ロジット分析とプロビット分析

田中 勝人（一橋大学）

1．二値選択モデル (Binary choice model)

単回帰モデル

$$y_i = \alpha + \beta x_i + u_i$$

において，被説明変数が 0 あるいは 1 の値しか取らないとする．このとき，

$$P(y_i = 1) = p_i = P(u_i = 1 - \alpha - \beta x_i), \quad P(y_i = 0) = 1 - p_i = P(u_i = -\alpha - \beta x_i)$$

となる．したがって，誤差項 $\{u_i\}$ については，

$$\mathrm{E}(u_i) = 0 \Rightarrow\ 0 < p_i = \alpha + \beta x_i < 1, \quad \mathrm{V}(u_i) = (\alpha + \beta x_i)(1 - \alpha - \beta x_i) > 0$$

という条件が課せられることになり，通常の回帰モデルは使えない．

そこで，次のモデルを考える．

$$y_i^* = \alpha + \beta x_i + u_i \tag{1}$$

ここで，誤差項 $\{u_i\}$ は古典的な仮定をみたすとする．また，$\{x_i\}$ は説明変数であり，ここでは説明の簡単化のためにスカラーの場合を考えるが，ベクトルへの拡張も考えられる．他方，y_i* は連続的であるが観測不可能な潜在変数 (latent variable) で，実際に観測されるのは

$$y_i = \begin{cases} 1 & (y_i^* > 0 \text{ のとき}) \\ 0 & (y_i^* \le 0 \text{ のとき}) \end{cases}$$

である．y_i は 0 あるいは 1 を取る二値確率変数である．このとき，(1) のモデルを **二値選択モデル**と呼ぶ．選択肢の数を任意の有限個にしたモデルは，**離散選択モデル**（discrete choice model）と呼ばれる．

二値選択の例： 自宅所有・非所有，結婚・非婚，就業・非就業，インフレ・デフレ政策，夫婦の子供の有無，試合の勝敗，**…**

(1) のモデルに対して，次の確率（**選択確率**）を求めたい．

$$p_i = P(y_i = 1) = P(y_i^* > 0) = P(u_i > -\alpha - \beta x_i) = 1 - F(-\alpha - \beta x_i)$$

ここで，$F(z)$ は，誤差項 u の分布関数であり，対称性が仮定される．したがって，

$$p_i = P(y_i = 1) = 1 - F(-\alpha - \beta x_i) = F(\alpha + \beta x_i)$$

となる．以下，(1) のモデルの推定を考える．

2．尤度関数 (Likelihood function)

データ $(x_1, y_1), \cdots, (x_n.y_n)$ が与えられた場合の尤度関数は，

$$\begin{aligned} l(\alpha, \beta) &= \prod_{y_i=1} p_i \prod_{y_i=0} (1 - p_i) = \prod_{i=1}^{n} \left\{ p_i^{y_i} (1 - p_i)^{1-y_i} \right\} \\ &= \prod_{i=1}^{n} \left\{ F(\alpha + \beta x_i)^{y_i} (1 - F(\alpha + \beta x_i))^{1-y_i} \right\} \end{aligned} \tag{2}$$

で与えられる.
• 分布関数としてよく使われるのは，ロジスティック分布と正規分布である.

3. ロジット分析 (Logit analysis)

選択確率がロジスティック分布を使って,

$$p_i = P(y_i = 1) = F(\alpha + \beta x_i) = \frac{e^{\alpha+\beta x_i}}{1 + e^{\alpha+\beta x_i}} \tag{3}$$

で表現されると仮定する.

$$f(z) = F'(z) = \frac{e^z}{(1+e^z)^2} \qquad \text{ロジスティック分布の密度関数}$$

$$1 - p_i = \frac{1}{1 + e^{\alpha+\beta x_i}} \qquad \text{非選択確率}$$

$$\frac{p_i}{1-p_i} = e^{\alpha+\beta x_i} \qquad \text{オッズ比（odds ratio）}$$

$$\log \frac{p_i}{1-p_i} = \alpha + \beta x_i$$

4. プロビット分析 (Probit analysis)

選択確率が正規分布を使って,

$$p_i = P(y_i = 1) = F(\alpha + \beta x_i) = \int_{-\infty}^{\alpha+\beta x_i} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz \tag{4}$$

で表現されると仮定する.

5. ロジットとプロビットの比較

• プロビットの方が，古典的な回帰分析の応用として自然である.

• しかし，ロジットの方が関数形が簡単なので，推定も容易である.

•	ロジスティック分布	正規分布
密度	$\frac{e^z}{(1+e^z)^2}$	$\frac{1}{\sqrt{2\pi}}e^{-z^2/2}$
平均	0	0
分散	$\pi^2/3 = 3.29 \approx (1.81)^2$	1

（注） ロジスティック分布の特性関数は,

$$\phi(\theta) = \int_{-\infty}^{\infty} e^{i\theta z} \frac{e^z}{(1+e^z)^2}\, dz = \Gamma(1 - i\theta)\,\Gamma(1 + i\theta) = \frac{\pi\theta}{\sinh \pi\theta}$$

で与えられる (Gradshteyn-Ruzhik, p.353)．$\Gamma(z)$ はガンマ関数

$$\Gamma(z) = \int_0^\infty x^{z-1}\, e^{-x}\, dx$$

である．これより，

$$k \text{ 次のモーメント} = m_k = \frac{1}{i^k} \left. \frac{d^k\, \phi(\theta)}{d\theta^k} \right|_{\theta=0} = \begin{cases} 0 & (k \text{ が奇数のとき}) \\ (-1)^{j-1}(2\pi)^{2j} \frac{B_{2j}}{2j} & (k = 2j \text{ のとき}) \end{cases}$$

ここで，B_{2j} はベルヌーイ数である．ベルヌーイ数 B_k $(k = 0, 1, \cdots)$ は，

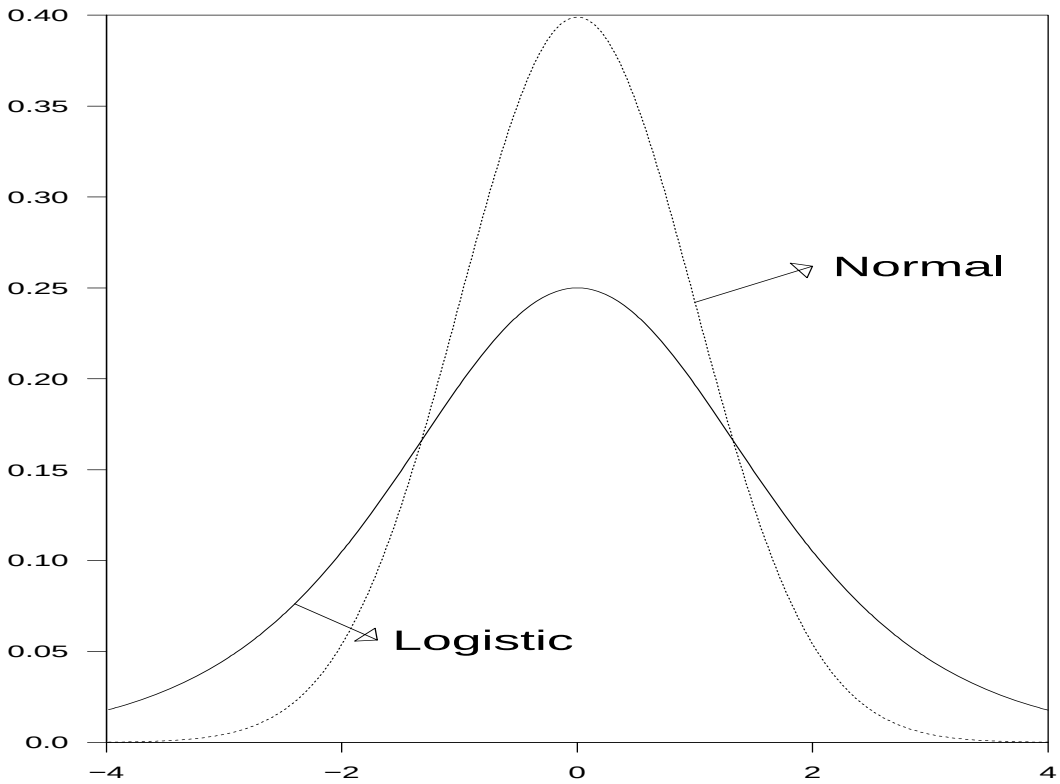
$$B_0 = 1, \quad \sum_{k=0}^{n} {}_{n+1}C_k\, B_k = 0 \quad (n = 1, 2, \cdots)$$

から得られる有理数であり，$B_1 = -1/2, B_2 = 1/6, B_3 = 0, B_4 = -1/30, \cdots$ となる．奇数次のベルヌーイ数は，B_1 以外はすべて 0 となる．

特に，分散は，$k = 2$ として，

$$\text{ロジスティック分布の分散} = m_2 = 4\pi^2 \times \frac{B_2}{2} = \frac{\pi^2}{3}$$

となる．



標準正規分布とロジスティック分布の密度関数

6. 推定 (Estimation)

データ $(x_1, y_1), \cdots, (x_n, y_n)$ に対して，対数尤度は，式 (2) から，

$$\begin{aligned} L(\alpha, \beta) &= \log l(\alpha, \beta) \\ &= \sum_{i=1}^{n} [y_i \log F(\alpha + \beta x_i) + (1 - y_i) \log \{1 - F(\alpha + \beta x_i)\}] \end{aligned}$$

で与えられる．$L(\alpha, \beta)$ を最大にする α と β，すなわち，

$$\max_{\alpha,\beta} L(\alpha, \beta) = L\left(\hat{\alpha}, \hat{\beta}\right)$$

となる $\hat{\alpha}$ と $\hat{\beta}$ が α と β の最尤推定量（MLE ）であり，次のことが成り立つ．

$$\begin{pmatrix} \sqrt{n}(\hat{\alpha} - \alpha) \\ \sqrt{n}(\hat{\beta} - \beta) \end{pmatrix} \quad \Rightarrow \quad \mathrm{N}(\mathbf{0}, \Omega^{-1})$$

ただし，

$$\Omega = \lim_{n\to\infty} \frac{-1}{n} \begin{pmatrix} \mathrm{E}\left(\frac{\partial^2 L}{\partial \alpha^2}\right) & \mathrm{E}\left(\frac{\partial^2 L}{\partial \alpha\, \beta}\right) \\ \mathrm{E}\left(\frac{\partial^2 L}{\partial \beta\, \alpha}\right) & \mathrm{E}\left(\frac{\partial^2 L}{\partial \beta^2}\right) \end{pmatrix} = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \begin{pmatrix} \frac{f_i^2}{F_i(1-F_i)} & \frac{f_i^2 x_i}{F_i(1-F_i)} \\ \frac{f_i^2 x_i}{F_i(1-F_i)} & \frac{f_i^2 x_i}{F_i(1-F_i)} \end{pmatrix}$$

ここで，$f_i = F'(\alpha + \beta x_i)$, $F_i = F(\alpha + \beta x_i)$.

• ロジスティック分布を仮定した場合と正規分布を仮定した場合では，当然のことながら，MLE の結果が異なる．しかし，5 節で述べた 2 つの分布の関係から，前者の推定値と標準誤差は，後者のほぼ $\pi/\sqrt{3} \doteq 1.81$ 倍となる．

• 尤度関数は，一般に concave．ただし，データに separation が起きる（$= x$ の値が低いグループと高いグループに対応する y の値が分離される）場合，フラットな尤度となり，推定が困難となる．

• 選択確率 $p_i = P(y_i = 1)$ に関する推定と検定については，次節で説明する．

7. 説明変数が複数個の場合

次の重回帰モデルを考える．

$$y_i^* = \boldsymbol{x}_i' \boldsymbol{\beta} + \varepsilon_i \tag{5}$$

ここで，$\boldsymbol{x}_i$ は定数項も含む p 次元ベクトル，ε_i は古典的な仮定をみたす誤差項である．他方，y_i^* は，単回帰の場合と同様に，観測不可能な latent variable で，実際には，2 値確率変数 y_i が観測される．

$$y_i = \begin{cases} 1 & (y_i^* > 0 \text{ のとき}) \\ 0 & (y_i^* \le 0 \text{ のとき}) \end{cases}$$

誤差項 $\{\varepsilon_i\}$ は対称な分布 F をもつと仮定すると，$\boldsymbol{\beta}$ の対数尤度関数は，

$$L(\boldsymbol{\beta}) = \sum_{n=1}^{n} [y_i \log F(\boldsymbol{x}_i' \boldsymbol{\beta}) + (1 - y_i) \log(1 - F(\boldsymbol{x}_i' \boldsymbol{\beta}))]$$

となり，$L(\boldsymbol{\beta})$ を最大にする値 $\hat{\boldsymbol{\beta}}$ が $\boldsymbol{\beta}$ の MLE となる．そして，中心極限定理

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \ \Rightarrow \ \mathrm{N}(\mathbf{0}, A^{-1})$$

が成立する．ここで，

$$A = \lim_{n\to\infty} \frac{-1}{n} \mathrm{E}\left(\frac{\partial^2 L}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\right) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \frac{f_i^2}{F_i(1-F_i)} \boldsymbol{x}_i \boldsymbol{x}_i'$$

- パラメータ $\boldsymbol{\beta}$ に関する検定

$$H_0 : R\boldsymbol{\beta} = \mathbf{0} \quad \text{vs.} \quad H_1 : \text{無制約} \quad (R \text{ のランクは } q)$$

尤度比検定を行うのが普通である．

$L_1(\hat{\boldsymbol{\beta}})$: 制約なしの最大対数尤度

$L_0(\tilde{\boldsymbol{\beta}})$: 帰無仮説のもとでの最大対数尤度

$$\lambda = 2 \times (L_1(\hat{\boldsymbol{\beta}}) - L_0(\tilde{\boldsymbol{\beta}})) \ \Rightarrow \ \chi^2(q) \quad (\text{帰無仮説のもとで})$$

- 選択確率 $p_i = P(y_i = 1$ の推定

$$\hat{p}_i = \hat{P}(y_i = 1) = F(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}) \ \sim \ \mathrm{N}(p_i, f_i^2 \boldsymbol{x}_i' A^{-1} \boldsymbol{x}_i)$$

- ロジット，プロビット分析の場合の決定係数は，

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{p}_i)^2}{\sum_{i=1}^{2} (y_i - \bar{y})^2}$$

で定義されるので，通常の回帰分析の場合よりは低めの値となる．

- 選択確率に関する検定

$$H_0 : \ p_i = p_i^0 \quad \text{vs.} \quad p_i \neq p_i^0$$

$$\frac{\hat{p}_i - p_i^0}{se(\hat{p}_i)} \ \sim \ \mathrm{N}(0,1) \quad (\text{帰無仮説のもとで})$$

ここで，$se(\hat{p}_i)$ は，$\hat{p}_i$ の標準誤差の推定値であり，

$$se(\hat{p}_i) = \sqrt{\hat{f}_i^2 \, \boldsymbol{x}_i' \hat{A}^{-1} \, \boldsymbol{x}_i}, \quad \hat{f}_i = f(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}})$$

で計算される．

8. モデルの例

- 既婚女性の就業行動

$$y^* = \beta_1 + \beta_2 x_1 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \varepsilon$$

$$y = \begin{cases} 1 & (y^* > 0\text{: 就労している}) \\ 0 & (y^* \leq 0\text{: 就労していない}) \end{cases}$$

x_2: 18 歳未満の子供の数　　x_3: 本人の年齢　　x_4: $= x_3^2$
x_5: 教育年数　　x_6: 夫の収入

9. 順序型ロジットと順序型プロビット（Ordered logit and ordered probit）

既婚女性の就業行動の分析において，就労形態が (i) 非就労 (ii) 短時間就労 (iii) 長時間就労 の 3 通りの場合を考える．これら 3 つの就労形態は，就労時間に関して順序付けされている．このような場合に，次のモデルを考えることができる．

$$y_i^* = \alpha + \beta x_i + u_i \tag{6}$$

$$y_i = \begin{cases} 0 & (y_i^* \leq 0 \text{ のとき}) \\ 1 & (0 < y_i^* \leq \mu \text{ のとき}) \\ 2 & (\mu < y_i^* \text{ のとき}) \end{cases}$$

このようなモデルを**順序選択モデル**（ordered choice model）という．この場合の選択確率は，次のようになる．

$$\begin{aligned} P(y_i = 0) &= P(y_i^* \leq 0) = P(\alpha + \beta x_i + u_i \leq 0) = F(-\alpha - \beta x_i) \\ P(y_i = 1) &= P(0 < y_i^* \leq \mu) = P(0 < \alpha + \beta x_i + u_i \leq \mu) \\ &= P(\alpha + \beta x_i + u_i \leq \mu) - P(\alpha + \beta x_i + u_i \leq 0) \\ &= F(\mu - \alpha - \beta x_i) - F(-\alpha - \beta x_i) \\ P(y_i = 2) &= P(\mu < y_i^*) = P(\mu < \alpha + \beta x_i + u_i) = 1 - F(\mu - \alpha - \beta x_i) \end{aligned}$$

- 尤度関数は，

$$l(\alpha, \beta, \mu) = \prod_{y_i=0} F(-\alpha - \beta x_i) \prod_{y_i=1} (F(\mu - \alpha - \beta x_i - F(-\alpha - \beta x_i)) \times \prod_{y_i=2} (1 - F(\mu - \alpha - \beta x_i)$$

となるので，これを最大にする MLE を求めればよい．

10. 多項ロジットと多項プロビット（Multinomial logit and multinomial probit）

選択に順序関係がない場合は，**多項選択モデル**（multinomial choice model）を考えることができる．

(例) 東京－大阪間の交通手段として，新幹線，飛行機，自動車を考えると，これらの選択の間には，特に順序関係がない．

今，ある事柄に関して，n 人の個人の各々が s 個の選択肢をもっているものとして，i 番目の個人 $(i = 1, \cdots, n)$ が 第 j の選択 $(j = 0, 1, \cdots, s-1)$ をした場合の効用を U_{ij} とする．そして，効用が

$$U_{ij} = \mu_{ij} + \varepsilon_{ij} \qquad (i = 1, \cdots, n;\ j = 0, 1, \cdots, s-1)$$

と表されるものとする．このとき，回帰モデルとして，

$$U^*_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\beta}_j + \varepsilon^*_{ij}$$

を考える．ここで，

$$U^*_{ij} = U_{ij} - U_{i0}, \quad \mu_{ij} - \mu_{i0} = \boldsymbol{x}'_{ij}\boldsymbol{\beta}_j, \quad \varepsilon^*_{ij} = \varepsilon_{ij} - \varepsilon_{i0}$$

であるが，実際には，観測不可能な U^*_{ij} の代わりに，y_i が観測される．例えば，$s = 3$ の場合には，次のようになる．

$$y_i = \begin{cases} 0 & (U^*_{i1} \le 0,\ U^*_{i2} \le 0 \text{ のとき}) \\ 1 & (U^*_{i1} \ge 0,\ U^*_{i1} \ge U^*_{i2} \text{ のとき}) \\ 2 & (U^*_{i2} \ge 0,\ U^*_{i2} \ge U^*_{i1} \text{ のとき}) \end{cases}$$

選択確率については，例えば．

$$\begin{aligned} P(y_i = 1) &= P(U^*_{i1} \ge 0,\ U^*_{i1} \ge U^*_{i2}) \\ &= P(\boldsymbol{x}'_{i1}\boldsymbol{\beta}_1 + \varepsilon^*_{i1} \ge 0,\ x'_{i1}\boldsymbol{\beta}_1 + \varepsilon^*_{i1} \ge \boldsymbol{x}'_{i2}\boldsymbol{\beta}_2 + \varepsilon^*_{i2}) \end{aligned}$$

となる．この値は，プロビットでは，

$$P(y_i = 1) = \int_{-a}^{\infty} \left\{ \int_{-\infty}^{b+x} f(x, y)\, dy \right\} dx$$

で計算される．ただし，$a = \boldsymbol{x}'_{i1}\boldsymbol{\beta}_1, b = \boldsymbol{x}'_{i1}\boldsymbol{\beta}_1 - \boldsymbol{x}'_{i2}\boldsymbol{\beta}_2$ であり，$f(x, y)$ は ε^*_{i1} と ε^*_{i2} の同時密度関数である．この積分計算は，選択肢の数 s が大きくなれば困難となる．

他方，ロジットでは，選択確率が

$$p_{ij} = P(y_i = j) = \frac{\exp(z_{ij})}{\sum_{k=0}^{s-1} \exp(z_{ik})}, \quad z_{i0} = 0, \quad z_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\beta}_j \ (j > 0)$$

と表されるものと仮定する．このとき，尤度関数

$$l(\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_{s-1}) = \prod_{y_i=0} p_{i0} \prod_{y_i=1} p_{i1} \cdots \prod_{y_i=s-1} p_{i,s-1}$$

を最大にすることにより，$\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_{s-1}$ の MLE が得られる．

• 多項ロジットの分析は容易であるが，選択確率は，IIA (Independence from Irrelevant Alternatives: 他の選択肢からの独立性）に基づいていることから，制約的である．他方，多項プロビットの選択確率は，IIA には基づかないので一般的であるが，計算は困難である．

11. 入れ子型ロジット（Nested logit）

類似性があるような選択肢の間では，IIA が成立しないと考えられる．例えば，東京－大阪間の交通手段として,「新幹線」,「A航空の飛行機」,「B航空の飛行機」から選択する場合，後者 2 つの選択肢は独立とは考えにくい．

• マクファーデン（D. McFadden）の方法は，誤差項の相関関係を考慮した分布（Gumbel7s type B extreme-value distribution と呼ばれる）を使って，多項ロジットの修正を行った．この場合の分析は，入れ子型ロジットと呼ばれる．詳しい説明は，縄田和満「Probit, Logit, Tobit」（『応用計量経済学　』牧，宮内，浪花，縄田共著，多賀出版，第 4 章に所収）を参照されたい．

12. ロジット，プロビット以外の選択モデル

• **トービット・モデル** (Tobit model)

被説明変数 y_i の取りうる値が連続的である点では，通常の回帰モデルと同じであるが，y_i が観測されるのは，ある条件をみたす場合に限る点が異なる．

（例） 耐久消費財への支出，既婚女性の労働時間，金融資産の保有量

このような場合のモデルとして，2 つの代表的なモデルがある．

(a) 途中打ち切り回帰モデル（censored regression model）

$$y_i^* = \boldsymbol{x}_i'\boldsymbol{\beta} + u_i, \quad y_i = \begin{cases} y_i^* & (y_i^* > 0 \text{ のとき}) \\ 0 & (y_i^* \leq 0 \text{ のとき}) \end{cases}$$

このモデルでは，被説明変数 y_i が観測されない（0 の場合）でも，説明変数 $\boldsymbol{x}_i$ は観測される．

(b) 切断回帰モデル（truncated regression model）

$$y_i^* = \boldsymbol{x}_i'\boldsymbol{\beta} + u_i, \quad y_i = y_i^* \ (y_i^* > 0 \text{ の場合})$$

このモデルでは，被説明変数 y_i が観測されない（0 の場合）と，説明変数 $\boldsymbol{x}_i$ も観測されない．

打ち切り回帰と切断回帰の尤度関数は異なる．したがって，$\boldsymbol{\beta}$ の推定量も異なることに注意．具体例等は，縄田「Probit, Logit, Tobit」を参照されたい．