統計・非集計モデルの基礎
―最尤法，t値，MNL，NL―

2016/04/25(月)

スタートアップゼミ#2

B4 後藤祥孝

目次

０．統計学導入

推測統計学

ある母集団からランダムにサンプリングされたデータを用いて母集団の特性値（=パラメータ）を推測したい

推定：　パラメータが未知の時に値をデータから求める.

検定：　パラメータに対して２つの仮説を立てた上でそのどちらを選ぶかを決定する.

１．最尤法

最尤法： 推定法の１つ.
　　尤度関数を最大化するようなパラメータを求める.

ランダムサンプル X_i とその値 x_i が与えられたときに
その生起確率は母集団の未知のパラメータθを用いて
　　$f_n(x_i;\theta)$　　と表現される.

n 個のサンプル （X_1,・・・,X_n) が与えられたとき
それらが同時に起こる確率は

$$f_n(x_1,\cdots,x_n:\theta) = f(x_1:\theta)\cdots f(x_n:\theta)$$

$$= \prod_{i=1}^{n} f(x_i:\theta)$$

１．最尤法

$$f_n(x_1, \cdots, x_n : \theta) = f(x_1 : \theta) \cdots f(x_n : \theta)$$
$$= \prod_{i=1}^{n} f(x_i : \theta)$$

→ x_i はデータが与えられたという意味で定数
　θは未知パラメータ
と考えると

$$L(\theta) = f_n(\theta : x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f(\theta : x_i)$$

としてθの関数 L(θ) が作られる.

１．最尤法

得られたデータは最も起こりやすいものが起こったと考える
↓
x_iを固定してθを動かしてL(θ)が最大となるθを求める

L(θ)：尤度関数　（θの尤もらしさを表す関数）
L(θ)を最大化するθ：最尤推定量

またL(θ)は積であることから対数をとり和に直されることもある.

$$l(\theta) = \ln L(\theta)$$
$$= \sum \ln f(\theta : x_i)$$　:対数尤度関数

２．　t 値の意味

t 値：　・推定などにより得られたパラメータを
　　そのパラメータの推定標準偏差で除したもの
　　・ t 値は t 分布に従う

t 検定は得られた t 値が t 分布上の棄却域に含まれているかどうかで検定を行う.
棄却域の値は求める有意水準とサンプル数によって決まる.

ex)　あるパラメータが十分なサンプル数から得られたとき

　t 値の絶対値が1.960を上回ると,
　得られたパラメータは９５％で有意となる.

３．非集計モデルの導入

確率効用最大化モデル

・個人が利用可能な選択肢から最も望ましい選択肢を選ぶ．（離散選択，効用最大化）

・効用は選択肢特性と個人属性によって決まるが，以下のことなどにより確率的に変動する．（確率効用）
  - 全てを観測するのは不可能
  - 観測誤差の存在
  - 行動者の認知誤差
  - 効用の関数形による誤差

以上のことから効用を定式化すると・・・

3．非集計モデルの導入

$$U_{in} = \underbrace{\beta_1 x_{1_{in}} + \beta_2 x_{2_{in}} + \cdots + \beta_K x_{K_{in}}} + \varepsilon_{in}$$

$$= \underbrace{V_{in}} + \underbrace{\varepsilon_{in}}$$

確定項　　確率項

U_{in}:確率効用

V_{in}:個人nの選択肢iに対する効用の確定部分

β_k:k番目の未知パラメータ

$x_{k_{in}}$:個人nの選択肢iに対するk番目の説明変数

ε_{in}:効用の確率項

3．非集計モデルの導入

2項選択モデル

$$P_n(i) = \Pr\left[U_{in} \geq U_{jn}\right]$$
$$= \Pr[V_i + \varepsilon_i \geq V_j + \varepsilon_j]$$
$$= \Pr\left[\varepsilon_i = \varepsilon, \varepsilon_j \leq \rho + V_i - V_j\right], -\infty < \varepsilon < \infty$$

ε_iがどのような確率分布に従うかによって
$P_n(i)$の式は異なる.

プロビットモデル

中心極限定理によって正規分布を仮定する.
しかし，選択確率に積分形が残ってしまい計算負荷が重い.
（5/16に説明予定）
→ロジットモデルの導入

4．MNL

ロジットモデル

- 誤差項ε_iにガンベル分布を仮定
- 選択確率の式に積分形が残らない（クローズドフォーム）

ガンベル分布

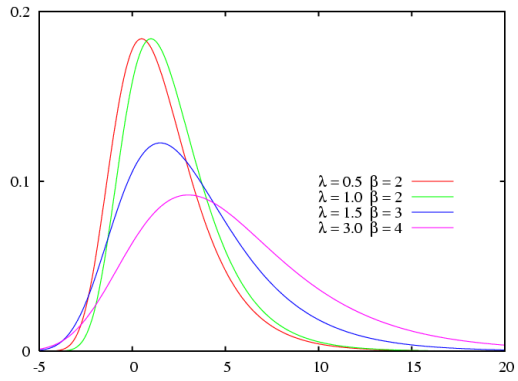
- 累積分布関数

$$F(\varepsilon) = \exp(-\exp(-\mu(\varepsilon - \eta)))$$

- 確率密度関数

$$f(\varepsilon) = F'(\varepsilon) = \mu \exp(-\mu(\varepsilon - \eta)) \exp(-\exp(-\mu(\varepsilon - \eta)))$$

μ:スケールパラメータ（εのばらつきの程度）
η:ロケーションパラメータ（分布の位置（=最頻値））
平均:$\eta + \gamma/\mu$
分散:$\pi^2/6\mu^2$　　($\gamma \approx 0.577$：オイラー定数)



４．MNL

ガンベル分布の性質

$$F(\varepsilon) = \exp(-\exp(-\mu(\varepsilon - \eta)))$$
$$f(\varepsilon) = \mu \exp(-\mu(\varepsilon - \eta)) \exp(-\exp(-\mu(\varepsilon - \eta)))$$

性質１：$\varepsilon_1, \varepsilon_2$が$(\eta_1, \mu), (\eta_2, \mu)$のガンベル分布に従うとき，
$\varepsilon = \varepsilon_1 - \varepsilon_2$は以下のロジスティク分布に従う．

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

性質２：$\varepsilon_1, \ldots, \varepsilon_i, \ldots, \varepsilon_I$がそれぞれ$(\eta_i, \mu)$に従うとき，
最大値$\max(\varepsilon_1, \ldots, \varepsilon_I)$はガンベル分布に従う
パラメータ：$\left(\frac{1}{\mu} \ln \sum_{i=1}^{I} \exp(\mu \eta_i)\ ,\ \mu\right)$

４．MNL

2項ロジットモデル続き（簡単のため$\eta = 0, \mu = 1$とする）

確率項の累積分布関数

$$\psi(\varepsilon) = \Pr[\varepsilon_1 \leq \varepsilon] = \exp[\exp(-\varepsilon)]$$

を適用すると

$$P_n(i) = \Pr[\varepsilon_1 = \varepsilon, \varepsilon_2 < \varepsilon + V_1 - V_2], -\infty < \varepsilon < \infty$$
$$= \Pr[\varepsilon_1 = \varepsilon] \Pr[\varepsilon_2 < \varepsilon + V_1 - V_2]$$
$$= \int_{-\infty}^{\infty} \psi'(\varepsilon)\psi(\varepsilon + V_1 - V_2)d\varepsilon$$
$$= \int_{-\infty}^{\infty} \exp(\varepsilon) \underbrace{\psi(\varepsilon)\psi(\varepsilon + V_1 - V_2)}_{= y\text{と置く}} d\varepsilon$$

４．MNL

すると

$$y = \exp[-\exp(-\varepsilon)(1 + \exp(V_2 - V_1))]$$

$$\frac{dy}{d\varepsilon} = \mathrm{y}\exp(-\varepsilon)(1 + \exp(V_2 - V_1))$$

より

$$P_n(i) = \int_{-\infty}^{\infty} y\exp(\varepsilon)\,d\varepsilon$$

$$= \int_0^1 \frac{y\exp(\varepsilon)}{y\exp(\varepsilon)(1 + \exp(V_2 - V_1))}\,dy$$

$$= \left[\frac{y}{1 + \exp(V_2 - V_1)}\right]_0^1$$

$$= \frac{1}{1 + \exp(V_2 - V_1)}$$

$$= \frac{\exp(V_1)}{\exp(V_1) + \exp(V_2)}$$

４．MNL

多項ロジットの導出

$$\mathrm{P}_n(i) = \Pr\left[U_{in} > U_{jn}, j \in J_n, i \neq j\right]$$
$$= \Pr\left[V_{in} + \varepsilon_{in} > V_{jn} + \varepsilon_{jn}, j \in J_n, i \neq j\right]$$
$$= \Pr[\underbrace{V_{in} + \varepsilon_{in} > \max_j\left(V_{jn} + \varepsilon_{jn}\right)}, i \neq j]$$

i以外の選択肢の中で最大の効用を与えるものよりもiの効用が大きい.

ガンベル分布の性質2より

$\max_j U_{jn} = U_n^*$とすると

U_n^*は$\left(\frac{1}{\mu}\ln\sum_{j\in J_n}\exp\left(\mu V_{jn}\right), \mu\right)$ のガンベル分布に従う.

$U_n^* = V_n^* + \varepsilon_n^*$とし$V_n^* = \frac{1}{\mu}\ln\sum\exp(\mu V_{jn})$とおく.

$\Rightarrow\varepsilon_n^*$はパラメータ$(0, \mu)$のガンベル分布に従う.

4．MNL

$$
\begin{aligned}
P_n(i) &= \Pr[V_{in} + \varepsilon_{in} \geq V_n^* + \varepsilon_n^*] \\
&= \Pr[\varepsilon_{in} - \varepsilon_n^* \geq V_{in} - V_n^*] \\
&= \frac{1}{1 + \exp(\mu(V_n^* - V_{in}))} \\
&= \frac{\exp(\mu V_{in})}{\exp(\mu V_{in}) + \exp(\mu V_n^*)}
\end{aligned}
$$

ガンベル分布の
性質１より

$$
= \frac{\exp(\mu V_{in})}{\exp(\mu V_{un}) + \exp(\mu \cdot \frac{1}{\mu} \ln \sum_{j \neq i} \exp(\mu V_{jn}))}
$$

$$
= \frac{\exp(\mu V_{in})}{\exp(\mu V_{in}) + \sum_{j \neq i} \exp(V_{jn})}
$$

$$
= \frac{\exp(\mu V_{in})}{\sum \exp(\mu V_{jn})}
$$

４．MNL

IIA特性（Independence of Irrelevant Alternative)

- 「選択確率比の文脈独立」とも呼ばれる.
- 無関係な選択肢から選択確率が独立であること.
- 例えば$\frac{P_{in}}{P_{jn}} = \exp(V_{in} - V_{jn})$となり選択肢$i, j$の効用確定項のみから決まり，$i.j$以外の選択肢から影響を受けない.

○長所

選択肢集合に含まれる全ての選択肢ではなく，部分集合を用いて推定しても推定値にバイアスが生じない.

×短所

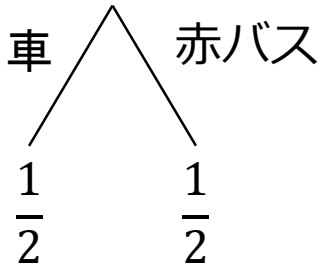
類似した選択肢が存在し，誤差項が独立であるという仮定が誤っているとき，類似した選択肢の選択確率が過大になってしまう.

＝＞赤バス・青バス問題

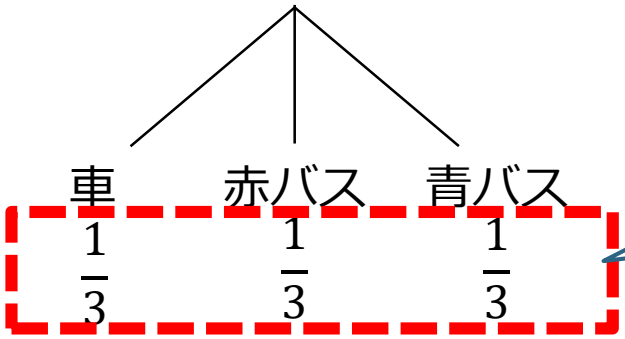
４．MNL

赤バス・青バス問題

車・赤バス：効用の確定項が全く同じ
が選択肢として存在する場合，選択確率は



青バス：車・赤バスと効用の確定項が全く同じ
を先ほどの選択肢に加えて導入すると，選択確率は

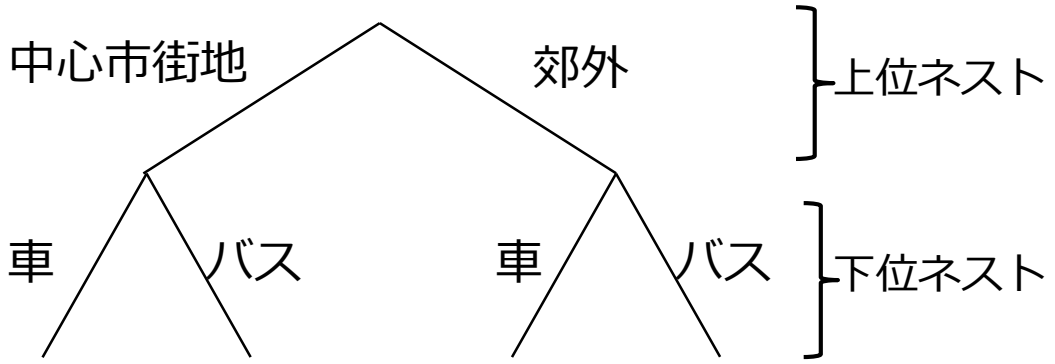


バス全体で効用は変化しないため
$\frac{1}{2}, \frac{1}{4}, \frac{1}{4}$
が正しいのでは？

5．NL

MNLのIIA特性を緩和したい.
→　すなわち，効用の誤差項に相関のありそうな場合について
　　考えたい

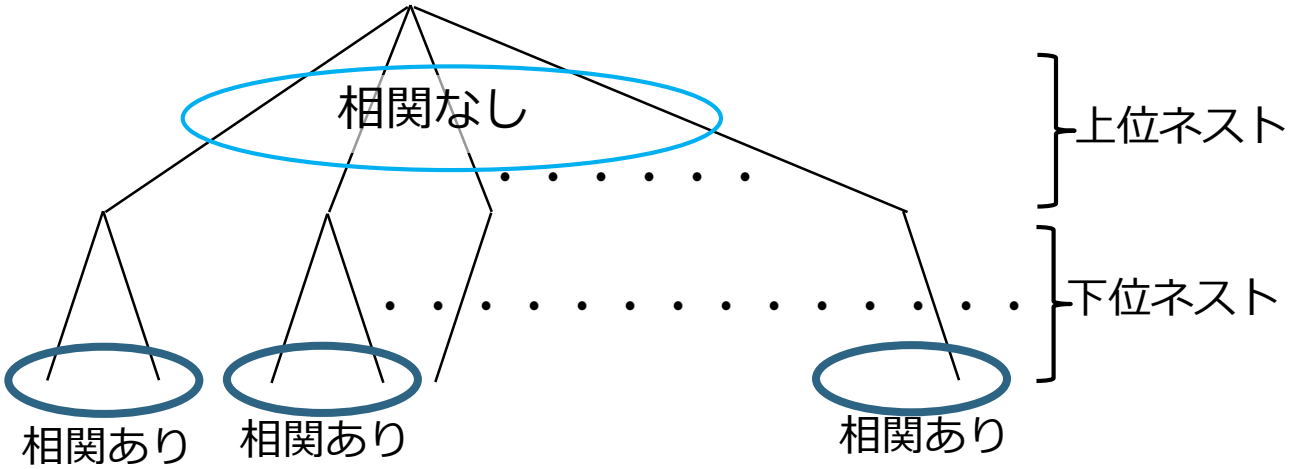
Ex)目的地と交通手段の組み合わせの選択問題



選択肢が｛市街地，車｝｛市街地，バス｝｛郊外，車｝｛郊外，バス｝
とすれば,
交通手段を選択するネストで誤差項の相関が生まれる.

→NL(Nested Logit)モデルの導入

5．NL



上位ネストの選択肢：　d
下位ネストの選択肢：　i
として下位ネストの選択肢間に誤差項の相関があると考える.

5．NL

選択肢 d i の組み合わせの効用は

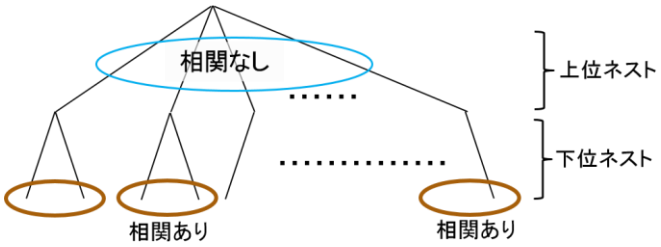
$$U_{di} = V_d + V_i + V_{di} + \varepsilon_d + \varepsilon_{di}$$

（個人を表す n の添え字は省略）

U_{di}：選択肢diの組み合わせによる効用
V_d ：選択肢dによる効用の確定項
V_i ：選択肢iによる効用の確定項
V_{di} ：選択肢diの組み合わせによる効用の確定項
ε_d ：選択肢dによる効用の確率項（$\max(U_{di})$がスケールパラメータμ_dを持つガンベル分布になるような分布に従うと仮定
ε_{di} ：　選択肢diによる効用の確率項（スケールパラメータμを持つ互いに独立なガンベル分布に従うと仮定

選択肢 di の選択確率は条件付き確率を用いて

$$P(d,i) = P(d|i) \cdot P(d)$$



上位ネストの選択肢：d
下位ネストの選択肢：i
として下位ネストの選択肢間に誤差項の相関があると考える.

5．NL

$$P(d) = \Pr\left[\max_i U_{di} > \max_i U_{d'i}, d' \neq i\right]$$

dを選んだ時の最大効用

d′を選んだ時の最大効用

$$= \Pr[V_d + \varepsilon_d + \max_i(V_i + V_{di} + \varepsilon_{di}) \geq V_{d'} + \varepsilon_{d'} + \max_i(V_i + V_{d'i} + \varepsilon_{d'i}), d' \neq d]$$

ε_{di}の分布の仮定により

$\max_i(V_i + V_{di} + \varepsilon_{di})$はスケールパラメータ$\mu$のガンベル分布に従う

⇒　ガンベル分布のロケーションパラメータV'_dとすると

$$V'_d = \frac{1}{\mu}\ln\sum(\exp(\mu(V_i + V_{di}))$$

ログサム変数

5．NL

$$P(d) = \Pr[V_d + V_d' + \varepsilon_d + \varepsilon_d' \geq V_{d'} + V_{d'}' + \varepsilon_{d'} + \varepsilon_{d'}']$$

$(\varepsilon_d' \equiv \max_i(V_i + V_{di} + \varepsilon_{di}) - V_d'$　とした)

⇒　確定項$V_d + V_d'$,誤差項$\varepsilon_d + \varepsilon_d'$

の離散選択問題

ε_d：　確率式内の不等式の左辺の確率項がスケールパラメータμ_dの
ガンベル分布に従うように仮定

⇒周辺確率$P(d)$は

$$P(d) = \frac{\exp(\mu_d(V_d + V_d')}{\sum_{d'} \exp(\mu_d(V_{d'} + V_{d'}')}$$

5．NL

$$P(i|d) = \Pr[U_{di} \geq U_{di'}, i' \neq i|d]$$
$$= \Pr[V_i + V_{di} + \underbrace{\varepsilon_{di}}_{} \geq V_{i'} + V_{di'} + \varepsilon_{di'}, i' \neq i]$$

スケールパラメータμの
ガンベル分布に従う

$$P(i|d) = \frac{\exp(\mu(V_i + V_{di}))}{\sum_i \exp(\mu(V_i + V_{di})}$$

5．NL

ここまでの$P(i|d)$, $P(d)$の結果を用いて

$$P(d,i) = P(i|d)P(d)$$

$$= \frac{\exp(\mu(V_i+V_{di}))}{\sum_i \exp(\mu(V_i+V_{di})} \cdot \frac{\exp(\mu_d(V_d+V_d'))}{\sum_d \exp(\mu_d(V_d+V_d'))}$$

６．まとめ

最尤法，ｔ値とモデルの関係（MNLを例に）

・MNLによって選択確率と効用の関係が表現された.
・データは各個人の選択結果と説明変数の値.
・求めたいのは説明変数に対するパラメータ.

個人ｎの選択肢iの効用確定項と選択確率を次のように表す.

$$V_{in} = \sum_{k=1}^{K} \theta_k X_{ink}$$

$$P_{in} = \frac{\exp(V_{in})}{\sum_{j \in J_n} \exp(V_{jn})}$$

６．まとめ

θ_kを求めるための尤度関数は次のようになる.

$$L(\theta_1, \ldots, \theta_K) = \prod_{n=1}^{N} \prod_{i \in I_n} P_{in}^{\delta_{in}}$$

δ_{in}は個人nがiを選択している場合１，そうでなければ０

対数尤度関数にすると

$$l(\theta_1, \ldots, \theta_K) = \ln L(\theta_1, \ldots, \theta_K)$$

$$= \sum_{n=1}^{N} \sum_{i \in I_n} \delta_{in} \ln(P_{in})$$

この式の最大化を考えてパラメータを求める.

６．まとめ

ｔ値を求めてパラメータの検定を行う.
→推定値θの推定標準偏差を求める必要がある.

推定値θの母分散共分散行列の推定値はヘッセ行列に推定値θを代入した値に等しくなることを利用する.

推定値θ'_kの母分散共分散行列

k番目の要素θ'_kの母分散$E(\theta'_k - \theta_k)^2$を$k$番目の対角要素

θ'_kとθ'_lの母共分散$E(\theta'_k - \theta_k)(\theta'_l - \theta_l)$を$(k, l)$要素とする行列

ヘッセ行列

$L(\theta_1, \dots, \theta_K)$に対して$\frac{\partial L}{\partial \theta_k \partial \theta_l}$を$(k, l)$要素に持つ行列