第1章　最尤法とは

ここでは最尤法の基本的な考え方を述べる．ここの説明は，おもに Hayashi (2000) *Econometrics*, Ch.7 による．ただし，最尤推定法に関係する簡単なところのみを扱う．式番号は Hayashi (2000) による．

1.1　Extremum Estimators （極値推定）

標本を用いて母集団の未知の特性値（パラメタ）の best guess を計算することを**推定**と呼んだ．基礎的な計量経済学では，線形関係を想定して，その係数を最小 2 乗法によって推定した．最小 2 乗法は残差の 2 乗和を最小にするような係数を推定値とする推定方法であるが，推定方法は最小 2 乗法に限らない．

推定すべきパラメタ（ベクトル）を θ と書こう．パラメタベクトルがとりうる範囲（**パラメタ空間**）を Θ と書くとき，Θ のなかで，θ のなんらかの関数 $Q_n(\theta)$ を最大化するような $\hat{\theta}$ をもって推定量とするような推定量を **extremum estimators** と呼ぶ．ここで，添え字の n はサンプルサイズが n であることを表しており，関数 $Q_n(\theta)$ は標本の関数である．標本に含まれるそれぞれの観測値のベクトルを w_i と書けば，extremum estimator は

$$\hat{\theta} \quad \text{maximizes} \quad Q_n(\theta; w_1, ..., w_n) \quad \text{subject to} \quad \theta \in \Theta$$

と特徴付けることができる．この最大化問題に解が存在しなければ推定はできないことになるが，ほとんどの応用において解の存在は仮定される（か，解が存在するための条件が満たされていると仮定される）．

2 つの Extremum Estimators

さまざまな推定量が extremum estimator の特殊形と位置づけられるが，ここではよく用いられる 2 種類だけに言及しておこう．ひとつは M 推定量（M-estimators），いまひとつが一般化積率法（GMM）である．最小 2 乗法も extremum estimator の特殊形のひとつである．

一般化積率法（GMM）では，なんらかの意味で定義された「距離」を最小化して推定する．通常の「距離」が，ベクトルの各要素の 2 乗和で定義されたことを思い出そう．標

本の K 値関数として $g_n(w_1, ..., w_n; \theta)$ が最小化されるべきベクトルであり，目的関数は

$$Q_n(\theta; w_1, ..., w_n) = \frac{1}{2} g_n(w_1, ..., w_n; \theta)' \hat{W} g_n(w_1, ..., w_n; \theta)$$
$$\text{with} \quad g_n(w_1, ..., w_n; \theta) \equiv \frac{1}{n} \sum_{i=1}^{n} g(w_i; \theta) \tag{7.1.3}$$

と書ける．他方，**M 推定量（M-estimators）**では，各観測値 w_i とパラメタ θ の実数値関数 $m(w_i; \theta)$ を最大化して推定する．すなわち，目的関数は

$$Q_n(\theta; w_1, ..., w_n) = \frac{1}{n} \sum_{i=1}^{n} m(w_i; \theta) \tag{7.1.2}$$

である．

最尤推定

最尤推定量（ML: Maximum Likelihood estimator）は M 推定量の代表例である．標本に含まれる観測値 w_i が互いに独立に同一の分布に従う（i.i.d.: independently and identically distributed）とし，その分布が有限のパラメタベクトル θ で表現できるとしよう．w_i の確率密度関数が $f(w_i; \theta)$ と表され，$f(.,.)$ の関数形は既知とする．w_i が i.i.d. だから，$f(.,.)$ の関数形は i に依存しない．θ の真の値を θ_0 と書けば，w_i が観測される真の確率密度は $f(w_i; \theta_0)$ である．$\theta_0 \in \Theta$ のとき，モデルは correctly specified である，という．

標本 $(w_1, w_2, ..., w_n)$ が観測される同時確率密度（joint density）は，w_i が i.i.d. だから，

$$f(w_1, w_2, ..., w_n; \theta) = \prod_{i=1}^{n} f(w_i; \theta_0) \tag{7.1.4}$$

である．さて，手許にある標本は「最も起きやすい状況」が起きた結果，と考えることは推定の発想として自然なものだろう．そのためには，関数形 f が分かっているのだから，標本 $(w_1, w_2, ..., w_n)$ を固定しておいて，パラメタベクトル θ を動かして，この同時確率密度を最大にするような θ を見つければよい．同時確率密度をパラメタベクトル θ の関数と捉え直すとき，この関数を尤度関数（likelihood function）と呼ぶ．対数変換は単調変換であるから，尤度関数を最大化する θ と，対数変換した尤度関数を最大化する θ は一致する．すなわち，最尤推定量とは，次の**対数尤度関数（log likelihood function）**を最大化する．

$$\log f(w_1, w_2, ..., w_n; \theta) = \sum_{i=1}^{n} \log f(w_i; \theta_0) \tag{7.1.5}$$

ここで，

$$m(w_i; \theta) = \log f(w_i; \theta_0) \tag{7.1.6}$$

と考えれば，最尤推定量が M 推定量の特殊形であることが分かるだろう．

ここでは，観測値が i.i.d. であるという仮定をおいて，尤度の最大化問題を個別の対数尤度の和の最大化問題として設定したが，観測値が i.i.d. でなければこのような変形はできない．たとえば観測値のあいだに時系列的・空間的な自己相関があるときには単純な個々の対数尤度の和の最大化問題として推定を行うことはできない．しかしそのような場合でも，自己相関のパターンが特定化できれば尤度を構成することはでき，最尤推定を行うことはできる．

一般に，最尤推定は対数尤度の和の最大化問題を収束計算によって解く．しかし，標本が与えられたとき，最尤推定のみが推定方法ではない．観測値のなんらかの関数の期待値がパラメタベクトル θ で表現できれば，GMM 推定を行うこともできる．

実際の (?) 推定では，対数尤度の最大化問題の収束計算までプログラミングする必要はない．Stata のような統計計量アプリケーションでは，最尤推定法の特殊例であるいくつかの（かなりの?）推定についてはコマンドラインが用意されていることが多いし，そのようなものがなくても，対数尤度を指定できれば最大化問題はお任せできる．もちろん，構造推定のようなややこしいものはそのかぎりではない．

条件付き最尤推定法

ここまで，観測値を表すベクトルをまとめて w_i と呼んできた．しかし，ほとんどの応用では，最小 2 乗推定のときのように，被説明変数 y_i と説明変数 x_i を考え，説明変数の変化が被説明変数の条件付き分布に与える効果を検討している．最尤推定法においては，ベクトル w_i を 1 つの被説明変数 y_i と複数の説明変数 x_i に分割して考える必要性は必ずしもないが，被説明変数と説明変数に分けて考えるほうが簡単なことも多い．

そこで，説明変数 x_i の条件付きの被説明変数 y_i の確率密度関数を $f(y_i|x_i;\theta)$ とし，説明変数ベクトル x_i の周辺密度関数を $f(x_i;\psi)$ と書こう．すると，条件付き確率密度の関係式から，

$$f(y_i, x_i; \theta, \psi) = f(y_i|x_i;\theta) f(x_i;\psi) \tag{7.1.9}$$

となる．いま，θ と ψ に関係がないとすると対数尤度は，

$$\sum_{i=1}^{n} \log f(w_i;\theta,\psi) = \sum_{i=1}^{n} \log f(y_i|x_i;\theta) + \sum_{i=1}^{n} \log f(x_i;\psi) \tag{7.1.10}$$

と分解できる．θ の値だけに興味があるときには，右辺の第 2 項の最大化問題の解が第 1 項の解に関係しないかぎり，第 2 項のことを考えなくてもよい．つまり，

$$m(w_i;\theta) = \log f(y_i|x_i;\theta_0) \tag{7.1.11}$$

とおいた M 推定量を考えればよい．

1.2 一致性

最尤推定量は，関数形の特定化が正しければ，一般的な条件のもとで一致性を持つ．しかし，θ を動かしても尤度が変化しないような妙な状況では，一致性は保証されない．真の θ_0

の近くで θ を動かすと尤度が変化する，という条件は identification 条件であり，Kullback-Leibler 情報不等式である．

1.3 漸近正規性

尤度関数は標本の関数だから，他の推定量と同じく，最尤推定においても標本が異なれば異なる推定値が得られる，すなわち最尤推定量は確率変数である．したがって，最尤推定量は分散を持つし，検定を行うこともできる．その準備として，推定量の漸近分布について考えよう．うすうす想像されるとおり，証明は中心極限定理の応用であり，推定量は一致性を持ち，漸近的に正規分布に従う．

M 推定量が最大化している目的関数は，

$$Q_n(\theta; w_1, ..., w_n) = \frac{1}{n}\sum_{i=1}^{n} m(w_i; \theta) \tag{7.3.1}$$

である．あとで使うので，ここで関数 $m(w_i;\theta)$ の 1 次微分と 2 次微分を導入しておこう．関数 $m(w_i;\theta)$ はベクトル θ の関数なので，θ のそれぞれの要素で偏微分したものを並べたベクトルを考えることができ，これを 1 次微分した関数とみなす．推定すべきパラメタが p 個，すなわち，ベクトル θ の次元が p であるとすると，1 次微分は p 次ベクトルで表される．このベクトルを $s(w_i;\theta)$ とおくと，

$$s(w_i;\theta) = \frac{\partial m(w_i;\theta)}{\partial \theta} \tag{7.3.2}$$

と書ける．このベクトル $s(w_i;\theta)$ を観測値 i についてのスコアベクトル（score vector）と呼ぶ．さて，この p 個ある関数のそれぞれを θ のそれぞれの要素で偏微分したものを並べた $p \times p$ の行列を考えることができる．一般に，実数値関数の 2 階微分行列をヘッシアン（Hessian）と呼ぶ．ここでは，

$$H(w_i;\theta) = \frac{\partial s(w_i;\theta)}{\partial \theta'} = \frac{\partial^2 m(w_i;\theta)}{\partial \theta \partial \theta'} \tag{7.3.3}$$

と書き，これを観測値 i についてのヘッシアン行列，あるいは**情報行列**と呼ぶ．

さて，目的関数 $Q_n(\theta)$ が 2 階連続微分可能であるとすると，最大化のための 1 階の必要条件は，微分してゼロ，だから，

$$\frac{\partial Q_n(\hat{\theta})}{\partial \theta} = \frac{1}{n} s(w_i;\hat{\theta}) = 0 \tag{7.3.4}$$

である．この式はベクトル θ の各要素で目的関数を偏微分して得られる p 本の連立方程式を表していることに注意しよう．他方，真の値 θ_0 と推定値 $\hat{\theta}$ のあいだのある値 $\bar{\theta}$ に対して中間値の定理が成り立ち，

$$\frac{\partial Q_n(\hat{\theta})}{\partial \theta} = \frac{\partial Q_n(\theta_0)}{\partial \theta} + \frac{\partial^2 Q_n(\bar{\theta})}{\partial \theta \partial \theta'}(\hat{\theta} - \theta_0)$$

が成り立つから，1 次微分と 2 次微分を代入してみると，

$$\frac{\partial Q_n(\hat{\theta})}{\partial \theta} = \frac{1}{n}\sum_{i=1}^{n} s(w_i;\theta_0) + \left[\frac{1}{n}\sum_{i=1}^{n} H(w_i;\bar{\theta})\right](\hat{\theta}-\theta_0) \tag{7.3.5}$$

最大化のための 1 階の条件から左辺はゼロだから，

$$\left[\frac{1}{n}\sum_{i=1}^{n} H(w_i;\bar{\theta})\right](\hat{\theta}-\theta_0) = -\frac{1}{n}\sum_{i=1}^{n} s(w_i;\theta_0)$$

ヘッシアンの平均値である左辺のカッコ内が逆行列を持つとすれば，その逆行列を左からかけて，両辺に $\sqrt{n}$ をかけると，

$$\sqrt{n}(\hat{\theta}-\theta_0) = -\left[\frac{1}{n}\sum_{i=1}^{n} H(w_i;\bar{\theta})\right]^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} s(w_i;\theta_0)$$

観測値が i.i.d. で，最大化の 1 階の条件が期待値で満たされている，

$$E\left[s(w_i;\theta_0)\right] = 0$$

のとき，Lindeberg-Levy の中心極限定理が成り立ち，

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} s(w_i;\theta_0) \xrightarrow{d} N(0,\Sigma), \quad \text{where} \quad \Sigma = E\left[s(w_i;\theta_0)s(w_i;\theta_0)'\right] \tag{7.3.12}$$

となるから，

$$\sqrt{n}(\hat{\theta}-\theta_0) \xrightarrow{d} N(0,\mathrm{Avar}(\hat{\theta})), \quad \text{where} \quad \mathrm{Avar}(\hat{\theta}) = -\left\{E\left[H(w_i;\theta_0)\right]\right\}^{-1} \tag{7.3.13}$$

を得る．つまり，最尤推定量の漸近的な分散共分散行列は $\mathrm{Avar}(\hat{\theta})$ で与えられる．ヘッシアン H はスコアベクトル s を微分したものだったから，いくつかのテクニカルな条件のもとで

$$-E\left[H(w_i;\theta_0)\right] = E\left[s(w_i;\theta_0)s(w_i;\theta_0)'\right]$$

が成り立つ．したがって，推定量の分散共分散行列の推定方法は 2 つある．ひとつはヘッシアンをそのまま推定するもので，

$$-\left\{\frac{1}{n}\sum_{i=1}^{n} H(w_i;\hat{\theta})\right\}^{-1} \tag{7.3.14}$$

を計算する方法である．いまひとつは，情報行列についての等式を用いて，

$$\left\{\frac{1}{n}\sum_{i=1}^{n} s(w_i;\hat{\theta})s(w_i;\hat{\theta})'\right\}^{-1} \tag{7.3.15}$$

を計算するものである．どちらがいいということはないが，尤度関数の 1 次微分が解析的に求めるのが困難であるときには，2 次微分を必要としない方法のほうが計算が簡単であろう．

1.4 有効性

最尤推定量は，一致性を持ち漸近的に正規分布に従う推定量のうちで最も分散が小さくなるとは限らない．

一般に対数尤度の 1 次微分をスコア（score）と呼び，

$$s(\theta) = \frac{\partial \log L}{\partial \theta} \tag{1.5.9}$$

と書く．また，

$$I(\theta) = E[s(\theta)s(\theta)'] \tag{1.5.10}$$

を情報行列（information matrix）と呼び，いくつかの条件のもとで対数尤度のヘッシアンのマイナスに等しい．すなわち，

$$I(\theta) = -E\left[H(\theta)\right] = -E\left[\frac{\partial^2 \log L}{\partial\theta\partial\theta'}\right] \tag{1.5.11}$$

が成り立つ．さて，一般に，不偏推定量の分散には，情報行列を用いて，

$$\mathrm{var}(\hat{\theta}) \geq I(\theta)^{-1}$$

が成り立つ．この不等式を**クラメール・ラオの不等式（Cramer-Rao inequality）**と呼び，右辺を**クラメール・ラオの下限（Cramer-Rao lower bound）**と呼ぶ．

1.5 検定

最尤推定量は確率変数だから，母集団が同じであっても標本が異なれば異なる推定値が得られる．最尤推定値の散らばりぐあいは標準誤差（分散共分散行列）によって評価され，その情報を用いて，最小 2 乗法のときと同じく，仮説検定を行うことができる．最尤推定のときによく用いられる検定は Wald 検定，尤度比検定（LR: Likelihood Ratio），ラグランジュ乗数検定（LM: Lagrange Multiplier）である．いずれも，帰無仮説が正しいときにはゼロになる計算式を用いており，漸近的に χ^2 分布に従う．自由度 m の χ^2 分布とは，m 個の独立した標準正規分布に従う確率変数の 2 乗和であることを思い出そう．帰無仮説が正しいときに r 次のベクトルがゼロになるはずであれば，標本が十分に大きければ，その 2 乗和（のようなもの）は χ^2 分布（のようなもの）に従うだろう，というのが検定の発想となる．

Stata のばあい，最尤推定のあとの `test` コマンドでは Wald 検定が，`lrtest` コマンドでは尤度比検定が行われる．3 つの検定量のいずれも漸近的に χ^2 分布に従い，その値の差は帰無仮説のもとでは標本サイズが大きくなるにつれてゼロに確率収束するから，どの検定を使うのが望ましい，ということはない．

帰無仮説は非線形でもよいが，ここでは線形の帰無仮説を考える．推定した p 次のパラメタを $\hat{\theta}$，検定したい帰無仮説を $r \times p$ の行列 R を用いて

$$R\theta = 0$$

と書こう．両辺は $r \times 1$ のベクトルになるから，この式は r 本の帰無仮説を同時に表している.「ある係数の値がゼロだ」といった，しばしば用いる検定では $r = 1$ である．行列 R は帰無仮説を指定しているから，既知の行列である．

Wald 検定は，帰無仮説 $R\theta = 0$ そのものを用いる．帰無仮説が真であれば，推定されたパラメタ $\hat{\theta}$ について $R\hat{\theta}$ もまたゼロの近くにいる．もちろん標本誤差があるのでゼロに等しくはならないが，ゼロからの距離は小さい．$R\hat{\theta}$ とゼロの距離は各要素の 2 乗和で表されるから，適当に共分散行列で標準化すれば，自由度 r の χ^2 分布に従う．

尤度比検定（LR: Likelihood Ratio）は，読んで字の如く，尤度の比（対数尤度の差）を用いる．すぐあとで述べるように，帰無仮説を制約条件として用いた尤度推定を行うことができる．そのようにして推定されたパラメタを $\tilde{\theta}$ と書こう．$\tilde{\theta}$ は，その推定方法から明らかなように，帰無仮説の条件を満たす．このときに最大化された尤度を $L(\tilde{\theta})$ としよう．他方，制約条件をおかない尤度推定によって得られた推定値をこれまでどおり $\hat{\theta}$ とし，最大化された尤度を $L(\hat{\theta})$ としよう．もし帰無仮説が正しければ，制約をおかずに推定された $\hat{\theta}$ と，制約をおいて推定された $\tilde{\theta}$ は同じような値になるはずだし，最大化された尤度 $L(\tilde{\theta})$, $L(\hat{\theta})$ も同じような値になるはずだろう．つまり，尤度の比 $L(\hat{\theta})/L(\tilde{\theta})$ は 1 になるはずであり，対数尤度の差 $\log L(\hat{\theta}) - \log L(\tilde{\theta})$ はゼロになるはずである．この対数尤度の差に $2n$ を乗じたものが LR 検定統計量であり，自由度 r の χ^2 分布に従う．

ラグランジュ乗数検定（LM: Lagrange Multiplier）は，制約条件つきの推定でのラグランジュ乗数を用いる．制約条件つきの最尤推定は，制約条件のもとでの対数尤度関数の最大化によって行う．つまり，

$$\max_{\theta} \log L(\theta) \quad \text{subject to} \quad R\theta = 0$$

という最大化問題を解けばよい．この最大化問題をラグランジュ乗数法によって解くことを考えよう．ラグランジュ乗数を γ とおく．制約条件が r 個あるのだから，γ もまた r 次のベクトルである．もし帰無仮説（=制約条件）が真であれば，制約条件をつけた対数尤度の最大化問題も，制約条件のない対数尤度の最大化問題も同じ解を与えるはずである．よって，制約つき最大化問題において制約条件は制約になっていない．したがって，そのラグランジュ乗数 γ はゼロに等しい．標準誤差があるから，適当に標準化すれば，自由度 r の χ^2 分布に従う LM 検定統計量を作ることができる．

Stata のような統計計量アプリケーションを用いた実際の応用では，推定されたパラメタがゼロに等しいという帰無仮説については，最小 2 乗法のときと同様に，検定統計量（z 値と書かれることが多い）と有意水準が自動的に出力される．その解釈は最小 2 乗法のときの t 統計量と同じと考えてよい．

第2章　離散選択モデル

ここでは代表的な離散選択モデルを解説する．Wooldridge (2002) *Econometric Analysis of Cross Section and Panel Data*, Ch.15 も参照せよ．ここで取り扱う内容は (15.1, 15.2, 15.3, 15.4, 15.5, 15.6, 15.9) の簡単なところのみである．式番号は Wooldridge (2002) による．

2.1　離散選択モデル

被説明変数が幾つかの限られた値を取るような状況では**離散選択（離散反応）モデル**（discrete choice, discrete response）が用いられる．もちろん厳密に言えば，所得や消費のようなデータでも整数の値しか取らないから離散的ではあるが，通常は多くても 10 程度の選択肢からひとつが選ばれるような状況を考える．離散選択モデルのなかでも，選ばれた値自身にはとくに意味のないモデルのことを**質的変数モデル**（qualitative variable）ともいう．離散選択モデルと質的変数モデルは同じものを指すようであるが，離散選択モデルには「回数」等の値に意味があるケースも含まれる．被説明変数が「Yes」「No」の 2 種類の値しか取らないケースは離散選択モデルの典型であり，とりうる値が 2 個の被説明変数をとくに 2 値変数（binary variable）とも呼ぶ．

離散選択モデルにもさまざまなモデルや推定法があるが，ここではそれぞれの離散選択が行われる確率を考え，その確率を最大にするようなパラメタを最尤法で推定するモデルに限定する．尤度あるいは条件付き尤度をそれぞれのモデルのもとで設定すれば，あとは最尤法によって推定が行われる．例として 2 値変数モデルを考えよう．被説明変数を y_i，説明変数ベクトルを x_i とする．y_i は 0 か 1 の値のみをとるとしよう．このとき，y_i が 1 となる確率を考え，

$$p(x_i) \equiv P(y_i = 1|x_i) \tag{15.1}$$

と書く．もちろん，y_i が 0 となる確率

$$P(y_i = 0|x_i) = 1 - P(y_i = 1|x_i) = 1 - p(x_i)$$

と書くことができる．ここで，確率分布関数 $P(.)$ の関数形を特定化できれば尤度を構成することができる．

離散選択モデルでしばしば興味の対象となるのは，説明変数 x_i の値の変化が被説明変数 y_i がある値を取る確率をどれほど変化させるかである．被説明変数自体は離散的な値しか

取らないし，質的変数モデルではその数値自体には意味がないから，期待値の解釈は難しいことに注意しよう．2 値変数モデルのばあい，説明変数の確率への変化は

$$\frac{\partial P(y_i = 1|x_i)}{\partial x_{ji}} = \frac{\partial p(x_i)}{\partial x_{ji}} \tag{15.2}$$

と表現できる．この値のことを**限界効果**（marginal effect）と呼ぶ．2 値変数モデルにおいてもこの限界効果を直接推定できることはあまりなく[1]，推定されたパラメタの値から一定の仮定をおいて計算することが多い．逆に言えば，モデルはしばしば非線形であるために，推定されたパラメタの値そのものが解釈しやすい意味を持つことはあまりなく，その符号，あるいは一定の仮定のもとでさらに計算された値を解釈することになる．

離散選択モデルの推定にあたっては，連続変数としての**潜在変数**（latent variable）を想定すると便利なことがある．すなわち，潜在変数がある**範囲**の値を取れば質的変数がある値を取ると考える．また，潜在変数と説明変数の関係には，最も簡単なケースでは線形の関係を仮定し，その係数を推定する．係数の推定には，Wald，LR，LM 検定が用いられる．

ここでは，離散選択モデルのうちでも基本的な probit モデルと logit モデルを取り扱う．まず probit モデルを説明したのち，その応用としての順序 probit モデルを扱う．次に logit モデルと，その応用としての多項 logit，nested logit モデルを扱う．最後に，潜在変数の応用として区間回帰モデルにふれる．順序モデルや多項モデルについては，順序 logit や多項 probit も考えられるが，説明の簡単さのためにそれらは省略する．

2.2 Probit モデル

Probit モデルでは，確率分布関数 $P(.)$ として正規分布を用いる．平均ゼロ，分散 1 の標準正規分布の分布関数を $\Phi(.)$，確率密度関数を $\phi(.)$ で表す．すなわち，

$$\Phi(z) = \int_{-\infty}^{z} \phi(v)dv \tag{15.10}$$

$$\phi(z) = (2\pi)^{-1/2} \exp(-z^2/2) \tag{15.11}$$

であり，この分布関数・密度関数について

$$\phi(z) = \Phi'(z), \quad \phi(z) = \phi(-z), \quad \phi'(z) = -z\phi(z)$$

が成り立つことが知られている．第 1 式は密度関数の定義，第 2 式は正規分布の対称性による．第 3 式は密度関数を微分すれば容易に求まる．

単純な probit モデル

単純な probit モデルとは，ここでは被説明変数が 2 値変数である 2 項選択モデルをいう．観測される変数 y_i は 0 か 1 の値しか取らない．ここで，0 と 1 という数値自体には意味が

[1] 2 値変数モデルのうち，**線形確率モデル**（linear probability model）では，推定された係数がほぼ限界効果に対応する．ただし，線形確率モデルは，当てはめ値が 0 と 1 のあいだにおさまらなくなる可能性があることなどから，あまり用いられていないようである．

ない．-1 と 3 の 2 種類の値しかとらない，と書いても議論はほとんど変わらないが，単に分かりにくくなるだけだろう．さて，観測される変数 y_i に対応して，観測されない潜在変数 y_i^* を考える．潜在変数 y_i^* は観測される変数 y_i が 1 をとるとりやすさの指標であり，潜在変数がある範囲の値を取れば質的変数が 1 となる，という関係にあるとする．ここでは，潜在変数が正の時には質的変数が 1，負の時には 0 の値を取るとしよう．すなわち，

$$\begin{aligned} &y_i = 1 \quad y^* > 0 \text{ のとき} \\ &y_i = 0 \quad y^* \leq 0 \text{ のとき} \end{aligned} \tag{15.9}$$

と書ける．もちろん，観測される変数の値が 0 から 1 へ変わる値（**閾値** threshold）はゼロでなくてもよいが，これをゼロ以外に設定しても定数項以外には影響しない．すなわち，説明変数から定数項を除外して，閾値を推定するという方法も考えられるが，係数についての結果は同じになる．

潜在変数は任意の実数値を取りうるとし，説明変数の線形関数であるとする．「線形関数である」とは推定されるパラメタベクトル β に対して線形であればよく，2 乗項・交差項・対数項等が入ってもよい．説明変数ベクトルを x_i とすると，

$$y_i^* = x_i\beta + u_i$$

と表現できる．u_i が互いに独立で同一の正規分布に従う誤差項であり，

$$u_i|x_i \sim N(0, \sigma^2)$$

とする．最小 2 乗推定のときには誤差項の分布の形状は仮定されていなかったことに注意しよう．

この設定のもとで，観測される変数が 1 あるいは 0 となる確率を求めよう．それぞれの観測値について質的変数が観測される確率（＝尤度）を求めることができれば，最尤法によって係数の推定を行うことができる．まず，観測される変数が 1 である確率は，

$$P(y_i = 1|x_i) = P(y^* > 0|x_i) = P(x_i\beta + u_i > 0|x_i) = P(u_i > -x_i\beta|x_i)$$

ここで，u_i は正規分布 $N(0, \sigma^2)$ に従うから，正規分布の対称性より

$$P(y_i = 1|x_i) = 1 - P(u_i < -x_i\beta|x_i) = 1 - (1 - P(u_i < x_i\beta|x_i)) = P(u_i < x_i\beta|x_i)$$

標準偏差で割って基準化すると，

$$P(y_i = 1|x_i) = P(u_i < x_i\beta|x_i) = \Phi\left(\frac{x_i\beta}{\sigma}\right)$$

同様に，観測される変数が 0 である確率は，

$$P(y_i = 0|x_i) = P(y^* \leq 0|x_i) = P(x_i\beta + u_i \leq 0|x_i) = P(u_i \leq -x_i\beta|x_i)$$

さきほどと同様の展開によって，

$$P(y_i = 0|x_i) = P(u_i \leq -x_i\beta|x_i) = 1 - \Phi\left(\frac{x_i\beta}{\sigma}\right)$$

u_i の独立性が仮定されれば，全体の尤度は各観測値の尤度の積で表現することができる．個々の条件付き尤度は

$$L_i = \left[\Phi\left(x_i\beta/\sigma\right)\right]^{y_i^*}\left[1-\Phi\left(x_i\beta/\sigma\right)\right]^{1-y_i^*}$$

と表現できる．したがって個々の条件付き対数尤度は

$$\log L_i = y_i^* \log\left[\Phi\left(x_i\beta/\sigma\right)\right] + \left(1-y_i^*\right)\log\left[1-\Phi\left(x_i\beta/\sigma\right)\right]$$

ここで，関数形から分かるとおり，対数尤度を最大化する係数推定値 $\hat{\beta}$ の値は誤差項の分散 σ に依存するが，σ の値自体は決まらない．そこで，$\sigma = 1$ と置いて最大化を行う．これは，probit モデルでは潜在変数が閾値より大きいか小さいかのみが問題になっており，潜在変数の「大きさ」は問題にならないことに対応している．

観測値間の独立性が仮定されれば，対数尤度は個々の条件付き対数尤度の和となり，これを最大化して推定を行う．推定されるべきパラメタは係数ベクトル β である．推定されたベクトル $\hat{\beta}$ の値はどのような意味を持っているのだろうか．Probit モデルでは，

$$P(y_i = 1|x_i) = \Phi\left(x_i\hat{\beta}\right)$$

であるから，$\hat{\beta}$ の符号は潜在変数の変化の方向を示すとしても，値はそのままでは分かりにくい．限界効果は，しばしばデータの平均値 $\bar{x}$ で評価され，

$$\frac{\partial}{\partial x_{ki}}P(y_i = 1|\bar{x}) = \phi\left(\bar{x}\hat{\beta}\right)\hat{\beta}_k \tag{15.13}$$

も報告される．限界効果は確率の変化分であるので，その大きさは**%ポイント**で表す．たとえば，説明変数に平均値を代入したときの確率 $P(y_i = 1)$ が 70 %であるときにある説明変数の限界効果が 5 %ポイントである，とは，説明変数の値が 1 増加したときに，確率 $P(y_i = 1)$ が 70 %から 75 %に変化する，ということである．説明変数がダミー変数であるばあいには，限界効果よりも確率の差

$$P(y_i = 1|x_{ki} = 1) - P(y_i = 1|x_{ki} = 0)$$

のほうがわかりやすいかもしれない．実際，Stata の `dprobit` や `mfx` では，ダミー変数の説明変数については確率の差が出力される．

順序 probit モデル

順序反応モデル（ordered response）とは，被説明変数に採用される変数がとりうる選択肢に明確な順序が存在するようなばあいに用いられる．たとえば，なにかの好みが被説明変数であるとき,「好き」「嫌い」の 2 択であれば 2 項選択であるが,「好き」「やや好き」「どちらでもない」「やや嫌い」「嫌い」であれば選択肢は 5 つであり，この 3 つには順序が存在する．債券の格付け等もこの例に当てはまるし，資産運用の方針が「国債中心」「国債と株式混合」「株式中心」というのも順序反応モデルの対象となりうる．

順序の決まった観測される変数 y_i を規定する連続な潜在変数 y_i^* を想定しよう．被説明変数のとりうる値を J とする．観測される被説明変数の取りうる値が「好き」「やや好き」「どちらでもない」「やや嫌い」「嫌い」だとすれば（$J=5$），潜在変数 y_i^* は「好き度」を示す連続変数である．潜在変数がある値（閾値）よりも大きな値となれば観測される被説明変数は「好き」となる，というように，潜在変数の一定の範囲に観測される被説明変数の値が対応していると考えよう．単純な probit モデルと同様に，この潜在変数が説明変数 x_i の 1 次関数で表現でき，

$$y_i^* = x_i\beta + u_i \tag{15.87}$$

と書けるとしよう．単純な probit モデルと同じく,「1 次関数である」とは推定されるパラメタベクトル β に対して線形であればよく，2 乗項・交差項・対数項等が入ってもよい．u_i は誤差項であり，正規分布に従うと仮定する．順序プロビットにおいても誤差項の分散は識別できないので，u_i は標準正規分布に従う，すなわち

$$u_i|x_i \sim N(0,1) \tag{15.87}$$

としよう．被説明変数のとりうる値の数は J 個だから，潜在変数の範囲を J 個に区切ってそれぞれに被説明変数の値が対応していると考えよう．J 個に区切るから区切りの数は $J-1$ 個であり，その値を小さいほうから $\alpha_1, \alpha_2, ..., \alpha_{J-1}$ としよう．対応する被説明変数の値をここでは $1, 2, ..., J$ とすると，潜在変数と被説明変数の対応は

$$\begin{aligned}
&y_i = 1 \quad \text{if} \quad y_i^* \le \alpha_1 \\
&y_i = 2 \quad \text{if} \quad \alpha_1 < y_i^* \le \alpha_2 \\
&y_i = 3 \quad \text{if} \quad \alpha_2 < y_i^* \le \alpha_3 \\
&\vdots \\
&y_i = J \quad \text{if} \quad \alpha_{J-1} < y_i^*
\end{aligned} \tag{15.88}$$

となる．ここで，閾値の値 $\alpha_1, \alpha_2, ..., \alpha_{J-1}$ も未知であることに注意しよう．被説明変数が 1 という値を取る確率は，単純な probit と同様に

$$P(y_i = 1|x_i) = P(y^* \le \alpha_1|x_i) = P(u_i \le \alpha_1 - x_i\beta|x_i) = \Phi(\alpha_1 - x_i\beta)$$

である．$y_i = 2$ となる確率も同じように考えると，

$$\begin{aligned}
P(y_i = 2|x_i) &= P(\alpha_1 < y_i^* \le \alpha_2|x_i) \\
&= P(\alpha_1 < x_i\beta + u_i \le \alpha_2|x_i) \\
&= P(u_i \le \alpha_2 - x_i\beta|x_i) - P(u_i < \alpha_1 - x_i\beta|x_i) \\
&= \Phi(\alpha_2 - x_i\beta) - \Phi(\alpha_1 - x_i\beta)
\end{aligned}$$

となる．$y_i = J$ については，

$$\begin{aligned} P(y_i = 2|x_i) &= P(\alpha_{J-1} < y_i^*|x_i) \\ &= P(\alpha_{J-1} < x_i\beta + u_i|x_i) \\ &= 1 - P(u_i < \alpha_{J-1} - x_i\beta|x_i) \\ &= 1 - \Phi(\alpha_{J-1} - x_i\beta) \end{aligned}$$

である．容易に分かるように，それぞれの値を取る確率を全て足すと 1 になる．被説明変数がそれぞれの値を取る確率が表現できたので，条件付き尤度関数を構成することができる．表現の簡単化のために，指標関数（indicator function）を導入しよう．指標関数とは，カッコの中の条件が満たされているときだけ 1 であり，満たされていないときにはゼロの値を取る関数であり，$\mathbf{1}(.)$ で表す．個々の条件付き尤度は

$$L_i = [P(y_i = 1|x_i)]^{\mathbf{1}(y_i=1)} \times [P(y_i = 2|x_i)]^{\mathbf{1}(y_i=2)} \times ... \times [P(y_i = J|x_i)]^{\mathbf{1}(y_i=J)}$$

だから，先に求めた確率を代入すると個々の条件付き対数尤度は

$$\begin{aligned} \log L_i = \mathbf{1}(y_i = 1)\log[\Phi(\alpha_1 - x_i\beta)] + \mathbf{1}(y_i = 2)\log[\Phi(\alpha_2 - x_i\beta) - \Phi(\alpha_1 - x_i\beta)] \\ +... + \mathbf{1}(y_i = J)\log[1 - \Phi(\alpha_{J-1} - x_i\beta)] \end{aligned} \tag{15.89}$$

観測値の独立性が仮定されれば，サンプル全体の条件付き対数尤度は $\log L_i$ の和であり，これを最尤推定することができる．ここで推定されるパラメタは，潜在変数の係数ベクトル β と閾値の値 $\alpha_1, \alpha_2, ..., \alpha_{J-1}$ であるが，説明変数ベクトル x_i に定数項が含まれているときには閾値の値のうちの 1 つが識別されない．潜在変数の値と閾値の相対関係だけが問題となるからである．Stata のばあい，説明変数ベクトルに定数項が含まれない代わりに閾値がすべて推定される．

推定されたベクトル $\hat{\beta}$ の解釈について考えよう．この係数ベクトルは潜在変数の値を決めるから，係数 β_k が正に推定されれば，説明変数 x_{ki} が大きくなれば潜在変数 y_i^* の当てはめ値が大きくなることを示しており，したがって被説明変数 y_i も「大きく」なる傾向があることを表している．それゆえ，限界効果を説明変数の平均で評価することにすれば，単純なプロビットモデルと同じく，

$$\begin{aligned} \frac{\partial}{\partial x_{ki}} P(y_i = 1|\bar{x}) &= -\phi\left(\alpha_1 - \bar{x}\hat{\beta}\right)\hat{\beta}_k \\ \frac{\partial}{\partial x_{ki}} P(y_i = J|\bar{x}) &= \phi\left(\alpha_{J-1} - \bar{x}\hat{\beta}\right)\hat{\beta}_k \end{aligned}$$

が成り立つし，「端」でない $y_i = j$ についても

$$\frac{\partial}{\partial x_{ki}} P(y_i = j|\bar{x}) = \left[\phi\left(\alpha_{j-1} - \bar{x}\hat{\beta}\right) - \phi\left(\alpha_j - \bar{x}\hat{\beta}\right)\right]\hat{\beta}_k$$

が成り立つ．「端」でない $y_i = j$ となる確率の変化には，x_i の変化によって「下から入ってくる」要因と「上へ出て行く」要因の両方が影響することに注意しよう．

区間回帰モデル

順序 probit モデルの応用として，区間データ（interval-coded data）の推定がある．区間データとは，所得階層のように，一定の範囲に入っていることだけが情報として入手可能なデータである．たとえば所得データのばあい，実際の所得 y_i^* とコード化された所得階層 y_i のあいだには

$$
\begin{aligned}
&y_i = 1 \quad \text{if} \quad y_i^* \leq a_1 \\
&y_i = 2 \quad \text{if} \quad a_1 < y_i^* \leq a_2 \\
&y_i = 3 \quad \text{if} \quad a_2 < y_i^* \leq a_3 \\
&\vdots \\
&y_i = J \quad \text{if} \quad a_{J-1} < y_i^*
\end{aligned}
$$

のような関係が成り立つ．この関係は順序選択モデルにおける潜在変数と被説明変数の関係に似ているが，区間データのばあいは閾値 $a_1, a_2, ..., a_{J-1}$ は既知の数値であり，y_i^* が明確な意味を持っているという点で異なる．さて，y_i^* と説明変数との関係を検討するため，線形関係

$$y_i^* = x_i\beta + u_i$$

を想定し，このパラメタ β を推定したいとしよう．このとき，誤差項 u_i が説明変数の条件付きで正規分布に従うと仮定すれば，順序 probit モデルと同様にして尤度関数を構成し，最尤推定を行うことができる．ただし，閾値 $a_1, a_2, ..., a_{J-1}$ は既知の数値なので推定の対象とならず，また，誤差項の分散は 1 に基準化できず，こちらは推定の対象となる．区間回帰モデルでは，観測できない y_i^* が明確な意味を持ち，閾値が観測可能だから，$\log(y_i^*)$ を被説明変数とする線形関係を想定した回帰分析も可能となる．この場合には閾値も対数変換する必要がある．

2.3 Logit モデル

Logit モデルでは，確率分布関数 $P(.)$ としてロジスティック分布を用いる．分布関数を $\Lambda(z)$ と書くと，

$$\Lambda(z) = \frac{\exp(z)}{1 + \exp(z)} \tag{15.12}$$

である．

単純な logit モデル

単純な logit モデルとは，ここでは 2 項選択モデルをいう．潜在変数を想定し，probit のときと同じく

$$y_i^* = x_i\beta + u_i$$

とすれば，誤差項 u_i が標準ロジスティック分布に従うときとも解釈できる．logit モデルでは，観測される変数が 1 である確率は，

$$P(y_i = 1|x_i) = \frac{\exp(x_i\beta)}{1 + \exp(x_i\beta)}$$

と表現され，推定されるべきパラメタは係数ベクトル β である．ここから対数尤度をただちに導くことができる．

Logit モデルと probit モデルはともに 2 項選択モデルである．最尤推定は定式化が正しければ推定量が一致性を持つから，逆に言えば，定式化が正しくなければ推定量の意義は怪しいものとなる．それゆえ，厳密に言えばサンプル (y_i, x_i) のデータ生成過程（DGP: data generating process）が logit モデルであるものを probit モデルで推定したり，その逆を行ったりすれば，推定量には信頼が置けないことになる．しかしじっさいには，分布の裾を除けば，いずれのモデルで推定しても，平均値周りで推定された限界効果は似たような値となることが多いし，最大化された対数尤度の値も似たようなものとなることが多い．それゆえ，単純なモデルを考えるかぎり，いずれのモデルを選択するかは実際の応用においてはほとんど問題とならない．もちろん，推定される係数ベクトル β の値は，関数形が異なるので，似たような数値にはならない．ただし，分布の中ほどについては，

$$\hat{\beta}_{\text{logit}} \simeq 1.6\hat{\beta}_{\text{probit}}$$

が成り立つことが知られている[2]．

多項 logit モデル

離散選択モデルのうち，被説明変数がとりうる値が 3 つ以上あるときに，一般に多項選択モデル（multinomial）と呼ぶ．ここではそのうち，とりうる選択肢に明確な順序が存在しない（unordered response）ばあいを考える．職業選択や交通手段選択，学校選択等，その例は数多い．一見すると順序があるように見える労働供給量の選択にも多項選択モデルは応用されている．

被説明変数 y_i がとりうる選択肢が J 個あるとし，その属性が説明変数ベクトル x_i で表現されるような主体 i が J 個の選択肢から 1 つを選ぶという状況を考える．ここでは，それぞれの選択肢の属性が主体の選択に与える効果は捨象している．また，説明変数ベクトル x_i には定数項が含まれている．このとき，多項 logit モデルではそれぞれの選択肢 j を選ぶ確率は

$$P(y_i = j|x_i) = \frac{\exp(x_i\beta_j)}{1 + \sum_{h=1}^{J-1}\exp(x_i\beta_h)} \quad \text{for } j = 1, ..., J-1 \tag{15.76}$$

と表される．経済主体は J 個の選択肢の中から 1 つ選んでいるから，J 個の選択確率の和は 1 に等しく，$j = J$ については

$$P(y_i = J|x_i) = \frac{1}{1 + \sum_{h=1}^{J-1}\exp(x_i\beta_h)}$$

[2] Cameron, A. Colin, Pravin K. Trivedi. 2005. *Microeconometrics: Methods and Applications.* Cambridge University Press., の (14.13) 式（p.473）による．

と基準化される．順序 probit モデルと同じく，指標関数 $\mathbf{1}(.)$ を用いればここから尤度関数を構成することができ，最尤推定を適用することができる．ここで推定されるパラメタはそれぞれの選択肢についての係数ベクトルたち $(\beta_1, ..., \beta_{J-1})$ であり，選択肢が 3 つあれば，2 つのベクトルが推定される．

推定されたパラメタ $(\beta_1, ..., \beta_{J-1})$ の解釈はなかなかめんどうである．説明変数 x_{ki} が連続変数であるとき，x_{ki} が 1 単位大きくなったときに選択肢 j が選ばれる確率の増分は

$$\frac{\partial}{\partial x_{ki}} P(y_i = j|x_i) = P(y_i = j|x_i) \left[\beta_{jk} - \frac{\sum_{h=1}^{J-1} \beta_{hk} \exp(x_i \beta_h)}{1 + \sum_{h=1}^{J-1} \exp(x_i \beta_h)} \right] \tag{15.77}$$

と計算される．もっと単純な解釈としては，相対的な選ばれ方（**オッズ比: odds ratio**）について

$$\frac{P(y_i = j|x_i)}{P(y_i = J|x_i)} = \exp(x\beta_j) \quad \text{for } j = 1, ..., J-1 \tag{15.78}$$

が成り立つから，この比の変分は $\beta_{jk} \exp(x\beta_i)\Delta x_k$ で近似される．また，同じことだが，対数オッズ比は線形結合 $x_i\beta$ で表現される．いずれにしても，β_{jk} の符号が正であれば，対応する説明変数 x_{ki} の値が大きくなれば選択肢 j が選ばれる確率が高くなることが分かる．また，定式化から明らかなように

$$P(y_i = j \text{ or } y_i = h|x_i) = P(y_i = j|x_i) + P(y_i = h|x_i)$$

なので，

$$P(y_i = j|y_i = j \text{ or } y_i = h, x_i) = \frac{\exp(x_i[\beta_j - \beta_k])}{1 + \exp(x_i[\beta_j - \beta_k])}$$

が成り立つ．

確率的選択モデル

ここまで述べてきた多項 logit モデルでは，選択肢のもつ属性の影響は考慮されておらず，選ぶ側の属性によってどの選択肢を選ぶ確率が高くなるか，のみが検討の対象であった．しかししばしば問題になるのは，選択肢の側の属性の影響であろう．どのような属性をもつ選択肢が選ばれやすいのか，という問題である．

このような問題は，（加法的）**確率効用モデル**（additive random utility）を基礎にした**確率的選択モデル**（probabilistic choice）によって分析される．いま，主体 i が選択肢 j を選んだときに得られる効用 y_{ij}^* が主体と選択肢ごとに定義される説明変数 x_{ij} と誤差項の線形関数で表されるとしよう．すなわち，

$$y_{ij}^* = x_{ij}\beta + a_{ij} \quad \text{for } j = 1, ..., J \tag{15.79}$$

y_{ij}^* は効用水準を表す潜在変数である．x_{ij} は主体ごと・選択肢ごとに異なる．たとえば x_{ij} は，個人 i が交通手段 j を選んだときの所要時間や，個人 i が病院 j に通うための交通費，

等である．主体ごと・選択肢ごとに異なる x_{ij} が観察でき，係数ベクトル β と誤差項 a_{ij} が決まれば，主体 i が選択肢 j を選んだときに得られる効用水準 y^*_{ij} が決まり，主体は効用が最も高くなるような選択肢を選ぶだろう．すなわち，

$$y_i = \text{argmax}(y^*_{i1}, y^*_{i2}, ..., y^*_{iJ})$$

である[3]．いま，誤差項 a_{ij} が独立に同一の**タイプ I の極値分布（type I extreme value distribution）**に従う[4]とすれば，選択肢 j が選ばれる確率は

$$P(y_i = j|x_i) = \frac{\exp(x_{ij}\beta)}{\sum_{h=1}^{J}\exp(x_{ih}\beta)} \tag{15.80}$$

となる（McFadden 1974）．この確率を用いるモデルは条件付き選択モデル（conditional logit）とも呼ばれる．この確率値を偏微分してみると，選択肢 j の k 番目の属性 x_{jk} が変化したときに選択肢 j が選ばれる確率の変分は

$$\frac{\partial}{\partial x_{jk}}P(y_i = j|x_i) = P(y_i = j|x_i)[1 - P(y_i = j|x_i)]\beta_k \tag{15.81}$$

と表される．

*多項 logit モデルの導出

上で述べた確率の導出は以下の通りである．極値分布の密度関数，分布関数はそれぞれ，

$$f(v) = \exp(-v - e^{-v})$$
$$F(v) = \exp(-e^{-v})$$

と表される．極値分布のモード（最頻値）はゼロだが，平均は e（オイラー数），中位値は $-\ln(\ln 2)$ である．いま，選ばれた選択肢 j の誤差項 a_{ij} となって選択肢 j が選ばれる確率は，$x_{ij}\beta = U_{ij}$ と書くと，

$$\left[\prod_{k\neq j} F(a_{ij} + U_{ij} - U_{ik})\right] f(a_{ij})$$

と表現できる．第 1 項は，その他の選択肢 k から得られる効用のほうが小さくなる確率，第 2 項は a_{ij} が実現する密度を表している．ここに極値分布の密度関数・分布関数を代入すると，

$$\prod_{k\neq j} \exp\left(-e^{-a_{ij}-U_{ij}+U_{ik}}\right) \times \exp(-a_{ij} - e^{-a_{ij}})$$

[3] argmax は，その引数を最大にするように選ばれたものを表す記号．max は最大化された値を表す．
[4] タイプ I の極値分布の密度関数は，$f(z) = \exp(-z)\exp(-\exp(-z))$ である．

変形のために対数をとると，

$$
\begin{aligned}
&-a_{ij} - e^{-a_{ij}} + \sum_{k \neq j} -e^{-a_{ij}-U_{ij}+U_{ik}} \\
&= -a_{ij} - e^{-a_{ij}} - e^{-a_{ij}-U_{ij}} + \sum_{k \neq j} e^{U_{ik}} \\
&= -a_{ij} - e^{-a_{ij}} \left(1 + e^{-U_{ij}} \sum_{k \neq j} e^{U_{ik}} \right) \\
&= -a_{ij} - e^{-a_{ij}} \left(e^{-U_{ij}} e^{U_{ij}} + e^{-U_{ij}} \sum_{k \neq j} e^{U_{ik}} \right) \\
&= -a_{ij} - e^{-a_{ij}} \left(e^{-U_{ij}} \sum_{k} e^{U_{ik}} \right)
\end{aligned}
$$

ここで，$e^{-U_{ij}} \sum_k e^{U_{ik}} = e^{\lambda_i}$ とおく．指数をとって元に戻すと，

$$\exp\left(-a_{ij} - e^{-a_{ij}} e^{\lambda_i}\right)$$

選択肢 j が選ばれる確率は，これを a_{ij} について積分すれば求められるから，

$$P(y_i = j) = \int_{-\infty}^{+\infty} \exp\left[-a_{ij} - e^{-(a_{ij}-\lambda_i)}\right] da_{ij}$$

ここで変数を変換する．$a'_{ij} = a_{ij} - \lambda_i$ とおくと，積分区間は変化せず（$+\infty - \lambda_i = +\infty$）

$$
\begin{aligned}
P(y_i = j) &= \int_{-\infty}^{+\infty} \exp\left[-a'_{ij} - \lambda_i - e^{-a'_{ij}}\right] da'_{ij} \\
&= e^{-\lambda_i} \int_{-\infty}^{+\infty} \exp\left[-a'_{ij} - e^{-a'_{ij}}\right] da'_{ij}
\end{aligned}
$$

積分のなかは極値分布の密度関数となっているから，$-\infty$ から $+\infty$ まで積分した値は 1 である．したがって，

$$P(y_i = j) = e^{-\lambda_i} = \frac{e^{U_{ij}}}{\sum_k e^{U_{ik}}}$$

多項 logit モデルとの比較

多項 logit モデルと確率的選択モデルは似たようなモデルであり，確率的選択モデルは多項選択モデルの一種とみなすこともあるが，選択肢の属性を考慮しているかしていないかという点で異なる．

多項 logit モデルのばあいには選択肢の属性は説明変数として含まれていないから，選択肢の属性が選択確率に与える影響は分析できない．したがって，選択肢の属性が重要でないか，興味の対象でないか，あるいは単に利用可能でないときに用いられる．家計のデータを集めて職業選択を分析する，といったばあいがこれにあたる．

確率的選択モデルには選択肢の属性が説明変数として含まれるから，家計や企業が観測可能な選択肢の属性に基づいて選択を行うとき，その観測可能な選択肢がどのような影響を与えるかが分析の対象となる．それゆえ，家計の購買行動（商品選択）や，仮想質問法の一種である**コンジョイント分析**（conjoint）に用いられる[5]．コンジョイント分析のばあいは観測可能な選択肢の属性を調査者が制御するが，観察データの場合にはデータの利用可能性に注意しなければならない．確率選択モデルの説明変数 x_{ij} は主体 i にとっての選択肢 j の値だから,「選ばれなかった選択肢」についての x_{ij} の値が必要であるからである．たとえば,「患者は近い病院に行く」という問題を確率的選択モデルで分析しようと思えば,行かなかったが選択肢に入っている病院を特定し，行かなかったその病院までの距離の情報を入手する必要がある．

近年の応用では，選択する主体の属性も説明変数に含んだより一般的なモデルを用いる．多項 logit モデルでの説明変数になるような主体属性を表す説明変数ベクトルを w_i とすれば，確率的効用は

$$y^*_{ij} = x_{ij}\beta + w_i\delta_j + a_{ij} \quad \text{for } j = 1, ..., J$$

と表され（ただし $\delta_J = 0$），これをもとに尤度が構成される．

IIA：他の選択肢からの独立性

確率選択モデルは主体の選択行動をモデル化するのに便利な定式化であるが，制約もある．その制約として最も強いといわれているのが**他の選択肢からの独立性**（IIA: independence from irrelevant alternatives）である．選択肢 j が選ばれる確率は

$$P(y_i = j|x_i) = \frac{\exp(x_{ij}\beta)}{\sum_{h=1}^{J}\exp(x_{ih}\beta)} \tag{15.80}$$

であったから，2 つの選択肢 j, h が選ばれる相対的な確率は

$$\frac{P(y_i = j|x_i)}{P(y_i = h|x_i)} = \frac{\exp(x_{ij}\beta)}{\exp(x_{ih}\beta)} = \exp[(x_{ij} - x_{ih})\beta] \tag{15.83}$$

となり，問題となっている 2 つの選択肢 j, h 以外の選択肢の選ばれやすさとは無関係である．

他の選択肢からの独立性は応用問題によっては深刻な問題となる．また，ある選択肢が利用可能でなくなったときの相対的な選択確率は変化しないので，政策分析にも制約となりうる．極端な例を考えてみよう（McFadden 1974）．交通手段の選択を考える．最初の選択肢は車と赤いバスの 2 つであり，それぞれの選択確率は 1/2 ずつであるとしよう．ここに 3 番目の選択肢として青いバスが加わり，青いバスと赤いバスの相対的な選択確率が等しいとすれば，IIA の仮定のもとでは，車・赤いバス・青いバスの選択確率はすべて 1/3 となる．

このような問題の解決法もいろいろ提案されている．ひとつは，確率的効用の誤差項が任意の相関を持つ J 次の多変量正規分布（multinomial normal）に従うと仮定する多項 probit

[5] コンジョイント分析のばあいには，選択している主体の属性を個別効果とみなして，panel logit モデルを用いるのが一般的のようである．

モデルである．多項 probit モデルは，単純な probit モデルからの素直な拡張のように見えるが，J 次の多変量正規分布に基づく尤度関数は，高次の積分を含むため複雑なものとなり，実際上の計算は非常に困難となる．通常の最尤推定法では，選択肢が 5 個以上の推定は実際上不可能であるとされる．

いまひとつの解決法は，選択肢の構造を階層化したモデル（hierarchical）であり，入れ子型選択モデル（nested logit）がその代表である．このモデルでは，最終的な選択肢はいずれかのグループに分けられ，選択を行う主体は，第 1 段階目ではまずグループを選び，次にそのグループに含まれる選択肢の中から 1 つを選ぶと仮定される．グループ内の選択肢同士については IIA の仮定が必要となるが，グループを超えた選択肢については IIA の仮定が緩められる．

厚生評価

確率的選択モデルでは，潜在変数として効用値を想定しているから，経済厚生の評価が可能となる．ある選択肢の属性が変化したときに，それに伴う選択の変化をも考慮した経済厚生の評価が理論的には可能であり，しばしば補償変分（compensated variation）が用いられる．すなわち，選択肢の属性が変化したのちに，変化前の最大化された効用水準を達成するために必要な所得額を計算すればよい．もちろん，選択肢の属性の変化とともに，所得額の変化は選択を変化させ，最大化された効用水準を変化させるので，その評価は必ずしも容易ではない．

2.4 推定の評価

離散選択モデルについても，推定された係数の値以外にいくつかの統計量を報告し，その推定の評価を行う必要がある．係数推定値についての検定は Wald・LM・LR 検定を用いることができるから，通常の最小 2 乗推定のばあいと同じく，各推定値の標準誤差，係数がゼロという帰無仮説に対する有意水準を報告するのが普通である．これらについては，Stata 等のパッケージアプリケーションでは自動的に出力される．また，最大化された対数尤度 L も報告されることが多い．

最小 2 乗推定の決定係数 R^2 に対応するものの 1 つとして，perfect correctly predicted が推定のよさの指標として報告される．これは，推定結果から各観測値についてそれぞれの選択肢を選ぶ確率を求め，その確率が最も高いものが実現したとした結果と，実際の被説明変数の値が一致しているものの比率である．

決定係数に類似したものはいろいろ提案されており，まとめて pseudo R^2 と呼ばれる．2 項選択モデルでは，McFadden（1974）が pseudo R^2 を提案している．定数項のみを説明変数として含むモデルを推定し，その最大化された対数尤度を L_0 とし，実際に最大化された対数尤度 L に対して，$1 - L/L_0$ を決定係数とするものである．この値はゼロと 1 の間に収まる．パッケージアプリケーションでは pseudo R^2 が自動的に出力されるので，確認しておく必要があろう．

2.5 「内生性」問題

最尤推定は，尤度関数が正しく特定化されているときに一致性を持つから，そうでなければ推定結果は怪しいものとなる．推定結果が怪しいものとなる要因は，おおむね最小 2 乗法のときとよく似ている．ここではありうる問題について考えよう．

Stock and Watson (2006) では，最小 2 乗推定が一致性を持たないような状況（内的妥当性がない状況）として，Omitted variables，Misspesification of the functional form，Errors in variables，Sample selection，Simultaneous causality が挙げられていた．いずれも誤差項と説明変数の相関をもたらし，一致性を失わせる要因となった．

ここで扱ったモデルでは，誤差項が独立に同一の分布に従うと仮定して最尤推定を行っており，条件付き尤度の導出では，

$$u_i|x_i \sim N(0,1)$$

という条件を用いている．この条件は $E(u_i|x_i) = 0$ を含意するから，やはり誤差項と説明変数間の相関が問題となる．それゆえ，離散選択モデルにおいても説明変数と（潜在変数にかかわる）誤差項の相関に注意する必要があるし，問題となりそうな状況は最小 2 乗推定のときとそれほど変わらない．最小 2 乗推定のときには，直交条件が満たされず内的妥当性がないばあいには，適切な操作変数を探してきて 2 段階最小 2 乗法を行うという解決方法があった．離散選択モデルにおいては，2 段階最小 2 乗法はそのままは適用できず，原因に応じていろいろな手法が提示されている．そのいくつかは，見落とされている原因を明示的に数式で表現し，最尤推定によって追加的なパラメタをも推定する方法を採る．

たとえば，省略変数があるばあい，その**無視された異質性**（neglected heterogeneity）を表す誤差項以外に追加し，異質性の分布を仮定してそのパラメタを推定する手法もある（Wooldridge 2002, ch 15.7.1）．説明変数が逆の因果性を持つばあいには，同時方程式体系を明示的に考慮して尤度関数を構成すれば解決できることもある（Wooldridge 2002, ch 15.7.2-3）．操作変数法を応用した推定方法も提案されているし，サンプルがパネル構造であれば，個別効果を考慮した推定方法もありうる．尤度関数があまりに複雑になるばあいには，シミュレーションをともなう推定（maximum simulated likelihood）も用いられる．

2.6 Stata code

Probit モデル・logit モデルは Stata では以下のようなコマンドである．

```
probit 被説明変数 説明変数
logit 被説明変数 説明変数
```

限界効果を求めるばあい，probit モデルでは `probit` のところを `dprobit` とすれば求めることができる．また，順序 probit モデル，多項 logit モデルのばあいは．

```
oprobit 被説明変数 説明変数
mlogit 被説明変数 説明変数
```

となる．

第3章　Tobitモデル

ここではTobitモデルと総称される推定方法を扱う．Wooldridge (2006) *Introductory Econometrics: A Modern Approach*, Ch.17を参考にしている．また，Wooldridge (2002) Ch.16, 17も参照せよ．式番号はWooldridge (2006) による．

3.1　Tobitモデルの使いみち

被説明変数がある限られた範囲の値しか取らない状況，あるいは，なんらかの条件に当てはまったとき（当てはまらなかったとき）にはデータが観測できない状況ではTobitモデルが用いられる．このような被説明変数を**制限従属変数**（LDV: Limited Dependent Variable）とも呼ぶ．広義のTobitモデルはさらに2つにわけて考えることができる．ひとつは，すべての観測値について説明変数も被説明変数も観測できるばあいである．このグループには，**端点解**（corner solution）のばあいや，トップコーディング等による**打ち切り**（censored）データのばあいが含まれる．いまひとつは，なんらかの条件に当てはまった観測値については被説明変数が観測できず，そのような観測値を推定に利用できないケースであり，**切断**（truncated）データと呼ばれる．このような状況はを**標本選択（サンプルセレクション）**（sample selection）とも呼ばれる．

データが打ち切られる，あるいは切断される観測できるための条件はあらかじめ分かっていることもあるし，分かっていないこともある．その条件が個々の観測値によって異なるばあいも，定数のばあいも，確率的に決まるばあいも考えられる．さらに，着目している推定式の説明変数にデータが観測できるかどうかが含まれるケース等も考えうる．広くTobitモデルと呼ぶときには，これらの同時方程式体系をも含む[1]．このようなTobitモデルについても，離散選択モデルと同様，潜在変数（latent variable）を想定すると便利なことが多い．

3.2　Type I Tobit

被説明変数があらかじめ決められた範囲の値しか取らない状況は，標準的な打ち切りTobitモデル（standard censored Tobit），あるいは**Type I Tobit**モデルと呼ばれる．この標準的なTobitモデルが適用できる典型的な状況は，家計や企業の最適化行動の結果として観測される消費量や生産量が，ミクロ経済学でいうところの端点解になっているばあ

[1] Tobitモデルの種々のタイプについては，Amemiya (1985) を参照せよ．Type IからType V Tobitという分類は雨宮健による．

いである．労働時間や，通常の財の購入はマイナスの値を取らないから，観測される被説明変数の値の範囲は非負の実数に限定される．タバコや酒等の嗜好財の消費を考えてみると，消費量は連続変数として観測できるものの，消費量がゼロという世帯も一定数存在するだろう．

被説明変数 y_i に対して，潜在変数 y_i^* を考えよう．潜在変数は制約がないときの最適解を表す．これまでと同じく，潜在変数は説明変数と誤差項の 1 次関数で表されるとし，誤差項は説明変数の条件付きで平均ゼロの正規分布に従うとしよう．つまり，

$$y_i^* = \mathbf{x}_i\beta + u_i, \quad u_i|\mathbf{x}_i \sim N(0, \sigma^2) \tag{17.18}$$

と表現できるとしよう．分散均一性（homoskedastisity）を仮定している．さて，観測される被説明変数の値は y_i はあらかじめ決められた範囲に限定されるが，ここでは正の値のみを取ることができるとする．すなわち，潜在変数が正の値をとったときにはその値がそのまま観測されるが，負の値となったときにはゼロとして観測される．y_i が取り得る範囲があらかじめ決まっていれば，同じようにして尤度を構成できる．ここでは，下限のみがあって，その下限がゼロのケースを扱っている．このとき，

$$y_i = \begin{cases} y_i^* & \text{if} \quad y_i^* \geq 0 \\ 0 & \text{if} \quad y_i^* < 0 \end{cases} \tag{17.19}$$

と書くことができる．潜在変数が負の値をとるときには被説明変数 y_i の値がゼロとなるので，標準正規分布の累積分布関数 $\Phi(.)$ を用いて

$$\begin{aligned} P(y_i = 0|\mathbf{x}_i) &= P(y_i^* < 0|\mathbf{x}_i) = P(\mathbf{x}_i\beta + u_i < 0|\mathbf{x}_i) \\ &= P(u_i < -\mathbf{x}_i\beta|\mathbf{x}_i) = P\left(\frac{u_i}{\sigma} < -\frac{\mathbf{x}_i\beta}{\sigma}\Big|\mathbf{x}_i\right) = \Phi\left(-\frac{\mathbf{x}_i\beta}{\sigma}\right) = 1 - \Phi\left(\frac{\mathbf{x}_i\beta}{\sigma}\right) \end{aligned}$$

と変形できる．ここで，誤差項の分布が説明変数 $\mathbf{x}_i$ と独立という条件を使っていることに注意しよう．潜在変数が正の値を取るときにはそのまま観測されるから，確率密度 は標準正規分布の確率密度関数 $\Phi(.)$ を用いて

$$Pr(y_i = \mathbf{x}_i\beta + u_i|\mathbf{x}_i) = Pr(u_i = y_i - \mathbf{x}_i\beta|\mathbf{x}_i) = \frac{1}{\sigma}\phi\left(\frac{y_i - \mathbf{x}_i\beta}{\sigma}\right)$$

したがって，サンプルが母集団からの無作為抽出で，誤差項が独立に同一の分布に従っているとすれば，各観測値の対数尤度は，

$$\log L_i = \mathbf{1}(y_i = 0)\log\left[1 - \Phi\left(\frac{\mathbf{x}_i\beta}{\sigma}\right)\right] + \mathbf{1}(y_i > 0)\log\left[\frac{1}{\sigma}\phi\left(\frac{y_i - \mathbf{x}_i\beta}{\sigma}\right)\right] \tag{17.22}$$

と表現され，最大化すべき対数尤度はこの個別の対数尤度の和となる．推定されるべきパラメタは，潜在変数の 1 次関数の係数パラメタ β，誤差項の標準偏差 σ である．統計アプリケーションでの出力は，通常の最小 2 乗推定に類似する．

係数の解釈

標準的な Tobit モデルで推定される係数 β は，説明変数 $\mathbf{x}_i$ のうちの対応する変数が 1 単位増加するときの潜在変数 y_i^* の期待値の増分を表している．通常の最小 2 乗推定では，これがすなわち，被説明変数 y_i の期待値の増分であるが，Tobit モデルではそうではない．これは，潜在変数が負の値を取るばあいには,「大きな」部分だけが期待値に反映されるためである．そこで，被説明変数の期待値の増分はどのように表されるのか考えてみよう．

まず，説明変数 $\mathbf{x}_i$ で条件付けたときの被説明変数の期待値 $E[y_i|\mathbf{x}_i]$ を求めよう[2]．被説明変数がとる値は場合分けされているから，この場合分けにしたがって，とりうる値とその値になる確率を計算する．つまり，

$$E[y_i|\mathbf{x}_i] = P(y_i = 0)E[y_i|y_i = 0, \mathbf{x}_i] + P(y_i > 0)E[y_i|y_i > 0, \mathbf{x}_i]$$

を計算すればよい．左辺第 1 項は，$y_i = 0$ のときのことだからゼロなので，

$$E[y_i|\mathbf{x}_i] = P(y_i > 0)E[y_i|y_i > 0, \mathbf{x}_i]$$

である．上で述べたように，$P(y_i = 0) = 1 - \Phi(\mathbf{x}_i\beta/\sigma)$ だから，

$$P(y_i > 0) = \Phi(\mathbf{x}_i\beta/\sigma)$$

である．次に，$E[y_i|y_i > 0, \mathbf{x}_i]$ を求めよう．$y_i > 0$ のとき $y_i = y_i^*$ だから，

$$E[y_i|y_i > 0, \mathbf{x}_i] = E[y_i^*|y_i^* > 0, \mathbf{x}_i] = E[\mathbf{x}_i\beta + u_i|\mathbf{x}_i\beta + u_i > 0, \mathbf{x}_i]$$

期待値の線形性から，$\mathbf{x}_i\beta$ と σ を期待値の外に出して，

$$E[y_i|y_i > 0, \mathbf{x}_i] = \mathbf{x}_i\beta + E[u_i|u_i > -\mathbf{x}_i\beta, \mathbf{x}_i] = \mathbf{x}_i\beta + \sigma E\left[\frac{u_i}{\sigma}\middle|\frac{u_i}{\sigma} > -\frac{\mathbf{x}_i\beta}{\sigma}, \mathbf{x}_i\right]$$

ここで，u_i/σ は標準正規分布に従う．標準正規分布において，ある値 c より大きな値をとることが分かっているときの確率密度は，その値 c より大きい部分を膨らませればよいので，

$$\frac{\phi(z)}{1 - \Phi(c)}$$

となるから，$z(\phi(z)/1 - \Phi(c))$ を c から ∞ について積分して，

$$E[z|z > c] = \frac{\phi(c)}{1 - \Phi(c)}$$

である[3]．$z = u_i/\sigma$，$c = -\mathbf{x}_i\beta/\sigma$ と読み替えると，

$$E\left[\frac{u_i}{\sigma}\middle|\frac{u_i}{\sigma} > -\frac{\mathbf{x}_i\beta}{\sigma}, \mathbf{x}_i\right] = \frac{\phi(u_i/\sigma)}{1 - \Phi(-\mathbf{x}_i\beta/\sigma)} = \frac{\phi(u_i/\sigma)}{\Phi(\mathbf{x}_i\beta/\sigma)}$$

[2] もちろん潜在変数の期待値は $E[y_i^*|\mathbf{x}_i] = \mathbf{x}_i\beta$ である．

[3] 標準正規分布の密度関数については $\phi'(z) = -z\phi(z)$ が成り立つから，$z\phi(z)$ の原始関数は $-\phi(z)$ であり，

$$\int_c^\infty \frac{z\phi(z)}{1 - \Phi(c)}dz = \frac{1}{1 - \Phi(c)}[-\phi(\infty) - (-\phi(c))] = \frac{1}{1 - \Phi(c)}[\phi(c)]$$

なので，

$$E[y_i|y_i > 0, \mathbf{x}_i] = \mathbf{x}_i\beta + \sigma\frac{\phi(u_i/\sigma)}{\Phi(\mathbf{x}_i\beta/\sigma)}$$

となる．ここで，$\lambda(c) = \phi(c)/\Phi(c)$ と定義する．この比は**逆ミルズ比（inverse Mill's ratio）**と呼ばれる．

$$E[y_i|y_i > 0, \mathbf{x}_i] = \mathbf{x}_i\beta + \sigma\lambda(\mathbf{x}_i\beta/\sigma) \tag{17.24}$$

この期待値は条件付き期待値（conditional expectation）とも呼ばれ，右辺第 2 項は，潜在変数の期待値 $\mathbf{x}_i\beta$ がマイナスで，かつ，y_i^* がプラスになる部分が期待値計算に含まれてしまう効果を反映している．逆ミルズ比と σ の積は $\mathbf{x}_i$ に依存していることに注意しよう．これらをまとめると，

$$E[y_i|\mathbf{x}_i] = \Phi(\mathbf{x}_i\beta/\sigma)\left[\mathbf{x}_i\beta + \sigma\frac{\phi(u_i/\sigma)}{\Phi(\mathbf{x}_i\beta/\sigma)}\right] = \Phi(\mathbf{x}_i\beta/\sigma)\mathbf{x}_i\beta + \sigma\phi(\mathbf{x}_i\beta/\sigma) \tag{17.25}$$

となり，$E[y_i|\mathbf{x}_i]$ は説明変数の非線形関数となる．仮定によりこの期待値は正の値をとる．最小 2 乗推定では被説明変数の条件付き期待値は説明変数の線形関数であり，誤差項の分散の大きさは期待値に影響しなかったことを思い出そう．

さて，説明変数が連続とすれば，説明変数 x_k が 1 単位大きくなるときの期待値の増分を計算できる．最小 2 乗推定ではその値は β_k で与えられるが，標準的な Tobit では簡単な形では表現できない．期待値の変分は，積の微分の公式から

$$\frac{\partial}{\partial x_k}E[y_i|\mathbf{x}_i] = \frac{\partial P(y_i > 0)}{\partial x_k}E[y_i|y_i > 0, \mathbf{x}_i] + P(y_i > 0)\frac{\partial}{\partial x_k}E[y_i|y_i > 0, \mathbf{x}_i] \tag{17.28}$$

それぞれの項目について計算してみよう．$P(y_i > 0) = \Phi(\mathbf{x}_i\beta/\sigma)$ だから，

$$\frac{\partial P(y_i > 0)}{\partial x_k} = (\beta_k/\sigma)\phi(\mathbf{x}_i\beta/\sigma) \tag{17.29}$$

また，$E[y_i|y_i > 0, \mathbf{x}_i] = \mathbf{x}_i\beta + \sigma\lambda(\mathbf{x}_i\beta/\sigma)$ だから，

$$\frac{\partial}{\partial x_k}E[y_i|y_i > 0, \mathbf{x}_i] = \beta_k + \beta_k\frac{d\lambda}{dc}(\mathbf{x}_i\beta/\sigma)$$

ここで，$d\lambda/dc = -\lambda(c)[c + \lambda(c)]$ より，

$$\frac{\partial}{\partial x_k}E[y_i|y_i > 0, \mathbf{x}_i] = \beta_k\left[1 - \lambda(\mathbf{x}_i\beta/\sigma)[\mathbf{x}_i\beta/\sigma + \lambda(\mathbf{x}_i\beta/\sigma)]\right] \tag{17.26}$$

これらを代入すると，

$$\begin{aligned}\frac{\partial}{\partial x_k}E[y_i|\mathbf{x}_i] = \frac{\beta_k}{\sigma}\phi\left(\frac{\mathbf{x}_i\beta}{\sigma}\right)\left(\mathbf{x}_i\beta + \sigma\lambda\left(\frac{\mathbf{x}_i\beta}{\sigma}\right)\right)\\ +\Phi\left(\frac{\mathbf{x}_i\beta}{\sigma}\right)\beta_k\left[1 - \lambda\left(\frac{\mathbf{x}_i\beta}{\sigma}\right)\left[\frac{\mathbf{x}_i\beta}{\sigma} + \lambda\left(\frac{\mathbf{x}_i\beta}{\sigma}\right)\right]\right]\end{aligned}$$

となるので，$\phi(c) = \Phi(c)\lambda(c)$ を使って整理すると，

$$\frac{\partial}{\partial x_k} E[y_i|\mathbf{x}_i] = \beta_k \Phi\left(\frac{\mathbf{x}_i\beta}{\sigma}\right) \tag{17.30}$$

となる．つまり，標準的な Tobit モデルを適用するのが適切な状況で OLS 推定を行うと，その傾きは $\Phi(\mathbf{x}_i\beta/\sigma)$ のぶんだけ過小評価される．この限界効果も説明変数 $\mathbf{x}_i$ に依存するので，効果を平均で評価した $\beta_k\Phi(\bar{\mathbf{x}}\beta/\sigma)$ を報告するか，効果の平均 $\beta_k \times n^{-1}\sum_{i=1}^{n}\Phi(\mathbf{x}_i\beta/\sigma)$ が報告される．説明変数がダミー変数の場合は，その変数が 1 のときの期待値と 0 のときの期待値の差が報告されることが多い．

特定化の問題

ここで扱っている標準的な Tobit モデルは，観測値の i.i.d. を仮定して対数尤度を構成しているから，この仮定が成り立たなければ推定量は一致性を持たない．たとえば，分散不均一性は OLS では一致性の問題とならないが，Tobit では一致性を損なう要因となる．また，離散選択モデルのときに示したように，内生性の問題があれば一致性は失われる．とはいえ，実際に真のデータ生成過程（DGP: Data Generating Process）は知りえないのだから，一致性を持つための条件がどれくらい尤もらしいか，のほうが問題となろう．

標準的な Tobit モデルにおける潜在的な重要な仮定は，$y_i^* > 0$ という条件付きでの $y_i > 0$ となる確率が，同じ条件付きでの y_i の期待値と強く連関している，という点であろう．y_i がゼロ以上となる確率と，y_i の条件付き期待値が別の要因で決まっているということもありうる．たとえば，生命保険の購入量と年齢の関係を考えてみよう．若いうちは生命保険に入る傾向が強くないだろうから，年齢と $y_i > 0$ となる確率は正の関係になろう．他方で，保険を購入した人たちだけに限ってみれば，年老いているほうが保険の価値は小さくなるから保険の購入量は小さくなり，$E[y_i|y_i > 0, x_i]$ と年齢とは負の関係を持つだろう．このような関係は標準的な Tobit モデルでは取り扱うことができない．

被説明変数が $y_i > 0$ となる確率と，同じ条件付きでの y_i の期待値が，異なる要因で決まると想定するモデルは，hurdle モデルや，two-part モデルと呼ばれるモデルで考えることができる．これらも広い意味では Tobit モデルとみなされている．

3.3 打ち切りデータ

標準的な Tobit モデルが想定する状況とよく似た状況に，データの**打ち切り**（censored）が起きている状況がある．これは，被説明変数があらかじめ決められた範囲を外れた値をとるときに，被説明変数が観測できない状況である．ただし，説明変数は観測できていることに注意しよう．このようなことが起きる典型的な例のひとつは，**トップコーディング**（top coding）の状況である．個票を用いた消費額や所得額を被説明変数にする回帰分析を行うとき，回答率の向上や個人情報保護の観点から，非常に大きな値は「〜万円以上」と一括してコーディングされることがある．この閾値をたとえば 2000 万円とすると，観測さ

れる所得 y_i は，実際の所得 y_i^* に対して，$y_i = \min(y_i^*, 2000\text{ 万円})$ となる．このとき，観測される被説明変数の値の範囲は 2000 万円以下と解釈できる．

標準的な Tobit モデルと同じく，被説明変数 y_i に対して，潜在変数 y_i^* を考えよう．潜在変数はトップコーディングの例であれば実際の所得や消費量に対応する．潜在変数は説明変数と誤差項の 1 次関数で表されるとし，誤差項は説明変数の条件付きで平均ゼロの正規分布に従う，すなわち

$$y_i^* = \mathbf{x}_i\beta + u_i, \quad u_i|\mathbf{x}_i \sim N(0, \sigma^2) \tag{17.36}$$

とする．さて，観測される被説明変数の値は y_i はあらかじめ決められた範囲に限定されるが，c_i を閾値としてトップコーディングが行われているとしよう．すなわち，潜在変数が c_i より小さい値をとったときにはその値がそのまま観測されるが，大きな値となったときには観測されない．ただし，c_i より大きいということは判別される．このようなケースを上からの打ち切り（censoring from above）もしくは right censoring と呼ぶ．y_i がある値より大きなときにのみ観測できる状況は，censoring from below，もしくは left censoring と呼ばれる．Left censoring でも同じようにして尤度を構成できる．さてこのばあい，

$$y_i = \begin{cases} y_i^* & \text{if} \quad y_i^* < c_i \\ c_i & \text{if} \quad y_i^* \geq c_i \end{cases} \tag{17.37}$$

と書くことができる．ここで，閾値 c_i は既知でさえあれば観測値ごとに異なっていてもよい．潜在変数が $y_i^* > c_i$ のときには被説明変数 y_i の値が c_i となるので，標準正規分布の累積分布関数 $\Phi(.)$ を用いて

$$\begin{aligned} P(y_i = c_i|\mathbf{x}_i) &= P(y_i^* > c_i|\mathbf{x}_i) = P(\mathbf{x}_i\beta + u_i > c_i|\mathbf{x}_i) \\ &= P(u_i > c_i - \mathbf{x}_i\beta|\mathbf{x}_i) = P\left(\frac{u_i}{\sigma} > \frac{c_i - \mathbf{x}_i\beta}{\sigma}\Big|\mathbf{x}_i\right) = 1 - \Phi\left(\frac{c_i - \mathbf{x}_i\beta}{\sigma}\right) \end{aligned}$$

と変形できる．ここで，誤差項の分布が説明変数 $\mathbf{x}_i$ と独立という条件を使っていることに注意しよう．標準的な Tobit モデルと同様に，サンプルが母集団からの無作為抽出で，誤差項が独立に同一の分布に従っているとすれば，各観測値の対数尤度は，

$$\log L_i = \mathbf{1}(y_i = c_i)\log\left[1 - \Phi\left(\frac{c_i - \mathbf{x}_i\beta}{\sigma}\right)\right] + \mathbf{1}(y_i < c_i)\log\left[\frac{1}{\sigma}\phi\left(\frac{y_i - \mathbf{x}_i\beta}{\sigma}\right)\right] \tag{17.39}$$

と表現され，最大化すべき対数尤度はこの個別の対数尤度の和となる．推定されるべきパラメタは，潜在変数の 1 次関数の係数パラメタ β，誤差項の標準偏差 σ である．

標準的な Tobit モデルと同じく係数 β は潜在変数 y_i^* の動きを示すパラメタである．しかし，打ち切りデータのばあいにはデータが観測できないのは主体の行動の結果ではなく，データ収集（data collection）の問題だから，期待値や限界効果は通常の最小 2 乗回帰と同じく，β で評価すればよく，期待値は説明変数の線形関数とみてよい．

打ち切りデータの 1 つの応用は，期間分析（duration analysis）である．期間分析は，事象（event）が起きるまでの期間の長さを被説明変数とする分析手法であり，しばしばデー

タの入手可能性から，事象がまだ起きていない観測値もサンプルに含まれる．そのようなばあい，事象が起きるまでの期間が観測できた期間よりも長いという情報しか使うことができず，打ち切りデータとみなすことができる．もっとも，期間分析は生存期間分析（survival analysis）の枠組みで検討されることも多い．

3.4 サンプル・セレクション・モデル

打ち切り（censored）と似たような状況に，データの切断（truncated）が起きている状況がある．切断データのばあい，一定の条件を満たさない観測値（データの一部）は観測されず，被説明変数の値の情報は利用できない．つまり母集団からランダムサンプリングしたサンプルのうちの一部分しか利用できないことになるので，**標本選択（sample selection）**が起きている，とも言う．

標本選択の典型例は，企業が申し出る（offer）賃金率の推定である．労働者は企業が offer する賃金率を受け入れて働くか，受け入れずに働かないか（take it or leave it）どちらかであるとすれば，観測される賃金率は受け入れられた賃金率だけであり，受け入れられなかった賃金率は観測されない．このようなとき，一致性のある推定量を得るにはどうしたらよいのだろうか．

標本選択は問題になるのか?

サンプルセレクションは必ずしも常に問題となる，つまり，最小 2 乗推定量の一致性をおびやかすというわけではない．たとえば，母集団から無作為抽出したサンプルからさらに無作為抽出してサンプルを作れば，実質的には小さな無作為抽出標本を作ったことになるので，最小 2 乗推定量は一致性を持つ．他方，標準的な Tobit モデルにおいて端点解をとる観測値を除去した標本を用いれば，(17.24) 式から容易に想像されるように，最小 2 乗推定量は一致性を失う．

ここで，選択変数（selection indicator）s_i を導入しよう．選択変数 s_i は，$(y_i, \mathbf{x}_i)$ の組合せが全て観測された観測値に対しては 1 の値をとる指標変数（indicator variable）であり，$s_i = 0$ であるような観測値 i は利用できない．母集団から 無作為抽出された 観測値について，被説明変数・説明変数・誤差項のあいだに，いつものような線形関係が成り立つとしよう．

$$y_i = \mathbf{x}_i\beta + u_i, \quad E[u_i|\mathbf{x}_i] = 0 \tag{17.42}$$

もし，s_i の値にかかわらず全ての i の y_i が観測できていたとすれば，その標本について最小 2 乗推定を行うことによって一致性のある推定量を得ることができる．標本選択が起きているときには $s_i = 1$ となる観測値しか利用できないが，それらに限っても直交条件が成り立っている，つまり，

$$E[u_i|\mathbf{x}_i, s_i = 1] = 0$$

が成り立っていれば，最小 2 乗推定推定量は一致性をもつ[4]．この条件が成り立つような状況はおもに 3 通り考えることができる．第 1 は，選択変数 s_i が説明変数 $\mathbf{x}_i$ の関数となっているばあい，つまり標本選択が説明変数だけで決まっているばあいである．このとき，$\{\mathbf{x}_i, s_i = 1\}$ が持つ情報は $\{\mathbf{x}_i\}$ が持つ情報と同じである．第 2 は，選択変数 s_i が説明変数 $\mathbf{x}_i$ とも誤差項 u_i とも独立であるばあいである．これは無作為抽出のケースに対応する．第 3 は，第 1 と第 2 のばあいの組合せで，選択変数 s_i が，説明変数 $\mathbf{x}_i$ と，誤差項と無相関の確率変数の関数となっているばあいである．いずれのばあいでも，選択変数は誤差項 u_i と無相関であることに注意しよう．

標本選択が最小 2 乗推定量の一致性を失わせるケースとして，打ち切りデータのうち打ち切られた観測値を利用しないばあいを考えてみよう．打ち切られるのは $u_i > c_i - \mathbf{x}_i\beta$ であるときだったから，選択変数は $s_i = \mathbf{1}(u_i < c_i - \mathbf{x}_i\beta)$ であり，誤差項 u_i と相関を持つ．したがって，このばあいには最小 2 乗推定量は一致性を持たない．

Heckman の 2 段階推定（Heckit）

サンプルセレクションに対処する基本的な方法のひとつが **Heckman の 2 段階推定（Heckit）**である．注目している回帰式はいつものような線形関係であり，

$$y_i = \mathbf{x}_i\beta + u_i, \quad E[u_i|\mathbf{x}_i] = 0 \tag{17.46}$$

とする．ここで，標本選択が起きていて，選択変数が

$$s_i = \mathbf{1}(\mathbf{z}_i\gamma + v_i \geq 0) \tag{17.47}$$

と表されるとする．$\mathbf{z}_i$ は標本選択を規定する変数群であり，外生変数とする．すなわち，注目している回帰式の誤差項 u_i と無相関であり，かつ，標本選択にまつわる誤差項 v_i とも無相関とする．Heckman の 2 段階推定が機能するためには，しばしば説明変数 $\mathbf{x}_i$ は $\mathbf{z}_i$ の部分集合である．つまり，全ての説明変数は $\mathbf{z}_i$ に含まれ，かつ，説明変数に含まれない外生変数が $\mathbf{z}_i$ に含まれている．のちに明らかになるように，$\mathbf{z}_i$ に含まれるが説明変数 $\mathbf{x}_i$ に含まれない変数は，2 段階最小 2 乗推定における除外された操作変数と同じ役割を果たす．これらの条件をまとめると，

$$E[u_i|\mathbf{x}_i, \mathbf{z}_i] = 0$$

[4] この条件は，次のように理解することもできる．いま，利用可能な観測値だけを使う回帰式を

$$s_i y_i = s_i\mathbf{x}_i\beta + s_i u_i \tag{17.44}$$

と書こう．利用できない観測値（$s_i = 0$）については全ての項がゼロになるので，β の推定のためになんの情報ももたらさない．さて，この回帰式を最小 2 乗推定して一致推定量を得るための条件は，誤差項と説明変数が直交する（内積がゼロ）だから，

$$\begin{aligned} &E[(s_i\mathbf{x}_i)(s_i u_i)] = E[s_i\mathbf{x}_i u_i] = 0 \\ &E[s_i u_i | s_i\mathbf{x}_i] = 0 \end{aligned} \tag{17.45}$$

が成り立てばよい．

である．選択変数にまつわる誤差項 v_i は変数 $\mathbf{z}_i$ と（したがって $\mathbf{x}_i$ と）独立であり，標準正規分布に従うと仮定する．誤差項 v_i は注目している回帰式の誤差項 u_i とは相関してもよく，その相関係数（のようなもの）を ρ とする．もし $\rho = 0$ であれば，標本選択によって最小 2 乗推定の一致性が失われない第 3 のケースに該当し，標本選択バイアスは発生しない．

さて，注目している回帰式（17.46）について，観測されるという条件付きの期待値をとることを考えよう．観測されるかどうかは $\mathbf{z}_i$ と v_i によって規定されるから，まず，これらについての条件付き期待値をとると，

$$E[y_i|\mathbf{z}_i, v_i] = E[\mathbf{x}_i\beta|\mathbf{z}_i, v_i] + E[u_i|\mathbf{z}_i, v_i]$$

説明変数 $\mathbf{x}_i$ は $\mathbf{z}_i$ の部分集合であり，また $\mathbf{z}_i$ は u_i と独立であると仮定しているから，

$$E[\mathbf{x}_i\beta|\mathbf{z}_i, v_i] = \mathbf{x}_i\beta, \quad E[u_i|\mathbf{z}_i, v_i] = E[u_i|v_i]$$

が成り立つから，

$$E[y_i|\mathbf{z}_i, v_i] = \mathbf{x}_i\beta + E[u_i|v_i]$$

となる．u_i と v_i が平均ゼロの 2 変量正規分布に従うとすると，パラメタ ρ について $E[u_i|v_i] = \rho v_i$ と書くことができ，

$$E[y_i|\mathbf{z}_i, v_i] = \mathbf{x}_i\beta + \rho v_i$$

さらに $s_i = 1$ について条件付き期待値をとると，

$$E[y_i|\mathbf{z}_i, s_i = 1] = \mathbf{x}_i\beta + \rho E[v_i|\mathbf{z}_i, s_i = 1]$$

右辺第 2 項は，$\mathbf{z}_i\gamma + v_i \geq 0$ のときの v_i の期待値だから逆ミルズ比となり，

$$E[y_i|\mathbf{z}_i, s_i = 1] = \mathbf{x}_i\beta + \rho\lambda(\mathbf{z}_i\gamma)$$

先に述べたように，もし $\rho = 0$ であれば右辺第 2 項が消えてしまい，標本選択の問題は発生しない．

さて，β を推定するためには，説明変数の項に加えて $\rho\lambda(\mathbf{z}_i\gamma)$ を追加して最小 2 乗推定すればよい．逆に言うと，標本選択の問題とは，$\lambda(\mathbf{z}_i\gamma)$ の項が省略変数（omitted variable）となる問題ともいえる．$\lambda(\mathbf{z}_i\gamma)$ は変数として観測されないが，v_i が正規分布に従うとすれば，probit 推定により，

$$P(s_i = 1|\mathbf{z}_i) = \Phi(\mathbf{z}_i\gamma) \tag{17.49}$$

を推定することができる．したがって，Heckman の 2 段階推定は以下のような手続きで行われる．

1. 全ての観測値を使って，被説明変数が観測できるかできないかの（s_i を被説明変数とする）probit 推定を行い，逆ミルズ比の推定値（$\hat{\lambda}_i = \lambda(\mathbf{z}_i\hat{\gamma})$）を計算する．

2. 被説明変数が観測される観測値を使って，y_i を $\mathbf{x}_i$ と $\hat{\lambda}_i$ に回帰した最小 2 乗推定を行う

標本選択によって最小 2 乗推定が一致性を失うかどうかは，2 段階目の推定での $\hat{\lambda}_i$ の係数 ρ がゼロであるかどうかを検定すれば分かる．もし $\rho = 0$ という帰無仮説が棄却できなければ，標本選択の問題は起きていないから，被説明変数が観測されたサンプルだけ用いて最小 2 乗推定を行っても一致性のある推定量が得られる．

Heckman の 2 段階推定の 2 段階目では被説明変数が観測される観測値を使って，

$$y_i = \mathbf{x}_i\beta + \rho\hat{\lambda}_i + e_i$$

を推定することになる．ここから，$\mathbf{z}_i$ について 2 つの注意すべき点が指摘できる．第 1 に，逆ミルズ比が一致推定されている必要があるから，$\mathbf{z}_i$ には全ての説明変数 $\mathbf{x}_i$ が含まれていたほうがよいということである．データが手元にあるのだから，全ての $\mathbf{x}_i$ を $\mathbf{z}_i$ に含ませることは難しくないだろうし，そうでなければ 1 段階目で省略変数バイアスをもたらすかもしれない．第 2 に，逆ミルズ比は $\mathbf{z}_i$ の非線形関数であるものの，しばしば線形に近いことから，$\mathbf{z}_i$ と $\mathbf{x}_i$ が同じ変数の組合せであれば，多重共線性の問題を発生しかねないということである．ちょうど，2 段階最小 2 乗法において除外された操作変数がなければ推定が行えないのと同じ状況である．それゆえ，選択変数についての 1 段階目の回帰に用いる $\mathbf{z}_i$ には，説明変数 $\mathbf{x}_i$ に含まれない外生変数が 1 つ以上，識別のために必要である．

Heckman の 2 段階推定は probit と最小 2 乗推定の組合せであるが，選択変数の誤差項 v_i と本来の誤差項 u_i が従う 2 変量正規分布の形状を規定できれば，直接に尤度関数を構成することもできる．Stata はデフォルトではこのような最尤推定を行う．

ここまで扱ってきたモデルは，**Type 2 Tobit モデル**とも呼ばれる．これは **Type 5 Tobit モデル**と呼ばれるより一般的なモデルの特殊形とみなすこともできる（Amemiya 1995）．

$$\begin{aligned} y_i &= \mathbf{x}_i\beta + u_i \\ &\quad \beta = \beta_1 \quad \text{if} \quad s_i = \mathbf{z}_i\gamma + v_i < 0 \\ &\quad \beta = \beta_2 \quad \text{if} \quad s_i = \mathbf{z}_i\gamma + v_i \geq 0 \end{aligned}$$

このモデルでは，s_i の値によって β がスイッチするので，**switching regression** とも呼ばれる．あるいは，s_i によってレジームが選択されるとも解釈できるので **self-selection model** と呼ばれることもある．さらに，s_i が観察できないばあいにも拡張でき，そのようなばあいには**未知レジームのスイッチング回帰**と呼ばれる．

3.5 Stata code

標準的な Tobit モデルは，以下のようなコマンドラインで推定することができる．

```
tobit 被説明変数 説明変数, ll(#) ul(#)
```

ここで，ll は left-censoring limit，rl は right-censoring limit である．
Heckman の 2 段階推定法は，以下のようなコマンドラインで推定することができる．

```
heckman 被説明変数 説明変数, select( 選択変数 )
```

上の表現を使えば，説明変数は x_i，選択変数は z_i である．heckman コマンドは，デフォルトでは最尤法を用いた推定をするので，あえて Heckman の 2 段階推定を行う場合には，twostep オプションを，select( ) のあとに加える必要がある．