統計的方法における最適化問題

安井　清一

最小二乗法が無制約最小化問題であるように，統計学の中には最適化問題が多く含まれている．解析法の構成方法には 2 通りがあるように思える．一つは，最小二乗法のように，誤差の二乗和の最小化という問題を解いて推定量を構成する方法である．もう一つは，算術平均で母平均を推定するといったように，まず統計量を構成しておいて，その最適性でもって解析法を作る方法である．いずれにしても，最適化問題が含まれている．統計学というと，数理的な印象を与えるかもしれないので，広くデータ解析を意味する目的として，統計学を統計的方法と呼ぶ．本稿ではその中に存在する最適化問題について考える．

キーワード：最小二乗法，無制約最小化問題，罰則付き最小二乗法，最良線形不偏推定量，等式制約付き最適化問題，一般化最小二乗法，最適計画

1. はじめに

統計学，統計的方法の目的は，平たく言うと，母集団からデータを収集して解析することによって，母集団の集団としての特徴を探ることである．目的はそうであっても，データ解析はある種の最適化問題を解くことで行われる場面が多く存在するし，解析法の性質を説明するうえでは最適性が重要になってくる．このように最適化問題は統計的方法の重要な手段である．統計的方法を説明するうえでは最適化問題を前面に押し出すことはないが，本稿では，最適化問題を意識して統計的方法を説明してみたい．

まず初めに馴染み深い単回帰モデルにおける最小二乗法から入り，重回帰モデル，リッジ回帰における最適化問題へ行く．その後，統計的方法における推定について，推定量の性質に関して最適化問題を考える．

2. 回帰モデルにおける最適化問題

2.1 最小二乗法による単回帰分析

統計的方法においても，OR においても最も基礎的な事項は単回帰モデルの回帰係数を最小二乗法によって求めることだろう．本稿ではまず，ここから始めたい．

バネのフックの法則の実験や，オーム抵抗の実験などで，実験や観察によってペアのデータ $(x_1, y_1), \ldots, (x_n, y_n)$ が取られているとしよう．単回帰分析では，ペアのデータ (x, y) 間に

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,\ i = 1, \ldots, n$$

という線形関係がある（線形モデル）と仮定して，パラメータ β_0 および β_1 の値を n 個のデータから求める．統計学においては，データからパラメータを求めることを推定という．また，n 個の誤差 ε_i，$i = 1, \ldots, n$ に確率分布（通常は正規分布）を仮定し，傾き β_1 があるか（$\beta_1 = 0$ でないかどうか）を調べるための検定，線形関係や誤差の仮定が妥当かどうかなどを調べる回帰診断と呼ばれる方法までを含めて，統計学では回帰分析と呼んでいる．詳しくは，佐和 [1] や Draper and Smith [2] などを参照してほしい．

パラメータ β_0 および β_1 の推定は，誤差の 2 乗和を最小とする β_0 および β_1 をそれらの推定値 $\hat{\beta}_0$ および $\hat{\beta}_1$ とする，おなじみの最小二乗法によって行われる．すなわち，

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg\min_{\beta_0, \beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

であり，無制約最小化問題である．これは凸関数の最小化なので，

$$\begin{cases} \dfrac{dL}{d\beta_0} = -2\displaystyle\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) \\ \dfrac{dL}{d\beta_1} = -2\displaystyle\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i \end{cases}$$

より，連立方程式

$$\begin{cases} \displaystyle\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \\ \displaystyle\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0 \end{cases}$$

の解 $(\hat{\beta}_0, \hat{\beta}_1)$ が (β_0, β_1) の推定値である．この連立方

やすい　せいいち
東京理科大学理工学部
〒 278–8510 千葉県野田市山崎 2641

程式を正規方程式という．これを解いて，

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}\ ,\ \hat{\beta}_1 = S_{xy}/S_{xx},$$

ただし，$\bar{x} = \sum_{i=1}^n x_i/n$，$\bar{y} = \sum_{i=1}^n y_i/n$，$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$，$S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$ である．

2.2 重回帰モデル

単回帰モデルにおいて，x_i を説明変数，y_i を目的変数という．単回帰モデルは説明変数が一つの場合であるが，通常はいくつか取ることができて，

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i,\ \ i = 1, \ldots, n$$

を重回帰モデルという．重回帰モデルの場合，説明変数が多いので行列表記が有用である．目的変数の値からなるベクトルを $\boldsymbol{y} = (y_1, \cdots, y_n)'$ とし，切片（の係数 “1”）と説明変数の値からなる行列

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix},$$

偏回帰係数からなるベクトル $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$，誤差のベクトル $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)'$ を用意すると，重回帰モデルは

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

のように書ける．偏回帰係数の推定値 $\hat{\boldsymbol{\beta}}$ は，$\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}$ を最小にする $\boldsymbol{\beta}$ で求める．すなわち，

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \qquad (1)$$

という無制約最小化問題である．これを幾何学的解釈によって解く方法もあるが，$S(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$ を $\boldsymbol{\beta}$ で微分して解く．一般的に $\partial(\boldsymbol{\beta}'\boldsymbol{a})/\partial\boldsymbol{\beta} = \boldsymbol{a}$，および，$\partial(\boldsymbol{\beta}'\boldsymbol{A}\boldsymbol{\beta})/\partial\boldsymbol{\beta} = 2\boldsymbol{A}\boldsymbol{\beta}$ であるので，$\partial S(\boldsymbol{\beta})/\partial\boldsymbol{\beta} = -2\boldsymbol{X}'\boldsymbol{y} + 2\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}$ となる．したがって，

$$\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{y}$$

を満たす $\hat{\boldsymbol{\beta}}$ が (1) の解である．これは正規方程式の行列表現である．$p+1$ 次正方行列 $\boldsymbol{X}'\boldsymbol{X}$ が正則であるとき，$\hat{\boldsymbol{\beta}}$ は一意に定まり，

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

である．

ところで，データ数 n が偏回帰係数の数 $p+1$ よりも少ないとき，$\boldsymbol{X}'\boldsymbol{X}$ は正則でない．また，説明変数間に相関が強いときや，線形関係が隠れているとき，$|\boldsymbol{X}'\boldsymbol{X}|$ が小さくなり，逆行列が計算上，不安定になる．計算上でなくても，誤差を確率変数としたとき，$\hat{\boldsymbol{\beta}}$ の分散が大きくなり，実際にはあまり有用でなくなる場合もある．また，$n = p+1$ ではすべてのデータを $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$ が通るため，すでに手持ちのデータに対しての当てはまりはよいが，将来，生じるデータに対する値については、かなり食い違うという現象も生じる．

このようなことは，ビッグデータのようなさまざまな種類の変数を大量に扱う場面では，しばしば生じるのではないかと思う．多くの異なる種類の説明変数を利用することに加え，それらの変数の 2 乗項や積項 $(x_i x_j, i \neq j)$，ラグなどを説明変数として利用し，さらに解析のリアルタイム性を重視すると $n \ll p$ となることも十分考えられる．実用に耐えうるモデルを構築するためには，変数選択などの一工夫が必要である．

回帰モデルの活用目的として，第 1 に予測がある．予測とは，将来生じるデータを推測することである．回帰分析においては，説明変数は定数なので，ある説明変数に対する将来生じる目的変数の値の推測である．上述した場面においては，予測の精度が低下するので，予測の精度に考慮した説明変数の選択（変数選択，モデル選択）も必要である．したがって，予測を目的にする場合は，予測の精度を最適化する説明変数の集合を求めるという問題になる．予測の精度にはいくつかの考え方があるが，赤池情報量規準 (AIC) が有名である．また，統計的機械学習の分野では予測の精度を汎化能力と呼び，一部のデータのみを用いてモデルを構築した後，残ったデータに当てはめを行うクロス・バリデーションという方法で汎化能力を最適化するものが多いように感じる．赤池情報量規準を代表とする情報量規準に基づいたモデル選択の詳細については小西と北川 [3] などがある．

以上のように，データから求めた重回帰モデルを実用において活用するために，変数選択は重要である．変数選択は p 個の説明変数からそれらの部分集合を求める組合せ最適化であるが，実際には 2^p 通りがモデルの候補ではない．モデルには解釈上の妥当性を考慮し，たとえば，二次項 x^2 を含むならば一次項 x も含ませるという階層性を入れる場合が多く，制約付きの組合せ最適化となる．

変数選択によって目的にあったモデルを構築するという方法以外に，罰則項を含んだ最小二乗法による方法がある．次節ではその代表格であるリッジ回帰を説明する．

2.3 リッジ回帰と罰則付き最小二乗法

ここでは多項式回帰について考える．多項式回帰とは，散布図に p 次多項式を当てはめることである．すなわち，データ (x_i, y_i), $i = 1, \ldots, n$ に対して，

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_p x_i^p + \varepsilon_i$$

を当てはめる．最小二乗法で（偏回帰）係数を求める場合，前節の話より，$n = p + 1$ のとき，すべての点 (x_i, y_i) を通る．明らかに，得られた関数は変動が大きく（上下運動が激しく），予測の精度は高くないことがわかる．もちろん，AIC などの規準を用いた変数選択によって，最適な次数を決めることも可能だが，誤差の 2 乗和に罰則項を追加した罰則付き最小二乗法で，得られる関数の変動を適度に抑えることも可能である．なお，罰則項は正則化項とも呼ばれ，そのとき，罰則付き最小二乗法は正則化最小二乗法と呼ばれる．

偏回帰係数の二乗和 $\sum_{j=1}^{p} \beta_j^2$ を罰則項としよう．定性的に考えると，より高次の項までモデルに取り入れるということは 0 でない β_j をより多く含ませるということだから，罰則項は大きくなる．実際にはすべての係数は非ゼロの値をもつが，大きさが調節されてモデル全体として適度な変動になると期待される．

定式化するためにモデルを少し変形する．切片である β_0 は，多項式の y 方向の位置を決めるだけであり，関数の変動とは関係ない．β_0 をなくしたモデルにするために，各 y_i から $\bar{y}$ を引き，さらに，各説明変数 x_i^j の平均値 $\bar{x}_j = \sum_{i=1}^{n} x_i^j / n$ を各説明変数 x_i^j において引く．加えて，説明変数 x_i^j のばらつきは，偏回帰係数の推定値に影響を与えるので，$s_j = \sqrt{\sum_{i=1}^{n} (x_i^j - \bar{x}_j)^2}$ で割っておく．すなわち，

$$y_i - \bar{y} = \beta_1 \frac{x_i^1 - \bar{x}_1}{s_1} + \cdots + \beta_p \frac{x_i^p - \bar{x}_p}{s_p} + \varepsilon_i$$

というモデルからスタートする．$z_{ij} = \frac{x_i^j - \bar{x}_j}{s_j}$ として，モデルの行列表記を新たに $\boldsymbol{y} = (y_1 - \bar{y}, \ldots, y_n - \bar{y})'$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$,

$$\boldsymbol{Z} = \begin{pmatrix} z_{11} & \cdots & z_{1p} \\ & \vdots & \\ z_{n1} & \cdots & z_{np} \end{pmatrix}$$

と再定義すると，

$$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

である．よって，罰則の大きさを調節するパラメータ $\lambda > 0$ を導入して，

$$S(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}'\boldsymbol{\beta}$$

を最小にする $\boldsymbol{\beta}$ が求めるべき値である．重回帰モデルのときと同様にして，$\partial S(\boldsymbol{\beta}) / \partial \boldsymbol{\beta} = -2\boldsymbol{Z}'\boldsymbol{y} + 2\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\beta} + 2\lambda\boldsymbol{\beta}$ なので，$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} S(\boldsymbol{\beta})$ とすると，

$$(\boldsymbol{Z}'\boldsymbol{Z} + \lambda \boldsymbol{I})\hat{\boldsymbol{\beta}} = \boldsymbol{Z}'\boldsymbol{y}$$

の解が求めるべき偏回帰係数の値である．ここで，$\boldsymbol{I}$ は p 次の単位行列である．以上より，

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{Z}'\boldsymbol{Z} + \lambda \boldsymbol{I})^{-1}\boldsymbol{Z}'\boldsymbol{y}$$

となり，これをリッジ回帰推定量という．

リッジ回帰は数理計画的な解釈ができる．罰則項は $(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta})$ の最小化に対して，偏回帰係数の値を束縛する働きがあることから，$t > 0$ を用いて，

$$\min_{\boldsymbol{\beta}} \quad (\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta})$$
$$\text{s.t.} \quad \boldsymbol{\beta}'\boldsymbol{\beta} \leq t$$

であると考えられる．λ を大きくすると罰則が大きくなるため，偏回帰係数 $\boldsymbol{\beta}$ は最適値として小さい値をとるようになるので，λ を大きくすることは t を小さくすることに対応している．この問題を解くにあたって実際は，制約条件がアクティブであると仮定される．それは，推定においては t を小さいほうから少しずつ動かして $\boldsymbol{\beta}$ を推定する，クロス・バリデーションなどの別の基準でアクティブでなくなる前に t の動きが止まる，といった理由からであると思われる．結局は等式制約 $\boldsymbol{\beta}'\boldsymbol{\beta} = t$ の下の最小化問題を考えることになる．よって，ラグランジュ関数は

$$L(\boldsymbol{\beta}, \theta) = (\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta}) + \theta(\boldsymbol{\beta}'\boldsymbol{\beta} - t), \ \theta > 0$$

であり，

$$\frac{\partial L}{\partial \boldsymbol{\beta}} = -2\boldsymbol{Z}'\boldsymbol{y} + 2\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\beta} + 2\theta\boldsymbol{\beta} = \boldsymbol{0}$$
$$\frac{\partial L}{\partial \theta} = \boldsymbol{\beta}'\boldsymbol{\beta} - t = 0$$

を解けばよい．第 1 行目の方程式から求められるリッジ回帰推定量 $\hat{\boldsymbol{\beta}} = (\boldsymbol{Z}'\boldsymbol{Z} + \theta\boldsymbol{I})^{-1}\boldsymbol{Z}'\boldsymbol{y}$ を第 2 式へ代入し，$\boldsymbol{y}'\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z} + \theta\boldsymbol{I})^{-2}\boldsymbol{Z}'\boldsymbol{y} = t$ を得る．よって，t に対するリッジ回帰推定量が得られるので，リッジ回帰における罰則付き最小二乗法は上述の数理計画によっても解釈される．

2.4 リッジ回帰の一般化

リッジ回帰の罰則項は $\sum_{j=1}^{p}\beta_j^2$ であった．2 乗和であると最小化問題の解を解析的に求めることができるが，罰則項としての役割だけを考えると必ずしも 2 乗和である必要がない．そこで，罰則項を $\sum_{j=1}^{p}|\beta_j|$ とした *Least Absolute Shrinkage and Selection Operator* (Lasso) が提案された．すなわち，Lasso では，

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} = (\boldsymbol{y}-\boldsymbol{Z\beta})'(\boldsymbol{y}-\boldsymbol{Z\beta}) + \lambda\sum_{j=1}^{p}|\beta_j|$$

を求めることになる．$\lambda = 0$ で通常の最小二乗法になるので，λ を大きいほうから 0 に向かって少しずつ動かしながら $\hat{\boldsymbol{\beta}}$ を求めるのだが，ある λ に対して正確に 0 になる $\hat{\beta}_j$ がいくつかあるところが Lasso の特徴である．すなわち，リッジ回帰のように罰則付き最小二乗法を行っているわけだが，同時に変数選択も行っているという構造になっている．λ を動かすと，正確に 0 になる $\hat{\beta}_j$ も変化するのだが，どの λ で止めるかは，クロス・バリデーションで予測の精度を最適化するなどの方法で決められる．

さらに，罰則項を $\sum_{j=1}^{p}|\beta_j|^{\alpha}$ のように一般化した罰則付き最小二乗法が考えられている．このときの数理計画表現は，

$$\begin{aligned}\min_{\boldsymbol{\beta}} \quad & (\boldsymbol{y}-\boldsymbol{Z\beta})'(\boldsymbol{y}-\boldsymbol{Z\beta}) \\ \text{s.t.} \quad & \sum_{j=1}^{p}|\beta_j|^{\alpha} \le t\end{aligned}$$

であり，リッジ回帰の一般化である．この辺りの解説は Hastie et al. [4] にまとめられている．

3. 推定に関する最適化問題

「ある製品の重さを n 回測定したら，$x_1,\ldots,x_n$ [g] のようなデータが得られた．その製品の真の重さ μ [g] はいくらか？」という問いに対して，真の重さを n 個のデータの平均値で求めようとする人がほとんどだろう．真の重さ μ をデータから言い当てることを推定といい，推定のための式を推定量，実際に計算した値を推定値という．推定量および推定値を $\hat{\mu}$ と書き，今ここでは，推定量と推定値とを特に区別しなくても不都合はないから，$\hat{\mu} = \bar{x} = \sum_{i=1}^{n}x_i/n$ である．しかしなぜ，平均値を選ぶのであろうか．子どもの頃から言われている，データを増やせば平均値は精密になる，すなわち，大数の法則が感覚としてある（どこかで刷り込まれた？）などが理由であるような気がする．本節では，推定量としての平均値へのこだわりについて考察する．

3.1 不偏推定量

製品の真の重さを μ としたとき，n 回測定して得られたデータを

$$x_i = \mu + \varepsilon_i,\ \ i = 1,\ldots,n \tag{2}$$

のように考える．$\varepsilon_1,\ldots,\varepsilon_n$ は測定誤差を表しており，互いに独立で同一な分布に従う確率変数であるとする．また，期待値と分散はそれぞれ $E[\varepsilon_i]=0$，$V[\varepsilon_i]=\sigma^2$ とする．この仮定の中で $E[\varepsilon_i]=0$ が特に重要であり，これは測定に偏りがないということを示しており，標準試料によって校正がきちんと行われているなどの現実的な意味が含まれている．測定に偏りがあると真の値を推定するのが困難になる．また，これらの各仮定は，校正に加え，測定時においても一定の手順（標準という）が定められており，それに従って熟達した人が行った結果であるという意味がある．たとえば，測定に未熟な人が行った場合，n 回測定している間に上達して，だんだんと分散が小さくなるといったようなことは起こらない，ということである．

さて，μ の推定量である $\bar{x}$ の期待値を求めてみよう．$\bar{x} = \mu + \sum_{i=1}^{n}\varepsilon_i/n$ であるので，

$$E[\bar{x}] = \mu + \frac{1}{n}\sum_{i=1}^{n}E[\varepsilon_i] = \mu$$

である．推定量の期待値が推定対象に一致する推定量のことを不偏推定量という．すなわち，$\bar{x}$ は平均的に推定対象 μ をとる推定量であり，また，大数の法則を参照すると，n を大きくする（データを増やす）と推定対象 μ に近づくという推定量である．

μ の不偏推定量は他にもある．たとえば，x_2, $x_4,\ldots,x_{2\lfloor n/2\rfloor}$ のような添え字が偶数のものだけを用いた平均値 $\bar{x}_1 = \sum_{k=1}^{\lfloor n/2\rfloor}x_{2k}/\lfloor n/2\rfloor$ もそうである．ゆえに，$\bar{x}$ は μ の不偏推定量だからよい，とは言い切れないのである．そこで次に推定量の分散を比較することとなる．

3.2 最小分散（最良）線形不偏推定量

平均値 $\bar{x}$ の分散は，

$$V[\bar{x}] = V\left[\sum_{i=1}^{n}\frac{\varepsilon_i}{n}\right] = \frac{1}{n^2}\sum_{i=1}^{n}V[\varepsilon_i] = \frac{\sigma^2}{n}$$

である．また，同様にして，$V[\bar{x}_1] = \sigma^2/\lfloor n/2\rfloor$ である．よって，$\bar{x}_1$ よりも $\bar{x}$ の分散のほうが小さいことが

わかる．どちらの推定量も $\sum_{i=1}^n a_i x_i$ という形をしており，線形推定量と呼ばれる．$\bar{x}$ はすべての a_i が $1/n$ であり，$\bar{x}_1$ は奇数添え字の a_i を 0，偶数添え字の a_i を $1/\lfloor n/2 \rfloor$ としたものである．また，これらは線形推定量かつ不偏推定量なので，線形不偏推定量と呼ばれる．ここで，線形不偏推定量の中で，最小の分散をもつ推定量，すなわち，最小の分散を与える a_i，$i=1,\ldots,n$ を求めてみよう．

$\sum_{i=1}^n a_i x_i$ が不偏推定量でなければならないので，$E\left[\sum_{i=1}^n a_i x_i\right] = \mu$ でなければならない．すなわち，$\sum_{i=1}^n a_i = 1$ の下で $V\left[\sum_{i=1}^n a_i x_i\right]$ を最小にする a_i，$i = 1,\ldots,n$ を求めることとなる．$V\left[\sum_{i=1}^n a_i x_i\right] = \sum_{i=1}^n a_i^2 \sigma^2$ だから，

$$\begin{aligned} \min_{a_1,\ldots,a_n} \quad & \sum_{i=1}^n a_i^2 \\ \text{s.t.} \quad & \sum_{i=1}^n a_i = 1 \end{aligned}$$

という等式制約付き最小化問題を解けばよい．形式的ではあるがラグランジュ未定係数法を適用する．$\boldsymbol{a} = (a_1,\ldots,a_n)$ として，ラグランジュ関数を $L(\boldsymbol{a},\lambda) = \sum_{i=1}^n a_i^2 - \lambda(\sum_{i=1}^n a_i - 1)$ とする．よって，

$$\begin{aligned} &\frac{\partial L(\boldsymbol{a},\lambda)}{\partial a_i} = 2a_i - \lambda = 0, \qquad i = 1,\ldots,n \\ &\frac{\partial L(\boldsymbol{a},\lambda)}{\partial \lambda} = \sum_{i=1}^n a_i - 1 = 0 \end{aligned}$$

を解けばよい．$2\sum_{i=1}^n a_i = n\lambda$，および，$\sum_{i=1}^n a_i = 1$ より $\lambda = 2/n$ である．ゆえに，$a_i = 1/n$，$i = 1,\ldots,n$ であり，$\bar{x}$ が μ の線形不偏推定量の中で最小の分散をもつことがわかった．このような推定量を最小分散線形不偏推定量，もしくは，最良線形不偏推定量という．

μ の推定量として感覚を通じて $\bar{x}$ を選んだが，$\bar{x}$ はデータのモデル (2) において μ に対する最小二乗法で得られる推定量（最小二乗推定量）になっている．一般的な重回帰モデルにおいても，どの誤差も期待値および分散が 0 および σ^2（等分散）であり，どの二つの誤差を見ても無相関であるという条件の下で，最小二乗推定量は最良線形不偏推定量であることがガウス・マルコフの定理で示される．詳しくは佐和 [1] などを参照されたい．

最後に，(2) において，誤差の仮定を一般的にしたモデルに対して，μ の最良線形不偏推定量を求めてみる．すなわち，$V[\varepsilon_i] = \sigma_i^2$，$i = 1,\ldots,n$，$Cov[\varepsilon_i,\varepsilon_j] \neq 0$ $(i \neq j)$ とする．行列を用いたほうが便利なので，データを行列表記すると，$\boldsymbol{x} = (x_1,\ldots,x_n)'$ とすると，

$$\boldsymbol{x} = \mu\boldsymbol{1} + \boldsymbol{\varepsilon},\ E[\boldsymbol{\varepsilon}] = \boldsymbol{0},\ V[\boldsymbol{\varepsilon}] = \boldsymbol{\Sigma} \tag{3}$$

となる．ただし，$E[\boldsymbol{\varepsilon}]$ は期待値 $E[\varepsilon_i]$ を列ベクトルに並べたものであり，$V[\boldsymbol{\varepsilon}]$ は対角要素に分散 $V[\varepsilon_i]$，非対角要素に共分散 $Cov[\varepsilon_i,\varepsilon_j]$ を配置した対称行列である．よって，分散の性質より，$\boldsymbol{\Sigma}$ は正値定符号行列である（非負値定符号行列の場合もあるが，ここでは正定性を仮定している）．任意の線形推定量は $\boldsymbol{a}'\boldsymbol{x}$ である．不偏性より，$E[\boldsymbol{a}'\boldsymbol{x}] = \boldsymbol{a}'E[\boldsymbol{x}] = \mu\boldsymbol{a}'\boldsymbol{1}$ であるから $\boldsymbol{a}'\boldsymbol{1} = 1$ でなければならない．推定量の分散 $V[\boldsymbol{a}'\boldsymbol{x}] = \boldsymbol{a}'V[\boldsymbol{x}]\boldsymbol{a} = \boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a}$ より，μ の最良線形不偏推定量は

$$\begin{aligned} \min_{\boldsymbol{a}} \quad & \boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a} \\ \text{s.t.} \quad & \boldsymbol{a}'\boldsymbol{1} = 1 \end{aligned}$$

の解より得られる．ラグランジュ関数は，

$$L(\boldsymbol{a},\lambda) = \boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a} - \lambda(\boldsymbol{a}'\boldsymbol{1} - 1)$$

であり，

$$\begin{aligned} &\frac{\partial L(\boldsymbol{a},\lambda)}{\partial \boldsymbol{a}} = 2\boldsymbol{\Sigma}\boldsymbol{a} - \lambda\boldsymbol{1} = \boldsymbol{0} \\ &\frac{\partial L(\boldsymbol{a},\lambda)}{\partial \lambda} = \boldsymbol{a}'\boldsymbol{1} - 1 = 0 \end{aligned}$$

を解けばよい．第 1 番目の式に左から $\boldsymbol{\Sigma}^{-1}$ をかけると，$2\boldsymbol{a} - \lambda\boldsymbol{\Sigma}^{-1}\boldsymbol{1} = \boldsymbol{0}$ が得られ，さらに左から $\boldsymbol{1}'$ をかけて第 2 番目の式に代入すると，$2 - \lambda\boldsymbol{1}'\boldsymbol{\Sigma}^{-1}\boldsymbol{1} = \boldsymbol{0}$ を得る．よって，$\lambda = 2/\boldsymbol{1}'\boldsymbol{\Sigma}^{-1}\boldsymbol{1}$ である．これを，$2\boldsymbol{a} - \lambda\boldsymbol{\Sigma}^{-1}\boldsymbol{1} = \boldsymbol{0}$ に代入すると，

$$\boldsymbol{a} = \frac{\boldsymbol{\Sigma}^{-1}\boldsymbol{1}}{\boldsymbol{1}'\boldsymbol{\Sigma}^{-1}\boldsymbol{1}}$$

である．よって，μ の最良線形不偏推定量は

$$\hat{\mu} = \boldsymbol{a}'\boldsymbol{x} = (\boldsymbol{1}'\boldsymbol{\Sigma}^{-1}\boldsymbol{1})^{-1}\boldsymbol{1}'\boldsymbol{\Sigma}^{-1}\boldsymbol{x}$$

である．なお，これも (2) に対する最小二乗推定量と同様に，

$$(\boldsymbol{x} - \mu\boldsymbol{1})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \mu\boldsymbol{1})$$

を最小にする μ であり，(3) に対する一般化最小二乗法による推定量（一般化最小二乗推定量）でもある．

3.3 再び単回帰分析（最適計画）

2.1 節の単回帰分析では，n 個のデータ (x_i, y_i) が与えられ，それに線形式を当てはめることが目的であっ

た．ここでは，その逆を考えたい．よい線形式を得るためにはどのようなデータをとればよいかを求める．

よい線形式というのを，回帰係数の推定量の分散が小さいものと定義する．単回帰モデルは $p=1$ の重回帰モデルなので，回帰係数の推定量は $p=1$ とした $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ である．推定量の分散は，$V\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}\right] = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$ である．推定量の分散は行列であり，大小比較ができないので，$V[\hat{\boldsymbol{\beta}}]$ の行列式にして比較することをよく行う．すなわち，$|\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}|$ を最小にする $\boldsymbol{X}$ を求める問題である．$\boldsymbol{X}$ の中身は，データ (x_i, y_i) の x_i からなるので，$|\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}|$ を最小にするために，どの x で y を得ればよいかという問題となる．この問題は x を自由にどの点にでも取ることができなければ成立しないので，x は制御できる変数でなければならない．つまり，これは x のある点で実験を行い y を観測するという場面に相当する．このことから，望ましい x，すなわち，望ましい実験点の集合 $\boldsymbol{X}$ を決める問題を最適計画を求める問題という．通常，実験は実験ができる範囲があったり，また，理論的に考える場合でも，実験可能領域を定めないと際限がないので，閉区間 $[-1,1]$ を実験点の領域とすることが標準的である．なお，ここでの計画は “design” であり，数理計画でいう “programming” とは漢字は同じだが意味は異なる．ここで，$\min_{\boldsymbol{X}} |\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}|$ は $\max_{\boldsymbol{X}} |\boldsymbol{X}'\boldsymbol{X}|$ と同じであることに注意すると，最適計画を求める問題は

$$\begin{aligned} \max_{\boldsymbol{X}} \quad & |\boldsymbol{X}'\boldsymbol{X}| \\ \text{s.t.} \quad & \forall i,\ x_i \in [-1,1] \end{aligned}$$

である．この問題から得られる最適計画は，$\boldsymbol{X}'\boldsymbol{X}$ の行列式 *Determinant* を目的関数（最適性の基準）とするところから “D-最適計画” と呼ばれる．目的関数を $\mathrm{tr}(\boldsymbol{X}'\boldsymbol{X})^{-1}$，つまり，推定量の分散の和 $V[\hat{\beta}_0]+V[\hat{\beta}_1]$ を最小にする計画を考えるものもあり，この問題で得られる計画を “A-最適計画” という．この他にも，E-最適性や I-最適性など，さまざまな基準がある．

さて，1 次元ではあるが単回帰モデルについて D-最適計画を求めよう．以下，話を簡単にするため n は偶数という仮定を追加する．$|\boldsymbol{X}'\boldsymbol{X}|$ は

$$\begin{aligned} \begin{vmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{vmatrix} &= n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2 \\ &= n\sum_{i=1}^n (x_i - \bar{x})^2 \\ &= n\boldsymbol{x}'\left(\boldsymbol{I} - \frac{\boldsymbol{J}}{n}\right)\boldsymbol{x} \end{aligned}$$

と変形できる．$\boldsymbol{J}$ はすべての要素が 1 である n 次正方行列である．$(\boldsymbol{I}-\boldsymbol{J}/n)$ はべき等・対称行列であるので，$|\boldsymbol{X}'\boldsymbol{X}| = ||\,(\boldsymbol{I}-\boldsymbol{J}/n)\,\boldsymbol{x}||^2$ である．また，$(\boldsymbol{I}-\boldsymbol{J}/n)$ は $\mathbf{1}$ が張る空間に対する直交補空間上への射影行列なので，$|\boldsymbol{X}'\boldsymbol{X}|$ を最大にするためには，$\boldsymbol{x}$ を $(\boldsymbol{I}-\boldsymbol{J}/n)$ の列空間上に取ればよい．このことは，最適な $\boldsymbol{x}$ を $\boldsymbol{x}^*$ とすると，$\mathbf{1}'\boldsymbol{x}^* = 0$，すなわち，最適な $\boldsymbol{x}$ における x_j，$j=1,\ldots,n$ の平均値は 0 であるということである．よって，$\boldsymbol{x}^* = (x_1^*,\ldots,x_n^*)'$ とすると，$|\boldsymbol{X}'\boldsymbol{X}| = n\sum_{i=1}^n (x_i^*)^2$ かつ $|x_i^*| \le 1$，$\forall i$ であるのですべての i について $x_i^* = \pm 1$ となり，$\mathbf{1}'\boldsymbol{x}^* = 0$ を満たすものは，n を偶数としたので，「x_i の半分を -1 に，もう半分を $+1$」にした計画が単回帰モデルにおける D-最適計画である．すなわち，実験領域の端で実験を行うのがよいということである．このことは，1 次項のみの線形モデルに対して，$p \ge 2$ においても基本的に成り立つ．

実験をもっと意識すると，データ数 n は実験回数，説明変数の数 p は実験に取り上げる因子の数である．n と p に対して最適計画を求める方法もあるが，「$\boldsymbol{X}$ の第 1 列目を $\mathbf{1}$ とし，残りの要素が ± 1 であり，各列が直交する行列」H_s を使う方法もある．このような行列 H_s に従う実験は，すべての実験点が超立方体の頂点であり，D-最適計画である．すべての実験点が超立方体の頂点であるということは，どの因子も -1 および 1 の 2 条件（これを 2 水準という）の実験になるので実験が行いやすい．そこで，H_s タイプの n 次正方行列，すなわち，「第 1 列目が $\mathbf{1}$，残りの要素が ± 1 のみである直交行列」H を見つけたいのだが，この問題は簡単ではない．行列 H はアダマール行列と呼ばれ，$n=1,2$，それ以上は n が 4 の倍数のときに存在する．さらに，「行または列の入れ換え」および「行または列の符号の反転」によって同一になるものを同型とするとき，$n=1,2,4,8,12$ においては一つしかなく，$n=16$ は五つ存在する．たとえば，$n=1,2,4$ に対しては，

(1)
$$\underset{n=1}{\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}} \quad \underset{n=2}{\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix}}$$

である．また，アダマール行列の構成法はいくつかあり，その一つとして

$$H_{2^k} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \otimes H_{2^{k-1}},\ k = 1, \ldots$$

（Sylvester タイプ）

がある．$\otimes$ はクロネッカー積である．この方法では $n = 12$ や $n = 16$ の残りの四つは作れない．しかし，アダマール行列を実験の計画および統計解析に用いたとき，Sylvester タイプとそれ以外のタイプとでは，統計的な性質が異なる．

最適計画の問題を含め，実験を効率的に行い，情報を効果的に得るためには，どのような実験点を取るか（実験の計画），また，どのように解析を行えばよいかを示した方法を実験計画法という．実験計画法の良書は多くあるが，山田 [5] は，基礎的な方法から最適計画のようなアドバンスな方法まで，内容が豊富である．実験計画法において忘れてはならないのが田口玄一である．アダマール行列は直交配列表実験として，田口以前から使用されていたが，実験用にアダマール行列を使いやすく書き直す工夫をすることにより，日本においては直交配列表実験が専門家以外の人々に普及し，製造業では一般的な実験方法となった．直交配列表実験とは実験回数が制約された中から，最大限に情報を得るための実験計画であり，製造条件の最適化などに役立ってきた．また，田口は製品の使用環境で生じるであろうノイズを製品開発時の実験で発生させ，ノイズの影響を緩和して，ノイズに強い製品を設計するため手法であるロバストパラメータ設計 (Robust Parameter Design) を開発した．さらに田口は製品開発・技術開発に関わる新しい実験・データ解析技術を多く開発し，実験計画法や統計的方法にも影響を与えた．それらの手法群はタグチメソッドと呼ばれており，世界でも知られている．タグチメソッドの入門書として立林 [6]，実験計画法・統計的方法の側面においては，日本では宮川 [7]，世界では Wu and Hamada [8] などがある．

4. おわりに

統計的方法の背後には最適化理論がある．サポートベクターマシンのように，統計的機械学習の手法においては，凸計画問題が陽の形で現れている．しかし，最適化理論は解析法の基礎をなすものの，統計的方法はそれ以外の部分も，等しく重要であることにも注意したい．たとえば，得られたデータが観察によるものなのか，実験によるものなのかによって，同じ回帰分析でも，最終的に導くことができる結果に違いが出てくる．ランダムな順序で行った実験からは取り上げた因子と特性（データ）との因果関係を統計的に検証できるが，そうでない観察から得られたデータからは，基本的に予測しか保証されていない．また，母集団の規定，サンプリング方法なども解析目的へどれだけ迫れるかに影響を与える．解析法，その基礎技術である最適化理論は，統計的方法の中心に位置するが，その他の要素が適切であってこそ，解析の目的を達成できることに，最後に触れておきたい．

参考文献

[1] 佐和隆光，『回帰分析』，朝倉書店，1979.
[2] N. R. Draper and H. Smith, *Applied Regression Analysis*, 3rd edition, John Wiley & Sons, 1998.
[3] 小西貞則，北川源四郎，『情報量規準』，朝倉書店，2004.
[4] T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning, Data Mining, Inference, and Prediction*, 2nd edition, Springer, 2009.
[5] 山田秀，『実験計画法 ―方法編―』，日科技連出版社，2004.
[6] 立林和夫，『入門タグチメソッド』，日科技連出版社，2004.
[7] 宮川雅巳，『品質を獲得する技術』，日科技連出版社，2000.
[8] C. F. J. Wu and M. S. Hamada, *Experiments, Planning, Analysis, and Optimization*, 2nd edition, John Wiley & Sons, 2009.

