメータから知ることができる
②交通現象の予測
→非集計レベルで推定したモデルを予測段階で集計化

集計方法

①数え上げ法(総当たり法)

選択肢iのマーケットシェアを$S(i)$とおく

$$\boxed{S(i) = \frac{1}{N}\sum_{n=1}^{N} P_n(i)\,, i = 1, \ldots, j}$$

N:サンプル数, $P_n(i)$:個人nの選択肢iの選択確率

※ $S(i)$は個人による選択確率の**重み付き総和**であって,
各々が選ぶ選択する確率を具体的に推測し求めるわけではない

集計方法

選択肢別に標本抽出した場合の$S(i)$

$$S(i) = \sum_{j=1}^{J} \left(\boxed{W_j} \cdot \frac{1}{N_j} \sum_{n=1}^{N} P_{nj}(i) \right) \ , i = 1, \ldots, j$$

W_j:アンケート調査時の実際のマーケットシェア

集計方法

②代表的個人法

…母集団の「代表的」な仮想の個人の説明変数値を計算し，その個人の選択確率をセグメントにシェアすること

(ここでは)年齢ごとや性別で分けたもの

- 「代表的」な変数値→モデル推定に用いたサンプル数の平均値

母集団の中で説明変数のバラつきが大きい場合予測精度は低下

↓

説明変数がばらつかない同質的なグループごとに「代表的個人」を立てる（マーケットセグメント）

しかし，グループ内が同質的になる保証がないため，あまり使われない

二項選択モデル

プロビットモデル

ランダム項が**正規分布**にしたがう

$\varepsilon_1 \sim N(0, \sigma_1^2)$　　$\varepsilon_2 \sim N(0, \sigma_2^2)$

$\varepsilon = \varepsilon_2 - \varepsilon_1$の分布

$\varepsilon \sim N(0, \sigma^2)$

※$\sigma^2 = \sigma_1^2 + \sigma_2^2 - 2\sigma_{12}$

$$F_\varepsilon(V_n) = \underline{\int_{-\infty}^{V_n} f(\varepsilon) d\varepsilon} = \Phi\left(\frac{V_n}{\sigma}\right)$$

$\Phi(x)$・・・標準正規分布関数
(平均0,分散1)

ロジットモデル(BNL)

ランダム項が**極値分布**にしたがう

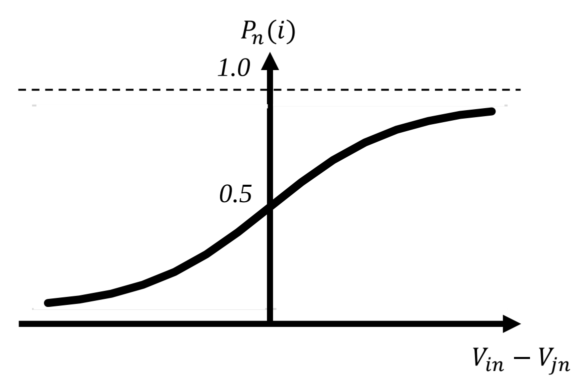
$\varepsilon_1 \sim EV(0, \mu)$　　$\varepsilon_2 \sim EV(0, \mu)$

$\varepsilon = \varepsilon_2 - \varepsilon_1$の分布

$\varepsilon \sim Logistics(0, \boxed{\mu})$

μ：スケールパラメータ[ε_nのばらつき]

$$F_\varepsilon(\varepsilon_n) = \frac{\exp(\boxed{\mu}\varepsilon_1)}{\exp(\boxed{\mu}\varepsilon_1) + \exp(\boxed{\mu}\varepsilon_2)} = \frac{1}{1 + \exp(-\boxed{\mu}\varepsilon_n)}$$



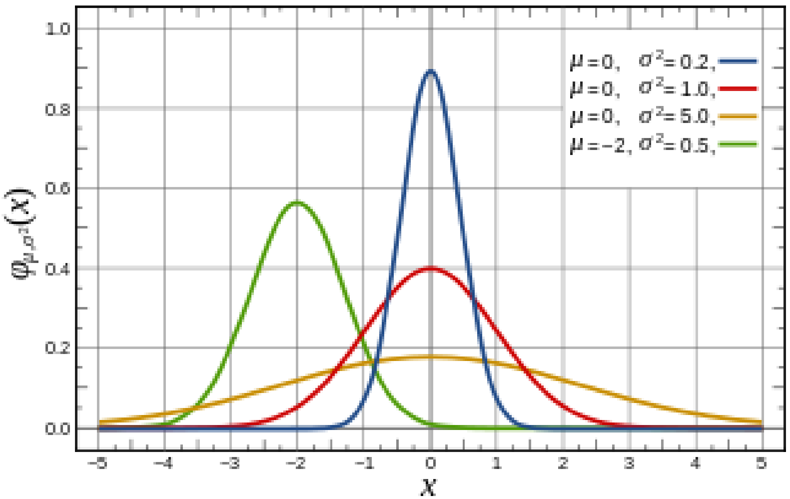
二項ロジットモデル(BNL)

- **ロジットモデルの方がよく利用される理由**

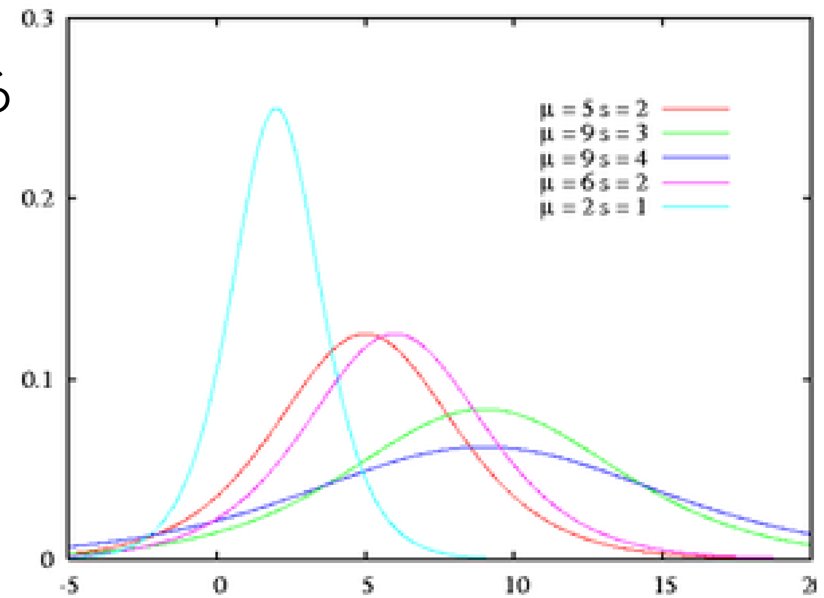
→プロビットモデルでは，選択確率に積分が残り取り扱いづらいため，積分の計算に工夫が必要

→個人の効用最大化説と整合

→正規分布よりゆるやかである



プロビットモデル



ロジットモデル

BNLの推定

- **最尤推定法**

行動モデル(未知パラメータ含む)が正しいことが前提で観測されたデータがもっともらしさ(尤度)が最大になるようにパラメータを定めている。

モデル推定においてパラメータ値を求める方法として利用されている

$$L = \prod_{n=1}^{N} \prod_{i=1}^{J} P_n(i)^{d_{in}}$$

d_{in}: 個人nが選択肢iを $\begin{cases} \text{選択した場合1} \\ \text{選択しない場合0} \end{cases}$

BNLの推定

- **最尤推定法**

実際のモデル推定では推定計算の簡便性(各項のかけ算→各項のたし算)のために上式の尤度関数に自然対数をとる

$$\ln L = \sum_{n=1}^{N} \sum_{i=1}^{J} d_{in} \ln P_n(i)$$

代わりに最大化する！！

BNLの推定

- **統計的検定**

→推定結果の値の妥当性を検証することが目的

→各項の係数の絶対値を対象とするt検定を利用

BNLの推定

- 帰無仮説と対立仮説について

①まず、帰無仮説「差はない」を立てる
②適当な指標を計算する
③その指標が起こる確率を計算する
④確率に基づいて、帰無仮説を採択するか、
　棄却するかを決める
⑤帰無仮説を採択した場合は、「意味のある差（有意差）はない」と結論する
⑥帰無仮説を棄却した場合は、対立仮説を採択し、「差はないとはいえない、つまり意味のある差（有意差）がある」と結論する

BNLの推定

t検定…各説明変数の効用値における影響を得られたパラメータ値から標準偏差で割った値(t値)で計る

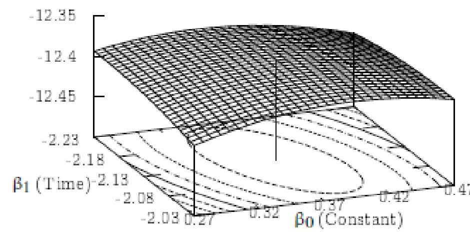
以下の順で検定をする

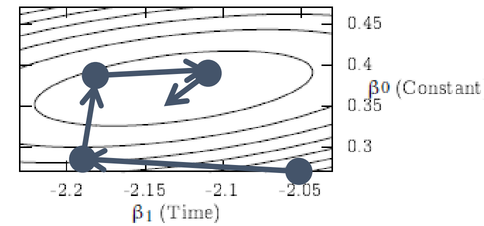
①帰無仮説と対立仮説を立てる

②有意水準を設定する(例えば5%とする)

③t値の絶対値が1.96以上であれば，そのパラメータ値が有意水準5%で有意であると言える

BNLの推定





- 非線形計画法

①反復法による最適解の探索

②目的関数が凸であれば，極大解を探索することで最大解を探すことができる

③2項ロジットモデルの対数尤度関数は凸であり，最尤解は一意に求まる

④一般的な探索終了条件は，最適解β^*において目的関数の勾配ベクトルが$\nabla L(\beta^*) = 0$になることを利用する

⑤パラメータのなかに識別不能なサンプルが含まれるとき，モデルは特異となり収束速度が著しく落ちるまたは収束しない場合がある(全サンプルで旅行時間が等しい，など)

BNLの推定

- **モデルの適合度**

モデルの確からしさの指標

➢ $L(0)$

すべてのパラメータを0とした無情報モデル（=当てずっぽうのモデル）の尤度、初期尤度

➢ $L(\beta^*)$

最大尤度モデル（=推定結果）の尤度

予測がすべて的中する完璧なモデルの尤度は1となる

BNLの推定

- **McFaddenの決定係数**

適合度指標は次のように表される

$$\rho^2 = 1 - \frac{\ln L(\beta^*)}{\ln L(0)} \quad (0 < \rho^2 < 1)$$

- **McFaddenの決定係数の欠点**

→自由度調整済み尤度比の導入(K:パラメータ数)

$$\bar{\rho}^2 = 1 - \frac{\ln L(\beta^*) - K}{\ln L(0)} \quad (0 < \bar{\rho}^2 < 1)$$

多くのパラメータを用いることでモデルが扱いにくくなることも考慮した尤度であるため，異なるモデルの比較に用いることができる

MNLの導出

多項ロジットモデルの場合にも応用するには…

- 定式化

$$P_n(i) = Pr\left[U_{in} \geq U_{jn}, \text{for } \forall j, i \neq j\right] = Pr\left[U_{in} \geq \max_{\forall j, j \neq i} U_{jn}\right]$$

また,

- 累積分布関数：

μ：ガンベル分布のスケールパラメータ[ε_nのばらつき]

$$F(\varepsilon) = exp[-exp\{-\boxed{\mu}(\varepsilon - \eta)\}]$$

- 確率密度関数：

η：ロケーションパラメータ[分布の位置(最頻値)]

$$f(\varepsilon) = \mu \cdot exp\{-\mu(\varepsilon - \boxed{\eta})\} \cdot F(\varepsilon)$$

MNLの導出

- ガンベル分布の性質

1.最頻値：$\underline{\eta}$,平均値：$\underline{\eta + \gamma/\mu}$ (γ:オイラー定数、0.577とする)

2.分散：$Var[\varepsilon_1 - \varepsilon_2] = \underline{\frac{\pi^2}{6\mu^2}}$

3.$\varepsilon_1, \varepsilon_2$がそれぞれ$(\eta_1, \mu), (\eta_2, \mu)$のパラメータを持つガンベル分布に従うとき差$\varepsilon = \varepsilon_1 - \varepsilon_2$は以下のロジスティック分布に従う

$$\underline{F(\varepsilon) = \frac{1}{1 + exp\{\mu(\eta_2 - \eta_1 - \varepsilon)\}}}$$

MNLの導出

$U_n^* \equiv \max_{\forall j, j \neq i}(V_{jn} + \varepsilon_{jn})$と定義する

$\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_j$がそれぞれ$(0, \mu)$を持つ互いに独立なガンベル分布に従うと仮定

$\rightarrow U_n^*$は$\left(\frac{1}{\mu}\ln\sum_{j\neq i}\exp(\mu V_{jn}), \mu\right)$を持つガンベル分布となる

$U_n^* = V_n^* + \varepsilon_n^*$とすると,

$$V_n^* = \frac{1}{\mu}\ln\sum_{j\neq i}\exp(\mu V_{jn})$$

MNLの導出

$P_n(i) = Pr\left[U_{in} \geq U_{jn}, \text{for } \forall j, i \neq j\right]$および，ガンベル分布の性質3.により，

$$
\begin{aligned}
P_n(i) &= Pr[V_{in} + \varepsilon_{in} \geq V_n^* + \varepsilon_n^*] \\
&= Pr[\varepsilon_{in} - \varepsilon_n^* \geq V_n^* - V_{in}] \\
&= \frac{1}{1 + \exp\left(\mu(V_n^* - V_{in})\right)} = \frac{\exp(\mu V_{in})}{\exp(\mu V_{in}) + \exp(\mu V_n^*)} \\
&= \frac{\exp(\mu V_{in})}{\exp(\mu V_{in}) + \exp\left(\ln \sum_{j \neq i} \exp\left(\mu V_{jn}\right)\right)} \\
&= \boxed{\frac{\exp(\mu V_{in})}{\sum_j \exp\left(\mu V_{jn}\right)}}
\end{aligned}
$$

←MNLの選択確率

※通常のスケールパラメータμの値は1とすることが多いが，関数的ランダム効果を出すため，NLなどでは1より小さな値とすることがある

MNLまとめ

- ε_{jn}: 独立かつ同一のガンベル分布$EV(0,\mu)$に従う

  累積分布関数：$F(\varepsilon) = exp[-exp\{-\mu(\varepsilon-\eta)\}]$

  確率密度関数：$f(\varepsilon) = \mu \cdot exp\{-\mu(\varepsilon-\eta)\} \cdot F(\varepsilon)$

- ランダム項の分散：$\frac{\pi^2}{6\mu^2}$

- 選択確率$P_n(i) = \frac{\exp(\mu V_{in})}{\sum_j \exp(\mu V_{jn})}$

MNLのIIA特性

- **IIA特性のメリット**
→選択肢集合が大きい場合，部分集合を用いて推定できる
- **IIA特性のデメリット**
→互いに相関している選択肢(例：赤バスと青バス)が含まれている場合，非現実的な選択確率を算出する可能性がある

↓

NL(Nested Logit Model)等の開発の動機となった