# ガウス—マルコフ-モデルにおける 最小2乗法による推定について

Prof. C. R. Rao*

私は OR に関する実際の経験がありません．私の勤務先は Hamaker 氏とは違いまして，研究所（Indian Statistical Institute）であります．そこには約 2,000 の人々註1) がおりまして研究を行っております．そこでは二つのものが生産されます．すなわち統計家と統計の公式（formula）であります．統計家はなかなか市場価値のある生産物であります．

さて今日は OR とは関係のない統計のお話をいたします．OR 学会の人々に OR に属さないことをお話しするのも，また有益であろうと思います．もしこれが数理統計の専門家の集りでしたら，OR のトピックをえらんだことでありましょう．

（以下講義の要旨）

ここでは最小2乗法による推定の問題を扱う．

最小2乗法は，分散分析，実験計画法，回帰分析等において広く用いられる．例えば一例として秤量問題がある．これについては日科技連のセミナーでお話した註2)．最小2乗法は測定の誤差をともなう場合のデータを取り扱う方法である．

$X_1, \cdots\cdots, X_n$ を観測される変数，$\theta_1, \cdots\cdots, \theta_m$，を未知のパラメターとするとき，

$$X_i = \sum a_{ij}\,\theta_j + \varepsilon_i, \qquad i = 1 \cdots\cdots n$$

と表わされるものとする．

$\varepsilon_i$ は平均0の確率変数で，その分散共分散行列は既知であるとする（以下においては簡単のために相関0とする.）

未知のパラメターの一次形式 $p_1\theta_1 + p_2\theta_2 + \cdots\cdots + p_m\theta_m$ を推定する問題を考える．その場合に最小2乗推定は，$p_1\hat{\theta}_1 + p_2\hat{\theta}_2 + \cdots\cdots + p_m\hat{\theta}_m$ で与えられる．ここに $\hat{\theta}_1, \cdots\cdots, \hat{\theta}_m$ は

$$\sum_{i=1}^{n} (X_i - a_{i1}\theta_1 - a_{i2}\theta_2 - \cdots - a_{im}\theta_m)^2$$

を $\theta_1, \cdots\cdots, \theta_m$ に関して最小にするように定めた値である．

最小2乗推定値は次のような性質を持つことはよく知られている．

1. $\varepsilon_i$ が正規分布に従うときは，それは最尤推定値に一致する．
2. $\varepsilon_i$ の分布型に仮定をおかなくても，推定値を $X_i$ に関して線型不偏なものに限れば，その中で最小2乗推定値は分散最小なものを与える．
3. $\varepsilon_i$ が正規分布をするときは，最小2乗推定値は，あらゆる不偏推定値の中で分散最小で

* 昭和35年6月13日 ISI セミナーにおける要旨．昭和35年10月27日受理．

註1 研究員以外の標本調査に従事する人々をふくむ．

註2 デミング，コクランの講義などと一諸に出版の予定．

ある．

しかし最小2乗推定値が，不偏推定値の中で一様に分散を最小にするものを与えるという性質は，$\varepsilon_i$ が正規分布をしないときには成り立たない．実際次のことが示される．

a） $\varepsilon_i$ は互いに独立で同一分布に従い，かつすべての次数のモーメントを持つ．

b） $\varepsilon_i$ の分散は $\theta_1, \cdots\cdots, \theta_m$ によらない．

c） 係数 $p_1, \cdots\cdots, p_m$ は，すべてが0または1ではない．

という条件の下で，最小2乗推定値が，一様最小分散不偏推定値であるための必要かつ十分な条件は，$\varepsilon_i$ が正規分布をすることである．

$\varepsilon_i$ の正規性の仮定を除いた場合に，最小2乗推定値を導く2の条件(ガウス＝マルコフの定理)は，線型推定，不偏性，最小分散の三つを仮定している．このような条件はあまりに厳しいように思われるので，それを変えようとする試みがいくつかなされている．例えばSavageは，不偏性の条件を除いて，

i） 線型

ii） 平均2乗誤差有界：$\sup_\theta E(c_1X_1+\cdots\cdots+c_mX_m-p_1\theta_1-\cdots\cdots-p_m\theta_m)^2<\infty$

iii） 平均2乗誤差最小

の三つの条件から最小2乗推定を導いた．

このほかにも結局最小2乗推定値をもたらすいくつかの方法があるが，それらはいずれもガウス＝マルコフの3つの条件のうちいくつかをある程度ゆるめたものにすぎない．

しかしここではこれらのものと全く別の考え方から，線型とか，不偏性とか，最小分散とかいう性質を全く用いることなく，最小2乗推定を導くことを考えよう．

例として，3つのものの重さを，天秤を4回用いて測定し，それから推定する問題を考えよう．$W_1, W_2, W_3, W_4$ を測定値，$\theta_1, \theta_2, \theta_3$ を重さとする．天秤を適当に用いて，

$$W_1=\theta_1+\theta_2+\theta_3+\varepsilon_1 \qquad W_3=\theta_1+\theta_2-\theta_3+\varepsilon_3$$
$$W_2=\theta_1-\theta_2+\theta_3+\varepsilon_2 \qquad W_4=\theta_1-\theta_2-\theta_3+\varepsilon_4.$$

と表わされるようになったものとしよう．

このとき，$\theta_1$，$\theta_2$，$\theta_3$ をどのように推定するかが問題である．それらの推定値は観測値の函数として，

$$\theta_1=T_1(W_1, W_2, W_3, W_4) \qquad \theta_3=T_3(W_1, W_2, W_3, W_4)$$
$$\theta_2=T_2(W_1, W_2, W_3, W_4) \qquad \theta_4=T_4(W_1, W_2, W_3, W_4)$$

と表わすことにする．

そうすると，まずこれらの推定値の満すべき条件としてもし測定の誤差が全くないならば，推定値は正確に真の値に一致しなければならない．すなわち

$$\varepsilon_1=\varepsilon_2=\varepsilon_3=\varepsilon_4=0 \quad \text{ならば}$$
$$T_1\equiv\theta_1 \quad T_2\equiv\theta_2, \qquad T_3\equiv\theta_3, \qquad T_4\equiv\theta_4$$

でなければならないことを要求するのは自然であろう．このような性質を，推定値 $T_i$ が方法的一致性 Methodical consistency を持つというように表現する．このことを一般的に定義すれば，一般の線型模型について，パラメターの一次形式 $p_1\theta_1+\cdots\cdots+p_m\theta_m$ の推定量 $T=T[X_1,\cdots\cdots,X_n]$ が methodically consistent であるとは，

$$T[E(X_1),\cdots\cdots,E(X_n)]=p_1\theta_1+\cdots\cdots+p_m\theta_m$$

となることである．

そうすると，推定量 $T$ が

（i） 線 型

（ii） methodically consistent

（iii） 平均2乗誤差最小

の三つの条件を満すならばそれは最小2乗推定に一致しなければならないことが証明される．

さらにこのことをもう少し拡張して，今 $W_5$ という第5の測定値があったとして，それは上の測定値とは独立で，かつその平均が0に等しいものとする．例えば重さの知られたものを量って，その測定値と真の値との差を $W_5$ とすればよい．そのとき，このような測定値は $\varepsilon$ の分散を推定するには役立つが，これが $\theta_1, \theta_2, \theta_3$ の推定値に影響を与えてはならないことは当然要求されるであろう．すなわち $T_i(W_1, W_2, W_3, W_4, W_5)\equiv T_i(W_1, W_2, W_3, W_4)$ でなければならない．

この $W_5$ のような測定値を補助測定 ancillary measurement と呼ぶ．

そうするともし補助測定にのみ誤差があるときには，推定値は正確な値を与えるものでなければならないことが要求されるのは当然であろう．

ところで，このような独立の測定でなくても，例えば $W_5=W_1-W_2$ とおくと，$E(W_5)=0$ であるから，このようなものも ancillary（補助統計量と記しておく）と考えることができる．

そうして，推定値はもし測定の誤差が，このような補助統計量にのみふくまれているならば，正確な値に一致するものでなければならないことを要求することにしよう．このような要求を満す推定量は，補助統計量と独立に方法的一致性を持つ．methodically consistent independently of ancillary と呼ぶ．

ここで，誤差が補助統計量にのみふくまれるということは，補助統計量と独立な誤差が全く存在しない場合，すなわち

$$W_i-\hat{W}_i\equiv 0$$

$$\hat{W}_i=W_i \text{ の } W_5 \text{ に対する回帰}=E(W_i|W_5)$$

である場合を意味する．

そうすると，methodically consistent である推定値は

$W_i=\hat{W}_i$ のとき $T_1\equiv\theta_1$，$T_2\equiv\theta_2$，$T_3\equiv\theta_3$ となるようなもの，すなわち

$$T_1(\hat{W}_1,\ \hat{W}_2,\ \hat{W}_3,\ \hat{W}_4)\equiv\theta_1$$

$$T_2(\hat{W}_1,\ \hat{W}_2,\ \hat{W}_3,\ \hat{W}_4)\equiv\theta_2$$

$$T_3(\hat{W}_1,\ \hat{W}_2,\ \hat{W}_3,\ \hat{W}_4) \equiv \theta_3$$

となるような推定値 $T_1, T_2, T_3$ を意味することになる．

そうすると，すべての補助統計量と独立であり，従って方法的に一致性を持つような推定値は，実は最小2乗推定値になることが証明されるのである．

このような考え方は，もし外から加わった誤差を取り除くならば，われわれは真の値を得なければならないという考え方によるものである．実験の大きさが $n$，推定すべきパラメターの数が $r$ のとき，実験の中には $n-r$ 個の補助統計量，すなわち外部的な誤差がかくされており，われわれの推定値はそれを取り除いたものでなければならないと考えるのである．

このような考え方は興味深く exciting なものである．

## 質疑応答

以下質問に入って（ここから食後の話）

〔問〕 Linear Programming の係数に誤差があるとそれは結果にどの程度影響するか．

〔答〕 理論的な研究については何も聞いていない．カルカッタではいくつかの例について計算している．

〔問〕 係数が確率変数である場合は調べられているが，しかしランダムでない変動の方が重要であろう．

〔答〕 係数に小さい誤差があっても，答に大きい違いをもたらすことがあるだろうと思う．

〔問〕 いろいろな consistency の概念の間の関係について，methodical consistency は線型モデルについてだけ定義されるものか．

〔答〕 基礎にある原則が重要である．それは観測に誤差がなく，完全な知識が得られれば，正しい値が得られなければならないということである．

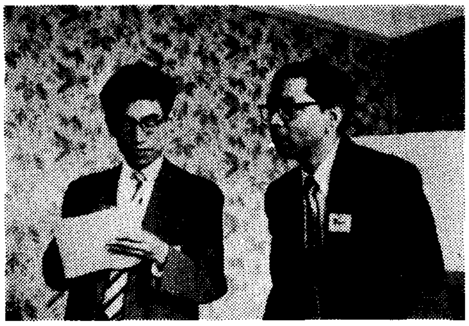
Fisher consistency（というのは Rao 氏の命名した概念である）は，分布が誤差なく知られる，すなわち経験分布が真の分布に一致するならば，パラメターの真の値が知られるということを意味する．このことの具体的に意味づけた一例が methodcal consistency である．

この考え方は，推定値を経験分布函数の functional と見なすことによって得られる．すなわち $S_n$ を標本数 $n$ の標本からの経験分布函数とし，$T$ を一つの functional として，$\hat{\theta}=T(S_n)$ とおく．（これによって $\hat{\theta}$ が $n$ と独立に定義できる．）

そうすると，$T$ が Fisher consistency を持つとは，$F$ を真の分布とするとき，

$$T(F) \equiv \theta$$

となることである．そうするとこのような性質を持つための条件は $T$ が Fréchet derivative である



ことである．

最尤推定値は Fisher consistency を持つことが知られている．

なお推定値の範囲を $S_n$ の explicit な functional に制限することにより，Le Cam の与えたような super-efficient な推定値，例えば正規分布の母平均 $\mu$ の推定において

$$\hat{\theta}=T_n=\begin{cases} \bar{x}, & |\bar{x}|>n^{-\frac{1}{4}} \\ =0, & |\bar{x}|<n^{-\frac{1}{4}} \end{cases}$$

とするようなものは除かれる．

註　この推定量の漸近分散を求めると

$$n\,\mathrm{Var}_\theta(\hat{\theta})=\sigma^2 \qquad \theta\neq 0$$
$$n\,\mathrm{Var}_0(\hat{\theta})=0 \qquad \theta=0$$

となるから，最尤推定量 $\bar{x}$ よりも一様に分散が小さくなる．このことを super-efficient と呼ぶ．

標本が $k$ 個のカテゴリーのどれかに属するような場合であるときには，それぞれに属する標本の数を $n_1, n_2, \cdots\cdots, n_k$ 個，$\sum n_i=n$ として

$$\hat{\theta}=T(n_1/n, \cdots\cdots, n_k/n) \text{ とおく．}$$

それぞれに入る確率を $\pi_1, \pi_2, \cdots\cdots, \pi_k$ とするとき，Fisher consistency は $n_1/n=\pi_1,\ n_2/n=\pi_2, \cdots\cdots,\ n_k/n=\pi_k$ のときには

$$T(\pi_1, \cdots\cdots, \pi_k) \equiv \theta$$

となることを要求するものである（これが本来 Fisher の与えた定義である．Statistical Methods and Scientific Inference を参照——竹内）

なお有限母集団については，Cochran が母集団全体の値を用いたときには真の値が得られるもの，という形で consistency を定義している．

例えば，母集団の値を $X_1, \cdots\cdots, X_N$ として，

$\mu=\frac{1}{N}\sum X_i$ とする．$\mu$ の推定値として

$$\hat{\mu}=T(x_1, \cdots\cdots, x_n)=\frac{1}{n}\sum x_i$$

をすべての $n$ について定義するとすれば，これはこの意味で consistent である．

なお普通の意味の consistency すなわち

$p \lim T_n \to \theta$ （$T_n$ が $\theta$ に確率収束）

についていえば，有限の $n$ についてこの意味で consistent なものは必らずしも Fisher consistent ではない．

## 再び Rao 氏の話

$x_1, \cdots\cdots, x_n$ が互いに独立に正規分布 $N(\mu, 1)$ に従うとするとき，$x_1-x_n$ は $N(0, \sigma^2)$ に従う．従って $x_1-x_n$ は補助統計量である．

そこで $S_n$ を経験分布函数とするとき，それをこのような補助統計量に関して補正したものを $\hat{S}_n$ とする．そのとき $T(S_n)\neq\theta$ でも

$$T(\hat{S}_n)\equiv\theta$$

となるようにすることが考えられる．ところで問題は一般的な場合にこれをどのように定式化するかである．

充足統計量 $T$ が存在し，かつそれが有界型完備ならば，補助統計量はすべてそれと独立になるから，その統計量 $T$ だけを考えればよい．(Basu の結果.)

しかしそうでない場合には，すべての補助統計量と独立な推定値というものを考えることがうまくゆかない．Basu の作った例によると，補助統計量の全体をとると，それが全標本と同等なものになってしまう場合がある．そのような場合にはすべての補助統計量と独立なものは定数しかない．Fisher は 2×2 の分割表の解析において，補助統計量を固定して考えることを主張して，周辺分布を固定した条件附分布をとったが，このような点から考えると Fisher の主張は正しくないといわねばならない．実際最大の補助統計量を求めることはむずかしい．例えば上の正規分布の例では $(x_1-x_n,\ x_2-x_n,\ \cdots\cdots)$ がそれであるが，それではコーシー分布についてはどうであろうか？

---

議義および質疑応答は極めて熱心な雰囲気のうちに行われ，かつその内容もレベルの高いもので，極めて充実した時間を過ごすことができた．

---

**附　記**

なお補助統計量と独立に方法的一致性を持つ推定量が最小2乗推定であるということは次のようにして示されるのである．

$X_1, \cdots\cdots, X_n$ を一次変換して，$Y_1, \cdots\cdots, Y_r$, $Z_1, \cdots\cdots, Z_{n-r}$ に移すことにより，

i) $Y_1, \cdots\cdots, Y_r$, $Z_1, \cdots\cdots, Z_{n-r}$ は互いに相関0

ii) $E(Y_i)$, $i=1, \cdots\cdots, r$ は $\theta_i$ の互いに独立な一次形式として表わされる．

iii) $E(Z_j)=0$, $j=1, \cdots\cdots, n-r$

すなわち，ここで $Z_1, \cdots\cdots, Z_{n-r}$ が補助統計量の全体を表わしているのである．そうするとよく知られているように，最小2乗推定値はつねに $Y_i$ の一次形式で表わされる．(このことは最小分散の前提か

ら直ちに導かれる．Bose の Least square aspect of analysis of variance では，$Y_i$ の張る空間が estimation space，$Z_j$ の張る空間が error space と呼ばれている.）

推定値 $T(X_1, \cdots\cdots, X_n)$ に対して，$Y_1, \cdots\cdots, Y_r$, $Z_1, \cdots\cdots, Z_{n-r}$ を代入して，

$T(X_1, \cdots\cdots, X_n) = U(Y_1, \cdots\cdots, Y_r,\ Z_1, \cdots\cdots, Z_{n-r})$

とすると，これが補助統計量と独立なことから，$U$ は実は $Z_j$ をふくまないこと，そうして methodically consistent は

$$U[E(Y_1), \cdots\cdots, E(Y_r)] = p_1\theta_1 + \cdots\cdots + p_m\theta_m$$

となることを意味する．

いま，

$p_1\theta_1 + \cdots\cdots + p_m\theta_m = q_1E(Y_1) + \cdots\cdots + q_rE(Y_r)$

と表わされたものとすると，

$U[E(Y_1), \cdots\cdots, E(Y_r)] = q_1E(Y_1) + \cdots + q_rE(Y_r)$

となるが，$E(Y_1), \cdots\cdots, E(Y_r)$ は一次独立で，かつそれらは任意の値をとり得るから，結局

$$U(Y_1, \cdots\cdots, Y_r) = q_1Y_1 + \cdots\cdots + q_rY_r$$

とならねばならない．すなわち$U$は最小２乗推定である．

---

（竹内　啓）