平成 27 年度 中級計量経済学・応用計量経済学
講義ノート 5 二項選択モデル

被説明変数が 1 または 0 の値を取る 2 項変数である場合を考える。この場合、回帰関数は、被説明変数の予測値ではなく、被説明変数が 1 をとる確率を表していると解釈できる。この節では、線形確率モデル（線形回帰モデルのことである）、プロビットモデル、ロジットモデルを紹介し、また最尤法による推定を紹介する。

5.1 被説明変数が 2 項変数である場合

経済学の実証分析では、被説明変数が 2 項変数である場合が多い。
例:

- 耐久消費財（家、車等）を購入するかしないか。
- （既婚女性が）働くかどうか。
- 大学に進学するかしないか。
- 喫煙するかしないか。
- 住宅ローンの審査に通るかどうか。

注：説明変数が 2 項変数である事も多い。その場合は講義ノート 2 で取り扱っており、特別なモデルを考える必要はない。ただ係数の解釈には注意が必要である。
被説明変数が 2 項変数であるモデルを、2 項選択モデルという。

5.2 線形確率モデル

線形回帰モデルで被説明変数が 2 項変数であるときには、そのモデルはどのように解釈すればよいのだろうか。
まず線形回帰モデルは、被説明変数の条件付期待値 $E(y|\mathbf{X})$ を線形にモデル化したものである。もし、y が 2 項変数なら、$E(y|\mathbf{X}) = 1 * Pr(y = 1|\mathbf{X}) + 0 * Pr(y = 0|\mathbf{X}) = \Pr(y = 1|\mathbf{X})$ である。つまり、線形回帰モデルで被説明変数が 2 項変数の場合は、$\Pr(y = 1|\mathbf{X})$ という条件付確率を線形にモデル化したものと考えることができる。したがって、以下のような被説明変数が 2 項変数の線形回帰モデルを線形確率モデルと呼ぶ。

$$E(y|\mathbf{X}) = Pr(y = 1|\mathbf{X}) = \beta'\mathbf{X} \tag{1}$$

- X_1 の係数 β_1 は X_1 を一単位変化させた時に $y = 1$ の確率がどのくらい変化するかを表わす。
- 線形回帰モデルで使われた統計的手法は、線形確率モデルでも、特に問題なく使用することが可能である。また、線形回帰モデルにおける理論的結果もそのまま適用できる。
- ただ、R^2 はそれほど意味をもたない。たとえば、被説明変数は 2 つの値しかとらないので、すべての観察点が回帰曲線上にあることはありえず、$R^2 = 1$ であることはまずありえない。

線形確率モデルの問題点　$\mathbf{X}$ の定義域に制限がない場合、モデルから予測される確率は、0 以下になったり、1 以上になったりする。そのため、線形確率モデルを使用するとモデルからの結果の解釈が困難になる場合がある。また、2 項選択モデルを使用して予測確率を計算し、それを他の統計手法を適用するときの材料に使うこともよくあり、その場合、予測確率が 0 と 1 の間に入らないと、その統計手法が使えない場合もある。

5.3　プロビット回帰とロジット回帰

2 項変数が被説明変数になっている場合に、よく使われる非線形モデルである、プロビットモデルとロジットモデルを紹介する。これらのモデルでは、常にモデルから予測される確率が 0 と 1 の間に入る。

プロビットモデル　$\Phi(\cdot)$ を標準正規分布関数とすると、

$$\Pr(y=1|\mathbf{X}) = \Phi(\beta'\mathbf{X}) \tag{2}$$

として、$\Pr(y=1|\mathbf{X})$ をモデル化するのが、プロビットモデルである。

なお、どの説明変数をモデルに含めるかについては、線形回帰モデルのときに議論したように、欠落変数からのバイアスを避けることを目的として、考えるべきである。

例:

- 説明変数が一つの場合。

  $$\Pr(y=1|X) = \Phi(\beta_0 + \beta_1 X). \tag{3}$$

  仮に $\beta_0 = -2$ で、$\beta_1 = 3$ であるとする。すると $X = 0.4$ の時の $Y = 1$ である確率は,

  $$\Pr(y=1|X=0.4) = \Phi(-2 + 3 \times 0.4) = \Phi(-0.8) \approx 21.2\% \tag{4}$$

  となる。

- 説明変数が二つの場合。

  $$\Pr(y=1|X_1, X_2) = \Phi(\beta_0 + \beta_1 X_1 + \beta_2 X_2). \tag{5}$$

  仮に $\beta_0 = -1.6$、$\beta_1 = 2$ で $\beta_2 = 0.5$ であるとする。説明変数の値が $X_1 = 0.4$ で $X_2 = 1$ のときの確率は

  $$\Pr(y=1|X_1 = 0.4, X_2 = 1) = \Phi(-0.3) \approx 38\% \tag{6}$$

  である。

X を変化させたときの影響　プロビットモデルは非線形モデルである。したがって、X を変化させたときの影響は、X の係数の値だけではとらえることができない。変化前と変化後の予測確率の変化を計算する必要がある。X_1 を ΔX_1 だけ変化させたときの影響は

$$\begin{aligned} & \Pr(y=1|X_1+\Delta X_1, X_2, \ldots, X_k) - \Pr(y=1|X_1, X_2, \ldots, X_k) && (7) \\ = \; & \Phi(\beta_0 + \beta_1(X_1 + \Delta X_1) + \beta_2 X_2 + \cdots + \beta_k X_k) - \Phi(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k) && (8) \end{aligned}$$

と計算できる。ただこれより簡単な式で書けない。

また、限界効果を計算することも多い。限界効果は次のように定義できる。

$$\frac{\partial \Phi(\beta' \mathbf{X})}{\partial X_k} = \phi(\beta' \mathbf{X}) \beta_k. \tag{9}$$

限界効果を論文などで掲載するときは、$\mathbf{X}$ の平均あるいは中央値での限界効果を計算して掲載することが多い。

ロジットモデル　プロビットモデルは、標準正規分布関数を使用した。そのかわりに標準ロジスティック分布関数を使用した条件付確率モデルをロジットモデルという。そのモデルでは、$y = 1$ の条件付き確率は

$$\begin{aligned} \Pr(y = 1 | X_1, \ldots, X_k) &= \Lambda(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k) & (10) \\ &\equiv \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k)}} & (11) \end{aligned}$$

となる。

- ロジットモデルとプロビットモデルは非常に似通った推定結果を与えることが多い。
- 歴史的には、ロジットモデルは推定時の計算がプロビットモデルに比べて早いためによく使われてきた。ただ、現代の計算機では気にするほどの違いではない。
- モデルをもっと複雑にしていくと二つのモデルの利点には違いが出てくる。具体的には、2 項選択ではなく、多項選択モデルを考えるとき、プロビットを拡張すると、色々な選択肢間の相関を簡単に表現できるが、推定量の計算が難しくなり、ロジットモデルを拡張した場合は、選択肢間の相関に制約を加えないとうまくモデル化できないが、計算は比較的容易になる。

線形確率モデルとプロビットやロジットモデルとの比較　どのモデルを使うべきかについては、特に決まった答えがあるわけではない。どのモデルがもっともデータと整合的かを調べる統計手法は存在するが、それほど頻繁には使われていない。

ただ、Angrist and Pischke (2008) の意見は一考の価値がある。線形確率モデルを使って推定をした場合、たとえ真の条件付期待値が線形でなかったとしても、なんらかの意味のあるパラメーターを推定することができる。たとえば、回帰変数が 2 項変数の場合、OLS 推定量は平均処置効果 $(E(y|X = 1) - E(y|X = 0))$ に収束する。しかし、プロビットやロジットの場合、もしモデルが間違っている場合には推定結果の解釈が難しくなる。したがって、線形確率モデルを使用するべきということである。

ただし、さきに見たとおり、線形確率モデルでは、確率としては不適切な値を確率の予測値をしてしまう可能性がある。

5.4　プロビットモデルの推定

プロビットモデルは係数に関して非線形なモデルである。したがって、OLS で推定することはできない。ここでは最尤法での推定を考える。

なお、ロジットモデルの場合も同じように推定することができる。

最尤推定法 (MLE)　尤度関数とは、実際に得られたデータを観察する同時確率を未知の係数の関数としてみたものである。最尤推定量とは、尤度関数を最大化する係数の値である。

アイデア:

- 実際に観察されたデータを最も高い確率で生成するような係数の値を、推定値として選ぶ。

最初に、説明変数が存在しない場合を考えよう。$p = \Pr(y=1)$ とする。すると、$y_i, i = 1, \ldots, n$ という標本を観察する確率は、

$$\prod_{i=1}^{n} p^{y_i}(1-p)^{1-y_i} \tag{12}$$

である。この確率をパラメーター p に関する関数としてみなしたものが、尤度関数である。対数尤度関数は

$$L(p) = \sum_{i=1}^{n} \{y_i \log p + (1-y_i)\log(1-p)\} \tag{13}$$

である。尤度関数は積の形になっているが、対数尤度は和の形になっているので、数学的に取り扱いやすく、通常の統計理論では対数尤度を考える。なお、対数変換は単調なので、最大値をとるパラメーターの値は変化しない。p の最尤推定量は $L(p)$ の最大値をもたらす p の値であり、この例では、$\hat{p} = \bar{y}$ である。

次に、プロビットモデルを考える。プロビットモデルでは、$\Pr(y_i = 1|\mathbf{X}_i) = \Phi(\mathbf{X}_i'\beta)$ である。このとき、X_i について条件付けた対数尤度関数は

$$L(\beta) = \sum_{i=1}^{n} \{y_i \log \Phi(\mathbf{X}_i'\beta) + (1-y_i)\log(1-\Phi(\mathbf{X}_i'\beta))\} \tag{14}$$

である。なお、X_i の分布が、β に依存しないのなら、これを尤度として考えても問題はなく、通常そのように仮定する（X_i は弱外生であるという）。そして、$L(\beta)$ を最大化することによって、β の最尤推定量を得る。ただし、プロビットモデルの最尤推定量は明示的な式で書くことができない。したがって、尤度関数の最大化は数値計算で行う。多くの統計ソフトで自動的に行うことができる。

5.5　最尤推定量の漸近的性質

最尤推定量は一般に明示的な式で書くことができず、その漸近的性質を直接証明することができない。

ここでは、どのようにして最尤推定量の漸近的性質を求めるのかを、かなりいい加減にみていく。数学的に厳密な証明は、Hayashi (2000) をはじめとする上級計量経済学の教科書に出ている。

まず、最尤推定量の一致性は、尤度関数の極限を考え、それが真のパラメーターの値で最大化されることを示し、最大値をもたらす統計量はその関数の極限の最大値をもたらす値に収束することを証明することで示す。つまり、

$$\frac{1}{n}L(\beta) \to_p E(y_i \log \Phi(\mathbf{X}_i'\beta) + (1-y_i)\log(1-\Phi(\mathbf{X}_i'\beta))) \tag{15}$$

ということを示し、$\beta_0 = \arg\max E(y_i \log \Phi(\mathbf{X}_i'\beta) + (1-y_i)\log(1-\Phi(\mathbf{X}_i'\beta)))$ ということを証明する。すると、

$$\hat{\beta} = \arg\max \frac{1}{n}L(\beta) \to_p \arg\max E(y_i \log \Phi(\mathbf{X}_i'\beta) + (1-y_i)\log(1-\Phi(\mathbf{X}_i'\beta))) = \beta_0 \tag{16}$$

ということが証明できる。

次に最尤推定量の漸近分布の導出の仕方を紹介する。基本的な方針は、テーラー展開によって最尤推定量を線形化し、そして中心極限定理や大数の法則を使用するというものである。

まず、$g(\cdot)$ を $L(\cdot)$ の一次微分の関数であるとする。すると最尤推定量は次の式を満たす。

$$g(\hat{\beta}) = 0. \tag{17}$$

この式を真のパラメーターの値の周りでテーラー展開をすると、

$$g(\beta_0) + H(\beta_0)(\hat{\beta} - \beta_0) + small = 0 \tag{18}$$

となる。ここで $H(\cdot)$ は $L(\cdot)$ の 2 回微分の関数である。この方程式を書き直すと、

$$\sqrt{n}(\hat{\beta} - \beta_0) = -\left(\frac{1}{n}H(\beta_0)\right)^{-1}\frac{1}{\sqrt{n}}g(\beta_0) + small \tag{19}$$

となる。ここで、$H(\beta_0)$ と $g(\beta_0)$ は、i.i.d. の確率変数の和で書くことができる。$H(\beta) = \sum_{i=1}^{n} H_i(\beta)$ かつ $g(\beta) = \sum_{i=1}^{n} g_i(\beta)$ と定義する。すると

$$\sqrt{n}(\hat{\beta} - \beta_0) \to_d N\left(\mathbf{0}, (E(H_i(\beta_0)))^{-1} E(g_i(\beta_0)g_i(\beta_0)')\,(E(H_i(\beta_0)))^{-1}\right) \tag{20}$$

となり、漸近分布が導出できる。

ところで、次の情報等式と呼ばれる式が成立する。

$$-E(H_i(\beta_0)) = E(g_i(\beta_0)g_i(\beta_0)'). \tag{21}$$

なお、この等式は、真の値 β_0 の元でしか成立しない。情報等式を使うと、

$$\sqrt{n}(\hat{\beta} - \beta_0) \to_d N\left(\mathbf{0}, -(E(H_i(\beta_0)))^{-1}\right) = N\left(\mathbf{0}, \left(E(g_i(\beta_0)g_i(\beta_0)')\right)^{-1}\right) \tag{22}$$

となる。

$\hat{\beta}$ の漸近分散の推定は二つの方法で行うことができる。

- $-\left(\frac{1}{n}H(\hat{\beta})\right)^{-1}$.
- $\left(\frac{1}{n}\sum_{i=1}^{n} g_i(\hat{\beta})g_i(\hat{\beta})'\right)^{-1}$. これは、BHHH (Berndt, Hall, Hall and Hausman) 推定量と呼ばれる。H の式を導出するのが難しい場合にはこの推定量は有用であるが、H が明示的に出るような場合には、避けたほうがよい。

5.6 2 項変数が被説明変数の場合の当てはまりの良さの指標

先にみたとおり、R^2 や $\bar{R}^2$ は被説明変数が 2 項変数の場合には、あまり適切な指標とは言えない。

- 当たり率: もし、予測確率が 0.5 より上なら、$y = 1$ を予測したとし、予測確率が 0.5 より下なら、$y = 0$ と予測したとして、予測が当たった割合を考える。$\hat{p}_i = \Phi(\mathbf{X}_i'\hat{\beta})$ として、

$$R_{hitrate} = \frac{1}{n}\sum_{i=1}^{n}(1 - |y_i - \mathbf{1}(\hat{p}_i > 0.5)|) \tag{23}$$

- ただ、当たり率はその解釈が難しい可能性がある。例えば、ほとんどの観測値で、$y_i = 0$ となっている場合、モデルに関わらず $y = 0$ を予測することで高い当たり率を得ることができる。かわりに次の指標を考える。$\hat{p}$ を $y_i = 1$ である観測値の割合であるとする。$wr_0 = min(\hat{p}, 1-\hat{p})$ とする。そして次の指標を考える。

$$R_p^2 = 1 - \frac{1 - R_{hitrate}}{wr_0} \tag{24}$$

- 疑似 R^2: L_0 を切片を除くすべての係数が 0 であるという制約のもとで最大化した対数尤度関数の値とする。その値は、$n_1 = \sum_{i=1}^{n} y_i$ として

$$L_0 = n_1 \log(n_1/n) + (n - n_1)\log(1 - n_1/n), \tag{25}$$

と書ける。疑似 R^2 は

$$pseudoR^2 = 1 - \frac{1}{1 + 2(L(\hat{\beta}) - L_0)/n} \tag{26}$$

として定義する。

- McFadden の R^2:

$$McFaddenR^2 = 1 - \frac{L(\hat{\beta})}{L_0}. \tag{27}$$

この指標は、1 が上限であり、その上限を達成することができる。なお、この指標も疑似 R^2 と呼ばれることがある。統計ソフトによって疑似 R^2 の定義が異なる場合があるので、注意が必要である。

5.7 データの生成過程をプロビットモデルで表現できる経済学的モデル

確率的効用モデル　2 項選択モデルに、経済学的なモデルによる理論的背景を立てることもできる。大学に行くかどうかの意思決定問題を考えよう。もし、個人 i が大学に行くとすると、$y_i = 1$ とし、行かないなら $y_i = 0$ とする。次のような経済学的モデルを考える。

- もし、個人 i が大学に行くなら、次の効用を得るとする。

$$u_{1i} = \mathbf{X}_i'\beta_1 + \epsilon_{1i}. \tag{28}$$

ここで、$\mathbf{X}_i$ は大学に行くことによる効用に影響を与える観測できる変数とし、ϵ_{1i} は効用に影響を与える他の要因を表すとする。

- 大学に行かないことから得られる効用も次のように定義する。

$$u_{0i} = \mathbf{X}_i'\beta_0 + \epsilon_{0i}. \tag{29}$$

- ϵ_{1i} と ϵ_{i0} は、同時に正規分布に従っているとする。
- 人々は効用最大化によって行動しているとする（$u_{0i} < u_{1i}$ なら大学に行く）。

すると、$y_i = 1$ を観測する確率は、

$$\Pr(y_i = 1|\mathbf{X}_i) = \Pr(u_{1i} > u_{0i}|\mathbf{X}_i) = \Pr(\epsilon_{0i} - \epsilon_{1i} < \mathbf{X}_i'(\beta_1 - \beta_0)) = \Phi\left(\mathbf{X}_i'\beta^*\right) \tag{30}$$

である。ここで、$\beta^* = (\beta_1 - \beta_0)/\sqrt{var(\epsilon_{0i} - \epsilon_{1i})}$ である。

つまり、確率的効用モデルによって人々の行動を記述できるとすると、データ生成過程はプロビットモデルで記述できる。この結果は、プロビットモデルの係数を解釈する際に役に立つ。例えば、

- 係数の大きさそのものは、あまり意味がない。（効用はその大きさそのものには意味がないので。）
- 係数は、ある変数が二つの選択肢それぞれから来る効用に与える影響の「差」として解釈できる。

確率的効用モデルによる 2 項選択モデルの基礎付けは、近年広く使われている構造推定におけるモデル構築の基礎をなすものであり、非常に重要である。

潜在変数モデル　潜在変数モデルにおいても、データ生成過程をプロビットモデルで記述できる。$u_i \sim N(0,1)$ として

$$y_i^* = \mathbf{X}_i'\beta + u_i \tag{31}$$

と定義する。y_i^* それ自体は観察できないが、$y_i^* > 0$ であるかどうかは観測できるとする。$y_i = \mathbf{1}\{y_i^* > 0\}$ と定義すると、データ生成過程はプロビットモデルで記述できる。

- 潜在変数モデルは、効用の差それ自体をモデル化したものと考えることもできる。
- 潜在変数モデルが適切である例としては、所得は観測できないが、貧困状態にあるかどうかはわかる場合などである。

5.8　分散不均一性

プロビットモデルにおいて、分散不均一性の議論をする際には、注意が必要である。まず、プロビットモデルは分散不均一なモデルであるが、かなり制約の強い分散不均一性を考えている。被説明変数の平均と分散は、

$$E(y_i|\mathbf{X}_i) = \Phi(\mathbf{X}_i'\beta), \tag{32}$$

$$var(y_i|\mathbf{X}_i) = \Phi(\mathbf{X}_i'\beta)(1 - \Phi(\mathbf{X}_i'\beta)) \tag{33}$$

である。つまり分散は、$\mathbf{X}$ の値に依存している。

また、潜在変数モデルにおける分散不均一性も考えることができる。次の潜在変数モデルを考える。

$$y_i^* = \mathbf{X}_i'\beta + u_i. \tag{34}$$

このモデルがプロビットモデルをもたらすためには、誤差項 u_i は分散均一である必要がある。u_i が分散不均一であるような拡張は可能である。しかし、そうした拡張は色々な理論上や計算上の注意が必要なことが知られており、この講義では取り扱わない。

- もし u_i が分散不均一なら、通常のプロビットモデルを最尤推定しても係数の一致推定量は得られない。
- もし u_i が分散不均一なら、y_i の条件付き期待地は $\Phi(\mathbf{X}_i'\beta)$ ではなくなる。

なお、線形確率モデルであれば、分散不均一性は通常の線形回帰モデルと同様に取り扱うことができる。なお、常に分散不均一になるので、頑健な漸近分散推定量を使用する必要がある。