

平成 22 年度 社会心理学研究法 III 木曜 5 時限 (統計解析, Statistics for Social Psychology)

1 授業の概要と目標 The Outline of the Lecture

心理学や社会学研究のデータ分析に必要な統計的な分析方法について、初等統計の復習を兼ねた講義を行う。特に複数の説明変数を用いる分散分析と重回帰分析、および 2 値変数を被説明変数とするロジスティック回帰の理解を講義の当面の目標とする。(内容は昨年度の授業とほぼ同じ予定であるが、クラスター分析と因子分析についてもできれば触れたい)

統計的な分析手法は、曖昧さを含むデータから信頼しうる結論を導くさまざまな方法を提案している。実際にデータを取り扱う場合には、2 つの異なる側面を考える必要がある。(1) ひとつは、分析者が得ようとする結論が本当にデータによって支持されるか否かを検討することであり、(2) もうひとつは複雑なデータのなかに気づかれずに潜んでいる情報を発見することである。通常の「統計学」やデータ分析方法の授業では、前者に重点をおいた内容であることが多い。この講義では、両方の側面について説明を行う。

推薦する教科書、参考書

- 山田剛史・村井潤一郎 (2004) よくわかる心理統計、ミネルヴァ書房。(分かりやすい入門書)
- 柳井晴夫・緒方裕光編著 (2006) SPSS による統計データ解析、現代数学社。(SPSS の利用例)
- Rothman, K.J. (2004) 矢野栄二・橋本英樹監訳。ロスマンの疫学 科学的思考への誘い (Epidemiology An Introduction の訳)、篠原出版社 (統計的をもちいた現実的な因果推論の方法についての解説。本来、医学分野の教科書だが心理学にも有益)
- Agresti, A. (2003) カテゴリカルデータ解析入門, An Introduction to Categorical Data Analysis. Wiley (1996) の訳。(大変優れた教科書。特に調査データの分析に関心のある人には強く推薦します。), もっと詳しく知りたければ 同じ著者の Categorical Data Analysis, Wiley (2002) がこの分野の定番の教科書(分厚い)。
- 南風原朝和 (2002) 心理統計学の基礎 統合的理解のために。有斐閣 (心理統計全般についての分かりやすい参考書, 山田・村井よりは難しい)
- 古川俊之監修 丹後俊郎著 (1993). 新版 医学への統計学。朝倉書店。(あつかっている内容は各種の検定、推定法にわたる。医学分野の詳細な具体例が豊富であり内容が充実している。かなり詳しい。)
- 丹後俊郎, 山岡和枝, 高木晴良 (1996), ロジスティック回帰分析: SAS を利用した統計解析の実際, 朝倉書店。(2 値データの回帰分析についての教科書。おそらく日本語では一番詳しい。)
- Kleinbaum, D.G. & Klein, M. (2002). *Logistic regression: a self-learning text*, Springer-verlag. (ロジスティック回帰についての丁寧な教科書。SAS と SPSS の利用法の説明がある。かなり詳しい話題までふれられている。主に医学、薬学の学生を想定しているが、社会科学にも有益。)
- 足立浩平 (2006), 多変量データ解析法 心理・教育・社会系のための入門, ナカニシヤ出版。(クラスター分析や因子分析など多変量の相互関係の記述的分析法の解説。極力数式を使わずに説明しようとしている。)
- 大津起夫 「調査データからの推論」、『統計科学のフロンティア』10 巻 3 章 (岩波書店) は記述的な多変量解析の理論的な性質を検討している (かなり技術的)。NLSY79 (米国の大規模な継時的調査) を題材に使っている。

ソフトウェア (無償でかつ Windows で利用可能なものを紹介します。ただし、この授業での実習は SPSS を利用の予定。)

R システム <http://www.r-project.org/>, <http://www.okada.jp.org/RWiki/> 本格的な Splus 類似の統計分析システム。

<http://www.rd.dnc.ac.jp/~otsu/doc/Rguide.pdf> は大津作のイントロ資料

Ggobi (データ解析のためのダイナミック 3D グラフィックス)。

<http://www.ggobi.org/>

2 基礎数学の復習

これぐらいは使わないと、話が混乱するので。高校の数学の復習と「行列 (matrix)」について説明する。

2.1 添字と和記号

$$\begin{aligned}\sum_{i=1}^M x_i &= x_1 + x_2 + \cdots + x_M \\ \sum_{i=1}^M \sum_{j=1}^N x_{i,j} &= \sum_{i=1}^M (x_{i,1} + \cdots + x_{i,N}) \\ &= x_{1,1} + \cdots + x_{1,N} + x_{2,1} + \cdots + x_{2,N} + \cdots + x_{M,1} + \cdots + x_{M,N} \\ &= \sum_{j=1}^N \sum_{i=1}^M x_{i,j}\end{aligned}$$

2.2 積記号・階乗・組み合わせ

n 個の数字 $(1, 2, \dots, n)$ を並べ替える方法は

$$n! = n \times (n-1) \times (n-2) \times \cdots \times 2 \times 1$$

通りある。これは n の階乗 (factorial) と呼ばれる。特に $0! = 1$ と決められている。実は階乗は理論的に自然な方法で自然数以外の正の実数に拡張される。ガンマ関数 $\Gamma(x) = \int_0^\infty y^{x-1} e^{-y} dy$ は x が自然数であるとき、 $\Gamma(x) = (x-1)!$ を満たす関数であり、階乗の拡張とみなせる。

積記号 \prod は要素の積を表す。

$$\prod_{i=1}^M x_i = x_1 \times x_2 \times \cdots \times x_M$$

積記号を用いると、

$$n! = \prod_{k=1}^n k$$

と書ける。

n 個の数字 $(1, 2, \dots, n)$ から r 個の数字を重複せずに選択する方法は

$$\binom{n}{r} = \frac{n!}{r!(n-r)!}$$

通りある。これは、 n 個の識別可能な対象を r 個のグループと $(n-r)$ 個のグループとに分割する方法の数でもある。

n 個の数字 $(1, 2, \dots, n)$ を n_1, n_2, \dots, n_k 個のグループに分割する方法は (ただし $n_1 + n_2 + \dots + n_k = n$)

$$\frac{n!}{n_1!n_2!\cdots n_k!}$$

通りある。

2.3 指数と対数

年間 100 パーセントの利子がつく預金を考える (こんなのは普通ないが)。1 年間に利子が元金に加えられる階数が 1 回であれば、1 年後の預金は 2 倍になり、2 年後には 4 倍、 n 年後には 2^n 倍になる。

また、半年に 1 回利子が元金に加えられるならば、半年あたりの利子は元金の $1/2$ 倍であるから、半年後には元金は 1.5 倍になり、1 年後には元金は 1.5^2 倍になる。同様に、 n 年後には $(1 + 1/2)^{2n}$ 倍になる。

もし、1 年間に利子が元金に加えられる回数が k 回であるならば、1 年後には元金は $(1 + 1/k)^k$ 倍になり、 n 年後には $(1 + 1/k)^{kn}$ 倍になる。

k をどんどん大きくしてゆくと (つまり瞬間複利計算を考えることになる)、 $(1 + 1/k)^k$ は発散せずにある数に収束する。これが自然対数の底と呼ばれるものであり、 $e = 2.71828\dots$ として表される。理論的に e は重要な数であり、多くの数学公式に現れる。 e を底とする指数関数 e^x は $\exp(x)$ とも表記される。つぎの公式が知られている。

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \cdots$$

一般的に実数 a, b と任意の正の数 c について、つぎの公式がなりたつ。

$$c^0 = 1$$

$$c^1 = c$$

$$c^{1/2} = \sqrt{c}$$

$$c^{-a} = 1/c^a$$

$$c^{a+b} = c^a c^b$$

$$(c^a)^b = c^{ab}$$

$$\lim_{a \rightarrow 0} a^a = 1$$

上の公式は、指数 (冪乗をあらわす数) が整数の場合には、 c が負の数であっても成立する。

対数関数 \log は指数関数の逆関数として定義される。一般的に正の数 c について

$$c^x = y \quad \text{ならば} \quad x = \log_c y$$

と表される。 c を対数の底 (テイ) と呼ぶ。

特に、 $\exp(x) = y$ ならば、 $x = \log y$ である。 e を底とする対数は自然対数とよばれる。工学系の文献では、 \log の代わりに \ln と表記し、 \log は 10 を底とする対数 (常用対数) を表すことが多い。原則として、この授業では \log は自然対数を表すことにする。2 の常用対数は $\log_{10} 2 = 0.30103$ である。つまり常用対数が 0.3 違うということはほぼ 2 倍の違いがあることを示す。

対数関数について、つぎの公式が成り立つ (c は正の数)。

$$\begin{aligned}\log_c 1 &= 0 \\ \log_c c &= 1 \\ \log_c(\sqrt{a}) &= \frac{1}{2} \log_c a \\ \log_c(1/a) &= -\log_c a \\ \log_c(ab) &= \log_c a + \log_c b \\ \log_c a^b &= b \log_c a \\ \log_c a &= \frac{\log a}{\log c} \\ \lim_{a \rightarrow 0} \log_c a^a &= 0\end{aligned}$$

対数は、日常接する多くの単位に用いられている。例えば

- 地震のマグニチュード (エネルギーの対数)
- 星の等級 (明るさ、光のエネルギーの対数)
- pH(ペーハー ; 酸性、アルカリ性の指標) (水素イオン濃度の対数)
- dB(デシベル : 音の大きさの指標) (音のエネルギーの対数) dB は常用対数の 10 倍を単位とする。3dB の違いはエネルギーが 2 倍違うことを意味する。

指数関数 $\exp(x)$ は x が大きくなるとき急速に増加する関数であり、一方 $\log(x)$ は x が増加するときゆっくりと増加する関数である (図 1 参照)。

2.4 三角関数

- $\sin \theta$: 正弦関数
- $\cos \theta$: 余弦関数
- 通常三角関数の角度の単位にはラジアンを使う。ラジアンでは、 2π が 360 度を意味する。
- $(\sin \theta)^2 + (\cos \theta)^2 = 1$: この式は θ の値によらず成立する。
- $\tan \theta = \frac{\sin \theta}{\cos \theta}$: 正接関数
- $\cot \theta = \frac{\cos \theta}{\sin \theta}$: 余接関数

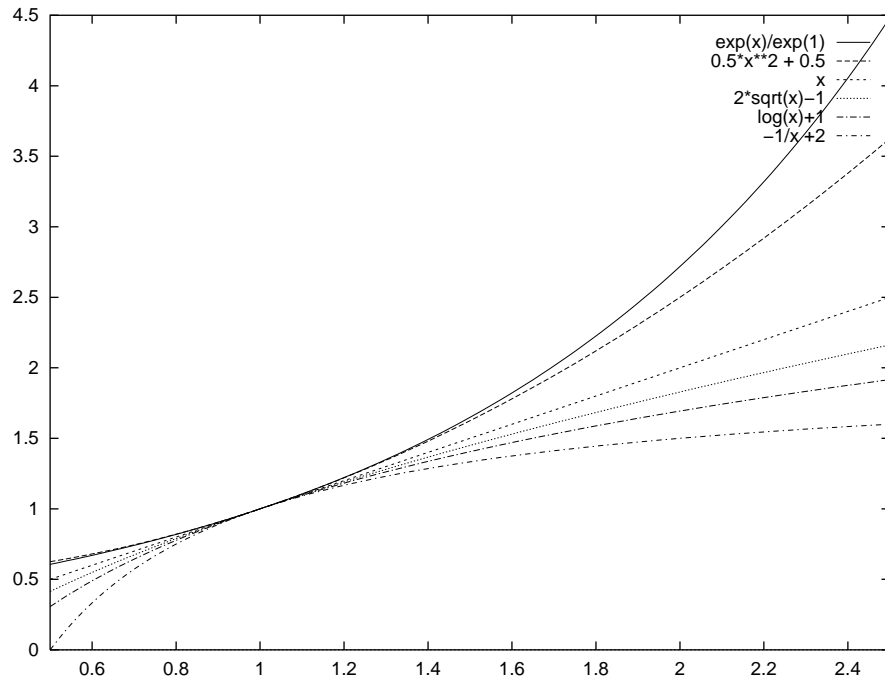


図 1: 関数の増加率の比較

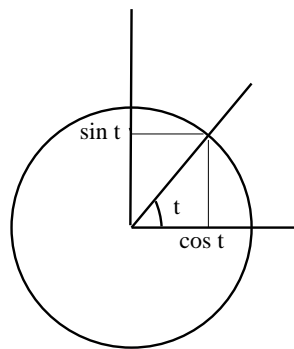


図 2: 三角関数の定義

2.5 ベクトル

$$\mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_k \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_k \end{pmatrix},$$

とする。

ベクトルの和

$$\mathbf{a} + \mathbf{b} = \begin{pmatrix} a_1 + b_1 \\ a_2 + b_2 \\ \vdots \\ a_k + b_k \end{pmatrix}$$

ベクトルのスカラー (定数) 倍

$$c\mathbf{a} = \begin{pmatrix} ca_1 \\ ca_2 \\ \vdots \\ ca_k \end{pmatrix}$$

ベクトルの内積

$$(\mathbf{a}, \mathbf{b}) = \sum_{i=1}^k a_i b_i$$

ベクトルのノルム (長さ) $\|\mathbf{a}\|$

$$\|\mathbf{a}\|^2 = (\mathbf{a}, \mathbf{a})$$

$$\|\mathbf{a}\| = \sqrt{(\mathbf{a}, \mathbf{a})}$$

ベクトルのノルムを用いると

$$(\mathbf{a}, \mathbf{b}) = \|\mathbf{a}\| \|\mathbf{b}\| \cos \theta$$

と書ける。ここで、 θ は2つのベクトルがなす角度である。

$(\mathbf{a}, \mathbf{b}) = 0$ のとき、 \mathbf{a} と \mathbf{b} とは直交するという。

3 行列と線形写像

この講義の目的は、多くの変数から情報を取り出す方法を学ぶことにあるが、多変数のデータを扱うための数学的な道具がベクトル (vector) と行列 (matrix) である。

3.1 行列の定義

行列は縦横に数値の並んだものである。

$$A = \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m,1} & a_{m,2} & \cdots & a_{m,n} \end{pmatrix} \quad (1)$$

この行列をさらに簡単に $A = (a_{i,j})$ などと書くこともある。行列の定数 (スカラー, scalar) 倍 αA は要素毎の定数倍 ($\alpha a_{i,j}$) であり、行列の和 $A + B$ は要素毎の和 ($a_{i,j} + b_{i,j}$) である。これだけでは、 $m \times n$ 次元のベクトルを考えるのと同じであるが、行列とベクトルおよび行列と行列の積を定義することにより、独自の意味を持たせることができる。ここで

$$\mathbf{x} = (x_1, x_2, \dots, x_n)^T, \quad \mathbf{y} = (y_1, y_2, \dots, y_m)^T$$

とする。ここでベクトルの肩についている T は、転置 (transpose, 行と列の交換) を意味する。つまり、ここで \mathbf{x} と \mathbf{y} はともに縦ベクトルである。ベクトルと行列の積はつぎのように定義される。

$$\mathbf{y} = A\mathbf{x} \quad (2)$$

要素で書き表すと、

$$\begin{aligned} y_1 &= a_{1,1}x_1 + a_{1,2}x_2 + \cdots + a_{1,n}x_n \\ y_2 &= a_{2,1}x_1 + a_{2,2}x_2 + \cdots + a_{2,n}x_n \\ &\vdots \\ y_m &= a_{m,1}x_1 + a_{m,2}x_2 + \cdots + a_{m,n}x_n \end{aligned}$$

である。

行列とベクトルの積の意味は、次のような例を考えればよい。

例1 \mathbf{x} の各要素 x_j は、商品 j の単価であり、行列の要素 $a_{i,j}$ は、顧客 i についての商品 j の売上数量とする。このとき、 y_i は顧客 i についての売上高である。

例2 x_j は、ある学生のセンター入試の科目 j の得点である。学生は、同一の試験によって複数の大学を受験することができる。各大学は、試験の素点ではなく、重みづけを行なって採点するものとし、行列要素の値 $a_{i,j}$ は大学 i における科目 j の重み係数であるとする。このとき y_i は、大学 i における重みづけの得点である。

行列 A は、 n 次元のベクトル全体を定義域 (domain) とし、 m 次元のベクトルを値域 (range) とする関数とみなすことができる。通常関数 ($\sin(x)$ や $x^2 - x + 3$ など) は、実数 (の一部分) を定義域とし、値域も実数であるが、ここではこれらをベクトル (つまり実数の組) に拡大して考えている。

行列とベクトルの積について、つぎの2つの性質が成り立つ。

$$A(\alpha x) = \alpha Ax \tag{3}$$

ここで α は実数である。

$$A(x_1 + x_2) = Ax_1 + Ax_2 \tag{4}$$

この2つの性質が成り立つ関数 (写像) のことを、線形関数 (linear function) または線形写像 (linear mapping) という。上の2つの性質を組み合わせると、

$$A(\alpha x_1 + \beta x_2) = \alpha Ax_1 + \beta Ax_2 \tag{5}$$

が成り立つことがわかる。

行列 B を $n \times k$ の大きさであり、各列ベクトルが b_1, \dots, b_k であるとする。行列 A と B の積 AB は

$$AB = (Ab_1, \dots, Ab_k) \tag{6}$$

として定められる。ここで、 (AB) は $m \times k$ の行列であり、任意の k 次元ベクトル x について、 $(AB)x = A(Bx)$ が成立する。一般的には、 AB と BA とは等しくない。また、 $(AB)^T = B^T A^T$ である。

3.2 行列のランク (階数)

ある行列が与えられたとき、それを特徴づける様々な値があるが、その一つにランク (rank) がある。ランクは階数とも呼ばれる。ランクは1次独立 (linearly independent) と呼ばれるベクトルの組についての性質に基づいている。

ベクトルの組 a_1, \dots, a_n が与えられているとする。これらについて

$$\alpha_1 a_1 + \dots + \alpha_n a_n = \mathbf{0} \tag{7}$$

となる実数の組 $(\alpha_1, \dots, \alpha_n)$ が存在し、しかも α_i ($i = 1, \dots, n$) 全てがゼロではない、つまり少なくとも一つの α_i はゼロでないとき、 a_1, \dots, a_n は1次従属 (linearly dependent) であるという。ここで $\mathbf{0}$ は全ての要素がゼロであるベクトルのことである。 x_1, \dots, x_n が1次従属ではないときすなわち、(7) が成立するのは $\alpha_1, \dots, \alpha_n$ がすべてゼロであるときに限られるとき、 x_1, \dots, x_n は互いに1次独立であるという。

m 行 n 列の行列 $A = (a_{ij})$ について、その n 個の列ベクトルを a_1, \dots, a_n と表記する。ある n 次元ベクトル x に A を作用させると、

$$y = Ax = x_1 a_1 + \dots + x_n a_n$$

となる。 x の値が様々に変化すると y の値もそれにつれて変わる。しかし、もし A が特別な値である場合には y の取り得る値は、ある特定の範囲に限定されたものであるかも知れない。極端な場合、 A の要素が全てゼロであれば y はゼロベクトル以外にはなり得ない。このようにして作られ

る y の全体を線形写像 A の像 (image) と呼び、 $\text{Im}A$ と表記する。 A のランク ($\text{rank}A$) は、 $\text{Im}A$ の次元 (1 次独立なベクトルを最大何個とれるか) によって定義される。これはまた、 a_1, \dots, a_n の中から選び出される 1 次独立なベクトルの最大の個数に等しい。いささか込み入った証明が必要であるが、 $\text{rank}A = \text{rank}A^T$ であることが示せる。また、 $\text{rank}A = \text{rank}AA^T = \text{rank}A^T A$ も成立する。

行列 A のカーネル (kernel) は $Ax = 0$ となる x の全体として定義され、 $\text{Ker}A$ と表記される。 x_1 と x_2 がともに A のカーネル $\text{Ker}A$ の要素ならば $\alpha x_1 + \beta x_2$ も $\text{Ker}A$ の要素である。

A の次元が $m \times n$ であるとき、 $\text{Im}A$ の次元と $\text{Ker}A$ の次元を足すと常に n となる。

3.3 連立 1 次方程式と逆行列

A を $p \times p$ の正方行列、 b と x を長さ p の縦ベクトルとする。 b は値がわかっているが、 x は未知であるとする。このとき式 $Ax = b$ は、 p 元連立 1 次方程式をあらわす。これについて次のようなことがわかっている。

1. $\text{rank}A = p$ であれば、 x の値は、常に一通りに定まる。このとき A は正則 (regular) であるという。
2. $\det A = 0$ である場合には、解が存在しない場合 (不能) と、解が複数存在する場合 (不定) とがある。 A が正則でないことを特異 (singular) であるという。解が複数存在するのは、 $\text{rank}A = \text{rank}(A|b) < p$ のときであり、解が存在しないのは $\text{rank}A < \text{rank}(A|b)$ の場合である。ここで $(A|b)$ は A の横に縦ベクトル b をならべた $p \times (p+1)$ の行列である。

A が正則であるときには、 $Ax = 0$ となるベクトル x はゼロベクトルのみである。また、このとき逆行列 (inverse matrix) A^{-1} が存在する。 A の逆行列とは、 $AX = I_p$ となる行列 X のことである。ここで I_p は、 p 次の単位行列 (identity matrix) を表す。単位行列とは対角成分が全て 1 であり、それ以外の成分が全てゼロである正方行列 (square matrix) (行と列の数が等しい) のことである。 X の列ベクトルを x_1, \dots, x_p とし、また I_p の j 列を e_j とする。方程式 $AX = I_p$ の両辺の j 列をとると、 $Ax_j = e_j$ である。 A が正則であれば、これらの p 個の連立 1 次方程式は必ず解を持つので、 x_j が求まり、 X を定めることができる。 $AX = I_p$ が成立すれば、つぎの議論から $XA = I_p$ であることも分かる。

$XA = Y$ とおいてみる。 $AXA = A = AY$ である。 $A = AI_p$ なので、 $A(I_p - Y)$ はゼロ行列である。 A は正則なので $I_p - Y$ の各列がゼロベクトルであり、 $Y = I_p$ となることがわかる。逆行列にはつぎのような性質がある。

1. $(AB)^{-1} = B^{-1}A^{-1}$
2. $(A^T)^{-1} = (A^{-1})^T$
3. 上三角行列 (upper triangular matrix) の逆行列は、上三角行列。
4. 下三角行列 (lower triangular matrix) の逆行列は、下三角行列。

3.4 行列式 (determinant)

$p \times p$ の正方行列 A について $\det A$ または $|A|$ という記法によって A の行列式をあらわす。 $A = (a_1, a_2, \dots, a_p)$ としよう。 A の行列式は A の p 個の列ベクトルによって構成される平行 (超)

多面体の符号付きの体積である。 $p = 2$ のときは、2つのベクトル a_1, a_2 の2つのベクトルによって形づくられる平行四辺形の面積、すなわち $0, a_1, a_1 + a_2, a_2$ の4点で囲まれた領域の面積である。ただし、ベクトルの位置関係によって符号が変わる。 a_1 から a_2 への方向の変化が逆時計回り(正の角度)のときには正符号であり、時計回り(負の角度)の時には負符号である。ただし回転の角度は180度以下とする。 $p = 3$ の時は3つのベクトルによって構成される平行6面体(直方体を歪めたもの)の体積であり、ただしこの場合もベクトルの向きによって符号が変わる。厳密には、行列式は次の性質を満たす正方行列の関数として定義される。

1. $\det I = 1$ ただし I は単位行列。
2. ある列を c 倍すると、行列式の値も c 倍になる。
3. A の第 j 列 a_j を別のベクトル b_j に置き換えた行列を B とする。また、 $a_j + b_j$ に置き換えた行列を C とする。このとき $\det C = \det A + \det B$ となる。
4. 2つの列を交換すると、行列式は絶対値が同じで符号が逆転する。

以上の性質を満たすものは、実は一通りしかない。上に述べた平行4辺形の符号つき面積や平行6面体の符号つき体積は、上の性質を満たすことが直観的にわかる。

また、行列式の定義からつぎの性質が導かれる。

1. 2つの列が同じベクトルであれば行列式はゼロになる。
2. ある列が他の列の定数倍であれば、つまり $a_i = \alpha a_j, (i \neq j)$ ならば、行列式はゼロである。
3. ある列を定数倍したベクトルを別の列に加えても、行列式の値は変わらない。
4. A を $p \times p$ の行列とする。つぎの4つの条件は全て同値である。このとき A は正則 (regular) であるという。

$\det A \neq 0$ であること

A の列が1次独立であること

$A_{p \times p}$ のランクが p であること

逆行列 A^{-1} が存在すること

5. $\det(AB) = \det(A) \det(B)$

6. $\det A^{-1} = 1/\det A$

7. $\det A = \det A^T$ (T は転置を示す。)

8. 上の定義に示した性質は、列を行と読み換えてもすべて成立する。

9. $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ のとき、 $\det A = ad - bc$

10. A が対角行列または三角行列の場合には、 $\det A = a_{11} \times a_{22} \times \cdots \times a_{pp}$ である。

多変数の微積分学や座標変換の計算で行列式はしばしば用いられる。

3.5 答えのない連立 1 次方程式と最小 2 乗法 (least square method)

連立 1 次方程式

$$Ax = b \quad (8)$$

において、この式を満足する x が存在するのは、 $\text{rank}A = \text{rank}(A|b)$ のときであると説明した。しかし、この条件が成立しない場合にも解に「近い」 x がどれであるかを定めることはできる。ここで、 $y = Ax$ とおく。 x の解としての良さを $\|b - y\|^2$ によって定義する。これはベクトルの次元を p とするとき、

$$\sum_{i=1}^p (b_i - y_i)^2 \quad (9)$$

となる。 x の値を調節して、これを小さくするような y をつくり出すことについて考えよう。

ここで、 A の次元は $n \times p (n \geq p)$ であり、 A のランクは p であるとする。つまり、 A の各列は 1 次独立なベクトルである。簡単のために $n = 3, p = 2$ の場合を考えてみよう。 $A = (a_1, a_2)$ とすると a_1, a_2 の各々は次元 3 のベクトルであり、 3 次元空間の中の同一直線上にはない。また、 $x_1 a_1 + x_2 a_2$ の全体 (x_1, x_2 が様々に変化する場合のベクトル全体) は、原点を通る一つの平面をなす。

より具体的に $a_1 = (1, 1, 1)^T$, $a_2 = (0, 1, 2)^T$ としよう。この場合 $b = (0, 2, 3)^T$ とおくと、 $Ax = b$ を満たす x は存在しない。多変数の微分を用いた計算から、実は

$$\hat{x} = (A^T A)^{-1} A^T b \quad (10)$$

が式 (9) を最小にすることがわかっている。これを使うと $\hat{y} = A\hat{x} = A(A^T A)^{-1} A^T b$ である。計算すると $A^T A = \begin{pmatrix} 3 & 3 \\ 3 & 5 \end{pmatrix}$ であり、 $(A^T A)^{-1} = \begin{pmatrix} 5/6 & -1/2 \\ -1/2 & 1/2 \end{pmatrix}$ である。これらを使うと $x_1 = 1/6, x_2 = 3/2$ のとき $\hat{y} = (1/6, 10/6, 19/6)^T$ であり、これが b に一番「近い」値であることがわかる。

4 多変数の微分

1 変数の微分の定義をつぎのようなものである。

$$f'(x_0) = \frac{df}{dx}(x_0) = \lim_{d \rightarrow 0} \frac{f(x_0 + d) - f(x_0)}{d} \quad (11)$$

また 2 次の微分はつぎのように定義される。

$$f''(x_0) = \frac{d^2 f}{dx^2}(x_0) = \lim_{d \rightarrow 0} \frac{f'(x_0 + d) - f'(x_0)}{d} \quad (12)$$

関数 f が複数の変数に基づいて定義される場合には、それぞれの変数について微分を考えることができる。 f がふたつの変数 x と y との関数であるとする。それぞれの変数についての微分は、本質的には 1 変数の場合と同様に定義されるが、複数の変数の関数であることを意識する場合にはつぎのような偏微分 (partial derivative) 記号を用いて表す。

$$\begin{aligned} \frac{\partial f}{\partial x}(x_0, y_0) &= \lim_{d \rightarrow 0} \frac{f(x_0 + d, y_0) - f(x_0, y_0)}{d} \\ \frac{\partial f}{\partial y}(x_0, y_0) &= \lim_{d \rightarrow 0} \frac{f(x_0, y_0 + d) - f(x_0, y_0)}{d} \end{aligned}$$

2 次の偏微分

$$\frac{\partial^2 f}{\partial x^2}, \frac{\partial^2 f}{\partial y^2}, \frac{\partial^2 f}{\partial x \partial y}$$

なども 1 変数と類似の方法で定義される。もし関数 f が十分に滑らかなら

$$\frac{\partial}{\partial x} \left(\frac{\partial f}{\partial y} \right) = \frac{\partial}{\partial y} \left(\frac{\partial f}{\partial x} \right)$$

が成り立つ (偏微分の順番によらない) ので、 $\frac{\partial^2 f}{\partial x \partial y}$ の微分を行っている変数の順番は問題ではない。ただし関数が滑らかではない特殊な場合には順番によって値が異なることもありうる。

ここで f が x の 2 次関数であり、つぎのような式で定義されているとしよう。

$$f(x) = \frac{ax^2}{2} + bx + c$$

このとき

$$\frac{df}{dx}(x) = ax + b, \quad \frac{d^2f}{dx^2}(x) = a$$

となる。また g が x と y のつぎのような 2 次関数であるとしてよう。

$$f(x, y) = \frac{a_{11}x^2}{2} + \frac{a_{22}y^2}{2} + a_{12}xy + b_1x + b_2y + c \quad (13)$$

このとき 1 次と 2 次の微分はそれぞれつぎのような式になる。

$$\frac{\partial f}{\partial x}(x, y) = a_{11}x + a_{12}y + b_1, \quad \frac{\partial f}{\partial y}(x, y) = a_{12}x + a_{22}y + b_2, \quad (14)$$

$$\frac{\partial^2 f}{\partial x^2}(x, y) = a_{11}, \quad \frac{\partial^2 f}{\partial y^2}(x, y) = a_{22}, \quad \frac{\partial^2 f}{\partial x \partial y}(x, y) = a_{12} \quad (15)$$

複数の変数をもつ関数の 1 次微分と 2 次微分はしばしば言及されるため特別な呼び名を持っている。1 次微分を並べたベクトル

$$\mathbf{g}(x, y) = \begin{pmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{pmatrix}$$

はグラディエント (傾斜) ベクトルと呼ばれる。グラディエントベクトルは関数の傾斜の方向 (山側) を示し、ベクトルの長さが大きいことは傾斜が急であることを示し、要素の値がゼロに近いことは傾斜が平坦であることを示している。また、2 次の微分を変数の順にならべた行列

$$\mathbf{H}(x, y) = \begin{pmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} \\ \frac{\partial^2 f}{\partial y \partial x} & \frac{\partial^2 f}{\partial y^2} \end{pmatrix}$$

を Hesse 行列と呼ぶ。

ここで (13) の表す関数がお椀型をしており最小点があると仮定しよう。最小点の近くでは関数はほぼ水平であるので、 x 方向も y 方向もともに 1 次微分がゼロになっているはずである。これはつぎのような連立 1 次方程式によって表される。ただしここで $a_{21} = a_{12}$ とする。

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

この条件を満たす x と y はつぎのように書ける。

$$\begin{pmatrix} x \\ y \end{pmatrix} = - \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}^{-1} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$$

5 確率変数・平均・分散

確率変数 (random variable) とは、値が確率的に変動する変数のことである。確率変数がどのような値をとるかは確率分布 (probability distribution) によって定義される。確率変数の数学的な表現は、変数が連続的 (continuous) に変化するものであるか、あるいは離散的 (discrete) な値をとるかによって異なる。ここでは、まず、離散的な値をとる場合を考える。

$Pr(X = x)$ によって、確率変数 X が値 x をとる確率を表すことにする。一般的には、この値は x の関数として定まるから、 $Pr(X = x) = p(x)$ と書ける。同様に $Pr(X = x, Y = y)$ によって、確率変数 X が値 x をとり、しかも確率変数 Y が値 y をとる確率を表す。2 つ以上の確率変数が同時にどのような値をとるかを指定する確率分布のことを同時分布 (simultaneous distribution) という。同時分布が与えられているとき、 $Pr(X = x) = \sum_y Pr(X = x, Y = y)$ となるが、これを X の周辺分布 (marginal distribution) と呼ぶ。 Y の周辺分布も同様にして定義される。

例 サイコロを 2 回投げることをし、 X を 1 回目の目の値とし、 Y を 2 回目の目の値とする。 $Z = X + Y$ と定義する。 X と Y の同時分布は $x = 1, 2, \dots, 6$ および $y = 1, 2, \dots, 6$ について、 $Pr(X = x, Y = y) \equiv 1/36$ で与えられる。一方 X と Z の同時分布は、 $x = 1, 2, \dots, 6$ 、 $z = x + 1, x + 2, \dots, x + 6$ について $Pr(X = x, Z = z) = 1/36$ となる。

X と Y の同時確率分布が与えられているとする。ここで、 $Pr(X = x) \neq 0$ であるならば、 $X = x$ となる場合に限って Y の分布を考えることができる。これを $X = x$ の下での Y の条件付き分布 (conditional distribution) と呼ぶ。条件付き分布の確率は $Pr(Y = y|X = x)$ の様に表され、つぎの公式が成り立つ。

$$Pr(Y = y|X = x) = \frac{Pr(Y = y, X = x)}{Pr(X = x)} \quad (16)$$

同時確率分布は、一般的には $Pr(X = x, Y = y) = p_{XY}(x, y)$ と表されるが、特に $Pr(X = x, Y = y) = p_X(x)p_Y(y)$ のように確率が 2 つの変数の関数の積として表されるとき、確率変数 X と Y とは統計的に独立 (independent) であるという。上式から、 X と Y とが独立であるならば、条件付き分布は

$$Pr(Y = y|X = x) = p_Y(y), \quad Pr(X = x|Y = y) = p_X(x)$$

となり、条件を与える変数の値によらず分布が一定であることがわかる。上のサイコロの例においては、 X と Y とは統計的に独立であるが、 X と Z とは独立ではない。

確率変数 X が連続である場合には、 X が特定の一つの値をとる確率はほとんどゼロである。意味があるのは特定の区間の範囲の値をとる確率である。 X が x 以下の値をとる確率が $F(x)$ であらわされるとき、 $F(x)$ を累積分布関数 (cumulative distribution function) または単に分布関数 (distribution function) と呼ぶ。確率変数 X が a よりも大きく b 以下である確率は、

$$Pr(a < X \leq b) = F(b) - F(a) \quad (17)$$

と表される。ここで、 $F(x)$ が微分可能のとき

$$f(x) = \lim_{\varepsilon \rightarrow 0} \frac{F(x + \varepsilon) - F(x)}{\varepsilon} = F'(x) \quad (18)$$

によって定義される関数 $f(x)$ ($F(x)$ の微分) を確率密度関数 (probability density function) と呼ぶ。確率密度関数を用いると

$$Pr(a < X \leq b) = \int_a^b f(x) dx \quad (19)$$

となる。 X と Y が連続である場合についても、離散値をとる場合と同様に、同時分布、周辺分布、条件付き分布を考えることができる。

確率変数 X の特徴を表す代表的な値は、期待値 (expected value)(または平均 (mean)) と分散 (variance) である。期待値は次の式によって定義される。

$$E(X) = \begin{cases} \sum_x xp(x) \\ \int xf(x)dx \end{cases} \quad (20)$$

また分散は

$$Var(X) = \begin{cases} \sum_x (x - \mu)^2 p(x) \\ \int (x - \mu)^2 f(x)dx \end{cases} \quad (21)$$

によって定義される。ここで $\mu = E(X)$ である。期待値は分布の位置 (確率変数の大小) についての情報を与え、分散は分布の広がりについての情報を与える。(21) において μ を X の平均以外の値にとると、分散よりも大きくなる。つまり、 μ_1 を平均以外の数とすると、

$$\int (x - \mu_1)^2 f(x)dx \geq \int (x - \mu)^2 f(x)dx \quad (22)$$

となる。離散的な場合についても同様である。

上にのべた平均と分散は、確率分布の特徴を与える量として定義されるものであり、観測されたデータの平均や分散とは異なることに注意する必要がある。特にこれらを区別する必要があるときには、前者 (確率分布の特徴) を、母平均 (population mean) ・母分散 (population variance) と呼び、後者 (データの特徴) を標本平均 (sample mean) ・標本分散 (sample variance) と呼ぶ。データが x_1, \dots, x_n で与えられているとき、標本平均 \bar{x} と標本分散 $\text{var}(x)$ とはそれぞれ

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \quad \text{var}(x) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = s^2 \quad (23)$$

と定義する。教科書によっては、標本の分散を

$$\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (24)$$

と定義しているものもあるが、この講義ではこの値を不偏分散 (unbiased variance) と呼んで区別することにする。ただし、これは正しくは母分散の不偏推定量 (unbiased estimator of population variance) と呼ぶべきものであり、不偏分散というのは正式な名称ではない。

母平均と母分散について、つぎの公式がなりたつ。ここで、 a, b は定数であり、 X と Y は確率変数である。

$$E(aX + bY) = aE(X) + bE(Y) \quad (25)$$

$$Var(aX + bY) = a^2 Var(X) + 2ab Cov(X, Y) + b^2 Var(Y) \quad (26)$$

ここで $Cov(X, Y)$ は、 X と Y の (母) 共分散であり、つぎの式で与えられる。

$$Cov(X, Y) = \begin{cases} \sum_x \sum_y (x - \mu_X)(y - \mu_Y)p(x, y) \\ \int \int (x - \mu_X)(y - \mu_Y)f(x, y)dx dy \end{cases} \quad (27)$$

ただし、 $\mu_X = E(X)$ 、 $\mu_Y = E(Y)$ であり、 $f(x, y)$ は X と Y との同時確率密度関数である。

また、標本平均と標本分散についても同様の公式がなりたつ。

$$\overline{ax + by} = a\bar{x} + b\bar{y} \quad (28)$$

$$\text{var}(ax + by) = a^2\text{var}(x) + 2abcov(x, y) + b^2\text{var}(y) \quad (29)$$

ここで、 $cov(x, y)$ は標本共分散

$$cov(x, y) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \quad (30)$$

である。

標本平均と標本分散は、それぞれデータの値の1次式および2次式の和として定義されるが、さらに一般してデータ3次式および4次式として表される量をデータの分布の特徴を表すものとして利用することもある。詳細な統計分析を行う場合に利用されるものとして、以下のものがある。

$$\text{平均値の周りの3次モーメント } \nu_3 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3$$

$$\text{平均値の周りの4次モーメント } \nu_4 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4$$

これらを標本分散 s^2 を用いて標準化した値が実際にはよく用いられる。

$$\text{歪度 (skewness)} \quad \frac{\nu_3}{(\sqrt{s^2})^3}$$

$$\text{尖度 (kurtosis)} \quad \frac{\nu_4}{(\sqrt{s^2})^4}$$

1変数の離散値をとる分布として代表的なものには、2項分布 (binomial distribution)、ポアソン分布 (Poisson distribution)、負の2項分布 (negative binomial distribution)、超幾何分布 (hyper geometric distribution) などがあるが、ここでは2項分布のみ説明する。ポアソン分布については、対数線形モデルと関連して後で説明する。

2項分布 (Binomial Distribution)

確率変数 X が2項分布 $Bin(p, 1)$ に従うとは、 $X = 1$ の生じる確率が p であり、 $X = 0$ の生じる確率が $1 - p$ であることを意味する。また、 Y が2項分布 $Bin(p, n)$ に従うとは、

$$Pr(Y = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad (k = 0, 1, \dots, n) \quad (31)$$

の確率を Y がとることをいう。各 $X_i, (i = 1, \dots, n)$ が $Bin(p, 1)$ の分布を持つ確率変数であり、しかも互いに独立であるとき $Y = X_1 + \dots + X_n$ として Y を定めると、 Y は $Bin(p, n)$ に従う。確率変数 $Y \sim Bin(p, n)$ の平均は np 、分散は $np(1-p)$ である。

1変数の連続確率分布としては、一様分布 (uniform distribution)、正規分布 (normal distribution)、カイ2乗分布 (χ^2 distribution)、 t 分布 (t distribution)、 F 分布 (F distribution) など様々なものがある。

一様分布 (uniform distribution)

一様分布は指定された区間 $[a, b]$ において、一定の大きさの確率密度関数を持つ分布のことである。確率密度関数はつぎのように定義される。

$$f(x) = \begin{cases} 1/(b-a), & a \leq x \leq b \text{ について} \\ 0, & \text{それ以外} \end{cases} \quad (32)$$

$$F(x) = \begin{cases} 0, & x < a \text{ のとき} \\ (x-a)/(b-a), & a \leq x \leq b \text{ のとき} \\ 1, & x > b \text{ のとき} \end{cases} \quad (33)$$

一様分布の平均と分散は

$$E(X) = (a+b)/2, \quad \text{Var}(X) = (b-a)^2/12 \quad (34)$$

であたえられる。

正規分布 (normal distribution)

正規分布の確率密度関数はつぎの関数であらわされる。

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \quad (35)$$

上の式で与えられる正規分布の平均は μ であり、分散は σ^2 である。特に、平均が 0、分散が 1 の正規分布を標準正規分布 (standard normal distribution) と呼ぶ。図 3 参照。正規分布の歪度はゼロ、尖度は 3 である。

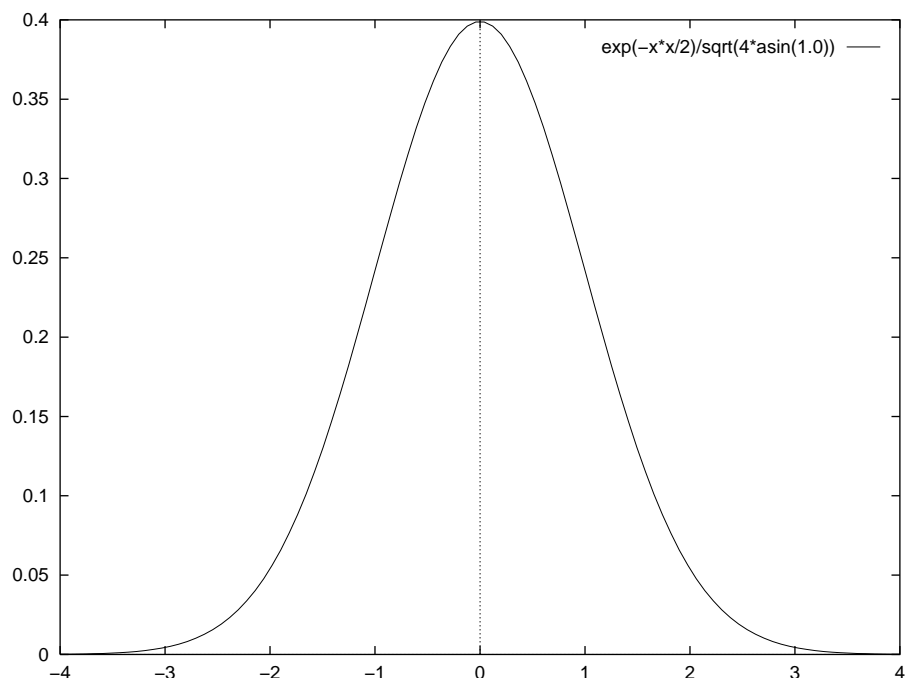


図 3: 標準正規分布の確率密度関数
確率密度関数は $\frac{1}{\sqrt{2\pi}} \exp(-x^2/2)$ である。

6 標本調査

ある母集団の平均や分散について推論するためには、その母集団から有限個の要素（標本, sample）を抽出し、それらの値に基づいて値を求める必要がある。

この場合、標本をどのように母集団から抽出するかが問題になる。通常は、標本が母集団（無限個の要素を持つ）から完全にランダムに抽出されたものとする。

しかし、母集団が有限個の場合には、抽出されたサンプルを母集団へ戻して再び抽出を行うか（復元抽出）、あるいは戻さないで抽出を行うか（非復元抽出）のどちらかによって、結果が異なる。

また、大規模な社会調査などでは推定値の安定を図るために、母集団を等質な複数の群に分割し、各々の群からそれらの大きさに比例した標本を抽出する方法も、しばしば用いられる（層別抽出）。母集団の対象の数を N 、標本（サンプル）の件数を n とする。母集団の平均は μ 、分散は σ^2 とする。

表 1: 標本平均の分布

抽出方法	\bar{x} の期待値	\bar{x} の分散	\bar{x} の標準偏差
復元抽出	μ	σ^2/n	σ/\sqrt{n}
非復元抽出	μ	$\sigma^2/n \times (1 - n/N)$	$\sigma/\sqrt{n} \times \sqrt{1 - n/N}$

比率の平均と分散 X が 0 – 1 値をとる変数であるとする。母集団での $X = 1$ の確率を π とする。標本の件数を n とする。

表 2: 比率の平均と分散

	平均	分散
母集団	π	$\pi(1 - \pi)$
復元抽出	π	$\pi(1 - \pi)/n$
非復元抽出	π	$\pi(1 - \pi)/n \times (1 - n/N)$

層別抽出

第 i 群の母集団の平均を μ_i 、分散を σ_i^2 とする。また i 群の全体における比率を w_i とする。母集団全体から無作為に n 件の標本を抽出する場合と、各層から w_i に比例した件数 n_i (合計で n 件) の標本を抽出する場合とを比較する。

表 3: 標本平均の分布 (層別)

	平均	分散
母集団	$\mu = \sum_i w_i \mu_i$	$\sigma^2 = \sum_i w_i [(\mu_i - \mu)^2 + \sigma_i^2]$
無作為復元抽出	μ	σ^2/n
層別復元抽出	$\mu = \sum_i (n_i/n) \mu_i = \sum_i w_i \mu_i$	$1/n \times \sum_i (n_i/n) \sigma_i^2 = 1/n \times \sum_i w_i \sigma_i^2$

7 2 群のデータの比較

(注意：以下では標本の件数を N と表記している)

7.1 標本平均の分布

定理：

母集団の平均が μ 、分散が σ^2 とする。これから独立 (independent) に抽出された N 個の標本が $\{x_1, \dots, x_N\}$ であるとする。これらの標本の平均 $\bar{x} = \sum_{i=1}^N x_i$ について次が成り立つ。

1. \bar{x} の期待値は μ である。つまり $E(\bar{x}) = \mu$.
2. \bar{x} の分散は σ^2/N であり、標準偏差は $\sqrt{\sigma^2/N}$ 。つまり、平均の標準偏差は 1 個の場合の $1/\sqrt{N}$ になる。

平均の精度を一桁あげようとする、標本数は 100 倍にする必要がある。

この定理は母集団がどのようなものであっても、分散が計算できるなら (有限の値を持つなら) 成立する。(分散が無限になる分布もある。この場合は、どんなに多くの標本を用いて平均を求めても、その分散は小さくならない。)

2 群のデータにおける統計量 (平均、分散など) を比較することは、最も基本的な統計的な分析の 1 つである。ここでは、Cleveland(1993) にとりあげられている例を紹介する。原典は Frisby & Clatworthy (1975) によるランダムドットステレオグラムの反応時間の測定実験である。Statlib の Data and story library (DASL) より入手したものを用いる。(http://lib.stat.cmu.edu/DASL/)

このデータは各行が 2 つの値 (1 列目が反応時間、2 列目が教示の種類) から構成されている。教示の種類は NV または VV のいずれかである。教示 NV はどのような図形が両眼の画像の融合によって生じるかを、言語的にあらかじめ説明するかまたは全く説明なしであることを意味し、VV は言語的な説明と実際に図形による説明の両者を行なったことを意味している。反応時間は、刺激図形 (ランダムドットステレオグラム) が提示されてから、被験者が立体画像を認識するまでの時間である。¹

¹M.Friendly による DASL のページにはデータ件数は 81 となっているが、実際に表示されているデータは 78 件であった。

データを以下に示す。

No.	Time	Condition	No.	Time	Condition
1	47.2	NV	44	19.7	VV
2	22.0	NV	45	16.2	VV
3	20.4	NV	46	15.9	VV
4	19.7	NV	47	15.4	VV
5	17.4	NV	48	9.7	VV
6	14.7	NV	49	8.9	VV
7	13.4	NV	50	8.6	VV
8	13.0	NV	51	8.6	VV
9	12.3	NV	52	7.4	VV
10	12.2	NV	53	6.3	VV
11	10.3	NV	54	6.1	VV
12	9.7	NV	55	6.0	VV
13	9.7	NV	56	6.0	VV
14	9.5	NV	57	5.9	VV
15	9.1	NV	58	4.9	VV
16	8.9	NV	59	4.6	VV
17	8.9	NV	60	3.8	VV
18	8.4	NV	61	3.6	VV
19	8.1	NV	62	3.5	VV
20	7.9	NV	63	3.3	VV
21	7.8	NV	64	3.3	VV
22	6.9	NV	65	2.9	VV
23	6.3	NV	66	2.8	VV
24	6.1	NV	67	2.7	VV
25	5.6	NV	68	2.4	VV
26	4.7	NV	69	2.3	VV
27	4.7	NV	70	2.0	VV
28	4.3	NV	71	1.8	VV
29	4.2	NV	72	1.7	VV
30	3.9	NV	73	1.7	VV
31	3.4	NV	74	1.6	VV
32	3.1	NV	75	1.4	VV
33	3.1	NV	76	1.2	VV
34	2.7	NV	77	1.1	VV
35	2.4	NV	78	1.0	VV
36	2.3	NV			
37	2.3	NV			
38	2.1	NV			
39	2.1	NV			
40	2.0	NV			
41	1.9	NV			
42	1.7	NV			
43	1.7	NV			



図 4: ランダムドット画像

表 4: 2 群データの比較 (無変換)

実験条件	件数 (N)	平均 (\bar{x})	分散 ($\hat{\sigma}^2$)	標準偏差
NV	43	8.56	65.73	8.09
VV	35	5.55	23.06	4.80

表 5: 2 群データの比較 (対数変換)

実験条件	件数 (N)	平均 ($\log \bar{x}$)	分散 ($\hat{\sigma}^2$)	標準偏差
NV	43	1.82	0.66	0.81
VV	35	1.39	0.67	0.82

t 値: 標本平均の差と分散 (標準偏差) の推定値との相対的な大きさの比

$$t = \frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{\sqrt{\frac{N_x + N_y}{N_x N_y (N_x + N_y - 2)} \left\{ \sum_{i=1}^{N_x} (x_i - \bar{x})^2 + \sum_{i=1}^{N_y} (y_i - \bar{y})^2 \right\}}} \quad (36)$$

ここで

$$ss_x = \sum_{i=1}^{N_x} (x_i - \bar{x})^2, \quad ss_y = \sum_{i=1}^{N_y} (y_i - \bar{y})^2$$

と表すことにすると、(36) は

$$t = \frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{\sqrt{\frac{N_x + N_y}{N_x N_y (N_x + N_y - 2)} (ss_x + ss_y)}}$$

となる。

この値を用いて、2 群の平均に顕著な差があるかどうかを調べるのが、 t 検定 (Student's t -test) である。この値を用いる方法は (通常の初等統計の教科書に書いてあるものである)、2 群の分散が等しいことを仮定している。 t 検定の発想は、標本平均の差の大きさの程度をその推定値の分散で標準化して評価しようとするものであり、それぞれの群の標本が $N(\mu_1, \sigma_x^2)$ および $N(\mu_2, \sigma_y^2)$ の

正規分布に独立に従うと仮定する。上の (36) の分子は $\bar{x} - \bar{y}$ でありこれは 2 群の平均の差の推定値である。第 1 群の平均の推定値 (標本平均) \bar{x} の分散は σ_x/N_x であり、 \bar{y} の分散は σ_y/N_y となる。平均の推定値の分散はそれぞれ

$$\hat{\sigma}_x^2 = \frac{SS_x}{N_x - 1}, \quad \hat{\sigma}_y^2 = \frac{SS_y}{N_y - 1}$$

で推定される。分子が「標本件数-1」になっているのは、推定値の平均が正しく母数と一致させるためである。さらに、2 群の分散が同一であることを仮定すると、その共通の分散は

$$\hat{\sigma}^2 = \frac{SS_x + SS_y}{(N_x + N_y - 2)}$$

によって推定される。

平均の差の推定値 $\bar{x} - \bar{y}$ の分散は $\sigma_x^2/N_x + \sigma_y^2/N_y$ である。分散が共通であることを仮定するとこれは $\frac{\sigma^2(N_x+N_y)}{N_x N_y}$ であり、(36) の分母はこの値の平方根を推定するものである。

もし σ^2 の推定が十分に正確であるなら、 t の値は平均ゼロ、分散 $\frac{\sigma^2(N_x+N_y)}{N_x N_y}$ の正規分布に従うと見なせるが、実際には $\hat{\sigma}^2$ の変動を無視できない。特に標本件数が少ない場合には、 $\hat{\sigma}^2$ の変動によって t の値はかなり変化する。検定に用いられる t 分布は、この変動を考慮したときに t の値がどのように分布するかを予測するものである。2 群の分散が等しければ (36) は「自由度 $N_x + N_y - 2$ の t 分布」という分布に従う。「自由度」とは t 分布の特徴を定める一つの数値であり、この値が大きいと正規分布に近づく。

2 群の分散が違う場合について類似の検定を行う方法が開発されており Welch の検定 (または Sattarthwaite の近似) と呼ばれている。

2005.6.22 修正 (2005 年の資料も間違い) この場合には、検定に用いる値は、

$$t = \frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{\sqrt{\hat{\sigma}_x^2/N_x + \hat{\sigma}_y^2/N_y}}$$

であり、これを t 分布で近似することができる。このとき、 t 分布の自由度は次の値を用いる。

$$df = \frac{(\hat{\sigma}_x^2/N_x + \hat{\sigma}_y^2/N_y)^2}{\frac{(\hat{\sigma}_x^2/N_x)^2}{N_x - 1} + \frac{(\hat{\sigma}_y^2/N_y)^2}{N_y - 1}} \quad (37)$$

2005.6.22 修正の終わり

7.2 平均と分散以外のものを用いる方法

平均、分散、標準偏差の他にデータの特徴について検討するための値
 データの位置 (全体の大小) について: 中央値 (median)、4 分位値 (点) (quartile)、分位点 (quantile)、刈り込み平均 (trimmed mean)
 データの広がり (散らばり具合) について: 範囲 (レンジ)、4 部位点間距離、刈り込み分散 (trimmed variance)

外れ値 (中央値から遠く離れた値) が含まれるデータについては、平均はかならずしも良い代表値ではないかも知れない。

7.3 図を用いた分布の比較

7.3.1 枝葉図

分布の特徴を簡単に把握するための関数として `stem` が用意されている。これは Tukey(1977) において枝葉図 (stem-and-leaf plot) と名付けられているグラフである。図 5 では、教示が "NV" であるも

```
> stem(fusionNV)

N = 43   Median = 6.9
Quartiles = 3.1, 10.3

Decimal point is at the colon

 1 : 779
 2 : 0113347
 3 : 1149
 4 : 2377
 5 : 6
 6 : 139
 7 : 89
 8 : 1499
 9 : 1577
10 : 3
11 :
12 : 23
13 : 04
14 : 7
15 :
16 :
17 : 4
18 :
19 : 7
20 : 4
21 :
22 : 0

High: 47.2
```

図 5: 枝葉図

のについて枝葉図を描いている。ここで、出力されているのは簡単なヒストグラムである。反応時間の整数部分が図の左側の見出しに描かれており、範囲に該当するデータの個数分の文字が、その行に並べて表示される。表示されるのは小数点 1 桁目の数値である。ほとんどの反応時間は 10 秒以下であるが、中には 47.2 秒かかるものもあり、正方向に裾が重いことが分かる。教示が "VV" であるものについても同様に、`stem(fusionVV)`、または直接 `stem(fusion$time[fusion$inst == "VV"])` とすれば枝葉図が表示される。図の上部に `Median` と表示されているのは中位数 (メディアン) であり、`Quartiles` とあるのは、第 1 四分位数と第 3 四分位数である。

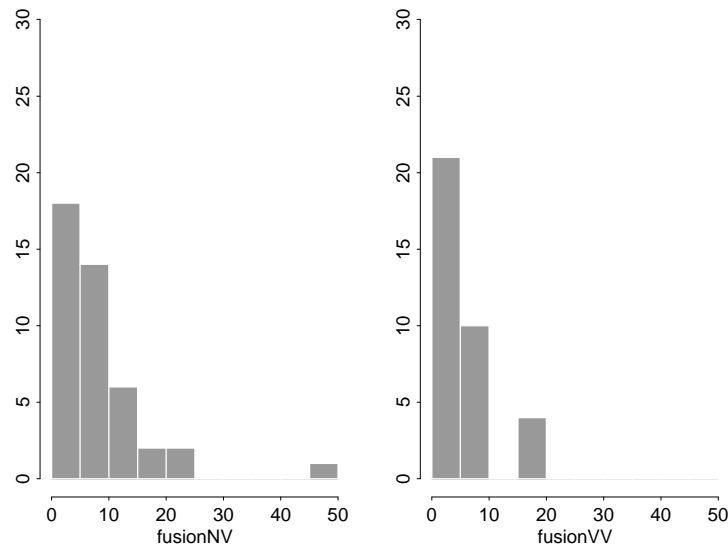


図 6: ヒストグラム

7.3.2 ヒストグラム

7.3.3 箱ヒゲ図

ヒストグラムより単純で、分布を比較するために効果的な方法として箱ヒゲ図 (boxplot) がある (Tukey,1977; Hoaglin et al.1983)。これらの図 7 の中心部分の箱は、データの 25 パーセント点 (第 1 四分位点) と 75 パーセント点 (第 3 四分位点) の範囲を示し、その中の横線はメディアンを表す。また、箱の両端から点線が伸びた先の線分は、ヒゲ (whisker) と呼ばれる。通常は四分位点から四分位点間距離の 1.5 倍の長さを取り、その範囲内で一番外側のデータの位置を示す。ヒゲより外側については、通常は全てのデータを表示する。これらの表示を検討することにより、データの大小、分布の広がり、歪み、最大・最小などを視覚的に確認できる。

標準正規分布では、第 3 四分位点の値は、0.6745 であり、これに四分位点間の距離の 1.5 倍を加えたヒゲの位置は 2.6980 となる。この上側確率は 0.0035 (0.35 パーセント) であり、かなり小さい。より裾の重い分布では、対応する確率はより大きくなる。例えば、自由度 4 の t 分布では、第 3 四分位点は 0.7407、ヒゲの位置は 2.9628 であり、上側確率は 0.0207 である。更に自由度 1 の t 分布 (Cauchy 分布) では、第 3 四分位点は 1、ヒゲの位置は 4、上側確率は 0.0780 であり、8 パーセントに近い。

7.3.4 分位点プロット

グラフによって分布の形を検討するための、もう一つの道具が分位点プロット (quantile plot) である。分位点プロットのなかでしばしば利用されるものには、分布の正規性を検討するための正規確率プロット (normal-probability plot) と、2 つの分布の関係をみるための分位点 - 分位点プロット (quantile-quantile plot / Q-Q plot) とがある。

図 8 の上 2 つは、fusion データ (NV 条件) の反応時間を正規確率プロットによって表示したものである。左は無変換のデータであり、右は対数変換後のデータを示している。正規確率プロット

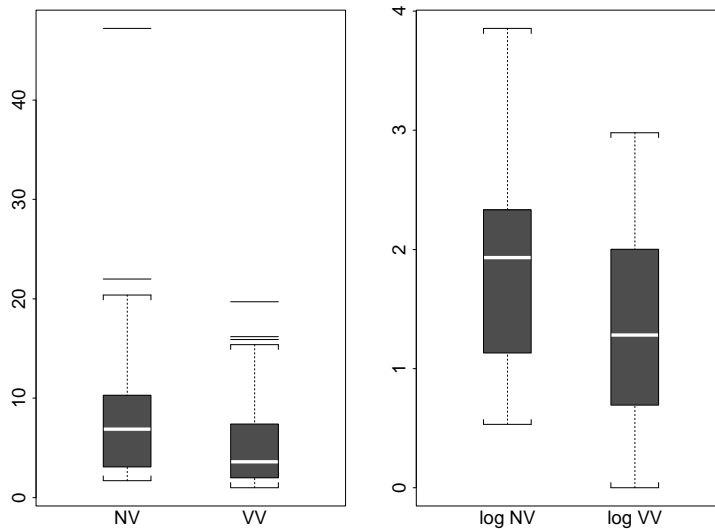


図 7: 両眼融合データの箱ヒゲ図 (左側：無変換、右側：対数)

は、データの各点について対応する下側確率を求め、その確率に対応する正規分布の分位点を横軸の値とし、データの値を縦軸にプロットする。もし、データの分布が正規分布に従っているならば、プロットされた点はほぼ直線上に並ぶはずである。左右の端で傾きが急になっているならば、これは正規分布に比べて裾の重い分布 (自由度の小さい t 分布がこれにあたる) であることを意味する。逆に傾きが緩やかなら、裾のつまった分布 (一様分布など) であることを示す。正規確率プロットに示した直線は第 1 四分位点と第 3 四分位点を結んでいる。変換前のデータはかなり正の方向に裾が重くなっているが、対数変換すると正規分布に近くなっている。ただし、小さい方の分布の裾が短い。

図 8 の下 2 つのグラフは、NV 条件のデータと VV 条件のデータを比較するための Q-Q プロットである。縦軸は先の正規確率プロットと同様にデータの分位点であるが、横軸も別のデータから求められた分位点の値である。もし、2 つのデータの分布が同一のものであるならば、表示される点はほぼ $X = Y$ の直線上にあるはずである。また、一方の分布が他方の分布を 1 次式で変換したものの (定数倍 + 定数) であれば、ほぼ直線上に点が表示されるはずである。左側は無変換のデータであり、右側は対数変換したものと士を比較している。表示してある直線は $X = Y$ を示す。対応する分位点同士を比較すると一貫して NV データが VV データより大きな値を取っていることが分かる。Q-Q プロットにおいては、2 群のデータの個数が異なる場合には、件数の少ない方に表示点を合わせ、もう 1 群の分位点は補間により推定した値を用いている。

これまでの検討からほぼ明らかだが、2 つのデータは無変換では分散がかなり異なる。しかし、対数をとることにより、正規性と等分散性の仮定は妥当なものとなる。Q-Q プロットから明らかのように、NV 条件の反応時間が明らかに大きい。無変換のデータに直接 t 検定を実行しても、有意とはならないが、対数変換すると帰無仮説が棄却される。

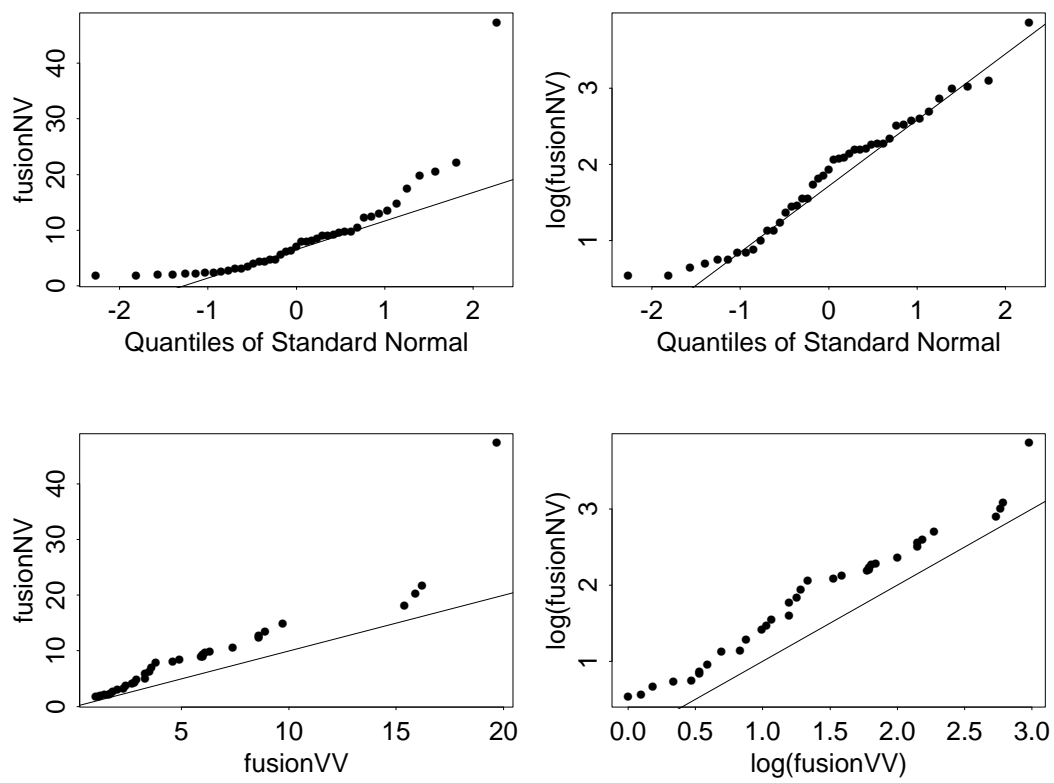


図 8: 正規確率プロットと分位点・分位点プロット

8 確率変数の性質

8.1 中心極限定理と正規分布

X_1, X_2, \dots, X_N を、それぞれ値 1 を確率 p で、値 0 を確率 $(1-p)$ でのる確率変数であるとする。また、それぞれが互いに独立に分布するものとする。つまり他の変数の値の大小によって値の生じ方に影響を受けることはない。

これらの和 $Y = \sum_{i=1}^N X_i$ は、平均 Np 、分散 $Np(1-p)$ (標準偏差 $\sqrt{Np(1-p)}$) となることが確率変数の性質から分かる。また、 N が大きくなると分布の形は次第に正規分布に近づいていく。つまり、 $Z = (Y - Np)/\sqrt{Np(1-p)}$ とすると、 Z は平均ゼロ、分散 1 と標準化されるが、 Z の分布は次第に平均ゼロ、分散 1 の正規分布 (標準正規分布) に近づく。

独立な確率変数の和の分布が、 N が増えるとともに正規分布に次第に近づくことは数学的に証明できるが、この定理は中心極限定理と呼ばれる。

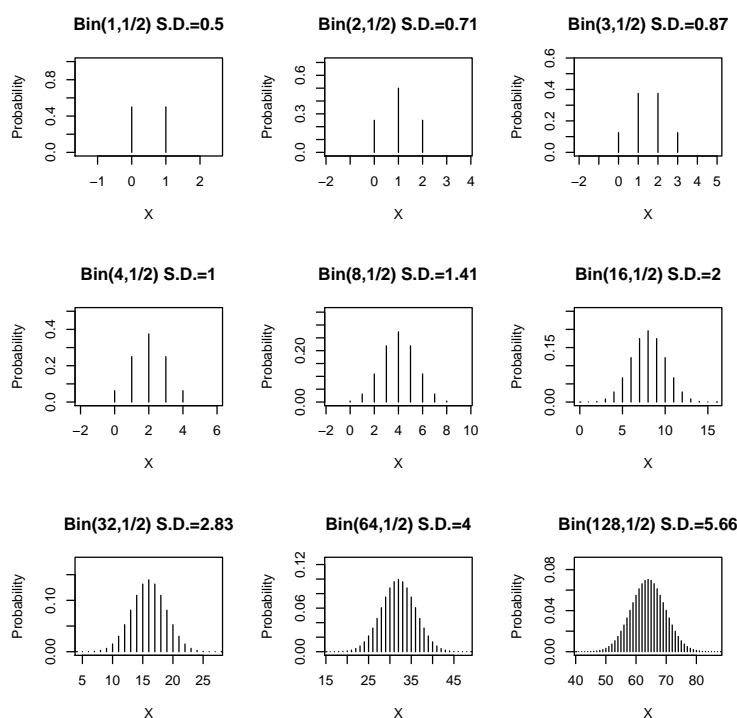


図 9: 2 項分布 $Bin(n, 1/2)$ (S.D. は標準偏差)

8.2 ヒストグラムと順序統計量

連続値をとる確率変数のヒストグラムは、確率密度関数の推定法の一つと考えられる。確率変数 X の確率密度関数を $f(x)$ とする。観察されたデータを x_1, x_2, \dots, x_N とすると、ヒストグラムは $\hat{f}(x)$ は次のように定義される。

$$\hat{f}(x) = \frac{n_i}{Nh}, \quad x \in (b_j, b_{j+1}]$$

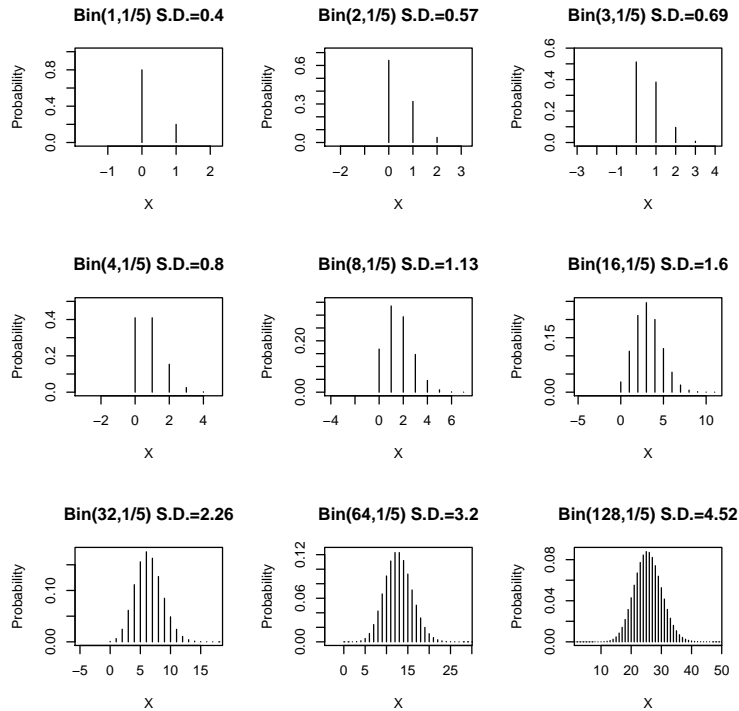


図 10: 2 項分布 $Bin(n, 1/5)$ (S.D. は標準偏差) 2006.06.23 修正

ここで、 $\{b_j\}$ はヒストグラムの区間の境界であり、 h は区間の長さである。また、 n_i は区間 $(b_j, b_{j+1}]$ 内にあるデータの件数を表す。

ヒストグラムは単純なものであり、特に注意深く検討するべき性質があるようには思わないかも知れないが、区間の設定は微妙な問題である。得られたヒストグラムがどれくらい「良い」ものであるかの一つの基準は、次の指標（積分 2 乗誤差）である。

$$ISE = \int_{-\infty}^{+\infty} [\hat{f}(x) - f(x)]^2 dx$$

問題は、区間を広げると真の確率密度関数 $f(x)$ を正確に表現することが難しくなり、区間を小さくすると 1 つの区間に含まれるデータの件数が減少し推定値が信頼できなくなることである。データの件数 N が大きく、 $f(x)$ が複雑な形をしているなら、区間幅を小さくした方が望ましく、その逆なら区間幅を大きくした方が良い。ISE は標本の値によって定まるものなので、得られた標本が異なれば違う値をとる。ヒストグラムの性質を分析すると、ISE の平均 $E_f(ISE)$ を最小にする区間幅は、 N の値が大きければほぼ、

$$h_0 = \left(\frac{6}{R}\right)^{1/3} N^{-1/3}$$

になる。この式中の R は

$$R = \int_{-\infty}^{+\infty} \{f'(x)\}^2 dx$$

を意味する。特に、正規分布の場合には、 $h_0 = 3.491\sigma N^{-1/3}$ となる。(Simonoff, J., (1998) *Smoothing Methods*, Springer)

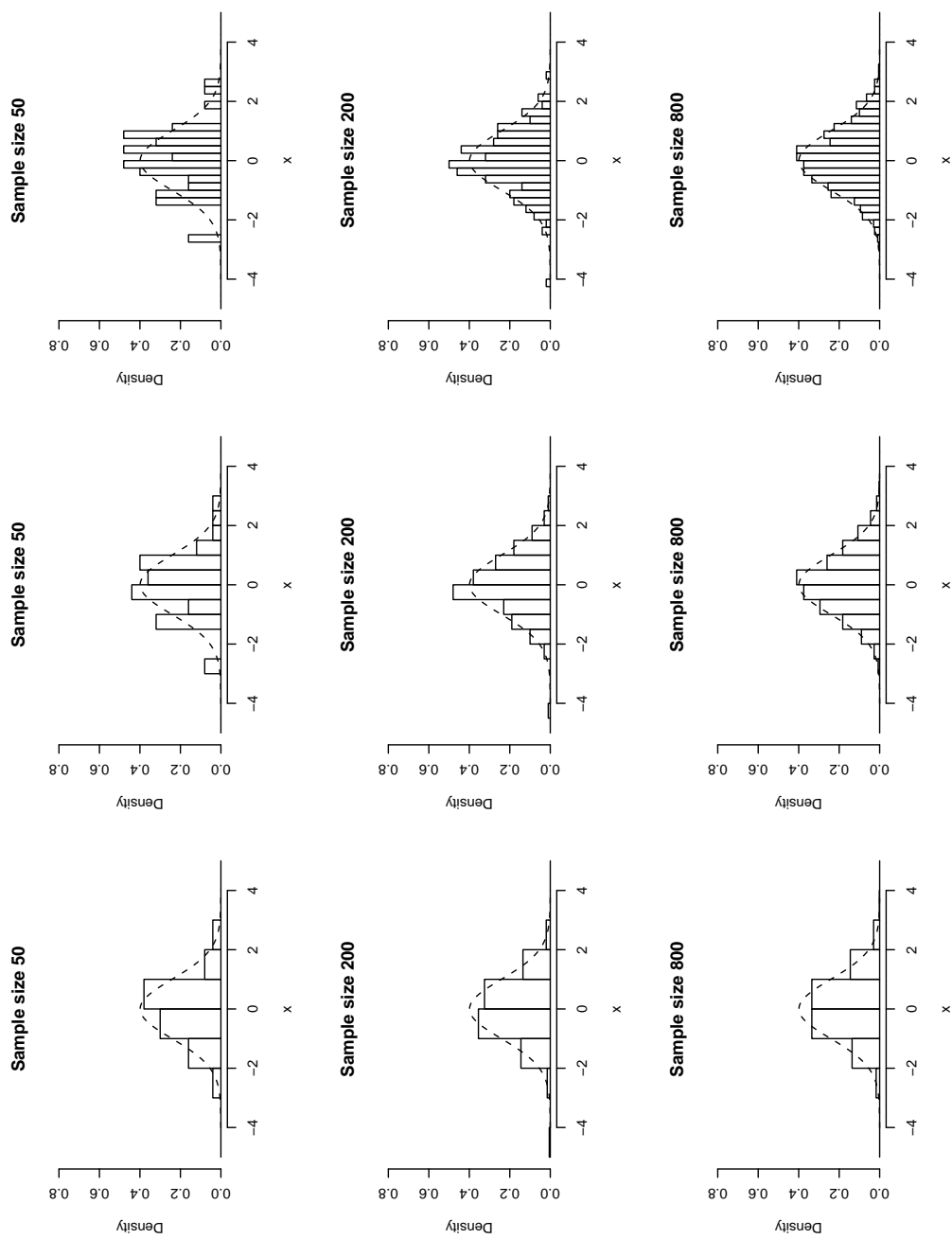


図 11: 標本件数とヒストグラムの凹凸

表 6: 標準正規分布の漸近的最適区間幅

N	h_0
50	0.95
200	0.60
800	0.38

8.3 順序統計量の分布

平均 μ 、分散 σ^2 の正規分布に従う、互いに独立な N 個の確率変数 X_1, \dots, X_N を考える。これらの平均 $\bar{X} = \frac{1}{N} \sum_{i=1}^N X_i$ の平均は μ 、分散は σ^2/N となる。一方、 N が大きいときメディアン（中位数）は平均 μ 、分散 $\frac{\pi\sigma^2}{2N}$ であり、分散は $\pi/2 = 1.571$ 倍大きい。しかし、正規分布より裾の重い分布（外れ値の生じやすい分布）の場合には、メディアンの分散の方が小さい場合もある。

t 検定は、2 群の比較を行うために、それぞれの平均と分散（共通、または個別に推定される）を用いるが、観測された数値の順序を用いる方法もある。2 つの群の観測値をそれぞれ $x_1, \dots, x_M, y_1, \dots, y_N$ とする。これら 2 群をまとめ、小さい方から順番にならべ、それらに順位をつける。ここで、 x_1, \dots, x_M に対応する数値につけられた順位の和を W とする。全体の順位の平均は $(M+N+1)/2$ であるから、2 群が同じ分布に基づくならば W が $M(M+N+1)/2$ より大きく異なることは少ないはずである。この W （順位和）を用いて 2 群の違いを識別する方法を Wilcoxon 順位和検定、または Mann-Whitney 検定と呼ぶ。

この方法の特徴は、(1) 例え外れ値があっても値は順位に変換されるため、結果が影響を受けづらいうこと、(2) 観測値の大小関係を変えない変換については、結果が同じであることである。

参考：確率変数 X の分布関数を $F(x)$ とする。これは $X \leq x$ となる確率が $F(x)$ であることを意味する。同一の分布に従う N 個の確率変数 X_1, \dots, X_N を考え、これらの観測値をそれぞれ x_1, \dots, x_N とする。確率変数 X_1, \dots, X_N の内、 k 番目に小さい値を $X_{(k)}$ とすると、最小値 $X_{(1)}$ の分布関数は、

$$1 - [1 - F(x)]^N$$

最大値 $X_{(N)}$ の分布関数は、

$$[F(x)]^N$$

となる。一般的に、 $X_{(k)}$ の分布関数は

$$\sum_{r=k}^N \binom{N}{r} [F(x)]^r [1 - F(x)]^{N-r} = \frac{N!}{(k-1)!(N-k)!} \int_0^{F(x)} u^{k-1} (1-u)^{N-k} du$$

であり、確率密度関数は

$$\frac{N!}{(k-1)!(N-k)!} [F(x)]^{k-1} [1 - F(x)]^{N-k} f(x) dx$$

となる。

分位点 ξ_p （下から確率 p に相当する点）の厳密な定義は、次のようなものである (Rao,1973)。

$X \leq \xi_p$ となる確率が p 以上であり、 $X \geq \xi_p$ となる確率が $1-p$ 以上の点。

Np を越えない最大の整数 (Np と同じであっても良い) を k_1 とし、 $N(1-p)$ を越えない最大の整数を k_2 とすると、標本から分位点の推定値 $\hat{\xi}_p$ を次のように定める (Rao,1973)。

$\hat{\xi}_p$ 以下の標本の件数が k_1 より大きく、 $\hat{\xi}_p$ 以上の標本の件数が k_2 より大きい値とする。

このように定義すると Np が整数でないときには、 $\hat{\xi}_p = x_{(k_1+1)}$ となり、 Np が整数のときは $[x_{(k_1)}, x_{(k_1+1)}]$ の区間の任意の値となる。

母集団の p 分位点を ξ_p とする。ここで ξ_p が一意に定まり（複数の値ではなく、一つの値だけが ξ_p となること）、確率密度関数が $f(\xi_p) > 0$ であれば、 $\sqrt{N}(\hat{\xi}_p - \xi_p)$ は N が増加するとともに、平均ゼロ、分散 $p(1-p)/f(\xi_p)^2$ の正規分布に近づく。

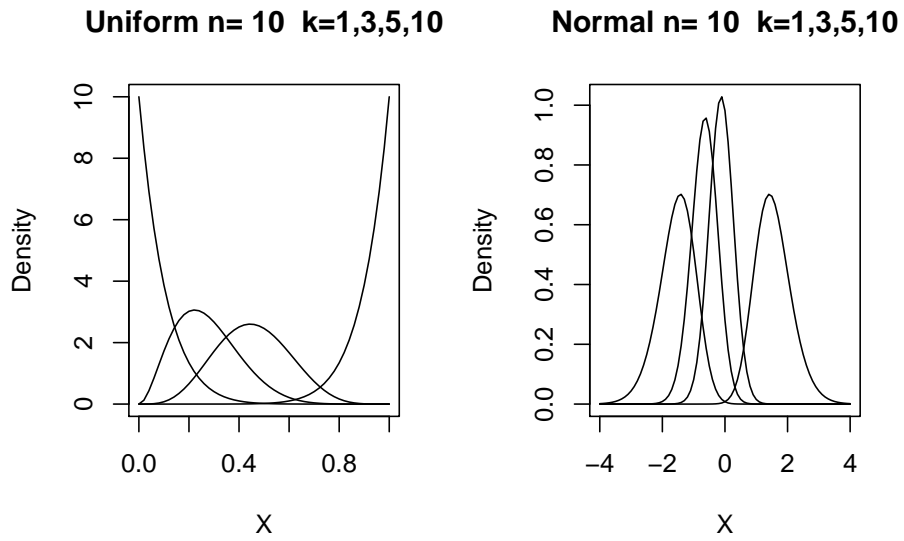


図 12: 順序統計量の確率密度関数

(証明はここでは省略。数理統計学の教科書を参照のこと。竹村彰通「現代数理統計学」創文社、C.R.Rao(原著,1973)「線形モデルとその応用」東京図書 など)。

9 1群および2群データの検定 (続き)

先のセクションで2群の平均の差についての t 検定の方法について説明した。復習すると、2群の平均の差 $\mu_x - \mu_y$ を $\bar{x} - \bar{y}$ によって推定し、2群の母集団に共通の分散を

$$\hat{\sigma}^2 = \frac{(ss_x + ss_y)}{N_x + N_y - 2}$$

によって推定する。さらに、これを用いて $\text{Var}(\bar{x} - \bar{y})$ を

$$\hat{\sigma}_{diff}^2 = \left(\frac{1}{N_x} + \frac{1}{N_y} \right) \hat{\sigma}^2$$

によって推定し、

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\hat{\sigma}_{diff}^2}}$$

を求め、これが帰無仮説 $H_0: \mu_x - \mu_y = 0$ の下で、自由度 $N_x + N_y - 2$ の t 分布に従うことを用いて検定を行った。

同様のアイデアは、2群の平均の差以外の場合にも適用することができる。もっとも単純なのは、1群の平均 μ_x がある特定の値 μ_0 に等しいか否かを検定する場合である。ここで帰無仮説を $\mu_x - \mu_0 = 0$ とすると \bar{x} の分散の推定値は

$$\hat{\sigma}_{mean}^2 = \frac{1}{N_x} \times \frac{ss_x}{N_x - 1}$$

であり、これを用いて

$$t = \frac{\bar{x} - \mu_0}{\sqrt{\hat{\sigma}_{mean}^2}}$$

を求める。母集団が正規分布であると仮定すれば、帰無仮説のもとで t が自由度 $N_x - 1$ の t 分布に従うので、この性質を用いて検定を行えばよい。

この公式を用いて2群の個々のデータに対応関係が存在する場合の検定を行うことができる。2群に対応があるとは、第1群のデータが右目の視力であり、第2群のデータが同一の被検者の左目の視力であるような場合を示す。

ここで、 i 番目の被験者の右目の視力 x_i と左目の視力 y_i に違いがあるかについて検討を行うものとする。右目の視力のデータを第1群、左目の視力のデータを第2群と考えると、対応のない場合の t 検定を用いることも可能ではあるが、対応を考慮して次のような分析を行う方が、差をより敏感に検出することが可能となる。

データに対応のある場合には、それぞれの差

$$d_i = x_i - y_i \quad (i = 1, \dots, N)$$

を求め、この平均がゼロであるか否かを検定すればよい。差の平均 \bar{d} の分散の推定値は

$$\hat{\sigma}_{mean}^2 = \frac{1}{N} \frac{ss_d}{N - 1}$$

であり、これを用いて

$$t = \frac{\bar{d} - 0}{\sqrt{\hat{\sigma}_{mean}^2}}$$

を計算し、これが帰無仮説のもとで自由度 $N - 1$ の t 分布に従うことを利用して検定を行う。

この方法の利点は、データの変動のうち2群に共通する部分が取り除かれるため、分析の対象となる部分の分散が小さくなり、平均の差を検出しやすくなることである。一方、不利な点は分散の推定を行う場合、自由度が小さくなるため、特にデータの件数が少ない場合には、分散の推定が難しくなる可能性のあることである。しかし、一般的にはこの欠点より、利点がまさる場合が多い。

9.1 順位に基く方法

上述の t 検定と同様の目的、つまり2群の位置母数（正規分布の場合は平均だが、平均の存在しない分布についても適用が可能であるので、より一般的に位置母数という用語を用いる）の違い、または対応のある2群の位置母数の違いを検定するために、データの順位に基く統計量を用いる方法がある。2群の位置母数の比較に用いられるのが、前述した Wilcoxon 順位和検定（Wilcoxon rank sum test, または Mann-Whitney 検定）であり、対応のある場合に用いられるのが Wilcoxon 符号付順位和検定（Wilcoxon signed rank sum test）と呼ばれる方法である。

9.1.1 Wilcoxon 順位和検定

重複するが、Wilcoxon 順位和検定の方法を示す。

1. 両群のデータを合わせ、大きさの順に並べ替え、それぞれのデータに順位を与える。1群の個数を N_x 、2群の個数を N_y とし、合計した件数を N とする。順位の総和は $R_{sum} = \sum_{k=1}^N k = (N+1)N/2$ となる。
2. もし、2群の分布が同一ならば、つまり帰無仮説が成立しているなら、第1群の順位の和 W_s の期待値は

$$E(W_s) = \frac{R_{sum} \times N_x}{N} = N_x(N+1)/2$$

となる。また、

$$\text{Var}(W_s) = \frac{N_x N_y (N+1)}{12}$$

となることも知られている。ただし、同順位のデータがあり、それらには同一である範囲の順位の平均値を与えることにすると（つまり、データが $(3, 5, 5, 9)$ であるとき順位は $(1, 2.5, 2.5, 4)$ となる）同順位に対応した順位和統計量 W_s^* の平均は $E(W_s)$ と同じであり、分散は

$$\text{Var}(W_s^*) = \frac{N_x N_y (N+1)}{12} - \frac{N_x N_y \sum_{k=1}^K (g_k^3 - g_k)}{12N(N-1)}$$

となる。ここで g_k は k 番目の同順位のグループに含まれるデータの個数である。データが $(3, 5, 5, 9)$ ならば $K=3$ であり、 $g_1=1, g_2=2, g_3=1$ となる。

3. 各群のデータの件数が大きければ、 W_s はほぼ正規分布に従うとみなせるので、上の性質を用いて W_s の期待値からの違いを元に検定を行うことができる。

9.1.2 Wilcoxon 符号付順位和検定

一方、対応のある場合の t 検定に相当するものが、Wilcoxon 符号付順位和検定である。これは次の手順をとる。

1. 2群のデータの対応する値の差 $d_i, (i = 1, \dots, N)$ を求める。
2. d_i の絶対値を求め、その絶対値の大きさの順に d_i を並び換え、順位を与える。
3. d_i の符号に従い、負であるなら順位をマイナスの数とする。正の場合にはそのままとする。
4. これらの符号付順位のうち、正の符号であるものの総和を V_s とし、負の符号であるものの（正になおした）順位の総和を V_r とする。ここで、 $N_x + N_y = N$ とすると、 $V_s + V_r = \frac{N(N+1)}{2}$ である。
5. 帰無仮説のもとで

$$E(V_s) = \frac{N(N+1)}{4}$$

および

$$\text{Var}(V_s) = \frac{N(N+1)(2N+1)}{24}$$

である。それぞれのデータの件数が大きければ V_s はほぼ正規分布に従うとみなせるので、この性質を用いて検定を行える。

また、 d_i の絶対値に同順位がある場合には、次のようになる。まず、両群での対応する値が同じで $d_i = 0$ となる組の数を g_0 とする。これ以外の $|d_i|$ について、同順位である組の個数を順に g_1, \dots, g_K とする。差がない組については値をゼロとし、それ以外の組に付いては、該当する範囲の順位の平均値を与えることにする（差がゼロである対も含めての順序を用いて計算する）。このとき、同順位に対応した符号付き順位和統計量 V_s^* の平均と分散とは、各々

$$E(V_s^*) = \frac{N(N+1) - g_0(g_0+1)}{4}$$

および

$$\text{Var}(V_s^*) = \frac{N(N+1)(2N+1) - g_0(g_0+1)(2g_0+1)}{24} - \frac{1}{48} \sum_{k=1}^K g_k(g_k-1)(g_k+1)$$

となる。

また、上の2つのいずれの方法についても、データの件数が小さい場合には、組み合わせ計算によって直接確率を求める。

10 無作為化試験と影響関係の推定

ポイント: 無作為化試験の重要性和観察研究の限界

10.1 調査の型

1. 探索的・記述的なもの

標本が無作為抽出される場合、標本平均の分散は標本数に反比例する（標本平均の標準偏差は標本数の平方根に反比例）。大規模な標本調査では、多段層別抽出が用いられることが多い。

政府機関によって行われる人口、経済調査（全数調査としては国勢調査、事業所統計など、標本調査としては家計調査など）

新聞社、テレビネットワーク、研究者が行う世論調査、投票行動調査

○ 文部省 統計数理研究所による「日本人の国民性調査」
生活観などを（一部）同一の質問文を用いて 1953 年より 5 年おきに全国調査している。被験者数は毎回数千人規模。

○ 生命保険文化センター「日本人の価値観調査」
データが公開されている（東大社情研 SSJ データアーカイブ）

○ 日本社会学会による「社会階層と社会移動」全国調査
1955 より 10 年おきに実施。1 万人近い被験者を対象とする国内の社会科学系の面接調査としては最大のもの。

2. 理論検証 (仮説検証) を目的とするもの

医学・薬学分野に典型例がある。

○ 疫学 (epidemiology) 調査 (意図的な介入を伴わないもの) → 状況証拠から病気の原因を追求する医学の分野。(食生活と病気、タバコの影響、環境中の毒物の影響など)

○ 新薬の臨床試験 (field trial) (介入を伴うもの) → 実験室で開発された薬物が本当に効果があるか否かの検討

10.2 Salk ワクチンの臨床試験

古典的事例：1954 年に米国で行なわれたソーク (J.Salk) ポリオワクチンの臨床試験が歴史的に有名。有効性と安全性が評価された。1950 年代の米国でのポリオの平均罹患率は 10 万人中 50 人程度のため、大量の被験者が必要とされた (Meier,1972)²。

1. 人口統計学的アプローチ：季節変動、地域変動、診断基準のあいまいさ、医師の知識の違いなどにより評価を行なうことが困難。

2. 観察対照法 (observed control approach): 一部採用

小学校の 2 年生にワクチンを投与。1 年生と 3 年生が対照群 (control group) となる。次のような問題点がある。

医師の判断基準のあいまいさ。

実験参加を志願したものの偏り (家庭環境、当人の健康状態などが全体の平均とは異なる可能性がある)。

3. 無作為化と偽薬 (プラシーボ) を用いる対照法 (無作為化試験, randomized trial)

2 重盲検法 (double-blind trial) または 2 重マスク試験 (double-masked trial) と呼ばれる。倫理性が問題にされることがあるが (どのような条件の下で無作為化試験が許されるか、また実験の途中で効果に差が認められる場合に実験を続行すべきか等)、現状では本当の効果を知らずの最良の方法。

²Meier,P (1972). The biggest public health experiment ever: The 1954 field trial of the Salk poliomyelitis vaccine. In Tanur,J.M. et al. (1989). *Statistics: A Guide to the Unknown, 3rd ed.*, Brooks/Cole Pub. Co.) 邦訳 (第 2 版) 安藤洋美監訳 (1984). 統計学講話, 現代数学社

10.3 未知の背景要因の影響

また外的な介入の有無は、影響関係の推論の確さと深い関係を持つ。一般的には、外的介入なしでの推論から決定的な結論を得るのは難しい。以下は、介入を伴わない(大量の)調査データから得られた結論が、他の証拠なしには、明快な解釈が困難であることを示している。

Smith et al. (1992)³ は、入学試験とは異なる分野の例ではあるが、調査データの結果の解釈に、慎重な注意が必要であることを指摘した論文である。ここで著者らは、大規模な危険要因についての調査 (MRFIT) を分析することによって、まず喫煙と自殺の間に関係が見られることを指摘している。MRFIT(Multiple Risk Factor Intervention Trial) は米国の 35 歳から 57 歳までの 361,662 人の男性を 12 年間にわたって追跡調査した研究である。年齢、人種、1 日の喫煙本数、糖尿病治療中であるか、心筋梗塞の有無、社会経済地位(居住地域から推定)がベースラインデータとしてとられており、これらの効果を考慮してもなお喫煙と自殺との間に関係は見られる。また、MRFIT 以外の他の大規模な複数の追跡調査でも同様の結果が得られている。

しかし、Smith et al.(1992) は、MRFIT について喫煙と他殺による死の間にも同様の関係が見られることを報告している。喫煙がその薬理作用によって自殺を増加させるという説明は、一見もっともらしく思えるが、他殺を増加させる原因となるとは考えづらい。喫煙は本当に自殺に影響を及ぼしているのだろうか？

これについての決定的な答えは現状では分かっていない。しかし、タバコが直接自殺の危険を高めるとこれらの調査結果から解釈するのは、困難であるように思える。研究者によっては捉えられていない何らかの心理的または社会的要因があり、これが喫煙の習慣と自殺傾向の両者に影響を及ぼしているとみなすことも可能と思われる。

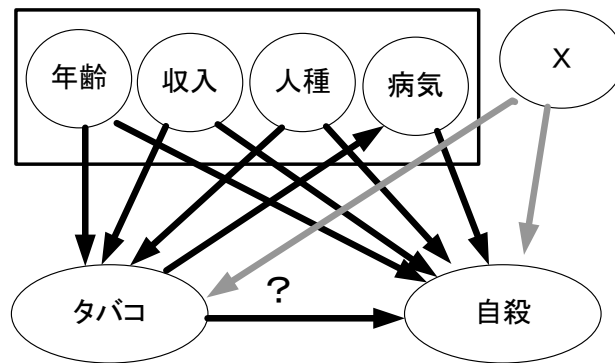


図 13: 喫煙の影響構造

参考文献

- [1] Doll,R., Peto R., Wheatley,K. Gray,R., & Sutherland,I. (1994). Mortality in relation to smoking: 40 years' observations on male British doctors. *British Medical Journal*, **309**, 901-11.

³Smith,G.D., Phillips,A.N. & Neaton,J.D. (1992). Smoking as "independent" risk factor for suicide: illustration of an artifact from observational epidemiology? *Lancet*, **340**, 709-12.

表 7: 喫煙行動と自殺率 (Smith et al.,1992)

タバコ/日	人数	自殺件数	自殺件数 1万人×年
0	228,545	291	1.09
1-19	29,333	50	1.47
20-39	72,200	166	2.00
40-59	27,844	78	2.46
60+	3,740	16	3.78

表 8: リスク要因と自殺の相対リスク (95% 信頼区間) (Smith et al. 1992)

リスク要因	年齢補正済み 相対リスク	完全補正済み 相対リスク
喫煙		
0	1.00	1.00
1-19	1.36(1.00-1.83)	1.36(1.00-1.84)
20-39	1.88 (1.55-2.27)	1.86(1.54-2.26)
40-59	2.31 (1.79-2.96)	2.27(1.77-2.92)
60+	3.44 (2.08-5.69)	3.33(2.01-5.52)
(χ^2 for trend)	(78.66, $p < 0.0001$)	(75.98, $p < 0.0001$)
収入		
低	1.35(1.13-1.60)	1.33(1.13-1.59)
他	1.00	1.00
人種		
黒人	0.85(0.60-1.21)	0.76(0.53-1.08)
他	1.00	1.00
心筋梗塞		
病歴あり	1.79(1.07-2.99)	1.73(1.03-2.90)
なし	1.00	1.00
糖尿病		
有病	1.66(0.99-2.77)	1.61(0.96-2.70)
なし	1.00	1.00

表 9: タバコと他殺による死 (Smith et al. 1992 に基づく)
総数 222, 人種と収入 (居住地域から推定) について補正された
10万人あたり死亡率の比

リスク要因 (一日あたり喫煙本数)	他殺による死亡の相対比率 (95% 信頼区間)
0	1.00
1-39	1.71 (1.29-2.28)
40-	2.04 (1.32-3.15)

- [2] Hemenway,D., Solnick,S.J. & Colditz,G.A. (1993). Smoking and suicide among nurses. *American Journal of Public Health*, 83, 249-251.
- [3] Meier,P (1972). The biggest public health experiment ever: The 1954 field trial of the Salk poliomyelitis vaccine. In Tanur,J.M. et al. (1989). *Statistics: A Guide to the Unknown*, 3rd ed., Brooks/Cole Pub. Co.) 邦訳 (第2版) 安藤洋美監訳 (1984). 統計学講話, 現代数学社
- [4] Moertel,C., Fleming,T., Creagan,E., Rubin,J., O'Connell,M. & Ames,M. (1985). High-dose vitamin C versus placebo in the treatment of patients with advanced cancer who have had no prior chemotherapy: A randomized double-blind comparison. *New England Journal of Medicine*, **312**, 137-141.
- [5] Smith,G.D., Phillips,A.N. & Neaton,J.D. (1992). Smoking as “independent” risk factor for suicide: illustration of an artifact from observational epidemiology?. *Lancet*, **340**, 709-12.
- [6] 重松逸造・柳川洋 監修 (1991). 新しい疫学, (財) 日本公衆衛生協会.
- [7] 丹後俊郎 (1998) 統計学のセンス, 朝倉書店.
- [8] Tverdal,A., Thelle,D., Stensvold,I. & Leren,P. (1993). Mortality in relation to smoking history: 12 years' follow-up of 68,000 Norwegian men and women 35-49. *Journal of Clinical Epidemiology*, **46**, 475-487.
- [9] Vartiainen,E.,Puska,P.Pekkanen,J.Tuomilehto,J., & Lönnqvist,J. (1994). Serum cholesterol concentration and mortality from accidents, suicide, and other violent causes. *British Medical Journal*,**309**, 445-447.

11 2値データの関連性とユール・シンプソンのパラドックス

2値データとは、(Yes,No), 合・否など2つの値をもつデータのことである。ここでは、表10のような2つの2値変数の関連性について検討する。

表 10: 死後の世界を信じるか? (性別集計)

性別	死後の世界についての考え	
	Yes	No または Undecided
女性	435	147
男性	375	134

Source: Data from 1991 General Social Survey (Agresti,1996 より引用)

上のデータは、死後の世界が存在することを信じるか、否かの回答を性別に集計したものである。ここで、性によって回答の頻度が異なるかどうかを検討する。もし、性別によって回答比率に違いがないとしても、サンプリングのばらつきによって、回答比率に多少の違いが生じる。では、どの程度の違いならば、性別が回答比率に違いをもたらしているといえるだろうか。

11.1 統計的独立

では、まず性別が回答比率に関係がない場合を考えよう。男性における Yes の (真の) 比率を π_M とし、女性の比率を π_F とする。性別で違いがないという仮説は

$$H_0 : \pi_F = \pi_M \quad (38)$$

と表される。ここで男性の総数を n_m 、また女性の総数を n_F とすると、4つの回答の期待値は

性別	Yes	No	計
男性	$\mu_{11} = n_M \pi_M$	$\mu_{12} = n_M (1 - \pi_M)$	n_M
女性	$\mu_{21} = n_F \pi_F$	$\mu_{22} = n_F (1 - \pi_F)$	n_F

となる。回答比率が性別によらないとき (ただし、 $\pi_M \neq 0, 1; \pi_F \neq 0, 1$)、次の関係が成立している。

$$\frac{\mu_{11}\mu_{22}}{\mu_{12}\mu_{21}} = 1 \quad (39)$$

上の関係が成立している場合には、性別が違っても回答には違いがない。また、逆に回答によって被験者を分類しても、Yes と答えたグループと No と答えたグループの間に性別比率の違いがないことになる。

ここで、 $\sum_{i=1}^2 \sum_{j=1}^2 \mu_{ij} = N$ とすると、上の条件が成立することは、

$$\frac{\mu_{ij}}{N} = \frac{\mu_{i1} + \mu_{i2}}{N} \times \frac{\mu_{1j} + \mu_{2j}}{N}, \quad (i, j = 1, 2) \quad (40)$$

がなりたつことに等しい。

このような関係が、2つの変数の間に成立することを統計的に独立 (statistically independent) または単純に独立と呼ぶ。より厳密に定義するとつぎのようになる。

確率変数 X と Y の同時分布が $p(x, y)$ であるとする。 Y を値 y に固定したときの X の条件つき分布 $p(x|y)$ が y の値によらず一定であるとき、同時分布はつぎのような形になる。 ($p(x|y) = p(x, y)/p(y)$ であることに注意。)

$$p(x, y) = p(x)p(y) \quad (41)$$

これより、上の条件が成立すれば、 X を x に固定したときの条件つき分布 $p(y|x)$ が x の値によらず一定であることも成立する。以上の性質が成立することを、 X と Y が統計的に独立であるという。 X と Y が独立であれば相関 (correlation) はゼロになるが、相関がゼロであっても独立とは限らない。また、 X と Y とがともに 2 値データの場合には、 (39) が 1 であれば、行と列の変数は独立である。

もし、 X と Y とが統計的に独立なら、 Y を予測するためには X は全く役にたたず、また X を予測するためには Y は役にたたない。

11.2 オッズ比 (odds ratio)

2 値をとるデータにおいて反応率の比 (上述のデータの場合には $\pi_M/(1-\pi_M)$ または $\pi_F/(1-\pi_F)$) のことをオッズ (Odds) という。オッズとは賭け率のことであり、たとえば Yes, No がワールドカップ日本・ジャマイカ戦の日本の勝ちを意味するなら、このときオッズは

$$\frac{\text{「日本の勝ち」の確率}}{\text{「ジャマイカの勝ち または 引き分け」の確率}}$$

となる。

さらに、 2×2 のデータにおいて、 2 つの条件についてのオッズの比、つまり

$$\theta = \frac{Pr(X=0, Y=0)/Pr(X=0, Y=1)}{Pr(X=1, Y=0)/Pr(X=1, Y=1)} = \frac{Pr(X=0, Y=0)Pr(X=1, Y=1)}{Pr(X=0, Y=1)Pr(X=1, Y=0)} \quad (42)$$

をオッズ比 (odds ratio) という。上に示した様に、オッズ比が 1 であることは、 X と Y とが独立であることを示す。

11.3 相対リスク

上のデータにおける π_F/π_M に相当するものを、相対リスク (relative risk) と呼ぶ。これは、医学関係からの用語である。つぎのような 2×2 データを考えよう。

実験条件	心筋梗塞	
	Yes	No
偽薬	189	10845
アスピリン	104	10933

Source :Preliminary Report: Findings from the Aspirin Component of the Ongoing Physicians' Health Study, *New England Journal of Medicine* , 318, 262-264 (1988), Agresti (1996) より

このとき、相対リスクは「偽薬 (統制群) における心筋梗塞の確率」と「アスピリン (実験群) における心筋梗塞の確率」の比のことであり、アスピリンを処方しないと、どれだけ心筋梗塞の確率が高いかを示す値である。

11.4 仮説からのずれの指標

いま、一般的に 2×2 のデータについて仮説がつぎのように表されるものとする。

$$H_0 : E(n_{ij}) = \mu_{ij} \quad (i, j = 1, 2) \quad (43)$$

ここで、左辺は該当するセルの頻度の期待値であり、期待頻度 (expected frequency) と呼ばれる。また、観測されたデータは n_{ij} , $i, j = 1, 2$ であるとする。

このとき、仮説 H_0 からのずれを検討するための指標として、つぎの2つが代表的である。

1. χ^2 統計量

$$X^2 = \sum \frac{(n_{ij} - \mu_{ij})^2}{\mu_{ij}} \quad (44)$$

2. 対数尤度比 (の2倍)

$$G^2 = 2 \sum n_{ij} \log \left(\frac{n_{ij}}{\mu_{ij}} \right) \quad (45)$$

いずれの値の分布も仮説が真ならば、各セルのデータの件数が増えるとともに、 χ^2 分布に近づく。全てのセルに含まれるデータの総件数 N が一定であるとする、ある固定された仮説 (μ_{ij}) に対して、これらの χ^2 分布の自由度は $(IJ - 1)$ となる。この性質を用いて、特定の仮説が成立しているか否かを検定できる。

ここで、自由度 k の χ^2 分布の平均は k 、分散は $2k$ である。

11.5 独立性の検定

X と Y (行と列) とが独立であることを検定するためには、上に類似の手続きを利用するが、少し変更を要する。統計的に独立であるということは、各セルの値が特定の固定されたものであることを必要とするわけではない。課せられている制約はもう少し緩やかなものである。

ここで標本の周辺度数を $\sum_i n_{ij} = n_{+j}$, $\sum_j n_{ij} = n_{i+}$, $\sum_i \sum_j n_{ij} = n_{++}$ とする。また、周辺度数の期待値を $\sum_i \mu_{ij} = n_{+j}$, $\sum_j \mu_{ij} = n_{i+}$, $\sum_i \sum_j \mu_{ij} = \mu_{++}$ と表記する。

独立性の仮定は、

$$\mu_{ij} = \frac{\mu_{i+} \mu_{+j}}{\mu_{++}} \quad (i, j = 1, 2) \quad (46)$$

が成り立つことに等しい。そこで、検定を行なうには、標本においてこの関係からのずれの大きさを測ればよい。

そこで、

$$\hat{\mu}_{ij} = \frac{n_{i+} n_{+j}}{n_{++}} \quad (i, j = 1, 2) \quad (47)$$

とおき、(44),(45) に代入すればよい。ただし、独立性の検定においては、 $\hat{\mu}_{ij}$ の値は固定されたものではなく、標本の値に応じて変更されるため、 χ^2 分布の自由度は1である。

より一般的に、 $I \times J$ の2重分類表の独立性の検定を行なうには、やはり、(47) によって $\hat{\mu}_{ij}$ を求める。このとき X^2, G^2 は漸近的に自由度 $(I - 1)(J - 1)$ の χ^2 分布に従う。

X^2 と G^2 を比較すると、 X^2 の方が χ^2 分布への収束が速い。 $N/IJ < 5$ のような条件では、 G^2 を χ^2 分布で近似することは難しい。

11.6 Fisher の正確検定 (Fisher's Exact Test)

標本数 N が小さい場合には、 X^2 や G^2 の χ^2 近似は難しい。この場合には、独立性の仮説のもとでの正確な分布を用いて検定を行なうことができる。

独立性の仮説が成り立つとき、周辺度数が固定されているとの条件のもとで (つまり n_{1+}, n_{+1} の両者が固定されている)、 n_{11} は超幾何分布 (hyper geometric distribution) に従う。この分布はつぎの式であらわされる。

$$p(n_{11}) = \frac{\binom{n_{1+}}{n_{11}} \binom{n_{2+}}{n_{+1} - n_{11}}}{\binom{n}{n_{+1}}} \quad (48)$$

超幾何分布は、つぎのような確率を表すものと考えればよい。壺のなかに白い玉が n_{+1} 個、赤い玉が n_{+2} 個含まれている。合わせて n 個である。ここで、壺のなかを見ずに、 n_{+1} 個の玉を取り出す。取り出したなかに含まれる白い玉の個数が n_{11} である。

超幾何分布の確率は上の式を用いて計算できるので、この分布と観測された n_{11} とを比較し、独立性の仮説を棄却できるかどうか検討すればよい。

ここで、 2×2 表の第 1 行の比率を p とおくと、 $n_{1+} = np$ となり、独立性の仮定のもとでの n_{11} の平均と分散は超幾何分布の性質からつぎの式で与えられる。これらの式は後述する Cochran-Mantel-Haenszel 検定で利用する。

$$E(n_{11}) = n_{+1}p = \frac{n_{1+}n_{+1}}{n} \quad (49)$$

$$\text{Var}(n_{11}) = \frac{n_{+1}n_{+2}n_{1+}n_{2+}}{(n-1)n^2} \quad (50)$$

11.7 対数オッズ比の分散

2 値データから得られるオッズ比がどの程度変動するかは、つぎのようにして知ることができる。

ここで、各セルの件数 n_{ij} は各々独立にポアソン分布 (後述する) に従うと仮定する。平均母数 μ のポアソン分布の分散は μ であるので標準偏差は $\sqrt{\mu}$ となる。オッズ比の対数 (対数オッズ比) はつぎの式で表される。

$$\log(n_{11}n_{22}/n_{12}n_{21}) = \log n_{11} + \log n_{22} - \log n_{12} - \log n_{21} \quad (51)$$

ここで $\log x$ の微分が $1/x$ であることを利用する。 n_{ij} がある程度以上大きければ、 $\log n_{ij}$ の分散は $1/n_{ij}^2 \times \text{Var}(n_{ij})$ で近似できる。これを用いると各 n_{ij} が大きい時には

$$\text{Var}\{\log(n_{11}n_{22}/n_{12}n_{21})\} \simeq \frac{1}{n_{11}} + \frac{1}{n_{22}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} \quad (52)$$

となる。

11.8 4 分点相関係数 (ϕ 係数)

2 つの変数 X と Y の間の Seearman の相関係数 (一般には単に「相関係数」または積率相関係数とも呼ばれる) はつぎの式で与えられる。

$$r = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^N (x_i - \bar{x})^2 \times \sum_{i=1}^N (y_i - \bar{y})^2}}$$

この値は、連続値を持つ2つの変数間の関係の指標としてしばしば用いられる。

Spearman の相関係数を2値データに適用すると、つぎのような値が得られる。

$$\phi = \frac{n_{11}n_{22} - n_{12}n_{21}}{\sqrt{n_{1+}n_{2+}n_{+1}n_{+2}}} \quad (53)$$

この値の2乗の n 倍は X^2 になっている。つまり $X^2 = n\phi^2$ である。

この値は4分点相関係数(または ϕ 係数)などと呼ばれる。相関係数を求める関数があればこの値は即座に求まるので便利ではあるが、周辺度数の影響を受けやすい。

もし、2値変数の背後に連続変数が存在し、その値が何らかの閾値によって0-1値に変換されていると仮定すると、上の係数はもとの連続変数の相関より0に近づくことが多い。また、閾値が中央値から外れると、背後の連続変数の相関が同一であっても、より絶対値が小さくなる。

つぎは、 X と Y とが相関係数 $1/\sqrt{2} = 0.7071068$ の2変量正規分布に従う場合の例である。各変数の平均は0、分散は1である。この分布から2000個の乱数を取り出した。標本相関係数を計算すると0.7178727である。つぎに、値が正であれば1、負であれば0として集計するとつぎのような結果が得られる。

	$Y = 0$	$Y = 1$
$X = 0$	766	222
$X = 1$	252	760

このとき $\phi = 0.5263392$ であり、オッズ比は10.40612である。

また、 X と Y の各々について、値が1より大きければ1、それ以下ならば0として集計するとつぎのような結果が得られる。

	$Y = 0$	$Y = 1$
$X = 0$	1524	154
$X = 1$	153	169

この場合の ϕ は0.4325268となる。オッズ比は10.93099である。

11.9 4分相関係数 (tetrachoric correlation)

分析対象となる 2×2 表の背後に、相関のある2変量正規分布を仮定し、2つの変数がある閾値によって分割されることにより、データが得られていると解釈する。この仮定にもとづいて、データにもっとも良く当てはまる分布を推定し、その相関係数を求める。このようにして得られる値を4分相関係数とよぶ。

11.10 Yule-Simpson のパラドックス

ここでは 2×2 の頻度データが複数ある場合について考える。先週の資料に示したアスピリンと心筋梗塞の例で考えると、一つの実験でこれだけのデータを集めることができるわけではない。実際には、実験者も実験場所も異なる複数のデータを集めたものである。ここで、薬の投与をあらわす変数を X 、心筋梗塞の有無を Y 、実験者(病院)を Z とする。関心対象になるのは、 X と Y の関係であるが、 Z はこの関係に影響を及ぼしているかも知れない。このように関心の対象となっている関係に影響を及ぼす可能性のある変数を、層別変数と呼び、各条件のことを層(layer)と呼ぶ。

層別にデータをみることが重要なのは、データ全体の集計結果と層別に検討した結果が異なる場合があるからである。表 11 のデータは、被告人の人種と死刑判決についてのデータである (Draper et al. (1992) より引用)。原論文の著者 (Radelet) は 1976 年から 1977 年にかけての米国フロリダ州の 20 の郡における 637 件の殺人事件のデータを収集している。フロリダ州全体には 67 の郡があるが、その内から人口を考慮し対象とする 20 郡が抽出された。637 件の事件のうち 311 件は、被告と被害者が顔見知りのケースであり、これを除いた 326 件が表に示したものである。

まず表 11 を見ると、被告人の人種と死刑判決の率の間には、強い関係はないように見える (白人被告 11.9%、黒人被告 10.2%)。しかし、被害者の人種別に同じデータを集計しなおして検討すると (表 12)、被告が黒人である場合の死刑判決の率が高いのが分かる。集計したデータでこのような傾向が消えるのは、データにつきのような特徴があるからである。

1. 白人の被告は白人を被害者とすることが多い
2. 黒人の被告は黒人を被害者とするが多い
3. 被害者が黒人であるときには死刑判決が少い

表 11: 殺人事件の被告人の人種と死刑判決

被告人の人種	死刑	死刑以外	計
白人	19	141	160
黒人	17	149	166
計	36	290	326

(オッズ比 1.18) *Source: Radelet, M. (1981). Racial characteristics and imposition of death penalty. American Sociological Review, 46, 918-927.*

表 12: 被告人の人種と死刑判決 (被害者の人種別)

被害者	被告人	死刑	死刑以外	計
白人	白人	19	132	151
	黒人	11	52	63
	小計	30	184	214
黒人	白人	0	9	9
	黒人	6	97	103
	小計	6	106	112
合計		36	290	326

(被害者が白人の場合のオッズ比 0.68)

上の例では、層別の傾向と全体の集計表における傾向とがかなり異なっている。極端な場合には、層別でのオッズ比と全体でのオッズ比が逆の傾向を示すことも有り得る。このような逆説的な例は、Yule-Simpson のパラドックス (Yule-Simpson's paradox) と呼ばれている。

表 13 はもう一つの例である。これは、1973 年のカルフォルニア大学バークレイ校への学部 (大学院) 別、性別の入学志願者数と合否結果である。全体を合計すると、男子の合格率が女子より大きい、学部別に検討すると B から F の学部では男女の合格率にほとんど違いがなく、また学部

表 13: 1973 年におけるカルフォルニア大学バークレイ校への学部別、性別の入学志願と可否の結果

性別	可否	学部						合計
		A	B	C	D	E	F	
男性								
	合格	512	353	120	138	53	22	1198
	不合格	313	207	205	279	138	351	1493
女性								
	合格	89	17	202	131	94	24	557
	不合格	19	8	391	244	299	317	1278
合計		933	585	918	792	584	714	4526

(Source: Freedman, D., Pisani, R. & Purves, R. (1978). *Statistics*, New York: W.W. Norton.

A では女子の方が合格率が高い。全体の傾向は、女子がより入学の難しい学部を受験する傾向が強いため生じている。

11.11 Cochran-Mantel-Haenszel 検定

そこで、層別のデータを対象として変数間の独立性の検定を行うための方法 (Cochran-Mantel-Haenszel 検定、以下では CMH 検定と略す) が開発されている。この方法が検定する帰無仮説は、「各層においてオッズ比がゼロ」であり、対立仮説は「各層がゼロでないある共通のオッズ比を持つ」というものである。

CMH 検定の基本的なアイデアは、つぎのようなものである。まず、各層ごとに n_{11} の帰無仮説のもとでの正確な分布 (超幾何分布) を考え、その平均と分散を求めるとつぎのようになる。

$$\mu_{11} = E(n_{11}) = \frac{n_{1+}n_{+1}}{n_{++}} \quad (54)$$

$$\text{Var}(n_{11}) = \frac{n_{1+}n_{2+}n_{+1}n_{+2}}{n_{++}^2(n_{++} - 1)} \quad (55)$$

ついで、各層における帰無仮説のもとでの n_{11k} の期待値 μ_{11k} からのずれの和を求め、これらを足し合わせる。この値は帰無仮説のもとで近似的に平均ゼロの正規分布に従うとみなせる。また分散の大きさを求めることもできる。この和の平均と分散は各層のデータが独立に分布することから、各々

$$E(n_{11+}) = \sum_k E(n_{11k}), \quad \text{Var}(n_{11+}) = \sum_k \text{Var}(n_{11k}) \quad (56)$$

となる。さらに、連続修正を加えたつぎの値が近似的に自由度 1 の χ^2 分布に従うことを用いて検定を行なう。

$$\frac{(|n_{11+} - E(n_{11+})| - 1/2)^2}{\text{Var}(n_{11+})} \quad (57)$$

CMH 検定は、対立仮説として各層における変数の関係がほぼ同じ傾向を持つものであることを仮定している。このため、層ごとに関係の方向が異なるデータにはうまく適用できない。

12 回帰分析とは何か

回帰分析 (regression analysis) とは、一般的には次のような関係の分析方法を示す。

被説明変数 (従属変数, dependent variable) Y と 1 つ以上の説明変数 (独立変数, independent variable) X_1, \dots, X_p を考える。これらについて、 Y の条件付平均 $E(Y|X_1, \dots, X_p)$ を説明変数の関数として表わしたい。もっともよく用いられるのは、この関数を説明変数の 1 次式によって推定するつぎの式である。

$$E(Y|X_1, \dots, X_p) = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p$$

回帰分析を用いる場合、次のような点が問題になる。

1. 説明変数の関数 (1 次式の場合は回帰係数 $\hat{\beta}_i, i = 1, \dots, p$) をどのようにして求めればよいか。
2. 回帰係数の推定値 $\hat{\beta}_i$ や予測値 \hat{Y} はどの程度の精度を持っているか。
3. 説明変数 X_1, \dots, X_p をどのように選べば良いか。
4. 説明変数が、実験などにおいて固定された変数として解釈できる場合と、社会調査などのように説明変数自体が外的な影響によって変動している場合とではどこが違うか。
5. 推定された回帰係数 $\hat{\beta}_i$ を、影響関係を表す値として解釈することは可能か。
6. 被説明関数と説明変数との間に非線形の関係がある場合には、どのようにすればよいか。

12.1 相関係数と回帰直線

ここでは、まず説明変数が 1 個の場合を考える。サンプル件数 (データの個数) を N とする。被説明変数 Y の値を $\{y_1, \dots, y_N\}$ とする。また説明変数 X の値を $\{x_1, \dots, x_N\}$ とする。求めようとしているのは、

$$E(Y|x) = \beta_0 + \beta_1 x$$

という 1 次式である。

回帰分析を考える場合、 X が確率変数と仮定するものと、分析の対象となる X の値 $\{x_1, \dots, x_N\}$ は、固定された値であるとするものがある。教科書によく引用されている、父親と息子の身長の場合では、 Y (息子の身長) と X (父親の身長) の両者が確率変数であるとみなせる。一方、実験データにおいて、 X が制御される変数を表す場合には、 $\{x_1, \dots, x_N\}$ はあらかじめ実験者によって指定された値である。ここでは、とくに断らない限り後者 (固定された x_i) を仮定する。

通常回帰分析は、つぎの確率モデルを仮定する。

$$Y_i = \beta_0 + x_i \beta_1 + \varepsilon_i, \quad (i = 1, \dots, N) \quad (58)$$

ここで ε は直線からのズレをあらわす誤差成分であり、通常は (1) 互いに独立で、(2) 平均ゼロ、また (3) 同一の分散 (σ^2) を持つ、(4) 正規分布に従うと仮定する。問題は、与えられたデータを用いて最も適切と思われる係数 β_0 および β_1 の値を求めることである。

そこで、係数を推定するための基準を

$$SS = \sum_{i=1}^N (y_i - \beta_0 - \beta_1 x_i)^2, \quad (59)$$

とし、この値を小さくする係数が望ましいと考える。

行列を用いるとつぎのように記述することができる。

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}, \mathbf{X} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_N \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$$

ここで、(59) は

$$SS = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 \quad (60)$$

となる。このような、目的からのズレの大きさの2乗和を最小化する推定方法は一般的に最小2乗法と呼ばれる。

この SS を最小にする $\boldsymbol{\beta}$ は、 $\mathbf{X}^T \mathbf{X} \boldsymbol{\beta} = \mathbf{X}^T \mathbf{y}$ の解になることが微分を検討することによってわかる。1変数の場合に、この式の値をより詳しく検討すると、つぎのようなことがわかる。

以下の説明で次の記号を用いる。

$$\bar{y} = \frac{1}{N} \sum_i y_i, \quad \bar{x} = \frac{1}{N} \sum_i x_i \quad (61)$$

$$\text{var}(x) = \frac{1}{N} \sum_i (x_i - \bar{x})^2, \quad \text{cov}(x, y) = \frac{1}{N} \sum_i (x_i - \bar{x})(y_i - \bar{y}) \quad (62)$$

$$\mathbf{C} = (\mathbf{X}^T \mathbf{X})^{-1} = \frac{1}{N^2 \text{var}(x)} \begin{pmatrix} \sum_i x_i^2 & -\sum_i x_i \\ -\sum_i x_i & N \end{pmatrix}, \quad \mathbf{X}^T \mathbf{y} = \begin{pmatrix} \sum_i y_i \\ \sum_i x_i y_i \end{pmatrix} \quad (63)$$

残差2乗和(60)を最小にする $\boldsymbol{\beta}$ は、

$$\hat{\boldsymbol{\beta}} = \mathbf{C} \mathbf{X}^T \mathbf{y} \quad (64)$$

であたえられる。これより

$$\hat{\beta}_0 = \bar{y} - \frac{\text{cov}(x, y)}{\text{var}(x)} \bar{x} = \bar{y} - \hat{\beta}_1 \bar{x}, \quad \hat{\beta}_1 = \frac{\text{cov}(x, y)}{\text{var}(x)} \quad (65)$$

となる。

$\hat{\mathbf{y}} = \mathbf{X} \hat{\boldsymbol{\beta}}$ とすると、次の式が成り立つ。

$$\|\mathbf{y} - \bar{\mathbf{y}}\|^2 = \|\mathbf{y} - \hat{\mathbf{y}}\|^2 + \|\hat{\mathbf{y}} - \bar{\mathbf{y}}\|^2 \quad (66)$$

ここで、 $\bar{\mathbf{y}}$ は、要素の全てが \bar{y} である長さ N のベクトルである。要素を用いて表すと、上式はつぎのようになる。

$$\sum_i (y_i - \bar{y})^2 = \sum_i (y_i - \hat{y}_i)^2 + \sum_i (\hat{y}_i - \bar{y})^2 \quad (67)$$

標本相関係数の定義は

$$r = \frac{\text{cov}(x, y)}{\sqrt{\text{var}(x) \text{var}(y)}} \quad (68)$$

であるが、相関係数と回帰による予測残差の間に、つぎのような関係がある。

$$r^2 = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \quad (69)$$

12.2 被説明変数の変換

実際に回帰分析を用いる場合、モデルの仮定が必ずしも成立しているとは限らない。特に、被説明変数の条件付期待値が説明変数の1次式によってはうまく表せない場合は少なくない。また、誤差分散が、説明変数の値によって異なる場合も多い。回帰分析の計算手順は、前節に述べた仮定に依存しているので、これらの条件が成立しなければ、得られた回帰式の推定値が望ましいものであることが疑わしくなる。

図14には自動車の速度 (mph) と停止距離 (ft) のデータを示す (出典 Ezekiel,1930; R-1.2 のライブラリより引用)。左側は生のデータであり、右側は被説明変数を停止距離の平方根にとっている。いずれも実線は回帰直線であり、破線は後述する lowess という手法で平滑化を行ったものである。

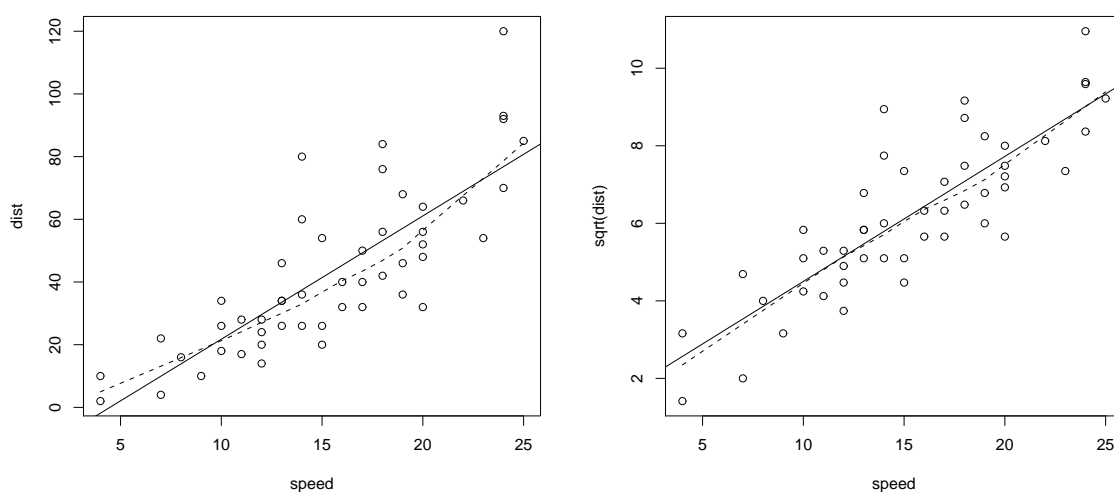


図 14: 自動車停止距離データ (Ezekiel;1930)

左側の図では、残差が中央部で負の方向に偏り、両端で正方向へ偏る傾向が見られる。また、大きな速度 (speed) においては残差の絶対値が大きくなる傾向が見られる。一方、右側の図ではこれらの傾向は、かなり改善されている。

統計モデルの仮定が成立しているか否かをデータに基いて検討することは、一般的にモデル診断と呼ばれる。多くの場合、残差の分布の傾向を検討することによって有益な情報が得られる。

また、変数とくに被説明変数に特定の変換を加えることにより、回帰分析モデルの仮定がより妥当であるようにすることができる。このために用いられる非線形変換の方法には、つぎのようなものがある。

1. 対数変換 ($V = \log Y$): 後述の Box-Cox 変換において、 $\lambda = 0$ の場合は対数変換として定義されている。
2. 平方根変換 ($V = \sqrt{Y}$): 被説明変数が頻度である場合には、ポアソン分布に従っているとみなせることが多い。平均 μ のポアソン分布の分散は μ であるので、頻度がある程度大きい数である場合には、平方根をとることにより、分散を一定に近付けることができる。

表 14: 分散安定化変換

ε の標準偏差が $\eta = E(Y|x)$ の関数であるとき、近似的に一定に近づける変換

ε の標準偏差		分散安定化変換
$\sigma_Y = f(\eta)$		
$\sigma_Y \propto \eta^k$	$(Y \geq 0)$	Y^{1-k}
$\sigma_Y \propto \eta$	$(Y \geq 0)$	$\log Y$
$\sigma_Y \propto \eta^{1/2}(1-\eta)^{1/2}$	$(0 \leq Y \leq 1)$	$\sin^{-1}(\sqrt{Y})$
$\sigma_Y \propto \eta^{-1}(1-\eta)^{1/2}$	$(0 \leq Y \leq 1)$	$(1-Y)^{1/2} - (1-Y)^{3/2}/3$
$\sigma_Y \propto (1-\eta^2)^{-2}$	$(-1 \leq Y \leq 1)$	$\log\{(1+Y)/(1-Y)\}$

3. Box-Cox 変換：冪乗による変換の一般化。つぎの式で \tilde{Y} は、 Y_1, \dots, Y_N の幾何平均 ($Y_1 \times Y_2 \cdots Y_N$)^{1/N} である。

$$V = \begin{cases} (Y^\lambda - 1)/(\lambda \tilde{Y}^{\lambda-1}) & \text{for } \lambda \neq 0, \\ \tilde{Y} \log Y & \text{for } \lambda = 0 \end{cases} \quad (70)$$

4. 逆正弦変換 ($V = \sin^{-1}(\sqrt{Y/m})$): Y が 2 項分布 $Bin(m, p)$ に従う場合、分散は $mp(1-p)$ である。 m がある程度大きければ、この変換を加えることによって p の値によらず分散を一定に近づけることができる。

13 重回帰分析

説明変数が複数ある場合に、モデルを拡張する。 $y_{N \times 1}$ は被説明変数であり、 $X_{N \times p}$ は説明変数とする。通常 X の第 1 列は定数 (全て 1) とする。

多変量の回帰分析のモデル (重回帰分析) はつぎのようなものである。

$$y = X\beta + \varepsilon \quad (71)$$

ここで、 ε の要素 ε_i ($i = 1, \dots, N$) は、平均ゼロ、分散 σ^2 の正規分布に互いに独立に従うものとする。

回帰係数 β はつぎの式で推定される。

$$\hat{\beta} = (X^T X)^{-1} X^T y \quad (72)$$

一般的に

$$E(Az) = AE(z), \quad Var(Az) = AVar(z)A^T \quad (73)$$

が成立するので、 $\hat{\beta}$ の平均は β であり、分散共分散行列は $\sigma^2(X^T X)^{-1}$ となることがわかる。

被説明変数の予測値は、 $\hat{y} = X\hat{\beta}$ で与えられ、また誤差分散の不偏推定値は、

$$\hat{\sigma}^2 = (y - \hat{y})^T (y - \hat{y}) / (N - p) = \|y - \hat{y}\|^2 / (N - p) \quad (74)$$

で与えられる。 $\hat{\sigma}^2$ は、モデルの仮定が正しければ $\sigma^2 \times \chi_{N-p}^2 / (N - p)$ に従う。また、理論的な検討から、 $\hat{\beta}$ と $\hat{\sigma}^2$ とはモデルの仮定の下で統計的に独立であることが知られている。

X が定数の列を含む場合、次の値を重相関係数の 2 乗 (R^2) と呼ぶ。 R^2 は r^2 と同様のつぎの式で定義される。次式は (69) と同一の内容をベクトルを用いて表記したものである。

$$R^2 = \frac{\|\hat{\mathbf{y}} - \bar{\mathbf{y}}\|^2}{\|\mathbf{y} - \bar{\mathbf{y}}\|^2} = 1 - \frac{\|\mathbf{y} - \hat{\mathbf{y}}\|^2}{\|\mathbf{y} - \bar{\mathbf{y}}\|^2} \quad (75)$$

R^2 は 0 から 1 までの範囲をとる。 $R^2 = 1$ は、誤差がゼロであること、すなわち $\mathbf{y} = \hat{\mathbf{y}}$ を意味する。また、 $R^2 = 0$ は、 $\hat{\mathbf{y}}$ の分散がゼロ、すなわち $\hat{y}_i = \bar{y}$ ($i = 1, \dots, N$) を意味する。

13.1 説明変数の選択

回帰モデルの性質として、多くの説明変数を含むモデルの推定を行なうと、回帰係数の推定精度が落ちることが知られている。多くの説明変数を用いて R^2 を求めると見掛け上大きな値が与えられることがあるが、別のデータに推定された $\hat{\beta}$ を適用すると予測の性能は概して悪い。もしデータが十分に大量にあればこのような問題は生じないが、複雑なモデルの推定について注意を払うことが不要であるほど大量のデータは普通には利用できない。このような問題が生じるのは、回帰係数がたまたま目前にあるデータについて日和見的に適合しているためである。これを避けるためには、モデルの適合度は保ちつつ、極力少数の説明変数によって予測を行なう必要がある。回帰分析における F 検定は、変数選択のための一つの方法である。

回帰係数の内、要素 $\beta_{q+1}, \dots, \beta_p$ がゼロであるか否かを検討することにしよう。 M_0 を $\beta_{q+1} = \dots = \beta_p = 0$ と制約を加えたモデル (説明変数の個数が q のモデル) とする。また、 M_1 を $\beta_{q+1}, \dots, \beta_p$ が必ずしもゼロでは無いモデルとする。モデル M_0 によって推定された \mathbf{y} の予測値を $\hat{\mathbf{y}}_0$ とし、 M_1 による予測値を $\hat{\mathbf{y}}_1$ とする。モデル M_0 (より単純なモデルといえる) が正しいとの仮定の下で、つぎのことが成立する。

$$\|\mathbf{y} - \hat{\mathbf{y}}_0\|^2 \sim \sigma^2 \chi_{N-q}^2 \quad (76)$$

$$\|\mathbf{y} - \hat{\mathbf{y}}_1\|^2 \sim \sigma^2 \chi_{N-p}^2 \quad (77)$$

$$\|\mathbf{y} - \hat{\mathbf{y}}_0\|^2 - \|\mathbf{y} - \hat{\mathbf{y}}_1\|^2 \sim \sigma^2 \chi_{p-q}^2 \quad (78)$$

また M_0 の仮定の下で、(77) と (78) とは独立である。また、このとき

$$F = \frac{\|\mathbf{y} - \hat{\mathbf{y}}_0\|^2 - \|\mathbf{y} - \hat{\mathbf{y}}_1\|^2}{\|\mathbf{y} - \hat{\mathbf{y}}_1\|^2} \times \frac{N-p}{p-q} \quad (79)$$

は、自由度 $(p-q, N-p)$ の F 分布に従う。もし、 M_0 が正しくなく、 M_1 が説明のために必要であるのならば、(76) の値は M_0 が成立する場合よりも大きくなるはずである。そこで、 F 値 (79) を F 分布の値と比較し、もし顕著に大きければ「 M_0 が成立している」という仮説を棄却することにする。通常、 F 値が分布の上位 5 パーセントに入れば、 M_0 を棄却することが多い。

また、回帰係数 β_j について、つぎのような検討を加えることが多い。以下で、行列 $(X^T X)^{-1}$ を $C = (c_{ij})$ と表記する。上に示したように、回帰係数 $\hat{\beta}_j$ の平均は β_j 、分散は $\sigma^2 \times c_{jj}$ である。残差の分散 σ^2 は未知であり、この推定値として $\hat{\sigma}^2$ を用いることにする。モデル M_1 の仮定の下で、

$$t = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 c_{jj}}} \quad (80)$$

は自由度 $N - p$ の t 分布に従う。これを用いて、平均値の信頼区間の場合と同様に、 β_j の信頼区間を求めることができる。もし、信頼区間がゼロを含めば、第 j 番目の変数を説明変数に加える必要は薄いとみなされる。

より単純で理論的に簡明なモデルの選択方法として、赤池情報量基準 AIC(Akaike's Information Criterion) を用いる方法がある。これは AIC とよばれる指標を最小化するモデルを最良モデルとみなす方法である。誤差分散が未知の回帰モデルの場合、AIC は次の式で与えられる。

$$\text{AIC} = N \log \frac{\|\mathbf{y} - \hat{\mathbf{y}}\|^2}{N} + 2q + \text{定数} \quad (81)$$

$$= N \log \|\mathbf{y} - \hat{\mathbf{y}}\|^2 + 2q + \text{定数}' \quad (82)$$

ここで、 q はモデルに含まれるパラメータの個数であり、誤差分散未知の回帰モデルの場合には $p + 1$ である。

また Schwarz(1978) は、ベイズ推定の立場からつぎのようなモデル選択基準 (のちに BIC と呼ばれるようになった) を提案している。

$$\text{BIC} = N \log \frac{\|\mathbf{y} - \hat{\mathbf{y}}\|^2}{N} + q \log N + \text{定数} \quad (83)$$

この他に伝統的にモデル選択の基準として利用されている指標に、自由度調整済み重相関係数 R^* がある。これは、(75) で定義される重相関係数 (の 2 乗) を修正した、次の式で定義される。ここで p は (定数を除く) 説明変数の個数を表す。

$$R^* = \sqrt{R^2 - \frac{p}{N - p - 1}(1 - R^2)} = \sqrt{1 - \frac{\|\mathbf{y} - \hat{\mathbf{y}}\|^2 / (N - p - 1)}{\|\mathbf{y} - \bar{\mathbf{y}}\|^2 / (N - 1)}} \quad (84)$$

14 モデル診断の方法

たとえデータがどんなものであっても、回帰分析を行えば解は得ることができる。しかし、回帰分析が仮定している条件が成立してないならば、解についての検討は不適切なものになってしまうかもしれない。得られた解についての検討は、残差と各々のサンプルが解に及ぼす影響を通じて行う。

14.1 残差の評価

回帰係数 β の推定値 $\hat{\beta}$ は、最小 2 乗基準を満す解であり、 \hat{y}_i は説明変数の値を固定したときの Y の平均、つまり $E(Y|x_i)$ を推定していると解釈できる。 x_i は X の第 i 行の行ベクトルである。また、 $e_i = y_i - \hat{y}_i$ ($i = 1, \dots, N$) を残差 (residual) とよび、これらを成分とするベクトル e を残差ベクトルと呼ぶ。

もし、 \hat{y}_i が十分に正確に $E(Y|x_i)$ を推定しており、さらに $\hat{\sigma}^2$ も σ^2 の推定値として十分に正確ならば、モデルの仮定の下で、

$$\frac{y_i - \hat{y}_i}{\hat{\sigma}} = \frac{e_i}{\hat{\sigma}}, (i = 1, \dots, N) \quad (85)$$

は、近似的に標準正規分布に独立に従う。

しかし、現実の多くの例では \hat{y}_i および $\hat{\sigma}^2$ の両者が確率的に変動する。そこで、もう少し厳密に e の分布を考えると、 $\hat{y} = X\hat{\beta} = X(X^T X)^{-1} X^T y$ であるので、

$$e = y - \hat{y} = (I - X(X^T X)^{-1} X^T) y \quad (86)$$

と表される。これより、

$$E(e) = \mathbf{0}, \quad \text{Var}(e) = \sigma^2 (I - X(X^T X)^{-1} X^T) \quad (87)$$

であることが導かれる。ここで、 $H = X(X^T X)^{-1} X^T$ とおき、この第 i 対角要素を h_{ii} とする。

$\hat{\sigma}^2/\sigma^2$ はモデルの仮定の下で自由度 $N - p$ の χ^2 分布に従うので残差をその分散の推定値の平方根で割って標準化をおこなう。

$$s_i = \frac{e_i}{\sqrt{\hat{\sigma}^2(1 - h_{ii})}} \quad (88)$$

この値を標準化残差 (standardized residuals) と呼ぶ。Draper & Smith (1998) では内的にスチューデント化された残差 (internally Studentized residuals) と呼んでいる。

ちなみに、

$$X = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ \vdots & \vdots \\ 1 & 10 \end{pmatrix}$$

とすると、 h_{ii} の値は $i = 1$ および $i = 10$ のとき 0.345、また $i = 5$ および $i = 6$ のとき 0.103 である。

さらに、 i 番目のデータの残差の標準化を行う場合に、データ全体から推定された $\hat{\sigma}^2$ ではなく、該当する残差の寄与を除いて推定された値を用いると、外れ値をより敏感に検出することが可能になり、また e_i と標準化のための分母が独立に分布する。 e_i の寄与分を除いて推定される残差分散はつぎの式で表される。

$$\hat{\sigma}^2(i) = \frac{(N - p)\hat{\sigma}^2 - e_i^2/(1 - h_{ii})}{N - p - 1} \quad (89)$$

この値を用いて標準化を行うと、

$$t_i = \frac{e_i}{\sqrt{\hat{\sigma}^2(i)(1 - h_{ii})}} \quad (90)$$

という値が得られる。この値をスチューデント化残差 (Studentized residuals) と呼ぶ。Draper & Smith(1998) では、外的にスチューデント化された残差 (externally Studentized residuals) と呼んでいる。

回帰分析の仮定は、

1. Y_i の平均 $E(Y|x)$ が説明変数 x の 1 次関数で表される。
2. Y_i ($i = 1, \dots, N$) は互いに独立。
3. Y_i の分散 $\text{Var}(Y|x)$ が一定。
4. Y_i は正規分布に従う。

というものである。これらの仮定が、実際に妥当なものであるか否かは、残差の分布の状態を図等によって確認することができる。もし、 $E(Y|x)$ が説明変数の非線形（1次関数ではないこと）な関数として表されるのなら、残差は平均ゼロではなく系統的な偏りを示す。また、 $Var(Y|x)$ が一定ではなく、 $E(Y|x)$ と関係をもっているならば、 \hat{y}_i と e_i の散布図を描くことにより、そのような傾向を発見できる。しばしば生じるのは、 $E(Y|x)$ が大きいと $Var(Y|x)$ もそれにつれて大きくなる場合である。このような傾向が生じたなら、 y_i の値を対数や平方根などによって変換したのち回帰分析を行なうと、より適切な分析が可能になる。

モデル診断にあたって注意すべき点はつぎのようなものである。

1. 残差に系統的な偏りがある場合には、本來說明変数によって説明されるべき変動が、直線からのズレとみなされ残差となる。このため、説明変数の影響が過小評価されているかも知れない。
2. 残差の分布が正規分布とはみなせない場合、とくに裾が重い（極端な値が生じやすい）場合にも、残差分散の推定が過大評価されやすい。回帰係数の検定には、残差分散の平方根が分母にあらわれるので、この値が過大評価されると、説明変数の効果があるのに見過ごす可能性が大きくなる。
3. 回帰係数の推定方法は残差分散が同一であることを仮定しているので、残差分散の大きさが説明変数の値によって異なる場合には、推定された回帰係数の値が妥当かどうか疑わしい。

14.2 各サンプルの解への影響評価 (感度分析)

残差の他に注意すべきことに、各データが推定値にどれだけの影響を及ぼしているかがある。大きな残差を持つデータは解に大きい影響を及ぼしがちであるが、たとえ残差が小さくとも、1個のデータの僅かな変動が推定値全体に大きな影響を及ぼす場合もある。

そこで i 番目のデータを除いた $N - 1$ 個のデータに回帰分析を適用して推定された N 個の推定値を $\hat{y}(i)$ とする。また、このときの回帰係数の推定値を $\hat{\beta}(i)$ とする。

ここで i 番目のデータの推定値に及ぼす影響をつぎの式で表す。

$$D_i = (\hat{y} - \hat{y}(i))^T (\hat{y} - \hat{y}(i)) / (p\hat{\sigma}^2) \quad (91)$$

$$= (\hat{\beta} - \hat{\beta}(i))^T X^T X (\hat{\beta} - \hat{\beta}(i)) / (p\hat{\sigma}^2) \quad (92)$$

この値は Cook の距離と呼ばれている。また、この値を e_i を用いてあらわすと、つぎのようになる。

$$D_i = \left\{ \frac{e_i}{\sqrt{\hat{\sigma}^2(1 - h_{ii})}} \right\}^2 \left\{ \frac{h_{ii}}{1 - h_{ii}} \right\} \frac{1}{p} \quad (93)$$

15 説明変数も確率変数である場合

ここでは被説明変数が、 $Y = X_1$ であり、説明変数は X_2, \dots, X_p であるとする。通常は X_1, \dots, X_p の同時分布が、多変量正規分布に従うと仮定して分析する。実際に得られる多変量のデータが多変量正規分布に従っているとの保証はないが、理論的な明解さからこのような仮定を置いて分析を進めることが多い。

独立に標準正規分布に従う p 個の確率変数 Z_1, \dots, Z_p を要素とする縦ベクトルを z とし、 X_1, \dots, X_p を要素とする縦ベクトルを x とする。 x が多変量正規分布に従うとは、各要素の平均を表すベクトル μ と、 $p \times p$ の行列 A があり、 $x = \mu + Az$ と書き表されることを意味する。

ベクトル x が多変量正規分布に従うなら、その要素のどんな 1 次式もまた正規分布に従う。特に、各要素の周辺分布も正規分布になる。

これまでの説明では、説明変数は実験者によって外的に定められた値をとることを仮定していたが、回帰の概念は説明変数が確率変数である場合にも適用できる。つまり、 $E(X_1|X_2, \dots, X_p)$ を X_2, \dots, X_p の式として求めることができるならば、これを回帰式とすればよい。

多変量正規分布の場合には、 X_1 の条件付分布はつぎの平均と分散を持つ正規分布であらわすことができる。ここで、 X_1 の分散を σ_{11} 、 X_1 と X_2, \dots, X_p の共分散からなる行ベクトルを $\Sigma_{1,2\dots p}$ 、 X_2, \dots, X_p の分散共分散行列を $\Sigma_{2\dots p,2\dots p}$ で表す。また、 Σ によって X_1, \dots, X_p 全体の分散共分散行列を示す。

$$E(X_1|X_2, \dots, X_p) = \mu_1 + \Sigma_{1,2\dots p} \Sigma_{2\dots p,2\dots p}^{-1} (X_2 - \mu_2, \dots, X_p - \mu_p)^T \quad (94)$$

$$\text{Var}(X_1|X_2, \dots, X_p) = \sigma_{11} - \Sigma_{1,2\dots p} \Sigma_{2\dots p,2\dots p}^{-1} \Sigma_{1,2\dots p}^T \quad (95)$$

$$(96)$$

上の (95) は、 X_1 の条件付分布の分散が X_2, \dots, X_p の値によらず一定であることを示している。これは多変量正規分布の著しい特徴である。また、(94) の $\Sigma_{1,2\dots p} \Sigma_{2\dots p,2\dots p}^{-1}$ が、回帰係数 β^T に相当する。

実際の計算においては、真の分散共分散の値は分からないので、データからつぎのように推定する。ここで x_i は変数 X_1, \dots, X_p の第 i サンプルにおける値を示す列ベクトルである。また \bar{x} は標本平均を示す。まず、つぎの式によって p 個の変数の分散共分散行列の推定を行う。

$$S = \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})(x_i - \bar{x})^T \quad (97)$$

分散共分散行列の $(N-1)$ 倍 $(N-1)S$ の分布は、ウィシャート (Wishart) 分布 $W(\Sigma, N-1)$ と呼ばれるものになる。

この S を用いて計算される回帰係数のベクトル $\hat{\beta} = S_{2\dots p,2\dots p}^{-1} S_{1,2\dots p}^T$ の正確な分布は複雑になるが、十分 N が大きければ S は真の分散共分散行列 Σ に収束するので、 $\hat{\beta}$ も真の回帰係数ベクトルに収束する。

15.1 相関係数の分布

変数 X_j と X_k の (母) 相関係数 ρ_{jk} は、つぎの式で定義される。

$$\rho_{jk} = \frac{\sigma_{jk}}{\sqrt{\sigma_{jj}\sigma_{kk}}} \quad (98)$$

標本相関係数 r_{jk} は、 S の要素を持ちいて

$$r_{jk} = \frac{s_{jk}}{\sqrt{s_{jj}s_{kk}}} \quad (99)$$

と定義される。ここでつぎの変換 (Fisher の z 変換) を考える。(添字は省略する。)

$$z = \frac{1}{2} \log \frac{1+r}{1-r}, \quad \zeta = \frac{1}{2} \log \frac{1+\rho}{1-\rho} \quad (100)$$

するとつぎの式が漸近的に成り立つことが知られている。

$$\sqrt{N-3} \left(z - \zeta - \frac{\rho}{2(N-1)} \right) \stackrel{\text{asym.}}{\sim} N(0, 1) \quad (101)$$

この近似は、分布が多変量正規分布の場合に成立するものであり、この仮定が成立しない場合には、妥当性を持たない。

また、2変量正規分布についての相関係数の正確な分布の公式は求められているが、かなり複雑になる。

16 回帰係数の解釈

回帰分析は、得られたデータの見掛け上の関係进行分析するものであり、計算の手續きにそれ以上の意味はない。しかし、多くの場合分析によって推定したいのは、変数間の影響関係である。これについて検討するには、計算手續きだけでなく、データ取得のプロセスや、変数間に想定される影響構造についての知識が必要になる。

回帰分析の結果(回帰係数)が説明変数が被説明変数におよぼす影響とは解釈できない場合を考えてみる。

1. 何らかの観測されない要因があり、これが説明変数と被説明変数の両者に影響をおよぼしている場合。

最初にあげた自動車の停止距離のデータにおいて、実験を低速度から高速度に順に行ったと仮定する。このとき初め路面が濡れており、時間とともに乾燥してきたとする。もし、路面の状態が観測されていなければ、速度の停止距離に及ぼす影響は、過小評価されるだろう。

2. 分析の対象データが、説明変数と被説明変数の両者の影響を受ける変数によって層別または選別されたものである場合。

大学入試の合格者について、家計収入と試験の成績の関係を調べるとする。大学の受験の決定および学力成績に家計収入が影響をおよぼしている可能性は大きい。合格者は、学力と家計収入の両者について選別されたデータとみなせる。合格者について家計収入と学力の関係を調べると、同世代全体における関係とはかなり異った傾向をもつ可能性が大きい。

実験においては、条件の割り当ては実験者によって制御され、この割り当ての方法が適切なら他の要因の影響を受ける可能性は小さい。先の自動車の停止距離の例で考えると、速度は停止距離に影響を及ぼす要因と関係を持つことがないように、無作為化または事前の計画を用いて設定することが可能である。特に無作為化を用いる場合には、たとえ未知の要因が存在したとしても、それと説明変数の関係が断ち切られるので、回帰係数の推定に偏りは生じない。この場合には、回帰係数を説明変数から被説明変数への影響を表すのものとして解釈することは妥当である。

しかし、社会調査などの実験者による介入を伴わない観察データにおいては、説明変数と被説明変数の両者に影響を及ぼしている変数は多く考えうる。これらのうちの幾つかは測定されたデータの内にあり、説明変数に加えることにより影響を除くことが可能かも知れない。しかしながら、考

慮すべき変数をすべて列挙しているという保証が存在しない。分析者が、主要な変数をすべて測定したと考えたとしても、その他に説明変数と被説明変数の両者に影響を及ぼしている変数を見落している可能性を否定できない。このため、観察データへ回帰分析を適用して得られる結果に影響関係の推定とみなすことは、変数間の影響構造についての明確な知識が予めなければ、困難になる。

17 非線形関係の推定

通常回帰分析のモデルでは、 $E(Y|x)$ が x についての 1 次式であることを仮定している。しかし、非線形関係を求める必要も多い。非線形関係を推定するためには、回帰分析のモデルを拡張する必要がある。このための方法としては、パラメトリックな方法とノンパラメトリックな方法との 2 つがある。

パラメトリックな方法とは、少数のパラメータで表される関数の集合を考え、この中で最適な関数関係を与えるパラメータを推定するものである。よく利用されるものに多項式回帰がある。これは

$$E(Y|x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_p x^p \quad (102)$$

のように、 Y の期待値を説明変数の多項式で表すものである。この場合は、通常の重回帰分析と同じ計算で解を求めることができる。より、複雑な関数関係を仮定する場合には、パラメータの推定のために非線形の数値最適化を用いる必要がある場合もある。一方、ノンパラメトリック回帰は、データに適合する関数を滑らかさについての制約を課しつつ推定しようとするものである。一般的には、推定される関数は少数のパラメータであらわすことができない。

ここでは、パラメトリックなモデルとして多項式より利用しやすい方法であるスプライン回帰と、ノンパラメトリック回帰手法の一つである lowess とについて紹介する。

17.1 スプライン回帰

多項式は次数をあげれば複雑な関数を表現できるが、実際に用いてみるとかなり扱いづらい特性をもっている。これは、一部の箇所での変更がそこから離れた箇所での関数の値に影響を与えるためである。そこで、より扱いやすい関数の組として工夫されたのがスプライン (spline) 関数である。スプライン関数は、滑らかに接続された区分的な多項式である。一般には、各区間内で 3 次式であらわさる関数を用いることが多く、これは 3 次スプライン (Cubic spline) と呼ばれる。各区間内で関数が k 次式であらわされる場合には、区間の端点で関数は $k - 1$ 次の微分までが連続になるように、つなぎあわされる。多項式が接続される箇所は節点とよばれる。

接点が z_1, z_2, \dots, z_m である k 次スプライン関数は

$$\beta_0 + \beta_1 x + \cdots + \beta_k x^k + \sum_{j=1}^m \theta_j (x - z_j)_+^k \quad (103)$$

とあらわされる。ここで下つきの添字 $+$ は、関数が正の時に式通りの値であり、負のときにはゼロであることを示す。

スプライン回帰は、上の式に現れる各項 $(x - z_j)_+^k$ を、各々新たな変数とみなして回帰を行う。計算は、多項式回帰と同じく線形であるので通常回帰と同じ手順で可能である。ただし、実際の計算においては上の形式では数値的に悪条件となることが多いので、理論的に同等でより扱いやすい B スプラインと呼ばれる関数の組を用いることが多い。

図に B スプライン関数の組の例を示す。左図では、節点は左右の両端と中央の 3 箇所にある。ただし、両端においては、 $k + 1 = 4$ 個の節点が重複している。節点が重複していると、その箇所での滑らかさの制約が弱められる。

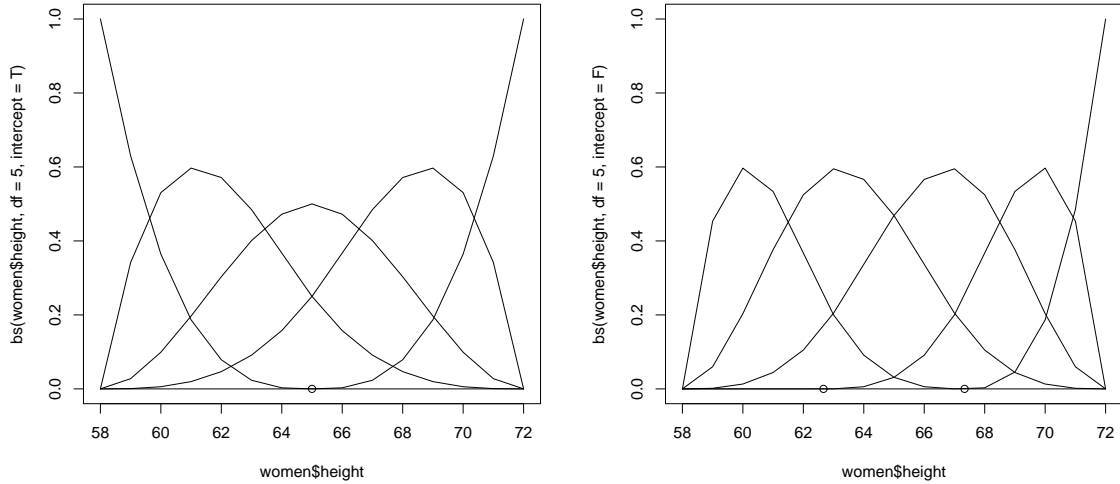


図 15: B スプライン関数の例 (右図は左端で関数がゼロと制約した場合)

17.2 局所回帰 (locally weighted regression; lowess)

この手法は W.S.Cleveland によって開発され、いくつかの統計ソフトウェアに散布図の平滑化曲線を描く方法として採り入れられている。lowess の計算手順はつぎのようなものである。

1. $\Delta_i(x) = |x_i - x|$ を大きさの順に並べかえて $\Delta_{(i)}(x)$ とする。
2. つぎの関数 (Tricube weight function) によって局所的な重みづけを行う。

$$T(u) = \begin{cases} (1 - |u|^3)^3 & \text{for } |u| < 1 \\ 0 & \text{otherwise.} \end{cases} \quad (104)$$

条件付き期待値を、つぎの局所的な重みづけを用いた回帰によって求める。

$$w_i(x) = T\left(\frac{\Delta_i(x)}{\Delta_{(q)}(x)}\right)$$

最小化すべき基準と予測式はつぎの 2 つの式になる。

$$\sum_{i=1}^N w_i(x)(y_i - a - bx_i)^2$$

$$\hat{g}(x) = \hat{a} + \hat{b}x$$

3. 上述の局所的な回帰による推定をつなぎあわせて回帰曲線を求める。

さらに Cleveland は、外れ値に対して頑健な回帰を求めるために Bisquare とよばれるつぎの関数を用いて重みづけの修正 r_i を課し、最小化する基準を

$$\sum_{i=1}^N w_i(x) r_i (y_i - a - bx_i)^2 .$$

とすることを提案している。ここで r_i は、つぎのように求めている。まず、残差の推定値 e_i のメディアンを s とする。Bisquare 関数はつぎのように定義される。

$$B(u) = \begin{cases} (1 - |u|^2)^2 & \text{for } |u| < 1 \\ 0 & \text{otherwise.} \end{cases} \quad (105)$$

ここで、 $r_i = B(e_i/(6s))$ と定義する。これにより大きな絶対値を持つ残差の影響が弱められる。また、この基準を最小化するには、反復計算が必要になる。

18 参考文献

Draper, N.R. & Smith, H. (1998). *Applied Regression Analysis, 3rd ed.*, Wiley. (回帰分析全般についてのわかりやすい解説。)

Chambers, J.M. & Hastie, T.J. eds. (1992) 柴田里程訳 (1994) *S と統計モデル: データ科学の新しい波*, 共立出版

Cleveland, W.S. (1993). *Visualizing Data*, Summit, New Jersey: Hobart Press. (loess の開発者による、平滑化を用いたデータ解析例)

Agresti, A. (1996). *An Introduction to Categorical Data Analysis*, Wiley. (ロジスティック回帰と対数線形モデルの解説, 邦訳あり)

Hastie, T.J. & Tibshirani, R.J. (1990). *Generalized Additive Models*, Chapman & Hall. (変数毎に非線形の変換を求める一般化加法モデルの教科書)

<http://www.r-project.org/Splplus> 類似のフリーソフト R のホームページ。英語圏の計算機統計の一流の研究者が開発に関わっている。回帰モデルを含め、lowess やスプライン回帰、一般化加法モデルなどが使える。

19 Splus による回帰分析

Splus を起動する (Unix のコマンドモードで、'Splus -e') と入力する。
以下は実行例

```
# データの読み込みとデータフレームの作成 (ヘッダなしの場合)
```

```
tb61 <- read.table("tb61.txt",row.names=NULL,  
  col.names=c("y","x1","x2","x3"))
```

```
# 回帰分析の実行
```

```
tb61.lm <- lm( y~ x1+x2+x3, data=tb61)
```

```
#読み込んだデータの表示
```

```
> tb61  
  y x1  x2 x3  
1 33 33 100 14  
2 40 47  92 15  
3 37 49 135 18  
4 27 35 144 12  
5 30 46 140 15  
6 43 52 101 15  
7 34 62  95 14  
8 48 23 101 17  
9 30 32  98 15  
10 38 42 105 14  
11 50 31 108 17  
12 51 61  85 19  
13 30 63 130 19  
14 36 40 127 20  
15 41 50 109 15  
16 42 64 107 16  
17 46 56 117 18  
18 24 61 100 13  
19 35 48 118 18  
20 37 28 102 14
```

```
# 分析結果の概要の表示
```

```
>summary(tb61.lm)
```

```
Call: lm(formula = y ~ x1 + x2 + x3, data = tb61)
```

```
Residuals:
```

```
  Min      1Q  Median      3Q      Max  
-10.34 -4.82  0.9897  3.855  7.909
```

Coefficients:

	Value	Std. Error	t value	Pr(> t)
(Intercept)	36.9601	13.0713	2.8276	0.0121
x1	-0.1137	0.1093	-1.0398	0.3139
x2	-0.2280	0.0833	-2.7377	0.0146
x3	1.9577	0.6349	3.0835	0.0071

Residual standard error: 5.956 on 16 degrees of freedom

Multiple R-Squared: 0.4805

F-statistic: 4.934 on 3 and 16 degrees of freedom, the p-value is 0.01297

Correlation of Coefficients:

(Intercept)	x1	x2	
x1	-0.2818		
x2	-0.6112	0.0735	
x3	-0.5823	-0.2021	-0.1587

```
> anova(tb61.lm)
```

Analysis of Variance Table

Response: y

SAS の TYPE I SS に相当

Terms added sequentially (first to last)

	Df	Sum of Sq	Mean Sq	F Value	Pr(F)
x1	1	3.8179	3.8179	0.107609	0.7471365
x2	1	183.9796	183.9796	5.185601	0.0368595
x3	1	337.3397	337.3397	9.508171	0.0071213
Residuals	16	567.6629	35.4789		

20 データの解説と課題

練習問題:

Dobson 表 6.1: ファイル <http://www.rd.dnc.ac.jp/~otsu/Toyo2005/DobsonTb61.txt>

インシュリン依存性糖尿病の 20 名の男性についての、(1)Y:炭水化物、(2)X₁:年齢、(3)X₂:体重と(4)X₃:蛋白質のデータ。20 名の被験者は 6 ヶ月間高炭水化物の食事をとっている。変数 Y は、総カロリーのうち炭水化物から得られたカロリーの比率をパーセントであらわしたものである。X₁ は被験者の年齢。X₂ は、理想体重からの相対的肥満度。X₃ は蛋白質からのカロリーの比率である。

1. 変数 Y を説明変数 X₁, X₂, X₃ の各々に対してプロットし (散布図; scatter plot を描き)、Y と各説明変数との関係を検討しなさい。

2. モデル

$$E(Y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}$$

をあてはめ、残差を調べてモデルと回帰分析の仮定の妥当性について検討しなさい。

3. つぎのモデルをあてはめ、上のモデルとの比較を行ない、帰無仮説 (null hypothesis) $H_0: \beta_1 = 0$ について検討しなさい。

$$E(Y_i) = \beta_0 + \beta_2 x_{i2} + \beta_3 x_{i3}$$

Dobson 表 6.8: ファイル (準備中) 血清中のコレステロール濃度が年齢とともに増加することはよく知られている。しかし、体重との関係は明らかではない。ファイル中の変数は (1) 血清コレステロール (ミリモル/l)、(2) 年齢、(3) 体重指標 (体重を身長²の2乗で割ったもの kg/m^2) である。体重指標が血清コレステロールに影響しているか否かを検討せよ。

課題

自分の関心のあるデータについて、重回帰分析を SAS、Splus などを用いて行い (各自の使いやすいものでよい) 分析結果を報告せよ。

専攻、学年、学生番号、氏名をレポート用紙に明記すること。また、データの出典を明らかにすること。自分で集めたデータの場合は、実験方法、調査方法を簡潔に解説する。データの特徴 (1 変量のヒストグラムや箱ひげ図、散布図などを含んでよい)、問題の背景、分析の方針、モデルの適合度と診断結果、結果の解釈などについて報告せよ。レポートは短いものであってもかまわないが、読んでよく理解できるものである必要がある。

レポートの項目だての例をあげる。1. はじめに (何を調べようとしているか。問題の背景の解説。) 2. データ (どのようにして得られたものか) 3. 分析 (重回帰分析などをどのように適用したか。その結果はどのようになったか。) 4. 結論。

21 回帰分析への補足 (1) AIC によるモデル選択

回帰分析の数値例 (人工データ) を用いて AIC の説明を行なう。

まず、つぎの式に従うデータを正規乱数を用いて人工的に生成する。

$$\mu_i = 1 + 0.5x_i + 0.5x_i^2, (i = 1, \dots, 101; x_i = -4 \text{ から } 0.08 \text{ 刻みで } +4 \text{ まで})$$

$$y_i = \mu_i + \varepsilon_i$$

ここで ε_i は平均ゼロ、標準偏差 4 の正規乱数である。

幾つか説明変数を変更して分析を行なった結果、AIC はつぎのようになった。

モデル	$\sum(\hat{y}_i - y_i)^2$	AIC	$\sum(\hat{y}_i - \mu_i)^2$
$1, x$	1892.6	768.1	729.1
$1, x, x^2$	1584.2	752.2	178.8
$1, x, x^2, x^3$	1582.7	754.1	180.4
$1, x, x^2, x^3, x^4$	1555.8	754.3	207.2
$1, x, x^2, x^3, x^4, x^5$	1553.4	756.2	209.6

上の各計算によって得られた推定値 \hat{y} を図 16 に示す。図中 1、2 などとあるのは、各々 1 次式による推定値、2 次式による推定値などを示す。曲線は μ の値であり、記号*は観測データの値である。この例では 2 次式から 5 次式による回帰にさして大きな違いはないが、AIC が最小である 2

次式による回帰モデルが「真の平均からの残差 2 乗和」 $\sum(\hat{y}_i - \mu_i)^2$ の値を最小にしていることが分かる。

SAS-GLM による上記データの分析

15:08 Tuesday, June 12, 2001

General Linear Models Procedure

Number of observations in data set = 101

General Linear Models Procedure

Dependent Variable: Y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	733.73497646	146.74699529	8.97	0.0001
Error	95	1553.43165674	16.35191218		
Corrected Total	100	2287.16663320			

R-Square	C.V.	Root MSE	Y Mean
0.320805	138.6342	4.0437498	2.9168496

Source	DF	Type I SS	Mean Square	F Value	Pr > F
X	1	394.54800693	394.54800693	24.13	0.0001
X2	1	308.43675236	308.43675236	18.86	0.0001
X3	1	1.52446716	1.52446716	0.09	0.7608
X4	1	26.85284466	26.85284466	1.64	0.2031
X5	1	2.37290537	2.37290537	0.15	0.7041

Source	DF	Type III SS	Mean Square	F Value	Pr > F
X	1	6.89243918	6.89243918	0.42	0.5178
X2	1	99.60866737	99.60866737	6.09	0.0154
X3	1	3.15430397	3.15430397	0.19	0.6615
X4	1	26.85284466	26.85284466	1.64	0.2031
X5	1	2.37290537	2.37290537	0.15	0.7041

Parameter	Estimate	T for H0: Parameter=0	Pr > T	Std Error of Estimate
INTERCEPT	0.3824110041	0.51	0.6135	0.75461302
X	0.4904424785	0.65	0.5178	0.75541548
X2	0.7149826718	2.47	0.0154	0.28968848
X3	0.0799270577	0.44	0.6615	0.18198112
X4	-.0254417681	-1.28	0.2031	0.01985347
X5	-.0037298503	-0.38	0.7041	0.00979119

General Linear Models Procedure

Number of observations in data set = 101

General Linear Models Procedure

Dependent Variable: Y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	702.98475928	351.49237964	21.74	0.0001
Error	98	1584.18187392	16.16512116		
Corrected Total	100	2287.16663320			

R-Square	C.V.	Root MSE	Y Mean
0.307361	137.8401	4.0205872	2.9168496

Source	DF	Type I SS	Mean Square	F Value	Pr > F
X	1	394.54800693	394.54800693	24.41	0.0001
X2	1	308.43675236	308.43675236	19.08	0.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
X	1	394.54800693	394.54800693	24.41	0.0001
X2	1	308.43675236	308.43675236	19.08	0.0001

Parameter	Estimate	T for H0: Parameter=0	Pr > T	Std Error of Estimate
INTERCEPT	0.9627744522	1.60	0.1119	0.60014410
X	0.8474025961	4.94	0.0001	0.17152576
X2	0.3592049865	4.37	0.0001	0.08223350

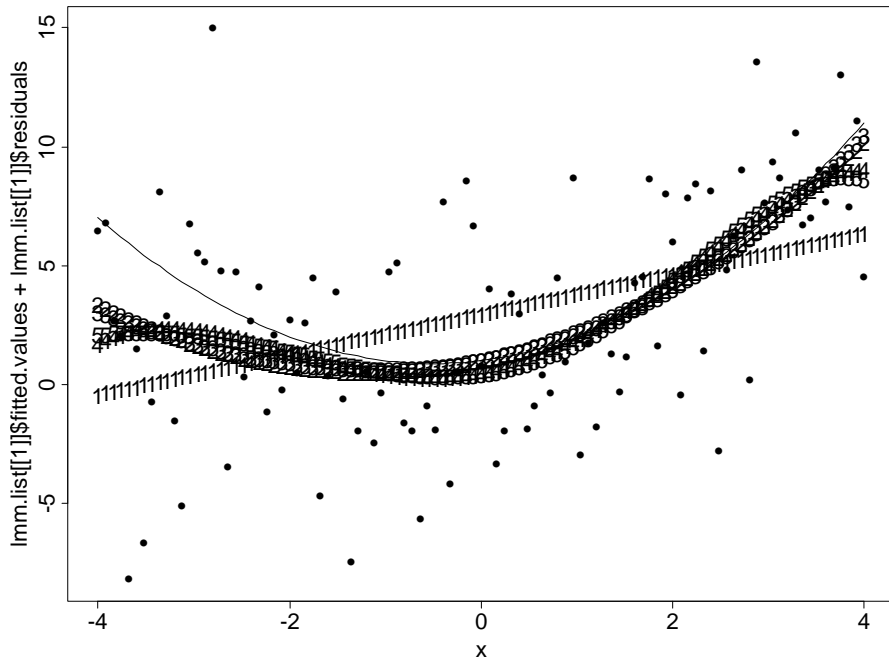


図 16: 多項式回帰の結果

22 分散分析

22.1 1元配置の分散分析

初等統計の教科書では、回帰分析と分散分析は別の手法として、取り扱われることが多いが、 X がダミー変数であることを仮定すると、共通のモデルとして取り扱うことができる。

ここでは、分散分析 (Analysis of Variance, ANOVA) のうち、最も簡単な場合である 1 元配置 (one-way design) について説明する。回帰分析において説明変数は連続な値をもつもの (身長、血圧など) であったが、分散分析では離散的な区分が説明変数である。ここで、3 種類の教育方法によって、生徒の成績に違いがあるか否かを検討することを想定する。3 つのクラス $j = 1, 2, 3$ がそれぞれ異なる教育方法の対象となっている。クラス j の生徒の成績を y_{ij} , ($i = 1, \dots, N_j; j = 1, 2, 3$) とする。また $N_1 + N_2 + N_3 = N$ とする。分散分析のモデルは、各クラス毎にデータが独立に同一の正規分布に従っていることを仮定する。つまり

$$y_{ij} = \mu_j + \varepsilon_{ij}, \quad (i = 1, \dots, N_j; j = 1, 2, 3) \quad (106)$$

であり、 $\varepsilon_{ij} \sim N(0, \sigma^2)$ である。検討すべき課題は μ_j , ($j = 1, 2, 3$) が互いに同一か否かである。

ここで、つぎのような行列を X として想定する。

$$X = \begin{bmatrix} 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \end{bmatrix} \quad (107)$$

最初の列のみが 1 である行は N_1 、1 列目と 2 列目が 1 である行は N_2 、また 1 列目と 3 列目が 1 である行は N_3 回繰り返されているものとする。この X を用い、 $\beta = (\mu_1, \mu_2 - \mu_1, \mu_3 - \mu_1)^T$ と置くことにより、分散分析のモデル (106) は、

$$y = X\beta + \varepsilon \quad (108)$$

と表わされる。ここで、 y は $y_{1,1}, \dots, y_{N_1,1}, y_{1,2}, \dots, y_{N_2,2}, y_{1,3}, \dots, y_{N_3,3}$ と y_{ij} の値を縦につないだベクトルである。平均に差があるか否かは、仮説 $H_0: \beta_2 = \mu_2 - \mu_1 = 0$ および $\beta_3 = \mu_3 - \mu_1 = 0$ を検定すればよい。

上に示した (107) の第 2 列と第 3 列目は、それぞれの行に対応する y_{ij} の値が、第 2 群または第 3 群に所属する場合に各々 1 となり、それ以外では 0 となる。このように、離散的なカテゴリーへの所属を 0 - 1 で表す変数をダミー変数 (dummy variable) と呼ぶ。

ここで、1 元配置の分散分析モデルについて、(107) とは異なる表現が可能であることを示す。(107) においては、 X の第 1 列目は定数項 (すべて 1) を表し、第 2 列目と第 3 列目は各々第 2 グループ ($j = 2$) と第 3 グループ ($j = 3$) への所属を表すダミー変数であった。このようにして構成される X を X_a とし対応する回帰係数ベクトルを $\beta_a = (\beta_{a0}, \beta_{a2}, \beta_{a3})^T$ とする。一般的に、分散

分析などの離散的な説明変数を含む回帰モデル(一般線形モデル)において、説明変数およびそれらのダミー変数から構成される X のことをデザイン行列 (design matrix) と呼ぶ。ここで示そうとしていることは、実質上同等なモデルが複数のデザイン行列によって表すことが可能なことである。つぎのような 2 つのデザイン行列を想定しよう。

1. X の第 1 列は第 1 グループへの所属を表すダミー変数とし、第 2 列、第 3 列を X_a と同じものとし、これを X_b と表す。回帰係数ベクトルは $\beta_b = (\beta_{b1}, \beta_{b2}, \beta_{b3})^T$ である。
2. X の第 1 列は定数項、第 2 列は第 1 グループへの所属を表すダミー変数、第 3 列は第 2 グループへの所属を表すダミー変数、さらに第 4 列は第 3 グループへの所属を表すダミー変数とする。これを X_c とする。 X_c は X_b の左側に定数ベクトルを並べたもの、つまり $(1|X_b)$ である。但し、 X_c に対応する回帰係数のベクトルを $\beta_c = (\beta_{c0}, \beta_{c1}, \beta_{c2}, \beta_{c3})^T$ として、次ぎの制約を満たすものとする。

$$\beta_{c1} + \beta_{c2} + \beta_{c3} = 0 \quad (109)$$

ここで、 X_a によるモデルの表現は、 X_c をデザイン行列とし、(109) の代わりに $\beta_{c1} = 0$ と制約を置いたものである。

これら 3 つのモデルはいずれも同等なものである。つまり、観測値 y に対して、推定値 $\hat{y}_a = X_a \hat{\beta}_a$ が得られるとすると、つねに $\hat{y}_a = X_b \hat{\beta}_b = X_c \hat{\beta}_c$ となる $\hat{\beta}_b$ と $\hat{\beta}_c$ が一通りに決まる。これら 3 つのモデルにおいて、推定される β の値は異なるが、推定値と残差 2 乗和 $\sum (y_i - \hat{y}_i)^2$ は同一である。ソフトウェアによっては、互いに異なるデザイン行列の表現を利用している場合がある。

22.2 サンプルが同数の場合

ここで、各群においてサンプルが同数の場合に計算式がどのようなものになるかを検討する。ここでは、上に示したデザイン行列 X_a を仮定する。各群のサンプル数は $N_1 = N_2 = N_3 = n$ であり、 $N = N_1 + N_2 + N_3$ とする。また、第 1 群に含まれる観測値の和と平均をそれぞれ $y_{.1}, \bar{y}_1$ とする。第 2 群、第 3 群についても同様である。また全ての観測値の和と平均を $y_{..}$ および \bar{y} とする。

まず、回帰係数について考えると、 $\hat{\beta} = (X^T X)^{-1} X^T y$ である。ここで、

$$X^T y = \begin{pmatrix} y_{..} \\ y_{.2} \\ y_{.3} \end{pmatrix}, \quad X^T X = \begin{pmatrix} N & n & n \\ n & n & 0 \\ n & 0 & n \end{pmatrix} \quad (110)$$

である。後者の逆行列を計算すると、

$$(X^T X)^{-1} = \frac{1}{n} \begin{pmatrix} 1 & -1 & -1 \\ -1 & (N-n)/n & 1 \\ -1 & 1 & (N-n)/n \end{pmatrix} \quad (111)$$

であり、これより、

$$\hat{\beta} = \begin{pmatrix} y_{.1}/n \\ (y_{.2} - y_{.1})/n \\ (y_{.3} - y_{.1})/n \end{pmatrix} \quad (112)$$

となる。推定値 \hat{y}_{ij} は、 $X \hat{\beta}$ であるので、所属する群 j によって値が定まり、第 1 群については、 $y_{.1}/n = \bar{y}_1$ 、第 2 群については $y_{.2}/n = \bar{y}_2$ 、また第 3 群については $y_{.3}/n = \bar{y}_3$ となる。

各群のサンプル数が異なる場合でも、同様に計算することにより、第 j 群についての予測値が $y_{.j}/N_j = \bar{y}_j$ となることが分かる。

22.3 帰無仮説の表現

回帰分析においては、「幾つかの回帰係数がゼロであること」を帰無仮説として取り上げた。分散分析のように、ダミー変数を説明変数として持つモデルにおいては、より多様な形式の帰無仮説を取り扱う必要が多い。次ぎの ANOVA モデルを考える。

$$y_{ij} = \mu_j + \varepsilon_{ij}, \quad (i = 1, \dots, N_j; j = 1, 2, 3) \quad (113)$$

ここで、通常もっともよく利用される帰無仮説は、 $\mu_1 = \mu_2 = \mu_3$ というものである。デザイン行列 (13) が設定されているとすると、回帰係数は $\beta_1 = \mu_1, \beta_2 = \mu_2 - \mu_1, \beta_3 = \mu_3 - \mu_1$ であり、上の帰無仮説は $H_0 : \beta_2 = \beta_3 = 0$ として表される。しかし、一般に検定の対象となる帰無仮説は、上のタイプのものには限らない。例えば、仮説 $\mu_3 - \mu_2 = 0$ は、上のデザイン行列では、 $\beta_3 - \beta_2 = 0$ と表現され、ある i について $\beta_i = 0$ という形式では表現されない。

上にのべた幾つかの帰無仮説はいずれも、回帰係数の (1 つまたは複数の) 1 次式として表される。例えば、 $\beta_2 = \beta_3 = 0$ という仮説は

$$Q^T = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad (114)$$

とおくと

$$Q^T \beta = 0 \quad (115)$$

と表記される。一般的に、 $Q^T \beta = \xi$ のような回帰係数への線形の制約式で表される帰無仮説を、線形仮説 (linear hypothesis) と呼ぶ。同じデザイン行列のもとで、仮説 $\beta_2 - \beta_3 = 0$ は

$$Q^T = \begin{pmatrix} 0 & 1 & -1 \end{pmatrix} \quad (116)$$

として、 $Q^T \beta = 0$ として表される。ここで、デザイン行列 $X_{N \times p}$ の各列ベクトルが 1 次独立であれば、 Q^T がどんなものであろうと $Q^T \beta$ の値は、データから一意的に推定できる。ところが、 X の列ベクトルが 1 次従属の場合、つまり $\text{rank} X < p$ の場合には、一つのデータ y について $Q^T \beta$ の値を一意的に推定できない場合が生じ得る。これを避けるためには、 Q^T の各行ベクトルが X の行ベクトルの 1 次結合として表される必要がある。この条件は、

$$\text{rank} \begin{pmatrix} X \\ Q^T \end{pmatrix} = \text{rank} X \quad (117)$$

と表現される。これはまた、 $\text{Ker} X \subseteq \text{Ker} Q^T$ とも表される。この条件が成り立つことを線形仮説が推定可能 (estimable) であるという。ある線形仮説が推定可能でないということは、指定されたデザイン行列がその仮説を検定するために適切ではないことを意味する。

ここでは、当面 X のランクは p であると仮定する。線形仮説 $H_0 : Q^T \beta = \xi$ を検定するためには、制約下での推定値と制約無しの下での推定値を求め、各々について残差の 2 乗和を求める。帰無仮説の制約下での Y の推定値を \hat{y}_0 とし、また制約無しでの推定値を \hat{y}_1 とする。 Q のランクを k とすると、もし帰無仮説が正しいのなら、

$$\|y - \hat{y}_0\|^2 - \|y - \hat{y}_1\|^2 \sim \sigma^2 \chi^2(k) \quad (118)$$

である。またこれとは独立な分布として

$$\|y - \hat{y}_1\|^2 \sim \sigma^2 \chi^2(N - p) \quad (119)$$

が得られる。回帰分析における変数選択の場合と同様に、次ぎの F 値が帰無仮説のもとで自由度 $(k, N - p)$ の F 分布に従うことを用いて検定を行なえる。

$$F = \frac{\|\mathbf{y} - \hat{\mathbf{y}}_0\|^2 - \|\mathbf{y} - \hat{\mathbf{y}}_1\|^2}{\|\mathbf{y} - \hat{\mathbf{y}}_1\|^2} \times \frac{N - p}{k} \quad (120)$$

22.4 2元配置の分散分析

分散分析においては、一つの離散値をとる説明変数のことを要因 (または因子)(factor) と呼び、要因に指定される各々の値を水準 (level) と呼ぶ。前節では実験が1要因の場合を考えたが、分散分析がその本領を発揮するのは、むしろ複数の要因が存在する場合である。

第1番目の要因 A が水準 A_1, A_2, \dots, A_J を持ち、第2番目の要因 B が水準 B_1, \dots, B_K を持つものとする。被説明変数は Y であり、この観測値は $y_{i,j,k}, i = 1, \dots, N_{jk}$ とする。ここで、添字 i, j, k は、実験条件 (A_j, B_k) における i 番目の測定値であることを表す。これらの要因を考慮して、被説明変数 y を説明する構造をどのように考えたら良いであろうか。一番単純なモデルは、どちらの要因も y には影響を与えず定数項のみを説明変数とするモデルである。これはつぎの式で表される。

$$Y_{i,j,k} = \mu_0 + \varepsilon_{i,j,k}, (i = 1, \dots, N_{j,k}; j = 1, \dots, J; k = 1, \dots, K) \quad (121)$$

この次ぎに考えるべきモデルは、要因 A のみ、または要因 B のみによって説明されるモデルである。これは

$$Y_{i,j,k} = \mu_0 + \mu_{A_j} + \varepsilon_{i,j,k} \quad (122)$$

$$Y_{i,j,k} = \mu_0 + \mu_{B_k} + \varepsilon_{i,j,k} \quad (123)$$

として表される。ここで、 $\sum_j \mu_{A_j} = 0$ 、および $\sum_k \mu_{B_k} = 0$ の制約が与えられているものとする (これらの制約がないとパラメータの一意性が失われる)。もう少し複雑なモデルは、要因 A と要因 B の両者が関係するものであり、

$$Y_{i,j,k} = \mu_0 + \mu_{A_j} + \mu_{B_k} + \varepsilon_{i,j,k} \quad (124)$$

である。ここでも、 μ_{A_j} と μ_{B_k} には前と同様の制約を加える。ここまで述べて来た説明変数によって説明される成分を主効果 (main effect) と呼ぶ。主効果は一つの要因のみに関連するダミー変数で説明される成分である。

2元配置の場合、もっとも複雑なモデルは $J \times K$ の各セル毎に平均が異なるというものである。単純な記述を用いると

$$Y_{i,j,k} = \mu_{A_j B_k} + \varepsilon_{i,j,k} \quad (125)$$

である。しかし、多くの場合は上に示したモデルとの比較を行なうので、そのために利用しやすい形を用いて、

$$Y_{i,j,k} = \mu_0 + \mu_{A_j} + \mu_{B_k} + \mu_{A_j B_k} + \varepsilon_{i,j,k} \quad (126)$$

と表現する。これは (125) と同等のモデルである。ただし、主効果には前と同様の制約をおき、また $\mu_{A_j B_k}$ の部分 (交互作用項 (interaction term)) には、つぎのような制約を置く。

$$\sum_{j=1}^J \mu_{A_j B_k} = 0, (k = 1, \dots, K) \quad (127)$$

$$\sum_{k=1}^K \mu_{A_j B_k} = 0, (j = 1, \dots, J) \quad (128)$$

このような単純なものから複雑なものに至るモデルの系列を用いて、データの分析を行なうことにより、どの要因が Y に影響をどの程度与えているかの検討を行なうことができる。

22.5 分散分析表

分散分析の検定は、現在ではコンピュータを用いて容易に計算が可能であるが、かつては手計算で結果を導いていた。バランスのとれた（サンプル数が各水準で均等な）実験計画の例が強調されるのは、それが水準間の差を検出するのに有利であるのと、手計算が容易であるためである。

分散分析での F 検定は、水準間の分散と水準内の分散との比を用いて計算するが、これらをわかりやすくまとめた表を分散分析表 (ANOVA table) と呼ぶ。これらの表の構成要素は、バランスのとれた実験計画においては、2元配置の場合にも簡単な計算で求まるが、それ以外の場合には一般的には連立方程式を解かなければならない。1元配置の分散分析について、つぎのような公式がなりたつ。ここで y_{ij} は第 j 群に含まれる i 番目のデータである。また \bar{y}_j は第 j 群の n_j 個のデータの平均、 \bar{y} はデータ全体の平均である。また $\sum n_j = N$ とする。

$$\sum_{j=1}^J \sum_{i=1}^{n_j} (y_{ij} - \bar{y})^2 = \sum_{j=1}^J \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2 + \sum_{j=1}^J n_j (\bar{y}_j - \bar{y})^2 \quad (129)$$

帰無仮説の仮定のもとで、上の式の左辺は自由度 $(N - 1)$ の χ^2 分布の σ^2 倍に従う。また、右辺第1項は、自由度 $(N - J)$ の χ^2 分布の σ^2 倍に従い、さらに第2項は、自由度 $(J - 1)$ の χ^2 分布の σ^2 倍に従う。

1元配置の分散分析表は表15のようになる。

表 15: 1元配置の分散分析表

	自由度 (df)	SS	$MS = SS/df$
級間残差 2 乗和	$J - 1$	$\sum_{j=1}^J n_j (\bar{y}_j - \bar{y})^2$	
級内残差 2 乗和	$N - J$	$\sum_{j=1}^J \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2$	
全体の残差 2 乗和	$N - 1$	$\sum_{j=1}^J \sum_{i=1}^{n_j} (y_{ij} - \bar{y})^2$	

2元配置の場合は、一般的には簡単な式にはならないが、各セルにおけるデータの個数が等しい場合には、つぎのような分散分析表 (表16) が得られる。ただし、要因 $A_j \times B_k$ に対応するデータは y_{ijk} , $i = 1, \dots, n_{jk}$ とする。ここで $n_{jk} = n$ (一定) を仮定しているので、 $n_{j\cdot} = N/J = Kn$ であり、また $n_{\cdot k} = N/K$ であることに注意。

22.6 2元配置分散分析の意義

2つの要因 A と B がある実験の場合、 A の効果を測るには A についての1元配置を考えれば十分であるように思われる。しかし、これではつぎの2つの問題が生じる可能性がある。

ひとつは、 A の各水準に対応するサンプルのなかに、かなり異質なものが含まれることである。つまり、 A_1 の水準に対応するサンプルの中には、 A_1B_1, A_1B_2, \dots など要因 B について水準の異なるものが含まれる。もし、要因 B の効果が大きいのなら、 A の一つの水準の中のサンプルの分散

表 16: 2 元配置の分散分析表 (同数サンプル)
 $n_{jk} \equiv n$ (各セル同数の場合)、総数 $N = J \times K \times n$

	自由度 (df)	SS	$MS = SS/df$
要因 A	$J - 1$	$\sum_{j=1}^J Kn(\bar{y}_j - \bar{y})^2$	
要因 B	$K - 1$	$\sum_{k=1}^K Jn(\bar{y}_{.k} - \bar{y})^2$	
交互作用	$(J - 1)(K - 1)$	$\sum_{j=1}^J \sum_{k=1}^K n(\bar{y}_{jk} - \bar{y}_j - \bar{y}_{.k} + \bar{y})^2$	
残差	$N - J \times K$	全体からの引き算で求める	
全体	$N - 1$	$\sum_{j=1}^J \sum_{k=1}^K \sum_{i=1}^n (y_{ijk} - \bar{y})^2$	

も大きなものになる。分散分析における F 検定は、級間の分散を級内分散 (残差分散) で割った値を用いる。A の一つの水準内の分散が大きくなるということは、 F 値の分母の値が大きくなることを意味する。これは検定が有意になりづらい、つまり A による効果を検出しづらくなるということである。2 元配置の分散分析は要因 A と要因 B の効果を同時に測定する。これを要因 A の効果の測定という側面からみるならば、要因 B の効果を除いたものについて、A の効果があるかどうかを検討していることになる。この場合、A の各水準における級内分散は小さくなるので、1 元配置での分析を行なう場合より、敏感に差の存在を検出できる。

これと類似の状況が、対応のない場合の t 検定と対応のある場合の t 検定についても生じる。対応のない場合は、要因 A だけを考えることに相当し、対応のある場合の検定は、各対が B の水準に相当していると考えればよい。

もうひとつは、サンプル数の不均等から生じる問題である。もし、 A_1 においては B_2 に対応するサンプルが多く、 A_2 においては B_1 のサンプルが多いとすると、 A_1 における従属変数 Y の傾向と A_2 における傾向の違いは、本当に A_1 と A_2 の効果の違いのせいなのか、あるいは B_2 と B_1 の違いによるものなのか判定がつかない。

A と B の両者の効果を者の効果を含むモデルを考慮すると、A と B の効果を各々測定することができる。もちろん、この場合サンプル数がバランスしている場合に比べると推定の精度は落ちる。デザイン行列を X とし、第 1 列を定数項 (要素は全て 1)、第 2 列は A_1 に対応するダミー変数、第 3 列は B_1 に対応するダミー変数とするとつぎのような式が得られる。

$$XX^T = \begin{pmatrix} N & n_{.1} & n_{.1} \\ n_{.1} & n_{11} & n_{11} \\ n_{.1} & n_{11} & n_{.1} \end{pmatrix} \quad (130)$$

ここで、 $C = (XX^T)^{-1}$ とすると、 $\text{Var}(\hat{\beta}) = \sigma^2 C$ である。

$(n_{11}, n_{12}, n_{21}, n_{22}) = (5, 5, 5, 5)$ とすると、 $\text{Var}(\hat{\beta}_2) = \text{Var}(\hat{\beta}_3) = \sigma^2/5$ であるが、 $(n_{11}, n_{12}, n_{21}, n_{22}) = (2, 8, 8, 2)$ とすると、 $\text{Var}(\hat{\beta}_2) = \text{Var}(\hat{\beta}_3) = 5\sigma^2/16$ である。

2 元の分散分析がしばしば利用されるもう一つのパターンは、2 つの要因の主効果が存在することは前提として、交互作用項が有意であるかどうかを検討するケースである。例えば、要因 A を刺激の種類とし、要因 B を教示、測定データは反応時間であるとする。刺激によって平均反応時間は異なり、また教示によっても異なることは分かっているものとする。ここで、教示を変更することにより、刺激への反応速度のパターンが変わるか否かを検討するとしてしよう。つまり、刺激の種

類により、反応が早くなる（傾向のある）教示が異なるかどうかを検討することである。このためには、主効果と交互作用項を含む分散分析モデルをあてはめ、交互作用項が有意であるかを検討すればよい。

このような利点があるとすれば、常に複数の要因をモデルに含めるべきであろうか。データが大量に利用できるならば、この方針は正しい。しかし多くの場合、利用可能なデータの量には制限があり、このため要因 A の差を検出するために他の要因をモデルに含めるべきか否かは、微妙な問題となることもある。モデルに含める要因を増やすことに伴う問題は、残差の自由度が減少することである。 A の効果を仮説検定するための F 値の分母は、残差を自由度で割ったものであり、 $\sigma^2 \chi_k^2 / k$ に従う。ここで k は自由度であるが、 k が小さいと推定される値の相対誤差は大きくなる。このため、不要な要因をモデルに含めると、却って検出力を減少させることになってしまう場合がある。また、残差の正規性からのズレが検定に及ぼす影響も大きくなる。これは、重回帰分析において説明変数をどのように選択すべきかということと同種の問題である。分散分析における F 検定は、このような問題に対応するための一つの方法である。つまり、要因 A の主効果を検出するために、要因 B の主効果が有意であればこれをモデルに取り込み、そうでなければ A のみをモデルに含める。このような方法によって、 A の効果を敏感に検出しようとするものである。

23 SAS と Splus による分散分析

以下のデータは Dobson の教科書の表 7.1 からとったものであり、3つの異なる条件下での作物の収量を表している。このデータを対象にして、実験区分を要因（説明変数）とする 1 元配置の分散分析を行なう。この実験においては要因数は 1 であり水準の数が 3 である。

表 17: Dobson 表 7.1 作物の収量

実験区分	データ									
1(対照群)	4.17	5.58	5.18	6.11	4.50	4.61	5.17	4.53	5.33	5.14
2(処理 A)	4.81	4.17	4.41	3.59	5.87	3.83	6.03	4.89	4.32	4.69
3(処理 B)	6.31	5.12	5.54	5.50	5.37	5.29	4.92	6.15	5.80	5.26

23.1 データの記入法

1 列目が要因、2 列目が反応値とする。SAS の入力法を工夫すれば色々な形式が可能であるが、このようなものが無難である。要因または説明変数が複数ある場合には、回帰分析の場合と同様に 1 行の中に並べる。変数の順番は input 文との対応がつけば、どのようであっても構わない。

```
1 4.17
1 5.58
1 5.18
....
2 4.81
```

2 4.17
....
3 6.31
3 5.12
3 5.54
3 5.50

23.2 プログラミング例

```
options linesize=80;

data tb71;
    infile 'tb7_1.dat';
    input tr resp;
proc print data=tb71;    % 確認のための表示
proc glm data=tb71;
    class tr;            % 重回帰の場合と違うのはここ。
    model resp = tr;
run;
```

注意事項

- 離散的な説明変数 (要因) については、GLM プロシジャのなかで、model ステートメントの前に、class ステートメントで宣言を行なう。
- 文字変数を入力する場合には、data ステップの input 文において、変数名のあとに\$マークを付ける。

23.3 例題データの所在

ディレクトリ

<http://www.rd.dnc.ac.jp/~otsu/lecture/cl14a0/Dobson> の中に、tb7_1.dat, tb7_4.dat, tb7_8.dat として、表 7.1、表 7.4、表 7.8 のデータがある。これらを自分のディレクトリにコピーして、それぞれ分析をおこなってみる。まず、最初にデータの形式を確認すること。SAS では表 7.1 については、上の例で実行できる。

表 7.4 は 2 要因 (3 × 2 水準) で、各条件について 2 件ずつデータが観測されている (人工データ)。第 1 番目と 2 番目の変数が要因を表す文字コードであり、3 番目の変数が観測値である。表 7.4 において、交互作用項を指定するには model y = a b a*b; の様に指定する。ここで a b は主効果を表し、a*b が交互作用項を表す。

表 7.8 はアチーブメントスコアのデータであり (出典 Winer,1971)、1 番目の変数 A が 3 水準の実験要因 (訓練法)、2 番目の変数 y がスコア (得点)、3 番目の変数 x は訓練開始前の測定された適性検査のスコアである。訓練法 A と適性検査スコア x を説明変数とし、y を被説明変数とするモデルを設定し分析を試みなさい。

23.4 SASによる分散分析(2元配置)

つぎのような2元配置のデータを考える(Dobson, 表7.4)。

2元配置のデータ(Dobson 表7.4)

要因 A	要因 B			
	B ₁		B ₂	
A ₁	6.8	6.6	5.3	6.1
A ₂	7.5	7.4	7.2	6.5
A ₃	7.8	9.1	8.8	9.1

このようなデータはつぎのようにしてファイルに記入する。

```
a1 b1 6.8
a1 b1 6.6
a1 b2 5.3
a1 b2 6.1
a2 b1 7.5
a2 b1 7.4
a2 b2 7.2
a2 b2 6.5
a3 b1 7.8
a3 b1 9.1
a3 b2 8.8
a3 b2 9.1
```

交互作用まで含めたモデルを当てはめる場合には、つぎのようなSASプログラムで分析すればよい。

```
options linesize=80;
data tb74;
    infile 'tb7_4.dat';
    input a$ b$ y;
proc print data=tb74;
proc glm data=tb74 ;
    class a b;
    model y = a b a*b /solution ss1 ss2 ss3;
run;
```

出力についての注意

1. この例ではType I のSSとType II, Type III のSSとがすべて同一になっているが、これは各セルのデータ数が同一でバランスしているためである。一般的には同じにはならない。
2. Type II のSSは、その要因を削除した場合の残差2乗和の増加量である。
3. 重回帰分析の場合には、Type III のSSとType II のSSは同一のものである。

4. 分散分析 (データの個数がバランスしない場合) には Type II の SS と Type III の SS とは必ずしも一致しない。
5. Type III の SS の意味は、マニュアルの解説によるといささか複雑であり、単に該当する要因を除いた場合の残差 2 乗和の減少分ではない。該当する要因に関わるより高次の交互作用項への影響も考慮した統計量となっている。SAS の GLM プロシジャでは、Type I、II、III、IV の 4 種の検定統計量が出力できるが、これらの用語が一般的にどれだけ用いられているものであるかは、良く分からない。
6. model ステートメントのオプション solution は、個別の水準に対応するパラメータの推定値を表示する指定である。

SAS による 2 元配置分散分析の出力例

General Linear Models Procedure
Class Level Information

Class	Levels	Values
A	3	a1 a2 a3
B	2	b1 b2

Number of observations in data set = 12

SAS システム

3

13:28 Thursday, May 27, 1999

General Linear Models Procedure

Dependent Variable: Y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	14.35000000	2.87000000	11.64	0.0048
Error	6	1.48000000	0.24666667		
Corrected Total	11	15.83000000			

R-Square	C.V.	Root MSE	Y Mean
0.906507	6.757217	0.4966555	7.3500000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
A	2	12.74000000	6.37000000	25.82	0.0011
B	1	0.40333333	0.40333333	1.64	0.2482
A*B	2	1.20666667	0.60333333	2.45	0.1672

Source	DF	Type II SS	Mean Square	F Value	Pr > F
A	2	12.74000000	6.37000000	25.82	0.0011
B	1	0.40333333	0.40333333	1.64	0.2482

Source	DF	Type III SS	Mean Square	F Value	Pr > F
A*B	2	1.20666667	0.60333333	2.45	0.1672
A	2	12.74000000	6.37000000	25.82	0.0011
B	1	0.40333333	0.40333333	1.64	0.2482
A*B	2	1.20666667	0.60333333	2.45	0.1672

Parameter	Estimate	T for H0: Parameter=0	Pr > T	Std Error of Estimate
INTERCEPT	8.950000000	B 25.48	0.0001	0.35118846
A a1	-3.250000000	B -6.54	0.0006	0.49665548
a2	-2.100000000	B -4.23	0.0055	0.49665548
a3	0.000000000	B .	.	.
B b1	-0.500000000	B -1.01	0.3529	0.49665548
b2	0.000000000	B .	.	.
A*B a1 b1	1.500000000	B 2.14	0.0766	0.70237692
a1 b2	0.000000000	B .	.	.
a2 b1	1.100000000	B 1.57	0.1684	0.70237692
a2 b2	0.000000000	B .	.	.
a3 b1	0.000000000	B .	.	.
a3 b2	0.000000000	B .	.	.

NOTE: The X'X matrix has been found to be singular and a generalized inverse was used to solve the normal equations. Estimates followed by the letter 'B' are biased, and are not unique estimators of the parameters.

23.5 Splus による分析

23.5.1 1元配置 (Dobson 表 7.1)

データの入力はずぎのようにおこなう。

```
> tb71 <- read.table("tb7_1.dat", col.names=c("condition","yield"),
  row.names=NULL)
> tb71$condition <- as.factor(tb71$condition)
```

分散分析の実行には関数 `aov` を用いる。結果 (分散分析表) を表示するには、関数 `anova` を用いる。

```
> tb71.aov <- aov(yield ~ condition, data=tb71)
> anova(tb71.aov)
Analysis of Variance Table
Response: yield
Terms added sequentially (first to last)
      Df Sum of Sq Mean Sq F Value Pr(F)
condition 2  3.76634 1.883170 4.846088 0.01590996
Residuals 27 10.49209 0.388596
```

パラメータ (係数 β) の推定値は、`coef` によって表示される。ここで、`condition1, condition2` と表示されているのは、もとのカテゴリーそのものには対応せず、対比 (contrast) とよばれる形

式に変換されたものである。これら、もとの要因の水準とどのような関係にあるかは、計算結果の contrasts 要素を表示することによって知ることができる。つぎの例で condition の 3 つの水準に対応するダミー変数を C_1 、 C_2 、 C_3 と表記すると、Splus の変数 condition1 は $-C_1 + C_2$ に対応し、condition2 は $-C_1 - C_2 + 2C_3$ に対応していることが分かる。

```
> tb71.aov$contrasts
$condition:
  [,1] [,2] # 1列が contrast1, 2列が contrast2
1  -1  -1 # 行は入力されたデータの水準
2   1  -1
3   0   2
> anova(tb71.aov)
Analysis of Variance Table
Response: yield
Terms added sequentially (first to last)
      Df Sum of Sq Mean Sq F Value Pr(F)
condition 2  3.76634 1.883170 4.846088 0.01590996
Residuals 27 10.49209 0.388596
> coef(tb71.aov)
(Intercept) condition1 condition2
      5.073      -0.1855      0.2265
```

23.5.2 2元配置 (Dobson 表 7.4)

以下で利用しているデータファイル tb7_4.dat はつぎの箇所にある。

<http://www.rd.dnc.ac.jp/~otsu/lecture/Dobson>

```
# データの読み込み
> tb74 <- read.table("tb7_4.dat", col.names=c("A", "B", "Y"))
> tb74
  A B  Y
1 a1 b1 6.8
2 a1 b1 6.6
3 a1 b2 5.3
4 a1 b2 6.1
5 a2 b1 7.5
6 a2 b1 7.4
7 a2 b2 7.2
8 a2 b2 6.5
9 a3 b1 7.8
10 a3 b1 9.1
11 a3 b2 8.8
12 a3 b2 9.1
```

```
> is.factor(tb74$A) # 文字データは要因として扱われる。
```

```
[1] T
```

```
> is.factor(tb74$B)
```

```
[1] T
```

数値変数を離散値をとる要因として取扱う場合には、

```
> tb74$A <- as.factor(tb74$A)
```

のように、関数 `as.factor` によって明示的に変数の属性を変更する。(上の例では `read.table` で入力された時点で要因となっているので、変化はしない)。

各要因について水準別の平均をグラフで確認することができる。

```
> X11()
```

```
> plot.design(tb74) # 水準毎の平均を表示
```

```
> plot.factor(tb74) # 各要因 (離散変数) 毎に水準別の箱ひげ図
```

分散分析を実行し、その結果を保存するには関数 `aov` を用いる。分散分析表を出力するには `summary` または `anova` を用いる。

```
> tb74.aov <- aov(Y ~ A*B,data=tb74) # 主効果+交互作用
```

```
> tb74.aov
```

```
Call:
```

```
  aov(formula = Y ~ A * B, data = tb74)
```

```
Terms:
```

	A	B	A:B	Residuals
Sum of Squares	12.74000	0.40333	1.20667	1.48000
Deg. of Freedom	2	1	2	6

```
Residual standard error: 0.4966555
```

```
Estimated effects are balanced
```

```
> summary(tb74.aov)
```

	Df	Sum of Sq	Mean Sq	F Value	Pr(F)
A	2	12.74000	6.370000	25.82432	0.0011274
B	1	0.40333	0.403333	1.63514	0.2482245
A:B	2	1.20667	0.603333	2.44595	0.1671644
Residuals	6	1.48000	0.246667		

もし主効果のみのモデルを当てはめるのなら、

```
tb74.aov <- aov(Y ~ A + B,data=tb74)
```

と記述する。また、Aのみの主効果を考えるのなら

```
tb74.aov <- aov(Y ~ A,data=tb74)
```

と指定する。

分散分析の結果得られる各水準 (および交互作用項) の対比 (contrast) に対応する係数は、関数 `coef` によって表示される。

```

> coef(tb74.aov)
(Intercept)  A1  A2  B  A1B  A2B
          7.35 0.475 0.675 -0.1833333 0.1 0.2166667
>
> coef(tb74.aov)["A1"] # 水準名または番号を指定できる。
A1
0.475

```

対比の内容は計算結果の contrasts というリスト要素に含まれている。

```

> tb74.aov$contrasts
$A:
  [,1] [,2]
a1  -1  -1
a2   1  -1
a3   0   2

$B:
  [,1]
b1  -1
b2   1

```

関数 `lm` の場合と同様に、`plot(tb74.aov)` と指定すると、適切なグラフを表示する。また、関数 `fitted.values`、`residuals` を適用できる。

説明変数に離散的な変数 (分散分析における要因) と連続変数が混在する場合の線形モデルによる分析を共分散分析 (Analysis of Covariance, ANCOVA) と呼ぶ。tb7_8.dat の分析を行ないなさい。

課題 ファイルの各列には3つの変数が含まれている。第1列は実験条件 (3水準) であり、教育方法を示す。第2列はアチーブメントテストの点数であり、これが被説明変数。第3列は授業 (実験) の開始前に行なった適性試験の得点である。第1列と第3列を説明変数とし、第2変数を被説明変数として、分析を行ないなさい。(SASではGLMで一部の変数についてclass宣言を行う。Splusでは関数 `lm` を用いて分析を行う。一部の変数を関数 `factor` を用いて、要因としての属性を与えておく。

24 多重比較 (multiple comparison)

1元配置の分散分析において、 F 検定を用いて検討されることは、帰無仮説 $H_0: \mu_1 = \mu_2 \cdots = \mu_J$ である。もし、 F 値が顕著に大きいなら、これが棄却され、いずれかの (μ_{j_1}, μ_{j_2}) の組において、 $\mu_{j_1} \neq \mu_{j_2}$ であるはずである。ところが、 F 検定自体はどの部分の平均が異なっているとみなされるかについては何も示さない。単純に考えると、各 (μ_{j_1}, μ_{j_2}) について、個別に t 検定または F 検定を行なえば違いのある箇所を発見できるはずである。しかし、これでは若干問題が生じてしまう。問題が生じるのは、複数の検定を繰り返し行なうためである。もし、各々の検定の有意水準が α であるとする、 K 回検定を繰り返して少くともそのうち1つが有意になる確率は、 α より遥かに大きくなる。

ここで、 $\alpha = 0.05$ であり、10 回検定を行なうとしよう。さらに各検定は統計的に独立であると仮定する (実際には、各検定統計量は互いに関係し合っているため独立であることはないが、ここでは簡単のためこのような仮定を置く)。10 回の検定がすべて棄却されない確率は $(1 - \alpha)^{10} \simeq 0.60$ である。つまり、帰無仮説の仮定の下で 10 回の検定のうち少なくとも 1 回が有意になる確率は約 40 パーセントになる。つまり複数の検定を同時に取り扱う場合には、有意水準を割り引いて考える必要がある。このように複数の群の比較を同時に行なう方法は、多重比較と呼ばれる。

表 18 は、3 水準および 5 水準の 1 元配置の分散分析をシミュレートした結果である。各実験条件におけるデータの個数を 5 個とし、で各水準の組合せについて両側 5% の有意水準で t 検定を行なった。データは帰無仮説 $\mu_1 = \dots = \mu_p$ に基づいて生成した。各々の条件について 1 万回 (分散分析を) 繰り返した。

表 18: t 検定への多重性の影響 (1 万回中)

水準	F 検定が有意	有意差の認められた対の数							1 対以上の t 検定が有意 合計
		1	2	3	4	5	6	7	
3	505	838	330	4					1172
5	484	1290	726	333	199	27	29	2	2606

多重比較のために非常に多くの検定方式が提案されているが、ここでは 4 つの方法について述べる。

24.0.3 ボンフェローニ (Bonferroni) の不等式

各検定に対応する検定統計量 (t 値や F 値など) を Z_k とし、その棄却域を R_k とする。 k 番目の検定において、帰無仮説が棄却されるのは $Z_k \in R_k$ であるとき、つまり Z_k が R_k の範囲に入る場合である。ここで、少なくとも一つの検定が有意になる確率は

$$P_1 = Pr(Z_1 \in R_1 \cup Z_2 \in R_2 \cup \dots \cup Z_K \in R_K) \quad (131)$$

と表される。確率論の一般的な公式 (Bonferroni の不等式) から、

$$P_1 \leq \sum_{k=1}^K Pr(Z_k \in R_k) \quad (132)$$

である。つまり各々の検定の有意水準が α/K 以下であれば、すべての検定において少なくとも一つが有意になる確率は α 以下である。

この性質は単純で分かりやすいが、多くの場合にあまりにも棄却の条件が厳しすぎる (検出力が低い) ことになる。

24.0.4 Fisher の LSD 法

R.A.Fisher によって提案された方法である (least significant difference procedure)。つぎのような手順をとる。

1. 以下の全ての検定を有意水準 α で行う。

2. 「群間には差がない」という帰無仮説を一元配置分散分析の F 検定で検定する。
3. もし、分散分析の有意差がなければ終了。
4. 分散分析で有意な差が認められたなら、必要な群の組み合わせについて、2 群の平均の差の t 検定を行う。

以上の手続きでは、最初に F 検定をおこなっているのので、最終的に有意な結果が得られる確率は、帰無仮説の下では α より必ず小さくなる。この手順は単純であり広く使われている。

24.0.5 Scheffé の方法

1 元配置の分散分析において、 p 個の水準の比較を行う場合を想定する。デザイン行列 X は $\beta_1 = \mu_1, \dots, \beta_p = \mu_p$ となるように設定されているものとし、つぎのような線形仮説を考える。

$$d = a_1\beta_1 + \dots + a_p\beta_p = 0 \quad (133)$$

ただしここで、 $a_1 + \dots + a_p = 0$ とする。このような線形仮説のことを対比 (contrast) という。2 水準の同等性の比較 ($\beta_1 - \beta_2 = 0$) などは対比の例である。さらにここで、各水準のサンプル数は等しく n であり、また $\|a\| = 1$ との仮定を置く。ここで、誤差の分散を σ^2 と仮定し、対比がゼロ ($d = 0$) であるという帰無仮説を検討する。ただし、個別の対比を考えるのではなく、すべての対比を同時に考える。この方法は Scheffé の多重比較と呼ばれる。

基本的な方針はつぎのようなものである。対比を検討するための統計量は

$$\hat{d} = a_1\hat{\beta}_1 + \dots + a_p\hat{\beta}_p \quad (134)$$

である。個別に仮説を検討するならば $\sqrt{n}\hat{d}/\hat{\sigma}$ が t 分布に従うことを用いて検定を行えばよい。ここで、 $\hat{\sigma}$ は残差から求められる誤差分散の不偏推定量である。上に示した \hat{d} の値は、 a の示す方向について、 β が原点からどれだけ離れているかを示すものである。

すべての対比を同時に考えるためには、個別の \hat{d} ではなく、 β の同時信頼区間 (領域) を考える。つまり $Pr(\beta \in D)$ の確率が、 $1 - \alpha$ となる領域 D を求める。ただし、この際 D の形はすべての対比の方向について対称とし、対比と無関係な方向については制約を与えないことにする。対比がゼロであるという帰無仮説 $d_a = 0$ は、 β の全体をあらゆる空間のなかでは、原点を通る $p - 1$ 次元の部分空間 (超平面) としてあらわされる。そして領域 D がこの超平面と重なっていなければ、 $d_a = 0$ が棄却されることにする。このような方針をとれば、信頼区間 (実際には多次元領域である) は 1 つ求められるだけなので有意水準が不適切になることはない。

制約を満たす対比の中で最大の絶対値を与えるものを \hat{d}_{max} とすると、帰無仮説のもとで次の分布に従うことが分かる。

$$\hat{d}_{max}^2/\hat{\sigma}^2 \sim (p-1)F(p-1, p(n-1)) \quad (135)$$

そこで、 F 分布の表を用いて、 $\hat{d}_a^2/(\hat{\sigma}^2(p-1))$ が自由度 $\{p-1, p(n-1)\}$ の F 分布の $1 - \alpha$ 点より大きければ、この仮説を棄却する。

24.0.6 Tukey の方法

Scheffé の方法はすべての対比を考慮したが、 $\beta_i - \beta_j$ という型の仮説に対比を限定すれば、検出力をあげることが (つまり差に敏感にすることが) できる。Tukey の多重比較の方法は帰無仮説の

もとの Student 化された範囲 (Studentized range)

$$\frac{\max_{i,j}\{\sqrt{n}|\hat{\beta}_i - \hat{\beta}_j|\}}{\hat{\sigma}} \quad (136)$$

の分布を求め、この値が大きくなる対比を有意とみなすものである。この方法は Scheffé の方法より検出力が高いので、しばしば利用されている。

25 因果関係は統計で分かるか?

25.1 無作為化の意義

t 検定や分散分析を行なって要因の効果を推定する場合、重要な前提はサンプルが各水準にランダムに割り当てられているということである。もし、何らかの意図的な操作によって、サンプルの割り当てを行なうならば、推定される実験の効果は誤った値であるかも知れない。例えば、英語教育法の効果を測定する実験を計画する場合を考えよう。教育法の要因を A とし、実施する方法により A_1 と A_2 の 2 種のクラス編成を行なうこととする。クラスへの生徒の割り当てが無作為に行なわれるならば、教育実施後の成績 Y の平均を比較することにより、 A_1 と A_2 2 つの方法の効果について正しく推定を行なうことができる。しかし、入学時の英語の成績を X_1 とし、 X_1 にもとづいて成績優秀者を A_1 に割り当てたなら、教育法の効果を正しく測定することはできない。

1954 年にアメリカで行なわれたソーク・ポリオワクチンの臨床試験は観察研究と無作為化の比較例として歴史的に有名である (Meier, 1989)。

実験の対象となった地域は 2 つの群に類別され、第 1 群では観察対照法が、第 2 群では偽薬を用いた 2 重マスク試験 (2 重盲検法) が採用された。つまり、第 1 群においては、2 年目の生徒にのみワクチンが接種され、1 年目と 3 年目には接種されない。ただし接種の対象となるのは希望した生徒たちである。第 2 群においては、同様に希望者が実験の対象となるが、そのうち無作為に選ばれた半数にはワクチンが投与され、残りの半数には生理食塩水が投与される。生徒当人も親も、また後にポリオ発病の診断にあたる医師も生徒が実際にワクチンを投与されたのか偽薬を投与されたのかは知らない。ポリオの発病は、衛生的で裕福な環境の子供に多い。また、実験への参加者は、非参加者に比べ親の教育水準が高く裕福であり、また学校の欠席率が高いことが後に分かった。実験の結果、ワクチンは特に麻痺性のポリオに顕著な効果のあることが確認された。ワクチンの効果は第 1 群でも、第 2 群でもほぼ同等に認められた。また発病率に参加者バイアスが存在することも確認された。

この実験手続き、特に第 2 群において真剣な努力の対象となっていることは、ポリオの発病とワクチンの投与の両者に同時に影響を与える要因を極力排除するということである。上の例では、たまたま 2 つの場合で効果の差がさほど表れていないが、一般的にはもし両者に影響を与える未知の変数が存在すれば、ワクチンの効果を正確に評価することはできない。無作為化は、たとえ未知の変数が存在したとしても、ワクチンの投与への影響を断ち切る。第 1 群と比べて第 2 群の実験が優れているのは、この影響の遮断が完全だからである。

ここに述べた問題は、回帰モデルの利用全般に関係する。もう一つ仮想的な例を考えて見よう。選挙の投票率を研究することを目的に、調査を行なう場合を想定する。被験者の属性として、年収の対数 X_1 、性別 X_2 (男 0, 女 1)、年齢 X_3 を調べ、選挙で投票するか否かを Y とし、回帰分析 (Y が 0-1 変数であるので、後述するロジスティック回帰を使う法がよい) を行なう。分析の結果、

$$E(Y|X) = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 \quad (137)$$

という回帰式が推定されたものとする(図 17)。ここで、 $\hat{\beta}_1$ の意味はどのように解釈できるだろうか。普通、教科書には「 X_1 が 1 単位増加したとき、 Y がどれだけ増加するかを意味する」という趣旨のことが書いてある。この場合、そのような解釈(年収が 1 割増加とすると投票率が $(\log 1.1 \times \beta_1)$ だけ増加する)は可能だろうか。このような解釈が可能になるためには、 X_1 と Y とに同時に影響を与えている未測定の変数(図 17 の U)が存在しないことが必要である。例えば、学歴が年収 X_1 と投票率 Y の両者に影響を及ぼしているにも係わらず、説明変数として用いられていない場合には、 $\hat{\beta}_1$ は学歴の影響をも含んだ値になる。この場合、回帰式は変数間の関係の記述としての意味はあるし、また予測のための道具としても有効であるが、影響関係の推定として解釈することは困難である。

もし、年収が本当に投票率にどのような影響を及ぼしているかを厳密に測定しようとするなら、被験者の年収をくじ引きで決定し(被験者 1 の年収は 400 万、被験者 2 は 500 万などのように収入をさだめ実際に支給する)、その上で調査を行なう必要がある。このような実験は現実には不可能であるが、年収のもつ効果を厳密に測定しようとするなら、このような方法を取らざるを得ない。この場合、学歴が未測定であったとしても、学歴と年収とは無関係になるので、推定された回帰係数 $\hat{\beta}_1$ が学歴(やその他の未知・既知両方の要因)の影響を受けることはない(図 18)。

ある変数に実験者が介入し無作為化するということは、観察によって受動的に得られるデータとは、本質的に分布の異なるデータが得られるということである。介入を受ける変数(X_1)の値は、実験者によって制御されているため、他の変数の値の影響を受けない。もし、 X_2 が X_1 に影響を及ぼしているのならば、 X_2 から X_1 への影響は実験を行なう場合には存在しなくなる(図には示していない)。また、実験者にとって未知の変数 U が X_1 と Y の両者に影響を及ぼしていたとしても、 U から X_1 への影響は存在しなくなる。一方、 X_1 が X_2 に影響を及ぼしている場合には、実験下であっても X_1 と X_2 との関係は、観察データの場合と変わらない。これは、 X_1 の値が変わることによって Y が被る影響が、 X_1 の値が無作為化によって定められている状況では正しく推定されうることを意味する。

しかし、現実には実験が不可能である場合は多い。無作為化実験を行わずに、 X_1 の Y への影響を正確に評価するためには、 X_1 と Y とに同時に影響を与える主要な要因(変数)をすべて観測する必要がある。もし、これが可能であるなら、これらの変数を全てモデルに取り込んだ上で、 β_1 の評価(および間接的な影響の評価)を行なうことにより、 X_1 の Y に及ぼす直接・間接の影響を評価することが可能である。調査データの分析の多くは、このような方針のもとで行なわれている。しかしながら、実際にこのような条件を実現するのは、かなり難しい。関係のありそうな要因をすべて測定したつもりであっても、実は未知の要因が影響を及ぼしている危険は常に存在する。

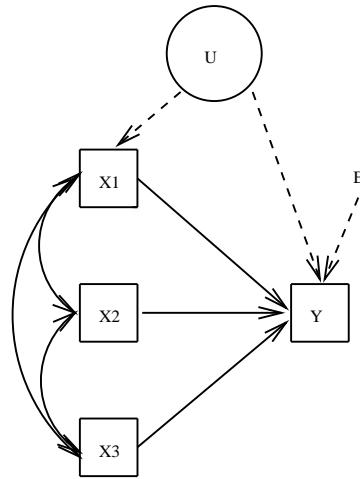


図 17: 回帰分析の概念

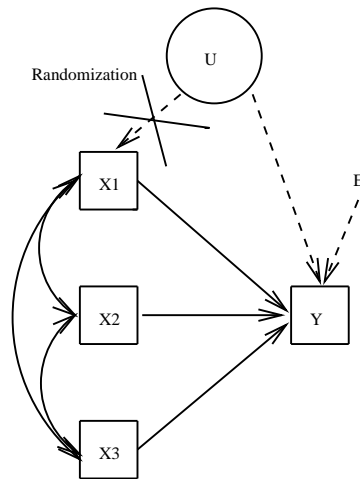


図 18: 無作為化の役割

26 StatLabs データ: 母親の喫煙と新生児体重

26.1 データの所在

<http://stat-www.berkeley.edu/users/statlabs/labs.html#babies>

にある。

データはブラウザで取得できる。

WEB ページ上で Birth weight II と表示されている箇所でもウスの右ボタンをクリックし、「名前をつけて保存」を選択する。

26.2 SAS によるデータ入力の例

```
data babies;
  infile 'babies.data' firstobs=2;
  input  bwt gestat parity age height weight smoke ;
  if bwt = 999 then bwt = . ;
  if gestat = 999 then gestat = . ;
  if parity = 9 then parity = . ;
  if age = 99 then age = . ;
  if height = 99 then height = . ;
  if weight = 999 then weight = . ;
  if smoke = 9 then smoke = . ;
run;
* proc print; /* 先頭の星をとると全件表示 */
proc means;
proc sort data=babies;
  by smoke;
proc means; by smoke;
proc glm data=babies;
  class parity smoke;
  model bwt = gestat parity age height weight smoke
        / ss1 ss3 solutions ;
run;
```

26.3 Splus/R によるデータ入力例

以下の内容は、対話的に入力することも可能だが、mule、emacsなどでファイルにあらかじめ作成しておいたほうが楽。ファイルに記述された内容を Splus/R で実行するには source("ファイル名") と入力する。

```
# 見出しがついているので、変数名が自動的につけられる。
babies <- read.table("babies.data",header=T)

# babies データフレームの変数を、変数名だけで参照可能にする。
attach(babies)

# 欠測値の処理を行う。
babies$bwt <- replace(bwt, bwt == 999, NA);
babies$gestation <- replace(gestation, gestation == 999, NA);
babies$parity <- replace(parity, parity == 9, NA);
babies$age <- replace(age, age == 99, NA);
babies$height <- replace(height, height== 99,NA);
babies$weight <- replace(weight, weight== 999,NA);
```

```

babies$smoke <- replace(smoke, smoke == 9, NA);

# 離散値をとる変数を factor として宣言し、カテゴリーに名前をつける。
babies$parity <- factor(parity, levels=c(0,1),
                        labels=c("First","Other"));
babies$smoke <- factor(smoke,levels=c(0,1),
                       labels=c("Nosmoke","Smoke" ) );

detach("babies") # R の場合、detach してから再 attach が
attach(babies)  # 必要のようだ。

# babies データフレームを attach しておく、変数名だけで参照できる。
# 下の例では data=baies を指定しなくても OK。
babies.lm <- lm(bwt ~ gestation + parity + age + height + weight + smoke,
               na.action=na.omit, data=babies)

summary(babies.lm) # 分析結果の概要
anova(babies.lm)  # 分散分析表 SS は SAS の TypeI に相当。

# factor 宣言された変数の回帰係数については注意が必要。
# どのようなデザイン行列 X が用いられているかは、
# つぎのようにして調べる。
# (R の最新版では SAS に類似のデザイン行列が採用されている。)

babies.lm$contrasts

```

26.4 共分散分析の適用

以上で説明したように、回帰分析と分散分析とは数学上は同一モデルとして表され、回帰係数の推定や検定もほぼ同一の手順によって行うことができる。これらの手法が異なるのは、モデルの特徴というよりは、むしろ応用の対象となるデータの性質であることが多い。ここで説明したモデルは、説明変数が実験者によって設定された値であることを前提としている。分散分析が応用される場面ではこの前提が当てはまることが多いが、実際に回帰分析の応用、特に社会科学ではこの前提は多くの場合当てはまらない。説明変数が確率的に変動する場合にも、ある種の条件を仮定すれば、予測値の推定は説明変数が実験者によって設定された場合と同様の手続きで行うことが可能である。最大の問題は、回帰係数の解釈が不明確になることにある。

上述の線形のモデルにおいて、説明変数に離散値と連続値の両者を含むものは共分散分析 (Analysis Covariance) と呼ばれる。以下に共分散分析の例を示すが、この例においてはデータは疫学調査に基づくものであり、研究者によって設定された値ではない。このような場合、回帰係数が影響の程度を表しているとみなせるためには、説明変数と非説明変数の両者に影響を及ぼす変数が全てモデルに取り込まれている必要がある。

共分散分析モデルには、次のような特徴がある。

1. 離散値をもつ説明変数 (要因) は X のいくつかの列におけるダミー変数として表現される。また、これらの交互作用もダミー変数となる。
2. 共分散分析においては、さらに連続値を持つ説明変数と離散値を持つ説明変数の交互作用も用いられる。これらの交互作用を表す変数は、離散値を表すダミー変数 (0-1 値) に、連続変数の値をかけたもの、つまりダミー変数が 1 の部分のみ連続変数の値を残し、他の部分はゼロとなる変数によって表される。

表 19: 出生時体重と母親の喫煙データ (Nolan & Speed, 2000 より)

変数	変数名	内容
出生時体重	bwt	新生児の出生時体重 (オンス)
妊娠期間	gestation	妊娠期間 (日)
パリティ	parity	第 1 子 (0,First) か否 (1,Other) か
年齢	age	母親の受胎時の年齢 (年)
身長	height	母親の身長 (インチ)
体重	weight	母親の妊娠前の体重 (ポンド)
喫煙	smoke	母親の喫煙 (Smoke), 非喫煙 (Nosmoke)

ここで X_{1b} と X_{1c} がある要因 (3 要因) を表すダミー変数であるとし, X_2 は別の連続値を持つ説明変数であるとする. これらの各要因に対応する標本を A 群, B 群, C 群と名付ける. X_{1b} は A 群と B 群における差に対応し, X_{1c} は A 群と C 群との差に対応する.

主効果のみのモデルは, 次の式で表される.

$$E(Y|X) = \beta_0 + \beta_{1b}X_{1b} + \beta_{1c}X_{1c} + \beta_2X_2$$

ここで, β_{1b} と β_{1c} とは, 離散変数であらわされる群間の Y の違いを表し, β_2 は各群に共通な連続変数 X_2 と Y との関係を表す.

離散変数と連続変数の交互作用を含むモデルは,

$$E(Y|X) = \beta_0 + \beta_{1b}X_{1b} + \beta_{1c}X_{1c} + \beta_2X_2 + \beta_{1b2}X_{1b}X_2 + \beta_{1c2}X_{1c}X_2$$

となる. ここで, β_{1b2} が非ゼロであることは, A 群と B 群において X_2 が Y に及ぼす効果が異なることを示す. A 群においては X_2 の回帰係数は β_2 であり, B 群においては $\beta_2 + \beta_{1b2}$ である. β_{1c2} についても同様のことが当てはまる.

3. 連続変数間の交互作用は, これらを単純に掛算した 2 次式によっては極めて限られた形の効果しか表現できない. スプライン関数とよばれる区分的多項式 (区分的 3 次式がよく用いられる) を用いると, 連続変数間の交互作用を柔軟に表現できるが, 現在のところ利用可能なソフトウェアは限られている.

26.4.1 分析の例

Nolan& Speed (2000) で引用されている新生児の体重と妊娠中の母親の喫煙に関するデータを分析対象とする. このデータは, 1960 年から 1967 年にサンフランシスコにおける健康調査によって得られた (Yerushalmy,1971). データは 1236 件であるが分析には欠測値のないもの 1174 件を用いた.⁴

ここで示されている変数は次の 7 つである.
データの一部

⁴データは <http://stat-www.berkeley.edu/users/statlabs/> において公開されている.

```

      bwt gestation parity age height weight  smoke
1 120      284 First  27    62    100 Nosmoke
2 113      282 First  33    64    135 Nosmoke
3 128      279 First  28    64    115 Smoke
4 123      NA First  36    69    190 Nosmoke
5 108      282 First  23    67    125 Smoke
...

```

以下はの分析は R-1.5.0 による．分散分析表は説明変数を逐次加えた場合の，残差 2 乗和の減少分を表している．標本件数のバランスした分散分析とは異なり，残差 2 乗和の減少量は変数の追加順に依存する．

モデル 1: 5 つの説明変数を用いたモデル．年齢は影響が小さいため説明変数から除いた．

Residuals:

```

      Min      1Q  Median      3Q      Max
-57.7164 -10.1500 -0.1594   9.6885  51.6199

```

Coefficients:

```

      Estimate Std. Error t value Pr(>|t|)
(Intercept) -80.71321   14.04465  -5.747 1.16e-08 ***
gestation    0.44408    0.02907  15.276 < 2e-16 ***
parityOther  -3.28762    1.06281  -3.093 0.00203 **
height       1.15497    0.20473   5.641 2.11e-08 ***
weight       0.04983    0.02503   1.991 0.04672 *
smokeSmoke  -8.39390     0.95117  -8.825 < 2e-16 ***
---

```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.82 on 1168 degrees of freedom

Multiple R-Squared: 0.2579, Adjusted R-squared: 0.2548

F-statistic: 81.2 on 5 and 1168 DF, p-value: < 2.2e-16

AIC は 9823.503

分散分析表

Response: bwt

```

      Df Sum Sq Mean Sq F value Pr(>F)
gestation 1 65450 65450 261.4292 < 2.2e-16 ***
parity    1 2345 2345 9.3658 0.002261 **
height    1 12554 12554 50.1437 2.462e-12 ***
weight    1 1801 1801 7.1941 0.007418 **
smoke     1 19497 19497 77.8779 < 2.2e-16 ***
Residuals 1168 292412 250
---

```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

モデル 2: 年齢以外の 5 つの変数，および妊娠期間と喫煙の交互作用項を用いたモデル．自由度調整済 R^2 は増加，AIC は減少している．いずれもモデルがより適切であることを示唆している．

ただし，実際の残差を検討してみると，非喫煙群で妊娠期間が極端なケース（短い場合と長い場合）があり，回帰係数に大きな影響を与えている（図.21）．

Residuals:

```

      Min      1Q  Median      3Q      Max
-57.87631 -9.95257 -0.01195   9.54773  53.49886

```

Coefficients:

```

      Estimate Std. Error t value Pr(>|t|)

```

```

(Intercept)      -59.07260   15.30129  -3.861  0.000119 ***
gestation        0.37020    0.03590  10.313 < 2e-16 ***
parityOther     -3.29621    1.05780  -3.116  0.001877 **
height          1.14888    0.20377   5.638  2.16e-08 ***
weight          0.04542    0.02494   1.821  0.068855 .
smokeSmoke     -66.82659   16.83114  -3.970  7.61e-05 ***
gestation:smokeSmoke  0.20970    0.06031   3.477  0.000525 ***

```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.75 on 1167 degrees of freedom
Multiple R-Squared: 0.2656, Adjusted R-squared: 0.2618
F-statistic: 70.33 on 6 and 1167 DF, p-value: < 2.2e-16

AIC は 9813.402

分散分析表

Response: bwt

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
gestation	1	65450	65450	263.9117	< 2.2e-16 ***
parity	1	2345	2345	9.4547	0.0021550 **
height	1	12554	12554	50.6198	1.952e-12 ***
weight	1	1801	1801	7.2624	0.0071424 **
smoke	1	19497	19497	78.6175	< 2.2e-16 ***
gestation:smoke	1	2999	2999	12.0910	0.0005253 ***
Residuals	1167	289413	248		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

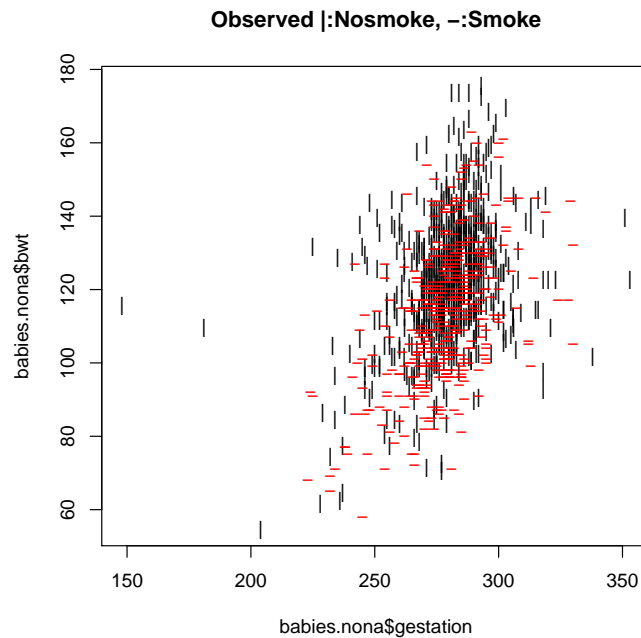


図 19: 妊娠期間と出生時体重 (観測データ)

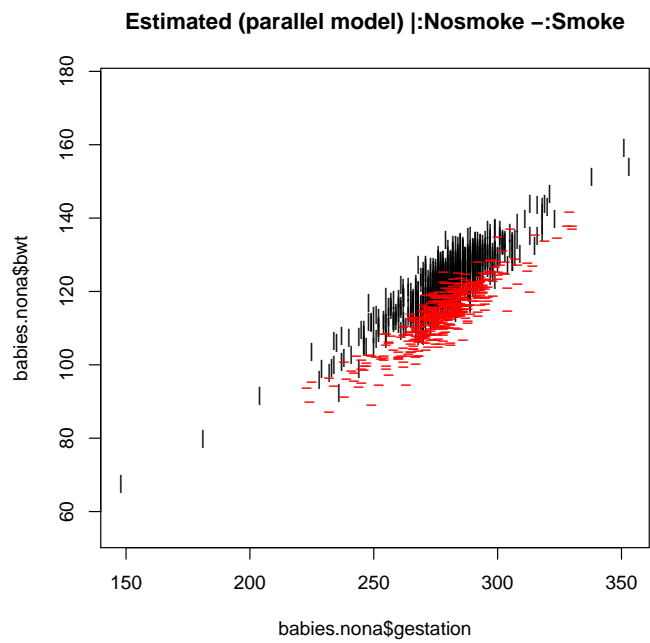


図 20: 妊娠期間と出生時体重 (モデル 1 の予測値)

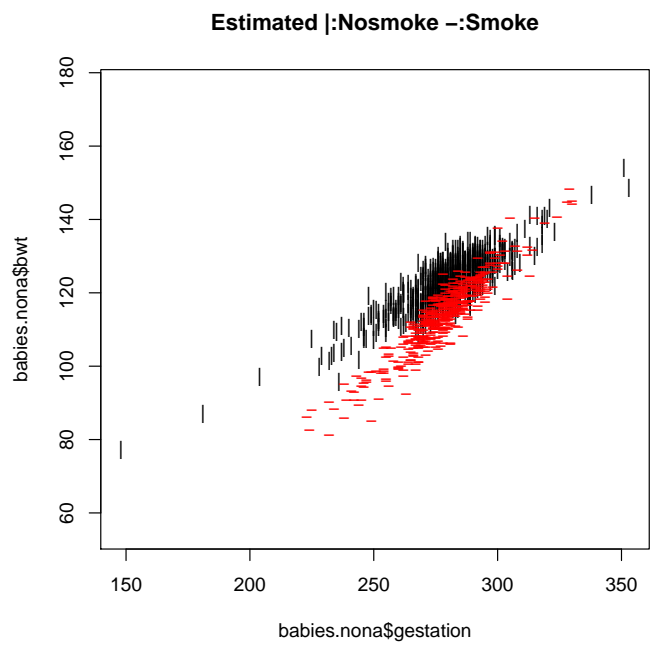


図 21: 妊娠期間と出生時体重 (モデル 2 の予測値)

27 最尤法によるパラメータの推定

線形の回帰分析の場合には、最小 2 乗基準によりつぎの式

$$\sum_{i=1}^N (y_i - \hat{y}_i)^2 \quad (138)$$

を最小化することによって、 $\hat{\beta}$ を求めた。

ここで、被説明変数が正規分布以外にも対応できるよう、推定方法をより一般的にしたい。 ε が正規分布以外の確率分布に従う場合にも、最小 2 乗法を使うことは妥当であろうか？

例えば、 Y_i が 2 項分布 $Bin(m, \pi)$ に従うとしよう。2 項分布 $Bin(m, \pi)$ は平均 $m\pi$ 、分散 $m\pi(1-\pi)$ である。また、ここで π は説明変数 x_1, x_2, \dots, x_p の関数 $\eta(\beta_1 x_1 + \dots + \beta_p x_p)$ として定まるとしよう。

単純に考えるなら、 i 番目の観測値に対応する π_i と y_i/m_i を考え、 π_i と y_i/m_i の残差 2 乗和を最小化する方法は妥当であるように思われる。しかし、もっとよい方法を理論的に構成することが可能である。

2 項分布の分散の大きさが m と π に依存して変化することを考慮する必要がある。現在、一般的に用いられている推定方法は最尤法 (maximum likelihood method) と呼ばれるものである。最尤法の一般論はつぎのようなものである。

データを生み出すと仮定される確率密度関数を $f(Y|\theta)$ とする。 Y は確率変数であり、 θ は分布を特徴づけるモデルパラメータ (母数) である。正規分布の場合であれば、平均と分散が母数であり、2 項分布の場合は反応確率 π が相当する。実際に観測されたデータを y_1, \dots, y_n とし、これらは互いに独立に分布していると仮定する。 $f(y_i|\theta)$ の値は、 y_i が指定された θ のもとで、どの程度生じやすいかをあらわすものと解釈できる。これを逆に θ の関数としてみるなら、特定の y_i について、その値を θ がどの程度生じやすいかの指標になっていると考えられる。これを尤度 (likelihood) と呼び多くの場合 $L(\theta|y_i)$ と表記する。 θ の推定値として尤度を最大にする値を採用する方法を最尤法 (maximum likelihood method) と呼ぶ。

観測データ全体について互いの独立性を仮定すると、これらの同時分布は $f(y_1|\theta) \times \dots \times f(y_n|\theta)$ であり、これを $L(\theta|y_1, \dots, y_n)$ と解釈する。しばしば、計算上の問題と理論的な要請から、尤度そのものではなく、尤度の対数が取り扱われる。これを対数尤度 (log-likelihood) と呼び、 $l(\theta|y_1, \dots, y_n)$ と表記する。対数尤度について、つぎのような式が成立する。

$$l(\theta|y_1, \dots, y_n) = \log L(\theta|y_1, \dots, y_n) = \sum_{i=1}^n \log f(y_i|\theta) \quad (139)$$

27.1 最尤推定値の例

正規分布の場合について考えてみよう。正規分布の確率密度関数はつぎの式で与えられる。

$$f(y|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \times \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right) \quad (140)$$

データ $\{y_1, \dots, y_n\}$ が与えられた場合の、対数尤度は

$$l(\mu, \sigma^2|y_1, \dots, y_n) = \text{定数} - \frac{n}{2} \log \sigma^2 - \frac{\sum_{i=1}^n (y_i - \mu)^2}{2\sigma^2} \quad (141)$$

となる。この値を最大化する μ と σ はそれぞれ、

$$\hat{\mu}_{ML} = \frac{1}{n} \sum_{i=1}^n y_i, \quad \hat{\sigma}_{ML}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{\mu})^2 \quad (142)$$

となる。つまり、母平均の推定値は標本平均であり、母分散の推定値が標本分散になる。正規分布の最尤法による推定を行なう場合、平均の推定値は不偏 (unbiased) であるが、分散の推定値の期待値は $\frac{n-1}{n}\sigma^2$ であり、 σ^2/n だけ過小評価となる。

つぎに 2 項分布にの場合を考える。2 項分布の確率分布は $y = 0, 1, \dots, m$ についてつぎの式で与えられる。

$$Pr(y|m, \pi) = \binom{m}{y} \pi^y (1 - \pi)^{(m-y)} \quad (143)$$

データ、 $\{y_1, \dots, y_n\}$ が与えられているとすると、 π についての対数尤度は

$$l(\pi|m, y_1, \dots, y_n) = \text{定数} + \sum_{i=1}^n \{y_i \log \pi + (m - y_i) \log(1 - \pi)\} \quad (144)$$

$$= \text{定数} + z \log \pi + (nm - z) \log(1 - \pi) \quad (145)$$

となる。ここで $z = \sum_{i=1}^n y_i$ である。この式を最大化する π は $\hat{\pi}_{ML} = z/(nm)$ となり、標本の全体の反応比率に一致する。

尤度が定義できるなら、式を解くかあるいは数値的な方法をもちいて、尤度を最大にするパラメータの値を計算することが可能であるので、非常に利用範囲の広い方法である。また、データが十分多い場合には、つぎのようにある意味で最尤推定法はデータのもつ情報を洩れなく利用する最善の推定方法であることが分かっている。

ここで、「最善」と述べた意味はつぎのようなものである。分布の母数 θ のある不偏推定値 (平均が真の値に等しいもの) T を考える。すると、つぎの不等式がかならずなりたつ。

$$Var(T) \geq i(\theta)^{-1} \quad (146)$$

ここで $i(\theta)$ は後述する Fisher 情報量と呼ばれるものである。最尤推定量については、データの件数が多い場合には、上の不等式において等号がほぼ成立する。

27.2 最尤法の特徴 (難)

最尤法の特徴は漸近的な性能の良さにある。ここで、漸近的 (asymptotic) とは、「標本数が増加するときにある性質が成立する」ということを意味する。以下では、いささか理論的になるが最尤法の特徴をまとめておく (全部は理解しなくてもよい)。

27.2.1 パラメータが 1 つの場合

観測された互いに独立なデータを $\{y_1, \dots, y_n\}$ としこれを \mathbf{y} であらわす。このサンプルについての対数尤度は

$$l(\theta|\mathbf{y}) = \sum_{i=1}^n \log f_{Y_i}(y_i|\theta) \quad (147)$$

となる。ここで、 f_{Y_i} は、確率変数 Y_i の確率密度関数である。

対数尤度の微分について、以下の公式が比較的緩い条件 (期待値をとるための積分と微分の交換が可能であること) のもとで、成立する。

$$\begin{aligned} E_{\theta_0} \left(\frac{\partial l}{\partial \theta} \Big|_{\theta_0} \right) &= 0 \\ E_{\theta_0} \left(\frac{\partial^2 l}{\partial \theta^2} \Big|_{\theta_0} \right) + \text{Var}_{\theta_0} \left(\frac{\partial l}{\partial \theta} \Big|_{\theta_0} \right) &= 0 \end{aligned} \quad (148)$$

期待値 (平均) および分散は、母数が θ_0 の分布についてのものであり、また微分も同じ母数 θ_0 における値であることに注意する。これらの公式は、確率密度関数のつぎの性質を用いて導かれる。

$$\int f_Y(y|\theta) dy \equiv 1 \quad (149)$$

さらに、いくつかの条件のもとで、つぎの公式 (Bartlett の等式) が成立する。

$$E_{\theta} \left(\frac{\partial^3 l}{\partial \theta^3} \right) + 3 \text{Cov}_{\theta} \left(\frac{\partial^2 l}{\partial \theta^2}, \frac{\partial l}{\partial \theta} \right) + E_{\theta} \left(\frac{\partial l}{\partial \theta} \right)^3 = 0 \quad (150)$$

対数尤度の微分 $U(\theta; y) = \partial l / \partial \theta$ はスコア統計量 (score statistic) と呼ばれることもある。この分散は

$$i(\theta) = \text{Var}_{\theta} \left(\frac{\partial l}{\partial \theta} \right) = - E_{\theta} \left(\frac{\partial^2 l}{\partial \theta^2} \right) \quad (151)$$

であるが、これは θ についてのフィッシャー情報量 (Fisher information) とよばれる。各サンプルが独立の場合には、つぎのようになり、サンプル全体に対応するフィッシャー情報量が各サンプルのフィッシャー情報量の和になることがわかる。

$$U(\theta|\mathbf{y}) = \sum_{i=1}^n \frac{\partial \log f_{Y_i}(y_i|\theta)}{\partial \theta}, \quad (152)$$

$$i(\theta) = \sum_{i=1}^n E_{\theta} \left(\frac{\partial \log f_{Y_i}(y_i|\theta)}{\partial \theta} \right)^2 \quad (153)$$

また、スコア統計量の分布について、つぎの式が漸近的に成り立つ。

$$i(\theta)^{-1/2} \left(\frac{\partial l}{\partial \theta} \right) \sim N(0, 1) + O_p(n^{-1/2}) \quad (154)$$

ここで、 $O_p(k_n)$ はある確率変数の列 X_n , ($n = 1, 2, \dots$) であって、任意の $\varepsilon > 0$ について、正の整数 m と正の数 a が存在し、 $Pr(|X_n/k_n| > a) < \varepsilon$ がすべての $n > m$ について成立することをいう。ただし、これは仮定されたモデルが正しい場合についてである。また微分は真のパラメータ値におけるものである。記法 O_p (確率変数の列) については Bishop et al. (1977) の Chap.14 を参照。

線形回帰を含む多くの統計モデルにおいては、尤度関数の形はパラメータの定義域の内部で単峰であり、尤度を最大にするパラメータの値は尤度関数の微分がゼロである点、つまり $U(\hat{\theta}|\mathbf{y}) = 0$ を満たす点として一意に定まる。フィッシャー情報量 $i(\theta)$ が大きければ (n が大きい場合にはこの条件は成立する) $\hat{\theta}$ の分布は、モデルが正しいとの仮定のもとでつぎのように近似される。

$$\hat{\theta} - \theta \sim N(0, i(\theta)^{-1}) \quad (155)$$

サンプル数 n が大きいとき、 $\hat{\theta}$ における対数尤度の値と、真のパラメータ値における対数尤度の値の違いは、つぎのような分布で近似される。

$$2l(\hat{\theta}|Y) - 2l(\theta|Y) \sim \chi_1^2 + O(1/n) \quad (156)$$

この値は対数尤度比 (log likelihood ratio) (尤度の比の対数だから) 統計量と呼ばれる。ここで、 χ_1^2 は自由度 1 の χ^2 分布をあらわす。また、 $O(k_n)$ は、ある数列 $x_n, (n = 1, 2, \dots)$ であって、 $|x_n/k_n|$ が有界 (有限の範囲にあること) であるものを示す。この近似は、(155) による正規近似が正確でない、比較的小さな n についても正確である。 $\hat{\theta}$ の $(1 - \alpha)$ 信頼区間は

$$2l(\hat{\theta}|\mathbf{y}) - 2l(\theta|\mathbf{y}) \leq \chi_{1,\alpha}^2 \quad (157)$$

をみたく θ の範囲で近似される。

ここで、注意すべきことは、 $\hat{\theta}$ はサンプル数 n が大きくなると、真の値 θ に収束するが、(156) の値はゼロには近付かないことである。 n が大きくなると尤度関数の傾斜も急峻になり、最尤推定値における対数尤度の期待値は真の値には収束しない。

また、(156) の近似は適当な定数 b を定めることにより、

$$2l(\hat{\theta}|Y) - 2l(\theta|Y) \sim (1 + b)\chi_1^2 + O(n^{-3/2}) \quad (158)$$

と改善されることが知られている。この修正は Bartlett 補正 (Bartlett adjustment) とよばれる。

27.2.2 複数のパラメータの場合

前述の (154), (155) における漸近的な結果は、多パラメータの場合にも同様に適用される。ただし、この場合、 $i(\theta) \rightarrow +\infty$ (この場合は行列) の意味は、行列 $i(\theta)$ の固有値が無限大になることを意味し、行列の個々の要素が大きくなることを意味しない。

パラメータベクトル θ が、2つの部分 $\theta^T = (\psi^T, \lambda^T)$ に分割されると仮定する。前半の部分が分析者にとって関心のある部分であり、後半は余分なパラメータ (局外母数 nuisance parameter と呼ばれる) であるとする。 θ についてのフィッシャー情報行列をつぎのように分割する。

$$i(\theta) = \begin{pmatrix} i_{\psi\psi} & i_{\psi\lambda} \\ i_{\lambda\psi} & i_{\lambda\lambda} \end{pmatrix} \quad (159)$$

また、この逆行列を

$$i(\theta)^{-1} = \begin{pmatrix} i^{\psi\psi} & i^{\psi\lambda} \\ i^{\lambda\psi} & i^{\lambda\lambda} \end{pmatrix} \quad (160)$$

とする。この逆行列の部分行列の逆行列はつぎのようになる (Rao, 1973, Sec. 1b の補足参照)。

$$(i^{\psi\psi})^{-1} = i_{\psi\psi} - i_{\psi\lambda} i_{\lambda\lambda}^{-1} i_{\lambda\psi} \quad (161)$$

これは、 $\hat{\psi}$ の共分散行列の逆行列をあらわしている。つまり、 λ が未知の場合の $\hat{\psi}$ の分散が $i^{\psi\psi}$ で近似されることを意味する。一方、真のパラメータの値 λ を固定した場合の $\hat{\psi}$ の分散行列は $i_{\psi\psi}^{-1}$ で近似される。これらを比較すると、 $i^{\psi\psi}$ の方が大きいことがわかる。つまり、同時に推定すべき未知のパラメータ λ があると、 $\hat{\psi}$ の分散は大きくなる。

多パラメータについての、真のパラメータ θ と最尤推定量 $\hat{\theta}$ における対数尤度比の分布は、つぎのように近似される。

$$2l(\hat{\theta}|Y) - 2l(\theta|Y) \sim \chi_p^2 + O(1/n) \quad (162)$$

ここで、 p は θ の次元である。また局外母数がある場合には、

$$2l(\hat{\psi}, \hat{\lambda}|Y) - 2l(\psi, \lambda|Y) \sim \chi_{p-q}^2 + O(1/n) \quad (163)$$

となる。ここで、 $p - q$ は ψ の次元、または (161) における i^{ψ} のランクである。パラメータ $\hat{\psi}$ の $(1 - \alpha)$ 信頼域は

$$2l(\hat{\psi}, \hat{\lambda}|\mathbf{y}) - 2l(\psi, \hat{\lambda}_\psi|\mathbf{y}) \leq \chi_{p-q, \alpha}^2 \quad (164)$$

で与えられる。

28 最尤法と赤池情報量基準 AIC(難)

以前の資料で、AIC と呼ばれるモデル選択のための基準が、線形モデルの場合には、

$$\text{AIC} = N \log \frac{\|\mathbf{y} - \hat{\mathbf{y}}\|^2}{N} + 2q + \text{定数} \quad (165)$$

と表されることを紹介した。ここで、 q はモデルに含まれるパラメータの個数である。AIC は、最尤法における対数尤度比検定と密接な関係を持っている。

28.1 期待対数尤度とモデルの良さ

対数尤度は

$$l(\boldsymbol{\theta}|Y) = \sum_{i=1}^N \log f(Y_i|\boldsymbol{\theta}) \quad (166)$$

として定義される。最尤法はこの値を大きくする $\boldsymbol{\theta}$ が良い推定値であろうという考えに基づいて、パラメータ推定を行なおうとする。では、どのような意味において、尤度を大きくするパラメータが望ましいといえるのか。

上の (166) を真の分布 $f(Y|\boldsymbol{\theta}_0)$ について平均すると次の式が得られる。

$$E_{\boldsymbol{\theta}_0}\{l(\boldsymbol{\theta}|Y)\} = \sum_{i=1}^N E_{\boldsymbol{\theta}_0}\{\log f(Y_i|\boldsymbol{\theta})\} \quad (167)$$

ここで、

$$E_{\boldsymbol{\theta}_0}\{\log f(Y_i|\boldsymbol{\theta})\} = \int f(Y_i|\boldsymbol{\theta}_0) \log f(Y_i|\boldsymbol{\theta}) dY_i \quad (168)$$

であるが、この値は、 $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ であるときに最大となることが分かっている。そこで、最大となる値からの差を求めると

$$E_{\boldsymbol{\theta}_0}\{\log f(Y_i|\boldsymbol{\theta}_0)\} - E_{\boldsymbol{\theta}_0}\{\log f(Y_i|\boldsymbol{\theta})\} = \int f(Y_i|\boldsymbol{\theta}_0) \{\log f(Y_i|\boldsymbol{\theta}_0) - \log f(Y_i|\boldsymbol{\theta})\} dY_i \quad (169)$$

となる。この値は $\boldsymbol{\theta}_0$ で表される分布と $\boldsymbol{\theta}$ で表される分布がどれくらい違っているかを、 $f(Y|\boldsymbol{\theta}_0)$ を基準点として評価した距離のようなものとみなすことができる。対数尤度を N で割った値

$$\frac{1}{N} l(\boldsymbol{\theta}|Y) = \frac{1}{N} \sum_{i=1}^N \log f(Y_i|\boldsymbol{\theta}) \quad (170)$$

は、 N が大きければ対数尤度の真の分布についての期待値を近似していると考えられるので、結局、最尤推定値は (169) の基準で真の分布と推定された分布の近さを測り、もっとも近いものを得ようとしていると解釈できる。

28.2 最尤推定量の真値からのずれ

前節において次のような式を示した（ここでは真のパラメータの値を θ_0 として記述している）。

$$2l(\hat{\theta}|Y) - 2l(\theta_0|Y) \sim \chi_p^2 + O(1/N) \quad (171)$$

自由度 p の χ^2 分布の平均が p であることを考慮すると、これは、モデルが p 個のパラメータを推定する必要がある場合、最尤推定値をモデルに代入して得られる対数尤度の 2 倍は、真の値 θ_0 を代入して得られる対数尤度の 2 倍より、 p だけ平均して大きいことを意味している。

上の式を真の分布 $f(Y|\theta_0)$ について平均すると、 N が十分大きければ、

$$E_{\theta_0}\{2l(\hat{\theta}_Y|Y)\} - E_{\theta_0}\{2l(\theta_0|Y)\} \simeq p \quad (172)$$

となる。これは $\hat{\theta}$ がサンプルデータセット毎に $Y = \{Y_i\}$ に依存して推定されると、このような差が生じることを意味する。

では、ある 1 つのサンプルデータセットから推定された $\hat{\theta}_1$ を他のサンプルデータセットに適用するとどうなるだろうか。パラメータの真値 θ_0 での対数尤度関数の形について考えると、1 つのデータセット $\{Y_i\}$ から得られる $2l(\theta|Y)$ の頂点は、データの変動のため θ_0 より若干ずれた箇所にある。頂点にあたるのが最尤推定値 $\hat{\theta}_1$ であり、この高さの違いの平均が p である。一方、真の分布の下での期待対数尤度 $E_{\theta_0}\{2l(\theta|Y)\}$ は、 $\theta = \theta_0$ において最大値をとる。2 つの関数 $2l(\theta|Y)$ と $E_{\theta_0}\{2l(\theta|Y)\}$ の 2 次微分の形を頂点の近くで比較すると、多くの通常利用されるモデルにおいては、 N が大きいときにはほぼ等しいと期待できる。期待対数尤度の $\hat{\theta}_1$ における値 $E_{\theta_0}\{2l(\hat{\theta}_1|Y)\}$ について考えると、次のことが導かれる。

1. $2l(\theta_0|Y)$ は $2l(\hat{\theta}_1|Y)$ より平均して p 小さい
2. $2l(\theta|Y)$ と $E_{\theta_0}\{2l(\theta|Y)\}$ との頂点付近の 2 次微分はほぼ等しい。
3. 前項から $E_{\theta_0}\{2l(\hat{\theta}_1|Y)\}$ は $E_{\theta_0}\{2l(\theta_0|Y)\}$ より平均して p 小さい。
4. これらをまとめると、 $E_{\theta_0}\{2l(\hat{\theta}_1|Y)\}$ は $2l(\hat{\theta}_1|Y)$ より平均して $2p$ 小さい。

$E_{\theta_0}\{2l(\hat{\theta}_1|Y)\}$ は、将来得られるであろう新たなデータセットの分布を $f(Y|\hat{\theta}_1)$ で予測したときの平均的な良さを表していると解釈できる。そして、これが最尤推定値を代入したときの対数尤度の 2 倍 $2l(\hat{\theta}_1|Y)$ より平均して $2p$ だけ小さいことが示された。AIC はこの減少分を考慮して、一般的に

$$AIC = -2l(\hat{\theta}|Y) + 2p \quad (173)$$

と定義される。この値は、最尤推定によって得られた分布が将来のデータをどれだけ良く予測するかの基準を与えるものと解釈できる。つまり AIC の小さい値を持つモデルが、より良いモデルであるとみなせる。AIC は最尤推定値における対数尤度が大きい程小さく、またモデルパラメータ数が小さければ小さくなる。

28.3 線形モデルの最尤推定

次の線形モデルを考える。

$$y = X\beta + \varepsilon \quad (174)$$

ここで、誤差成分 ε の各要素 $\varepsilon_i, (i = 1, \dots, N)$ は互いに独立に同一分散の正規分布 $N(0, \sigma^2)$ に従うものとする。この仮定のもとでモデルの尤度は次の式であらわされる。

$$f(\mathbf{y}|\boldsymbol{\beta}, \sigma^2) = \prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(y_i - \mu_i)^2}{2\sigma^2}\right) \quad (175)$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \exp\left(\sum_{i=1}^N \frac{-(y_i - \mu_i)^2}{2\sigma^2}\right) \quad (176)$$

ただし、ここで $\mu_i = \mathbf{x}_i\boldsymbol{\beta}$ である (\mathbf{x}_i は \mathbf{X} の第 i 行目)。

この値を最大にする $\boldsymbol{\beta}$ は、最小 2 乗法による次の推定量 $\hat{\boldsymbol{\beta}}$ である。

$$\hat{\boldsymbol{\beta}}_{ML} = \hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \quad (177)$$

また、 σ^2 については

$$\hat{\sigma}_{ML}^2 = \sum_{i=1}^N (y_i - \mu_i)^2 / N = \|\mathbf{y} - \hat{\mathbf{y}}_{ML}\|^2 / N \quad (178)$$

となる。これを (176) に代入すると、

$$f(\mathbf{y}|\hat{\boldsymbol{\beta}}, \hat{\sigma}_{ML}^2) = \left(\frac{1}{\sqrt{2\pi}}\right)^N (\hat{\sigma}_{ML}^2)^{-N/2} \exp\left(\sum_{i=1}^N \frac{-(y_i - \hat{y}_i)^2}{2\hat{\sigma}_{ML}^2}\right) \quad (179)$$

$$= \left(\frac{1}{\sqrt{2\pi}}\right)^N (\hat{\sigma}_{ML}^2)^{-N/2} \exp(-N/2) \quad (180)$$

この対数をとって 2 倍すると

$$2l(\hat{\boldsymbol{\theta}}|Y) = \text{constant} - N \log \frac{\|\mathbf{y} - \hat{\mathbf{y}}_{ML}\|^2}{N} \quad (181)$$

となる。これを (173) に代入すると、(165) が得られる。

また、もし σ^2 が既知であるとするなら、 $\hat{\mathbf{y}}_{ML}$ は上と同一であるので

$$2l(\hat{\boldsymbol{\theta}}|Y) = \text{constant} - \frac{\|\mathbf{y} - \hat{\mathbf{y}}_{ML}\|^2}{\sigma^2} \quad (182)$$

であり、AIC はつぎのようになる。

$$\text{AIC} = \frac{\|\mathbf{y} - \hat{\mathbf{y}}_{ML}\|^2}{\sigma^2} + 2q + \text{定数} \quad (183)$$

29 ロジスティック回帰分析 (Logistic Regression Analysis)

回帰分析 (分散分析・共分散分析を含む) は、被説明変数が連続な値を持つ場合の方法であるが、時には 0-1 反応または比率を被説明変数とする必要が生じる場合がある。ロジスティック回帰は、このような分析のための方法である。

線形の回帰分析は、説明変数 $\mathbf{X} = (X_1, \dots, X_p)$ を固定した場合の Y の平均をつぎのような式で表す。

$$E(Y|\mathbf{X}) = \mu_X = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p \quad (184)$$

さらに、 Y は μ_x を平均とする正規分布 $N(\mu_x, \sigma^2)$ に独立に従うと仮定している。

ロジスティック回帰は、上の2点をつぎのように変更したものである。まず、説明変数 X を固定した際、被説明変数 Y は2項分布 $Bin(m_x, \pi_X)$ に独立に従うとする。さらに $\pi_X = E(Y|X)$ が、つぎのような式で表されると仮定する。

$$\text{logit}(\pi_X) = \log \frac{\pi_X}{1 - \pi_X} = \eta_X = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \quad (185)$$

上の式は、ロジット (logit) 変換と呼ばれる。一方、 $\log \frac{\pi_X}{1 - \pi_X} = \eta_X$ の逆関数は、

$$\pi_X = \frac{\exp(\eta_X)}{1 + \exp(\eta_X)} \quad (186)$$

となる。これはロジスティック (logistic) 関数と呼ばれる。 η_X の値が $-\infty$ から $+\infty$ まで変化するとき、 π_X は0から1まで変化する。

29.1 最尤法による推定

つぎに考えるべきは、データからどのようにして $\hat{\beta}$ を推定すべきかである。線形の回帰分析の場合には、最小2乗基準によりつぎの式

$$\sum_{i=1}^N (y_i - \hat{y}_i)^2 \quad (187)$$

を最小化することによって、 $\hat{\beta}$ を求めた。単純に考えるなら、 i 番目の観測値に対応する π_i と y_i/m_i を考え、 π_i と y_i/m_i の残差2乗和を最小化する方法が考えられる。しかし、もっとよい方法を理論的に構成することが可能である。

2項分布 $Bin(m, \pi)$ は平均 $m\pi$ 、分散 $m\pi(1 - \pi)$ である。分散の大きさが m と π に依存して変化することを考慮する必要がある。現在、一般的に用いられている推定方法は最尤法 (maximum likelihood method) と呼ばれる方法である。最尤法の一般論はつぎのようなものである。

データを生み出すと仮定される確率密度関数を $f(Y|\theta)$ とする。ここで、 Y は確率変数であり、 θ は分布を特徴づけるモデルパラメータ (母数) である。正規分布の場合であれば、平均と分散が母数であり、2項分布の場合は反応確率 π が相当する。実際に観測されたデータを y_1, \dots, y_n とし、これらは互いに独立に分布していると仮定する。 $f(y_i|\theta)$ の値は、 y_i が指定された θ のもとで、どの程度生じやすいかをあらわすものと解釈できる。これを逆に θ の関数としてみるなら、特定の y_i について、その値を θ がどの程度生じやすいかの指標になっていると考えられる。これを尤度 (likelihood) と呼び多くの場合 $L(\theta|y_i)$ と表記する。 θ の推定値として尤度を最大にする値を採用する方法を最尤法 (maximum likelihood method) と呼ぶ。

観測データ全体を考えると、独立性を仮定するとこれらの同時分布は $f(y_1|\theta) \times \cdots \times f(y_n|\theta)$ であり、これを $L(\theta|y_1, \dots, y_n)$ と解釈する。しばしば、計算上の問題と理論的な要請から、尤度そのものではなく、尤度の対数を取り扱われる。これを対数尤度 (log-likelihood) と呼び、 $l(\theta|y_1, \dots, y_n)$ と表記する。対数尤度について、つぎのような式が成立する。

$$l(\theta|y_1, \dots, y_n) = \log L(\theta|y_1, \dots, y_n) = \sum_{i=1}^n \log f(y_i|\theta) \quad (188)$$

30 ロジスティック回帰の例

表 20 は、Dobson の教科書の表 8.2 である。このデータは薬品 (二硫化炭素ガス) を与えたあと、5 時間後の昆虫の死亡数を表している。この例では、説明変数は定数項の他には、薬品の投与量 (X_1) のみである。個体数が、2 項分布 $Bin(m, \pi)$ の試行数 m であり、昆虫の死亡数が被説明変数 Y である。

モデルはつぎの 2 つの性質を持つ。

1. $Y_i \sim Bin(m_i, \pi_i)$, ($i = 1, \dots, n$)
2. $\text{logit}(\pi_i) = \beta_0 + \beta_1 x_i$, ($i = 1, \dots, n$)

係数の推定値を表 21 に示す。

投与量 ($\log_{10} CS_2 mg/l$)	個体数	死亡数
1.6907	59	6
1.7242	60	13
1.7552	62	18
1.7842	56	28
1.8113	63	52
1.8369	59	53
1.8610	62	61
1.8839	60	60

	最尤推定値	標準誤差
β_0	-60.7175	5.1807
β_1	34.2703	2.9120

分析すべきデータを y_i, m_i, x_i ($i = 1, \dots, n$) とする。ここで、 y_i は i 番目のデータにおける反応件数 (成功など) であり、 m_i は i 番目のデータの総数である。また、 x_i は、説明変数の値を要素とする $p + 1$ 次元のベクトル $(1, x_{i1}, \dots, x_{ip})$ とする。

この場合モデルの尤度は

$$L(\beta | \mathbf{m}, \mathbf{y}) = \prod_{i=1}^n \left\{ \binom{m_i}{y_i} \pi_i^{y_i} (1 - \pi_i)^{m_i - y_i} \right\} \quad (189)$$

である。ただし、

$$\pi_i = \frac{\exp(\beta^T x_i)}{1 + \exp(\beta^T x_i)} \quad (190)$$

であり、これはつぎのように書き換えられる。

$$\log \frac{\pi_i}{1 - \pi_i} = \eta_i = \beta^T x_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} \quad (191)$$

また対数尤度は、

$$l(\beta|m, y) = \text{定数} + \sum_{i=1}^n \{y_i \log \pi_i + (m_i - y_i) \log(1 - \pi_i)\} \quad (192)$$

となる。これを最大にする β が最尤推定値 $\hat{\beta}$ となる。

30.1 仮説検定

ある回帰係数がゼロであるとの仮説 $H_0: \beta_{i_1} = \dots = \beta_{i_{p-q}} = 0$ を検定するには、(163) を用いる。つまり、制約されないモデルにおける最尤推定値を $\hat{\beta}$ とし、制約されたモデルにおける最尤推定値を $\tilde{\beta}$ とすると、

$$2l(\hat{\beta}|y) - 2l(\tilde{\beta}|y) \sim \chi_{p-q}^2 + O(1/n) \quad (193)$$

である。ここで、 p は制約されないモデルにおける推定パラメータの数であり、 q は制約されたモデルにおける推定パラメータの数である。この値を χ^2 分布と比較し、通常は上側 5% に入れば帰無仮説が棄却できるとする。

30.2 なぜロジット変換を用いるのか？

2項分布によって表現されるモデルの平均の構造を、いくつかの説明変数によって表現する場合、0-1 の区間を $-\infty$ から $+\infty$ までの区間に変換する関数としては様々なものが考えられる。ロジット変換のほかにもっとも代表的なのは、プロビット変換 $\Phi^{-1}(x)$ (probit function) である。これは正規分布の分布関数 (累積密度関数) の逆関数をリンク関数として用いるものである。

いくつか選択肢があるなかで、ロジット変換が多用される理由は計算が簡単であることと、つぎに述べるような特徴があるからである。ここで示す特徴とは、データのサンプリングに関係するものである。ロジット変換を用いたモデルは、前向きサンプリングと後ろ向き (回顧的 retrospective) サンプリングの両者に同時に適用しうるという性質を持っている。喫煙と肺ガンの関係を調べる場合、最初にある調査集団 (サンプル) を設定し、それらの被験者を追跡することによって肺ガンの発生率を調べるのが前向きサンプリングである。一方、ある時点で肺ガンの患者と健康な被験者を集め、これらの集団について喫煙の有無をさかのぼって調査するのが、後ろ向きサンプリングである。

被説明変数のカテゴリーを D と \bar{D} であらわすことにする。つまり $Y = 1$ であることを D とし、 $Y = 0$ であることを \bar{D} とする。また、ここでは簡単のために説明変数は 0-1 の 2 値を持つ説明変数 x が 1 つの場合を考える。前向きサンプリングでは、つぎのいずれかの方法でデータを得る。

1. 調査対象となる母集団から無作為に被験者をサンプリングし、それらの各被験者について、説明変数 X (喫煙の有無など) の値を調べる。この後、これらの被験者の被説明変数に対応する特徴 (肺ガンの発病) について調査する。
2. 調査対象母集団を説明変数 x で層別し、各層からあらかじめ決められた数の被験者をサンプリングする。これらの被験者について、被説明変数に対応する特徴を調査する。

このようなサンプルの特徴は、線形の回帰分析や分散分析において、被説明変数を比率に置き換えた場合に相当し、特に変わったところはない。ただし、当然のことであるが Y とサンプリング比率の間に関係があってはならない。

後ろ向きサンプリングでも、上に示した2つの方法（無作為サンプリングと層別サンプリング）の両者が考えられるが、通常は層別サンプリングが用いられる。これは、肺ガンなどの疫学調査では特定の病気の発病率は小さく、無作為サンプリングを用いて信頼に足る分析を行うためには、膨大な被験者がしばしば必要となるからである。これは、線形の回帰分析でいえば被説明変数 Y の値で集団を層別し、各層からあらかじめ決められた件数をサンプルとして採用することに相当する。線形の回帰分析において、このような分析はほとんど行われず、一般的にはこのようなサンプリングに基づくデータを説明するモデルが、前向きサンプリングの場合と同一のものである保証はない。

しかし、ロジスティック回帰においては、サンプリング方式の違いによらず、説明変数の影響力は等しく推定される。この理由を以下で説明する。

後ろ向きサンプリングを説明するために、サンプルとして採用されることをあらわすダミー変数を Z とする。 $Z = 1$ は想定している母集団のなかで、後ろ向きサンプリングによって採用されることをあらわし、 $Z = 0$ は採用されないことをあらわす。

後ろ向きサンプリングによって得られたデータが、表 22 のようなものであるとする。層別の前

表 22: データ件数

	\bar{D}	D	計
$X = 0$	n_{00}	n_{01}	$n_{0.}$
$X = 1$	n_{10}	n_{11}	$n_{1.}$
計	$n_{.0}$	$n_{.1}$	$n_{..}$

向きサンプリングにおいては $n_{0.}, n_{1.}$ が固定されており（層別しない場合にはこれらの値は確率変数になる）、 $n_{0.}, n_{1.}$ は確率変数になる。一方、層別の後ろ向きサンプリングの場合には、 $n_{.0}$ と $n_{.1}$ はあらかじめ決められた数であり、 $n_{0.}$ と $n_{1.}$ が確率変数になる。

後ろ向きサンプリングを行う場合、重要なことは \bar{D}, D いずれの層内においても説明変数 X とサンプリング確率の間には関係があってはならないということである。つまり、 $Pr(Z = 1|\bar{D})$ と $Pr(Z = 1|D)$ は当然異なり得るが、 \bar{D} および D の層の内部では、

$$Pr(Z = 1|\bar{D}) = Pr(Z = 1|\bar{D}, X = 0) = Pr(Z = 1|\bar{D}, X = 1) \quad (194)$$

および

$$Pr(Z = 1|D) = Pr(Z = 1|D, X = 0) = Pr(Z = 1|D, X = 1) \quad (195)$$

が成立していなければならない。もし、これらが成立しないのなら、被説明変数 Y と説明変数 X の間に、どのような関係でも見かけ上つくりだすことが可能になってしまう。説明変数 X の Y に及ぼす影響を正しく求めるためには、この条件は必須である。

ここで、 $Pr(Z = 1|\bar{D}) = p_0$, $Pr(Z = 1|D) = p_1$ としよう。後ろ向きサンプリングによって得られたデータに基づいて、 X の Y に及ぼす影響をロジスティック回帰で求めることは、つぎの式を求めることに相当する。

$$\log \frac{Pr(D|Z = 1, \mathbf{x})}{Pr(\bar{D}|Z = 1, \mathbf{x})} = \beta^T \mathbf{x} \quad (196)$$

ここで、条件付き確率についてのベイズの定理 (Bayes's theorem) を用いると、

$$Pr(D|Z = 1, \mathbf{x}) = \frac{Pr(Z = 1|D, \mathbf{x})Pr(D|\mathbf{x})}{Pr(Z = 1|D, \mathbf{x})Pr(D|\mathbf{x}) + Pr(Z = 1|\bar{D}, \mathbf{x})Pr(\bar{D}|\mathbf{x})} \quad (197)$$

$$= \frac{p_1 \pi_x}{p_1 \pi_x + p_0 (1 - \pi_x)} \quad (198)$$

$$= \frac{\exp(\alpha + \beta_0^T \mathbf{x})}{1 + \exp(\alpha + \beta_0^T \mathbf{x})} \quad (199)$$

となる。ここで、 $\pi_x = \exp(\beta_0^T \mathbf{x}) / \{1 + \exp(\beta_0^T \mathbf{x})\}$ であるとする。すなわち

$$\text{logit} \pi_x = \log \frac{Pr(D|\mathbf{x})}{Pr(\bar{D}|\mathbf{x})} = \beta_0^T \mathbf{x} \quad (200)$$

であり、 β_0 は母集団におけるロジスティック回帰係数を表す。また $\alpha = \log(p_1/p_0)$ である。

上の計算から、

$$\log \frac{Pr(D|Z=1, \mathbf{x})}{Pr(\bar{D}|Z=1, \mathbf{x})} = \alpha + \beta_0^T \mathbf{x} = \beta^T \mathbf{x} \quad (201)$$

となる。 β と β_0 で異なるのは定数項 β_0 (切片 intercept) だけであり、それ以外の回帰係数 β_i , ($i = 1, \dots, p$) は同一である。これは後向きの層別サンプリングによるデータについて、ロジスティック回帰を行なって得られる回帰係数は、前向きの (無作為または層別) サンプリングによって得られる回帰係数と、定数項以外は理論的には同一であることを表している (もちろん実際の推定値はサンプルに依存するので、完全に同一ではない)。

30.3 SAS PROC GENMOD による分析

表 20(Dobson 表 8.2) のデータを分析するプログラム例を示す。データは以下のファイルに記述してある。

/home1/otsu/cl14a0/Dobson/tb8_2.dat

SAS でロジスティック回帰を行なえるプロシジャには、LOGISTIC, CATMOD, および GENMOD があるが、ここでは汎用性の高い GENMOD を用いる。データの入力形式は、GLM プロシジャに準じる。

```

行
1 /* PostScript 出力の場合は、最初の 2 行を使う。
2 画面出力の場合は、3 行目を使う。
3 行頭の * はつぎにあらわれる ; までコメント行
4 であることを示す。
5 */
6 *filename gsasfile 'temp1.ps';
7 *goptions device=ps gaccess=gsasfile gsfmode=replace;
8 goptions device=xcolor;
9 options linesize=80;
10
11 data tb82;
12     infile 'tb8_2.dat';
13     input x m y;
14
15 proc genmod data=tb82;
16     make 'obstats' out=tb82out;
17     model y/m = x /
18     link=logit dist=binomial obstats
19     lrcl type1 type3;
20
21 proc print data=tb82out;

```

```

22
23 data grdat1;
24     set tb82;
25     set tb82out(keep=pred lower upper reschi);
26     ratio = y/m;
27 run;
28
29 /* PostScript の場合には color=black を指定する。*/
30 symbol1 color = yellow interpol=none value=dot;
31 symbol2 color = yellow interpol=spline value=none line=1;
32 symbol3 color = yellow interpol=spline value=none line=2;
33
34 axis1 label = ('Dose log10 CS2 mg/l') minor=none;
35 axis2 label = (angle=-90 rotate=90 'Response Ratio') minor=none;
36 axis3 label = (angle=-90 rotate=90 'Pearson Residual') minor=none;
37
38 proc gplot data=grdat1;
39     plot ratio*x=1 pred*x=2 lower*x=3 upper*x=3 /
40     overlay frame haxis=axis1 vaxis=axis2 ;
41 run;
42 proc gplot data=grdat1;
43     plot reschi*x=1 / vref=0.0
44     overlay frame haxis=axis1 vaxis=axis3 ;
45 run;

```

SAS プログラムに解説を加える。GENMOD プロシジャは、一般化線形モデル (generalized linear models) と呼ばれる統計モデルを用いた分析を行なうためのものである。このモデルの特殊ケースの一つがロジスティック回帰である。また、既に学んだ通常の線形モデル (重回帰と分散分析、共分散分析) も行なうことができる。さらに、後期で学習する対数線形モデルの分析も、このプロシジャを用いて行なえる。

1. 上の例では指定していないが、離散値をとる変数 (分散分析の要因にあたるもの) がある場合には、モデルステートメントの前に

```
class x1 x2;
```

などのように指定する。

2. 16 行目 proc genmod の make ステートメントは、'obstats' で指定される計算結果 (各オブザベーション i.e. サンプル) についての詳細な情報を SAS データセット tb82out に出力することを指定している。
3. 17 行目 model ステートメントの一般的な書式は

```
model 被説明変数 = 説明変数 1 説明変数 2 .... / オプション ... ;
```

というものである。ここで説明変数には、glm プロシジャの場合と同様に交互作用項 (x1 * x2 など) も指定することができる。

4. 17 行目 モデルステートメントにおける y/m はロジスティック回帰に特有の被説明変数の指定法である。 y/m で定義される比率が被説明変数であることを意味する。ただし、直接に比率を指定したのでは、そのオブザベーションにあてはめるべき 2 項分布の回数を指定する母数 ($Bin(m, \pi)$ の m) が分からない。ここに、示すような指定法で、 y と m に相当する 2 つの変数を特定することができる。等号 (=) の右辺は、説明変数である。

5. モデルステートメントの各オプションはつぎのような役割を持っている。

link=logit : ロジスティック回帰においては、 $\text{logit}(E(Y|x)) = \beta^T x$ という関係を仮定している。このような $E(Y|x)$ と x の 1 次式を関係づける関数をリンク関数 (link function) と呼ぶ。この例での指定は、リンク関数を logit に指定している。

dist=binomial : 被説明変数の分布 $Y|x$ を 2 項分布と指定する。

obstats : 観測値ごとの詳細情報を出力することを指定。

lrci : 尤度比から計算されるパラメータの信頼区間を出力する。

type1 : 線形モデル (GLM プロシジャ) の場合と同様に、逐次的に変数が増加させて検定を行なう。検定法は対数尤度比を用いた χ^2 検定。

type3 : 線形モデルの場合と同様に、ある変数およびその変数を含む交互作用項までも除いたモデルとの比較のための検定を行なう。検定法は対数尤度比を用いた χ^2 検定。

6. 23-27 行目 結果のグラフ出力をするために、あらたな SAS データセットをつくる。set ステートメントにより 2 つのデータセットに含まれる変数を一つのデータセットにまとめる。また、新たに変数 ratio を計算して求める。変数 pred は $\hat{\pi}_i$ であり、lower, upper は π_i の信頼区間の下限と上限である。(Y_i の信頼区間ではないことに注意。こちらは m_i に依存する。)

7. 30-32 行目 結果のグラフ出力のために、表示に用いる記号と線種を指定する。interpol は補間の指定であり、iterpol=none は点を結ばないこと、また iterpol=spline はスプラインとよばれる曲線による補間法によりデータをつなぐことを指定する。value は点の表示の指定であり、value=dot はデータの箇所に点を表示すること、value=none は特に記号を表示しないことを意味する。また、line=1 は補間の表示に実線を用いること、line=2 は破線を用いることを指定する。詳しくはマニュアル SAS/Graph Software を参照。

8. 34-36 行目 グラフ表示に用いる座標軸のラベルを指定している。

9. 38-40 行目 グラフの表示の実行 (データと推定値)。plot ステートメントの書式はつぎの通りである。

plot 縦軸の変数 * 横軸の変数 = 表示記号 ... /オプション

ここで、 = 表示記号の部分は特に指定しなくてもよい。オプション overlay は、複数のグラフを同一の画面に表示することを指定する。また frame は枠表示の指示である。

10. 42-44 行目 グラフ表示の実行 (残差)。plot ステートメントと vref=0.0 は縦軸の値 0.0 の位置に水平線をひく。横軸の値で垂直線をひく時には href=0.0 などとする。

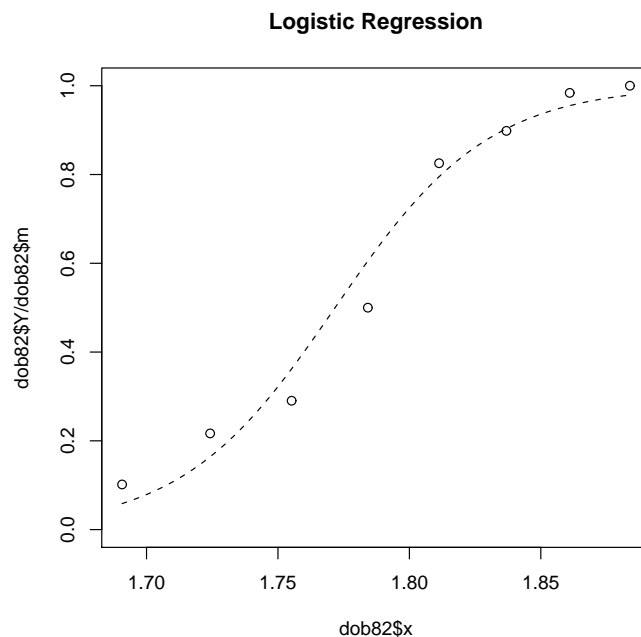


図 22: ロジスティック回帰の結果

以下は本格的な、線形モデルおよび対数線形モデルの教科書。

参考文献

- [1] Bishop, Y.M.M., Fienberg, S.E., & Holland, P.W. (1977). *Discrete Multivariate Analysis: Theory and Proactice*, Cambridge, MA: MIT Press.
- [2] McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models, 2nd ed.*, London: Chapman & Hall.
- [3] Rao, C.R. (1973). *Linear Statistical Inference and Its Applications, 2nd ed.*, New York: Wiley. 奥野他訳 (1977) 統計的推測とその応用, 東京図書.
- [4] 竹村彰通 (1991). 現代数理統計学, 創文社.

30.4 ロジスティック回帰の出力例

モデル適合度の項にあらわれる基準は、それぞれつぎのようなものである。線形の(重)回帰の場合と、ロジスティック回帰の場合を示す。

- Deviance: $D(\hat{\mu}; \mathbf{y})$

線形 :

$$\sum_i (y_i - \hat{\mu}_i)^2$$

ロジスティック :

$$2l(\hat{\pi}; \mathbf{y}) - 2l(\hat{\pi}; \mathbf{y})$$

ここで $\hat{\pi}$ は、 y_i/m_i を表す。 $l(\dots)$ は対数尤度をあらわす。

- Scaled Deviance: $D^*(\hat{\mu}; \mathbf{y})$

線形 :

$$\frac{\sum_i (y_i - \hat{\mu}_i)^2}{\phi}$$

。通常、 ϕ は残差分散の推定値。

ロジスティック :

$$\frac{2l(\hat{\pi}; \mathbf{y}) - 2l(\hat{\pi}; \mathbf{y})}{\phi}$$

ここで ϕ は過大分散パラメータと呼ばれるもの。特に指定しなければ 1。

- Pearson Chi-Square :

$$X^2 = \sum_i (y_i - \hat{\mu}_i)^2 / V(\hat{\mu}_i)$$

線形 : $V(\mu_i)$ は μ_i の値を特定した場合の、 Y_i の分散。線形の場合は、一般的には $\hat{\sigma}^2$ 。

ロジスティック : μ_i が、 $m_i \pi_i$ となる。ただし、この場合は $V(\hat{\mu}_i) = m_i \hat{\pi}_i (1 - \hat{\pi}_i)$ 。(比率の推定値の分散は、 $V(\hat{\pi}_i) = \hat{\pi}_i (1 - \hat{\pi}_i) / m_i$ である。)

- Scaled Pearson X2: X^2/ϕ を示す。ここで、 ϕ は過大分散パラメータ。特に指定しない場合は 1。

SAS システム

1

18:27 Monday, October 16, 2000

The GENMOD Procedure

Model Information

Description	Value	
Data Set	WORK.TB82	
Distribution	BINOMIAL	
Link Function	LOGIT	
Dependent Variable	Y	
Dependent Variable	M	
Observations Used	8	
Number Of Events	291	死亡総数
Number Of Trials	481	昆虫の総数

Parameter Information

Parameter	Effect	
PRM1	INTERCEPT	定数項
PRM2	X	

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	6	11.2322	1.8720
Scaled Deviance	6	11.2322	1.8720
Pearson Chi-Square	6	10.0268	1.6711
Scaled Pearson X2	6	10.0268	1.6711
Log Likelihood	.	-186.2354	.

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Std Err	ChiSquare	Pr>Chi
INTERCEPT	1	-60.7175	5.1807	137.3562	0.0001
X	1	34.2703	2.9121	138.4879	0.0001
SCALE	0	1.0000	0.0000	.	.

NOTE: The scale parameter was held fixed.

Likelihood Ratio Based Confidence Intervals For Parameters
(尤度比を用いたパラメータの信頼区間推定)

Two-Sided Confidence Coefficient: 0.9500

Parameter	Confidence Limits	Parameter Values	
		PRM1	PRM2
PRM1	Lower	-71.4423	40.2986
PRM1	Upper	-51.0789	28.8557
PRM2	Lower	28.8540	28.8540
PRM2	Upper	40.3005	40.3005

SAS システム

2

18:27 Monday, October 16, 2000

Observation Statistics

Y	M	Pred	Xbeta	Std	HessWgt	Lower
6	59	0.0586	-2.7766	0.2870	3.2548	0.0343
13	60	0.1640	-1.6286	0.2051	8.2274	0.1160
18	62	0.3621	-0.5662	0.1472	14.3213	0.2984
28	56	0.6053	0.4277	0.1318	13.3789	0.5422
52	63	0.7952	1.3564	0.1620	10.2610	0.7386
53	59	0.9032	2.2337	0.2147	5.1567	0.8597
61	62	0.9552	3.0596	0.2737	2.6534	0.9258
60	60	0.9790	3.8444	0.3338	1.2307	0.9605

Observation Statistics

Upper	Resraw	Reschi	Resdev
0.0985	2.5425	1.4093	1.2837
0.2268	3.1583	1.1011	1.0597

0.4310	-4.4514	-1.1763	-1.1961
0.6651	-5.8976	-1.6124	-1.5941
0.8421	1.9042	0.5944	0.6061
0.9343	-0.2909	-0.1281	-0.1272
0.9733	1.7778	1.0914	1.2511
0.9890	1.2570	1.1331	1.5940

31 対数線形モデル

31.1 ポアソン (Poisson) 分布

ここまで扱ったデータは、連続値や比率が主であった。連続値をとる確率変数のなかで最も代表的なものが正規分布であり、また2つの選択肢(0-1値)をとる確率変数の和が2項分布であった。

もう一種類のしばしば現れるデータは頻度であろう。確率的に変動する頻度を表す確率分布のなかで、最も代表的なものがポアソン分布である。

タイルが敷き詰められている舗道を考えよう。タイルはすべて10cm角、すなわち 100cm^2 の大きさであるとする。小雨が降っており、1分間あたり平均するとタイル1枚には μ 個の雨粒が落ちる。ある1分間に限ると、各々のタイルに落ちる雨粒の個数はどのようなものになるだろうか。ここで $\mu = 2$ と仮定してみよう。平均すると1枚あたりに落ちる雨粒の個数は2であるが、当然のことながらすべてのタイルに均等に2個ずつの雨粒が落ちる訳ではない。中には1個もあたらないものがあるかも知れないし、また中には3個、4個と多くの雨粒が落ちるものもあるだろう。この雨粒の個数を表す確率変数は、ポアソン分布と呼ばれる次の分布に従う。ここで x の値はゼロ以上の整数である。

$$Po(x|\mu) = e^{-\mu} \frac{\mu^x}{x!}, \quad (x = 0, 1, 2, \dots) \quad (202)$$

この分布は、2項分布 $Bin(m, \pi)$ において $m \times \pi$ の値を μ に固定し、その上で m の値 (試行数) を大きくしたものとみることができる。実際、2項分布は

$$Pr(x) = Bin(x|m, \pi) = \binom{m}{x} \pi^x (1 - \pi)^{(m-x)} \quad (203)$$

で与えられるが、上の条件を仮定すると $\pi = \mu/m$ であり、

$$Pr(x) = \left(\prod_{k=1}^x \frac{(m-k+1)}{km} \right) m^x (\mu/m)^x (1 - \mu/m)^{(m-x)} \quad (204)$$

であり、 m を十分大きくするとポアソン分布の式が得られることが分かる。

ポアソン分布には次のような特徴がある。

1. $Po(x|\mu)$ の平均は μ であり、分散もまた μ である。
2. 2つの確率変数 X と Y とが各々平均 μ_1 、 μ_2 のポアソン分布に独立に従うとするならば、2つの確率変数の和 $Z = X + Y$ は平均 $\mu_1 + \mu_2$ のポアソン分布に従う。

2番目の性質は、最初に紹介したポアソン分布の性質から直観的に分かる。片方のタイルに落ちる雨粒の個数が平均 μ_1 のポアソン分布であり、もう片方が平均 μ_2 であるとする、両者の合計は2枚を合わせたものの上に落ちる雨粒の個数とみなせるから、ポアソン分布になるだろうと推測がつく。1番目の性質からポアソン分布の標準偏差は、平均が大きい程大きく、また平均に関する相対的な標準偏差、つまり $\sqrt{\text{Var}(X)}/E(X) = 1/\sqrt{\mu}$ は、平均が大きい程小さくなることが分かる。

頻度を被説明変数とする分析には、後に説明する対数線形モデル(ポアソン回帰モデルと呼ばれることもある)を用いるのが標準的であるが、頻度を連続的な数値とみなして通常の回帰分析や分散分析を利用する場合もある。この場合、これらのモデルでは、被説明変数 Y の分散がどのような条件においても一定であると仮定していることに注意しなければならない。頻度データがポアソン分布に従うならば(これは多くの場合自然な仮定である)、期待値が大きいほど分散も大きくな

る。このため、 Y の値と直接使うのではなく、頻度の平方根 \sqrt{Y} を使った方がよい。 Y の期待値 μ が大きい場合には、近似的に次の式がなりたつ。

$$E(\sqrt{Y}) \simeq \mu^{1/2}, \quad \text{Var}(\sqrt{Y}) \simeq 1/4 \quad (205)$$

この変換により、近似的に分散が平均値によらず一定になることがわかる。

ポアソン分布は2項分布の極限として表されることを先に説明したが、逆に2項分布を2つのポアソン分布を用いて導くこともできる。ここで、 X と Y とが各々平均 μ_1, μ_2 のポアソン分布であり、互いに独立であるとする。 $Z = X + Y$ とおき、 $Z = m$ との制約下における X の条件つき分布は、平均 μ_1/μ_2 の2項分布 $Bin(m, \mu_1/\mu_2)$ になる。

31.2 一般化線形モデル

通常の回帰分析や分散分析で用いるモデル（線形モデル）は、次のような特徴を持っている。

1. 一つの被説明変数（従属変数） Y に、1つまたは複数の説明変数（独立変数） X_1, \dots, X_p が影響を及ぼしている。
2. 被説明変数の条件付き平均、つまり $E(Y|x)$ は、説明変数の1次式 $\mu_x = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$ であらわされる。
3. 被説明変数の条件付き分布は、一定の分散を持つ正規分布である。つまり、 $Y|x \sim N(\mu_x, \sigma^2)$ であり、 σ^2 は x の値による影響を受けない。ここで $Y|x$ は、 x を固定した条件の下での Y をあらわす。
4. Y の観測値 $y_i, (i = 1, \dots, N)$ は各々独立である。

先に説明したロジスティック回帰モデルは、上の2番目と3番目の仮定を次のように変更したものであった。

3. $\eta_x = \log(\mu_x)$ が、説明変数の1次式であらわされる。
4. Y を値が1と観測されたサンプル数とすると、 $Y|x, m_x$ は2項分布 $Bin(m_x, \mu_x)$ に従う。

対数線形モデル (log-linear model) は、これらの仮定を次のように変更したものである。

3. 被説明変数の条件つき平均 $\mu_x = E(Y|x)$ の対数 $\eta_x = \log \mu_x$ が、説明変数の1次式であらわされる。
4. $Y|x$ はポアソン分布 $Po(\mu_x)$ に従う。

つまり、説明変数が一定の値だけ変化すると、 Y の平均の比率がそれに対応して変化する。次の式がこの関係の表現である。

$$\eta_x = \log \mu_x = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p \quad (206)$$

このようなモデルを用いること理由は、次のようなものである。

1. 頻度データがポアソン分布すると仮定するのが自然である場合が多い。

2. 説明変数の効果は、頻度に線形の効果を持つと仮定するより、頻度の率に一定の効果を及ぼすと仮定する方が自然な場合が多い。

ためである。

線形モデル、ロジスティック回帰モデル、対数線形モデルの3つにおける条件付き平均の構造は、いずれも次のような形式で表現することができる。

$$\eta_x = g(\mu_x) = \beta^T x \quad (207)$$

ここで、 $\mu_x = E(Y|x)$ であるが、ロジスティック回帰のときは $\mu_x = E(Y/m_x|x)$ である。上にあげた3つのモデルは、いずれも一般化線形モデル (generalized linear model) と呼ばれるモデルの一例である。一般化線形モデルにおいて、被説明変数 Y の平均を変換する関数 g は、リンク関数 (link function) と呼ばれる。ロジスティック回帰においては、リンク関数はロジット関数であり、対数線形モデルにおいては対数関数である。

31.3 ロジスティック回帰による 2×2 表の分析

まず、 2×2 の頻度表を対数線形モデルによって分析する方法を示す。このデータにおいて性別と意見の関係を調べるには、独立性の χ^2 検定を行なう他に、ロジスティック回帰を用いることが可能であり、また対数線形モデルを用いることも可能である。

ロジスティック回帰によって分析するには、まず性別 X_1 を説明変数とする。 X_1 は男性のとき 0、女性のとき 1 をとるとする。また、回答 X_2 を被説明変数とし、Yes のとき 0、No または未決定のとき 1 とする。ここで、次のような平均の構造を仮定する。

$$\text{logit}(\mu_{x_1}) = \beta_0 + \beta_1 x_1 \quad (208)$$

μ_{x_1} は X_1 の値を x_1 とおいた場合の X_2 の平均、すなわち Yes の回答率である。もし性別と回答が独立ならば、性別が回答比率に影響を及ぼすことはない。これは $\beta_1 = 0$ が成立するというこ

$$\text{logit}(\mu_0) = \text{logit}(\mu_1), \text{すなわち } \log \frac{\mu_0}{1 - \mu_0} = \log \frac{\mu_1}{1 - \mu_1} \quad (209)$$

が成立することになる。これが成立するとオッズ比が 1 なので、性別と回答とが独立である。そこで、 x_1 がモデルを説明するために必要か否かの検定を行なえば、回答率と性別が独立であるか否かが分かる。

31.4 対数線形モデルによる 2×2 表の分析

対数線形モデルを用いて 2×2 表を分析するには、行と列を表す 2 つの変数 X_1 、 X_2 を説明変数とし、被説明変数 Y は、各セルにおけるデータの件数 n_{ij} とする。ここで、 X_1 と X_2 の値は 0, 1 を取るものとし、セルの添字もこれに合わせ $Y_{ij} = n_{ij}$ ($i, j = 0, 1$) と表記する。さらに Y_{ij} は平均 μ_{ij} のポアソン分布に従うとする。また、 X_1, X_2 によって分散分析の場合と同様の交互作用項を表すことにし、ここではこの 2 値変数を X_3 とおく。 $X_3 = X_1 \cdot X_2$ は X_1 が 1 でありかつ X_2 が 1 であるときにのみ 1 をとり、その他の場合には 0 である。

この例において、一番複雑なモデル (フルモデル) によって指定される平均の構造は

$$\eta_x = \log \mu_x = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \quad (210)$$

となる。より具体的に記述すると、

$$\eta_{00} = \beta_0 \quad (211)$$

$$\eta_{10} = \beta_0 + \beta_1 \quad (212)$$

$$\eta_{01} = \beta_0 + \beta_2 \quad (213)$$

$$\eta_{11} = \beta_0 + \beta_1 + \beta_2 + \beta_3 \quad (214)$$

となる。個々の係数の意味を検討すると、 β_0 は全体の件数の大きさを表し、 β_1 は女性の被験者が男性より全体でどれだけ多いかを表し、また β_2 は「No または未決定」の回答が Yes の回答より全体でどれだけ多いかを示す。交互作用項にあたる β_3 については、もし β_3 がゼロであるとする、次のような式が成り立つ。

$$\eta_{10} - \eta_{00} = \eta_{11} - \eta_{01} = \beta_1 \quad (215)$$

これを変形すると

$$\log \frac{\mu_{10}}{\mu_{00}} = \log \frac{\mu_{11}}{\mu_{01}} \quad (216)$$

となる。この式が成り立つとき、オッズ比は 1 である。すなわち、対数線形モデルにおける交互作用項の係数 β_3 がゼロであることが、 X_1 と X_2 とが独立であることに対応している。

31.5 対数線形モデルの推定

対数線形モデルの推定もロジスティック回帰と同様に最尤法によって求める。添え字 $i = 1, \dots, N$ がセルを表現するものとし、それに対応する頻度の観測値を y_i 、また説明変数 X_1, \dots, X_p の値を x_{i1}, \dots, x_{ip} とする。このとき、各セルに対応するポアソン分布の平均は、次の式であらわされる。

$$\mu_i = \exp(\eta_i) = \exp(\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}) \quad (217)$$

また、ポアソン分布の尤度は、各 y_i が独立であると仮定すると、

$$L(\boldsymbol{\mu}|\mathbf{y}) = \prod_{i=1}^N \exp(-\mu_i) \frac{\mu_i^{y_i}}{y_i!} \quad (218)$$

となる。この対数をとると、

$$l(\boldsymbol{\mu}|\mathbf{y}) = \sum_{i=1}^N (-\mu_i + y_i \log \mu_i - \log(y_i!)) \quad (219)$$

となる。ポアソン分布の平均 $\boldsymbol{\mu} = (\mu_1, \dots, \mu_N)$ は、 $\boldsymbol{\beta} = \beta_0, \dots, \beta_p$ の関数であるから、これらの値を調整して、 $l(\mathbf{y}|\boldsymbol{\mu})$ を最大化すれば、最尤推定値 $\hat{\boldsymbol{\beta}}$ が得られる。モデルが特定の条件を満たす場合には、 $\hat{\boldsymbol{\beta}}$ は \mathbf{y} の明示的な式として表されるが、一般的には簡単な式ではあらわされないため、コンピュータを用いて逐次近似により計算する。

死後の人生についての意見データに、交互作用項まで含むモデルを当てはめると、次のような推定値が得られる。この例は、交互作用項を含めるとフルモデルになるので、ある意味ではデータの書き換えともみなせる。上の例では、交互作用項がほぼ無視しうること、つまり性別で Yes の回答率に有意な差のないことがわかる。

係数	内容	推定値	標準誤差の推定値
β_0	全体の切片	5.9269	0.0516
β_1	女性	0.1484	0.0705
β_2	No または未決定	-1.0291	0.1006
β_3	交互作用項	-0.0558	0.1387

31.6 $2 \times 2 \times 2$ 表の分析

与えられているデータが、 $2 \times 2 \times 2$ の分割表である場合の、対数線形モデルによる分析法について説明を加える。

観測されたデータを n_{ijk} , ($i, j, k = 0, 1$) とする。説明変数は X_1, X_2, X_3 とすると、モデルによって表現される平均の構造は

$$\eta_{ijk} = \log E(\mu | X_1 = i, X_2 = j, X_3 = k) = \beta^T x \quad (220)$$

となる。ここで β と x とは、モデルに含まれる交互作用項の設定によって異なる。フルモデルすなわち、すべての変数とそれらの交互作用項を用いる場合には、要因は

$$\text{定数項, } X_1, X_2, X_3, X_1 * X_2, X_1 * X_3, X_2 * X_3, X_1 * X_2 * X_3 \quad (221)$$

の 8 つである。特に、 $2 \times 2 \times 2$ 表であれば各々 8 次元のベクトルになる。ここで、上の各要素の自由度は 1 である。

表 23: $2 \times 2 \times 2$ 表の要因

添え字	定数項	X_1	X_2	X_3	$X_1 * X_2$	$X_1 * X_3$	$X_2 * X_3$	$X_1 * X_2 * X_3$
(0,0,0)	1	0	0	0	0	0	0	0
(0,0,1)	1	0	0	1	0	0	0	0
(0,1,0)	1	0	1	0	0	0	0	0
(0,1,1)	1	0	1	1	0	0	1	0
(1,0,0)	1	1	0	0	0	0	0	0
(1,0,1)	1	1	0	1	0	1	0	0
(1,1,0)	1	1	1	0	1	0	0	0
(1,1,1)	1	1	1	1	1	1	1	1

より具体的にこの変数の値を記述すると、表 23 のようになる。対数線形モデルは、多くの場合、説明変数が離散的な変数（要因）であるデータについて適用されるが、これはモデルの持つ制限ではない。説明変数が連続変数であるとしても、上にしめした 0 - 1 変数を連続値に置き換えれば、分析可能である。離散変数と連続変数の交互作用は、連続値に 0 - 1 を掛け合わせた変数になる。

フルモデルの意味するものは、3 つの変数が互いに関係しあっているというものであり、これ以上の解釈がしやうがない。興味深いのは、3 次の交互作用項がゼロであり、その上で、2 次の交互作用項のいずれかがゼロとみなせる場合である。 $X_1 * X_3$ がゼロである場合は、 X_1 と X_2 は相互に関係し合い、また X_2 と X_3 とは関係しているが、 X_2 を固定すると X_1 と X_3 の間に直接の関係

はない。つまり、 X_2 を特定の値に固定した 2×2 表において、 X_1 と X_3 とは独立、すなわちオッズ比が 1 であることを示す。フルモデルにおいては、次のような式が成立する。

$$\eta_{000} = \log \mu_{000} = \beta_0 \quad (222)$$

$$\eta_{100} = \log \mu_{100} = \beta_0 + \beta_1 \quad (223)$$

$$\eta_{010} = \log \mu_{010} = \beta_0 + \beta_2 \quad (224)$$

$$\eta_{001} = \log \mu_{001} = \beta_0 + \beta_3 \quad (225)$$

$$\eta_{110} = \log \mu_{110} = \beta_0 + \beta_1 + \beta_2 + \beta_{12} \quad (226)$$

$$\eta_{101} = \log \mu_{101} = \beta_0 + \beta_1 + \beta_3 + \beta_{13} \quad (227)$$

$$\eta_{011} = \log \mu_{011} = \beta_0 + \beta_2 + \beta_3 + \beta_{23} \quad (228)$$

$$\eta_{111} = \log \mu_{111} = \beta_0 + \beta_1 + \beta_2 + \beta_3 + \beta_{12} + \beta_{13} + \beta_{23} + \beta_{123} \quad (229)$$

ここで、 $X_2 = 0$ における X_1 と X_3 の 2×2 表の対数オッズ比は、

$$\log \frac{\mu_{000}\mu_{101}}{\mu_{100}\mu_{001}} = \eta_{000} + \eta_{101} - \eta_{100} - \eta_{001} = \beta_{13} \quad (230)$$

であり、また $X_2 = 1$ における値は、

$$\log \frac{\mu_{010}\mu_{111}}{\mu_{110}\mu_{011}} = \eta_{010} + \eta_{111} - \eta_{110} - \eta_{011} = \beta_{13} + \beta_{123} \quad (231)$$

である。3 次の交互作用項がゼロ、つまり $\beta_{123} = 0$ であれば、2 つの 2×2 表でオッズ比は等しく、さらに $\beta_{13} = 0$ であれば、両者のオッズ比は 1 (対数オッズ比は 0) であり、 X_1 と X_3 とが、 X_2 を固定した各層において独立である。ここで、注意しなければならないのは、たとえ $\beta_{123} = \beta_{13} = 0$ が成立していたとしても、 X_2 によって層別を行なわない X_1 と X_3 の 2×2 表においては、オッズ比が一般的に 1 とはならないことである。

31.7 モデルの比較

対数線形モデルを用いて実際に分析を行うには、線形モデルの場合と同様に、2 つの点を考慮することになる。1 つは、特定のモデルを当てはめた結果得られる係数についての検討であり、もう一つは、ある変数をモデルから除いた際の、当てはまりの程度の違いについての検討である。

係数 β の推定は最尤法を用いて行うので、これらの推定値 $\hat{\beta}$ の分散の近似値が理論的に求まる。これを用いて、各々の $\hat{\beta}_i$ の分散を求めることができる。SAS などの出力において、パラメータの標準誤差 (standard error) として表示されているのは、このようにして求められたものである。ある変数 (または要因の水準) に対応する係数の値が、標準誤差の大きさを考慮するとさほど顕著に大きいとはいえないのなら、その変数が説明変数 (頻度) に影響を及ぼしているとは結論できない。係数の値が $\hat{\beta}_i$ とし、この係数の分散の推定値が b_i であるなら、真の値 β_i がゼロであるとき $\hat{\beta}_i^2/b_i^2$ は自由度 1 の χ^2 分布に従うはずである。この分布と比較によって得られる上側確率が、SAS の出力に現れている。

モデルの比較は、ロジスティック回帰におけるのと同様に、 χ^2 統計量を利用して行う。モデル 1 は定数項、および Z_1, \dots, Z_p をモデルに含むものとする。ここで Z_j は、変数の主効果である場合もあるし、交互作用項である場合もある。一方、モデル 0 はモデル 1 から、 Z_{p-q}, \dots, Z_p の q 個の変数から除いたもの、つまり $\beta_{p-q}, \dots, \beta_p$ をゼロに固定したモデルとする。モデル 1 を仮定して得ら

れた最尤推定値を $\hat{\beta}_1$ とし、モデル0の仮定のもとで得られた最尤推定値を $\hat{\beta}_0$ とする。それぞれの推定値に対応する対数尤度を、 $l(\hat{\beta}_1)$ 、 $l(\hat{\beta}_0)$ とする。係数がとりうる範囲はモデル1の方が広いので、 $l(\hat{\beta}_1) \geq l(\hat{\beta}_0)$ が常になりたつ。 G^2 はフルモデルと検討しているモデルとの対数尤度比の2倍、つまり

$$2 \times l(\text{フルモデル}) - 2 \times l(\hat{\beta}) \quad (232)$$

によって定義される。モデル0が真のモデルである場合には、 G^2 (対数尤度比統計量の2倍)の差が漸近的に自由度 q の χ^2 分布に従う。つまり、

$$G^2(M_0) - G^2(M_1) \stackrel{asym.}{\sim} \chi^2(q) \quad (233)$$

が成立する。この性質を用いて、モデルの比較を行うことができる。対数線形モデルの場合には、ある変数の係数がゼロであるという仮説が棄却されるということは、その変数が何らかの効果を被説明変数について及ぼしていることを意味する。また、交互作用項に対応する係数が有意であるということは、それに係わっている変数の間に関係のあることを意味する。

2×2 表の場合フルモデルには、{定数項, $X_1, X_2, X_1 * X_2$ } の4つの変数が存在する。これをモデル1とし、交互作用項 $X_1 * X_2$ を除いたモデルをモデル0とする。このとき、もしモデル0が真のモデルならば、

$$G^2(M_0) - G^2(M_1) \stackrel{asym.}{\sim} \chi^2(1) \quad (234)$$

となる。各セルの平均は次のような式で推定される。

$$\eta_{00} = \log \mu_{00} = \beta_0, \quad (235)$$

$$\eta_{10} = \log \mu_{10} = \beta_0 + \beta_1, \quad (236)$$

$$\eta_{01} = \log \mu_{01} = \beta_0 + \beta_2, \quad (237)$$

$$\eta_{11} = \log \mu_{11} = \beta_0 + \beta_1 + \beta_2 + \beta_{12} \quad (238)$$

$$(239)$$

この式から、オッズ比を求めると、

$$\theta = \frac{\exp(2\beta_0 + \beta_1 + \beta_2 + \beta_{12})}{\exp(2\beta_0 + \beta_1 + \beta_2)} = \exp(\beta_{12}) \quad (240)$$

が常に成立する。これは即ち β_{12} が対数オッズ比であることを示しており、 $\beta_{12} = 0$ であることと、行と列とが独立であることが同値であることがわかる。

31.8 過大分散 (over dispersion)

対数線形モデルは、ポアソン分布を用いてモデルの構築を行なう。ポアソン分布の特徴として、平均が μ であれば分散も μ であり、平均が定めれば分散の大きさも同時に定まることがある。このような特徴は、正規分布の場合にはない。正規分布は、平均と分散を別のパラメータとして持っている。

対数線形モデルのこのような性質が、ときには問題を引き起こすことがある。次のような分析を行なう場合を考えて見よう。

ある疾病の発生件数を説明するモデルを考える。市区町村毎に人口と発生件数のデータが得られていると仮定し、ある地域 i の発生件数を次のモデルで説明する。

$$\log \mu_i = \beta_0 + x_{i1} \quad (241)$$

ここで、 β_0 は定数項 (切片) であり、全データにおける発生率の水準に対応し、 x_{i1} は地域 i の人口の対数を表すものとする。

ここで、 x_{i1} は人口の影響を表しているが、地域の個別の特性が、発生率に影響しているとするものではない。もし、地域の別によらず、発生率を説明することができるのなら、上のモデルはデータによく当てはまるはずである。しかし、発生率に影響を及ぼす要因が他にあり、それが地域毎に少しずつ異なっているなら、モデルは厳密にはあてはまらなくなる。

他に発生率に影響を与える地域の属性を説明する変数があり、それをういてモデルを拡張すると、

$$\log \mu_i = \beta_0 + x_{i1} + \beta_2 x_{i2} \quad (242)$$

のようになる。この場合、推定される Y の分散は前の式よりは小さくなるが、なお真の平均がこの式からずれており、その要因を特定できない場合はしばしば生じる。

このような状況に対応するためには、次の 2 つの方法がある。

1. Y の分布がポアソン分布ではなく、平均の異なるポアソン分布の混合であるとして推定を行う。ポアソン分布の母数 μ が確率変数であり、ガンマ分布に従っていると仮定すると、 Y の分布は負の 2 項分布になる。
2. Y の厳密な分布は特定せず、ポアソン分布を仮定した場合と同様の平均構造を持ち、分散がポアソン分布の場合の定数倍であるとする。この定数倍をあらわすパラメータを過大分散パラメータ (overdispersion parameter) と呼ぶ。

後者の場合には、確率分布が厳密には特定されないので、最尤法の理論をそのまま適用することはできないが、最尤法とほぼ同様の計算手順によって、妥当な推定を行える。この方法は疑似尤度法 (quasi-likelihood method) と呼ばれる。

過大分散パラメータ ϕ の推定を、 β の推定と同時に進行。過大分散を指定するという事は、データに次のような構造を仮定することである。

$$\eta_x = g[\mu_x] = g[E(Y_x | \mathbf{x})] = \beta^T \mathbf{x} \quad (243)$$

$$\xi_x = \phi h[\mu_x] = \text{Var}(Y_x | \mathbf{x}) \quad (244)$$

ここで ϕ の値は、 x よらず、全体において一定であると仮定している。また関数 h は、過大分散パラメータを指定しない ($\phi = 1$) 場合の分散をあらわす関数である。ポアソン分布の場合には、 h は恒等関数 (値が変わらないもの) である。上の拡張されたモデルの下では、 $\phi \neq 1$ ならば、 Y_x の分布はもはやポアソン分布ではない。また、この仮定では Y_x の平均と分散が指定されるのみであり、分布が具体的には特定されないので、最尤法を利用することができない。しかし、最尤法と同様の計算によって良い推定を行うことが可能である。また、過大分散パラメータの大きさは

$$\hat{\phi} = \frac{1}{N-p} \sum_{i=1}^N \frac{(Y_i - \hat{\mu}_i)^2}{h[\hat{\mu}_i]} \quad (245)$$

によって推定する (p は定数項を含む説明変数の数)。この場合、推定される Y の条件付き平均は、過大分散パラメータを指定しない場合と等しいが、推定されたモデルの適合度とパラメータの標準誤差は異なる。

31.9 ロジスティック回帰と対数線形モデル

ロジスティック回帰は、対数線形モデルと密接な関係がある。

ロジスティック回帰における2値変数を Y とし、説明変数が X_1, X_2 の2つの2値変数である場合を考える。このとき、フルモデルは

$$\text{logit}\pi_x = \alpha_0 + \alpha_1x_1 + \alpha_2x_2 + \alpha_{12}x_1x_2 \quad (246)$$

となる。これを(222)~(229)と比較してみる。ここで、 X_3 を Y とすると、

$$\text{logit}\pi_{ij} = \alpha_0 + i\alpha_1 + j\alpha_2 + ij\alpha_{12} \quad (247)$$

であり、これは $\pi_{ij} = \mu_{ij1}/(\mu_{ij0} + \mu_{ij1})$ であることを考慮すると、

$$\text{logit}\pi_{ij} = \log \mu_{ij1} - \log \mu_{ij0} \quad (248)$$

$$= \beta_3 + i\beta_{13} + j\beta_{23} + ij\beta_{123} \quad (249)$$

となる。ここで、係数に注意すると

$$\alpha_0 = \beta_3, \alpha_1 = \beta_{13}, \alpha_2 = \beta_{23}, \alpha_{12} = \beta_{123} \quad (250)$$

の関係がなりたっていることが分かる。 Y の全般的な水準 α_0 が、対数線形モデルにおける $X_3 = Y$ の1反応の全体水準 β_3 に対応し、その他のロジスティック回帰における回帰係数は、対数線形モデルにおける Y と各変数との交互作用の大きさに対応している。

32 Rによる対数線形モデルの分析

SASでは、プロシージャGENMODまたはCATMODによって対数線形モデルの推定を行える。GENMODは一般化線形モデルの推定を汎用的におこなうためのプロシージャであり、連続変数を説明変数として含む対数線形モデル(ポアソン回帰とも呼ばれる)やロジスティック回帰も行うことが可能である。

SplusまたはRで対数線形モデルの分析を行なうには、関数glmかまたは関数loglinを用いる。glmはデータフレームを分析対象とし、一般化線形モデルの推定を行う関数である。連続変数および離散変数の両者を説明変数とすることが可能である。被説明変数を件数を表す変数とし、確率分布をポアソン分布に指定すると対数線形モデルの分析が行なえる(オプションでfamily=poissonと指定する。)

一方loglinは多重集計表を分析対象とするので、説明変数は離散変数に限られることになるが、利用方法はglmより簡単である。

多重集計表を作成するには、関数tableを用いる。ベクトルx、y、zがそれぞれ値1,2,10,20,30,"a","b","c","d"をとる同じ長さnの変数であるとき
tab1 <- table(x,y,z)とすると、tab1が2×3×4の集計表(24個のセルを含む)になり、それぞれのセルには該当するデータの件数が代入される。ベクトルの長さはnなので、セルの数を合計するとnになる。作成される表の見出し(ラベル)としては、それぞれの要因としてx、yなどの変数名が、水準名には変数の水準名または値が表示される。

以下は、多重集計表の例である。

以下はR-1.5.0でのglmを用いた分析例である。最初のデータフレームtb86dを作成し、ついで関数glmによって対数線形モデルの当てはめを行っている。

表 24: 自動車事故における乗員の負傷

性別	場所	シートベルト	負傷	
			無し	有り
女性	市街	無し	7287	996
		有り	11587	759
	郊外	無し	3246	973
		有り	6134	757
男性	市街	無し	10381	812
		有り	10969	380
	郊外	無し	6123	1084
		有り	6693	513

Agresti,A. (1996) *An Introduction to Categorical Data Analysis*, Wiley. 表.6.8 より
 Source: Dr Cristanna Cook, Medical Care Development, Augusta,Maine.

```
> tb86d <- read.table("seatbelt.dat",header=T)
> tb86d
  Gender Location SeatBelt Injury Count
1      F      U      N      n 7287
2      F      U      N      y  996
3      F      U      Y      n 11587
4      F      U      Y      y  759
5      F      R      N      n 3246
6      F      R      N      y  973
7      F      R      Y      n 6134
8      F      R      Y      y  757
9      M      U      N      n 10381
10     M      U      N      y   812
11     M      U      Y      n 10969
12     M      U      Y      y   380
13     M      R      N      n  6123
14     M      R      N      y  1084
15     M      R      Y      n  6693
16     M      R      Y      y   513

> tb86d.glm <- glm(Count ~ Injury*SeatBelt+ Injury*Location +Injury*Gender+
  SeatBelt*Location*Gender,family=poisson,data=tb86d)
> summary(tb86d.glm)
Call:
glm(formula = Count ~ Injury * SeatBelt + Injury * Location +
  Injury * Gender + SeatBelt * Location * Gender, family = poisson,
  data = tb86d)

Deviance Residuals:
    1      2      3      4      5      6      7      8
0.16160 -0.43483 -0.42327  1.69037 -0.15190  0.27851  0.51823 -1.44646
    9     10     11     12     13     14     15     16
0.21675 -0.76754  0.09329 -0.49684 -0.34700  0.83292 -0.05675  0.20564

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
```

```

(Intercept)          8.08784    0.01654 488.884 < 2e-16 ***
Injuryy              -1.21640    0.02649 -45.918 < 2e-16 ***
SeatBeltY            0.62713    0.02027  30.940 < 2e-16 ***
LocationU            0.80411    0.01966  40.891 < 2e-16 ***
GenderM              0.63640    0.02015  31.579 < 2e-16 ***
Injuryy:SeatBeltY   -0.81710    0.02765 -29.551 < 2e-16 ***
Injuryy:LocationU   -0.75806    0.02697 -28.105 < 2e-16 ***
Injuryy:GenderM     -0.54483    0.02727 -19.982 < 2e-16 ***
SeatBeltY:LocationU -0.15752    0.02441  -6.453 1.09e-10 ***
SeatBeltY:GenderM   -0.54186    0.02590 -20.925 < 2e-16 ***
LocationU:GenderM   -0.28274    0.02441 -11.584 < 2e-16 ***
SeatBeltY:LocationU:GenderM 0.12858    0.03228   3.984 6.78e-05 ***
---

```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

```

Null deviance: 61709.5207 on 15 degrees of freedom
Residual deviance: 7.4645 on 4 degrees of freedom
AIC: 184.92

```

Number of Fisher Scoring iterations: 3

出力の説明

Deviance Residuals は、ポアソン分布を仮定している場合には

$$\text{sign}(y_i - \hat{\mu}_i) \times \sqrt{2(y_i \log(y_i/\hat{\mu}_i) - y_i + \hat{\mu}_i)}$$

を表す。

Coefficients は、 β の値であり、Std Error は標準誤差 (推定値の標準偏差) を示す。z value は推定値を標準誤差で割ったもの。

Null deviance は、定数項のみのモデルを仮定した場合の G^2 の値。Residual deviance は指定したモデルについての G^2 。自由度 4 の χ^2 分布において 7.464 に対応する上側確率は 0.1133 である。

AIC は赤池情報量基準

33 層別データのロジスティック回帰による分析

先に層別の 2×2 分割表において、独立性の検定を CMH 統計量を用いて行い、また共通オッズ比を MH 統計量を用いて推定することを紹介したが、このような構造を持つデータをロジスティック回帰を用いて分析することも可能である。

各層の 2×2 の行を対照群 $x_0 = 0$ と実験群 $x_1 = 1$ を表すものとし、列が反応 ($Y = 0, Y = 1$) を表すものとする。さらに各層 ($k = 1, \dots, K$) は共通のオッズ比 $\exp(\beta)$ を持つものと仮定する。ここで、

$$\text{logit}(\pi_{jk}) = \alpha_k + \beta x_j, \quad j = 0, 1, k = 1, \dots, K \quad (251)$$

のようなロジスティック回帰モデルにより、層別された 2×2 表の構造を表すことができる。

もし、各層の各セルに十分な標本があれば最尤法を用いて推定することが可能であるが、層の数が多く各層内の標本数が小さい場合には最尤法によって良い推定値を得ることが難しい。特に 1 対

1 のケースコントロール研究の場合には、各層には 2 つの標本しか含まれず、最尤法による推定値 $\hat{\beta}^X$ は層の数が多いときほぼ $2\beta^X$ となることが知られている。

Mantel-Haenszel 推定量はこのような場合にもうまく共通オッズ比を推定するように工夫された方法であるが、ロジスティック回帰の推定法を工夫することにより、この問題に対応することが可能である。

基本的なアイデアは Fisher の正確検定のように周辺分布について条件を制約し、その制約下での正確な分布を求め、この尤度 (条件付き尤度 conditional likelihood) を最大化する β を求めるというものである。このような方法をとることの利点は、条件付き尤度は α_k をパラメータとして含まないので、推定すべきパラメータの数が層の数に依存しないことによる。そのため最尤法の漸近的な特徴が生かされて優れた推定を行うことができる。

ここで、より一般的に標本に通し番号 $i = 1, \dots, N$ がついているものとし、標本 i の説明変数の値を x_{ij} , ($j = 1, \dots, p$) とする。説明変数が実験群と対照群の違いだけなら $p = 1$ で x_{i1} は離散的な値 0-1 をとる。

標本 i における反応の確率は

$$\Pr(Y_i = y_i) = \frac{\exp \left[y_i (\alpha + \sum_{j=1}^p \beta_j x_{ij}) \right]}{1 + \exp(\alpha + \sum_{j=1}^p \beta_j x_{ij})} \quad (252)$$

さらに、 Y_1, \dots, Y_N が独立であるとし同時分布を考えると、

$$\Pr(Y_1 = y_1, \dots, Y_n = y_n) = \frac{\exp \left[(\sum_i y_i) \alpha + \sum_{j=1}^p (\sum_i y_i x_{ij}) \beta_j \right]}{\prod_i \left[1 + \exp(\alpha + \sum_{j=1}^p \beta_j x_{ij}) \right]} \quad (253)$$

ロジスティック回帰では、 β_j の値を観測された標本から推定する。尤度を β_j の関数とみると、 (y_i) の尤度への影響は $\sum_i y_i$ および $\sum_i y_i x_{ij}$, $j = 1, \dots, p$ の値によって決まる。このような統計量は十分統計量と呼ばれる。

ここで、 (y_i) の値をこれらの十分統計量が観測された値に一致するものに限定し、その条件のもとでの確率を考え、さらにそれについての尤度を検討することにする。もし X_j が連続値をもつものである場合には、このような (y_i) の集合を考えるのは難しいが、 X_j が 2 値である場合には、単純な組み合わせ問題 (件数は大きい) になる。

ここで、 $\sum_i y_i x_{i2} = t$ とし、この制約を満たす (y_j) の集合を $S(t) = (y_1^*, \dots, y_n^*)$ とおく。また $\sum_i y_i x_{ij} = t_j$, ($j = 1, \dots, p$)、 $\sum_i y_i = t_0$ とし、 $\sum_i y_i^* x_{ij} = t_j^*$, ($j = 1, \dots, p$)、 $\sum_i y_i^* = t_0^*$ とする。これらの集合に (y_i) を限定した場合の条件付確率を考えると、次の式で与えられることがわかる。

$$\Pr(Y_1 = y_1, \dots, Y_n = y_n \mid \sum y_i x_{i2} = t) = \frac{\Pr(Y_1 = y_1, \dots, Y_n = y_n)}{\sum_{S(t)} \Pr(Y_1 = y_1^*, \dots, Y_n = y_n^*)} \quad (254)$$

$$= \frac{\exp(t_0 \alpha + \sum_j t_j \beta_j) / \prod_i [1 + \exp(\alpha + \sum_j \beta_j x_{ij})]}{\sum_{S(t)} \exp(t_0^* \alpha + \sum_j t_j^* \beta_j) / \prod_i [1 + \exp(\alpha + \sum_j \beta_j x_{ij})]} \quad (255)$$

$$= \frac{\exp(t_0 \alpha + \sum_j t_j \beta_j)}{\sum_{S(t)} \exp(t_0^* \alpha + \sum_j t_j^* \beta_j)} \quad (256)$$

ここで分子と分母の指数関数の内部において $t_2^* = t_2$ が常に成立するので、この項が相殺されるため、上の式は β_2 には依存しないものになる。この条件付尤度を最大化するパラメータを求めるとすると、推定する必要のないパラメータを消去することにより、通常の最尤法より望ましい推定値を得ることができる。

このアイデアをケースコントロール研究データに適用すると、各層での反応 ($Y = 1$) の件数が観測値に一致するような y_i の組を想定することになる。ここで第 k 層のコントロールとケースの反応をそれぞれ y_{0k} および y_{1k} と表記し、これらの値の和を $s_k = y_{0k} + y_{1k}$ と置く。 $s_k = 0$ または $s_k = 2$ の場合には、取りうる値は一通りしかないので、複数の可能性を考える必要があるのは、 $s_k = 1$ の層のみとなる。ここで、各標本の反応確率は

$$\Pr(Y_{jk} = y_{jk}) = \frac{\exp[y_{jk}(\alpha_k + j\beta)]}{1 + \exp[(\alpha_k + j\beta)]} \quad (257)$$

であり、 $s_k = 1$ の層の条件付き確率は次の式となる。

$$\begin{aligned} \Pr(Y_{0k} = 0, Y_{1k} = 1 | s_k = 1) &= \frac{\exp(\alpha_k + \beta)}{\exp(\alpha_k + \beta) + \exp(\alpha_k)} \\ &= \frac{\exp(\beta)}{1 + \exp(\beta)} \\ \Pr(Y_{0k} = 1, Y_{1k} = 0 | s_k = 1) &= \frac{\exp(\alpha_k)}{\exp(\alpha_k + \beta) + \exp(\alpha_k)} \\ &= \frac{1}{1 + \exp(\beta)} \end{aligned}$$

ここで、全ての層についての条件付確率を考えると、 $s_k = 0$ および $s_k = 2$ の層については 1 である。ここで $s_k = 1$ となる層の個数を m とし、そのうち $y_{k1} = 1$ となる層の個数を m_1 とする。また $m - m_1 = m_2$ とする。

$$P(Y_{01} = y_{01}, Y_{11} = y_{11}, \dots, Y_{0K} = y_{0K}, Y_{1K} = y_{1K} | s_1, \dots, s_K) = \binom{m}{m_1} \left(\frac{\exp(\beta)}{1 + \exp(\beta)} \right)^{m_1} \left(\frac{1}{1 + \exp(\beta)} \right)^{m_2}$$

これは 2 項分布の尤度と同じものであるので、 $\exp(\hat{\beta}) / (1 + \exp(\hat{\beta})) = m_1 / m$ を満たす $\hat{\beta}$ が尤度を最大にする。これより、 $\hat{\beta} = \log(m_1 / m_2)$ となる。

参考文献

丹後俊郎、山岡和枝、高木晴良 (1996) ロジスティック回帰分析. 朝倉書店. ロジスティック解析についての、SAS の応用を中心とした解説。医学薬学分野の例が多い。ケースコントロール研究データの分析法についても解説してある。

Agresti, A. (1996). *An Introduction to Categorical Data Analysis*. Wiley. (英語であるが、大変わかりやすいロジスティック回帰と対数線形モデルの入門書。社会科学等応用分野の学生を対象としている。)

Agresti, A. (2002). *Categorical Data Analysis 2nd ed.* Wiley. 上の教科書より詳しい。カテゴリカルデータの分析法全般 (ロジスティック回帰および対数線形モデル) にわたって解説している。特に第 2 版ではケースコントロール研究データの分析の解説が増えた。大部 (700 ページ超) であるが、分かりやすく具体例が豊富。

Armitage, et al eds. (1999) *Encyclopedia of Biostatistics*, Wiley. 全 6 巻 (logistic regression と loglinear model の項、簡潔な解説)

Hosmer, D.W. & Lemeshow, S. (2000). *Applied Logistic Regression 2nd ed.*, Wiley. これもロジスティック回帰についての解説。モデル診断とケースコントロール研究の分析について詳しい。

Edwards, D. (2000). *Introduction to Graphical Modelling*, Springer. ここでは紹介しないが、この 15 年ほどの間に著しく理論が進展したグラフィカルモデル (変数間の相互依存関係を節点間を

結ぶ辺として幾何学的に表現するもの) についての解説。離散変数と連続変数とが混在する多変量データのためのモデル(条件付正規モデル)の性質と、分析ソフトウェア MIM の利用法の解説に重点をおいている。条件付き正規モデルは、離散変数に限定すれば対数線形モデルであり、また多変量分散分析(MANOVA)も特殊ケースとして含む。数式表記が一貫していないので読みづらいが、利用範囲は広いと思われる。MIM の学生版(データ件数の制約あり)は無償でダウンロードできる。<http://www.hypergraph.dk/>

引用文献

Dobson, A.J. 田中豊他訳(1993) 統計モデル入門 – 回帰モデルから一般化線形モデルまで – . 共立出版. (原著 *An Introduction to Generalized Linear Models*. Chapman and Hall (1990))

Draper, D. et al. (1992). *Contemporary Statistics, 1; Combining Information*, American Statistical Association.

34 クラスタ分析

これまでに説明した手法は、いずれも影響関係の強さを同定することを目的としているが、統計的な分析手法のもう一つの重要な役割として、複雑な相互関係を簡潔に記述することがある。複数の変数が分析対象となる統計的分析手法は、一般的に「(統計的)多変量解析法」とよばれるが、狭義にはこのような相互関係の記述のための手法を意味する。

広く用いられている多変量解析法には、クラスタ分析(クラスタリング)、主成分分析、因子分析、共分散構造分析(因子分析と回帰分析の統合モデル)、多次元尺度法、潜在クラス分析などがある。

ここでは、クラスタ分析のうち、特に「非階層的クラスタ分析」と呼ばれる手法(より具体的には k -means 法と呼ばれる方法)と、主成分分析法について解説する。

クラスタ分析とは、複数の(多くの場合は、かなり多数の)対象をそれらの類似性にもとづいて、互いに類似した群に類別する手法の総称である。階層的手法と非階層的手法とに大きく分類される。階層的手法は、生物分類に現れるような木構造を(非)類似性データに基づいて作成する。最近隣法、最遠隣法、ワード法など複数の方法がある。一方、非階層的方法は、木構造を作成せずに互いの類似性に基づいて対象をまとめることにより、複数の群を構成する。

ここでは、非階層的方法のうち代表的な手法である k -means 法について説明する。

k -means 法は、多変量の数値がそれぞれの対象について観測されている場合、これらの観測値に基づいて類別を行う。類別する群の個数は最初に指定する。観測されている変数が p 個あるものとし、これらの k 個の群の中心が次のベクトルで表されるものとする。

$$\boldsymbol{\mu}_j = \begin{pmatrix} \mu_{1j} \\ \mu_{2j} \\ \vdots \\ \mu_{pj} \end{pmatrix}, \quad (j = 1, \dots, k)$$

また、各対象の観測値は次のように表されるものとする。

$$\boldsymbol{x}_i = \begin{pmatrix} x_{1i} \\ x_{2i} \\ \vdots \\ x_{pi} \end{pmatrix}, \quad (i = 1, \dots, n)$$

ここで、各対象 i が k 個の群のうち、最も近い $\boldsymbol{\mu}_j$ によって代表される群に所属するものとする。このとき、これらの k 個の中心によって代表される群の全体としての当てはまりの悪さを

$$SS = \sum_i^n \|\boldsymbol{x}_i - \boldsymbol{\mu}_{j(i)}\|^2$$

によって表すことにする。ここで、 $j(i)$ は対象 i に最も近い $\boldsymbol{\mu}_j$ の添え字 j をあらわし、また記号 $\|\dots\|$ はベクトルの長さを表す。

k 個の $\boldsymbol{\mu}_j$ が設定されるならば、それぞれの \boldsymbol{x}_i に一番近い $\boldsymbol{\mu}_j$ は 2 点間の距離を計算すれば通常は一通りに定まるので、指標 SS の値を求めることができる。また、もう一方で $\boldsymbol{\mu}_j$ が不明であっても、もしそれぞれの \boldsymbol{x}_i の k 個の群への類別が定まっているならば、 $\boldsymbol{\mu}_j$ を群 j に属する \boldsymbol{x}_i の(多変量の)平均とおけば SS が最も小さくなるので、 $\boldsymbol{\mu}_j$ を決定することができる。しかしながら、一般的には SS を最小にする $\boldsymbol{\mu}_j$ と類別の方法を同時に知ることができない。

このため、実際の計算においては、適当に $\mu_j, (j = 1, \dots, k)$ を設定し、それに対する各 x_i の所属を求め、ついで SS を計算する。ついで、 x_i の所属を逐次変更することにより、それに伴って SS を最小にする μ_j も変更されるが、これらの更新により SS が逐次減少するように所属の変更を行う。厳密には SS を最小にするためには対象の所属について全ての組み合わせを検討するしかないが、これを実行するのは対象数が極めて少ない場合を除いて、組み合わせの数が極めて大きいために現実的ではない。そこで、対象の所属の更新方法に工夫を凝らし、多くの場合に妥当な解を与えるような手法が提案されている。

図 23,24 は、R.A.Fisher の著作に取り上げられている 3 つの品種のアヤメのガク長、ガク幅、花弁長、花弁幅の 4 つの変数のデータを示すものである。この分析には R システムのサンプルデータを利用した。図 23 はガク長、ガク幅であり、また図 24 は花弁長、花弁幅を示す。また、図中のシンボルはアヤメの 3 つの品種を表す。それぞれの品種は 50 個のサンプルからなっている。

ここで、 $k = 4$ として k -means 法によって得られた類別を表示色によって表している。

このような例では、正しい分類がデータから得られているので、あえて測定値からグループを推定する方法はないが、ここでは手法の特徴を示すための例として用いている。

ガクと花弁の長さ・幅から、種別を推定するには、ロジスティック回帰やこれを拡張した多項ロジスティック回帰、その他多様な手法が利用可能である。これらは、工学的なパターン認識の分野で、一般的に「教師あり学習」と呼ばれる。つまり、正しい結果がわかっているデータを教師（お手本）として分類規則を推測し、予測を行うものである。一方、クラスター分析などのように「正しい分類」についてのデータが存在しない状況のもとで、データの構造や規則性についての推測を行う手法は、「教師なし学習」と呼ばれる。

クラスター分析は、回帰分析や分散分析に比べて、分析の枠組みが必ずしも明確でない場合があるが、実用性は大きい。特に大量のデータを取り扱う場合には、まず少数のグループに（被験者などの）標本を分類し、それらのグループについて特徴を把握すると、全体の様子をよく理解できることが多い。特に、社会科学分野では、地理情報と各種の統計情報を組み合わせることにより、地域ごとの特性をパターンに分類することなどに利用できる。

比較的古い統計的多変量解析の文献では、階層的クラスター分析の手法が解説されていることが多いが、個人的経験からは非階層的な手法の方が、実用上の価値が大きいと思われる。

回帰分析や分散分析のような、明快な理論があるわけでないので、モデルのよさを評価するのは、一般的には難しい。特に、クラスターの数 k をどのように設定するかについて、様々な提案があるが、一般的な指針があるわけではない。また、観測値に基づく点間の距離に基づいて分類を行うために、それぞれの変数の尺度が分析結果に影響を及ぼすことに注意する必要がある。（回帰分析では、変数の尺度は F 検定などの結果に影響を及ぼさない）

社会科学でのクラスター分析の応用例の代表的なものとして倉沢・浅川（2004）、がある。これは、東京の地域別の各種の統計情報を地図上に濃淡表現し、またクラスター分析によって地域特性のグループを構成して表現したものである（旧版は 1984 年の出版）。

SPSS では、「分類」-「大規模ファイルのクラスタ」の指定が、 k -means 類似の手法を提供している（詳細アルゴリズムはマニュアルからではわからないが）。

最近のクラスター分析の先端的技術動向については、神島（2003a,2003b）が詳しい。

参考文献

- [1] 倉沢進、浅川達人（編）（2004）. 新編 東京圏の社会地図 1975 - 90, 東京大学出版会.

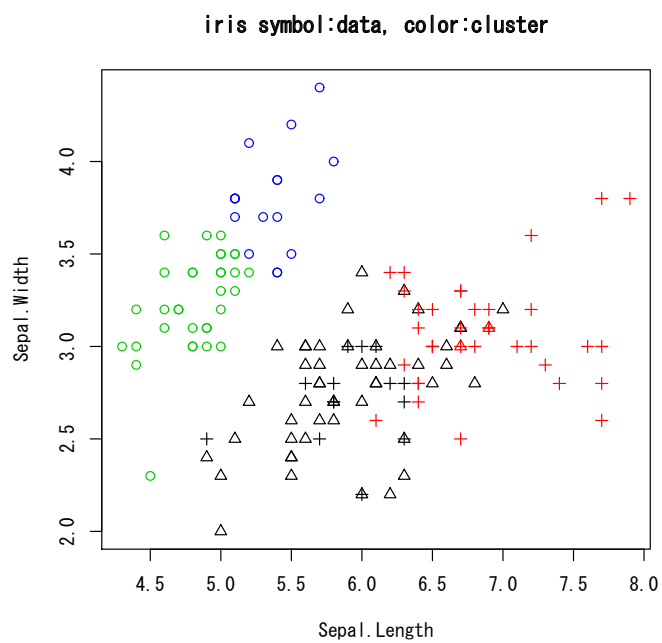


図 23: 横軸 ガク長、縦軸 ガク幅

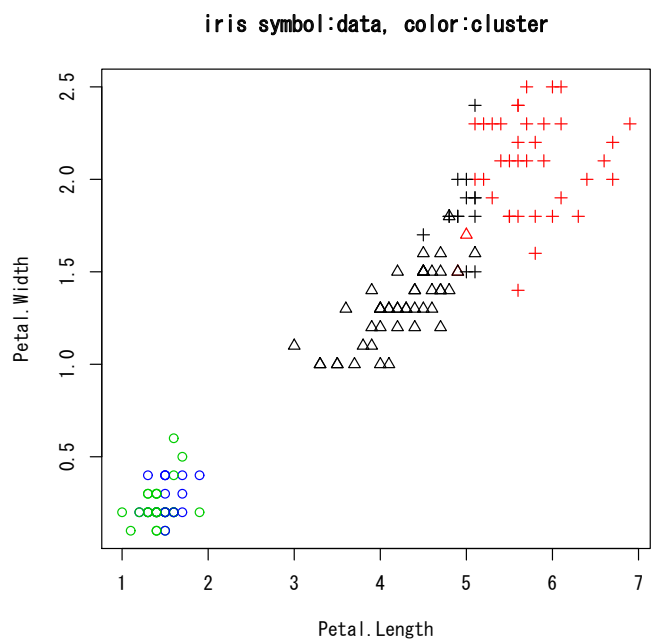


図 24: 横軸 花弁長、縦軸 花弁幅

- [2] 神高敏弘 (2003a) データマイニング分野のクラスタリング手法 (1) – クラスタリングを使ってみよう! –, 人工知能学会誌,18(1),59-65.
- [3] 神高敏弘 (2003b) データマイニング分野のクラスタリング手法 (2) – 大規模データへの挑戦と次元の呪いの克服 –, 人工知能学会誌,18(2),170-176.

35 主成分分析

主成分分析 (Principal Component Analysis, PCA) は, 多変量データの分析法としてもっとも広く用いられているものの一つである. PCA が求めるのは連続変数間の相互依存関係の簡潔な表現であり, 説明変数と被説明変数の区分的ないデータセットを対象として適用される.

分析対象となる変数を $X_j, (j = 1, \dots, p)$ と記述し, さらに得られているデータの値を $x_{i,j}, (i = 1, \dots, n; j = 1, \dots, p)$ と表す. ここで i はサンプル (上の例では被験者) を表し j は変数を表す. PCA の目的は, データ全体をよくあらわし得る少数の変数を求めることにある.

このために, 次の指標を最小化するような, X_1, \dots, X_p の 1 次式として表される $Z_k, (k = 1, \dots, q)$ を求める. ただし, 各 Z_k はまた q は p より (通常は大幅に) 少ない数である.

$$SSQ_q = \sum_j^p E(\|X_j - \sum_k^q a_{jk} Z_k\|^2) \quad (258)$$

式中の E は期待値を表す記号である. この値は, 観測変数 X_j を $Z_k, (k = 1, \dots, q)$ の 1 次式で予測した残差 2 乗和を, p 個の変数 $X_j, (j = 1, \dots, p)$ について足し合わせたものである.

ここで X_1, \dots, X_p を要素とする縦ベクトルを X とし, a_{jk} を要素とする $p \times q$ の行列を A とする. Z_1, \dots, Z_q を要素とする縦ベクトルを Z と表記する. また Z の各要素が X の 1 次式であるので, $q \times p$ の行列 B を用いて $Z = BX$ とかける. $AB = C$ とおくと $AZ = CX$ となる.

この記号を用いると上の式は, 次のように書き表される.

$$SSQ_q = \text{trace} E\{(X - AZ)(X - AZ)^T\} \quad (259)$$

$$= \text{trace} E\{(X - CX)(X - CX)^T\} \quad (260)$$

$$= \text{trace}(I_p - C)E(XX^T)(I_p - C)^T \quad (261)$$

主成分分析の手続きは, 観測変数の 1 次式を求め, それが元の観測変数を良く予測するように調整することに相当する.

上の式では A と Z が一意には定まらない. 実際 $q \times q$ の正則行列 G を一つ定めると, $A_1 = AU$ および $Z_1 = G^{-1}Z$ とすると, A_1 と Z_1 によっても同じ SSQ_q の値が得られる.

この条件を満たす $C = AB$ はつぎのようにして求められる. $E(XX^T) = \Sigma_{XX}$ とおく. この行列の固有値分解が次の式で与えられるとする.

$$\Sigma_{XX} = U\Lambda U^T \quad (262)$$

ここで U は直交行列であり, また Λ は固有値 $(\lambda_1, \dots, \lambda_p)$ を成分とする対角行列である. Λ の対角要素は降順に並んでいるものとする. このとき U の第 1 列から第 q 列を U_1 とすると, $A = B^T = U_1$ とおくことにより SSQ_q を最小化することができる. またこのようにして得られた Z_k を第 k 主成分と呼ぶ. 上の定義より $E(Z_k^2) = \lambda_k$ であることが分かる. また, $k \neq l$ ならば, $E(Z_k Z_l) = 0$ である.

各 Z_k の原点移動を許す場合には, $E(X) = E(AZ)$ であるときに SSQ_q が最小になることが計算によってわかる. これは各 X_j の平均がゼロとなるように原点移動して上述の SSQ_q を平均ゼロ

の Z について最小化することに等しい。また、主成分分析の結果は各 X_j の尺度に依存することに注意する必要がある。ある X_j の分散が際だって大きければ、分析結果は X_j を重視したものになる。

このようにして得られた Z は、 X の特徴を少数の変数によって表していると解釈できる。また、 A の j 行 a_j は変数 X_j のプロフィールと見做せる。つまり、 X_j と X_l とが互いに類似した変数であるなら（相関が大きく、分散も同程度なら）、 a_j と a_l とは類似した値を持つはずである。もし X_j と X_l との相関が -1 に近いなら、 a_j と a_l とは原点についてほぼ対称な値をとる。また、 a_j の長さの 2 乗 $a_j^T a_j$ は、 q 個の主成分によって表される X_j の分散の比率になっている。

35.1 SAS PROC FACTOR による主成分分析

主成分分析を SAS で行なうには、PROC PRINCOMP か PROC FACTOR を用いる。PROC FACTOR は本来、因子分析 (Factor Analysis) を行なうためのプロシジャであるが、オプションの指定によって主成分分析も行なえる。ここでは PROC FACTOR による方法を示す。

35.2 データ

ディレクトリ/home/otsu/cl14a0/sas につきのファイルがある。必要に応じてコピーして利用せよ。

- meth2pres.data 月別気圧データ (1000HP からの違い)
- meth2pres.doc 説明
- meth2pres.dat 同上、観測地名を含まず。
- meth2pres.row 観測地名のみのファイル
- meth2temp.data 月別気温データ
- meth2temp.doc 説明
- meth2temp.dat 同上、観測地名を含まず。
- meth2temp.row 観測地名のみのファイル

つぎが meth2pres.data の内容である (表頭の見出しはファイルには含まれていない)。

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Mean
WAZIMA	27	27	51	105	152	192	236	249	209	153	102	56	130
AIKAWA	30	28	52	104	149	188	231	248	211	157	107	61	131
NIIGATA	20	22	49	107	159	200	242	257	214	155	99	49	131
KANAZAWA	29	31	60	119	168	207	250	262	220	161	108	60	140
TOYAMA	21	24	55	115	165	204	246	258	215	156	103	53	135
NAGANO	-12	-4	31	102	155	194	236	245	199	133	72	19	114
TAKADA	19	19	45	111	161	202	245	257	214	153	98	49	131
UTUNOMIYA	14	24	57	116	164	200	236	249	209	151	93	39	129
HUKUI	25	29	61	123	172	210	253	263	220	158	104	55	139
TAKAYAMA	-23	-17	20	91	144	186	225	232	187	120	60	9	103
MATUMOTO	-10	-4	32	102	153	191	232	239	192	125	68	18	111
KARUIZAWA	-38	-34	-5	63	115	155	193	201	156	94	43	-11	78

MAEBASI	28	35	65	124	171	208	244	255	213	156	103	55	138
KUMAGAYA	31	39	70	127	174	210	246	259	217	159	105	55	141
MITO	25	31	62	117	162	196	233	248	210	154	101	49	132
TURUGA	40	42	70	127	173	210	254	266	225	167	117	69	147
GIHU	36	44	76	135	181	219	259	270	229	171	115	63	150
NAGOYA	36	43	75	135	180	217	256	268	228	169	114	62	149
IIDA	6	16	51	115	160	197	236	243	202	138	80	30	123
KOUHU	18	33	71	130	173	211	248	257	216	154	97	40	137
TYOUSI	58	63	89	134	170	197	228	248	228	182	137	87	152
TU	45	49	75	131	177	216	255	267	228	171	118	69	150
HAMAMATU	54	60	89	141	181	215	250	264	233	180	131	80	156
SIZUOKA	60	68	96	144	183	217	252	265	234	182	135	84	160
TOUKYOU	47	54	84	139	184	215	252	267	229	173	123	74	153
OWASE	57	63	90	140	177	211	248	259	228	176	127	80	155
YOKOHAMA	49	53	82	136	179	210	246	262	226	170	122	75	151
OOSIMA	62	64	86	130	168	197	230	245	217	170	131	90	149
HATIZYOUZIMA	103	104	124	163	192	220	252	266	248	208	172	130	182
SAIGOU	39	40	63	114	159	196	241	256	215	161	112	68	139

SAS を用いて分析するときには、meth2pres.data と meth2temp.data を用いればよい。他のファイルを用意したのは、あとで説明する xgobi を利用するためである。

つぎの SAS プログラムは、気圧データを対象とする分析の例である。

```

data pres;
  infile 'meth2pres.data';
  input site$ jan feb mar apr may jun    /* データの読み込み */
        jul aug sep oct nov dec myear;

proc print;                                /* 読み込み内容の確認 */

proc factor data=pres
  method=principal                        /* この指定で主成分分析を実行 */
  factors=3                               /* 抽出する主成分の数の指定 */
  score                                  /* 主成分スコアの出力の指定 */
  outstat= out2;                          /* 各種の情報を SAS データセットに出力 */
var   jan feb mar apr may jun            /* 分析対象とする変数の指定 */
      jul aug sep oct nov dec;

proc score data=pres score=out2 out=out3;
                                          /* 前の出力からスコアを計算 */

proc print data=out2;
proc print data=out3;
run;

```

36 Splus による主成分分析

主成分分析を行う場合には、`prcomp` と `princomp` の両者の関数が見えるが、`princomp` の方が使い易い。

ただし `princomp` で推定される共分散行列は最尤推定値 ($N - 1$ でなく N でわるもの) であるので、注意すること。

```
> meth2pres <- read.table("/home1/otsu/cl14a0/sas/meth2pres.data")
> dimnames(meth2pres)[[2]] <-
  c("Jan", "Feb", "Mar", "Apr", "May", "Jun", "Jul",
    "Aug", "Sep", "Oct", "Nov", "Dec", "Mean");
>
> meth2pres.princomp <- princomp(meth2pres[,1:12], score=T)
```

ここで `meth2pres.princomp` に主成分分析の結果が保存される。オプションの `score=T` は、主成分スコアを結果に保存するか否かの指定。T の場合には、保存する。

```
> names(meth2pres.princomp)
[1] "sdev"          "loadings"      "correlations" "scores"        "center"
[6] "scale"         "n.obs"         "call"          "factor.sdev"   "coef"
# 計算結果には、各種の情報が保存されている。
> meth2pres.princomp$sdev
  Comp. 1  Comp. 2  Comp. 3  Comp. 4  Comp. 5  Comp. 6  Comp. 7  Comp. 8
69.62844 17.58248 8.417505 3.208257 2.34953 1.349806 1.248197 0.9788613
  Comp. 9  Comp. 10  Comp. 11  Comp. 12
0.7796257 0.5001162 0.3501511 0.2876585
# 上は主成分得点の標準偏差。共分散行列の推定法の違いのため
# prcomp による値と若干異なっている。
>
> meth2pres.princomp$loadings
  Comp. 1  Comp. 2  Comp. 3  Comp. 4  Comp. 5  Comp. 6  Comp. 7  Comp. 8  Comp. 9
Jan  0.387   0.274  -0.111   0.384  -0.149           0.152  -0.453   0.269
Feb  0.383   0.174   0.226   0.170  -0.440           0.224
Mar  0.350           0.470           -0.317  -0.329   0.403  -0.325
Apr  0.257  -0.200   0.466           0.316   0.213  -0.305  -0.228   0.366
May  0.194  -0.325   0.259           0.428  -0.530   0.199
Jun  0.155  -0.415           0.144   0.116   0.263   0.594           -0.363
Jul  0.132  -0.487  -0.264   0.375  -0.127   0.472  -0.224   0.186   0.102
Aug  0.158  -0.414  -0.371           -0.294  -0.578  -0.124   0.119   0.140
Sep  0.233  -0.203  -0.191  -0.447           -0.270  -0.402
Oct  0.301           -0.190  -0.563  -0.102   0.124   0.189  -0.163  -0.257
Nov  0.353   0.202  -0.188  -0.252   0.214   0.124   0.254   0.579   0.498
Dec  0.376   0.293  -0.335   0.271   0.476           -0.308           -0.459
  Comp. 10  Comp. 11  Comp. 12
Jan  0.295  -0.266  -0.364
Feb -0.386   0.461   0.364
Mar  0.306  -0.128  -0.247
Apr -0.144  -0.329   0.349
May -0.187   0.274  -0.418
Jun  0.364  -0.109   0.260
Jul -0.271   0.141  -0.329
Aug           -0.282   0.337
Sep  0.402   0.511
Oct -0.468  -0.371  -0.212
Nov  0.142
Dec           0.203
```

計算結果の要素のうち `sdev` は、主成分スコアの標準偏差（固有値の平方根）であり、`loadings` は固有ベクトル（因子分析の用語を転用して因子負荷量 `factor loadings` とも呼ばれる）。固有ベクトルを表示すると絶対値の小さい箇所を自動的に省略する。行列には、すべての値が保持されている。

36.1 固有ベクトルのプロット表示 (PostScript 出力)

```
> postscript(file="meth2plot01.ps",horizontal=F,
  width=6,height=6)
> plot(meth2pres.princomp$loadings[,1],meth2pres.princomp$loading[,2],
  xlim=c(-1,1),ylim=c(-1,1),type="n")
> text(meth2pres.princomp$loadings[,1],meth2pres.princomp$loading[,2],
  dimnames(meth2pres.princomp$loadings)[[1]])
> dev.off()
Generated postscript file "meth2plot01.ps".
null device
      1
```

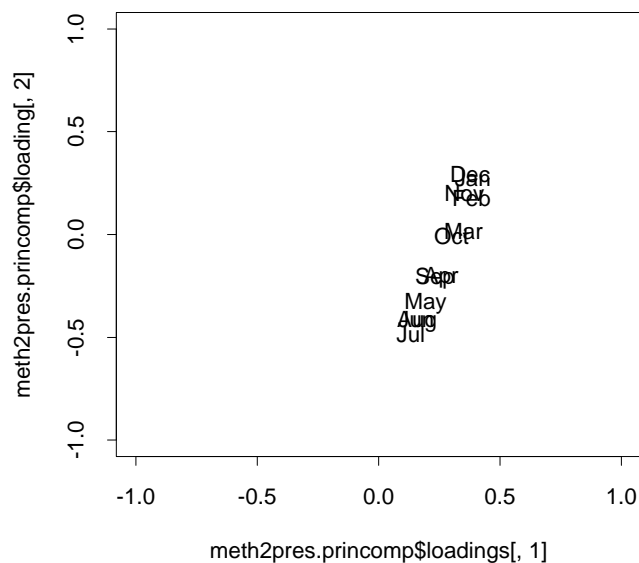


図 25: 固有ベクトル (因子負荷量) 第 1 主成分と第 2 主成分

36.2 サンプルスコアのプロット表示 (PostScript 出力)

```
> postscript(file="meth2plot02.ps",horizontal=F,width=6,height=6)
```

```

> plot(meth2pres.princomp$score[,1],meth2pres.princomp$score[,2],type="n")
> text(meth2pres.princomp$score[,1],meth2pres.princomp$score[,2],
+ dimnames(meth2pres.princomp$score)[[1]] )
> dev.off()
Generated postscript file "meth2plot02.ps".
null device

```

1

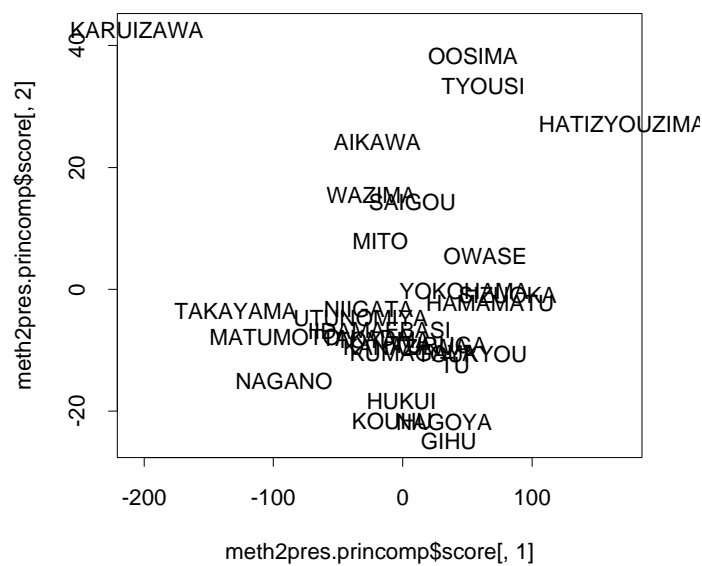


図 26: 主成分スコア 第 1 主成分と第 2 主成分

36.3 主成分分析のまとめ

観測データが p 個の変数 X_1, \dots, X_p からなるとする。また、これらの変数の共分散行列 をつぎのように表す。

$$S = \frac{1}{N-1}(w_{jk}), \quad w_{jk} = \sum_{i=1}^N (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$$

ここで、 N はデータの件数であり、 x_{ij} は変数 X_j の第 i 番目のサンプルにおける値を示す。また、 \bar{x}_j は X_j の標本平均を表す。

また、ここで S の固有値と固有ベクトルをそれぞれ

$$\lambda_1 \geq \lambda_2 \geq \dots \lambda_p; \mathbf{u}_1, \dots, \mathbf{u}_p$$

とする。これらは

$$S\mathbf{u}_j = \lambda_j\mathbf{u}_j, \quad (j = 1, \dots, p)$$

が成立していることを意味する。また固有ベクトルは各々長さ1、つまり $u_j'u_j = 1, (j = 1, \dots, p)$ であり、また互いに直交するように取られているものとする。

ある対称行列の異なる固有値に対応する固有ベクトルは、互いに直交することがつぎの様にして分かる。

$Su_1 = \lambda_1 u_1$ とし、また $Su_2 = \lambda_2 u_2$ としよう。ここで $\lambda_1 \neq \lambda_2$ とする。 $u_2'Su_1 = u_2'(\lambda_1 u_1)$ であるが、 λ_1 は定数(スカラー)であるから、この式は $\lambda_1 u_2'u_1$ である。また、 S が対称行列であることを考慮すると、 $u_2'S = (Su_2)'$ であるので、 $u_2'Su_1 = \lambda_2 u_2'u_1$ でもある。ここで、 $\lambda_1 \neq \lambda_2$ とすると、 $u_2'u_1 = 0$ でなければならない。これは、2つのベクトルが直交していることを意味する。

固有値が同一の場合には対応する固有ベクトルは、必ずしも直交するとは限らないが(つまりたがいに直交しないベクトルが固有ベクトルでありうる)、直交するような固有ベクトルの組を選ぶことは可能である。これ以降、固有ベクトルの組は互いに直交するように、また長さが1であるように選ばれていると仮定する。

行列 S の固有値を u_1, \dots, u_p とし、これらを並べた $p \times p$ の行列を $U = (u_1, \dots, u_p)$ とおく。この行列の各列の長さは1で、互いに直交しているので、直交行列(orthogonal matrix)になる。つまり $U'U = UU' = I_p$ である(I_p は p 次の単位行列)。

また、 U の定義から

$$SU = U\Lambda$$

である。ここで Λ は $\lambda_1, \dots, \lambda_p$ を要素とする対角行列である。上の式に、右から U' を掛けると

$$SUU' = S = U\Lambda U'$$

となる。つまり、 S が直交行列 U と対角行列 Λ の積に分解される。これを、対称行列の固有値分解(eigenvalue decomposition)と呼ぶ。

つぎに、主成分分析に関係する基礎的な性質をまとめる。

1. $Y_a = a_1 X_1 + a_2 X_2 + \dots + a_p X_p$ の分散は $a'Sa$ と表される。ここで $a' = (a_1, \dots, a_p)$ である。
2. ベクトル a の長さが1であるという制約のもとでは(つまり $a'a = 1$)、 $a = u_1$ のとき $a'Sa$ の値が最大値 λ_1 になる。
3. a の長さが1であり、かつ u_1 と直交するという制約のもとでは(つまり、 $a'a = 1, u_1'a = 0$)、 u_2 が $a'Sa$ の値を最大にする。
4. $Y_b = b_1 X_1 + b_2 X_2 + \dots + b_p X_p$ とし、 $b' = (b_1, \dots, b_p)$ とする。 Y_a と Y_b の共分散は $a'Sb$ である。
5. u_1 は(最大固有値に対応する)主成分直線の方角を表す。
6. (重要): 主成分直線はデータの重心(平均) $(\bar{x}_1, \dots, \bar{x}_p)$ を通る。
7. $P_1 = u_1 u_1'$ とおく。 P は $p \times p$ の対称行列である。 P_1 は直交射影子である。つまり対称でありかつ冪等 $P_1 P_1 = P_1$ である。 P_1 のランク(階数)は1である。
8. $I - P_1$ も直交射影子であり、ランクは $p - 1$ 。ここで I は単位行列。

9. 任意の長さ p のベクトル x について、 P_1x と $(I - P_1)x$ とは直交する。なぜなら、 $P_1x'(I - P_1)x = x'P_1'(I - P_1)x = x'(P_1 - P_1)x = 0$ であるから。
10. もし、 x が u_1 の定数倍ならば $P_1x = x$ である。
11. P_1x は、原点を通り u_1 の方向をもつ直線へ、点 x から垂線をおろした交点 (垂線の足) である。残差ベクトルは $x - P_1x = (I - P_1)x$ と表される。残差ベクトルと主成分直線とは直交する。
12. $SP_1 = \lambda_1P_1$ である。
13. (重要): これ以降、データは中心化されているとする (つまり平均ゼロになっている)。
14. u_{kj} を u_j の第 k 番目の成分とする。 $Z_j = u_{1j}X_1 + \dots + u_{pj}X_p$ と定義する。第 i 番目のデータを x_i (p 次元の縦ベクトル) で表すと、 i 番目のサンプルにおける Z_j の値は、 $u_j'x$ である。この値を z_{ij} と表す。これが第 i サンプルの第 j 主成分得点である。 Z_j の平均 ($\sum_{i=1}^N z_{ij}/N$) は ($j = 1, \dots, p$) について 0 である。
15. (重要): Z_j の平均はゼロ、分散は $u_j'Su_j = \lambda_j$ である。
16. $P_1x_i = u_1u_1'x_i = Z_{i1}u_1$ である。ここで、 Z_{i1} はスカラー (数値) なので、前に持ってくるができる。
17. 一般的に $v = (v_1, \dots, v_p)'$ の分散共分散行列が Σ であるとする。 A を $q \times p$ の行列とし、 $w = Av$ とする。 w は q 次元のベクトルである。このとき、 w の分散共分散行列は $A\Sigma A'$ となる。
18. 残差 $x - P_1x = (I - P_1)x$ の分散共分散行列は、 $(I - P_1)'S(I - P_1) = S - \lambda_1P_1$ である。
19. 2 番目以降の固有ベクトルについても、 $P_j = u_ju_j'$, ($j = 2, \dots, p$) とする。 $S = \lambda_1P_1 + \lambda_2P_2 + \dots + \lambda_pP_p$ である。したがって、残差の分散共分散行列は $S - \lambda_1P_1 = \sum_{k=2}^p \lambda_kP_k$ とかける。
20. 残差ベクトルの長さの 2 乗和は $\text{trace}(S - \lambda_1P_1) = \sum_{k=2}^p \lambda_k$ である。ここで trace (トレース、跡和) とは、正方行列の対角要素の和である。
21. $N \times p$ のデータ行列を X とおく。第 k 主成分得点は $z_k = Xu_k$ と表される。これは $N \times 1$ のベクトルである。また、平均はゼロである。
22. 前述の通り、 z_k の分散は λ_k である。また、 k_1 と k_2 が異なれば、 z_{k_1} と z_{k_2} の共分散はゼロである (したがって相関もゼロ)。

つぎに、主成分分析の解釈に関連する事項をまとめる。

1. $p \times p$ の共分散行列 S において、 q 個の顕著に大きな固有値が存在するものとする。上位 q 個の固有値に対応する固有ベクトルをならべた $p \times q$ の行列を $U_A = (u_1, \dots, u_q) = (u_{jk})$ とする。
2. もし、 X_{j_1} と X_{j_2} の相関が 1 に近ければ、 U_A の第 j_1 行と j_2 行とは、ほぼ正の定数倍になっている。
3. さらに X_{j_1} と X_{j_2} の分散がほぼ同じであるならば、 U_A の第 j_1 行と j_2 行とは、ほぼ等しくなる。

4. もし、 X_{j_1} と X_{j_2} の相関が -1 に近ければ、 U_A の第 j_1 行と j_2 行とは、ほぼ負の定数倍になっている。
5. さらに X_{j_1} と X_{j_2} の分散がほぼ同じであるならば、 U_A の第 j_1 行と j_2 行とは、原点をはさんでほぼ対称な座標を持つ。
6. X_j の分散は S の第 (j, j) 成分 s_{jj} であらわされる。この内、 q 個の主成分で表される分散は、 $\sum_{k=1}^q \lambda_k u_{jk}^2$ である。
7. $s_{jj} = \sum_{k=1}^q \lambda_k u_{jk}^2$ ならば、すなわち $u_{jk} = 0, (k = q + 1, \dots, p)$ ならば、 X_j の変動は q 個の主成分によって完全に説明される。
8. データの分散の標準化を行わなければ、主成分分析は変数の尺度によって本質的な影響を受ける。大きな分散を持つ変数が重視される結果が得られる。

36.4 Splus による主成分分析 (続き)

もし、すべての変数を平均ゼロ、標準偏差 1 として計算するには、つぎのようにする。

```
> data1 <- as.matrix(meth2pres[, 1:12])
> for(i in 1:12) {
  me <- mean(data1[, i])
  sd <- sqrt(var(data1[, i]))
  data1[, i] <- (data1[, i] - me)/sd
}>
```

これで、data1 の各列が標準化される。

つぎのようにして、data1 の平均と分散を確認してみる。

```
> apply(data1, 2, mean)
      Jan      Feb      Mar      Apr      May
-1.165734e-16 -2.035409e-17 -1.205055e-16  1.887379e-16  7.512509e-16
      Jun      Jul      Aug      Sep      Oct      Nov
 2.109424e-16 -7.179442e-16  6.652086e-16  8.174017e-16  1.44329e-16  1.702342e-16
      Dec
 7.956598e-17
> diag(var(data1))
[1] 1 1 1 1 1 1
```

各列の平均はほぼゼロ (計算誤差のため完全にゼロにはなっていない)、また分散も 1 となっている。var は各列を変数として、分散共分散行列を求める。diag はその対角成分 (ここでは各変数の分散) を求める関数である。

課題：十種競技 1995 年日本 50 傑のデータが

/home/otsu/cl114a0/Miyakawa/zyusyu0.data にある。被験者番号および変数名もファイルに含まれている。

つぎの手順を踏んで主成分分析と偏相関係数の検討を Splus を用いて行ないなさい。

1. `cp /home/otsu/cl14a0/Miyakawa/zyusyu0.dat .` などとしてファイルをコピーする。
2. ファイルの内容を `less` コマンドなどを用いて確認する。
3. つぎのようなコマンドを用いてデータを読み込む。

```
> zyusyu <- read.table("zyusyu0.dat",header=T)
```

データフレーム `zyusyu` は 11 変数を含み、1 列目は被験者番号 (順位) であるので、`prcomp` を適用する場合には、上の気圧データと同様に `as.matrix` を用いて変換する必要がある。

37 因子分析

37.1 因子分析のモデル

多くの変数を取り扱う統計的分析法は、一般的に多変量 (統計) 解析と呼ばれる。主成分分析 (PCA) は、多変量解析のなかで、最も広く使われている手法の一つである。計算は比較的単純であり (対称行列の固有値計算のためには信頼性の高い計算法が開発されている)、解釈も容易である。しかし、欠点としては変数の測定尺度に結果が依存することがあげられる。分析に用いられる変数の分散が異なれば、何らかの基準を用いて事前に標準化を行わなければ、結果は本質的に異なるものになる。

因子分析 (Factor Analysis) の利用目的は、主成分分析と類似したものであるが、主成分分析とはいささか異なるモデルと推定方法を用いている。

観測データが p 次元のベクトル $X = (X_1, \dots, X_p)'$ で表されているものとする。因子分析が仮定するモデルはつぎのようなものである。

$$X = \Lambda_{p \times m} f + \varepsilon \quad (263)$$

ここで f は次元 m の縦ベクトルであって ($m < p$)、直接には観察できない潜在的な変数 (共通因子, common factor) を表す。 Λ は $p \times m$ の行列であって、各変数がどの共通因子の影響を受けているかを表すものであり、因子負荷量 (factor loading) と呼ばれる。また、 $\varepsilon = (\varepsilon_1, \dots, \varepsilon_p)'$ は p 次元の縦ベクトルであり、知能テストにかかわる歴史的経緯から独自因子 (unique factor) と呼ばれる。上の式を要素毎にあらわすとつぎの式が得られる。

$$X_j = \lambda_{j1}f_1 + \dots + \lambda_{jm}f_m + \varepsilon_j, \quad (j = 1, \dots, p) \quad (264)$$

ここで、 f の分散行列 ($m \times m$) を Φ 、また ε の分散行列 ($p \times p$) を、 Ψ とおくことにする。さらに、 f と ε が統計的に独立であることを仮定すると、つぎの式が成り立つ。

$$\text{Var}(X) = \Sigma = \Lambda \Phi \Lambda' + \Psi \quad (265)$$

ここで、上式の左辺は観測データの分散行列によって推定できるが、右辺の各要素は直接には分からない。そこで、幾つかの仮定をおいて右辺の要素を推定し、これらを用いてデータの解釈を行なう。ここで、因子負荷量 Λ は主成分分析における固有ベクトルと類似した意味を持つ。

多くの場合につぎのような仮定をおいて分析を行なう。

1. f の各要素 f_i は互いに独立に、平均ゼロ、分散 1 の正規分布に従う。
2. ε の各要素 ε_j は互いに独立であり、また f の各要素とも独立に、平均ゼロ、分散 ψ_j の正規分布に従う。

これらの仮定のもとで、 X の分散は

$$\text{Var}(X) = \Sigma = \Lambda \Lambda' + \Psi \quad (266)$$

となる。ここで、 Ψ は ψ_1, \dots, ψ_p を要素とする対角行列である。

このモデルは、 p 個の変数の分散行列がランク (階数) m の対称行列 $\Lambda \Lambda'$ と、対角行列 Ψ の 2 つの成分の和として表されることを主張するものである。

37.2 推定方法

では、実際にデータが与えられているときに、どのようにして Λ と Ψ を求めたらよいのだろうか。ここで、データから得られる $p \times p$ の分散行列 S をつぎのようにして定義する。

$$S = \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})(x_i - \bar{x}) \quad (267)$$

ここで、 x_i は第 i 番目の標本ベクトルを表し、 \bar{x} は標本平均 $\sum_{i=1}^N x_i/N$ を表す。もし、 x の真の分布が平均 μ 、分散行列 Σ_0 の p 変量正規分布であるのならば、 S の値は標本数 N が増大するにつれて、 Σ_0 に次第に近付いてゆく。

もし因子分析モデル (266) が正しく、しかもデータの件数が十分にあるのなら、 $S \approx \Sigma_0$ が成立しているはずである。そこで、 $S \approx \Sigma = \Lambda\Lambda' + \Psi$ となるように Λ と Ψ を選ぶことによって、推定を行なえる。ここで問題となるのは、つぎの2点である。

1. どのような基準によって S と Σ の近さを定義するか。
2. どのような場合に、 Λ と Ψ の一意性が保証されるか。つまり、 $\Lambda_1 \neq \Lambda_2$ または $\Psi_1 \neq \Psi_2$ が成立しているにもかかわらず、 $\Sigma_1 = \Lambda_1\Lambda_1' + \Psi_1$ と $\Sigma_2 = \Lambda_2\Lambda_2' + \Psi_2$ とが等しくなることがないか。

最初の問題については、いくつかの基準が用いられている。ひとつは $\text{trace}(S - \Sigma)^2 = \|S - \Sigma\|^2 = \sum_{j=1}^p \sum_{k=1}^p (s_{jk} - \sigma_{jk})^2$ であり、通常これを $1/2$ 倍したつぎの値を重みなし最小2乗基準 (Unweighted Least Squares, ULS) と呼ぶ。

$$F_{\text{ULS}}(S; \Sigma) = (1/2)\text{trace}\{(S - \Sigma)^2\} \quad (268)$$

主因子法 (Principal Factor Analysis) と呼ばれている推定方法は、 Ψ を何らかの方法によって最初に推定し、 Λ を ULS と同等の基準によって推定するものである。また、 $\Psi = \mathbf{0}$ として ULS による推定を行なうと、主成分分析 (PCA) と同じ解が得られる。

また、より理論的に洗練された基準としては、一般化最小2乗基準 F_{GLS} とウィシャート最尤基準 F_{WML} とがある。これらは各々つぎの式で定義される。

$$F_{\text{GLS}}(S; \Sigma) = (1/2)\text{trace}\{(S - \Sigma)W^{-1}\}^2 \quad (269)$$

$$F_{\text{WML}}(S; \Sigma) = \log(\det \Sigma) + \text{trace}(S\Sigma^{-1}) - \log(\det S) - p \quad (270)$$

一般化最小2乗基準は、差の要素 $s_{jk} - \sigma_{jk}$ の分散と共分散を考慮した基準である。これが利用しているのは、正規性の仮定のもとで成立するつぎの2つの性質である。ここで、 σ_{jk} は真の分散行列の要素である。

1. s_{jk} の平均は σ_{ij} である。
2. s_{jk} と s_{lm} の共分散は $(\sigma_{jl}\sigma_{km} + \sigma_{jm}\sigma_{kl})/(N-1)$ である。

上の性質を考慮すると、 $W = \Sigma$ として F_{GLS} 基準を最小化することにより、優れた推定を行なえることが理論的に導かれる。実際には Σ は未知であるので、推定値として $W = S$ とおく。するとつぎのような式が得られる。

$$F_{\text{WGLS}}(S; \Sigma) = (1/2)\text{trace}\{(I_p - \Sigma S^{-1})^2\} \quad (271)$$

一方、最尤法は、 x が正規分布に従うとの仮定の下で、 S が Wishart 分布と呼ばれる分布 $W(\Sigma/(N-1), N-1)$ に従うことを用いて得られる。 S の確率密度関数は、正定値な S の値についてつぎのような式で与えられる (憶える必要はないが)。 S が正定値でないときには密度ゼロである。

$$w(S|\frac{1}{N-1}\Sigma, N-1) = \frac{(\det S)^{(N-p-2)/2} \exp\{-\frac{N-1}{2}\text{trace}(\Sigma^{-1}S)\}(N-1)^{p(N-1)/2}}{2^{p(N-1)/2}\pi^{p(p-1)/4}(\det \Sigma)^{(N-1)/2} \prod_{i=1}^p \Gamma[(N-i)/2]} \quad (272)$$

ここで Γ は Γ (ガンマ) 関数である。これは階乗 ($n!$) の連続関数への拡張とみなせる。実際 n が正の整数なら $\Gamma(n+1) = n!$ である。

上の式 (272) の対数を Σ の関数と見做すと、

$$-\frac{N-1}{2}\{\text{trace}\Sigma^{-1}S + \log(\det \Sigma)\} + \text{定数} \quad (273)$$

と書き表される。上述の F_{WML} は対数尤度を $-2/(N-1)$ 倍し、定数を加えて $S = \Sigma$ のときゼロになるよう調整したものである。

2つの基準 F_{WML}, F_{WGLS} のいずれも $S\Sigma^{-1}$ (またはこの逆行列の ΣS^{-1}) の関数である。実際、 F_{WML} には $\log(\det \Sigma) - \log(\det S)$ の項が現れるが、これは $\log(\det \Sigma / \det S) = \log\{\det(\Sigma S^{-1})\}$ と変形される。

行列の性質から、 D を正則な (逆行列のある) 対角行列とし、 $S_2 = DS_1D$ 、また $\Sigma_2 = D\Sigma_1D$ とするとつぎの等式が成立する。

$$F_{WML}(S_1; \Sigma_1) = F_{WML}(S_2; \Sigma_2) ; F_{WGLS}(S_1; \Sigma_2) = F_{WGLS}(S_2; \Sigma_2) \quad (274)$$

上の式は、変数の尺度が変わってもこれらの基準を用いた推定によって得られる結果が本質的には同じものであることを意味する。実際に得られる Λ や Ψ の値は変化するが、変数の尺度による影響を表すものである。

37.3 識別可能性

つぎに、 S が与えられたときに、 Λ と Ψ が一意的に決まるか否かについて検討する。

まず、 Λ については、 $\Lambda_2 = \Lambda_1 V$ (V は $m \times m$ の直交行列) とすると、 $\Lambda_2 \Lambda_2' = \Lambda_1 U U' \Lambda_1' = \Lambda_1 \Lambda_1'$ であるので、ケースの如何を問わず直交行列についての不定性は存在する。 Λ は $p \times m$ の大きさであるので、 $m \times p$ 個のパラメータを持つが、直交行列の不定性が $m(m-1)/2$ 個分あるので、実質 $mp - m(m-1)/2$ 個のパラメータを推定していると見做せる。共分散行列 Σ は対称であるので、 $p(p+1)/2$ 個分の変数で完全に表現される。 Ψ の分 p 個を含めて $(m+1)p - m(m-1)/2 \leq p(p+1)/2$ でなければ、一意的な解は求まらない。厳密には、「陰関数定理」と呼ばれる解析学の定理を用いて証明される。

簡単な $p = 2, m = 1$ のケースについて考えてみよう。

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}, \Lambda = \begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix}, \Psi = \begin{pmatrix} \psi_1 & 0 \\ 0 & \psi_2 \end{pmatrix}$$

であり、モデルが与える分散行列はつぎのようになる。

$$\Sigma = \Lambda \Lambda' + \Psi = \begin{pmatrix} \lambda_1^2 + \psi_1 & \lambda_1 \lambda_2 \\ \lambda_1 \lambda_2 & \lambda_2^2 + \psi_2 \end{pmatrix}$$

この場合推定すべきパラメータは4個(不定性は符号の正負のみ)であるが、共分散行列は3つのパラメータで表される。モデルパラメータの値は完全に自由に定まる訳ではないが、 $\lambda_1^2 < \sigma_{11}$, $\lambda_2^2 < \sigma_{22}$ の制約内で、 $\sigma_{12} = \lambda_1\lambda_2$ とする方法は無数に存在するため、パラメータを一意に決定することはできない。

変数が3個に増えた1因子モデルでは、パラメータの数は6個であり、 $p(p+1)/2 = 6$ にちょうど等しい。因子分析モデルがあてはまる範囲の Σ について、解は一意に求まる。例えば

$$\Sigma = \begin{pmatrix} 1 & & \\ 1/3 & 1 & \\ 1/5 & 2/5 & 1 \end{pmatrix}$$

とすると、 $\lambda_1 = \sqrt{1/6}$, $\lambda_2 = \sqrt{2/3}$, $\lambda_3 = \sqrt{6/25}$, $\psi_1 = 5/6$, $\psi_2 = 1/3$, $\psi_3 = 19/25$ とおくと、正確に $\Sigma = \Lambda\Lambda' + \Psi$ が成立する。

通常、因子分析を応用する場合には、多くの変数があり、因子数 m は p よりかなり小さいことを想定しているので、識別可能性の問題がしばしばおこる訳ではない。しかし、場合によっては上述の2変数の場合と類似の構造が、より大きな問題の一部にあらわれて、推定されるパラメータの値が不安定になる場合もある。一つの例は、ある因子に対応する因子負荷量が、2つの変数のみにおいてゼロでない値をとり、他の変数についてはゼロとなる場合である。この場合は、2変数1因子の場合と類似の条件であり、パラメータの一意性が失われる。

識別可能性については構造方程式モデル(後述)との係わりで、Bollen (1989) 4章、Johnston (1984) などにより詳しい説明がある。

37.4 因子の回転

先に、因子分析によって得られた因子負荷量 Λ には回転の不定性があることを指摘した。分析結果として考察されることが最も多いのは Λ のパターンであるが、もし、適当な回転(右からの $m \times m$ の直交行列を掛けること)によって、 Λ が解釈しやすくなるのなら、その方が望ましい。また、因子 f の各要素が独立であるとの制約を外すならば、直交行列にとどまらず、より一般的な正則行列による変形が可能になる。

ここで、最初に得られた因子得点を Λ_1 とする。因子 f の各要素 f_1, \dots, f_m が互いに独立で分散1であるとの仮定のもとで、つぎの式が成り立つ。

$$\text{Var}(\Lambda_1 \mathbf{f}) = \Lambda_1 \Lambda_1' \quad (275)$$

ここで、 T を $m \times m$ の正則行列とし、 $\mathbf{g} = T^{-1} \mathbf{f}$ とおく。また $\Lambda_1 T = \Lambda_2$ とする。

$$\Lambda_1 \mathbf{f} = \Lambda_1 T T^{-1} \mathbf{f} = \Lambda_2 \mathbf{g}$$

であるので、

$$\text{Var}(\Lambda_1 \mathbf{f}) = \text{Var}(\Lambda_2 \mathbf{g})$$

である。このとき $\Psi = \text{Var}(\mathbf{g}) = T^{-1} \text{Var}(\mathbf{f}) T^{-1'} = T^{-1} T^{-1'} = (T' T)^{-1}$ である。これは T が直交行列であるときのみ単位行列となる。

ここで T を適当に選ぶことによって、 Λ_2 を解釈しやすい形に変形する方法を考える。この操作のことを一般的に因子の回転(factor rotation)と呼ぶ。 T を直交行列に制限するものを、直交回転と呼び、特に直交行列に制限せず正則行列一般の範囲で考えるものを斜交回転(oblique rotation)

と呼ぶ⁵。また、特定のターゲット Λ_0 を指定しておき、 Λ_2 がなるべくこれに近くなるように T を選ぶことをプロクラステス変換 (Procrustes transformation) と呼ぶ。通常、ターゲット行列を Λ_3 とするとき、 $\Lambda_2 = \Lambda_1 T$ として、 $\text{trace}\{\Lambda_3 - \Lambda_2)(\Lambda_3 - \Lambda_2)'\}$ を最小化するように Λ_2 を選ぶ。これは要素毎の差の2乗和を最小化することを意味する。

では Λ がどのようなパターンをもっているならば、解釈しやすいといえるだろうか。このための基準として様々のものが可能であり、またそのための多くの方法が提案されているが、ここではバリマックス (Varimax) 法とプロマックス (Promax) 法の2つについて紹介する。

解釈を容易にする一つのパターンは、「 Λ の各列の要素が特定の行 (変数) についてのみ大きな絶対値をもち、他の多くの行についてはゼロに近い値を持つ」というものである。例えば、因子荷量がつぎのようなパターンである場合には、因子と変数の関係の解釈が容易である。

$$\Lambda = \begin{pmatrix} * & 0 & 0 \\ * & 0 & 0 \\ * & 0 & 0 \\ 0 & * & 0 \\ 0 & * & 0 \\ 0 & * & * \\ 0 & 0 & * \\ 0 & 0 & * \\ 0 & 0 & * \end{pmatrix} \quad (276)$$

バリマックス法はつぎの基準を最大化するように直交行列 T を求めるものである。

$$\frac{1}{p} \sum_{k=1}^m \sum_{j=1}^p (\lambda_{jk}^2 - \mu_k)^2 \quad (277)$$

ただしここで、 $\mu_k = \sum_{j=1}^p \lambda_{jk}^2 / p$ である。上の式 (277) は、 λ_{jk}^2 の各列の分散の和を示すものであるので、これを大きくするように直交行列 T を選ぶということは、 Λ の各要素を2乗したものの各列の分散の和が大きいことを意味する。もし、何らかの直交行列 T によって、(276) のようなパターンを得ることが可能なら、バリマックス基準 (277) を最大化するように T を選ぶことによって目的に近い Λ を得ることができるはずである。

プロマックス法は、バリマックス法と斜交プロクラステス変換を組み合わせた方法である。つぎの様な手順をとる。

1. 最初に得られた因子荷行列を Λ_1 とし、これをバリマックス法によって直交回転したものを $\Lambda_2 = \Lambda_1 T_1$ とする。
2. Λ_2 の大きさを標準化したあと絶対値の大小関係を冪乗によって強調したものを Λ_3 とする。
3. Λ_3 をターゲットとして、 Λ_2 の斜交プロクラステス変換を行ない、その結果得られた $\Lambda_4 = \Lambda_2 T_2 = \Lambda_1 T_1 T_2$ を因子荷行列とする。

ここで T_1 は直交行列であるが、 T_2 は一般の正則行列であるので、最終的に得られる変換行列 $T_3 = T_1 T_2$ は、斜交回転となる。

⁵よく考えると、斜交回転という呼び方はあまり適切でない。一般的に回転とは図形の直交する角を保つ変換であるから、直交回転以外の変換を「回転」と呼ぶのは誤解を招きやすい。

37.5 あてはまりの評価

因子分析モデルのデータへのあてはまりを評価する場合、つぎの2つの側面がある。

1. 各変数の分散のうち共通因子がどれだけを説明しているか。
2. データから得られた分散行列 S が因子分析モデル $\Sigma = \Lambda\Lambda' + \Psi$ にどの程度あてはまっているか。

1番目の基準は、2番目の基準がよくあてはまっていたとしても、大きいとは限らない。実際、 Λ の要素にくらべて Ψ の対角成分が大きい場合でも、正確に $S = \Sigma$ が成立することはありうる。

37.5.1 共通性

最初の基準を測る指標としては、共通性 (communality) とよばれるものがある。変数 X_j についての共通性はつぎの式で与えられる。

$$h_j^2 = \frac{\text{Var}(\sum_{k=1}^m \lambda_{jk} f_k)}{\text{Var}(X_j)} = \frac{\sigma_{jj} - \psi_j}{\sigma_{jj}}, \quad (j = 1, \dots, p) \quad (278)$$

これは X_j の分散のうち、共通因子によって説明される部分の比率をあらわす。各 f_j の分散が1で互いに統計的に独立であることを考慮すると、 $\text{Var}(\sum_{k=1}^m \lambda_{jk} f_k) = \sum_{k=1}^m \lambda_{jk}^2$ である。また、多くの推定法においては $s_{jj} = \hat{\sigma}_{jj}$ が成立するので、(278)は $\sum_{k=1}^m \lambda_{jk}^2 / s_{jj}$ とかける。文献によっては、比率ではなく $\text{Var}(\sum_{k=1}^m \lambda_{jk} f_k)$ そのものを共通性と呼ぶ場合もある。

厳密に考えれば、独自因子の分散 ψ_j には、その変数の特殊性による成分 (specificity) と、単なる測定誤差の両者が含まれている。測定誤差の大きさは繰り返しをとまなう測定などによって、因子分析とは別の手続きで評価しなければ分からない。

Harman(1976)では、観測変数の分散を1に基準化し、それをつぎの3つの部分に分割している。

$$\text{Var}(X_j) = 1 = h_j^2 + b_j^2 + e_j^2 \quad (279)$$

ここで、 h_j^2 が共通性、 b_j^2 が特殊性であり、 e_j^2 が測定誤差である。また、 $b_j^2 + e_j^2$ が独自性 (uniqueness) であり、 $1 - e_j^2$ を信頼性 (reliability) と呼んでいる。

37.5.2 モデル適合度

ウィシャート一般化最小2乗基準 (WGLS) とウィッシャート最尤基準 (WML) については、適合度指標 F_{WGLS} と F_{WML} の漸近的な性質を用いて検定を行なえる。これらの値の $(N-1)$ 倍は、つぎの基準が成立しているならば、標本数 N が大きくなるときいずれも漸近的に χ^2 分布に従う。

1. データは多変量正規分布に従う。
2. 分布の真の共分散行列は $\Sigma_0 = \Lambda_0\Lambda_0' + \Psi$ である。

このとき、 $(N-1)F_{WGLS}$, $(N-1)F_{WML}$ のいずれもつぎの自由度の χ^2 分布に漸近的に従う。

$$p(p+1)/2 - (\text{モデルの自由パラメータ数}) = p(p+1)/2 - \{pm + p - m(m-1)/2\} = (p-m)^2/2 - (p+m)/2 \quad (280)$$

この性質を用いて、モデルが適切か否かの検討を行なえる。つまり値が χ^2 分布について、小さな上側確率を持たば、モデルは適合していないことになる。

ここで、 χ^2 適合度検定の特徴として注意すべきことは、サンプル数 N が大きくなると、些細な Σ と S の違いが、次第に厳しく判定されるようになることである。現実のデータにおいては、厳密にモデルが成り立っていることはないので、サンプル数が極めて大きい場合の χ^2 検定の p 値は、このような傾向を持つものであることを、あらかじめ知っておく必要がある。

また、つぎのような適合度指標が最尤法の場合には、用いられる。

- AIC (Akaike's Information Criterion, AIC), (Akaike, 1973; 赤池, 1976):
 $-2 \times \text{最大対数尤度} + 2 \times (\text{自由パラメータ数})$
- BIC (Schwarz's Bayesian Information Criterion, BIC), (Schwarz, 1978):
 $-2 \times \text{最大対数尤度} + \log N \times (\text{自由パラメータ数})$

これらは最大対数尤度が大きくなると減少し、モデルパラメータ数が大きくなると増加する。AIC は赤池弘次によって、時系列モデルの次数選択のために開発された。また、BIC は AIC への批判としてベイズ事後確率の評価に基づいて Schwarz(1978) によって提案された。これらの指標を最小とするモデルが、良いモデルとされる。ここで、

$$-2 \times (\text{最大対数尤度}) = (N - 1)F_{WML} + \text{定数}$$

である。定数部分はモデルを変更しても分析対象となる変数が変わらなければ影響を受けないので、因子数の異なるモデルの良さを比較する場合には考慮する必要がない。

また、Jöreskog & Sörbom (1986) はつぎのように適合度指標 (Goodness of Fit Index, GFI) と、修正適合度指標 (Adjusted GFI, AGFI) を最尤法について定めた。

$$GFI_{WML} = 1 - \frac{\text{trace}\{(\hat{\Sigma}^{-1}S - I)^2\}}{\text{trace}\{(\hat{\Sigma}^{-1}S)^2\}} \quad (281)$$

$$AGFI_{WML} = 1 - \frac{p(p+1)}{2df} \times (1 - GFI_{WML}) \quad (282)$$

また、彼らは重みなし最小 2 乗法についても、つぎの基準を定義した。

$$GFI_{ULS} = 1 - \frac{\text{trace}\{(S - \hat{\Sigma})^2\}}{\text{trace}(S^2)} \quad (283)$$

$$AGFI_{ULS} = 1 - \frac{p(p+1)}{2df} \times (1 - GFI_{ULS}) \quad (284)$$

さらに Tanaka & Huba (1985) は一般化最小 2 乗法について、つぎの指標を提案した。

$$GFI_{WGLS} = 1 - \frac{\text{trace}\{(I - \hat{\Sigma}S^{-1})^2\}}{p} \quad (285)$$

$$AGFI_{WGLS} = 1 - \frac{p(p+1)}{2df} \times (1 - GFI_{WGLS}) \quad (286)$$

ここで df とは、推定の自由度 (つまり $p(p+1)/2 - \text{推定パラメータ数}$) のことである。Bollen(1989) の記述ではこれらは、おのおの GFI_{ML}, GFI_{GLS} などとなっている。

正規最尤基準は $F_{NML} = (N - 1)F_{WML}/N$ と定義されるが、ウィットシャート最尤基準と区別しての記述は、Browne & Arminger (1995) による。

38 因子得点の推定

因子分析の応用で最も利用されるのは因子負荷量 Λ のパターンであるが、因子得点 (f または回転後の g) もデータの解釈を行なうために有益である。主成分分析の場合と異なり、因子分析において因子得点は間接的に定義されている。モデルによって指定されているのは、 f の分布のみであり、特定のサンプルに対応する因子得点 f_i は、何らかの方法で推定しなければならない。このためのいくつかの推定法が提案されている。パートレット (Bartlett) 推定量、トムソン (Thomson) 推定量などと呼ばれるものがある。

被験者 i の因子スコアを f_i とし、これが未知の固定されたパラメータであるとみなす。因子分析のモデル

$$\mathbf{X}_i = \Lambda \mathbf{f}_i + \boldsymbol{\varepsilon}_i \quad (287)$$

を考えると、因子スコア f_i を推定する問題は、 $\boldsymbol{\varepsilon}$ の共分散行列 Ψ と Λ とが既知であるとすると、 Λ をデザイン行列とし、 f_i を未知変数とする重回帰分析の問題、つまり最小 2 乗法とみなせる。この場合、注意すべきは $\boldsymbol{\varepsilon}$ の各要素が、独立同分散ではないことである。上の条件を考慮すると、 f_i の最小 2 乗推定量は、観測値が \mathbf{x}_i であるとき

$$\hat{\mathbf{f}}_i = (\Lambda^T \Psi^{-1} \Lambda)^{-1} \Lambda^T \Psi^{-1} \mathbf{x}_i \quad (288)$$

で与えられる。この式で推定される因子スコアをパートレット推定量という。このとき、つぎの式が成立する。ここで、 $\Gamma = \Lambda^T \Psi^{-1} \Lambda$ とである。

$$E[(\hat{\mathbf{f}}_i - \mathbf{f}_i)(\hat{\mathbf{f}}_i - \mathbf{f}_i)^T] = (\Lambda^T \Psi^{-1} \Lambda)^{-1} = \Gamma^{-1} \quad (289)$$

一方、もし f_i を確率変数と考え、さらに \mathbf{X}_i との同時分布を考えると、つぎの式が成り立つ。ここで、 f_i と $\boldsymbol{\varepsilon}_i$ とは、独立であることに注意する。

$$\text{Var}(\mathbf{X}_i) = \Lambda^T \Phi \Lambda + \Psi \quad (290)$$

$$\text{Cov}(\mathbf{X}_i, \mathbf{f}_i) = \Lambda \Phi \quad (291)$$

$$\text{Var}(\mathbf{f}_i) = \Phi \quad (292)$$

$$(293)$$

ここで、 f_i と \mathbf{X}_i の同時分布が、多変量正規分布であると仮定すると、

$$E(\mathbf{f}_i | \mathbf{X}_i = \mathbf{x}_i) = \Phi \Lambda^T (\Psi + \Lambda \Phi \Lambda^T)^{-1} \mathbf{x}_i \quad (294)$$

$$= \Phi (\Phi + \Phi \Gamma \Phi)^{-1} \Phi \Lambda^T \Psi^{-1} \mathbf{x}_i \quad (295)$$

となる。この値を f_i の推定値と考え $\hat{\mathbf{f}}_i^*$ と表記する。上式は $\Phi = I$ を仮定すると、

$$\hat{\mathbf{f}}_i^* = (I + \Gamma)^{-1} \Lambda^T \Psi^{-1} \mathbf{x}_i \quad (296)$$

となる。 f_i と $\hat{\mathbf{f}}_i^*$ の同時分布を考えると、つぎの式が成立する。

$$E[(\hat{\mathbf{f}}_i^* - \mathbf{f}_i)(\hat{\mathbf{f}}_i^* - \mathbf{f}_i)^T] = (I + \Gamma)^{-1} \quad (297)$$

この値は、(289) よりも小さい。この推定量はトムソン (Thomson, 1951) による (Anderson, 1986, Sec. 14.7)。

39 実行例

ディレクトリ /home1/otsu/cl14a0/sas のなかのファイル test24.cor と test24.means にデータがある。説明は、test24.doc にまた長い変数名は test24.row にある。

以下はこれらのファイルを入力して因子分析を行なう例である。このプログラムは上記ディレクトリの test24.sas というファイルにある。下の例では相関行列と平均・標準偏差を別ファイルから入力し、分析用のデータセットをつくっているので操作が複雑になっているが、通常の原因データを入力し、proc factor で変数を指定することもできる。

```
data corr1 (type=corr);
  infile 'test24.cor';
  _type_='CORR';
  input _name_ $ X1-X24;

data mestd;
  infile 'test24.means';
  N = 147;
  input _name_ $ MEAN STD Reli;
;

proc transpose data=mestd
  out = mestd1 ;
  var n mean std;

data mestd2;
  set mestd1;
  _type_ = _name_ ;
  _name_ = '      ' ;

run;
data corr2 (type=corr); /* データステップが複雑になっているのは */
  set mestd2 corr1; /* 相関行列と平均、標準偏差データを */
/* 結合して入力データを作成しているため */
proc print; /* どのようなデータがつけられているか確認 */

proc factor data=corr2
  method=ml /* 最尤法の指定 */
  rotate=promax ; /* プロマックス回転の指定 */
run;
```

因子分析についての本、または関係することがらについて記述されている本をあげる。

参考文献

- Akaike,H. (1973). Information theory and an extersion of the maximum likelihood principle, in Prtrov & Csaki eds. 267-281.
- Anderson,T.W. (1984), *An Introduction to Multivariate Statistical Analysis, 2nd ed.*, Wiley.
- 赤池弘次 (1976). 情報量基準 AIC とは何か その意味と将来への展望, 数理科学, 153 (1976-3), 5-11.
- Arminger,G. Clogg,C.C., & Sobel,M. (1995). *Handbook of Statistical Modeling for the Social and Behavioral Sciences*, New York: Plenum Press.
- Bollen,K. A.(1989). *Structural Equations with Latent Variables*, Wiley.
- Browne,M.W. & Arminger,G. (1995), Specification and estimation of mean-and covariance-structure models. in Arminger et al. (1995) 185-249.
- Harman,H.H. (1976) , *Modern Factor Analysis, 3rd ed. revised*, Chicago: Univ. Chicago Press.
- Johnston,J. (1984), *Econometric Methods, 3rd ed.*, McGraw Hill.
- Jöreskog,K.G., & Sörbom (1986) *LISREL VI: Analysis of Linear Structural Relationships by Maximum Likelihood and Least Square Methods*. Mooresville, IN: Scientific Software,Inc. (現在では LISREL8 に版が上がっている。)

- Mardia, K.V., Kent, J.T. & Bibby, J.M. (1979), *Multivariate Analysis*, London: Academic Press.
- Schwarz, G. (1978), Estimating the dimension of a model, *The Annals of Statistics*, 6, 461-464.
- Tanaka, J.S. & Huba, G.J. (1985). A fit index for covariance structure models under arbitrary GLS estimation, *British Journal of Mathematical and Statistical Psychology*, 38, 197-201.
- 柳井晴夫、繁榎算男、前川眞一、市川雅教 (1990) 因子分析 - その理論と方法 -, 東京: 朝倉書店

40 PROC CALIS による共分散構造分析

共分散構造分析は、計量経済学で発展した構造方程式モデルに、直接には観測されない潜在変数 (因子分析における因子) を表現する機能を加えた方法である。共分散構造分析の利点は、分析者が想定する因子構造についての仮説を、統計的検定を用いて検討できることにある。

通常、次ぎの図に示すように観測可能な変数は四角で表現し、潜在変数は円または楕円で表現する。残差変数は囲まずに表現する。また、矢印は影響の因果関係を示す。両端に頭のある矢印は、相互的な関連 (相関関係) があることを示す。

共分散構造分析では、影響関係は線形のモデルに限定しているため、グラフは部分的な重回帰式の関係を組み上げたものと見做される。通常は、矢印の向きをたどっても、もとの変数に戻ることはないパターンにモデルを限定することが多い。このようなループ構造の存在しないモデルのことを逐次モデル (recursive model) と呼ぶ。

一般的には、変数間の条件付き独立 (偏相関がゼロ) の性質と、グラフの構造の関係は複雑であるが、すべての矢印が一方であり、しかもループが存在しない場合には (Directed Acyclic Graph, DAG と呼ばれる)、つぎの2つの性質は同値である。

1. 変数 X と変数 Y の間に直接の矢印が存在しない。
2. X と Y 以外からなる変数の集合 Z_1, Z_2, \dots, Z_k が存在し ($k = 0$ の場合も含む)、これらを固定すると X と Y とがつねに条件付き独立となる。

これより「共分散行列の逆行列をとってその (i, j) 要素がゼロならば、 X_i と X_j の間に直接の矢印は (偶然による場合を除いては) ない」という性質がなりたつ。しかし、逆は必ずしもなりたたない。つまり、矢印がないにも係らず逆行列の (i, j) 要素がゼロではない場合はありうる。つまり、全体の変数の共分散行列の逆において要素がゼロとならなくても、一部の変数間の共分散行列の逆において、要素がゼロとなることはあり得る。例えば、 X, Y, Z の3変数において、矢印が $X \rightarrow Y$ 、 $Z \rightarrow Y$ の2つであるとする。 X と Z の間に直接の矢印は存在しない。このとき σ_{XZ} はゼロ (固定する変数が空のとき) であるが、 Y を固定して得られる $\sigma_{XZ.Y}$ は一般的にはゼロにはならない。

図 27 のモデルは、Lord(1957) によるもので SAS のマニュアルから引用した。Lord のデータは語彙テストの成績についてのものである。 W と X は各 15 個の項目からなり、時間に厳しい制約は課していない。一方 Y と Z は各 75 項目からなるテストであって、こちらは厳しい時間制約のもとで回答させたものである。このモデルは、斜交回転を伴う2因子の因子分析モデルにおいて、いくつかの因子負荷量をゼロに制約したものと見做せる。

このモデルの分析プログラムの例を次ぎに示す。

```
option linesize=80;
```

```
data lord(type=cov);      /* データが共分散であることを指定している */
```

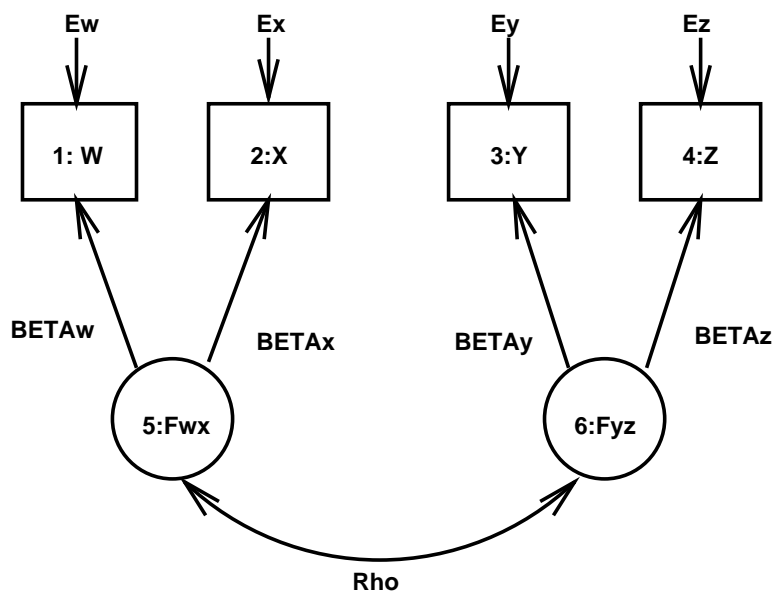


図 27: Lord(1957) フルモデル

```

input _type_ $ _name_ $ w x y z; /* _type_ に指定するものは既定 */
cards;
n . 649 . . . /* n: データ件数 */
cov w 86.3979 . . . /* cov: 共分散行列 */
cov x 57.7751 86.2632 . .
cov y 56.8651 59.3177 97.2850 .
cov z 58.8986 59.6683 73.8201 97.8192
;

title 'H4: unconstrained';
proc calis data=lord cov outram=ram4; /* 入力データタイプの出力指定 */
/* ram4 は出力結果を保存するファイルの名称 */
lineqs w = betaw fwx + ew, /* 構造方程式を記述する方式の一つ EQS 風 */
x = betax fwx + ex,
y = betay fyz + ey,
z = betaz fyz + ez;
std fwx fyz = 2*1, /* 標準偏差を指定 2*1 は1が2個 */
ew ex ey ez = vew vex vey vez; /* 残差の分散は推定する */
cov fwx fyz = rho; /* 因子間の共分散に名前を付ける */

proc print data=ram4; /* 計算結果の情報がすべて含まれる */
run;

```

上のプログラムは図の構造をそのまま表現したものであるが、さらにモデルに制約を加えて推定

を行なうことも可能である。つぎは、潜在変数が1個のモデルを指定したものである。

```
title 'H3: congeneric';
proc calis data=lord cov outram=ram3;
lineqs w = betaw f + ew,
        x = betax f + ex,
        y = betay f + ey,
        z = betaz f + ez;
std f = 1,
    ew ex ey ez = vew vex vey vez;
run;
```

また、つぎの例は2個の潜在変数が、各2個の観測変数に同じ係数で影響を及ぼしているとの仮説を表すものである。

```
title 'H2: Parallel';
proc calis data=lord cov outram=ram2;
lineqs w = betawx fwx + ew,
        x = betawx fwx + ex,
        y = betayz fyz + ey,
        z = betayz fyz + ez;
std fwx fyz = 2*1,
    ew ex ey ez = vew vex vey vez;
cov fwx fyz = rho;
run;
```

各々の例を実行し、どのモデルの当てはまりが良いか検討せよ。Akaike Information Criterion (AIC), Schwarz's Bayesian Criterion (BIC) などの指標を利用せよ。これらの値が小さいモデルが、データによく当てはまり、かつ簡潔なモデルである。AICの方が複雑なパラメータ数の多いモデルを選ぶ傾向がある。

CALIS プロシジャの詳細については豊田(1992)、SASのマニュアルなどを参照すること。

参考文献

Lord, F.M. (1957), A significance test for the hypothesis that two variables measure the same trait except for errors of measurement, *Psychometrika*, **22**, 207-220.

SAS Institute, (1990), *SAS/STAT User's Guide, 4th ed. Vol.1*, Cary, NC, USA: SAS Institute.

SAS Institute, (1996), *SAS/STAT Software: Changes and Enhancements*, Cary, NC, USA: SAS Institute.

豊田秀樹 (1992), SASによる共分散構造分析、東京大学出版会

41 多変量データにおける検定

前期にランダム(混合型)モデルについて紹介したが、その続きを少し説明する。余裕があれば、利用方法まで説明する予定であったが、簡単な説明にとどめる。

41.1 Hotelling の T^2

1 変数 1 標本の t 検定で用いられる t 値を 2 乗するとつぎのようになる。

$$\frac{(\bar{x}_1)^2}{(1/N)\hat{\sigma}^2} \quad (298)$$

この式を多変数への拡張は、つぎのように行う。まず \mathbf{X}_i ($i = 1, \dots, N$) を $N(\mathbf{0}, \Sigma)$ に独立に従う p 変量のベクトルとする。これらの平均は

$$\mathbf{d} = \frac{1}{N} \sum_{i=1}^N \mathbf{X}_i \quad (299)$$

であり、また分散共分散の不偏推定値は

$$\hat{\Sigma} = \frac{1}{N-1} \sum_{i=1}^N (\mathbf{X}_i - \mathbf{d})(\mathbf{X}_i - \mathbf{d})^T \quad (300)$$

で与えられ $\frac{1}{N-1}W(\Sigma, N-1)$ に従う。また \mathbf{d} と $\hat{\Sigma}$ とは、独立に分布することも知られている。 \mathbf{d} の分散共分散は $\frac{1}{N}\hat{\Sigma}$ で推定されるので、これらを用いて

$$T^2 = N\mathbf{d}^T \hat{\Sigma}^{-1} \mathbf{d} \quad (301)$$

という値を考えると、これが t 統計量の 2 乗の多変数への拡張とみなせる。これを Hotelling の T^2 (より正確には $T^2(p, N-1)$) と呼ぶ。2 群のデータが同一の多変量正規分布に従うという仮定のもとで、つぎの式が成り立つことが知られている。

$$T^2(p, m) \sim \frac{mp}{m-p+1} F(p, m-p+1) \quad (302)$$

上の例では、 $m = N-1$ であるので、右辺は

$$\frac{(N-1)p}{N-p} F(p, N-p)$$

となる。

つぎに、2 群の多変量データの平均ベクトルに差があるか否かを検定することを考えよう。

2 群の p 次元確率変数を \mathbf{X}_{i1} , ($i = 1, \dots, N_1$) および \mathbf{X}_{i2} , ($i = 1, \dots, N_2$) とする。2 群の p 次元ベクトルはおのおの $N(\boldsymbol{\mu}_1, \Sigma)$ 、および $N(\boldsymbol{\mu}_2, \Sigma)$ の多変量正規分布に独立に従うものとする。標準的な t 検定の場合と同様に、2 群の分散 (共分散行列) は同一であるとしている。ここで、 $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ および Σ の推定値はおのおの、次の式によって与えられる。

$$\hat{\boldsymbol{\mu}}_1 = \frac{1}{N_1} \sum_{i=1}^{N_1} \mathbf{X}_{i1} \quad (303)$$

$$\hat{\boldsymbol{\mu}}_2 = \frac{1}{N_2} \sum_{i=1}^{N_2} \mathbf{X}_{i2} \quad (304)$$

$$\hat{\Sigma} = \frac{1}{N_1 + N_2 - 2} \left\{ \sum_{i=1}^{N_1} (\mathbf{X}_{i1} - \hat{\boldsymbol{\mu}}_1)(\mathbf{X}_{i1} - \hat{\boldsymbol{\mu}}_1)^T + \sum_{i=1}^{N_2} (\mathbf{X}_{i2} - \hat{\boldsymbol{\mu}}_2)(\mathbf{X}_{i2} - \hat{\boldsymbol{\mu}}_2)^T \right\} \quad (305)$$

前の場合と同様に、 $\mathbf{d} = \hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2$ とおく。 \mathbf{d} の分散共分散行列の推定値は $(1/N_1 + 1/N_2)\hat{\Sigma}$ となる。これを用いてつぎのような値を求めると、 $T^2(p, N_1 + N_2 - 2)$ となることがわかる。

$$T^2 = \frac{N_1 N_2}{N_1 + N_2} \mathbf{d}^T \hat{\Sigma}^{-1} \mathbf{d} \quad (306)$$

従って、帰無仮説 $\mu_1 - \mu_2 = 0$ が成り立っているならば、

$$T^2 \sim \frac{(N_1 + N_2 - 2)p}{N_1 + N_2 - p - 1} F(p, N_1 + N_2 - p - 1) \quad (307)$$

となるので、この性質を使って多変量の平均の差の検定を行える。

41.2 多変量分散分析 (MANOVA) と混合型モデル

分散分析は、説明変数には多くの要因を仮定するが、被説明変数は1変量のモデルである。もし、被説明変数が多変量である場合には、モデルを拡張しなければならない。この様な必要は、被説明変数を独立な確率変数とみなすことができない場合に生じる。例えば、ある生徒の集団を幾つかのグループに分け、各々に異なる教育法を用いたとしよう。教育の成果を測るために、時期を変えて3回試験を行なうものとする。生徒 i の各回の成績を Y_{i1}, Y_{i2}, Y_{i3} と記述する。通常の分散分析では、これらは独立な確率変数と仮定するが、この例においては仮定が成立すると考えるのは難しい。1回目の成績がよい生徒は2回目も良い傾向があると考えるのが当然であるし、2回目の成績は3回目の成績と関係していると考えるのが普通であろう。生徒の個人の特性を説明要因としてモデルに取り込むことが、一つの方法として考えられるが(混合モデルの説明はこのような形をとる)、もう一つの対応策は Y_{i1}, Y_{i2}, Y_{i3} の間に相関が存在することを仮定することである。後者のモデルは、説明変数は分散分析と同じであるが、被説明変数は3次元のベクトルとみなせる。同様の問題は、医学分野では同一の患者についての経時的な測定を扱う場合に生じる。このような分散分析(または共分散分析)の拡張モデルを用いる分析法を多変量分散分析 (Multivariate ANalysis Of VAriance, MANOVA) と呼ぶ。MANOVA モデルを利用した分析は SAS の MIXED プロシジャを用いて分析することができる。

前期の説明と重複するが、以下に混合モデルの説明を再掲する。

MIXED(混合) という名がついている理由は、つぎのようなものである。MANOVA モデルはつぎのように記述することができる。

$$Y_{ij} = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \varepsilon_{ij}, (i = 1, \dots, n; j = 1, \dots, q) \quad (308)$$

ここで、 $\varepsilon_{ij}, (j = 1, \dots, q)$ は独立ではなく、相関があることを仮定している。ここで、もしこの相関行列がつぎのような1因子の因子分析モデルであらわされるものとしよう。

$$\begin{pmatrix} \sigma^2 & \rho & \cdots & \rho \\ \rho & \sigma^2 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & \sigma^2 \end{pmatrix} \quad (309)$$

つまり、対角要素は σ^2 であり、非対角要素は全て ρ である。このような相関構造は、個人毎の変動を表す確率変数 θ_i を導入することによって、つぎのように表現できる。

$$Y_{ij} = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \theta_i + \eta_{ij}, (i = 1, \dots, n; j = 1, \dots, q) \quad (310)$$

ここで、 $\eta_{ij}, (i = 1, \dots, n; j = 1, \dots, q)$ は全て独立同一分散である。このモデルは、説明変数に固定された値を持つもの X_j と、直接には観測されない確率変数 θ が混在するモデルとみなせる。このようなモデルを混合モデル (mixed model) と呼ぶ。

MANOVA は共分散分析の拡張であるが、同様の拡張をロジスティック回帰や対数線形モデルなどの一般化線形モデルについて行なう必要も生じる。このための分析法は SAS の GENMOD プロシジャの REPEATED ステートメントによって指定できる。