「計画数理II」講義ノート

吉野　隆

平成17年4月8日

このテキストは製作途中であり不備な部分が多くあります.

はじめに

このメモは東洋大学環境建設学科の３年生を対象とした授業「計画数理II」の講義ノートです．授業は，確率と統計の基礎から話を始めて，AIC(Akaike's Information Criterion) の使い方までを解説しています．教科書には坂本・石黒・北川の「情報量統計学」[1] を用いています．学生時代の僕は，この教科書を読み通すのにとても苦労しました．そのときに疑問に思ったことなどを活かして，授業では，なるべく初心者に解るように話をしています．話の展開はどうしても「情報量統計学」と同じになってしまいがちですが，難解な部分の解説を加えて，多少易しく話しているつもりです．話の厳密性が十分でないところもありますが，それは教科書を参照していただければと思います．

AIC との出会いがいつだったのかは忘れてしまいました．最小自乗法がパラメータ数を多くすればいくらでも残差平方和を小さくすることができることが不満だった自分にとって，はじめて AIC を知ったときの感動は忘れることができません．是非，たくさんの人に AIC の良さを知って頂きたいと思っています．

1　統計の基礎

1.1　１変数の場合

この副節では，主に，ある同一の事象を観測した結果として得られた数値の組 $(\ x_1,\ x_2,\ x_3, \cdots, x_n\ )$ の性質をいくつかの数値を用いて表現することについて考えていく．これらの集まりは何種類の数値を用いて表すべきだろうか？当然のことながら，データの個数分の数値を使えば，この数字の集まりは完全に表現する

ことができる．しかし，それでは性質を抽出したことにはならない．得られた結果の性質を表す量として，以下のような量が挙げられるだろう．

1.1.1 データの代表値

はじめに，得られた結果を全体の情報をひとつの数値として表す場合，どのようなものが最適であるかを考えよう．良く用いられるのは，以下の３つの量である．

1. 最頻値 (モード)：最大度数を与えるクラスの代表値.
2. 中央値 (メディアン)：データを大きさの順に並べ変えた時の中央にある値.
3. 平均値：データの総和をデータ数で割った値.

平均値 ($m(x)$) は

$$m(x) = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

で計算される．

おそらく平均値が最もポピュラーな代表値の決定方法だろう．平均値を代表値と仮定することの妥当性については，後で議論することになる．上記の３種類の値は，統計的に「タチが良い」データであれば同じような値をとる．逆に,「タチが悪い」データであると異なった値をとりどれを代表値とすれば良いかは一概には言えなくなる.「タチが悪い」データの例としては，ピークがふたつあるような分布を持つデータセットが挙げられる．

1.1.2 データのばらつき具合を与える量

得られたデータをふたつの数値で表現することを考えよう．ひとつめの数値としては，前述の「代表的な値」が選ばれるだろう．ふたつめの数値としては，どのような値が考えられるだろうか？

ふたつめの数値としては「ばらつきの程度を示す量」を考えよう．観測されるデータにはバラツキが伴われるのが普通なので，平均値が同じであっても，その平均値が意味している値の信頼度は異なることが考えられるからだ．このような量には，以下のようなものがある．

1. 範囲 (レンジ)：(データの最大値) − (データの最小値)
2. (標本) 平均偏差：平均値からのずれ (偏差) の絶対値の平均

$$d = \frac{1}{n}\sum_{i=1}^{n} |x_i - m(x)| \tag{2}$$

3. (標本) 分散：偏差の 2 乗の平均

$$\sigma^2(x) = \frac{1}{n}\sum_{i=1}^{n}\{x_i - m(x)\}^2 = m(x^2) - m(x)^2 \tag{3}$$

4. (標本) 標準偏差：分散の平方根

$$\sigma(x) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\{x_i - m(x)\}^2} \tag{4}$$

5. 変動係数：標準偏差を平均で割ったもの

$$v = \frac{\sigma(x)}{m(x)} \tag{5}$$

範囲は結果の幅を表現しているので，値が小さいほどデータのばらつきも小さくなると考えられる．しかし，偶然に，平均よりもかなり大きな値や平均よりもかなり小さな値がでたときに，その結果をひきずってしまう．

平均偏差 (または標本平均偏差) は平均値からのずれの大きさの平均なので，(範囲のようにふたつの特殊なデータに注目するのではなく) どのデータも等しい重みをおいてずれの程度を見ているという点でばらつきの程度を範囲よりはきちんと評価していると言える．問題は絶対値を用いていることにある．絶対値は計算結果の正負によって異なる取り扱いをしなければならないために，解析的に操作しにくい．

分散 (または標本分散) は平均偏差が持つ絶対値の取り扱いの繁雑さを避けるために平均からのずれを自乗した量を用いている．これによって場合分け計算を取り扱う繁雑さは無くなったもののバラツキの程度の示す量として，元データの単位とは異なる量を用いることになる (たとえば，元データが cm であった場合には，分散の値は cm^2 となる)．

標準偏差 (または標本標準偏差) は分散の平方根をとることで，ばらつきの指標を元データと同じ単位にしている．分散と同じくらい頻繁に使われる量ではあるが，平方根を用いるので取り扱いが多少繁雑になる．

変動係数は標準偏差を平均値で割ることで，ばらつきの度合を無次元で表現している．

1.2 ２変数の場合

次に，同時に計測できる二つの量 (x, y) を観測した $(x_1, y_1), (x_2, y_2), \cdots, (x_n)$ というデータの組を考える．このときに問題になるのは，二つの量がどのような関係にあるかである．

1.2.1 回帰直線

得られたデータの組を $y = a + bx$ に当てはめることを考える．最小自乗法によると，このとき，係数 a と b は，

$$a = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x)^2} \tag{6}$$

$$b = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2} \tag{7}$$

である．

最小自乗法は実測値と推定値との間のずれを残差平方和（least mean square）であるとし，この「ずれ」が最も少ないパラメータの値を最適値とする手法である．上記の実測値に対して，推定される曲線を $y = f(x; a, b, c, \dots)$ （$a, b, c, \dots$ はパラメータ）とおくと，残差平方和は，

$$S = \sum_{i=1}^{n} \{y_i - f(x_i; a, b, c, \dots)\}^2 \tag{8}$$

となり，最適なパラメータを得るための方程式は，

$$\begin{aligned} \frac{\partial S}{\partial a} &= 0, \\ \frac{\partial S}{\partial b} &= 0, \\ \frac{\partial S}{\partial c} &= 0, \\ &\vdots \end{aligned} \tag{9}$$

となる．なぜ残差平方和を最小にするパラメータが良いパラメータなのかは考察してみる価値があるが，これは後の「対数尤度」までの課題としておこうと思う．

前述の直線に近似する場合の式は

$$\begin{aligned} \frac{\partial}{\partial a}\left(\sum_{i=1}^{n}\{y_i - (a + bx_i)\}\right) &= 0, \\ \frac{\partial}{\partial b}\left(\sum_{i=1}^{n}\{y_i - (a + bx_i)\}\right) &= 0, \end{aligned} \tag{10}$$

を a と b について解くことで得られる．一度自分で導出することをお勧めする．

1.2.2 相関係数

簡単に言えば n 次元空間にある二つの単位ベクトルの内積を求める作業である.

$$r = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{x_i - m(x)}{\sigma(x)}\right]\left[\frac{y_i - m(y)}{\sigma(y)}\right] \tag{11}$$

$$= \frac{m(xy) - m(x)m(y)}{\sqrt{\sigma(x)^2\sigma(y)^2}} \tag{12}$$

第 1 週の課題

課題 1-1

次のデータについて，平均，分散，標準偏差，そして変動係数を求めよ (単位にも気をつけること).

290 kg, 296 kg, 445 kg, 497 kg, 487 kg, 414 kg, 313 kg, 334 kg, 510 kg, 395 kg, 363 kg, 522 kg.

課題 1-2

次のデータについて，回帰直線と相関係数を求めよ.

(8, 9), (8, 5), (6, 9), (6, 7), (6, 5), (4, 7), (4, 4), (3, 7), (3, 4), (2, 3)

2 確率と確率変数

2.1 事象と確率

2.1.1 標本空間

ある偶然を伴う実験の結果が，$\omega_1, \omega_2, \omega_3, \cdots, \omega_s$ のいずれかに属するとき，これらの結果すべての集合を**標本空間**と呼ぶ (以下では，これを Ω と表す).

標本空間を規定することは，以下のことを行うことを意味している.

1. 対象の限定
2. 結果の範囲の規定
3. 結果の記号化

2.1.2 事象

標本空間の部分集合を**事象**と呼ぶ (以下では，これを E や F で表す)．事象は

$$E = \{\omega | \omega \text{に関する条件}\} \tag{13}$$

という形で規定される．

2.1.3 事象の種類

1. 全事象 (U)：すべての結果の集合
2. 空事象 (ϕ)：どの結果も含まない集合
3. 余事象 ($\bar{E}$)：E に属さない事象の集合
4. 根本 (根元) 事象：ただひとつの結果からなる集合
5. 和事象 ($E \cup F$) ：E と F の少なくとも一方が属している事象
6. 積事象 ($E \cap F$) ：E と F の両方に属している事象
7. 排反事象：一方が起これば他方は決して起こらないという事象の関係．例えば， E と F が排反ならば $E \cap F = \phi$ である．

2.1.4 確率

事象 E に含まれる根元事象の数を $n(E)$ とするとき，事象 E の確率 $P(E)$ は

$$P(E) = \frac{n(E)}{n(\Omega)} \tag{14}$$

である．

2.2 確率の公理

2.2.1 確率の公理

確率論は，コルモゴロフが導入した「確率論の公理 (証明無しで成立を仮定した事項．今風の言葉で言えば「お約束」)」から始めるのが普通である[1]．コルモゴロフが導入した確率論の公理は以下の 3 点である．

[1]正しくはコルモゴロフの公理の前に「標本空間」と「σ 集合族についての定義があるが，今回の講義では応用を前提としているので「σ 集合族」の説明については省略している．しかし実際のところ， σ 集合族の話がないと，この公理のありがたさが解らないかもしれない．

1. 任意の事象 E について

$$0 \leq P(E) \leq 1 \tag{15}$$

2.

$$P(\Omega) = 1 \tag{16}$$

3. E_1, E_2, E_3, $\cdots$ が可算無限個の排反な事象ならば

$$P(E_1 \cup E_2 \cup \cdots) = P(E_1) + P(E_2) + \cdots \tag{17}$$

2.2.2 余事象の法則

$$P(\bar{E}) = 1 - P(E) \tag{18}$$

証明
$E \cup \bar{E} = \Omega$ かつ $E \cap \bar{E} = \phi$ より

$$P(E) + P(\bar{E}) = P(E \cup \bar{E}) = P(\Omega) = 1 \tag{19}$$

よって,

$$P(\bar{E}) = 1 - P(E) \tag{20}$$

2.3 条件つき確率と独立性

2.3.1 条件つき確率

ふたつの事象 E と F があり， $P(F) > 0$ であるとき

$$P(E|F) = \frac{P(E \cap F)}{P(F)} \tag{21}$$

を「事象 F を条件とする事象 E の条件つき確率」という.

2.3.2 独立

$P(E|F) = P(E)$ であるとき，事象 E と F は独立であるという.

2.4 ベイズの定理 (反転公式)

互いに排反で

$$\sum_{i=1}^{k} P(E_i) = 1 \tag{22}$$

となるような k 個の事象 E_i について，$P(F) > 0$ ならば

$$P(E_i|F) = \frac{P(E_i)P(F|E_i)}{\sum_{i=1}^{k} P(E_i)P(F|E_i)} \tag{23}$$

である．これを**ベイズの定理**という．

証明

定義より $P(E_i|F) = P(E_i \cap F)/P(F)$ である．また，$P(F|E_i) = P(F \cap E_i)/P(E_i) = P(E_i \cap F)/P(E_i)$ より，$P(E_i|F) = P(E_i \cap F)/P(F)$ の和をとって，$P(F) = P(E_1)P(F|E_1) + P(E_2)P(F|E_2) + \cdots P(E_k)P(F|E_k)$ である．よって，式 (23) が成立する．

第 2 週の課題

課題 2

3 人の死刑囚 A，B，C がいる．明日，3 人の囚人のうち 2 人が処刑されることがわかった．しかし，囚人達には誰が処刑されるかは知らされていない．囚人 A は看取に「2 人が処刑されるのだから B と C の少なくともひとりは処刑される．処刑される者のひとりの名前を教えてほしい」と要求したところ，看取は「B は処刑される」と言った．このとき，A が処刑される確率，および C が処刑される確率をベイズの定理を用いて求めよ．

2.5 確率変数と分布関数

標本空間 Ω 上で定義された実数値関数 $X(\omega)$ あって，任意の実数 x に対して $X(\omega)$ が x 以下となるような根源事象の集合 E によって決定される関数 $F(E)$ を一般に確率変数と呼んでいる．確率 ($P(E)$) は確率変数のひとつである．

確率変数が**離散型確率変数**であるとは，確率変数のとり得る値が有限個または可算無限個あるときをいう．このとき $X(\omega)$ が $X_i(i = 1, 2, \dots)$ という値をとる確率を

$$P_{x_i} = P(X = x_i) \tag{24}$$

と定義しておく．確率変数が**連続型確率変数**であるとは，確率変数のとり得る値が非可算無限個あるときをいう．連続型確率変数は，確率変数は必ずひとつの値を持つにもかかわらず，その値をとる確率が離散型確率変数のようには定義できないことに注意しなければならない．詳細は積分論にゆずって，ここでは立ち入らないことにする．そのかわり，以下に示すように分布関数を定義してその分布関数をもとにして連続型確率変数の表現方法について考えることにする．

分布関数とは，

$$F(x) = P(\omega|X(\omega) \leq x) \tag{25}$$

によって定義される関数である．これは事象 ω の実現値が x よりも低くなる確率を表している．離散型確率変数の分布関数は

$$F(x) = \sum_{x_i \leq x} p_{x_i} \tag{26}$$

と表され，連続型確率変数の場合には

$$F(x) = \int_{-\infty}^{x} f(\xi) d\xi \tag{27}$$

と表される．ここで f(x) は X の確率密度関数と呼ばれ，

$$f(x) \geq 0 \tag{28}$$

かつ

$$\int_{-\infty}^{\infty} f(x) dx = 1 \tag{29}$$

である．この密度関数を用いることによって，連続型確率変数は表現される．

離散型確率変数 X の期待値とは

$$\mu = E[X] = \sum_{t=1}^{\infty} x_i p_i \tag{30}$$

である．連続型確率変数 X の期待値は

$$\mu = E[X] = \int_{-\infty}^{\infty} x f(x) dx \tag{31}$$

である．

確率変数の関数 $g(x)$ の期待値も定義することができ，それぞれの確率変数について

$$E[g(X)] = \sum_{t=1}^{\infty} g(x_i) p_i \tag{32}$$

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) f(x) dx \tag{33}$$

である．とくに X の期待値 (平均とも言う) からのずれの平方の期待値は X の分散と呼ばれ，

$$\sigma^2 = E[(X-\mu)^2] = \sum_{t=1}^{\infty} g(x_i) p_i \tag{34}$$

$$\sigma^2 = E[(X-\mu)^2] = \int_{-\infty}^{\infty} g(x) f(x) dx \tag{35}$$

分散の平方根を標準偏差という．分散については

$$\sigma^2 = E[X^2] - \mu^2 \tag{36}$$

が成立する．

ここで，以前に説明した統計用語としての平均や分散と確率用語としての平均や分散の相違点について考えてみてほしい．統計が何をしようとしているのかが解るし，確率論から統計をサポートするためには何を示せば良いのかを考えるきっかけになる．

2.6 確率分布

任意の条件に対して確率を計算するために必要な式は何だろうか？離散型確率変数の場合には，個々の確率を数値で表すことができるのですべての独立事象の確率が与えられれば計算が可能である．また連続型確率変数の場合には，求める確率は確率密度関数を積分することによって与えられるために，確率密度関数の形が与えられれば計算が可能である．このため，離散型確率変数については全ての独立事象の確率を表す式を確率分布と呼び，連続型確率変数については確率密度関数を示す式を確率分布と呼ぶ．以下では，さまざまな確率分布について考えて行く．

2.6.1 ２項分布

２項分布とは確率 p の試行が n 回繰り返されたときに，着目した事象が k 回起こる確率を示す分布であり，

$$b(k|p) = {}_nC_k p^k (1-p)^{n-k}, \qquad k = 1, 2, \ldots, n \tag{37}$$

と表される．平均と分散はそれぞれ，

$$\mu = np \tag{38}$$

$$\sigma^2 = np(1-p) \tag{39}$$

である．平均については

$$\sum_{k=0}^{n} {}_nC_k p^k (1-p)^{n-k} = 1 \tag{40}$$

の両辺を p で微分して，両辺を $p(1-p)$ 倍することによって得られる式を変形すればよい．分散は平均の式の両辺を p で微分して変形を行うことで得られる．

2.6.2 ポワソン分布

ポワソン分布は

$$p(k|\lambda) = \frac{\lambda^k}{k!}e^{-\lambda} \tag{41}$$

という確率分布を持つ．これは二項分布に対して，期待値（$np = \lambda$）を一定のままで，$p \to 0$ かつ $n \to \infty$ という極限操作を行った結果として得られる．

2.6.3 多項分布

n 回の独立試行の結果が k_1, k_2, $\cdots$, k_n 回の E_1, E_2, $\cdots$ E_c という事象に分けられるとき，その確率分布を**多項分布**という．多項分布は

$$m(k_1, k_2, \cdots, k_c|p_1, p_2, \cdots, p_n) = \frac{n!}{k_1!k_2!\cdots k_c!}p_1^{k_c}p_2^{k_c}\cdots p_c^{k_c} \tag{42}$$

で与えられる．

2.6.4 一様分布

定義された範囲内にあるすべての値が同じ確率で生じる分布を**一様分布**という．その確率密度関数は，

2.6.5 正規分布

おそらく確率論のの中でも最も重要な分布だろう．次の 3 つの条件を満たす確率密度関数によって表される分布を**正規分布**という．

1. 真の値をとる確率は他の確率よりも大きい．
2. 実現回数を非常に多くしたときの期待値は真の値となる．
3. 真の値からずれが大きい値ほど実現しにくい．

正規分布の形

平均 (真の値) μ，分散 σ^2 の正規分布 $N(\mu, \sigma^2)$ の確率密度関数 $P(x)$ は，

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \tag{43}$$

である．特に $N(0, 1)$ を標準正規分布という．

正規分布の性質

積分が 1 になることは，$t = (x - \mu)/\sigma$ と変数変換をして極座標変換を行えば良い．平均値が μ になることは，上の性質から μ で一回微分して得られる．分散が

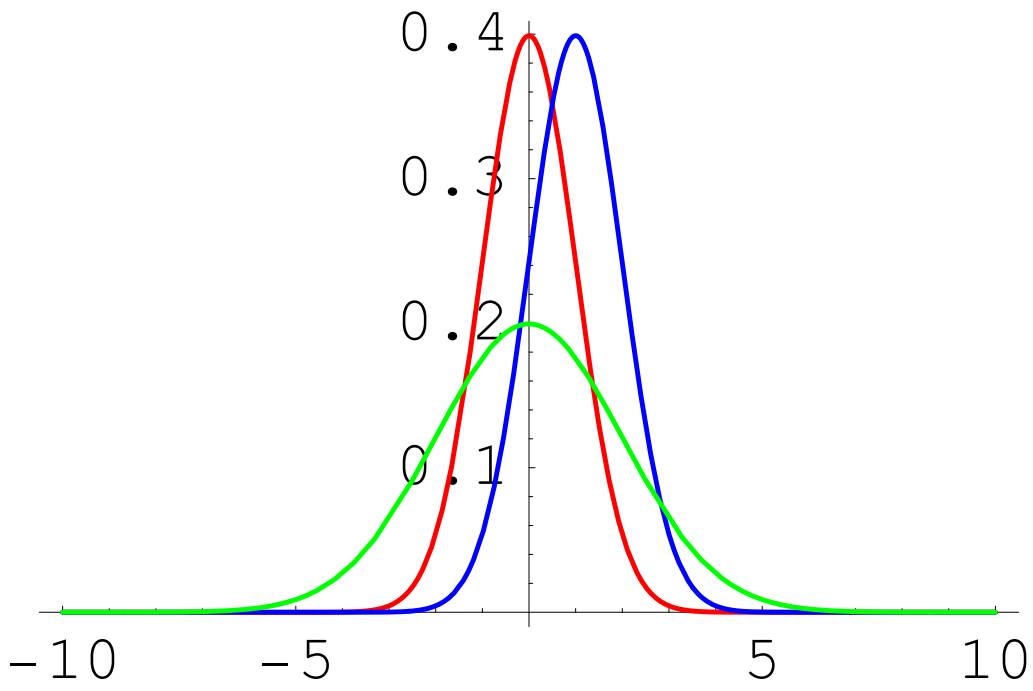


図 1: 正規分布の確率密度関数の例

σ^2 になることも，更に μ で微分してみることで得られる．変数変換 $t = (x-\mu)/\sigma$ によって，任意の正規分布が標準正規分布に焼き直すことができるために，この変換は重要である．例えば，

$$\int_a^b f(x|\mu,\sigma^2) = \int_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2} dt = \int_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}} f(t|0,1)dt \tag{44}$$

とできる．従って，$N(0,1)$ についてのデータがあれば，すべての積分が可能であることがわかる．実際に $N(0,1)$ の場合については数表や計算アルゴリズムが存在している．良く用いられる数値として，

$$\int_{-2.58}^{2.58} \sigma f(t|0,1)dt = 0.99 \tag{45}$$

$$\int_{-1.96}^{1.96} \sigma f(t|0,1)dt = 0.95 \tag{46}$$

がある．

2.6.6 カイ２乗分布

$X_1, X_2, \cdots, X_k$ が互いに独立に分布 N(0,1) に従うと仮定する．このとき，新しい確率変数 χ_k^2 を

$$\chi_k^2 = \sum_{i=1}^{k} X_i^2 \tag{47}$$

で定義する．このとき χ_k^2 の従う分布を**自由度 k のカイ 2 乗分布**と呼び，その密度関数は

$$f_k(\chi_k^2) = \frac{1}{2^{k/2}\Gamma(k/2)}(\chi_k^2)^{k/2-1}e^{-\chi_k^2/2} \tag{48}$$

である．ここで

$$\Gamma(k) = \int_0^\infty e^{-x}x^{k-1}dx \tag{49}$$

であり，**ガンマ関数**と呼ばれる．

χ_k^2 の期待値と $(\chi_k^2)^2$ の期待値は，

$$\int_0^\infty \chi_k^2 f_n(\chi_k^2)d\chi_k^2 = k \tag{50}$$

$$\int_0^\infty (\chi_k^2)^2 f_n(\chi_k^2)d\chi_k^2 = k(k+2) \tag{51}$$

から，平均が自由度に等しいこと，分散が自由度の 2 倍となることがわかる．

3 尤度と AIC

3.1 KL 情報量

データの当てはまりのよさを数値で表すことを考える．これを実現するために Kullbuck と Leibler が提案した量を Kullbuck-Leibler 情報量または KL 情報量と呼ぶ．離散型確率変数の場合，$\boldsymbol{p} = \{p_1, p_2, \cdots, p_n\}$ を真の確率分布，$\boldsymbol{q} = \{q_1, q_2, \cdots, q_n\}$ をモデルの確率分布とおくと，

$$I(\boldsymbol{p}, \boldsymbol{q}) \equiv \sum_{i=1}^{n} p_i \log p_i/q_i \tag{52}$$

ここで，n は確率変数の数であり，log は自然対数を表す．連続型確率変数の場合は，真の分布を $p(x)$ モデルの分布を $q(x)$ とすると，

$$I(p, q) \equiv \int_{-\infty}^{\infty} p(x) \log p(x)/q(x)dx \tag{53}$$

である．

KL 情報量は以下のような性質を持っている．

1. KL 情報量は常に 0 以上である．

2. モデルの分布が真の分布と一致するときに KL 情報量はゼロとなる．

この性質は KL 情報量をモデルの当てはまりのよさを与えるひとつのものさしとすることを支持している．また，KL 情報量が（負の）エントロピーに符号を逆にしたものであることも，ものさしとすることを支持するものである．上の性質から明らかなように，このものさしはあてはまりが良いほど小さい値を持つ．以下では，このものさしを用いてモデルの当てはまりのよさを比較する方法について考えてゆく．

3.2 対数尤度

KL 情報量は，離散的な場合，

$$\begin{aligned} I(\boldsymbol{p},\boldsymbol{q}) &\equiv \sum_{i=1}^{n} p_i \log p_i/q_i \\ &= \sum_{i=1}^{n} p_i \log p_i - \sum_{i=1}^{n} q_i \log q_i \qquad (54) \\ &= E[\log p] - E[\log q] \qquad (55) \end{aligned}$$

連続的な場合，

$$\begin{aligned} I(p,q) &\equiv \int_0^{\infty} p(x) \log p(x)/q(x) dx \\ &= \int_{-\infty}^{\infty} p(x) \log p(x) - \int_{-\infty}^{\infty} p(x) \log q(x) dx \\ &= E[\log p] - E[\log q] \qquad (56) \end{aligned}$$

と変形できる．どちらの式も，第一項は真の分布から得られる定数であり，モデルの当てはまりのよさを決めるのは第二項であることがわかる．この第二項の符号を逆転した式 $E[\log q]$ の近似値をデータから得ることを考える．

観測データを $\boldsymbol{x} = \{x_1, x_2, \cdots, x_m\}$ とする．離散型確率変数の場合で x_i が出現する確率を $q(x_i)$ ，連続型確率変数の場合で x_i についての確率密度関数をやはり $q(x_i)$ と書くと，上記の $E[\log q]$ の近似値は，

$$\frac{1}{m} \sum_{i=1}^{n} \log q(x_i) \qquad (57)$$

と表せそうである．これを**平均対数尤度**と呼ぶ．これをデータ数 m 倍した

$$\ell \equiv \sum_{i=1}^{n} \log q(x_i) \qquad (58)$$

を対数尤度（log likelihood）と呼ぶ．また，対数をとる前の

$$\prod_{i=1}^{n} q(x_i) \tag{59}$$

を尤度（likelihood）と呼ぶ．Likelihood（もっともらしさの度合い）というのは，この式が，考えている確率モデルにおけるそれぞれの観測データの積事象が観測される確率になっていることからも理解できる．

対数尤度や尤度は確率モデルのよさを測るものさしとなる．対数尤度と KL 情報量を比較してみると符号が逆転しているので，KL 情報量とは逆に，モデルの対数尤度が大きいほど，当てはまりがよくなることに注意する．

3.3 最尤法

3.3.1 定義

考えているモデルの尤度が最大になるようにパラメータの値を決めることによって，モデルが真の分布に最も近くなるように調整することを最尤法という．また，最尤法によって得られたパラメータの値を最尤推定量と呼び，変数名の上にハット（ $\hat{}$ ）をつけて表す．

3.3.2 正規分布

n 個のデータ $x_1, x_2, \ldots, x_n$ が正規分布する，すなわち x の確率密度関数が，

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2 \right\} \tag{60}$$

で表される仮定できるとする．このときにデータから得られる最尤推定量 $\hat{\mu}$ および $\hat{\sigma}^2$ についての式を求めることを考える．

このとき，対数尤度は，

$$\begin{aligned} \ell(\mu, \sigma^2) &= \sum_{i=1}^{n} \log f(x_i|\mu, \sigma^2) \\ &= \sum_{i=1}^{n} \left\{ \log \sqrt{\frac{1}{2\pi\sigma^2}} + \frac{1}{2} \left(\frac{x_i - \mu}{\sigma} \right)^2 \right\} \\ &= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \end{aligned} \tag{61}$$

となる．この量が最大となるのは，

$$\frac{\partial \ell}{\partial \mu} = 0, \qquad \frac{\partial \ell}{\partial \sigma^2} = 0 \tag{62}$$

のとき．この方程式の解 $\hat{\mu}$ と $\hat{\sigma}^2$ は，

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{63}$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2 \tag{64}$$

となる．すなわち，あるデータセットについて，それが正規分布すると仮定して最尤法によって推定されたパラメータの値は，平均の推定値が平均値，分散の推定値が標本分散になることを示している．

後のために必要になるので，このときの対数尤度 $\ell(\hat{\mu}, \hat{\sigma}^2)$ を求めておくと，

$$\ell(\hat{\mu}, \hat{\sigma}^2) = -\frac{n}{2}\log(2\pi\hat{\sigma}^2) - \frac{n}{2} \tag{65}$$

となることがわかる．

3.3.3 多項式回帰モデル

n データのペア $(x_1, y_1), (x_2, y_2), \cdots (x_n, y_n)$ を考える．これらのデータが，

$$y = a_0 + a_1 x + a_2 x^2 + \cdots + a_{m-1}x^{m-1} + a_m x^m \tag{66}$$

に従うと仮定し，このときの AIC を求めることを考える．そのために，実際に得られたデータは平均ゼロ・分散 σ^2 の誤差を伴うとしたした確率モデルを考える．このようなモデルを m 次多項式回帰モデルという．

このモデルの対数尤度は，誤差が正規分布に従うという確率モデルであることから，

$$\begin{aligned}\log q_i &= \log\left[\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{\frac{(y_i - a_0 - a_1 x_i - \cdots - a_m x_i^m)^2}{2\sigma^2}\right\}\right] \\ &= -\frac{1}{2}\log 2\pi - \frac{1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(y_i - a_0 - a_1 x_i - \cdots - a_m x_i^m)^2 \end{aligned} \tag{67}$$

と表されることにより，

$$\ell(a_0, a_1, \cdots, a_m, \sigma^2) = \sum_{i=1}^{n}\log q_i \tag{68}$$

$$= -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\{y_i - (a_0 + a_1 x_i + \cdots + a_m x_i^m)\}^2 \tag{69}$$

このとき，最尤推定量（$\hat{a}_0, \hat{a}_1, \cdots, \hat{a}_m, \hat{\sigma}^2$ ）は

$$\frac{\partial l}{\partial a_0} = \frac{\partial l}{\partial a_1} = \cdots = \frac{\partial l}{\partial a_m} = \frac{\partial l}{\partial \sigma^2} = 0 \tag{70}$$

の解である．この方程式は最初の $(m+1)$ 個について以下のように書き換えられる．

$$\begin{bmatrix} n & \sum x_i & \sum x_i^2 & \cdots & \sum x_i^m \\ \sum x_i & \sum x_i^2 & \sum x_i^3 & \cdots & \sum x_i^{m+1} \\ \vdots & \vdots & \ddots & & \vdots \\ \sum x_i^m & \sum x_i^{m+1} & \sum x_i^{m+2} & \cdots & \sum x_i^{2m} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_m \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \\ \vdots \\ \sum x_i^m y_i \end{bmatrix} \tag{71}$$

であり，この解は最小自乗法によって得られるものと同じである．また，最後の方程式は，

$$-\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{i=1}^{n} (y_i - \hat{a}_0 - \hat{a}_1 x_i - \cdots - \hat{a}_m x_i^m)^2 = 0 \tag{72}$$

の解であって，

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{a}_0 - \hat{a}_1 x_i - \cdots - \hat{a}_m x_i^m)^2 \tag{73}$$

である．

後のために必要になるので，このときの対数尤度を求めておくと，

$$\ell(\hat{a}_0, \hat{a}_1, \cdots, \hat{a}_m, \hat{\sigma}^2) = -\frac{n}{2} \log 2\pi - \frac{n}{2} \log \hat{\sigma}^2 - \frac{n}{2} = -\frac{n}{2} \left(\log 2\pi + 1 + \log \hat{\sigma}^2 \right) \tag{74}$$

3.4 AIC

3.4.1 AIC とは

n 次回帰モデルの最尤法の結果について改めて考えてみると，ある問題点に気が付く．それは次数が大きいほど残差平方和は小さくなることである．つまり，残差平方和が小さいことが正義だとすると，次数を大きくとることが正しいモデル選択の方法になってしまう．極論を言うと n 組のデータを用いて n 次回帰モデルを作ることが一番よいモデルであるということになる．しかし，この考え方はどうみてもおかしい．

赤池の情報量規準 (Akaike's Information Criterion: AIC) は，このような問題点を回避するためのものさしとなるものである．

3.4.2 AIC の導出

AIC の導出については参考図書を見てほしい．ここでは，何をもとにしてどんな仮定を行うことで導出されるのかを見るに留めておく．基本となるのは KL 情報量である．これを以下の条件（仮定）のもとで評価してみる．

- 最尤モデルの平均対数尤度を真のパラメータで Taylor 展開したときに 2 次の項まで考慮すれば十分である．
- 最尤推定量はデータの個数が無限大になるときに，真のパラメータとなる．

これらを仮定すると，パラメータ数 K の期待対数尤度（平均対数尤度のデータ $\boldsymbol{x}$ に関する期待値）$l_n^*(K) =$ つまり，K L情報量

$$I(p,q) = E[\log(p)] - E[\log(q)] \tag{75}$$

の第 2 項が,

$$\frac{(\text{最大対数尤度}) - (\text{パラメータ数})}{n} \tag{76}$$

で近似されることがわかる（n はデータ数)．そこで,（歴史的な経緯から）分子に (-2) を乗じたものを AIC(K) と呼び，この数値によってモデルのあてはまりの良さを評価する．すなわち，モデルのパラメータ数を K としたとき,

$$\mathrm{AIC(K)} = -2\ell(\cdots) + 2K \tag{77}$$

が小さいモデルほど良いモデルであるとして評価を行う．

3.5 AIC の使用例

3.5.1 硬貨の真偽

n 回の試行で r 回が表であった．このとき，この硬貨の真偽を判断してみる．ふたつのモデルが考えられる．ひとつは硬貨が本物であると考えるモデルである．これをモデルAと呼ぶことにする．モデルAでは表がでる確率は 0.5 とみなされる．もうひとつは硬貨が偽物であると考えるモデルである．これをモデルBと呼ぶことにする．モデルBでは表がでる確率 p が不明なので，最尤法を用いて p の値を求める．

表がでる確率が p であるとき，その対数尤度 $l(p)$ は,

$$l(p) = \sum_{i=1}^{n} \log p_i = \sum_{i=1}^{r} \log p + \sum_{i=1}^{n-r} \log(1-p) = r\log p + (n-r)\log(1-p) \tag{78}$$

$l(p)$ が最大になるように最尤推定量 $\hat{p}$ を設定するには,

$$\frac{\partial l}{\partial p} = \frac{r}{p} - \frac{n-r}{1-p} = 0 \tag{79}$$

を解いて,

$$\hat{p} = \frac{r}{n} \tag{80}$$

このとき，最大対数尤度は，

$$l(\hat{p}) = r \log \frac{r}{n} + (n-r) \log \frac{n-r}{n} \tag{81}$$

である．

この結果，モデルAの AIC は，

$$\mathrm{AIC}(0) = -2\{r \log 0.5 + (n-r) \log 0.5\} + 2 \times 0 = -2n \log 0.5 \tag{82}$$

モデルBの AIC は，

$$\begin{aligned} \mathrm{AIC}(1) &= -2 \left\{ r \log \frac{r}{n} + (n-r) \log \frac{n-r}{n} \right\} + 2 \times 1 \\ &= 2 \left\{ 1 - r \log \frac{r}{n} - (n-r) \log \frac{n-r}{n} \right\} \end{aligned} \tag{83}$$

考え方は逆になってしまうが，$\hat{p}$ を固定して，各試行回数における AIC を計算した結果を下の表に示した．100 回コインを投げて 60 回表になったら偽物と思った方がよさそうである．

0.55			0.6		
n	AIC(0)	AIC(1)	n	AIC(0)	AIC(1)
50	69.3	70.8	50	69.3	69.3
100	138.6	139.6	100	138.6	136.6
200	277.3	277.3	200	277.3	271.2
300	415.9	414.9	300	415.9	405.8
400	554.5	552.5	400	554.5	540.4

3.5.2 正規分布モデルの AIC

正規分布に関する確率モデルとしては以下の４つが考えられる．

- データはもともと与えられた μ と σ^2 に従う．
- 平均は最尤推定量 $\hat{\mu}$ が正しい．
- 分散は最尤推定量 $\hat{\sigma}^2$ が正しい．
- 平均，分散ともに最尤推定量 $\hat{\mu}$ ， $\hat{\sigma}^2$ が正しい．

これら４つのモデルをそれぞれ，モデルA，モデルB，モデルC，そしてモデルDとする．これらの AIC を AIC(0) ， AIC(1) ， AIC'(1) ， AIC(2) とし，3.3.2 節の結果を用いて求めることにする．

モデルAの AIC は，対数尤度の式 (61) を用いて，

$$\mathrm{AIC}(0) = -2 \times l(\mu, \sigma^2) + 2 \times 0 = n \log(2\pi\sigma^2) + \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \qquad (84)$$

となる．モデルB，モデルCの結果は一部に最尤推定量を使うので，それぞれ，

$$\begin{aligned} \mathrm{AIC}(1) &= -2 \times l(\hat{\mu}, \sigma^2) + 2 \times \\ &= n \log(2\pi\sigma^2) + \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \hat{\mu})^2 + 2 \\ &= n \log(2\pi\sigma^2) + n\frac{\hat{\sigma}^2}{\sigma^2} + 2 \end{aligned} \qquad (85)$$

$$\mathrm{AIC'}(1) = -2 \times l(\mu, \hat{\sigma}^2) + 2 \times 1 = \log(2\pi\hat{\sigma}^2) + \frac{1}{\hat{\sigma}^2} \sum_{i=1}^{n} (x_i - \mu)^2 \qquad (86)$$

となる．モデルDはすべて最尤推定量で構成されるので，最大対数尤度の式である式 (65) を用いて，

$$\mathrm{AIC}(2) = -2 \times l(\hat{\mu}, \hat{\sigma}^2) + 2 \times 2 = n \log(2\pi\hat{\sigma}^2) + n + 4 \qquad (87)$$

となる．例えば機械の故障の判定などは，これらの４つのモデルを比較することによって結論を得ることができる．

3.5.3 多項式回帰モデルの AIC

3.3.3 節の結果より，n データのペア $(x_1, y_1), (x_2, y_2), \cdots (x_n, y_n)$ を m 次回帰モデル

$$y = a_0 + a_1 x + a_2 x^2 + \cdots + a_{m-1} x^{m-1} + a_m x^m \qquad (88)$$

に当てはめたときの AIC は，

$$\begin{aligned} \mathrm{AIC}(m) &= -2l(\hat{a}_0, \hat{a}_1, \cdots, \hat{a}_m, \hat{\sigma}^2) + 2 \times (m+2) \\ &= n\left(\log 2\pi + 1 + \log \hat{\sigma}^2\right) + 2(m+2) \end{aligned} \qquad (89)$$

となる．

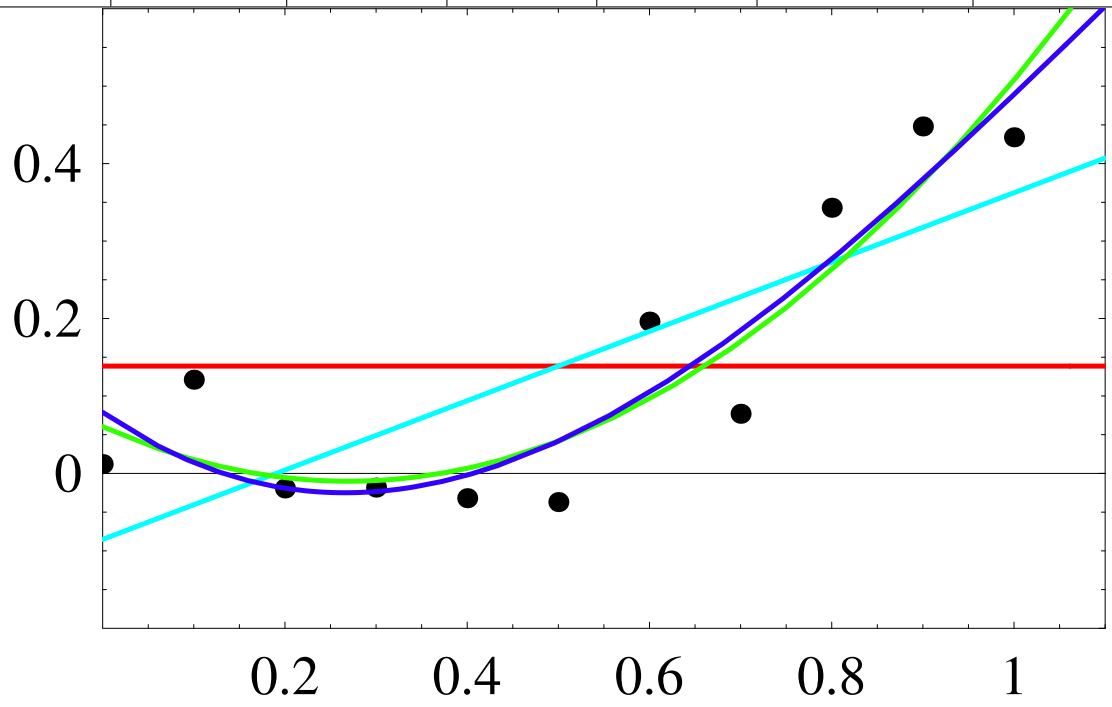
例えば，

i	1	2	3	4	5	6	7	8	9	10	11
x_i	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
y_i	0.012	0.121	-0.0194	-0.0183	-0.032	-0.037	0.196	0.077	0.343	0.448	0.434

というデータの組について，多項式回帰モデルを当てはめる．下の表に各次数における最尤推定量とそのときの最大対数尤度およびＡＩＣをまとめた．パラメー

タの数を多くすれば多くするほど分散の最尤推定量（残差平方和/データ数）が小さくなるので，対数尤度が大きくなりモデルを比較するのが困難であること．そして，ＡＩＣを用いるとその比較ができ，今回の場合，２次モデルが最良のモデルであると予想できることがわかる．実際にグラフに書いてみると２次と３次では曲線の形にほとんど違いがなく，パラメータの数を多くしてまで３次曲線で近似する必要がみられないことがわかる．

モデル	$\hat{a}_0$	$\hat{a}_1$	$\hat{a}_2$	$\hat{a}_3$	最大対数尤度	AIC
０次	0.138573				3.23417	-2.46835
１次	-0.0852364	0.447618			8.50008	-11.00017
２次	0.0602566	-0.522335	0.969953		13.37498	-18.74997
３次	0.078849	-0.817746	1.74463	-0.516453	13.53748	-17.07496



3.5.4 AR モデルの AIC

AR モデルとは，自己回帰モデルとも呼ばれ，等時間間隔で計測された時系列に対して，現在の自分を過去の自分の線形和として表すものである．これは過去に得られたデータ $x_1, x_2, \ldots, x_n$ をもとにして，現在の自分の値を

$$x_t = a_1 x_{t-1} + a_2 x_{t-2} \cdots + a_m x_{t-m} \tag{90}$$

$$= \sum_{i=1}^{m} a_m x_{t-m} \tag{91}$$

と近似して表そうというものである．このモデルには平均ゼロ，分散 σ^2 である誤差が伴うという確率モデルについて検討する．知りたいのは最適な m の値とそのときの最尤推定量 $\hat{a}_i$ $(i = 1, 2, \ldots, m)$ の値である．

最初にしなければならないのはデータの番号付けのやり直しである．もとのデータ $x_1, x_2, \ldots, x_n$ を $x_{1-m'}, x_{2-m'}, \ldots, x_0, x_1, x_2, \ldots, x_{n-m'}$ と置きなおす．このときの m' とは，比較検討する m' の値の最大値である．$x_{1-m'}, x_{2-m'}, \ldots, x_0$ というデータは観測データの個数には入れない．すなわち，$x_1, x_2, \ldots, x_{n-m'}$ を観測データとする．これは x_1 の推定値を計算するのに，それよりも過去のデータが必要なためである．以下ではデータの個数として，$n' = n - m'$ を用いて議論する．

誤差が正規分布すると仮定しているので，このモデルの対数尤度は，

$$l(a_1, a_2, \ldots, a_m, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{t=1}^{n'}\left(x_t - \sum_{i=1}^{m} a_i x_{t-i}\right)^2 \tag{92}$$

この式はデータとして x_{1-m} までを参照する．前述のようにデータの番号付けを変更する意味がここでわかる筈である．最尤推定量を求める方程式は，

$$\frac{\partial l}{\partial a_1} = \frac{\partial l}{\partial a_2} = \cdots = \frac{\partial l}{\partial a_m} = \frac{\partial l}{\partial \sigma^2} = 0 \tag{93}$$

となり，最後の方程式以外は，

$$\begin{bmatrix} C(1,1) & C(1,2) & C(1,3) & \cdots & C(1,m) \\ C(2,1) & C(2,2) & C(2,3) & \cdots & C(2,m) \\ \vdots & \vdots & \ddots & & \vdots \\ C(m,1) & C(m,2) & C(m,3) & \cdots & C(m,m) \end{bmatrix} \begin{bmatrix} \hat{a}_1 \\ \hat{a}_2 \\ \vdots \\ \hat{a}_m \end{bmatrix} = \begin{bmatrix} C(1,0) \\ C(2,0) \\ \vdots \\ C(m,0) \end{bmatrix} \tag{94}$$

とまとめられる．ここで $C(i,j)$ とは，

$$C(i,j) = \sum_{t=1}^{n'} x_{t-i} x_{t-j} \tag{95}$$

である．また最後の方程式を解くと，最尤推定量 $\hat{\sigma}^2$ を求めることができ，その結果は，

$$\hat{\sigma}^2 = \frac{1}{n'}\sum_{t=1}^{n'}\left(x_t - \sum_{i=1}^{m} a_i x_{t-i}\right)^2 = \frac{1}{n'}\left(C(0,0) - \sum_{i=1}^{m} \hat{a}_i C(i,0)\right) \tag{96}$$

となる．

3.5.5 重回帰モデルの AIC

重回帰モデルは，従属変数が複数の独立変数に対して線形な関係があると仮定するモデルである．独立変数の個数（変数の種類）を m とする．i 番目の独立変

数を x_i と表し，従属変数を y と表したときに，

$$y = a_1 x_1 + a_2 x_2 \cdots + a_m x_m \tag{97}$$

と表されると仮定する．このモデルには平均ゼロ，分散 σ^2 である誤差が伴うという確率モデルについて検討する．知りたいのは最尤推定量 $\hat{a}_i$ $(i = 1, 2, \ldots, m)$ の値である．

誤差が正規分布すると仮定しているので，このモデルの対数尤度は，

$$l(a_1, a_2, \ldots, a_m, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - a_0 - \sum_{j=1}^{m} a_j x_{ji}\right)^2 \tag{98}$$

最尤推定量を求める方程式は，

$$\frac{\partial l}{\partial a_1} = \frac{\partial l}{\partial a_2} = \cdots = \frac{\partial l}{\partial a_m} = \frac{\partial l}{\partial \sigma^2} = 0 \tag{99}$$

となり，最後の方程式以外は，

$$\begin{bmatrix} n & \sum x_{1i} & \sum x_{2i} & \cdots & \sum x_{mi} \\ \sum x_{1i} & \sum x_{1i}^2 & \sum x_{1i}x_{2i} & \cdots & \sum x_{1i}x_{mi} \\ \vdots & \vdots & \ddots & & \vdots \\ \sum x_{mi} & \sum x_{1i}x_{mi} & \sum x_{mi}x_{2i} & \cdots & \sum x_{mi}^2 \end{bmatrix} \begin{bmatrix} \hat{a}_1 \\ \hat{a}_2 \\ \vdots \\ \hat{a}_m \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_{1j}y_i \\ \vdots \\ \sum x_{mi}y_i \end{bmatrix} \tag{100}$$

とまとめられる．ただし，式中の $\sum$ はすべて $\sum_{i=1}^{n}$ の略である．これは以下でも同様である．また最後の方程式を解くと，最尤推定量 $\hat{\sigma}^2$ を求めることができ，その結果は，

$$\hat{\sigma}^2 = \frac{1}{n}\left\{\sum y_i^2 - a_0\sum y_i - \sum_{j=1}^{m} a_j \sum x_{ji}y_i\right\} \tag{101}$$

となる．最大対数尤度は，

$$l(\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_m, \hat{\sigma}^2) = -\frac{n}{2} - \frac{n}{2}\log\sigma^2 - \frac{n}{n} \tag{102}$$

となり，最終的に AIC は，

$$n(\log 2\pi + 1) + n\log\hat{\sigma}^2 + 2(m+2) \tag{103}$$

となる．

参考文献

[1] 坂本慶行，石黒真木夫，北川源四郎，情報量統計学，共立出版，1983