Technical Report YK-096
Sept. 18, 2026
Y. Karasawa

# 統計的推定の基本を学ぶ

## ～観測データを読み解く力を高めるために～

我々は物理的な空間に生きていても、背後にある仕組みや法則を直接見ることはできない。したがって、その成り立ちを理解するための手掛かりは、観測によって得られたデータに頼るほかない。 特に、複数の不確定要素が重なって生じる確率的な現象では、同じ環境条件を整えても異なる結果が得られる。そのため、観測データの統計的な性質を手がかりに、背後にある仕組みを推定する必要がある。このような状況で誤った結論を避けるために重要となるのが統計的推定である。 統計推定の方法は確率・統計学という数学理論に基づき、長い歴史の中で体系化されてきた。本レポートでは、その入り口として、得られたデータを見て誤った判断をしないための統計的手法の考え方をまとめる。 具体的には、① 最も確からしい値を求める最尤推定（点推定)、② 真の値が入り得る範囲を示す区間推定（信頼区間)、③ 複数のモデルの中から適切なものを選ぶための情報量基準（AIC）の入門編である。

### １．最尤推定

むかしむかし、ばらつきの大きい伝搬データに、ベテランがエイヤっと線を引いた。そこに名人芸を見た。でも、今の時代、これではいけない。間違った結論を出さないために、より良いモデルを見出すために統計的手法を身につけることが大事である。

#### 1.1 最尤推定とは

統計的推定とは、得られたデータから背後にある構造を統計的手法によって推定することである。その考え方（規範）の王道が最尤推定である。**最尤推定**（MLE）とは、結果を生み出す母集団の確率分布の形を定めた上で、観測されたデータの尤度（もっともらしさ；likelihood）が最大となるようなパラメータ値（母数）を推定する手法である。誤解されやすい言い方をあえてすれば、観測された結果は「その結果が最ももっともらしく見えるようなパラメータ値によって生じた」とみなして、そのパラメータを選ぶという考え方である。

#### 1.2 推定手順：母集団が正規分布である場合を例にして

具体的な推定手順を示すために、母集団の確率分布が平均$\mu$、分散$\sigma^2$の正規分布であるとしよう。正規分布であると言うことは既知、あるいは仮定としてスタートし、そのときのパラメータ$\mu$と$\sigma^2$の値（母数）を推定する。正規分布の確率密度関数(PDF)は次式で与えられる。

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \tag{1}$$

ステップ１：データを得る

得られたデータが $N$ 個, $x_1$〜$x_N$, だったとしよう。

ステップ２：尤度を求める

それぞれのデータに対する確率密度は $f(x_i)$ ($i$=1~$N$)である。独立試行によって $N$ 個のデータが生まれたと考え、その同時確率密度、すなわち、個々の確率密度の積 $L$ を求める。

$$L(\mu,\sigma^2)=\prod_{i=1}^{N} f(x_i)=\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^N \exp\left\{-\frac{1}{2\sigma^2}\sum_i^N\left(x_i-\mu\right)^2\right\} \tag{2}$$

上式で与えられる $L$ は、分布のパラメータ $\mu$ と $\sigma^2$ が変数になっており、**尤度関数**と呼ばれる。また、$L$ の値は**尤度** (likelihood)と呼ばれる。

ステップ３：尤度 $L$ の最大値（最大尤度）を与えるパラメータ $\mu$, $\sigma^2$ の値 $\hat{\mu},\hat{\sigma}^2$ を求め、これを母数の推定量（**最尤推定量**）とする。

$$\left(\hat{\mu},\hat{\sigma}^2\right)=\arg\max_{\mu,\sigma^2} L(\mu,\sigma^2) \tag{3}$$

最尤推定量を得る大局的なステップとしてはこれでお終いであるが、実際に求めようとすると、このままでは簡単ではない。そこで以下のステップに進む。

ステップ４：尤度関数の対数をとって**対数尤度関数** $l$ に代え、その**最大対数尤度**を得る最尤推定量を求める。最大値を与えるパラメータ値は尤度関数でも対数尤度関数でも変わらないので、解析を容易にするこの変換を行う。

$$\left(\hat{\mu},\hat{\sigma}^2\right)=\arg\max_{\mu,\sigma^2} L(\mu,\sigma^2)=\arg\max_{\mu,\sigma^2} l(\mu,\sigma^2) \tag{4}$$

$$\left(l(\mu,\sigma^2)\equiv\log L(\mu,\sigma^2)=\sum_{i=1}^{N}\log f(x_i)\right)$$

正規分布の場合で見てみよう。対数尤度関数は

$$l(\mu,\sigma^2)=-\frac{N}{2}\log\left(2\pi\sigma^2\right)-\frac{1}{2\sigma^2}\sum_i^N\left(x_i-\mu\right)^2 \tag{5}$$

であり、これを最大にするパラメータの最尤推定量は、以下の連立方程式で定められる。

$$\frac{\partial l(\mu,\sigma^2)}{\partial\mu}=\frac{1}{2\sigma^2}\sum_i^N\left(x_i-\mu\right)=0 \;\rightarrow\; \hat{\mu}=\frac{1}{N}\sum_i^N x_i=\bar{x} \quad \text{（標本平均）} \tag{6a}$$

$$\frac{\partial l(\mu,\sigma^2)}{\partial\sigma^2}=-\frac{N}{2\sigma^2}+\frac{N}{2\sigma^4}\sum_i^N\left(x_i-\mu\right)^2=0 \;\rightarrow\; \hat{\sigma}^2=\frac{1}{N}\sum_i^N\left(x_i-\bar{x}\right)^2 \quad \text{（標本分散）} \tag{6b}$$

結果だけ見ると、当たり前なものになっている。ただ、分散推定に関しては、教科書で習う不偏分散（(6b)式の 1/*N* の部分が 1/(*N*-1)に変わる）ではないのかと思うかもしれない。そうではない理由を[1]の 9 章 9.3.3 項末のティータイム欄で説明しているので、興味ある人は見てほしい。

最尤推定量をこのような綺麗な閉形式で示すことができるのは、ここで扱った正規分布やレイリー分布など極めて限られた分布になる。仲上-ライス分布、仲上m分布、ガンマ分布など、電波伝搬モデルに現れる大部分の分布については数式処理が難しく、数値解析で求めるしかない。

### 1.3 線形モデルにおける最尤推定と最小二乗法

量 $x$ に対して、出力 $y$ を生み出す物理現象があって、次式の線形モデルであるとしよう。

$$y = a_0 + a_1 x \tag{7}$$

実際にはこれに別の要因によって生まれる誤差（正規分布 $N(0,\ \sigma^2)$を仮定）が乗り、試行 $i$ における $x$ と $y$ の関係が$(x_i, y_i)$であったとしよう。この関係は誤差項$\varepsilon$を加えて次式で表される。

$$y_i = a_0 + a_1 x_i + \varepsilon_i \tag{8}$$

ここで、$N$ 組の標本値（$x_i, y_i; i=1\sim N$）から線形モデルの係数 $a_0, a_1$ を推定する問題を最尤推定で考えてみよう。

各 $y_i$ の条件付分布は

$$y_i \mid x_i \sim N(a_0 + a_1 x_i, \sigma^2) \ \rightarrow\ f(y_i \mid x_i, a_0, a_1, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{\left(y_i - a_0 - a_1 x_i\right)^2}{2\sigma^2} \right\}$$

試行が独立とすると、尤度関数 $L(a_0,a_1,\sigma^2)$は

$$L(a_0, a_1, \sigma^2) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{\left(y_i - a_0 - a_1 x_i\right)^2}{2\sigma^2} \right\}$$

であるから、対数尤度関数は次式になる。

$$l(a_0, a_1, \sigma^2) = \log L(a_0, a_1, \sigma^2) = \sum_{i=1}^{N} \left\{ -\frac{1}{2}\log\left(2\pi\sigma^2\right) - \frac{\left(y_i - a_0 - a_1 x_i\right)^2}{2\sigma^2} \right\}$$

上式右辺の第 1 項は定数で最尤推定量には影響しないので無視でき、最尤推定量は

$$\begin{aligned}(\hat{a}_0, \hat{a}_1) &= \arg\max_{a_0, a_1} l(a_0, a_1, \sigma^2) = \arg\max_{a_0, a_1} \left\{ -\frac{1}{2\sigma^2} \right\} \sum_{i=1}^{N} \left(y_i - a_0 - a_1 x_i\right)^2 \\ &= \arg\min_{a_0, a_1} \sum_{i=1}^{N} \left(y_i - a_0 - a_1 x_i\right)^2 \end{aligned} \tag{9}$$

となる。

上式最終辺の形は**最小二乗法**で定義される推定量そのものである。ゆえに、線形近似でかつ誤差が正規分布すると言う条件においては、最尤推定と最小二乗法による推定は同じ結果を導く、すなわち、正規分布誤差の場合は、最小二乗推定は最尤推定であると言うことである。このようにして得られた直線

$$\hat{y} = \hat{a}_0 + \hat{a}_1 x \tag{10}$$

は**回帰直線**と呼ばれる。

ここでは、最尤推定と最小二乗推定の関係を線形モデルの例で示したが、(10)式を多項式近似や任意の非線形モデルで置き換えることもできる。その際、誤差を正規分布で与える限りにおいて、最小二乗推定は最尤推定を維持している。

## ２．区間推定

### 2.1 区間推定とは

先に述べたように、最尤推定は、結果を生み出す母集団の確率分布の形を定めた上で、観測されたデータの尤度（もっともらしさ）が最大となるようなパラメータ値を推定する手法である。しかし、実際に得られたデータは母集団からのランダムサンプルであり、偶然によって推定値が真の値（母数）から大きく外れる可能性がある。 そこで、推定量が標本としてどのようにばらつくか（標本分布）を利用し、推定値を中心に「真の値を一定の信頼度で含むと考えられる範囲（**信頼区間**）」を構成する。このように、点推定だけでは判断を誤る可能性があるため、不確実性を幅として表現する推定法が**区間推定**である。

区間の幅は、あらかじめ定めた確率 $p$（信頼係数）によって決まる。このとき得られる区間は$p$%信頼区間と呼ばれる。直感的には「母数が $p$%の確率で存在する範囲」と言いたくなるが、頻度論的な意味ではこれは正確ではない。実際には、データセットを多数回取得し、その都度推定量から信頼区間を構成したとき、母数がその区間に含まれる割合が $p$%になる、という意味である。したがって、信頼区間は「推定方法の長期的な性質」を表すものであり、特定の一回の区間に確率を割り当てているわけではない。信頼係数としては 90%、95%、99% などが用いられるが、実務では 95% がよく採用される。信頼係数を大きくすると（例 : 99.9%）、確実性は高まる一方で、臨界値が大きくなるため区間幅が広くなり、特性の判断が難しくなる。

### 2.2 回帰直線と信頼区間

回帰直線については 1.3 節で説明したが、最尤推定値そのものも標本データセットの違いによってばらつきを持つ。このばらつきの大きさを求めるために、行列表示を用いて説明する。

$$y = X\beta + \varepsilon, \quad X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_N \end{pmatrix}, \quad \beta = \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} \tag{11}$$

最尤推定量は

$$\hat{\beta} = \left(X^{\mathrm{T}}X\right)^{-1} X^{\mathrm{T}} y \tag{12}$$

その分散の共分散行列は次式である。

$$\mathrm{Var}\left(\hat{\beta}\right) = \sigma^2 \left(X^{\mathrm{T}}X\right)^{-1}$$

$$X^{\mathrm{T}}X = \begin{pmatrix} N & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}, \ \left(X^{\mathrm{T}}X\right)^{-1} = \frac{1}{N\sum x_i^2 - \left(\sum x_i\right)^2} \begin{pmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & N \end{pmatrix}$$

これより、切片 $\hat{a}_0$ と傾き $\hat{a}_1$ の分散および $\hat{a}_0$ と $\hat{a}_1$ の共分散はそれぞれ次式になる。

$$\mathrm{Var}\left(\hat{a}_0\right) = \sigma^2 \left(\frac{1}{N} + \frac{\bar{x}^2}{S_{xx}}\right)$$

$$\mathrm{Var}\left(\hat{a}_1\right) = \frac{\sigma^2}{S_{xx}}$$

$$\mathrm{Cov}\left(\hat{a}_0, \hat{a}_1\right) = -\frac{\sigma^2 \bar{x}}{S_{xx}}$$

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i, \quad S_{xx} = \sum_{i=1}^{N} \left(x_i - \bar{x}\right)^2$$

これより、回帰直線上の $x$ に対する値 $\hat{y}$ の分散は、線形結合の分散公式

$$\mathrm{Var}\left(AX + BY\right) = A^2\mathrm{Var}(X) + B^2\mathrm{Var}(Y) + 2AB\,\mathrm{Cov}(X,Y)$$

を用いて、以下のように得られる。

$$\mathrm{Var}\left(\hat{y}(x_0)\right) = \mathrm{Var}\left(a_0 + a_1 x_0\right) = \sigma^2 \left(\frac{1}{N} + \frac{\bar{x}^2}{S_{xx}}\right) + x_0^2 \left(\frac{\sigma^2}{S_{xx}}\right) + 2x_0 \left(-\frac{\sigma^2 \bar{x}}{S_{xx}}\right)$$

$$= \sigma^2 \left( \frac{1}{N} + \frac{\left(x_0 - \bar{x}\right)^2}{S_{xx}} \right) \tag{13}$$

式中の母分散$\sigma^2$は知りえないものなので、標本データからの推定値（不偏分散）を用いる。

$$\sigma^2 \approx \hat{\sigma}^2 = \frac{1}{N-2} \sum_{i=1}^{N} \left( y_i - \hat{a}_0 - \hat{a}_1 x_i \right)^2 \tag{14}$$

ここで$N$-2 で割っているのは、本来ある自由度$N$のうち、$a_0$と$a_1$の推定に二つの自由度が使われ残りの自由度が$N$-2 になっているためである。（正規分布データの推定では、不偏分散は$N$-1 で割るが、その場合は平均値を推定するために自由度が 1 失われていると解釈するのと同じ理屈）。母分散$\sigma^2$を上記不偏分散$\hat{\sigma}^2$で置き換えると回帰直線上$x$の点における$\hat{y}$の分散は次式となる。

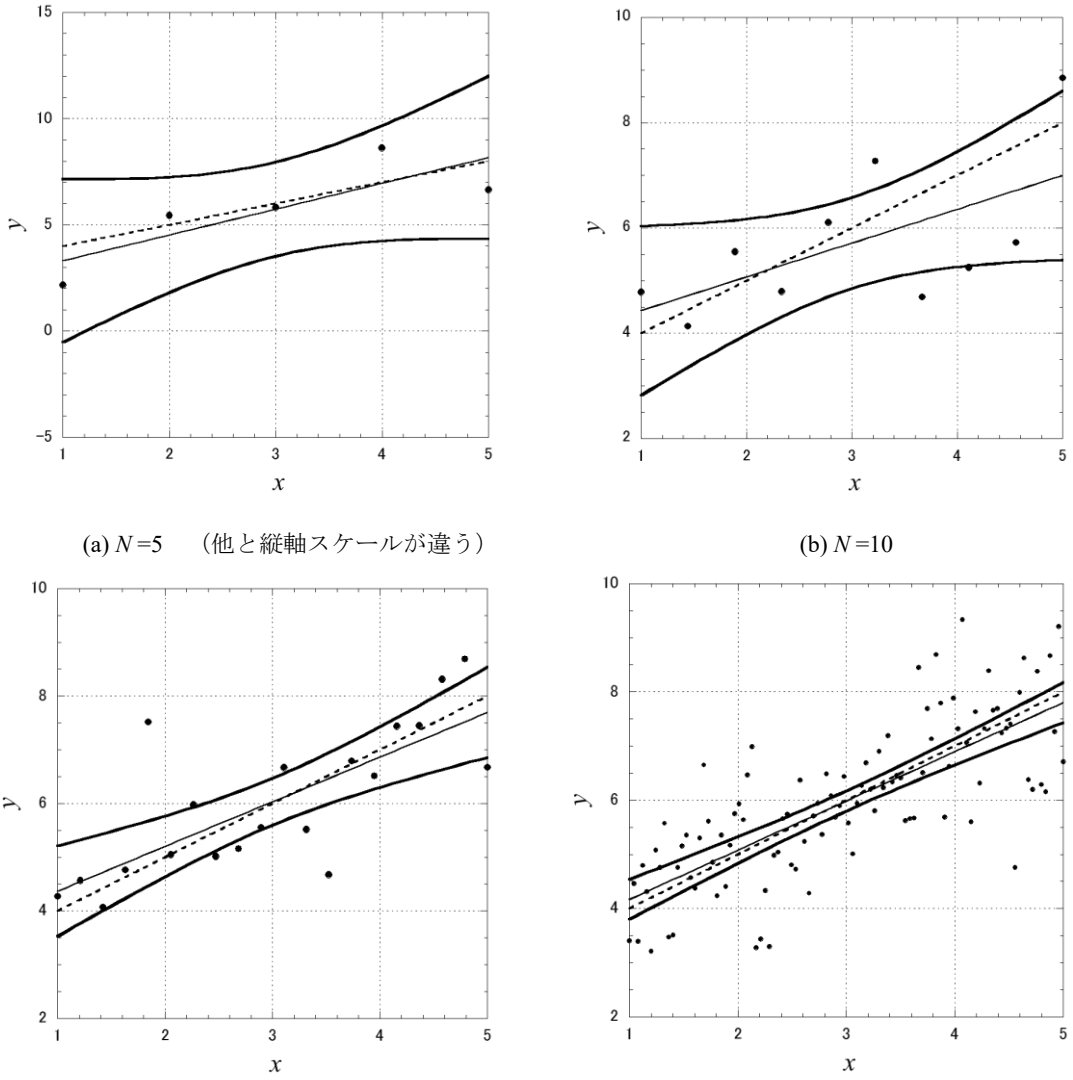
$$\mathrm{Var}\left(\hat{y}(x)\right) \approx \hat{\sigma}^2 \left( \frac{1}{N} + \frac{\left(x - \bar{x}\right)^2}{S_{xx}} \right) \tag{15}$$

$\hat{y}$の分布は、分散に母分散$\sigma^2$を使えば正規分布であるが、不偏分散$\hat{\sigma}^2$を使っていて不偏分散そのものが分布を持つので、自由度$N$-2 の t 分布になる。故に確率$p$%の範囲に入るのは累積確率で(1-$p$/100)/2（下限）～0.5+$p$/200（上限）の範囲になる。この範囲が$p$%信頼区間であり、次式で与えられる。

$$\begin{pmatrix} y_{\max} \\ y_{\min} \end{pmatrix} = \hat{a}_0 + \hat{a}_1 x \pm t(N-2,\alpha) \sqrt{\hat{\sigma}^2 \left( \frac{1}{N} + \frac{\left(x - \bar{x}\right)^2}{S_{xx}} \right)} \tag{16}$$

ここで$t$($N$-2,$\alpha$)は、自由度$N$-2 の t 分布変量で累積確率$\alpha$=0.5+p/200 の値（$p$=95 では$\alpha$=0.975）である。（分布形は$t$=0 を中心に対称）。t 分布は$N$の値が大きくなるほど標準正規分布$N$(0,1)に近づく。$p$ =95%の$t$の値は、$N$=10, 30, 100 に対して、2.23, 2.04, 1.99 であり、$N$=∞の漸近値（標準正規分布）は 1.96 である。

回帰分析の具体例を示す。元の特性は、$a_0$ =3, $a_1$ =1, $\sigma^2$ =1 で、$x$は 1～５で等間隔に$N$ =5, 10, 20, 100 としている。図 1 はこの分析結果を示す。図には、これまで述べてきた多くの特徴が表れており、それを感じ取ってほしい。信頼区間は、真の特性がこの範囲に存在する確率が高い部分を示す幅であって、個々の観測値$y_i$が分布する幅（ばらつき幅）ではないことに注してほしい（もちろん、区間幅にはばらつきの大きさが反映されているのではあるが）。




(a) $N$=5　（他と縦軸スケールが違う）　(b) $N$=10

(c) $N$=20　(d) $N$=100

図 1　回帰直線と 95%信頼区間（データ： $a_0$=3, $a_1$=1, $\sigma^2$=1; グラフ：点線：母集団特性、実線：回帰直線、太実線曲線：母集団特性の 95%信頼区間の上限と下限）

### 2.3 さらに理解を深めるために

#### （1）モデル選択の重要性

2.2 節では直線回帰の方法について述べたが、そもそも「どのモデルを採用するか」という判断が解析の出発点として極めて重要である。誤ったモデルを選んでしまうと、苦労して導いた結論が誤ったものになりかねない。そのため、散布図をよく観察し、データの特徴を見誤らないことが大切である。この心構えを示す例として、アンスコム（Anscombe, 1973）の有名な四つのデータセットがある（[1]の 10 章図 10.5）。これらは平均・標準偏差・相関係数・回帰直線がすべて同じであるにもかかわらず、散布図は全く異なる形を示す。アンスコムは、統計量だけに頼らず、散布図を必ず確認することの重要性を強調している。注意すべき点として、

1) 曲線的な関係があるデータに直線モデルを当てはめてしまう（モデル選択の誤り）。

2) ほぼ直線状のデータの中に一つだけ外れ値があり、それが回帰直線を大きく歪める場合がある。外れ値の原因を吟味する必要がある。
3) 外れ値ではないが、離れた一点が回帰直線に強い影響を与え、統計的性質を正しく反映しない場合がある。

繰り返しになるが、回帰分析では統計量だけで判断するのではなく、散布図を見てデータの傾向や外れ値の有無を確認することが不可欠である。

### （2）$x$ と $y$ の関係

線形回帰では、$x$ を原因、$y$ を結果とみなし、誤差は $y$ 軸方向にのみ発生すると仮定した。そのため、$x$ と $y$ が同じ確率分布を持っていても、回帰直線は 45° にはならず、相関係数に応じて傾きが緩やかになる（[2] の図 4.2）。

しかし、$x$ と $y$ に明確な原因・結果の関係がない場合、両者とも誤差を含むことになる。例えば、$x$ と $y$ の誤差分散が同程度であれば、誤差は $y$ 軸方向ではなく「点と回帰直線との距離」で評価する必要がある。このように、回帰分析では横軸と縦軸の関係性（主従関係の有無）を慎重に吟味することが重要である。

### （3）補足

本節では、回帰分析の基本である二変量の直線回帰（単回帰モデル）と、その信頼区間の定め方について自己完結的にまとめた。次のステップとしては、多項式・指数・対数などの一般関数への回帰分析、さらに多変量解析へと進むことになる。特に多変量解析は、ビッグデータが中心となる現代の AI 技術を支える基盤であり、データを見る目（＝分析技術）を養うことは非常に重要である。

最近の数値処理ソフト（Excel、KaleidaGraph など）には回帰分析機能が備わっており、計算自体は容易に行える。しかし、その結果の妥当性や物理的意味を正しく判断するためには、基礎原理を理解しておくことが不可欠である。

## ３．より良いモデルを選ぶための規範：情報量基準 AIC

最尤推定は、まずモデルの型を定め、そのパラメータ（母数）を推定する方法である。しかし、採用するモデルが異なれば推定結果も変わり、結論が変わる可能性がある。では、複数の候補モデルがあるとき、どのモデルを選ぶべきなのか。その選択の規範は何か。最大尤度が大きいモデルを選べばよいのか。この問いに答えるのが情報量基準であり、その代表が **AIC**（Akaike Information Criterion；**赤池情報量基準**）である。

AIC は、赤池弘次博士が 1970 年代初頭に数理統計学の理論を基盤として提案したモデル選択の枠組みである。無線通信、とくに自然現象を扱う電波伝搬では統計的モデリングが重要であり、適切なモデルの選択がシステム設計の高信頼化・効率化に大きく寄与する。AIC は、モデル評価の考え方も評価式そのものも非常にシンプルでありながら、応用範囲が極めて広い。また、

AIC を規範として、より良いモデル構築へと発展させることもできる。

本章では、この AIC の入り口部分を解説する。「AIC の使い方だけ知っておけば十分」という姿勢ではなく、かといって数学的に完全な理解を求めるものでもなく、直観的に理解できるレベルを目標としてまとめている。より詳しくは筆者の資料 [1] の第 9 章を参照してほしい。基礎から本格的に学びたい場合は、情報量基準の正統的な教科書 [3]、AIC の歴史的経緯を詳述した文献 [4] が参考になる。

### 3.1 最小二乗近似の回帰曲線で何が問題か

真の特性 $g(x)$（図２中の実線のカーブ）があり、$x_i$ でのサンプル毎に誤差が乗り、$y_i$ として得られたものが、同図に黒丸で示すデータであったとする。すなわち、

$$y_i = g(x_i) + \varepsilon_i \quad (i = 1, 2, \cdots, n)$$

である。目的は真の特性 $y=g(x)$の形を知りたいのだが、手持ちの情報は、図に Data として示した$(x_1,y_1)$, $(x_2,y_2)$, …, $(x_n,y_n)$の $n$ 組（図では $n$=11）のデータのみである。そこで、真の特性を予想するものとして、次式で与えられる $K$ 次の多項式近似を考える。

$$f(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_k x^K$$

この場合、次数 $K$ と係数 $a_0$〜$a_K$ を定めることになる。次数を仮決めすると、係数 $a_0$〜$a_K$ は次式で求める最小二乗法で定められる（前述のように最尤推定でもある）。最小二乗法はモデルと実現値の残差の平方和を最小にする規範で、

$$J = \sum_{i=1}^{n} \left( f(x_i) - y_i \right)^2$$

の $J$ が最小になるように $f(x)$を、すなわち $a_0$〜$a_K$ を定めればよい（その結果を $\hat{a}_0 \sim \hat{a}_K$ で表記）。最小二乗法による多項式近似の係数の定め方は市販の数値計算ソフト（Excel, KaleidaGraph など）に組み込まれているので、ここではそれは述べない。このようにして定められたカーブ

$$\hat{y} = f(x \mid \hat{a}_0, \hat{a}_1, \hat{a}_{2,} \cdots, \hat{a}_K)$$

は**回帰曲線**と呼ばれる。$K$=1 の場合は直線なので、**回帰直線**である。既に述べてきたように、回帰曲線は真のカーブ $g(x)$そのものではなく、真のカーブが回帰曲線を中心に $y$ 軸方向に正規分布すると予測する意味になる。

図３は、図２の Data （Data 1 と表記）に対して $K$=2, 4, 7 次で求めた回帰曲線を示している。図４は２種類の標本値（Data 1 と Data 2；Data 2 は Data 1 と誤差が独立な標本値）と Data 1 に対する回帰曲線との残差の平方和 $J$ を示している。これより、回帰曲線を作った Data 1 に対しては次数を上げるほどデータに近づいてゆく様子が分かる。では、推定において次数は高ければ高

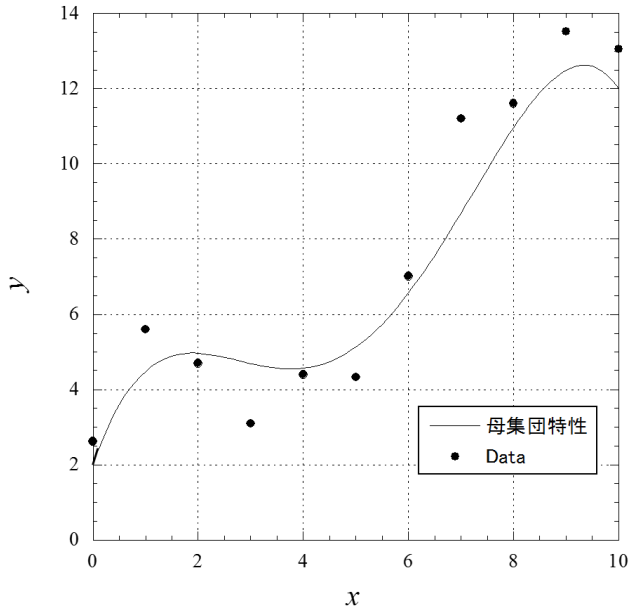



図２　実線で表される母集団特性（真の特性 $g(x)$）に誤差を含んで現れる標本値（Data）の例

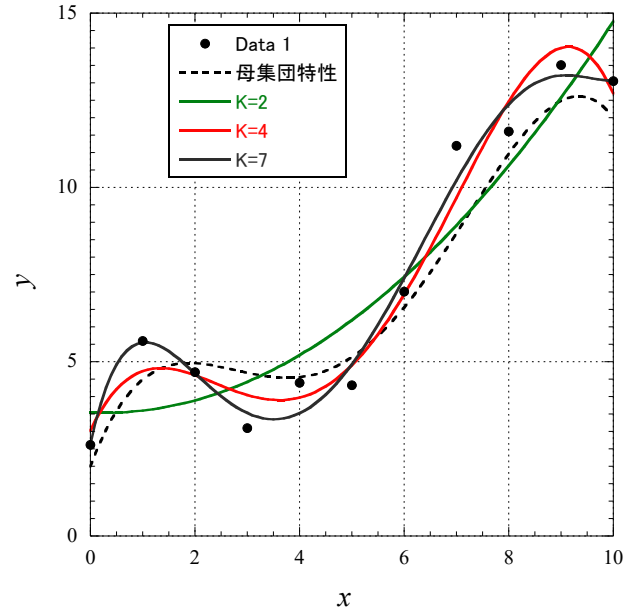



図３　標本値 Data 1（図２の Data と同じ）に対する回帰曲線（$K$=2, 4, 7）

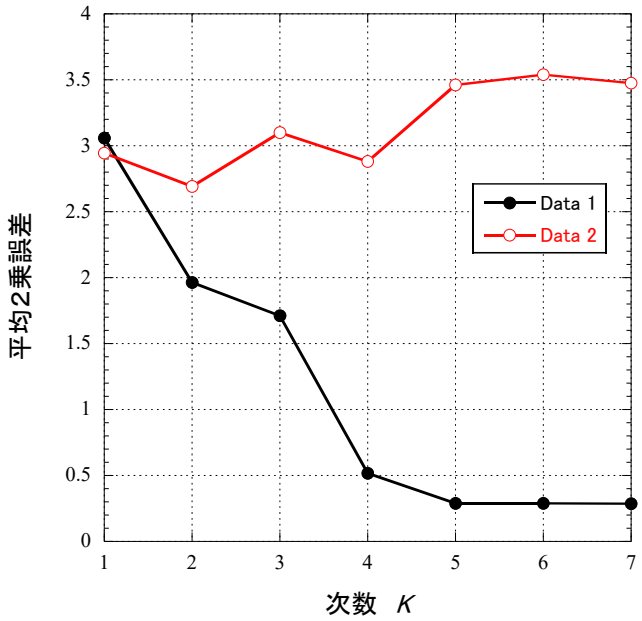



図４　Data 1 で作った回帰曲線に対する Data 1 と Data 2（Data 1 と誤差独立）の平均２乗残差の次数依存性

いほどよいのであろうか？別の言葉で言うと、再度同じ現象からのデータが得られたとき最初のデータで求めた回帰曲線は、新たなデータに対しても有効に働いているのかどうか、である。図４の Data 2 をみるとむしろ次数が高くなると残差が増える傾向が読み取れる。（注：乱数を替えて行うと、特に Data 2 については様々な結果が現われるので、図４は一例であるが、次数 $K$ が大きくなるところで誤差が増大する傾向は類似している。）

これより、次数が低すぎるとデータの構造を適切に表現できないが、高すぎるとデータの偶然変動に過敏になり、将来の現象予測にも誤った知見を与えることになる。この例にみられるように、データへの適合度とモデルの複雑さ（次数選択問題）をどのように折り合いをつけるかが鍵になるが、そのことに対して最小二乗法は答えを与えてくれない。モデルの評価の視点で見ると、Data 1 では標本値でモデルを作り、その標本値でモデルの精度を評価しているから、精度において過大評価になることは、大いに予測できることである。

これから学ぶ AIC は、このようなときのモデル選択の判断基準を与えるものである。よいモデルを作るためには次数（パラメータ数）の選択が鍵になりそうなことが分かったが、では、AIC はこれにどのように答えてくれるのであろうか？

### 3.2　二つの確率分布の近さを測る尺度：カルバック-ライブラー情報量

一方を基準とする確率分布 $g(x)$、もう一方を評価したい確率分布 $f(x)$とする。この二つの分布の近さを定量的に評価する尺度が求められる。たとえば、対数正規分布が、他の関数と組み合わせて使う場合の扱いにくさために、解析性に優れた、たとえば、ガンマ分布で置き換えたいというような場合である。あるいは、実測値の分布を得た場合に、それを特定の分布で近似したいと

いうような場合である。このような時、二つの分布の近さを測る尺度が必要になる。

最も簡単な尺度は分布の距離を求めるノルム（$|g(x)-f(x)|$や$\{g(x)-f(x)\}^2$の期待値）であるが、分布の値の大きい部分の差が結果を支配するので、確率の小さい部分も含めて全体を見たい場合には、次式で定義される**カルバック・ライブラー情報量**（Kullback-Leibler (KL) divergence ; KL 情報量）のほうが望ましい。

$$D(g:f)=\int_{-\infty}^{\infty} g(x)\log\left\{\frac{g(x)}{f(x)}\right\}dx \tag{17}$$

２つの分布 $g$ と $f$ が同じ場合に $D$=0、異なる場合に $D$>0 となり、ずれの程度に応じて値も大きくなる（$D\geq 0$ の証明は[3], p.28）。例えば、分布 $g$ の置換目的であれば、置き換えたい分布 $f$ のパラメータ値を調整して、$D$ が最も小さくなるように最適化するための規範として用いられる。

式の形を見ると、比が求められるので、ノルム評価とは逆に確率の値の小さい部分の影響を大きく受けそうにも見えるが、$g(x)$がかかって値の小さい部分の影響は押さえられているので、バランスが保たれている。(17)式を書き直すと

$$\int_{-\infty}^{\infty} g(x)\log\left\{\frac{g(x)}{f(x)}\right\}dx=\int_{-\infty}^{\infty} g(x)\log g(x)\,dx-\int_{-\infty}^{\infty} g(x)\log f(x)\,dx \tag{18}$$

となる。右辺第１項に負符号をつけたものが確率分布 $g$ のエントロピー、第２項の負符号を含んだものは**交差エントロピー**と呼ばれる量である。すなわち、

**KL 情報量 =［確率分布 $g$ と $f$ の交差エントロピー、］-［確率分布 $g$ のエントロピー］**

と整理される。

では、モデルの優劣も KL 情報量で評価すればよいと言うことになるのであるが、今の土俵では、肝心の母集団特性を表す $g$ が未知なため比較の基準がない。$g$ に代るものは何か、AIC の出発点がここになる。

### 3.3　AIC の概要

（１）出発点と目的地：KL 情報量ではなぜだめか？

AIC 構築の背景をなす大きな流れは以下である。

１）未知の母集団 $g(x|\theta_0)$から無作為に抽出した一組の標本がある：$x_1, x_2, \ldots, x_n$

２）その仕組みを推定するモデル（＝評価したいモデル）がある：$f(x;\boldsymbol{\theta})$ （複数のパラメータを想定しベクトル$\boldsymbol{\theta}$で表記）

３）AIC でモデルを評価する（複数のモデルが有る場合には AIC の数値を比較して良い方のモデルを選択する）

もし、母集団の特性（確率分布 $g(x|\theta=\theta_0)$）が分かっていれば、モデルの良さは KL 情報量 $D(g,f)$ で評価できる。しかし、それ（すなわち $g$）がわかっていないからどうするかが、AIC を生み出

す目的になる。すなわち、$g(x|\theta_0)$を知らずに、標本値だけで $f(x)$の良さをどう評価するか、と言う問題設定である。

（２）KL 情報量の中のモデル依存項に注目

KL 情報量は、式(17), (18)より、

$$D(g,f)=\int_{-\infty}^{\infty} g(x)\log g(x)\,dx-\int_{-\infty}^{\infty} g(x)\log f(x)\,dx$$

で表されたが、右辺の第一項（確率分布 $g$ のエントロピーの符号を反転させたもの）は、評価したい確率分布 $f$ に関係しない項であり、種々のモデル $f$ に対する比較をしたい場合には共通項として相殺されるものである。故に、右辺第２項の大小関係のみを調べることができれば、モデルの良さに関する比較評価が可能になる。すなわち、

$$D_0\equiv\int_{-\infty}^{\infty} g(x)\log f(x\,|\,\boldsymbol{\theta})\,dx=\left\langle \log f(x\,|\,\boldsymbol{\theta})\right\rangle_g \tag{19}$$

に着目すればよいということである。(18)式ではこの項に負符号をつけたものを交差エントロピーと呼んだが、次項ではこの量に着目する。しかし、この式にも未知の関数 $g(x)$が含まれているので、問題が解決されているわけではない。

（３）平均対数尤度と最大対数尤度：求めたいものと求められるもの

以下の２つの式を定義する。

$$la(\boldsymbol{\theta})\equiv nD_0=n\int_{-\infty}^{\infty}\log f(x\,|\,\boldsymbol{\theta})\,g(x)dx \tag{20}$$

$$l(\boldsymbol{\theta})\equiv\sum_{i=1}^{n}\log f(x_i\,|\,\boldsymbol{\theta}) \tag{21}$$

(20)式は (19)式と同じ形であるが、パラメータ $\boldsymbol{\theta}$ の関数として見ていて、平均対数尤度の $n$ 倍値である。一方、(21)式は対数尤度である。

(21)式を最大にするパラメータの最尤推定値 $\hat{\boldsymbol{\theta}}$ を求め、(20), (21)式に代入すると

$$la(\hat{\boldsymbol{\theta}})=n\int_{-\infty}^{\infty}\log f\left(x\,|\,\hat{\boldsymbol{\theta}}(x_1,x_2,\cdots,x_n)\right)g(x)dx \tag{22}$$

$$l(\hat{\boldsymbol{\theta}})=\sum_{i=1}^{n}\log f\left(x_i\,|\,\hat{\boldsymbol{\theta}}(x_1,x_2,\cdots,x_n)\right)\quad\text{（最大対数尤度）} \tag{23}$$

である。(23)式は最大対数尤度、(22)式は、厳密には「パラメータを最尤推定値で固定した場合の平均対数尤度の $n$ 倍値」であるが、以下ではこの値自体を「平均対数尤度 $n$ 倍値」と呼ぶ。こ

の二つについて、次のことが言える。

1）平均対数尤度 $n$ 倍値 $la$ は $n$ が十分大きいとき、大数の法則により、

$$\lim_{n\to\infty} la(\hat{\boldsymbol{\theta}}) = n\int_{-\infty}^{\infty} \log f(x\,|\,\boldsymbol{\theta}_0)\, g(x)dx$$

となり、求めたい量に確率収束する。このため、平均対数尤度 $n$ 倍値が求められれば、モデル間の比較評価が可能である。しかし、$g$ を知らないのでこれを求めることはできない。

2）一方、最大対数尤度 $l$ は標本値とモデルから計算できる。平均対数尤度 $n$ 倍値と最大対数尤度の関係が分かれば、最大対数尤度からモデル評価が可能になる。ただし、$n$ が十分大きい極限において両者が等しいのかどうかは式の上からは判然としない。

そこで、以下、最大対数尤度 $l(\hat{\boldsymbol{\theta}})$ と平均対数尤度 $n$ 倍値 $la(\hat{\boldsymbol{\theta}})$ の違いを調べる。

### （4）最大対数尤度 $l(\hat{\boldsymbol{\theta}})$ と平均対数尤度 $n$ 倍値 $la(\hat{\boldsymbol{\theta}})$ の差

平均対数尤度 $n$ 倍値 $la(\hat{\boldsymbol{\theta}})$ が分かれば、モデル評価に使える。しかし、それはわからず、最大対数尤度 $l(\hat{\boldsymbol{\theta}})$ なら計算できる。その差（補正値）はどのくらいか、それを定量的に示したのが AIC の真髄である。そして、その差はパラメータの数で近似できるという驚くべき結果を得るのである。

補正値の理論解析は数理統計学の知識を必要とし、筆者には手に負えない部分がある。そこで、この導出は文献[3]（第３章：一般的な導出過程が示されている）をみて学んで欲しい。（ページを割いてその部分をここに再掲するより、オリジナルを読んでもらったほうがよいと判断）。以下、その概要のみを述べる。

補正項＝最大対数尤度 – 平均対数尤度 n 倍値

$$= l\left(\hat{\boldsymbol{\theta}}(x_1, x_2, \cdots, x_n)\right) - la\left(\hat{\boldsymbol{\theta}}(x_1, x_2, \cdots, x_n)\right)$$

$$= \sum_{i=1}^{n} \log f\left(x_i\,|\,\hat{\boldsymbol{\theta}}(x_1, x_2, \cdots, x_n)\right) - n\int_{-\infty}^{\infty} \log f\left(x\,|\,\hat{\boldsymbol{\theta}}(x_1, x_2, \cdots, x_n)\right) g(x)dx$$

であり、これを３つの部分に分解して、それぞれの期待値（<・>）を求める。すなわち、

$$\langle \text{補正項} \rangle = \left\langle l\left(\hat{\boldsymbol{\theta}}(x_1, x_2, \cdots, x_n)\right) - l\left(\boldsymbol{\theta}_0\right)\right\rangle + \left\langle l\left(\boldsymbol{\theta}_0\right) - la\left(\boldsymbol{\theta}_0\right)\right\rangle + \left\langle la\left(\boldsymbol{\theta}_0\right) - la\left(\hat{\boldsymbol{\theta}}(x_1, x_2, \cdots, x_n)\right)\right\rangle$$

である。このとき、右辺第２項が 0 になることは容易に導かれる。一方、右辺の第１項と第３項の導出については種々の近似が必要になる。その結果、最終的に、第１項と第３項については

$$\left\langle l\left(\hat{\boldsymbol{\theta}}(x_1, x_2, \cdots, x_n)\right) - l\left(\boldsymbol{\theta}_0\right)\right\rangle \approx \left\langle la\left(\boldsymbol{\theta}_0\right) - la\left(\hat{\boldsymbol{\theta}}(x_1, x_2, \cdots, x_n)\right)\right\rangle \approx \frac{\text{パラメータ数}}{2}$$

となる。結果として、

$$\text{最大対数尤度の補正項} \approx \text{パラメータ数} \tag{24}$$

である。ここまでシンプルな結果が得られるのなら、もっと直感的に説明ができるのではと思うが、筆者はその方法を知らない。

なお、この結果が得られるためには、モデル $f(x)$が真の特性 $g(x)$を含むかそれに近い関係（図３に示した例題はまさにそういう典型例）のときである等の条件もある[3]。通常は、真実とあまりにかけ離れたモデルを想定すると言うことは無いであろうから、(24)式は大部分のケースに適用できると期待してよい。

（６）AIC のまとめ

以上をまとめて AIC は以下のように整理されている。

$$\text{AIC} = -2\times(\text{最大対数尤度}) + 2\times(\text{パラメータの数}) \tag{25}$$

$$\text{最大対数尤度}：l(\hat{\boldsymbol{\theta}}) = \sum_{i=1}^{n} \log f\left(x_i \mid \hat{\boldsymbol{\theta}}(x_1, x_2, \cdots, x_n)\right)$$

AIC では、この値が小さいほど良いモデル（＝選択すべきモデル）と言う規準を与えている。パラメータを多くすると対数尤度を大きくできるが、パラメータが増えること自体がペナルティであると解釈され、統計的モデリングにおいてはバランスの大事さ、すなわち、誤差が同じ程度ならパラメータ数の少ないモデルを選ぶべきということが強調される結果になっている。

図５は AIC による一連の評価ステップをまとめている。

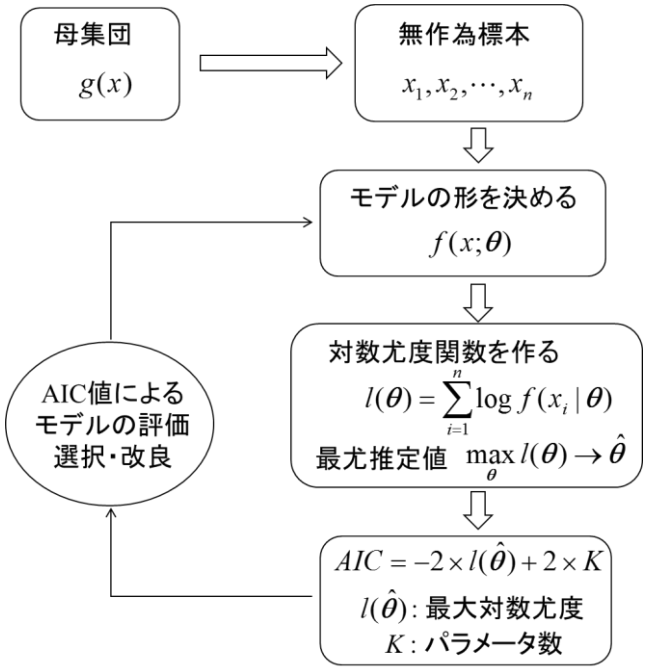



図５　AIC によるモデルの評価ステップ（評価・選択・改良）

### 3.4 AIC の応用例

得られた複数の標本値から、モデルとなる確率分布形を定め、そのパラメータの最尤推定値を得、そのモデルの良さを AIC で評価する方法は、図５で示した手順をそのまま実行すればよい。それを実際に行うときのネックは、モデルパラメーターの最尤推定量を標本値から定めるとき、それが解析的に求められない場合、コンピュータによる数値解法（ニュートン・ラフソン法などによる反復計算法）によらなければならない点である。正規分布やレイリー分布、指数分布のように解析解が得られる場合はよいが、ガンマ分布や仲上 m 分布など大部分は数値解法になる。この不便さは、近年流行のベイズ推定にも共通する（[1]第 12 章参照）。そのような場合の数値解法として MCMC（Markov Chain Monte Carlo）法が有力であろう。ここでは、回帰曲線との残差の分布が正規分布モデルに結び付けられる多項式近似モデルの例（図２で提起したその問題）を示す。

以下の式にしたがって生起される *n* 個の$(x_i,y_i)$標本値があったとする。

$$y_i = 2 + 4x_i - 1.8x_i^2 + 0.3x_i^3 - 0.015x_i^4 + \varepsilon_i \qquad \left(0 \le x \le 10;\ f_\varepsilon = N(0,1)\right)$$

式中の$\varepsilon_i$ は標本に含まれる誤差を与える量で、平均値 0、分散 1 の標本毎に独立な正規分布（$N(0,1)$）である。１節で示した Data とその母集団特性はこの式に基づいている。ここでは、*n*=11 と 100 で評価する。*n*=11 の例は図２で示しており、*n*=100 の場合を図６にプロットしている。

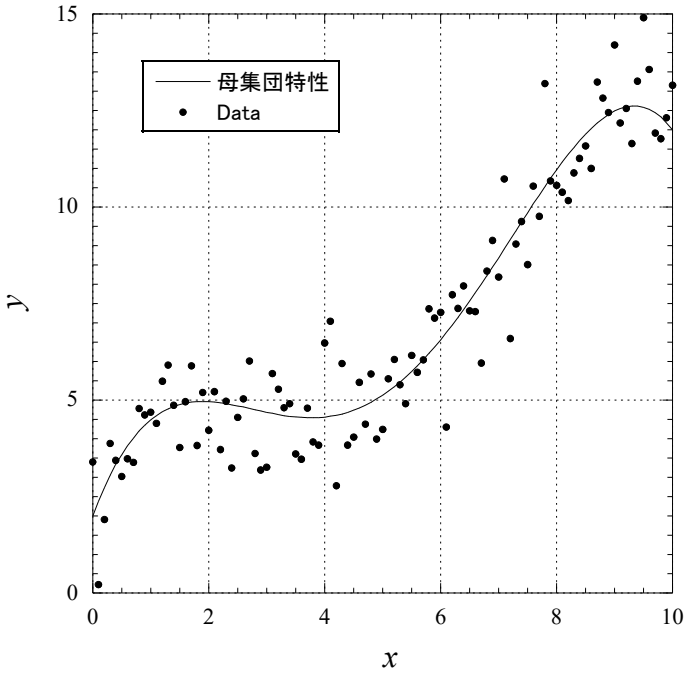



図６　AIC 評価用データ（*n*=100）（*n*=11 は図２）

このデータに対して以下の多項式モデルを評価する。

$$y = a_0 + a_1 x + a_2 x^2 + \cdots + a_K x^K + b \quad \left( f_b = N(0, \sigma^2) \right)$$

この場合のパラメータは $a_0$～$a_K$ と $\sigma^2$ の $K$+2 個である（注）。（注：$K$ 次多項式のモデルではモデルのパラメータ数は $a_0$ も入るので $K$+1。さらに、誤差の分散を含めて分布として考えれば、この分散が入り $K$+2 になる。分散を含めるかどうかは考え方の違いなので、考え方が統一されていればどちらでもよい。評価の際、すべてのモデルでの分散の扱いを共通にすれば、評価は大小関係の相対的なものなのでそれで問題はない）。基本的な考え方は、$a_0 + a_1 x + \cdots + a_K x^K$ を中心としてずれが正規分布するモデルのパラメータの最尤推定である。AIC によって、最適な次数（$K$ の値）を見つけ、良いモデルを選び出す問題である。

正規分布の場合、対数尤度関数は(23)式より

$$l(\boldsymbol{\theta}) = -\frac{n}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left( y_i - \sum_{k=0}^{K} a_i x_i^k \right)^2 \qquad \boldsymbol{\theta} = \left(a_0, a_1, a_2, \cdots, a_K, \sigma^2\right)$$

係数 $a_0$～$a_K$ の最尤推定値 $\hat{a}_0 \sim \hat{a}_K$ は尤度関数をそれぞれ $a_i$ で偏微分し 0 と置いて得られる $K$+1 個の連立方程式から求められる。

$$\frac{\partial}{\partial a_i}\sum_{i=1}^{n}\left( y_i - \sum_{k=0}^{K} a_i x_i^k \right)^2 = 0 \ \ (i = 1, 2, \cdots, n) \ \Rightarrow \ \hat{a}_0 \sim \hat{a}_K$$

これは最小二乗法であるので、最小二乗法の機能が具備されている市販ソフト（Excel, KaleidaGraph など）で容易に求めることができる。また、誤差分散の最尤推定値は、

$$\frac{\partial l\left(\boldsymbol{\theta}(\hat{a}_0, \cdots, \hat{a}_K, \sigma^2)\right)}{\partial \sigma^2} = 0 \ \Rightarrow \ \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left( y_i - \sum_{k=0}^{K} \hat{a}_i x_i^k \right)^2$$

となり、最大対数尤度は

$$l(\hat{\boldsymbol{\theta}}) = -\frac{n}{2}\log\left(2\pi\hat{\sigma}^2\right) - \frac{n}{2}$$

である。これより、AIC はパラメータ数が $K$+2 であるので次式となる。

$$AIC(K) = n\left(\log 2\pi + 1\right) + n\log\hat{\sigma}^2 + 2(K+2) \qquad (26)$$

表 1 は、$n$=11 と 100 の一例（図 2 と図 6）について、$K$=1～7 のモデルに対する平均二乗残差（分散の最尤推定値 $\hat{\sigma}^2$）、最大対数尤度（$l(\hat{\boldsymbol{\theta}})$）、AIC 値（$AIC$）および最小値との差分（$\Delta AIC$）を示している。同表より、$n$=11 では $K$=5 が、$n$=101 では $K$=4 が AIC 値の最小値になった。評価

を繰り返してみても、標本値の数 $n$ が大きくなると安定して $K$=4 を選択するが、標本数が少ない場合（$n$=11）では $K$=3, 4, 5 付近が選ばれる結果を得ている（表 1 では $n$=11 で $K$=5 が選ばれているが、この試行の偶然の結果である）。母集団特性の次数を $K$=4 で与えているので、AIC が妥当な次数のモデルを選んでくれたと言える。

正規分布に帰着できるモデルの評価においては、回帰曲線との残差の二乗平均値は次数 $K$ と共に小さくなるが、ある値を過ぎると変化ほとんどなくなる。この変化がない部分は最大対数尤度も変化が小さくなり、結果として、パラメータの増加に対するペナルティ分のみが AIC 値を押し上げる働きをする。このことから、二つのモデルが同程度の残差（＝標本値と回帰曲線との残差の二乗和）を与えるならば、パラメータの少ないほうが良いモデルであるとする原理が示されたのである。確率分布の評価であれば、３パラメータの一般化ガンマ分布を持ち出せば、尤度は高くなるであろうが、本当のよさ（＝選ぶべきモデル）は尤度だけではないですよと言うことである。

ここでは、AIC の適用例として、机上のモデルを示したが、筆者は、日本の強い雨の降り方に関する長期的な変化を調べるために種々のモデルを用い、どれを選ぶべきかの評価に AIC を活用している[5]。

表 1　$n$=11 と 101 の標本値に対するモデルの最大対数尤度と AIC 評価値

| | $n=11$ | | | | $n=101$ | | | |
|---|---|---|---|---|---|---|---|---|
| $K$ | $\hat{\sigma}^2$ | $l(\hat{\theta}^2)$ | $AIC$ | $\Delta AIC$ | $\hat{\sigma}^2$ | $l(\hat{\theta}^2)$ | $AIC$ | $\Delta AIC$ |
| 1 | 3.0594 | -21.759 | 49.517 | 17.951 | 2.4465 | -188.49 | 382.99 | 83.827 |
| 2 | 1.9648 | -19.323 | 46.646 | 15.080 | 1.4410 | -161.76 | 331.52 | 32.362 |
| 3 | 1.7104 | -18.560 | 47.120 | 15.554 | 1.4363 | -161.60 | 333.20 | 34.038 |
| 4 | 0.51739 | -11.984 | 35.968 | 4.4022 | 1.0053 | -143.58 | 299.16 | 0 |
| 5 | 0.28911 | -8.7831 | 31.566 | 0 | 1.0041 | -143.52 | 301.04 | 1.8795 |
| 6 | 0.28928 | -8.7863 | 33.573 | 2.0067 | 0.99152 | -142.88 | 301.77 | 2.6057 |
| 7 | 0.28748 | -8.7520 | 35.504 | 3.9380 | 0.99111 | -142.86 | 303.72 | 4.5634 |

### 3.5 まとめ

本章では、AIC が生まれた背景、基本的な考え方、そして具体的な評価方法について述べた。AIC の基礎理論は数学的に高度であるものの、実際に使う立場から見ると、非常に簡潔で理にかなった評価法であることが理解できたと思う。一般的な分布に対する最尤推定には数値計算が必要となるが、正規分布を前提としたモデル化であれば解析的に解ける場合も多く、AIC は実務上きわめて有用な手法である。

情報量基準の本質である平均対数尤度は、母集団の未知の分布を含むため直接求めることはできない。しかし、標本データとモデルから得られる最大対数尤度に一定の補正を加えることで、平均対数尤度の推定量として利用できる。その補正量がモデルのパラメータ数になるという、理論的に非常に美しい結果が得られている点こそが、AIC の核心である。 この結果から、同程度

の尤度を与えるモデル同士であれば、パラメータ数の少ないモデルを選ぶべきであるという原理が導かれ、これはまさにオッカムの剃刀の精神に一致する。

以下に、AIC に関する補足的な事項をまとめる。 **1）モデリングの目的は「良いモデル」を得ることであり、真のモデルを求めることではない。**現実の真のモデルは極めて複雑であり、主要なパラメータが適切に選ばれていれば十分である。 **2）AIC は真の次数とモデルの次数が一致することを要求しない。**データ数が少ない場合には、予測の安定性の観点から低次モデルが適切となる場合がある。 **3）AIC は相対評価の指標であり、絶対的な近似度（KL 情報量）を直接評価するものではない。**しかし、モデル最適化の目的関数として利用することで、適応的なモデル構築にも応用できる。 **4）AIC を源流として、情報量基準は TIC、GIC などへと発展してきた。**それでも、仮定するモデルが現象の主要な特性を含む場合には、AIC の補正項（パラメータ数）は十分に妥当であり、現在でも AIC は高い有用性を保ち続けている。

**参考文献**

[1] 唐沢好男, 電波伝搬モデルの基礎：確率分布と統計的推定, ネット公開私製本, 2025.
TR-YK-078_Probability_Distributions.pdf
[2] 唐沢好男, *無線通信物理想技術へのアプローチ*, コロナ社, 2021.
[3] 小西貞則, 北川源四郎, *情報量基準*, 朝倉書店, 2004,
[4] 赤池弘次他, *赤池情報量基準 AIC：モデリング・予測・知識発見*, 共立出版, 2007.
[5] Y. Karasawa, “Long-term variations of extreme rainfall in Japan for predicting the future trend of rain attenuation in radio communication systems,” Climate, vol. 2005, 13, 145, pp. 1-19, 2025.
https://www.mdpi.com/2225-1154/13/7/145

読者の皆さんへ
本レポートはネット公開私製本「電波伝搬モデルの基礎：確率分布と統計的推定」の補足資料４になっています。