最尤法[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．

[1] 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．式番号は Hayashi (2000) による．

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．

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．

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．

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．

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．